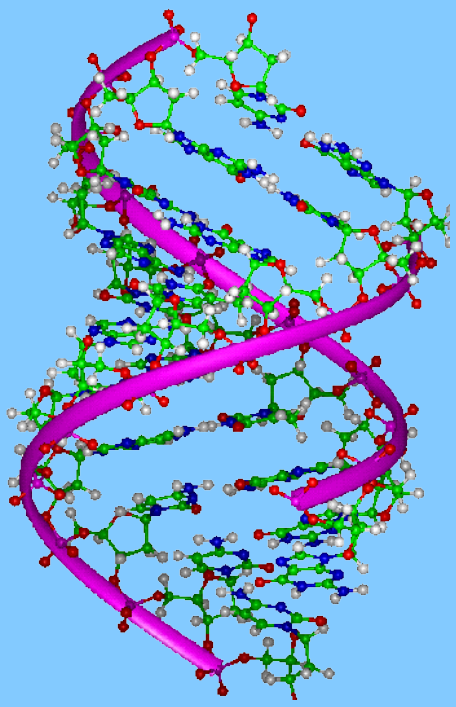
Basic Bioinformatics

A Practical User Introduction

Tuesday 29 March 2016



This material is intended to be used for self instruction or in a more formal class setting. When offered as a set of class exercises, I describe my intentions as follows:

Overview

This is an **entry level** course aimed that those with a reasonable biological background but **no significant experience with bioinformatics**. The course is broadly based around a series of exercises in which a combination of simple analytical tools and publicly available databases is applied to the investigation of a single human gene. The training manual for the course is comprised of detailed instruction for the tasks undertaken. Included are, questions (with answers) and discussion of and the interpretation of the results achieved. The manual is under constant development, a recent version is available online.

Outline

Participants are asked to imagine an interest in the disease **aniridia**. Course exercises then provide extremely detailed instruction leading participants to discover the gene primarily associated with this disease and all that is interesting about that gene and its protein products.

Objectives

This course is not intended to fully meet the requirements of many modern research projects. The intention is more to examine and practice the basic tools that underpin many current bioinformatic solutions and have changed little over the years. Specifically, we set out to show how many answers can be simply be looked up in an appropriate information source and/or how it is also usually a trivial matter to compute the similar/identical answers using readily available software.

We set out to explain the operation of the various programs used in the exercise, but only to the extent that allows a user to select parameters intelligently and to interpret results fully.

Timetable

The main topics covered, in order, are as follows:

Simple Information and Data (i.e. sequence) retrieval from public resources (primarily the **NCBI** and **EBI**).

Genome Databases (primarily **Ensembl** including access using **Biomart**)

Pairwise Sequence Alignment (**dotplots**, **global** and **local** alignments)

Database searching (using elements of the **blast** family)

Primer design (primarily using the **NCBI** services)

Secondary Structure Prediction (**GOR**, **JPred**)

Multiple Sequence Alignment (clustal, muscle, t-coffee …)

PSI-blast (an explicit example, but incorporated into a number of the previous topics)

3D Structure (a very superficial look specific to the exercise protein)

The timing can vary depending on circumstances. A very rushed **3 days** to a relatively leisurely **5 days**.

DPJ – 2016.03.03

Course Objectives

The aim is to present a hands-on introduction to the ways computer resources can be utilised to assist the molecular biological researcher. Our target audience is the informed biologist who is new to using the computer for such a purpose. We cannot cover "everything" but hope to introduce a range of computing facilities available to answer simple molecular biological enquiries. Where appropriate, how programs work will be discussed, but only as far as is required to understand the parameters that a user is required to set and the results generated. We will assume that a brief introductory talk mentioning the more important databases has been delivered before offering these exercises.

There will be a mixture of talks and frequent practical sessions. During the practical sessions you are encouraged to go through the exercises in the book, as far as is practical, at your own pace. Please ask for assistance at any time you are unsure how to proceed or require clarification on any point.

At various points in the exercises, you will be asked to answer simple questions. The purpose of the questions is to draw your attention to some aspect of what you are doing rather than to test your understanding. Sometimes, the answers you note down will be useful later in the exercise, so try not to skip over the questions. If you cannot answer any question, please just ask[1].

It is now less usual than previously for individual researchers to meticulously analyse small volumes of data. Sequence data is typically generated and analysed by large, well equipped, Institutes. Whole genomes are processed in a single project. Sequence data, analytical results and associated, often automated, interpretation are centrally organised for easy access. In consequence, many investigations can be completed by simply locating answers stored in information rich databases. Accordingly, our exercises will be roughly split into two sections. In the first we will take a topic and see how it is possible to discover effectively all one would wish to know from various combinations of appropriate information resources over the internet. In the second section we will look at how many of the same conclusions might have been reach by applying analytical software to raw sequence data. The software tools we will use are essentially those used to generate the pre-processed "answers" achieved in the first section so it should not be surprising that the results will be very similar.

Of course, we will cheat! Our topic will be a well researched one to ensure that there is a full set of pre-processed answers to find in the databases we interrogate. But the principles we will use are universal. As completely as we were able, we have built all the exercises around a single starting point. That is, a need to find out all there is to know about the human disease of the eye called **aniridia**.

If your research does not involve humans, please do not be concerned. Using a human example ensures a well researched topic, but very few of the resources we will visit are specifically human. The basic principles of how to proceed really are not organism specific.

DPJ/PDFJ – 2016.03.03

1 If you are working from the **PDF** version of these notes, some "model answers" can be accessed via the link that is behind the **?** or **.** that terminates each question. This is "work in progress", I am far from completion at the time of typing.

Searching for ready made answers

To recap, our objective is to discover what we can about the disease **aniridia** and its causes. There are many ways we could begin, including the ultimately lazy (but often very effective) universal strategy of "google". In an attempt to be slightly more directed, and as our example query involves a human disease, we will start with **GeneCards**.

GeneCards

If we assume the disease might be genetically linked, a sensible starting service for our investigations might be **GeneCards**, a famous service maintained in Israel by the **Weizmann Institute**. **Genecards** is:

> " ... a searchable, integrated, database of human genes that provides concise genomic, transcriptomic, genetic, proteomic, functional and disease related information on all known and predicted human genes"

Information from many sources is stored for each human[2] gene in a "card". These cards can be searched using clues (**Keywords**) specifying your interests to determine the gene(s) that might be of interest. In addition to summarising what is known about a gene, each **GeneCard** offers links to pertinent internet resources for further investigation. Effectively, **GeneCards** has searched tens of databases for each gene and saved the results as links. Thus, one search of **GeneCards** can substitute for many individual searches of other resource sites.

Hopefully you have a browser window open at this stage, if not, open one up and go to:

http://www.genecards.org

Leave the type of search set to set to Keywords, enter the *Search Term* **aniridia** and click the button. You will be rewarded with a list of links to **GeneCards** in order of "relevance"[3].

		Symbol	Description	Category	GIFtS	GC Id	Score
1	⊕	PAX6	Paired Box 6	Protein Coding	61	GC11M031806	59.98
2	⊕	WT1	Wilms Tumor 1	Protein Coding	66	GC11M032365	18.21
3	⊕	ELP4	Elongator Acetyltransferase Complex Subunit 4	Protein Coding	49	GC11P031487	15.40
4	⊕	DEL11P13	Wilms Tumor, Aniridia, Genitourinary Anomalies And Mental Retardation Syndrome	Uncategorized	4	GC11U901781	14.72
5	⊕	LUZP2	Leucine Zipper Protein 2	Protein Coding	47	GC11P024518	14.67
6	⊕	FOXC1	Forkhead Box C1	Protein Coding	56	GC06P001610	14.46
7	⊕	PITX2	Paired-Like Homeodomain 2	Protein Coding	57	GC04M110617	13.01
8	⊕	FOXE3	Forkhead Box E3	Protein Coding	43	GC01P047416	13.01
9	⊕	CYP1B1	Cytochrome P450, Family 1, Subfamily B, Polypeptide 1	Protein Coding	63	GC02M038034	12.94
10	⊕	LGR4	Leucine-Rich Repeat Containing G Protein-Coupled Receptor 4	Protein Coding	52	GC11M027345	12.91
11	⊕	KRT12	Keratin 12, Type I	Protein Coding	49	GC17M040861	12.33
12	⊕	OTX2	Orthodenticle Homeobox 2	Protein Coding	58	GC14M056799	12.33
13	⊕	PAX4	Paired Box 4	Protein Coding	55	GC07M127610	12.33
14	⊕	IGF2	Insulin-Like Growth Factor 2	Protein Coding	63	GC11M002130	11.80

The score for each **GeneCard** suggests how close a match has been found between your search term(s) and the gene. The top hit should be a clear "winner".

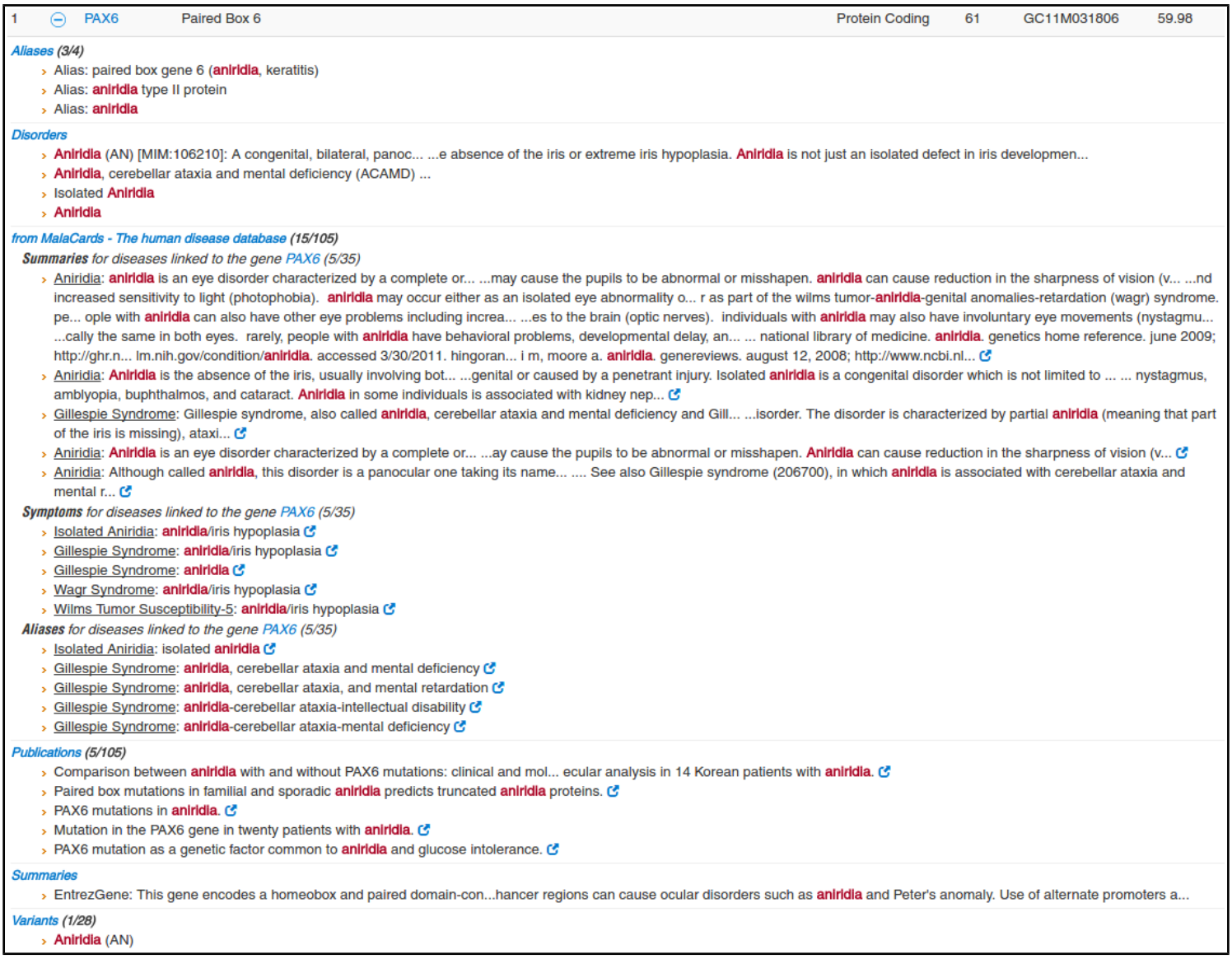
Words matching search terms are marked in red.

What do you conclude to be the gene most relevant to **aniridia**? _____

2 There is also a **mouse genecards**. Maybe more to follow?

3 A computer's idea of "relevance" may not match yours. Here, the **Score** used to rank hits takes account of the frequency of the search term(s) weighted by where terms occur (e.g. in a "**Title**" would imply more importance than in a "**Comment**" field, perhaps?) and the specificity of the term (e.g. "**the**" is not specific, "**aniridia**" is). Only if you are really fascinated by this sort of thing try:
`https://www.elastic.co/guide/en/elasticsearch/guide/current/scoring-theory.html`

To confirm the top hit as the best match, click a few of the ⊕ buttons by the gene names for more information.



1 ⊖ PAX6 Paired Box 6 Protein Coding 61 GC11M031806 59.98

Aliases (3/4)
- Alias: paired box gene 6 (**aniridia**, keratitis)
- Alias: **aniridia** type II protein
- Alias: **aniridia**

Disorders
- **Aniridia** (AN) [MIM:106210]: A congenital, bilateral, panoc... ...e absence of the iris or extreme iris hypoplasia. **Aniridia** is not just an isolated defect in iris developmen...
- **Aniridia**, cerebellar ataxia and mental deficiency (ACAMD) ...
- Isolated **Aniridia**
- **Aniridia**

from MalaCards - The human disease database (15/105)

***Summaries** for diseases linked to the gene PAX6 (5/35)*
- Aniridia: **aniridia** is an eye disorder characterized by a complete or... ...may cause the pupils to be abnormal or misshapen. **aniridia** can cause reduction in the sharpness of vision (v... ...nd increased sensitivity to light (photophobia). **aniridia** may occur either as an isolated eye abnormality o... r as part of the wilms tumor-**aniridia**-genital anomalies-retardation (wagr) syndrome. pe... ople with **aniridia** can also have other eye problems including increa... ...es to the brain (optic nerves). individuals with **aniridia** may also have involuntary eye movements (nystagmu... ...cally the same in both eyes. rarely, people with **aniridia** have behavioral problems, developmental delay, an... ... national library of medicine. **aniridia**. genetics home reference. june 2009; http://ghr.n... lm.nih.gov/condition/**aniridia**. accessed 3/30/2011. hingoran... i m, moore a. **aniridia**. genereviews. august 12, 2008; http://www.ncbi.nl...
- Aniridia: **Aniridia** is the absence of the iris, usually involving bot... ...genital or caused by a penetrant injury. Isolated **aniridia** is a congenital disorder which is not limited to ... ... nystagmus, amblyopia, buphthalmos, and cataract. **Aniridia** in some individuals is associated with kidney nep...
- Gillespie Syndrome: Gillespie syndrome, also called **aniridia**, cerebellar ataxia and mental deficiency and Gill... ...isorder. The disorder is characterized by partial **aniridia** (meaning that part of the iris is missing), ataxi...
- Aniridia: **Aniridia** is an eye disorder characterized by a complete or... ...ay cause the pupils to be abnormal or misshapen. **Aniridia** can cause reduction in the sharpness of vision (v...
- Aniridia: Although called **aniridia**, this disorder is a panocular one taking its name... .... See also Gillespie syndrome (206700), in which **aniridia** is associated with cerebellar ataxia and mental r...

***Symptoms** for diseases linked to the gene PAX6 (5/35)*
- Isolated Aniridia: **aniridia**/iris hypoplasia
- Gillespie Syndrome: **aniridia**/iris hypoplasia
- Gillespie Syndrome: **aniridia**
- Wagr Syndrome: **aniridia**/iris hypoplasia
- Wilms Tumor Susceptibility-5: **aniridia**/iris hypoplasia

***Aliases** for diseases linked to the gene PAX6 (5/35)*
- Isolated Aniridia: isolated **aniridia**
- Gillespie Syndrome: **aniridia**, cerebellar ataxia and mental deficiency
- Gillespie Syndrome: **aniridia**, cerebellar ataxia, and mental retardation
- Gillespie Syndrome: **aniridia**-cerebellar ataxia-intellectual disability
- Gillespie Syndrome: **aniridia**-cerebellar ataxia-mental deficiency

Publications (5/105)
- Comparison between **aniridia** with and without PAX6 mutations: clinical and mol... ecular analysis in 14 Korean patients with **aniridia**.
- Paired box mutations in familial and sporadic **aniridia** predicts truncated **aniridia** proteins.
- PAX6 mutations in **aniridia**.
- Mutation in the PAX6 gene in twenty patients with **aniridia**.
- PAX6 mutation as a genetic factor common to **aniridia** and glucose intolerance.

Summaries
- EntrezGene: This gene encodes a homeobox and paired domain-con...hancer regions can cause ocular disorders such as **aniridia** and Peter's anomaly. Use of alternate promoters a...

Variants (1/28)
- **Aniridia** (AN)

Follow the link to the most relevant **GeneCard**.

At the top of the **GeneCard** is a table offering the opportunity to **Jump to** any of the **section**s of the **GeneCard** easily.

Jump to section									
	Aliases	Disorders	Domains	Drugs	Expression	Function	Genomics	Localization	Orthologs
	Paralogs	Pathways	Products	Proteins	Publications	Sources	Summaries	Transcripts	Variants

Discover some of the properties of **PAX6**[4] using the **Jump to section** menu.

Cytogenetic location (**Jump to** the ***Genomics* section**). _____

Number of **UniProt** isoforms (**Jump to** the ***Proteins* section**). _____

Jump to the ***Transcripts* section**. Notice the number of estimated **PAX6** transcripts is inconsistent. In particular:

How many transcripts are predicted by matches to mRNAs in **REFSEQ**? _____

How many transcripts are predicted by the **A**lternative **S**plicing **D**atabase (**ASD**)? _____

How many transcripts are predicted by **Ensembl**? _____

How would you rationalize the discrepancies? _____

4 OK ... so I am not good at mysteries!

Within the **Transcripts section**, there is an **Additional mRNA sequences:** subsection listing mRNA sequences from **Genbank**[5] that match **PAX6**. Follow the link to the **Genbank** entry with accession code **BX640762.1**[6]. You will be taken to the entry as it is stored in the **Genbank** database. The sequence and its annotation is displayed in **Genbank** format and surrounded by useful links that we will ignore for the time being.

How many times does the term **PAX6** occur in this entry's annotation[7]? _____

Move back to the **PAX6 GeneCard**.

Why might the number of **Additional mRNA** matches not match the number of **PAX6** transcripts? _____

_____

Jump to the ***Variants*** **section**.

Listed are the first **5** variations in the genomic region of **PAX6** according to the two variant databases **dbSNP** and **Humsavar**.

Sequence variations from dbSNP and Humsavar for PAX6 Gene

Filter: (results) See all 1,621 »

SNP ID	Clin	Chr 11 pos	Sequence Context	AA Info	Type	MAF
rs1506 5 43	--	31,788,750(-)	ACAGC(A/T)GGGTG		utr-variant-3-prime	
rs592859 5 43	--	31,797,787(-)	TTATC(C/G)TGGGG		intron-variant	
rs608293 5 43	--	31,786,732(-)	ATGGT(A/G)AACAA		utr-variant-3-prime	
rs628224 5 43	--	31,797,626(+)	AGTTC(A/G)TTACT		intron-variant	
rs640258 5 43	--	31,792,182(-)	GCAGG(C/G)CCTCA		intron-variant	

Click on the **See all** link to show all relevant variations. Move to the bottom of the list where you will find the, relatively few, **Humsavar** database entries. Follow one or two of the **Humsavar** links to arrive at pages from ExPASy in Switzerland.

What sort of variations are recorded in the **Humsavar** database? _____

The **dbSNP** variants are linked to corresponding **Ensembl** data, where such exists, by the superscript **5**. Try a couple of these links.

Note that all **SNP**s available from **dbSNP** are also available from **Ensembl**. Why might that be? _____

_____

Filter with the term **rs358**. You should find three **dnSNP** database entries.

Sequence variations from dbSNP and Humsavar for PAX6 Gene

Filter: rs358 (Filtered 3 out of results)

SNP ID	Clin	Chr 11 pos	Sequence Context	AA Info	Type	MAF
rs35821697 5 43	--	31,797,364(+)	ACATT(C/T)TTATC		intron-variant	
rs35883677 5 43	--	31,794,440(+)	CACAC(-/A)CACAC		intron-variant	
rs35840358 5 43	--	31,810,042(+)	AGAGC(C/G)CGGGG		intron-variant, utr-variant-5-prime	

Why might it be considered odd that **rs35883677** be included in **dbSNP**? _____

_____

GeneCards provides links to entries that relate to **PAX6** in many other databases (e.g. **Ensembl** , **OMIM**). You can access entries in any of these databases from **GeneCards**. Mostly for later reference, note:

The **Ensembl** Accession Number for the **aniridia** gene (**Jump to** the ***Aliases*** **section**). _____

The number of human **PAX** genes (**Jump to** the ***Paralogs*** **section**). _____

What **Orthologues** exist in **Mouse** and **Drosophila** (**Jump to** the ***Orthologs*** section)? _____

_____

What functions are suggested for **PAX6** (**Jump to** the ***Summaries*** **section** and/or **Jump to** the ***Function*** **section**)? _

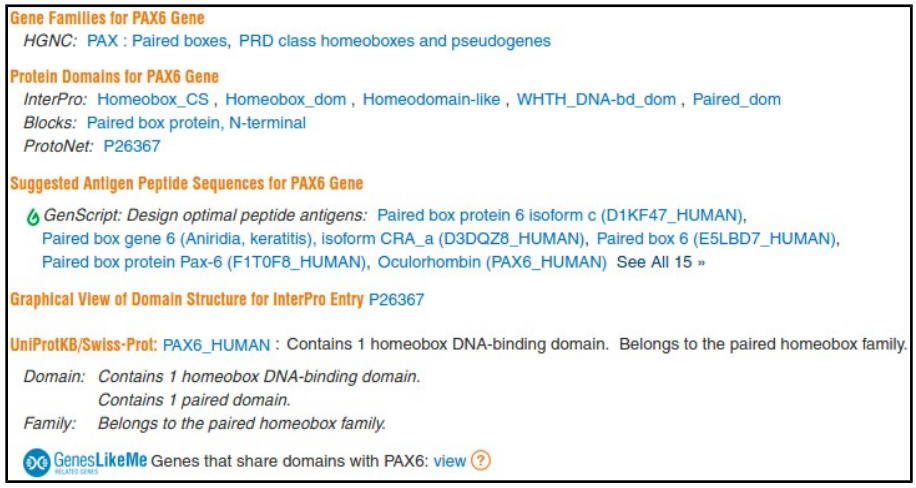
_____

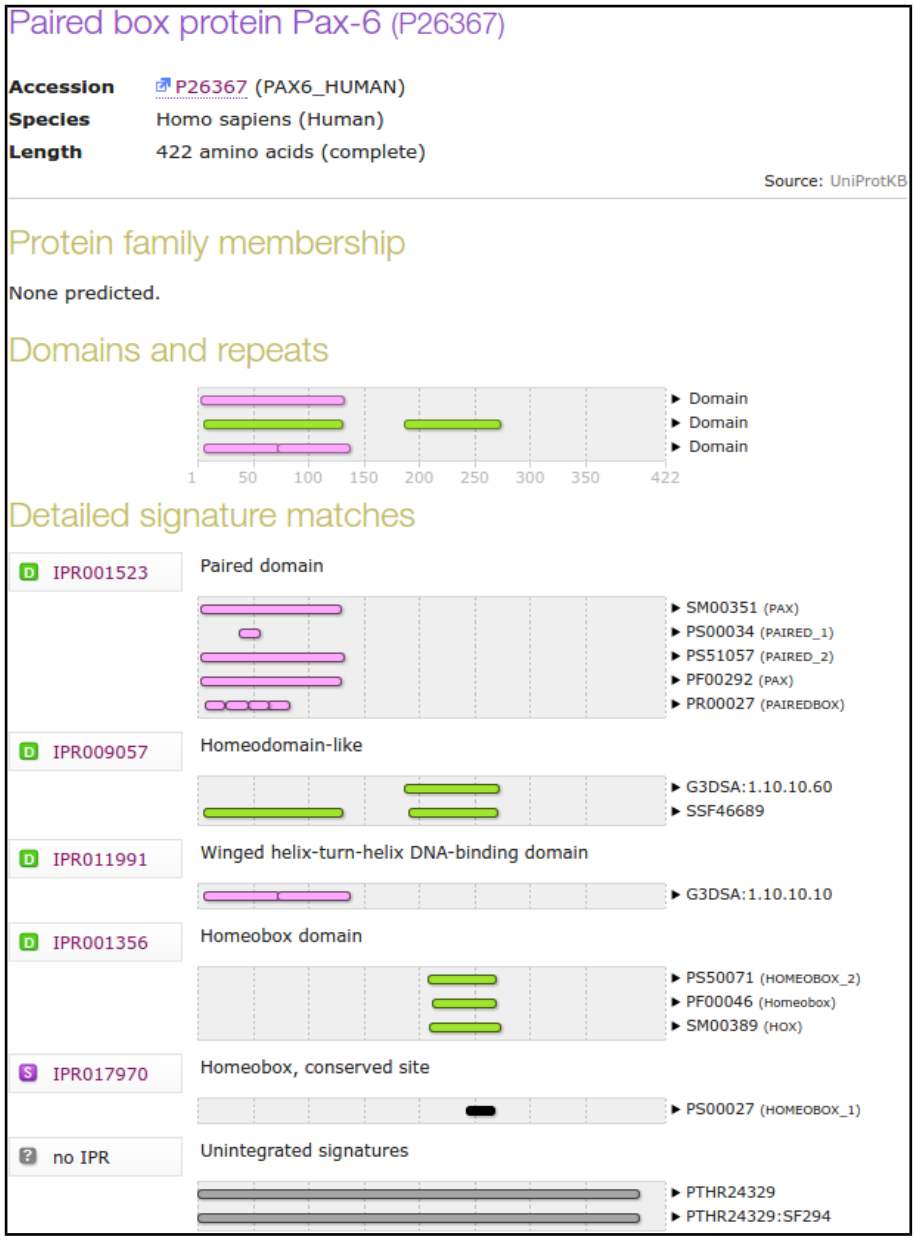
5 You will need to expand the display **See All** the **Additional mRNA sequences** before you can access the link you need. Sequences submitted to **EMBL** (Europe), **Genbank** (America) or **DDBJ** (**D**NA **D**ata**B**ase of **J**apan) are exchanged daily, so these sequences are not specific to **GenBank**.

6 Interpreting the accession code is not straight forward. The prefix letter(s) indicates the type of sequence and the database of first submission. The number before the full stop is there to make the code unique. The number after the full stop is a version number.

7 Search the web page for **PAX6** and ignore all hits in the web page annotation down the right of the page. Only hits in the data entry itself count.

To view the domain structure of the **PAX6** protein, **Jump to** the ***Domains* section**. In the **Protein Domains for PAX6 Gene** section, the **InterPro** and **Blocks** families to which **PAX6** belongs suggest a **Paired Box** domain[8] at the **N** terminal, and a **Homeobox** domain.





Try the Graphical View of Domain Structure for InterPro Entry P26367. This is a graphic generated by the **Interpro** database. See that the presence of the two main domains mentioned in the textual report is confirmed.

See the membership evidence for the two **Interpro** families that suggest a **Paired Box** followed by a **Homeobox**.

We will return here later for a closer look. For now, given that the **Signatures** whose identifiers begin **PS** represent **Prosite** matches and those beginning **PR** represent **Prints** matches. Can you explain:

Why there are two **Prosite** predictions for both the **Homeobox** and the **Paired Box** domains? _____

_____

Why **Prints** appears to predict four very small **Paired box** domains instead of the single larger domain indicated by all the other predictions? _____

A quick look at OMIM 

8 **PA**ired bo**X** (hence the name **PAX**) at the start (**N** terminal) of the protein. We investigate further later on.

Retrieving and Examining DNA Sequence Data

A good start would be to retrieve some of the DNA sequences that are associated with **PAX6**, the gene most associated with the disease **aniridia**, as was discovered from **GeneCards**. This could be done in a number of ways at any of several sites. The most popular choice must be to use **Entrez** at the **NCBI**.

First, go to the **Home Page** of the **The National Center for Biotechnology Information (NCBI)** (if necessary, try your **Bookmarks**, as a last resort, type in the **URL** "**www.ncbi.nlm.nih.gov**").

You will arrive at a page offering access to the many **NCBI** resources available to you. Currently, you only require to search for **DNA** sequences, specifically those that relate to **aniridia**, so first set the database selection field of the **Search** facility at the top of your page to **Nucleotide**. and click on the **Search** button.

You are now offered the subset of **NCBI** resources specific to **Nucleotide**s. You need to search the databases with more sophistication than the basic search offered, so click on the Advanced search option button.

Then in the Nucleotide Advanced Search Builder, change All Fields to Title in the pull down menu associated with the first search field and type in the keywords:

chromosome 11

In the second search field, again change All Fields to Title and type in the keyword:

pax6

You are asking **Entrez** to search for all **Nucleotide** database entries that contain the terms **chromosome 11** and **pax6** in the section of their annotation intended to be a succinct brief description (I.e. **Title**) of the entry. Click on the Search button to start the search going.

There is just one matching entry which is

```
LOCUS       NG_008679              40170 bp    DNA     linear   PRI 11-AUG-2014
DEFINITION  Homo sapiens paired box 6 (PAX6), RefSeqGene on chromosome 11.
```

arrayed before you in **Genbank** format, very neat!! It was the **DEFINITION** line that you searched by selecting the **Field** value **Title**. I needed a few tries to get the right search to find just what was needed, and was a bit surprised at the simplicity and accuracy of the final search. You are looking at a **RefSeqGene** (a subset of the **RefSeq** database) entry. As such, it represents a genomic sequence for a "well-characterised gene", in this case **PAX6**.

Take a brief tour of the **FEATURES** for this entry and you will see that there are actually two genes associated with this sequence. **PAX6**, of course, and **ELP4** on the strand that is the complement of that represented here.

```
gene            5001..38170
                /gene="PAX6"
                /gene_synonym="AN; AN2; D11S812E; MGDA; WAGR"
                /note="paired box 6"
                /db_xref="GeneID:5080"
                /db_xref="HGNC:8620"
                /db_xref="MIM:607108"
gene            complement(38437..>40170)
                /gene="ELP4"
                /gene_synonym="AN; C11orf19; dJ68P15A.1; hELP4; PAX6NEB;
                PAXNEB"
                /note="elongator acetyltransferase complex subunit 4"
                /db_xref="GeneID:26610"
                /db_xref="HGNC:HGNC:1171"
                /db_xref="MIM:606985"
```

```
join(16551..16560,20128..20258,21186..21401,22106..22271,
28174..28332,28848..28930,29160..29310,29409..29524,
32102..32252,32943..33028)
/gene="PAX6"
/gene_synonym="AN; AN2; D11S812E; FVH1; MGDA; WAGR"
/note="isoform a is encoded by transcript variant 1;
paired box homeotic gene-6; oculorhombin; aniridia type II
protein"
/codon_start=1
/product="paired box protein Pax-6 isoform a"
/protein_id=" NP_000271.1 "
/db_xref="GI:4505615"
/db_xref="CCDS: CCDS31451.1 "
/db_xref="GeneID: 5080 "
/db_xref="HGNC: HGNC:8620 "
/db_xref="MIM: 607108 "
/translation="MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISR
ILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEI
RDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGSWGTRP
GWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFT
QEQIEALEKEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRR
QASNTPSHIPISSSFSTSVYQPIPQPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPS
FTMANNLPMQPPVPSQTSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMNSQPMGTSGTT
STGLISPGVSVPVQVPGSEPDMSQYWPRLQ"
```

At the top of your page, **Analyze this sequence** by clicking on the **Highlight Sequence Features** option. The **Co**ding **S**equence (**CDS**) feature for **PAX6** is displayed for you by highlighting the relevant parts (the coding **exons**) of the sequence and displaying the **CDS** details including the DNA regions that need to be **join**ed to form the **CDS** and the **translation** of the **CDS**.

Use the controls at the bottom of your page to look at the other features of this entry (select feature **number** and then click on the **Feature** button).

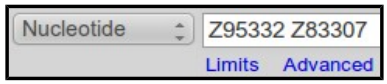
What were the features that you found? _____

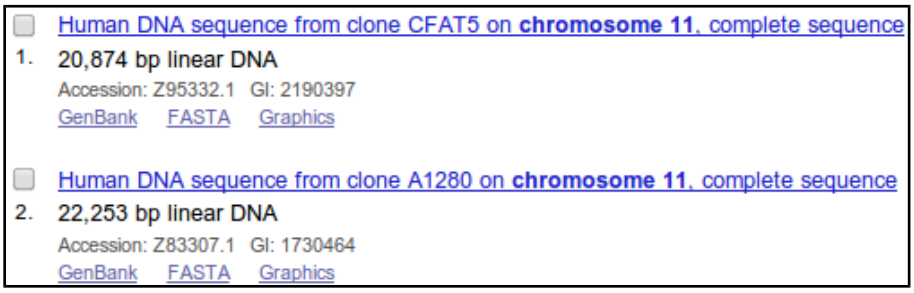
Why might you have expected more features than there were? _____

```
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff in
            collaboration with Isabel Hanson, David FitzPatrick. The reference
            sequence was derived from Z95332.1 and Z83307.1.
            This sequence is a reference standard in the RefSeqGene project.

PRIMARY     REFSEQ_SPAN         PRIMARY_IDENTIFIER PRIMARY_SPAN       COMP
            1-18852             Z95332.1           2023-20874
            18853-40170         Z83307.1           105-21422
```

Take a look at the **COMMENT** and **PRIMARY** sections just above the **FEATURES**. This entry is suggested to be constructed from two sequences from **GenBank**. That is, the products of two sequencing projects.

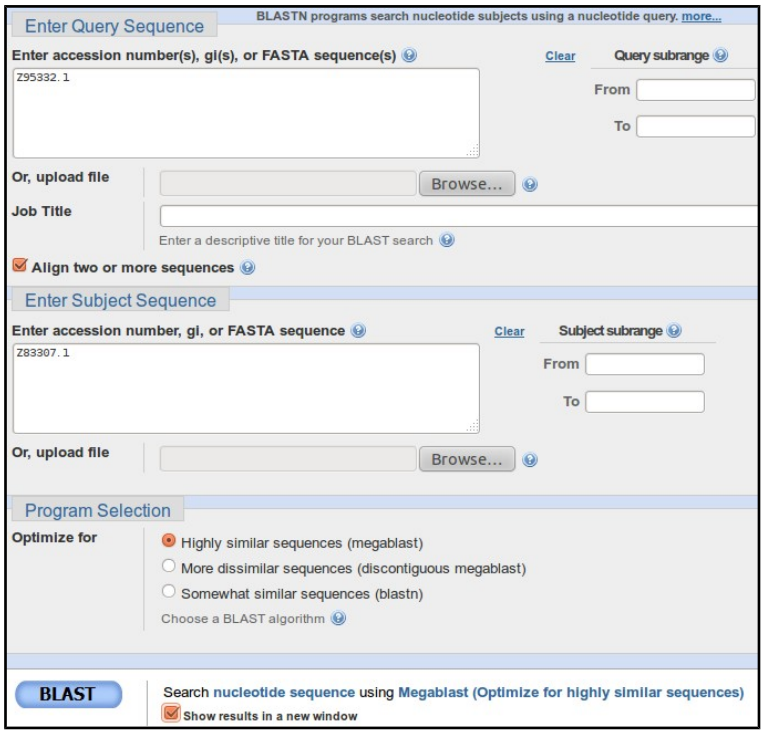
Take a quick look at the **GenBank** entries by entering their **ACCESSION** numbers into the **Search** box at the top of your page. Click on the Search button.





Lo and behold, the two **GenBank** entries are summoned forth. Take a look at one or both. Not particularly illuminating I think[9]. These are clones sequenced as part of the **H**uman **G**enome **P**roject (**HGP**). They served to cover regions of **Chromosome 11** and have little biological significance in themselves.

Move back to the list, as illustrated. Elect to **Analyze these sequences**, selecting from the extensive range of possibilities **Run BLAST**. We will look at **blast** properly later, the idea here is to simple prove that these two sequencing clones really do overlap in the fashion suggested by the evidence so far. So, elect to **Align two or more sequences**[10]. Cut and paste one of the sequencing clone **accession numbers** from the **Enter Query Sequence** box to the **Enter Subject Sequence** section of the form. Elect to **Show results in a new window**[11]. Firmly address the BLAST button.



Just one region of overlap should be identified.

```
Query  20771  GATCCGGAGCGACTTCCGCCTATTTCCAGAAATTAAGCTCAAACTTGACGTGCAGCTAGT  20830
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1      GATCCGGAGCGACTTCCGCCTATTTCCAGAAATTAAGCTCAAACTTGACGTGCAGCTAGT  60

Query  20831  TTTATTTTAAAGACAAATGTCAGAGAGGCTCATCATATTTTCCC  20874
              |||||||||||||||||||||||||||||||||||||||||||
Sbjct  61     TTTATTTTAAAGACAAATGTCAGAGAGGCTCATCATATTTTCCC  104
```

How does the alignment you generated match up with the annotation of the original **RefSeq** entry you discovered? _

---

9 The annotation is very sparse which makes these entries very hard to find directly. The **EML-Bank** versions include some links to the **Ensembl** codes you identified when looking at **GeneCards**. These would have been helpful but are not part of the official **I**nternational **N**ucleotide **S**equence **D**atabase **C**ollaboration (**INSDC**) annotation that should be consistent between **GenBank**, **E**uropean **N**ucleotide **A**rchive (**ENA**), which includes **EML-Bank**, and **D**NA **D**ata **B**ank of **J**apan (**DDBJ**).

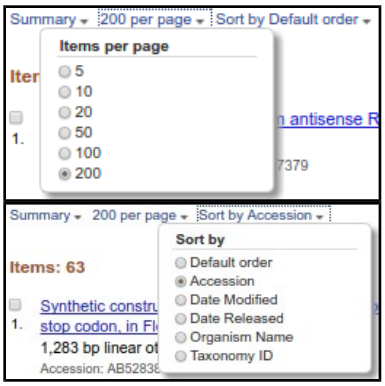
10 As opposed to comparing each of the two clones against an entire sequence database.

11 Just because its neater. In my, significantly less then humble, opinion anyway.

Now for an entirely new search. The easiest way to get a fresh start is to move back to your browser tab displaying the **GenBank** Search results, and then click on the **Advanced** option of the **Search** facility at the top of the page. You should arrive back at the **Nucleotide Advanced Search Builder** offering a fresh start.

Set up a new search as illustrated and set it going. Ultimately simple this time. You have requested all **Human** sequences that are centrally associated with the gene **PAX6**.

You should achieve a list of **60** or so sequences, all clearly claiming **PAX6** association and most proclaiming their humanity loudly in Latin.

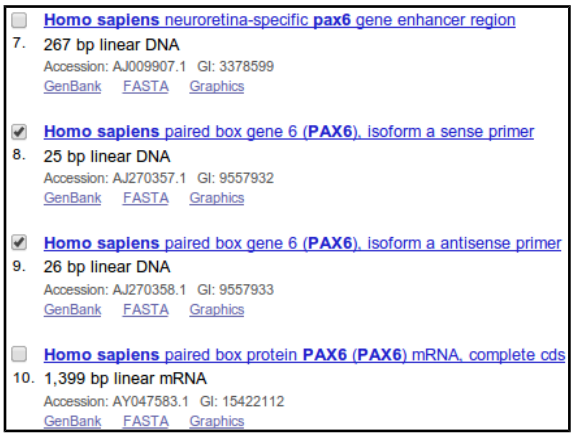


You will have more hits than can be displayed in one go. Also, the hits are arranged in a "**Default**" order which has thus far defied all my attempts to associate with any definition of logic!

To deal with both of these issues, use the display control pull down menus at the top of your page to set the items **per page** to something big and the **Sort by** option to something sane.

The list shows matches between the terms entered and the ***annotation*** of DNA sequences. Not all relevant sequences will be present. For example, the **mRNA** with accession number **BX640762** was justifiably referenced in the **PAX6 Genecard** will not be in this list. **PAX6** appears nowhere in the entire annotation of **BX640762** let alone just its **DESCRIPTION** (or **Title**) field.

A little way down the list you should see two primer sequences. Their **Description**s suggest they are a pair of PCR primers used for picking out the **PAX6** gene. Select both by clicking in their selection boxes.




```
LOCUS       AJ270357                25 bp    DNA     linear   PRI 26-JUL-2000
DEFINITION  Homo sapiens paired box gene 6 (PAX6), isoform a sense primer.
ACCESSION   AJ270357
VERSION     AJ270357.1  GI:9557932
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 25)
  AUTHORS   Palm,K., Salin-Nordstrom,T., Levesque,M.F. and Neuman,T.
  TITLE     Fetal and adult human CNS stem cells have similar molecular
            characteristics and developmental potential
  JOURNAL   Brain Res. Mol. Brain Res. 78 (1-2), 192-195 (2000)
   PUBMED   10891600
REFERENCE   2  (bases 1 to 25)
  AUTHORS   Palm,K.
  TITLE     Direct Submission
  JOURNAL   Submitted (04-OCT-1999) Surgery, Cedars Sinai Medical Center, 8700
            Beverly Blvd., Los Angeles, CA 90048, US
COMMENT     Related entry: NM_000280.
FEATURES             Location/Qualifiers
     source          1..25
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
     misc_feature    1..25
                     /note="PCR sense primer for paired box gene 6 (PAX6),
                     isoform a"
ORIGIN
        1 ccagccagag ccagcatgca gaaca
//
```

Click on the **sense primer**. Properly, you would read all the **References** carefully. Instead, note the length looks about right and return to your list with the **Back** button.

It will be good to investigate these primers later, so find the diminutive 

button which is near the bottom of your page and use it. **Choose** your **Destination** to be **File** and set the **Format** of that file to be **FASTA**. Strike the **Create File** button with a firm and confident click of your every ready mouse. With irritating presumption, the choice of file name is made for you. Your sequences are stored in a file named:


```
sequence.fasta
```

The **NCBI** is justifiably not famed for its understanding of poetry! Do whatever it takes to rename this file to be called:

```
pax6_primers.fasta
```

Retrieving and Examining Protein Sequence Data

For protein sequences, we will move from the **NCBI** and use **UniprotKB**. First move to the home of **UniProtKB (www.uniprot.org)**.

Enter the **GenBank** accession number **Z83307** that you discover at the **NCBI** into the **Query** box at the top of you page and then click the button.

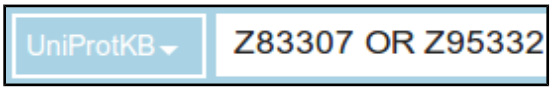
You should achieve a list of around **10** protein sequences all associated with the genes **PAX6** and **ELP4**. Most of the list will be **unreviewed** entries from **UniProtKB/TrEMBL**.

To simplify, click on the button that reduces the list to just the **Reviewed** entries from **UniProtKB/Swiss-Prot**.

Entry	Entry name		Protein names	Gene names	Organism	Length
P26367	PAX6_HUMAN		**Paired box protein Pax-6** (Aniridia type II protein) (Oculorhombin)	**PAX6**, AN2	Homo sapiens (Human)	422
Q96EB1	ELP4_HUMAN		**Elongator complex protein 4** (hELP4) (PAX6 neighbor gene protein)	**ELP4**, C11orf19, PAXNEB	Homo sapiens (Human)	424

Increase the information in the **Protein names** column by clicking the button.

Where have you seen these genes mentioned previously? _____



Now edit the contents of the **Query** box to be both the **Genbank** accession numbers that you discovered at the **NCBI** joined by the boolean operator **OR** and then **click the** button.

Q15293	RCN1_HUMAN		**Reticulocalbin-1**	**RCN1** RCN	Homo sapiens (Human)	331
E9PLM2	E9PLM2_HUMAN		**Reticulocalbin-1**	**RCN1**	Homo sapiens (Human)	20
H0YER5	H0YER5_HUMAN		**Reticulocalbin-1**	**RCN1**	Homo sapiens (Human)	57
H0YDA4	H0YDA4_HUMAN		**Reticulocalbin-1**	**RCN1**	Homo sapiens (Human)	28
E9PP27	E9PP27_HUMAN		**Reticulocalbin-1**	**RCN1**	Homo sapiens (Human)	58

Your list of hits should be a little longer this time. Order the list by Gene names by clicking in the appropriate column heading. The extra proteins should be at the bottom of your list associated with a third gene to which introductions are still pending. Meet **RCN1**. Note that all but one of the proteins associated with this gene are **unreviewed**.

These entries are in **UniProtKB/TrEMBL** (indicated with a as opposed to a which indicates a **reviewed UniProtKB/Swiss-Prot** protein).

For a nice tidy list, again click on the button that reduces the list to just the **Reviewed** entries from **UniProtKB/Swiss-Prot**.

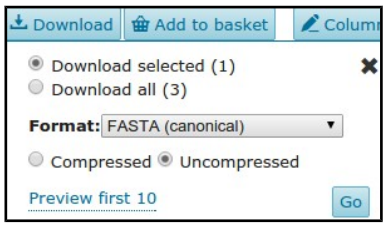
Entry	Entry name		Protein names	Gene names	Organism	Length
Q96EB1	ELP4_HUMAN		**Elongator complex protein 4** (hELP4) (PAX6 neighbor gene protein)	**ELP4**, C11orf19, PAXNEB	Homo sapiens (Human)	424
P26367	PAX6_HUMAN		**Paired box protein Pax-6** (Aniridia type II protein) (Oculorhombin)	**PAX6**, AN2	Homo sapiens (Human)	422
Q15293	RCN1_HUMAN		**Reticulocalbin-1**	**RCN1**, RCN	Homo sapiens (Human)	331

How is that this is the first occasion that the gene **RCN1** has been apparent? _____

How is it that we have found any protein sequences at all by looking at the barren annotation of the two clones **Z83307** and **Z95332**? _____

Select the entry **PAX6_HUMAN** by clicking in the selection box on the extreme left.

Click on the Download button near the top of your page. Ensure **Uncompressed** and **FASTA (canonical)** are selected, as illustrated, and click on the Go button. Do whatever it takes to get the canonical isoform of the selected sequence into a file on your **Desktop** called:

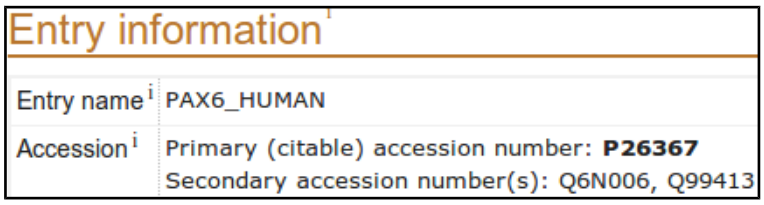


pax6_human.fasta

Deselect the entry **PAX6_HUMAN**. Click on the Accession number (**P26367**) to view the database entry for the protein you have just saved to file in all its **UniProtKB** splendour.

Note the navigation bar down the left hand side of the entry that enables easy access to the various sections of the database record.

Move to the Entry information section.



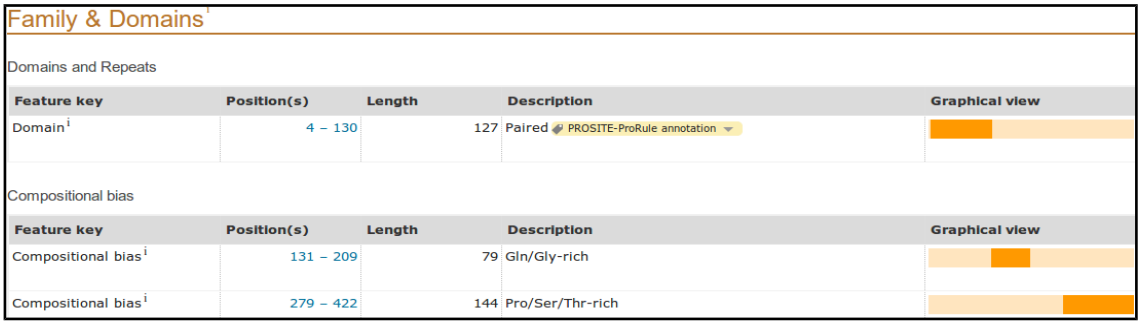
Make a note of the first **UniProtKB Accession number** for the **PAX6** protein. _____

Why do you suppose there is more than one **Accession number** for this protein? _____

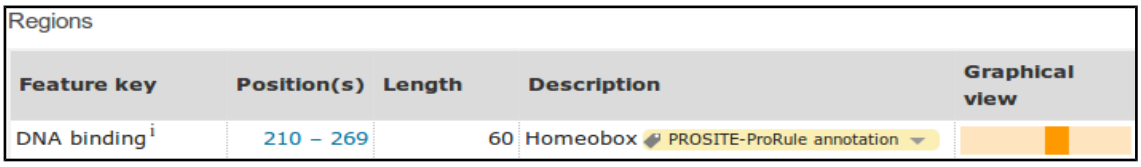
Make a note of the **UniProtKB** Identifier (or entry name)[12]? _____

Move to the Family & Domains section.

Look in the **Domains and Repeats** section for the two major domains suggested by **GeneCards**. Only the **Paired** box **Domain** is present?



Now move to the Function section. Look in the **Regions** section.



You should see the **Homeobox** recorded as a **DNA binding Functional** region. Both the **Paired** box and the **Homeobox** are, of course, **Domains** and both are **DNA binding Functional** regions. It is messy that one is listed as a **Domain**, the other as **DNA binding**[13]? Nevertheless ...

What are the start and end positions of the **Paired** domain? _____

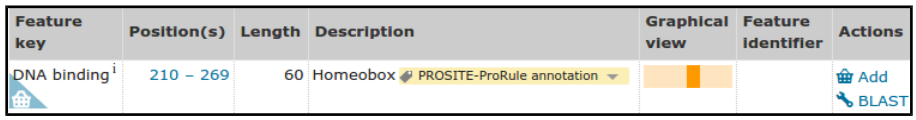
What are the start and end positions of the **Homeobox** domain? _____

Note the range of the **Proline**, **Serine**, **Threonine** rich region at the end of the protein. _____

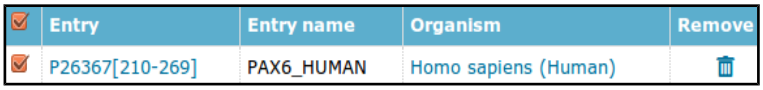
12 **UniprotKB** still sometimes uses 2 names for each entry, an **Identifier** (or **Entry Name**) and an **Accession** code.

13 Apparently, the reasons are "Historic"? An illogicality is admitted, but it is not trivial to fix issues such as this. It is unlikely in the near future therefore, that both **PAX6** domains will be recorded equally as **Domains** that possess the function of **DNA Binding**, which I would suggest to be the ideal.

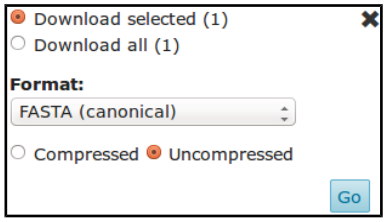
Whilst you have it in view, download the sequence of just the **Homeobox** region for analysis later on. The way to do this involves baskets? As in shopping baskets? OK, ours not to reason why ... just do whatever works. In this instance, click on the Add associated with the **Homeobox region**.



Go to the top of the page and click on Basket 1. Select the only entry.



Click on the Download button. Accept the default settings as illustrated and click on the Go button.



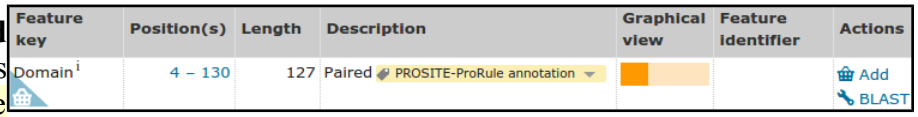
Do whatever it takes to get the **Paired** box sequence into a file on your **Desktop** called:

homeobox_domain.fasta

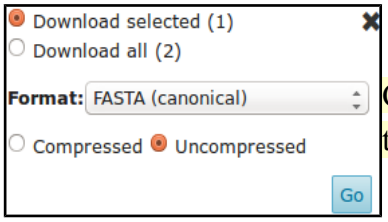
Finally, open up the file you just created in an appropriate text editor and change the first line[14] from:

`>sp|P26367|210-269` to `>Homeobox-Domain P26367(210-269)`

It would also be good to save the **Paired box** region for later analysis whilst it is nearby. So move once more to the Family & Domains section. Click on the Add associated with the **Paired** box **Domain**.



Go up the page and click on Basket 2. Select the **Paired box** entry only.



Click on the Download button. Accept the default settings as illustrated and click on the Go button.

Do whatever it takes to get the **Paired** box sequence into a file on your **Desktop** called:

pax_domain.fasta

As for the **Homeobox** file, open up the new file in an appropriate text editor and change the first line from:

`>sp|P26367|4-130` to `>Pax-Domain P26367(4-130)`

To complete the job tidily, Click on the Clear button to empty your **Basket**. Confirm that you ***really really*** want to tip all the **entries** out of the stupid **Basket**. Enough shopping for today?!?

14 Not strictly necessary, to be honest. However, the | symbol occasionally causes problems, so I suggest removing it.

Consider next the Pathol./Biotech section. In particular, the **Natural variants**.

The mention of **AN** in the description implies that this variant is a cause for the disease **Aniridia**. All the variants recorded here are described as causal for either **AN** or **PETAN**. **PETAN** refers to another disease (**Peters anomaly**).

One might suppose that **all** the variants recorded as substitutions[15] must **all** be associated with one or more **SNP**(s) in the genome.

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier
Natural variant	17 – 17	1	N → S in AN. 1 Publication		VAR_003808
Natural variant	18 – 18	1	G → W in AN. 1 Publication		VAR_003809
Natural variant	19 – 19	1	R → P in AN. 1 Publication		VAR_047860
Natural variant	22 – 26	5	Missing in AN. 2 Publications		VAR_008693
Natural variant	26 – 26	1	R → G in PETAN. 2 Publications		VAR_003810
Natural variant	29 – 29	1	I → S in AN. 1 Publication		VAR_008694
Natural variant	29 – 29	1	I → V in AN. 1 Publication		VAR_003811
Natural variant	33 – 33	1	A → P in AN. 1 Publication		VAR_008695
Natural variant	37 – 39	3	Missing in AN. 1 Publication		VAR_008696

However, only the variation at **375** is associated with a **dbSNP** entry?

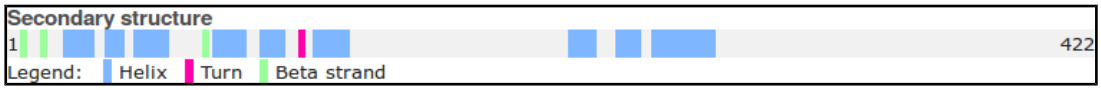
Natural variant	375 – 375	1	P → Q in AN; reduced DNA binding ability. 1 Publication Corresponds to variant rs200015827 [dbSNP | Ensembl].		VAR_015066

Why do you suppose this is the case? _____

_____

Click on the A → P link associated with position **33**. This substitution is the disease causing **Natural variant** that is the focus of many of the exercises that follow. You will be taken to entry **VAR_008695** of a variant database (glanced at when considering **Humsavar** whilst looking around **GeneCards**) at **Expasy** in Switzerland (to be revisited). This entry very probably represents a short cut to much we will more ponderously discover by other means. Too easy!! take a quick look and then move back to **PAX6_HUMAN**.

Move to the Structure section.



Click on the **Show more details** button.

Feature key	Position(s)	Length	Description	Graphical view
Beta strand	6 – 8	3	Combined sources	
Beta strand	14 – 16	3	Combined sources	
Helix	23 – 34	12	Combined sources	
Helix	39 – 46	8	Combined sources	
Helix	50 – 63	14	Combined sources	
Beta strand	77 – 79	3	Combined sources	
Helix	81 – 93	13	Combined sources	
Helix	99 – 108	10	Combined sources	
Turn	114 – 116	3	Combined sources	
Helix	120 – 133	14	Combined sources	
Helix	219 – 229	11	Combined sources	
Helix	237 – 246	10	Combined sources	
Helix	251 – 275	25	Combined sources	

Describe the arrangement of Helices within **PAX6**. _____

_____

Note the start position of the middle helix of each set of three. _____

_____

Note the end position of the third helix of each set of three. _____

_____

Note the position of the Beta strands relative to the helix groups. _____

_____

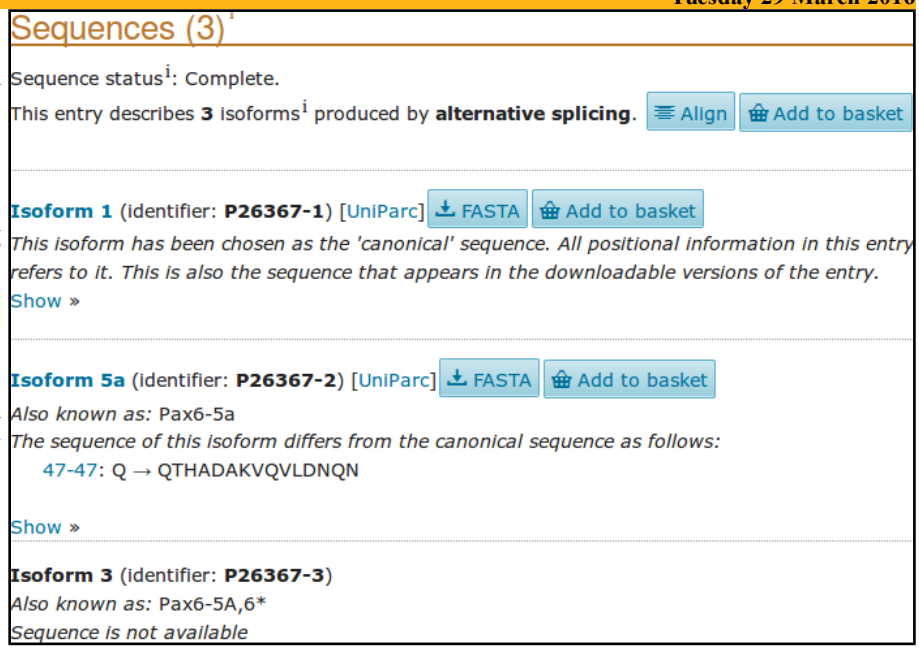
15 That is, all but those at positions **22-26** and **37-39**, which are **deletions**.

Click on the link to the Sequences section. This section confirms the **Uniprot** isoform count of **3** recorded in **GeneCards**.

The sequence of **isoform 1**, the 'canonical' sequence, is the one you will have saved in a file called **pax6_human.fasta**. Having admired this sequence, displayed by default, send it away by clicking on the nearest « Hide button.

The second isoform, **Isoform 5a**, is shown here to have **14** extra amino acids after position **47** (a **Q**) of the canonical sequence.

The **3rd** isoform, **Isoform 3**, declares "The sequence of this isoform is not available", indicating that although the literature records Western blots suggesting a third protein product[16], the protein sequence has yet to be determined.



Compare the two isoforms with sequence. Click on the Align link[17]. The best isoform alignment is computed and displayed by **ClustalW**, a popular tool for aligning sequences, that you will use again later. The isoforms are extremely similar. The alignment highlights the only difference well. The extra amino acids in the second isoform are clear to see. Experiment with the **Annotation** possibilities. I choose to show the two major domains and the helices, showing clearly the way the helix triplets are arranged in the **PAX** and **Homebox** domains.

Describe the arrangement of Helices within the two major domains of **PAX6**. _____

_____

16 "No experimental evidence..." in this section means that the splice variant is only identified as a result of a match with a single cDNA or EST which is regarded as insufficient to assign an amino acid translation. All **UniProt/SwissProt** splice variants supported by matches with 2 or more cDNA/ESTs are checked to ensure they are not actually just errors due to frameshifts, intron retention etc.

17 The **Select** button allows choice of isoforms. Redundant here as we wish to compute an alignment of the only two isoforms of known sequence.

Move back to the page showing **P26367**. You should still have the Sequences section in view.

Note the extra sequence in **P26367-2** and where it starts. _____

Move to the Cross-references section and see that the names of several relevant sequences are recorded, including the genomic two genomic sequencing clones.

Select the link destinations:	
EMBL	M77844 mRNA. Translation: AAA59962.1.
GenBank	M93650 mRNA. Translation: AAA36416.1.
DDBJ	AY047583 mRNA. Translation: AAK95849.1.
	BX640762 mRNA. Translation: CAE45868.1.
	Z95332, Z83307 Genomic DNA. Translation: CAG38363.1.
	Z83307, Z95332 Genomic DNA. Translation: CAG38087.1.
	BC011953 mRNA. Translation: AAH11953.1.

How would you rationalise the reference to the mRNA entry **BX640762** here? _____

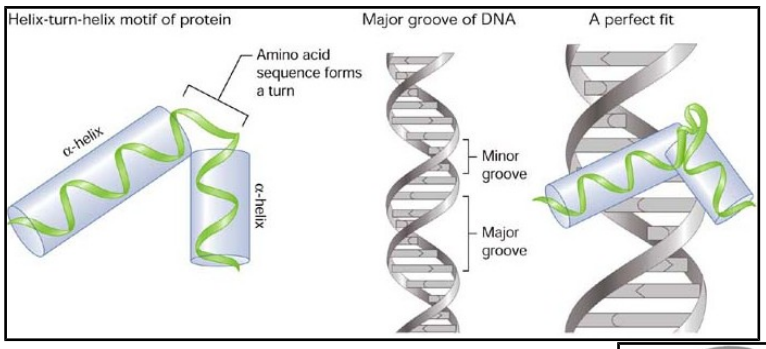
_____

PROSITE	PS00027. HOMEOBOX_1. 1 hit.
	PS50071. HOMEOBOX_2. 1 hit.
	PS00034. PAIRED_1. 1 hit.
	PS51057. PAIRED_2. 1 hit.
	[Graphical view]

At the bottom of the Cross-references section are links to **Family and domain databases**, including **PROSITE** database. Click on the link to the third **PROSITE** entry down, the first for **PAIRED**.

You will be presented with the documentation page for the **PROSITE** pattern/profile for a **Paired domain** at the top of which is a general **Description** of a **Paired domain**. I found this quite a difficult read. Please do not struggle with it too much. My summary of the parts of the message relevant to these exercises is:

- A **Paired domain** is about **126** amino acids long
- It is generally found at the beginning of a protein
- It is often followed by a homeodomain (as here) and/or an octopeptide (of unspecified properties?)
- There is often a **Pro-Ser-Thr-rich C-terminus** (as in this case)
- A paired domain is a DNA binding domain that has 2 binding regions each of which involves a helical triplet
- The second and third helices of each helical triplet form **helix-turn-helix** (**HTH**) motifs
- The **HTH** regions bind the DNA major groove[18]



- The first helical triplet is preceded by a **β-turn** and **β-hairpin** ("wing") that participate in the DNA binding

- The linker region between the two helical triplets can bind the **DNA minor groove**

All of these properties have already been suggested and/or will be discovered variously as the exercise progresses.

18 If, like me, you have conceptual problems with major and minor groves. Try this **animated picture**. Helped me at least. As did the image above.

Move down to the **Technical section**. **PROSITE** stores representations of conserved protein features (**motifs**). It uses two methods. One requires a bit of thought, but the other is quite intuitive. That which you are now considering is a simple **Consensus pattern** defined clearly at the top of this section. The suggestion here is that this pattern is to be found in all known **Paired domain** sequences[19].

Technical section

PROSITE methods (with tools and information) covered by this documentation:

PAIRED_1, PS00034; Paired domain signature (PATTERN)

- Consensus pattern:
  R-P-C-x(11)-C-V-S
- Sequences in UniProtKB/Swiss-Prot known to belong to this class: 58
  - detected by PS00034: 58 (true positives)
  - undetected by PS00034: 0 (false negative or 'partial')
- Other sequence(s) in UniProtKB/Swiss-Prot detected by PS00034:
  7 false positives.

What is the **Consensus pattern** for a **Paired domain**? _____

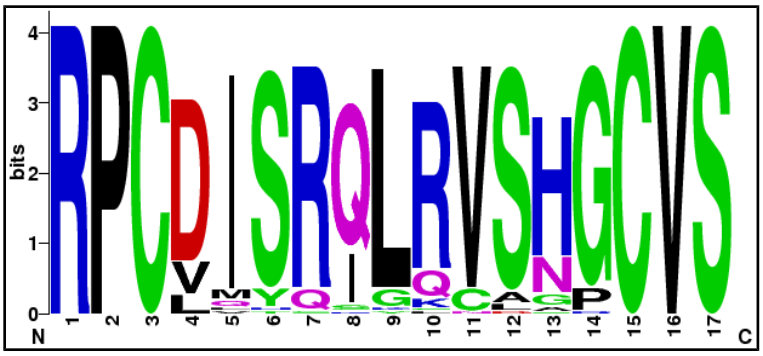
Where in a **Paired domain** should the **Consensus pattern** occur? _____

How would you interpret this **Consensus pattern**? _____

_____

How effective does the **Technical section** imply this pattern to be**?** _____

_____



There are a number of links to follow here. I would suggest you Retrieve the sequence logo from the alignment. The logo is representing the degree to which each position of the **Consensus pattern** is conserved and how. The higher the letter, the more prominent is that letter in that position[20]. The particular value of the logo here is to illustrate that the region of the **Consensus pattern** is not only conserved over the **3** positions at either end. It is well conserved over its entire length. Not well enough, however, to be described effectively using such a simplistic strategy as this sort of **Consensus pattern**.

Move back and up to the **Description** of a **Paired domain**. Follow the link to the **Prosite** documentation for **homeobox** embedded here (**PDOC00027**) .

How well does the secondary structure suggested by **Prosite** match that recorded by **Uniprot**? _____

_____

How many **helix-turn-helix** motifs would you expect in the **homeobox** domain? _____

Where (in amino acid positions) would you expect all the **PAX6 HTH** motifs to be? _____

19 This is not strictly true, as you will discover later.

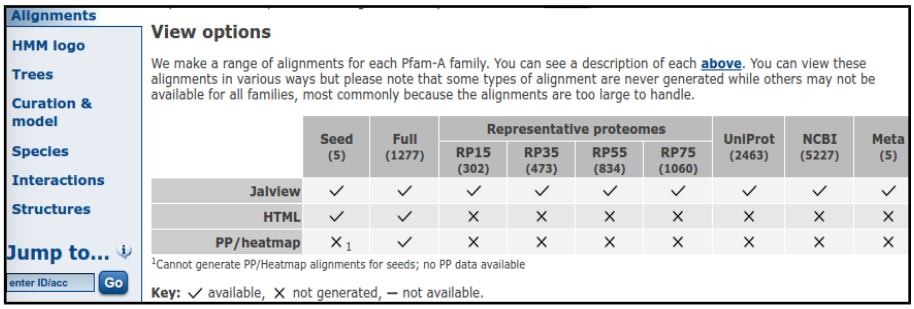
20 A fuller description is provided from the **Logo** page. Click on the **sequence logo help document** link.

Pfam [i]	PF00046. Homeobox. 1 hit. PF00292. PAX. 1 hit. [Graphical view]

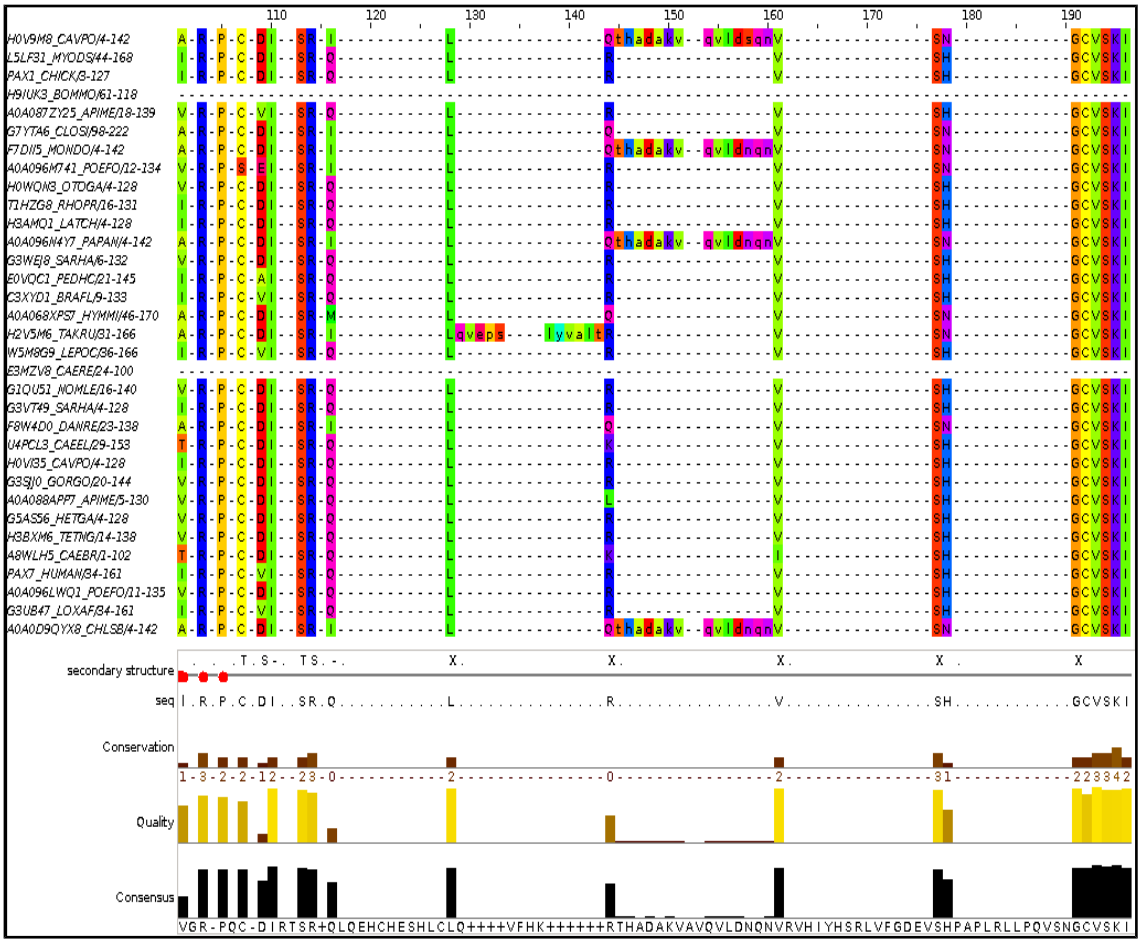
Move back to the page displaying the **UniprotKB/SwissProt entry** for the human **PAX6** protein. Find the **Pfam** links in the **Family and domain databases** section (use the Cross-references button again if necessary). Unsurprisingly, there are two links to **Pfam** .

The **Pfam** database is a collection of protein domain families[21]. Each family is represented by multiple sequence alignments and **H**idden **M**arkov **M**odels (**HMM**s)[22].

Click on the accession code for the **Pfam** entry for **PAX** (**PF00292**). The requested **Pfam** entry is displayed offering access to much **PAX** related information in other databases. Now, click on the **Alignments** link on the left of the page. In the **View options** section, click on the tick in the **Full** column of the **Jalview**[23] Row. A new window will thrust its way onto your screen displaying (possibly after a little persuasion) the requested alignment displayed by the **Jalview** applet.



More **Jalview** functionality is available when running **Jalview** via **Java Web Start**, so click on the start Jalview via Java Web Start button[24]. In a new window, you should now see the same alignment even more garishly coloured for your delight[25]. Use the scroll bar on the right to scroll up and down the full set. The **Seed** alignment for this family is carefully constructed and manually curated, however the **Full** alignment is transparently not. The alignment is automatically generated by the program **HMMER3**. The alignment includes a number of **PAX** domains that align badly at the beginning. The region illustrated is that around the **isoform 5a 14** amino acid insertion.



Allowing for the distorted numbering of the alignment, how would you interpret the extra **14** or so amino acids that some sequences appear to have around position **100-190**?

How might you interpret the way that **HMMER3** suggests that some sequences have a similar insertions in a slightly different places to others?

21 For a more complete description of **Pfam** go to: **http://pfam.sanger.ac.uk/help**.

22 **HMM**s are essentially mathematical representations of the alignments. Not very pretty for humans to look at, but currently considered the best way to present an alignment to a computer program so that alignments can be compared with each other and/or other protein sequences.

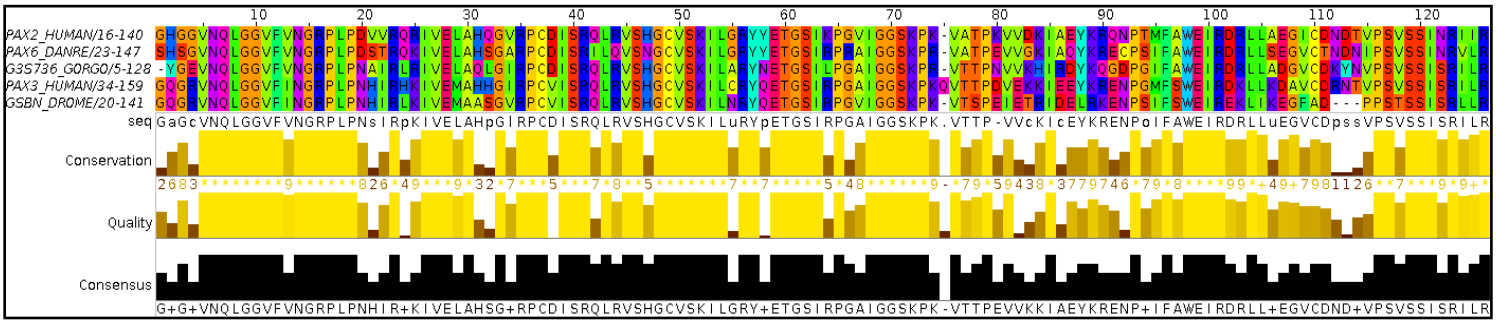
23 A very nice Java tool for viewing alignments that we will use again.

24 Exactly what you have to do next should be intuitive (mostly a matter of replying affirmatively to a series of foolish questions), but can vary according to operating system and browser. Whatever is required to display the alignment – **do it**.

25 On some systems, there can be problems getting **Java Web Start** to behave properly. Ask if you have any difficulty.

For a more restrained view, move back to the **Pfam PAX** entry page, move back to the **Pfam PAX** entry page. In the **View options** section, click on the tick in the **Seed** column of the **Jalview** Row.

Click on the start Jalview via Java Web Start button to start the **Java Web Start** version of **Jalview**. In view are the aligned sequences of the **Seed** alignment from which the profile **HMM** for **PAX** is calculated. None of the **5** seed sequences include the **14** extra amino acids noted previously[26]. Human **PAX6** is not a seed sequence.



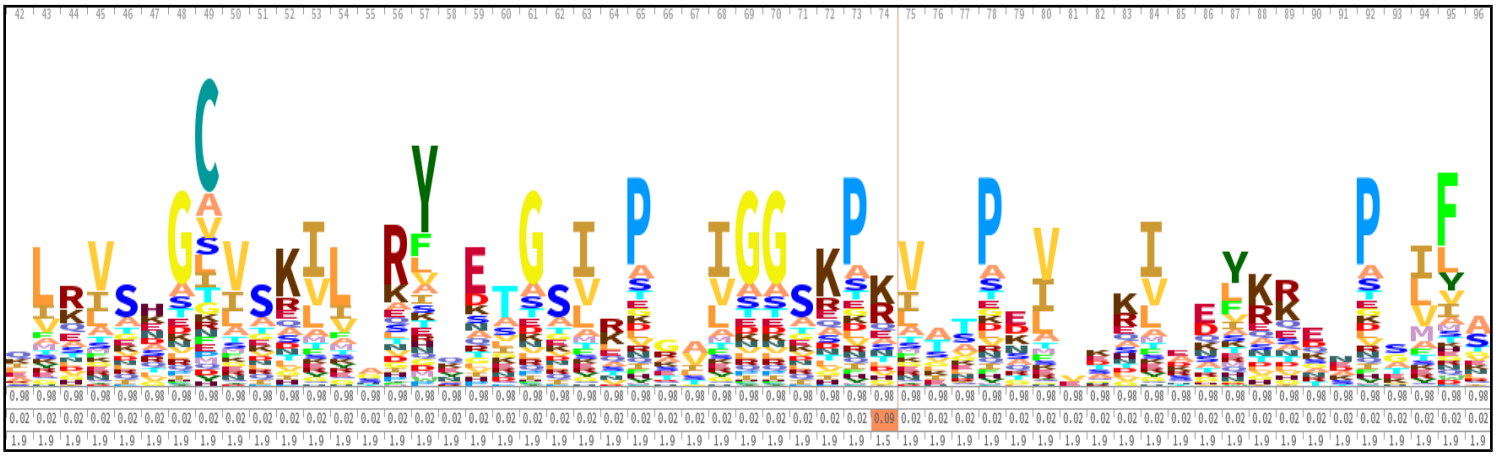
Note down the position in the alignment where all but one seed sequence has a gap. _____

Note the consensus character at this point and its most logical interpretation. _____

Back again to the **Pfam PAX** entry page. Click on the **HMM Logo** link on the left of the page. This is a way of visualising the **HMM** profile computed from the seed sequence alignment you have just been viewing. The logos are indubitably very beautiful. There is a link their documentation just above the picture.

How is the heavily gapped position of the seed alignment represented in its **HMM Logo**? _____

_____



How would you interpret the **Logo** in this region? _____

_____

PRINTS[i] PR00027. PAIREDBOX. Go back now to the **PAX6_HUMAN UniProtKB/Swiss-Prot** entry. Find the **PRINTS** links in the **Family and domain databases** section (use the Cross-references button again if necessary). In **PRINTS**, domains are represented by **fingerprints**. A **fingerprint**, in this context being an ordered group of conserved motifs.

How many links to **PRINTS** would you expect? _____

What if anything, do you think is missing? _____

Well, if you thought all was **OK**, you were jolly well wrong! Should be a link to a **Homeobox** domain surely? You will see why it is missing later, when we look at the tools used to search **PRINTS** (now a supplementary exercise).

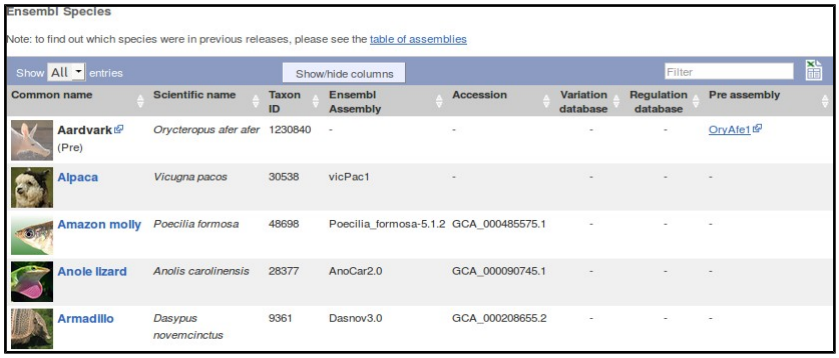
Other databases from the European Bioinformatics Institute (EBI)

26 **Full** alignment columns that are not represented in the seed alignment (and so do not contribute to the calculation of the **HMM**), are shown in lower case. As has been noted previously, including the **14** extra positions referenced here.

Browsing Genomes with Ensembl

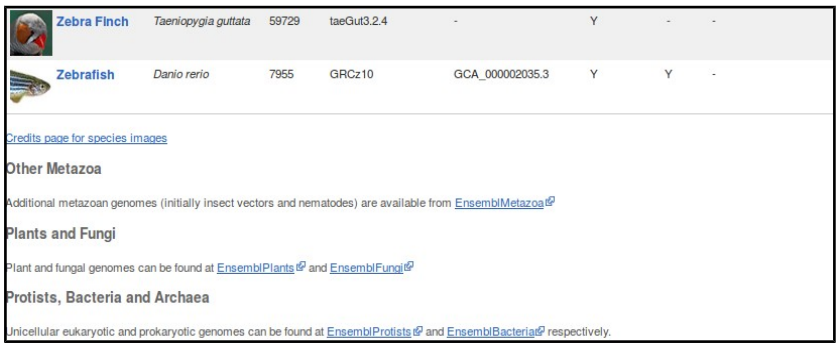
The objective now is to examine **PAX6** and related genes in the context of the entire human genome. You will use the genome browser **Ensembl**, one of several well known facilities allowing easy access to genomic information. Similar databases and browsers can be found at **NCBI** and the **U**niversity of **C**alifornia, **S**anta **C**ruz (**UCSC**).

Go to the **Ensembl** home page (**`www.ensembl.org`**). Choose to **View full list of all Ensembl species** using the link just under the **Select a species** menu.



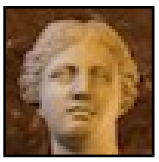
Note that **Ensembl** offers far more than just the Human Genome.

In particular, note the links to **EnsemblPlants**, **EnsemblFungi**, **EnsemblBacteria etc. Ensembl** databases at the bottom of the list.



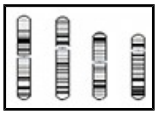
During this exercise, you will only look at the Human genome, by far the most fully developed. However, all the other **Ensembl** genomes are behind the same interface. The techniques required to examine the Human genome are broadly those required to examine any **Ensembl** genome.

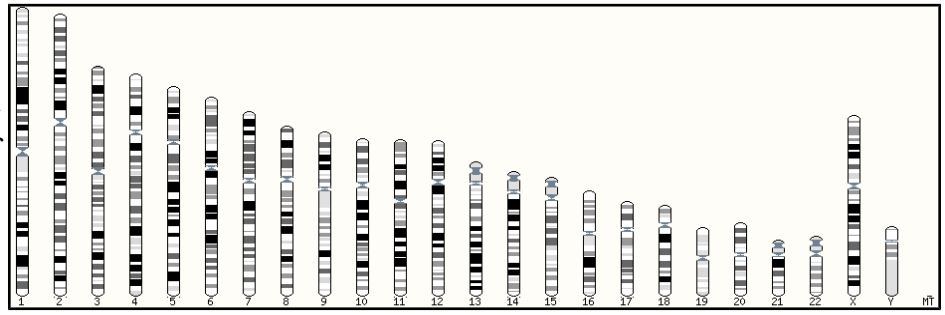
Try looking at the **Species tree** by clicking on the **View the full Ensembl species tree** link (at the top of the **Species list** page). This shows the main (i.e. vertebrate) **Ensembl** species arranged according to their probable evolution.



Move back to the **Ensembl** Home page. Click on the Human genome icon.

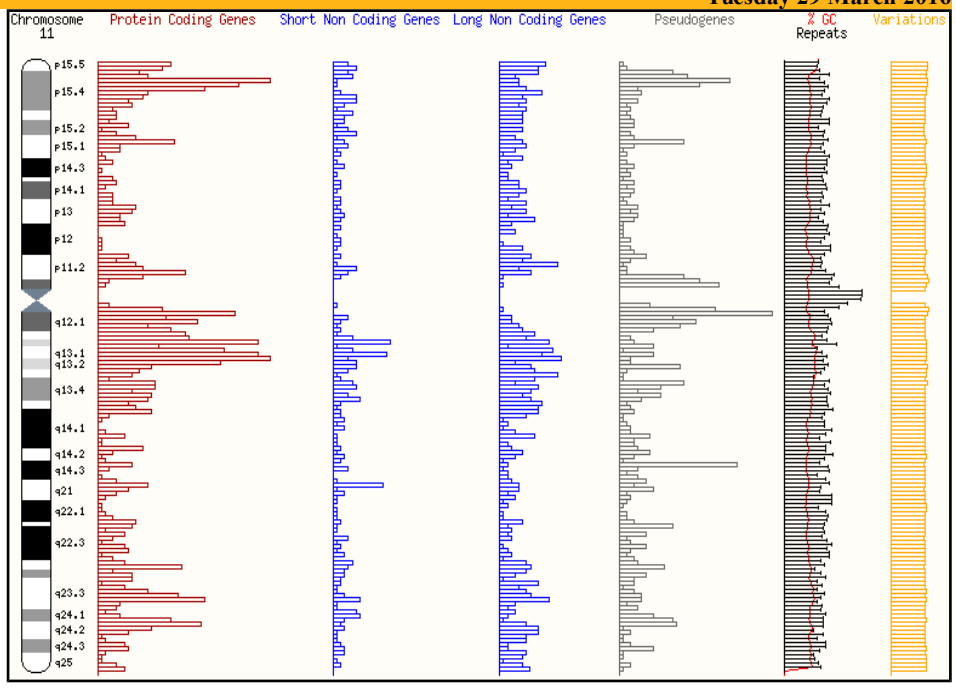
Select the **View karyotype** link. Your reward will be an image of the banding patterns of all the human chromosomes.





Notice the links to the major alternative genome database browsers, particularly the two very popular options from America. Try one or both of these. You will be linked to a view from the alternative browser as close to the **Ensembl** page you are viewing as is possible. The **NCBI** site, in this instance, offers a very similar view of all the chromosomes. An easy route to these alternative services can be very useful. No one provider is "best" at everything.

From the **Ensembl Karyotype** view, select the **PAX6** chromosome and then **Chromosome Summary** from the menu presented. Look at the **PAX6** chromosome region.

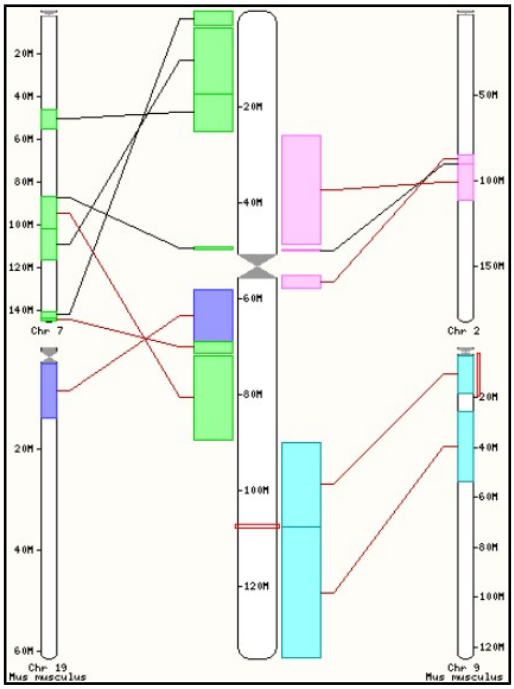


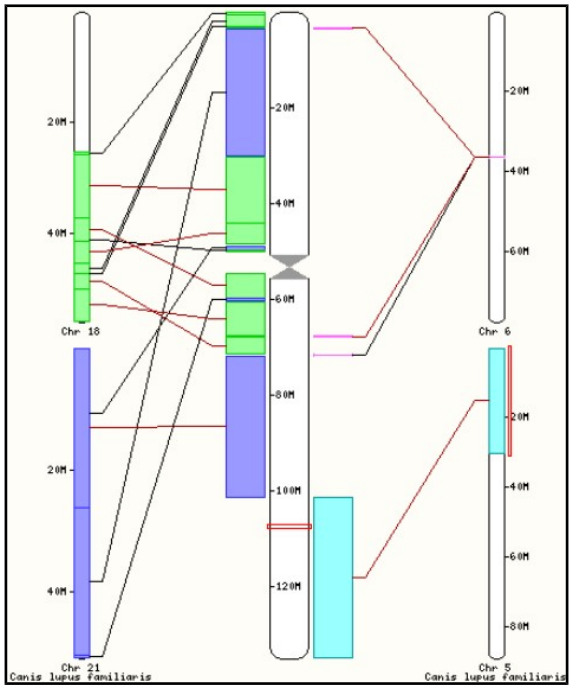
Is it a gene dense region?_____

_____

What about Variation density?_____

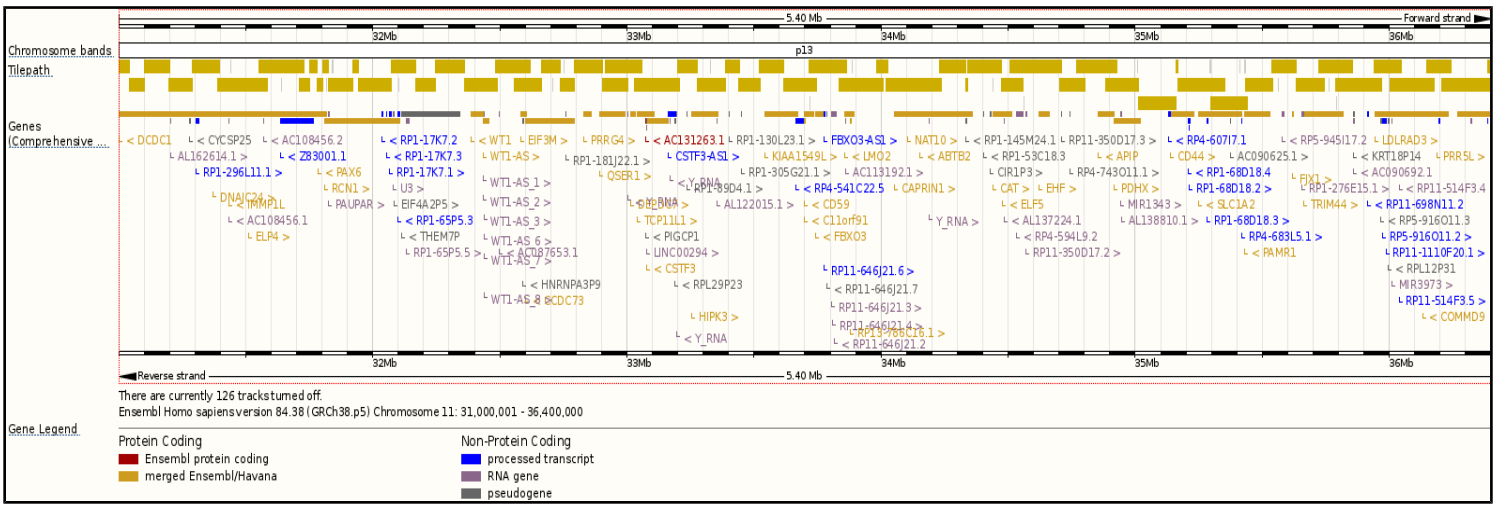
_____





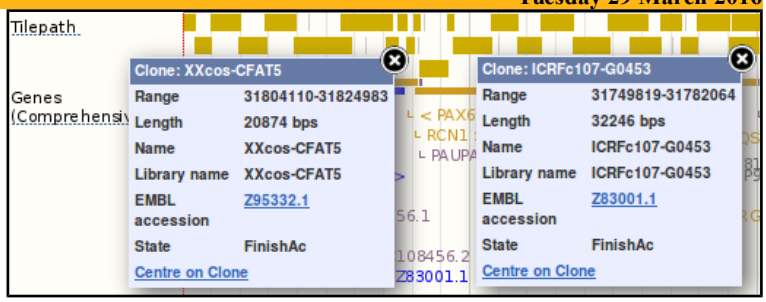
A short diversion. Click on the **Synteny** link on the left hand side of the page. A graphic will appear showing all regions (over **100,000 bp**) of Human Chromosome **11** that are very similar to regions of the mouse genome. You can choose from a number of other **Ensembl** species. I choose **mouse** and **dog**. Diversion over, move back to the **Chromosome summary** of Human Chromosome **11**.

Click on your chosen band and select the proffered link to the region. A **Region overview** is generated. Look in the section marked **Genes (Comprehensive ...** you should be able to locate **PAX6**.

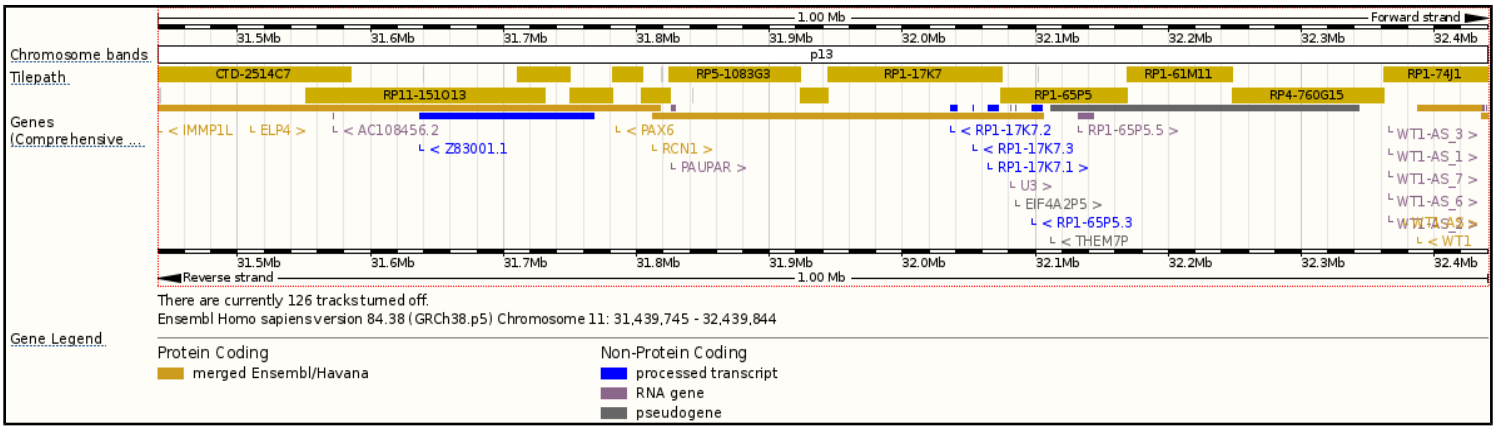


Can you also see the two other genes you might have expected to be in this region? _____

The gold rectangles of the **Tilepath** track (sometimes referred to as the "**Golden Path**") represent the way selected sequencing clones from the **Human Genome Project** overlap to define the entirety of **Chromosome 11**. If you click with care, you should be able to identify the two clones you found in **Genbank** that cover the **PAX6** region.

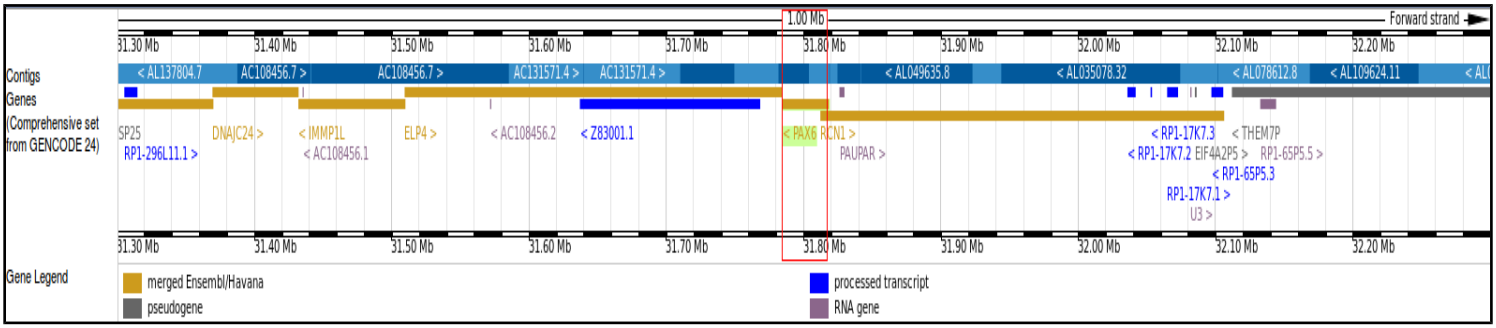


In order to see a detailed view of a particular gene, it is necessary to click on its representation at the top of the **Genes (Comprehensive …** track. That is, the relevant portion of one of the rectangles in the two rows just below the three rows of the **Tilepath**. This is practically impossible at the resolution you have currently, so zoom in around the **PAX6** region by clicking on the < PAX6 gene name and selecting the **Centre here** link of the **Location** menu.



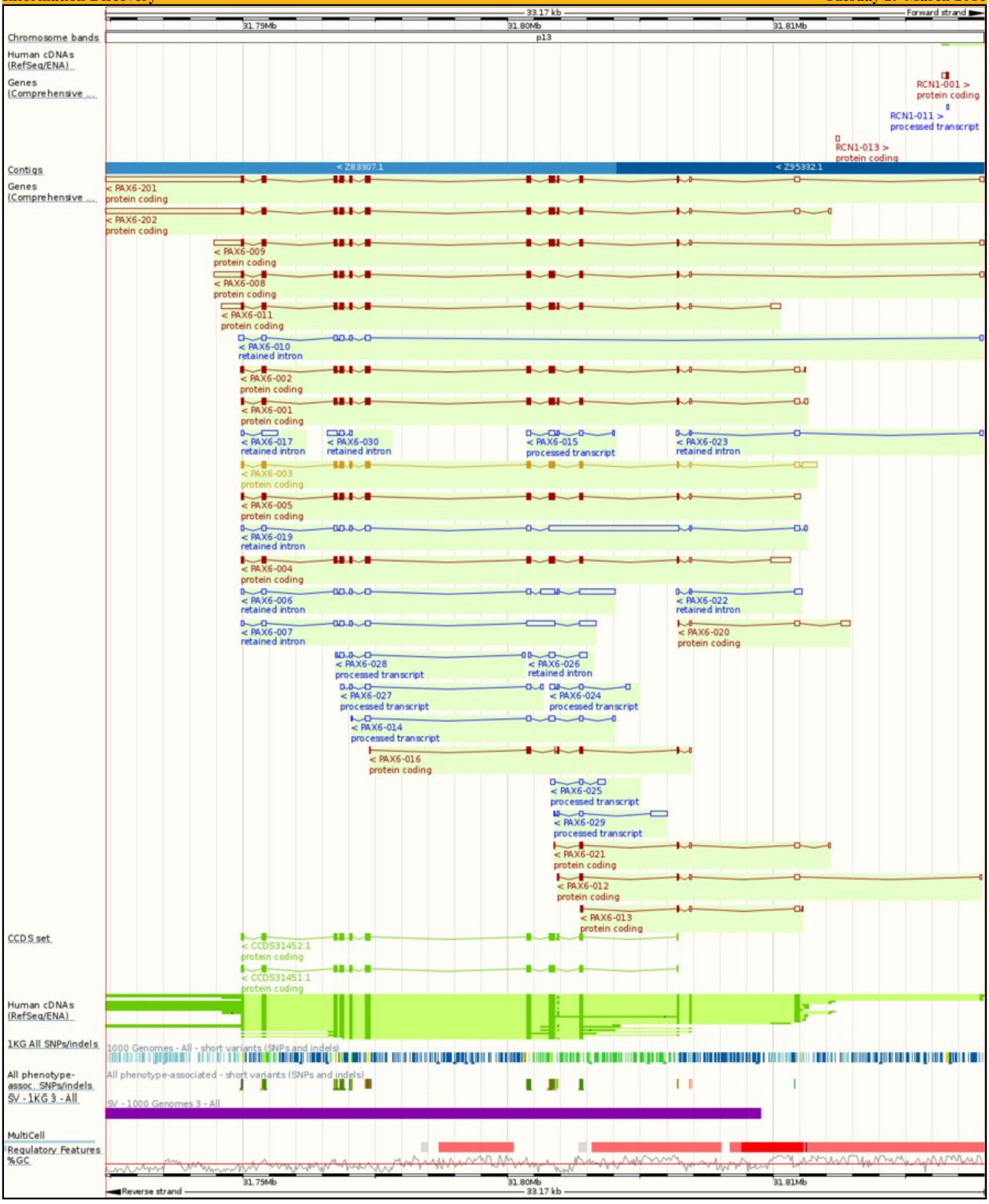
The **Region Overview** is redrawn for a more specific region, with the **PAX6** gene in the centre (well, nearly?). If you note that the ⌊ symbol indicates the leftmost extent of the gene with which it is associated, you should now be able to identify the gene block that corresponds to **PAX6**. Be careful though, as the **PAX6** gene block actually runs into the **ELP4** gene block! The genes do not overlap, so their blocks are both at the same level. As there is too small a gap between them the two genes cannot be represented as distinct blocks. You need to aim at the right hand extremity of the very long block that represents **ELP4**, **PAX6** and other genes even further to the left. Not the clearest of graphics I suggest!

So, having found it, click on the **PAX6** gene block and choose the **Location** option. **Ensembl** zooms into a **Region in detail** view centred (exactly this time!) around **PAX6**. This view is comprised of two illustrations. The first, more general view, shows the location of **PAX6** and immediately neighbouring features, predominantly genes. The colours of the genes represented in this display indicate the gene type (see **Gene Legend**). The second illustration shows a more detailed view of the region exactly spanned by **PAX6**[27].

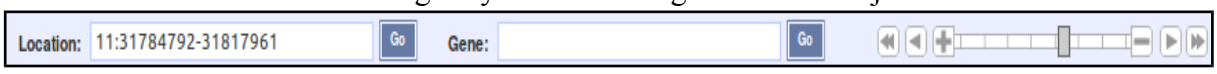


In both views, in the **Contigs** track, the "**Golden Path**" is now represented as thick alternately light and dark blue lines. The red rectangular box in the more general view indicates the region of the chromosome represented in the more detailed view.

27 The simplest (most sensible?) route to his point would have been to use the search facility offered from the **Ensembl** home page. From here you could have specified **Human** as your organism of choice and **PAX6** as a search term to indicate your area of interest. This would get you a list of matching genes with **PAX6** at its head. From here, select the target gene to arrive at a textual summary. Then just jump to the gene **Location** view that now enlivens your life. Faster and more direct? True … but where is the poetry? Try it, if you have the time and inclination.

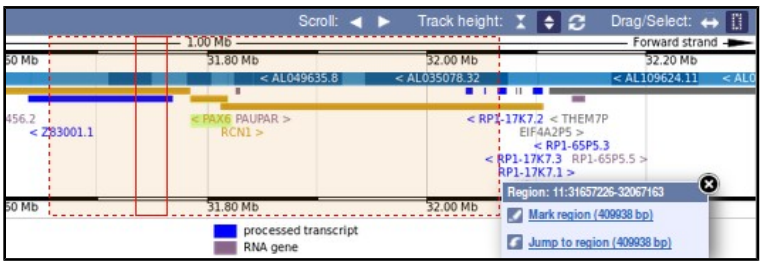


The more detailed view **Location** bar indicates the region you are viewing and allows adjustment and zoom. In the detailed view, the forward strand is described above the light/dark blue line, the reverse strand below. **PAX6** is on the reverse strand and so its predicted transcripts appear below the blue line and should be read from right to left. The filled boxes of the transcripts represent coding exons, the unfilled boxes non-coding exons. The wiggly lines joining the boxes are the introns. As should not surprise you, many of the transcripts of **PAX6** span two **contigs**.



There are a number of ways to customize the region displayed in the **Region in detail** view. The most dramatic is to Scroll: ◄ ► the the red box in the more general view left or right. The area of focus (governed by the size of the red box) remains constant, but is adjusted horizontally. When you release you mouse, you are asked whether you wish your detailed view to reflect the new logical position of your red box[28] (**Update this image**), or whether you wish to bottle out and return from whence you came (**Reset scrollable image**). Try it.

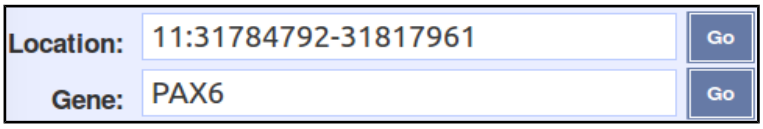
To scroll in a more reserved fashion, again in the more general view, select the **Scroll to a region** part of the Drag/Select: ↔ button. Now you can scroll your view left or right by clicking image and moving your mouse left and/or right. Once you have done this, **Ensembl** checks that you really like where you ended up in the same way it did for more abandoned scrolling. Try it.

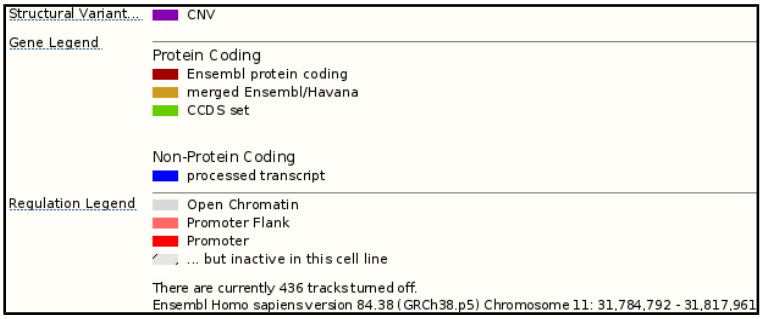


To change the size of the red box defining the area displayed in detail, click on the **Select a region** part of the Drag/Select: ↔ button. Then drag your mouse over the region you wish to be displayed. When you release the mouse button, choose to **Jump to region**. Try it.

If you just click anywhere in the more general view, you will be asked if you wish to **Centre here**. If you do this, both views will be re-centred as requested without any adjustment to the size of the region in view. Try it.

For single base accuracy, you could set the **Location** values in the more detailed view, but I wonder when one would to do this? These possibilities examined, use the **Gene** field of the to get **PAX6** centrally back into focus.





By default, in the more detailed view, transcripts confidently predicted by the **Ensembl** pipeline and/or the **Vega**/**Havana** project are shown. The colour of a transcript indicates its type[29].

The **Human cDNAs (RefSeq/ENA)** and CCDS **set**[30] tracks show the prime evidence for the transcript predictions.

The **Human cDNAs (RefSeq/ENA)** track represents very close matches between good quality human **cDNA**/**mRNA** sequences and the genome. Dark green boxes indicate matched regions and suggest the position of all exons. Light green boxes indicate regions of the genome between matches, suggesting introns.

The **CCDS set** track represents very close matches between the most assured coding DNA sequences and the genome. Dark green boxes indicate matched regions and suggest the position of coding exons. Wiggly lines join **CCDS** matches, suggesting the positions of introns between coding exons.

A good match between a high quality **cDNA**/**mRNA** sequence and the genome can suggest all the exons of a transcript. A corresponding good match between a **CCDS** sequence and the genome can suggest which of that transcript's exons code for protein.

What are the (familiar?) contig numbers containing all of **PAX6**? _____

Explain the visible differences between the coding exons of transcripts **PAX6-008** and **PAX6-009**?_____

_____

28 That is, the centre of the display, the red box actually moves with the display.

29 See the **Gene Legend** at the bottom of the display, and illustrated here.

30 **CCDS (Consensus CoDing Sequence)** is a collaborative effort to identify a core set of human and mouse protein coding regions that are consistently annotated and of high quality. The long term goal is to support convergence towards a standard set of gene annotations.

From the left hand side of the browser page, select **Configure this page**. Turn on the features:

Genscan predictions (from **Genes and transcripts → Prediction transcripts**, choose **with labels**)

Proteins (mammal) from UniProt (from **mRNA and protein alignments → Protein alignments**, choose **Normal**)

Then click on **SAVE and close** (the  in the top right hand corner). The **Detailed view** will reassemble showing where any **Mammalian Uniprot proteins** match, extremely closely, translated regions of the genomic sequence. As do the **CCDS set** track, these matches indicate where protein coding exons exist in the genome. Essentially, the **CCDS set** track and the **Proteins (mammal) from Uniprot** tracks serve the same purpose. That is they both identify the coding exons of each transcript prediction. The **CCDS set** track is far more targeted and effective, to the extent that a **Protein Comparison** track is no longer included in the default display. The **CCDS** evidence is also far more efficiently generated and is used to reduce the search space of the less efficient searches that follow in the pipeline.

The **Genscan prediction** track shows where a program called **Genscan** predicts gene structures directly from the genomic sequence. **Genscan** bases its judgement on models derived from other known human gene structures. **Genscan** is very good, but it does get over excited at times. It misses little, but tends to predict some spurious exons[31]. **Genscan** and similar programs will be discussed briefly later (in a **Supplementary Exercise**). **Genscan** predictions are not sufficiently accurate to be taken, in detail, at face value (hence they are not included in the default display), however, they are very fast and accurate enough to be used to reduce the search space for subsequent gene searches. **Genscan** is run ahead of even the **CCDS** searches in the pipeline.

What **PAX6** exon of note has **Genscan** omitted to predict?_____

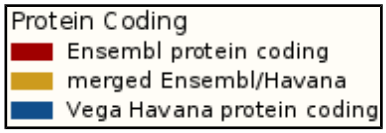
Select **Configure this page** once again. Turn on the features:

Vega Havana (from **Genes and transcripts → Genes**, choose **Expanded with labels**)

EST-based (from **Genes and transcripts → Genes**, choose **Expanded with labels**)

Then click on **SAVE and close**. You should now have in view three types of transcript prediction including purple **EST-based genes** that are suggested only by matches with **E**xpressed **S**equence **T**ags (**ESTs**[32]). Of course, the **Vega Havana** transcript predictions are a subset of the predictions displayed originally by default. On the other hand, the **EST-based** transcript predictions are not considered of sufficient reliability to include in the default view.

From the **Gene Legend** at the bottom of your display, you learn that:



Red transcripts are **Ensembl protein coding**. A rough translation being "Protein coding transcripts predicted either by the **Havana/Vega** people or by the **Ensembl pipeline** of programs, ***but not both***".

Gold transcripts, and there is currently only one, are **merged Ensembl/Havana**. A rough translation being "Protein coding transcripts predicted, in all important respects, by ***BOTH*** by the **Havana/Vega** people ***AND*** by the **Ensembl pipeline** of programs".

Dirty blue? transcripts are **Vega Havana protein coding**. A rough translation being "Protein coding transcripts predicted by the **Havana/Vega**". These are the ones you have just caused to be added to your display.

Some questions now, but .. why not just cheat and look the answers up straight away for these? Whilst worthy, they could easily take up more time than they deserve.

Is a **merged Ensembl/Havana** transcript also a **Vega Havana protein coding** transcript?_____

Which **Ensembl protein coding** transcripts were only predicted by the **Ensembl** pipeline?_____

For pedants only. Can you see how to identify the **Ensembl protein coding** predictions only predicted by the **Ensembl** pipeline by the way they are numbered? _____

_____

31 Particularly as it is run in the **Ensembl** pipeline. Here its purpose is only to identify where genes are likely to be. Sequence comparison between just these regions and coding sequences (**CCDS** sequences, proteins, cDNAs etc.) are then used to predict transcript detail. **Ensembl** only requires **Genscan** to work approximately to reduce the volume of sequence comparisons rather than to generate "stand alone" gene predictions.

32 An **E**xpressed **S**equence **T**ags (**EST**) is a short (**200-500** nucleotide) sequence representing a single reads of a cDNA. As it is generated from a single sequencing read, it will not usually be of very high quality.

The features now displayed represent just a few of those offered by **Ensembl**[33]. Feel free to try more.

Before looking at some of the **gene** and **transcript** properties **Ensembl** offers, it would be good to save the genomic sequence of this region for analysis later on.

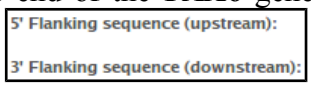
To effect this, first click on the 

link on the left hand side of the page.

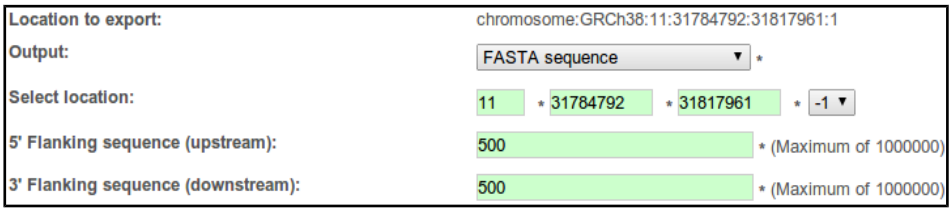
PAX6 is on the negative strand of the genomic region in view, so in the **Select location** section of the **Export data** page the will fly forth, you must change the strand selection from 

to 

.

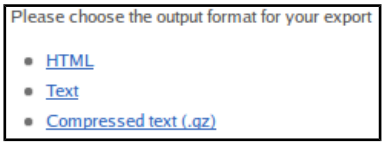
Ask for **500** base pairs of extra sequence at either end of the **PAX6** gene. That is, set both 

to **500**.



Click on the 

button.



Choose **Text** as the **output format** for the sequence to be saved.

In you browser you should now have the genomic region of the **PAX6** gene, with **500** base pairs of flanking sequence on either end, in **FASTA** format.

on your **Desktop**. Then use a suitable text editor to change the rather clumsy first line to:

`pax6_genomic.fasta`

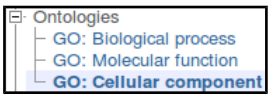
>pax6-genomic sequence

Move back to the **Ensembl** view of the **PAX6** genomic region.

33 My display is telling me that I have have well over **400** tracks turned off!

Click on the gold **Ensembl/Havana gene Ensembl** transcript in the more detailed view and then on the link to the **Gene**[34]. You now see the **Gene summary** for **PAX6**[35].

Click on the **GO: Cellular component** link (from **Gene-based displays → Ontologies**). You are offered a tabular view of a part of the ontology for this transcript.



Accession	Term	Evidence	Annotation Source	Transcript IDs	
GO:0000790	nuclear chromatin	IDA	UniProtKB/Swiss-Prot:P26367	ENST00000379111 ENST00000241001 ENST00000379107 ENST00000606377 ENST00000379123 ENST00000419022 ENST00000379109 ENST00000379115 ENST00000379129 ENST00000379132	• Search BioMart • View on karyotype
GO:0005622	intracellular	IEA	Propagated from Mus_musculus ENSMUSP00000087870 by orthology	ENST00000419022	• Search BioMart • View on karyotype
GO:0005634	nucleus	IDA	UniProtKB/Swiss-Prot:P26367	ENST00000379111 ENST00000241001 ENST00000379107 ENST00000606377 ENST00000379123 ENST00000423822 ENST00000419022 ENST00000379109 ENST00000524853 ENST00000455099 ENST00000438681 ENST00000379115 ENST00000379129 ENST00000379132	• Search BioMart • View on karyotype
GO:0005654	nucleoplasm	IDA	UniProtKB/Swiss-Prot:P26367	ENST00000379111 ENST00000241001 ENST00000379107 ENST00000606377 ENST00000379123 ENST00000419022 ENST00000379109 ENST00000379115 ENST00000379129 ENST00000379132	• Search BioMart • View on karyotype
GO:0005737	cytoplasm	IDA	UniProtKB/Swiss-Prot:P26367	ENST00000379111 ENST00000241001 ENST00000379107 ENST00000606377 ENST00000379123 ENST00000419022 ENST00000379109 ENST00000379115 ENST00000379129 ENST00000379132	• Search BioMart • View on karyotype

I choose to direct you to the **Cellular component** section of the **Gene Ontology**, purely because it is the smallest, and thus easiest to represent here. Feel free to glance at the other two sections. As you can see, each table records those properties of the protein(s) associated with the selected gene (**PAX6** in this case).

I only aim to impart an awareness of the **Gene Onotogy**. To covered **GO** properly would require more space than can be afforded here. Click here to discover more for yourselves.

How might the **Gene Ontology** improve sequence database text searching? _____

Why is the **Transcript Ids** column necessary? _____

Follow one or two of the **GO accession code** links in any of the **3** tables. These will take to **GO** entries from the **Gene Ontology Home** site.

Generally, **GO** terms are more prolific here than in **Uniprot**[36] but the message is the same concerning the domains and functions of the **PAX6** gene transcripts.

34 It does not really matter which protein coding transcript you choose. All the **PAX6** transcripts correspond to the same gene. However, to consistently reproduce exactly the results I got, you need to choose the one gold transcript currently on offer.

35 Or you might have taken the more pedestrian route of clicking on the **Gene: PAX6** tab waiting patiently at the top of the page? This would give you a view of the gene disassociated from any particular transcript.

36 Because **Ensembl** searches for **GO** terms in both **Uniprot** and **Refseq**.

Click on the Show transcript table button to display the transcript information textually. Transcripts are not necessarily in the same order in the table as in the **Regional** view.

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
PAX6-201	ENST00000419022	6922	436aa	Protein coding	CCDS31452	F1T0F8 P26367	NM_001258462 NM_001310158 NM_001310161 NP_001245391 NP_001297087 NP_001297090	TSL:1 GENCODE basic
PAX6-202	ENST00000606377	6860	436aa	Protein coding	CCDS31452	F1T0F8 P26367	NM_001258463 NM_001310161 NP_001245392 NP_001297090	TSL:1 GENCODE basic
PAX6-009	ENST00000379129	2616	436aa	Protein coding	CCDS31452	F1T0F8 P26367	-	TSL:5 GENCODE basic
PAX6-011	ENST00000379107	2591	436aa	Protein coding	CCDS31452	F1T0F8 P26367	-	TSL:5 GENCODE basic
PAX6-008	ENST00000379132	2574	422aa	Protein coding	CCDS31451	P26367 Q66SS1	NM_001127612 NP_001121084	TSL:5 GENCODE basic APPRIS P1
PAX6-003	ENST00000379123	2160	422aa	Protein coding	CCDS31451	P26367 Q66SS1	NM_000280 NM_001258464 NP_000271 NP_001245393	TSL:1 GENCODE basic APPRIS P1
PAX6-001	ENST00000379115	1763	436aa	Protein coding	CCDS31452	F1T0F8 P26367	NM_001604 NP_001595	TSL:1 GENCODE basic
PAX6-002	ENST00000241001	1631	422aa	Protein coding	CCDS31451	P26367 Q66SS1	-	TSL:1 GENCODE basic APPRIS P1
PAX6-005	ENST00000379111	1627	422aa	Protein coding	CCDS31451	P26367 Q66SS1	NM_001258465 NP_001245394	TSL:1 GENCODE basic APPRIS P1
PAX6-004	ENST00000379109	2157	422aa	Protein coding	-	P26367 Q66SS1	-	CDS 3' incomplete TSL:2 APPRIS P1
PAX6-020	ENST00000525535	677	2aa	Protein coding	-	-	-	CDS 3' incomplete TSL:3
PAX6-021	ENST00000524853	574	57aa	Protein coding	-	E9PKM0	-	CDS 3' incomplete TSL:4
PAX6-012	ENST00000423822	567	61aa	Protein coding	-	B1B1I9	-	CDS 3' incomplete TSL:3
PAX6-016	ENST00000455099	497	124aa	Protein coding	-	B1B1J0	-	CDS 3' incomplete TSL:5
PAX6-013	ENST00000438681	455	38aa	Protein coding	-	B1B1I8	-	CDS 3' incomplete TSL:2
PAX6-029	ENST00000533156	847	No protein	Processed transcript	-	-	-	TSL:5
PAX6-014	ENST00000471303	782	No protein	Processed transcript	-	-	-	TSL:5
PAX6-027	ENST00000531910	643	No protein	Processed transcript	-	-	-	TSL:3
PAX6-015	ENST00000481563	613	No protein	Processed transcript	-	-	-	TSL:3
PAX6-028	ENST00000530373	572	No protein	Processed transcript	-	-	-	TSL:4
PAX6-025	ENST00000530714	567	No protein	Processed transcript	-	-	-	TSL:4
PAX6-024	ENST00000534353	540	No protein	Processed transcript	-	-	-	TSL:4
PAX6-019	ENST00000533333	6173	No protein	Retained intron	-	-	-	TSL:2
PAX6-006	ENST00000470027	2842	No protein	Retained intron	-	-	-	TSL:2
PAX6-007	ENST00000494377	2460	No protein	Retained intron	-	-	-	TSL:2
PAX6-010	ENST00000464174	979	No protein	Retained intron	-	-	-	TSL:5
PAX6-017	ENST00000474783	702	No protein	Retained intron	-	-	-	TSL:2
PAX6-030	ENST00000532916	627	No protein	Retained intron	-	-	-	TSL:3
PAX6-026	ENST00000534390	578	No protein	Retained intron	-	-	-	TSL:4
PAX6-023	ENST00000532175	524	No protein	Retained intron	-	-	-	TSL:3
PAX6-022	ENST00000527769	487	No protein	Retained intron	-	-	-	TSL:3

Note the number of **RefSeq mRNA**s and **RefSeq** proteins associated with **Ensembl** transcripts predictions. _____

_____

Which of the **24 RefSeq mRNA**s reported by **GeneCards** can be seen here? _____

_____

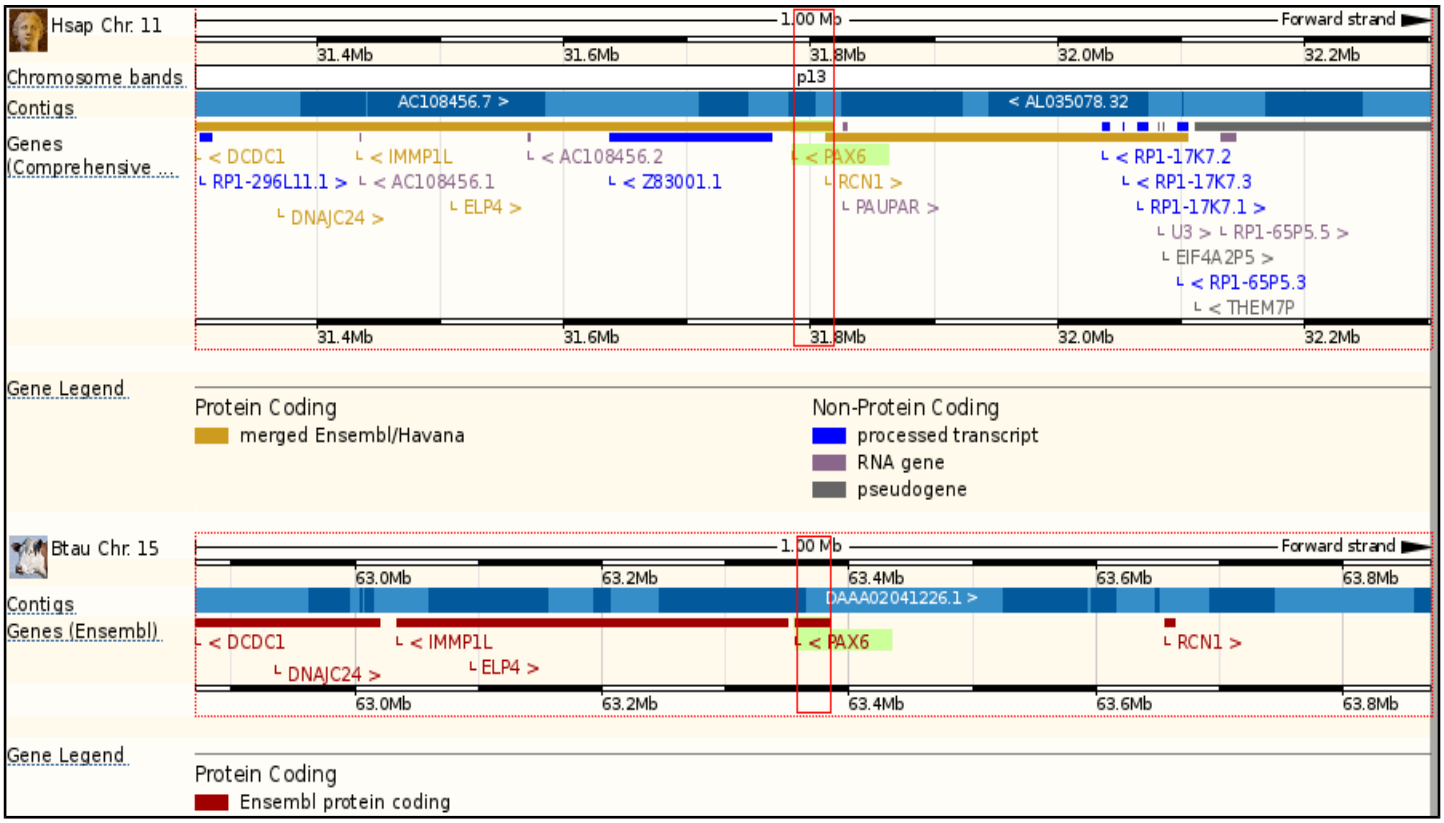
Why would you suppose these **mRNA**s were selected and the others ignored? _____

_____

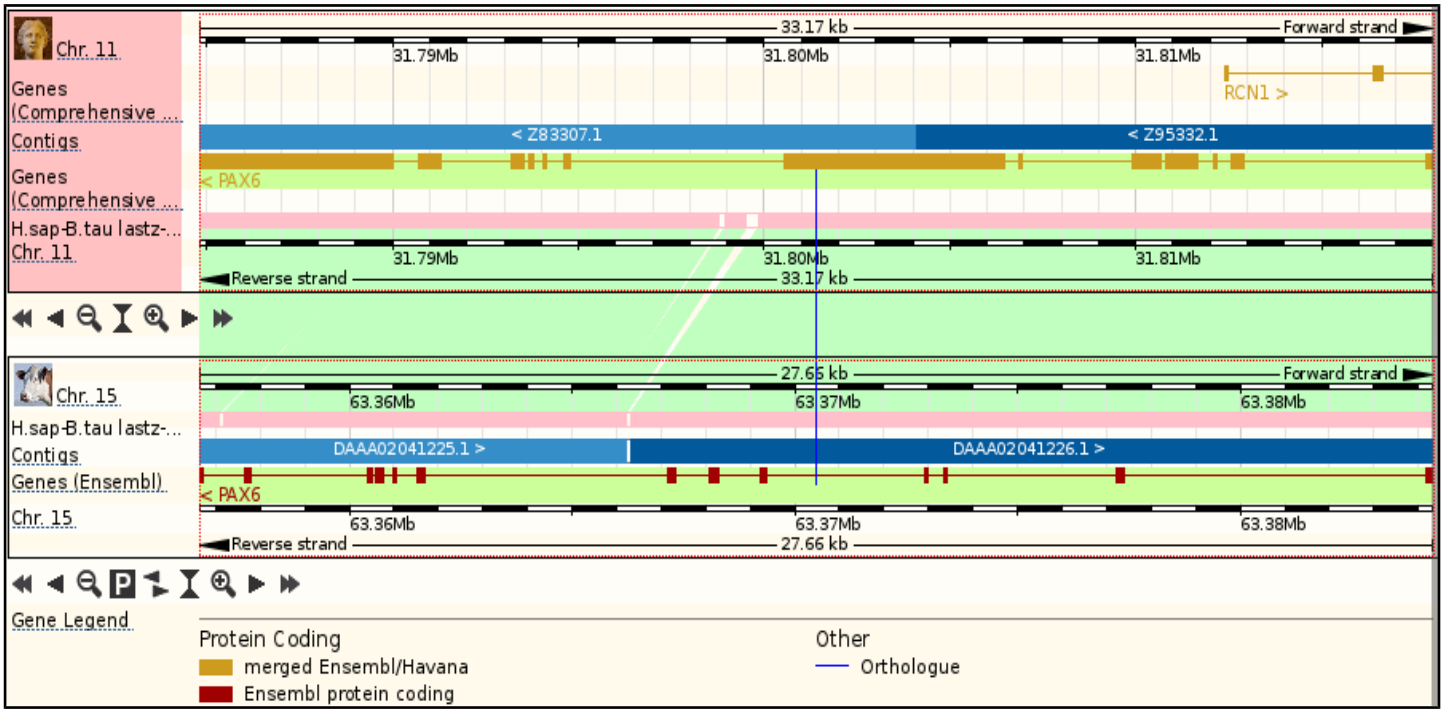
Click on the **Othologues** link in the left hand side of your browser page. Take a look at some of the alignments providing support for the homologous relations. The protein alignments are the more informative.

Using the evidence of the protein alignments, which **PAX6** isoforms do the fruitfly orthologues most resemble? ____

Move back to the othologue list and try one or more of the **Region Comparison** links. I selected the one for **Cow**. Not too far and not too near I mused? In the more general regional comparison, you can still see familiar neighbouring genes such as **ELP4** and **RNC1**. Not so sure I wish to be quite this similar to a cow?



In the more detailed view, the similarities are even more worryingly apparent!



Once your curiosity is completely sated, move back to the othologue list.

Click on the **Paralogues** link. **Paralogues** here should match those reported by **GeneCards** as **GeneCards** obtains its **Paralogues** report from **Ensembl**.

Do the **Paralogue** reports of **GeneCards** and **Ensembl** agree? _____

Try some • Alignment (protein) links to view an alignments between a **PAX6** isoform and its **paralogues**.

Which isoform of **PAX6** has been chosen for the alignments, and why would you suppose it was selected? _____

_____

Which isoform is most common amongst the **paralogues**? _____

Move back to the list of **paralogues**. Try one or more of the **Region Comparison** links. Not nearly so interesting as for the orthologues as the only expectation of similar regions here would be the two gene **paralogues** and not the surrounding region.

Move back to the list of **paralogues**. Go to the **Gene tree** representing proteins potentially related to **PAX6**. This picture gets more outrageously huge each time I look at it! I will trust you get to see it for yourselves. If not, click here. Or, if you are off line, here.

To expand any subtree, click on the horizontal triangles of your choice and select **expand this sub-tree**. To expand the entire tree, click on any of the horizontal triangles , stand well back and select **expand all sub-trees**.

You can download this tree in various textual forms using the button from the menu at the top of the graphic. **Newick** format, accepted by most tree drawing programs, is probably the most important of the many offered.

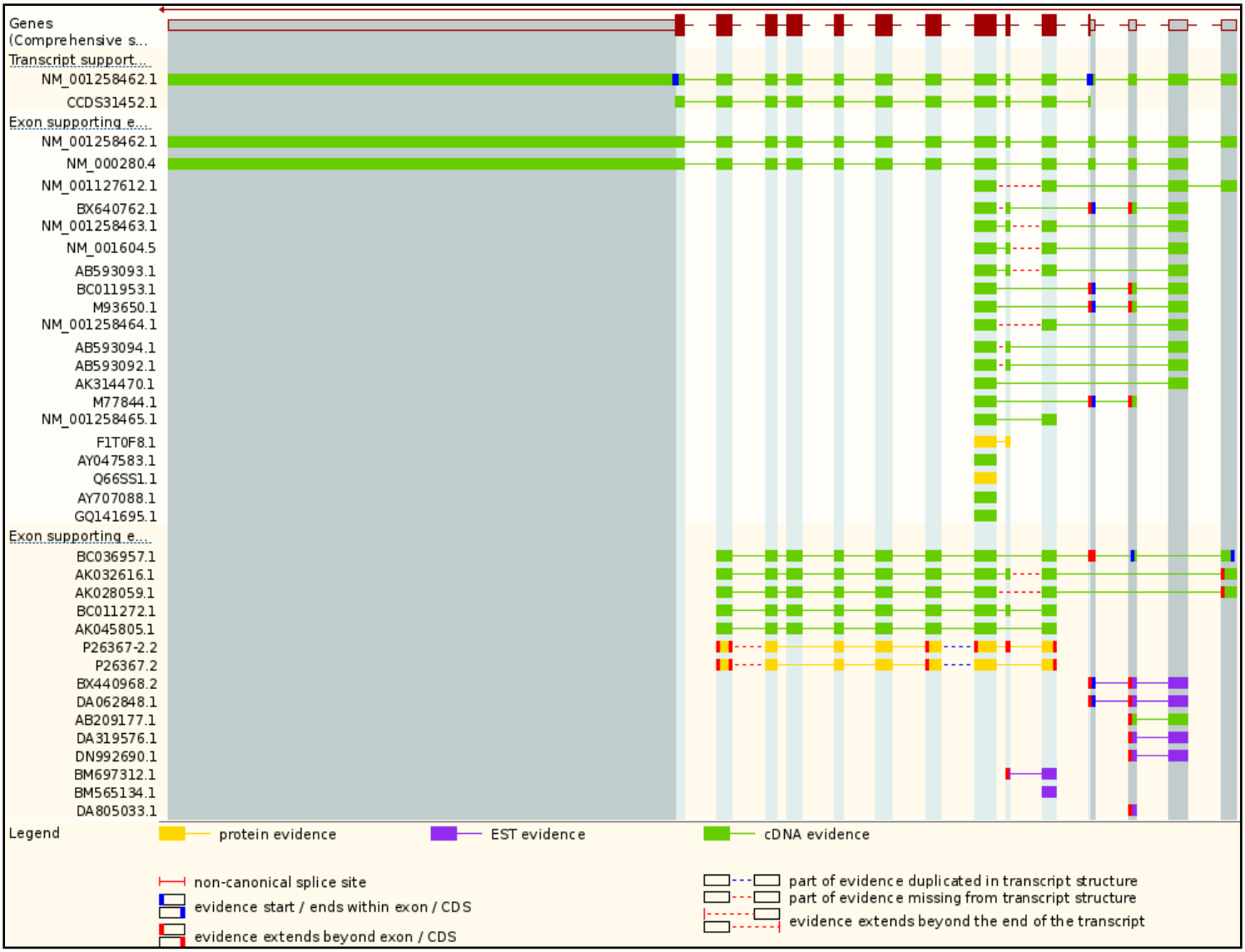
Click on the **Supporting evidence** link in the left hand side of your browser page.

Transcript	CDS support	UTR support	Other transcript support	Exon supporting features
ENST00000241001 [view evidence]	[align] CCDS31451.1			35
ENST00000379107 [view evidence]	[align] CCDS31452.1			35
ENST00000379109 [view evidence]				35
ENST00000379111 [view evidence]	[align] CCDS31451.1			35
ENST00000379115 [view evidence]	[align] CCDS31452.1			35
ENST00000379123 [view evidence]	[align] CCDS31451.1		[align] NM_000280.4	34
ENST00000379129 [view evidence]	[align] CCDS31452.1			37
ENST00000379132 [view evidence]	[align] CCDS31451.1			37
ENST00000419022 [view evidence]	[align] CCDS31452.1	[align] NM_001258462.1		34
ENST00000423822 [view evidence]				30
ENST00000438681 [view evidence]				21
ENST00000455099 [view evidence]				26
ENST00000464174 [view evidence]				8
ENST00000470027 [view evidence]				12
ENST00000471303 [view evidence]				13
ENST00000474783 [view evidence]				2
ENST00000481563 [view evidence]				8
ENST00000494377 [view evidence]				8
ENST00000524853 [view evidence]				30
ENST00000525535 [view evidence]				21
ENST00000527769 [view evidence]				13
ENST00000530373 [view evidence]				8
ENST00000530714 [view evidence]				6
ENST00000531910 [view evidence]				8
ENST00000532175 [view evidence]				21
ENST00000532916 [view evidence]				6
ENST00000533156 [view evidence]				8
ENST00000533333 [view evidence]				29
ENST00000534353 [view evidence]				7
ENST00000534390 [view evidence]				4
ENST00000606377 [view evidence]	[align] CCDS31452.1	[align] NM_001258463.1		34

For each transcript **Ensembl** offers evidence to support the existence of the **CDS**[37], **UTRs**[38] and **Exons**[39]. Take a look at some of the **[align]**ments and sequences[40] providing the support for the **CDS**s and **UTR**s.

37 **Co**ding **S**equence, typically a match with a protein sequence and/or a **CCDS** entry.
38 **UnT**ranslated **R**egions, typically matches with **mRNA** sequences.
39 Matches with protein and/or coding DNA sequences.

Click on the **view evidence** link to view the support for **ENST00000419022**. In the **Transcript supporting evidence** section, sequence matches supporting the entire transcript are recorded. Here there is match with a **RefSeq** mRNA suporting all exons of the transcript[41] and **CCDS** (and sometimes, but rarely these days, protein matches) suggesting which exons are coding. In the **Exon supporting evidence** sections, extra matches with a variety of sequences from different sources are offered to support individual exons of the transcript are illustrated.



Some matching sequences offer no support for the 6th exon from the right. Why do you suppose this is? _____

_____

Why is there no protein or **CCDS** evidence for exons 1, 2 and 3? _____

_____

Until recently (mid-2014), the **RefSeq** mRNA match in the **Transcript evidence** section did not match all exons. Was this logical? _____

_____

40 Those starting **NP_** are **Refseq** proteins. Those starting **NM_** are **Refseq mRNA** sequences. **RefSeq** links will take you to the **NCBI**.

41 This need not necessarily be the case. In fact, until recently (mid-2014) this transcript was supported by a **RefSeq** mRNA that had the **6th** exon missing! Essentially, **Ensembl** does not take the **RefSeq** mRNAs as defining a transcript by themselves. This is reasonable as these mRNA sequences do not necessarily represent a single carefully sequenced mRNA. **RefSeq** mRNAs are typically constructed from many imperfect sequences assumed to be of the same cDNA/mRNA and cannot therefore be assumed to be always **100%** accurate. Oh that life was just a little more straight forward?

Click the **Exons** link (from **Transcript-based displays → Sequence**). **Exons**, **Introns** and **Variations** within **Exons** are clearly displayed[42].

How many exons are there in this transcript? _____

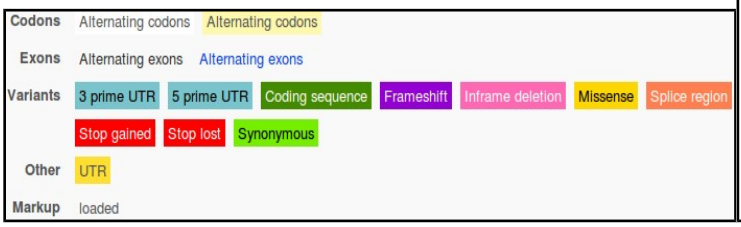
What are the first two bases and what are the last two bases of nearly every intron? _____

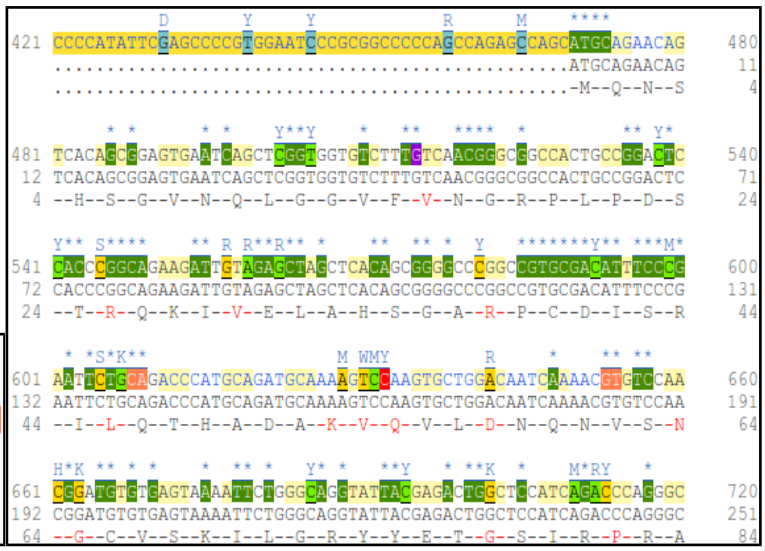
How long is the sixth exon and why would this concur with your expectations? _____

Explain the **Start Phase** and **End Phase** columns? _____

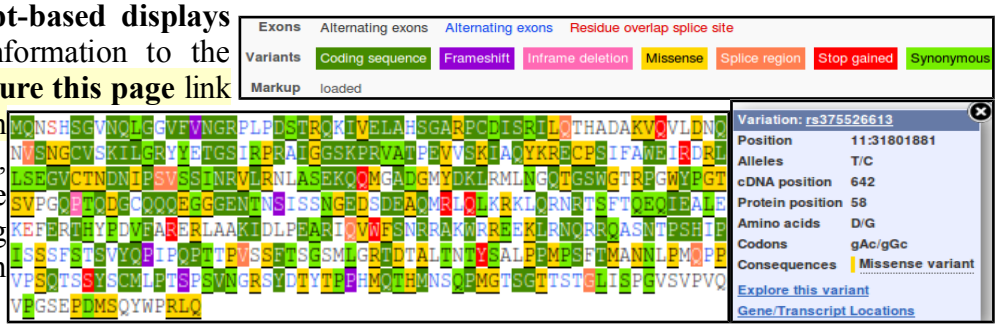
Click on the **cDNA** link (**Transcript-based displays → Sequence**). The default view of the transcript **cDNA** is ornate, exhibiting all the features described in the **Key**.

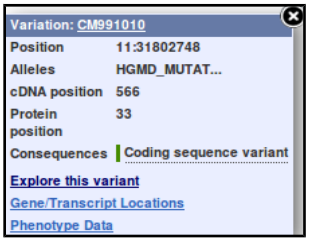
The default view displays most available information. Click on the **Configure this page** link and choose **deselect all**. Click on **Save and close**. Now you see a plain view intended to allow easy export of the sequence by copying/pasting from your browser window.





Click on the **Protein** link (**Transcript-based displays → Sequence**). Add the variation information to the default view by clicking on the **Configure this page** link and ticking **Show variants**. Click on **Save and close**. A number of features, including sequence variations, are shown (details in the **Key**). By clicking on variations, further information can be obtained, try a few.

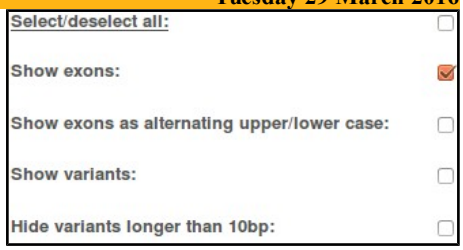




Many of the variations come from the **HGMD_MUTATION** database. Much of this database is not in the public domain. For the commercial section of the **HGMD_MUTATION** database, mutation details are not available to all. They are only revealed to people who pay for the database. Here, they just make a mess of what would otherwise be a useful display.

42 Exactly which variants are displayed can be customised using the **Configure this page** facility, try it if you have time. The types of variation can be determined from the **Key**, a version of which is illustrated above (with the **cDNA** view).

Click again on the **Configure this page** link and do whatever it takes to select just the **Show exons** option. Click on **Save and close**.




```
MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQ
NVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRL
LSEGVCTNDNIPSVSSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGSWGTRPGWYPGT
SVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALE
KEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIP
ISSSFSTSVYQPIPQPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPP
VPSQTSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMNSQPMGTSGTTSTGLISPGVSVPVQ
VPGSEPDMSQYWPRLQ
```

This view should be plain but showing clearly where all exons start and end.

Where is the start and end of the **Prosite Paired Box** pattern (**R-P-C-x(11)-C-V-S**)? _____

Where, in relation to the pattern, are the extra **isoform 5a** amino acids? _____

Why might the positions of these two features be significant? _____

Click on the **Domains & features** link (from **Transcript-based displays → Protein Information**). Look at the domains of **Domain type Smart**[43]. There are two. Predictably, a **Paired box** and a **Homeobox** domain.

Domain source	Start	End	Description	Accession	InterPro
PANTHER	1	434	-	PTHR24329	-
PANTHER	1	434	-	PTHR24329:SF294	-
Gene3D	7	86	-	1.10.10.10	-
Gene3D	87	150	-	1.10.10.10	-
Gene3D	201	284	-	1.10.10.60	-
Prosite_profiles	222	282	Homeobox domain	PS50071	IPR001356 [Display all genes with this domain]
Smart	224	286	Homeobox domain	SM00389	IPR001356 [Display all genes with this domain]
Pfam	226	281	Homeobox domain	PF00046	IPR001356 [Display all genes with this domain]
Prosite_patterns	257	280	Homeobox, conserved site	PS00027	IPR017970 [Display all genes with this domain]
Superfamily	6	143	Homeodomain-like	SSF46689	IPR009057 [Display all genes with this domain]
Superfamily	205	283	Homeodomain-like	SSF46689	IPR009057 [Display all genes with this domain]
Pfam	4	142	Paired domain	PF00292	IPR001523 [Display all genes with this domain]
Smart	4	142	Paired domain	SM00351	IPR001523 [Display all genes with this domain]
Prosite_profiles	4	144	Paired domain	PS51057	IPR001523 [Display all genes with this domain]
PRINTS	8	23	Paired domain	PR00027	IPR001523 [Display all genes with this domain]
PRINTS	26	44	Paired domain	PR00027	IPR001523 [Display all genes with this domain]
PRINTS	60	77	Paired domain	PR00027	IPR001523 [Display all genes with this domain]
PRINTS	78	95	Paired domain	PR00027	IPR001523 [Display all genes with this domain]

Where do each of the **SMART** domains start and end?_____

Do the regions match your earlier recording? If not, why not?_____

What are the **Interpro**[44] database accession codes for the two major **PAX6** domains?_____

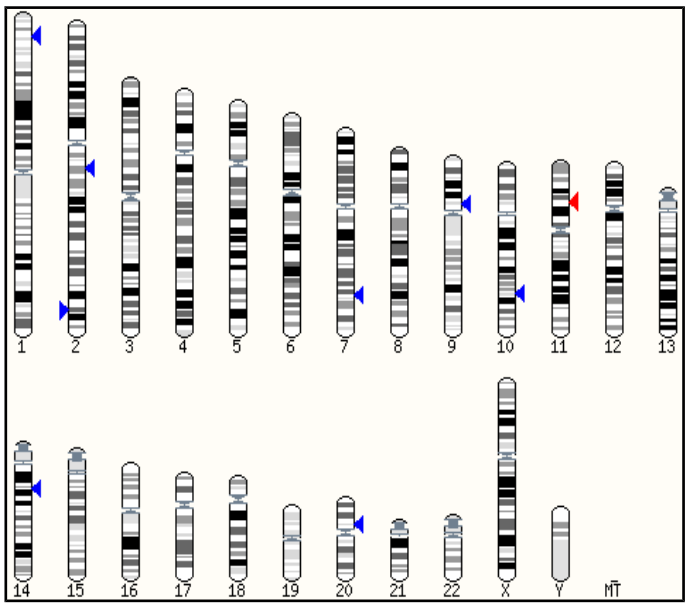
Why does **Prints** appear to predict four **Paired_domains**?_____

43 **SMART** (a **S**imple **M**odular **A**rchitecture **R**esearch **T**ool) allows the identification and annotation of genetically mobile domains and the analysis of domain architectures. More than 500 domain families found in signalling, extracellular and chromatin-associated proteins are detectable. These domains are extensively annotated with respect to phyletic distributions, functional class, tertiary structures and functionally important residues. Each domain found in a non-redundant protein database as well as search parameters and taxonomic information are stored in a relational database system. User interfaces to this database allow searches for proteins containing specific combinations of domains in defined taxa.

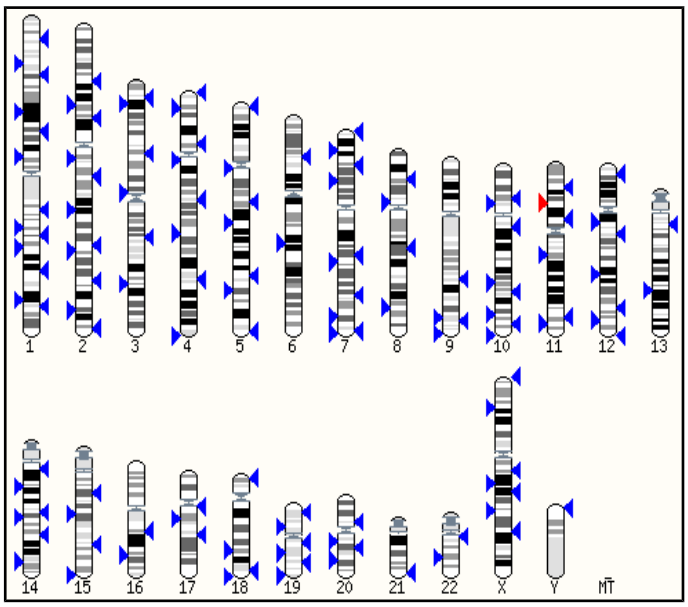
44 **InterPro** is a database of protein families maintained at the **EBI**.

Click on **Display all genes with this domain** for the **Paired domaim** and **Homeobox domain InterPro** families. The locations of all genes including each domain will be displayed graphically and textually. **PAX6** is shown in red.

Paired domain - IPR001523



Homeobox domain - IPR001356



Which domain, **Paired domain** or **Homeobox domain** is more common in humans? _____

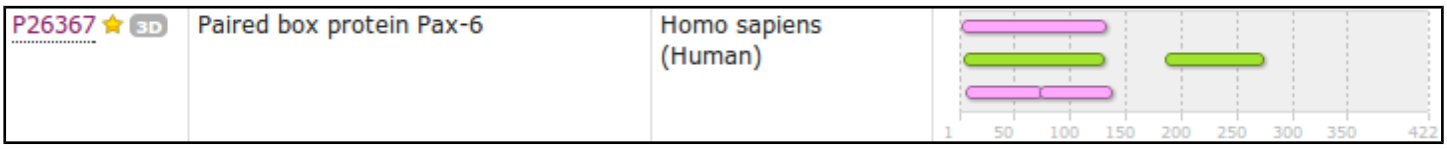
How many human **PAX** genes are there? _____

Are all the **PAX** genes on **Chromosome 11**? _____

How does **Interpro** match with the **PAX6 Paralogues** reported by **Ensembl/GeneCards** earlier? _____

_____

Move back to the **Domains & features** display. Link to the **InterPro** database entry for **Paired domain**. Here you will find much confirmation of that which we have already discovered. Click on the **Proteins matched** link. You will see listed a number of representations of proteins that, according to **InterPro**, include a **Paired domain**. Amongst these will be **P26367**[45]. There are links provided to entries in a number of relevant databases for each listed protein.

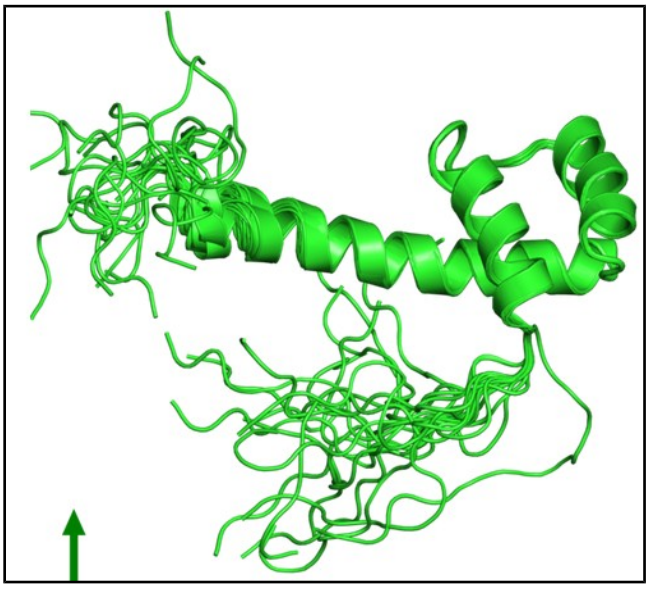


45 Third from the bottom of the first page, last time I counted.

Click on the **Structures** link in the top left hand corner of the page. **InterPro** will offer links to relevant entries in the **PDBe**, **SCOP** and **CATH**[46] databases. Click on the link to the **6pax** entry in the **PDBe** database. You will arrive at the entry for **6pax** in **PDBe**, the European version of **PDB** maintained at the **EBI**. Views of this structure are offered on the right hand side of the page. Click on the largest image which shows the paired box protein domain binding DNA rather beautifully. Once you have admired this image sufficiently, move back to the **6PAX PDBe** entry. From the Quick links on the right of the page, select the **3D Visualisation** option.

The **Prosite** documentation you read earlier suggested two paired box subdomains, each of which " ... form a three-helical fold, with the most C-terminal helices comprising a **helix-turn-helix** (**HTH**) motif that binds the **DNA major groove**". Move your image around to confirm this assertion.

The same **Prosite** documentation claims the subdomain nearer the N terminal " ... encompasses an N-terminal **beta-turn** and **beta-hairpin**, also named '**wing**', participating in DNA-binding. The linker can bind into the **DNA minor groove**". Manipulate you image to investigate the veracity of these assertions.



Once you have seen all there is to see of **6PAX**, move back to the **Ensembl Domains & features** display. Try the same tricks with the **InterPro Homeobox domain**. This time, it is difficult to find **P26367** in the huge list[47] **Proteins matched**, but you do not need to in order to link to the **Stuctures**. There are many more structures to choose from this time. I suggest you go for **2cue**. You have to imagine the DNA this time.

Can you explain the strangely frayed ends displayed in some of the representations of the **2cue** 3D structure? _____

_____

Further Features of Ensembl

46 **PDB** is the main database of **3D** protein structures. **SCOP** and **CATH** are also **3D** structure databases, we will very superficially visit both later.

47 If you really wanted to, the best approach is to search for **P26367** in the search box at the top of the page and then look for the **Homeobox domain** entry in the **Detailed signature matches** list.

Mining Ensembl with Biomart

Used from the **Ensembl** pages, **Biomart** allows the intuitive construction of queries of the entire **Ensembl** database. **Biomart** can be used for various purposes, some of which will be investigated here. The first objective is to use **BioMart** to identify human genes that share specific properties with **PAX6**.

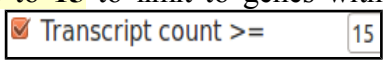
Return to the **Ensembl** homepage (**http://www.ensembl.org**). Click on the **BioMart** link in the banner at the top of the page.

The Start page will enquire which dataset and organism you wish to search. Set **CHOOSE DATABASE** to the latest **Ensembl Gene** build (**Ensembl Genes 84**, as I type). Select **Homo sapiens genes** from the **CHOSE DATASET** menu..

Click on the Filters link to specify the genes of interest. On this page, enter the following data:

REGION: Click on the '+' to view the options, but select nothing to search the entire genome. Click on the '-' to tidy the **REGION** options away.

GENE: Click on the '+'. Check **Transcript count >=** and set the value to **15** to limit to genes with many transcripts only. A trifle arbitrary, but never mind. Click on the '-' to tidy the **GENE** options away.

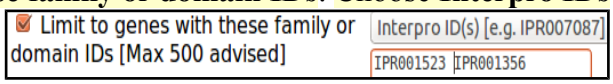


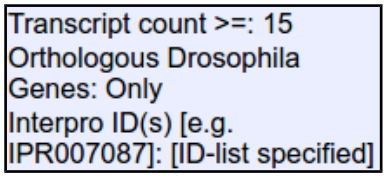
MULTI SPECIES COMPARISONS:

Click on the '+'. Check **Homologue filters** and set the value to **Orthologous Drosophila Genes** to limit matches to only genes with fly orthologues. Click on the '-' to tidy the **MULTI SPECIES COMPARISONS** options away.

PROTEIN DOMAINS AND FAMILIES:

In this section, turn on **Limit to genes with these family or domain IDs**. Choose **Interpro IDs** as the means of selection. Enter the two **Interpro ID(s)** of the **PAX6** gene that you noted earlier[48] in a **comma** or **space** separated list. This translates to all proteins with either a **paired box** or a **homeobox** domain, or both.

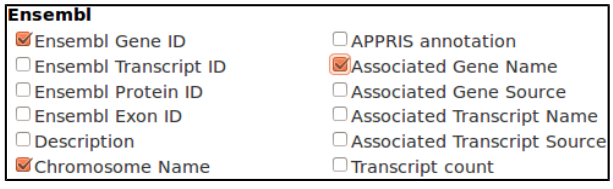




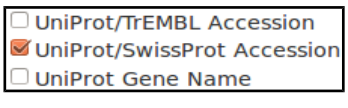
On the left of your page, **BioMart** will have summarised the search you have requested in its Filters section. Given the search criteria match your expectations, click on the **Count** button. You should see that your filtering has selected around **13** genes from the **66,000** or so to which **Ensembl** admits existence.

Now click on the Attributes button to specify what properties you would like to retrieve for the genes you have selected. Enter the following:

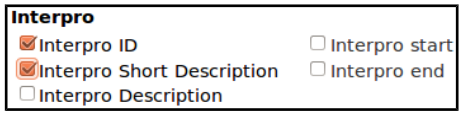
GENE:
Check **Ensembl Gene ID** (checked by default)
Turn off **Ensembl Transcipt ID** (they are far too copious!)
Check **Chromosome Name**
Check **Associated Gene Name**



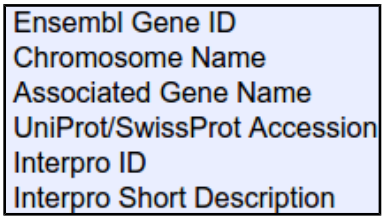
EXTERNAL: Check **UniProt/SwissProt Accession**



PROTEIN DOMAINS AND FAMILIES:
Check **Interpro ID**
Check **Interpro Short Description**



48 OK … just in case you have forgotten! **IPR001523** and **IPR001356**.



On the left of your page, **BioMart** summarises the search you have requested in its Attributes section. Given the selected attributes match your expectations, click on the Results button. By default, only **10** filtered entries are displayed. Note that some appear to be the same!! They are not however. Some hits differ in features that you have not chosen to display and so give rise to apparent duplications. This can be eliminated by checking **Unique results only**, try it.

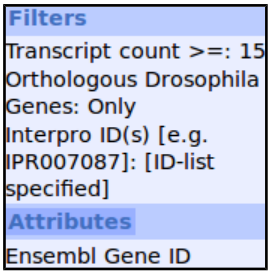
In order to see all your hits, elect to lengthen your results list to **200 rows**. Your list is now complete but will include more entries than the gene count **Ensembl** declared to match your search criteria. Look closely and you will see that only the predicted number of genes occur in the **Ensembl Gene Name** column, but some occur more than once.

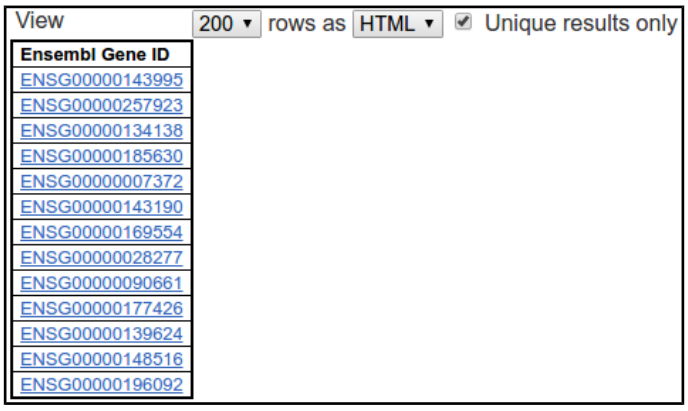
Duplication occurs when an **Ensembl Gene** matches differing combinations of properties that have been selected for display. For example, the **PAX6** protein (**Gene ID**: **ENSG00000007372**) will match both the chosen **Interpro** families[49] resulting in **2** list entries. The same gene also corresponds to proteins uniquely from **Uniprot/TrEMBL**, resulting in list entries with blank **UniProt/SwissProt Accession** entries[50]. So there will be at least **3 ENSG00000007372** entries in the list.

Ensembl Gene ID	Chromosome Name	Associated Gene Name	UniProt/SwissProt Accession	Interpro ID	Interpro Short Description
ENSG00000143995	2	MEIS1		IPR001356	Homeobox_dom
ENSG00000143995	2	MEIS1	O00470	IPR001356	Homeobox_dom
ENSG00000257923	7	CUX1	P39880	IPR001356	Homeobox_dom
ENSG00000134138	15	MEIS2	O14770	IPR001356	Homeobox_dom
ENSG00000134138	15	MEIS2		IPR001356	Homeobox_dom
ENSG00000185630	1	PBX1	P40424	IPR001356	Homeobox_dom
ENSG00000185630	1	PBX1		IPR001356	Homeobox_dom
ENSG00000007372	11	PAX6	P26367	IPR001356	Homeobox_dom
ENSG00000143190	1	POU2F1	P14859	IPR001356	Homeobox_dom
ENSG00000143190	1	POU2F1		IPR001356	Homeobox_dom
ENSG00000169554	2	ZEB2	O60315	IPR001356	Homeobox_dom
ENSG00000169554	2	ZEB2		IPR001356	Homeobox_dom
ENSG00000028277	19	POU2F2	P09086	IPR001356	Homeobox_dom
ENSG00000028277	19	POU2F2		IPR001356	Homeobox_dom
ENSG00000090661	19	CERS4		IPR001356	Homeobox_dom
ENSG00000090661	19	CERS4	Q9HA82	IPR001356	Homeobox_dom
ENSG00000177426	18	TGIF1		IPR001356	Homeobox_dom
ENSG00000177426	18	TGIF1	Q15583	IPR001356	Homeobox_dom
ENSG00000139624	12	CERS5		IPR001356	Homeobox_dom
ENSG00000139624	12	CERS5	Q8N5B7	IPR001356	Homeobox_dom
ENSG00000148516	10	ZEB1	P37275	IPR001356	Homeobox_dom
ENSG00000007372	11	PAX6	P26367	IPR001523	Paired_dom
ENSG00000007372	11	PAX6		IPR001523	Paired_dom
ENSG00000196092	9	PAX5	Q02548	IPR001523	Paired_dom
ENSG00000196092	9	PAX5		IPR001523	Paired_dom

You have identified and displayed information concerning a number of genes, each of which has at least **15** transcripts, includes either a **paired box** domain or a **homeobox** domain (or both) and has **orthologues** in the **Drosophila** genome. It is probably best to suppress the burning question "Why?".

Finally, for this search, move once more to the Attributes and turn off everything except the **Ensembl Gene ID**.





Click on the Results button once more. Set the **View** of **200 rows as CSV**[51]. Set **Unique results only**, although there will be no duplications this time, of course.

In the **Export all results to** section, set the format to **CSV**, choose **Unique results only** and then click Go.

Do whatever it takes to end up with a file on your **Desktop** called:

Ensembl_Genes_List.txt

containing your gene list.

49 As it contains both a **Paired Box** and a **Homeobox**. The list duplication could be "solved" by not asking for the **Interpro IDs** or **Interpro Short Descriptions** to be displayed. But … enough!

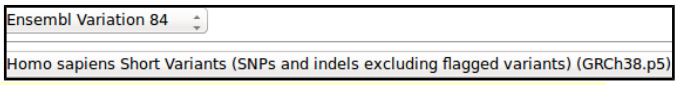
50 Where the **UniProt/SwissProt Accession** column is blank, the existence of one or more independent **UniProt/trEMBL** entries exist. You could have included these by asking for the **UniProt/TrEMBL Acession** codes to be displayed, but then the list would be very much longer.

51 **C**omma **S**eparated **V**alues

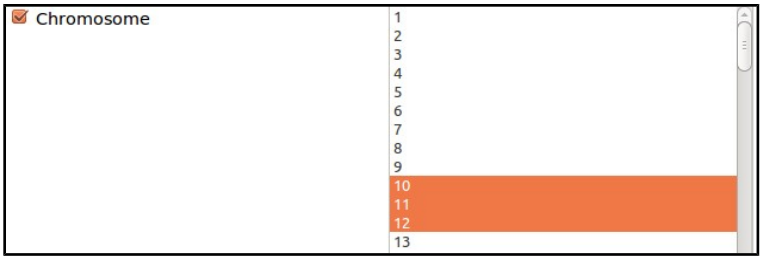
Now try a search of the variations section of the database. The object this time is to list all the **splice acceptor** variations associated with any of the genes listed in the file you have just made from specified chromosomes. A very simple search to set up as you need only specify the file of **Ensembl Gene Id**s, the type of variation that is of interest and the chromosomes of interest.

Start afresh by clicking on the New button in the top left hand corner of your page.

Set **CHOOSE DATABASE** to the latest **Ensembl variation** release (**Ensembl Variation 84** at the time of typing)



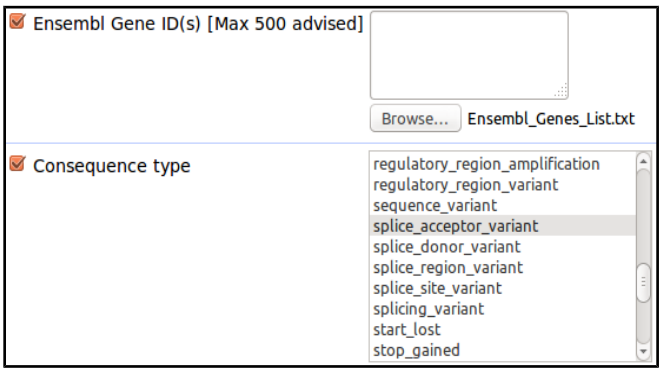
Set **CHOOSE DATASET** to **Homo sapiens Short Variation (SNPs and indels excluding flagging variants)**.



Click on the Filters link.

Open the **REGION** section.

Check **Chromosome** and select just chromosomes **10**, **11** and **12**.



Open **GENE ASSOCIATED VARIATION FILTERS**.

Check **Ensembl Gene ID(s)**
Browse for the file `Ensembl_Genes_List.txt`

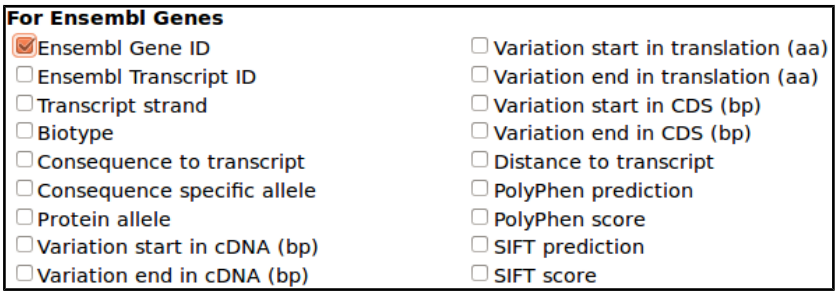
Check **Variant consequence**
Select **splice_acceptor_variant**

Click **Count**. You should have around **5** hits.

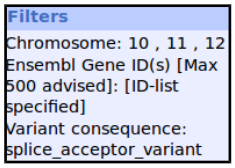
Click on the Attributes link.

Open **GENE ASSOCIATED INFORMATION**.

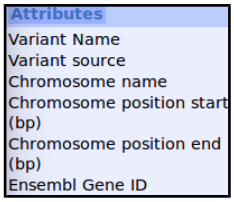
Select **Ensembl Gene ID**



Given your



and



are as you expect, click the Results button.

Set the **View** to **200** rows, **HTML**, **Unique results only**.

Your list represents all the variants available from **Ensembl** that affect the **splice acceptor** sites of the genes that you discovered in your previous **BioMart** search that occur on Chromosome **10**, **11** or **12**.

Variant Name	Variant source	Chromosome name	Chromosome position start (bp)	Chromosome position end (bp)	Ensembl Gene ID
rs747921459	dbSNP	12	50167244	50167243	ENSG00000139624
rs778443322	dbSNP	12	50130645	50130644	ENSG00000139624
rs779328216	dbSNP	10	31527082	31527081	ENSG00000148516
TMP_ESP_10_31816010_31816012	ESP	10	31527082	31527084	ENSG00000148516
rs772066684	dbSNP	10	31527097	31527096	ENSG00000148516

Only **3** of the genes listed in the file you created matched the location filters. One in each of the chosen chromosomes. Including **PAX6** (in chromosome **11**). No suitable variations found for **PAX6** here though. A pity, but never mind.

Lastly, to demonstrate the extraction of sequences from **Ensembl**. You have already made a file containing the genomic sequence of the gene **PAX6** for further analysis. This was easily achieved using the **Ensembl** browser interface. However, had you wanted the sequences of all the genes associated with a **homeobox** domain (not far short of **300**) using the browser method would be quite intolerable. With **Biomart**, such an enterprise is trivial, as we will now demonstrate.

Once again, start afresh. Click on the New button at the top of the page.

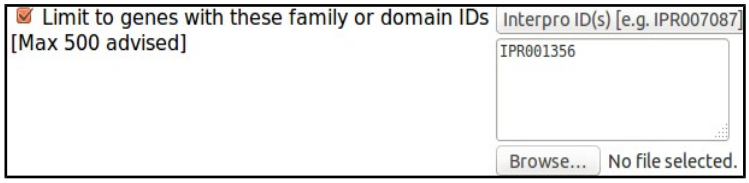
Set **CHOOSE DATABASE** to the latest **Ensembl Gene** build (**Ensembl Genes 84**, as I type). Select **Homo sapiens genes** from the **CHOSE DATASET** menu.

Click on the Filters link to specify the genes of interest.

Open **PROTEIN DOMAINS AND FAMILIES**.

Check **Limit to genes with these family or domain Ids**.

Choose **Interpro IDs** as the means of selection.

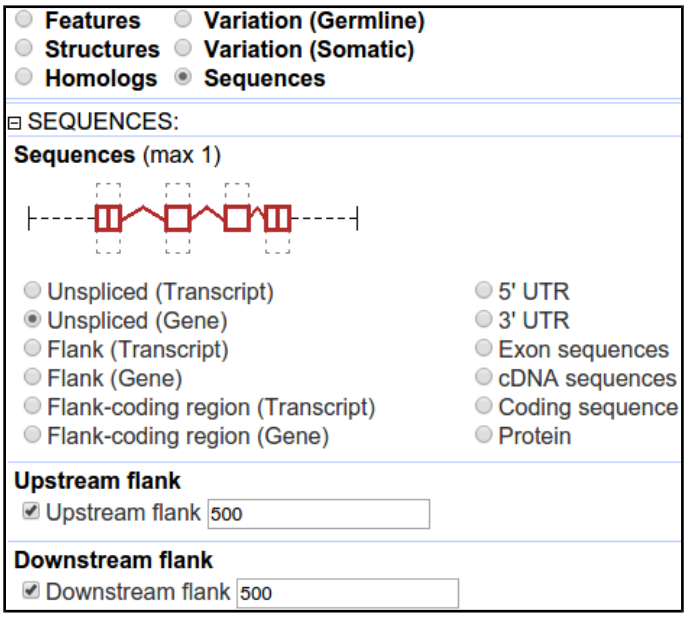


Enter the **Interpro ID** for **homeobox** domains (**IPR001356**).

Click **Count**. You should have around **250** hits.

Move to the **Attributes** section and select **Sequences**.

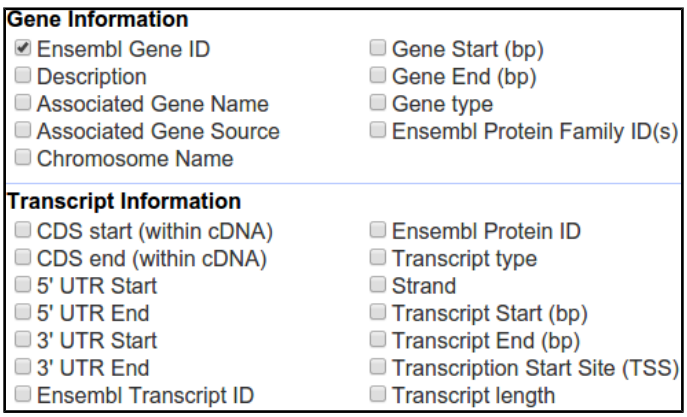
Open the **SEQUENCES** section.
ask for **Unspliced(Gene)**,
with **500** base pairs of **Upstream flank**,
and **500** bases of **Downstream flank**.



That is, all the relevant sequence with an extra bit of on each end.

Now move to the **Header Information** section. Uncheck **Ensembl Transcript ID**.

To include the identifiers of all the transcripts that **Ensembl** predicts for all these genes is far to unwieldy.



Click on the **Results** button

The first **10** 0f the requested sequences will appear in **Fasta** format with just the **Ensembl Gene ID** in the identification line.

```
>ENSG00000004848
CCGACGCCCGTCCGGCCTAGAGAGGACGGCAATTGAGCGCCAGGGAGGTCAAGGCGGTGT
CCTGGGAGCCCAATATCGGCTTTGAAAAGTGCCGAGTGAGTCTGAACGTGAGCCTCGAGG
CCTCGCCTCCAGGGAGCGAGGTCCTCGGGTTGGTGGAAGGGGCGCACGCCACCCAGATGG
CATTCACAGGCCTGGCATCCCAATAAGCTAGAACTTCGGCCAACACTAAAGGGTCAAAGG
AGGAGCTGCAGGAAGAGGATAGCGGACTTAGAAAATTGGTAACTTAAAAAAAGAAAGAAA
GAAAAGGAAAATTGGCTCTGGTGCGTGCCCGCTGCCCCCACCCCCGCCTGCCCCTCTGGA
ATCCAGTCCGGGCTTTGCGCCGCGCCCACAGGCCGACGCAGCCCGGCCTCTGGCGAGAGC
CAATCAGAGGGCGCCTCTCAGCACGTGGAGGAGAGAGACTCCAGAGCTCAGCGCCCGCTG
CTCACTACACTTGTTACCGCTTGTCCTGAGCGCGGAGAGGGCGAGCTCGGGCCGCGGGCA
GGGCGGGAGCCGGCAGCCGGCAACCAAGGGAGGCAGAAAGGCACAAAGATCGCAATAATA
TCCGTTATAACCCGCTATCTAACCCCACCCCCAACACACACCCATCCATCCCACCCTCCG
GGAGAGGCAGCCGGCGATCCGCTCTCTGCGCCCTGGGAAAAAGCCCCAGCCATGAGCAAT
CAGTACCAGGAGGAGGGCTGCTCCGAGAGGCCCGAGTGCAAAAGTAAATCTCCAACTTTG
CTCTCCTCCTACTGCATCGACAGCATCCTGGGCCGGAGGAGCCCGTGCAAAATGCGGTTG
CTGGGAGCCGCGCAGAGCTTGCCTGCTCCGCTGACCAGCCGCGCCGACCCGGAAAAGGCC
GTGCAAGGTAAGGATGCTCCCGTCAGGCACTTACTAAGGGCATTGGGCCCTGATTTGGAT
GTTTGGTGTTCGGGGGCCAGTGGCCTGGAATTGTCAATTTGGAGAGGAAGGAAGGAGAGG
```

In the **Export all results to** section, keep the default option of creating a **File** in **FASTA** format. Choose **Unique results only** and then click Go.

Do whatever it takes to end up with your **homeobox** sequences in a file on your **Desktop** called:

`human_genetic_hox.fasta`

Sequence Analysis

The overall target remains to discover all we can about **Aniridia** and its genetic causes. Here we will see how information can be derived from analysis of sequence data directly. Many of the analysis tools and methods we will use and discuss will be those that were used to generate the "ready made" answers you have already investigated. The conclusions of this section should not therefore contradict those you have already reached.

To start the analysis section, you should have data files including:

Two PCR primers associated with the **PAX6** gene in a way we have not yet fully determined, in a file called:

pax6_primers.fasta

The sequence of the Paired box domain of the most prolific human **PAX6** isoform in a file called:

pax_domain.fasta

The genomic sequence of the whole of the **PAX6** gene, plus an extra 500 bases in each direction, in a file called

pax6_genomic.fasta

The canonical human **PAX6** isoform in a file called:

pax6_human.fasta

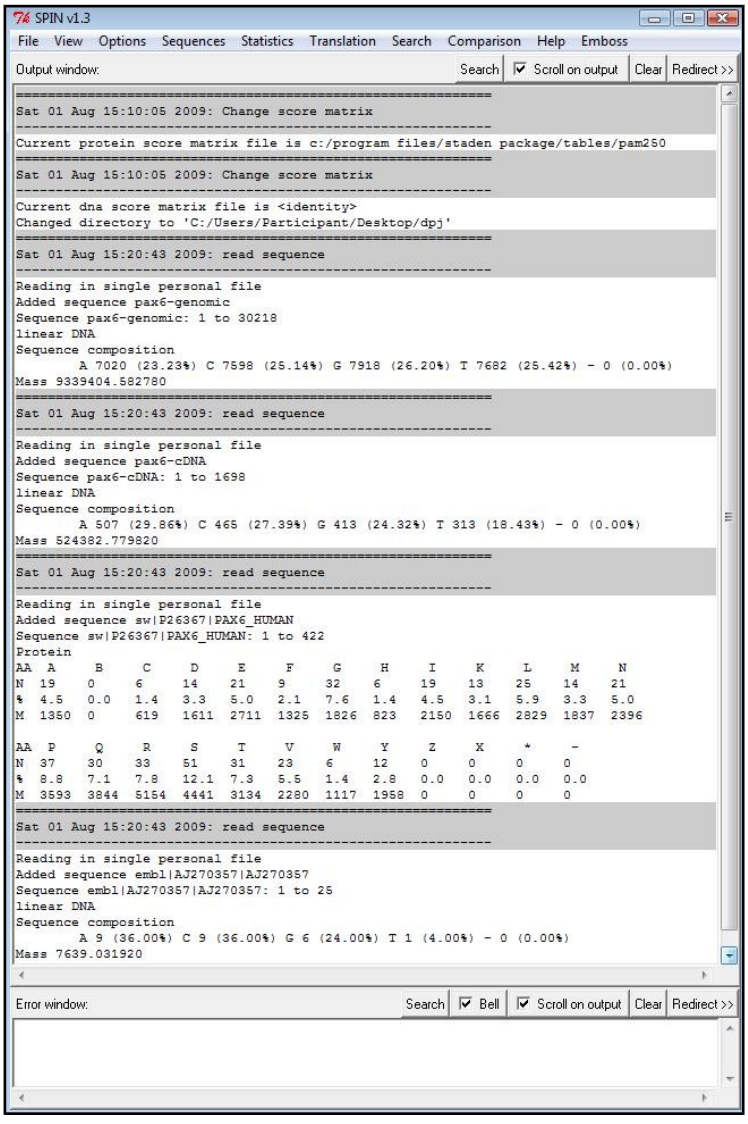
The file **pax6_genomic.fasta** contains the sequence of **PAX6** from a person(s) who, it can be assumed, does not suffer from **Aniridia**. Given a sequence from a person who **does** suffer from **Aniridia**, an obvious first step of investigation might be to compare the sequence from the affected source with the wild type. It would be reasonable to speculate that differences might explain the disease. Here we have to cheat a bit. I have provided an mRNA sequence we will suppose you sequenced from an **Aniridia** patient. It is stored in a file called:

pax6_cdna.fasta

which you will find in a directory called **Working Directory** on your **Desktop**.

Internet resources will be employed for many analyses. Software that runs under windows on your workstation will be used very occasionally. All windows software has been installed, without **Administrator** privileges, using an Installer that will be made available to you.

The **EMBOSS** package[52], using a program called **spin**[53] to provide a **G**raphical **U**ser **I**nterface (**GUI**) will be used for most of the workstation analyses. This is not ideal, particularly as support for **spin** is defunct, but the options[54] are few.



Start **Spin** in the usual windows way (it is part of the **The Staden Package**, which is in the **Start menu** under **Sequence Analysis Tools Installer**). Proceed by setting **spin** to manage its files somewhere sensible. Strangely, its default inclination is to create user output files in its installation directory? This is silly, so select **Change directory** from the **File** pull down menu and select a more sensible default directory. I suggest you might make a directory in your desktop named as you please for this purpose[55].

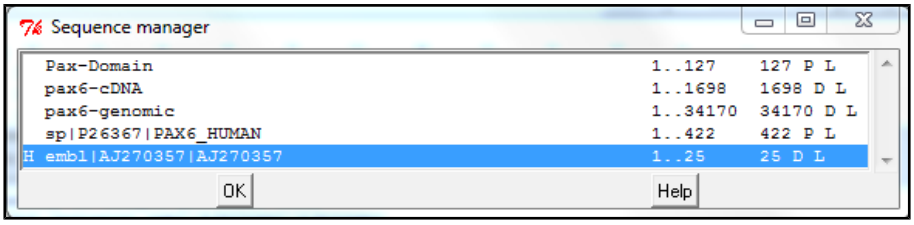
52 Documentation pages for **EMBOSS** can be accessed at the **EMBOSS** home page at: **http://emboss.sourceforge.net**

53 A part of the **Staden Package**.

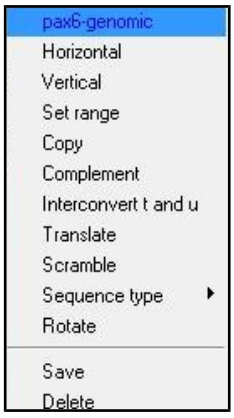
54 That is **FREE** options!

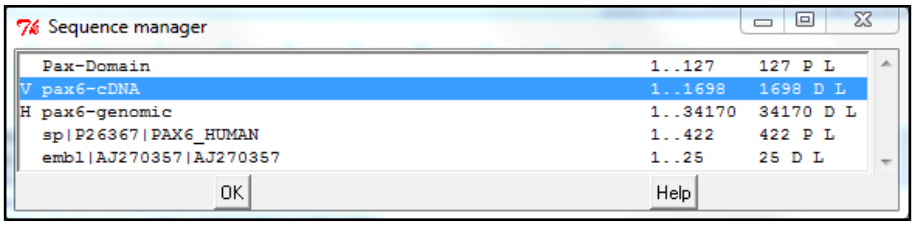
55 It would be very sensible to move **pax6_cdna.fasta** and all the files you have created thus far (**pax6_genomic.fasta**, **pax6_primers.fasta**, **pax_domain.fasta** and **pax6_human.fasta** in particular) into this directory.

Begin by loading all the sequence files you have gathered thus far. From the **File** drop down menu select **Load Sequence** and then **Simple**. This will present a new dialogue box for you to specify a sequence to be loaded. Leave the choice of **Database** as the default **Personal File**, as the all the sequences needed initially are stored in local files. To the right a text box awaits the file path for a sequence file. Use the **Browse** button to find and enter the directory you set up as your default[56]. Select all the sequence files listed above[57] and click **Open** and then **OK** in the **Load Sequence** window. From this point onwards the package will refer to these sequences by their **fasta** titles (e.g. **pax6-cDNA** and **pax6-genomic)**[58].



Take some time to become acquainted with the **spin Sequence manager**. This provides access to all loaded sequences. Its main purpose is to enable sequences to be selected for analysis. From the **File** drop down menu, select **Sequence Manager**. A window will appear listing the sequences showing their start and end, length and type (**D** for DNA or **P** for Protein). The last sequence to be loaded will be labelled **H** (for **H**orizontal)[59]. The sequence set to **H**orizontal is the one that **spin** will analyse by default. Ensure that **pax6-genomic** is set to **H**orizontal by right clicking its name and selecting **Horizontal**. Right clicking an entry brings forth a menu offering several more possibilities than we will use in these exercises. For example, it is possible to **Set range**, that is specify regions of entries to be analysed separately. Also it is possible to **Translate** or **Compliment** DNA sequences to create new entities for analysis.





In similar fashion, make **pax6-cDNA** the **V** (for **V**ertical) sequence[60].

Leave the **Sequence manager** by clicking **OK**. You are now be ready to proceed.

56 If you did things correctly, you should be in the right place without having to move anywhere.

57 You can select multiple files in a number of ways including holding the **Ctrl** key down whilst clicking on each of the files you require.

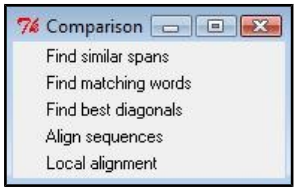
58 **spin** will describe all the sequences it has loaded. Note that only the first of the primer sequences was loaded, sadly **spin** does not recognise that **pax6_primers.fasta** is a multiple sequence file. This will not matter to us as the exercises stand at present.

59 This refers, somewhat bizarrely, to the placement of the sequence on Cartesian axes.

60 This only really makes complete sense when you are computing dotplots, which you are just about to do. **spin** is limited in that it allows a maximum of 2 sequences to be selected. When only one sequence is required, the **H**orizontal selection is used.

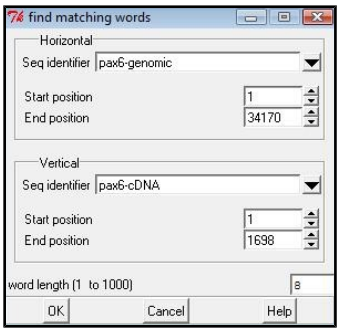
Pairwise Sequence Comparison

You will be using options from **spin**'s **Comparison** menu for the next few steps. To set up that menu for easy access, click on **spin**'s **Comparison** button and then on the dotted line of the menu that will appear. Position the menu window that "tears off" conveniently.

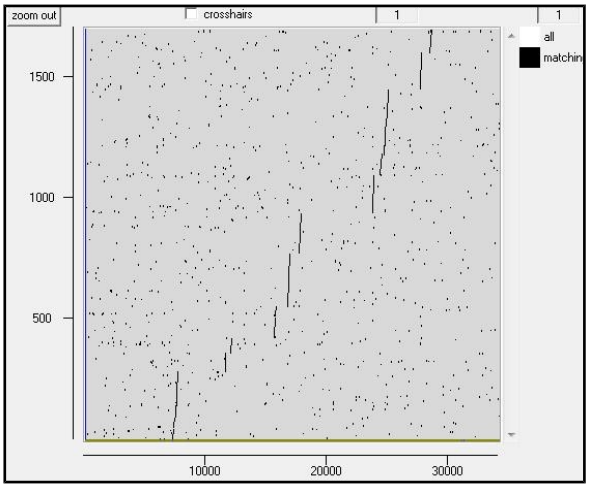


Graphical pairwise sequence comparison - DotPlots

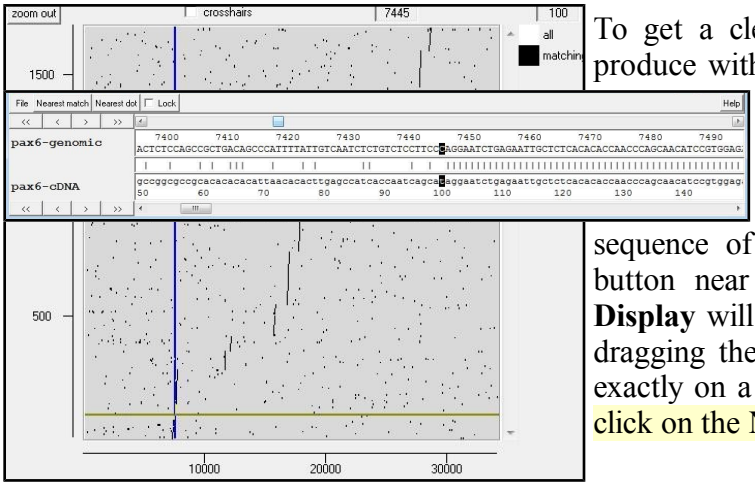
An intuitive graphical representation of the comparison between two sequences is a **dotplot**. One sequence is represented on each of two Cartesian axes and significant matching regions are indicated as a plot. Further detail will be provided by presentation. These plots are used to indicate which sections of sequence might align with each other convincingly and thus warrant further investigation.



From your **Comparison** menu, select **Find matching words**, which offers a simple and fast way to draw a dot plot. Check that you have the sequences you expect in the **Horizontal** and **Vertical** positions[61]. Accept the default **word length** of **8**[62] and click **OK**. A **dotplot** will burst forth with the genomic sequence along the bottom axis, and the cDNA sequence on the vertical axis, as illustrated. As far as the resolution of the picture allows one to judge, there appears to be **11** matching regions.

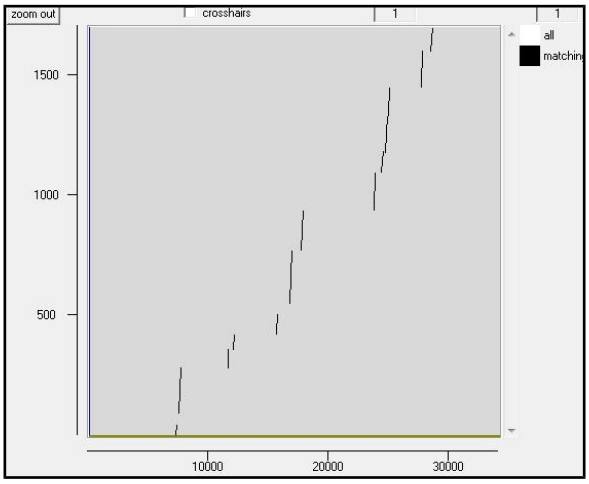


What do you suppose these regions represent?_____



To get a clearer view of the features of any of the graphics you produce with **spin**, you can always make the picture window bigger, use the **X** and **Y** controls at the bottom to reshape the view, hold the **Ctrl** key down and select a region with your right mouse button to **zoom in** to a feature[63]. Try any or all of these possibilities. You can also view the sequence of any feature by double clicking your left hand mouse button near that feature. Try this and a **Sequence Comparison Display** will appear. You can control the sequence region in view by dragging the cross hairs that will appear around. To fix the display exactly on a given feature, get as near as you can manually and then click on the **Nearest match** button.

Using a **word length** as small as **8** shows the meaningfully matching regions quite well in this case, but there is quite a bit of unnecessary "noise"[64]. The features we wish to be shown are long, so it is possible to use a bigger word size without losing matches of importance.

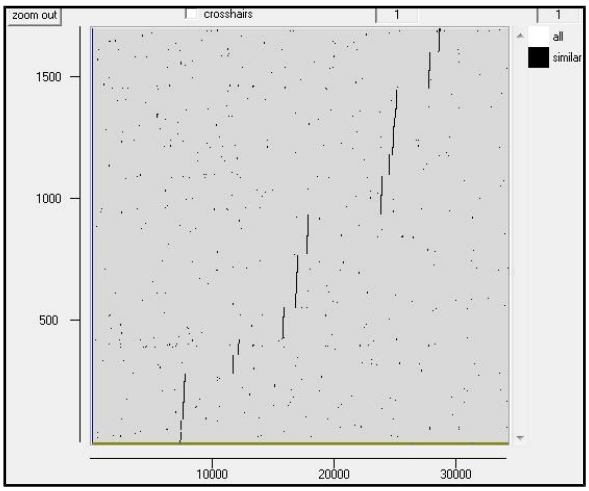


Re-run **Find matching words**, exactly as before but this time, select a **word length** of **50**. Your new dotplot will be drawn on top of you first and so will thus be difficult to see. A plot can be repositioned by moving around its configuration button (the corresponding square in the top right hand corner) with the middle mouse button. You could move the new dotplot so it is above or below the first or into a separate window (the best choice) by dragging its configuration button into an unoccupied place on your desktop. Your second plot is essentially as the first, without the noisy background. Now it looks more like twelve matching regions. We compare genomic sequence with cDNA, so it is reasonable to assume the matching regions of this plot represent the exons found in a transcript of **PAX6**.

61 If you need to, use the pull down menus provided to select from any of the sequences in your **Sequence manager**.
62 Thus asking the program to indicate the presence of exactly matching regions of **8** base pairs between the chosen sequences.
63 There is a **zoom out** button at the top of the display to undo any **zooming in** you try.
64 I.e. dots that do not represent anything interesting.

Is the number of matching regions consistent with the number of exons you might expect?_____

If not, can you explain the discrepancy?_____

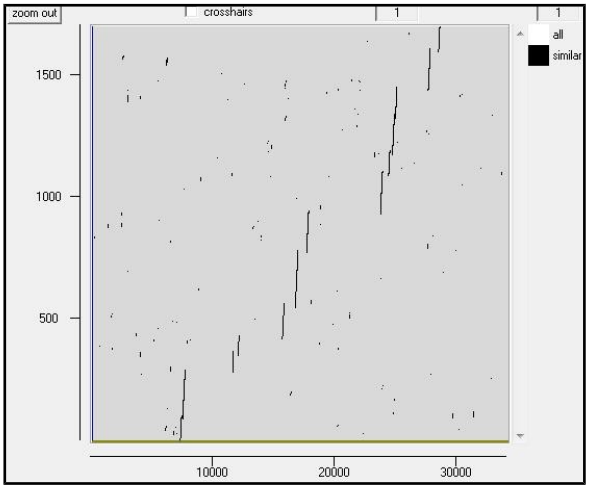


The second plot is clearer. Noise has been eliminated without loss of "real" features. The choice is not always so clear. It can be necessary to experiment using different parameters.

The algorithm you have used thus far is fast but crude. Adequate for these sequences between which all real matches are long and almost exact. However, in most circumstances a more meticulous slower[65] approach might be required in which "similar" as well as exactly matching words are recognised. From your **Comparison** menu, select **Find similar spans**, accepting the default **window length** of **11** and **minimum score** of **10**[66], click **OK**[67]. Your plot should be as illustrated.

Again, from your **Comparison** menu, select **Find similar spans**. This time, set the **window length** to **50**. **spin** automatically resets its **minimum score** field to a new default of **28**. Click **OK**. Separate out your plots so that they can be seen clearly. No surprises? A less noisy background is the most obvious effect. Also you should be able to see that the "real" features appear to have grown a little.

Why do you suppose that might be?_____



The dotplot can be useful as a rapid overview of the similarity between two sequences. Alone, it provides insufficient detail. For this, we need to generate textual alignments.

spin produces helpful graphics when textually aligning sequences. These graphics are most informative when superimposed over a corresponding dotplot. Here, the most useful dotplot is the simplest. The one you generated using the "**Find matching words**" algorithm with a word length of **50**. If you have not already done so, throw all the others away. If necessary, recreate the required dotplot.

Pairwise textual sequence alignment

Now make some alignments that will show the detail of the dotplot comparisons. The algorithms used are more rigorous than those used for searching databases, often producing more revealing alignments. Thus it can be a good idea to use these tools on similar sequences identified by database similarity searches

Global sequence alignment

A global alignment is one that aligns two sequences over their entire lengths. In **spin**'s **Comparison** menu, the option **Align sequences**[68]. Click on **spin**'s **Align sequences** option. Click **OK** accepting the defaults[69]:

score for match	**(4)**	value added to the alignment score for each correctly matched base
score for mis-match	**(-2)**	value added to the alignment score for each incorrectly matched base
penalty for starting gap	**(8)**	value subtracted from the alignment score for each gap
penalty for each residue in gap	**(1)**	value subtracted from the alignment score for each element of a gap beyond the first

spin represents the computed alignment graphically, over the dotplot. It should suggest that it has successfully aligned the group of exons around **23,500** to **25,000** in the genomic sequence and **940** to **1,450** in the mRNA[70], but then failed with the rest.

65 Not that the speed difference will be apparent with sequences of the size we are comparing here.

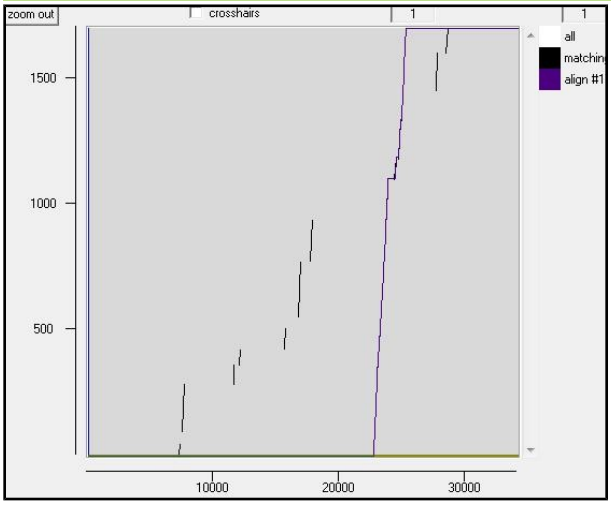
66 Thus matched windows of length **11** in which **10** of the aligned bases are identical are regarded as significant. By default, **spin** awards **1** point to matching bases and **0** points to mismatched bases. The equivalent program in the **EMBOSS** package, **dotmatcher**, awards matching bases **+5** and penalises mismatched bases **-4**. The full **EMBOSS** default DNA comparison scoring matrix can be found in an appendix.

67 Depending how you left you displays, you might need to separate out your dotplots in order to see them clearly.

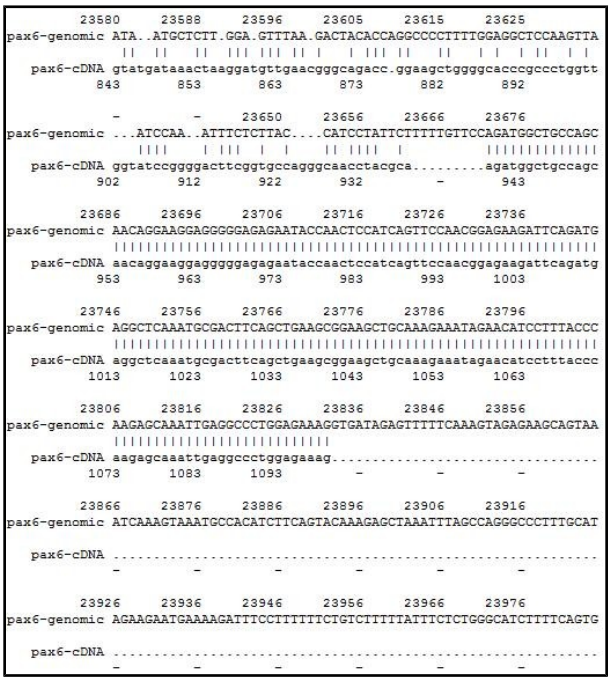
68 The **EMBOSS** package offers the program **needle**, which is a rigorous implementation of the Needleman-Wunsch algorithm for global alignment. **EMBOSS** also offers a less rigorous (i.e. faster and sloppier) program called **stretcher**.

69 A complete description of these parameters will be included in a presentation. Soon, if it has not already occurred.

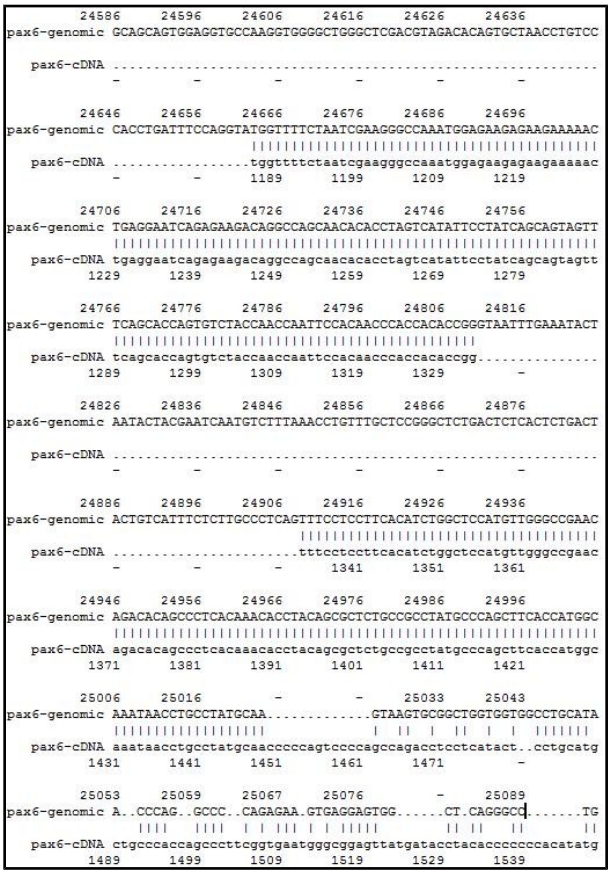
70 Click in the **crosshair** box and use the crosshair to get a good idea of where features are in your plots. I think the cross-hair is very irritating to leave active, so I suggest you click it off again once its purpose is fulfilled.



Examine how your alignment graphic matches the dotplot by right clicking its configuration button and using the **Hide**/**Reveal** option.



Now inspect the textual alignment in **spin**'s **Output window**. As the graphic suggests, the mRNA has been prefixed with over **22,000** minus signs to make it of equal length to the genomic sequence. There follows an entirely unconvincing (given the evidence of the dotplots) alignment of the first few exons of the mRNA with the region around **23,000** in the genomic sequence.



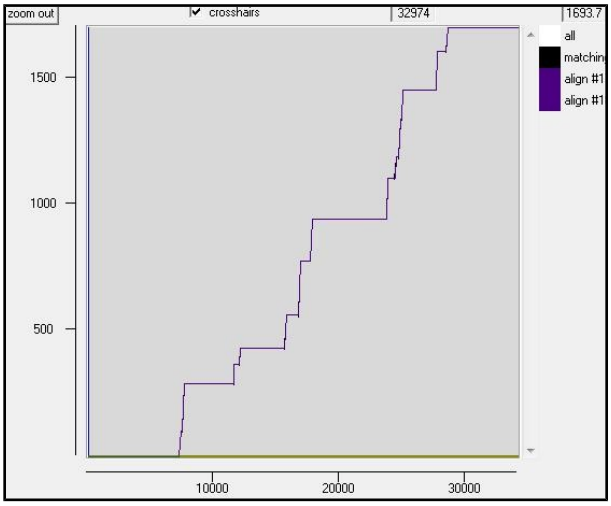
The **4** exons around **23,500-25,000** in the genomic sequence and **940-1,500** in the mRNA are aligned convincingly. The final mRNA exons are aligned, with an excess of poetry, with the bases after **25,000** of the genomic sequence.

The mRNA is then desperately padded to reach the end of the alignment. The bizarre padding at each end of the mRNA is required to fulfil the requirement of a global alignment to be “end to end”. The entirety of both sequences must be included in the alignment.

How many convincingly aligned regions did you see?_____

Roughly how many did you expect?_____

Clearly, this alignment is not correct. Can you explain why?_____



Try the alignment again, this time with much "cheaper" gaps. First **Hide** the plot associated with your first global alignment attempt. This should leave the dotplot in clear view. Now click on **spin**'s **Align sequences** option once again. This time, set the **penalty for each residue in gap** to **0**. This amounts to asking that all gaps be penalised exactly **8** points no matter how long they be. Click **OK**.

Now the alignment graphic shows the alignment dutifully passing though all the matching regions identified by the dotplot. Take a look at the textual output associated with this second alignment.

How many matching regions are there this time?_____

Is the count **now** roughly as you would expect?_____

These global alignments show how misleading it can be to run programs without carefully considering their assumptions and parameter values. The second time around, you achieved the "correct" alignment, but only by making the cost of gaps so cheap that huge introns could be mistaken for normal insertion/deletion events. This only worked because the mRNA/genomic sequences came from the same organism and the corresponding exons were effectively identical[71]. If the comparison was between homologous sequences from different organisms[72], it would usually be necessary to use gap penalties that reflected the way real insertion/deletion events occurred in the exons. Declaring almost free gaps to cope with introns would rarely succeed[73].

The best solution is to accept that the alignment between a cDNA/mRNA and genomic sequence is a special case. A general alignment program will not make the right assumptions for such alignments and is thus the wrong tool for the job. There is a program in the **EMBOSS** package, called **est2genome**, which is specifically designed for the alignment of cDNA/mRNA and genomic sequences. **est2genome** (and similar programs) may assume much more about the sequences to be aligned than can a general purpose alignment program. Gaps representing introns can be placed far more accurately if they are **known** to represent introns. Programs such as **est2genome** seek the highly conserved bases that occur at intron/exon boundaries, **C**/**T** rich intronic regions, **polyA** regions and **Stop**/**Start** codons to assist its detection of exons and gene structures.

est2genome is a fine program, but there are two other options offered at the **NCBI** in America that do the same job, I think, somewhat more nicely. Of these, I choose the program called **splign** for this exercise on the grounds it is the more sophisticated service[74]. To investigate, go to the home of **splign** at:

http://www.ncbi.nlm.nih.gov/sutils/splign

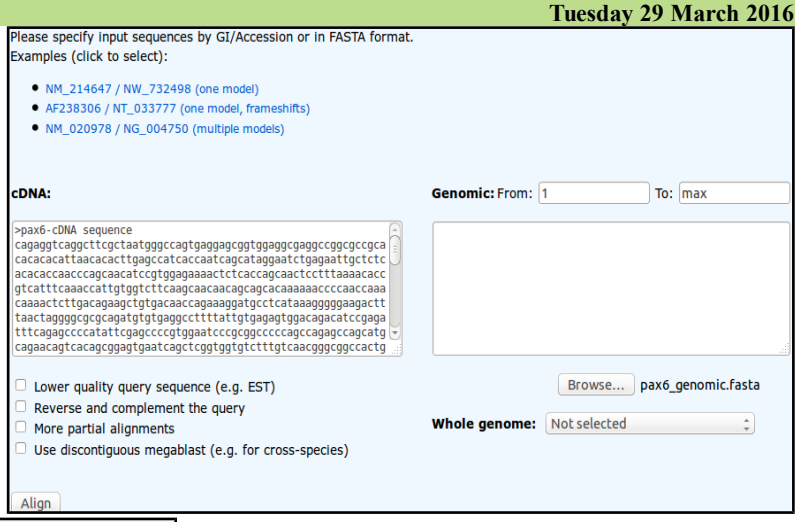
71 and so would align correctly given almost any gap penalties, or other parameter settings that might have been chosen.

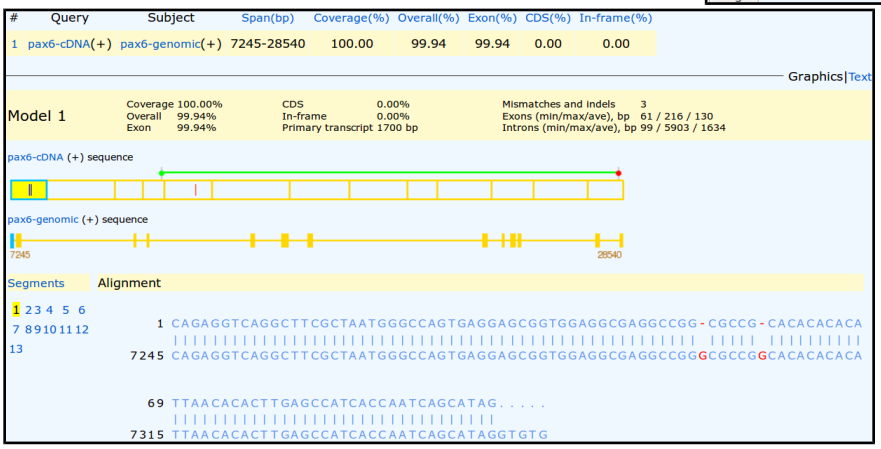
72 Comparing mRNA from one organism with genomic sequence from another is one way of investigating gene structure in newly sequenced genomic sequence.

73 One very simple solution to this problem offered by the **GCG** (now defunct) package version of **needle** (a program called **gap**) was to offer a gap penalty ceiling. The program would compute a gap penalty in the normal way until a given "cut off" was reached. **gap** would then assume it was trying to stretch over an intron rather than allow for an insertion/deletion, so it would allow the gap to increase at will without further raising the penalty. This worked much better than my intuition suggested it should!

74 The other is called **spidey**. The **spidey** home page is:
http://www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey/
It is pretty straight forward to use, just follow your nose. **est2genome**, **splign** and **spidey** should all give the same answer for this simple alignment and very similar answers for all such problems.

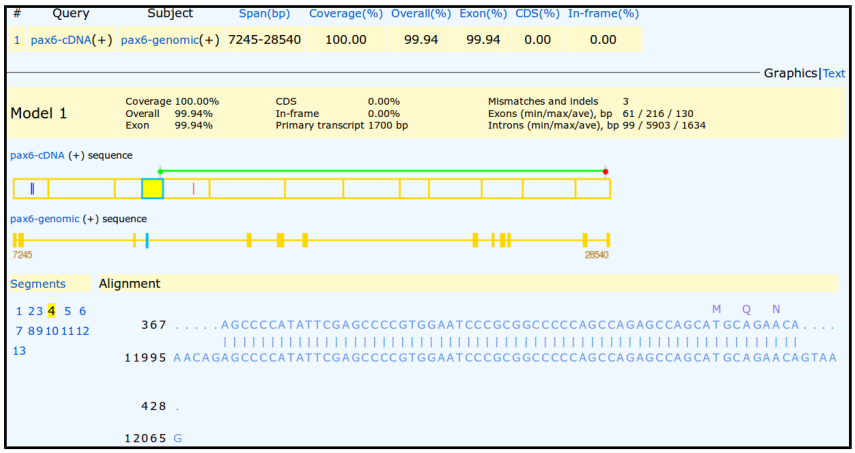
Click on the **Online** button. In the **Genomic** section, **Browse** to upload **pax6_genomic.fasta**. In the **cDNA** section, paste the sequence **pax6_cdna.fasta**. Where **cDNA** and **Genomic** sequences share exons that are nearly identical, **splign** uses the comparison algorithm **megablast** (default choice). Where exons are less similar (e.g. when the **cDNA** and **Genomic** sequences are from different organisms) the more sensitive option **discontinuous megablast**, might be a better choice[75]. Note the option to compare your **cDNA** with a **Whole genome** (including Human). Today, the default options are fine. Click the **Align** button.

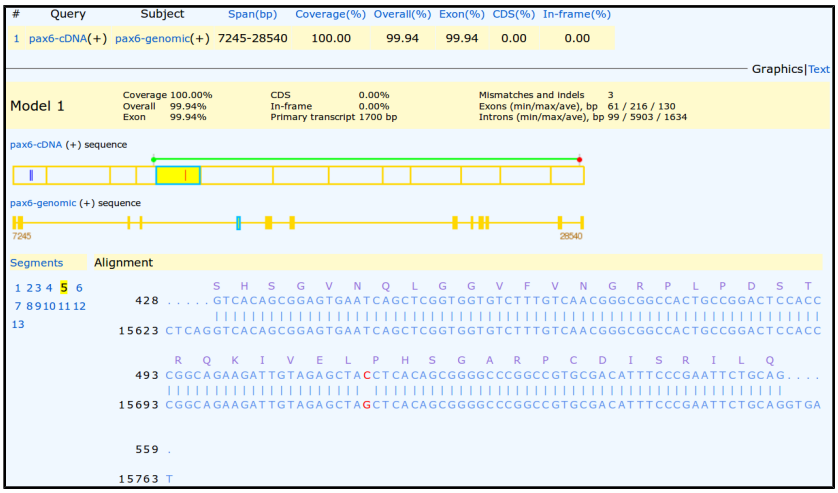




Your results will appear showing the cDNA split into **13** sections (the predicted exons) corresponding to **13** regions of the genomic sequence indicated by yellow rectangles. The first exon alignment is displayed showing two **cDNA** deletions. Though these are in a non-coding region, they could easily still be very significant. However, for the purposes of this exercise, let us assume they are not. The **Start** (green) and **Stop** (red) codons of the cDNA are illustrated by the bar above the cDNA display.

Click on the exon including the green **Start** codon (the fourth). The first coding exon is now displayed. The statistics at the top of the display include the claim that there are three discrepancies (**Mismatches and indels**[76]) between the **cDNA** and **Genomic** sequences. Two of these are the deletions we have already seen in the first exon of the cDNA. The third is indicated by the red bar in the fifth exon of the **cDNA** display.



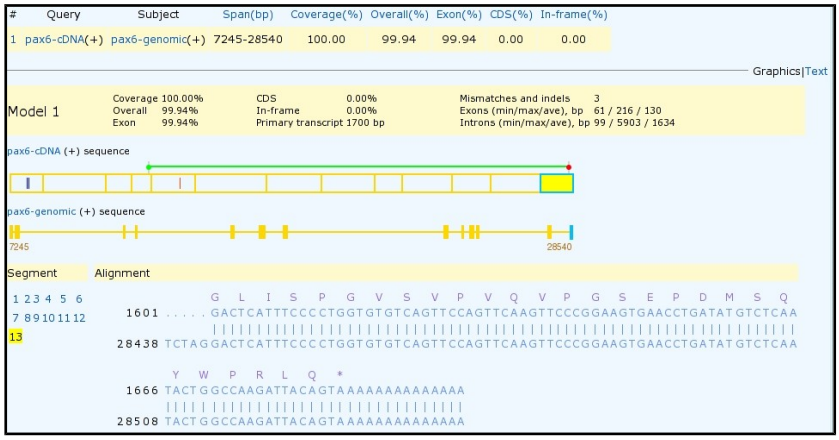


Click on the fifth exon section of the cDNA display.

The third difference, a substitution, should be clear to see. Given it changes the coded protein, this substitution is likely to be the most significant.

What is the amino acid corresponding to this position in the mRNA of the aniridia patient[77]?_____

75 Why this is so will be considered later when we look at the database searching program **blast**.
76 An **indel** is either an **in**sertion or a **del**etion, depending upon which of the aligned sequences you are considering.
77 How this substitution affects the protein is not clear. Translating both the **Genomic** sequence and the **cDNA** might have helped? The **Standard Genetic Code** is in an Appendix, or look back to the **Uniprot** feature table for **PAX6_HUMAN**. It is the **Natural Variant** at position **33**.



Click on the last exon section in the cDNA display. You should now see the final exon of the cDNA with the **Stop** codon and polyA region.

Finally, click on the **Text** link to view the textual summary of the **splign** results.

#	Query	Subject	Span(bp)	Coverage(%)	Overall(%)	Exon(%)	CDS(%)	In-frame(%)
1	pax6-cDNA(+)	pax6-genomic(+)	7245-28540	100.00	99.94	99.94	0.00	0.00

Graphics|Text

#	Query	Subject	Idty	Len	Q.Start	Q.Fin	S.Start	S.Fin	Type	Details
+1	pax6-cDNA	pax6-genomic	0.981	103	1	101	7245	7347	<exon>GT	M53IM5IM43
+1	pax6-cDNA	pax6-genomic	1	188	102	289	7447	7634	AG<exon>GT	M188
+1	pax6-cDNA	pax6-genomic	1	77	290	366	11537	11613	AG<exon>GC	M77
+1	pax6-cDNA	pax6-genomic	1	61	367	427	12000	12060	AG<exon>GT	M61
+1	pax6-cDNA	pax6-genomic	0.992	131	428	558	15628	15758	AG<exon>GT	M86RM44
+1	pax6-cDNA	pax6-genomic	1	216	559	774	16686	16901	AG<exon>GT	M216
+1	pax6-cDNA	pax6-genomic	1	166	775	940	17606	17771	AG<exon>GT	M166
+1	pax6-cDNA	pax6-genomic	1	159	941	1099	23674	23832	AG<exon>GT	M159
+1	pax6-cDNA	pax6-genomic	1	83	1100	1182	24348	24430	AG<exon>GT	M83
+1	pax6-cDNA	pax6-genomic	1	151	1183	1333	24660	24810	AG<exon>GT	M151
+1	pax6-cDNA	pax6-genomic	1	116	1334	1449	24909	25024	AG<exon>GT	M116
+1	pax6-cDNA	pax6-genomic	1	151	1450	1600	27602	27752	AG<exon>GT	M151
+1	pax6-cDNA	pax6-genomic	1	98	1601	1698	28443	28540	AG<exon>AA	M98

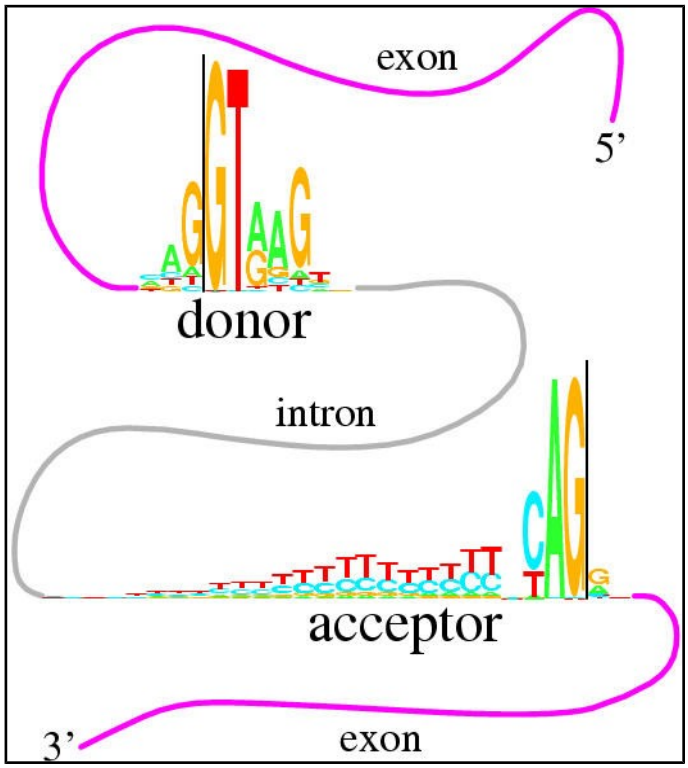
How do you interpret the **Details** column for exons 1 and 5?_____

Where is the substitution in the aniridia patient mRNA?_____

Where is the substitution in the Genomic Sequence?_____

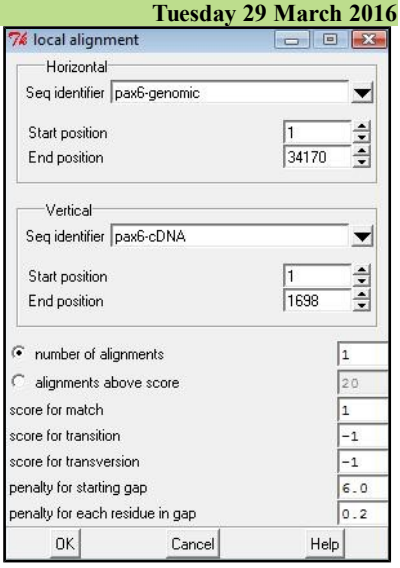
Compare the predicted **splign** intron/exon boundaries with the conservation suggested by the logo[78]?_____

What deviation(s) from the model suggested by the logo can you see?_____



78 The original label for this very nice graphic is:
This figure shows two ''sequence logos'' which represent sequence conservation at the 5' (donor) and 3' (acceptor) ends of human introns. The region between the black vertical bars is removed during mRNA splicing. The logos graphically demonstrate that most of the pattern for locating the intron ends resides on the intron. This allows more codon choices in the protein-coding exons. The logos also show a common pattern ''**CAG|GT**'', which suggests that the mechanisms that recognize the two ends of the intron had a common ancestor. See R. M. Stephens and T. D. Schneider, "Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites", J. Mol. Biol., 228, 1124-1136, (1992).

Local sequence alignment

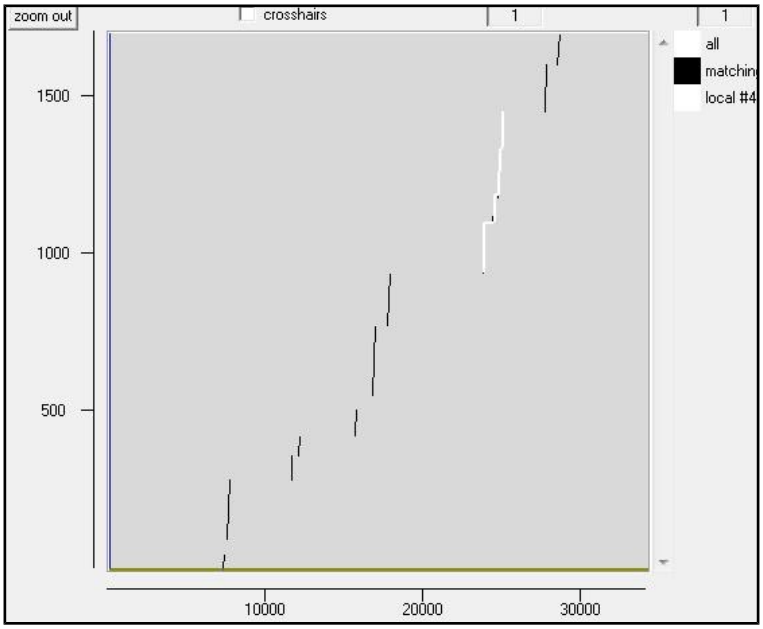


Global sequence alignment algorithms align sequences over their entire lengths. **Local** alignment methods searches for regions of similarity and need not include the entire length of the sequences. It is important to select the type of alignment that makes best sense for your sequences. Here, we expect an ordered sequence of matching regions (the exons) to occur in both sequences but spaced very differently. With an effort (or preferably appropriate software), we can get reasonably sensible alignments using a global or a local approach. However, if our sequences represented multi-domain proteins that shared some domains but not others, or the same domains but in differing orders, or proteins where there have been regions of duplication, a single sensible global alignment would not exist. A local approach would be essential.

Move back to **spin**, tidy away all previous graphics plots except one clear **dotplot** between **pax6-genomic** and **pax6-cDNA**. From **spin**'s **Comparison** menu, select **Local alignment**. Note that:

- in common with most local alignment programs, the default **number of alignments** to be reported is **1**[79]
- **transversions** and **transitions** may be penalized differently
- gaps are slightly cheaper than for **gobal** alignments

Accept all the defaults and click **OK**. The graphic suggests that the single "best" region that **spin** has elected to locally align, is the region around **25,000** in the genomic sequence where there are **4** exons that are close together. It was this region around which the global alignment algorithm chose to build its first attempt.

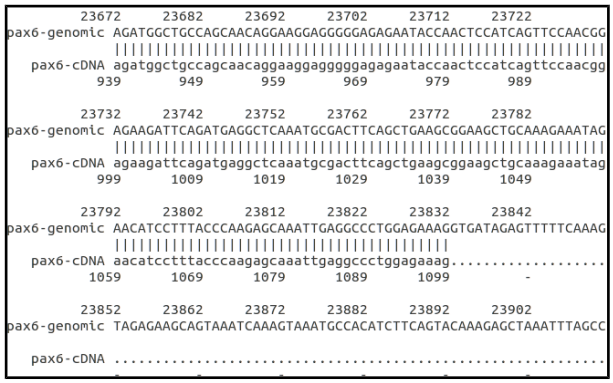


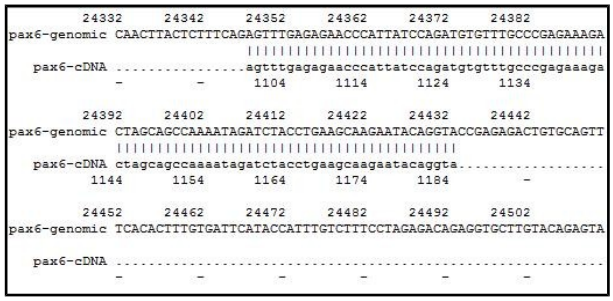
Check that the textual and graphic outputs agree.

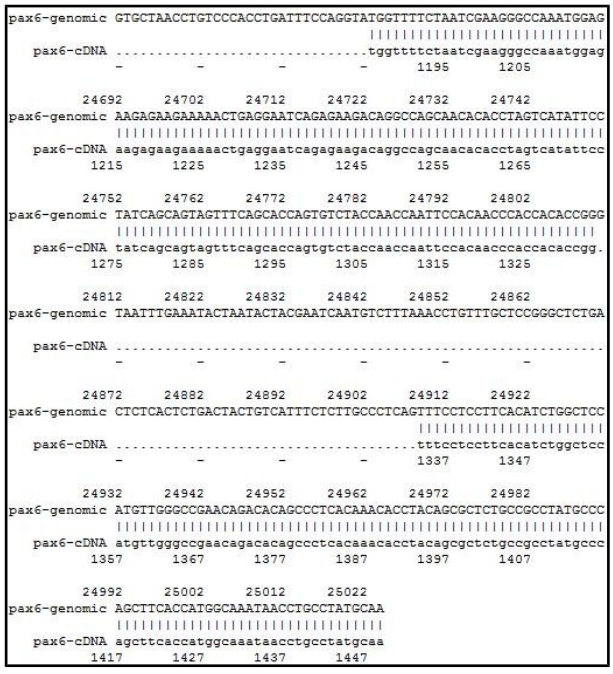
How might the gap around **24,600** in the genomic sequence been positioned more intelligently? _____

_____

The choice of this region is primarily due to the choice of gap penalties. If bigger gap penalties were selected (increase **penalty for each residue in gap** to **1**, say) **spin** would cease to regard the **4** exons around **25,000** as one entity. Their individual alignment scores would look less attractive. Eventually, the region around **16,700** (the longest single exon) would easily outscore the rest. If you have time, try it.

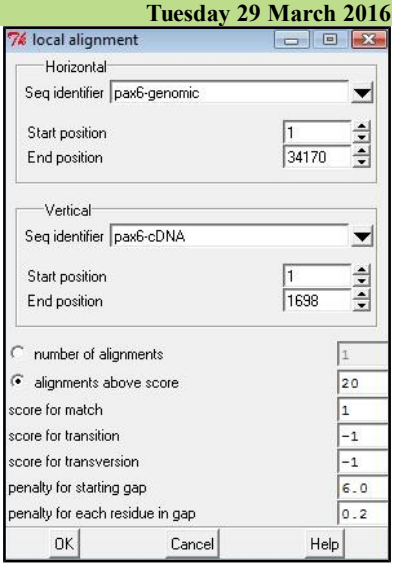






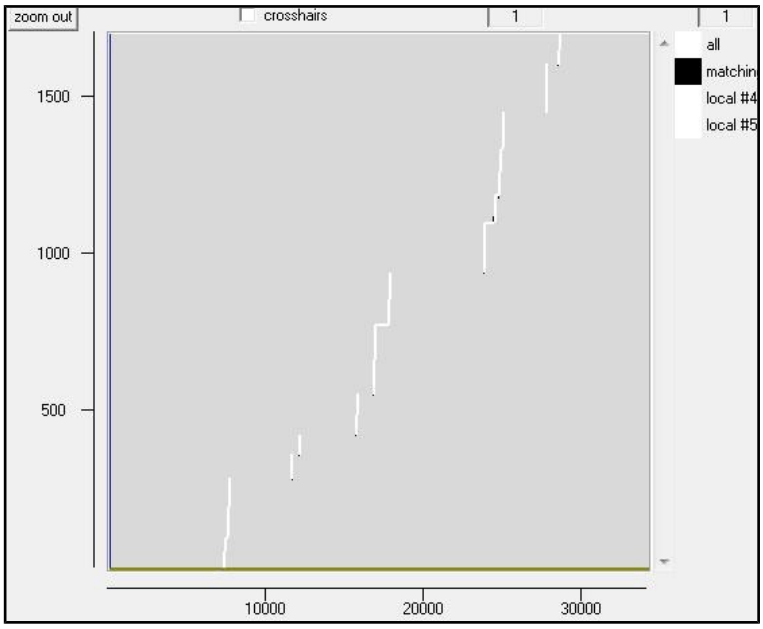
79 I.e. report the best matching region and ignore any others, however deserving they might be? This has always seemed an odd default to me.

From **spin**'s **Comparison** menu, select again **Local alignment**. Clearly, there is more than **1** meaningful local alignment. To see more, **spin** offers two options.



You could guess the **number of alignments** value. However, this is not trivial. You know the number of exons (and so the number of logical alignments expected) with some certainty by now, but this does not necessarily correspond to the expected number of local alignments. **spin** will combine close exons into single alignments if gap penalties are low enough. **spin** assumes that your choice of **number of alignments** is exactly correct. Too low and you will miss real alignments. Too high and **spin** will show alignments of ever increasing fantasy.

It is normally better, therefore, to specify the alignments you wish **spin** to display by providing a quality cut off rather than a volume cut off. To do this, turn on the **alignments above score** option. Use the default value of **20**.



It is not obvious how one might choose a sensible value. In such circumstances, the "lazy" option of just accepting the default is fine. Maybe someone sensible chose it? If not, one can always iterate to a value that works. I.e. if you would like more alignments, lower the value and try again. Too many? raise it. Repeat until the program agrees with what you wished to be correct in the first place. Is not science wonderful? With hope in your heart, click **OK**.

Looking at the graphic, I would say we got lucky this time[80]. There is a local alignment that covers every exon identified by the **dotplot** and no extra results requiring explanation.

Scan your results to find the exons. You should see that all have been correctly aligned.

Why do you suppose your aligned exons are not presented in the correct positional order?_____

Further features of Spin

ORF Identification and Translation

80 If you wish to be meticulously honest. You should **Hide** the graphic of your first **Local alignment** (right click its configuration box, choose **Hide**). Not that it will change anything as you will have found exactly the same alignment the second time round – along with the others.

Searching for sequence similarities in databases.

The most popular way to investigate a sequence has always been to compare it with one of the sequence databases now accessible from sites all over the world. When sequences databases were more sparsely populated than now, the objective was to search hopefully, not always with success, for any convincingly similar sequence(s). When such a match was discovered, it could be supposed that known properties of the "similar" database sequence might provide insight to the properties of the query sequence. Now, the databases are full of sequences representative of most interesting conditions. Similarity searches are conducted in the expectation of finding many close "hits" for almost any sequence. Fewer database searches are conducted in complete ignorance of what the query sequence might be.

Here, take the **PAX6** genomic DNA sequence retrieved from **Ensembl** and conduct two searches analogous to those run in the **Ensembl** pipeline. Results should confirm that which has already been discovered using other sources.

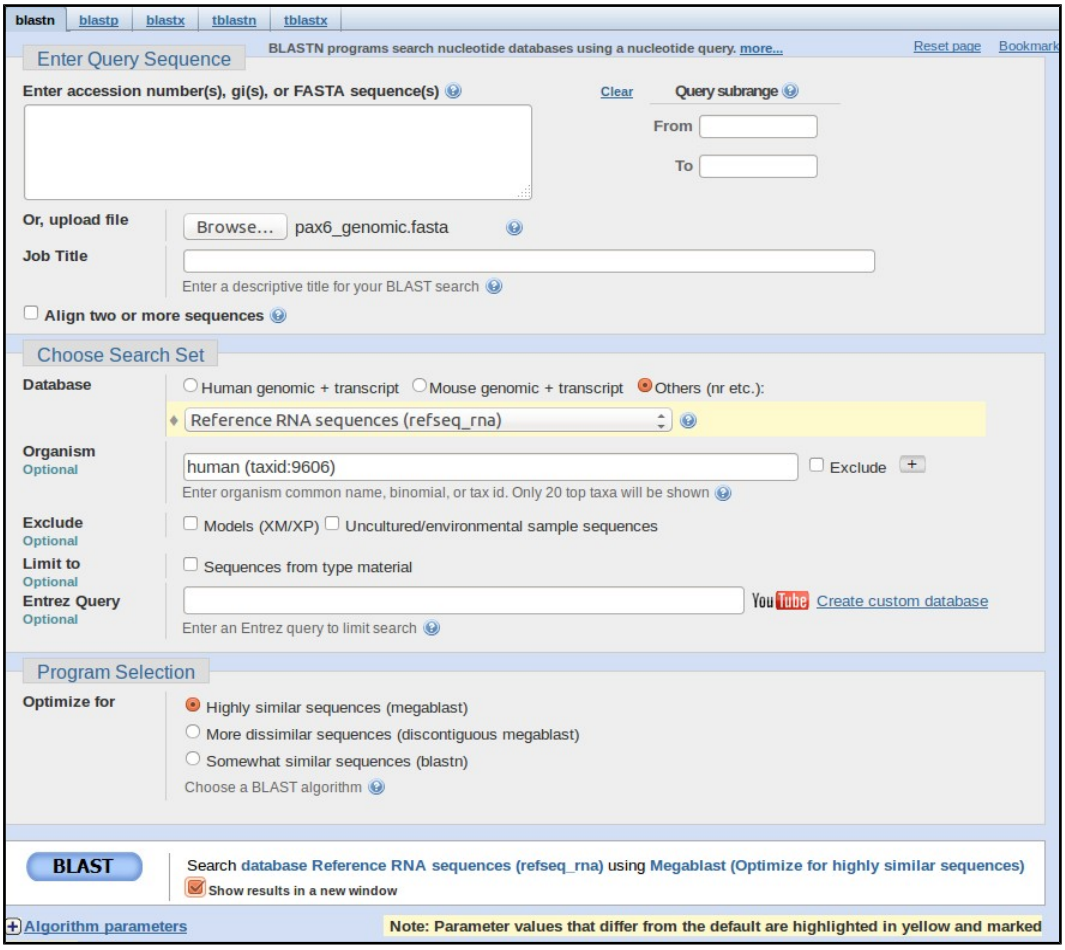
blast is not the only sequence database searching program available, but it is the most popular by a very long way. **blast** searches are offered in many forms by many servers all over the world, but the most comprehensive and reliable service has to be that offered by the **NCBI**.

Go to the **NCBI** homepage at:

`http://ncbi.nlm.nih.gov`

Select the **BLAST** option (from the **Popular Resources** list). In the **Basic BLAST** section, select **nucleotide blast**. Use the **Enter Query Sequence Browse** (or **Choose File**) button to upload the file:

pax6_genomic.fasta.



For results like those used by **Ensembl** to predict **PAX6** transcripts, you must compare your genomic sequence to a reliable set of human mRNA/cDNA (or similar) sequences.

In the **Choose Search Set** section, set the **Database** to **Reference RNA sequences (refseq_rna)**.

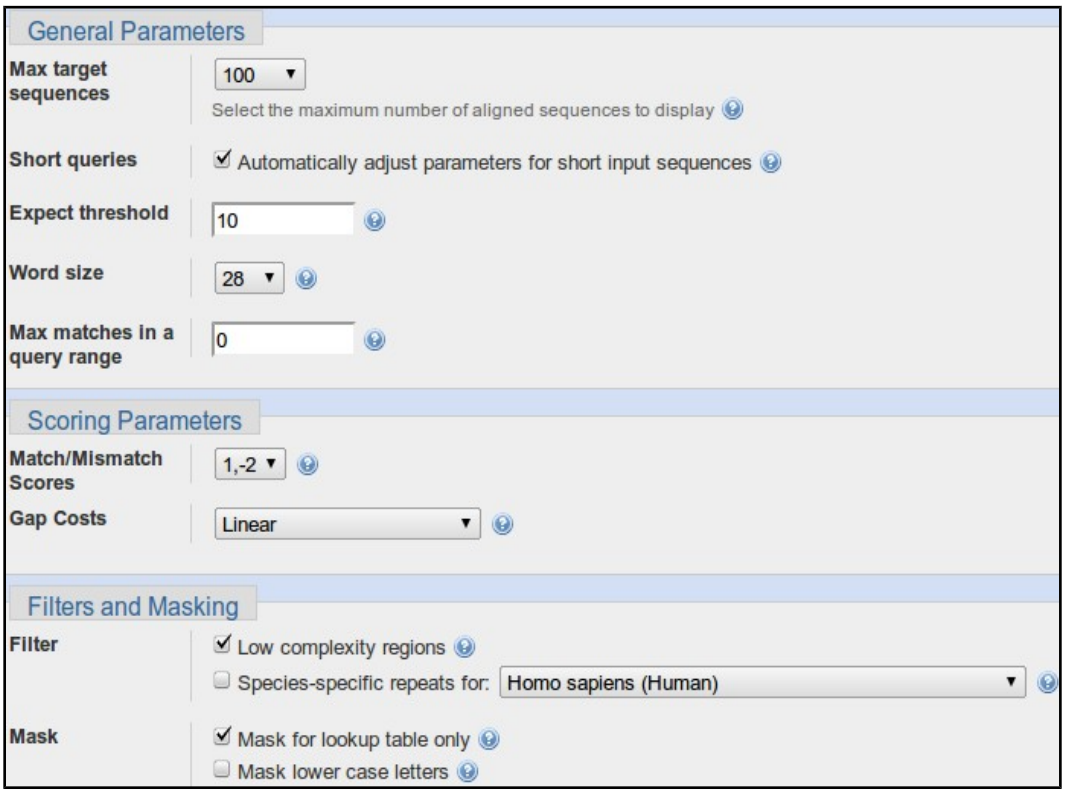
You are now able to specify an **Organism**, choose **Human**.

blast is now set to compare the **PAX6** genomic region with all **Human** mRNA sequences in **RefSeq**.

Note that the default **Program Selection** is **Highly similar sequences (megablast[81])**, which seems appropriate here as all the mRNA that correctly match should surely do so almost perfectly.

81 **megablast** is a less sensitive but even faster version of **blast** only suitable when, as now, almost identical matches are sought.

Click on the **Algorithm Parameters** button. The defaults are fine here, but before starting your search, try changing the **Program Selection** and observing the different **Algorithm Parameters**.



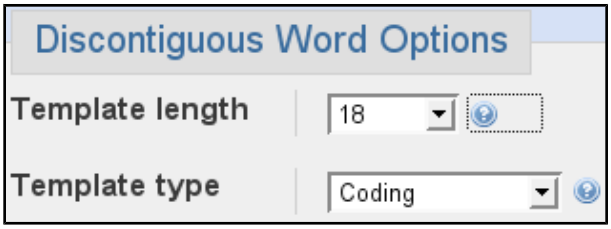
The default settings of all shared parameters are identical for the two slower more sensitive **Program Selections**.

There are differences for **megablast**, where speed is of the essence and sensitivity can be sacrificed.

Smaller **Word sizes** slow searches but increase sensitivity. For **megablast** the default **Word size** is **28** otherwise it is **11**.

Gapped alignment is time consuming and, by default, considered more crudely by **megablast** than the other two algorithms[82].

Filtering and Masking matches with organism specific repeats and/or low complexity regions takes time, and so only avoiding **Low complexity regions**[83] is on by default for all **Program Selections**.



When **discontinuous megablast** is selected, an extra options section appears. Discussing how this flavour of **blast** works is a little beyond the scope of these note, but briefly. Unlike the other **Program Selections**, **discontinuous megablast** does not just look for exactly matching "words" of given size as a first step towards identifying matching regions between sequences. It looks for a pattern of matching bases within a word. For example, the default choice assumes your query is **coding** and looks for **11** matching bases within a word of **18**. Approximately, every third base is allowed not to match. Biologically, this can be justified as allowing for third codon position wobble. For more detail, use the appropriate button. Notice there are buttons by every parameter selection. Try one or two. In the process, discover:

When would **Mask lower case letters** be a useful thing to do? _____

_____

Automatically adjust parameters for short input sequences is independent of **Program selection**, and so remains unaltered.

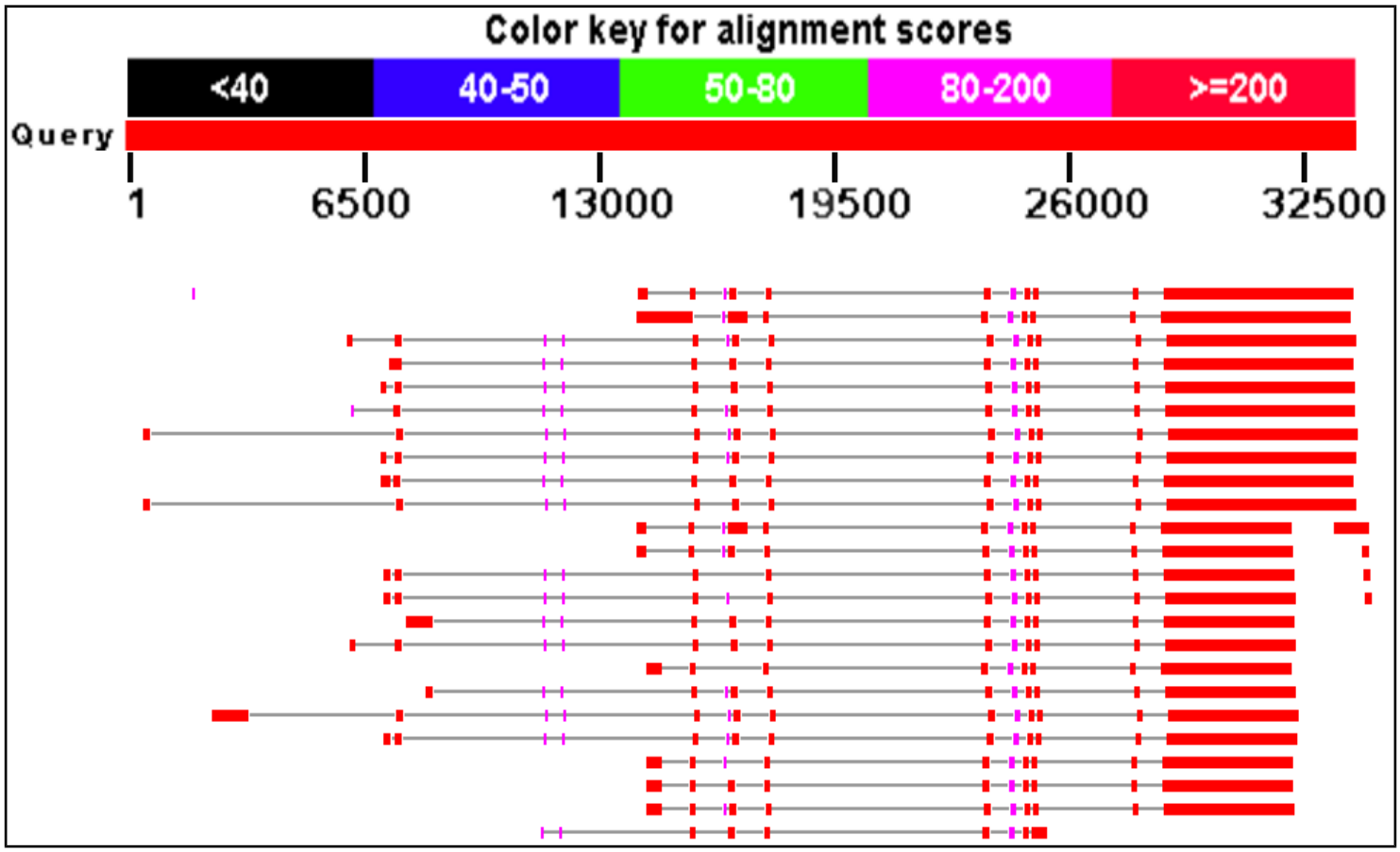
Which parameters would **blast** need to **automatically adjust** to cater for short input sequences (such as primers being tested for uniqueness), and why? _____

_____

82 By default, **megablast** uses **Linear Gap Costs**. That is, it just multiplies the size of the gap with the **Mismatch** penalty. The other two algorithms employ the more common **Affine** strategy, using **Existence** and **Extension** penalties. For more about **Gap Penalties**, go **here**.

83 This filter avoids finding "hits" supported only by matches in regions not specific to the query. For example, a polyA tail cannot help to identify a specific mRNA as it is present is all mRNAs. The use of this filter will be evident when we look at the **blast** output.

Finally, ensure all the defaults are back in place[84] and the **megablast** is the **Program Selection**, ask **blast** to **Show results in a new window** and then click on the BLAST button. Impressively swiftly, you will have results. At the top of which will be a graphical overview.



This graphic implies that there are **24** full length matches between your genomic sequence and mRNAs in **RefSeq**. The **RefSeq** entries had to be "gapped" in order to compensate for the introns that are represented in the genomic sequence but removed from the mRNA sequences. The **red blocks** therefore represent very closely matching (>=200 brownie points) exons, the lines joining the **red blocks** represent introns that have been spliced out. All **24** hits match reasonably uniformly except for the first few exons, implying significant variation in the **5´ UTR**.

Why do you suppose that a few of the exons do not achieve the maximum score? _____

_____

Explain why one exon in the reasonably consistent region, does not appear in all of the transcript matches? _____

_____

GeneCards reported that there were **24 PAX6** transcripts recorded in **RefSeq**, **11** high quality **NM_** entries plus a further **13 XM_ PREDICTED** transcripts. **Ensembl** claimed to have used **10** or the **11** high quality **NM_ RefSeq** sequences to aid its transcript predictions, but ignored the **13 XM_ PREDICTED**, less certain **RefSeq** sequences. **blast** just sees sequences and cannot be influenced by the quality of the support for their existence, so **blast** reports that all **24 RefSeq PAX6** mRNAs match the **PAX6** genomic region convincingly.

Perfect consistency between three sources of information! Wonderful, but this is not always the case. When **RefSeq** acquires extra sequences, **Ensembl** will not notice until its next "**genebuild**" event. When **Ensembl** finally gets into line with the changes in **RefSeq**. **GeneCards** will not respond until its next update. **RefSeq** has been relatively static of late, with regards to this particular gene. It is rather a nice change to be looking at a neat and consistent picture at this point. Sadly, it will not last.

84 If you have any non-default settings, they should be highlighted in yellow.

In summary, if you hover over the graphical hits, their origin will be displayed above the graphic[85]. The facts are:

- The top **10** and the bottom **1** full length hits are of the best quality (i.e. **NM_** entries with good supporting evidence). **11** in all.
- From the **11th** full length hit, there are **13** entries all labelled "**PREDICTED**". We have already concluded that **Ensembl** is clever enough not to rely on **PREDICTED RefSeq** entries alone to justify an **Ensembl** transcript prediction. **GeneCards** does count them as sufficient to indicate **RefSeq** transcript predictions however.
- There are **4** small hits to the extreme right of the graphics at the same level as the top **4 PREDICTED** hits. These are the ends of **mRNA**s for the **ELP4** gene and are exactly where you should expect them to be given previous discussion. Reject these contemptuously, they do not pertain to our investigation of **PAX6**.
- The tiny smudge match to the left of the top hit is "**uncharacterized**" and fails to fit in with my story, so I ignore it!

So, this **blast** search suggests the existence of **24 PAX6** transcripts supported by **RefSeq** data, as is reported by **GenCards**. Also, the results are consistent with the information discovered in **Ensembl**.

Move down a trifle and you will find a simple list of the **29** matches represented in the graphic.

Description	Max score	Total score	Query cover	E value	Ident	Accession
Homo sapiens paired box 6 (PAX6), transcript variant 11, mRNA	9659	12484	19%	0.0	100%	NM_001310161.1
Homo sapiens paired box 6 (PAX6), transcript variant 10, mRNA	9659	15161	24%	0.0	100%	NM_001310160.1
Homo sapiens paired box 6 (PAX6), transcript variant 8, mRNA	9659	12929	20%	0.0	100%	NM_001310158.1
Homo sapiens paired box 6 (PAX6), transcript variant 7, mRNA	9659	12729	20%	0.0	100%	NM_001258465.1
Homo sapiens paired box 6 (PAX6), transcript variant 6, mRNA	9659	12761	20%	0.0	100%	NM_001258464.1
Homo sapiens paired box 6 (PAX6), transcript variant 5, mRNA	9659	12737	20%	0.0	100%	NM_001258463.1
Homo sapiens paired box 6 (PAX6), transcript variant 4, mRNA	9659	12862	20%	0.0	100%	NM_001258462.1
Homo sapiens paired box 6 (PAX6), transcript variant 2, mRNA	9659	12833	20%	0.0	100%	NM_001604.5
Homo sapiens paired box 6 (PAX6), transcript variant 1, mRNA	9659	12942	20%	0.0	100%	NM_000280.4
Homo sapiens paired box 6 (PAX6), transcript variant 3, mRNA	9659	12791	20%	0.0	100%	NM_001127612.1
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X13, mRNA	6613	10063	15%	0.0	100%	XM_005252958.3
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X12, mRNA	6613	9439	14%	0.0	100%	XM_011520153.1
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X11, mRNA	6613	9329	14%	0.0	100%	XM_006718246.2
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X10, mRNA	6613	9410	14%	0.0	100%	XM_011520152.1
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X9, mRNA	6613	10507	16%	0.0	100%	XM_005252956.3
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X8, mRNA	6613	9783	15%	0.0	100%	XM_005252955.3
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X7, mRNA	6613	9091	14%	0.0	100%	XM_011520151.1
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X6, mRNA	6613	9637	15%	0.0	100%	XM_011520150.1
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X5, mRNA	6613	11324	17%	0.0	100%	XM_011520149.1
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X4, mRNA	6613	9814	15%	0.0	100%	XM_005252954.3
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X3, mRNA	6613	9172	14%	0.0	100%	XM_011520148.1
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X2, mRNA	6613	9502	15%	0.0	100%	XM_011520147.1
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X1, mRNA	6613	9576	15%	0.0	100%	XM_011520146.1
PREDICTED: Homo sapiens elongator acetyltransferase complex subunit 4 (ELP4), transc	1775	1775	2%	0.0	100%	XM_005252865.2
Homo sapiens paired box 6 (PAX6), transcript variant 9, mRNA	647	2630	4%	0.0	100%	NM_001310159.1
Homo sapiens elongator acetyltransferase complex subunit 4 (ELP4), transcript variant 3, r	433	433	0%	6e-118	100%	NM_001288726.1
Homo sapiens elongator acetyltransferase complex subunit 4 (ELP4), transcript variant 2, r	433	433	0%	6e-118	100%	NM_001288725.1
Homo sapiens elongator acetyltransferase complex subunit 4 (ELP4), transcript variant 1, r	433	433	0%	6e-118	100%	NM_019040.4
Homo sapiens uncharacterized LOC440034 (DKFZp686K1684), long non-coding RNA	141	141	0%	4e-30	100%	NR_033971.1

85 Or you could just read the textual list that follows the graphic if you wish to insist on the simplistic.

Why were you not surprised to discover **24 PAX6** transcripts in **Refseq** matching this sequence? _____

_____

Which of the **Refseq PAX6** transcripts corresponds to **isoform 5a**? _____

Moving further down the results you will come to the alignments between **pax6-genomic** and the matching database entries. All similarity searches use local alignment strategies[86], so you should not be surprised to see a number of alignments for each "hit" in the list. Here we have a genomic query sequence aligned exclusively with mRNA sequences from **RefSeq**. The expectation is therefore to find an alignments corresponding to exons. The alignments are ordered by quality, though you are provided with a **Sort by:** menu to alter the order to taste[87].

Look at the first alignment for the best matching **PAX6** transcript. It is the alignment of the very last exon of a **RefSeq** transcript with the end of the gene you exported from **Ensembl**.

Notice the lower case string of '**a**'s. The case indicates that they were ignored (**filtered**) as a **Low complexity region** whilst **megablast** was looking for identically matching words that might suggest matching regions[88]. By themselves, the '**a**'s are not sufficient evidence that a biological match exists. Only because the surrounding sequence is compellingly similar, can it be assumed that such a match does exist. The '**a**'s are replaced (lower case to indicate they were filtered) when the final alignment is computed. If you look a little further down the same alignment, you will see several other runs of '**a**'s and **'t**'s for which the same explanation applies.

Score	Expect	Identities	Gaps	Strand
9659 bits(5230)	0.0	5230/5230(100%)	0/5230(0%)	Plus/Plus

```
Query  28441  AGGACTCATTTCCCCTGGTGTGTCAGTTCCAGTTCAAGTTCCCGGAAGTGAACCTGATAT  28500
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1500   AGGACTCATTTCCCCTGGTGTGTCAGTTCCAGTTCAAGTTCCCGGAAGTGAACCTGATAT  1559

Query  28501  GTCTCAATACTGGCCAAGATTACAGTaaaaaaaaaaaaaaaaaaaaaaGGAAAGGAAAT  28560
              |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1560   GTCTCAATACTGGCCAAGATTACAGTAAAAAAAAAAAAAAAAAAAAAAAGGAAAGGAAAT  1619

Query  28561  ATTGTGTTAATTCAGTCAGTGACTATGGGGACACAACAGTTGAGCTTTCAGGAAAGAAAG  28620
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1620   ATTGTGTTAATTCAGTCAGTGACTATGGGGACACAACAGTTGAGCTTTCAGGAAAGAAAG  1679
```

86 To use a global approach would be to imply that you were only interested in database entries that matched your query sequence from end to end. Generally, this is not true. You would usually be interested in a database sequence that was similar over any significant region.

87 Why not try them? End up with the alignments for the top hit in **E value** order.

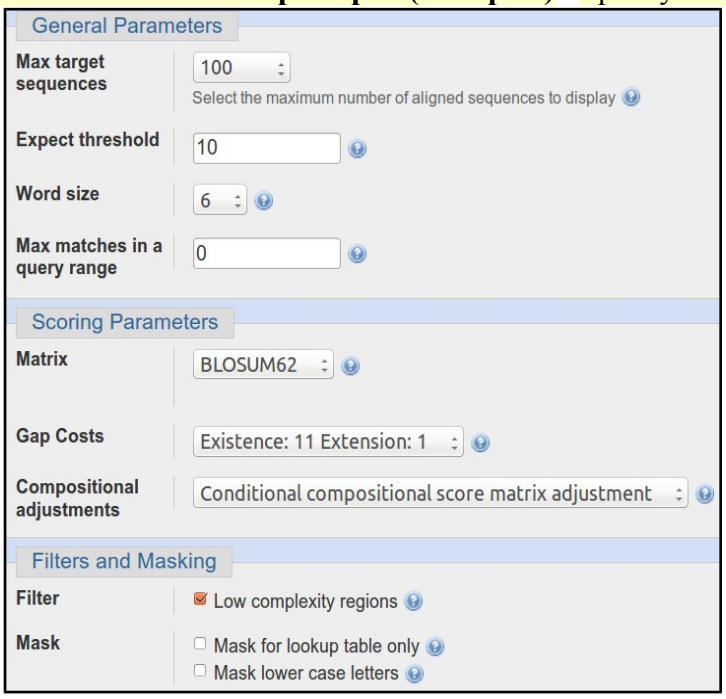
88 The mRNA in the file **pax6_cdna.fasta** ends at this polyA region. I wonder about the long 3' UTR suggested by some of the **RefSeq** entries?

Now use a version of **blast** (called **blastx**) to compare your genomic sequence with a protein database. **blastx** will translate a DNA query sequence in all six reading frames and compare each translation with a protein sequence database. Thus, in a similar fashion to that employed by the **Ensembl** pipeline, protein coding regions of the genomic DNA can be identified. For clarity, we will use only the well annotated human proteins of the **SwissProt** section of **Uniprot**. First go to the home of **blast** at:

`http://blast.ncbi.nlm.nih.gov/Blast.cgi`

In the **Basic BLAST** section, select **blastx**. Use the **Enter Query Sequence Browse** (or **Choose File**) button to **upload file pax6_genomic.fasta**.

In the **Choose Search Set** section, set the **Database** to **UniProtKB/Swiss-prot prot(swissprot)**. Specify the **Organism** as **Human**.



Take a look at the **Algorithm parameters**[89].

The **Word size** choice is **2, 3** or **6** . The default is **6**. We seek very close matches here, so the largest **Word size** would seem appropriate.

The default scoring matrix is **BLOSUM62**, but choices from both the **BLOSUM** and **PAM** families are offered.

Low complexity regions will be filtered by default.

The **Max matches in a query range** is worth knowing about. Use the appropriate button to discover more.

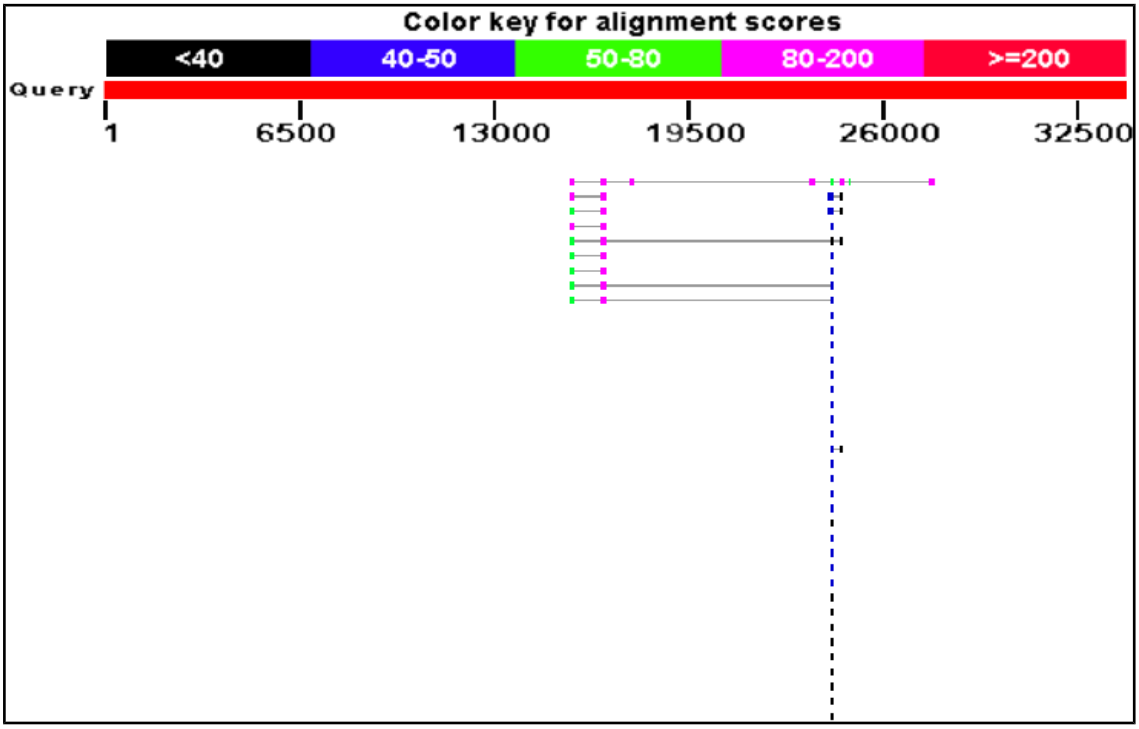
In what circumstances would you imagine that the **Max matches in a query range** parameter might be set to something other than its default value of **0**? _____

_____

_____

Change nothing other than to ask **blast** to **Show results in a new window**. and click the **BLAST** button.

After minimal thought, **blastx** will thrust its conclusions before you. Hover over the graphical hits for identification.



What are the **9** stronger matches around base position **16,000**?_____

Why would you expect exactly **9** matches around this point?_____

What do you make of the plethora of matches around **24,000**?_____

89 Here I will assume we have talked about these parameter and you are reasonably well informed of the issues.

Move down to the textual list of the matches. Hopefully as you fully expected you will find the expected number of **Paired box** matches at the top of the list followed by many many **Homeobox** matches.

Alignments Download GenPept Graphics

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	RecName: Full=Paired box protein Pax-6; AltName: Full=Aniridia type II protein; AltName: Full=Oculc	160	767	3%	2e-40	97%	P26367.2
☐	RecName: Full=Paired box protein Pax-2	131	214	1%	8e-31	74%	Q02962.4
☐	RecName: Full=Paired box protein Pax-8	131	208	1%	1e-30	76%	Q06710.2
☐	RecName: Full=Paired box protein Pax-5; AltName: Full=B-cell-specific transcription factor; Short=BS	128	211	1%	6e-30	74%	Q02548.1
☐	RecName: Full=Paired box protein Pax-4	117	258	1%	2e-26	67%	O43316.1
☐	RecName: Full=Paired box protein Pax-9	112	179	1%	5e-25	69%	P55771.3
☐	RecName: Full=Paired box protein Pax-1; AltName: Full=HuP48	111	177	1%	4e-24	69%	P15863.4
☐	RecName: Full=Paired box protein Pax-3; AltName: Full=HuP2	107	219	1%	7e-23	65%	P23760.2
☐	RecName: Full=Paired box protein Pax-7; AltName: Full=HuP1	105	217	1%	3e-22	68%	P23759.4
☐	RecName: Full=Retinal homeobox protein Rx; AltName: Full=Retina and anterior neural fold homeol	48.9	84.7	0%	1e-04	46%	Q9Y2V3.2
☐	RecName: Full=Retina and anterior neural fold homeobox protein 2; AltName: Full=Q50-type retinal	46.2	80.5	0%	2e-04	48%	Q96IS3.1
☐	RecName: Full=Homeobox protein aristaless-like 4	47.4	47.4	0%	4e-04	68%	Q9H161.2
☐	RecName: Full=Paired mesoderm homeobox protein 1; AltName: Full=Homeobox protein PHOX1; A	45.8	45.8	0%	7e-04	68%	P54821.2
☐	RecName: Full=Paired mesoderm homeobox protein 2; AltName: Full=Paired-related homeobox prc	45.8	45.8	0%	7e-04	68%	Q99811.2
☐	RecName: Full=Dorsal root ganglia homeobox protein; AltName: Full=Paired-related homeobox prot	45.8	45.8	0%	8e-04	71%	A6NNA5.1
☐	RecName: Full=Homeobox protein ARX; AltName: Full=Aristaless-related homeobox	46.6	46.6	0%	0.001	68%	Q96QS3.1

Why do you suppose the **Paired box** matches precede the **Homeobox** matches? _____

How do you suppose the **Max matches in a query range** parameter might be of value if this order was reversed? __

_____

Take a look at the alignments. You will see many places where regions have been filtered as non-informative. I suggest the one illustrated was filtered because it would match anywhere that was sufficiently **Serine** rich.

Score	Expect	Method	Identities	Positives	Gaps	Frame
81.3 bits(199)	5e-29	Compositional matrix adjust.	51/52(98%)	51/52(98%)	0/52(0%)	+3

```
Query  24654  FQVWFSNRRAKWRREEKLRNQRRQASNtpshipisssfstsVYQPIPQPTTP  24809
              QVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTP
Sbjct  254    IQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTP  305
```

How does this "non-informative" region match expectations suggested by **Prosite** and the **Feature table** of **Uniprot** for **PAX6_HUMAN**? _____

Primer Design

To determine the presence or absence of the mutation we have detected, a test based on restriction maps could be employed. This approach is investigated in one of the supplementary exercises at the end of this book. In that extra exercise, it is shown there is more than one restriction enzyme whose cut site is dependant upon the mutation. With a little more work (with the same programs), we could easily ascertain exact restriction fragment sizes expected for selected enzyme(s) with the mutation and without it. As long as the differences were sufficiently unsubtle, a **R**estriction **F**ragment **L**ength **P**olymorphism (**RFLP**) test could be designed.

For a variety of reasons, including the ready availability and low cost of sequencing, this is typically not the preferred way to proceed. It is normally preferable to use PCR to isolate the region around the mutation and sequence all individuals under examination. To do this, the first step would be to design suitable PCR primers. One program, in many different forms, is almost exclusively used for this purpose. The program is **primer3**. It is free and can be downloaded and run under linux and windows (at least). It is available as part of the **EMBOSS** package (**eprimer3**) and from a number of websites, including at the **M**assachusetts **I**nstitute of Technology (**MIT**)[90]:

`http://frodo.wi.mit.edu/`

This site is popular with many users offering complete control over the various options offered by **primer3**.

Another excellent **primer3** web interface (linked from the **MIT** site) developed in the Netherlands is available at:

`http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi`

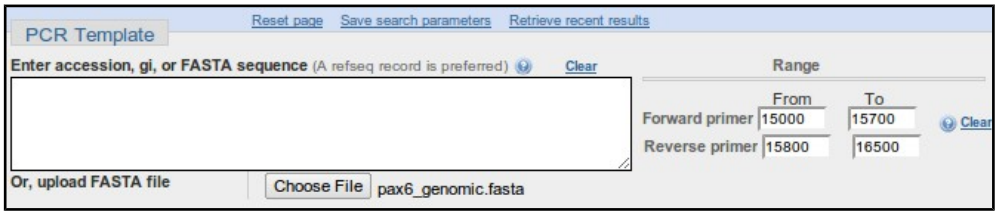
The site incorporates access to a **blast** search to check the uniqueness of the selected primers (important if unwanted PCR products are to be avoided).

Mostly because of its completely seamless inclusion of a **blast** search to compare potential primers with appropriate sequence collections, I suggest we here use **primer3** as implemented at the **NCBI**, even though it offers less than complete control over the execution of **primer3** itself. Go to:

`http://www.ncbi.nlm.nih.gov`

Click on the **BLAST** option. Select **Primer-BLAST** from the **Specialized BLAST** section.

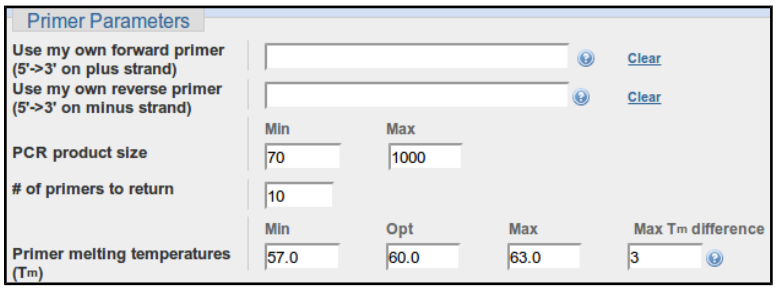
Upload your genomic **PAX6** sequence using the **Browse** (or **Choose File**) button for the **PCR Template**.



You have established that the mutation of greatest interest is the **G/C** substitution at position **15714** of the genomic sequence copied from **Ensembl**. It is logical therefore to specify that this feature be included in the PCR product not too near either end. Accordingly, request the **Forward primer** to be chosen **From** the region starting at base pair **15000** and continuing **To** base pair **15700**. Set the range for the **Reverse primer** to be **From 15800** and **To 16500**.

The default **PCR product size** is specified in the **Primer Parameters** section as between **70** and **1000** base pairs. This seems fine.

I would not presume to advise you on the melting temperatures that were most suitable[91]. For this exercise, the defaults work splendidly.



By default, **primer-BLAST** will report the best **10** primer pairs it can find (**# of primers to return**). This is plenty for the exercise.

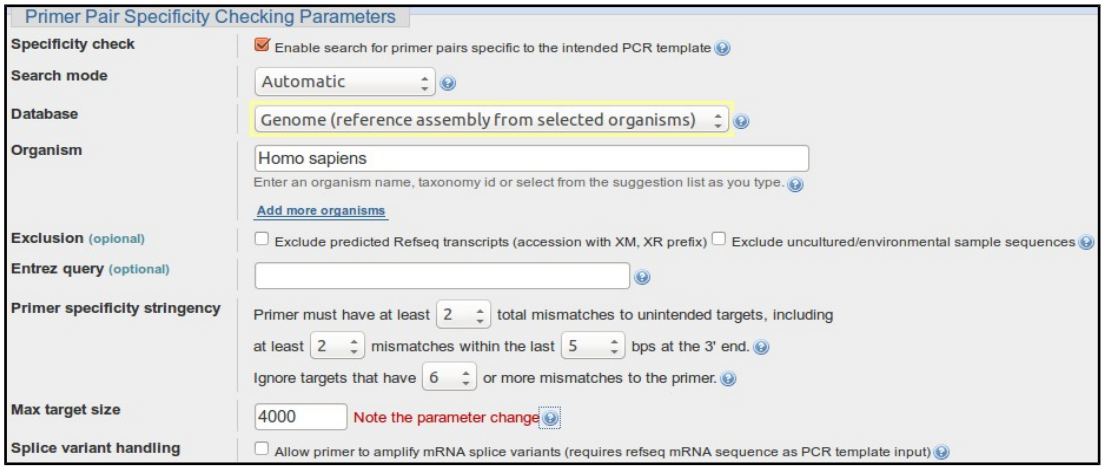
Do you think **10** primer pair suggestions is sufficient? If not, what number would you choose? _____

90 The **MIT** now offer a new version of **primer3** (**version 0.4.0**, soon **primer4** maybe?). Its URL is: **http://bioinfo.ut.ee/primer3-0.4.0/**. I have yet to investigate this version fully.

91 My policy has been to not discuss parameters that pertain to the experimental conditions. I think now that there are some such parameters that do need a little discussion. I will include this in future versions of these notes. In the mean time, I recommend going to the **MIT** site (or the **Wageningen** site) and making use of the very readable explanations linked from every parameter. The full **primer3** manual can be found here.

In addition to running **primer3** to suggest primers, **Primer-BLAST** checks against the possibility of unwanted PCR products by comparing potential primers against an appropriate sequence database with **blast**.

In the **Primer Pair Specificity Checking Parameters** section, set the **Database** selection to **Genome (reference assembly from selected organisms)**. Leave the **Organism** set as **Homo sapiens**.

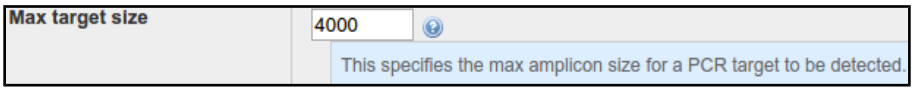


You thus request each potential pair of PCR primers to be compared to the entire human genome.

Unintended products, similar in size to the intended product, can so be identified.

The ideal conclusion is "just one product will be produced, on chromosome **11**, in the region of the **PAX6** gene".

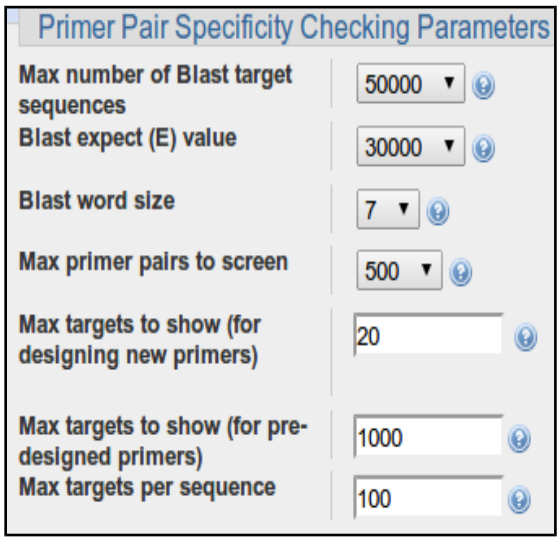
Use the appropriate button to discover the purpose of the **Max target size** parameter.



This is a new parameter replacing a very different parameter, the purpose of which was somewhat less obvious. The reason for the **Max target size** parameter is surely pretty transparent, so maybe there is now less requirement to wake up its button? For the present, the maximum size of any proposed PCR product, in this instance, is **1,000** base pairs (the form default). So the greatest size of an unwanted product that might be a problem (the **Max target size**) must be small enough to potentially be mistaken for a real product of **1,000** base pairs. **4,000** base pairs seems a bit cautious to me? However, unless you feel strongly about the matter, accept the default value of **4000**.

What value would you choose here if you were looking for uncluttered results? _____

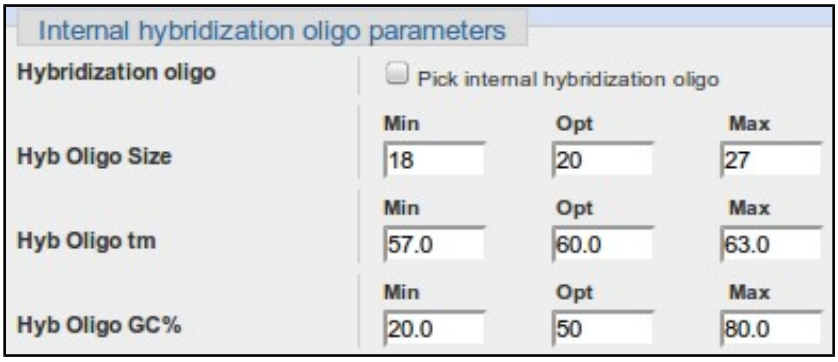
Before setting **primer-BLAST** going, click on the **Advanced parameters** button. Not really so **Advanced**? More **Avoidable** by those in a hurry. At the top are the **Primer Pair Specificity Checking Parameters** that control the way that **blast** is run. Note the buttons offering explanation.



Note the very high default **Blast expect (E) value**, suggesting you will be interested in matches with your primers that might occur up too **30000** times by chance! This does make sense as the primers will be very short and so many good, even exact, "chance" matches might be expected against a large database.

Comment upon the small default value for the **Blast word size**? _____

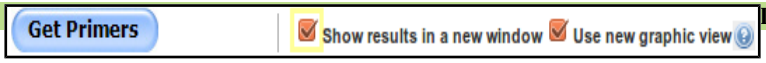
_____



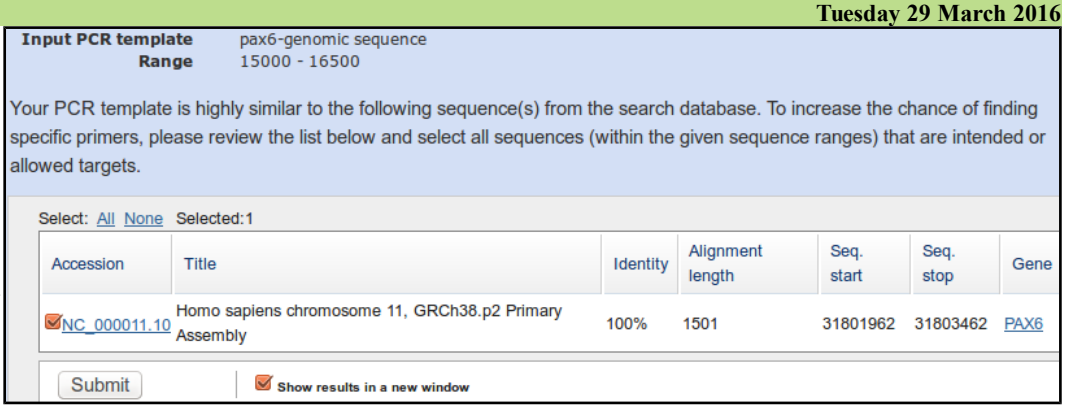
Note that you could get **primer-BLAST** to suggest an **Internal hybridisation oligo**, but decline the invitation this time.

Accept all the **Advanced parameters** as they are. Ask **primer-BLAST** to **Show results in a new window**.

Click on the **Get Primers** button.



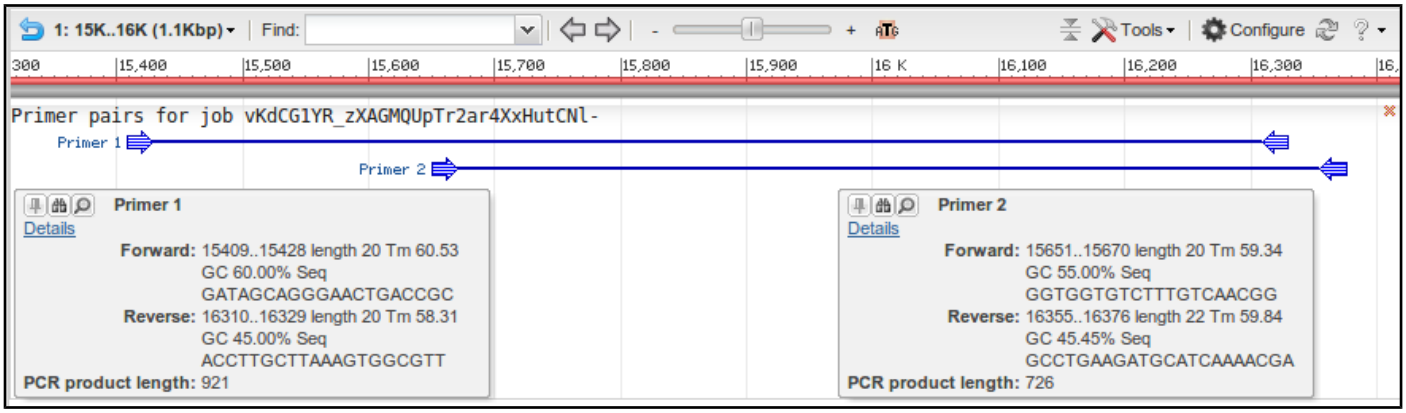
After a few moments of deep thought, **primer-BLAST** will notice that the template sequence you are using is **highly similar** (identical in fact) to part of an entry in the database being searched. Hardly surprising if one was to think about it.



You are invited to select all listed regions (just one this time) where matches with primers are likely to be the intended product. In this case, that is the whole list of one, so click on the **All** button.

Every pair of primers that **primer3** selects ***must*** match this region of **Chromosome 11** as it is precisely the region investigated by **primer3** in the first place. This process avoids **blast** reporting intended products as unintended products. Finally, all is ready, so ask to **Show results in a new window** and then click on the **Submit** button.

Once you have revelled in the opportunity to twiddle the fingers and scratch the ear(s) whilst **primers3** and **blast** go merrily about their appointed tasks, you will receive your results. These should look disarmingly like mine if all has gone well, in **Summary** and in **Detail**.



Primer pair 1

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	GATAGCAGGGAACTGACCGC	Plus	20	15409	15428	60.53	60.00	3.00	2.00
Reverse primer	ACCTTGCTTAAAGTGGCGTT	Minus	20	16329	16310	58.31	45.00	5.00	1.00
Product length	921								

Products on intended target

>NC_000011.10 Homo sapiens chromosome 11, GRCh38.p2 Primary Assembly

```
product length = 921
Features associated with this product:
   paired box protein Pax-6 isoform X6

   paired box protein Pax-6 isoform X1

Forward primer  1         GATAGCAGGGAACTGACCGC  20
Template        31803053  ....................  31803034

Reverse primer  1         ACCTTGCTTAAAGTGGCGTT  20
Template        31802133  ....................  31802152
```

Primer pair 2

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	GGTGGTGTCTTTGTCAACGG	Plus	20	15651	15670	59.34	55.00	5.00	3.00
Reverse primer	GCCTGAAGATGCATCAAAACGA	Minus	22	16376	16355	59.84	45.45	8.00	2.00
Product length	726								

Products on intended target

>NC_000011.10 Homo sapiens chromosome 11, GRCh38.p2 Primary Assembly

```
product length = 726
Features associated with this product:
   paired box protein Pax-6 isoform X6

   paired box protein Pax-6 isoform X1

Forward primer  1         GGTGGTGTCTTTGTCAACGG  20
Template        31802811  ....................  31802792

Reverse primer  1         GCCTGAAGATGCATCAAAACGA  22
Template        31802086  ......................  31802107
```

Just **two** solutions met the default criteria for success used by **primer3**. Up to **10** were permitted[92]. Hovering over the graphical results will bring forth textual summaries. Try it. Note the rather ugly job identification! Clearly, the poetry generated for your results is extremely unlikely to be the same as illustrated.

Neither of your two suggested primer pairs should be associated with any unintended products, even with the very generous suggestion that products **4000** bases long should be considered a potential problem[93].

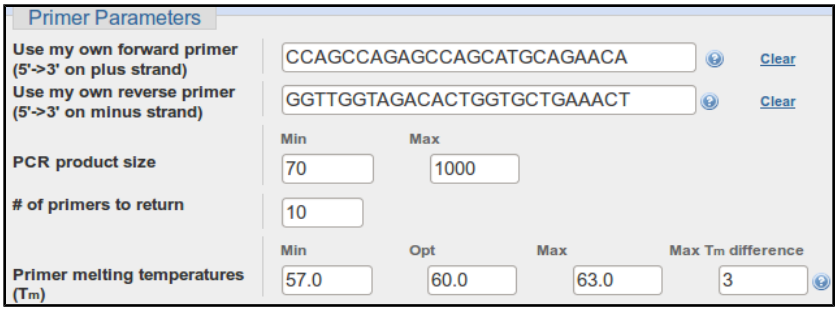
92 Which rather makes mock of all the deep thought employed deciding upon the most sensible maximum number of predictions to be reported.

93 This was not true until very recently. **Primer-BLAST** reported many more primer pair suggestions and quite a few unintended products for each. The previous parameter restriction the length of unintended products was somewhat more generous.

As well as suggesting primers for PCR (or other purposes) and (optionally) suggesting hybridisation oligos, **primer-BLAST** can be used to evaluate user-selected primers. Earlier, you saved a pair of primer sequences associated with **PAX6** when searching the nucleotide databases at the **NCBI**. It would be interesting to discover the product these might produce. To do this you need an unsullied **Primer-BLAST** page. Go again to:

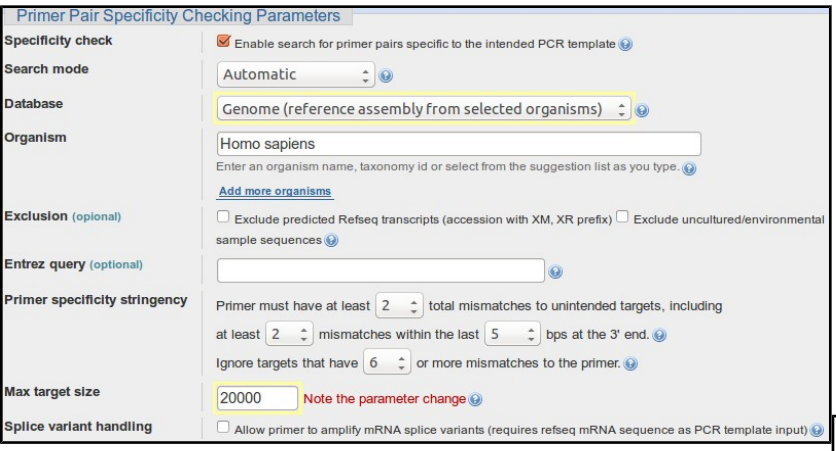
`http://www.ncbi.nlm.nih.gov`

Click on the **BLAST** option. Select **Primer-BLAST** from the **Specialized BLAST** section. Upload your genomic **PAX6** genomic sequence using the **Browse (or Choose File)** button for the **PCR Template**.



Open up the file you made containing the primers from **GenBank** (**pax6_primers.fasta**) in a text editor.

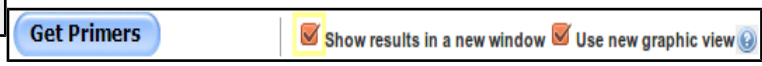
Copy and **Paste** the two primer sequences into the **Use my own forward primer** and **Use my own reverse primer** boxes as appropriate.



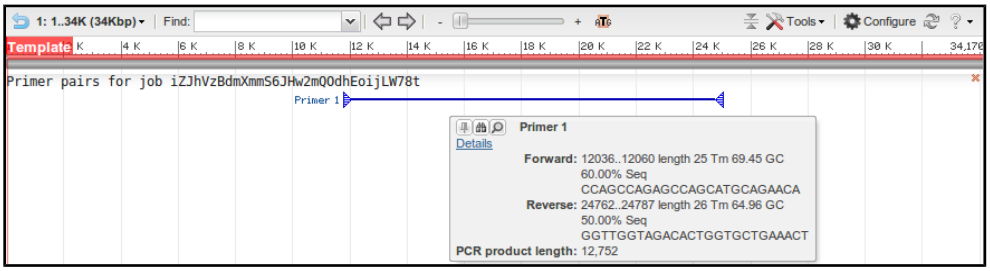
In the **Primer Pair Specificity Checking Parameters** section, set the **Database** selection to **Genome (reference assembly from selected organisms)**.

Leave the **Organism** as **Homo sapiens**.

Raise the **Max target size** parameter from **4000** to **20000**. You will be checking for enormous unintended unintended products with this run of **Primer-BLAST**. The reasons for this will become apparent soon.



Ask **primer-BLAST** to **Show results in a new window**.



Primer pair 1

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	CCAGCCAGAGCCAGCATGCAGAACA	Plus	25	12036	12060	69.45	60.00	6.00	0.00
Reverse primer	GGTTGGTAGACACTGGTGCTGAAACT	Minus	26	24787	24762	64.96	50.00	4.00	1.00
Product length	12752								

Products on potentially unintended templates

>NC_000011.10 Homo sapiens chromosome 11, GRCh38.p2 Primary Assembly

```
product length = 12752
Features associated with this product:
   paired box protein Pax-6 isoform X6

   paired box protein Pax-6 isoform X1

Forward primer  1         CCAGCCAGAGCCAGCATGCAGAACA  25
Template        31806426  .........................  31806402

Reverse primer  1         GGTTGGTAGACACTGGTGCTGAAACT  26
Template        31793675  .........................  31793700
```

Click on the **Get Primers** button.

After a short thrill filled pause, you will receive a result that should again looks more that a trifle like mine.

On the face of it, a fine match, an excellent result. Even the single **potentially unintended product** reported is actually the **intended product**. For some reason that is not immediately apparent to me, **Primer-BLAST** does not protect against discovering intended products when looking for unintended ones when examining specific primers[94]?

Success! However, applying a small measure of sober reflection, one has to wonder at a PCR product of **12,752** base pairs? I suspect that to be just a tad on the boastful side of probable[95]? Clearly, **primer-BLAST** is convinced, but mayhap a look at the references that came with these primer sequences would be advised before accepting this result at face value.

94 I have asked the guys at NCBI to explain. No full answer as yet, further prodding required.

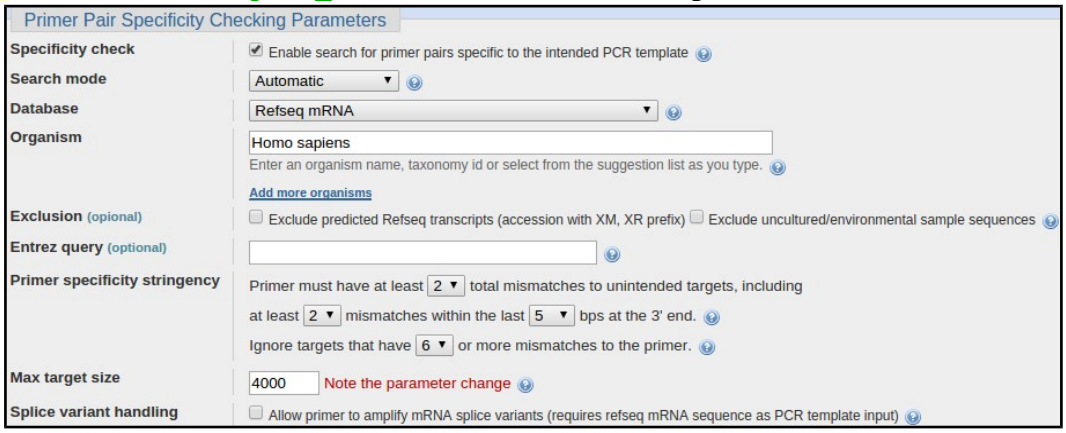
95 Apparently, such a PCR product is possible! However, above 5,000 base pairs would be slow, require very close attention and be prone to errors.

Unfortunately, the only paper referenced does not explain what might be going on particularly clearly. However, there is a hint that the primers you saved were designed for use with mRNA/cDNA data. Therefore it might be interesting to run **primer-BLAST** one last time with **pax6_cdna.fasta** as the **PCR Template**.

Simply move back to your last **primer-BLAST** launch page. This time, load **pax6_cdna.fasta** as the **PCR Template**.

In the **Primer Pair Specificity Checking Parameters** section, set the **Database** selection set to **Refseq mRNA** and leave the organism set to **Homo sapiens**.

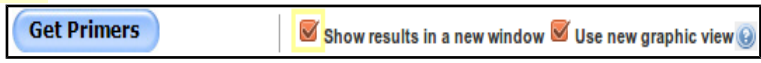
Set the **Max target size** back to its default value of **4000**, you should expect much smaller mRNA products this time, so no need for extending this maximum beyond **4000**.

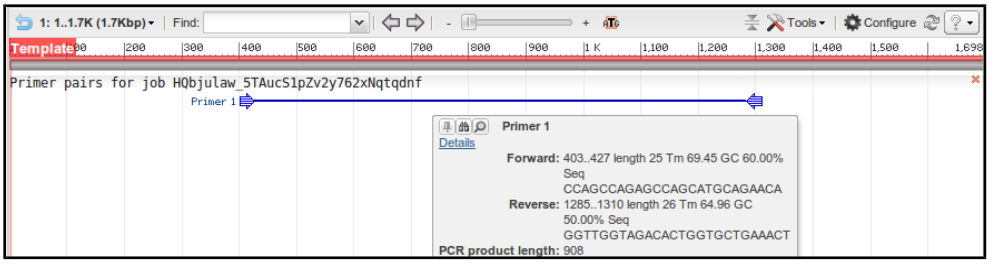


These selections suppose that the design of PCR product was for selection from a library of all human cDNAs.

As ever, ask **primer-BLAST** to **Show results in a new window**.

Click on the **Get Primers** button.





Primer pair 1

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	CCAGCCAGAGCCAGCATGCAGAACA	Plus	25	403	427	69.45	60.00	6.00	0.00
Reverse primer	GGTTGGTAGACACTGGTGCTGAAACT	Minus	26	1310	1285	64.96	50.00	4.00	1.00
Product length	908								

The result is a much more reasonable **Product length** of just **908** base pairs,reinforcing the theory that these primers were indeed designed for use with a cDNA library.

Products on potentially unintended templates

>NM_001310159.1 Homo sapiens paired box 6 (PAX6), transcript variant 9, mRNA

```
product length = 908
Forward primer  1     CCAGCCAGAGCCAGCATGCAGAACA  25
Template        114   .........................  138

Reverse primer  1     GGTTGGTAGACACTGGTGCTGAAACT  26
Template        1021  .........................  996
```

>NM_001310158.1 Homo sapiens paired box 6 (PAX6), transcript variant 8, mRNA

```
product length = 950
Forward primer  1     CCAGCCAGAGCCAGCATGCAGAACA  25
Template        496   .........................  520

Reverse primer  1     GGTTGGTAGACACTGGTGCTGAAACT  26
Template        1445  .........................  1420
```

>XM_006718246.2 PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X11, mRNA

```
product length = 707
Forward primer  1     CCAGCCAGAGCCAGCATGCAGAACA  25
Template        457   .........................  481

Reverse primer  1     GGTTGGTAGACACTGGTGCTGAAACT  26
Template        1163  .........................  1138
```

>XM_011520152.1 PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X10, mRNA

```
product length = 749
Forward primer  1     CCAGCCAGAGCCAGCATGCAGAACA  25
Template        457   .........................  481

Reverse primer  1     GGTTGGTAGACACTGGTGCTGAAACT  26
Template        1205  .........................  1180
```

>XM_005252956.3 PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X9, mRNA

```
product length = 908
Forward primer  1     CCAGCCAGAGCCAGCATGCAGAACA  25
Template        876   .........................  900

Reverse primer  1     GGTTGGTAGACACTGGTGCTGAAACT  26
Template        1783  .........................  1758
```

Before moving on, afford a quick glance at the report offered concerning possible unintended products. Here **primer-BLAST** warns against human mRNAs that might be cloned along with the intended target.

The first thing to note is that the intended target is not generated from a **RefSeq** mRNA. It comes from an mRNA taken from an **aniridia** patient directly. Therefore, there is no unintended product that we can ignore because it is really the intended product discovered by a different route, even though no filtering of the **RefSeq** database was undertaken.

All the unintended products could/would potentially be generated by the primers under investigation and have the potential to cause confusion. If you look down the list, you should conclude that the **16** unintended products come from **16** of the **24 RefSeq PAX6** transcripts first noted by **GeneCards**[96] and then confirmed later by **blast**.

96 Well, at least they will be noted by **GeneCards** as soon as they bring themselves up to date!

9 of the **11 NM_** good quality transcripts are detected. **7** of the **13** poorer quality **XM_** "**PREDICTED**" transcripts are also present. So **16** of the **24 PAX6** transcript sequences in **RefSeq** were detected.

Why do you suppose **blast** did not pick up all the transcripts? _____

_____

Note that the intended product is **908** base pairs long. Note that all the unintended products except two, near the top of the list are either **908** long or **950** long. A difference of **42**.

How would you tell quickly which isoform was represented by each mRNA listed here? _____

_____

Some fairly redundant questions to finish this section. I think I have already answered them all. But maybe you might wish to differ?

Is the number of "**potentially unintended products**" as you would you expect, given the evidence from **GeneCards**, **Ensembl** and **blast**?_____

_____

For all the "**potentially unintended products**", the selected primers match exactly. Can you explain this? _____

_____

The "**potentially unintended products**" are of different sizes. Can you explain the difference between the possible product lengths? ___

_____

Are the numbers of "**potentially unintended products**" of each possible length consistent with your **blast** results? _____

_____

>XM_005252955.3 PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X6, mRNA

```
product length = 908
Forward primer  1     CCAGCCAGAGCCAGCATGCAGAACA  25
Template        481   .........................  505

Reverse primer  1     GGTTGGTAGACACTGGTGCTGAAACT  26
Template        1388  .........................  1363
```

>XM_011520150.1 PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X6, mRNA

```
product length = 950
Forward primer  1     CCAGCCAGAGCCAGCATGCAGAACA  25
Template        366   .........................  390

Reverse primer  1     GGTTGGTAGACACTGGTGCTGAAACT  26
Template        1315  .........................  1290
```

>XM_011520149.1 PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X5, mRNA

```
product length = 950
Forward primer  1     CCAGCCAGAGCCAGCATGCAGAACA  25
Template        1275  .........................  1299

Reverse primer  1     GGTTGGTAGACACTGGTGCTGAAACT  26
Template        2224  .........................  2199
```

>XM_005252954.3 PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X4, mRNA

```
product length = 950
Forward primer  1     CCAGCCAGAGCCAGCATGCAGAACA  25
Template        457   .........................  481

Reverse primer  1     GGTTGGTAGACACTGGTGCTGAAACT  26
Template        1406  .........................  1381
```

>NM_001258465.1 Homo sapiens paired box 6 (PAX6), transcript variant 7, mRNA

```
product length = 908
Forward primer  1     CCAGCCAGAGCCAGCATGCAGAACA  25
Template        429   .........................  453

Reverse primer  1     GGTTGGTAGACACTGGTGCTGAAACT  26
Template        1336  .........................  1311
```

>NM_001258464.1 Homo sapiens paired box 6 (PAX6), transcript variant 6, mRNA

```
product length = 908
Forward primer  1     CCAGCCAGAGCCAGCATGCAGAACA  25
Template        443   .........................  467

Reverse primer  1     GGTTGGTAGACACTGGTGCTGAAACT  26
Template        1350  .........................  1325
```

>NM_001258463.1 Homo sapiens paired box 6 (PAX6), transcript variant 5, mRNA

```
product length = 950
Forward primer  1     CCAGCCAGAGCCAGCATGCAGAACA  25
Template        393   .........................  417

Reverse primer  1     GGTTGGTAGACACTGGTGCTGAAACT  26
Template        1342  .........................  1317
```

>NM_001258462.1 Homo sapiens paired box 6 (PAX6), transcript variant 4, mRNA

```
product length = 950
Forward primer  1     CCAGCCAGAGCCAGCATGCAGAACA  25
Template        455   .........................  479

Reverse primer  1     GGTTGGTAGACACTGGTGCTGAAACT  26
Template        1404  .........................  1379
```

>NM_001604.5 Homo sapiens paired box 6 (PAX6), transcript variant 2, mRNA

```
product length = 950
Forward primer  1     CCAGCCAGAGCCAGCATGCAGAACA  25
Template        443   .........................  467

Reverse primer  1     GGTTGGTAGACACTGGTGCTGAAACT  26
Template        1392  .........................  1367
```

>NM_000280.4 Homo sapiens paired box 6 (PAX6), transcript variant 1, mRNA

```
product length = 908
Forward primer  1     CCAGCCAGAGCCAGCATGCAGAACA  25
Template        541   .........................  565

Reverse primer  1     GGTTGGTAGACACTGGTGCTGAAACT  26
Template        1448  .........................  1423
```

>NM_001127612.1 Homo sapiens paired box 6 (PAX6), transcript variant 3, mRNA

```
product length = 908
Forward primer  1     CCAGCCAGAGCCAGCATGCAGAACA  25
Template        455   .........................  479

Reverse primer  1     GGTTGGTAGACACTGGTGCTGAAACT  26
Template        1362  .........................  1337
```

Gene Identification Software – Specifically Genscan 

Computational Protein Analysis

In this section, the plan is to look exclusively at the protein of **PAX6**. The object is to use various software items to confirm what has already discovered from the various web resources. Often the software you will use will be exactly that which was used to determine the pre-computed results you browsed previously.

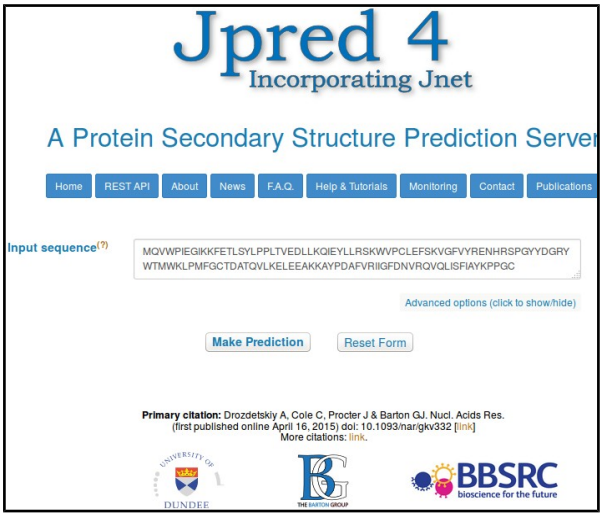
Predicting Protein Secondary Structure.

Beta strand	6 - 8	3
Beta strand	14 - 16	3
Helix	23 - 34	12
Helix	39 - 46	8
Helix	50 - 63	14
Beta strand	77 - 79	3
Helix	81 - 93	13
Helix	99 - 108	10
Turn	114 - 116	3
Helix	120 - 133	14
Helix	219 - 229	11
Helix	237 - 246	10
Helix	251 - 275	25

A first step is to look at to look at ways to predict the protein secondary structure of the **PAX6** protein. Evidence from various sources suggests that the **PAX6** protein has **9** helices arranged in triplets, plus a few beta strands.

To save time, I show here the relevant section from the **Uniprot Feature Table**. The helical triplets are involved in binding. **2** triplets are to be found in the paired box region, the other in the homeobox a little further along. Here we will try one of the most sophisticated methods available, to predict, essentially from from the primary sequence, what we already know. If you really wish, I also offer a supplementary exercise based around one of the earlier prediction methods, still used, but although faster, significantly less accurate than more modern methods.

Early Secondary Structure Prediction Methods - GOR



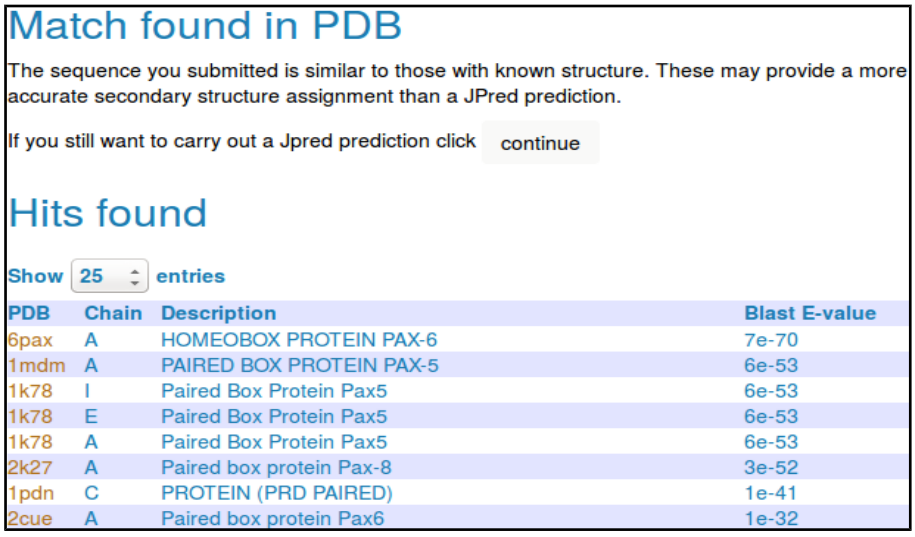
The service considered by many to offer the most effective method of predicting secondary structure is called **Jpred**. This is developed by the Barton group now located at Dundee University. Over **80%** accuracy is claimed for **Jpred** predictions. Due to the inherent imprecision in defining the end positions of secondary structure elements, **80%** is pretty much as good as is practically possible.

Go to the **Barton Group** web site at:

http://www.compbio.dundee.ac.uk

and follow the link to the Jpred 4 Incorporating Jnet server. Copy and paste the **PAX6** protein into the appropriate text box. Click on **Make Prediction**.

With alacrity, **JPred** will report several hits with proteins of known **3D** structure (using **blast**). Links are offered to a number of entries in the **PDB** structure database. At least **2** of the **PDB** entries listed should be familiar.



Jpred offers the suggestion that it really does not make sense to continue. After all, if the **3D** structure is effectively known, why predict (guess?) the **2D** structure? The answer to this challenge being a petulant "**Because we want to!**"

Click purposefully on the **Continue** button. **JPred**, with a small sigh of exasperation, will submit your job and let you know how busy it is. **Jpred** typically takes a while as it has much to consider[97].

Jpred will use **PSI-Blast** to align your sequence with all sequences deemed to be homologous, from a particularly appropriate database. **Jpred** can then make its structure predictions based on an aligned "family" of proteins, rather than just one individual sequence. Intuitively at least, this has to be a fine idea. A multiple alignment of related proteins will typically represent far more evidence for prediction than any single protein.

97 If the wait becomes unbearable, consider opening a new window/tab and moving on to the next section, returning to **Jpred** later.

JPred presents the results of running two secondary structure predictions, using the program **JNET**, based on two different representations of the alignment (**HMM** and **PSSM**, similar ideas that will be discussed at some point). Predicted helices are represented as red blocks, predicted beta sheets as green arrows. A consensus prediction is presented (**jnetpred**) as is an indication of prediction confidence (**JNETCONF**). Algorithms are also run to predict **coiled coils** (**Lupas**, with window sizes **21**, **14**, **28**). The first view of the results offered is a graphical overview aligned with your original single sequence.

The annotation bars below the alignment are as follows:

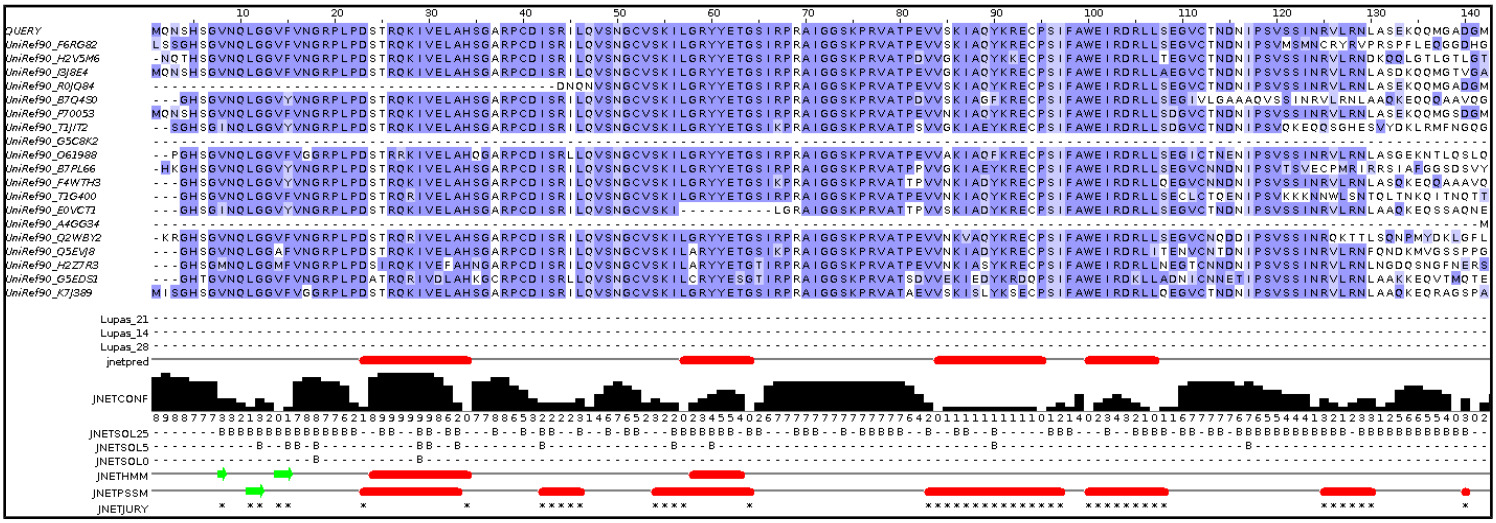
- Lupas_21, Lupas_14, Lupas_28
  Coiled-coil predictions for the sequence. These are binary predictions for each location.
- JNETSOL25,JNETSOL5,JNETSOL0
  Solvent accessibility predictions - binary predictions of 25%, 5% or 0% solvent accessibility.
- JNetPRED
  The consensus prediction - helices are marked as red tubes, and sheets as dark green arrows.
- JNetCONF
  The confidence estimate for the prediction. High values mean high confidence. prediction - helices are marked as red tubes, and sheets as dark green arrows.
- JNetALIGN
  Alignment based prediction - helices are marked as red tubes, and sheets as dark green arrows.
- JNetHMM
  HMM profile based prediction - helices are marked as red tubes, and sheets as dark green arrows.
- jpred
  Jpred prediction - helices are marked as red tubes, and sheets as dark green arrows.
- JNETPSSM
  PSSM based prediction - helices are marked as red tubes, and sheets as dark green arrows.
- JNETFREQ
  Amino Acid frequency based prediction - helices are marked as red tubes, and sheets as dark green arrows.
- JNETJURY
  A '' in this annotation indicates that the JNETJURY was invoked to rationalise significantly different primary predictions.*

The full key to all the abbreviations used (and more information about **JNet**) can be displayed by clicking on the **details on acronyms used** link.

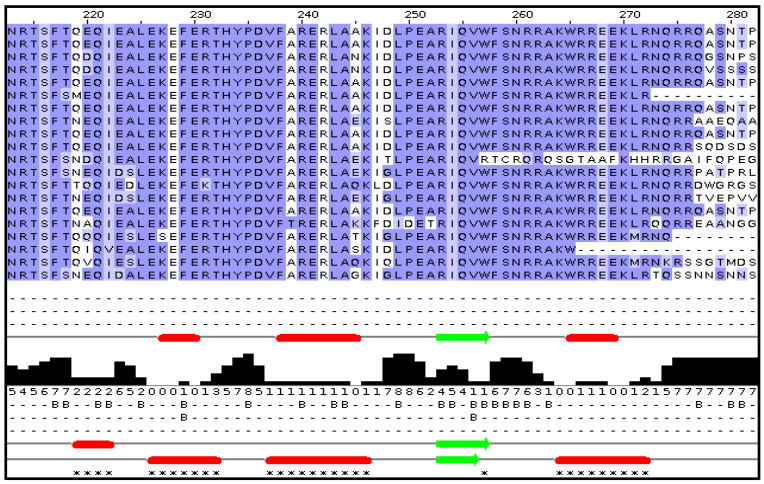
For a fuller view, elect to **View results in Jalview**[98]. You will arrive at a page inviting you to select from various viewing options. The options are explained clearly, but to save you time reading and pain deciding, I suggest you go for **Option1** for the clearest view. This option does not confuse the picture by gapping your query sequence (and thus making it more difficult to associate structure predictions with regions of the **PAX6** protein) and does not force you to look at the entire, huge, **MSA** generated by **PSI-Blast**.

Jalview presents something very similar to the original view of the **Jpred** results. This time though, the most significant part of the **PSI-Blast MSA** from which the predictions were computed is displayed.

Here I have included the **Jalview** version of the predictions around the **PAX** region



and those around the homeobox.

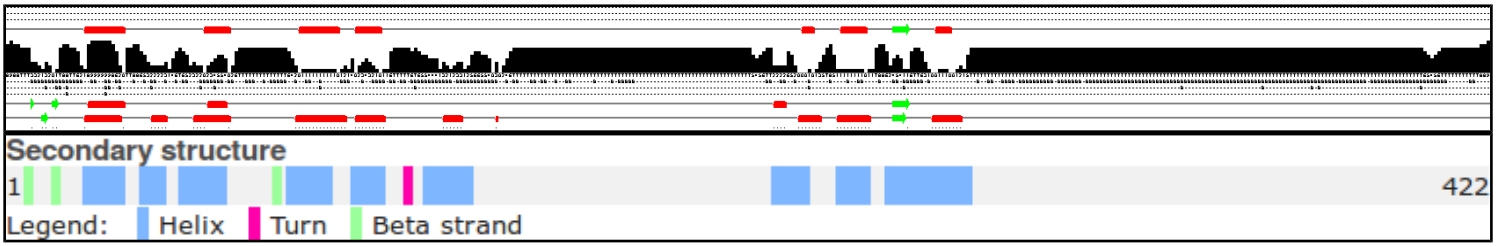


What protein database has **Jpred** chosen to search for protein sequences for the alignment upon which its predictions will be based?_____

Why do you suppose this database was used in preference to, say **UniprotKB**?_____

98 Should that not work, try **Full HTML**.

Also, I have lined up the entire prediction with the **Uniprot** Feature Table graphic as previously.



It would appear the helices predicted least confidently by **Jpred** are the same ones with which **GOR IV** (investigated in a supplementary exercise) had problems.

How would you rate the **Jpred** prediction overall?

Domain & Motif prediction.

You will have discovered from, several information sources, that the **PAX6** human protein has two DNA binding domains. A paired box at the **N terminal** and a homeobox a little further along. Both of the domains include **H**elix-**T**urn-**H**elix (**HTH**) motifs. In this exercise, I invite you to investigate how you might have discovered these domains and motifs using the various freely available domain databases (discussed previously) and other feature prediction programs. Clearly, this is superfluous for this particularly, well documented protein, but a valuable option in other circumstances.

One approach would be to consider each relevant domain database in turn. Each major domain database has its own Home web site and customised software to take **Query** protein sequences, compare those sequences with domain representations (typically based on **H**idden **M**arkov **M**odels) and to report convincing matches. This would work, but would be tedious as there are many viable databases to consider. It would be dangerous to rely on too few of the databases available as none is perfect. You need a consensus prediction to be sure you miss nothing.

Also, you would need to know which databases are particularly appropriate for each domain you considered might be present. All databases cannot be optimised for all types of domain (for example, the **SMART** database specialises only in domains that occur in signalling proteins).

So, let us not search individual domain databases in the main part of these exercises. Instead, I offer a supplementary exercise showing how you might search a representative selection individually for the domains you know to be present in **PAX6** human. I selected the **Prosite**, **Pfam** and **PRINTS** domain databases, If you get time to do this exercises, I particularly your attention to the **PRINTS** section. It illustrates how **PRINTS** just fails to see one of the domains that it should have found.

In this supplementary exercise also, I invite you to look at a simple **EMBOSS** program that tries (without complete success!) to discover **HTH** motifs.

Domain & Motif prediction with Prosite, Pfam, HTH & PRINTS 

Here, let us just look at using **Interpro** to do the whole job. **Interpro** will search for all domains using all the appropriate domain databases. Essentially, it takes all the tedium and decision out of using the miscellany of domain searching resources.

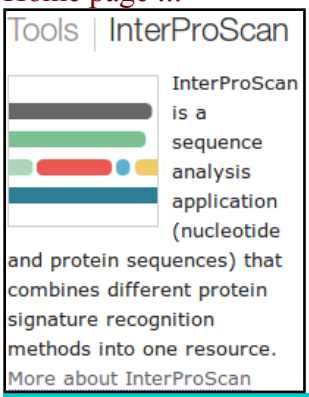
Searching 

Interpro defines protein families according to the way that proteins match elements of a wide range of protein family databases, including all those we have discussed thus far. **Interpro** provides a search tool that will search all or any of the major protein family databases and assign **Interpro** family associations to the query protein(s) accordingly. To have a look at some of the possibilities offered by **Interpro**, Go to:

http://www.ebi.ac.uk/interpro/

If you were to enter the **PAX6** human protein into the obvious place on the **InterPro** home page, you would produce almost exactly the results you saw many pages back, when you were looking at **GeneCards**. Do this if you have the time and inclination.

By implication, **InterPro** offers a fuller experience via the **InterProScan** search tool. Other than the opportunity not to search **ALL** the domain databases, and having the results arranged slightly differently, I am unsure what the extra effort brings? Never mind, there are many things of which I am unsure, so, from the **InterPro** Home page ...

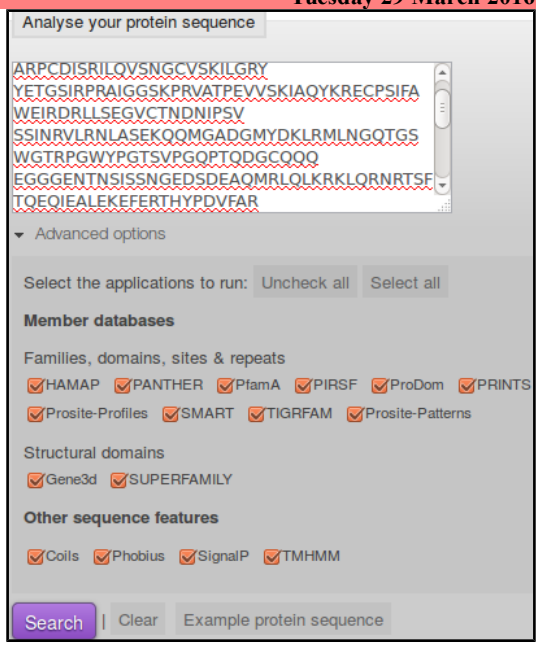


Select the **InterProScan** link. Here you will be offered the opportunity to download the **InterProScan** program.

I am not sure this is too useful an offer for most? But it is there.

For now, chose the online **Sequence search**.

You will arrive at a page that looks very similar to that from which you started, as far as the offer to run a domain search is concerned? Except! We now have **Advanced options**. Click on the **Advanced options**.



The **Advanced options** only allow you to choose which databases you wish to search and which feature prediction programs you wish to run. The default is to use all the databases and to run all the predictor programs. I struggle to imagine an occasion I would want to save the **EBI** servers a few cycles by considering which options to deselect, but it so nice to know I could if I wished to.

In passing, the offer to run the feature predictor programs in the **Other sequence features** section is relatively new. Of course, all these programs could be run individually from their home websites (follow the links behind the program names), in the same way as the domain databases can be searched individually. **Interpro** just aims to make thing easy for the user. The programs currently offered are:

- **Coils** is a program for predicting **coiled coils**. **Jpred** employs a program called **Lupas** to do this. Possibly the same program, given the author of **Coils** is someone called **Lupas, A**.
- **Phobius** & **TMHMM** are programs to predict **Transmembrane regions** (essentially **hydrophobic**, **uncharged** regions). There is no reason to expect any **Transmembrane regions** in this protein.
- **SignalP** predicts the presence and location of **signal peptide cleavage sites** in amino acid sequences from different organisms. I am pretty certain that there is no reason to expect signal peptides in this protein.

Do you think it a good idea for **Interpro** to offer feature prediction programs as well as domain database searches?

Paste the human **PAX6** sequence into the patiently waiting box. Accept the default to "***do everything***". Click on the **Search** button.

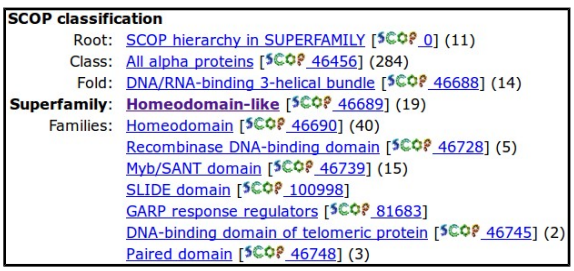
After several moments of deep thought, filtering and validating, you will be presented with a table of results looking very much like the one your saw when investigating **GeneCards**. There is, however, one important difference. In the **Unintegrated signatures** section, you will see that a **coiled coil** has been detected by the program **Coils**.

This was not included in the **GeneCards** information as **Interpro** has only recently included analysis using **Coils**. **GeneCards** will very probably catch up next time it is updated. It is also worth noting the **Jpred** does not predict a coiled coil anywhere in the protein?

The prediction matches the position of the first two helices of the homeobox. As you have seen, these lay along side each other.

Do you think the Coil prediction might be correct?

Notice that **Interpro** assigns both the **PAX** domain and the **Homeobox** domain of human **PAX6** to the **Interpro** family **Homeodomain-like**. Both of these associations are based on the hit behind the link **SSF46689**.

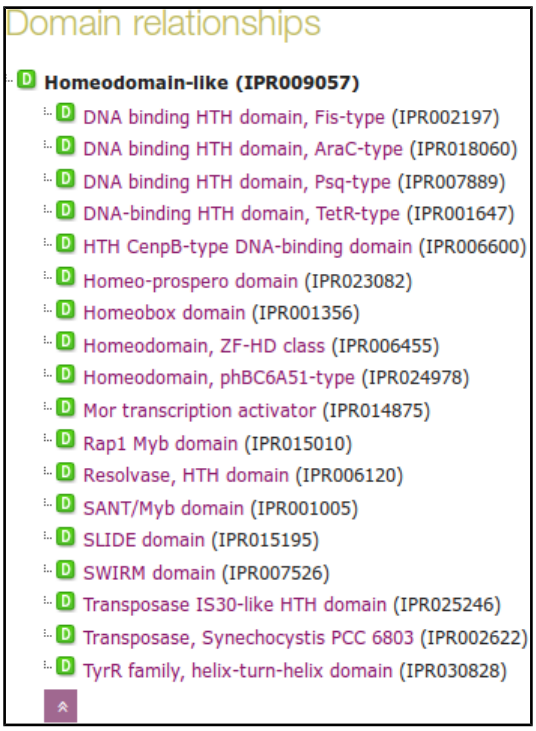


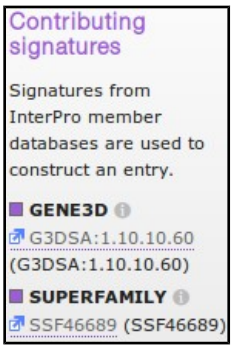
Follow this link and you will see it leads to the **Homeodomain-like superfamily** of the *Superfamily* 1.75 HMM library and genome assignments server database that specialises in very general ("at the **SCOP**[99] **superfamily** level) protein classifications. One **Superfamily** entry will typically correspond to a number of more specific domain definitions in other domain databases. Here you can see that the **Superfamily** domain **Homeodomain-like** includes both the **Homeodomain** & **Paired domain Families**.

99 **S**tructural **C**lassification **O**f **P**roteins, we will visit here later.

The **Gene3D** database is similar to **superfamily** but based on the **CATH** database[100]. It suggests the two **HTH** motifs of the paired box are both **Winged helix-turn-helix**. A different type of **HTH** to that of the **homeobox.** Possibly this is why the program **HTH** (see supplementary exercise) found the **homeobox HTH** easily, but really failed with the **paired box** features? **CATH** also appears to have an entry that includes the **homeobox**, but not the **paired box**?

Move back to your **InterProScan** results. Follow the link to the **Interpro** family **Homeodomain-like** (IPR009057). Click on the  button in the **Domain relationships** section to show the full list of **Homeodomain-like Interpro** domains.





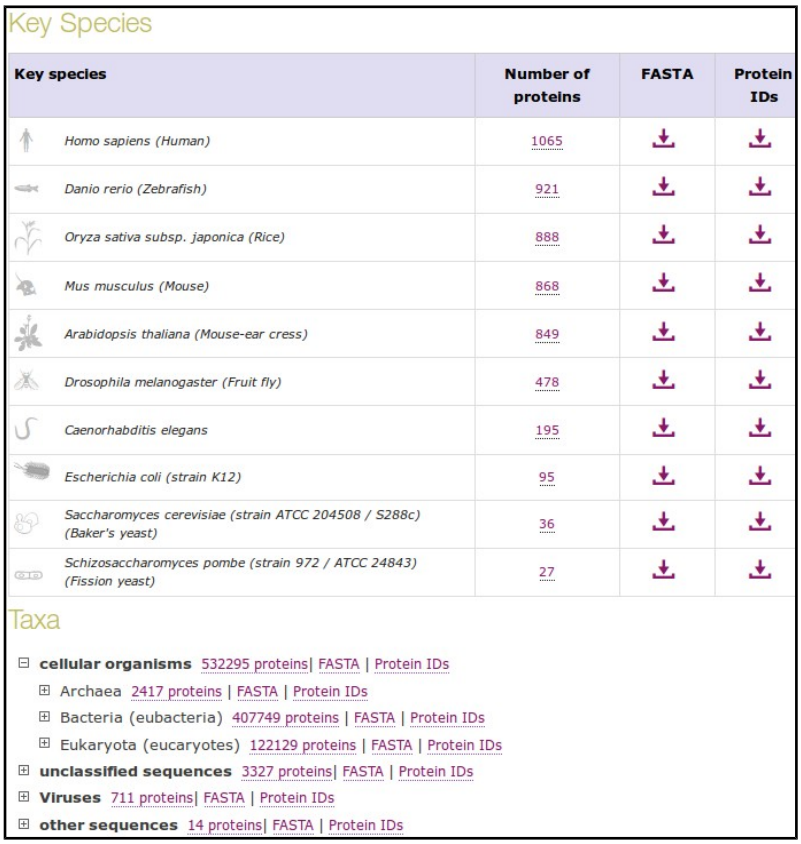
Note also the **Contributing signatures** in the top right hand corner of the page. Here is listed the domain databases that are searched to determine the presence of an **Interpro Homeodomain-like** domain.

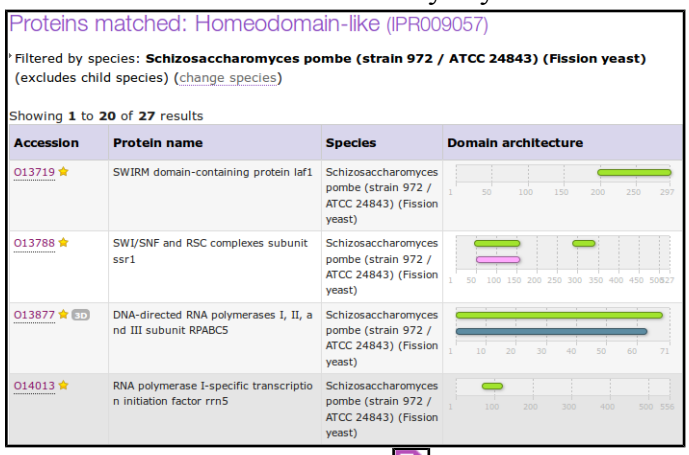
Essentially, if **Gene3D** finds a match with its **Homeodomain-rel** domain and/or **Superfamily** finds a match with its **Homeodomain-like** domain, then **Interpro** acknowledges a match with its **Homeodomain-like** domain (IPR009057).

None of the other domain databases **Interpro** searches are used to determine membership of (IPR009057).

To obtain an impression of how widely spread throughout nature is this domain. Click on the **Species** button on the left hand side of the page.

As you can see, this is a very popular domain. You can make this list enormous by injudicious employment of the expansion buttons. Why not? It amused me for a few moments anyway.





By clicking on the appropriate  button, you can get to either the protein sequences in **Fasta** format or list their accessions codes. Try a few, but be careful! It really does get you **ALL** the sequences, and that is often quite a lot, which will take time.

100 **CATH** is similar to **SCOP** in that it is another Structural classification database. We will look at it further, if and when I can ever make sense of it. It does also appear to have some parts that are not working properly? I investigate, I seek better persons than I to explain.

Multiple Sequence Alignment

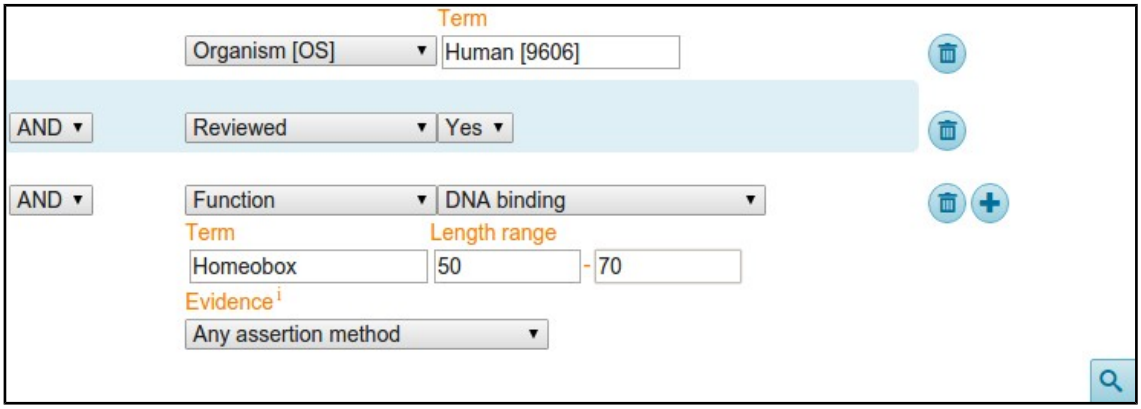
Here we will look at some software tools to align some protein sequences. Before we can do that, we need some sequences to align. I propose we try all the human **homeobox** domains from the well annotated section of **UniprotKB**. Getting the sequences is a trifle clumsy, so concentrate now! There used to be a much easier way, but that was made redundant by foolish people intent on making the future ever more tricky!!

So, begin by going to the home of **Uniprot**:

http://www.uniprot.org/

Choose the Advanced option of the Search button.

First specify that you are only interested in **Human** proteins. To do this, set the first field to **Organism [OS]** and **Term** to **Human [9606]**.



Set the second field selector to **Reviewed** and the corresponding **Term** to **Yes** (that is, choose to find only **SwissProt** entries).

Click on the button to request a further field selection option. Set the new field to **Function**. Set the type of **Function** to **DNA binding**. Set the **Term** selection to **Homeobox**.

From previous investigations, you should be aware that a **Homeobox** domain is **generally 60** amino acids in length. To avoid partial and/or really weird **Homeobox** proteins, set the **Length** range settings to recognise only **homeobox**s between **50** and **70** amino acids long.

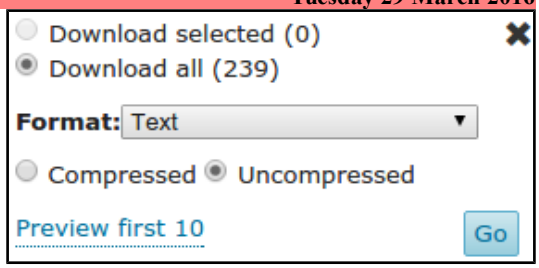
Leave the **Evidence** box as Any assertion method, one does not wish to be too fussy! Address the button with authority to get the search going.

BLAST | Align | Download | Add to basket | Columns — 1 to 25 of 239 — Show 25

Entry	Entry name	Protein names	Gene names	Organism	Length
Q9H2P0	ADNP_HUMAN	**Activity-dependent neuroprotector h...**	**ADNP**, ADNP1, KIAA0784	Homo sapiens (Human)	1,102
Q96G23	CERS2_HUMAN	**Ceramide synthase 2**	**CERS2**, LASS2, TMSG1	Homo sapiens (Human)	380
O43186	CRX_HUMAN	**Cone-rod homeobox protein**	**CRX**, CORD2	Homo sapiens (Human)	299
P39880	CUX1_HUMAN	**Homeobox protein cut-like 1**	**CUX1**, CUTL1	Homo sapiens (Human)	1,505
P35548	MSX2_HUMAN	**Homeobox protein MSX-2**	**MSX2**, HOX8	Homo sapiens (Human)	267
Q9H9S0	NANOG_HUMAN	**Homeobox protein NANOG**	**NANOG**	Homo sapiens (Human)	305
P43699	NKX21_HUMAN	**Homeobox protein Nkx-2.1**	**NKX2-1**, NKX2A, TITF1, TTF1	Homo sapiens (Human)	371
P52952	NKX25_HUMAN	**Homeobox protein Nkx-2.5**	**NKX2-5**, CSX, NKX2.5, NKX2E	Homo sapiens (Human)	324
P23760	PAX3_HUMAN	**Paired box protein Pax-3**	**PAX3**, HUP2	Homo sapiens (Human)	479
P26367	PAX6_HUMAN	**Paired box protein Pax-6**	**PAX6**, AN2	Homo sapiens (Human)	422

A fine miscellany of sequences will assemble upon you screen. Most seem to declare themselves in possession of a **Homeobox** or two (including **PAX6_HUMAN**), so I suggest a declaration of success.

Now save the entire list into a file using the Download button. Set the download to **uncompressed**. Make sure you have **all** sequences selected and that **Text** (i.e. **EMBL** or **SwissProt**) format selected. Press the **Go** button and do whatever it takes to ensure your results end up in a file residing on your **Desktop** called:



human_homeobox_proteins.emb

```
ID   ADNP_HUMAN              Reviewed;        1102 AA.
AC   Q9H2P0; E1P5Y2; O94881; Q5BKU2; Q9UG34;
DT   01-NOV-2002, integrated into UniProtKB/Swiss-Prot.
DT   01-MAR-2001, sequence version 1.
DT   22-JUL-2015, entry version 131.
DE   RecName: Full=Activity-dependent neuroprotector homeobox protein;
DE   AltName: Full=Activity-dependent neuroprotective protein;
GN   Name=ADNP; Synonyms=ADNP1, KIAA0784;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC   Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [MRNA].
RC   TISSUE=Fetal brain;
RX   PubMed=11013255; DOI=10.1074/jbc.M007416200;
RA   Zamostiano R., Pinhasov A., Gelber E., Steingart R.A., Seroussi E.,
RA   Giladi E., Bassan M., Wollman Y., Eyre H.J., Mulley J.C.,
```

Take a swift look at the file you have just created. Your neat list of **Human Homeobox** sequences will have transformed into a flood of **many SwissProt** format **UniProtKB** entries. Ugly, but what is required.

Search (**Control F**) for the term **DNA_BIND**.

It should occur many times (at least once per sequence) in the **F**eature **T**ables and most often refer to a **Homeobox** region.

In the **DNA_BIND** **F**eature **T**able entries, the position of the **Homeobox**es are recorded and will be used by the next program to isolate the sequence of the **Homeobox**es.

```
FT   CHAIN           1     430       Pre-B-cell leukemia transcription factor
FT                                   2.
FT                                   /FTId=PRO_0000049237.
FT   DNA_BIND      244     306       Homeobox; TALE-type.
FT                                   {ECO:0000255|PROSITE-ProRule:PRU00108}.
FT   COMPBIAS      137     145       Poly-Ala.
FT   MOD_RES       136     136       Phosphoserine.
FT                                   {ECO:0000269|PubMed:24275569}.
FT   MOD_RES       151     151       Phosphoserine.
FT                                   {ECO:0000269|PubMed:18220336,
FT                                   ECO:0000269|PubMed:18669648,
FT                                   ECO:0000269|PubMed:20068231}.
FT   MOD_RES       330     330       Phosphoserine.
FT                                   {ECO:0000269|PubMed:18669648,
FT                                   ECO:0000269|PubMed:21406692}.
FT   CONFLICT      393     393       M -> I (in Ref. 1; CAA42503).
FT                                   {ECO:0000305}.
SQ   SEQUENCE   430 AA;  45881 MW;  EF2FFA158C4DAF68 CRC64;
     MDERLLGPPP PGGGRGGLGL VSGEPGGPGE PPGGGDPGGG SGGVPGGRGK QDIGDILQQI
     MTITDQSLDE AQAKKHALNC HRMKPALFSV LCEIKEKTGL SIRSSQEEEP VDPQLMRLDN
     MLLAEGVAGP EKGGGSAAAA AAAAASGGGV SPDNSIEHSD YRSKLAQIRH IYHSELEKYE
     QACNEFTTHV MNLLREQSRT RPVAPKEMER MVSIIHRKFS AIQMQLKQST CEAVMILRSR
     FLDARRKRRN FSKQATEVLN EYFYSHLSNP YPSEEAKEEL AKKCGITVSQ VSNWFGNKRI
     RYKKNIGKFQ EEANIYAVKT AVSVTQGGHS RTSSPTPPSS AGSGGSFNLS GSGDMFLGMP
     GLNGDSYSAS QVESLRHSMG PGGYGDNLGG GQMYSPREMR ANGSWQEAVT PSSVTSPTEG
     PGSVHSDTSN
//
```

Now to extract from the whole protein sequences you have saved in a file, the sequences of just the **Homeobox** domains. One way of doing this (possibly not the best), is to use an **EMBOSS** package program called **extractfeat**. This can be found in many places, including the Bioinformatics server at **Wageningen** in the Netherlands. Go to:

http://emboss.bioinformatics.nl/

EDIT
aligncopy
aligncopypair
biosed
codcopy
cutseq
degapseq
descseq
entret
extractalign
extractfeat

Find the program **extractfeat** (in the **EDIT** section), and set it going.

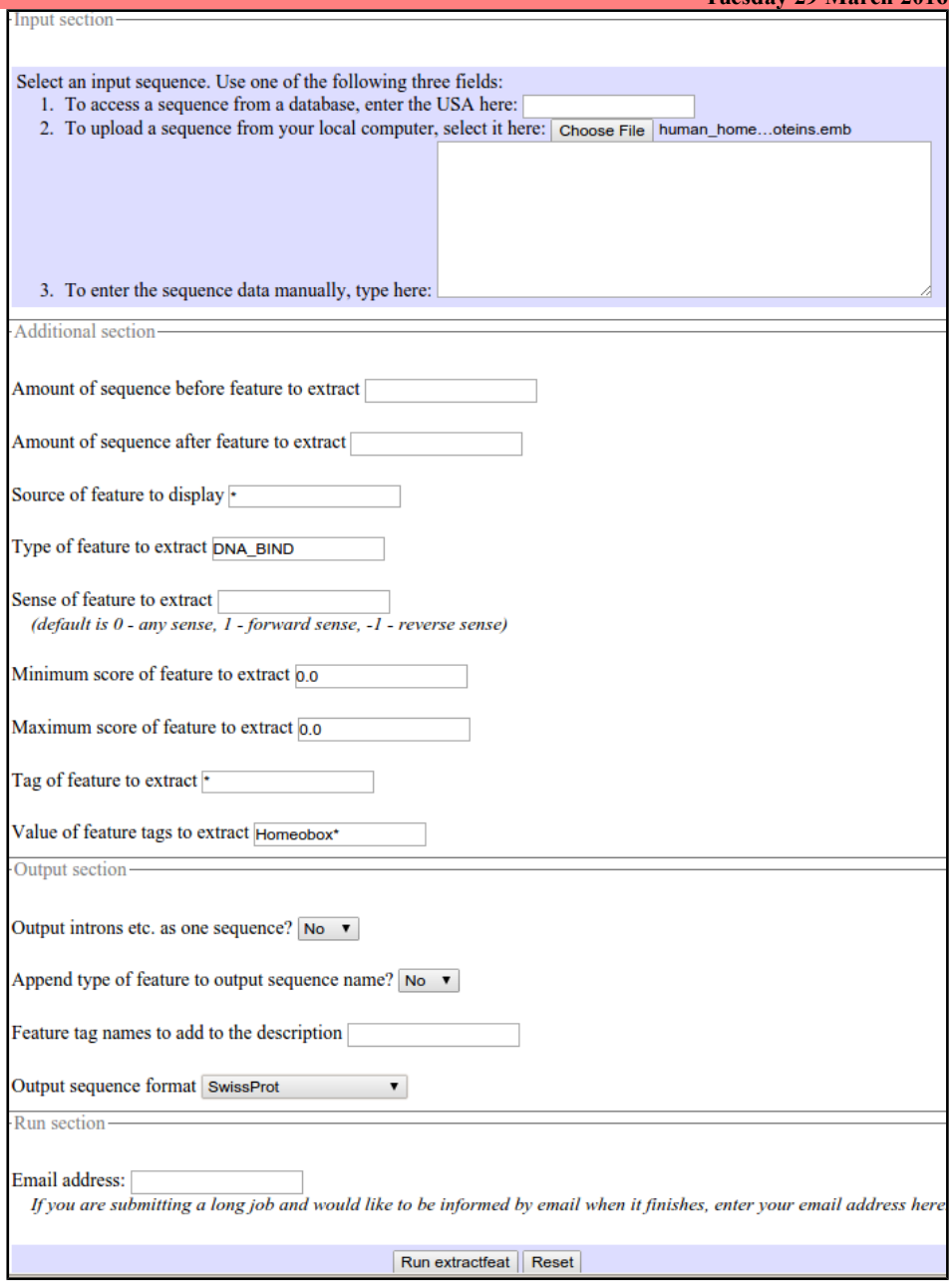
Use the **Choose File** button to **upload** the **SwissProt** format sequences from **UniProtKB** that you saved in **the file** `human_homeobox_proteins.emb`.

Set **Type of feature to extract field** to **DNA_BIND** (Make sure you remove the "*").

Set **Value of feature tags to extract** to **Homeobox*** (Make sure you append the "*" to ensure hits with, for example "**homeoboxes**).

Set the **Output sequence format** to **SwissProt** (**Fasta** would do, but **SwissProt** retains more annotation).

Click on the Run extractfeat button to start **extractfeat** going. Many sequences of **60** amino acids (or so) in length will leap into view.



OUTPUT FILE outseq

```
ID   ADNP_HUMAN_754_814     Unreviewed;         61 AA.
DE   [dna_bind] RecName: Full=Activity-dependent neuroprotector homeobox protein; AltName: Full=Activity-dependent neuroprotective protein;
SQ   SEQUENCE    61 AA;   7330 MW;  2A08F4F3C18785D8 CRC64;
     LDPKGHEDDS YEARKSFLTK YFNKQPYPTR REIEKLAASL WLWKSDIASH FSNKRKKCVR
     D
//
ID   CERS2_HUMAN_67_128     Unreviewed;         62 AA.
DE   [dna_bind] RecName: Full=Ceramide synthase 2; Short=CerS2; AltName: Full=LAG1 longevity assurance homolog 2; AltName: Full=SP260; AltName: Full=Tumor metastasis-suppressor gene 1 protein;
SQ   SEQUENCE    62 AA;   7373 MW;  B94732A59CA60B9F CRC64;
     LLNIKEKTRL RAPPNATLEH FYLTSGKQPK QVEVELLSRQ SGLSGRQVER WFRRRRNQDR
     PS
//
ID   CRX_HUMAN_39_98     Unreviewed;         60 AA.
DE   [dna_bind] RecName: Full=Cone-rod homeobox protein;
SQ   SEQUENCE    60 AA;   7369 MW;  B8E43274B30EBAC6 CRC64;
     QRRERTTFTR SQLEELEALF AKTQYPDVYA REEVALKINL PESRVQVWFK NRRAKCRQQR
//
```

Right click the outseq button and select **Save Link as...** . Do whatever it takes to save all your **Homeobox** domains into a file residing on your **Desktop** called:

`homeobox_human.emb`

Finally, we have some sequences with which to investigate the multiple sequence alignment programs.

Take a look at the file you have created. You should have many human **homeobox** domains in **SwissProt** format, looking rather as they did in your browser window. Happily **ClustalX**, the first multiple alignment program to be investigated, accepts multiple sequence **SwissProt** format files as input.

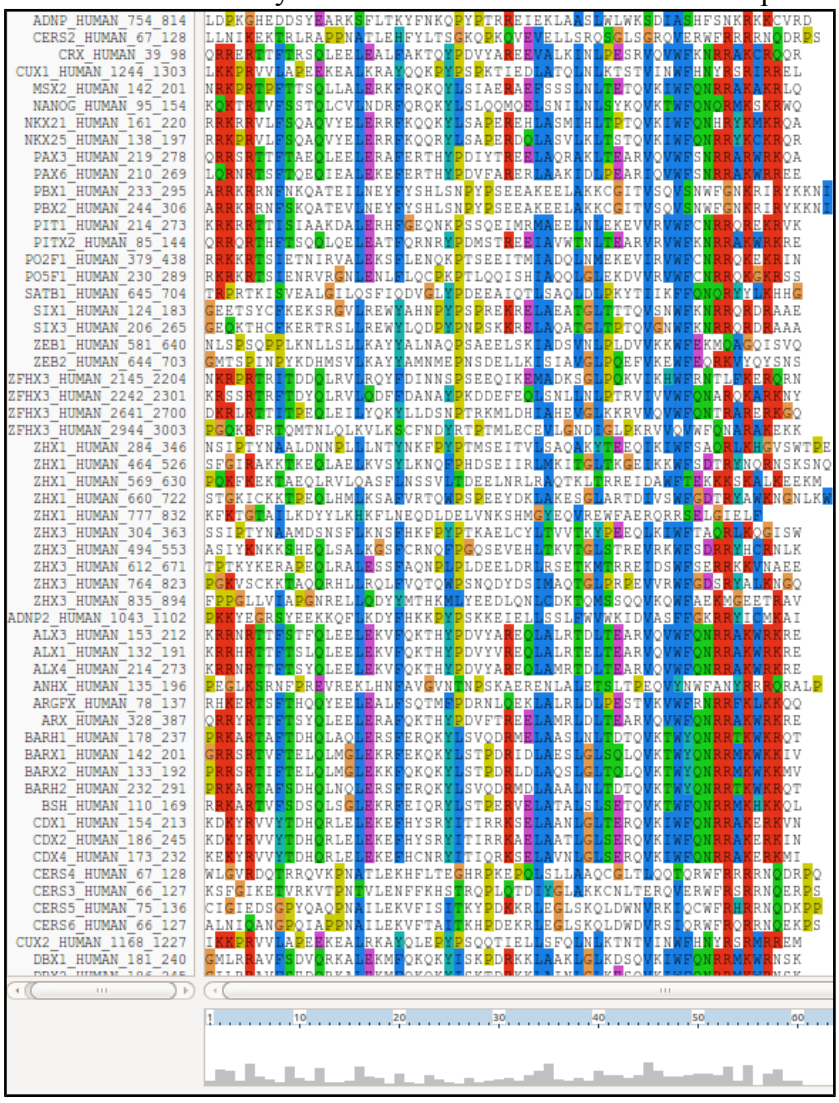
ClustalX is a part of the mostly widely known family of **m**ultiple **s**equence **a**lignments (**msa**) programs, originating in the **1980**s. Until relatively recently, it was the only real option. **ClustalX** still has merit, although it lacks some of the sophistication of more recent programs. **ClustalX** runs on effectively all workstations and has a nice **G**raphical **U**ser **I**nterface (**GUI**). A good place for us to start. It is installed on your workstations.

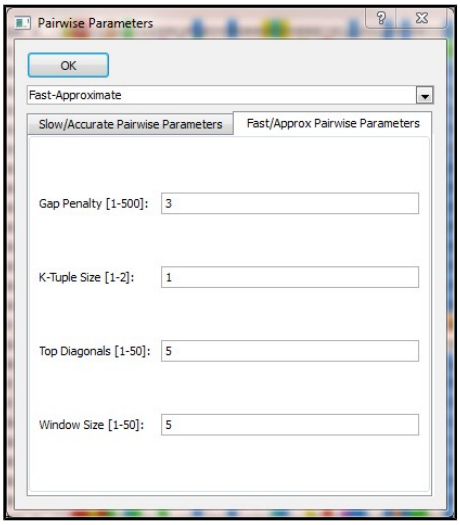
Start up the program **ClustalX**[101]. The **ClustalX** **G**raphical **U**ser **I**nterface (**GUI**) will regally mount your screen.

Select **Load Sequences** from the **File** pull down menu and load your file of **homeobox** domains.

The sequences will arrange themselves colourfully. Many of the **homeoboxes** are similar enough to look convincing even before alignment. Note the "Manhattan skyline" under the sequences indicating the degree of conservation.

You might like to increase the **Font** size from the minute default setting designed for Hawks and Eagles, to something more comfortable. **24** works tolerably well for me.





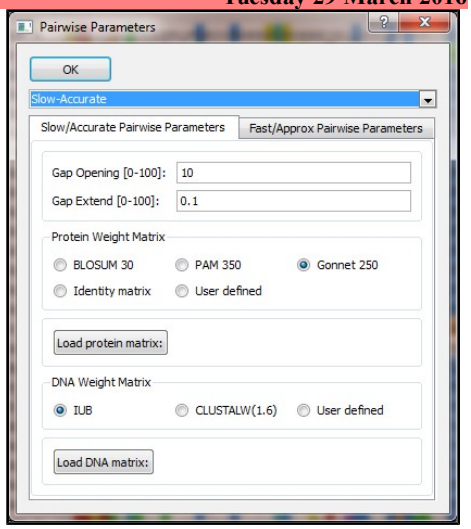
From the **Alignment** pull down menu, go to the **Alignment parameters** menu and select **Pairwise Alignment Parameters**. Just for a moment, change the setting from **Slow-Accurate** to **Fast-Approximate**. Bring the corresponding parameters into view by clicking on **Fast/Approx Pairwise Parameters** tab[102].

Hopefully, we will have discussed the way **ClustalX** (and similar multiple alignment tools) work. Intuitively, it should not make a lot of difference how the initial pairwise comparison stage is conducted. However, it very often does.

Specifically for this set of proteins, as well as generally, **ClustalX** will give a noticeably better alignment if the initial pairwise alignment stage is done carefully. Accordingly, reverse your whimsical setting change by moving back from **Fast-Approximate** to **Slow-Accurate**.

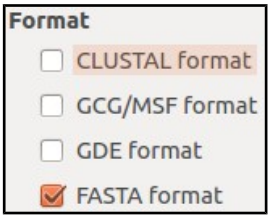
101 Of course, you could run **Clustal** from websites all over the world if you wished. Specifically, it is available both at the **EBI** and the Bioinformatics server at **Wageningen**. Try it if you have time. You get the same results but will, sadly, lose the pretty interface.
http://www.ebi.ac.uk/Tools/clustalw2/index.html
http://www.bioinformatics.nl/tools/clustalw.html

102 The **Fast-Approximate** algorithm is essential that which the database searching program **fasta** employs. Assuming we have discussed how **fasta** works, it should require no further explanation here.

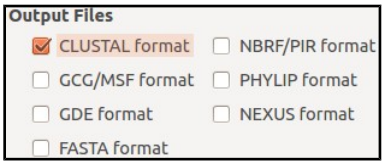


Click on the **Slow/Accurate Pairwise Parameters** tab for a final look at the default parameters to be used. The **Slow-Accurate** option is essentially a version of **Global Alignment** algorithm we will have discussed previously. Hopefully, all the parameter options will therefore be familiar to you.

I will assume both sets of parameters at least seem familiar? If not please ask. The default **Slow/Accurate Pairwise Parameters** you now have in view are fine. Click the **OK** button to dismiss the **Pairwise Parameters** window.



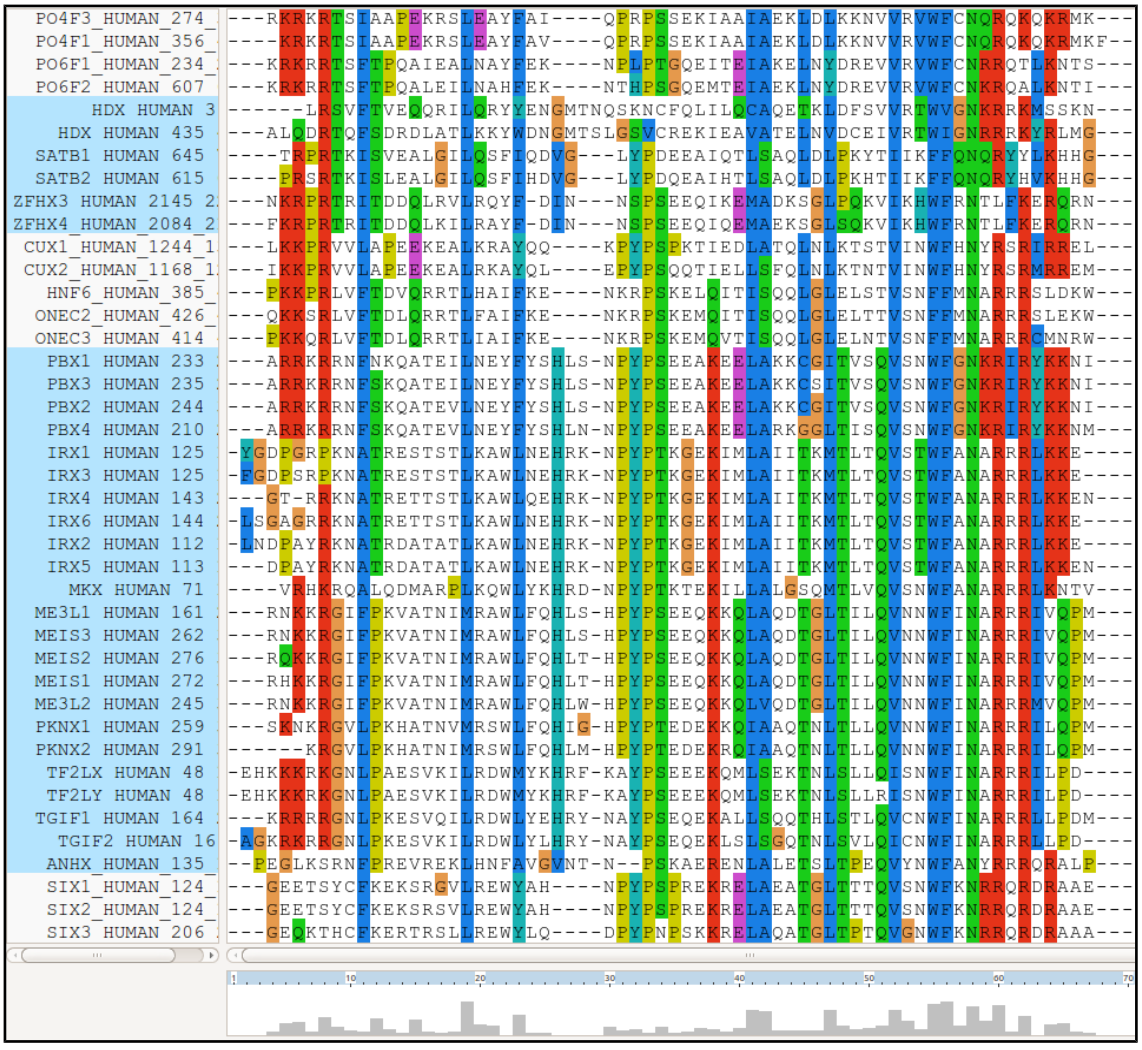
Before proceeding, save the **homeobox** sequences in **FASTA** format, which will better suit the other **msa** programs we will try. Do this by selecting **Save sequences as...** from the **File** pull down menu. Deselect **CLUSTAL format**, select **FASTA format**. Click **OK**. A file called `homeobox_human.fasta` will be created. Take a look to check it is as you would expect.

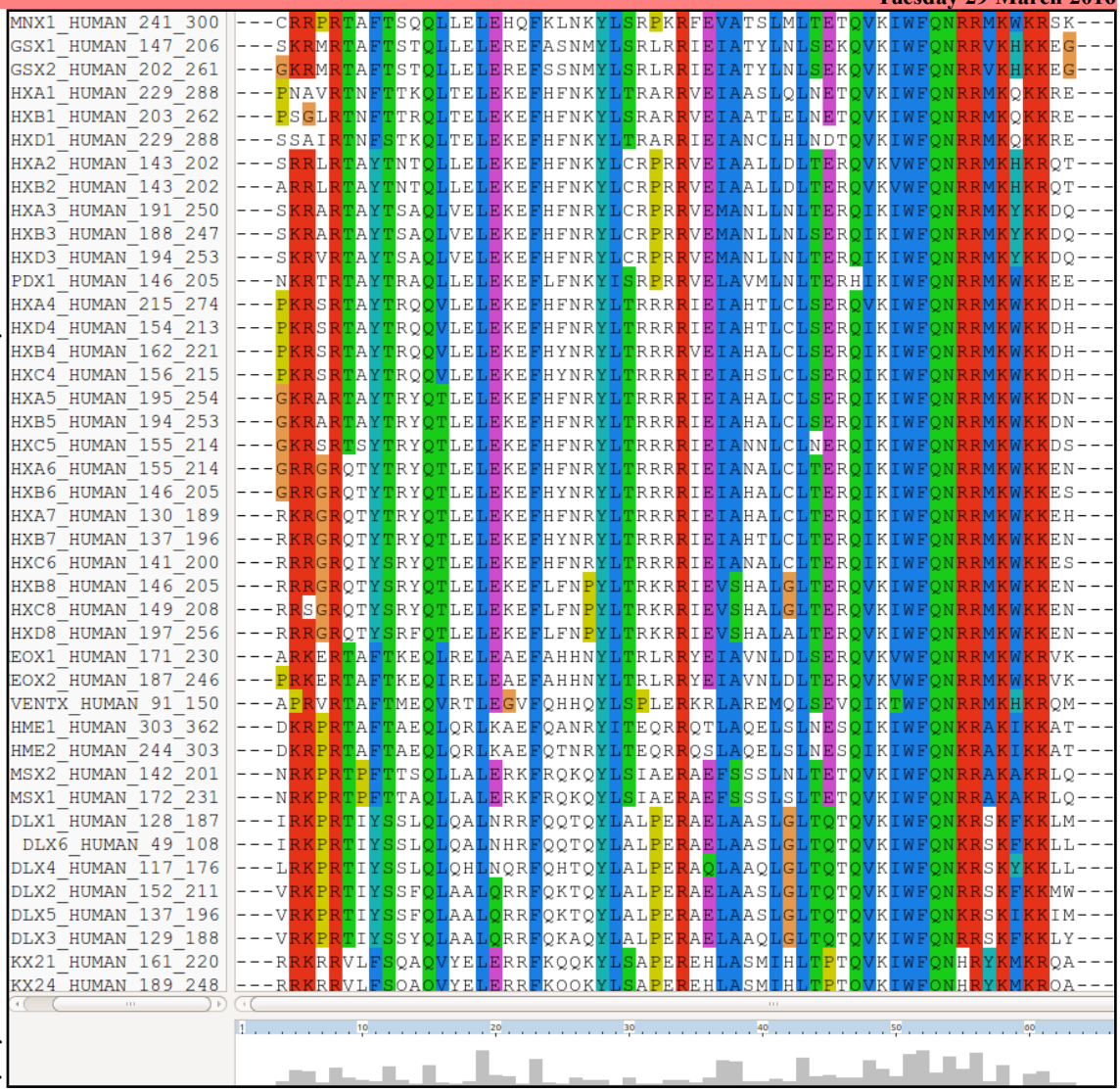


Strangely, saving your sequences in **FASTA** format convinces **clustalx** that it should now output its alignments in **FASTA** format. To prevent this, again select **Output Format Options** from the **Alignments** pull down menu. Deselect **FASTA format** and select **CLUSTAL format**. Click **OK**.

From the **Alignment** pull down menu, select **Do Complete Alignment**. Accept the default names for output files and click on the **OK** button. **ClustalX** will start to think deeply and eventually come up with it view of how the **homeobox** domains should be aligned.

Not a bad first try. From an entirely non scientific, cosmetic viewpoint, the ragged ends offend a trifle, as does the gap just before position **30**!





In reality, these features might be very interesting, but here I go for pretty!

So, just to investigate what is possible, select all the **homeobox** sequences that are causing the gap around position **30** by clicking on their names (quite a lot of them I fear). Hold the **Ctrl** key down to allow multiple selection.

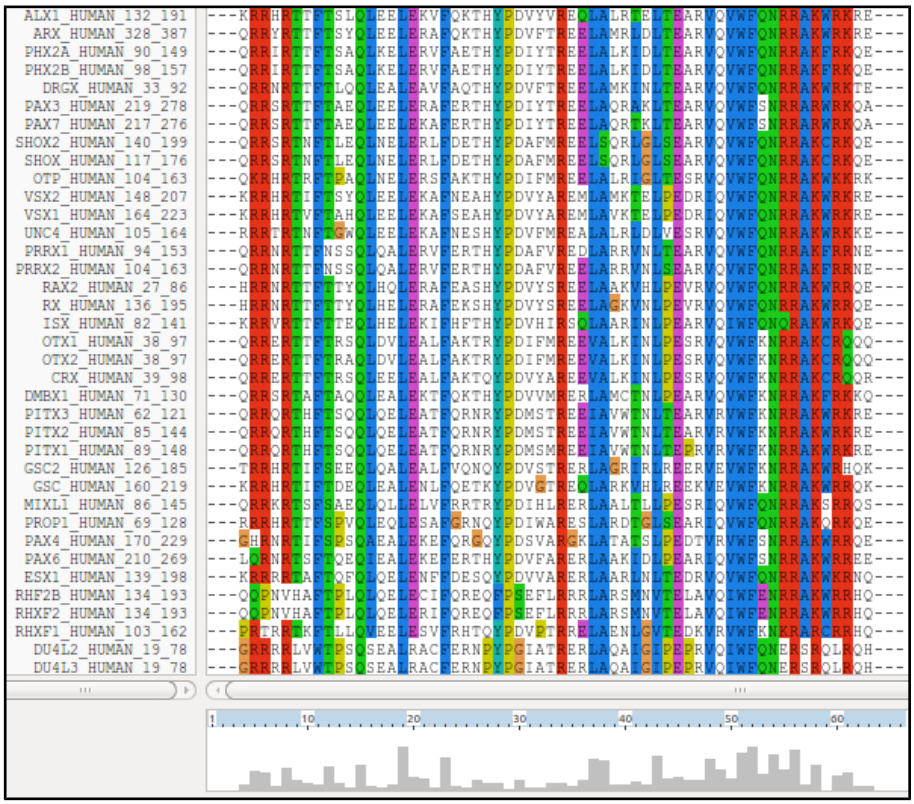
Once you have them all, go to the **Edit** pull down menu and select **Cut Sequences**. Then select **Remove Gap-Only columns** from the **Edit** pull down menu. Nasty gap gone … along with all scientific credibility, but … never mind.

You could recompute the alignment from scratch with the reduced sequence set, but you should end up with the same answer, of course. Just for the sake of it, select **Select All Sequences** from the **Edit** pull down menu. Then select **Remove All gaps** from the **Edit** pull down menu and confirm your intentions. You are now back where you started, but without the sequences that mess up the alignment intolerably!

Save your filtered set of sequences in a file. From the **File** pull down menu select **Save Sequences as...** . Choose **FASTA** format only.

The default file name is OK, even though it will overwrite the original sequences. I am convinced the sequences eliminated would not be aligned convincingly with any of the tools we have at hand. Let us lose them! Press the **OK** button.

From the **Alignment** pull down menu, select **Output Format Options** and select **CLUSTAL format** only. Again, from the **Alignment** pull down menu, select **Do Complete Alignment**. Accept the default names for the output files. This will overwrite your previous efforts, but no matter. More deep thought. Well, I got back to where I was, no gaps around position **30** but still with ragged ends!



You will recall from earlier that the **Prosite** pattern for a **homeobox** is the ever memorable:

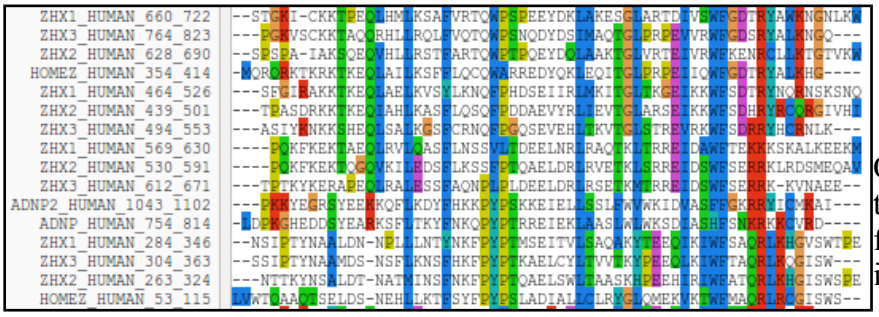
[LIVMFYG]-[ASLVR]-x(2)-[LIVMSTACN]-x-[LIVM]-{Y}-x(2)-{L}-[LIV]-[RKNQESTAIY]-[LIVFSTNKH]-W-[FYVC]-x-[NDQTAH]-x(5)-[RKNAIMW]

This corresponds to positions **37** to **60** in this alignment. See that the "Manhatten Skyline" is encouraging in the parts of this region that matter.

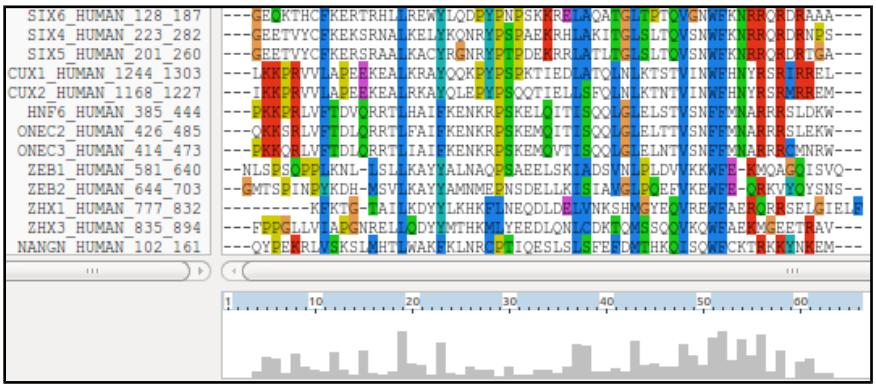
Note that the profile **Tryptophan**, in position **51**, is **very** consistent, but not quite **100%** as suggested by the **Prosite** pattern[103]. The **W** was even conserved in the sequences that were cosmetically removed.

Position **53** is not conserved ("**-x-**") according to the **Prosite** pattern. In the alignment segment offered here, it looks like a pretty consistent **Q**. However, the "**Manhattan skyline**" at this position is very low, suggesting that the sequences in view might not be typical of the whole alignment set. Which, upon checking …. they are not!

Looking through this alignment, I get the feeling I could design a better, stricter pattern for the region between **37** and **60**. Possibly true, but remember the pattern in **Prosite** aims to represent the conservation of **Homeobox** domains in **ALL** organisms. Here we have only sequences from **Human**.



Of course, things are not quite so convincing throughout. If you look at the top and bottom few sequences, you will see that **ClustalX** had its moments of uncertainty.



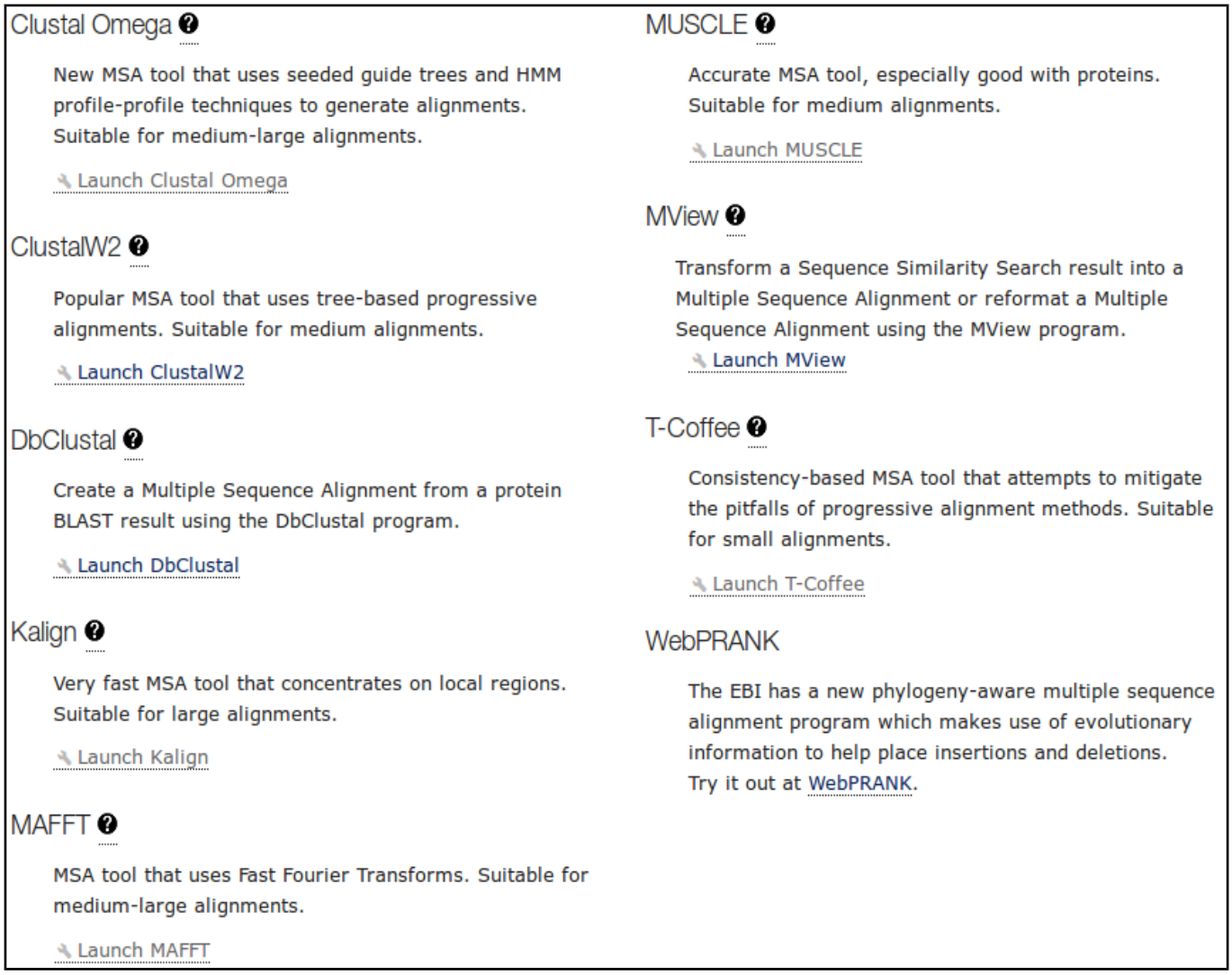
Note, however, the consistent **W** in position **50** despite the surrounding crumble.

103 From the "**Manhattan Skyline**", you can see the conservation is less than **100%**. Less conserved than the **F** that immediately follows in fact? Look at your alignment, the "**Manhattan Skyline**" does not seem to reflect reality? The **W** is **very** well conserved, although the scoring matrices would regard any deviation from **W** as serious? I need to find out more about how the **Skyline** is computed.

Now to show existence of some **msa** program options available on the web. There are many. They are available from a number of server sites. An obvious place to start has to be the **EBI** page dedicated to **MSA**. Go to:

`http://www.ebi.ac.uk/Tools/msa/`

Offered here is an impressively full selection of popular, current generation **msa** tools. Each is accompanied by advice to guide the choice of tool to best fit the circumstances. Each tool is provided with a link to its **Launch** interface. All the **Launch** interfaces are very consistent. Once you have run one of the **msa** options, you should have no trouble running any of the others.

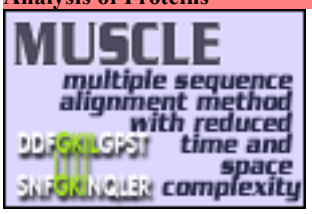


Here I intend to align again the human **homeboxes** with just one of the tools on offer. Then take a quick look at how the machine generated multiple alignment can be manually edited using **jalview**, a program that is installed on your workstation and that you have already used as an alignment viewer when investigating **Pfam**.

Then I will invite you to try a few of the other options for yourself and see that they do not all produce the same alignment! Differences reflect not only the parameters selected, which we will have discussed, but also the particular objectives of the program selected. For example, a multiple protein sequence alignment optimal for investigating conservation of protein structure might well not be identical to one best representing protein evolution.

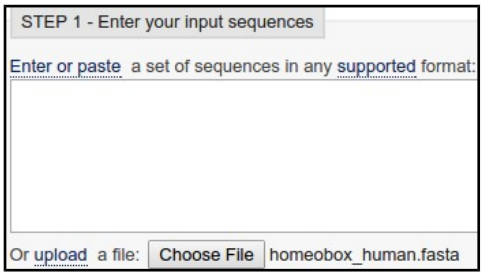
Used to align the **Homeobox** sequences used in this exercise, I do not expect you will see much difference between the outputs of any of these options. They will all work sufficiently on such a simple data set.

The program whose use I choose to describe carefully, leading on to a short **jalview** exercise is **MUSCLE**. I choose thus as **MUSCLE** is now the first choice of most of the people with whom I work. Almost as popular are **Clustal Omega**, **MAFFT** and, for **phylogeny**, **PRANK**.



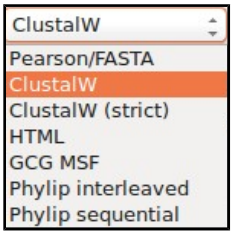
So the plan now is to use **MUSCLE**[104] to align again the **homeobox** sequences previously aligned with **clustalX**. **MUSCLE** works in a way similar to **clustalX** but it takes rather more care in the generation of the **Guide Tree** used to control the order of pairwise construction of the final multiple alignment[105]. Particularly for more difficult alignments, **MUSCLE** should do a better job than **clustalX**. The alignment you will generate here will certainly be different. I leave you to judge for yourselves whether it is better.

Start by requesting to Launch MUSCLE.



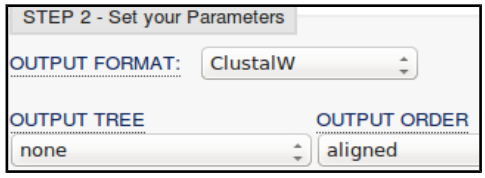
Use the Choose File button to upload the file containing the **FASTA** format **homeobox** sequences, **homeobox_human.fasta**. This file should no longer included the sequences with a mess around position **30**.

Take a look at the **Set your Parameters** section of the page.



There are a number of **OUTPUT FORMATS** offered. For a quick glance at your results, both **ClustalW** or **HTML** are fine. Here I suggest it would be nice to generate an output that can be downloaded and viewed in **Jalview**[106]. The default **ClustalW** or **Pearson/FASTA** serve for this purpose. As **ClustalW** looks more like an alignment in the web page, I choose **ClustalW**[107].

The **EBI** site speaks the truth when it claims that *"The default settings will fulfill the needs of most users and, for that reason, are not visible."*. But this is obscenely patronizing advice! so click the More options... button anyway.



I confess myself confused at the lack of any meaningful options to consider? I was expecting at least the **gap open** and **gap extension penalty** options (available elsewhere, including **Wageningen**), plus a way to change the **scoring matrix**. I have inquired why things are as they are. No practical issue here, as I intended to suggest the defaults whatever they were.

Look at the range of settings for the **OUTPUT TREE** parameter. **none** is indeed the thinking persons choice, but …

How do the options for the **OUTPUT TREE** relate to the output files of **ClustalX** and the difference between the way that **ClustalX** and **muscle** work?

Comment on how one might choose between the range of options offered for the aligned parameter?

104 Available from a variety of websites in addition to the **EBI**, including the Bioinformatics server at **Wageningen**:
http://www.bioinformatics.nl/tools/muscle.html

105 As discussed, superficially at least, previously. I hope.

106 The java alignment editor and viewer you used to look at the **Pfam** alignments earlier.

107 But feel free to try the others. **HTML** is the default at **Wageningen**. The **Phylip** formats are the best if you are going to analyse your output further with the phylogeny programs of the **PHYLIP** package.

After considering these enigmas, or before if you prefer, Click on the Submit button and sit back to admire **muscle** in action.

The alignment that is computed is, superficially at least, similar to that offered by **ClustalX**.

The alignment is irritatingly split into two sections. A nice extra parameter might have been "How wide would you like your alignment to be"? A problem with the format rather than the program, to be fair.

```
ISL2_HUMAN_191_250        --TTRVRTVLNEKQLHTLRTCYAANPRPDALMKEQLVEMTGLSPRVIRVWFQNKRCKDKK
ISL1_HUMAN_181_240        --TTRVRTVLNEKQLHTLRTCYAANPRPDALMKEQLVEMTGLSPRVIRVWFQNKRCKDKK
LHX2_HUMAN_266_325        --TKRMRTSFKHHQLRTMKSYFAINHNPDAKDLKQLAQKTGLTKRVLQVWFQNARAKFRR
LHX9_HUMAN_267_326        --TKRMRTSFKHHQLRTMKSYFAINHNPDAKDLKQLAQKTGLTKRVLQVWFQNARAKFRR
LHX6_HUMAN_219_278        --AKRARTSFTAEQLQVMQAQFAQDNNPDAQTLQKLADMTGLSRRVIQVWFQNCRARHKK
LHX8_HUMAN_225_284        --AKRARTSFTADQLQVMQAQFAQDNNPDAQTLQKLAERTGLSRRVIQVWFQNCRARHKK
ZFHX3_HUMAN_2641_2700     --DKRLRTTITPEQLEILYQKYLLDSNPTRKMLDHIAHEVGLKKRVVQVWFQNTRARERK
ZFHX4_HUMAN_2560_2619     --DKRLRTTITPEQLEILYEKYLLDSNPTRKMLDHIAREVGLKKRVVQVWFQNTRARERK
ZFHX2_HUMAN_1857_1916     --DKRLRTTILPEQLEILYRWYMQDSNPTRKMLDCISEEVGLKKRVVQVWFQNTRARERK
ZFHX2_HUMAN_2065_2124     --QRRYRTQMSSLQLKIMKACYEAYRTPTMQECEVLGEEIGLPKRVIQVWFQNARAKEKK
ZFHX3_HUMAN_2944_3003     PGQKRFRTQMTNLQLKVLKSCFNDYRTPTMLECEVLGNDIGLPKRVVQVWFQNARAKEKK
ZFHX4_HUMAN_2884_2943     --HKRFRTQMSNLQLKVLKACFSDYRTPTMQECEMLGNEIGLPKRVVQVWFQNARAKEKK
LMX1A_HUMAN_195_254       --PKRPRTILTTQQRRAFKASFEVSSKPCRKVRETLAAETGLSVRVVQVWFQNQRAKMKK
LMX1B_HUMAN_219_278       --PKRPRTILTTQQRRAFKASFEVSSKPCRKVRETLAAETGLSVRVVQVWFQNQRAKMKK
LHX1_HUMAN_180_239        --RRGPRTTIKAKQLETLKAAFAATPKPTRHIREQLAQETGLNMRVIQVWFQNRRSKERR
LHX5_HUMAN_180_239        --RRGPRTTIKAKQLETLKAAFAATPKPTRHIREQLAQETGLNMRVIQVWFQNRRSKERR
LHX3_HUMAN_157_216        --AKRPRTTITAKQLETLKSAYNTSPKPARHVREQLSSETGLDMRVVQVWFQNRRAKEKR
LHX4_HUMAN_157_216        --AKRPRTTITAKQLETLKNAYKNSPKPARHVREQLSSETGLDMRVVQVWFQNRRAKEKR
                                          :  :            .         :  :

ZHX1_HUMAN_777_832        GIELF
ZHX3_HUMAN_835_894        AV---
ZEB2_HUMAN_644_703        NS---
ZEB1_HUMAN_581_640        VQ---
HOMEZ_HUMAN_53_115        SWS--
ZHX2_HUMAN_263_324        SWSPE
ZHX3_HUMAN_304_363        SW---
ZHX1_HUMAN_284_346        SWTPE
NANGN_HUMAN_102_161       EM---
ADNP_HUMAN_754_814        D----
ADNP2_HUMAN_1043_1102     AI---
ZHX1_HUMAN_569_630        KEEKM
ZHX2_HUMAN_530_591        MEQAV
ZHX3_HUMAN_612_671        EE---
```

At the very bottom of the page, **muscle** whines:

PLEASE NOTE: Showing colors on large alignments is slow.

So click the **Show Colors** button at the top of the page and try to live with the pain of such gross Trans-Atlantic inept spelling in a European site!!! Good Grief! They get everywhere!!

Well, an improvement I suppose? Colours are very useful (even slow ones) in the interpretation of alignments. Various colour schemes are used to clarify the message of alignments. Colouring can indicate shared amino acid properties not immediately evident when the letter representations differ.

```
ISL2_HUMAN_191_250        --TTRVRTVLNEKQLHTLRTCYAANPRPDALMKEQLVEMTGLSPRVIRVWFQNKRCKDKK
ISL1_HUMAN_181_240        --TTRVRTVLNEKQLHTLRTCYAANPRPDALMKEQLVEMTGLSPRVIRVWFQNKRCKDKK
LHX2_HUMAN_266_325        --TKRMRTSFKHHQLRTMKSYFAINHNPDAKDLKQLAQKTGLTKRVLQVWFQNARAKFRR
LHX9_HUMAN_267_326        --TKRMRTSFKHHQLRTMKSYFAINHNPDAKDLKQLAQKTGLTKRVLQVWFQNARAKFRR
LHX6_HUMAN_219_278        --AKRARTSFTAEQLQVMQAQFAQDNNPDAQTLQKLADMTGLSRRVIQVWFQNCRARHKK
LHX8_HUMAN_225_284        --AKRARTSFTADQLQVMQAQFAQDNNPDAQTLQKLAERTGLSRRVIQVWFQNCRARHKK
ZFHX3_HUMAN_2641_2700     --DKRLRTTITPEQLEILYQKYLLDSNPTRKMLDHIAHEVGLKKRVVQVWFQNTRARERK
ZFHX4_HUMAN_2560_2619     --DKRLRTTITPEQLEILYEKYLLDSNPTRKMLDHIAREVGLKKRVVQVWFQNTRARERK
ZFHX2_HUMAN_1857_1916     --DKRLRTTILPEQLEILYRWYMQDSNPTRKMLDCISEEVGLKKRVVQVWFQNTRARERK
ZFHX2_HUMAN_2065_2124     --QRRYRTQMSSLQLKIMKACYEAYRTPTMQECEVLGEEIGLPKRVIQVWFQNARAKEKK
ZFHX3_HUMAN_2944_3003     PGQKRFRTQMTNLQLKVLKSCFNDYRTPTMLECEVLGNDIGLPKRVVQVWFQNARAKEKK
ZFHX4_HUMAN_2884_2943     --HKRFRTQMSNLQLKVLKACFSDYRTPTMQECEMLGNEIGLPKRVVQVWFQNARAKEKK
LMX1A_HUMAN_195_254       --PKRPRTILTTQQRRAFKASFEVSSKPCRKVRETLAAETGLSVRVVQVWFQNQRAKMKK
LMX1B_HUMAN_219_278       --PKRPRTILTTQQRRAFKASFEVSSKPCRKVRETLAAETGLSVRVVQVWFQNQRAKMKK
LHX1_HUMAN_180_239        --RRGPRTTIKAKQLETLKAAFAATPKPTRHIREQLAQETGLNMRVIQVWFQNRRSKERR
LHX5_HUMAN_180_239        --RRGPRTTIKAKQLETLKAAFAATPKPTRHIREQLAQETGLNMRVIQVWFQNRRSKERR
LHX3_HUMAN_157_216        --AKRPRTTITAKQLETLKSAYNTSPKPARHVREQLSSETGLDMRVVQVWFQNRRAKEKR
LHX4_HUMAN_157_216        --AKRPRTTITAKQLETLKNAYKNSPKPARHVREQLSSETGLDMRVVQVWFQNRRAKEKR
                                          :  :            .         :  :

ZHX1_HUMAN_777_832        GIELF
ZHX3_HUMAN_835_894        AV---
ZEB2_HUMAN_644_703        NS---
ZEB1_HUMAN_581_640        VQ---
HOMEZ_HUMAN_53_115        SWS--
ZHX2_HUMAN_263_324        SWSPE
ZHX3_HUMAN_304_363        SW---
ZHX1_HUMAN_284_346        SWTPE
NANGN_HUMAN_102_161       EM---
ADNP_HUMAN_754_814        D----
ADNP2_HUMAN_1043_1102     AI---
ZHX1_HUMAN_569_630        KEEKM
ZHX2_HUMAN_530_591        MEQAV
ZHX3_HUMAN_612_671        EE---
```

But any decoration available here is far short of what can be achieved with **Jalview**, so click on the **Download Alignment File** button to save you alignment in a file on your **Desktop** called:

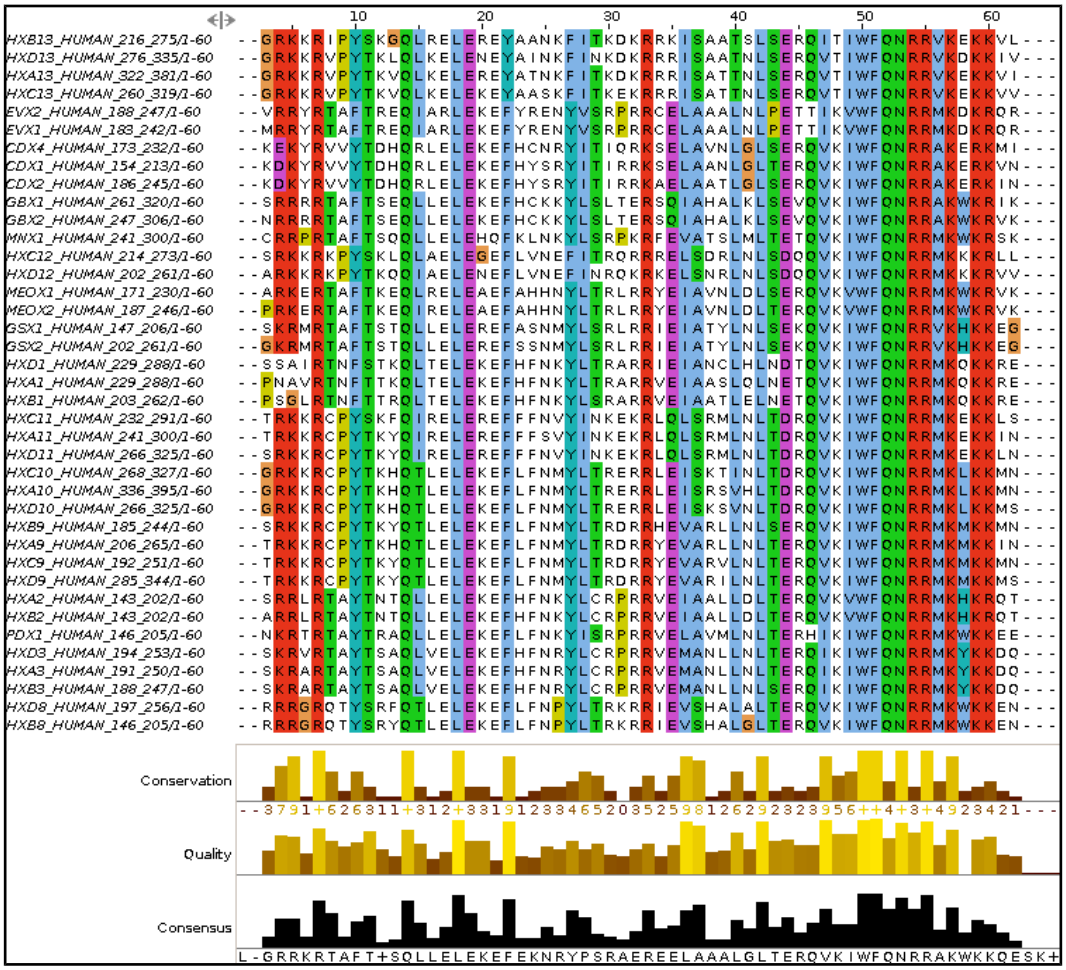
`homeobox_human_muscle.aln`

Start up **Jalview** from the **Windows Start** menu[108]. Close down all the example outputs it sees fit to show you on start up. From the **File** pull down menu choose **from File** from the **Input Alignment** option. Locate and load the file `homeobox_human_muscle.aln`.

The default view is a trifle bland. Try a few of the options from the **Colour** pull down menu.

You could try the same colour scheme used by **ClustalX**, for example.

Now the **MUSCLE** and massaged **ClustalX** alignments look even more similar! In the nicely aligned regions at least.



There are many **Jalview** features that merit investigation. Have a look around if you have time. In particular, **Jalview** will compute simple phylogenetic trees for you employing a number of methods (**Calculate Tree** from the **Calculate** pull down menu). Try it, but be aware this is only sensible if you were very sure of your alignment (and have a few less sequences maybe?).

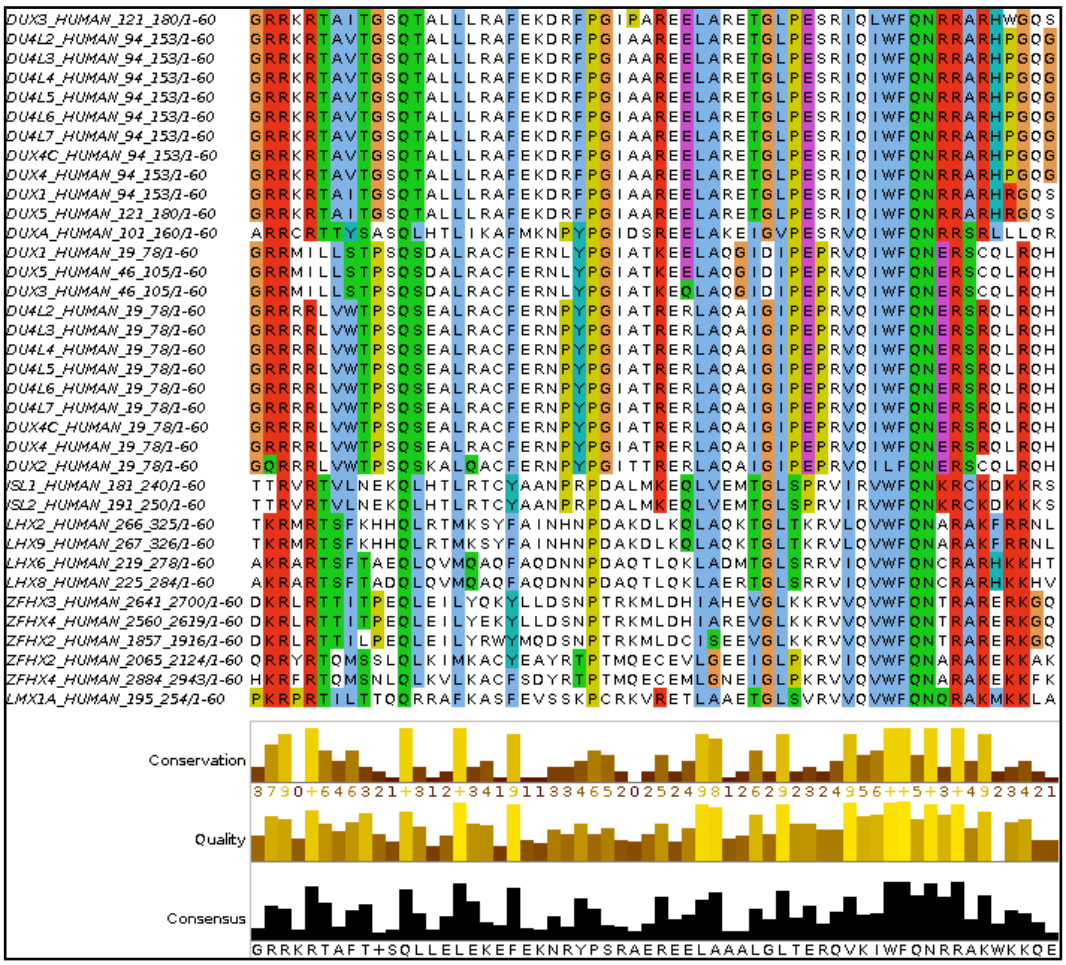
Jalview is made by the same group as produce **JPred**. You could send your alignment for **Secondary Structure Prediction** via the **Web Service** pull down menu, if you wished.

A very important purpose of **Jalview** is to allow users to edit alignments as well as just to view them. For example, hold down the **Shift** key, click and hold on any amino acid at the edge of a gap, slide left and right and see that you can introduce and/or alter the position of gaps. It is very important to be able to edit alignments generated by even the best of programs. As I hope has been made clear, the alignment algorithms are crude. If you know something about the sequences you are aligning it is very reasonable to suppose you can improve upon the computer's alignments. **Jalview** tries to make this possibility easy. Look through some of the other **Edit** pull down menu options, it does not matter how much you mangle your alignment, you can always make another one.

Finally, take a look at the **Jalview** "**Manhattan Skyline**" for the highly conserved **W** at position **50**. This seems a much more reasonable representation of the "truth" than **clustalX** managed? I am not sure what I believe at this point.

108 *All Programs -> Sequence Analysis Tools Installer -> Jalview -> Jalview*

You can also **Select** and **Cut** sequences in a way similar to that you employed with **clustalx**. I could not resist it! I removed all the ugly sequences that caused the gaps at the start and finish of the alignment (just select their names and then select **Cut** or **Delete** from the **Edit** menu). I achieved the gap-free beautiful alignment illustrated.



Of course, **Jalview** does not compute alignments, so once I had removed all the unfortunate proteins, I had to use an **Edit** option to tidy up my meddling. I used **Remove Empty Columns** to get rid of the gap columns at the start of the alignment. The gaps at the end just melted away once the sequences that supported their presence were removed.

Science is easy! Once you remove the need for honesty that is.

Try some of the other **msa** tools offered by the **EBI**. Note the differences in the alignments computed. These differences are important and are not just the result of different parameters applied to exactly the same algorithms. Alignments are generated for different purposes in a variety of subtly different ways. A full consideration of **msa** is beyond the scope of these exercises[109].

109 We do run specialist training in the generation and interpretation of **multiple sequence alignments** in Cambridge however. All are welcome.

PSI-BLAST

This program is used to find a comprehensive set of relatives of a protein. First, **BLAST** is used to find closely related proteins. From an alignment of these proteins a general "profile" (a **P**osition **S**pecific **S**coring **M**atrix - **PSSM**) is computed. A **PSSM** is very similar in concept and purpose to an **HMM** profile in that it summarises significant features present in the sequences it represents.

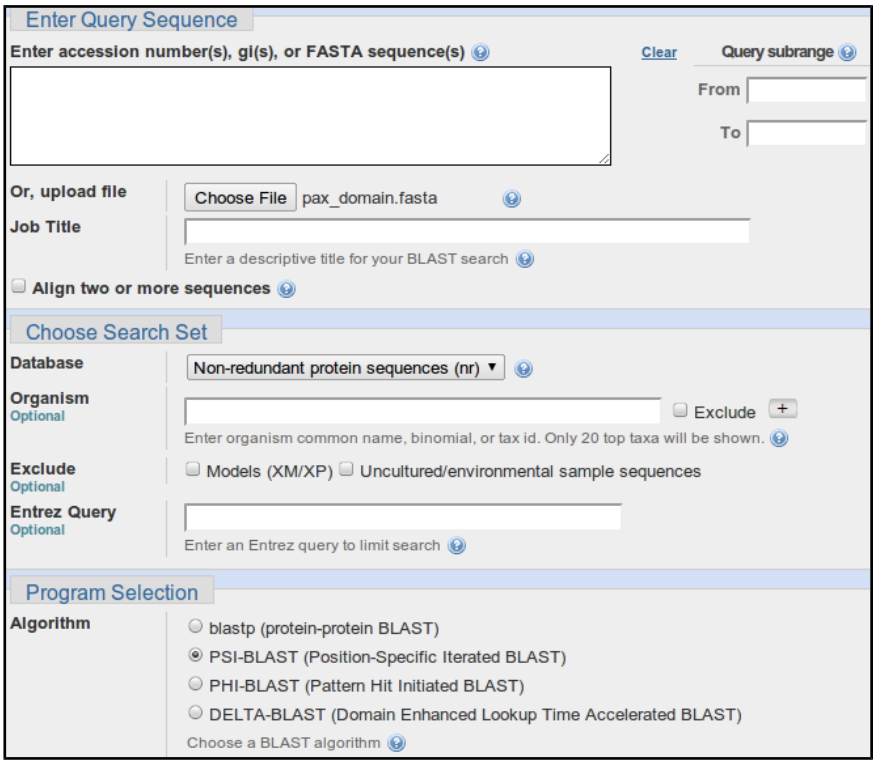
A query against the protein database is then run using the **PSSM**, and a larger more widely associated group of proteins is found. This larger group is used to construct another **PSSM**, and the process is repeated until no more significantly matching new sequences can be detected, or the user tires of the whole process.

You have used **PSI-BLAST** integrated into **Jpred** already and similar ideas were used to create the **PFAM** alignments. Here we will use **PSI-BLAST** explicitly at the **NCBI** on the **Paired DOMAIN** of the **PAX6** protein that you saved in a file earlier. It should be possible to detect a large family of **PAX** domains and to eventually multiply align them generating something like the **Full** alignment from the **PFAM** database viewed earlier[110].

To investigate **PSI-BLAST** go first to the **NCBI** Home page at:

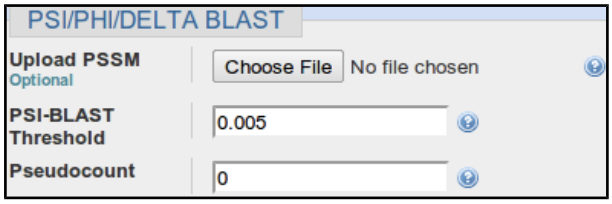
http://www.ncbi.nlm.nih.gov/

Click on the **BLAST** option. Select **protein BLAST** from the **Basic BLAST** section. Upload the **PAX6** paired box domain sequence (stored in the file **pax_domain.fasta**) using the appropriate **Browse** button.



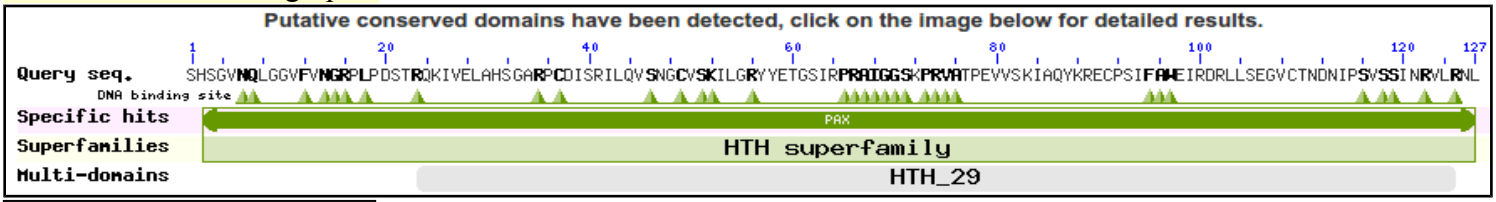
Select **PSI-BLAST** from the **Program Selection** section. Leave all the others options at their default settings, particularly the option to search all the proteins available.

Before you set **PSI-BLAST** going, click on the **Algorithm parameters** link and take a look at the **PSI/PHI/DELTA BLAST** section. Here is offered the option to use a **PSSM** from a previous run **PSI-BLAST**, potentially on a different database (but with the same query sequence). Accept the default that database entries scoring better than an **Expect Threshold** of **0.005** be offered for inclusion into the **PSSM** of each successive **PSI-BLAST** iteration. Remember the buttons.



What do you suppose the choice of **Pseudocount** might influence? _____

Elect to **Show results in a new window** and then click on the BLAST button. After several moments of deep thought, **PSI-BLAST** will come back with its first set of results, at the top of which is a report that (unsurprisingly) matches have been detected between the query sequence and several domain databases. For more detail, click on the **Conserved Domains** graphic.



110 But hopefully a mite more credible!

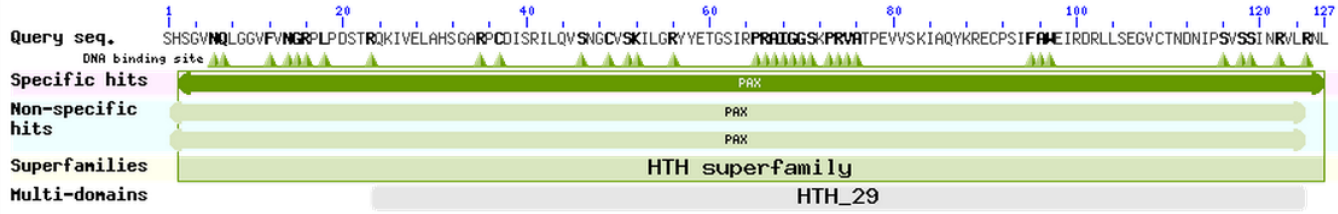
SMART, **Pfam** and the **NCBI Conserved Domains** database hits are reported. None should be a surprise.

Conserved domains on [lcl|Query_2485]

View: Standard Results

Pax-Domain P26367(4-130)

Graphical summary | Zoom to residue level | show extra options »



Search for similar domain architectures | Refine search

List of domain hits

Name	Accession	Description	Interval	E-value
[+] PAX	cd00131	Paired Box domain	2-127	5.03e-80
[+] PAX	smart00351	Paired Box domain;	1-125	2.30e-81
[+] PAX	pfam00292	'Paired box' domain;	1-125	2.38e-81
[+] HTH_29	pfam13551	Winged helix-turn helix; This helix-turn-helix domain is often found in transferases and is ...	23-125	1.15e-04

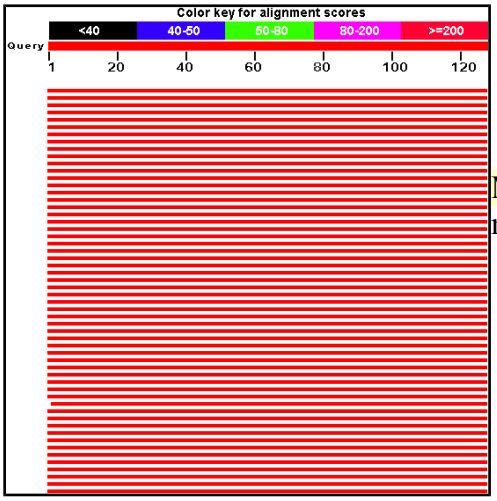
Blast search parameters

Data Source: Live blast search **RID =** XCKC805Y014

User Options: **Database:** CDSEARCH/cdd v3.14 **Low complexity filter:** no **Composition Based Adjustment:** yes **E-value threshold:** 0.01 **Maximum number of hits:** 500

References:

- Marchler-Bauer A et al. (2015), *"CDD: NCBI's conserved domain database."*, **Nucleic Acids Res.43**(D)222-6.
- Marchler-Bauer A et al. (2011), *"CDD: a Conserved Domain Database for the functional annotation of proteins."*, **Nucleic Acids Res.39**(D)225-9.
- Marchler-Bauer A et al. (2009), *"CDD: specific functional annotation with the Conserved Domain Database."*, **Nucleic Acids Res.37**(D)205-10.
- Marchler-Bauer A, Bryant SH (2004), *"CD-Search: protein domain annotations on the fly."*, **Nucleic Acids Res.32**(W)327-331.



Moving back to the main **PSI-BLAST** results, you will see that there are many high quality hits covering the whole length of the query sequence.

The best **500** of these are listed.

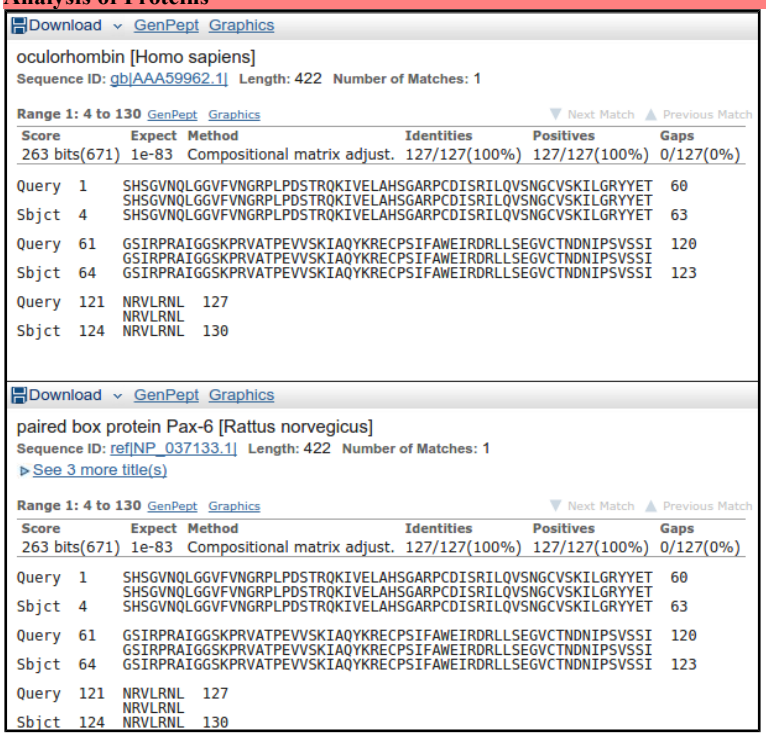
All the listed hits are selected for inclusion into the **PSSM** for the next iteration. Unless you feel strongly about any particular entry, leave them all selected.

Sequences producing significant alignments with E-value BETTER than threshold

Select: All None Selected:0

Alignments | Download | GenPept | Graphics | Distance tree of results | Multiple alignment

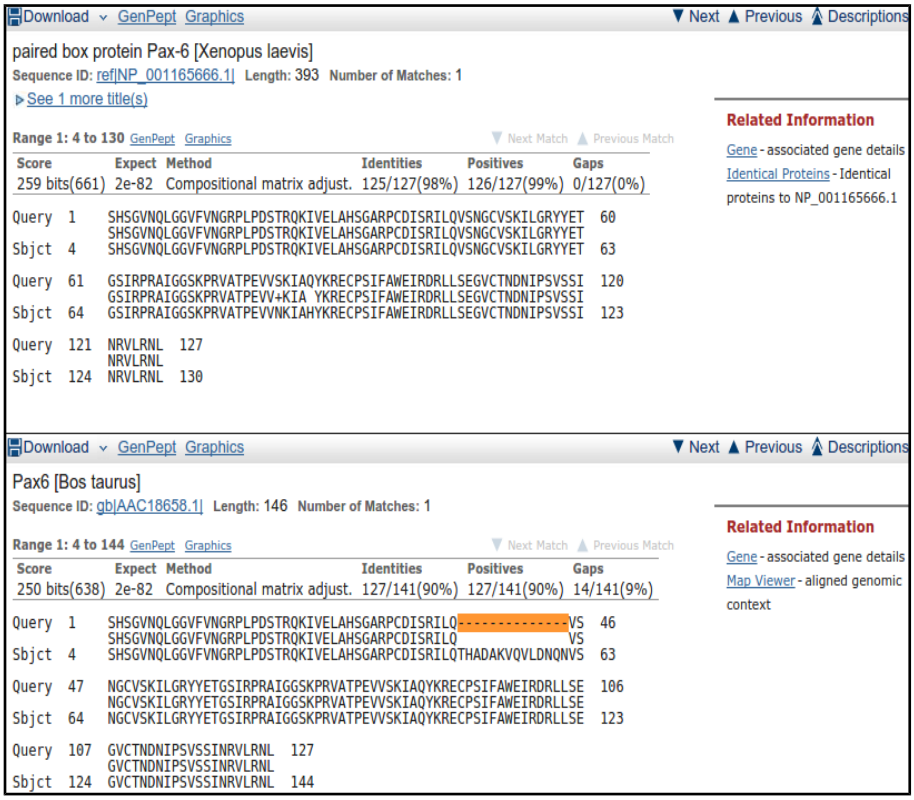
Description	Max score	Total score	Query cover	E value	Ident	Accession	Select for PSI blast	Used to build PSSM
PREDICTED: paired box protein Pax-6 isoform X4 [Macaca nemestrina]	262	262	100%	1e-83	100%	XP_011722295.1	☑	
PREDICTED: paired box protein Pax-6 isoform X2 [Ursus maritimus]	263	263	100%	1e-83	100%	XP_008685073.1	☑	
oculorhombin [Homo sapiens]	263	263	100%	1e-83	100%	AAA59962.1	☑	
paired box protein Pax-6 [Rattus norvegicus]	263	263	100%	1e-83	100%	NP_037133.1	☑	
PREDICTED: paired box protein Pax-6 isoform X2 [Fukomys damarensis]	263	263	100%	1e-83	100%	XP_010638711.1	☑	
PREDICTED: paired box protein Pax-6 isoform X2 [Cavia porcellus]	263	263	100%	1e-83	100%	XP_003464531.1	☑	
PREDICTED: paired box protein Pax-6 isoform X2 [Aotus nancymaae]	263	263	100%	1e-83	100%	XP_012307699.1	☑	
PREDICTED: paired box protein Pax-6 isoform X2 [Callorhinchus milii]	263	263	100%	1e-83	100%	XP_007885973.1	☑	
PREDICTED: paired box protein Pax-6 isoform X2 [Heterocephalus glaber]	263	263	100%	1e-83	100%	XP_004851665.1	☑	
PREDICTED: paired box protein Pax-6 isoform X2 [Octodon degus]	263	263	100%	1e-83	100%	XP_004638029.1	☑	
PREDICTED: paired box protein Pax-6 [Poecilia reticulata]	261	261	100%	1e-83	98%	XP_008404092.1	☑	



Move down to the **Alignments** section of the results and you will see that many of the top hits match the query exactly.

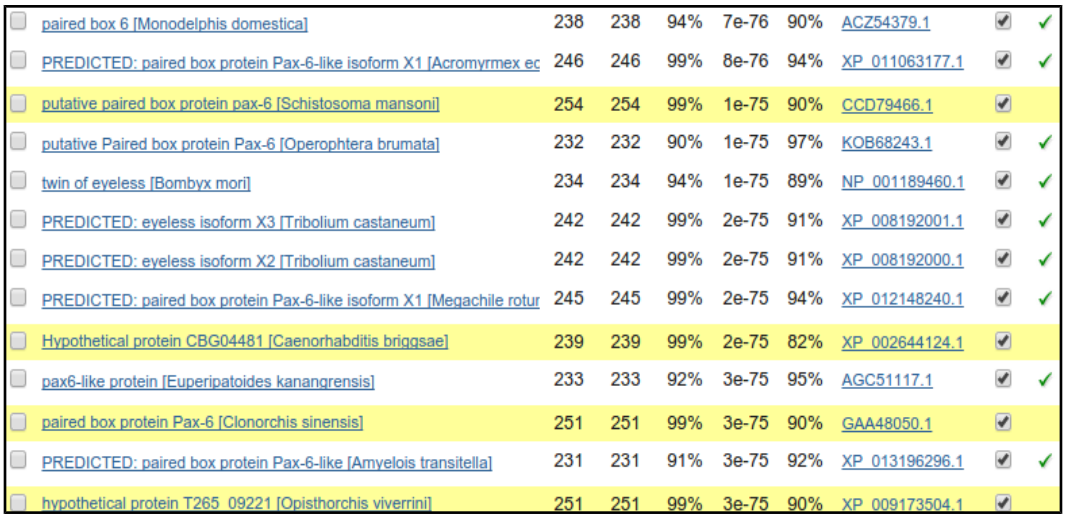
Note that many of the top hits come from the GenPept database (roughly equivalent to the TrEMBL section of UniProtKB).

How might the inclusion of relatively poor quality sequences and the presence of so much duplication have been minimised?_____



Move down far enough and you will see less perfect matches, some of which involve proteins with the extra 14 amino acids of **isoform 5a** of **PAX6_HUMAN**.

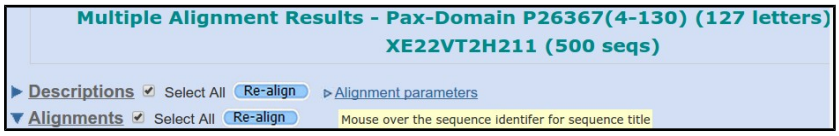
Having browsed your results sufficiently, click on the Go button to **Run PSI-Blast iteration 2**. It is at the bottom of the hit list.



After a few moments, **PSI-BLAST** will return with the results of searching through the database again using the **PSSM** derived from the hits of the first iteration(☑ed). This time the top of the list will be predominantly filled with hits that have already been incorporated into the **PSI-BLAST PSSM**. However, look far enough down the list and you will find some new ones, highlighted yellow.

Once more, click on the Go button to **Run PSI-Blast iteration 3**. That is probably enough! It took 4 iterations before there were no more new sequences suggested for inclusion into the **PSMM** when I ran this last, so if you really want to take things to their logical conclusion, it should not detain you long.

Next, move to the top of the **Descriptions** list and **Select All**. Click on the **Multiple Alignment** button. You have elected to use the **NCBI** multiple alignment program **Cobalt** to align all the **PAX** domain sequences of your final **PSI-BLAST** iteration that match with an **Expect** score better than **0.001**. In an impressively short time, your alignment will appear.



Move past the long list of proteins that have been aligned (the easiest way is to hide the **Descriptions** view).

At the top of the actual alignment, set **View Format** to **Plain Text** (…. and then hide the **Descriptions** again??), this being the easiest format to understand in a hurry. The alignment will have very ragged ends, but the important region of **120** or so amino acids representing the **PAX** domain is really quite impressive. In particular, the i**soform 5a** insertion is very convincing[111].

XP_003977912	52	TRQKIVELAHSGARPCDISRILQTHDA--VQVLDSEKV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	114
XP_009296159	26	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNENV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	90
XP_003246075	54	TRQKIVELAHSGARPCDISRILQ-------------V----------SNGCVSKILG----RYYETGSIKPRAIGGSK	104
XP_012793883	41	TRQRIIELAHSGARPCDISRILQ-------------V----------SNGCVSKILC----RYYETGSIRPKAIGGSK	91
XP_005991286	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDIQNV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
EFX75780	37	TRQKIVELAHSGARPCDISRILQ-------------V----------SNGCVSKILG----RYYETGSIRPRAIGGSK	87
ABB43131	25	TRQKIVELAHSGARPCDISRILQ-------------V----------SNGCVSKILG----RYYETGSIRPRAIGGSK	75
ETN66652	41	TRQKIVELAHSGARPCDISRILQ-------------V----------SNGCVSKILG----RYYETGSIKPRAIGGSK	91
XP_006128959	56	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	120
XP_010874560	44	TRQKIVELAHSGARPCDISRILQTHDDSKVQVLDNENV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	108
AFJ24746	53	TRQRIVELAHSGARPCDISRILQ-------------V----------SNGCVSKILC----RYYETGSIRPKAIGGSK	103
XP_007885968	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVVDNRKV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
BAA24024	42	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDSQNV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	106
XP_012307695	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
CBY09679	55	TRQKIVELAHSGARPCDISRILQ-------------V----------SNGCVSKILA----RYYETGSIKPRAIGGSK	105
XP_007181079	82	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	146
CAF29075	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDSENV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
XP_004264009	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
XP_009184622	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
AAW24017	55	TRQKIVELAHSGARPCDISRILQ-------------V----------SNGCVSKILA----RYYETGSIKPRAIGGSK	105
XP_008547741	26	TRQKIVELAHSGARPCDISRILQ-------------V----------SNGCVSKILG----RYYETGSIRPRAIGGSK	76
XP_012162452	50	TRQKIVELAHSGARPCDISRILQ-------------V----------SNGCVSKILG----RYYETGSIKPRAIGGSK	100
XP_006975926	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNENV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
KDR14710	21	TRQKIVELAHSGARPCDISRILQ-------------V----------SNGCVSKILG----RYYETGSIRPRAIGGSK	71
XP_005530321	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
ABI98847	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
XP_010794780	44	TRQKIVELAHSGARPCDISRILQTHDE--VQVLDSEKV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	106
NP_001103907	26	TRQKIVELAHSGARPCDISRILQ-------------V----------SNGCVSKILG----RYYETGSIRPRAIGGSK	76
XP_010356630	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
XP_010638709	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
XP_005064878	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNENV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
NP_038655	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNENV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
XP_005401829	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
XP_004638028	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
BAM74254	32	TRQRIVELAHSGARPCDISRILQ-------------V----------SNGCVSKILG----RYYETGSIRPRAIGGSK	82

111 Much more so than the **Full** alignment offered by **PFAM**, I would contend. Although, it has to be admitted, the **Pfam** alignment included more sequences and I suspect they would have gone for a less closely homologous set of sequences. Even so … I think the alignment illustrated here is **MUCH** more beautiful!!

Protein Tertiary Structure

Protein Data Bank (PDB)

The **P**rotein **D**ata **B**ank (**PDB**) archive is the major repository of information about the 3D structures of biological molecules, including proteins and nucleic acids. Structures in the archive range from tiny proteins and bits of DNA to complex molecular machines like the ribosome.

In 1998, the **R**esearch **C**ollaboratory for **S**tructural **B**ioinformatics (**RCSB**) became responsible for the management of the **PDB**.

In 2003, the **wwPDB** formed to maintain a single **PDB** archive of macromolecular structural data that is freely and publicly available to the global community. It consists of organizations that act as deposition, data processing and distribution centres for **PDB** data.

PDBe is the European resource for the collection, organisation and dissemination of data on biological macromolecular structures. In collaboration with the other Worldwide Protein Data Bank (**wwPDB**) and **EMDataBank** partners, they work to collate, maintain and provide access to the global repositories of macromolecular structure data (the Protein Data Bank (**PDB**) and Electron Microscopy Data Bank (**EMDB**)).

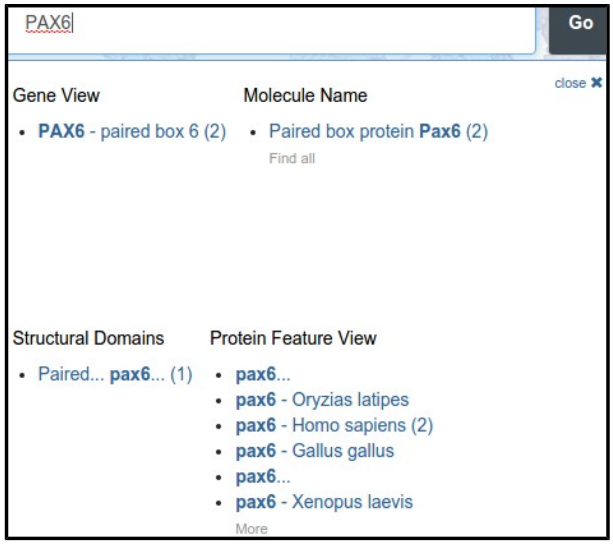
In the course of the exercises undertaken to this point, you will have already had a look at the 3D structures for the 2 major domains of the human **PAX6** protein. You might have taken a more direct route to these structures by asking for them directly from the **RCSB PDB** site as follows.

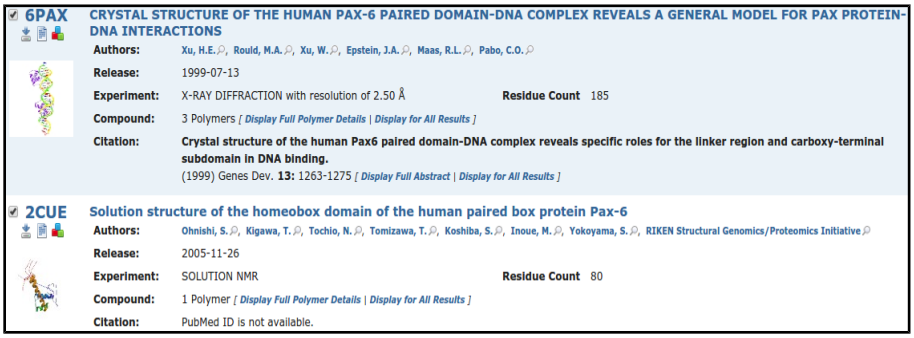
Go to:

http://www.rcsb.org

Enter **PAX6** in the **Search** box and click on the **Go** button.

Click on the link under the **Molecule Name** title..





The two **PDB** structure hits will, hopefully, be familiar. Links are provided with each hit to view the structure with **Jmol** (a java based structure viewer), view the textual **PDB** entry and download the **PDB** entry to a file.

Take a look at the **Jmol** view of the **6PAX PDB** entry. This you have seen this previously, but now I suggest a very quick visualisation of the main mutation that causes aniridia occurs in the **PAX** protein. The idea is to locate and highlight the **Alanine** that mutates to a **Proline** in an aniridia sufferer. As you have discovered, this is the residue **33** in the canonical protein, as recorded by **UniProtKB**. It is residue **30** in the protein as visualised here, the difference being explained by **post translational modification** which, in this instance, removes the first three amino acids.

Instructions for using **Jmol** can be found in many places. For a **Quick Guide**, you might try:

http://blc.arizona.edu/courses/mcb184/graphics/JmolQuickReferenceSheet.pdf

One place for the full manual is:

http://jmol.sourceforge.net/docs/JmolUserGuide/

Please note **Jmol** is not the only structure visualisation option available to you, nor is it the most sophisticated. It is just the one used by **PDB**. Here we look at just the minimum of **Jmol** skills to see what is required. First notice you can zoom in and out with the wheel of your mouse. You can also rotate the image in all directions using your left hand mouse button. Use these two tricks as needed.

To proceed any further, you really need a console window into which you can type commands. To get a console, choose **Console**. From the right hand mouse button pull down menu.

In the lower section of the console, select the **30th** amino acid with the command:

select 30

Then to make the selected residue more evident, type in the two commands:

spacefill

color cyan

and then manipulate the structure until the selected amino acid can be best observed.

LJM/DPJ/PDFJ 2015.08.23

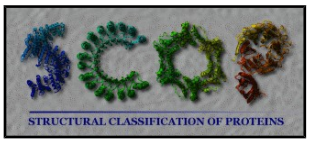
This page is "Work in progress"

As SCOP and CATH still crop up, I wish to include them, probably as an extra exercise.

Also maybe some other structure discussion I am not really qualified to write?

This is just some remnants left over from previous versions.

Structural Classification of Proteins (SCOP)



The **S**tructural **C**lassification **O**f **P**roteins (**SCOP**) database is a largely manual hierarchical classification of protein structural domains based on similarities of their amino acid sequences and three-dimensional structures.

For a quick look at **SCOP**, go to:

http://scop.mrc-lmb.cam.ac.uk/scop/

You could go directly to all **SCOP** has to offer concerning **PAX** proteins by using a **Keyword search of SCOP entries**. For a better view of the way **SCOP** is structured, click on the **top of the hierarchy** link.

The top level **SCOP** classifications is listed. It should be possible to view sample structures for each class, but your browser is unlikely to be suitably configured[112], so instead, click on the **All alpha proteins** link[113].

Classes:

1. All alpha proteins [46456] (284)
2. All beta proteins [48724] (174)
3. Alpha and beta proteins (a/b) [51349] (147)
   Mainly parallel beta sheets (beta-alpha-beta units)
4. Alpha and beta proteins (a+b) [53931] (376)
   Mainly antiparallel beta sheets (segregated alpha and beta regions)
5. Multi-domain proteins (alpha and beta) [56572] (66)
   Folds consisting of two or more domains belonging to different classes
6. Membrane and cell surface proteins and peptides [56835] (58)
   Does not include proteins in the immune system
7. Small proteins [56992] (90)
   Usually dominated by metal ligand, heme, and/or disulfide bridges
8. Coiled coil proteins [57942] (7)
   Not a true class
9. Low resolution protein structures [58117] (26)
   Not a true class
10. Peptides [58231] (121)
    Peptides and fragments. Not a true class
11. Designed proteins [58788] (44)
    Experimental structures of proteins with essentially non-natural sequences. Not a true class

Next, the **DNA/RNA binding 3-helical bundle** category looks the most tempting, so click it. From here options abound! **Homeodomain-like** has appeal, but **"Winged helix" DNA-dinding domain** suggests so much more adventure. Either/both works. You decide which way to descend down **SCOP**'s hierarchy. Note the opportunities to branch off to other relevant databases.

112 You need either **rasmol** and/or **Chime** to be suitably configured. Both are good, but obsolete, structure viewers.
113 Well ... they were not very big beta sheets!

Appendix I: Sequence symbols

Nucleotide symbols, their complements, and the standard one-letter amino acid symbols are shown below in separate lists. The letter codes for amino acid codes and nucleotide ambiguity were proposed by IUB (Nomenclature Committee, 1985, Eur. J. Biochem. 150; 1-5)

Nucleotides

The meaning of each symbol and its complement are shown below.

IUB/GCG	Meaning	Complement
A	A	T
C	C	G
G	G	C
T/U	T	A
M	A or C	K
R	A or G	Y
W	A or T	W
S	C or G	S
Y	C or T	R
K	G or T	M
V	A or C or G	B
H	A or C or T	D
D	A or G or T	H
B	C or G or T	V
X/N	G or A or T or C	X
.	not G or A or T or C	.

Amino Acids

Here the standard one and three letter amino acid codes, synonymous codons and IUB codes are shown. Codons following semicolons (;) are not sufficiently specific to define a single amino acid even though they represent the best possible back-translation into the IUB codes!

Amino Acid NOTATION

Symbol	3-letter	Meaning	Codons	IUB Depiction
A	Ala	Alanine	GCT,GCC,GCA,GCG	!GCX
B	Asp,Asn	Aspartic,Asparagine	GAT,GAC,AAT,AAC	!RAY
C	Cys	Cysteine	TGT,TGC	!TGY
D	Asp	Aspartic	GAT,GAC	!GAY
E	Glu	Glutamic	GAA,GAG	!GAR
F	Phe	Phenylalanine	TTT,TTC	!TTY
G	Gly	Glycine	GGT,GGC,GGA,GGG	!GGX
H	His	Histidine	CAT,CAC	!CAY
I	Ile	Isoleucine	ATT,ATC,ATA	!ATH
K	Lys	Lysine	AAA,AAG	!AAR
L	Leu	Leucine	TTG,TTA,CTT,CTC,CTA,CTG	!TTR,CTX,YTR;YTX
M	Met	Methionine	ATG	!ATG
N	Asn	Asparagine	AAT,AAC	!AAY
P	Pro	Proline	CCT,CCC,CCA,CCG	!CCX
Q	Gln	Glutamine	CAA,CAG	!CAR
R	Arg	Arginine	CGT,CGC,CGA,CGG,AGA,AGG	!CGX,AGR,MGR;MGX
S	Ser	Serine	TCT,TCC,TCA,TCG,AGT,AGC	!TCX,AGY;WSX
T	Thr	Threonine	ACT,ACC,ACA,ACG	!ACX
V	Val	Valine	GTT,GTC,GTA,GTG	!GTX
W	Trp	Tryptophan	TGG	!TGG
X	Xxx	Unknown		!XXX
Y	Tyr	Tyrosine	TAT,TAC	!TAY
Z	Glu,Gln	Glutamic,Glutamine	GAA,GAG,CAA,CAG	!SAR
*	End	Terminator	TAA,TAG,TGA	!TAR,TRA;TRR

Appendix II: Sequence comparison scoring matrices

Default EMBOSS DNA Scoring Matrix.

	A	T	G	C	S	W	R	Y	K	M	B	V	H	D	N	U
A	**5**	-4	-4	-4	-4	1	1	-4	-4	1	-4	-1	-1	-1	-2	-4
T	-4	**5**	-4	-4	-4	1	-4	1	1	-4	-1	-4	-1	-1	-2	5
G	-4	-4	**5**	-4	1	-4	1	-4	1	-4	-1	-1	-4	-1	-2	-4
C	-4	-4	-4	**5**	1	-4	-4	1	-4	1	-1	-1	-1	-4	-2	-4
S	-4	-4	1	1	-1	-4	-2	-2	-2	-2	-1	-1	-3	-3	-1	-4
W	1	1	-4	-4	-4	-1	-2	-2	-2	-2	-3	-3	-1	-1	-1	1
R	1	-4	1	-4	-2	-2	-1	-4	-2	-2	-3	-1	-3	-1	-1	-4
Y	-4	1	-4	1	-2	-2	-4	-1	-2	-2	-1	-3	-1	-3	-1	1
K	-4	1	1	-4	-2	-2	-2	-2	-1	-4	-1	-3	-3	-1	-1	1
M	1	-4	-4	1	-2	-2	-2	-2	-4	-1	-3	-1	-1	-3	-1	-4
B	-4	-1	-1	-1	-1	-3	-3	-1	-1	-3	-1	-2	-2	-2	-1	-1
V	-1	-4	-1	-1	-1	-3	-1	-3	-3	-1	-2	-1	-2	-2	-1	-4
H	-1	-1	-4	-1	-3	-1	-3	-1	-3	-1	-2	-2	-1	-2	-1	-1
D	-1	-1	-1	-4	-3	-1	-1	-3	-1	-3	-2	-2	-2	-1	-1	-1
N	-2	-2	-2	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2
U	-4	5	-4	-4	-4	1	-4	1	1	-4	-1	-4	-1	-1	-2	5

Appendix III: list files

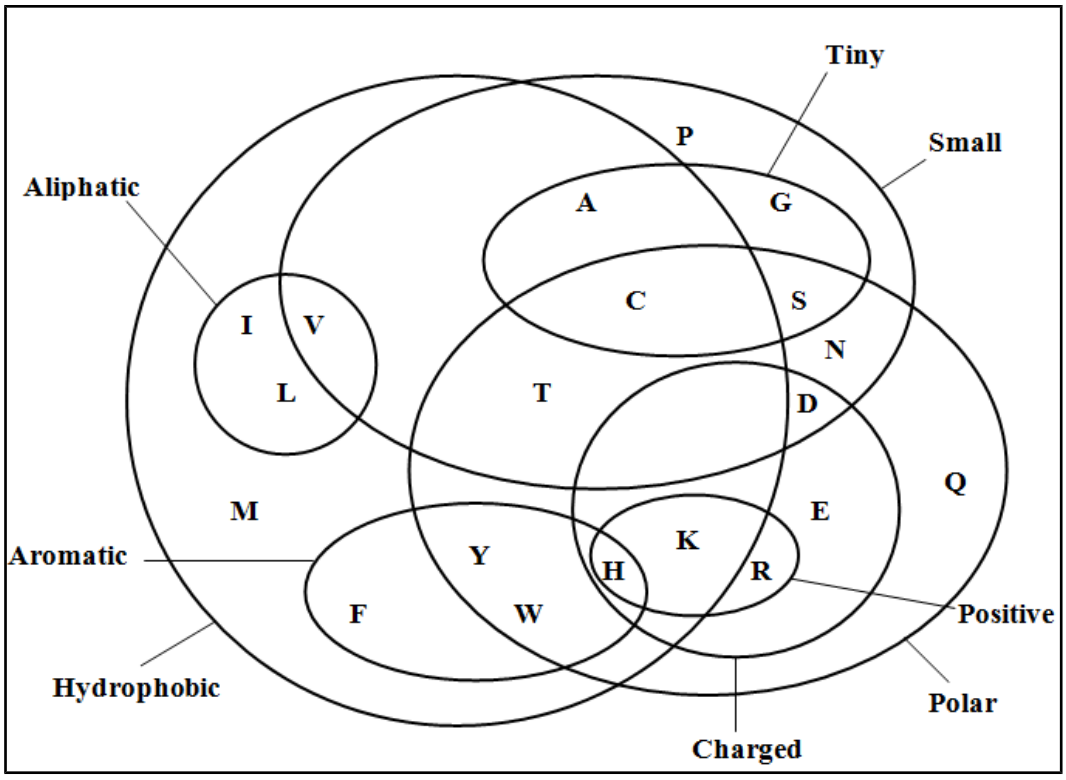
A list file is a text file of sequence file names or database references. It is an excellent and flexible way for you to produce a specific mini-database, e.g. to do ***fasta*** searches against or look for patterns in. Each entry is on a separate line. These can be sequences either in your Unix account, in a sequence database or even in another list file. You can even put in comments if you preface them with an exclamation mark (**!**). The **@** symbol in front of your list file tells EMBOSS programs that the file is a file of filenames and not a sequence. For example:

seqret @listfile

```
embl:hsfau          ! an entry in the embl database
rabbit.seq          ! a file in your directory
swissprot:pax6*     ! all SwissProt entries whose identifier starts with "pax6"
@hedgehog.list      ! another list file called "hedgehog.list"
```

Appendix IV: Amino Acid Properties

Amino acid properties important in the determination of protein tertiary structure.



A quick look at OMIM

OMIM is the **Online** version of Victor McKusick's **Mendelian Inheritance in Man**. It is a database of human disease phenotypes, with a substantial genetic component. There are many ways to get to the **OMIM** entries relating to **aniridia** and **PAX6**, including directly from the **PAX6 GeneCard**.

Jump to the **Disorders section**. The first **OMIM** link (607108) is to the gene **PAX6** entry. There are other **OMIM** links for **PAX6** disorders, including one for **Aniridia** (106210).

OMIM: 607108 **disorders:** 106210 604229 604219 148190 136520 120430 165550 120200 206700

UniProtKB/Swiss-Prot: PAX6_HUMAN, P26367

- Defects in PAX6 are the cause of aniridia (AN) [MIM:106210]. A congenital, bilateral, panocular disorder characterized by complete absence of the iris or extreme iris hypoplasia. Aniridia is not just an isolated defect in iris development but it is associated with macular and optic nerve hypoplasia, cataract, corneal changes, nystagmus. Visual acuity is generally low but is unrelated to the degree of iris hypoplasia. Glaucoma is a secondary problem causing additional visual loss over time

Click on the **Aniridia** link. Here the **Gene map locus** is confirmed. Click on the **Location** link for the **OMIM** map of Chromosome **11** around **PAX6**[114]. You will see a fuller representation in **Ensembl** later.

Location (genomic start, cyto location) (from NCBI)	Gene/Locus	Gene/Locus name	Gene/Locus MIM number	Phenotype	Phenotype MIM number	Pheno map key	Comments	Mouse symbol (from MGI)
11:31,284,170 11p13	DCDC1	Doublecortin domain-containing protein 1	608062					
11:31,391,376 11p13	DPH4	DPH4, S. cerevisiae, homolog of	611072					Dnajc24
11:31,453,948 11p13	IMMP1L, IMP1	Inner mitochondrial membrane peptidase, subunit 1, S. cerevisiae, homolog of	612323					Immp1l
11:31,531,296 11p13	ELP4, PAX6NEB	Elongation protein 4, S. cerevisiae, homolog of	606985					Elp4
11:31,806,339 11p13	PAX6, AN2, MGDA	Paired box homeotic gene-6	607108	Aniridia	106210	3		Pax6
				Cataract with late-onset corneal dystrophy	106210	3		
				Coloboma of optic nerve	120430	3		
				Coloboma, ocular	120200	3		
				Foveal hyperplasia	136520	3		
				Gillespie syndrome	206700	3		
				Keratitis	148190	3		
				Morning glory disc anomaly	120430	3		
				Optic nerve hypoplasia	165550	3		
				Peters anomaly	604229	3		

Move back to the **aniridia OMIM** entry and follow the link to the **OMIM** entry for **PAX6 (607108)** which you will find in the **TEXT** section. From the **Table of Contents** menu, select **Allelic Variants**. Click on **Table View**.

PAIRED BOX GENE 6; PAX6

Allelic Variants (Selected Examples):

Number	Phenotype	Mutation	dbSNP
.0001	ANIRIDIA	PAX6, 2-BP INS	-
.0002	ANIRIDIA	PAX6, EXON G DEL	-
.0003	ANIRIDIA	PAX6, GLN116TER	[rs121907912]
.0004	PETERS ANOMALY ANIRIDIA, INCLUDED	PAX6, ARG26GLY	[rs121907913]
.0005	ANIRIDIA	PAX6, ARG103TER	[rs121907914]
.0006	CATARACTS, CONGENITAL, WITH LATE-ONSET CORNEAL DYSTROPHY	PAX6, SER353TER	[rs121907915]
.0007	ANIRIDIA	PAX6, IVS12DS, G-C, -1	-
.0008	ANIRIDIA	PAX6, ARG203TER	[rs121907916]
.0009	ANIRIDIA	PAX6, ARG240TER	[rs121907917]
.0010	ANIRIDIA	PAX6, IVS11AS, A-G, -2	-
.0011	KERATITIS, AUTOSOMAL DOMINANT	PAX6, IVS10AS, A-T, -2	-
.0012	FOVEAL HYPOPLASIA, ISOLATED	PAX6, ARG125CYS	[rs121907918]
.0013	ANIRIDIA, ATYPICAL	PAX6, VAL126ASP	[rs121907919]
.0014	FOVEAL HYPOPLASIA AND PRESENILE CATARACT SYNDROME	PAX6, GLY64VAL	[rs121907920]

Several variants of this protein causing **aniridia** are listed, many are associated with a **dbSNP** entry, none of which is the one we will be investigating. As is admitted, the list is just selected examples[115]

What do you notice about all the variants that are associated with a **dbSNP** entry?_____

Does this surprise you?_____

114 Note the way you can move along the Chromosome map. I went **Forward 3** entries to get my picture.

115 The full (known) list (and much more) is available in a database specific to human **PAX6** allelic variants. This can be found at: http://lsdb.hgu.mrc.ac.uk/

Other databases from the European Bioinformatics Institute (EBI)

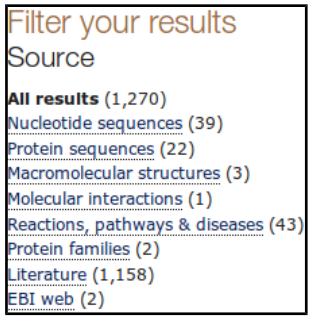
In a new browser window, go to the **EBI** homepage:

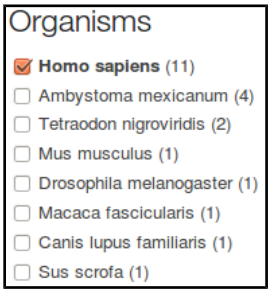
http://www.ebi.ac.uk

For this exercise, you started to investigate **aniridia** at the **Genecards** site in Israel. There are many alternative starting points, including the **EBI**[116], where many of the databases you have been browsing are maintained, or at least mirrored. The **EBI** is not restricted to any particular organism and so would be a better choice for inquiries that are not specific to Homo Sapiens. The resources available for searching are displayed just underneath the search box.

You have already discovered much about **anirida**, certainly enough to consider using the **PAX6** protein accession code **P26367** as a search term. However, if you were really starting from the beginning, you would more probably be inclined to type the keyword **aniridia** into the **Explore the EBI** box and click the **Search** button.

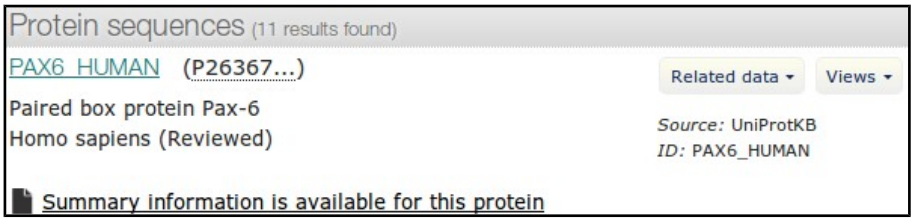
The **EBI Search** finds matches to **aniridia** with the many entries in the many databases of the **EBI**. Take a quick look at the hit list that is generated. From here, this exercise might divert in a numerous directions. For now, the focus is still the **PAX6** human protein. So, **Filter your results** to show just the **Protein sequences**.

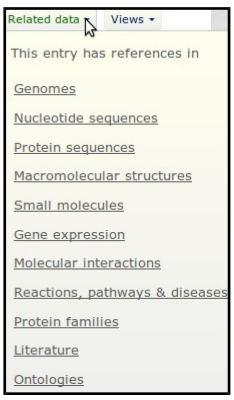




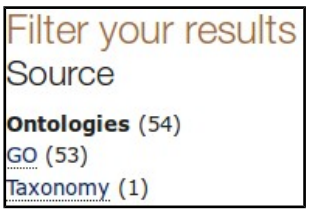
Further refine your list by specifying just the **Organism Homo sapiens**.

At the top of your list, you shoud not be surprised to see **PAX6_HUMAN**.





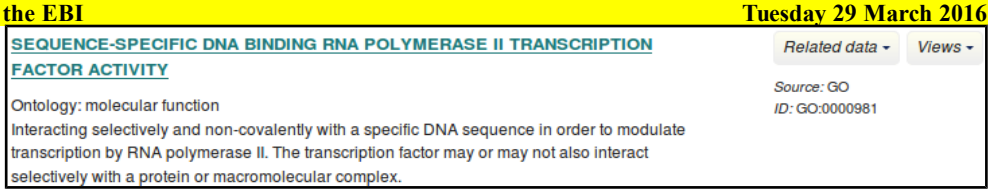
It is possible to link to the corresponding entries in other **EBI** databases. From the **Related data** pull down menu, select the **Ontologies** option.



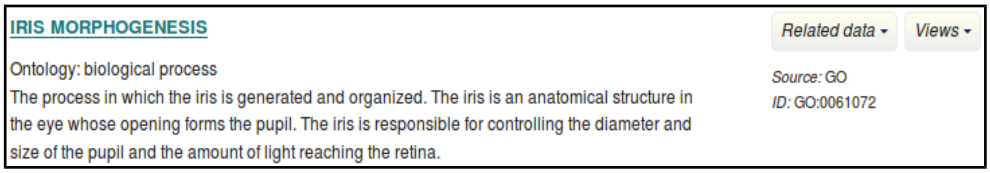
Filter your results to include just entries from the **G**ene **O**ntology (**GO**) database.

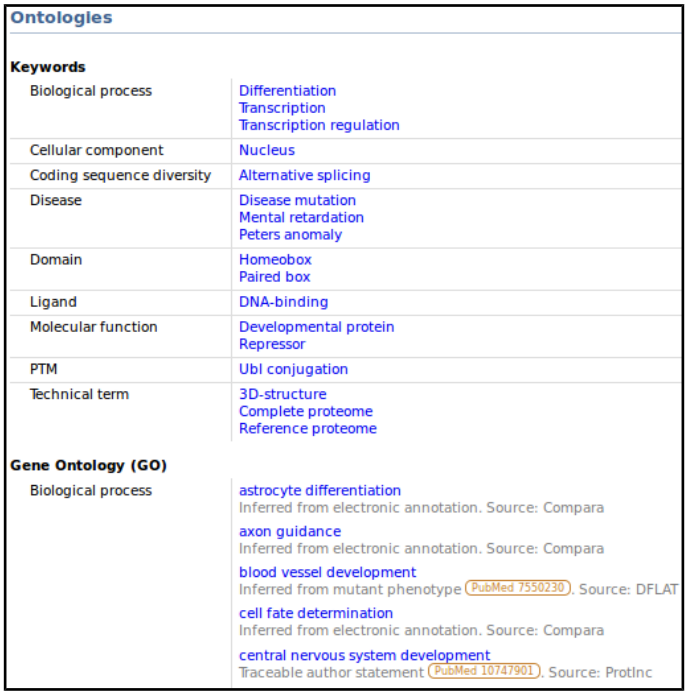
116 And, of course, the **NCBI** in the USA.

Many of the entries found should cause you little surprise[117].



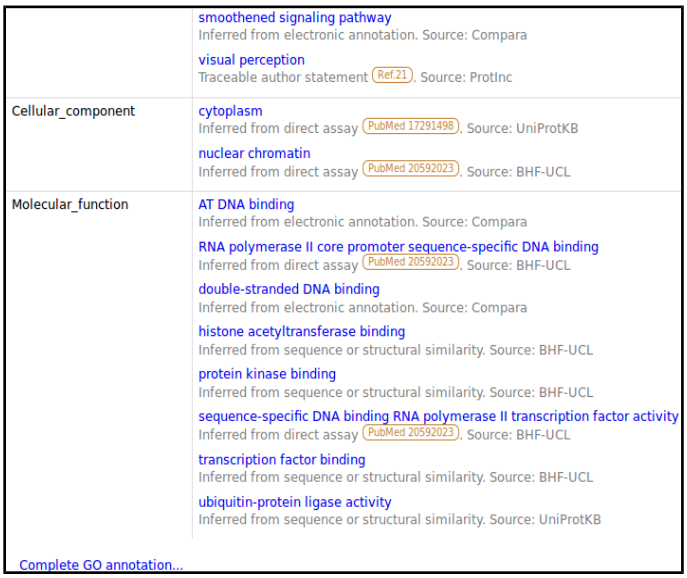
Of course, you could find links to the same **GO** entries within the **UniprotKB** database entry.





Move to the bottom of your list of **GO** entries. Follow the link back to **PAX6_HUMAN**.

Move to the **Ontologies** Section. You will see confirmation of the major domains of **PAX6** and its molecular function suggested by **Keywords** and **Gene Ontology**. Stricter use of keywords must reduce the failures of annotation searches as were experienced previously.



Following the **Keywords** are a large number of relevant **GO** terms classified as **Biological process**, **Cellular component** or **Molecular function**.

117 I include two examples I thought particularly pertinent. The first is a **molecular function**, the second is a **biological process**. Both were on the first page of **GO** entries.

Move to the top of the page. Here there are links to **Clusters** of proteins with sequences **100%**, **90%** and **50%** identical that of **PAX6**. These clusters, constructed from **UniprotKB**, are organised into the databases **UniRef100**, **UniRef90** and **UniRef50**. Comparing proteins with cluster databases produces more succinct results that would be generated by a search against **Uniprot** itself. Follow the links to each cluster report. Note that you can **Customize** the display. Fun, but maybe it is easier just to follow the link to the entire entry in order to obtain more information.

Cluster ID	Status	Cluster name	Size	Cluster member(s)	Organisms	Length	Identity
UniRef100_P26367	★	Cluster: Paired box protein Pax-6	10	P26367 P63015 Q66SS1 F2Z5M7 F6S4R0 F7C9R7 G1P774 H0XKU3 I7G9J6 D3DQZ8	Homo sapiens (Human) Mus musculus (Mouse) Sus scrofa (Pig) Callithrix jacchus (White-tufted-ear marmoset) Macaca mulatta (Rhesus macaque) Myotis lucifugus (Little brown bat) Otolemur garnettii (Small-eared galago) (Garnett's greater bushbaby) Macaca fascicularis (Crab-eating macaque) (Cynomolgus monkey)	422	100%

Cluster ID	Status	Cluster name	Size	Cluster member(s)	Organisms	Length	Identity
UniRef90_P26367	★	Cluster: Paired box protein Pax-6	56	P26367 P63015 Q66SS1 F2Z5M7 F6S4R0 F7C9R7 G1P774 H0XKU3 I7G9J6 +46	Homo sapiens (Human) Mus musculus (Mouse) Sus scrofa (Pig) Callithrix jacchus (White-tufted-ear marmoset) Macaca mulatta (Rhesus macaque) Myotis lucifugus (Little brown bat) Otolemur garnettii (Small-eared galago) (Garnett's greater bushbaby) Macaca fascicularis (Crab-eating macaque) (Cynomolgus monkey) Latimeria chalumnae (West Indian ocean coelacanth) +15	422	90%

Cluster ID	Status	Cluster name	Size	Cluster member(s)	Organisms	Length	Identity
UniRef50_P26367	★	Cluster: Paired box protein Pax-6	231	P26367 P63015 Q66SS1 F2Z5M7 F6S4R0 F7C9R7 G1P774 H0XKU3 I7G9J6 +221	Homo sapiens (Human) Mus musculus (Mouse) Sus scrofa (Pig) Callithrix jacchus (White-tufted-ear marmoset) Macaca mulatta (Rhesus macaque) Myotis lucifugus (Little brown bat) Otolemur garnettii (Small-eared galago) (Garnett's greater bushbaby) Macaca fascicularis (Crab-eating macaque) (Cynomolgus monkey) Latimeria chalumnae (West Indian ocean coelacanth) +123	422	50%

Move again to the **UniRef100** cluster report.

Follow the link to the **UniRef100** entry.

What was the **seed sequence** upon which this cluster was built? _____

What do you imagine a **seed sequence** might be (click for **Help**)? _____

What is the **Representative sequence** protein for this cluster? _____

What do you imagine a **Representative sequence** might be (click for **Help**)? _____

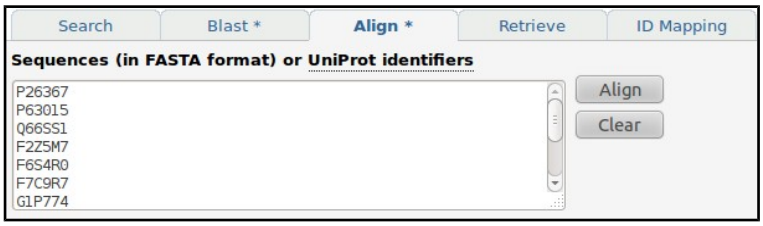
_____

How many sequences are from **UniProtKB/Swiss-Prot** and how many from **UniprotKB/TrEMBL** (Look at the **Dataset** pull down Menu)? _____

Given your last answer, how would you interpret the colours of the stars in the **Status** column? _____

Cluster member(s)
P26367
P63015
Q66SS1
F2Z5M7
F6S4R0
F7C9R7
G1P774
H0XKU3
I7G9J6
D3DQZ8

Select all the sequences of the **UniRef100** entry.

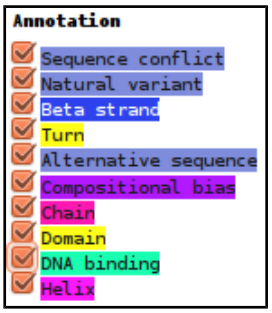


Click the **Align** tab at the very top of the page.

Click the **Align** button.

Try some **Annotation** options. Even all of them!!

```
27   QKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSK   86   P26367   PAX6_HUMAN
27   QKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSK   86   P63015   PAX6_MOUSE
27   QKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSK   86   Q66SS1   Q66SS1_HUMAN
27   QKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSK   86   F2Z5M7   F2Z5M7_PIG
27   QKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSK   86   F6S4R0   F6S4R0_CALJA
27   QKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSK   86   F7C9R7   F7C9R7_MACMU
27   QKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSK   86   G1P774   G1P774_MYOLU
27   QKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSK   86   H0XKU3   H0XKU3_OTOGA
27   QKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSK   86   I7G9J6   I7G9J6_MACFA
61   QKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSK   120  D3DQZ8   D3DQZ8_HUMAN
     ************************************************************
```



Which sequences react to the **Annotation** request, and why? _____

It is clear from the alignment that all the **UniRef100** entry sequence are ***not***, in this case at least, identical!

```
1    ----------------------------------MQNSHSGVNQLGGVFVNGRPLPDSTR   26   P26367   PAX6_HUMAN
1    ----------------------------------MQNSHSGVNQLGGVFVNGRPLPDSTR   26   P63015   PAX6_MOUSE
1    ----------------------------------MQNSHSGVNQLGGVFVNGRPLPDSTR   26   Q66SS1   Q66SS1_HUMAN
1    ----------------------------------MQNSHSGVNQLGGVFVNGRPLPDSTR   26   F2Z5M7   F2Z5M7_PIG
1    ----------------------------------MQNSHSGVNQLGGVFVNGRPLPDSTR   26   F6S4R0   F6S4R0_CALJA
1    ----------------------------------MQNSHSGVNQLGGVFVNGRPLPDSTR   26   F7C9R7   F7C9R7_MACMU
1    ----------------------------------MQNSHSGVNQLGGVFVNGRPLPDSTR   26   G1P774   G1P774_MYOLU
1    ----------------------------------MQNSHSGVNQLGGVFVNGRPLPDSTR   26   H0XKU3   H0XKU3_OTOGA
1    ----------------------------------MQNSHSGVNQLGGVFVNGRPLPDSTR   26   I7G9J6   I7G9J6_MACFA
1    MCEAFYCESGQTSEISGNPIFEPRGIPRPPARASMQNSHSGVNQLGGVFVNGRPLPDSTR   60   D3DQZ8   D3DQZ8_HUMAN
                                       ***************************
```

Can you rationalize why one of the sequences is allowed to be different to all the others? _____

_____

After viewing this **UniRef100** entry, how "non-redundant" would you say was **UniprotKB**?_____

Why do you suppose it might be useful to have identical sequences in **UniprotKB**?_____

Further Features of Ensembl

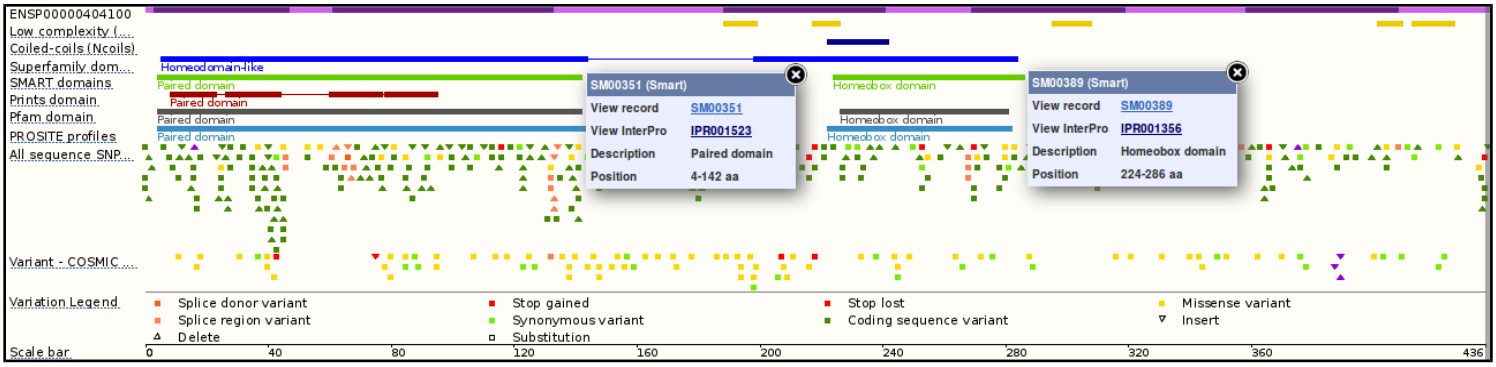
This is really some final steps I removed from the main **Ensembl** investigation in order to speed things along a trifle. Not sufficiently vital to leave in place, but still "interesting"? So I make them a supplementary exercise. I hope to add some more material covering the extended variation features of **Ensembl** later.

These instructions assume you are looking at the PAX6-201 transcript of the human PAX6 gene. If that is not where you are, the URL is:

http://www.ensembl.org/Homo_sapiens/Transcript/Summary?db=core;g=ENSG00000007372;r=11:31806340-31839509;t=ENST00000419022

Just click on the link and you should be taken to the appropriate page. Then … proceed as follows:

Click on the **Protein Summary** link (from **Transcript-based displays → Protein Information**). The graphic shows the **SMART** domains you recently noted. Click on the **SMART** features. Note that feature start/end points and **Interpro** links are also available here.



Click on the **Gene: PAX6** tab. Click on the **Variation Table** link (from **Gene-based displays → Gene Variation**).

Number of variant consequences		Type	Description
0	-	Transcript ablation	A feature ablation whereby the deleted region includes a transcript feature (SO:0001893)
318	Show	Splice donor variant	A splice variant that changes the 2 base region at the 5' end of an intron (SO:0001575)
374	Show	Splice acceptor variant	A splice variant that changes the 2 base region at the 3' end of an intron (SO:0001574)
114	Show	Stop gained	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript (SO:0001587)
54	Show	Frameshift variant	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three (SO:0001589)
9	Show	Stop lost	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript (SO:0001578)
0	-	Start lost	A codon variant that changes at least one base of the canonical start codon (SO:0002012)
0	-	Transcript amplification	A feature amplification of a region containing a transcript (SO:0001889)
10	Show	Inframe insertion	An inframe non synonymous variant that inserts bases into in the coding sequence (SO:0001821)
0	-	Inframe deletion	An inframe non synonymous variant that deletes bases from the coding sequence (SO:0001822)
0	-	protein altering variant	A sequence_variant which is predicted to change the protein encoded in the coding sequence (SO:0001818)
1016	Show	Missense variant	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved (SO:0001583)
670	Show	Splice region variant	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron (SO:0001630)
0	-	Incomplete terminal codon variant	A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed (SO:0001626)
585	Show	Synonymous variant	A sequence variant where there is no resulting change to the encoded amino acid (SO:0001819)
0	-	Stop retained variant	A sequence variant where at least one base in the terminator codon is changed, but the terminator remains (SO:0001567)
2740	Show	Coding sequence variant	A sequence variant that changes the coding sequence (SO:0001580)
0	-	Mature miRNA variant	A transcript variant located with the sequence of the mature miRNA (SO:0001620)
200	Show	5 prime UTR variant	A UTR variant of the 5' UTR (SO:0001623)
375	Show	3 prime UTR variant	A UTR variant of the 3' UTR (SO:0001624)
3358	Show	Non coding transcript exon variant	A sequence variant that changes non-coding exon sequence in a non-coding transcript (SO:0001792)
15257	Show	Intron variant	A transcript variant occurring within an intron (SO:0001627) (WARNING: table may not load for this number of variants!) View list in BioMart
0	-	NMD transcript variant	A variant in a transcript that is the target of NMD (SO:0001621)
8409	Show	Non coding transcript variant	A transcript variant of a non coding RNA gene (SO:0001619) (WARNING: table may not load for this number of variants!) View list in BioMart
6631	Show	Upstream gene variant	A sequence variant located 5' of a gene (SO:0001631) (WARNING: table may not load for this number of variants!) View list in BioMart
7114	Show	Downstream gene variant	A sequence variant located 3' of a gene (SO:0001632) (WARNING: table may not load for this number of variants!) View list in BioMart
38149	Show	ALL	All variations (WARNING: table may not load for this number of variants!) View list in BioMart

You are rewarded by a table offering to show various subsets (or **ALL**) of the **Variations** coincident with the **PAX6** region of the human genome. In the next section of the exercise, you will also generate a list of **SNPs** for **PAX6**, but this option allows for much greater refinement. Click on some of the **Show** options offered. I would imagine the **Splice donor variant**s and **Splice acceptor variants** that occur in the critical start and end base pairs of introns would be of particular interest in "Real Life".

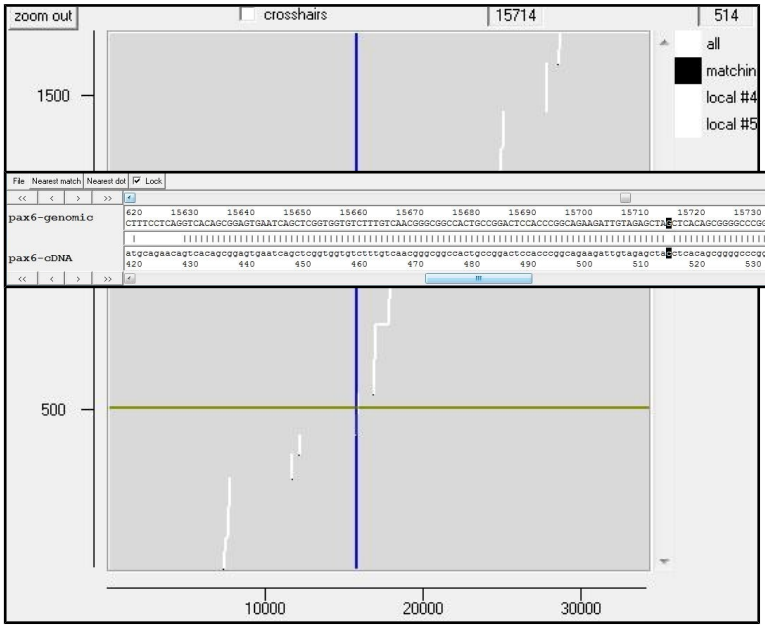
You will see from the lists you produce, many of these mutations are from the **HGMD** and reveal little, unless you pay. There are more useful entries elsewhere however. Have a look around. For my illustration, starting with the list of all **Splice acceptor variant**s, I removed some of the table columns that did not contained obviously interesting information using the Show/hide columns option at the top of the table.

COSM4032358	11:31790861	C/T	somatic_SNV	COSMIC	Splice acceptor variant Non coding transcript variant	ENST00000494377
COSM926339	11:31790861	C/A	somatic_SNV	COSMIC	Splice acceptor variant Non coding transcript variant	ENST00000494377
COSM4032358	11:31790861	C/T	somatic_SNV	COSMIC	Splice acceptor variant Non coding transcript variant	ENST00000533333
COSM926339	11:31790861	C/A	somatic_SNV	COSMIC	Splice acceptor variant Non coding transcript variant	ENST00000533333
COSM4032358	11:31790861	C/T	somatic_SNV	COSMIC	Splice acceptor variant	ENST00000606377
COSM926339	11:31790861	C/A	somatic_SNV	COSMIC	Splice acceptor variant	ENST00000606377
CS068282	11:31793554	HGMD_MUTATION	sequence_alteration	HGMD-PUBLIC	Splice acceptor variant	ENST00000241001
CS982311	11:31793554	HGMD_MUTATION	sequence_alteration	HGMD-PUBLIC	Splice acceptor variant	ENST00000241001
CS068282	11:31793554	HGMD_MUTATION	sequence_alteration	HGMD-PUBLIC	Splice acceptor variant	ENST00000379107
CS982311	11:31793554	HGMD_MUTATION	sequence_alteration	HGMD-PUBLIC	Splice acceptor variant	ENST00000379107
CS068282	11:31793554	HGMD_MUTATION	sequence_alteration	HGMD-PUBLIC	Splice acceptor variant	ENST00000379109
CS982311	11:31793554	HGMD_MUTATION	sequence_alteration	HGMD-PUBLIC	Splice acceptor variant	ENST00000379109
CS068282	11:31793554	HGMD_MUTATION	sequence_alteration	HGMD-PUBLIC	Splice acceptor variant	ENST00000379111
CS982311	11:31793554	HGMD_MUTATION	sequence_alteration	HGMD-PUBLIC	Splice acceptor variant	ENST00000379111
CS068282	11:31793554	HGMD_MUTATION	sequence_alteration	HGMD-PUBLIC	Splice acceptor variant	ENST00000379115
CS982311	11:31793554	HGMD_MUTATION	sequence_alteration	HGMD-PUBLIC	Splice acceptor variant	ENST00000379115

Try the **Variation Image** link. This offers the same information, dressed up as an enormous graphic! Indubitably pulchritudinous, but the poetry is a tad over blown for a simple lad such as I.

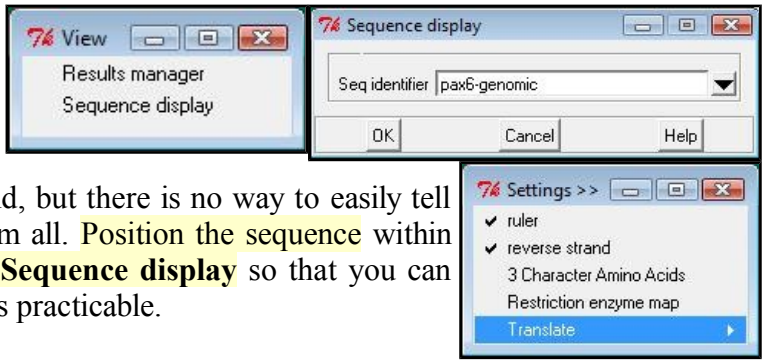
Further Features of Spin

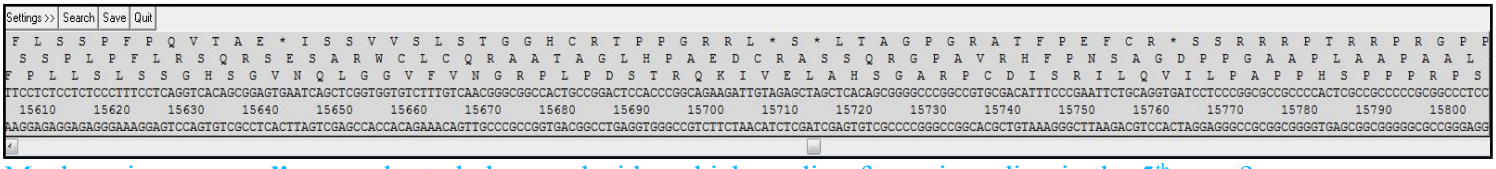
This **Supplementary Exercise** assumes you to be exactly as you would be if you had just finished the main pairwise alignment exercise with **spin**, ending with a look at **Local Alignment**. I moved it here to try and make things a little shorter. It really only shows a few extra features of **spin** whilst, obsessively, trying to get **spin** to translate the unaffected sequence to show explicitly that the amino acid that is mutated into a **Proline** is an **Alanine**. I only partially and clumsily succeeded with the latter and you knew the answer anyway! So it only just deserves to be a **Supplementary Exercise**. Anyway, for what it might be worth, here it is.



The fact that your alignments are not in positional order does make it a little difficult to find things. In particular, the mutation in the coding exon. **splign** suggests this is in the **5th** exon. One way to view it therefore would be to double click on the **5th** exon of the graphical display. This will bring the **Sequence Comparison Display** into view. Click on the **Nearest match** button to view the alignment for the **5th** exon. Click on the **Lock** button so that the cDNA and genomic sequences stay in alignment. Make the **Sequence Comparison Display** as wide as you can. Move the display along until you have the mutation in view. Wonderful, but we have not persuaded any of the software to declare what the translation of the wild type sequence would be at the mutation site[118].

From **spin**'s **View** menu, select **Sequence display**. Check that the genomic sequence is selected. Click **OK**. You should now be viewing the genomic sequence. To add amino acid translations, from the **Settings** menu, select **Translate** and choose to **Translate + frames**. You know the coding reading frame must be on the top strand, but there is no way to easily tell which of the three frames it might be[119], so translate them all. Position the sequence within your display so that the mutation is in view. Shape the **Sequence display** so that you can easily see all three translations and as much sequence as is practicable.

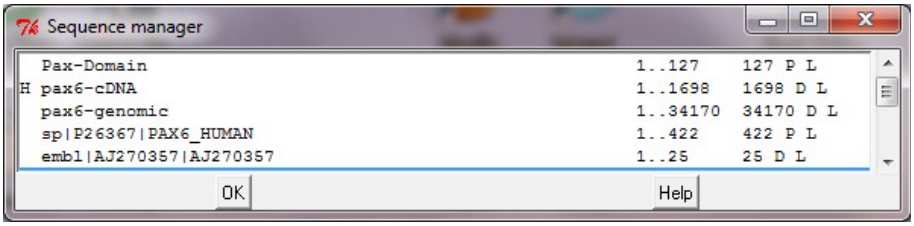




Maybe using your **splign** results to help you decide, which reading frame is coding in the 5th exon?_____

What is the amino acid corresponding to the mutated base in the **PAX6** genomic sequence?_____

At this point, the unworthy cynic might note that we have known all this from the time we looked at the feature table of the **Uniprot** entry. True, but how much more rewarding to have worked it all out for ourselves!



Time to tidy up. Remove as many of your textual outputs from **spin**'s **Output window** as your patience allows. You have saved all that might be useful later. Dispose of **spin**'s graphics windows and **Sequence displays**, all were very beautiful in their time but have now served their purpose fully. In the next section, you will be analysing the mRNA (**pax6-cDNA**), so set it as **spin**'s **H**orizontal sequence in the **Sequence manager**. From this point on, you will be using a lot of **EMBOSS** programs from **spin**. All **EMBOSS** programs are selected from **spin**'s **Emboss** menu, even if I do not explicitly state this each and every time.

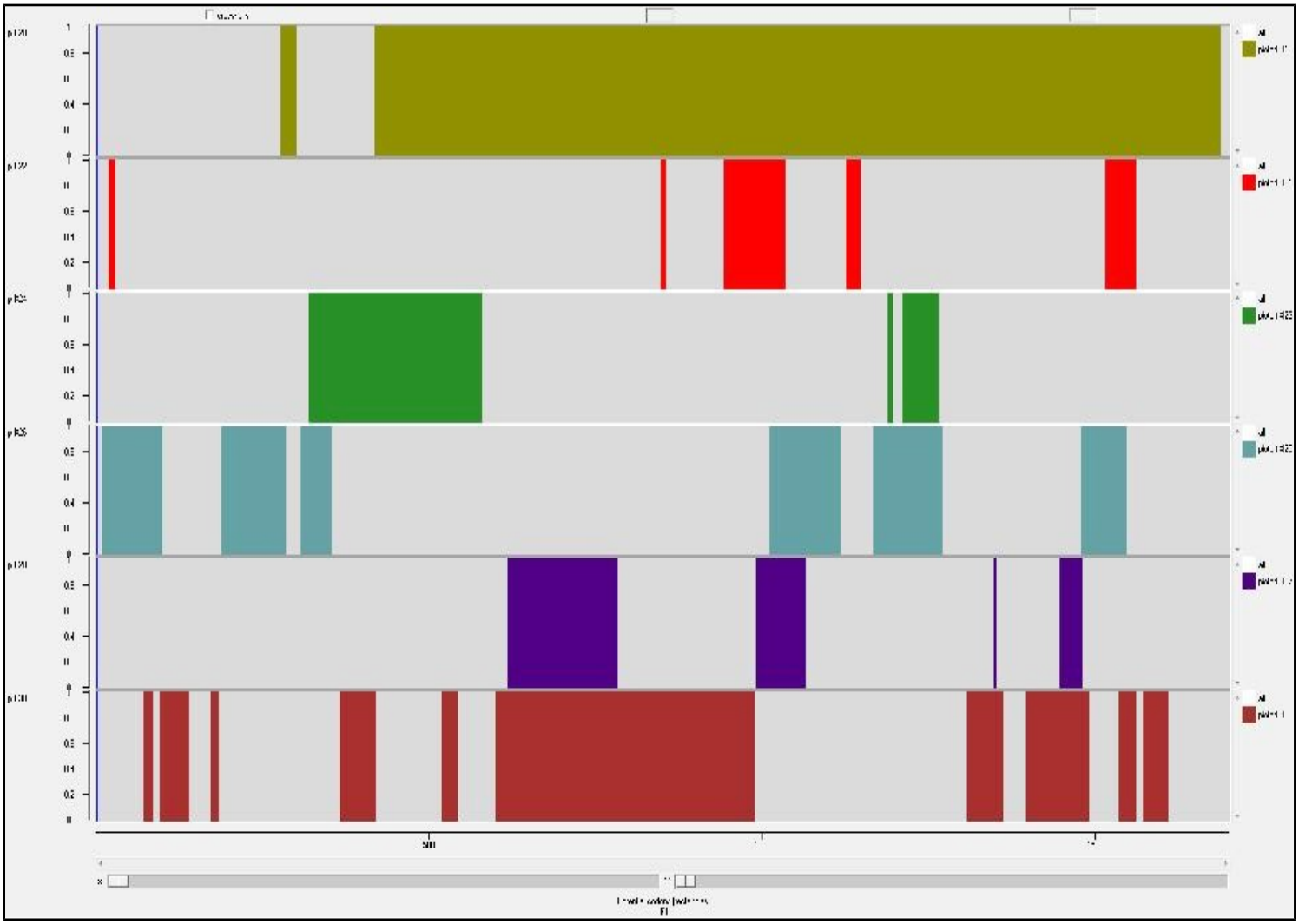
118 Yes I **know** we could work it out using the table in the appendix ... but that would be like "reading the manual"!! Exclusively for wimps!

119 Well, OK there is .. but let us be lazy.

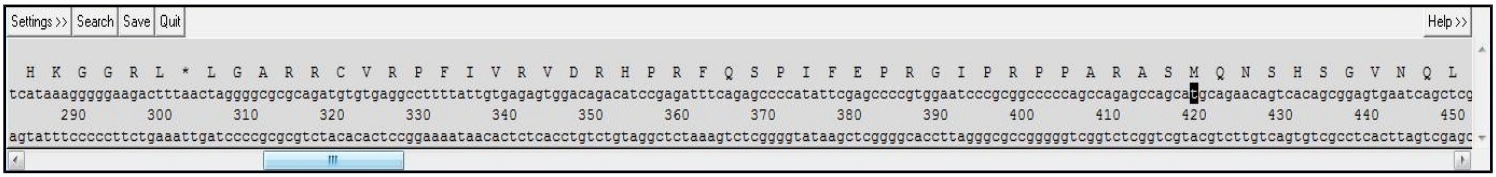
ORF Identification and Translation

Now identify and translate the coding portion of the cDNA sequence into protein. The **EMBOSS** program **plotorf** provides a preliminary graphic of the distribution of ORFs in six frames**.**

From **spin**'s **Emboss** menu, select **plotorf** from the **Nucleic/Gene Finding** menu. Use the default start/stop codons (**Advanced section**). Click **OK**.



You will see a graphical output that shows the potential open reading frames (ORF) in all six-frames. You will notice that the longest ORF occurs in the top section of the graphic (*i.e.* forward sense, frame 1) starting at around **400** bases and ending around **1700**. Try using the cross-hair to obtain a more accurate estimate of the start of the longest ORF. Double click in the graphic, **spin** will produce a sequence display in which you can request a translation of the appropriate reading frame. From this display, it is easy to determine the extent of the ORF and the coding region[120].



You have passed over several opportunities to record the exact start and stop positions of the coding region of your cDNA. Maybe this is the moment to make a note. You need to know that in the **spin Sequence Display** the numbered base is below the least significant digit. Thus both the **340**$^{\textbf{th}}$ base and the **380**$^{\textbf{th}}$ base are **A**s. Also, the last **Stop** (*****) codon (**TAA**) before the coding region of this mRNA spans bases **301** to **303**.

At what base position does the coding sequence of the mRNA commence?_____

Move the **Sequence Display** along to the region in which **plotorf** suggests the coding sequence ends.

What is the base position of the last base of the coding sequence?_____

120 The crosshair and the sequence display are both **Staden** features that have been made to work for many of the **EMBOSS** programs.

Other ways to determine the whereabouts of coding regions include use of the **EMBOSS** program **getorf**[121]. Select **getorf** from the **Nucleic/Gene finding** menu. In the **Additional section**, ensure that the **Code to use** is set to **Standard**. Use your **plotorf** results to select a sensible **Minimum nucleotide size of ORF to report**[122]. Select **Translation of regions between START and STOP codons** for the **Type of output**. In the **Output section**, change the **Filename** to **getorf_results.txt** and hit **OK**.

Look at your **getorf** results file with an editor. The number of "answers" **getorf** suggests will depend upon the minimum **ORF** size you chose. **plotorf** suggests that the correct (most likely at least) **ORF** will be in the region **400** to **1700**. Find the relevant prediction in your **Sequence Display** output. You will (again) see that the translation is from **418** to **1683**.

```
>pax6-cDNA_1 [318 - 578]
MCEAFYCESGQTSEISEPHIRAPWNPAAPSQSQHAEQSQRSESARWCLCQRAATAGLHPA
EDCRATSQRGPAVRHFPNSAGVQRMCE
>pax6-cDNA_2 [418 - 1683]
MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELPHSGARPCDISRILQVSNGCVSKILGRY
YETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSV
SSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQ
EGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALEKEFERTHYPDVFAR
ERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIP
QPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQTSSYSCMLPT
SPSVNGRSYDTYTPPHMQTHMNSQPMGTSGTTSTGLISPGVSVPVQVPGSEPDMSQYWPR
LQ
>pax6-cDNA_3 [781 - 617] (REVERSE SENSE)
MTRLVCYRWYRPPRTVICLGFPKQRWTGTPAYTGLFCLQLLESLLSVYYHRLPWV
>pax6-cDNA_4 [986 - 600] (REVERSE SENSE)
MELVFSPPPSCCWQPSCVGCPGTEVPGYQPGRVPQLPVCPFNILSLSYMPSAPICCFSLA
RLRRTLFIDDTLGMLSLVQTPSDSNLSRISQAKMDGHSRLYWAILLTTSGVATLGLLPPI
ALGLMEPVS
```

To save the translation of this region, pick the **EMBOSS** program **transeq**[124] from the **Nucleic/Translation** menu. Move to the **Additional section,** and set the **Regions to translate** to **418-1683**. Ensure **Translation Frames** is set to **1,** and **Code to use** to **Standard**. In the **Output** section set the **Filename field** to **pax6.pep**. Hit **OK**.

In **spin**'s **Sequence Manager**, ensure **pax6_cDNA** is the **Horizintal** sequence[123].

Using an appropriate editor, alter the comment line in your **pax6.pep** file to read **"pax6 conceptual translation from mutant cDNA"** and save you handiwork for perpetuity. Your file should appear similar to that illustrated.

```
>pax6 conceptual translation from mutant cDNA
MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELPHSGARPCDISRILQVSNGCVSKILGRY
YETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSV
SSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQ
EGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALEKEFERTHYPDVFAR
ERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIP
QPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQTSSYSCMLPT
SPSVNGRSYDTYTPPHMQTHMNSQPMGTSGTTSTGLISPGVSVPVQVPGSEPDMSQYWPR
LQ
```

Now, you are able to compare the protein sequence from your cDNA with the **PAX6** database entry. This requires a sequence alignment to make apparent any discrepancies between sequences.

Align your translated **pax6.pep** sequence with the database entry. Choose the **spin** alignment option you consider most logical (global or local, both will work, but one makes more sense than the other). Load in the protein sequence you saved in the file **pax6.pep**[125] and the **PAX6** human protein from the **Uniprot** sequence database (you will recall that you saved this in a disk file called **pax6_human.fasta** in your working directory[126]). The default settings should do, just set the program going and view the results.

What is the single amino acid difference between the two sequences?_____

What is the position of the difference?_____

121 Using **getorf**, as we do here, is ponderous compared with just reading along the **Sequence Display** of **spin**. However, unlike any native aspect of **spin**, **getorf** can be run without any graphics. It can be run many times from a simple script without user intervention. If you had to detect coding regions in many sequences, rather than just one, **getorf** (plus a simple script) would be the sane option.

122 You want to eliminate all the very small ORFs but not the one you consider might be the genuinely coding region (i.e. the biggest one). In "real life" one might select a rather large value to select only the longest ORF. For the exercise, maybe a smaller value to see the longest few regions would be appropriate? How about **150**?

123 **spin** has an irritating tendency to assume you wish to analyse the output of the last option you executed.

124 This program may not be strictly necessary if you have already obtained the correct protein by using **getorf**.

125 You will need to use the **Load Sequence/Simple** option of the **spin File** pull down menu as it is not yet in the **Sequence manager**.

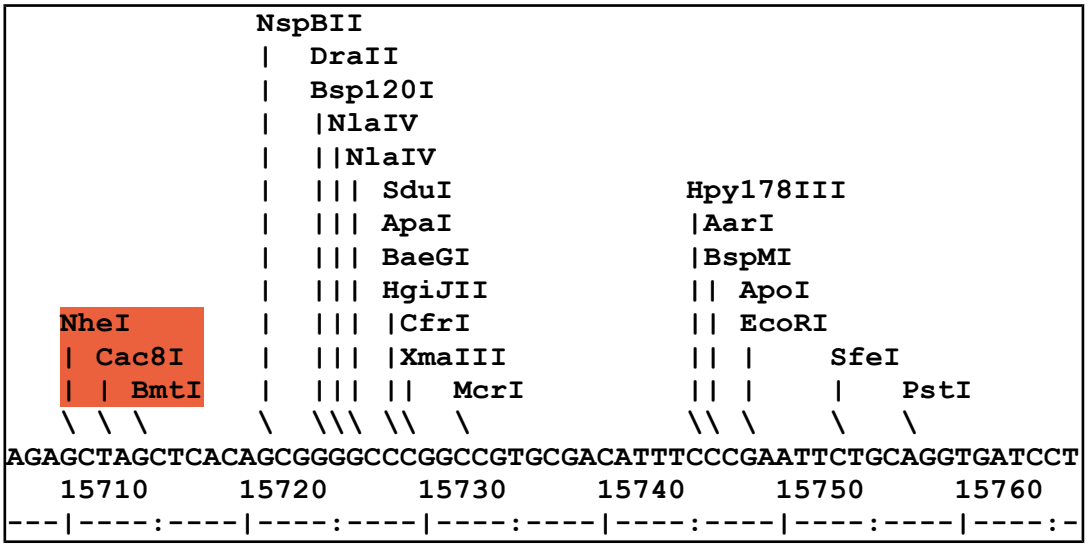
126 It should be in the Sequence manager as **sw|P226367|PAX6_HUMAN**, or you could load it directly from the **Uniprot** database at the **EBI** using the **SRS** capabilities built into **EMBOSS/spin**.

Restriction Maps

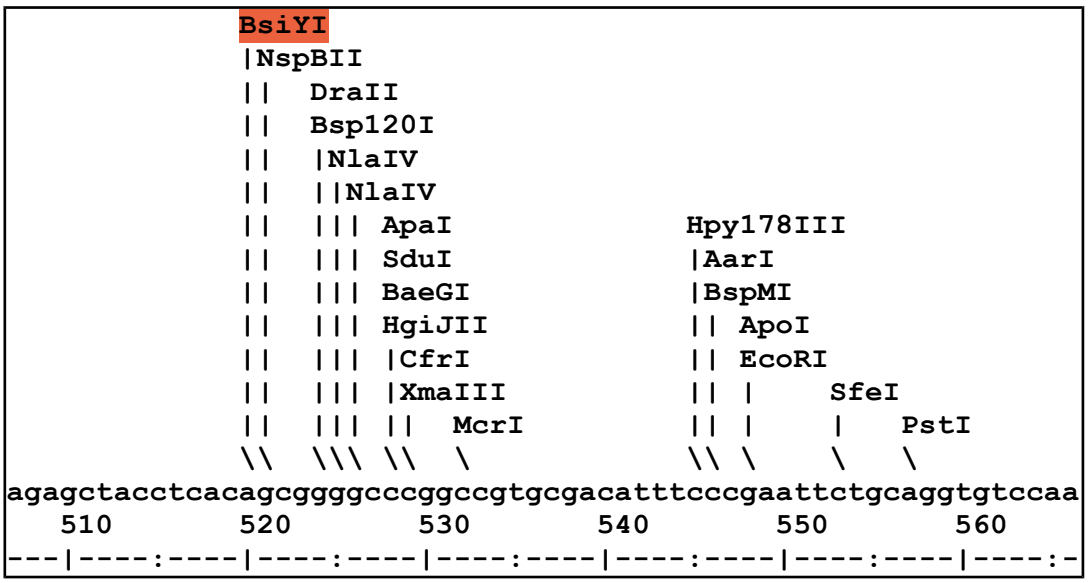
Particularly when looking at the **NCBI** program **splign**, you recorded the position and type of all the differences between your aniridia patient's mRNA and the genomic sequence from **Ensembl**. The most interesting mismatch is between base position **514** of the mRNA (**C**) and **15714** of the genomic sequence (**G**). You could check these positions by looking at your second global alignment (saved in **global_results02.txt**) and/or your second local alignment (saved in **local_results02.txt**), or you could just believe!

Restriction mapping is trivial using computers. Use the **EMBOSS** program **remap** to make a restriction map of each sequence in the area where the difference is seen in the alignment. Any differences in the two restriction maps could be used to further investigating the discrepancy between the two sequences[127].

First map the relevant region of the genomic sequence. Choose **remap** from the **Nucleic/Restriction** menu and load the genomic DNA sequence. Set the **Start position** and **End position** fields to **15707** and **15766**. In the **Required section**, leave the **Comma separated enzyme list** as **all** but set the **Minimum recognition site length** to **6**. In the **Output section**, you know the translation and sense, so turn off **Display translation** and **Display cut sites and translation of reverse sense**[128]. Set the **Output filename** to **remap_genomic.txt**. Leave all else and click on **OK**. Your output should begin as illustrated.



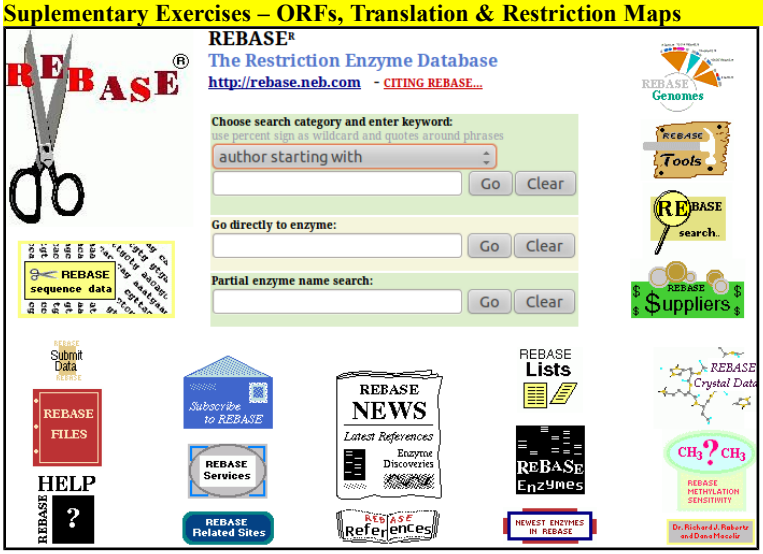
Run **remap** again in the fashion outlined above for the corresponding part of the **PAX6** mutated cDNA. The region to map is from **507** to **566**. This time, set the **Output filename** to **remap_cdna.txt**.



The one base pair difference has altered the way a number of restriction enzymes cut. Differences include the deletion of the **3** restriction sites of the genomic map and the creation of the **1** site in the cDNA map. Complete information about these restriction enzymes, can be found at the **REBASE** (a publicly available database of restriction enzyme data) web site.

127 If there is a single restriction enzyme that cuts differently due to the mutation, it might be possible to design a **R**estriction **F**ragment **L**ength **P**olymorphism (**RFLP**) test to detect the mutation in individuals. Not a likely choice these days. **PCR** would generally be regarded as a much better option. Treat **RFLP** as an excuse for doing some restriction maps if you wish.

128 ***BUG!!*** I fear there is a bug here. These 2 options are turned **ON** even thought they appear to be turned **OFF**. To really turn them **OFF**, you need to turn them **ON** and then turn them **OFF** again!! Sorry, I will try to get this fixed.

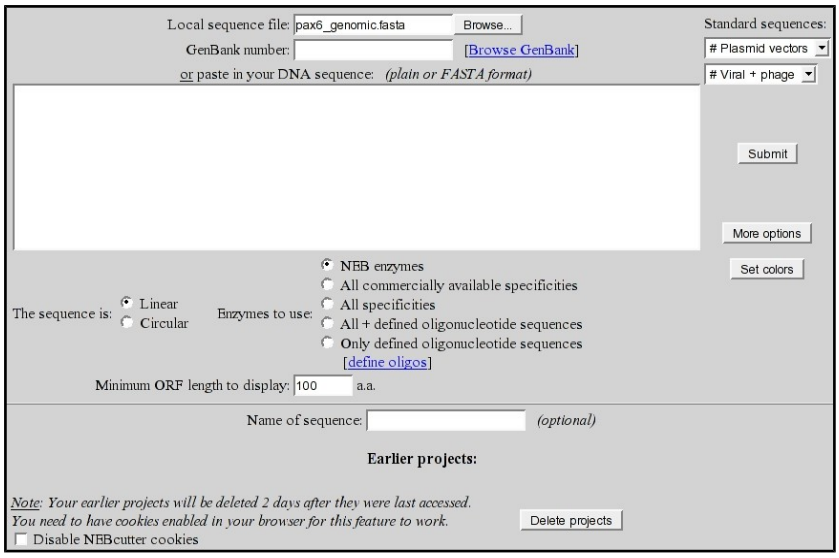


Before moving on, take a quick look around the home of **REBASE**. Make a new navigator window and go to:

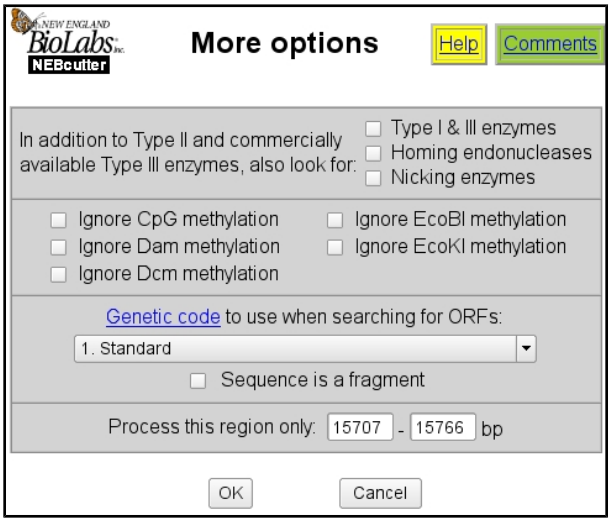
`http://rebase.neb.com/`

You could download the **REBASE** in a large number of different formats (including that required by the **EMBOSS** package). To examine the possibilities, click on the **REBASE FILES** link.

For a really pretty and interactive map, starting from the **REBASE** home page once more, click on the **REBASE Tools** link.



Select **New England Biolabs NEBcutter**. **Load sequence file** **pax6_genomic.fasta** and ask for **More options**.

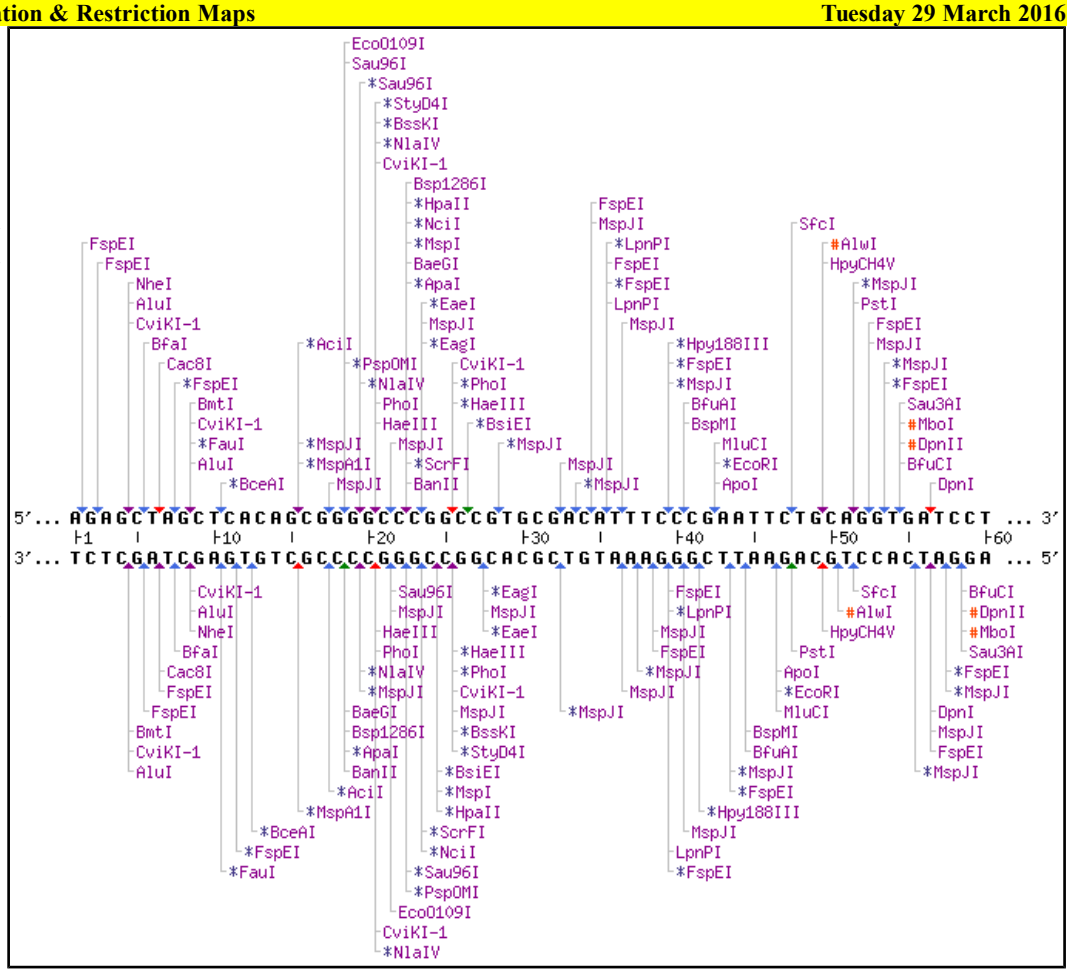


Set the **Genetic code** to **Standard**.

Elect to **Process this region only** from **15707 – 15766**. Click **OK** in the **More options** window.

Click on Submit in the main **NEBCutter** window.

You will be rewarded by a beautiful map of the same genomic **PAX6** region investigated with **remap**. You can link to the individual **REBASE** entries and customize the map in many ways. This must be a better way to ponder the map of a specific sequence.



In order to make the map you generated look a little more like the one produced by **remap**, you would need at least to restrict the number of enzymes mapped. If you recall, you told **remap** to consider only enzymes with a recognition site of **6** base pairs or more.

Click on **Custom digest** (amongst **Main options**). You will see you map in tabular form. Click on **Enzymes with a particular site length**, choose **6 bp site** and click **OK**. Your table of enzymes will shrink dramatically to meet your specification.

Pick all the listed enzymes and then click on the green **Digest** button at the bottom of the page to create a graphical map roughly equivalent of the one you made with **remap**.

Note that to find out about an enzyme you can hover your mouse over its name, or click on the name for fuller information. There is much more to investigate at this site. I leave you to discover for yourselves.

[3 bp site] - [4 bp site] - [4.5 bp site] - [5 bp site] - [6 bp site]

Site length between 6 and 6 OK

Pick all	Enzyme	Specificity	Cuts	% activity in 1	2	3	4
☑	ApaI	G_GGCC^C	1	25	50	0	100
☑	BfuAI	ACCTGCNNNN^NNNN_	1	0	75	100	10
☑	BmtI	G_CTAG^C	1	25	100	25	50
☑	BspMI	ACCTGCNNNN^NNNN_	1	?	?	100	?
☑	EagI	C^GGCC_G	1	10	25	100	10
☑	EcoRI	G^AATT_C	1	100	100	100	100
☑	NheI	G^CTAG_C	1	100	100	10	100
☑	PspOMI	G^GGCC_C	1	25	25	10	100
☑	PstI	C_TGCA^G	1	75	75	100	50

There are quite a few less enzymes mentioned in the map you have just made compared to that generated by **remap**. Can you speculate what the main reason for this might be?_____

Some enzymes in this map appear in the same place as remap predicted, but have different names. Can you explain why**?** _____

Gene Identification Software – Specifically Genscan

These programs are used to find promoters, splice sites, coding versus non-coding regions and polyadenylation signals in novel DNA sequences. These features can be combined to form complete gene model predictions.

When looking for splice sites in genomic DNA, some programs use the best currently accepted consensus (e.g. **AG** dinucleotide at intron/exon boundary) to scan the unknown sequence. Other programs are trained with a set of splice sites in known sequences (neural network approach). These applications often over predict. Prediction accuracy ranges between 50% and 95%. Recent programs for gene identification use Hidden Markov Models. There are many GeneID programs. One of the best single programs for *de novo* gene prediction is **Genscan**.

Genscan can recognise more than one potential gene in a sequence. It can also recognise partial genes. **Genscan** has been primarily designed for vertebrates. It may be less accurate for other organisms.

Go to the **Genscan** server at:

http://genes.mit.edu/GENSCAN.html

Use the default **Organism** and **Suboptimal exon cutoff**. Set **Print options** to **Predicted CDS and peptides**. Copy an paste, ***just the PAX6 genomic sequence*** from **pax6_genomic.fasta** into the approriate text box. **Genscan** does not understand **FASTA** format. Were you to use the **Upload your DNA sequence file** option, or to copy and paste the whole of this file, **18** of the **22** characters of the first line "**>pax6-genomic sequence**" would be regarded as part of the sequence[129], rendering the **Genscan** output impossible to interpret. Click on **Run GENSCAN**.

In the results that are generated, predicted genes and sub-features such as exons are numbered (**Gn.Ex**) and listed in a table displaying their **Type**, strand (**S**), start (**Begin**) and end (**End**) positions, and length (**Len**). A probability (**P**) is assigned to each predicted exon. High probability exons (**P > 0.99**) are nearly always correct, those with **0.50 < P < 0.99** are correct most of the time, the rest are not reliable.

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.01	Init	+	5509	5600	92	0	2	65	24	139	0.462	3.17
1.02	Intr	+	7682	8297	616	2	1	37	65	397	0.257	24.65
1.03	Intr	+	12000	12060	61	2	1	96	98	38	0.917	3.91
1.04	Intr	+	12789	13073	285	1	0	92	107	50	0.900	4.61
1.05	Intr	+	15326	15442	117	0	0	71	52	55	0.567	0.64
1.06	Intr	+	15628	15758	131	2	2	141	58	149	0.999	17.61
1.07	Intr	+	16550	16591	42	1	0	62	85	51	0.608	0.64
1.08	Intr	+	16686	16901	216	2	0	104	94	112	0.995	12.10
1.09	Intr	+	17606	17771	166	1	1	76	93	72	0.974	6.03
1.10	Intr	+	17856	17927	72	1	0	88	42	58	0.553	0.58
1.11	Intr	+	23674	23832	159	2	0	108	54	194	0.919	17.96
1.12	Intr	+	24348	24430	83	1	2	29	86	85	0.945	1.76
1.13	Intr	+	24660	24810	151	2	1	72	58	76	0.958	2.74
1.14	Intr	+	24909	25024	116	1	2	113	108	94	0.977	14.07
1.15	Intr	+	27602	27752	151	1	1	115	99	80	0.992	11.54

S.Start	S.Fin
7245	7347
7447	7634
11537	11613
12000	12060
15628	15758
16686	16901
17606	17771
23674	23832
24348	24430
24660	24810
24909	25024
27602	27752
28443	28540

Illustrated here are the results I got alongside the very authoritative predictions from **splign** that you computed from the same data earlier. Overall, the prediction is more than reasonable, given the approximate nature of the training set. Observations include:

- **9** of the **10** genes suggested by **splign** between bases **12000** and **28540** were predicted exactly by **Genscan**.

129 The ">", "-", "**6**" and the **space** would be thrown away. Many, but not all, of the other **18** characters are valid DNA ambiguity codes (see **Appendix I**). **GENSCAN** seems willing to attach meaning to ***all*** the letters, even the ones not part of the **IUB** alphabet?

- **Genscan** missed the final exon (**28443-2850**) and so failed to predict a complete gene.
- Of the **4** exons **Genscan** predicts in this region but **splign** denies, one is the **42** base pair exon that is only present in **isoform 5a**. This is genuine, but **splign** would not see it as the **cDNA** used was not an **isoform 5a** sequence.
- The other **3** "extra" exons in this region were all predicted with suspiciously low probability scores (**17856-17927 P=0.553**, **16550-16591 P=0.608**, **15326-15442 P=0.567**).
- **Genscan** has not made a particularly impressive job of predicting the non-coding exons at the **5'** end of the gene.

A full explanation of the output table used to be included at the bottom of the results page. It has now been removed, but as I think it was useful, I include it here. Note that **Intr** is not short for **Intr**on, as most sensible people might guess, it is short for **Inter**nal exon!! I suspect a warped sense of humour here?

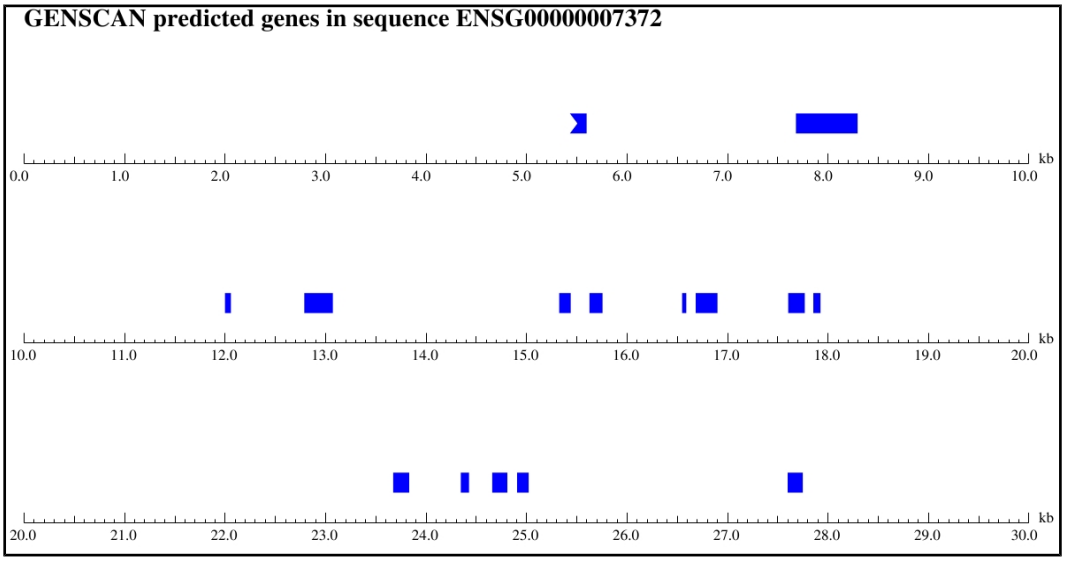
```
Explanation

Gn.Ex  : gene number, exon number (for reference)
Type   : Init = Initial exon (ATG to 5' splice site)
         Intr = Internal exon (3' splice site to 5' splice site)
         Term = Terminal exon (3' splice site to stop codon)
         Sngl = Single-exon gene (ATG to stop)
         Prom = Promoter (TATA box / initation site)
         PlyA = poly-A signal (consensus: AATAAA)
S      : DNA strand (+ = input strand; - = opposite strand)
Begin  : beginning of exon or signal (numbered on input strand)
End    : end point of exon or signal (numbered on input strand)
Len    : length of exon or signal (bp)
Fr     : reading frame (a forward strand codon ending at x has frame x mod 3)
Ph     : net phase of exon (exon length modulo 3)
I/Ac   : initiation signal or 3' splice site score (tenth bit units)
Do/T   : 5' splice site or termination signal score (tenth bit units)
CodRg  : coding region score (tenth bit units)
P      : probability of exon (sum over all parses containing exon)
Tscr   : exon score (depends on length, I/Ac, Do/T and CodRg scores)
```

Another way to gain an impression of how well **Genscan** has performed, is to align the predicted protein with the protein from **UniProtKB** that you have stored in **pax6_human.fasta**. To save time, I have done that for you with the **EMBOSS** program **needle** (sensitive global alignment option). **needle** was run with default parameters from the **EBI** web pages. I have edited the output in this illustration, but only to reduce its size.

Genscan has regarded all the exons it predicts as coding, hence the messy beginning. The effects of the missing internal exons are very clear to see, including the familiar **14** amino acids of **isoform 5a**, missing in the **UniprotKB** entry, but correctly predicted by **Genscan**.

The consequences of the missing terminating exon is also clear to see.

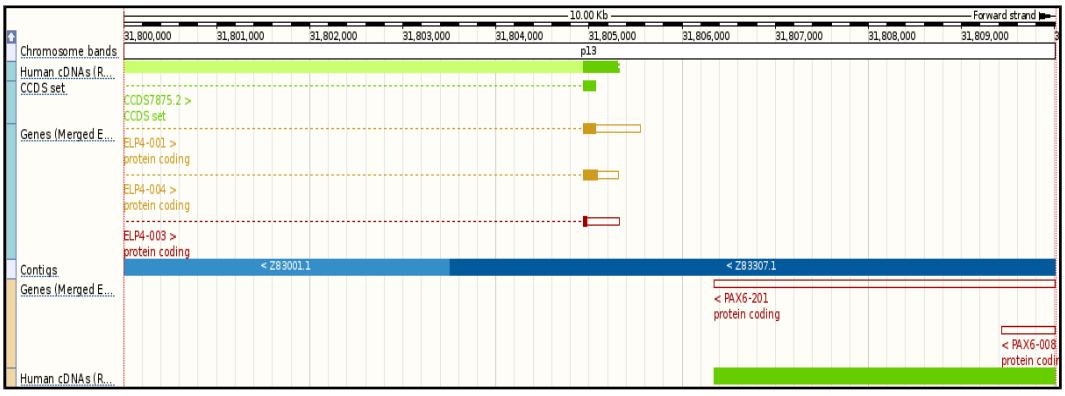
```
GENSCAN pep    1 MWPGGPGRLAASSLSRLPGAAATRPVSSEFRQVSPFLPPSFLPSRRSPPP  50
GENSCAN pep   51 LLRRESWAGEVSTQEPKLSTSSESPRRRATAGERERRRARGGRPAQPQRQ 100
GENSCAN pep  101 RQRQRGLSSREEGRLQARQQPRLLAPASPNAPDPREKTKRLCGAQGPGTA 150
GENSCAN pep  151 AESNRCRLLCPARGPANAQRAGAPRTRRGSAEGTCSPRLRPPAGPRALCR 200
GENSCAN pep  201 TGLRRRRVSRHSSAEGGTAPATSPSGVSRQGSRTLESPIFEPRGIPRPPA 250
GENSCAN pep  251 RASMQNNPVGSHPPSCSWWVLVRGRASLELRPGSLEAQRNQARFEVTVEA 300
GENSCAN pep  301 HGVGAQLNSKCGKARVFQQPALAEPHRLKEGPQHLIHLFWKRRSPPGASG 350
GENSCAN pep  351 AVQWAGARLLGLGCPMSGMAGGPVVLYLIDSRELTAEVGTGHSGVNQLGG 400
                                                 :...|||||||||
PAX6_HUMAN     1                                 MQNSHSGVNQLGG  13
GENSCAN pep  401 VFVNGRPLPDSTRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNVS 450
                 ||||||||||||||||||||||||||||||||||              ||
PAX6_HUMAN    14 VFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ--------------VS  49
GENSCAN pep  451 NGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFA 500
                 ||||||||||||||||||||||||||||||||||||||||||||||||||
PAX6_HUMAN    50 NGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFA  99
GENSCAN pep  501 WEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGADGMYDKLRML 550
                 ||||||||||||||||||||||||||||||||||||||||||||||||||
PAX6_HUMAN   100 WEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGADGMYDKLRML 149
GENSCAN pep  551 NGQTGSWGTRPGWYPGTSVPGQPTQGLPCLGNVMGETAFEGLGHIDHIVY 600
                 |||||||||||||||||||||||.
PAX6_HUMAN   150 NGQTGSWGTRPGWYPGTSVPGQPTQD------------------------ 175
GENSCAN pep  601 GCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEAL 650
                 ||||||||||||||||||||||||||||||||||||||||||||||||||
PAX6_HUMAN   176 GCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEAL 225
GENSCAN pep  651 EKEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQR 700
                 ||||||||||||||||||||||||||||||||||||||||||||||||||
PAX6_HUMAN   226 EKEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQR 275
GENSCAN pep  701 RQASNTPSHIPISSSFSTSVYQPIPQPTTPVSSFTSGSMLGRTDTALTNT 750
                 ||||||||||||||||||||||||||||||||||||||||||||||||||
PAX6_HUMAN   276 RQASNTPSHIPISSSFSTSVYQPIPQPTTPVSSFTSGSMLGRTDTALTNT 325
GENSCAN pep  751 YSALPPMPSFTMANNLPMQPPVPSQTSSYSCMLPTSPSVNGRSYDTYTPP 800
                 ||||||||||||||||||||||||||||||||||||||||||||||||||
PAX6_HUMAN   326 YSALPPMPSFTMANNLPMQPPVPSQTSSYSCMLPTSPSVNGRSYDTYTPP 375
GENSCAN pep  801 HMQTHMNSQPMGTSGTTSTX                              820
                 ||||||||||||||||||
PAX6_HUMAN   376 HMQTHMNSQPMGTSGTTSTGLISPGVSVPVQVPGSEPDMSQYWPRLQ    422
```



At this point I would love to suggest you take a look at the **PDF** and/or postscript images using the links at the top of your results page. Unfortunately, for some time now they have not worked. In compensation, I offer the picture you would get if the links still worked[130]. The pictures are very beautiful, but it should be remembered that making predictions look pretty does not improve their accuracy.

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.01	Init	+	5509	5600	92	0	2	65	24	139	0.462	3.17
1.02	Intr	+	7682	8297	616	2	1	37	65	397	0.257	24.65
1.03	Intr	+	12000	12060	61	2	1	96	98	38	0.917	3.91
1.04	Intr	+	12789	13073	285	1	0	92	107	50	0.900	4.61
1.05	Intr	+	15326	15442	117	0	0	71	52	55	0.567	0.64
1.06	Intr	+	15628	15758	131	2	2	141	58	149	0.999	17.61
1.07	Intr	+	16550	16591	42	1	0	62	85	51	0.608	0.64
1.08	Intr	+	16686	16901	216	2	0	104	94	112	0.995	12.10
1.09	Intr	+	17606	17771	166	1	1	76	93	72	0.974	6.03
1.10	Intr	+	17856	17927	72	1	0	88	42	58	0.553	0.58
1.11	Intr	+	23674	23832	159	2	0	108	54	194	0.919	17.96
1.12	Intr	+	24348	24430	83	1	2	29	86	85	0.945	1.76
1.13	Intr	+	24660	24810	151	2	1	72	58	76	0.958	2.74
1.14	Intr	+	24909	25024	116	1	2	113	108	94	0.977	14.07
1.15	Intr	+	27602	27752	151	1	1	115	99	80	0.992	11.54
1.16	Term	+	28405	28442	38	2	2	68	48	31	0.666	-5.40
1.17	PlyA	+	29242	29247	6							1.05
2.03	PlyA	-	29529	29524	6							1.05
2.02	Term	-	35069	34938	132	2	0	74	50	134	0.825	6.29
2.01	Intr	-	38749	38623	127	0	1	109	30	61	0.727	3.08

I tried giving **Genscan** a longer portion of the genomic sequence (an extra **6000** base pairs downstream) and found I got a prediction with a **Term**inal exon (close, but not quite right) and a **polyA**. The extra sequence must offer better context.



I also got a partial gene on the opposite strand. A quick look with **Ensembl** at the region towards the end of the **PAX6** suggests this could be the last part of **ELP4**?

The longer genomic sequence is available if you wish to experiment further (or get it yourself with **Ensembl**).

Genscan predictions are dependent upon the choice of the set of training sequences used. This website (and all the other **Genscan** sites I have used) offer a very limited range of training set choices. In the circumstances, the prediction is surprisingly accurate. For serious use, it would be necessary to install **Genscan** locally and to use customized sets of training data.

130 I have asked whether these pictures will ever again be available, I await a reply. So long ago, my expectations have all but expired.

Early Secondary Structure Prediction Methods - GOR

A number of simple methods for investigating protein secondary structure were developed from as far back as the late **1960**s, early **1970**s. Significantly better methods emerged as the **1970**s progressed, claiming accuracy of around **60%** or more. The method we will look at here is one of those improved methods. We will look at the method due to **G**arnier, **O**sguthorpe and **R**obson (**GOR**), originally published in **1978**, but developed/improved well beyond that time. The latest version being **Version V** (published **2002**, a new server announced **2005**). **GOR V** incorporates more modern approaches and has been merged into other prediction strategies. **Wikipedia** offers an informative article on protein secondary structure that is well worth a look if this is an area of particular interest for you.

The older methods, such as **GOR**, are not as accurate as the more modern approaches investigated in the main exercise, however, they do have the advantage of being very quick and easy to run. They provide a reasonable prediction (usually) with little fuss and so still have a role to play, I would suggest.

First a quick glance at the implementation of the **GOR** method in the **EMBOSS** package. This is the method as it was first published. We run it here only to contrast with the better options of the main exercise and more recent version of the same program!

There are a number of ways to access and run the programs of the **EMBOSS** package. Here we will use a web interface called **Emboss Explorer**, as implemented at the **Wageningen Bioinformatics Webportal**, a major Bioinformatics Service Centre of the Netherlands.

Go first to the **Wageningen EMBOSS service** (**http://emboss.bioinformatics.nl/**).

PROTEIN 2D STRUCTURE
garnier
helixturnhelix
hmoment
pepcoil
pepnet
pepwheel
tmap
topo

Select **garnier** from the **PROTEIN 2D STRUTURE** section of the program list.

Use the **Browse** facility **To upload a sequence from your local computer**. The required sequence is the one in **pax6_human.fasta**. Default parameter settings are fine. Click on the **Run garnier** button.

This version of **GOR** suggests the most likely of four types of secondary structure: Alpha Helices (**H**); Beta Sheets (**E**); Coils (**C**) and Turns (**T**), at each position of the protein sequence.

Some features are composed of only one or two residues. Especially in the case of larger structures such as helices or sheets, such predictions are dubious. Later versions of the algorithm avoid the more foolish predictions.

```
          .   10    .   20    .   30    .   40    .   50
      MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSN
helix                              HH     H
sheet EE       EEEEEEEE           EEEEE         EEEEEE
turns                        T          TTTT TTT      TTT
 coil   CCCCCCC         CCCCC CC            C
          .   60    .   70    .   80    .   90    .  100
      GCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAW
helix                                 HHHHHHHHH        HHH
sheet   EEEEEEEE      EEEE       EEE
turns TT         TTTTT     TTT   T             TTTT TT
 coil                C       CCC                   C  C
          .  110    .  120    .  130    .  140    .  150
      EIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGADGMYDKLRMLN
helix HHH                           HHHHHHHHHHHHHHHHH
sheet                      EEEEEEEEE                  EEE
turns    TT TTTTTTTTT   T
 coil      C         CCC          CC                   C
          .  160    .  170    .  180    .  190    .  200
      GQTGSWGTRPGWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEA
helix                                                HHH
sheet                             EEE
turns TT  TT   TTT TTT  TT  TTTTT    TT
 coil   CC  CCC   C   CC  CC        C  CCCCCCCCCCCCCC
          .  210    .  220    .  230    .  240    .  250
      QMRLQLKRKLQRNRTSFTQEQIEALEKEFERTHYPDVFARERLAAKIDLP
helix HHHHHHH            HHHHHHHHHHHH     HHHHHHHHHHHHHH
sheet          E
turns        TT TTTT                 T   T
 coil               CCCCC              CCC
          .  260    .  270    .  280    .  290    .  300
      EARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIP
helix HHHH           HHHH    H
sheet     EEE                          EEEEE
turns           TT        TTTT TTTTT   T       TTTTTT
 coil        CCC  CCC              CCC      CC       CCCC
          .  310    .  320    .  330    .  340    .  350
      QPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQ
helix
sheet     EEEEE       EE EEEE          EEEEE        EEE
turns  T                      TTTT          TT  TT     TT
 coil C CC     CCCCCCC  C        CCCCC       CC  CC
          .  360    .  370    .  380    .  390    .  400
      TSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMNSQPMGTSGTTSTGLISPG
helix
sheet     EEEE                    EEE              EEEE
turns TTTT    TT  T  TTTTTTT  TT     T   T             T
 coil           CC CC        CC     CC CCC CCCCCCC    CC
          .  410    .  420
      VSVPVQVPGSEPDMSQYWPRLQ
helix
sheet EEEEEEE
turns         T   TTTTT  TTT
 coil        C CCC     CC
```

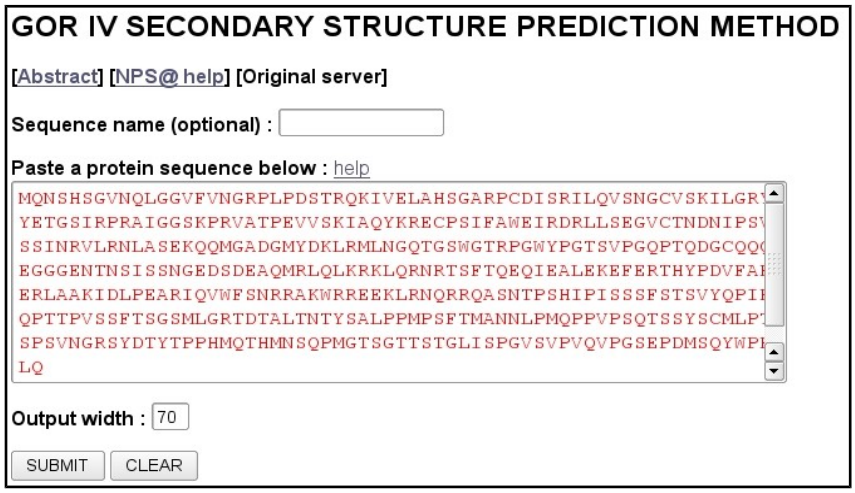
How credible would you say was the prediction at amino acid 33? _____

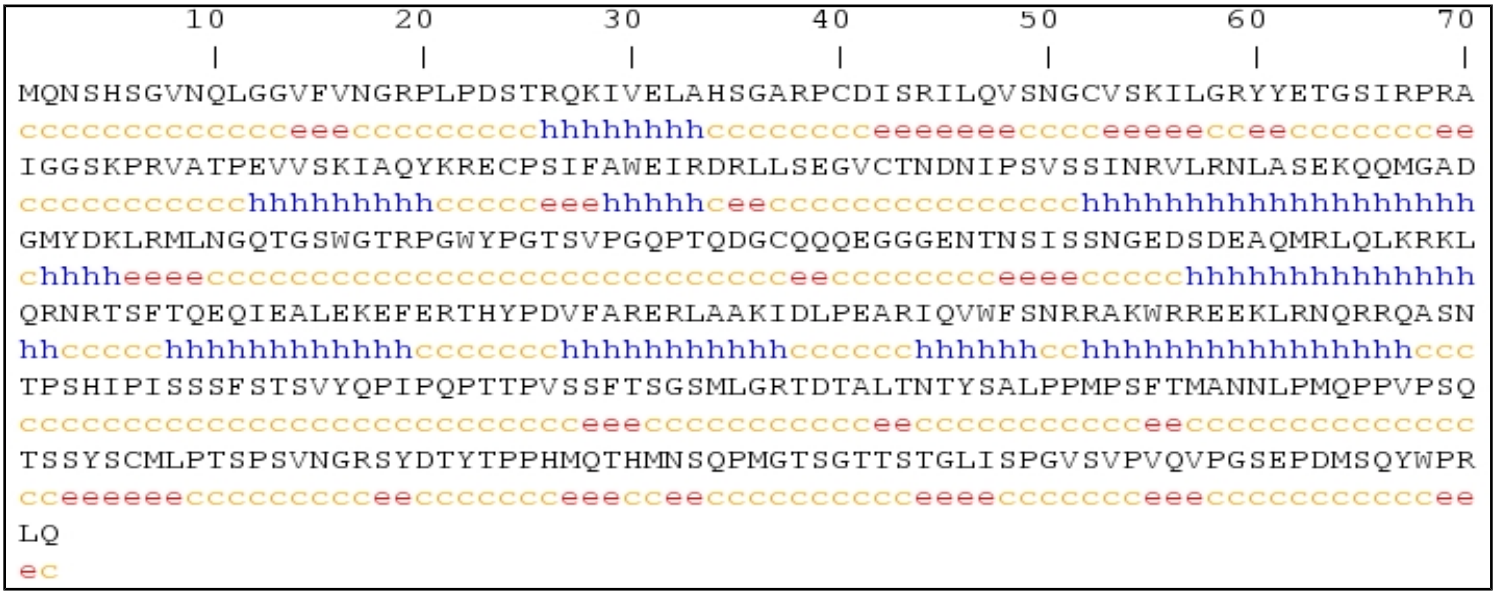
How would you rate the prediction overall? _____

Although the latest version of **GOR** is **GOR V**[131], **GOR IV**[132] is more available, reliable and recognisably similar to the original **GOR**. So try **GOR IV** at:

http://npsa-pbil.ibcp.fr

Pôle Bioinformatique Lyonnais (**PBIL**) (Lyons University) offering a wide range of interesting services including a number of ways to predict protein secondary structure. Move down to the **Secondary structure prediction** and click on the link to **GOR IV**. Copy and paste the **PAX6** protein from the file **pax6_human.fasta** into the appropriate text box (just the amino acid codes, this server does not like **fasta** format). Click on the **SUBMIT** button.



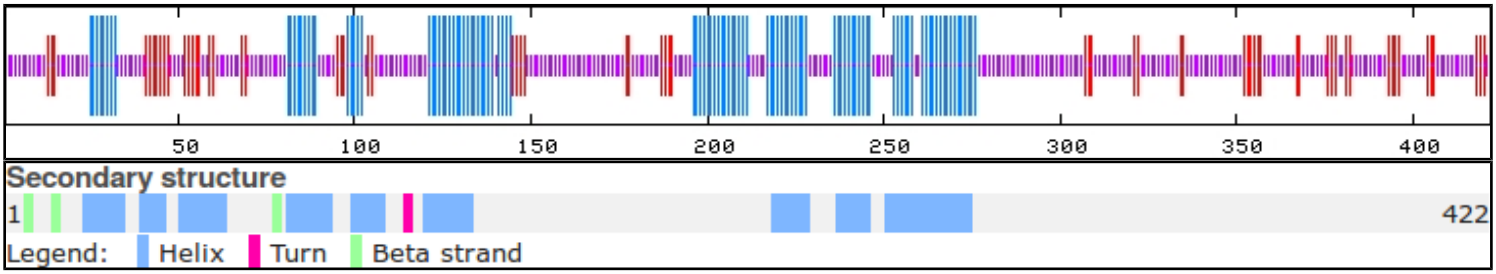


GOR IV, unlike the **EMBOSS garnier** program (**GOR I**), does not try to differentiate between **C**oil and **T**urn. It chooses for each position the most likely from **H**elix, **E**xtended (i.e. beta sheet) or unstructured (i.e. **C**oil).

Also unlike the initial **GOR** implementation of the **garnier** program, **GOR IV** only predicts helices of at least four residues and beta sheets of at least two residues.

How does the prediction at position **33** compare with that of **garnier**?_____

Further down your output is a graphic showing how the structure predictions are positioned along the protein sequence. This can usefully be compared to part of the graphic you considered earlier, showing where these features lie in the view of **Uniprot**.



How would you rate the prediction overall?_____

131 Read all about it at: **http://bioinformatics.oxfordjournals.org/cgi/content/full/21/11/278.**

132 **J. Garnier. J.F. Gibrat and B.Robson in Methods in Enzymology, vol 266, p 540-553 (1996)**. **GOR IV** uses a data base of **267** proteins to validate its predictions which it claims to have a mean accuracy of **64.4%** for a three state prediction (**H**, **E** or **C**).

Domain & Motif prediction with Prosite, Pfam, HTH & PRINTS

In this supplementary exercise, we will search for **protein sequence motifs** and **protein domains** using individual database searches and specific simple specialised software. Just in case you still have doubts, I have linked the references above to the wikipedia pages that clearly define both motifs and domains in this context. You may have gathered, I am a big fan of **Wikipedia**.

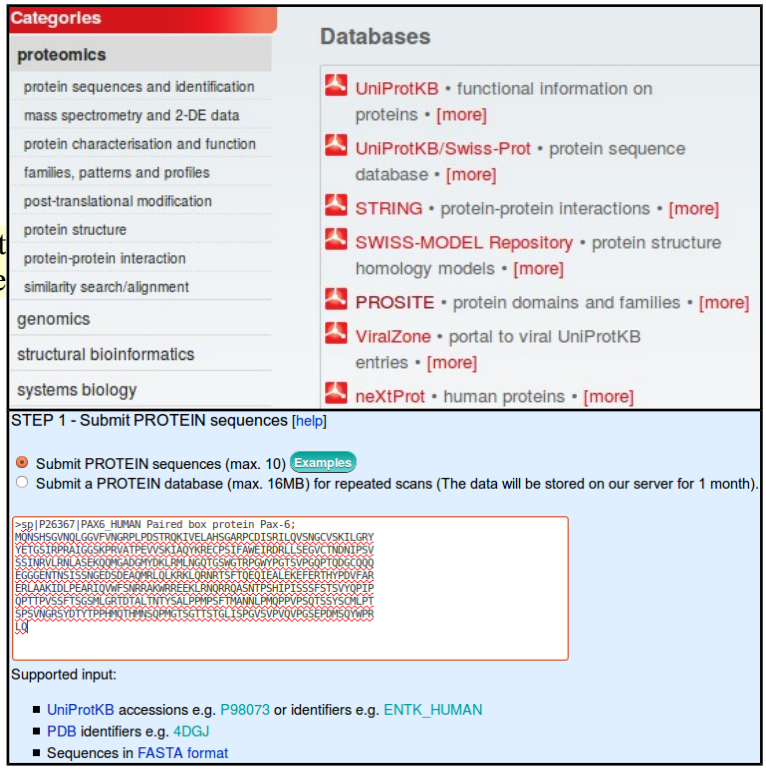
Searching

A major database for both motifs and domains is **PROSITE**. Many sequence motifs in this database are described using a simple pattern description language. They include some very common, simple motifs, many only a few residues long, that indicate possible sites for post-translational modifications (e.g. glycosylation or phosphorylation).

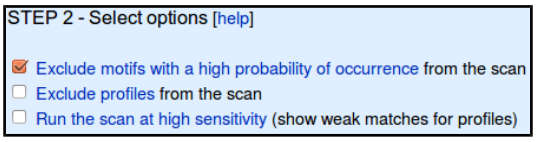
Go to the **ExPASy**[133] site at:

http://www.expasy.org

Select **proteomics** from the list of **Categories**. Select **PROSITE** from the **Databases** section. Click on the **ScanProsite** link at the top of your page.

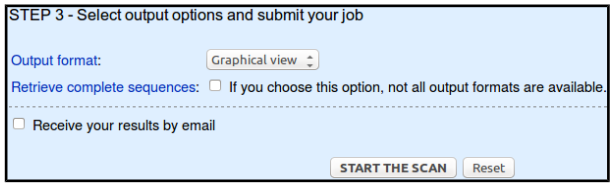


Enter **pax6_human** in the **STEP 1 - Submit PROTEIN sequences** section.



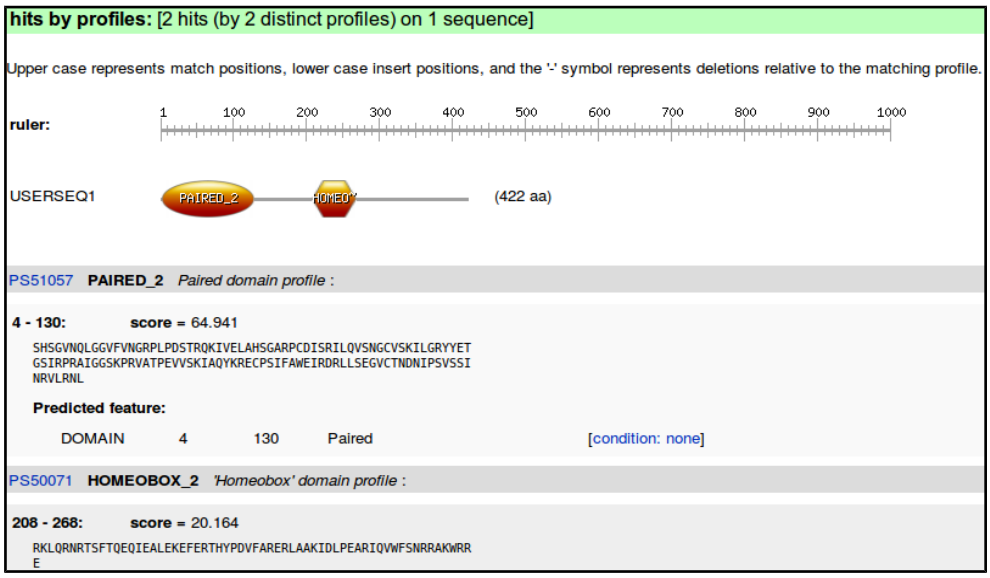
In the **STEP 2 - Select options** section, ensure that the **Exclude motifs with a high probability of occurrence** box is ticked.

The defaults offered in the **STEP 3 - Select output options and submit your job** section are fine so just click on the **START THE SCAN** button. In but a few moments, your results will burst forth.

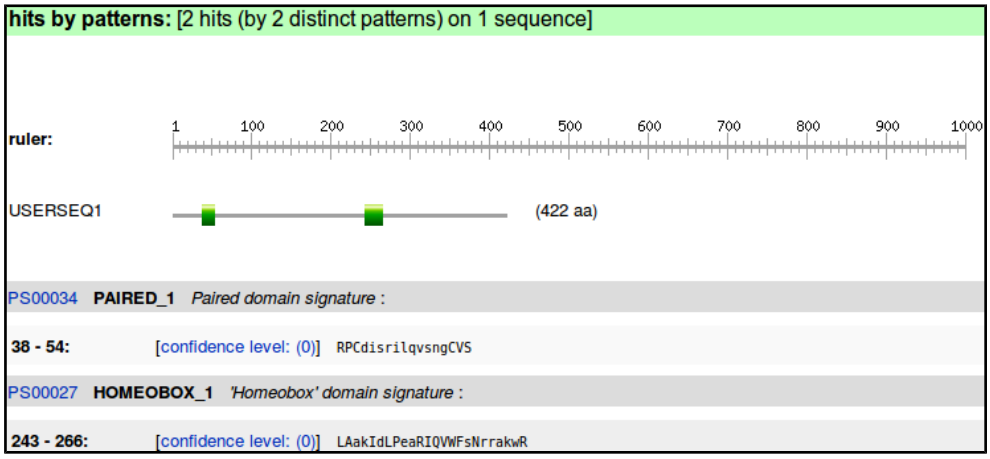


133 **Expasy** is a major site for protein based research in Switzerland. As the all knowing **Wikipedia** puts it:
"**ExPASy** is a **bioinformatics** resource portal operated by the **S**wiss **I**nstitute of **B**ioinformatics (**SIB**) and in particular the **SIB Web Team**. It is an extensible and integrative portal accessing many scientific resources, databases and software tools in different areas of life sciences. Scientists can access a wide range of resources in many different domains, such as **proteomics**, **genomics**, **phylogeny/evolution**, **systems biology**, **population genetics**, and **transcriptomics**."

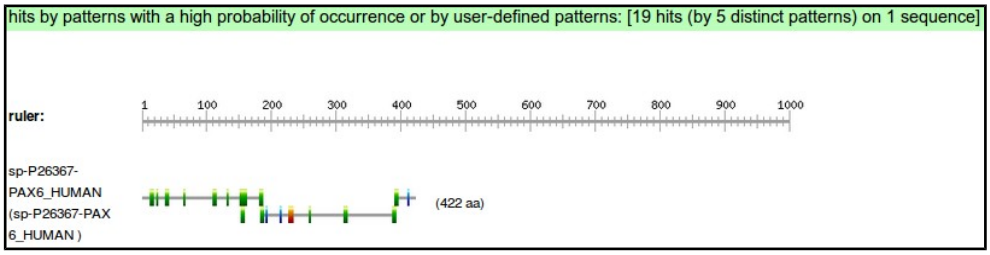
Two hits with **PROSITE** profiles suggesting the familiar domains in their familiar places.

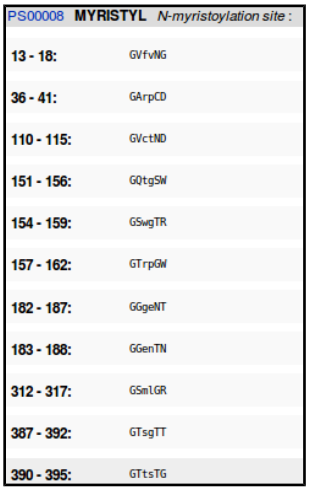


Two hits with **PROSITE** patterns confirming the same domains by matching their highly conserved subregions. Confirmation of what has already been discovered more than once, but this time, discovered by running a database search program manually. Exactly this program was used to generate the annotation read earlier.



Move back to the search submission page. In **Step 2**, deselect **Exclude patterns with a high probability of occurrence**. **START THE SCAN** again.





This time you will see many more hits with patterns.

Follow the link to the documentation for an **N-myristoylation site** (**PS00008**).

What is the signature pattern for **N-myristoylation** site? _____

How would you interpret this pattern? _____

How many **N-myristoylation** sites did **ScanProsite** suggest there might be in **PAX6_HUMAN**? _____

How many real **N-myristoylation** sites would you guess there might be in **PAX6_HUMAN**? _____

Searching Pfam

Go to the home of **Pfam** at:

http://pfam.sanger.ac.uk/

Select the **VIEW A SEQUENCE** option. Enter **pax6_human** (or the corresponding accession code) into the proffered space and press the **Go** button. You will be taken to the sequence and offered links to view (again) typical 3D structures for the domains of this protein. Also, you are offered the opportunity to generate easily a phylogenetic tree based upon **PAX6** from the **TreeFam** database, which is fun if nothing else. We will not be seriously covering phylogeny in the course of these exercises, but why not try it anyway by clicking on the **TreeFam** link.

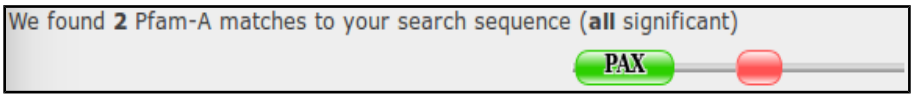
Fine, but you are just looking at what has already been decided. Here we set out to discover, by analysis. How could you use **Pfam** for a sequence that has yet to be annotated.

Go back to the home of **Pfam** at:

http://pfam.sanger.ac.uk/

This time select the **SEQUENCE SEARCH** option. Copy and paste the sequence of **PAX6_HUMAN** into the appropriate box. Click on the **Go** button.

You should discover nothing you did not expect. This same conclusions, but via investigation of the sequence rather than database lookup.



Significant Pfam-A Matches

Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To					
PAX	'Paired box' domain	Domain	CL0123	4	128	4	128	1	125	125	238.8	8.5e-72	n/a	Show
Homeobox	Homeobox domain	Domain	CL0123	211	267	212	267	2	57	57	79.7	9.3e-23	n/a	Show

Have a look around generally, but in the course of your investigations, Click on one of the **CL0123** links. You will see that both the **PAX** and **Homeobox Pfam** families belong to a collection of families (a **Clan**) all of which contain **helix-turn-helix** motifs and are mostly involved in DNA binding. Unsurprisingly, the clan in question is the **Helix-turn-helix** clan.

Would you have found anything of interest had you chosen "to search for Pfam-B matches" (you will need to run the search again in order to answer this one)? _____

Searching for Helix-Turn-Helix motifs

When looking at the **Prosite PAX** and **homeobox** documentation earlier, you will have recorded three positions where there should be **Helix-Turn-Helix** motifs that bind the DNA major groove. The **EMBOSS** package includes a program, **helixturnhelix**, that searches for these motifs. **helixturnhelix**[134] uses a scoring matrix derived from alignments of known **HTH** motifs. The original matrix of 1987 looked for motifs spanning 20 amino acids, the later (1990) improved matrix was based on an alignment of 91 **HTH** motifs and was 22 amino acids wide.

There are a number of ways to access and run the programs of the **EMBOSS** package. Here we will use a web interface called **Emboss Explorer**, as implemented at the **Wageningen Bioinformatics Webportal**, a major Bioinformatics Service Centre of the Netherlands.

PROTEIN 2D STRUCTURE
garnier
helixturnhelix
hmoment
pepcoil
pepnet
pepwheel
tmap
topo

Go first to the **Wageningen EMBOSS service** (**http://emboss.bioinformatics.nl/**). Select **helixturnhelix** from the **PROTEIN 2D STRUTURE** section of the list of programs.

Use the **Browse** facility **To upload a sequence from your local computer**. The required sequence is the one you saved in the file **pax6_human.fasta**, of course. The default parameter settings are fine for a first try. Click on the **Run helixturnhelix** button.

helixturnhelix will suggest just one part of the sequence that might be a **Helix-Turn-Helix**.

To which, if any, of the expected **HTH** motifs does this prediction correspond? _____

_____

```
#=======================================
#
# Sequence: PAX6_HUMAN     from: 1    to: 422
# HitCount: 1
#
# Hits above +2.50 SD (972.73)
#
#=======================================

Maximum_score_at at "*"

(1) Score 1109.000 length 22 at residues 238->259
            *
 Sequence: FARERLAAKIDLPEARIQVWFS
           |                     |
         238                   259
 Maximum_score_at: 238
 Standard_deviations: 2.96
```

```
#=======================================
#
# Sequence: PAX6_HUMAN     from: 1    to: 422
# HitCount: 2
#
# Hits above +2.50 SD (972.73)
#
#=======================================

Maximum_score_at at "*"

(1) Score 1086.000 length 20 at residues 39->58
            *
 Sequence: PCDISRILQVSNGCVSKILG
           |                   |
          39                  58
 Maximum_score_at: 39
 Standard_deviations: 2.89

(2) Score 1028.000 length 20 at residues 240->259
            *
 Sequence: RERLAAKIDLPEARIQVWFS
           |                   |
         240                 259
 Maximum_score_at: 240
 Standard_deviations: 2.69
```

As the expectation was that three **HTH** motifs might be found, try again. This time, in the Additional Section of **heleixturnhelix**'s menu window, choose to **Use the old (1987) weight data**.

This time you will find **2 HTH** motifs, essentially the one you found previously, plus another. Note that the length of the predictions have shrunk from **22** amino acids to **20**.

To which, if any, of the expected **HTH** motifs does the new prediction correspond? _____

_____

134 This method is described in: Dodd I.B., Egan J.B. (1987) "Systematic method for the detection of potential lambda cro-like DNA-binding regions in proteins." J. Mol. Biol. 194: 557-564. Revised in Dodd I.B., Egan J.B. (1990) "Improved detection of helix-turn-helix DNA-binding motifs in protein sequences." Nucleic Acids Res. 18: 5019-5026.
You can download the more recent paper from: **http://nar.oxfordjournals.org/cgi/reprint/18/17/5019.pdf**

It seems a bit odd to get the better result from the older scoring matrix, so try once more, using the default scoring matrix. This time, in the **Additional section**, set the **Minimum SD** to **2.0** before you click **OK**.

```
#=======================================
#
# Sequence: PAX6_HUMAN     from: 1   to: 422
# HitCount: 2
#
# Hits above +2.00 SD (825.93)
#
#=======================================

Maximum_score_at at "*"

(1) Score 1109.000 length 22 at residues 238->259
           *
 Sequence: FARERLAAKIDLPEARIQVWFS
           |                     |
         238                   259
 Maximum_score_at: 238
 Standard_deviations: 2.96

(2) Score 827.000 length 22 at residues 37->58
           *
 Sequence: ARPCDISRILQVSNGCVSKILG
           |                     |
          37                    58
 Maximum_score_at: 37
 Standard_deviations: 2.00
```

By asking to see hits just 2 standard deviations away from the average (random) score, you see a second reasonable prediction. Lowering your standards further will be to no avail with either scoring matrix. Either the third suggested **HTH** is not present in **PAX6_HUMAN**, or **helixturnhelix** is simply not sufficiently clever to find it.

Can you suggest anything particular about the third putative **HTH** motif that might explain its reluctance to be discovered? _____

_____

Searching PRINTS

The **PRINTS** database defines functional protein families. Domains are identified by a number of short, ordered, well-conserved regions. A full match to one of these "fingerprints" will match all the relevant short regions in the correct order. A partial match is recorded if some are missing or if they occur in an incorrect order. **PRINTS** can be searched using the **fingerPRINTscan** program.

Go to the **fingerPRINTscan** home page:

http://umber.sbs.man.ac.uk/fingerPRINTScan/

Select the **FPScan** link[135] and paste in the **PAX6_HUMAN** sequence in raw format. Leave all defaults and hit the **Send Query** button.

Highest scoring fingerprints for your query			
Fingerprint	**E-value**	**GRAPHScan**	**Motif3D**
PAIREDBOX (relations)	1.499643e-43	Graphic	

The top hit is with the **PAIREDBOX fingerprint**. No surprise here. Move down to the list of the best **10** hits.

Ten top scoring fingerprints for your query									
Ancestry	**Fingerprint**	**No. of Motifs**	**SumId**	**AveId**	**PfScore**	**Pvalue**	**Evalue**	**GRAPHScan**	
PAIREDBOX	PAIREDBOX	4 of 4	3.5e+02	87	3213	1.3e-49	1.5e-43	IIII	Graphic
HTHREPRESSR	HTHREPRESSR	2 of 2	75.92	37.96	586	5.3e-08	0.17	II	Graphic
POUDOMAIN	POUDOMAIN	2 of 5	65.80	32.90	577	1.7e-07	0.39	...II	Graphic
HOMEOBOX	HOMEOBOX	2 of 3	102.06	51.03	724	3e-07	1.2	.II	Graphic
PRICHEXTENSN	PRICHEXTENSN	3 of 8	102.84	34.28	664	1.2e-05	20	.iIi....	Graphic
POAALLERGEN	POAALLERGEN	2 of 8	42.41	21.20	393	7e-05	1.7e+02	.....i.i	Graphic
7TM-->GPCRCLAN-->GPCRRHODOPSN-->LTBRECEPTOR-->LTB1RECEPTOR	LTB1RECEPTOR	2 of 6	71.96	35.98	371	0.00032	8.4e+02	..I.I.	Graphic
PROTEINF153	PROTEINF153	2 of 5	52.81	26.40	458	0.00038	6.9e+02	i...i	Graphic
ACONITASE	ACONITASE	2 of 9	63.61	31.80	336	0.00047	1.5e+03	.....iI.	Graphic
GLIADGLUTEN-->GLIADIN	GLIADIN	2 of 9	73.82	36.91	396	0.0013	3.7e+03	.I.....I	Graphic

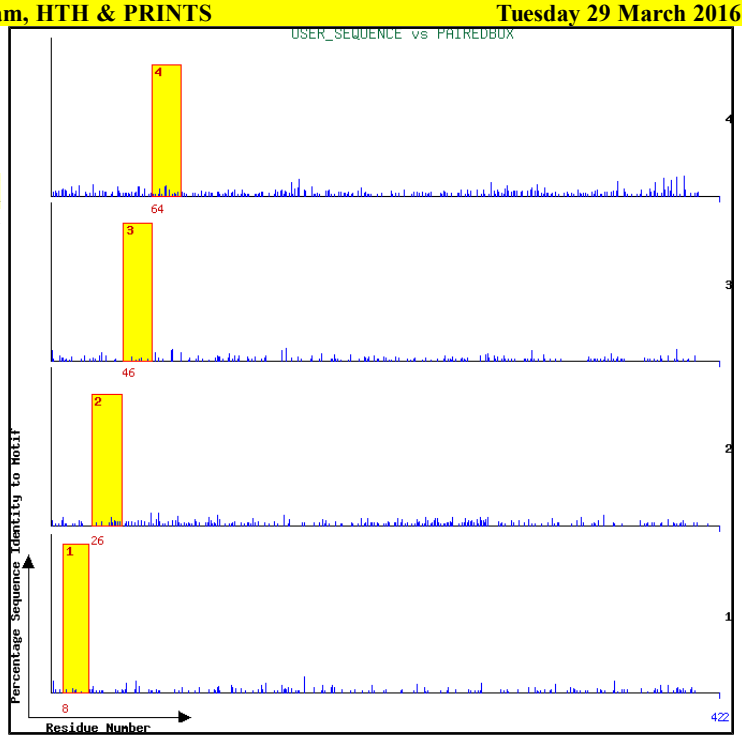
In the list of **Ten top scoring fingerprints**, there is a second **fingerprint** that matches all elements in the correct order. This is the **HTHREPRESSR**. Click on the **HTHREPRESSR** link and from the documentation you can confirm that an **HTHREPRESSOR** is an **HTH** motif of the sort detected by the **EMBOSS** program **helixturnhelix**. Move back to your **fingerPRINTscan** results.

Ten top scoring fingerprints for your query. Detailed by motif									
FingerPrint Name	Motif Number	IdScore	PfScore	Pval	Sequence	Length	low	Pos	high
PAIREDBOX	1 of 4	93.82	815	1.01e-12	VNQLGGVFVNGRPLPD	16	0	8	0
	2 of 4	82.91	821	6.08e-13	RQKIVELAHSGARPCDISR	19	0	26	0
	3 of 4	87.39	809	2.95e-12	LQVSNGCVSKILGRYYET	18	0	46	0
	4 of 4	83.08	768	6.99e-14	GSIRPRAIGGSKPRVATP	18	0	64	0
HTHREPRESSR	1 of 2	32.91	134	3.98e-02	ARERLAAKID	10	0	239	0
	2 of 2	43.00	452	1.34e-06	DLPEARIQVWFSNRRAK	17	0	248	0

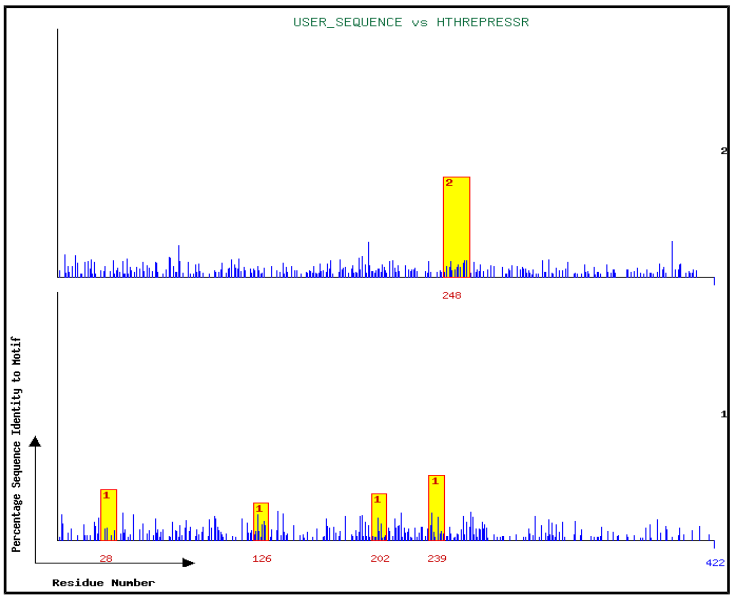
From the Position information included in the **Detailed by motif** table, you can see that the **HTH** motif that **fingerPRINTscan** finds is the same one that the **EMBOSS** program **helixturnhelix** found with default settings.

135 Alternatively you can use the mirror at the **EBI http://www.ebi.ac.uk/printsscan/**

Take a look at the **GRAPHScan** for the **PAIREDBOX** prediction and see that is is good! Four out of four very positive motif matches are shown.



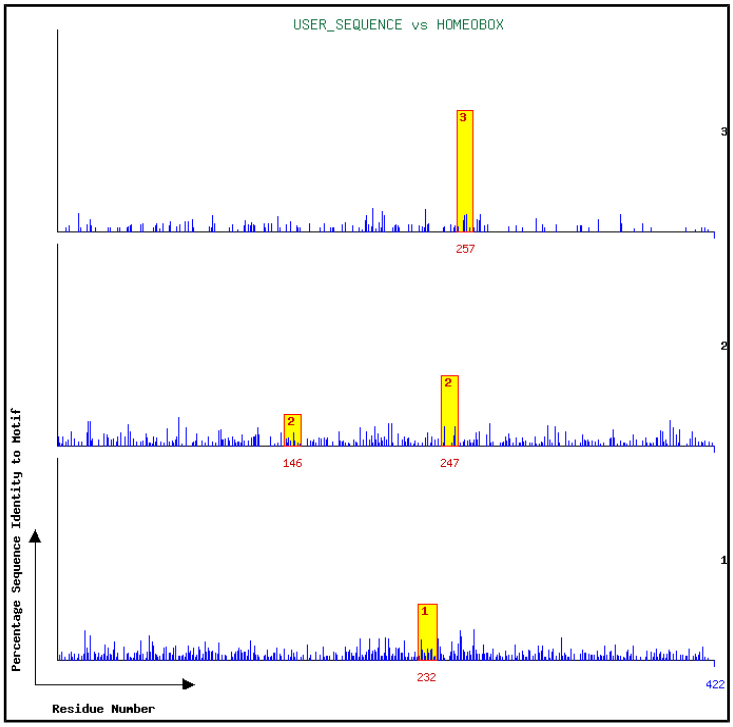
Each motif by itself might not be significant. Together, in a **fingerprint**, they constitute a confident prediction for a **Paired Box domain**.



Click on the **Graphic** link for the **HTHREPRESSOR** hit. The best (highest) of the four **motif 1** hits plus the single **motif 2** hit is the finger print that justifies the **HTHREPRESSOR** prediction.

Move back to the **Ten top scoring fingerprints** table. Notice that, whilst there is a prediction for a **HOMEOBOX**, it is an incomplete prediction. Only two of the required motifs were detected and so no prediction of a **HOMEOBOX** would have been made automatically by **fingerPRINTscan**. This explains why there is no **PRINTS** prediction for a **HOMEOBOX** in the **Uniprot Feature Table** for **PAX6_HUMAN**.

However, if you click on the **Graphics** link for the " 2 out of 3" motif hit for **homeobox**, you will see that **fingerPRINTscan** only missed the **HOMEOBOX** by a whisker!



From the Top ten scoring fingerprints table, you can see that **fingerPRINTscan** considers the first motif to be missing ("**.II**"). But I can see a fairly healthy motif 1 in the graphic. I think I would be inclined to give the **HOMEOBOX** the benefit of the doubt, would you not? Programs can be so very picky!!! Its a hit

Model Answers to Questions in the Instructions Text.

Notes:

For the most part, these "**Model Answers**" just provide the reactions/solutions I hoped you would work out for yourselves. However, sometime I have tried to offer a bit moer back ground and material for thought? Occasionally, I have rambled off into some rather self indulgent investigations that even I would not want to try and justify as pertenent to the objective of these exercises. I like to keep these meanders, as they help and entertain me, but I wish to warn you to only take regard of them if you are feeling particularly strong and have time to burn. Certainly not a good idea to indulge here during a time constrained course event!

Where things have got extreme, I am going to make two versions of the answer. One starting:

Summary:

Which has the answer with only a reasonably digestible volume of deep thought. Read this one.

The other will start:

Full Answer:

Beware of entering here! I do not hold back. Nothing complicated, but it will be long and full of pedantry.

This makes the Model answers section very big. **BUT**, it is not intended for printing or for reading serially, so I submit, being long and wordy does not matter. Feel free to disagree.

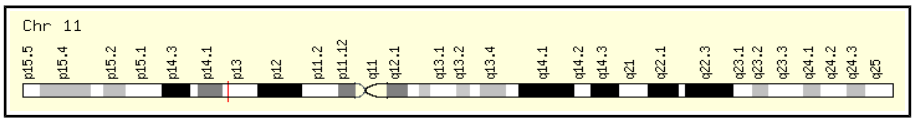
From your investigations using **GeneCards**:

What do you conclude to be the gene most relevant to **aniridia**?

> The gene most relevant to the keyword **aniridia** is **PAX6**. I imagine, in "real life", one might take an interest (with good reason, as you will see) in some of the other significantly scoring genes. However, this is but an exercise, in which the complexities of reality can be shovelled to one side whenever it suits. So **PAX6** it is then!

Properties suggested by **GeneCards** for the gene of interest were:

Cytogenetic location.



The Cytogenetic location of **PAX6** is shown as **11p13**

Three sources are quoted for this information, **Ensembl, Entrez Gene** and **HGNC**. Happily, and expectedly for such a basic prediction, they all agree.

Number of **UniProt** isoforms.

Protein details for PAX6 Gene (UniProtKB/Swiss-Prot)

Protein Symbol: P26367-PAX6_HUMAN *Recommended name:* Paired box protein Pax-6
Protein Accession: P26367 *Secondary Accessions:* Q6N006 Q99413

Protein attributes for PAX6 Gene

Size: 422 amino acids
Molecular mass: 46683 Da
Quaternary structure:
Interacts with MAF and MAFB. Interacts with TRIM11; this interaction leads to ubiquitination and proteasomal degradation, as well as inhibition of transactivation, possibly in part by preventing PAX6 binding to consensus DNA sequences.

Three dimensional structures from OCA and Proteopedia for PAX6 Gene

2CUE (3D) 6PAX (3D)

Alternative splice isoforms for PAX6 Gene

UniProtKB/Swiss-Prot: P26367-1 P26367-2 P26367-3

3 isoforms are suggested by the **UniProtKB** protein sequence database.

How many transcripts are predicted by matches to mRNAs in **REFSEQ**?

Summary:

(24) REFSEQ mRNAs : NM_000280.4 NM_001127612.1 NM_001258462.1 NM_001258463.1 NM_001258464.1 NM_001258465.1 NM_001310158.1 NM_001310159.1 NM_001310160.1 NM_001310161.1 NM_001604.5 XM_005252954.3 XM_005252955.3 XM_005252956.3 XM_005252958.3 XM_006718246.2 XM_011520146.1 XM_011520147.1 XM_011520148.1 XM_011520149.1 XM_011520150.1 XM_011520151.1 XM_011520152.1 XM_011520153.1 See Less «

There are **24** matches of the **PAX6** gene with **mRNA**s in **REFSEQ** implying **24** alternative transcripts.

Full Answer:

REFSEQ proteins: NP_000271.1 NP_001121084.1 NP_001245391.1 NP_001245392.1 NP_001245393.1 NP_001245394.1 NP_001297087.1 NP_001297088.1 NP_001297089.1 NP_001297090.1 NP_001595.2

In the **Protein Section**, only **11 REFSEQ** proteins are reported? But every protein coding **RefSeq** transcript/mRNA has an associated (not necessarily unique) **RefSeq** protein? So **11 RefSeq** proteins implies only **11** alternative **RefSeq** transcripts, which contradicts the assertion of **24** above?

Until fairly recently, the two reports **did** agree because **GeneCards** only reported the **RefSeq** mRNAs whose accession codes commenced with **NM_** (of which there are exactly **11**). These are slightly better evidenced than the remaining **13** whose accession codes commenced with **XM_**. The **NM_** transcripts correspond **1 to 1** with the (reported) **RefSeq** proteins whose accession codes commence **NP_**. The **XM_** transcripts correspond **1 to 1** with (unreported) **RefSeq** proteins whose accession codes commence **XP_**.

Now **GeneCards** have decided to report the less well evidenced mRNAs (whose accession codes begin **XM_**) but still not to report their corresponding **RefSeq** proteins (which do exist and whose accession codes begin **XP_**). Hence the seemingly illogical discrepancy between the two reports.

It maybe **GeneCards** has a good reason to only report **NP_** proteins? Or, it could be a mistake. I go for the mistake theory. I have asked and received the response below, which supports the mistake theory I suggest.

NP_ proteins and **XP_** proteins comes from different sections of **RefSeq**. This is explained in the manual as illustrated. For now, consider just the **11 NP_** proteins reported.

Definitions:

- **Model RefSeq**: RNA and protein products that are generated by the eukaryotic genome annotation pipeline. These records use accession prefixes XM_, XR_, and XP_.
- **Known RefSeq**: RNA and protein products that are mainly derived from GenBank cDNA and EST data and are supported by the RefSeq eukaryotic curation group. These records use accession prefixes NM_, NR_, and NP_.

As noted above, **11 RefSeq** proteins does not imply **11** distinct protein sequences. **RefSeq** translates the coding regions of all **11 RefSeq mRNA** sequences and records the results, even if most are identical. In this case, there really should be at most **3** distinct protein sequence possibilities, given that **UniprotKB** predicts just **3** isoforms.

To illustrate the veracity of this claim, I have retrieved all **11 RefSeq PAX6** protein from the databases at the **NCBI** and multiply aligned them with a program called **COBALT** (again at the **NCBI**). I did this using strategies very similar to those you will investigate in the next few pages of these exercises.

The resultant alignment (see below) used to only include **7** of the top **8** proteins. When this was true, I was able to claim that you should be able to see that just **2** distinct proteins are evident (a 14 amino acid **Insertion**/**Deletion** near the start of the protein being the only difference). Why there were only **2**, rather than the promised **3** distinct sequences will become clear shortly.

Now, as is so tediously often the case, **4** more proteins have emerged, according to **RefSeq**. **3** of these dare to upset the poetry of my interpretations! I retire in petulance. **Uniprot** actually claims there are **2** proteins of known sequence and **1** of unknown sequence. The current alignment suggests there are **4** isoforms, all of known sequence? Well, it is acceptable to disagree, these conclusions are all only predictions after all.

So, I hear you ask, why pursue this sort of issue to such depth? It is not really important unless you are are really researching **PAX6** and/or **aniridia**. ***TRUE***! but the more general objective is to expose the fact that sometimes the discrepancies/anomalies you will come across are not due to alternative biological interpretations but to the way the various information resources are administered.

Plus, of course, I am an incorrigible pedant and cannot leave alone the minutiae, however minute they may be.

```
NP_001595       1   MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQthadakvqvldnqnVSNGCVSKILGRYYETGSI  80
NP_001297087    1   MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQthadakvqvldnqnVSNGCVSKILGRYYETGSI  80
NP_001245391    1   MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQthadakvqvldnqnVSNGCVSKILGRYYETGSI  80
NP_001245392    1   MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQthadakvqvldnqnVSNGCVSKILGRYYETGSI  80
NP_000271       1   MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ--------------VSNGCVSKILGRYYETGSI  66
NP_001121084    1   MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ--------------VSNGCVSKILGRYYETGSI  66
NP_001245393    1   MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ--------------VSNGCVSKILGRYYETGSI  66
NP_001245394    1   MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ--------------VSNGCVSKILGRYYETGSI  66
NP_001297088    1   MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ--------------VSNGCVSKILGRYYETGSI  66
NP_001297089        ------------------------------------------------------------------------------
NP_001297090        ------------------------------------------------------------------------------

NP_001595      81   RPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGADGMYDKL  160
NP_001297087   81   RPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGADGMYDKL  160
NP_001245391   81   RPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGADGMYDKL  160
NP_001245392   81   RPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGADGMYDKL  160
NP_000271      67   RPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGADGMYDKL  146
NP_001121084   67   RPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGADGMYDKL  146
NP_001245393   67   RPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGADGMYDKL  146
NP_001245394   67   RPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGADGMYDKL  146
NP_001297088   67   RPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGADGMYDKL  146
NP_001297089   1    ----------------------------------------------------------------------MGADGMYDKL  10
NP_001297090   1    ----------------------------------------------------------------------MGADGMYDKL  10

NP_001595      161  RMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALE  240
NP_001297087   161  RMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALE  240
NP_001245391   161  RMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALE  240
NP_001245392   161  RMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALE  240
NP_000271      147  RMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALE  226
NP_001121084   147  RMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALE  226
NP_001245393   147  RMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALE  226
NP_001245394   147  RMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALE  226
NP_001297088   147  RMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALE  226
NP_001297089   11   RMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALE  90
NP_001297090   11   RMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALE  90

NP_001595      241  KEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTPV  320
NP_001297087   241  KEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTPV  320
NP_001245391   241  KEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTPV  320
NP_001245392   241  KEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTPV  320
NP_000271      227  KEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTPV  306
NP_001121084   227  KEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTPV  306
NP_001245393   227  KEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTPV  306
NP_001245394   227  KEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTPV  306
NP_001297088   227  KEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTPV  306
NP_001297089   91   KEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTPV  170
NP_001297090   91   KEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTPV  170

NP_001595      321  SSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQTSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMNSQPM  400
NP_001297087   321  SSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQTSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMNSQPM  400
NP_001245391   321  SSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQTSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMNSQPM  400
NP_001245392   321  SSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQTSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMNSQPM  400
NP_000271      307  SSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQTSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMNSQPM  386
NP_001121084   307  SSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQTSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMNSQPM  386
NP_001245393   307  SSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQTSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMNSQPM  386
NP_001245394   307  SSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQTSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMNSQPM  386
NP_001297088   307  SSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQ------------------VSAAGGGLHNPGPREVRSGSGPA  367
NP_001297089   171  SSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQTSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMNSQPM  250
NP_001297090   171  SSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQTSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMNSQPM  250

NP_001595      401  GTSGTTSTGLISPGVSVPVQVPGSEPDMSQYWPRLQ--  436
NP_001297087   401  GTSGTTSTGLISPGVSVPVQVPGSEPDMSQYWPRLQ--  436
NP_001245391   401  GTSGTTSTGLISPGVSVPVQVPGSEPDMSQYWPRLQ--  436
NP_001245392   401  GTSGTTSTGLISPGVSVPVQVPGSEPDMSQYWPRLQ--  436
NP_000271      387  GTSGTTSTGLISPGVSVPVQVPGSEPDMSQYWPRLQ--  422
NP_001121084   387  GTSGTTSTGLISPGVSVPVQVPGSEPDMSQYWPRLQ--  422
NP_001245393   387  GTSGTTSTGLISPGVSVPVQVPGSEPDMSQYWPRLQ--  422
NP_001245394   387  GTSGTTSTGLISPGVSVPVQVPGSEPDMSQYWPRLQ--  422
NP_001297088   368  DLIGCVCTLESFSHYSDWLD----QSSRRQSIPSLLsd  401
NP_001297089   251  GTSGTTSTGLISPGVSVPVQVPGSEPDMSQYWPRLQ--  286
NP_001297090   251  GTSGTTSTGLISPGVSVPVQVPGSEPDMSQYWPRLQ--  286
```

2015.05.23: I have now received a reply from the **GeneCard** folk concerning the discrepancy between the number of proteins and the number of transcripts. Clearly not a vital issue, except that this sort of thing should make the user wary when using these sorts of resources. As is universally true of all things human construction, they fall sadly short of perfect. I think … they give me a slightly cloudy admission of "a bug" that will be fixed soon. Make you own mind up, but only if you feel so inclined, mostly for my benefit, I record their reply here:

" … **GeneCards** presents **RefSeq** based transcripts and proteins. In addition to transcripts and proteins, **RefSeq** has also predicted transcripts and predicted proteins ('**NM**' is for transcripts, while '**XM**' is for predicted transcripts. Similarly '**NP**' is for proteins while '**XP**' is for predicted proteins). The number of not predicted proteins matches the not predicted transcripts, same goes for the predicted items. In **GeneCards** current version we display the predicted transcripts, and do not display the predicted proteins, which causes the discrepancy you mention. ***This issue will be fixed in a future version***."

2015.07.08: Now a new version of **GeneCards**. No change as yet.

2015.08.31: Yet another new version of **GeneCards**. Still no change. Time for another moan?

2016.03.03: Further update to **Genecards**. No change. I give up.

How many transcripts are predicted by the Alternative Splicing Database (ASD)?

Well, it rather looks as if the **ASD** has come to the end, although its final release is still available from the **EBI**.

GeneCards claim they will be revising their policy to refer to **ASD** sometime in the future. They say:

> "We currently do not have any links to **ASD**, as their data has been frozen and their site taken down. We plan to upgrade this subsection."

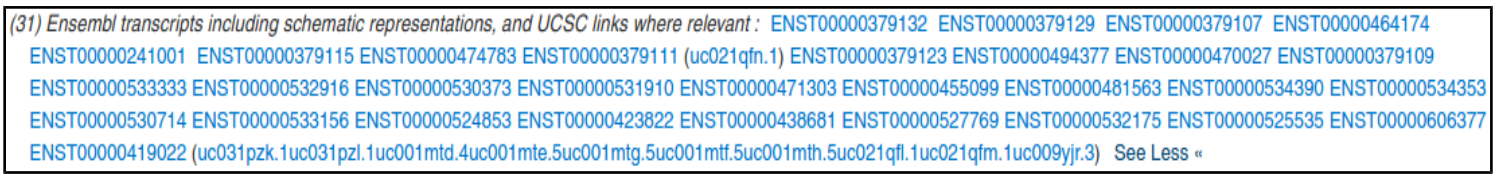
They also offer a brief explanation of how to interpret their graphic.

Genecards computes its graphic using information from both **ASD** and from a Korean resource (still active) called **Ecgene** (Genome Annotation for Alternative Splicing). So maybe this item will not entirely disappear?

As far as this discussion is concerned, **ASD** represents a resource that should be inclined to be quite generous about accepting the credibility of possible splice variants. In contrast, before they started to admit less well evidence transcripts at least, one might expect **RefSeq** to be more fussy and therefore to predict less possibilities? Yes … quite right … if one includes the **RefSeq PREDICTED** mRNAs, this is no longer true.

Ignoring the dismal background, The raw observation remains that according to **ASD** there are **21** splice patterns for the **PAX6** gene.

How many transcripts are predicted by **Ensembl**?



The **Ensembl** genome database predicts **31** transcripts for **PAX6**. The winner! True, but **Ensembl**, uniquely amongst the databases we have considered here, does not restrict itself to transcripts that are protein coding. Now that is simply cheating!

How would you rationalize the discrepancies?

The conflicting numbers of transcripts reported from assorted sources by **GeneCards** are all only predictions based on different evidence and with variant certainty criteria. The "real" number of transcripts is not known with certainty. Some single sources will predict different numbers of transcripts dependant upon credibility levels and evidence sources that can be set by the user. **Ensembl** is such a source, as you will see later.

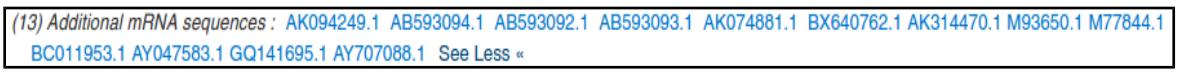
It is also important to realise that the motivations of the various prediction source varies. The **A**lternative **S**plicing **D**atabase (**ASD**), for example set out to report transcripts that could occur given any reasonable splice variation, **RefSeq** transcripts (in principle at least) require a much more demanding level of evidence. **Ensembl** includes transcripts that are not protein coding, as you will soon see.

How many times does the term **PAX6** occur in this entry's annotation?

The term **PAX6** does not occur at all in the **Genbank** entry for **BX640762.1**[136]. Clearly, this will affect the way text searches involving the key **PAX6** work, as will be seen later.

The annotation of the entries of the early databases (**EMBL**, **GenBank**, **DDBJ**) was left to the submitter and so, rather haphazard. It would seem reasonable that this entry should mention the formal name of the gene it represents. We will touch on what is being done to improve this situation later..

Why might the number of **Additional mRNA** matches not match the number of **PAX6** transcripts?



13 Additional mRNA sequences are recorded.

Genbank contains only entries that have been directly sequenced. If an **mRNA** was sequenced more than once it could be represented in **Genbank** more than once. If an **mRNA** exists that has never been sequenced directly, it will not be included in **GenBank**. So, there is no expectation that the **mRNA**s of a particular gene in **GenBank** should match any prediction of the number of transcripts there might be.

RefSeq attempts to represent the biological truth by including just one sequence for every predicted transcript. This often requires the inclusion of sequences that were not directly sequenced, but rather compiled from several directly generated sequences. **RefSeq** also attempts to remove redundancy where possible. So **RefSeq** should contain the same number of **mRNA** sequences as it supposes there to be transcripts.

136 As it would not in the corresponding **EMBL** or **DDBJ** entries as the annotation content in all three databases is identical (well nearly, there can be differences as you will discover later). Only the annotation format is supposed to differ between these databases.

What sort of variations are recorded in the **Humsavar** database?

Summary:

Humsavar entries record the presence of mutations in the protein. The position of the mutations are specified as amino acid positions relative to the canonical protein of the gene under investigation. Most of the **Humsavar** entries for **PAX6** appear to be **Disease variants**.

Full Answer:

Humsavar is a simple text file providing an index of "**Human polymorphisms and disease mutations**". It currently (**2015.08.31**) offers around **71,000** entries divided into three classifications.

Disease variants:	**26874**	{Variants have been found in patients and disease-association is reported in literature. However, this classification is not a definitive assessment of variant pathogenicity.}
Polymorphisms:	**38105**	{No disease-association has been reported.}
Unclassified variants:	**6816**	{Variants have been found in patients but disease-association remains unclear.}
Total:	**71795**	

This is in contrast to the **dbSNP** variations which refer to variations in the genomic whose position is specified as a number of base pairs from the start of the chromosome and are, generally, **NOT** disease related as it is not the prime purpose of **dbSNP** to record such variations (see later discussion).

Note that all **SNP**s available from **dbSNP** are also available from **Ensembl**. Why might that be?

Something of a trick question I fear. **Ensembl** regularly copies in all of the **dnSNP** database entries. Therefore, of course, all **dbSNP** entries will also be in **Ensembl**. I included this question as it does indicate an encouraging example of sharing data/information.

Why might it be considered odd that **rs35883677** be included in **dbSNP**?

Summary:

Because this variation is not a **SNP** (**S**ingle **N**ucleotide **P**olymorphism). It is an **InDel** (**In**sertion/**De**letion). It suggests that in the position indicated there is sometimes no base ("**-**") and sometimes there is an **A**. Sensibly, **dbSNP** is not mindlessly exclusive to **SNP**s. Other interesting variants are also recorded.

Sequence variations from dbSNP and Humsavar for PAX6 Gene

Filter: rs358 (Filtered 3 out of results)

SNP ID	Clin	Chr 11 pos	Sequence Context	AA Info	Type	MAF
rs358 21697 5 43	--	31,797,364(+)	ACATT(C/T)TTATC		intron-variant	
rs358 83677 5 43	--	31,794,440(+)	CACAC(-/A)CACAC		intron-variant	
rs358 40358 5 43	--	31,810,042(+)	AGAGC(C/G)CGGGG		intron-variant, utr-variant-5-prime	

Full Answer:

In passing, **rs35883677** is an **InDel**. But it is not labelled as such in the **Type** column? Not sure I understand why not. I have inquired.

2015.07.08: Searching specifically for **indels** seems to be something that might be very useful, but is rather inconsistent and illogical I think. If you filter in this region for "**indel**" at **GeneCards**, it tries to convince you there are only **2 indels** in this region. Nonsense! If one searches for "**-/**" one gets **76** hits which I suggest are **indels**, but still not enough my intuition tells me. No clear answer from **GeneCards** help desk, they say "talk to **NCBI**" so I do. **NCBI** suggest a way to search their copy of the database which gives **216** hits for roughly the same region. Reasonable, but the search term is "**ind del**"? The space in the middle I would never have guessed. To be fair, it is possible to select the search term from a web site link, in which case you do not need to know about the whimsical space, which is essential if you use the **Advanced Search** in a similar fashion to the way you will experience in the exercise soon. They also offer an alternative way to search for **indels**:

> " … Another route with a more graphical bent is to start in the Gene database record:
>
> http://www.ncbi.nlm.nih.gov/gene/5080
>
> Then go to the **Variation** section (**Table of Contents**), then to See **Variation Viewer** (**GRCh38**). There are a lot of tracks you can remove in the **Configure** menu, and **Filters** on the left; one each for deletion and insertion. When I check **Source = dbSNP** and the **deletion** and **insertion** boxes, I end up with **107** variants."

which seems quite logical to me, but generates yet another different number of hits? The possible explanation offered is:

> "… I think the **dbSNP** page might include merged or removed records …"

which sounds quite likely, if a trifle depressingly unpredictable.

Ensembl call **indels Insertions** or **Deletions** (as far as I can tell), but I can see no way to search for them using **Biomart**. I continue to investigate.

2015.07.09: I suspect the problem with search for **indels** in **GeneCards** is that they have mislabelled the **Molecular consequence dbSNP** field as the **Variation Type** field. This will not work surely? A variation of type **Insertion** could easily have a **molecular consequence** that does not involve the word **Insertion** (or **Deletion**, or **indel**, or "**in del**"). I have asked for elucidation.

2015.08.31: Nothing so far, asked again. Pretty certain my theory is correct.

2016.03.04: Still nothing. OK chaps, I give up. So we just live with these inconsistencies I suppose.

The **Ensembl** Accession Number for the **aniridia** gene.

The **Ensembl** accession code for **PAX6** is: **ENSG00000007372**

Not a vital piece of information at for now, but you will meet this **Accession Number** several times in the next few pages, so I thought I would get the formal introductions over early.

Note the outrage of leading zeroes before the quite modest number that really identifies the gene. It will be a good idea to remember that there are **7** of them at this stage. It makes my eyes hurt and the rest of me weary to count them every time I type them in.

This **Accession Number** was designed in the days when the deep thinkers were estimating many many more genes than there actually turned out to be. They did not manage to think deeply enough to predict that splice variation would provide such variety as it does. With one gene capable of producing many variant protein forms, significantly fewer genes are required.

The trouble is that once one has designed a field capable of representing a very very big number, it is not trivial/practical to redesign it when one realizes that said field is a good four zeroes over provided. So we live with lots of leading zeroes with nothing much to do. Quite amusing really. In a perverse sort of way.

The number of human **PAX** genes.

According to **GeneCards PAX6** is one of **9** human **paralogues**. This information **Genecards** copies from **Ensembl**. As before, you could follow the **5** superscript links to see the **Ensembl** view of each paralogue.

What **Orthologues** exist in **Mouse** and **Drosophila**?

Summary:

Orthologs for PAX6 Gene (?) See less «

Organism	Taxonomy	Gene	Similarity	Type	Details
cow (Bos Taurus)	Mammalia	PAX6 [35]	97 (n) 99.76 (a)		286857 NM_001040645.1 NP_001035735.1
		PAX6 [36]	100 (a)	OneToOne	15:63356631-63384294
mouse (Mus musculus)	Mammalia	Pax6 [35]	94.34 (n) 99.77 (a)		18508 NM_001244198.1 NP_001231127.1
		Pax6 [16]			105668900
		Pax6 [36]	100 (a)	OneToOne	2:105668900-105697364
chimpanzee (Pan troglodytes)	Mammalia	PAX6 [35]	99.15 (n) 99.49 (a)		737387 XM_003954364.1 XP_003954413.1
		PAX6 [36]	99 (a)	OneToOne	11:31718289-31729470

Ask to **See all** the **orthologues** listed, the default list is rather withered!

In **Mouse**, as in a number of other organisms, the othologous gene has the same name as in **Human** (**PAX6**).

fruit fly (Drosophila melanogaster)	Insecta	ey [37]	93 (a)		
		Poxn [37]	64 (a)		
		sv [37]	73 (a)		
		toy [37]	50 (a)		
		toy [35]	58.48 (n) 65.94 (a)		43833 NM_079899.4 NP_524638.3
		ey [36]	30 (a)	ManyToMany	4:718315-741787
		toy [36]	48 (a)	ManyToMany	4:1010351-1028548

In **Drosophila** there are several **orthologues**. **ey** (**ey**eless) and **toy** (**t**win **o**f **ey**eless) being the most important. The naming of the **Drosophila** genes is a little more whimsical than elsewhere.

Full Answer:

Drosophila is an important model organism for the study of **Aniridia**. You will see later another **Drosophila** gene mentioned as an important **orthologue** of **PAX6** in human. The orthologue not mentioned here is **prd** (**p**ai**red**) and it is indeed a genuine **orthogue**. I am not entirely clear why it does not appear here (despite several explanations from experts). Here is the nearest I got (from the **Flybase** team) to an answer, sort of:

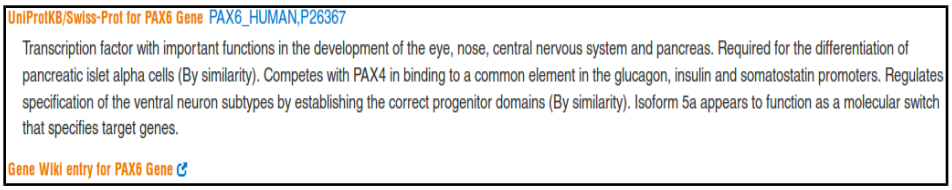
> "As you know, there is often a one-to-many relationship with these things. **prd** is considered (according to **OrthoDB**, which **FlyBase** uses) to be orthologous to **Pax1**, **Pax2**, **Pax3**, **Pax4**, **Pax5**, **Pax6**, **Pax7**, **Pax8**, and **Pax9**. More info here: **http://flybase.org/reports/FBgn0003145.html** in the **Orthologs -> Human orthologs** section.
>
> More info on how we calculate these relationships can be found **here**.
>
> I have no idea what **Ensembl** are doing. They are often slightly out of sync with some aspects of our data.
>
> **ey** and **toy** are related, and are also both orthologous to **Pax1**, **2**, **3**, **4**, **5**, **6**, **7**, **8**, and **9**.
>
> So, on the basis of sequence relationship, **ey**, **toy**, and **prd** are **PaxX** orthologues."

This seems to say that the **FlyBase** chaps think the **Ensembl** fellows have missed something?

One could continue to speculate on the mention here of the **Poxn** gene, or the **sv** gene, but I do not feel sufficiently bold.

What functions are suggested for **PAX6**?

In the **Summary** section:


UniProtKB/Swiss-Prot for PAX6 Gene PAX6_HUMAN,P26367

Transcription factor with important functions in the development of the eye, nose, central nervous system and pancreas. Required for the differentiation of pancreatic islet alpha cells (By similarity). Competes with PAX4 in binding to a common element in the glucagon, insulin and somatostatin promoters. Regulates specification of the ventral neuron subtypes by establishing the correct progenitor domains (By similarity). Isoform 5a appears to function as a molecular switch that specifies target genes.

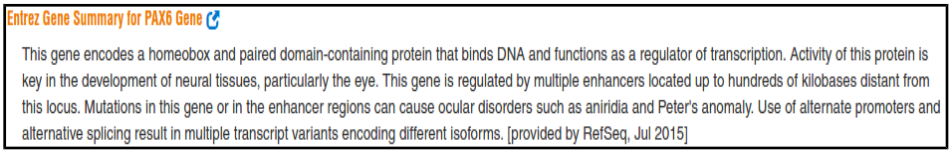
Gene Wiki entry for PAX6 Gene

Genecards, referring to **UniProtKB**, says it all in the first two words. Centrally, **PAX6** is a transcription factor.

More general information is available from the Gene Wiki entry for PAX6 Gene. Here it claims, with confidence, there to be just four **Drosophila** orthologues for **PAX6**.

> "Of the four **Drosophila Pax6** orthologues, it is thought that the **eyeless** (**ey**) and **twin of eyeless** (**toy**) gene products share functional homology with the vertebrate canonical **Pax6** isoform, while the **eyegone** (**eyg**) and twin of **eyegone** (**toe**) gene products share functional homology with the vertebrate **Pax6**(**5a**) isoform. **Eyeless** and **eyegone** were named for their respective mutant phenotypes."

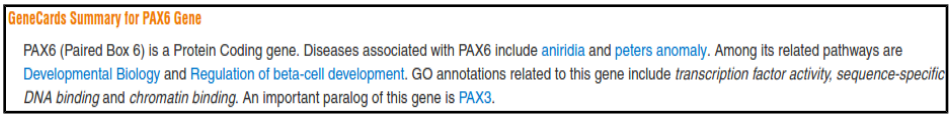
No mention here of **prd**, **sv** or **Poxn**, plus we meet **toe** and **eyg** for the first time? It is time to remind ourselves that the discovery of details about the example gene **PAX6** is ***NOT THE POINT*** here! It is, however, valuable to see just how far from a clear, stable and singular picture is available from the various resources visited. I suspect this will not surprise many of you. After all, we are mostly sifting amongst predictions, as opposed to concrete facts. What we find is laced with opinion and differing interpretation. Life is not straight forward. I am never quite sure whether that is a good thing or a bad thing, but it certainly can be frustrating at times.


Entrez Gene Summary for PAX6 Gene

This gene encodes a homeobox and paired domain-containing protein that binds DNA and functions as a regulator of transcription. Activity of this protein is key in the development of neural tissues, particularly the eye. This gene is regulated by multiple enhancers located up to hundreds of kilobases distant from this locus. Mutations in this gene or in the enhancer regions can cause ocular disorders such as aniridia and Peter's anomaly. Use of alternate promoters and alternative splicing result in multiple transcript variants encoding different isoforms. [provided by RefSeq, Jul 2015]

The information from **Entrez** offers a better description of the expected domains.

It notes that mutations in the gene can result in **Aniridia** and similar disorders. **Entrez** also mentions multiple isoforms, as suggested elsewhere.


GeneCards Summary for PAX6 Gene

PAX6 (Paired Box 6) is a Protein Coding gene. Diseases associated with PAX6 include aniridia and peters anomaly. Among its related pathways are Developmental Biology and Regulation of beta-cell development. GO annotations related to this gene include *transcription factor activity, sequence-specific DNA binding* and *chromatin binding*. An important paralog of this gene is PAX3.

GeneCards has, relatively recently, offered its own version! It seems a bit sparse, but maybe there will be more later.

Uniquely, this descriptions mentions pathways and annotation based on **G**ene **O**ntology (**GO**, discussed later) associations. The **GeneCards Summary** only refers to **one** of the **eight** paralogues? Maybe **PAX4** is more important than the others?

I found I had to read all the descriptions several times before I could see that they were all roughly agreeing. Different views of the same topic? There are many ways of looking at anything after all.

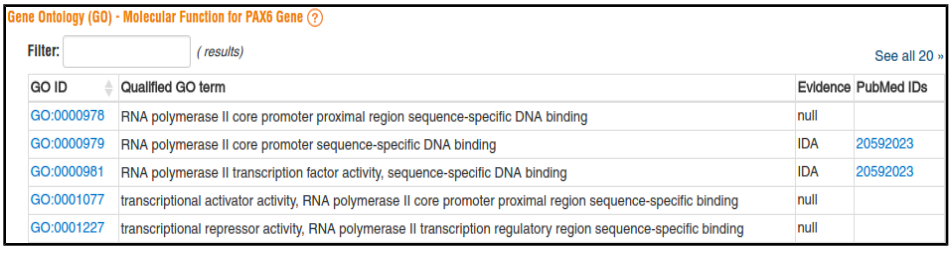
In the **Function** section:


Molecular function for PAX6 Gene

GENATLAS Biochemistry:	paired box (DNA binding) containing protein 6,with homeo domain,expressed in the central nervous system and endocrine pancreas,key regulator of eye development and regulator of glial precursors in the ventral neural tube PAX6
UniProtKB/Swiss-Prot Function:	Transcription factor with important functions in the development of the eye, nose, central nervous system and pancreas. Required for the differentiation of pancreatic islet alpha cells (By similarity). Competes with PAX4 in binding to a common element in the glucagon, insulin and somatostatin promoters. Regulates specification of the ventral neuron subtypes by establishing the correct progenitor domains (By similarity). Isoform 5a appears to function as a molecular switch that specifies target genes. PAX6_HUMAN,P26367

GeneCards refers to **GENATLAS** and **UniProtKB** for information about function.

The contribution from **UniProtKB** is identical to that quoted in the **Summary** section. **GENATLAS**, another human gene database, finds a different way to say similar things.


Gene Ontology (GO) - Molecular Function for PAX6 Gene

Filter: (results) See all 20 »

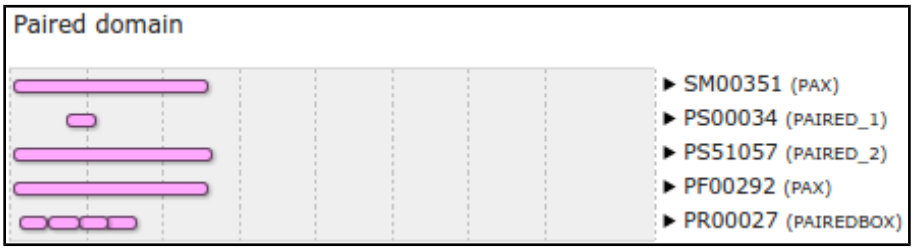
GO ID	Qualified GO term	Evidence	PubMed IDs
GO:0000978	RNA polymerase II core promoter proximal region sequence-specific DNA binding	null	
GO:0000979	RNA polymerase II core promoter sequence-specific DNA binding	IDA	20592023
GO:0000981	RNA polymerase II transcription factor activity, sequence-specific DNA binding	IDA	20592023
GO:0001077	transcriptional activator activity, RNA polymerase II core promoter proximal region sequence-specific binding	null	
GO:0001227	transcriptional repressor activity, RNA polymerase II transcription regulatory region sequence-specific binding	null	

GeneCards also lists all the **G**ene **O**ntology terms associated with the molecular function of **PAX6**. If you look at all **22**, they seem mostly consistent with that which we have discovered thus far.

Why there are two **Prosite** predictions for both the **Homeobox** and the **Paired Box** domains?

All **Prosite** accession codes begin **PS**. For most domains, **Prosite** includes a **H**idden **M**arkov **M**odel (**HMM**) entry representing the entire domain and a simple **Pattern** entry representing a smaller highly conserved signature within the domain. In most instances, a single protein domain will match both a **Prosite HMM** and a **Prosite Pattern** entry giving the initial impression that there might be two domains, a large one with a smaller one contained within it.

In this example, **PS51057** is the **HMM** representing the conservation apparent over the entirety of most **Paired Box** domains. **PS00034** is the **Pattern** representing the extremely high conservation evident over a very small (**17** amino acids) region of most **Paired Box** domains.

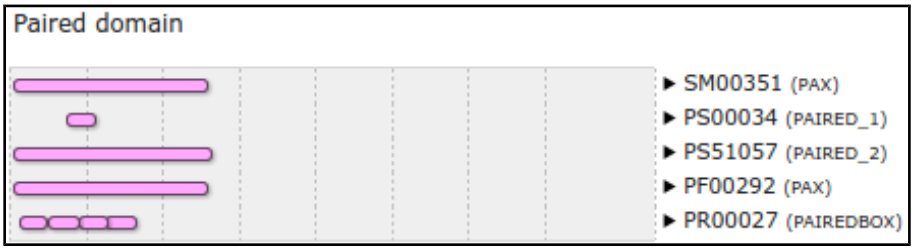


PS50071 is the **HMM** representing the conservation apparent over the entirety of most **Homebox** domains. **PS00027** is the **Pattern** representing the extremely high conservation evident over a significantly smaller region of most **Homebox** domains. The **Pattern** in this case is not considered strong enough to be taken too seriously. It is demoted to a "**conserved site**" and appears in a separate graphic.

Why **Prints** appears to predict four very small **Paired box** domains instead of the single larger domain indicated by all the other predictions?

All **Prints** accession codes begin **PR**. **Prints** entries do not normally represent domains as a single conserved feature. Instead they define a number of features (each of which may not be significant considered in isolation) that must all occur in a given order and within a specified distance of each other. Matches with each element of a **Prints** entry are represented separately, so a single domain detected by a match with a **Prints** entry can erroneously appear to be a match with a series of very small domains.

In this example, the **Prints** entry for a **Paired Box** domain is **PR00027**. For **Prints** to recognize a **Paired Box** domain, **4** relatively small conserved signatures must be found in a specific order. Each of the **4** matches is represented separately although **Prints** is only claiming to have discovered a single **Paired Box** domain.



From your investigations using **OMIM**:

What do you notice about all the variants that are associated with a **dbSNP** entry?

They are all substitutions rather than **in**sertions or **del**etion**s** (**indels**). At first glance, this would seem logical as, if accurately named, the **dbSNP** database really should contain only **S**ingle **N**ucleotide **P**olymorphisms (**SNP**s, i.e. substitutions). However, there are lots of **indels** recorded in the **dbSNP** database, so maybe the observation here is just chance or reflects only that the majority of **dbSNP** database entries are substitutions? Maybe this was not such a good question to ask after all.

Does this surprise you?

I am no longer sure. Maybe it surprises me more than it used to? Maybe many things surprise me more than they used to?

From your investigations using **Entrez**:

What were the features that you found?

Summary:

The first feature was the **Co**ding **S**equence (**CDS**) for a **PAX6** isoform.

The other three features were the coding sequences for three **ELP4** isoforms.

```
complement(39424..>39569)
/gene="ELP4"
/gene_synonym="AN; C11orf19; dJ68P15A.1; hELP4; PAX6NEB;
PAXNEB"
/inference="similar to AA sequence (same
species):RefSeq:NP_001275654.1"
/exception="annotated by transcript or proteomic data"
/note="isoform 2 is encoded by transcript variant 2;
elongator complex protein 4; PAX6 neighbor gene protein;
elongation protein 4 homolog"
/codon_start=3
/product="elongator complex protein 4 isoform 2"
/protein_id=" NP_001275654.1 "
/db_xref="GI:570359562"
/db_xref="GeneID: 26610 "
/db_xref="HGNC: 1171 "
/db_xref="MIM: 606985 "
/translation="MAAVATCGSVAASTGSAVATASKSNVTSFQRRGPRASVTNDSGP
RLVSIAGTRPSVRNGQLLVSTGLPALDQLLGGGLAVGTVLLIEEDKYNIYSPLLFKYF
LAEGIVNGHTLLVASAKEDPANILQELPAPLLDDKCKKEFDEDVYNHKTPESNIKMKI
AWRYQLLPKMEQIGPVSSSRFGHYYDASKRMPQELIEASNWHGFFLPEKISSTLKVEP
CSLTPGYTKLLQFIQNIIYEEGFDGSNPQKKQRNILRIGIQNLGSPLWGDDICCAENG
GNSHSLTKFLYVLRGLLRTSLSACIITMPTHLIQNKAIIARVTTLSDVVVGLESFIGS
ERETNPLYKDYHGLIHIRQIPRLNNLICDESDVKDLAFKLKRKLFTIEWVQDNYLRQE
RNIYPPGFSYLLKQKDSAWGEGSLQHSTFLMSFLAKATAFASRLVRHSEPLKQNGSGR
IRQAAGPRLWHDGRRQEAPGLLGIPP"
```

```
complement(39438..>39569)
/gene="ELP4"
/gene_synonym="AN; C11orf19; dJ68P15A.1; hELP4; PAX6NEB;
PAXNEB"
/inference="similar to AA sequence (same
species):RefSeq:NP_061913.3"
/exception="annotated by transcript or proteomic data"
/note="isoform 1 is encoded by transcript variant 1;
elongator complex protein 4; PAX6 neighbor gene protein;
elongation protein 4 homolog"
/codon_start=1
/product="elongator complex protein 4 isoform 1"
/protein_id=" NP_061913.3 "
/db_xref="GI:91208435"
/db_xref="CCDS: CCDS7875.2 "
/db_xref="GeneID: 26610 "
/db_xref="HGNC: 1171 "
/db_xref="MIM: 606985 "
/translation="MAAVATCGSVAASTGSAVATASKSNVTSFQRRGPRASVTNDSGP
RLVSIAGTRPSVRNGQLLVSTGLPALDQLLGGGLAVGTVLLIEEDKYNIYSPLLFKYF
LAEGIVNGHTLLVASAKEDPANILQELPAPLLDDKCKKEFDEDVYNHKTPESNIKMKI
AWRYQLLPKMEIGPVSSSRFGHYYDASKRMPQELIEASNWHGFFLPEKISSTLKVEPC
SLTPGYTKLLQFIQNIIYEEGFDGSNPQKKQRNILRIGIQNLGSPLWGDDICCAENGG
NSHSLTKFLYVLRGLLRTSLSACIITMPTHLIQNKAIIARVTTLSDVVVGLESFIGSE
RETNPLYKDYHGLIHIRQIPRLNNLICDESDVKDLAFKLKRKLFTIERLHLPPDLSDT
VSRSSKMDLAESAKRLGPGCGMMAGGKKHLDF"
```

```
complement(39533..>39569)
/gene="ELP4"
/gene_synonym="AN; C11orf19; dJ68P15A.1; hELP4; PAX6NEB;
PAXNEB"
/inference="similar to AA sequence (same
species):RefSeq:NP_001275655.1"
/exception="annotated by transcript or proteomic data"
/note="isoform 3 is encoded by transcript variant 3;
elongator complex protein 4; PAX6 neighbor gene protein;
elongation protein 4 homolog"
/codon_start=2
/product="elongator complex protein 4 isoform 3"
/protein_id=" NP_001275655.1 "
/db_xref="GI:570359564"
/db_xref="GeneID: 26610 "
/db_xref="HGNC: 1171 "
/db_xref="MIM: 606985 "
/translation="MAAVATCGSVAASTGSAVATASKSNVTSFQRRGPRASVTNDSGP
RLVSIAGTRPSVRNGQLLVSTGLPALDQLLGGGLAVGTVLLIEEDKYNIYSPLLFKYF
LAEGIVNGHTLLVASAKEDPANILQELPAPLLDDKCKKEFDEDVYNHKTPESNIKMKI
AWRYQLLPKMEIGPVSSSRFGHYYDASKRMPQELIEASNWHGFFLPEKISSTLKVEPC
SLTPGYTKLLQFIQNIIYEEGFDGSNPQKKQRNILRIGIQNLGSPLWGDDICCAENGG
NSHSLTKFLYVLRGLLRTSLSACIITMPTHLIQNKAIIARVTTLSDVVVGLESFIGSE
RETNPLYKDYHGLIHIRQIPRLNNLICDESDVKDLAFKLKRKLFTIEAGVQWHDLGSR
QPRLLGSSNSPASASLVAGITGAHHHTQLIFVFLVEMGFHHVGQAGLELLTSGDSSAS
ASQSAGITGMSYRARPRALYFKENKSKVGARQLLETREEHLSSRLLILTQAERLCMGR
RFFTAFHIFNELPCKGDCICLQTCQTQ"
```

Full Answer:

Note that only the final coding exon of **ELP4** is within this **RefSeq** sequence, which is defined as the genomic region for **PAX6**. This is clear from the length of the **translations** offered. The exon referenced is only long enough to code for just over **40** amino acids which is far short of any of the three isoform sequences offered here.

```
join(16551..16560,20128..20258,21186..21401,22106..22271,
28174..28332,28848..28930,29160..29310,29409..29524,
32102..32252,32943..33028)
/gene="PAX6"
/gene_synonym="AN; AN2; D11S812E; FVH1; MGDA; WAGR"
/note="isoform a is encoded by transcript variant 1;
paired box homeotic gene-6; oculorhombin; aniridia type II
protein"
/codon_start=1
/product="paired box protein Pax-6 isoform a"
/protein_id=" NP_000271.1 "
/db_xref="GI:4505615"
/db_xref="CCDS: CCDS31451.1 "
/db_xref="GeneID: 5080 "
/db_xref="HGNC: HGNC:8620 "
/db_xref="MIM: 607108 "
/translation="MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISR
ILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEI
RDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGSWGTRP
GWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFT
QEQIEALEKEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRR
QASNTPSHIPISSSFSTSVYQPIPQPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPS
FTMANNLPMQPPVPSQTSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMNSQPMGTSGTT
STGLISPGVSVPVQVPGSEPDMSQYWPRLQ"
```

```
gene            5001..38170
                /gene="PAX6"
                /gene_synonym="AN; AN2; D11S812E; MGDA; WAGR"
                /note="paired box 6"
                /db_xref="GeneID:5080"
                /db_xref="HGNC:8620"
                /db_xref="MIM:607108"
```

```
gene            complement(38437..>40170)
                /gene="ELP4"
                /gene_synonym="AN; C11orf19; dJ68P15A.1; hELP4; PAX6NEB;
                PAXNEB"
                /note="elongator acetyltransferase complex subunit 4"
                /db_xref="GeneID:26610"
                /db_xref="HGNC:HGNC:1171"
                /db_xref="MIM:606985"
```

Note also that this final coding exon of **ELP4** (stretching from **39438** to **39569** of this **RefSeq** entry) does **not** overlap the coding region of the **PAX6** gene itself (stretching from **16551** to **33028** of this **RefSeq** entry)[137].

In fact, the two entire genes do not overlap according to the evidence here. The entire **PAX6** gene extends from **5001** to **38170**. The portion of the **ELP4** gene that is included in this entry extends from **40170** (the end) to **38437** (in the opposite direction). This give a gap between the two genes stretching from **38171** to **38436**.

As you will see later, **Ensembl** will confirm the lack of overlap between these two genes graphically as well as their relative positions.

Note that **ELP4** was associated with **aniridia** by **GeneCards**. However, I believe only because of its proximity to **PAX6**.

137 The features here only represent the **CDS** regions of the genes. I wonder why not the entire transcript length?

Why might you have expected more features than there were?

Summary:

GeneCards (referring to **UniProt**) suggested that **PAX6** has three isoforms. This would lead me to expect three features here related to **PAX6**?

Full Answer:

As you will discover later, one of the **PAX6** isoforms could not be represented in the fashion displayed here, but I would still expect there to be two **PAX6** isoform features. The explanation from the **NCBI** is that this sort of **RefSeq** entry is intended to be used as a template against which sequences from an individual can be mapped to seek variations. Only a token **CDS** feature is included to indicate the position of the gene. For such an entry, recording every isoform is not essential.

This sounded convincing to me, Until I began to wonder why there were three **CDS** features for **ELP4** which is not even the gene primarily represented by this entry? Maybe I will ask more questions if and when I ever have the strength. In the meantime, mostly for my information, I record their exact explanation here.

" ... note that **RefSeqGene** defines genomic sequences to be used as reference standards for well-characterized genes. These sequences serve as a stable foundation for reporting mutations, for numbering exons and introns, and for defining the coordinates of other variations. We normally select one **RefSeq** transcript to serve as a reference standard. The goal is not to record all introns and exons of all isoforms, but just to choose one representative to help define the locus. Therefore, most of our **RSG** records have only a single **RefSeq** as reference standard. If an **LSDB** manager or other stakeholder requests that other **RefSeqs** be added as alternate standards, this can easily be done (with the complication that, if a public **LRG** exists, the **RefSeqGene** record is fixed). We receive requests from stakeholders to include **RefSeqs** that represent all known exons, or **RefSeqs** that have become community standards. Often, after creating an **RSG** using our own internal criteria, we receive stakeholder requests to change or add transcripts. Many of these requests come from the **LRG** project regarding transcripts to be included on the **LRG** records.

Generally, **RefSeq** accessions can be added or removed without reversioning, unless a transcript is upgraded or a new one defined that extends beyond the bounds of the **RSG**, or matches a new build of the genome, in which case the **RSG** will be extended and reversioned as needed.

Regarding the chromosomal locus, our standard range is 5 kb upstream from the 5' end and 2 kb downstream from the 3' end of the mRNAs with the greatest extent. For this calculation, we do indeed use all available **RefSeq** (**NM_**) accessions. If the database manager or stakeholder has information on promoters or other upstream or downstream regulatory regions, we can certainly extend the **RefSeqGene** locus to accommodate these.

Regarding mismatches, the goal is to exactly match the current build of the genome, unless there is overwhelming transcript and EST evidence that a mismatch should be retained.

Regarding the confusing subject of exon numbering, exon numbers are currently provided only on **RSG** genomic records based on a subset of available transcript **RefSeqs** for the gene. These are often those selected by locus-specific databases as reference sequence reporting standards. You can find an explanation of how exons are numbered here:

http://www.ncbi.nlm.nih.gov/refseq/rsg/faq/#exon

You will find links to more information on **RefSeqGenes** on the home page for the **RefSeqGene** project:

http://www.ncbi.nlm.nih.gov/refseq/rsg/

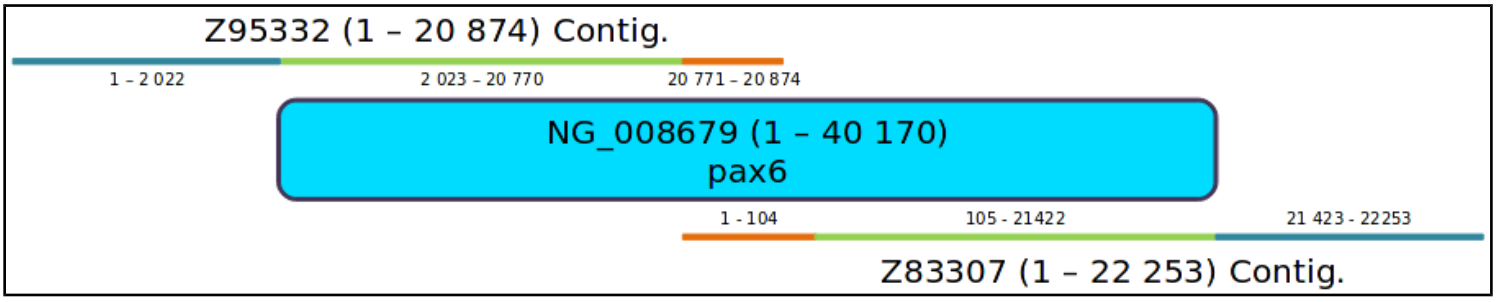
Regarding the **PAX6 RSG** sequence, only difference I see between **NG_008679.1** and the current build of the genome (**GRCh38**) is an extra '**G**' beyond the 3'-UTR of the **PAX6** transcripts (at **NC_000011.10:g.31,819,125**). ... "

Yes, well I think I followed most of that? and that my interpretation is broadly correct? In summary, there are no fixed rules.

How does the alignment you generated match up with the annotation of the original **RefSeq** entry you discovered?

Summary:

The most intuitive way of encapsulating graphically the way these two sequencing clones overlap was donated by **Cecilia Pinto** (**Oeiras**, **2013.12.09-12**). Much better than my rambling attempts, that I keep for sentimental reasons in the "Full Answer". Thank you Cecilia.

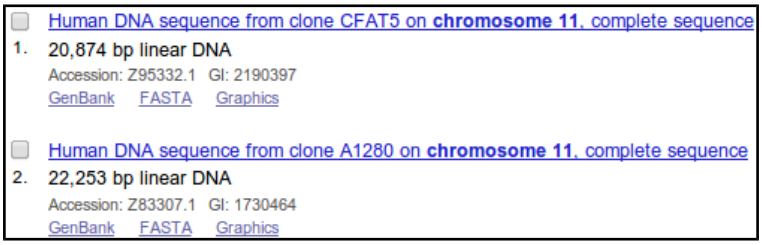


Full Answer:

Do not spend to much time working this one out, the picture above should be more than sufficient. I just needed to see it all balanced ... then I can sleep soundly?

If you do want to read on, I strongly suggest you look at the picture contributed by Cecilia (now promoted to the "**Summary** Answer") first. So simple! I have to admit I cannot follow my own wonderful table at all now ... at least, not without bleeding! Although, it did feel good at the time?

So …




```
Query  20771  GATCCGGAGCGACTTCCGCCTATTTCCAGAAATTAAGCTCAAACTTGACGTGCAGCTAGT  20830
Sbjct  1      GATCCGGAGCGACTTCCGCCTATTTCCAGAAATTAAGCTCAAACTTGACGTGCAGCTAGT  60
Query  20831  TTTATTTTAAAGACAAATGTCAGAGAGGCTCATCATATTTTCCC  20874
Sbjct  61     TTTATTTTAAAGACAAATGTCAGAGAGGCTCATCATATTTTCCC  104
```

The Query sequence is **Z95332** (Length **20,874**)

The Subject sequence is **Z83307** (Length **22,253**)

PRIMARY	REFSEQ_SPAN	PRIMARY_IDENTIFIER	PRIMARY_SPAN	COMP
	1-18852	Z95332.1	2023-20874	
	18853-40170	Z83307.1	105-21422	

NG_008679 Range Start	NG_008679 Range End	NG_008679 Range	Z95332 Range Start	Z95332 Range Start	Z95332 Range	Z83307 Range Start	Z83307 Range End	Z83307 Range
-	-	-	1	2022	2022	-	-	-
1	18748	18748	2023	20770	18748	-	-	-
18749	18852	104	20771	20874 (end)	104	1	104	104
18853	40170 (end)	21319	-	-	-	105	21422	21318
-	-	-	-	-	-	21423	22253 (end)	831
		40171			20874			22253

Legend:

Not used in construction of **RefSeq** entry **NG_008679**

Non-overlapping **GenBank** entry used in construction of **RefSeq** entry **NG_008679**

Overlapping **GenBank** entry used in construction of **RefSeq** entry **NG_008679**

Total entry lengths

The **RefSeq** entry was thus constructed by overlapping the two **Genbank** entries and then manually trimming away the edges to form a biologically meaning region. If I was a bit brighter, I think I might have come to that conclusion without the fuss above? Oh well, one has to use what one has.

I refer you again to the far more intuitive way of encapsulating the same message graphically, donated by **Cecilia Pinto** that is now the "**Summary** Answer" above. Much better! Thank you Cecilia.

From your investigations using **UniProtKB**:

Where have you seen these genes mentioned previously?

PAX6 and **ELP4** were the genes mentioned in the annotation of the **RefSeq** entry **NG_008679**. Part of both of these genes are contained in the clone you are looking at here, **Z83307**.

How is that this is the first occasion that the gene **RCN1** has been apparent?

To avoid mention up to this point, the **RNC1** gene must be entirely in the portion of the sequencing clone **Z95332** that does not include any part of the sequencing clone **Z83307**. That is, between bases **1 and 20770** of **Z95332** (see illustrations from a few answers back).

When we look at the location views in **Ensembl**, you will see clearly how the three genes **PAX6**, **ELP4** and **RNC1** share the same region of **Chromosome 11** and how they overlap the **2** sequencing clones. You will also be able to easily identify the genes and investigate them in precisely the manner we are investigating **PAX6**, should the urge fall upon you.

How is it that we have found any protein sequences at all by looking at the barren annotation of the two clones **Z83307** and **Z95332**?

The only way that the **UniProtKB** search facility can find proteins from the sequencing clone accession numbers you supplied, is to follow information it finds in the annotation of those clones. But … there was no such information! At least, not when you looked at these clones (**Z83307** and **Z95332**) at the **NCBI**.

This search references the same clones stored in **EMBL-Bank**. The annotations of **GenBank** and **EMBL-Bank** are not identical. The main content, established by the **I**nternational **N**ucleotide **S**equence **D**atabase **C**ollaboration (**INSDC**), will be the same in all but format. However, the **NCBI** (**GenBank**) and the **EBI** (**EMBL-Bank**), will occasionally, whimsically add content to ensure that life remains distant from any user aspiration for consistency.

```
ID   Z83307; SV 1; linear; genomic DNA; STD; HUM; 22253 BP.
XX
...
DR   ENA-CON; GL000101.
DR   Ensembl-Gn; ENSG00000007372; homo_sapiens.          { Ensembl Genes
DR   Ensembl-Gn; ENSG00000109911; homo_sapiens.          {
DR   Ensembl-Scaffolds; Z83307.1:1-22253; homo_sapiens.
DR   Ensembl-Tr; ENST00000241001; homo_sapiens.          {
DR   Ensembl-Tr; ENST00000350638; homo_sapiens.          {
DR   Ensembl-Tr; ENST00000379107; homo_sapiens.          {
DR   Ensembl-Tr; ENST00000379109; homo_sapiens.          {
DR   Ensembl-Tr; ENST00000379111; homo_sapiens.          {
DR   Ensembl-Tr; ENST00000379115; homo_sapiens.          {
DR   Ensembl-Tr; ENST00000379123; homo_sapiens.          {
DR   Ensembl-Tr; ENST00000379129; homo_sapiens.          { Ensembl Transcripts
DR   Ensembl-Tr; ENST00000379132; homo_sapiens.          {
DR   Ensembl-Tr; ENST00000379163; homo_sapiens.          {
DR   Ensembl-Tr; ENST00000395934; homo_sapiens.          {
DR   Ensembl-Tr; ENST00000419022; homo_sapiens.          {
DR   Ensembl-Tr; ENST00000455099; homo_sapiens.          {
DR   Ensembl-Tr; ENST00000474374; homo_sapiens.          {
DR   Ensembl-Tr; ENST00000606377; homo_sapiens.          {
...
DR   GOA; B1B1I8.                                        {
DR   GOA; B1B1I9.                                        {
DR   GOA; B1B1J0.                                        { Gene Ontology Links
DR   GOA; E9PKM0.                                        {
DR   GOA; E9PPJ9.                                        {
DR   GOA; G5E9D4.                                        {
DR   HGNC; HGNC:1171; ELP4.                              { Human Genome Organisation (HUGO)
DR   HGNC; HGNC:8620; PAX6.                              { Gene Nomenclature
DR   InterPro; IPR001523; Paired_dom.                    {
DR   InterPro; IPR008728; Elongator_complex_protein_4.   { Interpro Links
DR   InterPro; IPR009057; Homeodomain-like.              {
DR   InterPro; IPR011991; WHTH_DNA-bd_dom.               {
DR   UniProtKB/TrEMBL; B1B1I8; B1B1I8_HUMAN.             {
DR   UniProtKB/TrEMBL; B1B1I9; B1B1I9_HUMAN.             {
DR   UniProtKB/TrEMBL; B1B1J0; B1B1J0_HUMAN.             { UniProtKB/TrEMBL Links
DR   UniProtKB/TrEMBL; E9PKM0; E9PKM0_HUMAN.             {
DR   UniProtKB/TrEMBL; E9PPJ9; E9PPJ9_HUMAN.             {
DR   UniProtKB/TrEMBL; G5E9D4; G5E9D4_HUMAN.             {
...
```

In this instance, the **EMBL-Bank** versions of the clones, **Z83307** and **Z95332**, include links to genes and transcripts determined by **Ensembl** plus links to other databases including **UniprotKB** and **Interpro**. It is these extra **D**atabase Cross-**R**eference (**DR**) lines that are being used to identify the **UniProtKB** entries.

Note that the annotation does not relate the genes and transcripts to which it refers to any particular regions of the genomic DNA? For this we must wait until we investigate **Ensembl**.

In readiness of that heady experience, I note that the **Ensembl** codes for the three genes are:

ENSG00000007372 - which is **PAX6**, of course, and noted in both clones.
ENSG00000109911 - which is **ELP4,** the second gene mentioned in the clone **Z83307**
ENSG00000049449 - which is **RCN1,** the second gene mentioned, but only in the clone **Z95332**

Note that there is, of course, no mention of **ENSG00000049449** in the illustrated annotation of **Z83307**.

Make a note of the first **UniProtKB Accession number** for the **PAX6** protein.

The first (**Primary**) **PAX6** protein accession code is **P26367**

Not a challenging question, just a hint that this code will occur regularly as the pages drift by.

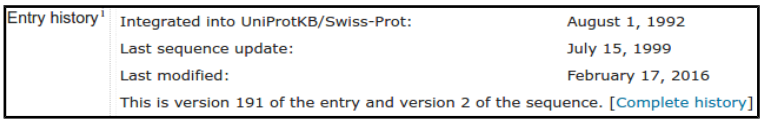
Why do you suppose there is more than one **Accession number** for this protein?

The **PAX6** protein has three accession codes in total. **P26367**, **Q6N006** and **Q99413**.

Before publication, a protein sequence must be allocated a **UniProtKB** accession number. Every **UniProtKB** entry has an accession number. As **UniProtKB** evolves, sufficiently similar entries can be merged to save space. For example, it might be considered sensible to merge two very similar variants of a single protein (**isoforms**) and to record their differences as a **Var**iation **Seq**uence (**VAR_SEQ**) **F**eature **T**able (**FT**) entry. Each merge event creates a redundant accession code. However, accession codes must not be lost. It must always be possible to find a sequence using the accession code allocated when originally published. To make this possible, accession codes made "redundant" by entry merges are retained as **Secondary** accession codes. The accession code thought to be most important is kept as the **Primary** accession code.

In this case, **P26367** is the **Primary** accession code. **Q6N006** and **Q99413** are **Secondary accession codes**. Searching **UniProtKB** for any of these three **accession codes** will work.

There is even a **Complete history** button for those wishing for more detail.



Make a note of the **UniProtKB** Identifier (or entry name).

The **UniProtKB Entry name** of the **PAX6** protein is **PAX6_HUMAN**. Unlike sequences in the **DNA** databases, a sequence in **UniProtKB** can have two names, an **Entry name** and an **Accession Code**. This works as there is no naming conflict with any other protein sequence database. Both the **Entry name** and the **Accession Code(s)** uniquely identify the database entry. Strictly only one or the other is required. The **Entry Names** are intended to be more memorable for human users.

Again, not a question to challenge, just a hint that this **Identifier** will occur regularly it that which follows.

What are the start and end positions of the **Paired** domain?

These next three "questions" are really not attempting to query, but to get you to just be aware of the values I have included in the text. Hopefully, you will see why later. No real need to write them down again, but … just in the name of pedantry:

UniProtKB claims the **Paired** domain of the **PAX6** Human protein extends from residues **4** → **130**.

What are the start and end positions of the **Homeobox** domain?

UniProtKB claims the **Homeobox** domain of the **PAX6** Human protein extends from residues **210** → **269**.

Note the range of the **Proline**, **Serine**, **Threonine** rich region at the end of the protein.

UniProtKB claims the **Proline**, **Serine**, **Threonine** compositionally biased region of the **PAX6** Human protein extends from residues **279** → **422**. This region will be detected by software you use later and it properties given due regard.

Why do you suppose this is the case?

The **S**ingle **N**ucleotide **P**olymorphism **d**ata**b**ase (**dbSNP**) database has entries for a variety of classes of polymorphisms, from a variety of organisms. As ever, **Wikipedia** provides a pleasingly simple description.

There seems to be no hard and fast rules about which variations are to be included, however, reading from various sources, the main focus is on variations for which the **M**inor **A**llele **F**requency (**MAF**) is >= than 1%[138]. **Wikipedia** suggests the purpose of the database to be:

> "physical mapping, **population genetics**, investigations into evolutionary relationships, as well as being able to quickly and easily quantify the amount of variation at a given site of interest. In addition, **dbSNP** guides applied research in **pharmacogenomics** and the association of genetic variation with phenotypic traits."

I read this to mean primarily population studies and similar, where variations that do not cause major problems but are present at significant levels are the major interest.

Specifically concerning disease causing variations, from the **NCBI** pages I find:

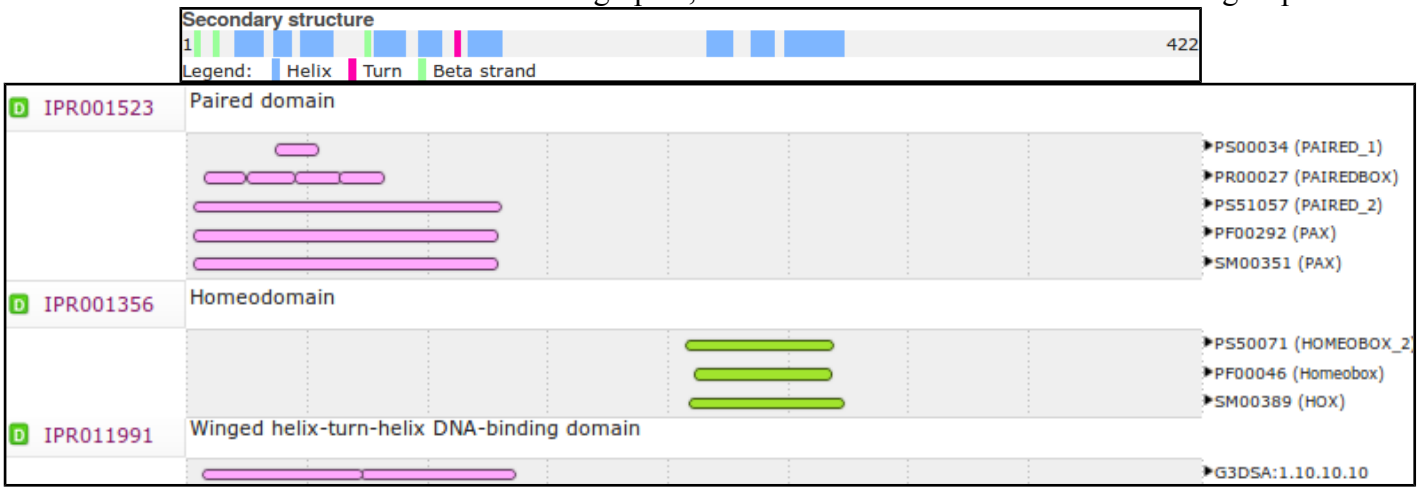
> "Originally, the great majority of data in **dbSNP** was collected and defined as variations simply using sets of co-aligned genomic or DNA sequences. Because this process typically had little to no focus on disease condition, only about **250** records in **dbSNP** were successfully associated with phenotype-causing variations or a clinical outcome in **OMIM**.
>
> Starting in the Spring of 2008, however, **dbSNP** began accepting submissions of **Clinical/LSDB** variations as well as annotations to existing variations (including phenotype) ... Currently there are a total of **1266** records in **dbSNP** that were submitted as **Clinical/LSDB** variations ... and **1134** records submitted as **Clinical/LSDB** variations that also have **OMIM** links … "

I read this to mean there might well be a far greater focus on disease causing variations in the future, but for now, they will be sparse. This is, I submit, what we see evidenced here.

In passing, I regard the patchy evidence provided here for that which I believed in the first place as unsatisfactory. If anyone can suggest anything better, I would be ever so grateful.

Describe the arrangement of Helices within **PAX6**.

From the evidence of the textual table and the graphic, there are **nine** helices in all that occur in groups of **three**.



Aligning the graphical representation of the positions of these helices with the **Interpro** domain prediction graphics similar to that found via **GenCards** earlier, it is clear that the first two of the helical triplets lie in the **PAX** domain and the third is in the **Homeobox** domain.

Note the start position of the middle helix of each set of three.

I have included this information in the text. Again, this is not really a "question" but an invitation to notice something you might confirm by analysis later on. No real need to write anything, but there again - why not:

According to **UniProtKB**, the second helix of the first triplet starts at residue **39**
According to **UniProtKB**, the second helix of the second triplet starts at residue **99**
According to **UniProtKB**, the second helix of the third triplet starts at residue **237**

138 "the **snp135Common table** will only contain **SNP**s with a minor allele frequency >= 1%"

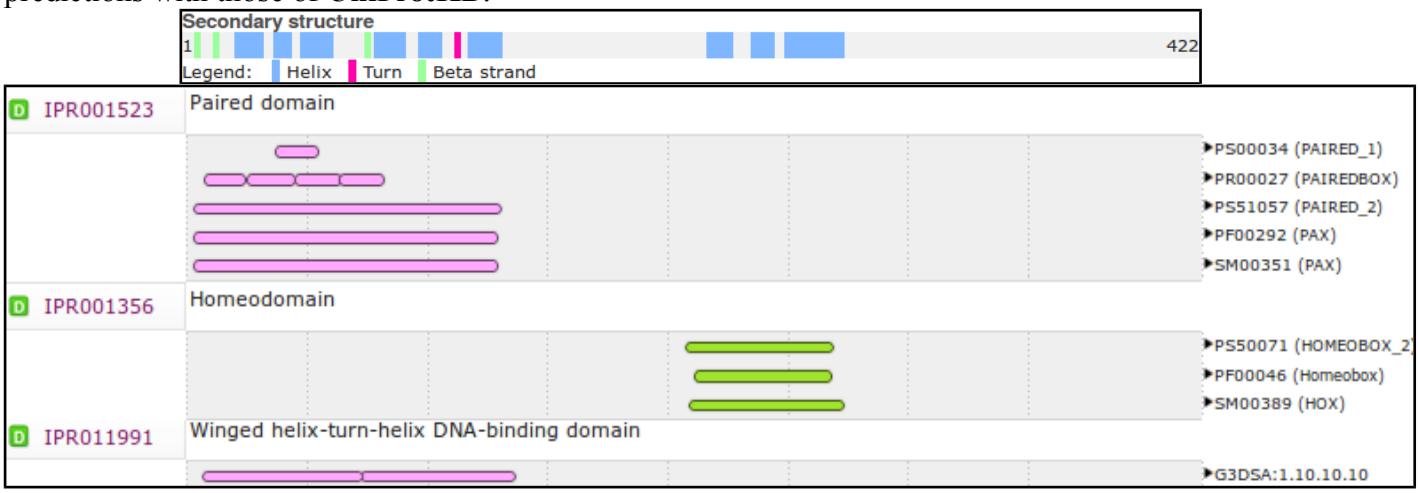
Note the end position of the third helix of each set of three.

According to **UniProtKB**, the first helical triplet ends at residue **63**
According to **UniProtKB**, the second helical triplet ends at residue **133**
According to **UniProtKB**, the third helical triplet ends at residue **275**

There is a point to recording these numbers … honest! The second two helices of each triplet form a special sub-domain that we will read of (and search for with the relevant analytical software) later. Here we see the first indication of where these features may be.

Note the position of the Beta strands relative to the helix groups.

There are **3** Beta strands, according to **UniProtKB**. All are within the **Paired Box** domain. Two before the first helical triplet (residues **6** to **8** and **14** to **16**), and one in before the second helical triplet (residues **77** to **79**). Later you will run programs to predict the secondary structure of this protein. It will be interesting to compare your predictions with those of **UniProtKB**.



Beta strands precede the **2** helical triplets in the **Paired Box** domain but not that in the **Homeobox** domain. As will be seen just a little later, each of the three helical triplets should include a **Helix-turn-helix** motif. A **Helix-turn-helix** preceded by Beta sheet structure(s) can be a **Winged helix-turn-helix**.

I think the absence of Beta Sheet(s) in the **Homeobox** may well be the reason that **GeneCards** (quoting **Interpro**) reported only two **Winged Helix-turn-helix** motifs, rather than three. That is, there is an **HTH** in the **Homeobox**, but not a **Winged HTH**.

Describe the arrangement of Helices within the two major domains of **PAX6**.

I detect some repetition creeping into these questions! Forgive me, they have evolved rather than have been planned as a coherent whole.

You have noted already the **nine** helices that occur in **three** groups of **three**. The highlighted alignment offers another way to confirm that **two** helical triplets coincide with the **Paired** domain and **one** helical triplet with the **Homeobox** domain. Well, you have to allow for a little bit of overflow here and there, but nothing is perfect is it.

Note the extra sequence in **P26367-2** and where it starts.

In the rather stark text version[139] of **PAX6_HUMAN**, the **Seq**uence **Var**iation (**VAR_SEQ**) that define the difference between the canonical form of the **PAX6** protein and **isoform 5a** is recorded as:

```
FT   VAR_SEQ         47       47          Q -> QTHADAKVQVLDNQN
```

which translates literally to:

> "the single amino acid, **Q**, in the **47th** residue position of the canonical protein is replaced by the **15** amino acids, **QTHADAKVQVLDNQN**, in the same position to form **isoform 5a**"

A more biologically orientated way of putting this might be to say that **isoform 5a** is define by an insertion of an extra **14** amino acids, **THADAKVQVLDNQN**, after the **47th** amino acid of the canonical protein. As you will see clearly later, the difference between the two isoforms is determined by the inclusion or exclusion of a single small exon of **42** base pairs during the construction of the mRNA.

This insertion is in a rather critical position, as you will discover later.

How would you rationalise the reference to the mRNA entry **BX640762** here?

As you have already established, **BX640762** is genuinely a **PAX6** sequence and deserves its place in this list. However, you have also already established that this sequence would not be found by an annotation search for the most obvious search term "**PAX6**". So, I conclude this list was constructed either by using an annotation search with different keyword(s) or by using a sequence search producing a list of sequences that are very similar to the **PAX6_HUMAN** sequence, independently of what might be found in their annotations.

139 You can follow the link I have built into this **PDF**, or you can view the **PAX6_HUMAN** entry in a number of formats using the links at the top of the entry page, text | xml | rdf/xml | gff | fasta . Either way, you will see the **PAX6_HUMAN** entry displayed in **EMBL** format. All the lines beginning **FT** describe the **F**eature **T**able for this entry. The line being considered here is amongst them somewhere … Honest!

A quick look at **Prosite**:

What is the **Consensus pattern** for a **Paired domain**?

R-P-C-x(11)-C-V-S

Where in a **Paired domain** should the **Consensus pattern** occur?

The documentation declares the **Consensus pattern** to be derived from the amino acids between **34** and **50** of the **Paired** domain.

> We use the region spanning positions 34 to 50 of the paired domain as a signature pattern. This conserved region spans the DNA-binding HTH located in the N-terminal subdomain. We also developed a profile that covers the entire paired domain, including the PAI and RED subdomains and which allows a more sensitive detection.

UniprotKB claims the **Paired** domain to start at amino acid **4**.

☐ Domain	4 - 130	127	Paired

This leads me to conclude the **Consensus pattern** to be between residues **38** and **54** of the entire protein. Which is exactly correct.

```
          10         20         30         40         50
MQNSHSGVNQ LGGVFVNGRP LPDSTRQKIV ELAHSGARPC DISRILQVSN
          60         70         80         90        100
GCVSKILGRY YETGSIRPRA IGGSKPRVAT PEVVSKIAQY KRECPSIFAW
```

Knowing the rough location of this pattern will be useful later on.

How would you interpret this pattern?

The **pattern** is matched where there is:

An **Argenine**(**R**) - followed by a **Proline**(**P**) - followed by a **Cystine**(**C**)
followed by exactly **11** amino acids of any type **x(11)**
followed by a **Cystine**(**C**) - followed by a **Valine**(**V**) - followed by a **Serine**(**S**)

The syntax for these patterns is slightly richer than implied by this example. Specifically:

- It is possible to use square brackets to indicate possible variation in a position. For example:
  [VIL] would match if a position that was any of a **Valine**(**V**), **Isoleucine**(**I**) or **Leucine**(**L**).
- Curly brackets can be used to indicate “anything except”. For example:
  {P} would match if a position was anything except a **Proline**(**P**).
- Round brackets can be used to indicate ranges. For example:
  [FY](2,4) matches if at least **2**, but no more than **4** positions were either **Phenylalanine**(**F**) or **Tyrosine**(**Y**).
- **L(0,2)** or **L(,2)** would match if there was between **0** and **2 Leucine**(**L**)s.
- **A(2,)** would match if there were **2** or more **Alanine**(**A**)s

The minus signs between the elements of a pattern are optional. Well, they are for all the software I can recall using with these patterns anyway. Some believe the minus signs improve readability.

How effective does the **Technical section** imply this pattern to be?

The **Technical section** claims that there are **58** sequences in **Swiss-Prot** that include a **Paired Box** domain and so should match this pattern. This claim assumes that all the sequences of **Swiss-Prot** are **100%** accurately annotated. A little optimistic, but one must have faith in something.

- Sequences in UniProtKB/Swiss-Prot known to belong to this class: 58

The **Technical section** notes that every one of these **58 Paired boxes** includes the pattern.

- detected by PS00034: 58 (true positives)

The **Technical section** further notes, with some small redundancy, that none of the "known" **Paired Boxes** go undetected either because they do not contain the pattern or because they are partial sequences.

- undetected by PS00034: 0 (false negative or 'partial')

The **Technical section** finally opines that there are **7 Swiss-Prot** sequences that definitely do ***not*** include a **Paired Box** that ***do*** match the **Paired Box** pattern.

- Other sequence(s) in UniProtKB/Swiss-Prot detected by PS00034: 7 false positives.

Not bad for such a simple strategy I would suggest, but not as good as the **PROSITE MATRIX** (i.e. **H**idden **M**arkov **M**odel, **HMM**) that is evaluated just below the pattern. This also finds all the genuine **Paired Boxe**s, but does not pick up any false positives.

PAIRED_2, PS51057; Paired domain profile (MATRIX)

- Sequences in UniProtKB/Swiss-Prot known to belong to this class: 58
  - detected by PS51057: 58 (true positives)
  - undetected by PS51057: 0 (false negative or 'partial')
- Other sequence(s) in UniProtKB/Swiss-Prot detected by PS51057: NONE.

Also, the performance of this pattern is not as immaculate as is suggested here, as will be touched on in another answer still to come.

How well does the secondary structure suggested by **Prosite** match that recorded by **Uniprot**?

Well, the homeobox is predicted by **UniProt** to be from residues **210-269**. **60** amino acids exactly as suggested by the **PROSITE** documentation.

DNA binding[i]	210 – 269	60	Homeobox

Three homeobox helices are predicted by **UniProt** to be at:

Helix[i]	219 – 229	11
Helix[i]	237 – 246	10
Helix[i]	251 – 275	25

Relative to the start of the protein	**Relative to the start of the homeobox**
219 – 229	009 – 019
237 – 246	027 – 036
251 – 275	041 – 065

No turn is predicted by **UniProt**. This is not too surprising as turns are not easy to predict. **PROSITE** suggests there should be a turn between just two helices, forming a **helix-turn-helix** region though which the domain binds **DNA**. **PROSITE** makes no mention of a third helix in front of the **helix-turn-helix** motif.

Diagrammatically:

```
UniProt:  xxxxxxxxHHHHHHHHHHHxxxxxxxHHHHHHHHHHxxxxHHHHHHHHHHHHHHHHHHHHHHHHHH
          |        |         |         |         |         |         |
          1        10        20        30        40        50        60
PROSITE:  xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxHHHHHHHHttHHHHHHHHHxxxxxxxxxx
          |        |         |         |         |         |         |
          1        10        20        30        40        50        60
```

PROSITE is short a helix and **UniProt** is short a turn, but what is left is broadly in the right place? If a trifle over long here and there? What an imperfect world we occupy.

How many **helix-turn-helix** motifs would you expect in the **homeobox** domain?

Just one, according to the **PROSITE** documentation (and all the other evidence discovered thus far) at least.

Where (in amino acid positions) would you expect all the **PAX6 HTH** motifs to be?

Helix[i]	23 – 34	12
Helix[i]	39 – 46	8
Helix[i]	50 – 63	14
Beta strand[i]	77 – 79	3
Helix[i]	81 – 93	13
Helix[i]	99 – 108	10
Turn[i]	114 – 116	3
Helix[i]	120 – 133	14
Helix[i]	219 – 229	11
Helix[i]	237 – 246	10
Helix[i]	251 – 275	25

HTH motifs are functional elements of one sort of **DNA-binding domain**.

In all, there are three **HTH**s in the **PAX6** protein for human. The second and third helix of each of the three helical triplets of the protein form an **HTH** motif of one sort or another. Computing just from the **UniProtKB** secondary structure predictions alone therefore, the three **HTH**s will be around:

```
039 → 063
099 → 133
237 → 275
```

The first two **HTH** motifs are in the **Paired** domain. The **PROSITE** documentation for **Paired** domain alone is not sufficiently specific to predict the position of these **HTH**s with residue number precision.

> The crystal structures of prd and Pax proteins show that the DNA-bound paired domain is bipartite, consisting of an N-terminal subdomain (PAI or NTD) and a C-terminal subdomain (RED or CTD), connected by a linker (see). PAI and RED each form a three-helical fold, with the most C-terminal helices comprising a helix-turn-helix (HTH) motif that binds the DNA major groove. In addition, the PAI subdomain encompasses an N-terminal beta-turn and beta-hairpin, also named 'wing', participating in DNA-binding. The linker can bind into the DNA minor groove. Different Pax proteins and their alternatively spliced isoforms use different (sub)domains for DNA-binding to mediate the specificity of sequence recognition [4,5].

The third **HTH** is in the **homeobox** domain. The **PROSITE** documentation for a **homeobox** domain suggests this **HTH** starts at position **31** of the **homeobox** and continues to position **50**. **UniProtKB** suggests the **homeobox** starts at residue **210**. So, an alternative suggestion of the position of the third **HTH** would be:

A schematic representation of the homeobox domain is shown below. The helix-turn-helix region is shown by the symbols 'H' (for helix), and 't' (for turn).

```
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxHHHHHHHHttHHHHHHHHHxxxxxxxxxx
|       |       |       |       |       |       |
1       10      20      30      40      50      60
```

DNA binding[i]	210 – 269	60	Homeobox

```
240 → 259
```

Close enough? All this so you can check the predictions of the **HTH** positions you might compute later (now a **Supplementary Exercise**), with only partial success, as it is not an easy prediction. Also, there are a number of different types of **HTH**. The software is optimised for just one of the possibilities.

A quick look at **Pfam**:

Extra **Pfam** Notes:

Not associated with any specific question, and not vital to the main exercise. That is, history, mostly for me.

Originally the **Pfam** database had two sections: **Pfam-A** (manual annotated seed alignments, high quality) and **Pfam-B** (automatically generated alignments of much less quality). This observation is mostly interesting as it reflects the organisation of many of the databases considered in these exercises. That is, a quality, manually maintained section plus a machine generated, less reliable, section to try and include "everything".

The **Pfam** people no longer generate their **Pfam-B** section as it has proved to be of little value. They say (**2015.08.10**):

> " … we are not longer producing **Pfam-B**s, largely because most of the clusters not covered by **Pfam** are rarely meaningful potential new domains. Our recommendation now, is to take the piece of sequence of interest and run it using the **HMMER** webserver.
>
> **http://www.ebi.ac.uk/Tools/hmmer/search/phmmer**
>
> This will then produce a set of results that essentially provide you with the same information as a **Pfam** entry."

As in the above quote, they **Pfam** team seem to use **Pfam** and **Pfam-A** to mean the same thing (i.e. the entire **Pfam** database) these days. Just in case of meeting any old references, **Pfam-A** entries (that is all of them these days) have accession codes starting **PF**, the old **Pfam-B** accession codes start with **PB**.

Confirmed (**2016.03.10**):

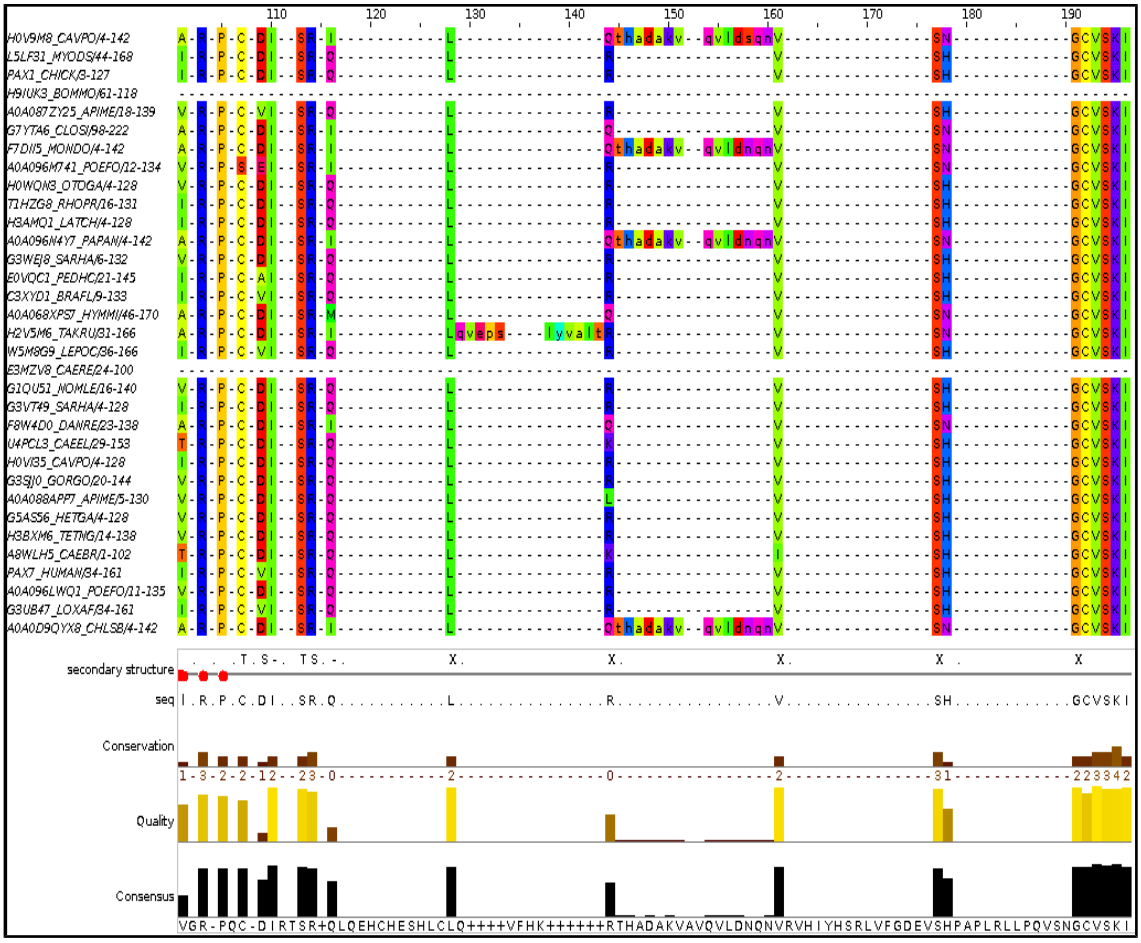
> " … You can assume **Pfam** == **PfamA** (there are still some reference to '**PfamA**' internally but that's historical, due to the reason to distinguish **A/B**)."

Allowing for the distorted numbering of the alignment, how would you interpret the extra **14** or so amino acids that some sequences appear to have around position **100-190**?

These must be the extra **14** amino acids that define the second **PAX6** isoform. The sequences that include the extra amino acids must represent **isoform 5a** proteins.

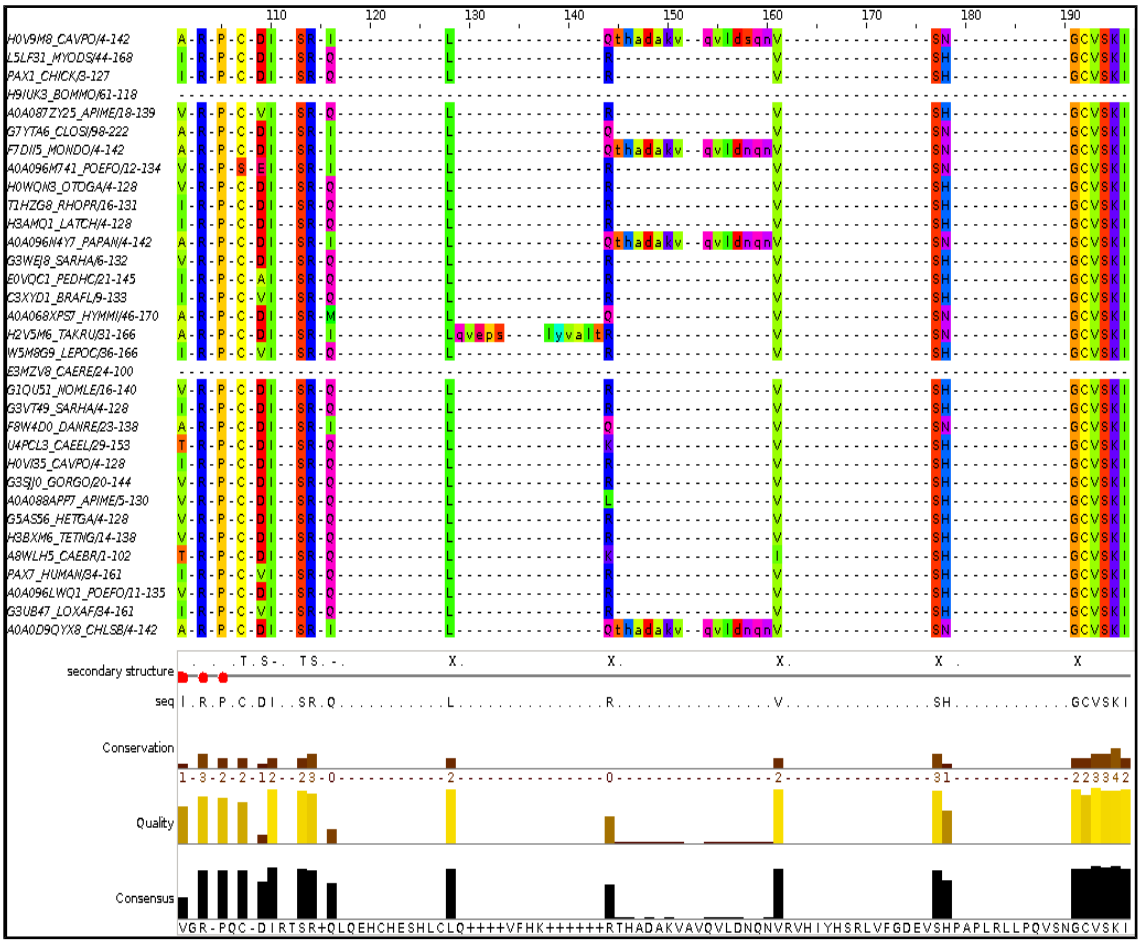
All the residues represented by lower case letters are insertions, relative to the **Seed HMM** profile for a **Pfam Paired** domain.

The worry is that all the insertions should surely be in the same place? The alignment illustrated appears to offer **5** or so alternative insertion positions?



How might you interpret the way that **HMMER3** suggests that some sequences have a similar insertions in a slightly different places to others?

Summary:



At first glance, this looks like a scrappy alignment!

The expectation of a neat **single** insertion point for the extra **14** amino acids of **isoform 5a** in the region illustrated is denied!

Actually, it is not nearly as bad as it would seem. I investigate in the **Full Answer:** section, and conclude that the reason this alignment looks so poor is due mostly to the inclusion of just a few "difficult" sequences. If one was to remove these (quite possibly not scientifically supportable, but … what do I care!) one can make an alignment that is much easier upon the intuition.

As I wish to encourage you ***not*** to go through the **Full Answer:** until you have more time, I make the central point here that messing around trying to understand all there is to know about this particular alignment and **PAX** is absolutely ***NOT*** the point. What I would like you to realise is that **Pfam** alignments (and others) are offered to you in an alignment editor (**jalview** in this case). This means you can do so much more than just look at it in wonder. You can manipulate any alignment in ways similar to those I used below, see the information from different angles and come to more enlightened conclusions. Not so very earth shattering, in this case, but another time? Who knows.

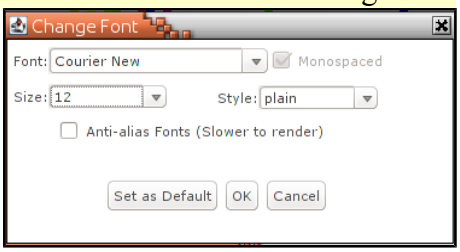
Full Answer:

Practically, the indifferent quality of this alignment really does not matter that much. It is the **HMM profile** computed from the **Seed alignment** that is used to represent a **Pfam Paired** domain and this it does very effectively. The alignment here is just for show. It still seems a shame that it is not more immediately convincing.

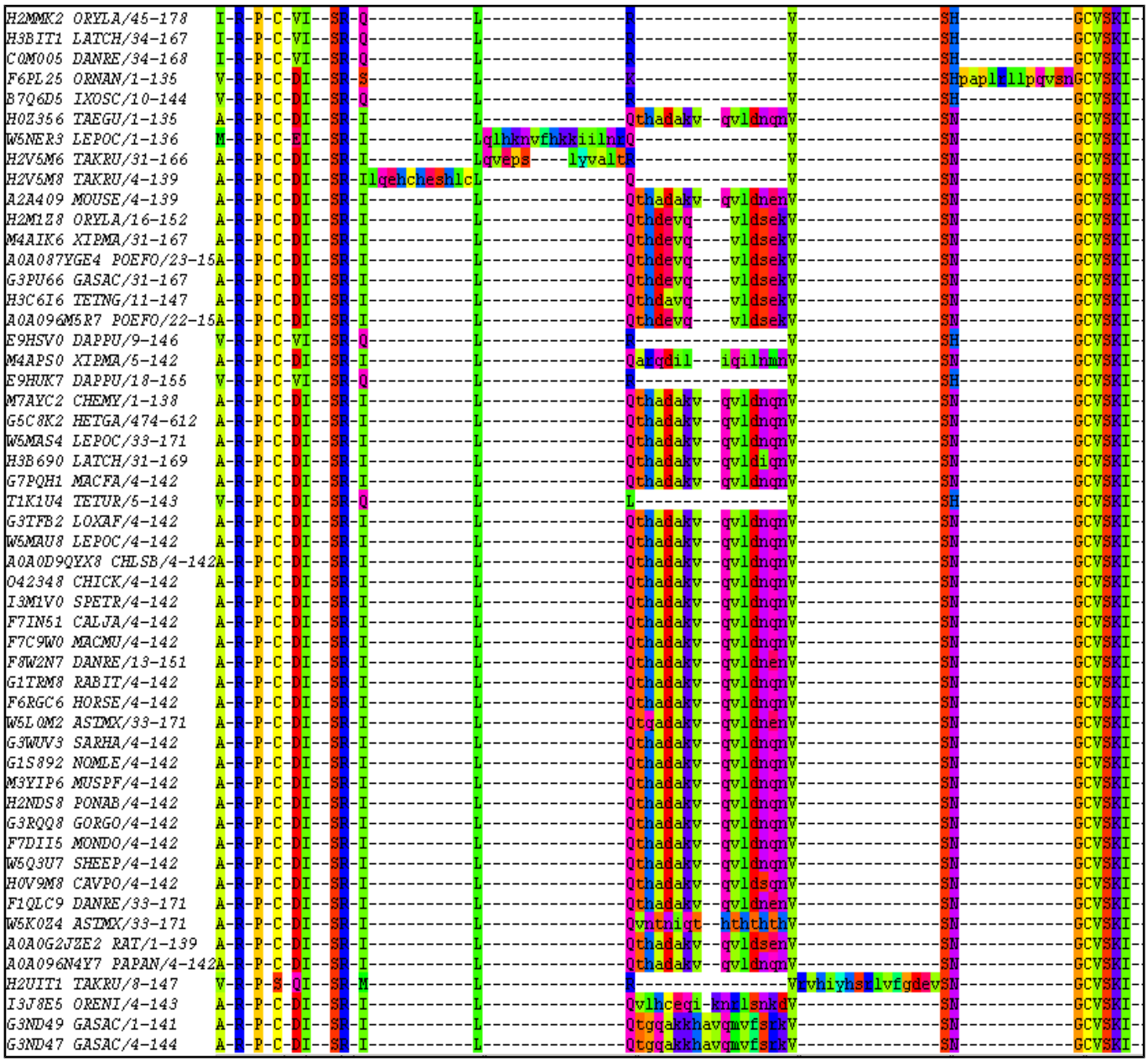
The objective here is to discover just why the alignment for the full **PAX** family so poorly represents what should be a single point of insertion for the **14** amino acids that differentiate the two **PAX** isoforms of known sequence. The messy alignment above is particularly distressing as the equivalent generated previously by **HMMER2 (**the father of **HMMER3**) was really quite convincing. The **Pfam** team explain that **HMMER3** runs much faster (**200** times faster apparently) than **HMMER2** did. But surely this is irrelevant is the quality of the alignment suffers significantly?

The first step must be to try and make the situation a little clearer. I aim to bring all the **isoform 5a** protein sequence together in the alignment, so they can be all viewed at the same time. **Jalview** makes this trivial. In order to produce the next view of the alignment I did the following:

- Ordered the sequences by length. I reason the **isoform 5a** sequences should generally be longer than the **canonical isoform** sequences. **Calculate → Sort → By Length**. Then move to the bottom of the alignment where the longer sequences will be.



- Changed the **Font** to **Courier New** and the **Font Size** to **12**. **Format →** Fill in the form as illustrated. Personal preferences I admit, but I think these choices do make things a little clearer.



Now all the isoform 5a sequences can be seen together, the overall effect is much more encouraging. HMMER3 has aligned the sequences sensibly. It seems to be the inclusion of just a few sequences that, for whatever reason, are not behaving themselves very well, that has made the overall effect so ugly.

OK, so why not put scientific justification to one side and aim purely for a "nice effect"? This I did by simply removing all the sequences that did not fit from the alignment. Click on the identifiers of all the sequences that are problematic, holding the **Ctrl** key down all the time. Then remove those selected. **Edit → Delete**. To gain most of the effect I craved, I had to remove only a few, well under **10**, sequences.

Finally, remove from the alignment all columns that are now purely padding characters ("-") due to the deletions. **Edit → Remove Empty Columns**. The resultant alignment is far nearer the expectations encouraged by that which has been learned about this family thus far.

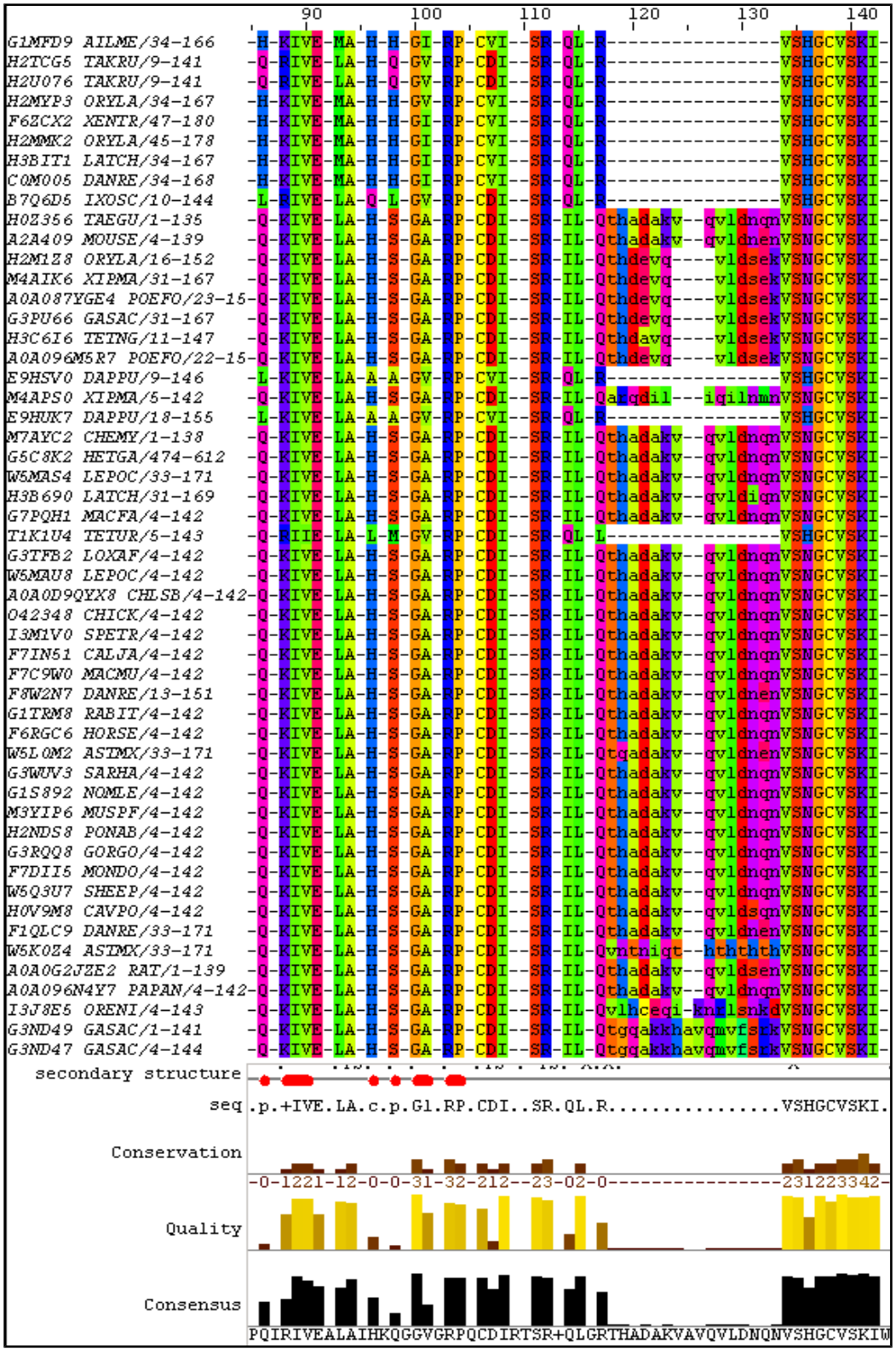
I stop at this point, but take a swift glance at the positions of the highly conserved residues of the **PROSITE** pattern for **Paired** domain.

The **RPC** starts around residue **103** (depending how many sequences were deleted). Placing your mouse over the **Consensus** bar for these residues shows **85%** conservation. Well short of the **100%** claimed by **PROSITE**.

Further, there exists evidence in my alignment that some **PAX** sequences have a single amino acid insertion between the **P** and the **C**!

Further still, investigating the **CVS** residues a little further along (**138-140** in my illustration), the suggestion is around **90%** conservation. Impressive, but still not **100%**.

I am sure these (and other) anomalies could be massaged into a closer approximation to "perfect" with a little effort. In particular, if you look at the top of the alignment (where the suspiciously short sequences reside). Here I see many many candidates for exclusion from the alignment on the grounds of "convenience".



Also note, the **isoform 5a** insertion occurs in the middle of the **PROSITE** pattern for a **Paired** domain. As will be discussed again later, this has to make the pattern **RPCx(11)CVS** very much less than useful at detecting **isoform 5a Paired** box sequences!

So, I set out to investigate why the **HMMER3** initial alignment was so much uglier than the **HMMER2** equivalent. I conclude that either one or both of the following is true:

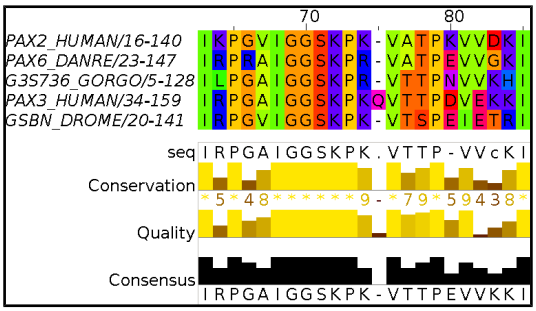
- **HMMER3** is a mite more generous than **HMMER2** with regards to the **PAX** sequences it accepts.

- More distantly related **PAX** sequences have been included in the databases searched by **Pfam** since the far off days of **HMMER2**.

Finally, it is encouraging to reflect that **HMMER3** goes **200** times faster then **HMMER2**, that has to be truly impressive. Never mind the alignment of all the **PAX**s we have been contemplating here, **HMMER3** is exceptionally effective at its main task of matching domain representations to protein sequences.

Note down the position in the alignment where all but one seed sequence has a gap.
Note the consensus character at this point and its most logical interpretation.

In residue position **75** of the seed alignment, **4** of the **5** protein sequences are gapped. Only one sequence, **PAX3_HUMAN**, has an amino acid recorded, a **Q (**Glutamine).

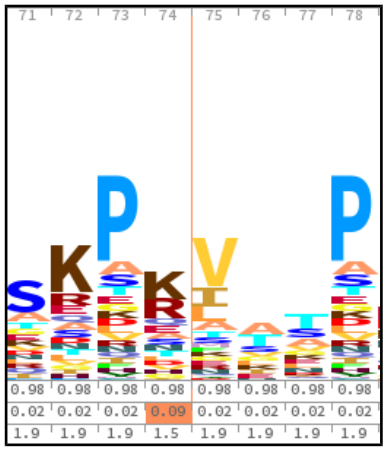


The **Consensus** character at this point is "**-**". **Jalview** has it own way to calculate the **Consensus**. Read the documentation for the official explanation. Informally: for positions where there is no dominant amino acid code, + means "more than one possibility", - means "predominantly a gap".

How is the heavily gapped position of the seed alignment represented in its **HMM Logo**?
How would you interpret the **Logo** in this region?

The heavily gapped position of the seed alignment is position **75**. In this position, **4** of the **5** aligned sequences have been gapped, the remaining sequence has a **Q**.

This position does not appear in the **Logo** (although there is a position **75** … which relates to position **76** of the alignment … which seems a bit silly to me!). This implies that the **HMM** represents the data at position **75** thus:



"Generally not present, but a relatively high chance of an insertion which is most likely to be a **Q**"

The alternative, equivalent, representation would be:

"Generally a **Q**, but a relatively high probability of a deletion"

Had the second alternative been selected, the **Logo** would have shown a healthy **Q** at position **75**.

A thin brownish line is placed in the **Logo** to indicate where position **75** was omitted. The **Logo** is not a precise enough representation to clearly show all the details of the chance of an insertion and whether that the insertion is likely to be a **Q** …. but this will be recorded in the **HMM** itself.

A quick look at **Prints**:

How many links to **PRINTS** would you expect?

Two surely. It should be clear by now that we are investigating a protein with **two** major domains, a **Paired Box** and a **Homeobox**. It is reasonable to expect that all the domain database searches would find both these domains. **PRINTS** shows only the presence of the **Paired Box** domain.

What if anything, do you think is missing?

So, what is missing is evidence of the **Homeobox** domain. **PRINTS** has detected the **Paired Box** but has entirely missed the **Homeobox** domain. All the domain searches work slightly differently and none always pick up everything that they should. This is why the best strategy is not to rely on just one domain search but to always run them all[140]. That one search occasionally misses something is to be expected, for all the searches to fail together should really not happen (well, rarely at least!).

That it is **PRINTS** that fails in this instance should not be taken to mean that **PRINTS** is an inferior domain database. As you will see later on in the exercises, **PRINTS** only fails to detect the **Homeobox** by a whisker. The presence of a **Homeobox** domain is detected, but at a strength below that required for **PRINTS** to record a hit.

140 **Interpro** allows you to do this easily, as hopefully we have discussed and as you will see for yourself later on.

A quick look at the **cluster databases**:

What was the **seed sequence** upon which this cluster was built?

D3DQZ8

What do you imagine a **seed sequence** might be?

The seed sequence is the longest sequence in a cluster.

The UniRef databases are generated in a hierarchical fashion where:

- UniRef100 clusters are generated first using sequences from UniProtKB and UniParc databases
- UniRef90 clusters are generated using UniRef100 seed sequences (longest sequence in a cluster)
- UniRef50 clusters are generated using UniRef90 seed sequences

What is the **Representative sequence** protein for this cluster?

P26367

What do you imagine a **Representative sequence** might be?

The representative sequence is the top sequence in a cluster.

All the proteins in each cluster are ranked to facilitate the selection of a biologically relevant representative for the cluster. The representative sequence (ranked first) is selected based on the following criteria:

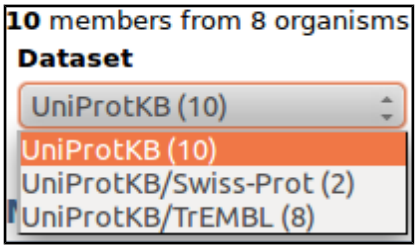
1. quality of annotation: order of preference is a member from UniProtKB/Swiss-Prot then UniProtKB/TrEMBL and last is UniParc
2. meaningful name: members with protein names that do not contain words such as hypothetical, probable etc. are preferred
3. organism: entries from model organisms are preferred
4. sequence length: longest sequence is preferred

How many sequences are from **UniProtKB/Swiss-Prot** and how many from **UniprotKB/TrEMBL** (Look at the **Dataset** pull down Menu)?

UniProtKB(10)

UniProtKB/Swiss-Prot(2)

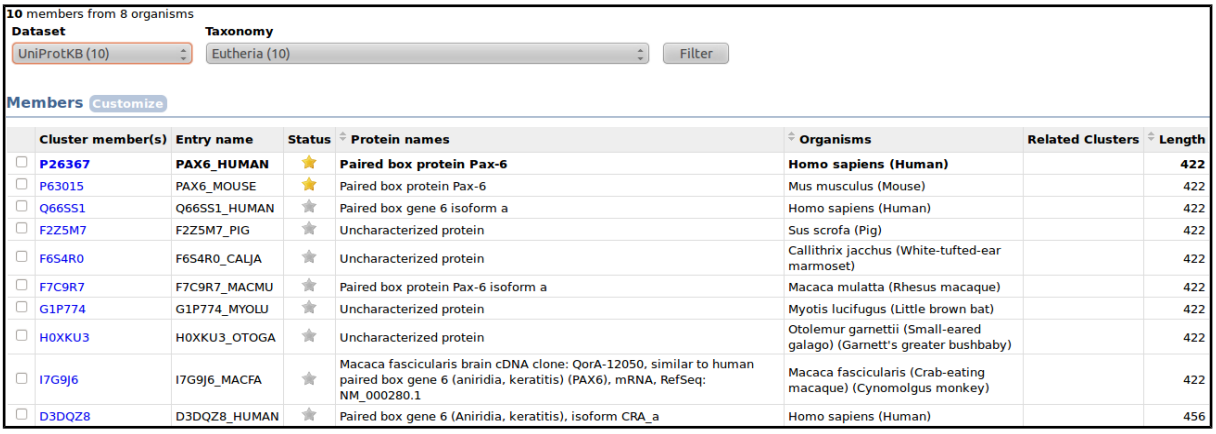
UniProtKB/TrEMBL(8)



Given your last answer, how would you interpret the colours of the stars in the **Status** column?

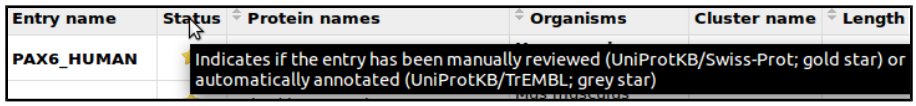
Swiss-Prot

TrEMBL



Cluster member(s)	Entry name	Status	Protein names	Organisms	Related Clusters	Length
P26367	**PAX6_HUMAN**	★	**Paired box protein Pax-6**	**Homo sapiens (Human)**		**422**
P63015	PAX6_MOUSE	★	Paired box protein Pax-6	Mus musculus (Mouse)		422
Q66SS1	Q66SS1_HUMAN	☆	Paired box gene 6 isoform a	Homo sapiens (Human)		422
F2Z5M7	F2Z5M7_PIG	☆	Uncharacterized protein	Sus scrofa (Pig)		422
F6S4R0	F6S4R0_CALJA	☆	Uncharacterized protein	Callithrix jacchus (White-tufted-ear marmoset)		422
F7C9R7	F7C9R7_MACMU	☆	Paired box protein Pax-6 isoform a	Macaca mulatta (Rhesus macaque)		422
G1P774	G1P774_MYOLU	☆	Uncharacterized protein	Myotis lucifugus (Little brown bat)		422
H0XKU3	H0XKU3_OTOGA	☆	Uncharacterized protein	Otolemur garnettii (Small-eared galago) (Garnett's greater bushbaby)		422
I7G9J6	I7G9J6_MACFA	☆	Macaca fascicularis brain cDNA clone: QorA-12050, similar to human paired box gene 6 (aniridia, keratitis) (PAX6), mRNA, RefSeq: NM_000280.1	Macaca fascicularis (Crab-eating macaque) (Cynomolgus monkey)		422
D3DQZ8	D3DQZ8_HUMAN	☆	Paired box gene 6 (Aniridia, keratitis), isoform CRA_a	Homo sapiens (Human)		456

Hover over the column heading **Status** to confirm this conclusion.



Which sequences react to the **Annotation** request, and why?

The **SwissProt** sequence react, the **TrEMBL** ones do not. Only the **SwissProt** entries have quality annotation.

Can you rationalize why one of the sequences is allowed to be different to all the others?

xxxxx

After viewing this **UniRef100** entry, how "non-redundant" would you say was **UniprotKB**?

xxxxx

Why do you suppose it might be useful to have identical sequences in **UniprotKB**?

xxxxx

From your investigations using **Ensembl**:

Is it a gene dense region?

Just in case you had forgotten, the region in question is **P13**.

Is it gene dense? A bit less that average I would say. Not a vital question, in the particular, but it is interesting how gene density seems vary quite substantially. **P12**, for example, is a bit of a desert! In contrast, several parts of **Q13** are very well endowed, particularly with protein coding genes.

What about Variation density?

Summary:

It is the **Variations** column to which I draw your attention here. Uniform density over most of the chromosome is suggested I think? Well, that is barring a couple of small peaks each side of a barren centromere. This was a much more interesting question before **Ensembl** included the entire dbSNP database form the **NCBI**. Variations were relatively sparse before this event and the density varied with greater whimsicality.

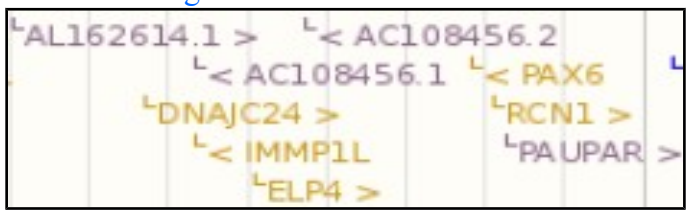
Full Answer:

2015.06.01: This display was broken! The cause was that a massive increase in the size of the **dbSNP** database exceeding the capacity of the display!. The boys (and girls) of **Ensembl** suggested they may have to move to a logarithmic scale in order to cope.

2015.07.18: All fixed. Not sure if the logarithmic scale was implemented, but the picture now looks quite convincing, if bland. The picture in the book is now (fairly) current. There are now so many variations it is difficult to read much into this picture at all? Superficially, the distribution of variations looks even more uniform throughout the chromosome than it did previously. This impression may well be exaggerated by the use of a logarithmic scale? The "peaks" around the centromere are certainly more subdued.

Can you also see the two other genes you might have expected to be in this region?

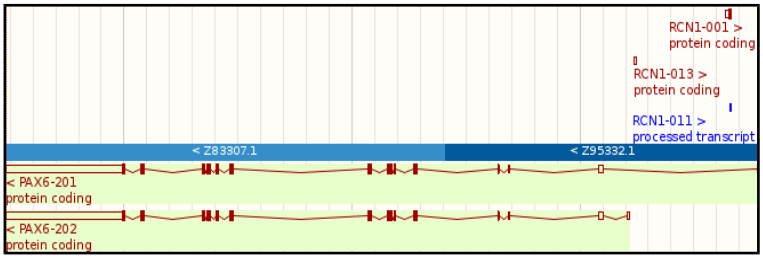
PAX6, **ELP4** and **RCN1** can be clearly seen in positions predicted by the annotations you found earlier[141].



The ">" characters for **ELP4** and **RCN1** both indicate that these genes are on the strand opposite to that where **PAX6** resides. This you will also have expected.

What are the (familiar?) contig numbers containing all of **PAX6**?

The portions of the genomic clones **Z83307** and **Z95332** needed to represent the longest transcript of **PAX6** are clear to see, represented as alternating dark and light blue bars.



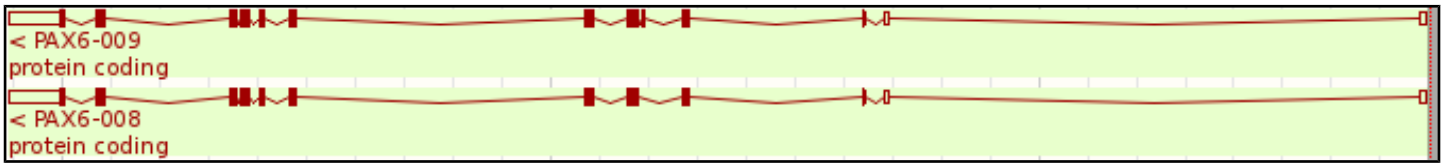
Note the "<" characters in each bar indicating that both clones were reversed and complimented (relative to their representation in **GenBank**/**EMBL-Bank**) in order to be assembled with all the other contig contributions to the entire sequence of **Chromosome 11**. In the **EMBL-Bank** version of both these clones, **PAX6** was suggested to be on the forward strand. In **Ensembl**. **PAX6** is shown below the blue contigs bar, implying that it is on the reverse strand of **Chromosome 11** as a whole.

Note the beginnings of some **RCN1** transcripts on the forward strand. **Ensembl** suggests, mostly between **UTR**s and never involving coincident exons, an overlap between the **5' UTR** of some transcripts of **PAX6** and the **3' UTR** of some transcripts of **RNC1**.

There is no sign of **ELP4** in this view. The part of this gene included in **Z83307** does not overlap with the **PAX6** region and is so does not appear here.

141 You need to allow that the gene names occupy much more space than the genes at this scale. The next picture gives a more useful impression.

Explain the visible differences between the coding exons of transcripts **PAX6-008** and **PAX6-009**?



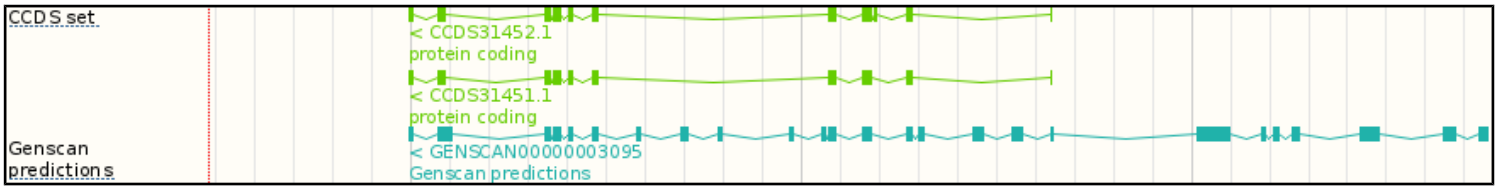
Of course, one cannot be certain that all differences between these transcripts can be observed from a picture.

For complete accuracy, one would need to view the textual representations, however, there is one clear difference observable here. There is an extra coding exon in **PAX6-009**, the **5th**, to be particular (remember you must read the transcript from right to left as **PAX6** is on the reverse strand of the Chromosome).

The inclusion/omission of the **5th** exon, a very small coding exon, must result in alternative protein products (isoforms) for each of the two transcripts. Maybe this small exon is the source of the extra **14** amino acids that define **isoform 5a** of **PAX6**? This should be clearly confirmed/denied when we get to look at the transcripts in more detail. It must be a good bet though? There are only two isoforms with sequence after all.

Or … maybe the answer to the next question will be enough to convince you?

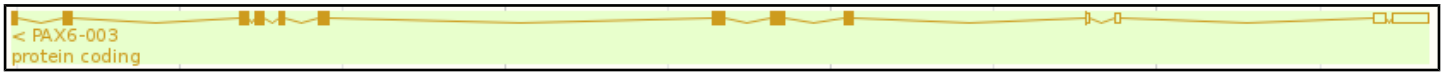
What **PAX6** exon of note has **Genscan** omitted to predict?



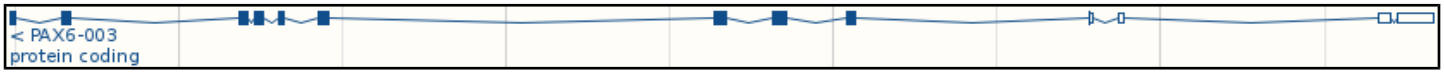
There are two **CCDS** set hits, one for each isoform. The extra coding exon that defines **isoform 5a** can be seen clearly in the upper **CCDS** hit (**3rd** coding exon in from the right, as we are looking at an analysis of the reverse strand of the gene here, that means the **3rd** coding exon from the start of the gene). This extra coding exon is missed by **Genscan**, probably because of its size. **Genscan** does predict all the other coding exons (plus extras) and makes a reasonable shot at the non-coding exons also. You could see **Genscan** in action again in an optional exercise.

Is a **merged Ensembl/Havana** transcript also a **Vega Havana protein coding** transcript?

Well, of course! Here it is being a gold **merged Ensembl/Havana** transcript from the default display:

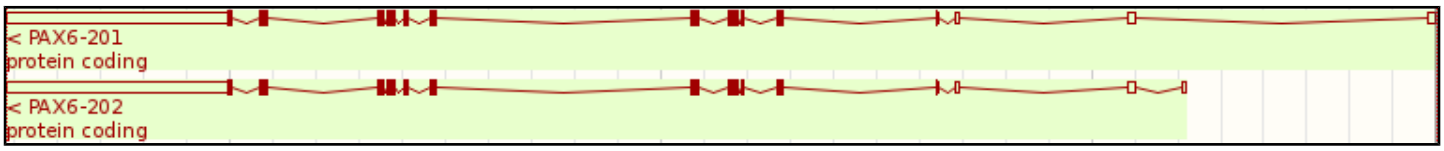


and here it is again being a blue (I think) **Vega Havana protein coding** transcript from the extra track you added to your display just showing **Vega** predictions exclusively.



Which **Ensembl protein coding** transcripts were only predicted by the **Ensembl** pipeline?

There are just **2**. You could verify this (but please do not!) by comparing carefully the **Vega Havana** track that you introduced into the display with the **Genes (Comprehensive …** track from the default display. Or you could read the answer to the next question and take a short cut.



All the Vega Havana track predictions, that are protein coding, should appear also in the **Genes (Comprehensive …** track. Only the protein coding transcript predictions made exclusively by the **Ensembl** pipeline will only be in the **Genes (Comprehensive …** track.

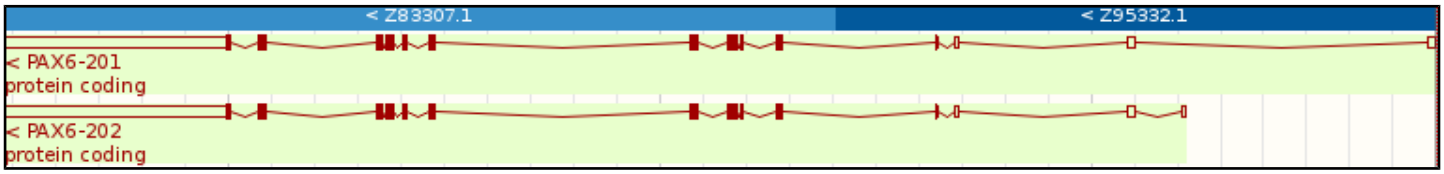
Can you see how to identify the **Ensembl protein coding** predictions only predicted by the **Ensembl** pipeline by the way they are numbered?

The numbering for transcripts predicted by the **Vega-Havana** team commences at **001**.

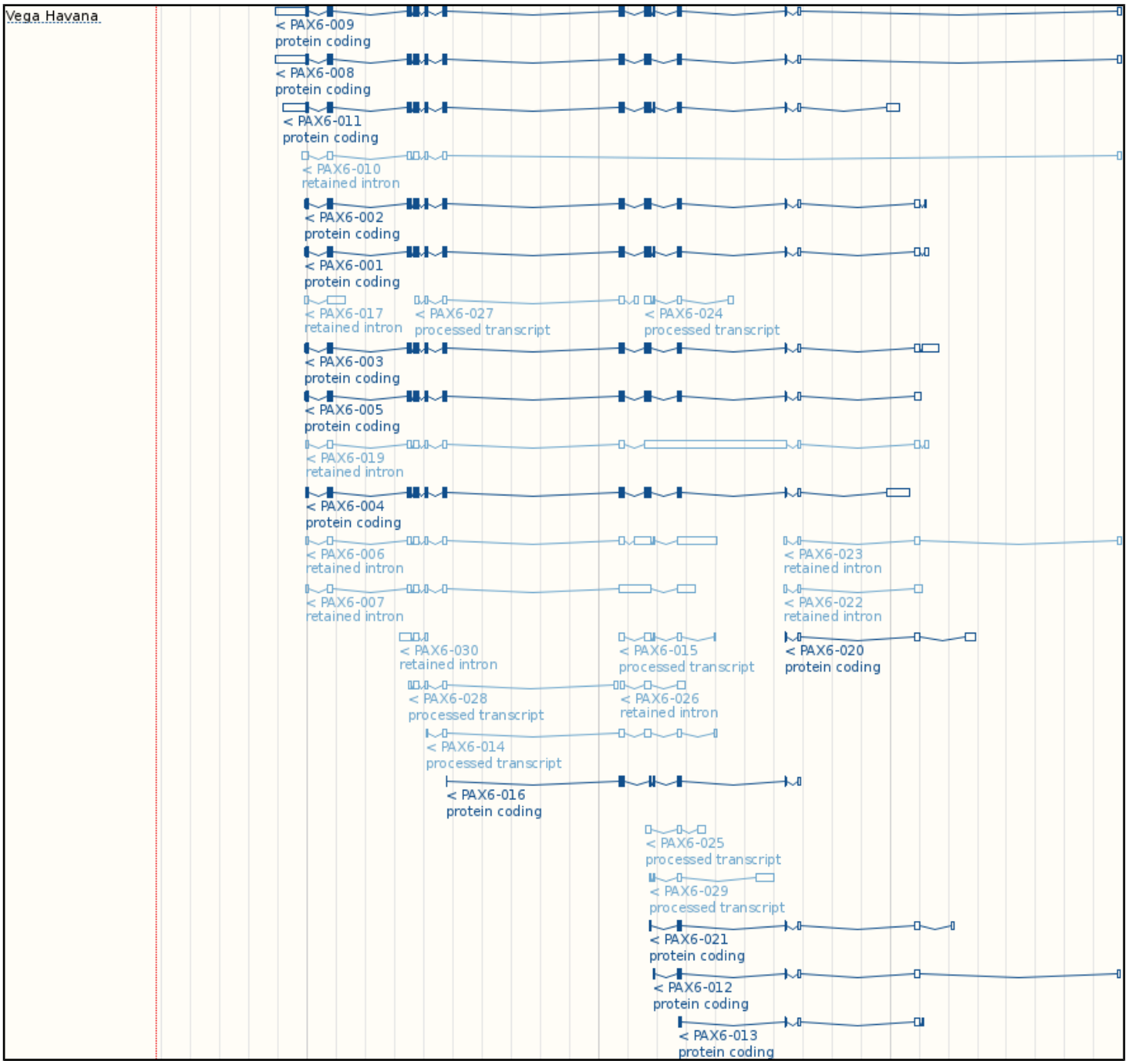
The numbering for transcripts predicted by the **Ensembl** pipeline commences at **201**.

When a transcript is agreed upon by both predictive methods, the **Vega-Havana** numbering takes precedence.

In this example, there are just two transcripts that are predicted by the **Ensembl** pipeline but not by the **Vega-Havana** team. They are the transcripts **PAX6-201** & **PAX6-202**.



You can confirm this (if you really must!) by looking for differences between the **Vega-Havana** only track and the **Genes (Comprehensive …** track. **PAX6-201** & **PAX6-202** are the only protein coding prediction only in the **Vega-Havana** track.



How might the **Gene Ontology** improve sequence database text searching?

As you will have noticed, there are problems with search for sequences using textual search terms against the original annotation content, historically provided by the sequence submitter and not curated. Such annotation could be rather whimsical. Sequences that should match sensibly chosen search keys often do not. Genuine matches are thus inappropriately missed.

Rigorous use of **GO** terms and accession codes, such as those in the tables you have been viewing, must radically improve search efficacy.

Of course, there remains the problem of which **GO** terms apply to which sequences. **RefSeq** and **UniProtKB** do not always agree, for example. However, **GO** annotation has to be a significant step in the right direction.

Why is the **Transcript Ids** column necessary?

The tables you have been viewing relate to the **PAX6** gene. However, individual **GO** terms correspond to a particular protein. Many genes will relate to more that one protein isoform (as is the case for **PAX6**). Individual **GO** terms must be associated with specific transcript(s) (effectively specific protein products). The **Transcript ID** column exists to make the link between a **GO** term and the particular transcript(s) to which it applies.

Extra **Gene Ontolgy** Notes:

Until recently (approximately **September 2015**) **Ensembl** offered, from transcript view pages, links to illustrations of the hierarchical network of **GO** terms associated with particular isoforms (transcripts) of a gene. The illustrations were thought to be of little value for efficiently informing a user of the function of a protein, so they were removed. I think I agree with this move, however I did find the pictures conveyed a quick and intuitive understanding of the way the **GO** database was structured as a network of terms.

When the illustrations were available I asked a question that tried to bring this out. I retain the question here as I think it might still serve a purpose,. An alternative source of enlightenment can be found at the **GO** site here.

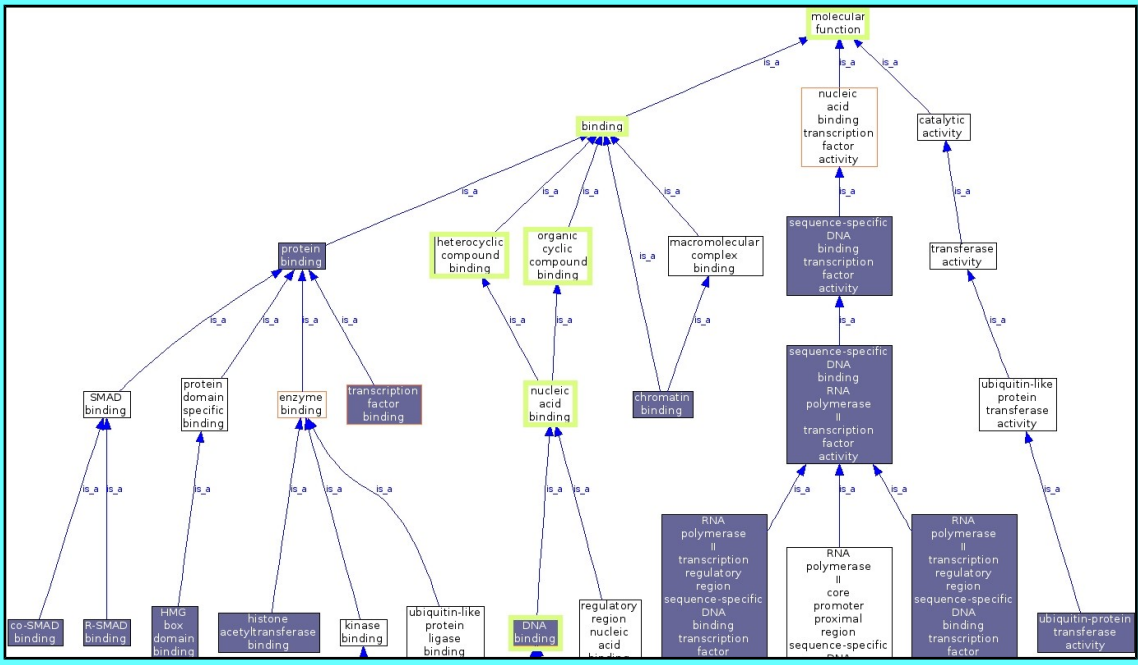
The original questions (and answers) followed instruction to look at the illustrations and were:

Give an example of the most specific terms represented?
Give an example of the most general terms?

I would love to put the whole graph here, but … they are a bit big, so I will restrain myself to two sections.

The **Gene Ontology (GO)** terms are of three categories defining separately the **molecular function**, **cellular component** and **biological process** of a protein. So, there are three graphs.

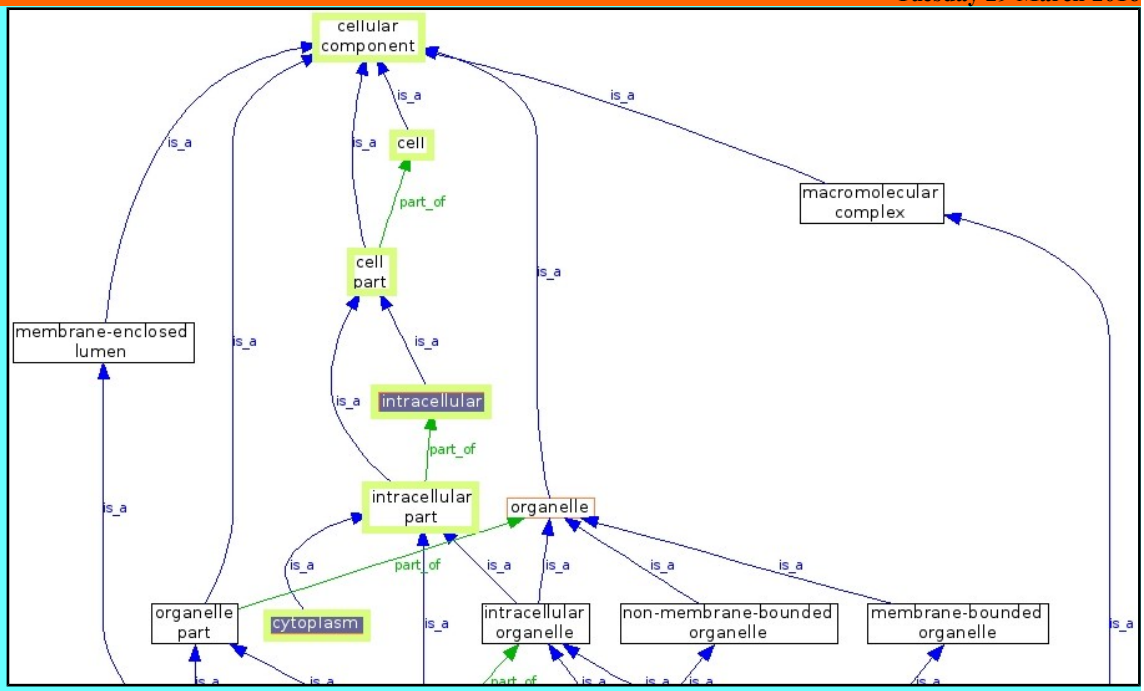
Considering a small portion of the **molecular function** graph, going from most specific to least specific.



The protein in question is considered to be:

> **DNA binding** which implies it *is_a* **nucleic acid binding** protein so it *is_a* **heterocyclic compound binding** & **organic cyclic compound binding** protein and thus *is_a* **binding protein** which *is_a* **molecular function**

Similarly for the **cellular component** graph.



The protein in question is considered to be a:

component of the **cytoplasm** which implies it ***is_a***
component of an **intracellular part** which implies it ***is_a***n
intracellular component which implies it ***is_a*** component of a **cell part** which implies it ***is_a***
cell component which implies it ***is_a*** **cellular component**

In a very similar fashion you could follow a path or two through the **biological process** maze. This one is too scary for me folks, you are on your own!

Note the number of **RefSeq mRNA**s and **RefSeq** proteins associated with **Ensembl** transcripts predictions.
Which of the **24 RefSeq mRNA**s reported by **GeneCards** can be seen here?
Why would you suppose these **mRNA**s were selected and the others ignored?

Summary:

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
PAX6-201	ENST00000419022	6922	436aa	Protein coding	CCDS31452	F1T0F8 P26367	NM_001258462 NM_001310158 NM_001310161 NP_001245391 NP_001297087 NP_001297090	TSL:1 GENCODE basic
PAX6-202	ENST00000606377	6860	436aa	Protein coding	CCDS31452	F1T0F8 P26367	NM_001258463 NM_001310161 NP_001245392 NP_001297090	TSL:1 GENCODE basic
PAX6-009	ENST00000379129	2616	436aa	Protein coding	CCDS31452	F1T0F8 P26367	-	TSL:5 GENCODE basic
PAX6-011	ENST00000379107	2591	436aa	Protein coding	CCDS31452	F1T0F8 P26367	-	TSL:5 GENCODE basic
PAX6-008	ENST00000379132	2574	422aa	Protein coding	CCDS31451	P26367 Q66SS1	NM_001127612 NP_001121084	TSL:5 GENCODE basic APPRIS P1
PAX6-003	ENST00000379123	2160	422aa	Protein coding	CCDS31451	P26367 Q66SS1	NM_000280 NM_001258464 NP_000271 NP_001245393	TSL:1 GENCODE basic APPRIS P1
PAX6-001	ENST00000379115	1763	436aa	Protein coding	CCDS31452	F1T0F8 P26367	NM_001604 NP_001595	TSL:1 GENCODE basic
PAX6-002	ENST00000241001	1631	422aa	Protein coding	CCDS31451	P26367 Q66SS1	-	TSL:1 GENCODE basic APPRIS P1
PAX6-005	ENST00000379111	1627	422aa	Protein coding	CCDS31451	P26367 Q66SS1	NM_001258465 NP_001245394	TSL:1 GENCODE basic APPRIS P1

Most of the top protein coding transcripts are associated with **RefSeq mRNA**s.

The **RefSeq mRNA**s shown here to match the **Ensembl** transcript predictions are **10** of the **11** whose accessions codes commence **NM_**[142]. These appear with their corresponding protein sequences whose accession codes start **NP_**.

GeneCards records the presence in of further **13 RefSeq mRNA**s with accession codes beginning **XM_** and associated proteins with accession codes starting **XP_**. These **13 mRNA**s do not appear here.

The **NM_/NP_** sequences are supported by better evidence than the **XM_/XP_** sequences. **Ensembl** regards only the **NM_/NP_ RefSeq** entries of sufficient quality to predict transcripts.

> • NCBI RefSeq aims to provide a comprehensive set of mRNA and proteins. Manually annotated, reviewed proteins have an ID beginning with NP, known protein, while mRNAs in this category begin with NM. Predicted proteins and mRNA transcripts are not used in the Ensembl Genebuild, and begin with XP and XM, respectively.

Full Answer:

It is curious that the **RefSeq** mRNAs do not map 1 ⟷ 1 with **Ensembl** transcript predictions? I conclude this is just the two predictive strategies disagreeing to a small extent? **Ensembl** appears to be happy to predict just 6 transcripts to match the **11 RefSeq** suggests?

I suggest the fact that several of the **Ensembl** transcript predictions correspond to more than one **RefSeq** mRNA is due to that fact that **Ensembl** will have used just one **RefSeq** mRNA to support the existence of the transcript as a whole and the rest to support individual exons. This supposition seems to be reflected in the **Supporting Evidence** views you will see in a few pages time.

142 I suspect the reason only **10** of the **11** eligible sequences are mentioned in this table is that the **RefSeq** column lists do not set out to be comprehensive. When you view the **Supporting Evidence**, in a page or two, more **RefSeq mRNA** sequences are mentioned.

Using the evidence of the protein alignments, which **PAX6** isoforms do the fruitfly orthologues most resemble?

The protein used to represent **PAX6** human is consistently **ENSP00000404100**. This can most easily be confirmed by clicking on the • Alignment (protein) link for each of the **2 Fruitfly** orthologues in turn to view the relevant orthologous protein alignments. This is the protein sequence of **isoform 5a**, probably chosen as it is the longer option (**436** amino acids as opposed to **422**) and so (from the crude informatics viewpoint) represents more information.

As discovered from **GeneCards**, there are two **Fruitfly** orthologues with the gene names **ey** and **toy**. **Ensembl** agrees, which should not be surprising as **GeneCards** consults **Ensembl** for orthologue information. Looking at the first few lines of the protein alignments for these genes, it is clear that that **14** amino acid insert that defines **isoform 5a** (**THADAKVQVLDNQN**) is not present in either. It is therefore reasonable to conclude that the representative fly proteins are both closest to the canonical protein sequence of **PAX6** human (**isoform 1**).

Protein alignment for **ey**

```
ENSP00000404100/1-436   ------MQN-------------------------SHSGVNQLGGVFVNGRPLPDSTRQ
FBpp0099810/1-898       GKPSPTMEAVEASTASHPHSTSSYFATTYYHLTDDECHSGVNQLGGVFVGGRPLPDSTRQ
                              *:                        .***********.**********


ENSP00000404100/1-436   KIVELAHSGARPCDISRILQTHADAKVQVLDNQNVSNGCVSKILGRYYETGSIRPRAIGG
FBpp0099810/1-898       KIVELAHSGARPCDISRILQ-------------VSNGCVSKILGRYYETGSIRPRAIGG
                        ******************             *************************


ENSP00000404100/1-436   SKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASE
FBpp0099810/1-898       SKPRVATAEVVSKISQYKRECPSIFAWEIRDRLLQENVCTNDNIPSVSSINRVLRNLAAQ
                        *******.******:******************.*.*********************::
```

Protein alignment for **toy**

```
ENSP00000404100/1-436   -MQN----------------------SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHS
FBpp0088249/1-543       MMLTTEHIMHGHPHSSVGQSTLFGCSTAGHSGINQLGGVYVNGRPLPDSTRQKIVELAHS
                         * .                     .***:*******:********************


ENSP00000404100/1-436   GARPCDISRILQTHADAKVQVLDNQNVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATP
FBpp0088249/1-543       GARPCDISRILQ-------------VSNGCVSKILGRYYETGSIKPRAIGGSKPRVATT
                        ************             ******************:*************.


ENSP00000404100/1-436   EVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGADG
FBpp0088249/1-543       PVVQKIADYKRECPSIFAWEIRDRLLSEQVCNSDNIPSVSSINRVLRNLASQKEQQAQQQ
                         **.***:******************** **..*****************:*:* . :
```

Well, maybe also it is not that simple? I would not be surprised If there were isoforms for **ey** and/or **toy** that were roughly equivalent to human **isoform 5a**. The alignment displayed could well reflect the relatively arbitrary choice of **Ensembl** as to which isoform it decides to use for the alignments, rather than any deep and meaningful biological truth. Already you can see that **Ensembl** prefers the (presumably) less important human isoform, merely because it is longer (more letters to match). Again, useful though these displays are, caution is required before reading too much "biology" into them.

Ensembl does not pick up the **prd** fruitfly homologue to **PAX6** mentioned elsewhere (or the others noted previously)? Again, I wonder why. Mind you, **Ensembl** does only claim "**Selected orthologues**"? Still **prd** is a pretty important one to pass over!

Do the **Paralogue** reports of **GeneCards** and **Ensembl** agree? If not, can you explain the discrepancies?

Well, yes they do agree, so there are no discrepancies to explain. Both sources now agree that there are **9 PAX** genes in the human genome. As for orthologues, hardly surprising as **GeneCards** consults **Ensembl** for paralogue information.

I left the question here as recently there was disagreement due to misinterpretation of the **Ancestral taxonomy** field. I wanted to make the point that such small discrepancies are not uncommon and one should therefore be wary. A small price to pay for such immediate access to such volumes of information I suggest.

Which isoform of **PAX6** has been chosen for the alignments, and why would you suppose it was selected?

For these alignments, as for the ortholgues, **Ensembl** consistently uses the same protein (**ENSP00000404100**). This is **isoform 5a** which is again used, because it is longer. Never mind the biology, for the computer's purposes, longest is best. It is more likely to match things as there is more of it. Sophisticated what!

The example is the alignment with **PAX7**.

```
ENSP00000404100/1-436   -------------------------MQN-----SHSGVNQLGGVFVNGRPLPDSTRQKI
ENSP00000364524/1-520   MAALPGTVPRMMRPAPGQNYPRTGFPLE-VSTPLGQGRVNQLGGVFINGRPLPNHIRHKI
                                                ::      .:. ********:******:  *:**


ENSP00000404100/1-436   VELAHSGARPCDISRILQTHADAKVQVLDNQNVSNGCVSKILGRYYETGSIRPRAIGGSK
ENSP00000364524/1-520   VEMAHHGIRPCVISRQLR-------------VSHGCVSKILCRYQETGSIRPGAIGGSK
                        **:** * *** *** *:             **:******* ** ******* ******


ENSP00000404100/1-436   P-RVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPS--VSSINRVLRNL--
ENSP00000364524/1-520   PRQVATPDVEKKIEEYKRENPGMFSWEIRDRLLKDGHCDRSTVPSGLVSSISRVLRIKFG
                        * :****:* .** :**** *.:*:********.:* * ...:**  ****.****
```

Which isoform is most common amongst the **paralogues**?

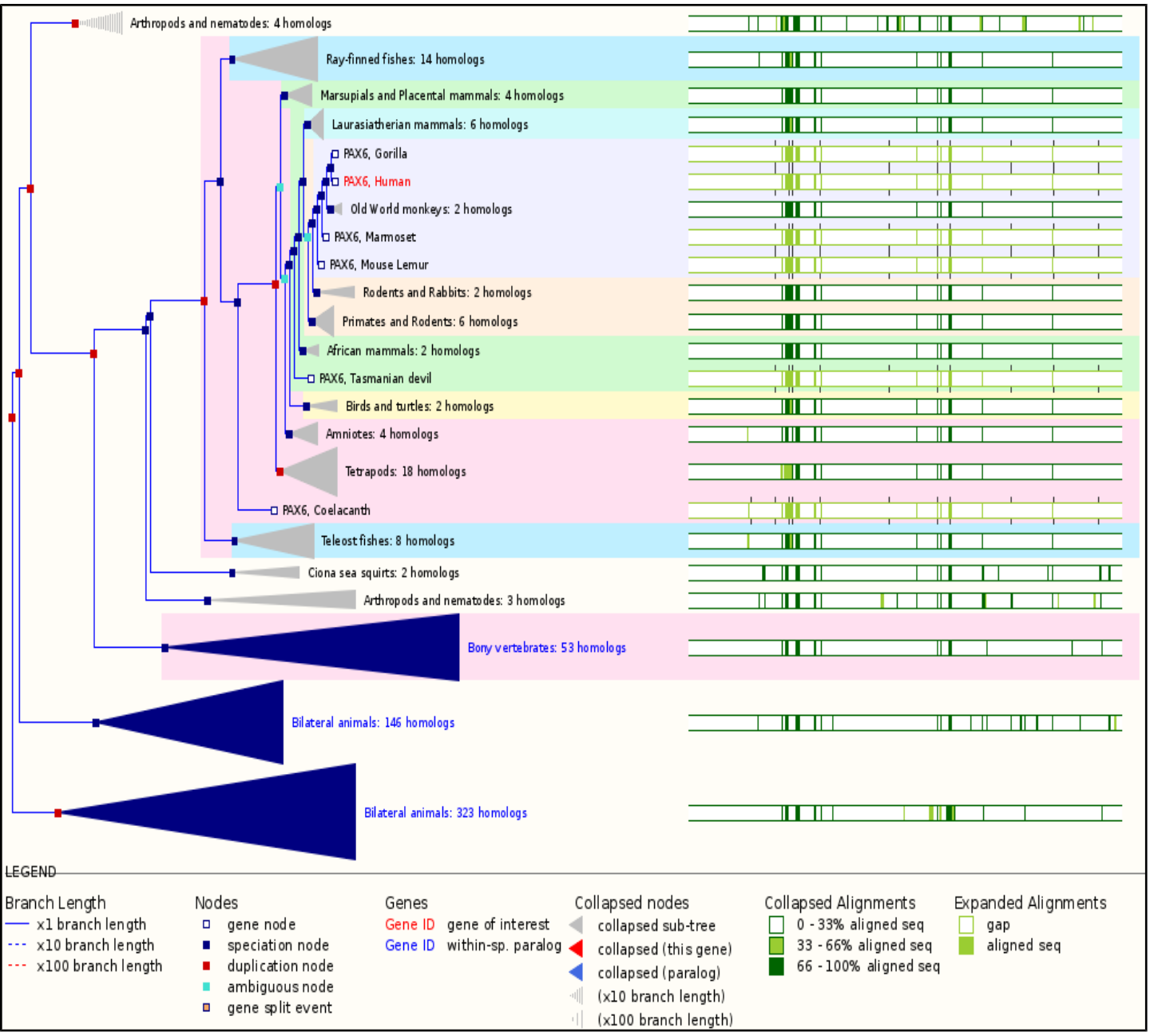
What a cruel and pointless question! As for the orthologues, there used to be a link that would summon forth all the alignments at once. Sadly, this seems to have been retired? So the only way to answer this question with confidence is to investigate each alignment in turn. Do not do that … just read and believe.

Scanning quickly down the eight pairwise protein alignments, you would be able to see that all involve **isoform 5a** of **PAX6** with paralogue sequences that lack the **14** amino acids insert. This proves nothing beyond the fact that **Ensembl** chooses to use what is probably the canonical form for the eight paralogues. There could easily also be an **isoform 5a** equivalent for every paralogue.

An unanswerable question, I suggest? Well, a question for which a complete answer would require further investigation at least. I leave it in, as it does provoke some thought … possibly?

In passing, before the "Show me all the alignments together" link was removed, I had hoped I could view all the orthologues/paralogues compared with each other in one go (i.e. a multiple alignment). No such luck, although, upon reflection, I can see why. That alignment would be very messy outside the **PAX** regions. Neither the orthologues nor the paralogues claim to be similar over their entire length after all.

The Gene Tree, as promised. Click here to return to the Instructions.



Some matching sequences offer no support for the 6th exon from the right. Why do you suppose this is?

The **6th** exon from the beginning of the transcript (the right of the display) is that whose inclusion defines **isoform 5a**. All **mRNA/cDNA/protein** sequences representing the canonical **PAX6** human protein will not include this exon. This will be made absolutely apparent when you look at the textual representation of the exons very shortly.

However, canonical sequences can be useful as supporting evidence for all individual exons, excluding the **6th**, so several are included in the **Exon supporting evidence** tracks.

Why is there no protein or **CCDS** evidence for exons 1, 2 and 3?

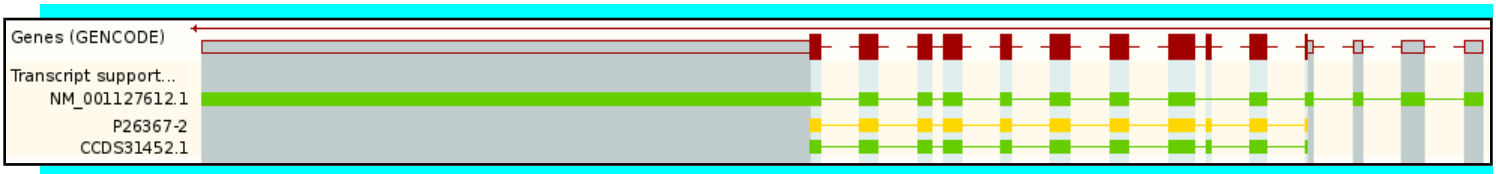
CCDS stands for **C**onsensus **CoD**ing **S**equence. All **CCDS** sequences are just the coding regions of genes. They do not include introns or **3'/5' UTR**s. Accordingly, there should not be (and is not) any **CCDS** (or protein) evidence for exons **1**, **2** or **3** as these exons include no protein coding regions.

Until recently (mid-2014), the **RefSeq** mRNA match in the **Transcript evidence** section did not match all exons. Was this logical?

A rather desparate question to make a quite interesting observation?

This is a problem that has gone away now with the discovery of a **RefSeq** mRNA that exactly supports this transcript. Previously, the best that could be found was the **RefSeq** mRNA **NM_001127612**, which lacked the **6th** exon (I,e, a canonical version of this transcript).

I include here the, now obsolete answer to this question. Perhaps a little pedantic, but I want to retain the particular point that **RefSeq** mRNAs are, justifiably, not perfect and the general point that we are looking at good, but still flawed, predictions here. One should always retain a modicum of scepticism when using these resources.



Not really. If this transcript is real, there must be an **mRNA** that matches its every exon, including the **6th, isoform 5a**, exon. It would seem there is no such **mRNA** sequence in **RefSeq** however. Ideally, there would be.

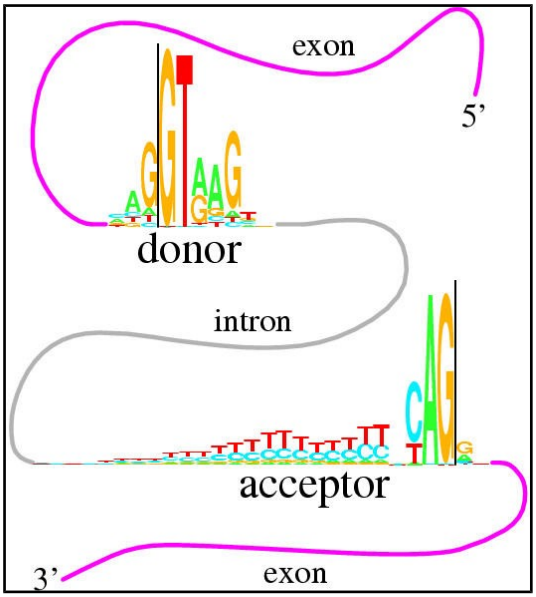
The problem is that **RefSeq mRNA**s are not usually the sequence of a single **mRNA**. They are a composite of a number of sequences assumed to represent the same biological entity. Things can go wrong, as they probably have here. The **RefSeq** construct suggests a canonical **mRNA**, **Ensembl** suggests an **isform 5a mRNA**. Well, one of them is probably right.

Ensembl first uses **CCDS** matches to establish the presence of the gene and all its possible isoforms. The **RefSeq** matches are primarily required to establish various forms of the **UTR**s for each prediction of a coding region. Here **Ensembl** has chosen to believe the combination of the **UTR**s suggested by **NM_001127612** and the coding region suggested by **CCDS31452**. **Ensembl** must be assuming that the **RefSeq mRNA** is not **100%** correct. This is clearly possible, but I do not regard the reasoning of **Ensembl** as particularly transparent, particularly as there is another **CCDS** match that would entirely support the **RefSeq** evidence?

How many exons are there in this transcript?

14, I think the relevance of this question has reduced over the aeons. I certainly cannot recall clearly why I asked?

What are the first two bases and what are the last two bases of nearly every intron?



As you are probably well aware, introns are highly conserved at each end. They typically begin with **GT** and end with **AG**. This rule is obeyed by all but one of the introns of this transcript (**intron 3-4** starts **GC** rather than **GT**).

As the cartoon suggests, the conservation does not apply just to the first and last two bases, but that is where the conservation is most strict. So strict that when exceptions from this rule were sought in the databases, it was thought most of the deviations were due to annotation error!

The cartoon also suggests that introns have C/T rich regions towards their ends (the **Polypyrimidine tract**). This too is clearly evident in most of the introns of this transcript, even though only small parts of the introns are displayed.

How long is the sixth exon and why would this concur with your expectations?

It is **42** base pairs long, so it codes for **14** amino acids. Specifically, it codes for the **14** extra amino acids that define **isoform 5a**.

Explain the **Start Phase** and **End Phase** columns?

An exon/intron boundary can occur anywhere in a codon. The **Start** and **End Phases** record how an intron has been inserted into a coding region with respect to the coding reading frame.

If an exon ends at the end of a codon, then its **End Phase** is **0**.
Clearly, the next exon must begin at the start of a codon. Its Start Phase is also **0**.

If an exon ends after the first base of a codon, then its **End Phase** is **1**.
Clearly, the next exon must begin after the first base of a codon. Its **End Phase** is also **1**.

If an exon ends after the second base of a codon, then its **End Phase** is **2**.
Clearly, the next exon must begin after the second base of a codon. Its **End Phase** is also **2**.

I attempt a picture, though I am sure that is clear? I just like pictures, and lots of colours.

```
    S  A  I  L  Q                T  H  A  D  A  K  V  Q  V  L                V  S  N ...
...TCCCGAATTCTGCAG← intron →ACCCATGCAGATGCAAAAGTCCAAGTGCTG← intron →GTGTCCAAC...
      End Phase 0              Start Phase 0      End Phase 0          Start Phase 0

    S  A  I  L                Q  T  H  A  D  A  K  V  Q  V  L                V  S  N ...
...TCCCGAATTCTGC← intron →AGACCCATGCAGATGCAAAAGTCCAAGTGCT← intron →GGTGTCCAAC...
     End Phase 1              Start Phase 1      End Phase 2          Start Phase 2
```

Where is the start and end of the **Prosite Paired Box** pattern (**R-P-C-x(11)-C-V-S**)?
Where, in relation to the pattern, are the extra **isoform 5a** amino acids?
Why might the positions of these two features be significant?

The pattern **RPCxxxxxxxxxxxCVS** is pretty easy to spot, but the **14** amino acid insertion of **isoform 5a** (**THADAKVQVLDNQN**, corresponding to the entire 3rd coding exon) has landed right in the middle of the pattern!

```
MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQ
NVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRL
LSEGVCTNDNIPSVSSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGSWGTRPGWYPGT
SVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALE
KEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIP
ISSSFSTSVYQPIPQPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPSFTMANNLPMQPP
VPSQTSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMNSQPMGTSGTTSTGLISPGVSVPVQ
VPGSEPDMSQYWPRLQ
```

This is very significant as far as the efficacy of the **PROSITE** pattern is concerned. Despite the claim of the **PROSITE** documentation that the pattern picks up all known **Paired** domains, it is not going to work on any **isoform 5a** protein. This claim relies on the fact that all the **PAX** proteins in **SwissProt** are represented by their canonical sequence. The very common **isofrom 5a** is consistently recorded as a **FEATURE**, not search by the pattern matching software.

In order to detect just the **PAX isoform 5a**, the pattern would have to be:

R-P-C-x(25)-C-V-S

To detect both isoforms, using just one pattern:

R-P-C-x(11,25)-C-V-S

would work, but would be insufficiently specific and would generate far too many false positives. These sort of patterns are useful, but only with caution. They are valuable because of their simplicity, but very fragile.

Where do each of the **SMART** domains start and end?
Do the regions match your earlier recording? If not, why not?

SMART predicts a **Paired_dom** from amino acid positions **4** → **142**.

Smart	4	142	Paired domain	SM00351

Previously you noted that **UniProtKB** claims a **Paired** domain extending from residues **4** → **130**.

Domain[i]	4 – 130	127	Paired

SMART predicts a **Homeodomain** from amino acid positions **224** → **286**.

Smart	224	286	Homeobox domain	SM00389

Previously you noted that **UniProtKB** predicts a **Homeobox** extending from residues **210** → **269**.

DNA binding[i]	210 – 269	60	Homeobox

So no agreement in nomenclature or position. It would be nice if we could all agree what to call things, but I suppose that is not likely to happen in this world of personalised whimsy. There is more justification for the variation in predicted position of the various domains.

SMART and **UniProtKB** will be using marginally different methods and models to predict domains. It is exceedingly good news that there is pretty comprehensive agreement as to which domains present and roughly where they are. However, it would be unreasonable to expect differing prediction strategies to come up with precisely the same answers, correct to the amino acid position. After all, I wonder if any two human experts would exactly agree where a particular feature began and ended, even given all possible evidence?

What are the **Interpro** database accession codes for the two major **PAX6** domains?

There are three independent predictions (from **Prosite_profiles**, **Smart** & **Pfam**) of a **Homeobox domain**. Each and all suggest that this protein belongs to the **Homeobox domain Interpro** family which has the **Accession** number **IPR001356**.

Prosite_profiles	222	282	Homeobox domain	PS50071	IPR001356
Smart	224	286	Homeobox domain	SM00389	IPR001356
Pfam	226	281	Homeobox domain	PF00046	IPR001356

There are four independent predictions (from **Prosite_profiles**, **Smart Pfam** & **Prints**) of a **Paired domain**. Each and all suggest that this protein belongs to the **Paired domain Interpro** family which has the **Accession** number **IPR001523**.

No matches with **Prosite_patterns** are here to support the presence of either the **Paired domain** or the **Homeobox domain**.

Pfam	4	142	Paired domain	PF00292	IPR001523
Smart	4	142	Paired domain	SM00351	IPR001523
Prosite_profiles	4	144	Paired domain	PS51057	IPR001523
PRINTS	8	23	Paired domain	PR00027	IPR001523
PRINTS	26	44	Paired domain	PR00027	IPR001523
PRINTS	60	77	Paired domain	PR00027	IPR001523
PRINTS	78	95	Paired domain	PR00027	IPR001523

The **Prosite_pattern** for a **Homeobox domain** does match and is accepted as sufficient evidence by **Interpro** to indicate a **Homeobox, conserved site**, which has its own **Interpro** entry. **Accession** number **IPR017970**. It is not, however, of sufficient significance to contribute to the evidence for the **Homeobox domain** itself.

Prosite_patterns	257	280	Homeobox, conserved site	PS00027	IPR017970

These results are entirely consistent with the **Interpro** results you have seen previously ***except*** that there is no match here with the **Prosite_pattern** for a **Paired domain**. This is because the transcipt under incestifation here represnts an **isofrom 5a** protein. This means the **Paired domain Prosite_pattern** will not match because of the extra **14** amino acids (see answer to previous question).

Had you investigated the transcript for a canonical protein (for example, **ENST00000379123**, the gold transcript you looked at earlier) the pattern would have matched and been counted as further evidence for a **Paired domain**.

Pfam	4	128	Paired domain	PF00292	IPR001523
Smart	4	128	Paired domain	SM00351	IPR001523
Prosite_profiles	4	130	Paired domain	PS51057	IPR001523
PRINTS	8	23	Paired domain	PR00027	IPR001523
PRINTS	26	44	Paired domain	PR00027	IPR001523
Prosite_patterns	38	54	Paired domain	PS00034	IPR001523
PRINTS	46	63	Paired domain	PR00027	IPR001523
PRINTS	64	81	Paired domain	PR00027	IPR001523

Why does **Prints** appear to predict four **Paired_domains**?

Prints does not find the **Homeobox_domain** at all. You already will be aware that this search fails from looking at the **UniProtKB** predictions.

Prints appears to find **FOUR** **Paired_domains**. Of course, this is only because of the way **Prints** works. **Prints** finds **FOUR** signatures that together indicate **ONE Paired_domain**. I think we might have been here before? Possibly too many times?

Which domain, **Paired domain** or **Homeobox domain** is more common in humans?
How many human **PAX** genes are there?

As you will have expected, there are but **9 Paired_doms** in the Human genome. There are many more **Homeobox_doms**.

Are all the **PAX** genes on **Chromosome 11**?

Of course not? What a stupid question!

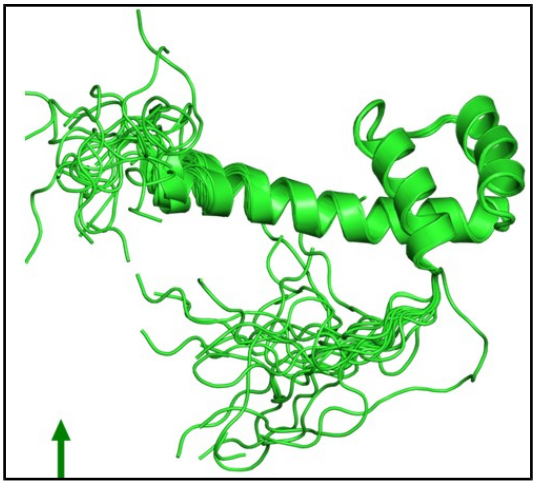
Well, I suppose they could all be on **Chromosome 11**? By chance … or maybe design … who knows, the lack of predictable pattern in all this business never ceases to astound me.

But, philosophy aside, the answer is **NO**.

How does **Interpro** match with the **PAX6 Paralogues** reported by **Ensembl/GeneCards** earlier?

The evidence from both **GeneCards** and **Ensembl** is that there are **9 PAX** paralogues in Human. Yep, we all agree and … these questions are becoming a trifle repetitive one feels!

Can you explain the strangely frayed ends displayed in some of the representations of the **2cue** 3D structure?



2cue is a 3D structure determined by **N**uclear **M**agnetic **R**esonance (**NMR**).This is a process that does not involve immobilizing the target as a crystal (as is the case with structures determined by **X-ray crystallography**). Parts of the protein will still be moving around whilst its structure is being determined.

I think of **NMR** as analogous to taking a long exposure photograph of a group of children. Each child will appear in many different places! The frayed ends represent various positions in which the ends of the **homeobox** were detected during the **NMR** process.

In some views, including the one you were offered to move around, all the possible positions are averaged out before the structure is stored. I prefer the fuzzy view ... much more fun.

I broadly believe that which I have just typed, however, I must stress that my understanding of **NMR** is tragically incomplete. If anyone would like to offer a better explanation, I am very willing to hear it.

From your investigations of **Global Alignment**:

What do you suppose these regions represent?

Exons

Is the number of matching regions consistent with the number of exons you might expect?

Well, I think the only point left to this question is to say that the resolution of the graphics is insufficient to make a proper judgement concerning how many exons there might be. From looking at **Ensembl**, I would expect about **13** or **14**? From looking at this picture … I am simply not sure? I seem to have counted **12** when I first wrote this section.

I leave the question, but the need to know the exact number of exons, at this stage has become vanishingly small over the eons of evolution endured by these exercises. Let us say … **12**! Which I think is right.

If not, can you explain the discrepancy?

What discrepancy? I pause for thought, but why would I have an expectation of exon count here anyway? Depends which **transcript** and which **isoform**.

So … I decline to try to explain anything further. Stupid question!

Why do you suppose that might be?

The larger the **window size**, the less likely it is that small features, or noise, will be detected. The signal from a small feature (noise or otherwise) will be too weak to influence the general impression of a relatively large window. Therefore. Ther longer the **window size** you choose, the stronger/longer the feature has to be before it will be noticed.

The presence of an extended strong feature will be detected as long as some part of it remains in a scoring window. The longer the **window**, the further the **window** must be from the feature before its influence is insignificant. Therefore, the longer the **window**, the longer strong features will appear to be.

How many convincingly aligned regions did you see?

4

Roughly how many did you expect?

12 or so … one per exon anyway.

Clearly, this alignment is not correct. Can you explain why?

This alignment algorithm only wishes to maximise an alignment score. It sees ***ALL*** the high scoring exon regions, however, as the gaps between many of the exons (introns that is) are so long that the penalties for representing them correctly are greater than the gain achieved by the inclusion the extra exons in the alignment. Arithmetically, it is better to align all the exons either side of the **4** exons that were aligned sensibly, in the biologically improbably fashion shown. Arithmetically the best alignment, biologically ridiculous!

This behaviour is exaggerated because this program regards the enormous gaps in has suggested at the start and end of the alignments as “free”. Some global alignment programs offer the option of penalising the ends gaps in the same way as for internal gaps. Normally, not penalising end gaps is sensible as it allows for the sequences to have slightly different lengths. In this case, penalising end gaps should have resulted in a better alignment.

Had you used stretcher (the faster, less vigorous algorithm offered by the **EMBOSS** package) you would have got a much improved answer in this case (but not generally). This is because **stretcher** works in a way far closer to the way an informed human might think. **stretcher** does not mindlessly insist of the highest alignment score. Instead, it looks for all the high scoring regions (i.e. all the exons) and then computes the best way to link them together. The result is a far more convincing alignment, but not the arithmetically best scoring answer.

How many matching regions are there this time?

Where you to trawl though your textual output carefully (or simply take my immaculate word for it), you would find **13** perfectly (or nearly so) aligned regions, implying **13** exons.

To be pedantic, the nicely aligned regions do not match the exons exactly (as will become apparent later), so it is not possible to claim definite evidence for any particular number of exons. However, **13** has to be a pretty confident estimate.

Is the count **now** roughly as you would expect?

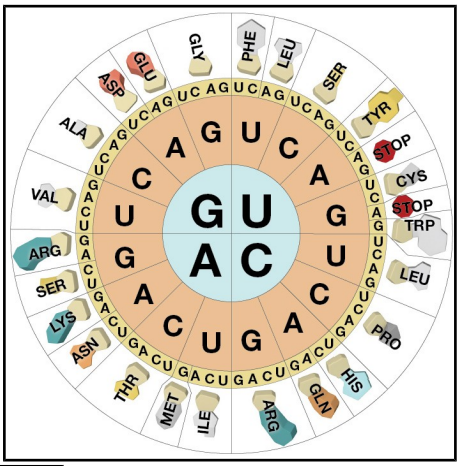
Yes, roughly as suggested by **Ensembl**.

From your investigations comparing mRNA/cDNA with genomic DNA:

What is the amino acid corresponding to this position in the mRNA of the aniridia patient?

This is easy to answer. The top sequence is the mRNA from the **aniridia** patient. **spline** is kind enough to explicitly inform us that the mutated codon, **CCT**, will be expressed a **P**roline.

So, why not translate the wild type genomic sequence also **spline**?! Easy enough to look up. But I resent having to do so!



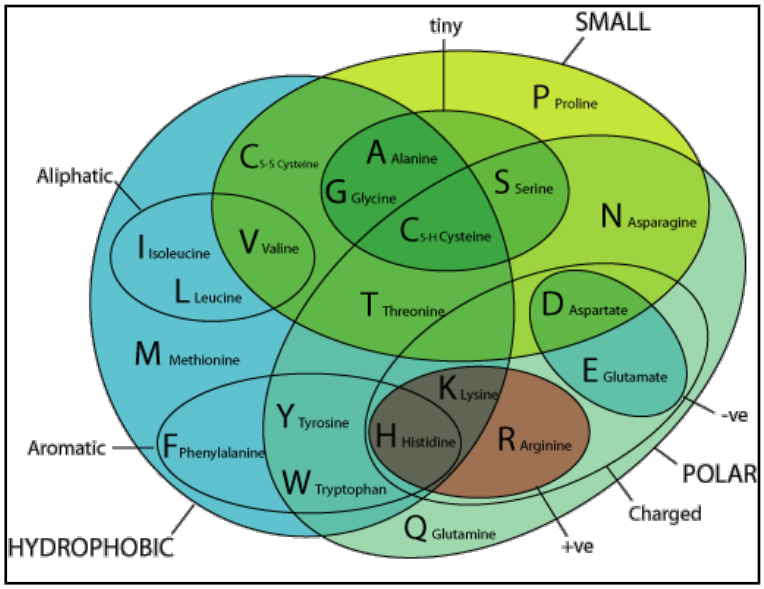
From this rather beautiful representation of the **Genetic Code**, I conclude:

Patient **CCT** → **Proline** (**P**)
Wild Type **GCT** → **Alanine** (**A**)

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier
Natural variant	17 – 17	1	N → S in AN. 1 Publication		VAR_003808
Natural variant	18 – 18	1	G → W in AN. 1 Publication		VAR_003809
Natural variant	19 – 19	1	R → P in AN. 1 Publication		VAR_047860
Natural variant	22 – 26	5	Missing in AN. 2 Publications		VAR_008693
Natural variant	26 – 26	1	R → G in PETAN. 2 Publications		VAR_003810
Natural variant	29 – 29	1	I → S in AN. 1 Publication		VAR_008694
Natural variant	29 – 29	1	I → V in AN. 1 Publication		VAR_003811
Natural variant	33 – 33	1	A → P in AN. 1 Publication		VAR_008695
Natural variant	37 – 39	3	Missing in AN. 1 Publication		VAR_008696

Or, looking back at the relevant **Uniprot Feature Table**, I come to an identical conclusion. This is the Natural Variant at position **33**, as you could easily confirm with a bit of counting along the alignment.

Either way, it could be a quite serious mutation. **Alanine** and **Proline** having quite distinct properties.



Also, according to the font of all easily packaged knowledge, Wikipedia:

> "**Proline** and **glycine** are sometimes known as "**helix breakers**" because they disrupt the regularity of the α helical backbone conformation; however, both have unusual conformational abilities and are commonly found in **turns**."

The **helices** of this protein are of particular importance to its vital **DNA Binding** role.

How do you interpret the **Details** column for exons 1 and 5?

Summary:

The **Details** column shows the alignments of each exon in a compressed format described in the **spline** documentation as illustrated.

11. Alignment transcript	Alignment transcript represents full details of the alignment in a form of a string composed of characters 'M', 'R', 'I' and 'D' where each character corresponds to an elementary command (Match, Replace, Insert or Delete) needed to transform the query segment into the subject segment. The string is encoded with RLE.

The majority of the exon alignments are trivial.

#	Query	Subject	Span(bp)	Coverage(%)	Overall(%)	Exon(%)	CDS(%)	In-frame(%)
1	pax6-cDNA(+)	pax6-genomic(+)	7245-28540	100.00	99.94	99.94	0.00	0.00

Graphics|Text

#	Query	Subject	Idty	Len	Q.Start	Q.Fin	S.Start	S.Fin	Type	Details
+1	pax6-cDNA	pax6-genomic	0.981	103	1	101	7245	7347	<exon>GT	M53IM5IM43
+1	pax6-cDNA	pax6-genomic	1	188	102	289	7447	7634	AG<exon>GT	M188
+1	pax6-cDNA	pax6-genomic	1	77	290	366	11537	11613	AG<exon>GC	M77
+1	pax6-cDNA	pax6-genomic	1	61	367	427	12000	12060	AG<exon>GT	M61
+1	pax6-cDNA	pax6-genomic	0.992	131	428	558	15628	15758	AG<exon>GT	M86RM44
+1	pax6-cDNA	pax6-genomic	1	216	559	774	16686	16901	AG<exon>GT	M216
+1	pax6-cDNA	pax6-genomic	1	166	775	940	17606	17771	AG<exon>GT	M166
+1	pax6-cDNA	pax6-genomic	1	159	941	1099	23674	23832	AG<exon>GT	M159
+1	pax6-cDNA	pax6-genomic	1	83	1100	1182	24348	24430	AG<exon>GT	M83
+1	pax6-cDNA	pax6-genomic	1	151	1183	1333	24660	24810	AG<exon>GT	M151
+1	pax6-cDNA	pax6-genomic	1	116	1334	1449	24909	25024	AG<exon>GT	M116
+1	pax6-cDNA	pax6-genomic	1	151	1450	1600	27602	27752	AG<exon>GT	M151
+1	pax6-cDNA	pax6-genomic	1	98	1601	1698	28443	28540	AG<exon>AA	M98

For example:

For **Exon 2,** **spline** informs us **M188**, meaning "There are **188** bases aligned and they all **M**atch perfectly".

For **Exon 3**, **spline** informs us **M77**, meaning "There are **77** bases aligned and they all **M**atch perfectly".

The only **2** interesting entries are those were there are some disagreements. That is, the entries for **Exons 1** and **5**, which, following the documentation, I translate thus:

Exon 1 – M53IM5IM43

An alignment of **103** bases, the first **53** of which **M**atch perfectly (**M53**), there then follows an **I**nsertion (**I**), a further **5 M**atched bases(**M5**), a second **I**nsertion (**I**) all finished off with **43 M**atched bases (**M43**).

Exon 5 – M86RM4R

An alignment of **131** bases, the first **86** of which **M**atch perfectly (**M86**), there them follows a **R**eplacement (**R**) and a further **44 M**atched bases(**M44**).

Full Answer:

From the individual **Exon 1** display, it can be inferred that the declaration of an **I**nsertion or a **D**eletion is made to describe the type of variation required to transform the **cDNA** (**Query**) sequence into the **genomic** (**Subject**). Hence the two **indels** (**I**nsertions or **D**eletions) are considered to be **I**nsertions.

```
   1 CAGAGGTCAGGCTTCGCTAATGGGCCAGTGAGGAGCGGTGGAGGCGAGGCCGG-CGCCG-CACACACACA
     ||||||||||||||||||||||||||||||||||||||||||||||||||||| ||||| ||||||||||
7245 CAGAGGTCAGGCTTCGCTAATGGGCCAGTGAGGAGCGGTGGAGGCGAGGCCGGGCGCCGGCACACACACA
```

Not that it is a vital issue, but I would have thought the other way around was more logical? That is, to consider the **genomic** sequence as the **reference** against which a particular **mRNA** might vary. In other words, what we see here would surely be more relevantly recorded as "This **mRNA/cDNA** has two **D**eletions relative to the **genomic** sequence which, presumably, attempts to represent the norm in the general population"? Just the reflection of an irretrievable pedant, but I am right, nevertheless!!!

In the documentation it enigmatically states "The string is encoded with **RLE**.". Just in case, **RLE** stands for **Run-length encoding** which is succinctly defined by **Wikipedia**. In a nutshell, it is a very simple form of data compression that recognizes that:

```
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

can be compressed to:

```
60X
```

which has to be very effective for any data that has runs of identical characters of significant length. This is certainly the case here where one would expect long stretches of **M**s in most alignments. Of course, life would get tricky if the data included numeric characters, but that is not an issue here[143].

I think it worth mentioning, that this way of representing an alignment is a simplification of **CIGAR** format[144]. This format is used for **SAM** (**S**equence **A**lignment **M**ap) and **BAM** (**B**inary **A**lignment **M**ap, exactly the same as **SAM**, except compressed) files. You will be engulfed in **SAM/BAM** files if you ever do any **N**ext **G**eneration **S**equencing (**NGS**).

So, straight form the **SAM/BAM Format Specification** I copy the table of **CIGAR** enlightenment.

CIGAR: CIGAR string. The CIGAR operations are given in the following table (set '*' if unavailable):

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

- H can only be present as the first and/or last operation.
- S may only have H operations between them and the ends of the CIGAR string.
- For mRNA-to-genome alignment, an N operation represents an intron. For other types of alignments, the interpretation of N is not defined.
- Sum of lengths of the M/I/S/=/X operations shall equal the length of SEQ.

Note, in particular, the extended range of **Op**erators and the different meaning associated with the operator '**M**'. The operators '**=**' and '**X**' are such that any '**M**' is either an '**=**' or and '**X**' but never both. Which leaves one pondering whem one might use '**M**' in preference to either an '**=**' or an '**X**'?

Where is the substitution in the aniridia patient mRNA?
Where is the substitution in the Genomic Sequence?

spline makes one work quite hard to answer this one! Unless I am missing something.

From the alignment of **Exon 5**, the exon including the **R**eplacement, with a bit of squinting, it can be confirmed that the **R**eplacement is at:

```
        R  Q  K  I  V  E  L  P  H  S  G  A  R  P  C  D  I  S  R  I  L  Q
  493 CGGCAGAAGATTGTAGAGCTACCTCACAGCGGGGCCCGGCCGTGCGACATTTCCCGAATTCTGCAG....
      ||||||||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||||||
15693 CGGCAGAAGATTGTAGAGCTAGCTCACAGCGGGGCCCGGCCGTGCGACATTTCCCGAATTCTGCAGGTGA
```

Base pair position **514** of the aniridia patient's **mRNA**

Base pair position **15714** of the **genomic** sequence

It might also have been relevant to ask which amino acid position corresponded to the **R**eplacement. To discover this one would need to look at the alignment of **Exon 3**, where the coding begins.

```
                                                          M  Q  N
  367 .....AGCCCCATATTCGAGCCCCGTGGAATCCCGCGGCCCCCAGCCAGAGCCAGCATGCAGAACA....
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
11995 AACAGAGCCCCATATTCGAGCCCCGTGGAATCCCGCGGCCCCCAGCCAGAGCCAGCATGCAGAACAGTAA
```

More squinting, and I conclude the **A** of the **ATG** representing the initial **Methionine** of the protein coding region is at position **418**. That is, the **3' UTR** ends at position **417**. So the **R**eplacement is at:

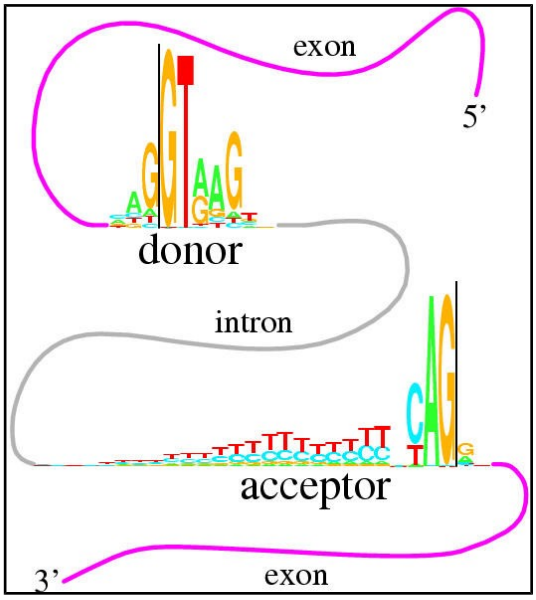
Base position **514** – **417** = **97** of the protein coding region of the **mRNA**.

As **97** / **3** is **32** remainder **1**, the **R**eplacement is at codon position **1** of the **33rd** amino acid of the protein.

Which you knew already, of course! Cannot help thinking that **spline** might have helped a bit more here?

143 The **Wikipedia** article shows how this complication might be overcome.
144 There may or may not be some justification for calling the format **CIGAR**, but if there is, I have no idea what it might be.

Compare the predicted **splign** intron/exon boundaries with the conservation suggested by the logo?
What deviation(s) from the model suggested by the logo can you see?



You may have gathered, I rather like this logo, although I rather think it is leading me to make the same point a trifle to often?

The logo is in almost **100%** agreement with the predictions of **spline**.

As you will have noted previously, when looking at the **Ensembl** predictions of exons locations of a similar transcript of the **PAX6** human gene, there is a single exception.

Type
<exon>GT
AG<exon>GT
AG<exon>GC
AG<exon>GT
AG<exon>GT
AG<exon>GT
AG<exon>GT
AG<exon>GT
AG<exon>GT
AG<exon>GT
AG<exon>GT
AG<exon>GT
AG<exon>

The easiest way to show this in the **spline** output is to look at the **spline** text output again.

The **Type** column records the type of all the **<exon>** alignments it predicts. It also records **2 flanking intron base pairs**.

It is clear that the only time the spline prediction deviates from the model suggested by the logo is at the end of the **3rd** exon. Here there is **GC** rather then **GT**. Well, nothing is perfect!

From your investigations of **Local Alignment**:

How might the gap around **24,600** in the genomic sequence been positioned more intelligently?

spin has positioned a gap in this region merely to maximize the overall alignment score. There is more than one way of achieving this simple goal. However, if it were to be recognized that the gap to be positioned was to represent an intron, then one of the arithmetically equivalent options becomes far more attractive than the others. This "best" option is not the one chosen by **spin**, which is forgiveable as **spin** had nor reason to expect an intron and was not written to understand the properties of introns anyway.

The alignment chosen for this region by spin was:

```
          24392     24402     24412     24422     24432     24442
pax6-genomic CTAGCAGCCAAAATAGATCTACCTGAAGCAAGAATACAGGTACCGAGAGACTGTGCAGTT
             |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
   pax6-cDNA ctagcagccaaaatagatctacctgaagcaagaatacaggta..................
          1144      1154      1164      1174      1184          -

                                    ...

          24632     24642     24652     24662     24672     24682
pax6-genomic GTGCTAACCTGTCCCACCTGATTTCCAGGTATGGTTTTCTAATCGAAGGGCCAAATGGAG
                                       ||||||||||||||||||||||||||||
   pax6-cDNA ..............................tggttttctaatcgaagggccaaatggag
             -         -         -         -      1195      1205
```

Shifting the gap 3 places to the left neither changes the size of the gap nor the perfection of the alignment either side of the gap and so does not affect the alignment score. However, it does mean the gap begins with an **GT** and ends with a **AG** which is what one might expect if it were known that the gap represented an intron. So, if **spin** was a little better informed, the improved alignment would have been:

```
          24392     24402     24412     24422     24432     24442
pax6-genomic CTAGCAGCCAAAATAGATCTACCTGAAGCAAGAATACAGGTACCGAGAGACTGTGCAGTT
             |||||||||||||||||||||||||||||||||||||||||
   pax6-cDNA ctagcagccaaaatagatctacctgaagcaagaatacag.....................
          1144      1154      1164      1174      1184          -

                                    ...

          24632     24642     24652     24662     24672     24682
pax6-genomic GTGCTAACCTGTCCCACCTGATTTCCAGGTATGGTTTTCTAATCGAAGGGCCAAATGGAG
                                     |||||||||||||||||||||||||||||
   pax6-cDNA ............................gtatggttttctaatcgaagggccaaatggag
             -         -         -         -      1195      1205
```

This is the alignment that the customized program **splign** chose as **splign** understands something of the expected properties of introns. **spin** was confused because it is a general alignment program concerned only with the simple arithmetic of maximising alignment scores.

Why do you suppose your aligned exons are not presented in the correct positional order?

To **spin**, the logical order in which to present the alignments is that governed by quality rather than position. So, the highest scoring alignment, rather than the first exon alignment, will be at the top of the list. I think this is generally logical. Once again, the program **splign**, knowing it was looking for an ordered set of exons, was more obliging.

From your investigations of **ORF detection**:

At what base position does the coding sequence of the mRNA commence?

From the illustration in the notes above, it can be determined (with a bit of pain filled arithmetic) that the **A** base of the **ATG Methionine** codon that starts the coding region of this mRNA is at position **418**.

Note that the nearest **Stop** Codon before the **Methionine** is well before the start of the ORF indicated by **plotorf**. This shows that the definition of an ORF being used by **plotorf** is **Start → Stop**, rather than **Stop → Stop**. As you can see from the output below (generated by another **Emboss** program called **showorf**).

```
            ---------|---------|---------|---------|---------|
       301 TAActaggggcgcgcagatgtgtgaggccttttattgtgagagtggacag 350
F1       1 *  L  G  A  R  R  C  V  R  P  F  I  V  R  V  D  R  16

            ---------|---------|---------|---------|---------|
       351 acatccgagatttcagagccccatattcgagccccgtggaatcccgcggc 400
F1      17  H  P  R  F  Q  S  P  I  F  E  P  R  G  I  P  R  P 33

            ---------|---------|---------|---------|---------|
       401 ccccagccagagccagcATGcagaacagtcacagcggagtgaatcagctc 450
F1      34   P  A  R  A  S  M  Q  N  S  H  S  G  V  N  Q  L   49
```

What is the base position of the last base of the coding sequence?

Slide along the sequence display and you should be able to conform the coding region of this mRNA ends at position **1683**. To complete the dreadful pedantic excess, I support his with more **showorf** output.

```
            ---------|---------|---------|---------|--------
      1651 cctgatatgtctcaatactggccaagattacagTAAaaaaaaaaaaa 1698
F1     450 P  D  M  S  Q  Y  W  P  R  L  Q  *  K  K  K  K   4
```

Leaving only the mystery of what **showorf** might be trying to convey by the **"4**" at the end of the display? I find my concern drifting somewhat.

What is the single amino acid difference between the two sequences?

It has been established that there is but one base pair difference in the coding region of this cDNA and the wild type genomic sequence. There can therefore be, at most, one amino acid that is different between the two protein sequences. Surely a case for a global rather than a local alignment strategy? Not that the choice should be of any real consequence with sequences that are this similar.

```
                    1        11        21        31        41        51
sp|P26367|PAX6_H MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRY
                    ********************************:***************************
          pax6 MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELPHSGARPCDISRILQVSNGCVSKILGRY
                    1        11        21        31        41        51
```

From the alignment I generated with the global option of **spin**, there is a **A → P** (**Alanine → Proline**) substitution evident. **Alanine** and **Proline** are amino acids with quite different properties, so it is reasonable to suppose that the this substitution will be significant.

What is the position of the difference?

The substitution is at residue **33**, as you will have seen reported by various sources previously.

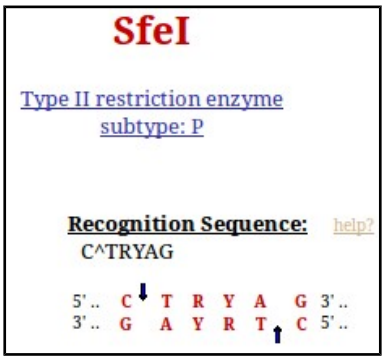
From your investigations of **Restriction Mapping**

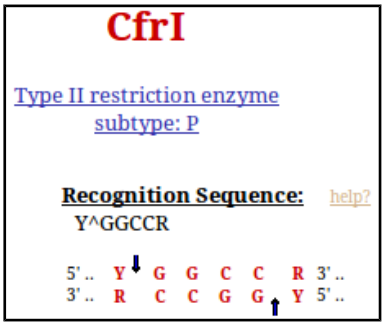
There are quite a few less enzymes mentioned in the map you have just made compared to that generated by **remap**. Can you speculate what the main reason for this might be?

Probably differences in the definition of a **6** cutter.

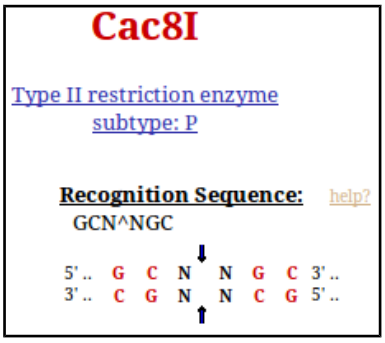
E.g. **BfuAI** has a recognition site of **ACCTGCNNNN_NNNN-**, is this a **6** base pair recognition site or **14**?

Enzymes in remap maps but not reported by nebcutter include:





6 cutters according to **remap**? Surely **R** & **Y** only count **0.5**? I agree with **nebcutter**, these are **5 cutter**s.



Oh come on **remap**!! you cannot count **N**s!! This is a **5 cutter** also.

Some enzymes in this map appear in the same place as **remap** predicted, but have different names. Can you explain why?

Restriction mapping programs, by default, only map one member of each **isoschizomer** family. There is no consistency between programs in the choice of the representative enzyme.

For example:

Commercially Available:

Enzymes	Cloned	Sequenced	Recognition Sequence	Suppliers
BseX3I	-	-	C↓GGCCG	IV
BstZI	-	-	C↓GGCCG	R
EagI	yes	yes	C↓GGCCG	N
EclXI	-	-	C↓GGCCG	MS
Eco52I	-	-	C↓GGCCG	FK

Count: 5

Not Commercially Available:

Enzymes	Cloned	Sequenced	Recognition Sequence
AaaI	-	-	C↓GGCCG
BsoDI	-	-	CGGCCG
SenPT16I	-	-	C↓GGCCG
TauII	-	-	CGGCCG
Tsp504I	-	-	CGGCCG
XmaIII	-	-	C↓GGCCG

Count: 6

remap chooses to show just **XmaIII**

nebcutter chooses **EagI**, probably because it is commercially available.

They both have the same recognition site and so would both cut at the indicated position.

From your investigations of **Searching for sequence similarities in databases**

When would **Mask lower case letters** be a useful thing to do?

Generally, whenever one might suspect the automatic masking algorithms of **blast** might miss a non informative region in a specific query sequence, obviously.

A specific example might be when a query sequence contained a significant informative region that was known to be common amongst the sequences being searched. If this region was left unmasked, **blast** would pick up so many similar matches to this one region that other interesting similarities might be obscured. By manually masking such a region by changing it to lower case, its matches would not be seen by **blast** and matches with other regions of the query sequence should be more apparent.

Which parameters would **blast** need to **automatically adjust** to cater for short input sequences (such as primers being tested for uniqueness), and why?

The **word size**: Clearly, if you are trying to find matches for a primer (for example) of around **20** base pairs, it would be pretty silly to use a **word size** of **28** (default for **megablast**). A **word** the same size as the primer would find only exact matches. A **word** of about **7** would allow a couple of mismatches and would probably be most generally appropriate.

The **expect score**: As good chance matches between between a short query sequence and a large database will be abundant, it would not be sensible to choose a demanding (i.e. small) **expect score** to represent the limit of significance. In particular, a primer sized query sequence of around **20** base pairs might easily exactly match more than **10** times (generally the default maximum expect score for a significant match) just by chance. After all, there are only **4** bases, a string of **20** is not that long and the databases can be huge! Typically **blast** chooses very high **expect score** cut off for short query sequences, effectively removing the **expect score** filter altogether.

Earlier versions of **blast** did not automatically adjust these parameters. When a short query sequences were selected, suitable adjustment was left to the user. Without sensible parameter adjustment, results could be greatly confusing. For example, a **21** base pair primer could easily match perfectly more than **10** times against a large DNA sequence database. **blast** is set to ignore matches that are expected to occur more than **10** times by chance. Thus even exact matches with such a small sequences would be ignored! Now automatic parameter adjustment is undertaken by **blast**, the user does not really have to think too hard. However, it does seem to be a good idea to know what **blast** is doing and why.

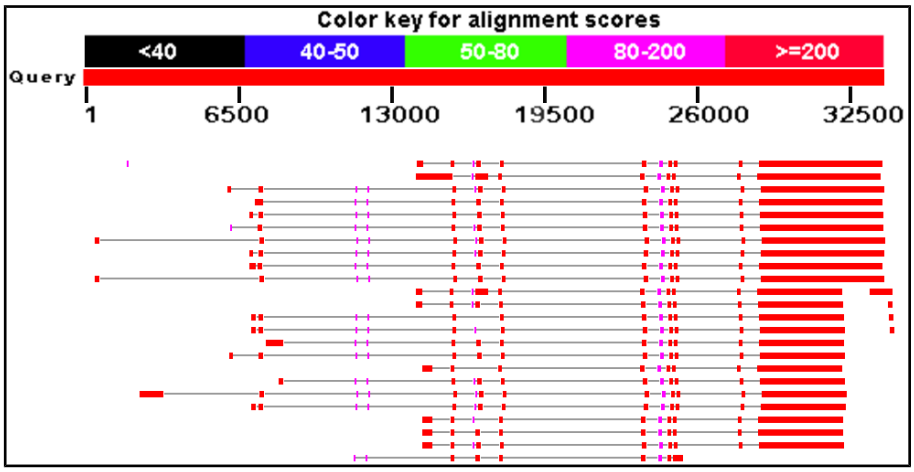
Why do you suppose that a few of the exons do not achieve the maximum score?

Summary:

Each local region of significant alignment between a database entry and a query sequence is scored independently. The scoring method that governs the alignment score colour in this graphic, reflects both the quality of the match ***and*** its length. Unless a particular region is of sufficient length, it cannot achieve the **200 bit** threshold even if the alignment is perfect. Note that is is the shorter regions that fail to reach the >=200 status. All of the illustrated local alignments associated with **PAX6** transcripts are essentially perfect.

Full Answer:

In common with most database searching programs, **blast** compares query sequences with database entries using a local strategy. The overall evaluation of a particular query sequence is taken to be the highest local score.



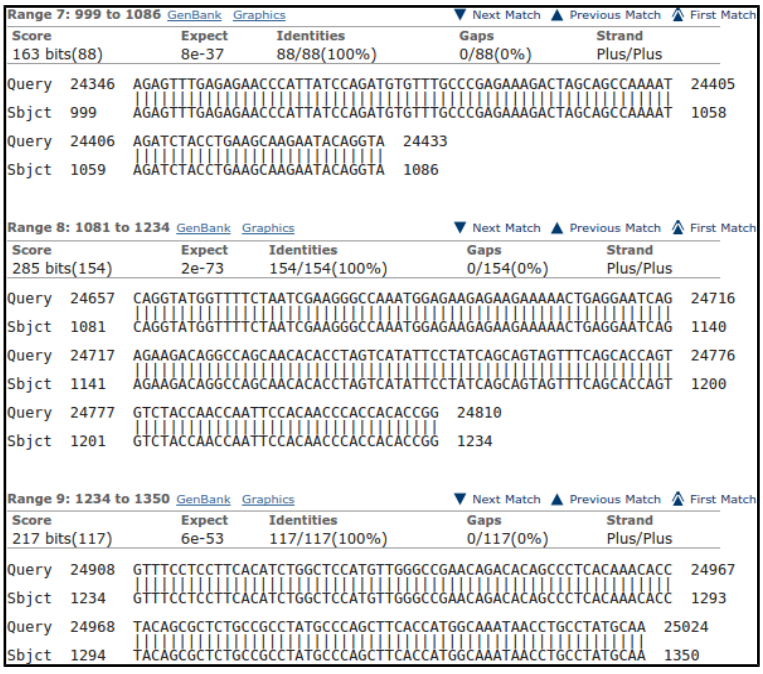
Individual local matches are coloured according to individual quality. In this query, all true matches should be perfect, or very nearly so. Scores might therefore be expected to be maximal (>=200). However, they are not? Some only manage a score in the range 80-200.

The score referenced for this purpose is the **bit score**. For a full, no holds barred definition of this score, try here. I prefer this somewhat gentler version:

> "The **bit score** gives an indication of how good the alignment is; the higher the score, the better the alignment. In general terms, this score is calculated from a formula that takes into account the alignment of similar or identical residues, as well as any gaps introduced to align the sequences. A key element in this calculation is the "substitution matrix ", which assigns a score for aligning any possible pair of residues. The **BLOSUM62** matrix is the default for most **BLAST** programs, the exceptions being **blastn** and **MegaBLAST** (programs that perform **nucleotide–nucleotide** comparisons and hence do not use protein-specific matrices). Bit scores are normalized, which means that the bit scores from different alignments can be compared, even if different scoring matrices have been used."

Still too scary? The important things to note are that:

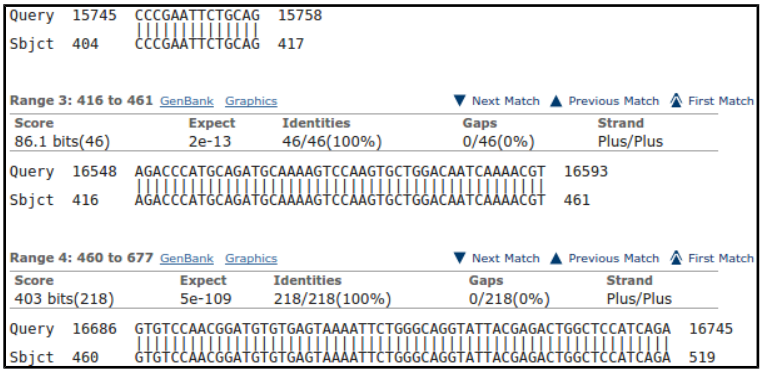
- These scores are based on a simple DNA scoring matrix (**1** for a match, **-2** for a mismatch by default for **megablast**), plus penalties for gaps. So scores will be limited by the length of the alignment, ignoring gaps.
- The scores reflect penalties for **indel**s (**in**sertions or **del**etions).
- The scores are normalised so that they do not depend on the chosen scoring matrix. This allows bits scores from searches using different scoring matrices to be compared.



This being so, **bit scores** will reflect the length of an alignment as well as its quality. If an alignment is very short, it might be perfect but still not achieve a very high value. **bit scores** are designed to reflect significance, not just local quality. A short perfect match clearly can be less significant than a longer less perfect match. That is what you see illustrated here.

You can see evidence of what is occurring in the alignments further down your results. Here is illustrated one of the 80-200 exons that occur in all transcripts at position **24,346**[145]. The match is perfect, but the length of the exon is consistently just to short to get to the heady >=200 level.

Note how imperfectly **blast** finds exon/intron boundaries. If the start of an intron happens to match the start of the next exon, **blast** will included the bases in two alignments[146]. It is not looking for exons and introns as was **spline**, it just mindlessly seeks matches.



For a further example, look at the exon that is found only in the **isoform 5a** transcripts. It is tiny (**42** base pairs) and scores well below >=200 even thought it is a perfect match.

Note that the alignment is **46** base pairs long due to **blast** adding on two bases either side that are actually the highly conserved intron start and end base pairs. As you can see, these extra base pairs occur in the preceding and succeeding alignment also.

Explain why one exon in the reasonably consistent region, does not appear in all of the transcript matches?

Oh dear oh dear! Not this again.

Well I refer to the **isoform 5a** exon, of course. The tiny inconsistent one about **9** exons in from the right (when it exists). This will, clearly, only occur in **isoform 5a** transcripts. One day I will tidy these questions up a trifle!

Why were you not surprised to discover **24 PAX6** transcripts in **Refseq** matching this sequence?

Repetitive? True. This question is more "interesting" when the resources you have visited do not agree. That is, most of the time. Whilst the situation is in balance, the answer is:

Because **GeneCards** says there should be **11** quality and a further **13** less supported **PAX6 mRNA** sequences in **RefSeq**. A total of **24 PAX6** implied transcripts in total. In passing, you could have discovered the number of **PAX6 mRNA** sequences in **RefSeq** but asking **RefSeq** directly. Probably a more sensible and certainly a more reliable approach.

145 In order to make this illustration, I needed set **Sort by**: (top of the alignments) to **Query start position**.

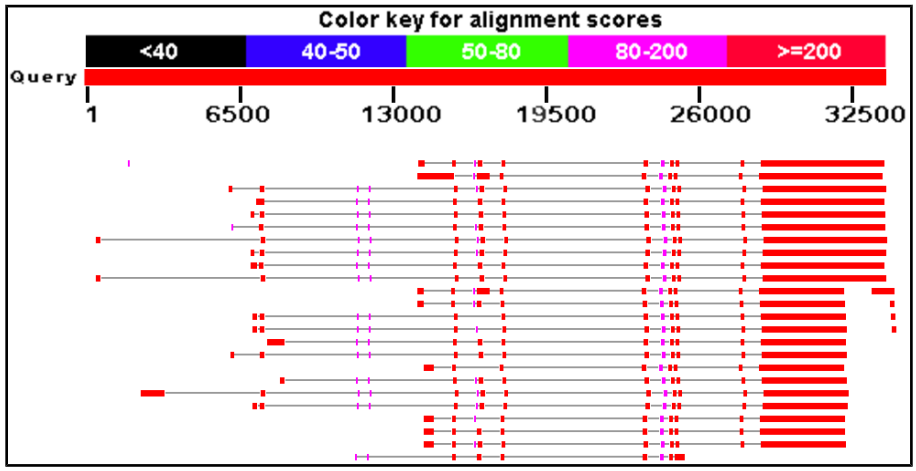
146 **6** base pairs (**Sbjct**: **1081-1086**, **CAGGTA**) occur in both the first two matches illustrated. Just **1** base pair is shared between the **2nd** and **3rd** match (**Sbjct**: **1234**, **G**).

Which of the **Refseq PAX6** transcripts corresponds to **isoform 5a**?

Summary:

As I am sure you are tired of noting by now, all the transcripts with the extra tiny exon around position **1,600** in the genomic sequence are **isoform 5a** transcripts.

Full Answer:



The **isoform 5a** transcripts can be spotted most easily from the graphic. They are the ones with the extra small exon slightly to the left of middle (around base position **1,600**). For example, the **first**, **second** and **third blast** matches displayed. If you hover over these matches with your mouse, you will see that they are t**ranscript variants 1**, **2**, **3**, **6**, **7**, **8**, **11**, **12**, **14**, **18**, **19**, **20**, **21** and **23** (in the vertical order of the graphic).

Stated with the unequalled poetry of **RefSeq Accession Code** and lyrical **Title** Line, that becomes:

TITLE	ACCESSION CODE
Homo sapiens paired box 6 (PAX6), transcript variant 11, mRNA	NM_001310161.1
Homo sapiens paired box 6 (PAX6), transcript variant 10, mRNA	NM_001310160.1
Homo sapiens paired box 6 (PAX6), transcript variant 8, mRNA	NM_001310158.1
Homo sapiens paired box 6 (PAX6), transcript variant 5, mRNA	NM_001258463.1
Homo sapiens paired box 6 (PAX6), transcript variant 4, mRNA	NM_001258462.1
Homo sapiens paired box 6 (PAX6), transcript variant 2, mRNA	NM_001604.5
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X13, mRNA	XM_005252958.3
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X12, mRNA	XM_011520153.1
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X10, mRNA	XM_011520152.1
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X6, mRNA	XM_011520150.1
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X5, mRNA	XM_011520149.1
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X4, mRNA	XM_005252954.3
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X3, mRNA	XM_011520148.1
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X1, mRNA	XM_011520146.1

Yes well, that was fun? When I first wrote this question, there was only **5** or **6 RefSeq** mRNAs! The message of the question was to ensure you could see how to spot the **isoform 5a** transcripts (again!), not to list them! But, never mind, doing so was in fine tune with the ennui of the moment.

Additional Meanderings:

That really should not detain anyone but me? They belong after the consideration of masking the run of As for the **megablast** you ran. I just enjoyed this detour, so I keep it somewhere low profile. The whole journey leads nowhere of any note, so, should you decide to read, expect little!

The **500** base pairs of 3' flanking sequence added on to the **Ensembl** sequence for "good measure", is not part of the alignment (as would be expected). This can be seen easily if you look at the end of the alignment illustrated above, which is the alignment of the last exon of a transcript.

The entire length of this transcript is **6,732** base pairs.

Homo sapiens paired box 6 (PAX6), transcript variant 11, mRNA
Sequence ID: ref|NM_001310161.1| Length: 6732 Number of Matches: 11

The entire length of the genomic query sequence is **34,170** base pairs.

Query ID	lcl|Query_71179
Description	pax6-genomic sequence
Molecule type	nucleic acid
Query Length	34170

The alignment ends at position position **6,729** of the mRNA (**3** from the end) and position **33,670** of the genomic sequence (exactly **500** base pairs from the end).

```
Query  33601  ATTTGACATCCTGGCAAATCACTGTCATTGATTCAATTATTCTAATTCTGAATAAAAGCT  33660
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  6660   ATTTGACATCCTGGCAAATCACTGTCATTGATTCAATTATTCTAATTCTGAATAAAAGCT  6719

Query  33661  GTATACAGTA  33670
              ||||||||||
Sbjct  6720   GTATACAGTA  6729
```

The **3** missing base pairs of the mRNA are all **A**s, due to **polyadenylation**. Position **6,729** being recorded as a **polyA** site by **RefSeq**. A very short **polyA** tail surely? But there is no telling what stage of the mRNA is recorded on **RefSeq**. **Wikipedia** says:

```
polyA_site      2495
                /gene="PAX6"
                /gene_synonym="AN; AN2; D11S812E; FVH1; MGDA; WAGR"
```

```
    6541 aaaaaaatag aataagaaac ctgattttta gtactaatga aatagcgggt gacaaaatag
    6601 ttgtcttttt gattttgatc acaaaaaata aactggtagt gacaggatat gatggagaga
    6661 tttgacatcc tggcaaatca ctgtcattga ttcaattatt ctaattctga ataaaagctg
    6721 tatacagtaa aa
//
```

> "The tail is shortened over time, and, when it is short enough, the mRNA is enzymatically degraded."

Of course, the neatness of this observation does reflect less some profund biological truth than it does that this mRNA just happens to be one that extends furthest to the right in the genome, and there is no chance match between the **polyA** tail and the extra **500** bases of genomic sequence you added on when extracting it from **Ensembl**.

The journey was fun even though the destination was dubious. Much the way of a considerable portion of life in general one might reflect?

In what circumstances would you imagine that the **Max matches in a query range** parameter might be set to something other than its default value of **0**?

The default value for this parameter, **0**, means that there be no limit to the number of matches listed with any given region of the query sequence. This can mislead when one region is prolifically and strongly conserved in the database being searched. In such a case, it is possible that so many matches with one region are found that there is not space in the list of hits for other, weaker but significant, matches with other regions of the query sequence.

This option allows a user to say:

> "Once you have found **50** (say) similar matches with a single part of my query, I ***have got the message!!***, list no more matches unless they pertain to a different part of my query and thus tell me something new."

All this you can convey by just changing that **0** to **50**!! Are not computers wonderful?

Note of honesty: Whilst I believe all the above, I have yet to make this option work satisfactorily. I think I know what might be amiss, but I spare you the details. Inquiry with **NCBI** initiated **2016.03.29**.

What are the **9** stronger matches around base position **16,000**?

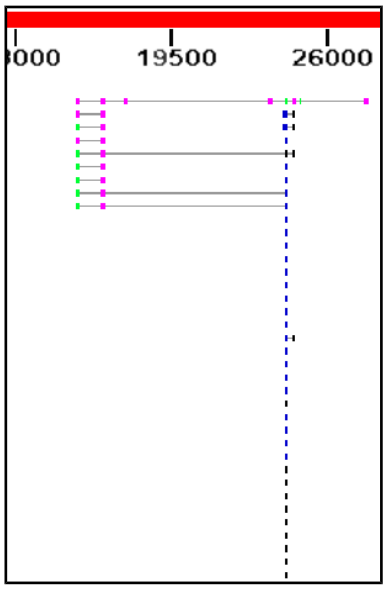
Matches between the regions of genomic DNA encoding **Paired Box** domains.

Why would you expect exactly **9** matches around this point?

Because that is how many **Paired box** domains are suggested to be in the human genome by counting the number of quality **mRNA** sequences in **RefSeq** claiming to include a **Paired box** coding region. There is **PAX6** plus its **8** paralogues, imaginatively all named:

PAX1, **PAX2**, **PAX3**, **PAX4**, **PAX5**, **PAX6**, **PAX7**, **PAX8** & **PAX9**

What do you make of the plethora of matches around **24,000**?



These are matches between human **mRNA** sequences with the regions of genomic DNA encoding **Homeo Box** domains. As you discovered from **Interpro**, there are many of these.

The thin line joining features implies that those features relate to the same database entry.

Notice that **4** of the **9** proteins matching a **Paired box** genomic region also match a **Homeo box** region. the remaining **5** do not. This implies that **4** of the **9** proteins corresponding to the hits detected here have a **Paired box** domain near the start of the protein and a **Homeo box** domain further along. This is exactly as was suggested by the **PROSITE** annotation you examined.

Why do you suppose the **Paired box** matches precede the **Homeobox** matches?

Because they score more highly and so, in the opinion of **blast**, are more worthy. Primarily, they score more highly because they are longer. The list is ranked by **E Value**. Good matches with long sequence are less likely to occur by chance than equally good matches with shorter sequences.

Possibly a more interesting question[147] might have been: "Why are not all the hits which include both domains at the top of the list?". Surely they should be, as they match over a longer proportion of the query sequence and so must, in general at least, be of the greatest significance.

They do not always come at the top of the list because **blast** scores each matching region individually and uses the ranking scores associated with the single region with the highest **E Value** to evaluate the similarity of the entire database entry with the query. This has to be a dubious practice surely? But, it appears to work, so why complain.

To justify this last assertion, Look at your top hit.

Description	Max score	Total score	Query cover	E value	Ident	Accession
RecName: Full=Paired box protein Pax-6; AltName: Full=Aniridia type II protein; AltName: Full=Oculorhombin	160	767	3%	2e-40	97%	P26367.2

E Val = **2e-40**, **Max score** = **160**, **Total score 767** associated with the whole of **P26367.2**

RecName: Full=Paired box protein Pax-6; AltName: Full=Aniridia type II protein; AltName: Full=Oculorhombin
Sequence ID: sp|P26367.2|PAX6_HUMAN Length: 422 Number of Matches: 8

Range 1: 46 to 123 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
160 bits(406)	2e-40	Compositional matrix adjust.	76/78(97%)	78/78(100%)	0/78(0%)	+3

```
Query  16680  MQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDR  16859
              +QVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDR
Sbjct  46     LQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDR  105

Query  16860  LLSEGVCTNDNIPSVSSL  16913
              LLSEGVCTNDNIPSVSS+
Sbjct  106    LLSEGVCTNDNIPSVSSI  123
```

Range 2: 254 to 305 GenPept Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
81.3 bits(199)	5e-29	Compositional matrix adjust.	51/52(98%)	51/52(98%)	0/52(0%)	+3

```
Query  24654  FQVWFSNRRAKWRREEKLRNQRRQASNtpshipisssfstsVYQPIPQPTTP  24809
               QVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTP
Sbjct  254    IQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTP  305
```

Range 3: 312 to 344 GenPept Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
70.5 bits(171)	5e-29	Compositional matrix adjust.	33/33(100%)	33/33(100%)	0/33(0%)	+2

```
Query  24926  GSMLGRTDTALTNTYSALPPMPSFTMANNLPMQ  25024
              GSMLGRTDTALTNTYSALPPMPSFTMANNLPMQ
Sbjct  312    GSMLGRTDTALTNTYSALPPMPSFTMANNLPMQ  344
```

Now look at the first few individual regional alignments for this hit.

As you can see, the **E Value** and **Max score** values used to evaluate the whole protein were computed from just the best (ranked by **E Value**) local alignment! Crude, but never mind.

The **Total score** for the entire protein is the sum (rounded up to the nearest integer) of all the bit scores for all **8** local alignments computed for this protein (I suggest you just trust me on this assertion).

147 That I did not ask, because I only just thought of it.

How do you suppose the **Max matches in a query range** parameter might be of value if this order was reversed?

If **Paired boxes** had been more prolific, then the number of **Paired box** matches might have filled the **blast** hit list before the highest scoring **Homeo box** hit was registered.

If **Homeo boxes** were longer, and so justified a better **E value**, then the number of **Homeo box** matches might have filled the **blast** hit list before the highest scoring **Paired box** hit was registered.

Either of these situations would be very unfortunate, but easily avoided by setting the **Max matches in a query range** parameter to something sensible (**50** say). This would ensure that only the top **50** items in the **blast** hit list would be dominated by the strongest hit.

For further discussion of the parameter, see above.

How does this "non-informative" region match expectations suggested by **Prosite** and the **Feature table** of **Uniprot** for **PAX6_HUMAN**?

blast identifies two non-informative regions. I only discussed the prettiest one above. The region discussed is comprised largely of **Serines**, **Prolines**, **Threonines** & **Isoleucines** the **15** residues between **294-308**.

Score	Expect	Method	Identities	Positives	Gaps	Frame
81.3 bits(199)	5e-29	Compositional matrix adjust.	51/52(98%)	51/52(98%)	0/52(0%)	+3

```
Query  24654  FQVWFSNRRAKWRREEKLRNQRRQASNtpshipisssfstsVYQPIPQPTTP  24809
              QVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTP
Sbjct  254    IQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTP  305
```

The second (to be found much further down your **blast Alignments** output) is comprised entirely of **Arginines**, **Luecines** and **Lysines** and **Glutamines**, the **10** residues between **203 - 212**.

Score	Expect	Method	Identities	Positives	Gaps	Frame
85.9 bits(211)	3e-16	Compositional matrix adjust.	56/66(85%)	58/66(87%)	5/66(7%)	+3

```
Query  23649  YHPILFVP-----DGCQQQEGGGENTNSISSNGEDSDEAQMrlqlkrklqrNRTSFTQEQ  23813
              ++P   VP     DGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQ
Sbjct  162    WYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQ  221

Query  23814  IEALEK  23831
              IEALEK
Sbjct  222    IEALEK  227
```

Uniprotkb also suggests there are two **compositionally biased regions**.

Compositional bias	131 – 209	79	Gln/Gly-rich
Compositional bias	279 – 422	144	Pro/Ser/Thr-rich

Well, hardly an exact match, but there is approximate agreement? One would certainly suppose that **blast** is only willing to mask fairly severe cases of **compositional bias**. It is also probable that **blast** has a rather more mechanistic (i.e. non-biological) interpretation of what **computational bias** is?

PROSITE also predicts the more obvious region of **computational bias**, rather more generally:

> "An octapeptide and/or a homeodomain can occur C-terminal to the paired domain, as well as a Pro-Ser-Thr-rich C-terminus"

From your investigations of **Primer Design**

Do you think **10** primer pair suggestions is sufficient? If not, what number would you choose?

Until very recently, the default here was **5**. That seemed rather low to me. I included this question to solicit opinion rather than to impart knowledge. A default of **10** seems more in line with my instincts, but people who use this program seriously mostly tell me that they can select suitable primers from the first **2** or **3** suggestions of the program. So, **5** would seem a good choice and **10** would be moving towards cautiously overdoing things.

On the whole, informed opinion suggests that **10** suggestions will be more than enough in most circumstances.

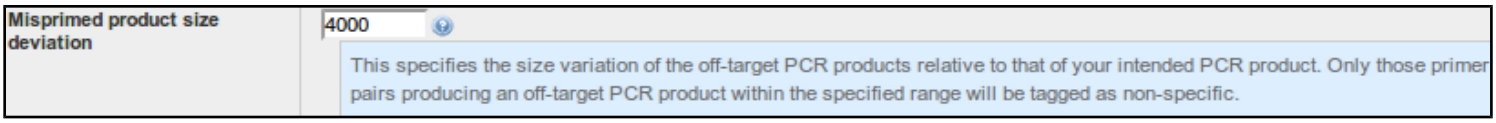
What value would you choose here if you were looking for uncluttered results?

Summary:

Clearly, the smaller the number chosen, the shorter will be the list of spurious products. However, pick something too small and you risk including unintended product(s) that could cause confusion. The size selected must be sufficient that larger unwanted PCR product(s) could easily be spotted by other means (simply by size?).

Full Answer:

Well, mostly for me, and just in case you were curious, when I first wrote the question, the parameter was very different and not so easy to understand. Pure self indulgence, I know, but here is the history. The parameter explained itself, via the button, thus:



I interpreted this to mean that only **blast** predicted products of up to **X+4,000** base pairs, where **X** base pairs is the length of the intended target, will be given any regard. It is thus assumed that a difference of **4,000** base pairs between an intended PCR product (predicted by **primer3**) and a spurious product (detected by **blast**) can easily be detected simply by size difference.

Of course this parameter also will reject unwanted **blast** predicted products that are less than **Y–4,000** base pairs, where **Y** base pairs is the length of the intended target, will be given any regard. Given the largest possible **primer3** suggestion will be **1,000** base pairs (the form setting for the exercise specifies products of between **100**[148] and **1,000** base pairs), this is hardly an issue here.

Comment upon the small default value for the **Blast word size**?

By default, **blast** will be looking for aligned exactly matching blocks of **7** nucleotides when identifying where a primer might match a database entry. The entire primer match with the template sequence does not have to be exact for the primer to be acceptable. The entire primer is typically only around **20** bases long. And word size much more that **7** would clearly miss too much to be effective.

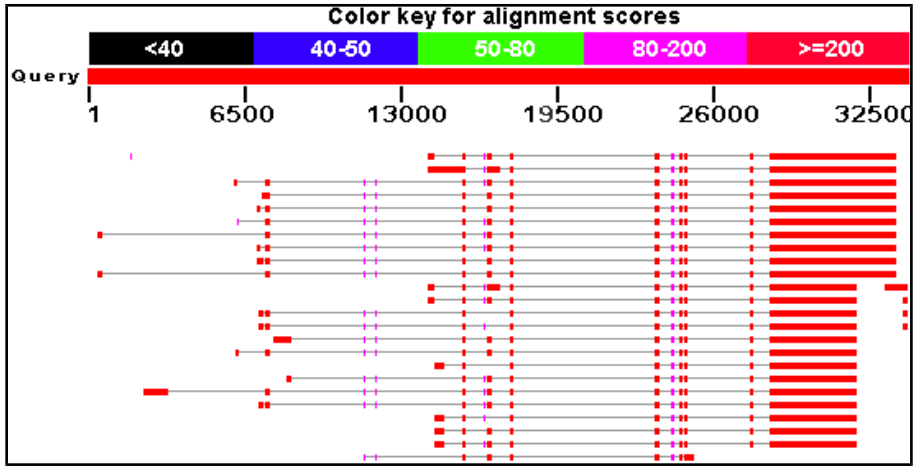
148 The form explicitly declares a minimum of **70**, but the ranges from which the **forward** & **reverse** primers must come (**15000-15700** & **15800-16500**) make the smallest possible **primer3** prediction **100** base pairs long.

Why do you suppose **blast** did not pick up all the transcripts?

Summary:

Well, the simple answer is that the **PREDICTED** transcripts that were not detected as unwanted products cannot include either the forward primer, or the reverse primer, or both. This is, almost, the only possible explanation.

Full Answer:



Of course, for this run, you did specify that you were not interested in products longer than **4,000** base pairs, so it could be that one or more products were possible but longer than that? I suspect this would only be feasible if there were retained introns involved, but previous **blast** results do no suggest this to be the case. I would say the only possible candidate for an over-length product might be the second hit down in the graphical representation generated previously by **blast**. The first and third exons from the left look a bit bloated, but not really sufficiently to cause a problem.

Description	Max score	Total score	Query cover	E value	Ident	Accession
Homo sapiens paired box 6 (PAX6), transcript variant 11, mRNA	9659	12484	19%	0.0	100%	NM_001310161.1
Homo sapiens paired box 6 (PAX6), transcript variant 10, mRNA	9659	15161	24%	0.0	100%	NM_001310160.1
Homo sapiens paired box 6 (PAX6), transcript variant 8, mRNA	9659	12929	20%	0.0	100%	NM_001310158.1
Homo sapiens paired box 6 (PAX6), transcript variant 7, mRNA	9659	12729	20%	0.0	100%	NM_001258465.1
Homo sapiens paired box 6 (PAX6), transcript variant 6, mRNA	9659	12761	20%	0.0	100%	NM_001258464.1
Homo sapiens paired box 6 (PAX6), transcript variant 5, mRNA	9659	12737	20%	0.0	100%	NM_001258463.1
Homo sapiens paired box 6 (PAX6), transcript variant 4, mRNA	9659	12862	20%	0.0	100%	NM_001258462.1
Homo sapiens paired box 6 (PAX6), transcript variant 2, mRNA	9659	12833	20%	0.0	100%	NM_001604.5
Homo sapiens paired box 6 (PAX6), transcript variant 1, mRNA	9659	12942	20%	0.0	100%	NM_000280.4
Homo sapiens paired box 6 (PAX6), transcript variant 3, mRNA	9659	12791	20%	0.0	100%	NM_001127612.1
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X13, mRNA	6613	10063	15%	0.0	100%	XM_005252958.3
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X12, mRNA	6613	9439	14%	0.0	100%	XM_011520153.1
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X11, mRNA	6613	9329	14%	0.0	100%	XM_006718246.2
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X10, mRNA	6613	9410	14%	0.0	100%	XM_011520152.1
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X9, mRNA	6613	10507	16%	0.0	100%	XM_005252956.3
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X8, mRNA	6613	9783	15%	0.0	100%	XM_005252955.3
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X7, mRNA	6613	9091	14%	0.0	100%	XM_011520151.1
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X6, mRNA	6613	9637	15%	0.0	100%	XM_011520150.1
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X5, mRNA	6613	11324	17%	0.0	100%	XM_011520149.1
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X4, mRNA	6613	9814	15%	0.0	100%	XM_005252954.3
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X3, mRNA	6613	9172	14%	0.0	100%	XM_011520148.1
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X2, mRNA	6613	9502	15%	0.0	100%	XM_011520147.1
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X1, mRNA	6613	9576	15%	0.0	100%	XM_011520146.1
PREDICTED: Homo sapiens elongator acetyltransferase complex subunit 4 (ELP4), transc	1775	1775	2%	0.0	100%	XM_005252865.2
Homo sapiens paired box 6 (PAX6), transcript variant 9, mRNA	647	2630	4%	0.0	100%	NM_001310159.1

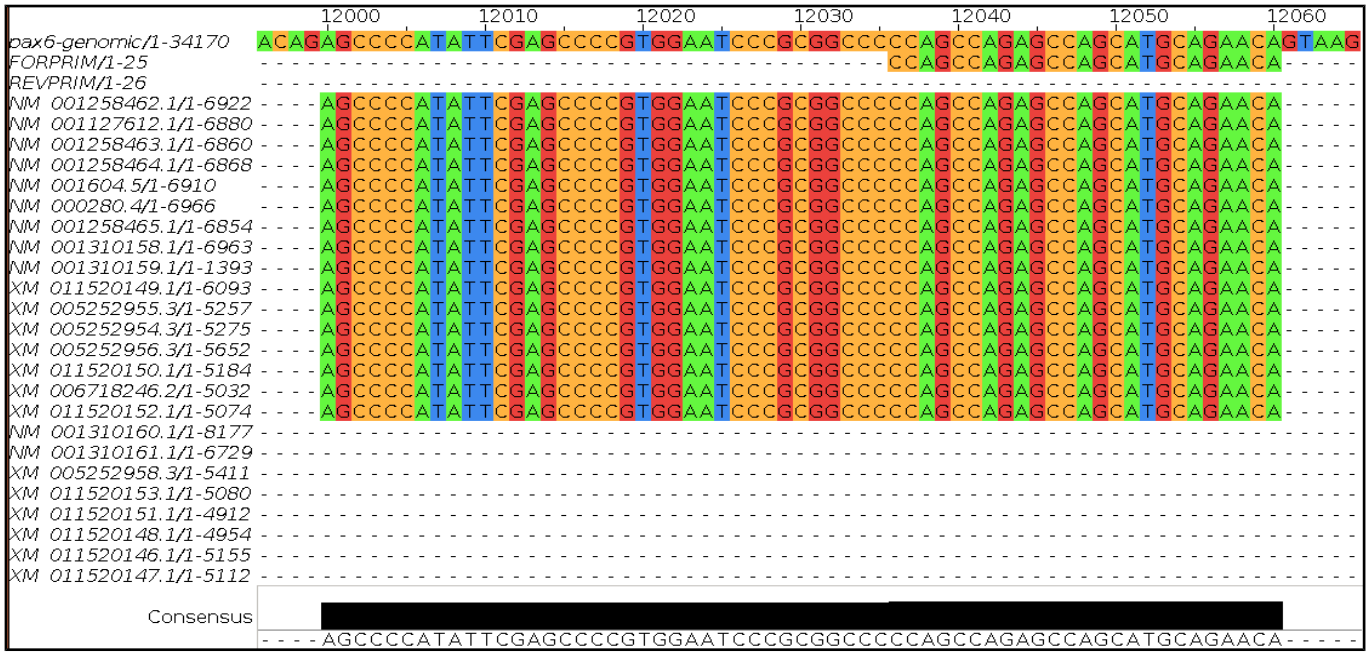
I might also be that unwanted PCR products are eliminated/introduced due to variations in the predicted transcripts. However, this can be ruled out as previous experiments, **blast** assures us that all **24** potential transcripts match the genomic sequence exactly.

Enough! Only because I want to, I will compute the alignments to prove the missing primer matches. Read no further unless you are truly in the mood. Much of the reason for recording the rest of this answer is that, apart from enjoying the pursuit of irrelevant detail, I also wanted to remember how I made the alignments and certainly feel I could have made both these, and my point much more simply? Suggestions welcome.

OK, I started by computing an alignment that was a mapping of all **24** transcripts onto the **PAX6** genomic regions as represented in the file **pax6_genomic.fasta**. I used a program called **gmap**, which like **spline** (used in the exercise) is designed to align cDNA/mRNA sequences with corresponding genomic sequences. The version of **gmap** I used runs under **linux** from the command line. It has the advantage over **spline** that is will align more than one cDNA/mRNA sequence against the genome in one run. Unfortunately, it does not generate an output format that can be easily displayed in the way I required here. I did try to persuade a couple of general multiple alignment programs (**clustalw** & **muscle**) to make me a usable alignment, but ran into the same difficulties we experienced in the exercise. I failed to find gap penalties that would get the programs to gap the larger introns. Even if I had succeeded to get the gaps in the right place, I would not have believed them to be placed with sufficient accuracy for the same reasons this was not possible when we tried the same trick with general alignment software for just one cDNA sequence against the genome in the exercise.

So, I made a rough alignment with **clustalw** and edited it to exactly what was suggested by **gmap** using **jalview**. This took **HOURS**. There has to be a better way!! You have already used all the software mentioned except **gmap** and **clustalw**. You will use **clustalw** and see how **jalview** can be use to edit, as well as just view, alignments a little later.

All that effort to show that the region around the forward primer looks like this:

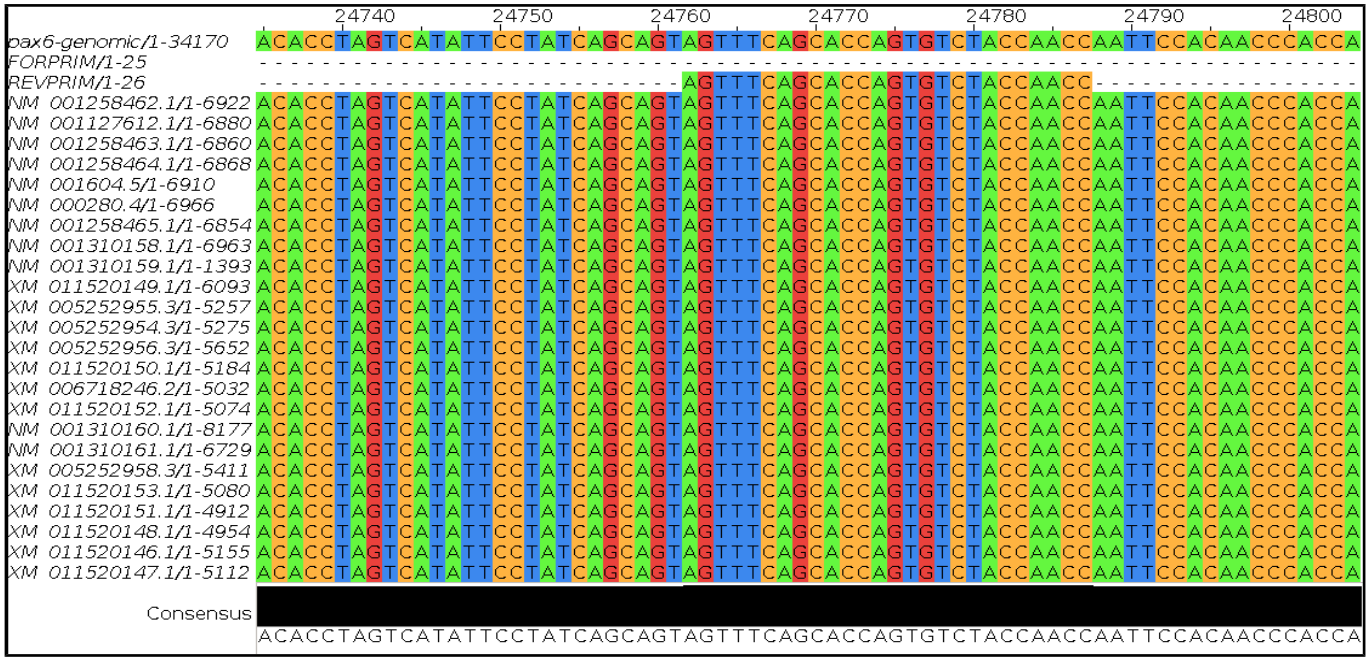


Showing clearly that the **8** transcripts:

NM_001310160.1
NM_001310161.1
XM_005252958.3
XM_011520153.1
XM_011520151.1
XM_011520148.1
XM_011520146.1
XM_011520147.1

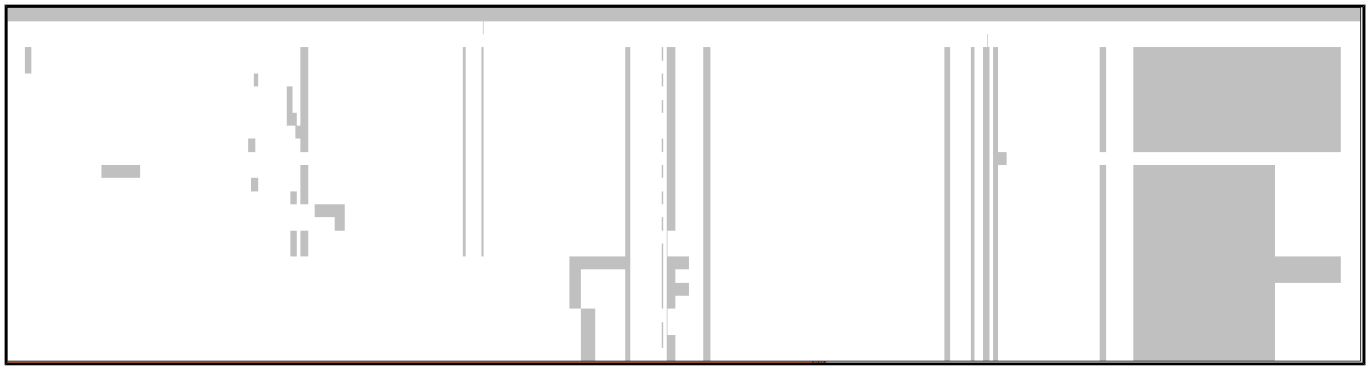
Have the exon that includes the forward primer spliced out and so will not produce any PCR product. Feel free to check this by comparing the textual results of your **blast** of the genomic sequence against the **RefSeq** mRNAs and the results of **PRIMER-BLAST**. I did, it was lots and lots of fun and I ended up content that all was logically consistent.

The alignment around the reverse primer looks like this:



All **24** putative transcripts match the reverse primer perfectly. So blast should indeed find **16** of the **24** transcripts sequences in **RefSeq**, and it does.

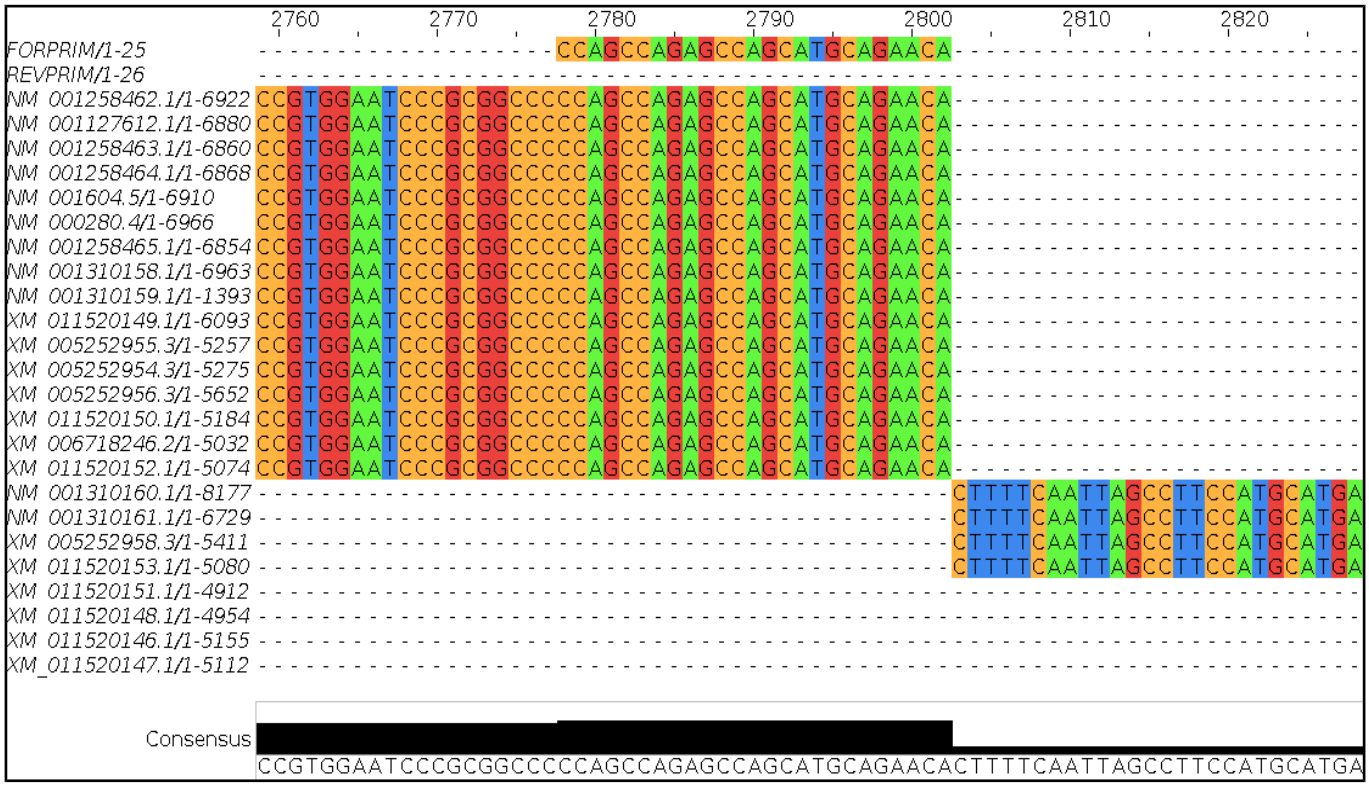
Jalview offers an overview of the entire alignment. The top row shows the genomic sequence. The second row shows the position of the forward primer. The third row shows the position of the reverse primer.



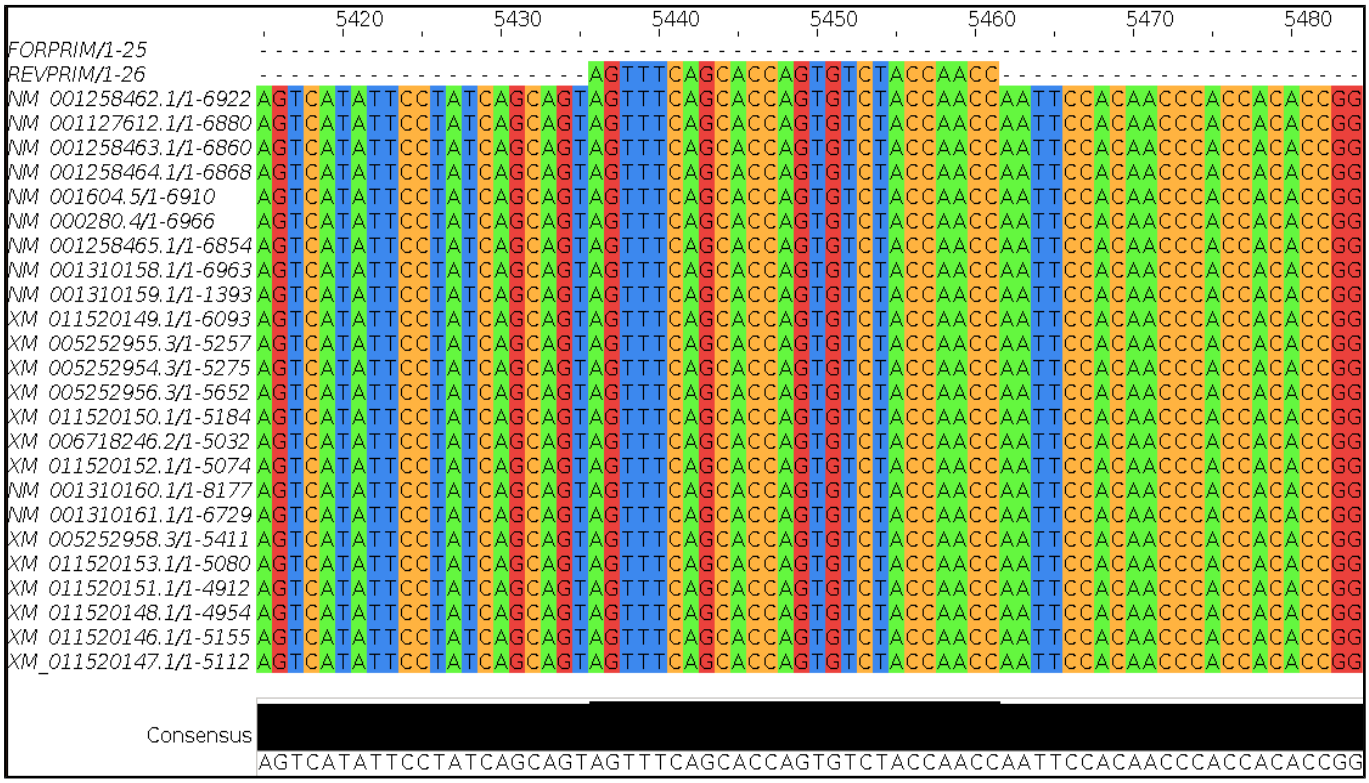
Except for the order of the transcripts, this view is very similar to the overview graphic generated by **blast**. The transcripts missing the forward primer, which isoform each transcripts represents and the fact that all transcripts match the reverse primer should be very clear.

Finished Dave? Well no, not quite. I wondered why I had included the genomic sequence in my alignment. Finding no answer to that question, I tried to make an alignment of just the primer sequences and the mRNAs. I thought this would be easy. I was wrong. The general programs are still going to get the gaps wrong whatever penalties are used. Some transcripts have exons entirely missing in all other transcripts leaving no clues as to which way round they should be aligned. The scaffold provided by the genomic sequence was essential. So, I made a mRNA only alignment by editing the alignment discussed above with **jalview**. This was easy (although you would not think so given the time it took me to work out how to do it!). I loaded the alignment into **jalview**, deleted the genomic sequence and then removed all empty columns (that is, all columns with no bases in them due the the removal of the genomic sequence). Clever eh? Just because it is there, here are the pictures.

Forward primer region (the primer is right at the end of an exon):

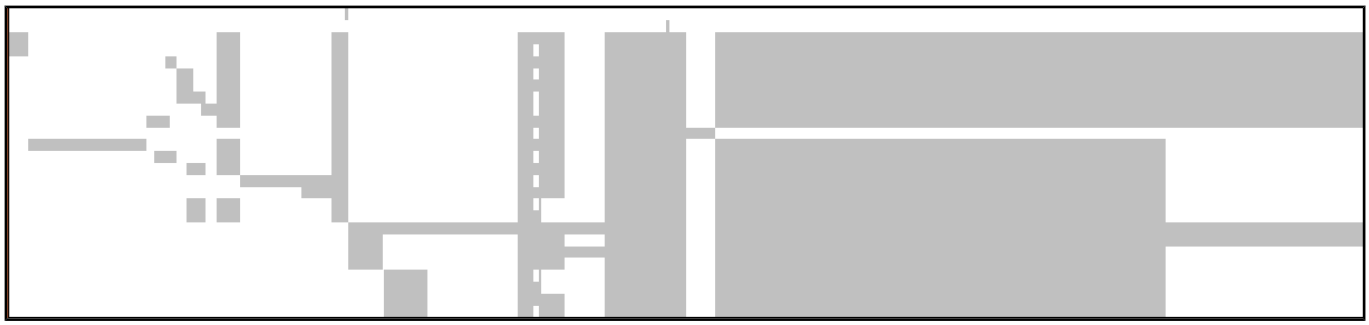


Reverse primer region:



Overview:

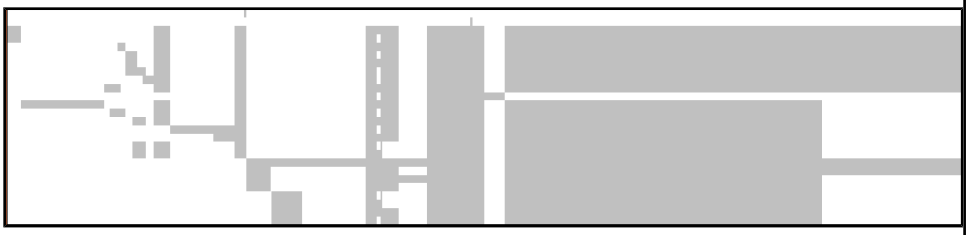
Without the evidence of the genomic sequence, the two leftmost exons could logically swap position. There is no transcript that includes both these exons and no overlap between either and any other exon in any transcript (most clearly verified from the previous **Overview** plot). Thus, there is no exon evidence of the order in which the two should appear.



Now I am done! This has to be the most over the top answer yet, but at least it kept me out of trouble for a while.

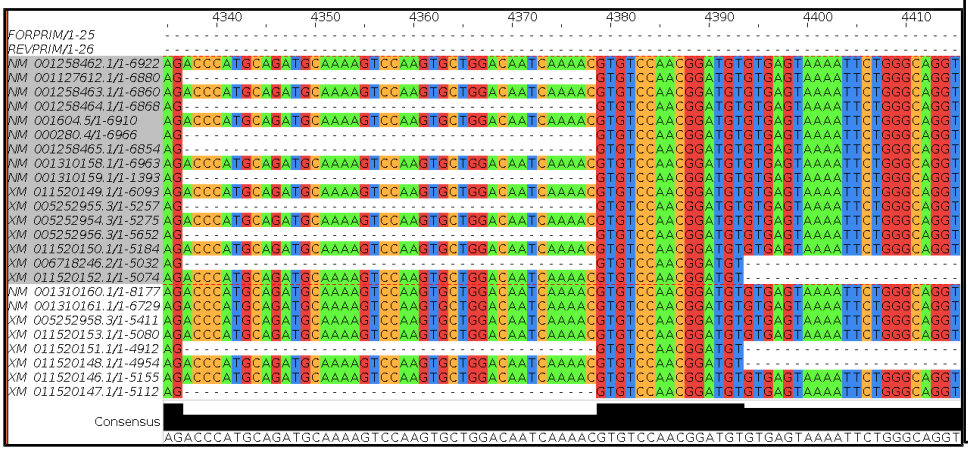
How would you tell quickly which isoform was represented by each mRNA listed here?

From the illustrations of the last answer (in particular, the **jalview** overviews), it is clear that all the mRNAs that produce a product include the region that determines which isoform is represented. That is, all are one isoform or the other.



The last two of the mRNAs that produce a product, have a bit chewed out just after the isoform defining region (an exon spliced out, if you prefer). It is logical to suppose these would be the mRNAs from which the two shorter products were generated.

Indeed, looking at the relevant part of the mRNA only alignment shows them to be **XM_006718246** (product length **707**, excluding the **isoform 5a** exon that suggest it codes for a **canonical** protein) and **XM_011520152** (product length **749**, including the extra **42** base pairs suggesting it codes for an **isoform 5a** protein).



Products on potentially unintended templates

>NM_001310159.1 Homo sapiens paired box 6 (PAX6), transcript variant 9, mRNA

```
product length = 908
Forward primer  1     CCAGCCAGAGCCAGCATGCAGAACA  25
Template        114   .........................  138

Reverse primer  1     GGTTGGTAGACACTGGTGCTGAAACT  26
Template        1021  ..........................  996
```

>NM_001310158.1 Homo sapiens paired box 6 (PAX6), transcript variant 8, mRNA

```
product length = 950
Forward primer  1     CCAGCCAGAGCCAGCATGCAGAACA  25
Template        496   .........................  520

Reverse primer  1     GGTTGGTAGACACTGGTGCTGAAACT  26
Template        1445  ..........................  1420
```

>XM_006718246.2 PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X11, mRNA

```
product length = 707
Forward primer  1     CCAGCCAGAGCCAGCATGCAGAACA  25
Template        457   .........................  481

Reverse primer  1     GGTTGGTAGACACTGGTGCTGAAACT  26
Template        1163  ..........................  1138
```

>XM_011520152.1 PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X10, mRNA

```
product length = 749
Forward primer  1     CCAGCCAGAGCCAGCATGCAGAACA  25
Template        457   .........................  481

Reverse primer  1     GGTTGGTAGACACTGGTGCTGAAACT  26
Template        1205  ..........................  1180
```

>XM_005252956.3 PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X9, mRNA

```
product length = 908
Forward primer  1     CCAGCCAGAGCCAGCATGCAGAACA  25
Template        876   .........................  900

Reverse primer  1     GGTTGGTAGACACTGGTGCTGAAACT  26
Template        1783  ..........................  1758
```

>XM_005252955.3 PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X8, mRNA

```
product length = 908
Forward primer  1     CCAGCCAGAGCCAGCATGCAGAACA  25
Template        481   .........................  505

Reverse primer  1     GGTTGGTAGACACTGGTGCTGAAACT  26
Template        1388  ..........................  1363
```

>XM_011520150.1 PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X6, mRNA

```
product length = 950
Forward primer  1     CCAGCCAGAGCCAGCATGCAGAACA  25
Template        366   .........................  390

Reverse primer  1     GGTTGGTAGACACTGGTGCTGAAACT  26
Template        1315  ..........................  1290
```

>XM_011520149.1 PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X5, mRNA

```
product length = 950
Forward primer  1     CCAGCCAGAGCCAGCATGCAGAACA  25
Template        1275  .........................  1299

Reverse primer  1     GGTTGGTAGACACTGGTGCTGAAACT  26
Template        2224  ..........................  2199
```

>XM_005252954.3 PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X4, mRNA

```
product length = 950
Forward primer  1     CCAGCCAGAGCCAGCATGCAGAACA  25
Template        457   .........................  481

Reverse primer  1     GGTTGGTAGACACTGGTGCTGAAACT  26
Template        1406  ..........................  1381
```

All other transcripts that generate PCR products generate products of length either **908** or **950**. Given the difference (**42** base pairs) is exactly the size of the **isoform 5a** exon, it is reasonable to assume the transcripts generating PCR products of length **908** represent **canonical** proteins, whereas the transcripts generating PCR products of length **950** represent **isoform 5a** proteins.

Is the number of "**potentially unintended products**" as you would you expect, given the evidence from **GeneCards**, **Ensembl** and **blast**?

Yes, I think so, given you accept my investigation (see above) as to why there were only **16** "**potentially unintended products**" when you might have expected **24**, given your **blast** results. Once **GeneCards** catches up with **RefSeq**, it too will encourage an initial expectation of **24** "**potentially unintended products**". **Ensembl** only uses the higher quality **RefSeq** mRNAs. Currently, **Ensembl** implies there to be **7** good quality **RefSeq** mRNAs. I have faith that **Ensembl** will use all **11** when it is updated next.

For all the "**potentially unintended products**", the selected primers match exactly. Can you explain this?

Well, of course they do??? All the transcripts found are generated from the same region of genomic DNA and therefore will be identical in all shared regions, including the primer regions. I suppose, in other instances, it would be possible to have transcripts with variation in the regions matching the primers insufficient to stop the primers working? But not in this case.

One might conclude there are no genuinely "unintended" products? All are real **PAX6** transcripts of varying certainty. A genuine unintended product would come from an entirely different part of the genome and would not necessarily match exactly with respect to the primers. They would just need to be "good enough to work".

The "**potentially unintended products**" are of different sizes. Can you explain the difference between the possible product lengths?

Are the numbers of "**potentially unintended products**" of each possible length consistent with your **blast** results?

Yes yes yes! I think both these questions made a bit more sense a few generations of these notes ago. We have already answered them sufficiently I suggest. I refer you to the answers above.

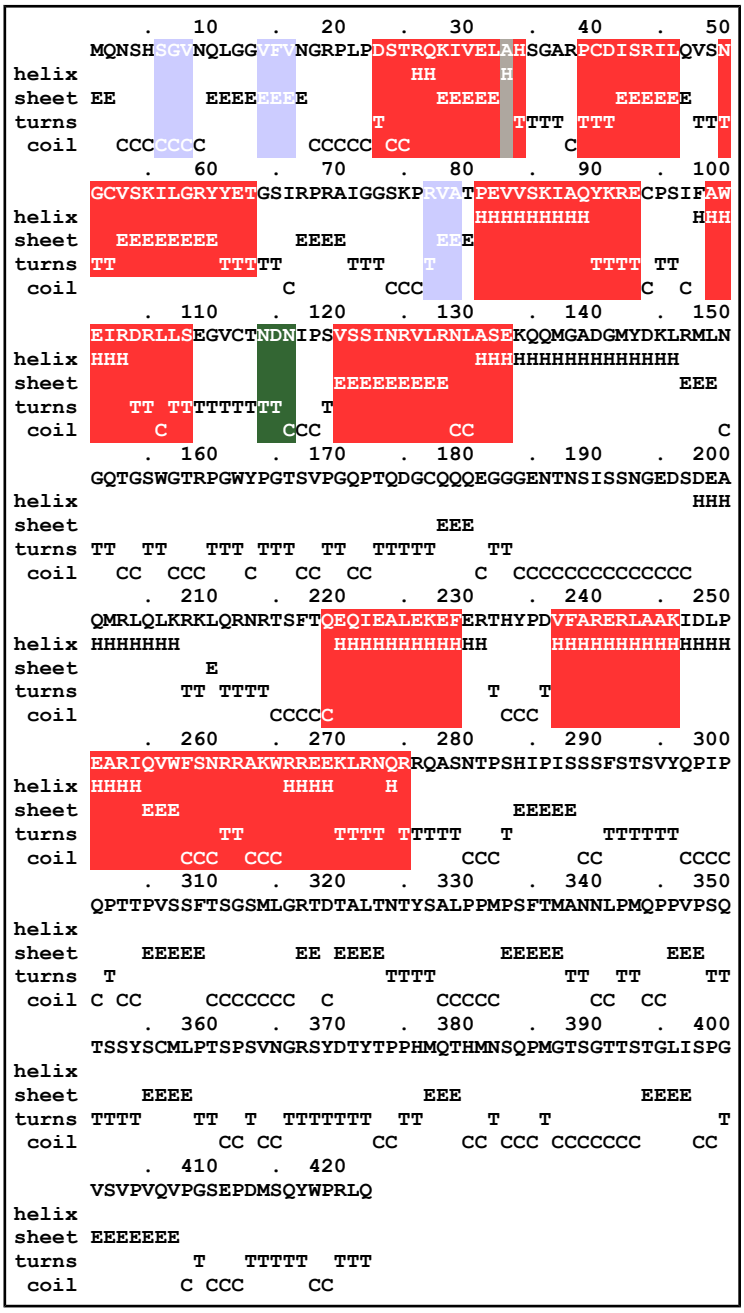
From your investigations of **Protein Secondary Structure Prediction with GOR I**

How credible would you say was the prediction at amino acid 33?

The original **GOR** predicts a Helix of length **1** at position **33**! This is just daft! A single amino acid cannot form a helix all by itself, it must have a few friends to make a believable Helix … or Beta Sheet, come to that.

Later versions of **GOR** made such silly predictions impossible, preferring to take the second best possibility for all insane predictions. In this case, positions **26**, **27** & **33** would all probably be predicted as Beta Sheet (**E** – for **E**xtended). Wrong (according to **UniprotKB**) but more logical.

Position **33** is, as you have discovered, the amino acid that, when mutated from an **A**lanine to a **P**roline, is the major cause of **Aniridia**. As you will confirm later, it is at the end of a Helix critical to to the DNA binding properties of the protein. **A**lanines are fine in Helices. **P**rolines are not.



How would you rate the prediction overall?

Beta strand	6 - 8	3
Beta strand	14 - 16	3
Helix	23 - 34	12
Helix	39 - 46	8
Helix	50 - 63	14
Beta strand	77 - 79	3
Helix	81 - 93	13
Helix	99 - 108	10
Turn	114 - 116	3
Helix	120 - 133	14
Helix	219 - 229	11
Helix	237 - 246	10
Helix	251 - 275	25

Referring to the structure according to **UniProtKB**, I colour the Helices Red and the Beta Sheets Blue. Position **33** I colour Grey. I do not believe the turn predictions, there are far too many! However, as the one predicted does, almost, match the **UniProtKB** assertion, I will highlight it with Green.

The **C**oils are boring, so I do not colour them at all!

Well, what do you think? It might reach the **60%** or so accuracy claimed by the version of **GOR**, but what does that really mean? There are only four possibilities so **25%** "accuracy" can be expected by chance. Add a small insight into the relative abundance of each possibility and you could improve on that before doing anything clever.

My, intuitive only, impression is that the program has found some useful insight into the problem, but I cannot say I am overly impressed. Of course, this is just one example. It is unreasonable to make harsh general judgement based on such light evidence. This is, after all, the crudest implementation of a simple method and I have seen it do much better with more compliant proteins.

From your investigations of **Protein Secondary Structure Prediction with GOR IV**

How does the prediction at position **33** compare with that of **garnier**?

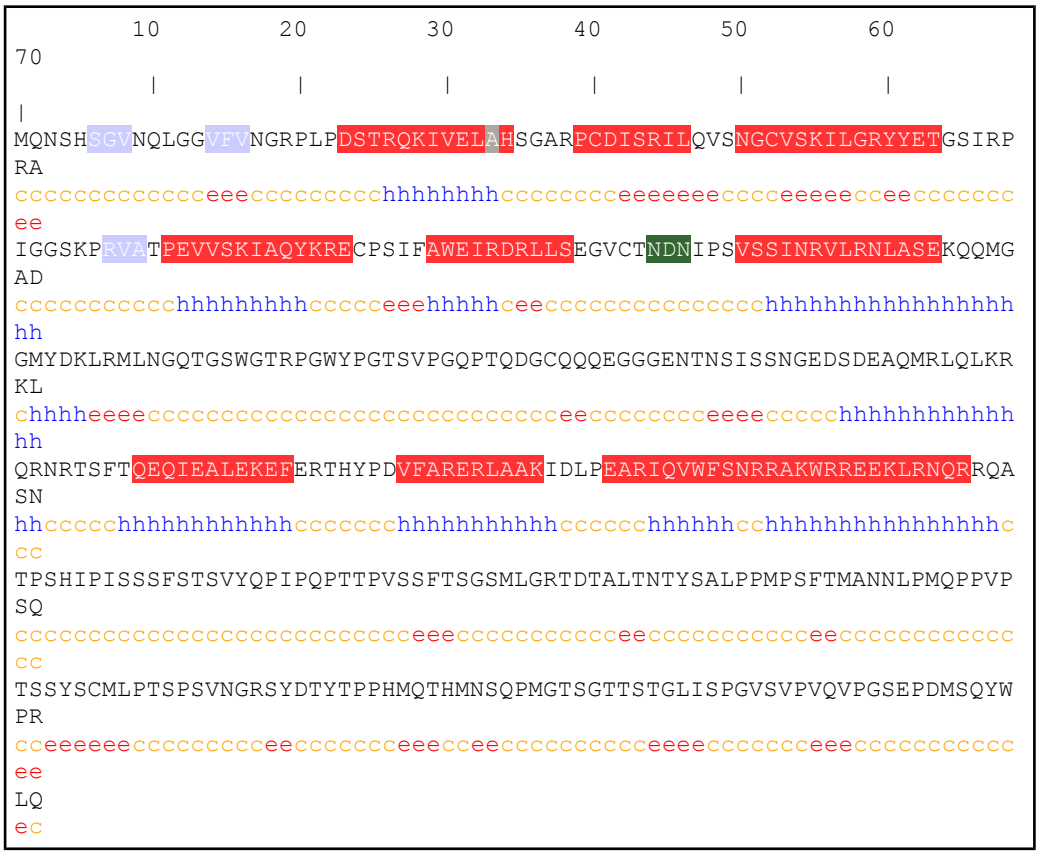
Well, at least there is no attempt to suggest a one amino acid helix in this case! **GOR IV**, with justice, does not regard this as possible. Such silly predictions suggested by the arithmetic are rejected and replaced by more sensible alternatives computed by considering the likelihoods computed for other possibilities.

In a nutshell, **GOR IV** predicts position **33** to be at the end of a helix whose existence is supported reasonably well by **UniProtKB**. This is good! **GOR I** predicts position **33** to be a helix of length **1** amino acid in a region weakly suggested to have some sort of structure. This is weak, to say the least.

How would you rate the prediction overall?

I have coloured up the **GOR IV** results in the same way as I did for the **GOR I** results discussed above. Colouring the turn was unnecessary as **GOR IV** does not even try to predict turns.

This prediction is clearly loads better than the original **GOR** managed. In particular:

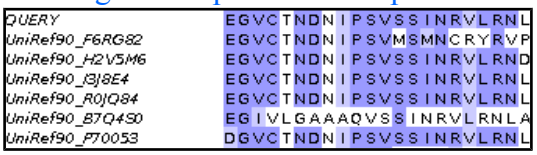


- The region after the last helix (position **280** or so onwards) is, give or take a few spurious attempts at Beta Sheets, predicted accurately as unstructured (i.e. Coil).
- The predictions for the Beta sheets are slightly worse that **GOR I** managed. Both programs got the second of the three and missed the first entirely, but **GOR I** did pick two of the three positions for the third Beta sheet, **GOR IV** missed it altogether!
- Neither program got the helices quite right, but **GOR IV** was much closer to the right answer (according to **UniProtKB**). The two helices **GOR IV** missed altogether were predicted as largely structured (i.e. Beta Sheet instead of Helix). Generally, **GOR** is pretty reliable at picking structure, but not always the right structure. Knowing this, you could claim that predicting the helices as Beta Sheets beats predicting them as Coil.

I conclude, this is a usefully accurate prediction (as long as you do not take it ***TOO*** seriously) and certainly a big improvement on **GOR I**.

From your investigations of **Protein Secondary Structure Prediction with Jpred**

What protein database has **Jpred** chosen to search for protein sequences for the alignment upon which its predictions will be based?



The database **Jpred** instructed **PSI-blast** to use to seek proteins homologous to the **PAX6** query can be determined by looking at the sequence identifiers displayed down the left hand side of the alignment in **Jalview**. The identifiers are constructed from the name of the database and the entry identifier separated by an underline character. So the database is the **UniRef90** cluster database built from the **UniProtKB** database.

The **UniRef** cluster databases comprise entries that are not individual protein sequences, but cluster of similar sequences. In the case of the **UniRef90** database, each entry includes all sequences 90% identical to a given seed sequence. A representative sequence is elected as the only one of the cluster to be considered by such as **PSI-blast**, but clearly, a hit with any representative sequence implies significant similarity with all the sequences of its cluster.

I offer a supplementary exercise to investigate these cluster databases for those to whom they might be of particular interest.

Why do you suppose this database was used in preference to, say **UniprotKB**?

The reason **Jpred** runs **PSI-blast** is to identify sequences representing as wide a family of proteins as possible, to which a **Query** sequence belongs. For the purpose of structure prediction, there is little value in this collection including many sequences that are essentially identical. A wide variety of sequences, as long as they still are likely to be homologous, is of far greater value than a huge number of sequences. Using a **UniRef** database allows that only the **Representative** sequence of each cluster of very similar sequences will be recognised and aligned by **PSI-blast**. This allows the **PSI-blast MSA** to include an extensive range of variation without being bloated by sequences too similar to be individually interesting.

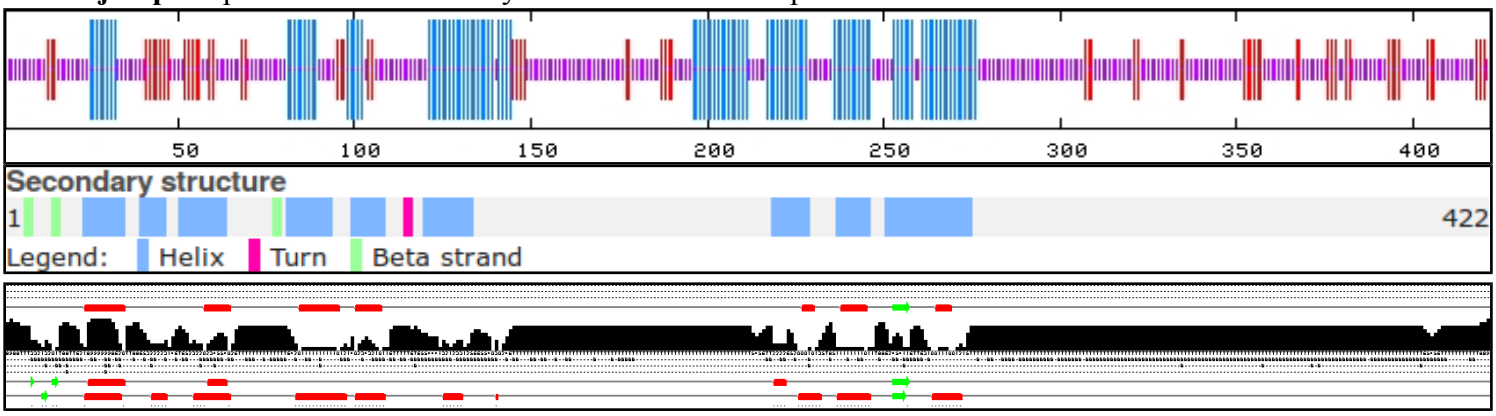
How would you rate the **Jpred** prediction overall?

Well, frankly, not as wonderful as I was expecting. Better than **GOR IV** (investigated in a supplementary exercise), but still leaves room for improvement? **jnetpred** (essentially the answer) is reasonable. It misses a couple of helices including one that **GOR IV** also overlooks. However, it has considerably less false positive prediction tendencies. The **JNETHMM** predictions are particularly poor, saved by the much more accurate deliberations of **JNETPSSM**.

JNETHMM is a prediction computed from the **H**idden **M**arkov **M**odel (**HMM**) representation of the final **PSI-blast MSA**.

JNETPSSM is a prediction computed from the **P**osition **S**pecific **S**coring **M**atrix (**PSSM**) representation of the final **PSI-blast MSA**. **PSI-blast** uses **PSSM**s of the **MSA** of each iteration of its search as a **Query** for the next iteration.

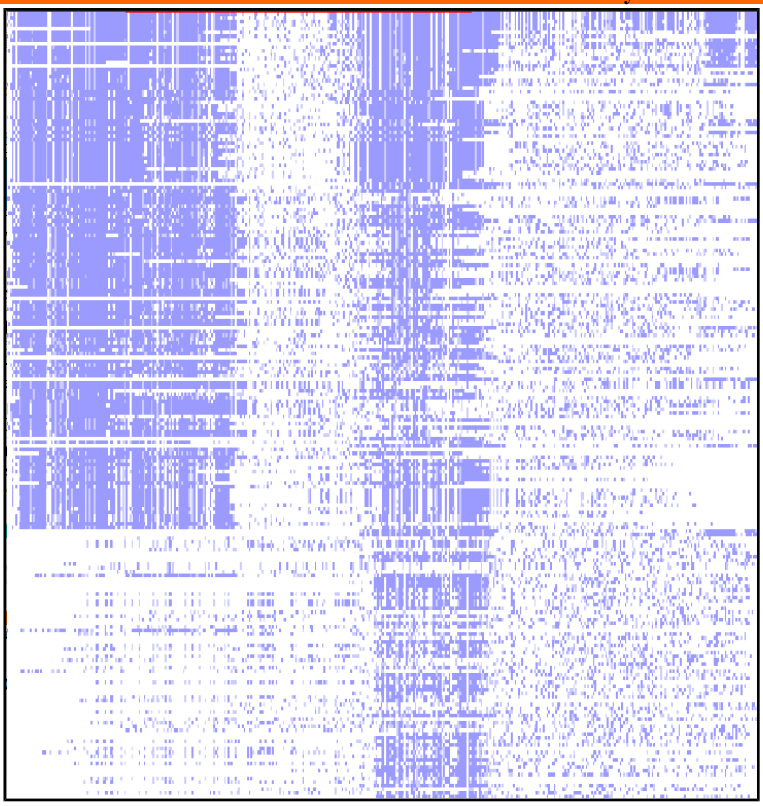
The **jnetpred** prediction is effectively the consensus of the predictions of **JNETHMM** and **JNETPSSM**.



So, can the prediction be improved? **Jpred** is better than this result suggests!

On reflection, maybe just throwing in the entire sequence of **PAX6_HUMAN** and hoping for the best was a little crude? Our protein has two major domains whose secondary structure one might expect to be conserved. **PSI-blast** will gather together a mountain of sequences that have one, or the other, or both of the domains and try to align them as if they were homologous over their entire length (a **global alignment**). ***BUT***, they are not all

globally homologous! This means that the alignment of both the domains regions will be polluted by sequence that represent proteins that do not include that domain. This must substantially reduce the quality of the prediction?



This phenomena can be illustrated by choosing to view the **Jalview Overview Window** (available from the **View** pull down menu).

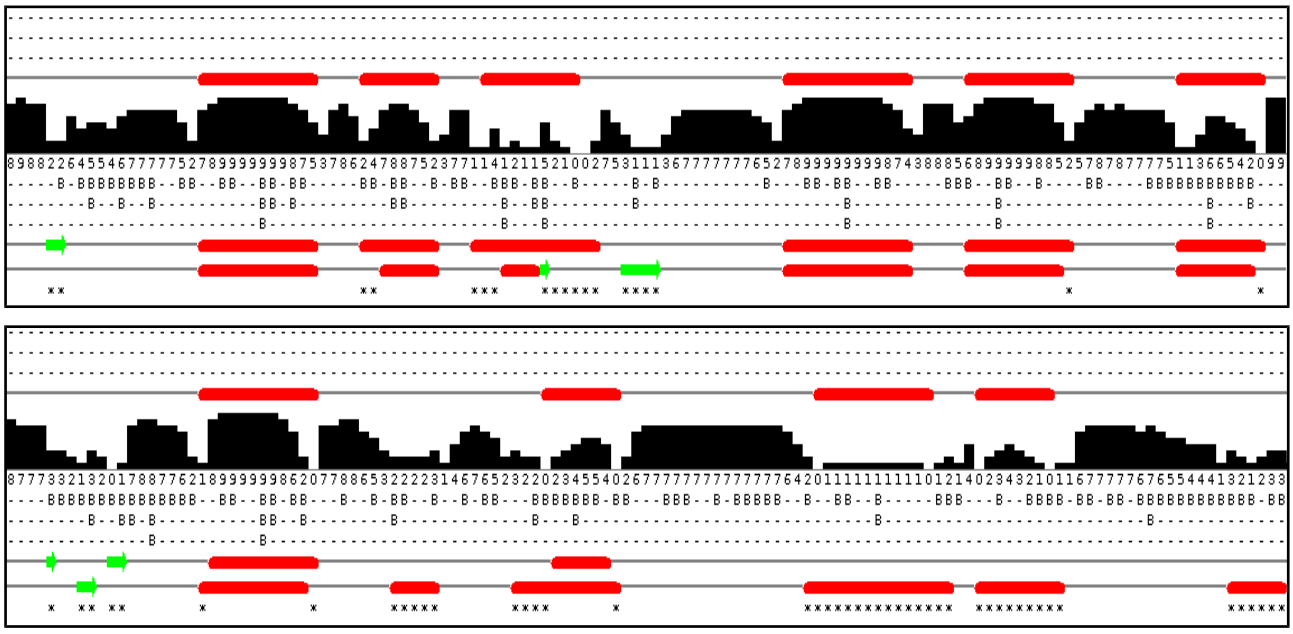
The wider column of blueness at the start of the alignment represents the **paired box** domains. The picture suggests about one third of the aligned sequences do not have a **paired box** domain, but those sequences will have unrelated sequence in that region that will reduce the degree to which the alignment represents the properties of a **paired box** and so also the likelihood of a sensible structure prediction.

The problem for the more common **homebox** domain looks less severe, however, the alignment clearly includes many sequences that do not look to have a **homeobox** domain.

So, what to do? I suggest the two domains might be investigated separately? Why not run **Jpred** twice, once with just the **PAX6_HUMAN** paired box region and then again with just the **homebox** region.

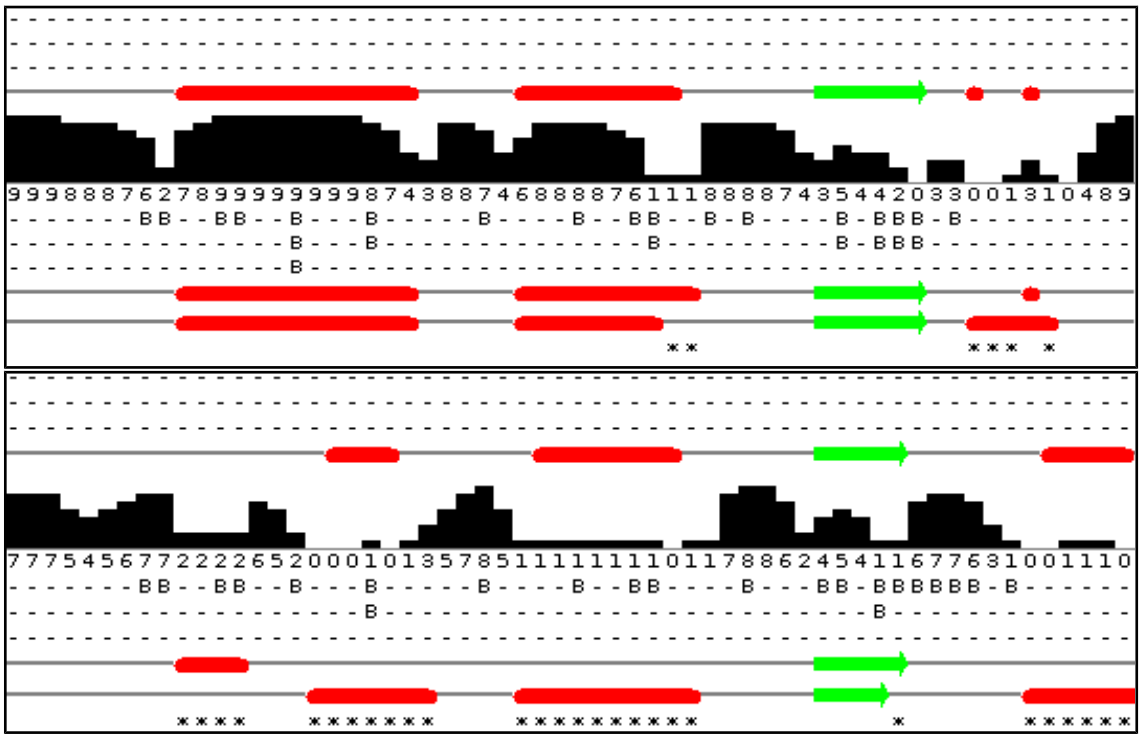
I have done this for you and will now show you the results, however, should you wish to try it yourself, you already have the isolated sequence of both domains saved in local files. The sequence of the **paired box** region should be in a file called **pax_domain.fasta**. The **homeobox** sequence should be in a file called **homeobox_domain.fasta**. Just run **Jpred** again with each sequence and you should get results very similar to mine.

First the new **paired box** prediction (top) compared to the original (bottom).



Massively improved I would suggest. All helices present and accurately placed. The **JPREDHMM** prediction, in particular, is very much improved. The Beta Sheet predictions seem weak? It finds only one (accurately) of the three that **UniProtKB** suggests to be present. I wonder why, but the helices for the paired box domain specific prediction are excellent.

And so to the **homeobox** specific results. Once more, the new **homeobox** prediction (top) compared to the original (bottom).



As the **homeobox**s are significantly more numerous than the **paired box**s, less interference from sequences not including a **homeobox** might have been expected. I imagined the improvement in prediction would be minimal. However, it is very much better! All three helices are predicted in the correct positions, although **Jpred** appears to be a little reluctant about the third helix? There is a rather strong beta sheet prediction that is unsupported by **UniProtKB**. There is no reason to suppose that **UniProtKB** is 100% correct, of course, but nothing I can find suggests that a beta sheet should appear in the middle of a homeobox. An enigma for another day.

So I conclude that this sort of protein analysis requires a little bit more than just throwing an entire sequence at a dumb program and assuming something marvellous will occur. In this case, considering the regions of the protein that are expected to be homologus separately is a very logical thing to do (and entirely obvious, retrospectively at least). Geoff Barton, whose group is responsible for **Jpred** agrees. He says:

> " … Always split proteins into domains when searching. … "

So for both domains the prediction of the helices is far more accurate when each domain is considered separately. However, it is not just the red bars indicating the position of the helical predictions that should be noted. Look also at the confidence histogram. It indicates clearly that with more specific data to work on, better predictions can be made with much improved confidence (i.e. likelihood of being correct!).

Searching PROSITE

What is the signature pattern for **N-myristoylation** site?

From the database entry, it can be seen that the pattern is **6** positions wide. **2** of those positions can be any amino acid. Only one position is fully specified. Not too demanding on the whole. I would expect this to match most proteins of any size and not always because there was an **N-myristoylation** site.

MYRISTYL, PS00008; N-myristoylation site (PATTERN with a high probability of occurrence!)

- Consensus pattern:
  G-{EDRKHPFYW}-x(2)-[STAGCN]-{P}[GistheN-myristoylationsite]

How would you interpret this pattern?

The pattern is explained in the database thus.

- The N-terminal residue must be glycine.
- In position 2, uncharged residues are allowed. Charged residues, proline and large hydrophobic residues are not allowed.
- In positions 3 and 4, most, if not all, residues are allowed.
- In position 5, small uncharged residues are allowed (Ala, Ser, Thr, Cys, Asn and Gly). Serine is favored.
- In position 6, proline is not allowed.

The description is not entirely an honest reflection of the information to which the scanning software will respond. The software is given to understand that **ANY** amino acid can occur in positions **3** and **4**. The software has no way to know that "**Serine** is favoured" in position **5**! Maybe you think that my pointing out these transparent truths makes me an intolerable pedant? Well … so is the computer!

How many **N-myristoylation** sites did **ScanProsite** suggest there might be in **PAX6_HUMAN**?

11, is the short answer. A site every **40** amino acids or so.

PS00008 MYRISTYL *N-myristoylation site* :

13 - 18:	GVfvNG
36 - 41:	GArpCD
110 - 115:	GVctND
151 - 156:	GQtgSW
154 - 159:	GSwgTR
157 - 162:	GTrpGW
182 - 187:	GGgeNT
183 - 188:	GGenTN
312 - 317:	GSmlGR
387 - 392:	GTsgTT
390 - 395:	GTtsTG

How many real **N-myristoylation** sites would you guess there might be in **PAX6_HUMAN**?

It is not really possible to answer this question from the evidence of this exercise. Intuitively, I would expect a large number of false positives from as weakly specified motif as this one. It has been suggested of this **PROSITE** pattern, by researchers looking at more sophisticated detection methods, that:

> "**PS00008** of **PROSITE** constructed from a small dataset … produces a great number of not only false positive but false negative predictions."

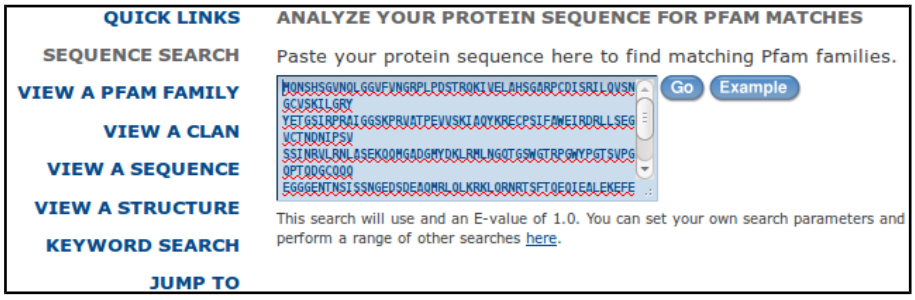
Good enough for me not to believe the majority of these predictions to be reliable. Not good enough for me to hazard a meaningful guess as to how many real sites I would expect in this particular protein.

Searching Pfam

Would you have found anything of interest had you chosen "to search for Pfam-B matches"?

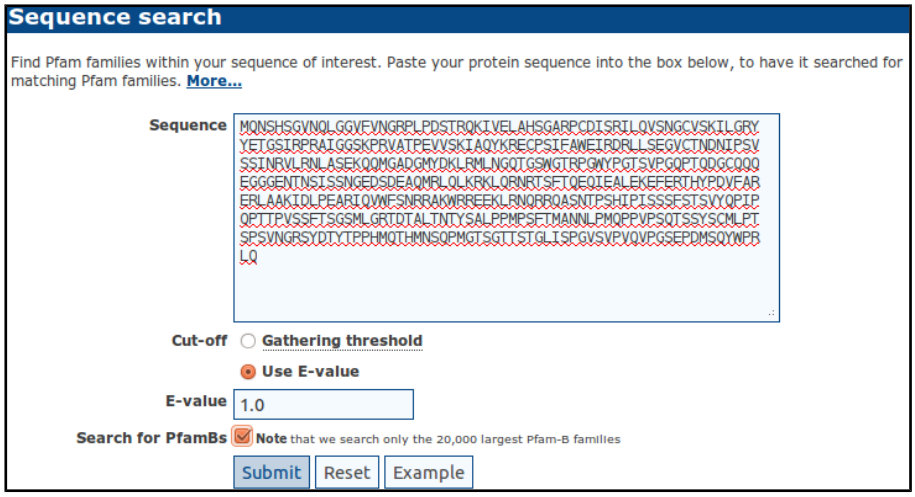
I did not find it straight forward to persuade the search to include **Pfam-B** entries.

In the end, I set up the search as required just to search **Pfam-A.** Then, click on the truly tiny **here** link at the end of the sentence suggesting *"You can set up your own search parameters ..."*.



A new search form! I pause to feel badly treated in that I have to re-enter my sequence! I click the Search for **PfamB**s box, reflecting as I do so, how nice that button would have looked on the previous search form?

With a triumphant flourish, I click on the **Submit** button.



NOTHING!! Well one use to find a **PfamB** hit that made all the fiddling around make a very small amount of sense, but now??? **NOTHING**!!! They claim the hit one got previously was not a worthy hit … but … but … it was a hit!!! AND, I had a very nice story about how it fitted in with all that we have discovered previously. One, in common with dear Victoria, is NOT amused!

2015.08.08: So you thought that was bad Dave Mawr!!! … Now the little button has gone altogether and cannot search **PfamB** at all???? Query sent to Help.

2015.08.10: In reply to the question " … Searching **PfamB** - Is this no longer possible? … "

" … Correct - we are not longer producing **Pfam-B**s, largely because most of the clusters not covered by **Pfam** are rarely meaningful potential new domains. Our recommendation now, is to take the piece of sequence of interest and run it using the **HMMER** webserver.

http://www.ebi.ac.uk/Tools/hmmer/search/phmmer

This will then produce a set of results that essentially provide you with the same information as a **Pfam** entry. … "

This makes sense only if you change **really** to **rarely** in line 2? I assume this is what was meant (indeed it is, confirmed later). I also assume **PfamB** still to exists, in spirit at least, … as a source to discover new **PfamA** entries? But, users must effectively run **HMMER** themselves if they wish to search beyond what is offered by **PfamA** … now consistently referred to as **Pfam**? Hmm, maybe I try to clear this up a bit sometime?

Searching for Helix-Turn-Helix motifs

To which, if any, of the expected **HTH** motifs does this prediction correspond?

Using all the defaults, there is just one prediction:

```
Sequence: FARERLAAKIDLPEARIQVWFS
            |                   |
          238                 259
```

Looking again at the reported secondary structure in **UniProtKB**, I would expect there to be 3 **HTH** motifs roughly at:

```
39  →  63
99  → 116
237 → 275
```

So here we are looking at the second expected **HTH**. That is, the one associated with the helical triplet of the **HomeoBox** domain.

STRAND	6	8	3
STRAND	14	16	3
HELIX	23	34	12
HELIX	39	46	8
HELIX	50	63	14
STRAND	77	79	3
HELIX	81	93	13
HELIX	99	108	10
TURN	114	116	3
HELIX	120	133	14
HELIX	219	229	11
HELIX	237	246	10
HELIX	251	275	25

To which, if any, of the expected **HTH** motifs does the new prediction correspond?

The new prediction is

```
Sequence: PCDISRILQVSNGCVSKILG
            |                 |
           39                58
```

Near enough to the **HTH** expected for the first helical triplet of the **PairedBox** domain for me.

Can you suggest anything particular about the third putative **HTH** motif that might explain its reluctance to be discovered?

Nope …. which is why one should answer the questions straight away! I must have had a theory at one time?

It is possible to find this third **HTH** by using the old scoring matrix and an **SD** limit of **0.8**

However, by loosening things up this far, one gets seven answers instead of three. Fours out of the seven are wrong (well, one supposes so at least).

From your investigations of **Domain & Motif identification** using **Interpro**

Do you think it a good idea for **Interpro** to offer feature prediction programs as well as domain database searches?

Well … why not? The purpose of **InterProScan** is to associate regions of query proteins with **Interpro** domains. This was originally achieved was, exclusively, by simply comparing a query sequence with all entries of relevant individual domain databases. These entries being representations of alignments of examples of specific domains constructed by homology searching (i.e. **blast** and similar).

I would suggest including a few predictor programs would provide extra evidence gathered from more general, more theoretical definitions of domains. I would imagine the inclusion of these program has improved and widened the picture provided by **InterProScan**.

Searching domain databases, typically composed of **HMM profiles**, such as **Pfam**, **Prosite** and **PRINTS** is quite different to running the predictor programs. As I cannot improve on the justification of this claim offered to me by Geoff Barton (Head of the group resonsible for **Jalview**, **Jpred**, **Jnet** and much more), I will just reproduce his explanation here:

> " … The main difference is that with an **HMM profile** you have a "specific" example of a domain or motif whereas with something like **COILS**, you have something trained across all examples.
>
> For example, for secondary structure prediction, you could (a) do predictions of alpha-helix and beta-strand just by aligning a sequence to a protein of known structure, or an **HMM** from a family of aligned proteins of known structure. This is a specific case of secondary structure in the context of one protein family. Or (b) you can train a predictor from **ALL** protein families and then apply this. The advantage of (a) is it is very specific to the individual protein family and so should be more accurate for that family. The disadvantage is that it does not generalise to proteins that are not very like the specific example. The advantage of (b) is that it will work with any protein but will likely be less accurate than (a) for proteins that fit into the (a) category. … "

Do you think the Coil prediction might be correct?

I do not recall anything in what we have discovered thus far that would directly suggest there should be a **coiled coil** here, in the middle of the **HTH**. However, wikipedia does suggest **coiled coils** are associated with **transcription factors** (which **pax6_human** is).

" ... Many **coiled coil**-type proteins are involved in important biological functions such as the regulation of **gene expression**, e.g. **transcription factors**. ... "

I think I would not be overly convinced by this prediction, but I would not make that judgement with any great confidence. The all knowing **wikipedia** says:

" … **Coiled coils** usually contain a repeated pattern, ***hxxhcxc***, of **h**ydrophobic (***h***) and **c**harged (***c***) amino-acid residues, referred to as a **heptad repeat**. … "

Geoff Barton comments:

> " … Sometimes the pattern that is particular to **coiled-coils** also turns up in other helices that pack against each other. You would need to look at some examples of coiled-coil structures to see if the example you are using fits structurally. … "

Which seems very reasonable. The **heptad repeat** pattern could easily occur just by chance. **COILS** surely cannot predict the structure of the helices well enough to make an assured judgement? **COILS** offers a suggestion the user must follow up with other resources.

There is also the evidence that **Jpred**, possibly using the **COILS** program disguised as **LUPAS**, did not detect any coiled coils. This could be for a number of reasons. Possibly **LUPAS** is not the same program as **COILS**, or it is a different version, or **Jpred** runs **COILS**, but with different parameters.

Not many clear and confident answers in Bioinformatics are there!

From your investigations of **Multiple Sequence Alignment**

Clustalx

How might you have saved the need to recompute the alignment by selecting an **Edit** option other than **Remove All Gaps**?

Pretty sure using **Remove Gap-Only columns** rather than **Remove All Gaps** would have got to the second alignment in one step. I struggle to see how this could fail, but cannot convince myself it would be absolutely certain to work. Maybe some of the sequences that were removed would subtly alter the alignment calculations? Doubtful, given how crude the whole thing is. Anyway, it would have been good enough for me in "Real Life" for this data. I did not suggest this approach as we had to make a **Fasta** format file with the entire sequence set to be aligned by the alternative software as a by product of working with **clustalx**. This would not be easily possible if the **Remove Gap-Only columns** short cut was taken.

Muscle

How do the options for the **OUTPUT TREE** relate to the output files of **ClustalX** and the difference between the way that **ClustalX** and **muscle** work?

Comment on how one might choose between the range of options offered for the aligned parameter?

Interpro:

Relationships: PARENT/CHILD; CONTAINS/FOUND IN

The PARENT/CHILD relationship is used to indicate family/subfamily relationships defined by the member database methods in the entries. To be a CHILD, >75% of the protein set of the CHILD must be represented in the PARENT. If an InterPro PARENT has more than one InterPro CHILD a protein sequence should not be found in the match table of more than one of these children. If one InterPro entry is described as the child of another InterPro entry, this implies that the child entry is more specific than the parent, and that in all cases a protein sequence match to the child entry implies a match to the parent. Signatures for the parent and child entries must overlap by >50% of their sequence and there must be no adjacent signatures to the CHILD that are also covered by one or more of the PARENT signatures. A list of the PARENT/CHILD relationships by InterPro entry accession is available from the ftp site: ParentChildTreeFile.txt.

The CONTAINS/FOUND IN relationship is used to indicate features of the entry that are not defined by Parent/Child relationships, these include: Regions, Domains, Repeats and Sites. For a CONTAINS/FOUND IN relationship to be established, 40% or above of the proteins in the InterPro entry must contain the feature. Some features can be found in more than one type of protein or family of proteins, but is not children in the family sense. Features may be structural or functional and can be found in proteins with different domain organisations. The CONTAINS/FOUND IN relationship, is therefore useful in linking InterPro entries which are associated by composition but are not related hierarchically.

Taxonomy coverage:

Taxonomy Coverage

The Taxonomy Coverage aims to provide 'at a glance' view of the taxonomic range of the sequences associated with each InterPro entry and the number of sequences associated with each lineage. The taxonomic lineages are 'clickable' and provide a pop-up, which displays the tax-ID, the taxonomy and taxonomic subgroup(s)/species having matches to proteins, the protein match counts and a FASTA link. Clicking on the taxonomy or taxonomic subgroup(s)/species links to the protein overview matches for the selected taxonomy. Clicking on the FASTA box will download the complete set of FASTA sequences for the selected taxonomy of the entry.

The lineages were carefully selected to provide a view of the major groups of organisms. The circular display has the taxonomy-tree root as its centre. Selected model organisms populate the outer most circle. Nodes of the taxonomy-tree are placed on the inner circles. Radial lines lead to the description for each node. No significance is attached to the position of the node on a particular inner-circle, other than convenience, though some attempt has been made to group nodes. The nodes themselves are either true taxonomy nodes and have a NCBI taxonomy number or are artificial nodes created for this display; of which there are three: **Unclassified**, **Other Eukaryotes** (Non-Metazoa) and the **Plastid Group**.

Artificial Taxon: **Unclassified** contains the following NCBI taxon groups:

- Taxonomy:12884 Viroids
- Taxonomy:12908 unclassified sequences
- Taxonomy:28384 other sequences

The Eukaryota (TAXONOMY:2759) comprises 29 taxons, these have been grouped into two artificial taxons and one existing taxon:

Fungi/Metazoa (TAXONOMY:33154); Node **Metazoa**

Artificial Taxon; **Plastid Group**, this contains the following NCBI taxon groups:

- Taxonomy: 2763 Rhodophyta
- Taxonomy: 2830 Haptophyceae
- Taxonomy: 3027 Cryptophyta
- Taxonomy: 33090 Viridiplantae
- Taxonomy: 33630 Alveolata
- Taxonomy: 33634 stramenopiles
- Taxonomy: 33682 Euglenozoa
- Taxonomy: 38254 Glaucocystophyceae
- Taxonomy: 339960 Katablepharidophyta

Each taxonomic group within this artificial taxon contains organisms that have a plastid.

Artificial Taxon; **Other Eukaryotes** (Non-Metazoa), this comprises the following NCBI taxon groups:

- Taxonomy: 5719 Parabasalidea
- Taxonomy: 5752 Heterolobosea
- Taxonomy: 66288 Oxymonadida
- Taxonomy: 136087 Malawimonadidae
- Taxonomy: 154966 Nucleariidae
- Taxonomy: 193537 Centroheliozoa
- Taxonomy: 207245 Diplomonadida group
- Taxonomy: 543769 Rhizaria
- Taxonomy: 554296 Apusozoa
- Taxonomy: 554915 Amoebozoa
- Taxonomy: 556282 Jakobida

Each taxonomic group within this artificial taxon are the remaining taxonomic groups of the NCBI taxon:2759, which are not in the Plastid Group and are not Fungi/Metazoa (TAXONOMY:33154).

Overlapping Interpro entries

Overlapping InterPro Entries

This section displays entries that share more than 70% of their proteins. Such overlaps define PARENT/CHILD and CONTAINS/FOUND IN relationships between InterPro entries.

IPR009007	Numbers of overlapping proteins	Average numbers of overlapping amino acids

In the above example, InterPro entry IPR011969 contains proteins which are also found in IPR009007 as a result of the protein signatures of the two entries overlapping.

The two entries have been compared firstly by counting the number of proteins which are common to both, the results of which are displayed in the Venn diagram on the left, and secondly by calculating the average overlap of the protein signatures, in amino acids, with the results displayed in the bar diagram on the right.

Venn diagram display of the overlap of proteins common to both entries:

- The purple intersection contains the number of overlapping proteins common to both IPR009007 and IPR011969, which is 31 in this case.
- The pink section on the left is the number of proteins found in IPR009007 but not IPR011969, which is 35378.

Bar diagram display of the average amino acid overlap between the protein signatures:

The average number of amino acids overlapping in the sequences of the 31 proteins common to both entries is then calculated, with the results displayed in the bar diagram on the right. The bar diagram display is only shown for 'Domain - Domain' relationships.

- The purple segment in the middle shows the average number of amino acids overlapping between IPR009007 and IPR011969 for the 31 proteins, in this case 104.
- The pink segment shows the average number of amino acids found in IPR009007, but not IPR011969, for the 31 proteins, which is 0.
- The blue segment shows the average number of amino acids found in IPR011969, but not IPR009007, for the 31 proteins, which is 15.

The results of these comparisons are used to calculate the percentage overlap score, with all scores greater than 70% displayed on the InterPro pages. In this example, since all proteins found in IPR011969 are also found in IPR009007, and all the amino acids from IPR009007 overlap with those from IPR011969, the percentage overlap score is 100%.

Muscle:

- The blue section on the right is the number of proteins found in IPR011969 but not IPR009007, which is 0; i.e. all proteins associated with IPR011969 occur in IPR009007.

How do the options for the **OUTPUT TREE** relate to the output files of **ClustalX** and the difference between the way that **ClustalX** and **muscle** work?_____

none
From first iteration
From second iteration

Comment on how one might choose between the range of options offered for the aligned parameter?_____

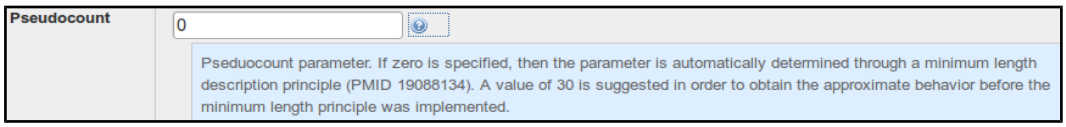
Clearly?

aligned

PSI-Blast

What do you suppose the choice of **Pseudocount** might influence?

I clicked with confidences upon the link to the help. It opined as illustrated.



I suppose the next step is to read PMID 19088134? There is most certainly no elucidation amongst the strangle of words offered here?

The article **Abstract** says:

"**P**osition **s**pecific **s**core **m**atrices (**PSSM**s) are derived from multiple sequence alignments to aid in the recognition of distant protein sequence relationships. The **PSI-BLAST** protein database search program derives the column scores of its **PSSM**s with the aid of **pseudocounts**, added to the observed amino acid counts in a multiple alignment column. In the absence of theory, the number of **pseudocounts** used has been a completely empirical parameter. This article argues that the minimum description length principle can motivate the choice of this parameter. Specifically, for realistic alignments, the principle supports the practice of using a number of **pseudocounts** essentially independent of alignment size. However, it also implies that more highly conserved columns should use fewer **pseudocounts**, increasing the inter-column contrast of the implied **PSSM**s. A new method for calculating **pseudocounts** that significantly improves **PSI-BLAST**'s; retrieval accuracy is now employed by default."

The article itself, continues in like vein ….. how about we close our eyes and accept the defaults? I would just wonder why the whole thing does not commence with, at least an attempt, to answer the question in the forefront of my inquiry, which is .. "**WHAT, in the current context, IS a pseudocount?**". I do not believe it is as tricky as they appear to wish us to believe. I will try again later, when my view of the world is less storm infested.

******

**********NOTES**********

HelixTurnHelix

helixturnhelix uses the method of Dodd and Egan and finds helix-turn-helix nucleic acid binding motifs in proteins.

The helix-turn-helix motif was originally identified as the DNA-binding domain of phage repressors. One alpha-helix lies in the wide groove of DNA; the other lies at an angle across DNA.

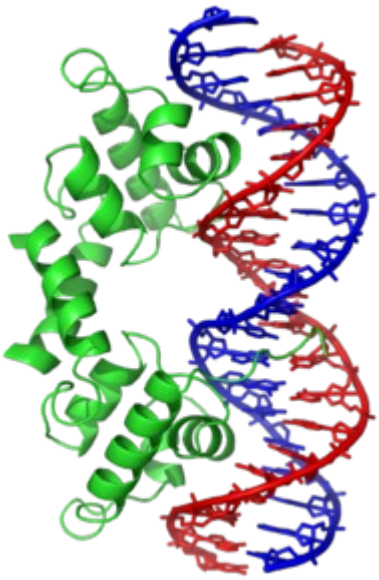
The old (1987) data has a motif length of 20 residues, whilst the default data (Ehth.dat) has a motif length of 22 residues.

With care these can be replaced to suit your data sets. If the files are placed in the following directories they will be used in preference to the files in the EMBOSS distribution data directory:

- . (your current directory)
- .embossdata
- ~/ (your home directory)
- ~/.embossdata

Here is the default file:

```
# Amino acid counts for 91 Helix-turn-helix (presumed) protein motifs
# from Dodd IB and Egan JB (1990) Nucl. Acids. Res. 18:5019-5026.
#
Sample: 91 aligned sequences
#
# R  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 Total Exp
# - -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- -- ----- ---
  A  2  1  3 14 10 12 75  6 15  9  1  1  4  3  8 15  4  4  4 11  0 10   212 995
  C  0  0  1  1  0  0  0  0  0  3  3  1  1  0  0  0  0  0  0  1  0  3    14 106
  D  0  1  0  1 14  0  0 14  1  0  5  0  1  2  0  0  0  0  1  1  0  2    43 556
  E  4  5  0 11 26  0  0 16  9  3  3  0  3 12 13  0  0  2  0  1 13  6   127 669
  F  4  0  4  0  0  4  0  1  0 10  0  0  0  0  1  0  0  1  1  1 22  0    49 358
  G  9  7  1  4  0  0  8  0  0  0 50  0  6  0  7  1  0  3  1  1  0  4   102 761
  H  4  3  1  1  2  0  0  3  2  0  5  0  3  3  0  2  0  2  4  5  0  2    42 225
  I 10  0 13  3  2 15  0  4  9  4  0 17  0  2  0  1 31  1  4  8 16  1   141 583
  K  4  4  6 11 12  1  1 14 11  0  5  2  2  7  2  1  0  5  8  4  5 15   120 516
  L 16  1 17  0  1 35  0  3 12 31  0 22  0  2  1  1 22  1  1 12 20  0   198 954
  M  7  0  2  1  1  1  0  0  5  7  1 10  0  0  2  0  2  0  0  2  0  1    42 275
  N  0  8  0  1  0  0  0  2  1  1 14  0  8  1  4  2  0  4  9  0  0 11    66 383
  P  1  6  0  1  0  0  0  0  0  0  0  0  3 13  7  0  0  0  0  0  0  3    34 403
  Q  2  1 21  9 11  0  0  9  8  0  0  2  1 17  7 12  0  3 12  5  3  9   132 437
  R  9 10 14  9  5  0  1 16 10  0  1  0  1 17  8  7  0 17 28  3  0 16   172 609
  S  2 17  0  8  4  1  6  1  2  2  3  0 37  1 25  5  0 29  3  0  1  5   152 552
  T  6 24  3 12  1  5  0  2  2  4  0  5 20  4  3 39  0  4  1  0  4  3   142 512
  V  7  3  1  1  2 16  0  0  2 12  0 29  0  5  3  3 32  0  7  8  7  0   138 724
  W  2  0  0  0  0  0  0  0  0  1  0  1  0  0  0  0  0  0  2 21  0  0    27 105
  Y  2  0  4  3  0  1  0  0  2  4  0  1  1  2  0  2  0 15  5  7  0  0    49 267
```



The λ repressor of bacteriophage lambda employs a helix-turn-helix (top; green) to bind DNA (bottom; blue and red).

In proteins, the helix-turn-helix (HTH) is a major structural motif capable of binding DNA. It is composed of two α helices joined by a short strand of amino acids and is found in many proteins that regulate gene expression. It should not to be confused with the helix-loop-helix domain.

Its discovery was based on similarities between the genes for Cro, CAP, and λ repressor, which share a common 20-25 amino acid sequence that facilitates DNA recognition. In particular, recognition and binding to DNA is done by the two α helices, one occupying the N-terminal end of the motif, the other at the C-terminus. In most cases, such as in the Cro repressor, the second helix contributes most to DNA recognition, and hence it is often called the "recognition helix". It binds to the major groove of DNA through a series of hydrogen bonds and various Van der Waals interactions with exposed bases. The other α helix stabilizes the interaction between protein and DNA, but does not play a particularly strong role in its recognition.

DPJ – 2015.09.14