ELB17S

Entry Level Bioinformatics

06-10 November 2017

(Second 2017 run of this Course)

Basic Bioinformatics Sessions

Practical 1: Databases and Tools

Sunday 29 October 2017

Investigating the gene(s) associated with Aniridia

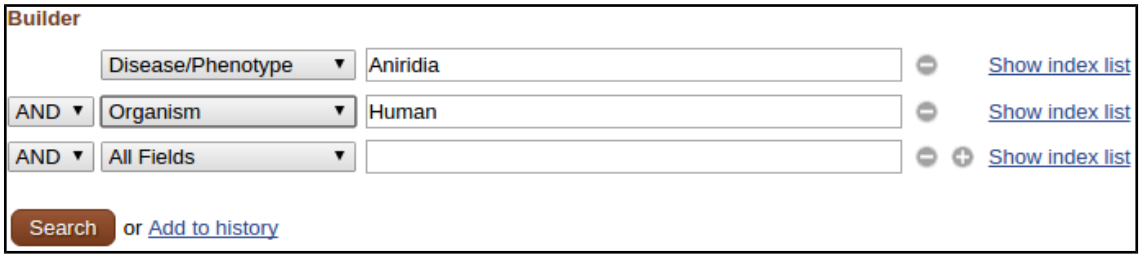
As a starting point for this exercise, imagine you have a vital interest in discovering and investigating the main human gene responsible for the terrible disease of the eye, **Aniridia**. There are many ways (including **google**!) you could discover what this gene might be. I choose to delve into the vast seas of knowledge so generously proffered by the **The National Center for Biotechnology Information (NCBI)**.

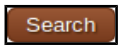
So, begin by going to the **Home Page** of the **The National Center for Biotechnology Information (NCBI)** ("**http://www.ncbi.nlm.nih.gov/**").

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

A fine list of genes will emerge, including those sought. However, our interest is specific to Human, so the search should really be organism specific. To do this, one needs to execute an **Advanced** search. So, click on the **Advanced** button of the **Search** tool.

Now you can specify the precise field(s) of the annotation you wish to interrogate. In this case, set the **Disease/Phenotype** field to **Aniridia** and the **Organism** field to **Human**. As the two conditions are linked by **AND**, both must be true for any gene to be listed.



Click on the pretty red  button.

Name/Gene ID	Description	Location	Aliases	MIM
WT1 ID: 7490	Wilms tumor 1 [*Homo sapiens* (human)]	Chromosome 11, NC_000011.10 (32387775..32435535, complement)	AWT1, EWS-WT1, GUD, NPHS4, WAGR, WIT-2, WT33	607102
PAX6 ID: 5080	paired box 6 [*Homo sapiens* (human)]	Chromosome 11, NC_000011.10 (31784792..31817961, complement)	AN, AN2, ASGD5, D11S812E, FVH1, MGDA, WAGR	607108
ELP4 ID: 26610	elongator acetyltransferase complex subunit 4 [*Homo sapiens* (human)]	Chromosome 11, NC_000011.10 (31509729..31784525)	AN, AN2, C11orf19, PAX6NEB, PAXNEB, dJ68P15A.1, hELP4	606985
TRIM44 ID: 54765	tripartite motif containing 44 [*Homo sapiens* (human)]	Chromosome 11, NC_000011.10 (35662692..35811053)	AN3, DIPB, HSA249128, MC7	612298
DEL11P13 ID: 100528024	Wilms tumor, aniridia, genitourinary anomalies and mental retardation syndrome [*Homo sapiens* (human)]		C11DELp13, WAGR	194072

Just a few genes survive. All should really be examined, but this is just an exercise, so trust me … it is **PAX6** that is the most interesting gene[1], in this context. This is the one to follow up by clicking on the link to its details.

From the **Summary** section one can conclude (sticking to the features that pertain to this exercise) that:

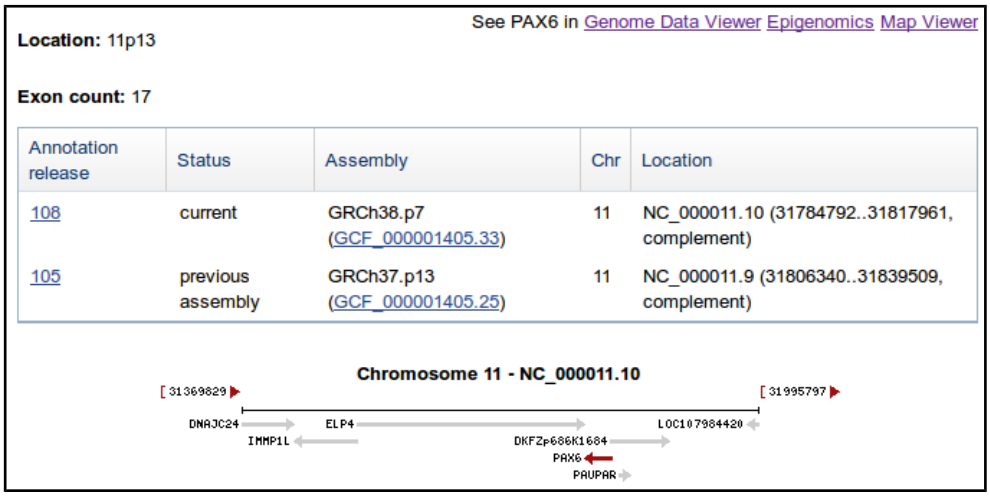
- there are two major domains, a paired domain and a homeobox, both of which bind DNA
- the gene regulates transcription (is a transcription factor)
- there is more than one protein isoform, and thus more than one transcript variant.

Summary 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]

1 This despite **WT1** being at the top of the list? This is a new promotion for **WT1**. For years it has been but a close second to **PAX6**. Whilst congratulations are clearly in order, this elevation is jolly inconvenient for the story I wish to reveal. So … I intend to ignore it!

From the **Genomic context** section it can be seen that:

- **PAX6** is situated on **Chromosome 11**, band **p13**
- **PAX6** is on the complementary strand relative to that chosen by **Map Viewer** to represent **Chromosome 11**
- **ELP4** (another gene in the list of human genes associated with **Aniridia**) is exceedingly close, on the opposite strand to **PAX6**. This might be worthy of a glance, at a later time?
- There are **17** exons for **PAX6**. Jolly good, but I really wanted to know how many transcripts there were according to the **NCBI**? That is, how many different ways it is thought that nature spliced the **17** exons together. I would also like to discover how many distinct **isoforms** the **NCBI** imagines to result from however many **transcripts**. I proceed with impatience!



Click either the Genome Data Viewer or the Map Viewer link. Both offer essentially the same story, the choice really is cosmetic. Do you like your genomes vertical or horizontal. I am a horizontal man myself, so I prefer the Genome Data Viewer. The data is from the **Map Viewer Genome Database**, whichever choice you make.

I reproduce both views here. The **Genome Data Viewer** picture is included in the **PAX6** gene page for free, so maybe the **MapViewer link** is the best one for you to choose? Or both, of course! First consider the marginally clearer and simpler **Genome Data Viewer** picture.

So, if I tell you the region displayed is the entire **PAX6** region of **Chromosome 11** and the green lines labelled on the right as something beginning with **NM_** represent the different transcripts, can you now say how many transcripts there are according to this view? In passing, the blobs along each line represent the exons. Dark blobs are coding exons. Light blobs represent the exons that form the **3'/5' UTR** regions of each transcript. The Introns are the pale green lines joining the blobs together.

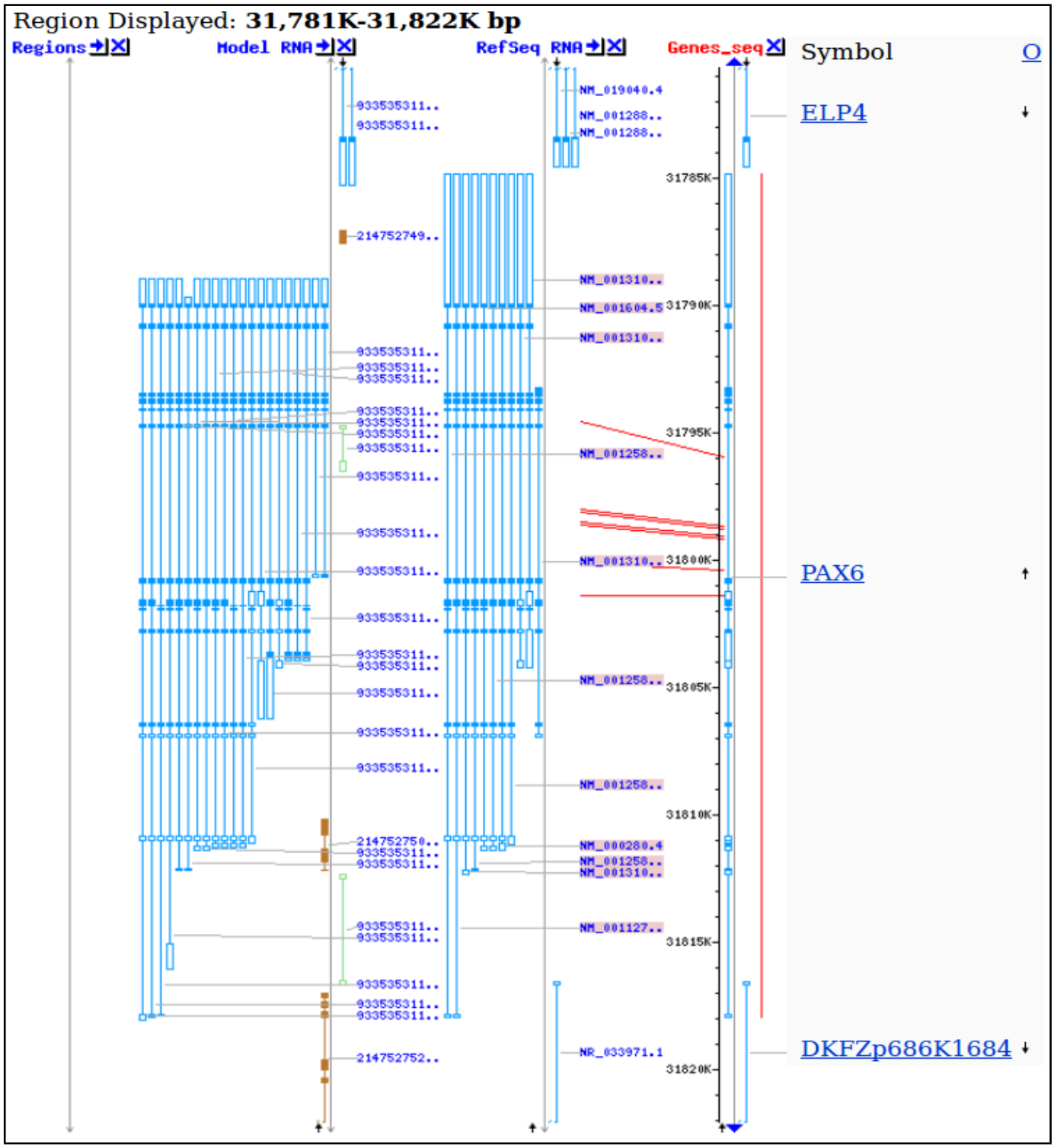
The prediction of the transcripts shown here are based on database searches of all Human mRNA sequences stored in **RefSeq** against this region of the genome. The theory is that every human mRNA sequence must match (nearly)

perfectly somewhere in the human genome. Where it matches, there must be the genomic DNA from which the mRNA was transcribed. How charmingly simple!

To differentiate between coding and non-coding exons of a transcript, why not compare all human proteins with the genome (after suitable translation to amino acid codes in all six reading frames). They too must match near perfectly somewhere, identifying the **Co**ding **S**equence (**CDS**) of each transcript. Transcript fully located. Job done! Of course, it does not always work so very neatly, but we need not admit that for the moment at least.

Comparing proteins with the genome is clumsy, compute intensive, slow. For major organisms (currently just Human and Mouse), specially comprehensive databases of extremely reliable **DNA Coding Sequences** have been constructed. Searching with these databases enables much more efficient searching for coding exons and so is very much preferred.

And so to the **Map Viewer** version of exactly the same region of **Chromosome 11**.

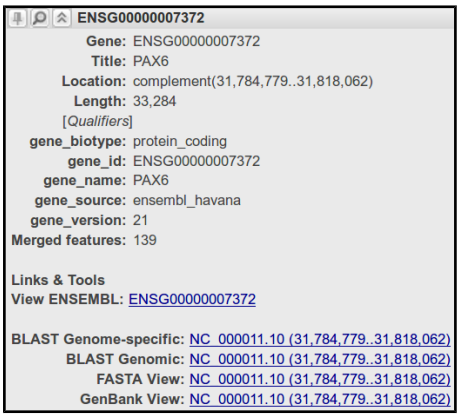


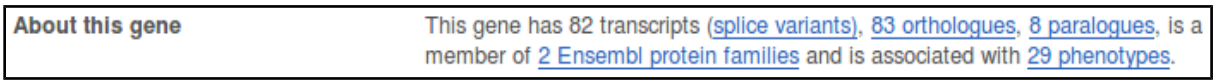
OK, times up, how many transcripts are predicted for **PAX6** by **MapViewer**?

11 being the correct answer. Obviously? Exactly as suggested by the **Genome Data Viewer**! True, but this is not always the case. The transcript count (and much else) depends on the version of the data used to build the views.. Quite recently, these two viewers displayed the interpretation of different data versions. **Mapviewer** being slightly behind the times. When this was the case, the transcript count depended upon which viewer was chosen. This vitally illustrates that many of the "facts" presented by these services are but ***predictions*** that will vary as more/better data become available. Pretty good predictions, but nevertheless, ***predictions***!

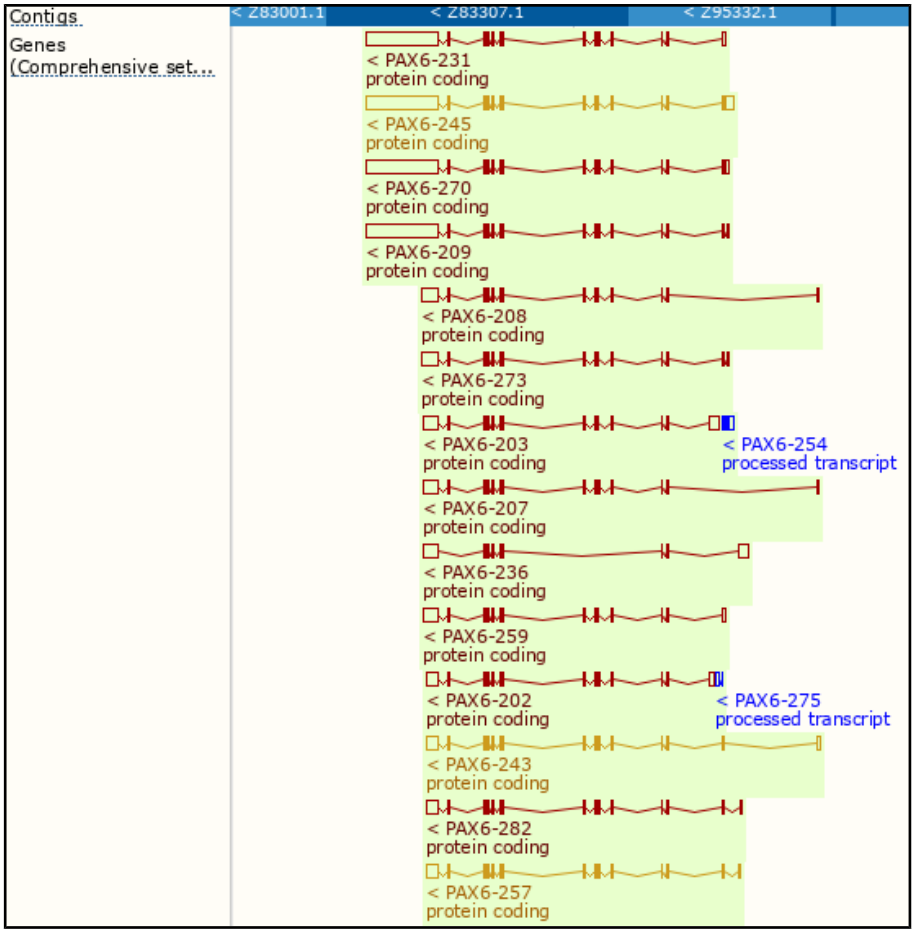
In passing, the reason that there used to be a difference in transcript counts between the two viewers was that **MapViewer** used an older version of **RefSeq** than the **Genome Data Viewer**. The older **RefSeq** included some extra mRNA sequences of less certainty than the ones you see represented above. Clearly, the evidence for these extra mRNA sequences was proved insufficient and they were removed in the newer **RefSeq**. Where they exist, such less certain **RefSeq** mRNA sequences can be recognised easily as their labels (**Accession Codes**) which begin with **XM_** rather than **NM_**. I make a point of mentioning this as the inclusion of data of varying credibility, in databases such as **RefSeq**, is very common. Usually, the difference in confidence is that between database entries that are only detected by computer programs (questionable) and those that have been properly investigated by human experimenters/investigators (less questionable).

Even without database version variation, seemingly trivial inquiries such as "how many transcripts are there?" can still yield conflicting answers depending upon where the question is asked. Move back to the page describing the **PAX6** gene. In the familiar graphic at the top of the **Genome regions, transcripts and products** section you will find routes to corresponding information from the **Ensembl Genome Database**. Hover over the **PAX6** (also known as **ESNG00000007372**, by **Ensembl** and close friends) green line in the bottom half of the picture. You will be rewarded by cheery gray box full of links to **Ensembl** and other exciting places.



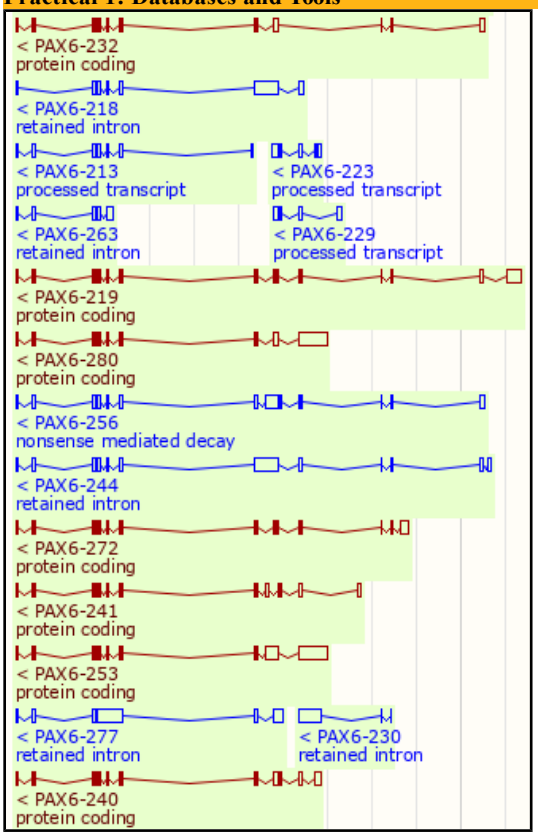


Use the link labelled **View ENSEMBL**:. A view of the region of **Chromosome 11** similar to those you have already considered will leap forth. As before, the exons for each transcript are represented by blobs (filled for coding, empty for **UTR** regions). Introns are represented by wiggly lines joining the blobs. Notice first that there are considerably more than **11** transcripts represented here! At the top of the page, in tiny letters it claims **82!** (a massive increase even from the **31** transcripts predicted by a recent previous version of **Ensembl!**).



You ***could*** check this assertion by counting all the transcripts represented in the graphic, but I would not recommend doing so. Sometimes it is best just to believe. There are indeed **82**.

The colour scheme used for the transcripts we might discuss in overview later. For now, just know that the gold transcripts are supported by better evidence than the red ones. Once more a database that offers data items of varying credibility.



Looking a little further down the transcripts displays, you will see that an increasing proportion of the transcripts are not **protein coding** (the blue ones). Both of the displays you examined at the **NCBI** only represented protein coding transcripts. This partially explains why **Ensembl** appears finds so many more transcripts that its broad alternatives.

So a further reason for not finding a consistent answer to the simple question "How many transcripts are there for the **PAX6** gene" is variation in the ***definition*** of a transcript.

Also, and more importantly, **Ensembl** and **MapViewer** use different strategies to predict transcripts (and just about everything else!). Both use database searches in roughly the manner described previously and (for the human genome at least) the same basic assemblies of the genome and sequence databases. It is the interpretation of the data and analytical results that varies.

The database searches used as the fundamental strategy to identify transcripts take a very long time to execute, even given the immense computing resources available to the **NCBI** and the **Ensembl** teams. Some clever strategies are employed to minimise the time spent on these searches. It would be good to consider these, specifically with respect to their implementation by **Ensembl**, at least superficially.

Show/hide columns (1 hidden) Filter

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
PAX6-245	ENST00000638914.1	7300	422aa	Protein coding	CCDS31451	P26367 Q66SS1	NM_000280 NM_001258464 NP_000271 NP_001245393	TSL:1 GENCODE basic APPRIS ALT1
PAX6-270	ENST00000640368.1	6975	436aa	Protein coding	CCDS31452	F1T0F8 P26367	-	TSL:5 GENCODE basic APPRIS P4
PAX6-231	ENST00000606377.6	6901	436aa	Protein coding	CCDS31452	F1T0F8 P26367	-	TSL:5 GENCODE basic APPRIS P4
PAX6-209	ENST00000419022.6	6888	436aa	Protein coding	CCDS31452	F1T0F8 P26367	NM_001604 NP_001595	TSL:1 GENCODE basic APPRIS P4
PAX6-203	ENST00000379109.7	3182	422aa	Protein coding	CCDS31451	P26367 Q66SS1	-	TSL:2 GENCODE basic APPRIS ALT1
PAX6-273	ENST00000640610.1	2730	422aa	Protein coding	CCDS31451	P26367 Q66SS1	-	TSL:1 GENCODE basic APPRIS ALT1
PAX6-259	ENST00000639916.1	2622	422aa	Protein coding	CCDS31451	P26367 Q66SS1	NM_001258465 NP_001245394	TSL:1 GENCODE basic APPRIS ALT1
PAX6-243	ENST00000638903.1	2620	436aa	Protein coding	CCDS31452	F1T0F8 P26367	NM_001258462 NP_001245391	TSL:1 GENCODE basic APPRIS P4
PAX6-207	ENST00000379129.7	2614	436aa	Protein coding	CCDS31452	F1T0F8 P26367	-	TSL:5 GENCODE basic APPRIS P4
PAX6-202	ENST00000379107.7	2579	436aa	Protein coding	CCDS31452	F1T0F8 P26367	-	TSL:5 GENCODE basic APPRIS P4
PAX6-208	ENST00000379132.8	2576	422aa	Protein coding	CCDS31451	P26367 Q66SS1	-	TSL:5 GENCODE basic APPRIS ALT1
PAX6-282	ENST00000640975.1	2553	436aa	Protein coding	CCDS31452	F1T0F8 P26367	NM_001310158 NP_001297087	TSL:1 GENCODE basic APPRIS P4
PAX6-257	ENST00000639409.1	2450	436aa	Protein coding	CCDS31452	F1T0F8 P26367	NM_001258463 NP_001245392	TSL:1 GENCODE basic APPRIS P4
PAX6-201	ENST00000241001.13	1736	422aa	Protein coding	CCDS31451	P26367 Q66SS1	NM_001127612 NP_001121084	TSL:1 GENCODE basic APPRIS ALT1
PAX6-235	ENST00000638629.1	2844	286aa	Protein coding	-	A0A1W2PRA8	NM_001310160 NP_001297089	TSL:2 GENCODE basic

For a more detailed view of the predicted transcripts, click on the Show transcript table link. The transcript predictions are now presented in the form of a table. The protein coding transcripts are all at the top of the table. I counted **56**, but I would not claim to be completely accurate, I got a bit confused half way down the list! Lots more than the **NCBI** anyway.

Ensembl uses both the sequences of **RefSeq** mRNAs and those of their protein products (the **RefSeq** entries whose **Accession Code**s commence **NP_**) to predict transcripts, however, **Ensembl** appears to have less blind faith in the accuracy of these data than the **NCBI**.

We should briefly discuss the significance of the first **14** transcripts associating with entries in the **CCDS** database. Particularly noting that **14** is far closer to **11** and there are just **2** distinct **CCDS** entries referenced here.

Note that there is no "one to one" correspondence between **RefSeq** mRNAs and transcript predictions. All **11** **RefSeq** mRNAs are referenced, but ***two*** are used to support the single first transcript in the list. If **Ensembl** regarded **RefSeq** mRNAs as "perfect" (as the NCBI appears to do) this would clearly be a nonsense!

PAX6-220	ENST00000525535.2	875	3aa	Protein coding	-	-	-	CDS 3' incomplete TSL:5
PAX6-260	ENST00000639920.1	676	72aa	Protein coding	-	A0A1W2PR58	-	CDS 3' incomplete TSL:5
PAX6-256	ENST00000639394.1	1988	163aa	Nonsense mediated decay	-	A0A1W2PQW3	-	TSL:5
PAX6-227	ENST00000533156.2	848	No protein	Processed transcript	-	-	-	TSL:5
PAX6-213	ENST00000464174.6	846	No protein	Processed transcript	-	-	-	TSL:5
PAX6-222	ENST00000530373.6	785	No protein	Processed transcript	-	-	-	TSL:5
PAX6-223	ENST00000530714.6	650	No protein	Processed transcript	-	-	-	TSL:5
PAX6-267	ENST00000640251.1	649	No protein	Processed transcript	-	-	-	TSL:5
PAX6-229	ENST00000534353.5	540	No protein	Processed transcript	-	-	-	TSL:4
PAX6-254	ENST00000639203.1	532	No protein	Processed transcript	-	-	-	TSL:5
PAX6-233	ENST00000638278.1	417	No protein	Processed transcript	-	-	-	TSL:NA
PAX6-275	ENST00000640617.1	412	No protein	Processed transcript	-	-	-	TSL:4
PAX6-279	ENST00000640819.1	368	No protein	Processed transcript	-	-	-	TSL:5
PAX6-228	ENST00000533333.5	6173	No protein	Retained intron	-	-	-	TSL:2
PAX6-216	ENST00000474783.2	4392	No protein	Retained intron	-	-	-	TSL:5
PAX6-214	ENST00000470027.7	3587	No protein	Retained intron	-	-	-	TSL:2
PAX6-265	ENST00000640172.1	2525	No protein	Retained intron	-	-	-	TSL:5

We should discuss why it is reasonable to not regard a match of a **RefSeq** mRNA with the **Genome** as, by itself, sufficient evidence to uniquely predict a transcript.

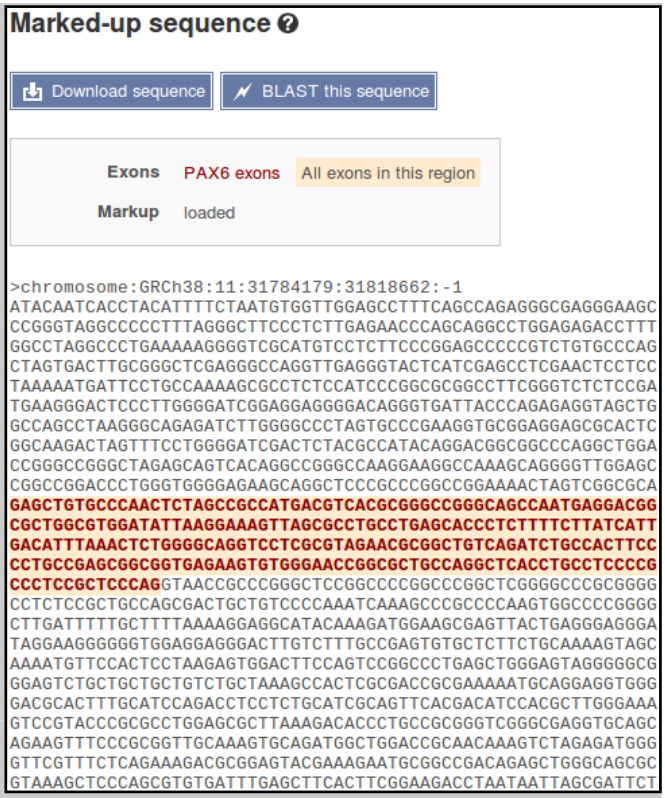
Looking further down the list you will see that many **Ensembl** protein coding transcripts are predicted without reference to any **RefSeq** entry.

Hover over the evidence **Flags** associated with the transcript predictions towards the end of the list. How reliable would you judge these predictions to be?

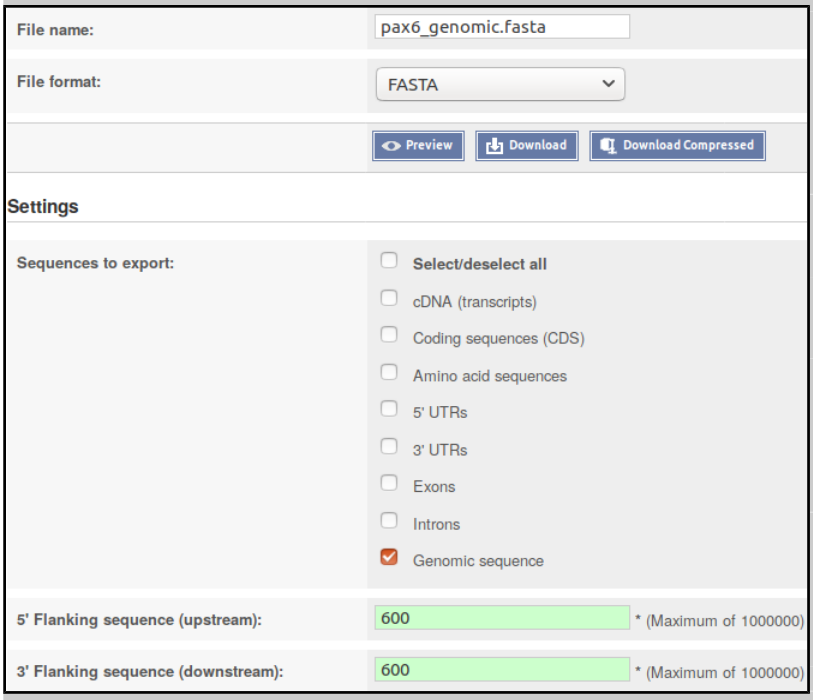
We could go on. Other sources (not necessarily **Genome Databases**) would count the transcripts differently again. Perhaps the best answer to the question "How many transcripts are there for the **PAX6** gene" is "**Several**".

Before leaving **Ensembl**, it would be good to save the genomic sequence of this region for analysis later on.

To do this, first click on the **Sequence** link on the left hand side of the page. Under the transcript table the sequence of the **PAX6** region of the genome will be displayed. The exons will be tastefully highlighted for you delectation. The display includes **600** base pairs of **Flanking Sequence** (**3'** and **5'**) which are included (by default) when the sequence is downloaded.



Now chose to Download sequence. The **Download sequence** form will burst into view.



Set the **File name:** to `pax6_genomic.fasta`

Set the **File format:** to **FASTA**

Accept the default **600** base pairs for both the **5'Flanking sequence (upstream):** and the **3´ Flanking sequence (downstream):**.

Finally, click on the Download button and do whatever it takes to move the file you create to somewhere sensible on your **Desktop**.

Using whatever text editor is most convenient, edit your file to:

1. Remove the many blank lines at the top of the file. These serve no purpose, but are not really a problem. They are, however ugly!
2. Change the first word of the first line of the file to contain information, from **11** to **pax6_genomic**. This first word is defined as the sequence identifier in **FASTA** format (as, I hope, will be explained at some point). **pax6_genomic** is a far more informative identification than **11** (simply the Chromosome number).

The next investigation might be too discover "How many protein isoforms might there be for **PAX6**?".

Well, whilst the **Ensembl** transcript list is still in view, glance down the **Protein** column which displays the size of the protein products for each transcript. Clearly insufficient evidence for a serious **isoform** count, but enough to set a lower limit, as the same **isoform** cannot be more than one length! If there were not so very many! One might count how many different lengths of proteins were listed. I tried to do this, but I gave up around **twenty-something**. Let us be content to declare that there are **lots**. The most likely looking ones are either **422** or **436** amino acids long. Some of the others might cause a raised eyebrow or two, especially the one that is **3** amino acids long (second to last **Protein coding** entry in the list)? But, who are we to question! **Lots** is the informal **Ensembl** minimum total.

Click your way back to the NCBI PAX6 gene entry. So, now to discover the number of protein products (**isoforms**) that the **NCBI** predicts. This view makes this simple question clumsy to answer as the protein products of each transcript are reported separately (as they are by **Ensembl**), even when they are identical???

However, it can be done. Click on the NCBI Reference Sequences (RefSeq) link in the **Table of contents** on the right hand side of the page. Focus on the **mRNA and Protein(s)** sub-section. Skim down the entries for every transcript. Check the different isoform names. I see:

```
01 - NM_000280.4     → NP_000271.1     paired box protein Pax-6 isoform a
02 - NM_001127612.1 → NP_001121084.1  paired box protein Pax-6 isoform a
03 - NM_001258462.1 → NP_001245391.1  paired box protein Pax-6 isoform b
04 - NM_001258463.1 → NP_001245392.1  paired box protein Pax-6 isoform b
05 - NM_001258464.1 → NP_001245393.1  paired box protein Pax-6 isoform a
06 - NM_001258465.1 → NP_001245394.1  paired box protein Pax-6 isoform a
07 - NM_001310158.1 → NP_001297087.1  paired box protein Pax-6 isoform b
08 - NM_001310159.1 → NP_001297088.1  paired box protein Pax-6 isoform c
09 - NM_001310160.1 → NP_001297089.1  paired box protein Pax-6 isoform d
10 - NM_001310161.1 → NP_001297090.1  paired box protein Pax-6 isoform d
11 - NM_001604.5     → NP_001595.2     paired box protein Pax-6 isoform b
```

I count **4** different isoforms, imaginatively named **Isoform a**, **Isoform b**, **Isoform c** and **Isoform d**. One associated with each transcript description. Look carefully at the annotations and there is more information. In particular:

Description:

Isoform b is also known as **Isoform 5a**. Why this is interesting will become apparent in a page or so.

Isoform b is also reported to be longer than **Isoform a**.

Description	Transcript Variant: This variant (5) differs in the 5' UTR and includes an alternate in-frame exon in the 5' coding region, compared to variant 1. The encoded isoform (b, also known as 5a) is longer than isoform a. Variants 2, 4, 5 and 8 encode the same isoform (b).

Conserved Domains:

Conserved Domains (2) summary

smart00351 Location:4 → 128	PAX; Paired Box domain
pfam00046 Location:214 → 266	Homeobox; Homeobox domain

Both **Isoform a** and **Isoform b** are recorded as having two domains. A Paired Box Domain at the beginning, and a Homoebox Domain further along.

Conserved Domains (2) summary

smart00351 Location:4 → 142	PAX; Paired Box domain
pfam00046 Location:228 → 280	Homeobox; Homeobox domain

Both **Paired Box Domains** are primarily indicated by a hit with the relevant entry in the SMART database. Both **Homeobox Domains** are supported by matches with Pfam database entries. Other domain databases will almost certainly provide supporting evidence, but reference to just one match is sufficient here.

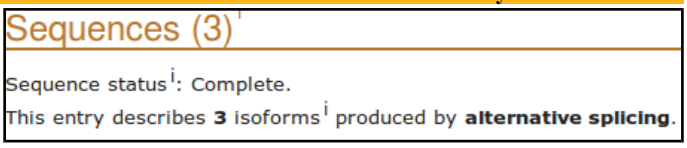
From the location information, the **Paired Box** of **Isoform b** appears to include an extra **14** amino acids.

UniprotKB offers yet another version of this story. Just for a for a few clicks, let us intrude into the **UniProtKB** section of your course.

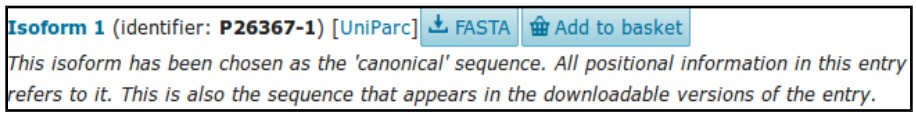
At the very bottom of the current page, you will find a link to **UniprotKB**. Use it.

Protein Accession	Links	
	GenPept Link	UniProtKB Link
P26367.2	GenPept	UniProtKB/Swiss-Prot:P26367.2

Lo! the **PAX6** human protein as seen and understood by **UniProtKB**. Click on the Sequences (3) button on the left hand side of the page. **UniProtKB** declares **3** isoforms! At least, **3** that it is willing to admit to publicly.

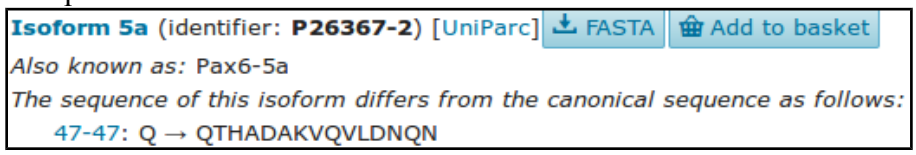


There is **isoform 1**, also known as **isoform a** in America. Note that this is the "*canonical sequence*" for this protein. That is, this is the isoform used to represent this protein in **UniProtKB**. The sequence(s) of all other isoform(s) are recorded as elements of the annotation.



Also we have **Isoform 5a** (or **PAX6-5a**), also known as **isoform b** in America (where it also answers to **Isoform 5a** when pressed). Note that the entry declares the sequence difference to be:

47-47: Q → QTHADAKVQVLDNQN



Literally:

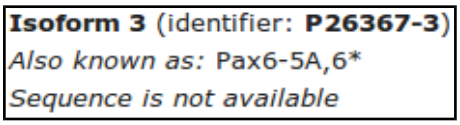
> "The amino acid at **position 47** is a **Q** in the canonical sequence. In **isoform 5a** this is replaced by the **15** amino acids **QTHADAKVQVLDNQN**".

More coherently this amounts to:

> "**isoform 5a** differs from the canonical **isoform 1** in that it has an insertion of **14** amino acids after the **47th** amino acid (a **Q**) of the canonical protein".

It is significant to note that position **47** is right in the middle of the **Paired Box Domain** that occurs in both isoforms. This confirms that which was noticed at the **NCBI**.

Finally **UniProtKB** proudly presents the somewhat ephemeral **isoform 3** (or **PAX6-5A,6*** for those who enjoy formality). But, this one has no known sequence? Not much that Bioinformatics can offer here methinks.

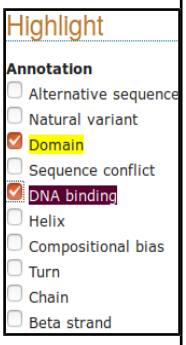


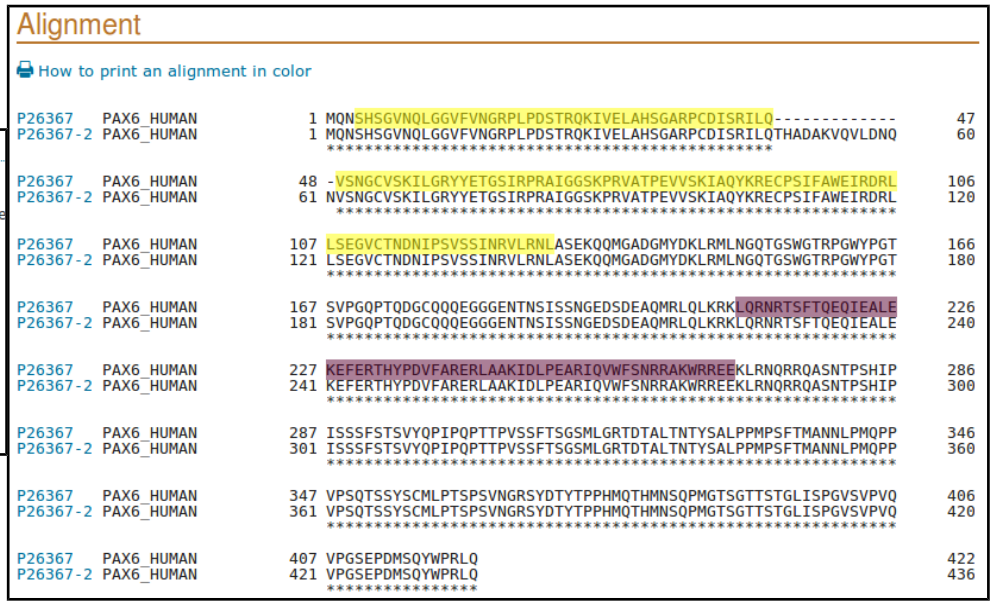
So I hope you will agree that the **UniProtKB** count stands at a very modest **2**, plus a ghost.

To visualise the differences between the **2** isoforms with sequence, click on the Align button at the top of the **Sequences** section. After deep thought and much fumbling, **UniProtKB** will multiply align all the isoform sequences for you. As there are only **2** in this case, this will appear very similar to a **Pairwise** alignment. Highlight the **DNA binding** regions and the **Domains**.

I leave the interpretation of this splendid display to you, and later short discussion if required.

The extra **14** amino acids of **isoform 5a** are due to the inclusion of a tiny extra (**42** base pair) exon in some transcripts.





Can you see the evidence for this assertion in the regional genomic maps of a few pages back?

We need to save a some protein sequences for future analysis. This is easiest from **UniProtKB** so now is good. To declare your intention to save the entire canonical version of the **PAX6** protein to a file, move back from your alignment. Move to the top of the page where you will find the bizarre invitation to Add to basket ? Just do it.

You also need to download the sequences of both domains is separate files, via your basket. First the **Paired Box**. Click the Family & Domains button on the left of the page. Then use the Add button adjacent to the **Paired** entry. Its now in your basket you will be ecstatic to know.

Feature key	Position(s)	Description	Actions	Graphical view	Length
Domain	4 – 130	Paired PROSITE-ProRule annotation	Add BLAST		127

As they are so conveniently in view, take note of the **Compositional bias** features. They will be of interest when we look at database searching.

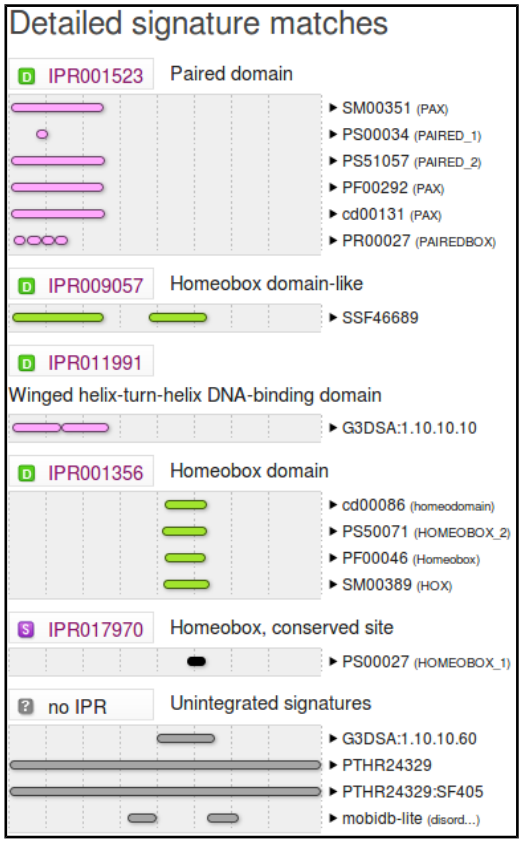
Feature key	Position(s)	Description	Actions	Graphical view	Length
Compositional bias	131 – 209	Gln/Gly-rich	Add BLAST		79
Compositional bias	279 – 422	Pro/Ser/Thr-rich	Add BLAST		144

Natural variant (VAR_008694)	29	I → S in AN.	1 Publication
Natural variant (VAR_003811)	29	I → V in AN.	1 Publication
Natural variant (VAR_008695)	33	A → P in AN.	1 Publication
Natural variant (VAR_008696)	37 – 39	Missing in AN.	1 Publication
Natural variant (VAR_008697)	42	I → S in AN; mild.	1 Publication
Natural variant (VAR_008698)	43	S → P in AN.	1 Publication
Natural variant (VAR_003812)	44	R → Q in AN.	1 Publication

Then take an excursion to glance at the Pathol./Biotech section. Note the many **Natural variant**s recorded as responsible for **AN** (**AN**iridia, that is). Particularly those around amino acid positions **29** to **44** and specifically that at position **33**.

Looking at PCR Primer Design later, you will be attempting to create a **PCR** products from patients that, when sequenced, will determine the presence or absence of this variant.

Next, skip nimbly to the Family & Domains section. Concentrate on the **Family and domain databases** sub-section. Here are displayed the results of comparing the **PAX6** protein with many of the available **Domain/Motif Databases**, including those of the Interpro Consortium, collectively.



Are the results broadly as you might expect?

For an effective graphic summary, link to View protein in InterPro for the **Interpro** graphical results. If the detail is not entirely transparent, this result will be discussed further when you generate it for yourselves using **Interpro**.

The results you are looking at are computed, largely automatically, by the **UniProtKB/Interpro** annotation system. However, running many of the same analyses manually is trivial. Maybe you will do some in the course of these exercises?

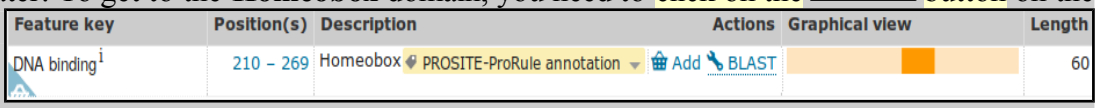
Finally, return to the **UniProtKB PAX6** page and move to the Structure section.

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

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

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

Back to saving sequences for later! To get to the **Homeobox** domain, you need to click on the Function button on the left hand side of the page.



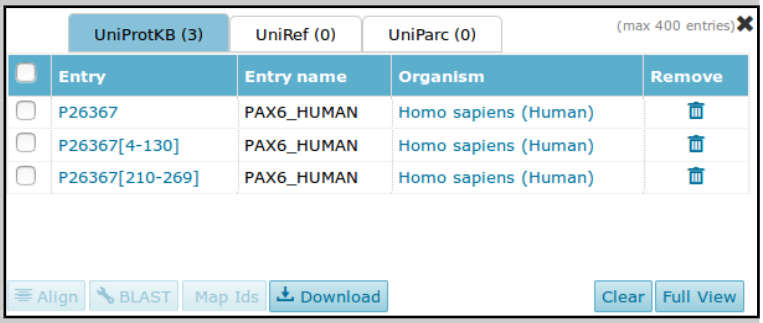
A valid question at this point might be "Why is the **Homeobox** domain a **Function** (specifically a **DNA binding** feature), but the **Paired** domain is a **Domain** feature?" To which the answer is "*History, dear boy, history*" to paraphrase a disputed quote of dear Harold (Macmillan that is).

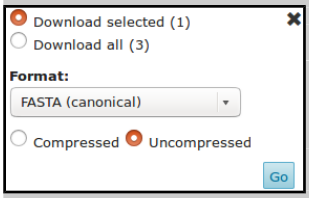
In fact, both are **Domains**, and both are **DNA binding**. The illogicality of them being recorded in different places is accepted, however, to fix this early mistake now would not, it is claimed, be trivial. So, we live with it. So doing, click on the appropriate Add button and then prepare to head for the checkout desk (Good Grief! I am beginning to get used to this!).

Shimmy back to the top of the page. You should have Basket 3 things in your basket.

Click on the basket to view your booty.

For each of the 3 items in turn (not all at once or you get all sequences in one file), select and Download .



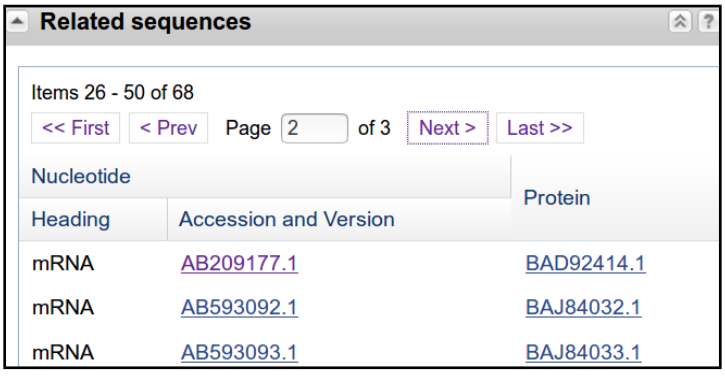


Each time ensure the download parameters are set to **Uncompressed** and **FASTA (canonical)**. Then click the Go button.

The next few steps, as before, are very browser/OS dependant. Just do whatever it takes to save the three sequences in files called, as appropriate:

```
pax6_human.fasta

pax_domain.fasta

homeobox_domain.fasta
```

Now move back to America to the **NCBI** view of the **PAX6** gene. If you have problems getting there … click here.

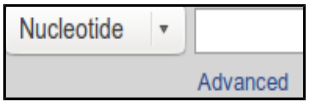


Near the bottom of the page, there is a section called **Related sequences**. Move to the second page (of three) of the list of sequences. Click on the top entry, the mRNA called AB209177.1. You will be rewarded by a **GenBank** entry in GenBank format.

Formats are tedious, but we will discuss them briefly at some point. You have already seen FASTA format. We will bump into EMBL format at some point. The other **137** or so formats are to be ignored!

Can you see the official gene name **PAX6**, mentioned in this entry? The **Gene Name** field (where **PAX6** should most certainly be mentioned) is entirely missing! If you searched **GenBank** (or **EMBL** come to that) for this sequence using the most obvious search **Keyword**, that is **PAX6**, do you think you would find this **PAX6** mRNA? You clearly should! A case for more consistent annotation? Perhaps something to consider further when we superficially mention the Gene Ontology project later.

Next, search the **Nucleotide** databases, by textual **Keyword**, for **PAX6** related sequences and down load one or two for investigation. To achieve this worthy goal, move to the top of the current page and note that the database selection has changed from **Gene** to **Nucleotide**. 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:

paired box 6

You are asking Entrez to search for all **Nucleotide** database entries that contain the terms "**chromosome 11**" and "**paired box 6**" 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 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**.

```
LOCUS       NG_008679              40170 bp    DNA     linear   PRI 11-SEP-2017
DEFINITION  Homo sapiens paired box 6 (PAX6), RefSeqGene (LRG_720) on
            chromosome 11.
```

Take a look at the FEATURES for this entry. You will see that there are **three** genes mentioned. **PAX6**, of course. Also, on the strand that is the complement of that represented here, there is **PAX6-AS1** and **ELP4**.

```
gene            complement(<1..6396)
                /gene="PAX6-AS1"
                /gene_synonym="DKFZp686K1684"
                /note="PAX6 antisense RNA 1"
                /db_xref="GeneID:440034"
                /db_xref="HGNC:HGNC:53448"
gene            5001..38170
                /gene="PAX6"
                /gene_synonym="AN; AN2; ASGD5; D11S812E; FVH1; MGDA; WAGR"
                /note="paired box 6"
                /db_xref="GeneID:5080"
                /db_xref="HGNC:HGNC:8620"
                /db_xref="MIM:607108"
```

```
gene            complement(38437..>40170)
                /gene="ELP4"
                /gene_synonym="AN; AN2; 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"
```

Can you find the additional genes **PAX6-AS1** and **ELP4** in the genome displays you have looked at so far?

```
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 protein Pax-6; 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="CCDS: CCDS31451.1 "
/db_xref="LRG:p1"
/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, Analyse **this sequence** by clicking on the Highlight Sequence Features option. The **Co**ding **D**NA **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).

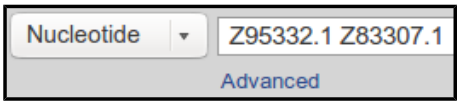
What were the features that you found?

Why might you have expected more features than there were?

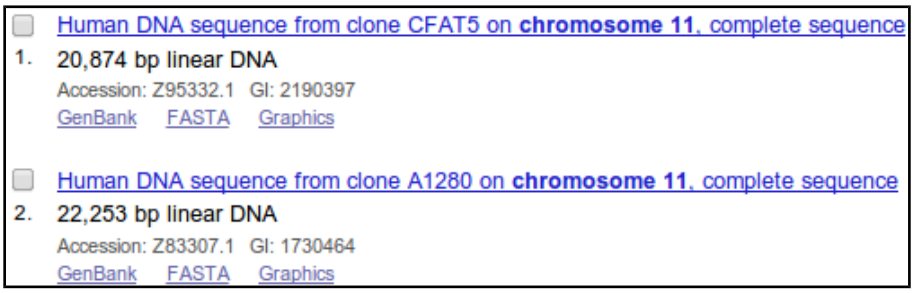
```
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff in
            collaboration with Isabel Hanson, David FitzPatrick. The reference
            sequence was derived from Z95332.1 and Z83307.1.
            This sequence is a reference standard in the RefSeqGene project.

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 the alignment of two sequences from GenBank. The two aligned sequences being "**contigs**", that is products of two individual sequencing projects of separate portions of the **PAX6** genomic region. We should discuss role of "**contigs**" in the human genome project, a little.

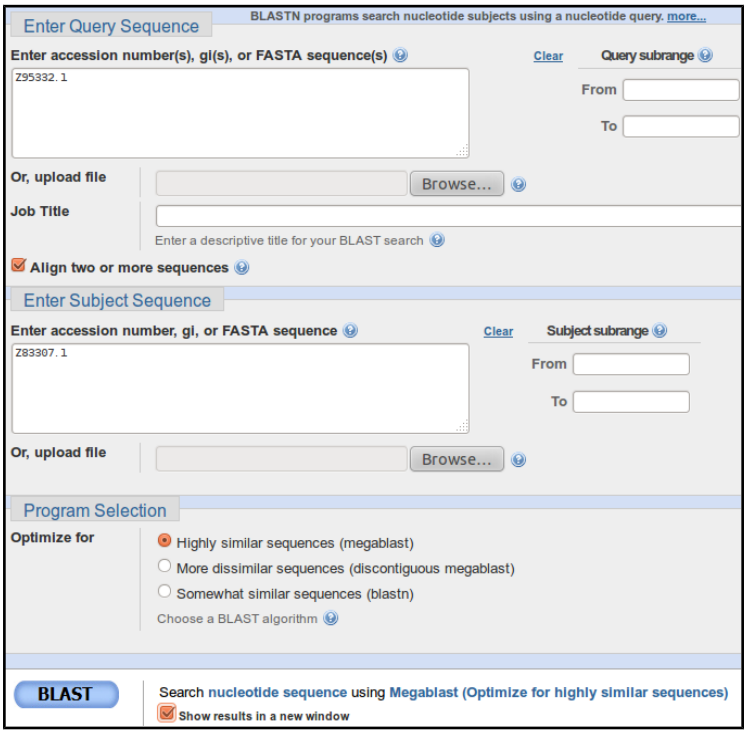
Take a quick look at the **GenBank** entries by entering their ACCESSION numbers (be sure to include the "**.1**", the version number, at the end to avoid unwanted hits) into the **Search** box at the top of your page. Click on the Search button.



Lo and behold, the two GenBank entries are summoned forth. Take a look at one or both. Not particularly illuminating I think[2]. These are clones sequenced as part of the Human Genome Project (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 **Analyse 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**[3]. 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**[4]. 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
```

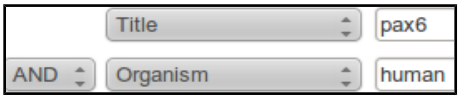
How does the alignment you generated match up with the annotation of the original **RefSeq** entry you discovered?

2 The annotation is very sparse which makes these entries very hard to find directly. The EML-Bank versions include some links to Ensembl codes. These would have been helpful but are not part of the official International Nucleotide Sequence Database Collaboration (INSDC) annotation that should be consistent between GenBank, European Nucleotide Archive (ENA), which includes EML-Bank, and DNA Data Bank of Japan (DDBJ).

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

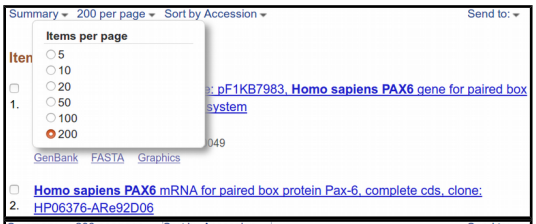
4 Just because its neater. In my, significantly less then humble, opinion anyway.

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.

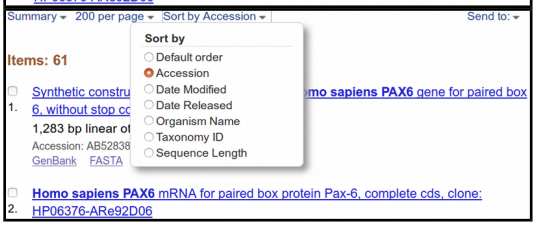


Set up a new search as illustrated and set it going. Ultimately simple this time. You have requested all **Human** sequences that are centrally associated with the gene **PAX6**.

A list of **60** or so sequences, all clearly claiming **PAX6** association and announcing their humanity loudly in Latin, will tumble forth.



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



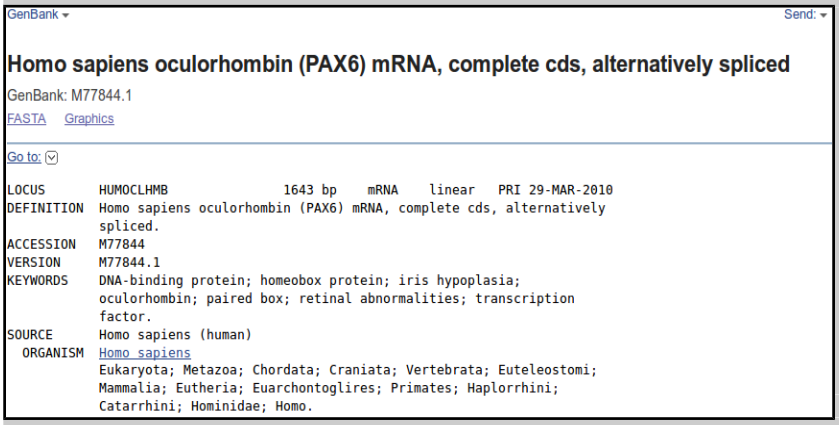
To deal with both of these issues, use the display control pull down menus at the top of your page to set the items **per page** to something big and the **Sort by** option to something sane.

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 AB209177 was justifiably referenced in the **PAX6 Gene** entry but will not be in this list. **PAX6** appears nowhere in the annotation of AB209177 including its **DESCRIPTION** (or **Title**) field.

Move far down the list, you will come to the **RefSeq PAX6** mRNAs of a few pages back. Just before these entries is M77844.1. Save this one for later analysis. I choose M77844.1 as it includes a few variations that will add interest. Select the target sequence.

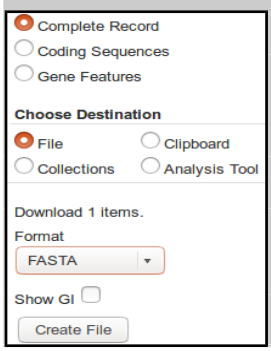
You could now use the diminutive 

button which is near the bottom of your page to download all the selected sequences into a single file.



However, as there is only one sequence, and it would be so nice to be introduced properly before such intimacies as "downloading". Click on the link to the database entry to see it in all its **GenBank Format** glory.

The sequence is for analysis rather than decoration, so use the format menu at the top of the page (currently set **GenBank**), and ask for **FASTA** format.



Now click the tiny **Send to:** button and **Choose Destination** to be **File**.

Strike the **Create File** button with a firm resolve. With irritating presumption, the choice of file name is made for you. Your sequence will be 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_mrna.fasta`

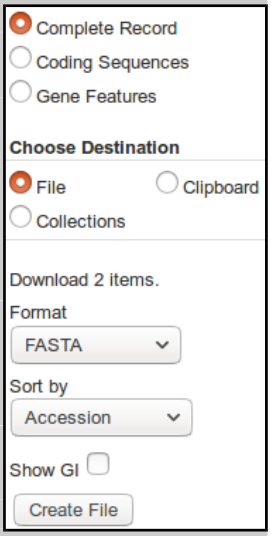
One last file to save. Move back to your list of hits and deselect the mRNA that you have already saved.

Near the top of the list you should find 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 Send: button which is at the top 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 confident click of your every ready mouse. Once more, the choice of file name is made for you. Your sequences are stored in a file named:

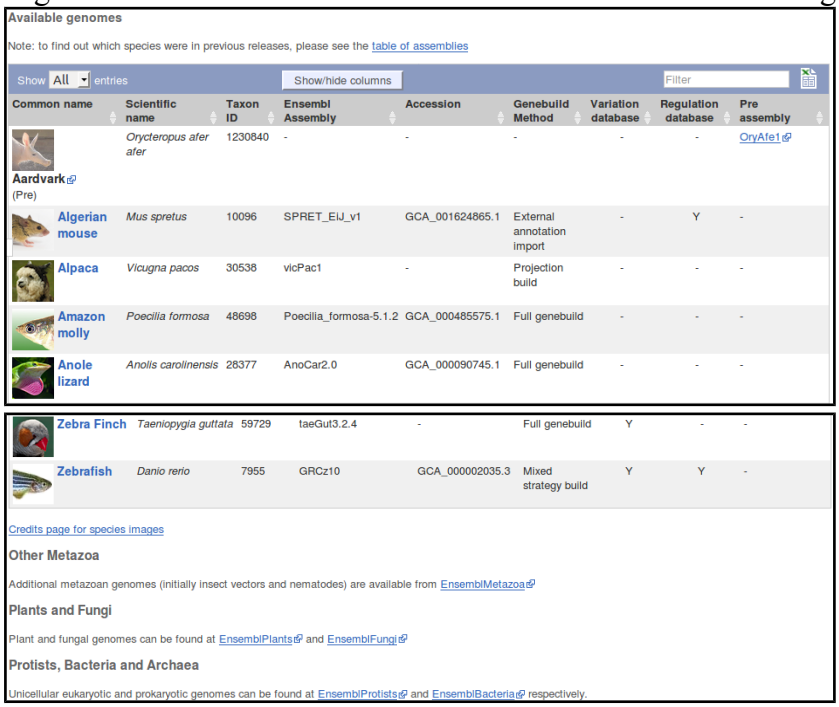



```
sequence.fasta
```

Do whatever it takes to rename this file to be called:

```
pax6_primers.fasta
```

Back to **Ensembl**. More with the objective of looking at more sources of information via **Ensembl** than becoming expert **Ensembl** users.



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** (and **MapViewer**, of course) 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 completely recorded. 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.

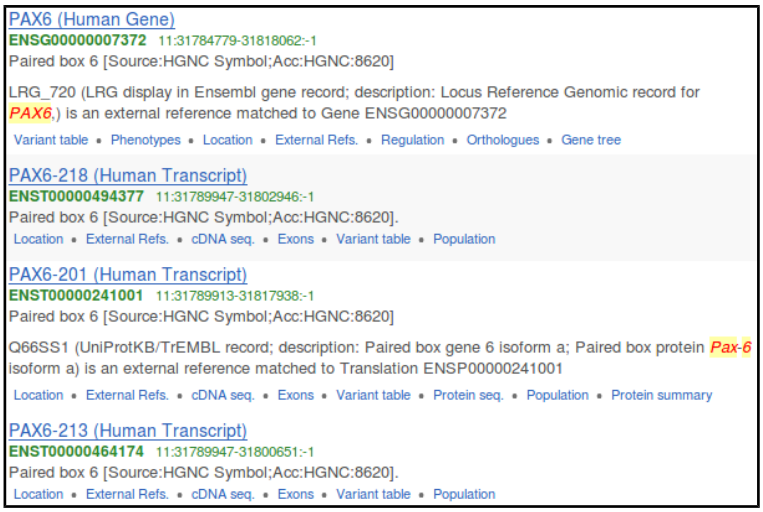
Move back to the **Ensembl** home page and go to the **Human PAX6** gene information by setting the **Search** fields as shown and clicking the **Go** button boldly.

The target gene is at the top of the hit list.

Click on the link to the **PAX6 (Human Gene)**.

You should recognise the view you now see. The list of transcripts and view of the genomic region exactly as you examined via the **NCBI**.

There is much to investigate here, but maybe that should wait for a specialised **Ensembl** course. They are run regularly in Cambridge and elsewhere.



To make a bit more space, elect to Hide transcript table .

Begin by taking a look at how **Ensembl** sees the **Homologues** of **PAX6**. First the **Orthologues** and then the **Paralogues**. 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 (from the **View Sequence Alignments** menu, select **View Protein Alignment**).

Which human **PAX6** isoform has been chosen to search for orthologues?

How do you suppose this choice might have been justified?

Once your curiosity concerning orthologue alignments is completely sated, click on the Paralogues link. View some of the protein alignments between the **PAX6** isoform and its **paralogues**.

How many **PAX** protein paralogues are there for human? Suggest a prettier naming scheme than **PAX1**, **PAX2**, **...**

Some paralogues seem to have two regions of high similarity (e.g. **PAX4** or **PAX2**), others only one (e.g. **PAX1**)? Can you explain?

Next look at some transcript specific features as they are recorded in **Ensembl**. To do this, one must first select a transcript, so Show transcript table once more and select ENST00000419022 (**PAX6-209**). Again, to make a bit more space, why not Hide transcript table away.

Now click the Exons link (from **Transcript-based displays → Sequence**). **Exons**, **Introns** and **Variations** within **Exons** are clearly displayed.

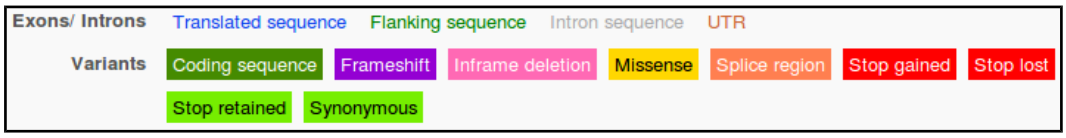
Intron 2-3	31,810,827	31,806,926			3,902	gtgagtccgcttctttcttctcgct.....ttttttctccttctgttttgtcttag
ENSE00001098662	31,806,925	31,806,849	-	-	77	GGGGAAGACTTTAACTAGGGGCGCGCAGATGTGTGAGGCCTTTTATTGTGAGAGTGGACA GACATCCGAGATTTCAG
Intron 3-4	31,806,848	31,806,463			386	gcaagttctgtggtggctgctttgg.....ttaactcctattttcttgctaacag
ENSE00002523992	31,806,462	31,806,402	-	1	61	AGCCCCATATTCGAGCCCCGTGGAATCCCGCGGCCCCCAGCCAGAGCCAGCATGCAGAAC A
Intron 4-5	31,806,401	31,802,835			3,567	gtaagtgcctctggtctttctggga.....tttcctctcctctccctttcctcag
ENSE00003602163	31,802,834	31,802,704	1	0	131	GTCACAGCGGAGTGAATCAGCTCGGTGGTGTCTTTGTCAACGGGCGGCCACTGCCGGACT CCACCGGCAGAAGATTGTAGAGCTAGCTCACAGCGGGGCCCGGCCGTGCGACATTTCCC GAATTCTGCAG
Intron 5-6	31,802,703	31,801,913			791	gtgatcctcccggcgccgccccact.....ttgaaggtatatttttgtgttatag
ENSE00003512677	31,801,912	31,801,871	0	0	42	ACCCATGCAGATGCAAAAGTCCAAGTGCTGGACAATCAAAAC
Intron 6-7	31,801,870	31,801,777			94	gtaagcttgtcattgtttaatgcat.....ttttctgtccacttcccctatgcag
ENSE00003523920	31,801,776	31,801,561	0	0	216	GTGTCCAACGGATGTGTGAGTAAAATTCTGGGCAGGTATTACGAGACTGGCTCCATCAGA CCCAGGGCAATCGGTGGTAGTAAACCGAGAGTAGCGACTCCAGAAGTTGTAAGCAAAATA GCCCAGTATAAGCGGGAGTGCCCGTCCATCTTTGCTTGGGAAATCCGAGACAGATTACTG TCCGAGGGGGTCTGTACCAACGATAACATACGAAGC

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

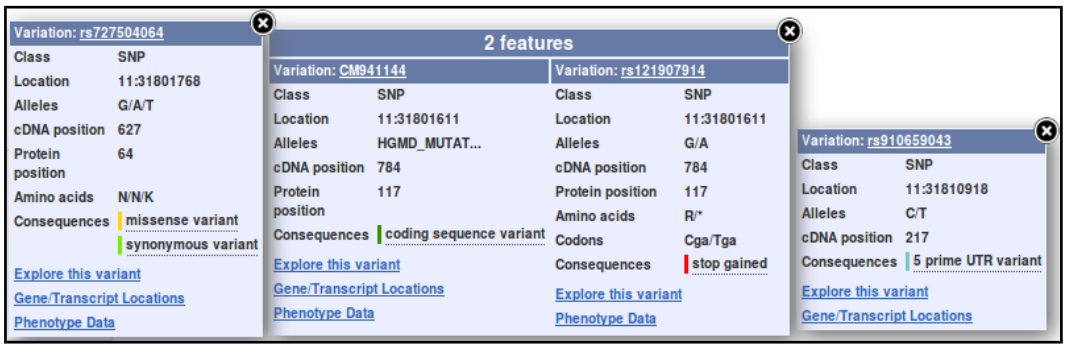
How long is the sixth exon and why would this concur with your expectations?

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

Click on some of the colourful variation locations. The colours are explained in the legend at the top of the display.



The variations come from a number of databases, including **dbSNP**. The **dbSNP** entries are those whose names begin with "**rs**". dbSNP can be investigated directly at the NCBI, of course, but it very convenient to have all the variation information built into **Genome Databases** such as **Ensembl**.



Click on the **Domains & features** link (from **Transcript-based displays → Protein Information**).

Domain source	Start	End	Description	Accession	InterPro
PANTHER	1	411	-	PTHR24329	-
PANTHER	1	411	-	PTHR24329:SF294	-
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]
Superfamily	6	143	Homeodomain-like	SSF46689	IPR009057 [Display all genes with this domain]
Gene3D	201	284	Homeodomain-like	1.10.10.60	IPR009057 [Display all genes with this domain]
Superfamily	205	283	Homeodomain-like	SSF46689	IPR009057 [Display all genes with this domain]
Prosite_patterns	257	280	Homeobox, conserved site	PS00027	IPR017970 [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]
Gene3D	7	86	Winged helix-turn-helix DNA-binding domain	1.10.10.10	IPR011991 [Display all genes with this domain]
Gene3D	87	150	Winged helix-turn-helix DNA-binding domain	1.10.10.10	IPR011991 [Display all genes with this domain]

Are you surprised that the precise location of the **PAX6** Homeobox domain is not identically predicted by the **SMART** and **Pfam Domain Databases**? If not, why not?

How is that all the predictions, of different domain databases, for a **Paired domain** have the same **Interpro** identifier?

Why does **PRINTS** appear to predict four **Paired_domain**s?

Click on the link to the **SMART** entry for the **Paired domain** (**SM00351**).

Here you will find (quoted from **Interpro**) a **Description** of a **Paired domain**.

Where would you expect a **Paired domain** to occur in a protein?

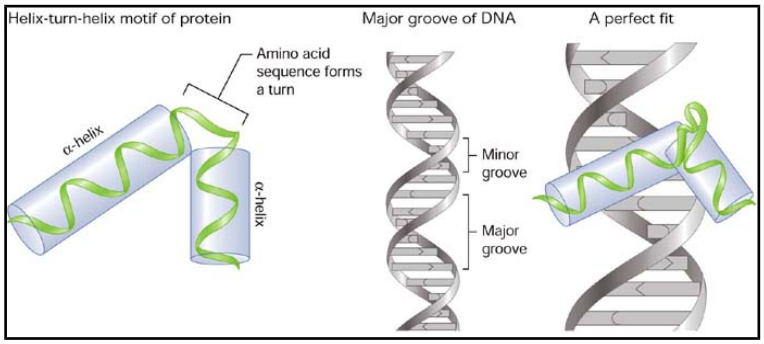
What expectations do you have concerning what typically follows a **Paired domain**?

The paired domain is an approximately 126 amino acid DNA-binding domain, which is found in eukaryotic transcription regulatory proteins involved in embryogenesis. The domain was originally described as the 'paired box' in the Drosophila protein paired (prd) [(PUBMED:2877747), (PUBMED:3123319)]. The paired domain is generally located in the N-terminal part. An octapeptide [(PUBMED:10811620)] and/or a homeodomain can occur C-terminal to the paired domain, as well as a Pro-Ser-Thr-rich C terminus.

Paired domain proteins can function as transcription repressors or activators. The paired domain contains three subdomains, which show functional differences in DNA-binding. 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. 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 [(PUBMED:11103953), (PUBMED:15148315)].

The second paragraph of the **Description** claims, in gross summary:

- 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[5]



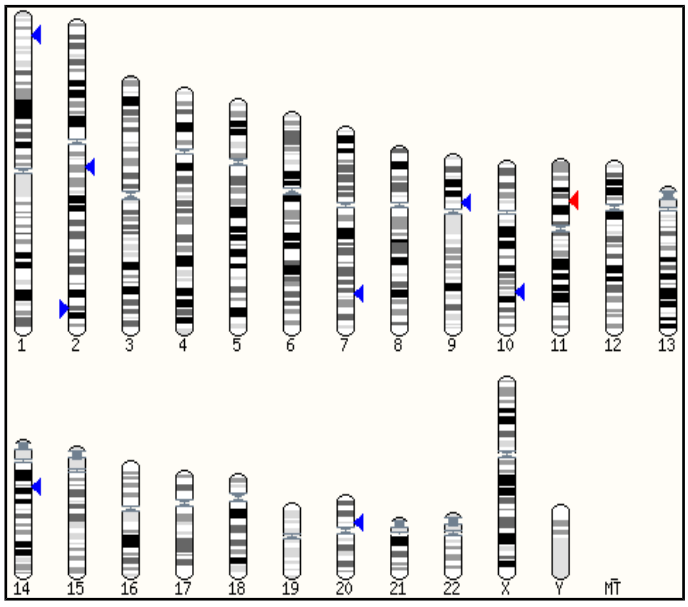
- The first helical triplet is preceded by a β-turn and β-hairpin ("wing") that participate in the DNA binding

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

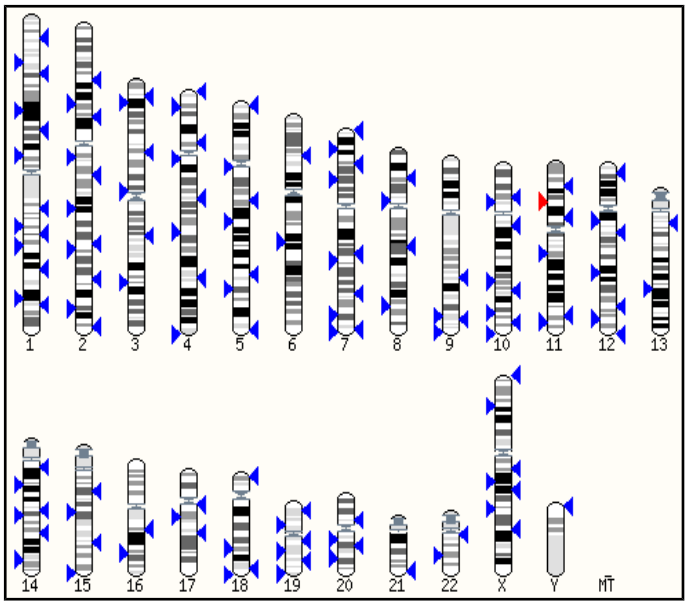
Bear this in mind when looking at the 3D structures a couple of pages on.

Click on Display all genes with this domain for the **Paired domain** 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



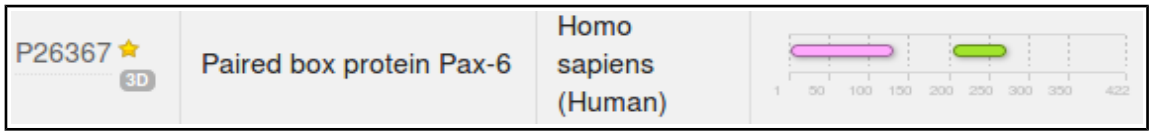
Which domain, **Paired domain** or **Homeobox domain** is more common in humans?

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

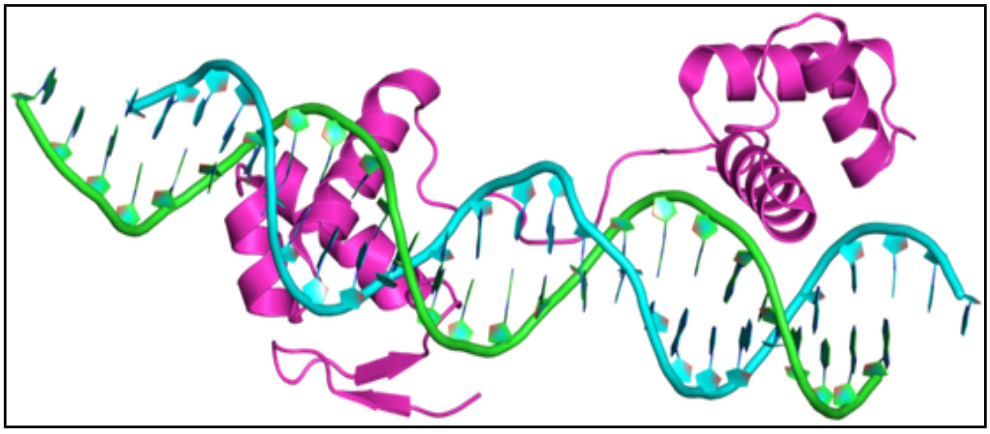
Are all the **PAX** genes on **Chromosome 11**?

5 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 back to the **Domains & features** display. Link to the **InterPro** database entry for **Paired domain**, also know as IPR001523. Here you will find the origins of the **SMART** documentation. 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 the human **PAX6** protein, also known as **P26367**[6].

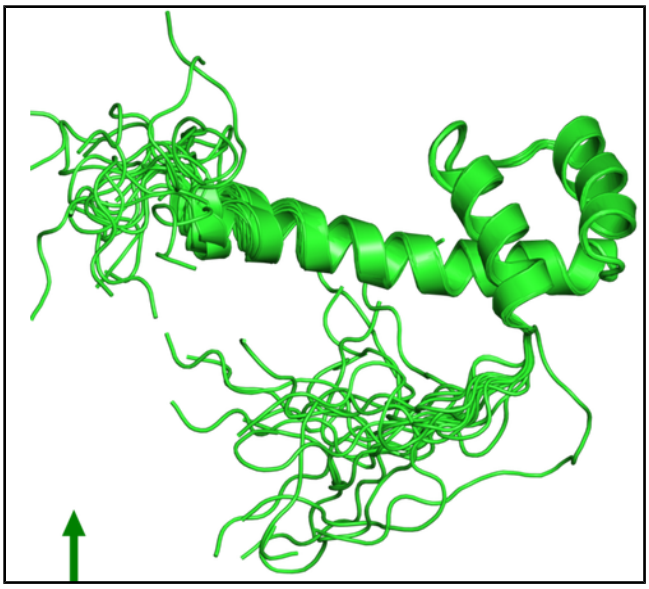


Click on the **Structures** link in the top left hand corner of the page. **InterPro** will offer links to relevant entries in the PDBe, SCOP and CATH[7] databases. Click on the link to the 6pax entry in the **PDBe** database. You will arrive at the entry for **6pax** in **PDBe**, the European version 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, in all its various guises, sufficiently, move back to the **6pax PDBe** entry. From the Quick links on the right of the page, select the 3D Visualisation option.



The **SMART** documentation you read earlier suggested two paired box sub-domains, 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 **SMART** documentation claims the sub-domain 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[8] **Proteins matched**, but you do not need to in order to link to the **Structures**. There are many more structures to choose from this time. I suggest you go for **2cue**. You have to imagine the DNA this time.

It looks rather as if the **Homeobox domain** also includes a helical triplet including a **Helix-Turn-Helix**. You could have confirmed this by reference to the relevant SMART documentation (as you did for the **Paired box** domain). It is the **HTH** that the **Homeobox** uses to bind to DNA.

InterPro did not detect the **Homeobox HTH** as it did the **Paired box HTH**. Have you any thoughts as to why this might be?

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

6 Third from the bottom of the first page, last time I counted.
7 **PDB** is the main database for **3D** protein structures. **SCOP** and **CATH** are also **3D** structure related databases.
8 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.

To end, a gesture towards demonstrating that you could quite easily have computed most of the information you have been accessing, ready packed, from various databases. There are many way this objective could be achieved, I choose to search for the features of the **PAX6** protein.

As has been discovered from several information sources, 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, you will investigate how you might discover 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 in domains that occur in signalling proteins).

So, let us not search individual domain databases. I am sure you could find your own way through using most of the major searches, if you wished. Notes on using the Prosite, **Pfam** and **PRINTS** domain databases appear in the discussion sections of appropriate exercises, but should not take up significant class practical time I feel. Investigating each individually turn does have some merit however. **Prosite** illustrates how widely domain matches can vary in significance, **Pfam** gives and opportunity to superficially discuss **HMM**s and searching **PRINTS** illustrates the small margin between a positive and a negative result.

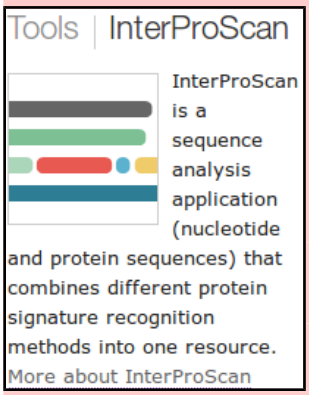
Here, use just **Interpro** to do the whole job. **Interpro** will search for all domains using the appropriate domain databases, thus removing the tedium of interrogating a miscellany of domain searching resources individually.

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 and click the **Submit** button, you would produce exactly the results you saw many pages back, when you were investigating **UniProtKB**[9]. 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 ...

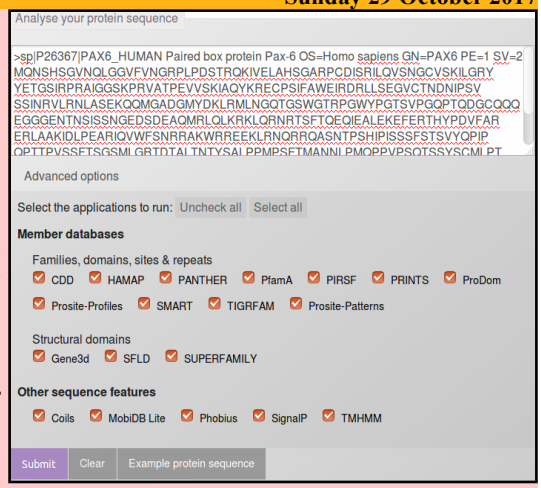


Select the **InterProScan** link. Here you will be offered the opportunity to download the **InterProScan** program.

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

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

9 Not surprising as **UniProtKB** simply links to **Interpro** to show you its graphic.



You will arrive at a page 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**.

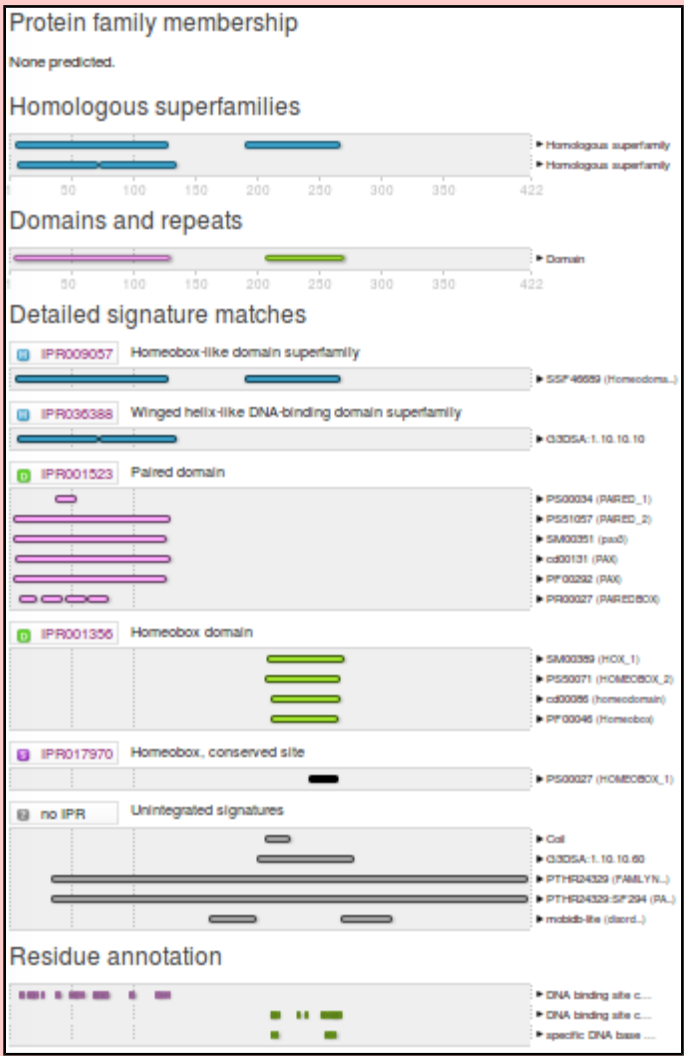
The **Advanced options** only allow you to choose which databases you wish to search and which feature prediction programs you wish to run. The default is to use all the databases and to run all the predictor programs. I struggle to imagine an occasion I would want to save the **EBI** servers a few cycles 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**.
- MobiDB Lite is a method of **Fast and highly specific consensus prediction of intrinsic disorder in proteins**. A new facility for **Interpro**. It uses **MobiDB**, a database of annotations of **intrinsic protein disorder**. **Protein disorder** being a structural features characterising large sets of proteins with prominent members that are **intrinsically disordered proteins**.
- **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 (from the file you made earlier called **pax6_human.fasta**). Accept the "***do everything***" default. 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 earlier when looking around **UniProtKB**.

There is, however, at least one significant 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 **UniProtKB** information, maybe as **Interpro** has only recently included analysis using **Coils? UniProtKB** might catch up next time it is updated.

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 **Homeobox domain-like**. Both of these associations are based on the hit behind the link SSF46689.

SCOP classification

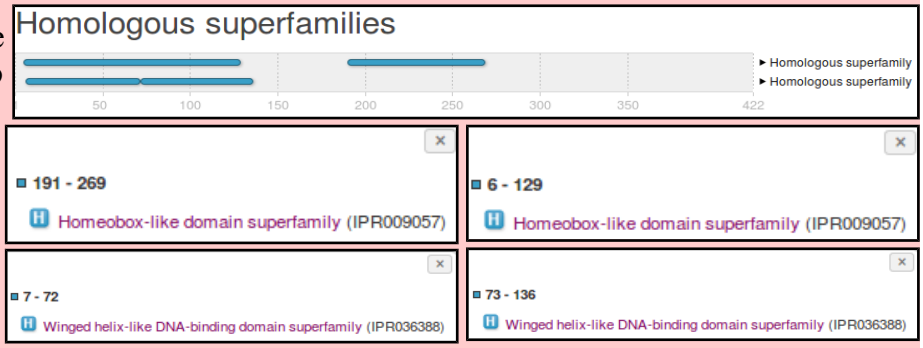
Root:	SCOP hierarchy in SUPERFAMILY [SCOP 0] (11)
Class:	All alpha proteins [SCOP 46456] (284)
Fold:	DNA/RNA-binding 3-helical bundle [SCOP 46688] (14)
Superfamily:	**Homeodomain-like** [SCOP 46689] (19)
Families:	Homeodomain [SCOP 46690] (40)
	Recombinase DNA-binding domain [SCOP 46728] (5)
	Myb/SANT domain [SCOP 46739] (15)
	SLIDE domain [SCOP 100998]
	GARP response regulators [SCOP 81683]
	DNA-binding domain of telomeric protein [SCOP 46745] (2)
	Paired domain [SCOP 46748] (3)

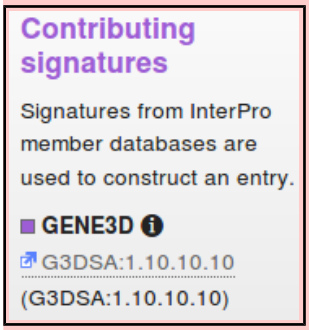
Follow this link and you will see it leads to the Homeodomain-like superfamily of the Superfamily 1.75 HMM library and genome assignments server database that specialises in very general (**SCOP**[10] **superfamily** level) protein classifications. One **Superfamily** entry will typically correspond to a number of more specific **SCOP** classifications. Here you can see that the **Superfamily** domain **Homeodomain-like** includes both the **Homeodomain** & **Paired domain Families**.

Return to your **Interpro** results page. The links beginning "**GD3SA**" point to **Superfamily** domains defined by the CATH Protein Structure Classification database. CATH is similar to to **SCOP** in that it is another Structural classification database. **CATH Superfamilies** are to be found in the Gene3D database[11]. One such link suggests two regions that belong to a **Winged helix-like DNA-binding domain superfamily**. These seem to correspond to the two **Helix Triplets** of the **Paired domain**. Note that the **Helix Triplet** in the **Homeobox domain** is not detected by **Gene3D**? Possibly because of the lack of **Beta Sheet** "**Wings**" in the **Homeobox domain**?

Interpro provides a unified report of all the superfamilies detected either by reference to the **SCOP** or **CATH** databases.

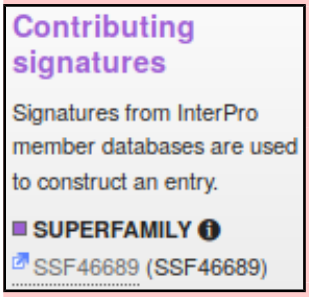
Click on the region bars and you will be offered links to the relevant **Interpro** entries.





Follow one of the links to the **Interpro** family **Winged helix-like DNA-binding domain superfamily** (IPR036388). Note the **Contributing signatures** in the top right hand corner of the page. Here is listed the domain database entries that are used to determine the presence of an **Interpro Winged helix-like DNA-binding domain superfamily**

Essentially, if **GENE3D** finds a match with its **Winged helix-like DNA-binding domain superfamily** (G3DSA:1.10.10.10), then **Interpro** records a match with its **Winged helix-like DNA-binding domain superfamily** (IPR036388).



Move back to your **Interpro** graphic and follow one of the links to the **Interpro** family **Homeodomain-like domain superfamily** (IPR009057). Again, note the **Contributing signatures**.

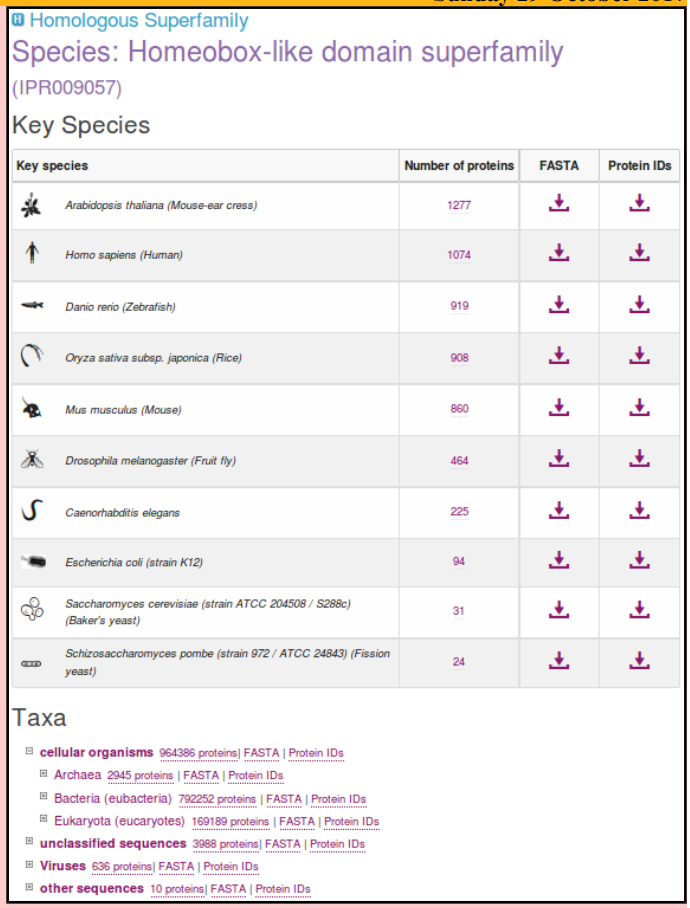
This time it is stated that, if **Superfamily** finds a match with its **Homeodomain-like superfamily** (SSF46689), then **Interpro** records a match with its **Homeodomain-like** domain (IPR009057)[12].

I conclude the **Homologous superfamilies** and **Domains and Repeats** sections of the graphic simply summarise and confirm information from the **Detailed signature matches** section.

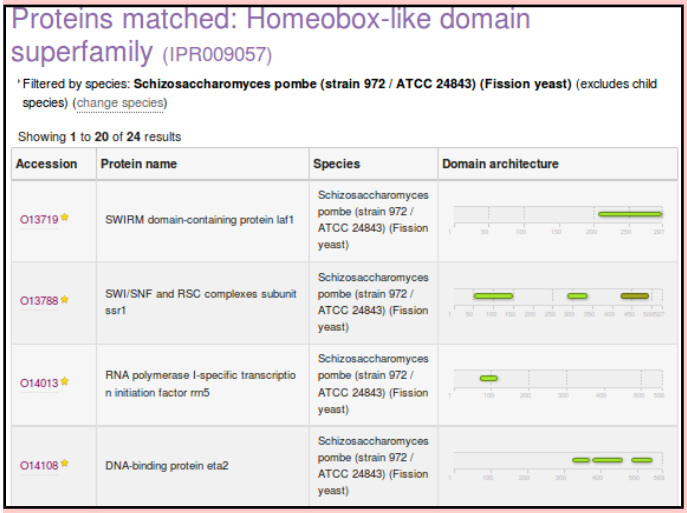
10 **S**tructural **C**lassification **O**f **P**roteins.
11 Broadly, **CATH** is to **Gene3D** as **SCOP** is to **Superfamily**.
12 Until recently, matches with Gene3D entries were also regarded as significant here.

While you have the **Interpro Homeobox-like domain superfamily** in view, it is easy to obtain an impression of how widely spread throughout nature is this domain family. You have already established that there are a fair few in human proteins.

Click on the **Species** button on the left hand side of the page.



As you can see, this is a very popular domain. 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 can take time.



You can make this list enormous by injudicious employment of the expansion buttons (the **Number of protein** links). Why not? It amused me for a few moments anyway.

Finally, return again to you **Interpro** graphic. Notice that the **Paired domain** prediction is supported by matches with **six** different domain databases. Only **four** of these support the **Homeobox domain** prediction. The missing two database matches are with **Prosite patterns** (identifier begins **PS** and typically matches the domain partially where it is best conserved) and with **PRINTS** (identifier begins **PR**).

Why do you suppose there is no match from **PRINTS** or **Prosite** patterns to support the **Homeobox domain** prediction for this protein?

What do you suppose the **Homeobox conserved site** might be?

THE END

DPJ – 2017.10.29

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 more background 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 pertinent 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.

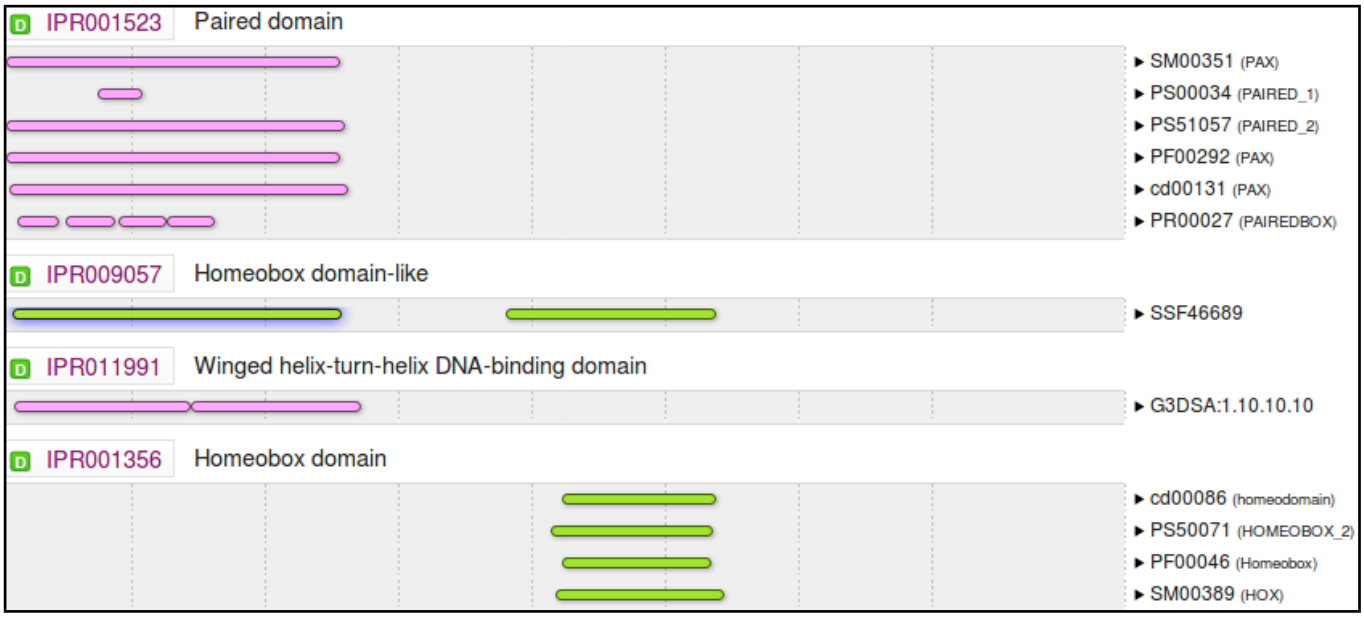
This makes the Model answers section very big. **BUT**, it is not intended for printing or for reading serially, so I submit, being long and wordy does not matter. Feel free to disagree.

From your investigations using **UniProtKB**:

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 (discovered via **UniProtKB** earlier), it is clear that the first two of the helical triplets lie in the **Paired** domain and the third is in the **Homeobox** domain.



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 canonical isoform a. The **NCBI** say they omit the other isoform(s) as they do not aspire to "completeness" but just an indication of structure with the **RefSeq** entries.

The other three features were the coding sequences for three **ELP4** isoforms. Why more than one for this gene then? Possibly because they are "more different" representing interesting variation in gene structure?

```
complement(39424..>39569)
/gene="ELP4"
/gene_synonym="AN; AN2; 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="CCDS: CCDS73271.1 "
/db_xref="GeneID: 26610 "
/db_xref="HGNC: HGNC:1171 "
/db_xref="MIM: 606985 "
/translation="MAAVATCGSVAASTGSAVATASKSNVTSFQRRGPRASVTNDSGP
RLVSIAGTRPSVRNGQLLVSTGLPALDQLLGGGLAVGTVLLIEEDKYNIYSPLLFKYF
LAEGIVNGHTLLVASAKEDPANILQELPAPLLDDKCKKEFDEDVYNHKTPESNIKMKI
AWRYQLLPKMEQIGPVSSSRFGHYYDASKRMPQELIEASNWHGFFLPEKISSTLKVEP
CSLTPGYTKLLQFIQNIIYEEGFDGSNPQKKQRNILRIGIQNLGSPLWGDDICCAENG
GNSHSLTKFLYVLRGLLRTSLSACIITMPTHLIQNKAIIARVTTLSDVVVGLESFIGS
ERETNPLYKDYHGLIHIRQIPRLNNLICDESDVKDLAFKLKRKLFTIEWVQDNYLRQE
RNIYPPGFSYLLKQKDSAWGEGSLQHSTFLMSFLAKATAFASRLVRHSEPLKQNGSGR
IRQAAGPRLWHDGRRQEAPGLLGIPP"
```

```
complement(39438..>39569)
/gene="ELP4"
/gene_synonym="AN; AN2; 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="CCDS: CCDS7875.2 "
/db_xref="GeneID: 26610 "
/db_xref="HGNC: HGNC:1171 "
/db_xref="MIM: 606985 "
/translation="MAAVATCGSVAASTGSAVATASKSNVTSFQRRGPRASVTNDSGP
RLVSIAGTRPSVRNGQLLVSTGLPALDQLLGGGLAVGTVLLIEEDKYNIYSPLLFKYF
LAEGIVNGHTLLVASAKEDPANILQELPAPLLDDKCKKEFDEDVYNHKTPESNIKMKI
AWRYQLLPKMEIGPVSSSRFGHYYDASKRMPQELIEASNWHGFFLPEKISSTLKVEPC
SLTPGYTKLLQFIQNIIYEEGFDGSNPQKKQRNILRIGIQNLGSPLWGDDICCAENGG
NSHSLTKFLYVLRGLLRTSLSACIITMPTHLIQNKAIIARVTTLSDVVVGLESFIGSE
RETNPLYKDYHGLIHIRQIPRLNNLICDESDVKDLAFKLKRKLFTIERLHLPPDLSDT
VSRSSKMDLAESAKRLGPGCGMMAGGKKHLDF"
```

```
complement(39533..>39569)
/gene="ELP4"
/gene_synonym="AN; AN2; 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="CCDS: CCDS73272.1 "
/db_xref="GeneID: 26610 "
/db_xref="HGNC: 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 shorter than any of the three entire isoform sequences offered here.

Note also that this final coding exon of **ELP4** (stretching from **39424/39438/39533** 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).

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**.

```
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 protein Pax-6; 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="CCDS: CCDS31451.1 "
/db_xref="LRG:p1"
/db_xref="GeneID: 5080 "
/db_xref="HGNC: HGNC:8620 "
/db_xref="MIM: 607108 "
/translation="MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISR
ILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEI
RDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGSWGTRP
GWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFT
QEQIEALEKEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRR
QASNTPSHIPISSSFSTSVYQPIPQPTTPVSSFTSGSMLGRTDTALTNTYSALPPMPS
FTMANNLPMQPPVPSQTSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMNSQPMGTSGTT
STGLISPGVSVPVQVPGSEPDMSQYWPRLQ"
```

RefSeqGenes, comprise the entire gene plus **5,000** "extra" base pairs in either direction. The overlap here is entirely within the "extra" base pairs.

```
gene            5001..38170
                /gene="PAX6"
                /gene_synonym="AN; AN2; D11S812E; FVH1; MGDA; WAGR"
                /note="paired box 6"
                /db_xref="GeneID:5080"
                /db_xref="HGNC:HGNC:8620"
                /db_xref="MIM:607108"
```

Careful study of any of the three **Genome Database** displays visited earlier (**Genome Data Viewer**, **Map Viewer**, **Ensembl**) will confirm the relative positions of **PAX6** and **ELP4**. The view offered by Map Viewer is the clearest offered in this document

```
gene            complement(38437..>40170)
                /gene="ELP4"
                /gene_synonym="AN; AN2; 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"
```

The annotation (specifically the **gene_synonym**s) of **ELP4** associate this gene with **PAX6**. However, as the **ELP4** gene annotation to the right attests, only because of its proximity.

General protein information

Preferred Names
elongator complex protein 4

Names
PAX6 neighbor gene protein
elongation protein 4 homolog

Why might you have expected more features than there were?

Summary:

All the evidence has suggested that **PAX6** has at least **2** isoforms. This would lead me to expect at least **2** CDS features here related to **PAX6**?

Full Answer:

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/

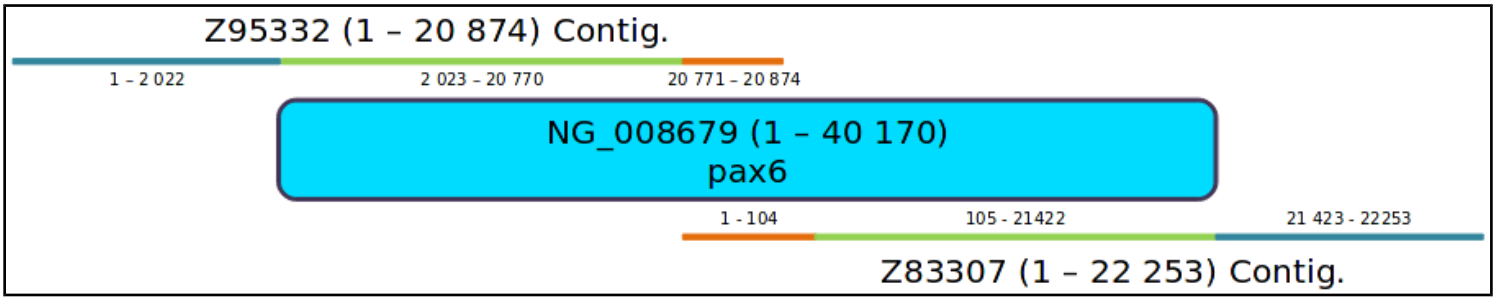
Regarding the **PAX6 RSG** sequence, only difference I see between **NG_008679.1** and the current build of the genome (**GRCh38**) is an extra **'G'** beyond the 3'-UTR of the **PAX6** transcripts (at **NC_000011.10:g.31,819,125**). ... "

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.

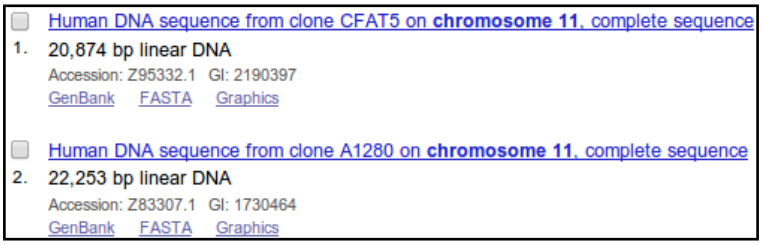


Full Answer:

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.

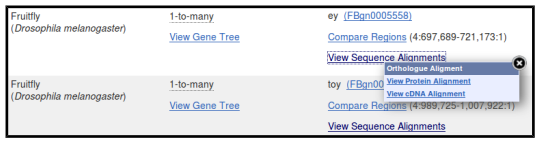
I refer you again to the far more intuitive way of encapsulating the same message graphically, donated by **Cecilia Pinto** that is now the "**Summary** Answer" above. Much better! Thank you once more Cecilia.

From your investigations using **Ensembl**:

Which human **PAX6** isoform has been chosen to search for orthologues?
How do you suppose this choice might have been justified?

The protein used to represent **PAX6** human is consistently **ENSP00000492024**. At least, this was the choice for the alignments I looked at. 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.

Only if you have the time, take a quick look at the two Fruitfly orthologues recorded here. The Fruitfly genes are named **ey** (**ey**e) and **toy** (**t**win **o**f **ey**e). 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. Is it therefore reasonable to conclude that both fly proteins are both closest to the canonical protein sequence of **PAX6** human (**isoform 1**)?



Protein alignment for **ey**

```
ENSP00000492024/1-436  ------------------------------MQ-N-SHSGVNQLGGVFVNGRPLPDSTRQ
FBpp0099810/1-898      GKPSPTMEAVEASTASHPHSTSSYFATTYYHLTDDECHSGVNQLGGVFVGGRPLPDSTRQ
                                                    :  : .************.**********

ENSP00000492024/1-436  KIVELAHSGARPCDISRILQTHADAKVQVLDNQNVSNGCVSKILGRYYETGSIRPRAIGG
FBpp0099810/1-898      KIVELAHSGARPCDISRILQ--------------VSNGCVSKILGRYYETGSIRPRAIGG
                       ********************              ***************************

ENSP00000492024/1-436  SKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASE
FBpp0099810/1-898      SKPRVATAEVVSKISQYKRECPSIFAWEIRDRLLQENVCTNDNIPSVSSINRVLRNLAAQ
                       *******.******:*******************.*.**********************::
```

Protein alignment for **toy**

```
ENSP00000492024/1-436  ----------------------MQN-SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHS
FBpp0088249/1-543      MMLTTEHIMHGHPHSSVGQSTLFGCSTAGHSGINQLGGVYVNGRPLPDSTRQKIVELAHS
                                           .. .***:******:**********************

ENSP00000492024/1-436  GARPCDISRILQTHADAKVQVLDNQNVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATP
FBpp0088249/1-543      GARPCDISRILQ--------------VSNGCVSKILGRYYETGSIKPRAIGGSKPRVATT
                       ************              ******************:**************.

ENSP00000492024/1-436  EVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQMG---
FBpp0088249/1-543      PVVQKIADYKRECPSIFAWEIRDRLLSEQVCNSDNIPSVSSINRVLRNLASQKEQQAQQQ
                        **.***:********************* **..******************:*:* .
```

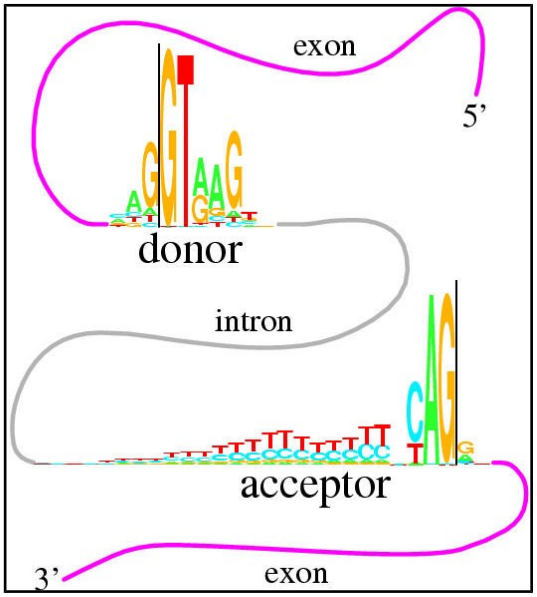
Well, maybe it is not that simple? I would not be surprised If there were isoforms for **ey** and/or **toy** that were roughly equivalent to human **isoform 5a**. The alignment displayed could well reflect the relatively arbitrary choice of **Ensembl** as to which fruitfly isoform it decided to use when the search for matches with **ENSP00000492024** was executed. Maybe, the more "important" canonical isoform was preferred?

As always, it is not the detail that is of real interest here, I merely try to point out that the information you browse I this sort of database often reflects as much informatics convenience as Biological profundity. It is still of vital interest, of course, but should possibly not be over interpreted?

Finally, **Ensembl** does not find all fruitfly homologues of **PAX6**? I suppose **Ensembl** does only claim "**Selected orthologues**"? Even so, **prd** in particular, is a pretty important gene to pass over! The **Flybase** people explained to me why **prd** might have been omitted. I will keep this explanation to myself until I fully understand it.

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?

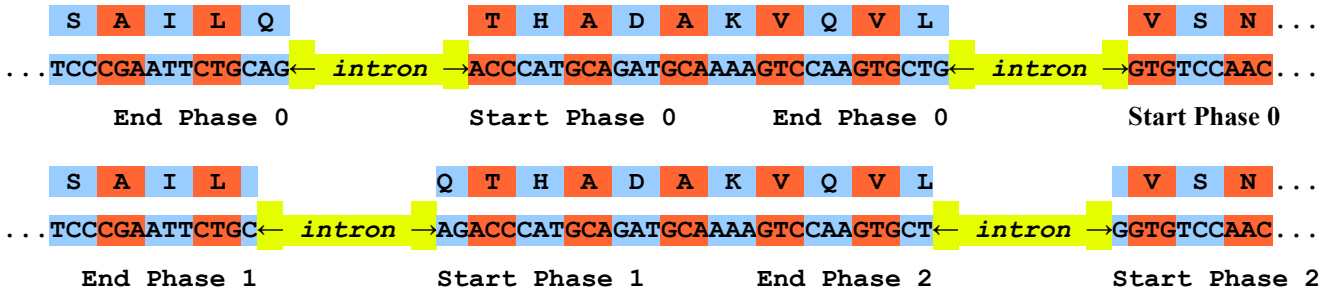
An exon/intron boundary can occur anywhere in a codon. The **Start** and **End Phases** record how an intron has been inserted into a coding region with respect to the coding reading frame.

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.



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

Prints does not find the **Homeobox_domain** at all. If you were to investigate by using the Prints search carefully, you will find it nearly does, but the evidence is not quite strong enough. As has been discussed, none of these systems are perfect. They all occasionally fail. That is why it is always best to use **Interpro** to consult them all and deliver a consensus answer.

Prints appears to find **FOUR Paired_domains**. This is only because of the way **Prints** works. **Prints** finds **FOUR** signatures (or **motifs**) that together indicate **ONE Paired domain**. **Prints** searches for ordered series of **motifs** that together indicate **domains**. Here it reports each of four motifs separately, but it is only claiming one **Paired 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]

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 domain**s in the Human genome. There are many more **Homeobox domain**s.

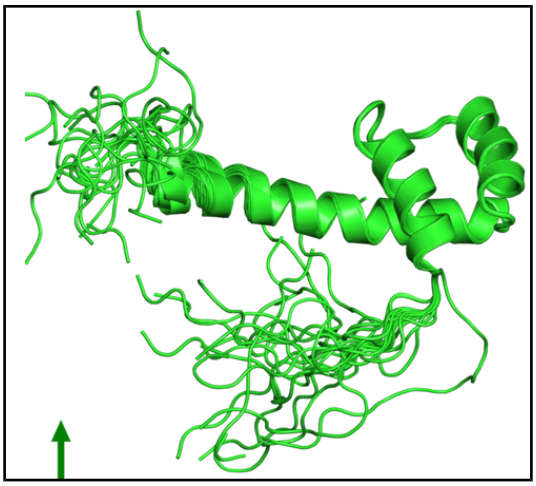
Are all the **PAX** genes on **Chromosome 11**?

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**.

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 **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, 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 programs 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 responsible 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** (a system for secondary structure prediction that you will meet later), possibly using the **COILS** program disguised as **LUPAS**, does 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!

DPJ – 2017.10.29

Discussion Points and Casual Questions arising from the Instructions Text.

Notes:

Work in progress I fear.

The intention is to provide a full consideration of some issues skimmed over in the exercise proper.

If you are attending a "supervised" presentation of the exercise, I would hope to have conducted a live discussion of all these issues to an extent that reflects:

- the depth that seems appropriate
- the time available
- the degree to which the issues seem to match the interests of the class
- how many of you are awake

Here, I hope to write out very full answers were such a response exists. Accordingly, I suggest you will not need to read much of many of these discussions. There will be much detail of interest to rather few of you. Possibly a bit self indulgent, but I wish to make a note of all the background I have discovered while writing these exercises.

In a nutshell, the exercises are trying to make very general points avoiding too much detail. Nevertheless, I record the detail outside the main exercise text, just in case it might be if interest. Some of the answers to the "**Casual Questions**" are exceedingly trivial. Some of the "**Discussion Points**" are exceedingly long and rambling. You have been warned.

Can you now say how many transcripts there are according to the **Genome Data Viewer**?

11, count the transcript prediction lines of blobs and wiggly lines.

How many transcripts are predicted for **PAX6** by **MapViewer**?

11 … he noted unnecessarily!

Discussion of the **Ensembl** transcript colour and numbering schemes.

Introduce Ensembl pipeline

Introduce Vega … for a number of vertebrates

Havana = group feeding Vega for human/Mouse and similar … not all genomes of Vega

GENCODE … amalgamation of Vega and ensembl pipeline … source of Ensembl transcript predictions

Conclusion: gold … agreed between pipeline and Vega

red … either Vega or pipeline, used to be able to tell which by the transcript number (>=200 pipeline, <200 Vega) but now all numbers 200+

blue … non-protein coding

The naming/numbering of transcripts is being improved. Current temporary. Future a method representing prediction quality.

Source … Latest gems from Ben of Ensembl

Strategies employed to minimise the time spent on searches employed to determine gene structures, specifically with respect to their implementation by **Ensembl**.

In particular:

first genscan … find most genes

then CCDS (CDS agreed by pipeline, Vega and NCBI … Human/Mouse specific at present) search on genscan hits only …. reveals coding regions accurately

then mRNA (RefSeq and other high quality data/predictions) only on CCDS hits … reveals UTRs accurately

The significance of the first **14** transcripts associating with entries in the **CCDS** database. Particularly noting that **14** is far closer to **11** and there are just **2** distinct **CCDS** entries referenced here.

2 CCDS entries? This should be logical when you have investigated the number of protein isoforms perhaps?

14 is nearer to 11 … but not the same!! See next discussion for partial elucidation

Why it is reasonable to not regard a match of a **RefSeq** mRNA with the **Genome** as, by itself, sufficient evidence to uniquely predict a transcript.

If RefSeq mRNAs really were the result of sequencing individual mRNA carefully, it might be reasonable to regard a hit with the genome as conclusive. However, they are not. They are assemblies of single pass, poor quality, cDNA sequences. Sequencing masses of these was very popular early this century.

Ensembl regards these sequences as good evidence …. but not conclusive by themselves.

NCBI appears to rely more on the reliability of RefSeqs mRNA sequences

How reliable would you judge these predictions to be?

Needs thought and investigation here but … main message is that there is huge variance in quality between these predictions! Far from binary announcement of existence or otherwise.

Specifically, there is only **APPRIS** support where there is **CCDS** matching. This makes sense as a **CCDS** hit implies a relationship to a confident protein isoform that is very likely to have **orthologues**. This will make more sense when we have considered how many **PAX6** isoforms there might be.

Sequence formats, specifically **FASTA** format.

Indeed, sequence formats will be discussed, but a little further down. Until then, try to contain thy urgent thirst for elucidation.

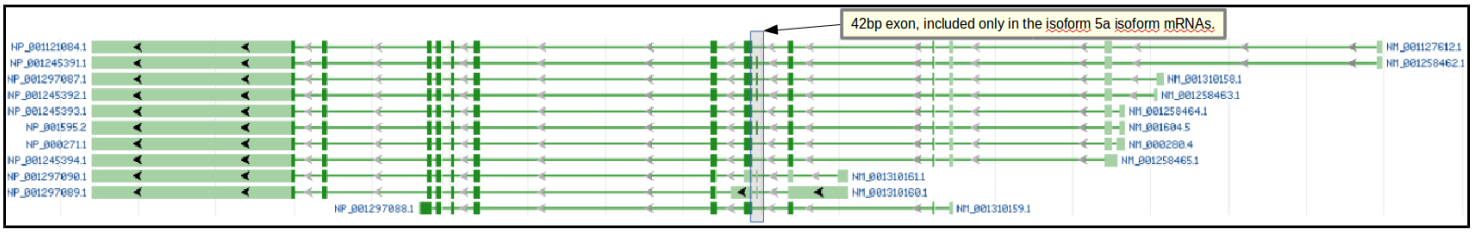
Discussion of the isoform alignments.

Not much to say? … see the inserted 14 amino acids in the middle of the PAX domain?

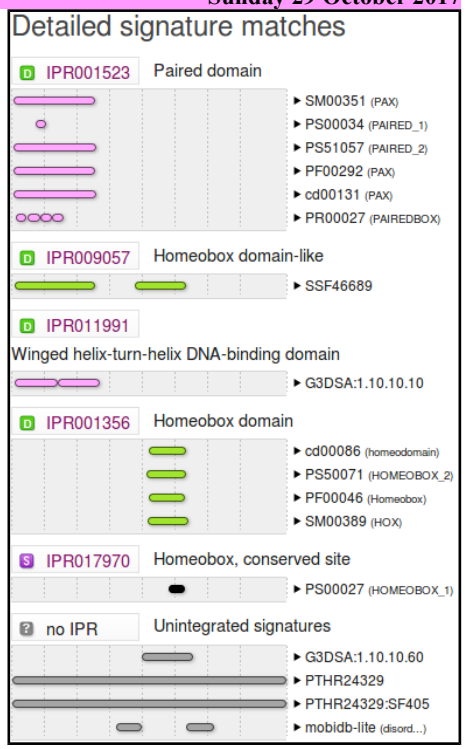
Refer to silly domain / DNA Binding confusion, although I think I do that elsewhere.

Can you see the evidence for this assertion in the regional genomic maps of a few pages back?

Yep … need a picture here … has to be an American one as **Ensembl** pictures are too crushed. I choose the **Genome Data Viewer version**.



Are the **Interpro** results broadly as you might expect?



Yep … two domains, **homeo** & **PAX**, as suggested by **NCBI**. Here more domain/motif databases are quoted, but the conclusion is the same both sides of the Atlantic.

Note the inconsistent naming of the domains! Is this really necessary one muses? Life is muddled enough already surely. How long does it take to choose a single name?

This graphic will be considered in more detail later when we look ay **Interpro** more closely.

Sequence formats.

2 varieties required. For analysis (FASTA) or for storage in a database with annotation (GenBank, EMBL).

FASTA for all the sequences saved so far, minimal annotation, just enough for identification. The sequence is the issue.

```
>NAME Description
Sequence … … …
>NAME Description
Sequence … … …
```

Genbank or **EMBL** (why two!!?) where the annotation is the primary focus (although a bit silly without the sequence!). Formats for the databases. Pity there is two, but to expect too much sanity between **EMBL** and **NCBI** is clearly asking too much. Here we look at **Genbank**, later we will see **EMBL**. I will not elaborate, both have online manuals (GenBank, EMBL). The basics are intuitive (I hope).

Some reference to the times of many many formats here???

Can you see the official gene name **PAX6**, mentioned in this entry?

No … **PAX6** occurs several times in the page (try searching with **Ctrl F**) but only in the page annotation, not in the databases entry!

If you searched **GenBank** (or **EMBL** come to that) for this sequence using the most obvious search **Keyword**, that is **PAX6**,

Do you think you would find this **PAX6** mRNA?

Absolutely not!!

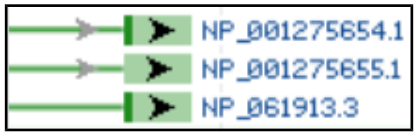
A superficial mention of the **Gene Ontology Project**.

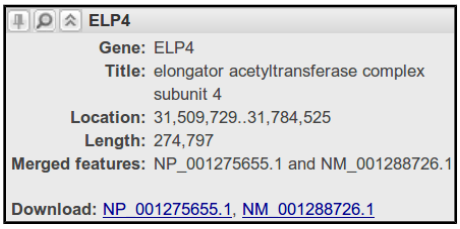
Very superficially … maybe I paste in some of the stuff I deleted here? Or maybe it still lurks somewhere else?

Can you find the additional genes **PAX6-AS1** and **ELP4** in the genome displays you have looked at so far?

ELP4 is clearly visible in the **Map Viewer** version of this region of the genome.

The **Genome Data Viewer** picture prefers to show the location of the **RefSeq** mRNAs that support the existence of **ELP4**.

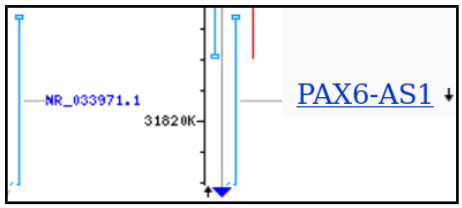




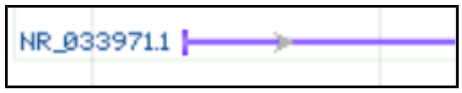
Hover of any of these and the link to **ELP4** is revealed.

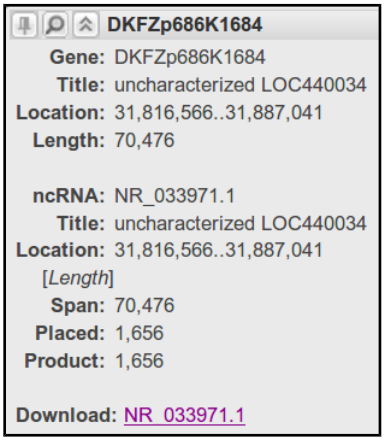
The **Ensembl** display you viewed earlier clearly includes all its predictions of transcripts for **ELP4**.

PAX6-AS1 is clearly visible in the **Map Viewer** version of this region of the genome.

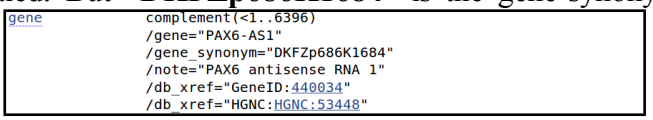


The **Genome Data Viewer** picture shows the location of the **RefSeq** RNA that support the existence of **PAX6-AS1**.





Hover over the RNA reference. An association with a gene called "**DKFZp686K1684**" is revealed. But "**DKFZp686K1684**" is the gene-synonym of **PAX6-AS1**. So the gene is discovered, if indirectly.



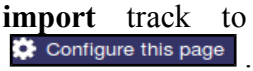
This **gene synonym** implies that this gene was originally identified by the **German Cancer Research Centre (DKFZ)**.

No mention of **PAX6-AS1** Though? Unless you **Download NR_033971.1** and look at the **FASTA** description line which reads:

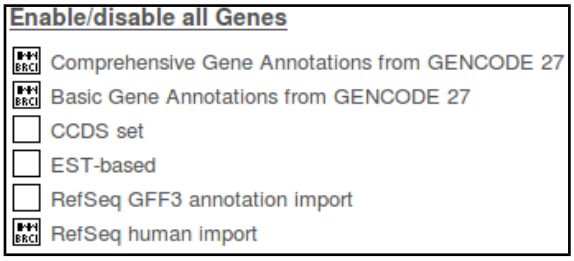
```
>gi|300068930|ref|NR_033971.1|:1-1656 Homo sapiens PAX6 antisense RNA 1 (PAX6-AS1), long non-coding RNA
```

Declaring that this **RefSeq RNA** is a feature of a non-coding gene called **PAX6-AS1**. The only reason for naming it such being that it slightly overlaps the **PAX6** gene on its antisense strand.

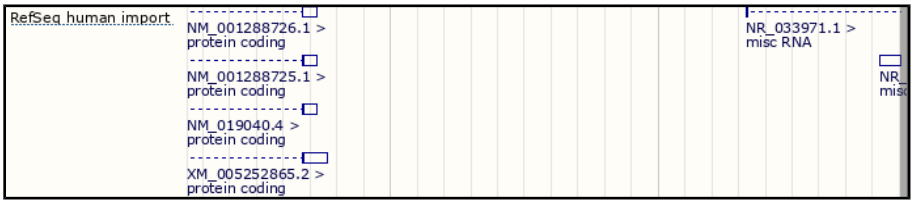
PAX6-AS1 is also represented in the **Ensembl** view of the **PAX6** region. However, it is not so easy to find. Certainly there is no obvious evidence is the view as you examined it previously.

To find **PAX6-AS1** (even disguised as **DKFZp686K1684**), should you really want to, try the following. First add the **RefSeq human import** track to your display. To achieve this, elect to 

. In the **Genes** subsection of the **Genes and transcripts** section, turn on **RefSeq human import**, choosing **Expanded with labels**.

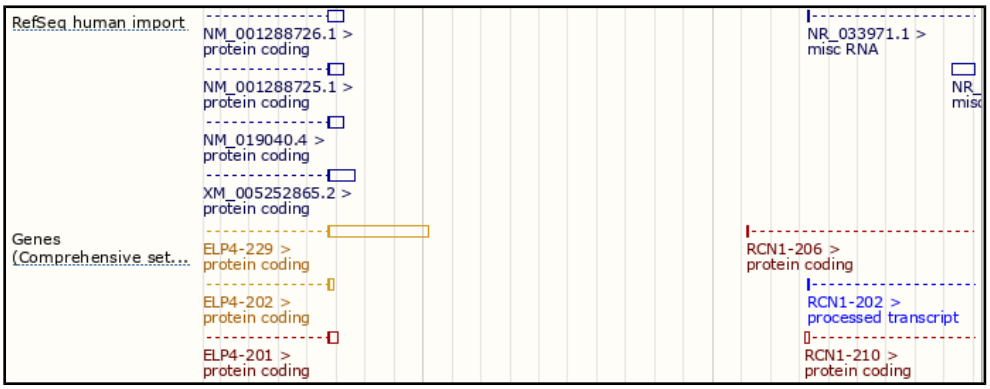


Finally click on the  in the top right hand corner to **Save and close** your selections.



Essentially, you have asked for some **RefSeq** based predictions from the **NCBI** to be added to the display. Amongst these (top right) is the misc RNA prediction based on the match between the **RefSeq** sequence **NR_033971.1**

Job done? Well … yes I think, but having travelled so far! Let us proceed to the tortuous end.

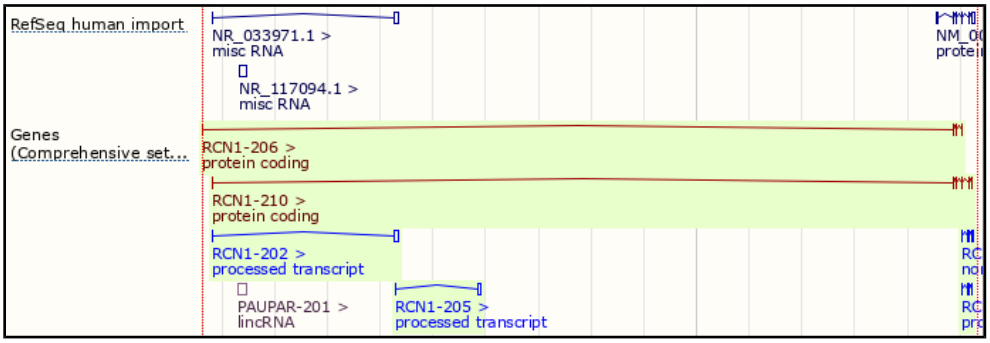


Notice the suspicious matching of some of the **RCN1** gene transcripts and the new **misc RNA** from the **NCBI** (I removed the **Genes (Basic set from GENCODE 27)** track from the display to enhance clarity).

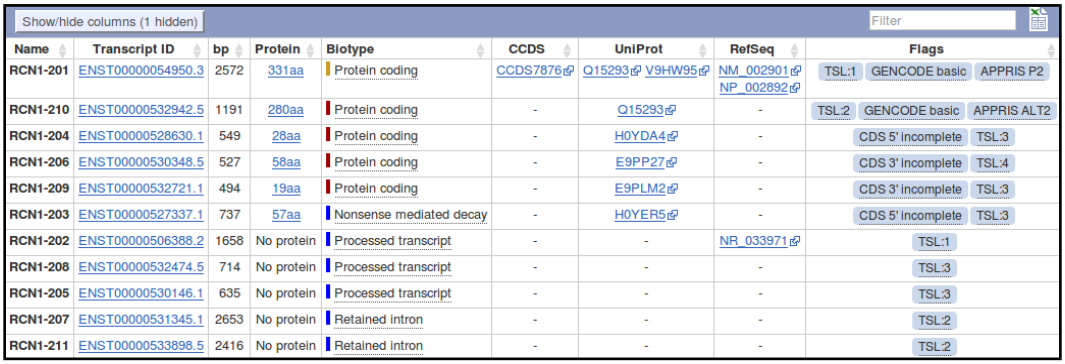
For a closer look, adjust the Region in view to expose the whole of that covered by **NR_033971.1**

To do this 

and then blunder about with the **zooming** and **sliding** and **dragging** tools until you get a view something like mine ... ***OR***, you could use the trick it has just taken me an hour to discover, of selection the area desired by entering the **RCN1** gene name in the appropriate place!!!



Now I am even more convinced by the similarity of structure of the **processed transcript RCN1-202** and the **misc RNA NR_033971.1**



Show/hide columns (1 hidden) Filter

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
RCN1-201	ENST00000054950.3	2572	331aa	Protein coding	CCDS7876	Q15293 V9HW95	NM_002901 NP_002892	TSL:1 GENCODE basic APPRIS P2
RCN1-210	ENST00000532942.5	1191	280aa	Protein coding	-	Q15293	-	TSL:2 GENCODE basic APPRIS ALT2
RCN1-204	ENST00000528630.1	549	28aa	Protein coding	-	H0YDA4	-	CDS 5' incomplete TSL:3
RCN1-206	ENST00000530348.5	527	58aa	Protein coding	-	E9PP27	-	CDS 3' incomplete TSL:4
RCN1-209	ENST00000532721.1	494	19aa	Protein coding	-	E9PLM2	-	CDS 3' incomplete TSL:3
RCN1-203	ENST00000527337.1	737	57aa	Nonsense mediated decay	-	H0YER5	-	CDS 5' incomplete TSL:3
RCN1-202	ENST00000506388.2	1658	No protein	Processed transcript	-	-	NR_033971	TSL:1
RCN1-208	ENST00000532474.5	714	No protein	Processed transcript	-	-	-	TSL:3
RCN1-205	ENST00000530146.1	635	No protein	Processed transcript	-	-	-	TSL:3
RCN1-207	ENST00000531345.1	2653	No protein	Retained intron	-	-	-	TSL:2
RCN1-211	ENST00000533898.5	2416	No protein	Retained intron	-	-	-	TSL:2

Time to look at the transcript table for **RCN1** for the detail.

Go to the **RCN1 Ensembl** gene page using the main search option at the top of your current page. Make sure the transcript table is in view and take a look at the entry for the **processed transcript RCN1-202**.

By the Lord Harry! **processed transcript RCN1-202** is based exclusively on the evidence of the match between **NR_033971.1** and the genome!

I conclude that what the **NCBI** predict as the **non-coding gene PAX6-AS1** with a single transcript based upon the **RefSeq** RNA **NR_033971.1**, **Ensembl** predicts as a non-coding transcript of the protein coding gene **RCN1**.

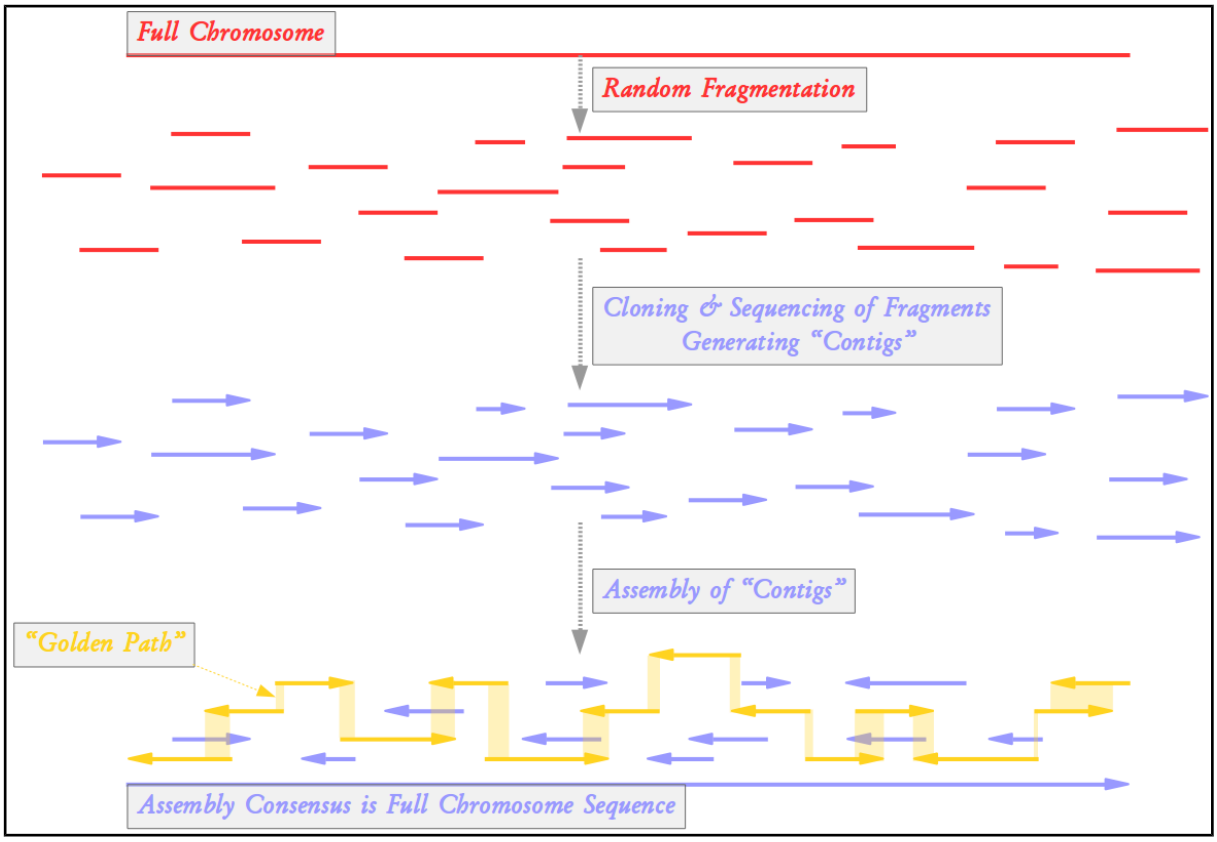
All this nonsense achieves rather little, in the context of the exercise, I suppose. I certainly do not want to suggest you follow your way through the pain I have just endured. However, I hope this little diversion into pedantry does illustrate how using multiple sources of imperfect information can be less than straight forward. To obtain a complete picture often requires lot of effort and patience. Never mind, it will ever get better … possibly.

The role of "**contigs**" in the human genome project.

The objective here is to establish some understanding of what these two sequences that you have found are. To do this it is necessary to understand how the Human Genome was determined using the sequencing technologies available at the turn of the century.

Broadly, the **Human Genome** was considered to big to sequence in one step. Each **Chromosome** was therefore processed separately.

However, even the smallest **Human Chromosome** was too large to be efficiently sequenced as a single entity. Accordingly, **Chromosomes** were fragmented randomly into manageable sections (**20-40Mb** at the start of the project, up to **150Mb** by the end). Each fragment was cloned and sequenced separately. The sequences determined for the chromosome fragments are, in this context, referred to as **Contigs**. The **Contigs**, once reassembled, determined the sequence of each entire **Chromosome**. Time for President Clinton to, somewhat optimistically, announce the task completed.

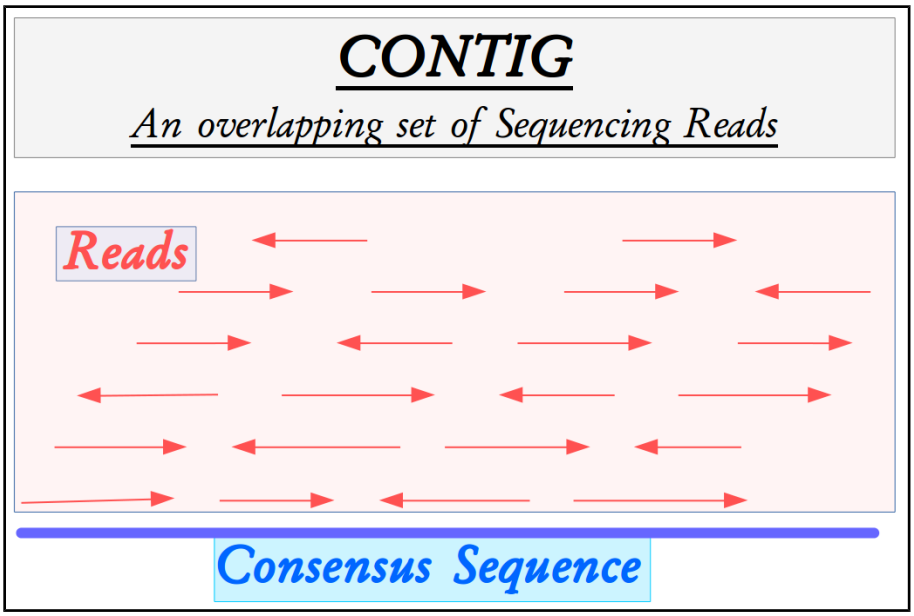


All the individual **Contig** sequences are retained in specialist databases. A minimal selection of the **Contigs** are stored in more general databases such as those you are searching in this exercise. The selected **Contigs** form a "**Golden Path**" through the assembly of all **Contigs**. The "**Golden Path**" is such that the entire **Chromosome** is represented using the smallest set of **contigs** practical.

Clearly, just the **contigs** of the "**Golden Path**" would be insufficient to reliable determine the **Full Chromosome Sequence**. "**Golden Path**" elements might overlap only by a few tens of base pairs. Such an overlap would not be credible except for the knowledge that it is supported by many other **contigs** stored elsewhere.

So, you are looking at the two "**Golden Path**" **contigs** whose overlap fully encompasses the entire **PAX6** gene. Your next task is to use blast to compute the overlap between the two **contigs**.

To conclude, a final note on the term "**Contig**".



Contig (short for **Contig**uous) was a term introduced by Rodger Staden to mean an overlapping set of sequencing reads.

Once assembled, any overlapping set of sequencing reads (**Contig**) will acquire a **Consensus Sequence** that is its single best representation.

The ultimate objective of any sequencing project is to create a single **Contig** that represents the entire target region. The **Consensus Sequence** of this final **Contig** will be "**The Answer**".

Inevitably, due to incomplete data and/or insufficiently clever software, the initial assemblies generated many partial region **Contigs**. Sequencing and assembling must continue until a whole region **Contig**, of acceptable quality, is computed.

For reasons of convenience, the term **Contig** has come to mean the **Consensus sequence** associated with a **Contig**uous set of sequencing reads. This is the meaning I have used in the preceding discussion.

How many **PAX** protein paralogues are there for human? Suggest a prettier naming scheme than **PAX1**, **PAX2**, **...**
Some paralogues seem to have two regions of high similarity (e.g. **PAX4** or **PAX2**), others only one (e.g. **PAX1**)? Can you explain?

Clearly, there are **9 PAX paralogues** for Human (according to **Ensembl**, and all the other sources I have come across). They are **PAX1 PAX2 PAX3 PAX4 PAX5 PAX6 PAX7 PAX8** and, last but by no means least, **PAX9**.

The obvious way to decide which regions of the aligned proteins have been best conserved is to examine the alignments. Rather than plough through all **8** separate pairwise paralogue alignments on offer here, why not gather all the sequences together and construct a single **multiply alignment** (we will consider the issues of **M**ultiple **S**equence **A**lignment, **MSA**, in a separate exercise, later)? To save time, I will do this for you. I will record most of what I did, but suggest you just look at the results, unless you have plenty of time to spare, of course).

Species	Gene ID	Peptide ID	Peptide length	% identity (Protein)	% coverage	Genomic location
Human (*Homo sapiens*)	ENSG00000007372	ENSP00000492024	436 aa	33 %	65 %	11:31784779-31818062
Human (*Homo sapiens*)	ENSG00000075891	ENSP00000359319	396 aa	36 %	72 %	10:100735603-100829941

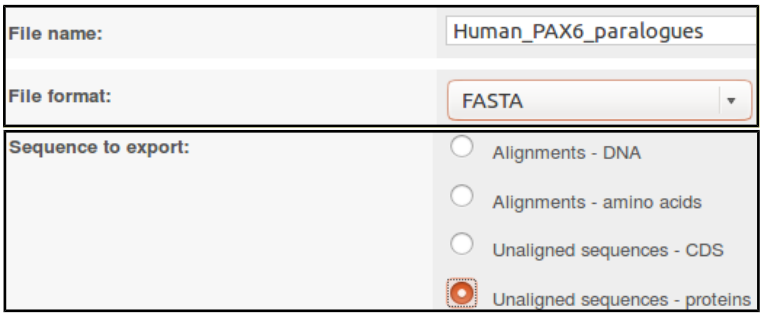
```
CLUSTAL W (1.81) multiple sequence alignment


ENSP00000492024/1-436  ------------MQN-SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ
ENSP00000359319/1-396  MDMHCKADPFSAMH-PGHGGVNQLGGVFVNGRPLPDVVRQRIVELAHQGVRPCDISRQLR
                             *:  .*.****************** .**:******.*.******* *:


ENSP00000492024/1-436  THADAKVQVLDNQNVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRE
ENSP00000359319/1-396  --------------VSHGCVSKILGRYYETGSIKPGVIGGSKPKVATPKVVDKIAEYKRQ
                               **:****************:* .******:****:**.***:***:
```

First note that the protein **ENSP00000492024**, used to find the **orthologues** to the **PAX6** protein, was also used to find the **paralogues**. You could prove this to yourself by looking at a few of the pairwise **paralogue protein alignments**.

Starting from the list of **8 paralogues**, click on 

. Choose **FASTA** for your format.

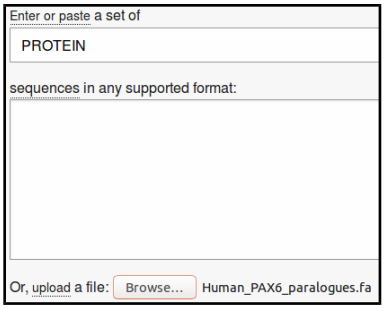
You want **Unaligned sequences – proteins**. Select accordingly.



Click the  to start the download. Your sequences will arrive … somewhere … in a file called:

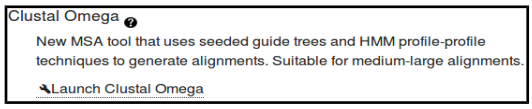
`Human_PAX6_paralogues.fa`

If you entertain the slightest doubts as to what might be in the file, take a look! Note that there are **9** protein sequences, **PAX6** (**ENSP00000492024**) plus its **8 paralogues**.



Now, go to the **MSA** services of the **EBI** to crudely align the **9 Human PAX paralogues** (no need for great accuracy at this point).

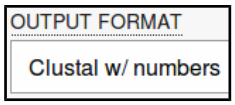
Select an **MSA** program. Any of those on offer will do. I chose **Clustal Omega**.



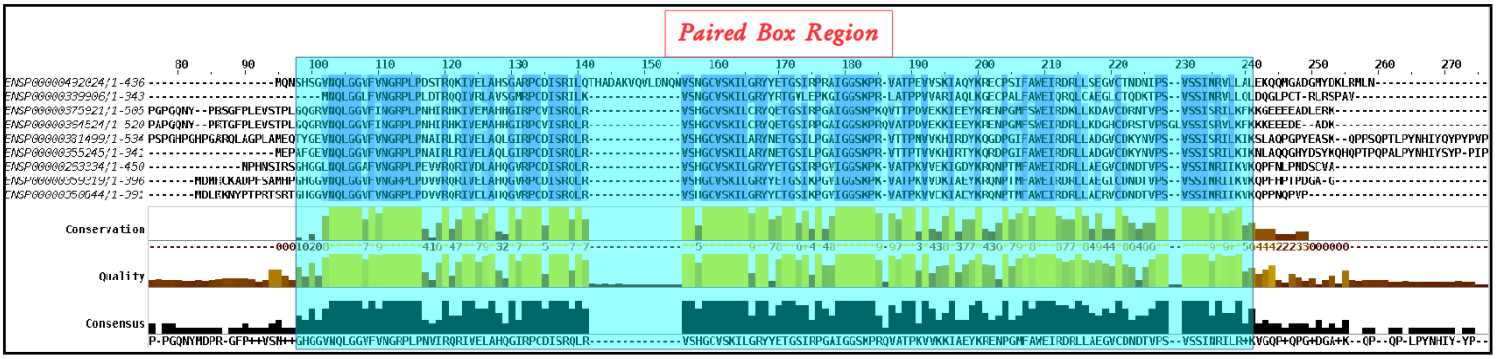
Choose to enter the sequences to be aligned from:

`Human_PAX6_paralogues.fa`

Select an **OUTPUT FORMAT**, I suggest the clearest is **Clustal w/ numbers**. Click the **Submit** button.

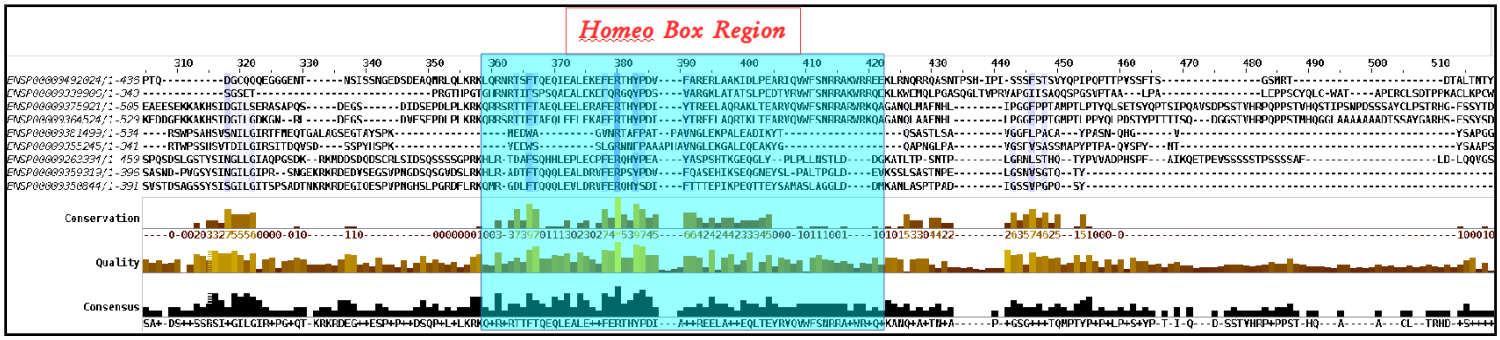


The results I show you were computed by **Clustal Omega**, but I put the **Clustal w/o numbers** output through a program called **Jalview** (which you will meet later) to make them prettier.



All the aligned proteins are **Paired Box** proteins. By definition, they must all include a **Paired Box Domain**. It should not therefore be surprising that the region of this multiple alignment coincident with the **Paired Box domain** of the **PAX6** protein (the top one) is very highly conserved between all the aligned proteins.

Note that only the **PAX6** protein is represented by the **isoform 5a**, all the others are canonical **isoform 1** proteins. I am sure that does not mean that only **PAX6** has an **isoform 5a**. I suspect it is simply that the longer protein is best for searching databases that will present only the canonical shorter isoform for matching.



There is something odd around the region of the **PAX6 HomeoBox**. There is high conservation between some of the **Paired Box proteins** (the top **4** maybe) but not all of them (specifically. the bottom **5**).

Well, these are ***Paired Box proteins***. They are all obliged to have a **Paired Box Domain**, however, nowhere in the rule book does it insist they also have a ***Homeo Box Domain***! It would appear, some do and some do not. Which is fine This observation will be confirmed by some of the documentation you will read soon and also during the exercise in which we investigate features of **blast**.

And finally, … a prettier naming scheme than **PAX1**, **PAX2**, … ? Well it surely must be an easy task to surpass such a low standard of imagination! The whimsy of my exceedingly good self, for example?

I would name the first two after my Mother, so we have **Edith** and **Lillian**. My father would demand the next three to be **Percival**, **Francis** and **Herbert**. Then there is sweet **Keri** from the Valleys of South Wales who stole my weary heart in 1966 … and has yet to return it. Plus, of course, dear **Gwenllian**, from a similar era, whose spirit was as full of poetry as her name. To conclude must, of course, be **Muddy** (as in Waters) and **Lightening** (as in Hopkins)!

So, my nomenclature suggestions for the happy band of **9 Human PAX paralogues**, in ascending numerical order -

Edith Lillian Percival Francis Herbert Keri Gwenllian Muddy and **Lightening –** MUCH better!!

Are you surprised that the precise location of the **PAX6** Homeobox domain is not identically predicted by the **SMART** and **Pfam Domain Databases**? If not, why not?

Both **Smart** and **Pfam** predict the locations of protein domains. They both use similar, but not identical, methods. In this case, both predict a **Homeobox domain** where it is very likely that there is a **Homeobox domain**. This is surely very good news. Should we really expect the predicted locations to be identical? These are just predictions after all and it is questionable whether domains really have precise amino acid specific locations. It is doubtful that all human experts would agree on the most probable exact domain location. Why would we expect computer programs to do better?

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]

How is that all the predictions, of different domain databases, for a **Paired domain** have the same **Interpro** identifier?

Interpro does not have its own domain models. It defines domains by the predictions of other domain databases including **Prosite_profiles**, **Smart** and **Pfam**. So if, as here, a **Homeobox domain** is detected by **Prosite_profiles** (**PS50071**), **Smart** (**SM00389**) and **Pfam** (**PF0046**), there exists **3** pieces of evidence to encourage **Interpro** to declare it to believes there to be a **Homeobox domain** (**IPR001356**).

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]

Any one of the **Prosite_profiles**, **Smart** or **Pfam** hits would have been sufficient for **Interpro** to assign membership of this domain to its **Homeobox** classification **IPR001356**.

Where would you expect a **Paired domain** to occur in a protein?
What expectations do you have concerning what typically follows a **Paired domain**?

The **Paired domain** is here said to be found "*generally in the N-terminal part*" of the protein. That is certainly true of all the examples we have met so far.

> The paired domain is an approximately 126 amino acid DNA-binding domain, which is found in eukaryotic transcription regulatory proteins involved in embryogenesis. The domain was originally described as the 'paired box' in the Drosophila protein paired (prd) [(PUBMED:2877747), (PUBMED:3123319)]. The paired domain is generally located in the N-terminal part. An octapeptide [(PUBMED:10811620)] and/or a homeodomain can occur C-terminal to the paired domain, as well as a Pro-Ser-Thr-rich C terminus.
>
> Paired domain proteins can function as transcription repressors or activators. The paired domain contains three subdomains, which show functional differences in DNA-binding. 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. 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 [(PUBMED:11103953), (PUBMED:15148315)].

The claim here that "*An octapeptide and/or a homeodomain can occur C-terminal to the paired domain, as well as a Pro-Ser-Thr-rich C terminus*" confirms what was seen from the **Human PAX paralogue alignments** Previously. That is, sometimes there is a **homeodomain C-terminal** to the **Paired domain**, but not always.

From the **UniProtKB** documentation, you saw that Human **PAX6** at least has "*a Pro-Ser-Thr-rich C terminus*"

Feature key	Position(s)	Description	Actions	Graphical view	Length
Compositional bias	131 – 209	Gln/Gly-rich	Add BLAST		79
Compositional bias	279 – 422	Pro/Ser/Thr-rich	Add BLAST		144

Note the mention of the important **prd** Drosophila gene here, overlooked in the **Ensembl** presentation of orthologues to Human **PAX6**?

InterPro did not detect the **Homeobox HTH** as it did the **Paired box HTH**. Have you any thoughts as to why this might be?

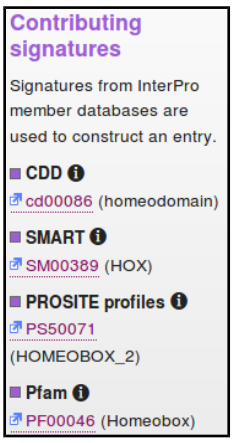
The documentation from **SMART**, which really originates from **Interpro**, clearly claims the presence of an **HTH** as the DNA binding element. However, **Interpro** does not predict the presence of an **HTH**, as it did for the **Paired Box**?

> The homeobox domain or homeodomain was first identified in a number of drosophila homeotic and segmentation proteins, but is now known to be well-conserved in many other animals, including vertebrates [(PUBMED:2568852), (PUBMED:1357790)]. Hox genes encode homeodomain-containing transcriptional regulators that operate differential genetic programs along the anterior-posterior axis of animal bodies [(PUBMED:12445403)]. The domain binds DNA through a helix-turn-helix (HTH) structure. The HTH motif is characterised by two alpha-helices, which make intimate contacts with the DNA and are joined by a short turn. The second helix binds to DNA via a number of hydrogen bonds and hydrophobic interactions, which occur between specific side chains and the exposed bases and thymine methyl groups within the major groove of the DNA. The first helix helps to stabilise the structure.
>
> The motif is very similar in sequence and structure in a wide range of DNA-binding proteins (e.g., cro and repressor proteins, homeotic proteins, etc.). One of the principal differences between HTH motifs in these different proteins arises from the stereo-chemical requirement for glycine in the turn which is needed to avoid steric interference of the beta-carbon with the main chain: for cro and repressor proteins the glycine appears to be mandatory, while for many of the homeotic and other DNA-binding proteins the requirement is relaxed.

I cannot be certain why, however, **HTH**s are difficult to detect just with computer programs. I used to include an exercise that tried for this protein. It proved impossible to obtain a complete picture. One of the reasons being that there are a number of different types of **HTH**. Any given program will typically only search effectively for one type.

Not a very satisfactory answer!

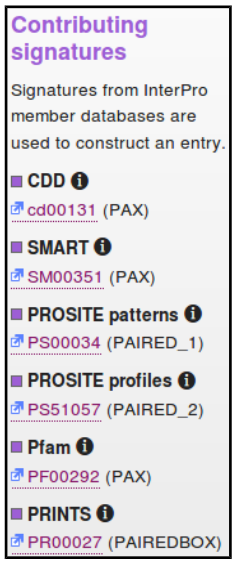
Why do you suppose there is no match from **PRINTS** or **Prosite** patterns to support the **Homeobox domain** prediction for this protein?
What do you suppose the **Homeobox conserved site** might be?



Well, the short and rather boring answer to the first part of this question is that **Interpro** did not interrogate **PRINTS** or **Prosite patterns** when it considered the existence of a **Homeobox domain** in this protein!

To prove this you only need to follow the links to the relevant **Interpro** entries and look at the **Contributing signatures**.

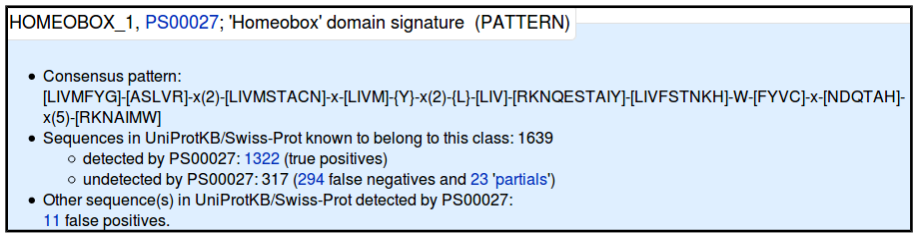
Both **PRINTS** and **Prosite patterns** are used to determine the presence of a **Paired domain**. Neither is used to detect a **Homeobox domain**.



More interestingly, in the case of **Human PAX6** at least, it would not have made any difference had **PRINTS** and/or **Prosite patterns** been considered for the **Homeobox domain** prediction.

Interpro did actually register a match between the **PAX6 human protein** and the relevant **Prosite pattern**. However, **Interpro** judged this match as too weak (i.e. the probability of a false positive is too high) to be regarded as viable evidence for predicting a **Homeobox domain**. **Interpro** records the match as a conserved site, as you can see from your **Interpro** graphic.

Were you to look at the relevant **Prosite** entry (the illustration is a link), you would see that the **Prosite** pattern is quite long, but rather non-specific (the pattern syntax will be fully explained somewhere else). It misses **317** of the **1,639 Homeobox** domains in **SwissProt**! And incorrectly claims a **Homeobox** where no **Homeobox** exists on **11** occasions (according to **SwissProt**, which is assumed immaculate in this context). I think **Interpro** is correct to take a hit with this pattern rather lightly.



PRINTS has a domain model for **HOMEOBOX**, however, it does not match the **PAX6 Human Homeobox domain** sufficiently well to register as a hit! Nearly, but not quite good enough. One might speculate that, in the judgement of **Interpro** at least, the chance of a false ***negative*** is too high to consider the **PRINTS** model seriously for **Homeobox** detection.

You could just believe me when I claim the **PRINTS HOMEOBOX model** does not work in this instance? Instead just speed read the next two pages concentrating only on the last bit which covers the struggles of **PRINTS** to find a **HOMEOBOX** (**recommended**), or you could prove all for yourself by doing the search. Just for the few doubters and those of you who have nothing better to do, I offer full instructions here (although I feel sure you could work it all out for yourselves).

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[13]:

http://130.88.97.239/PRINTS/

Select the **FPScan** link 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

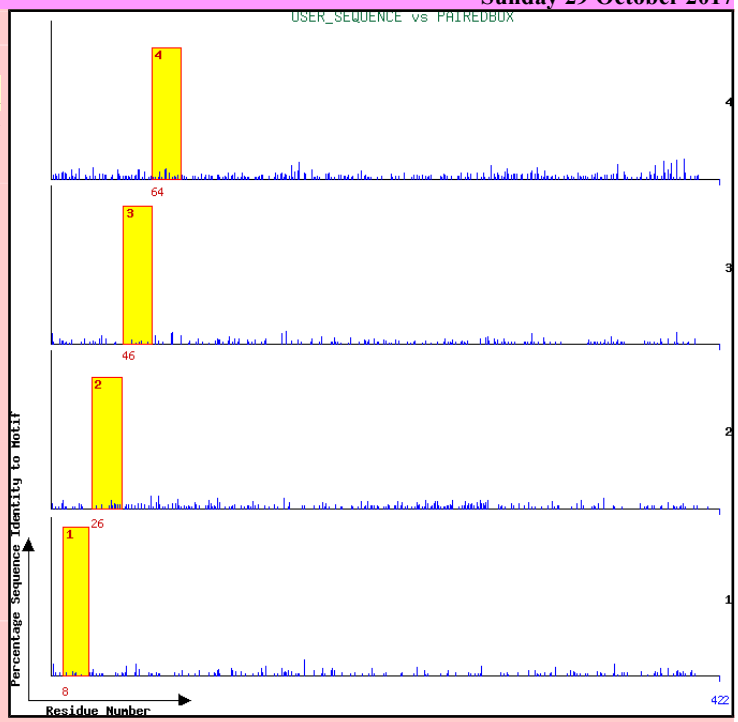
In the list of **Ten top scoring fingerprints**, there is a second **fingerprint** that matches all elements in the correct order. This is the **HTHREPRESSR**. Click on the **HTHREPRESSR** link and from the documentation you can confirm that an **HTHREPRESSOR** is an **HTH** motif of which you might have reasonably expected three? Move back to your **fingerPRINTscan** results. Shimmy down to the **Ten top scoring fingerprints**.

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

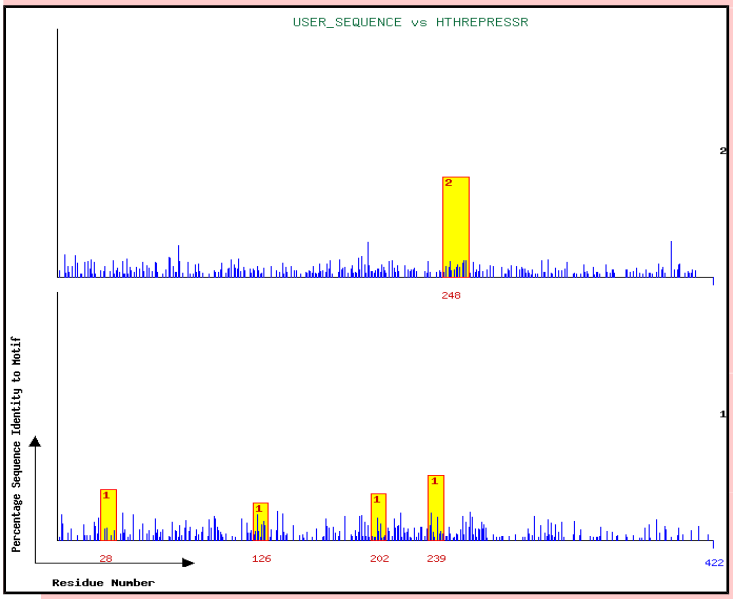
From the Position information included in the **Detailed by motif** table, you can see that the **HTH** motif that **fingerPRINTscan** finds is the one that is part of the **HOMEOBOX** domain it fails to fully detect. **PRINTS** does not see the **HTH**s in the **PAIREDBOX** domain.

13 Despite the inexplicitly undecorated URL, this is the currently official **PRINTS** home page provided by **Manchester** University.

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.

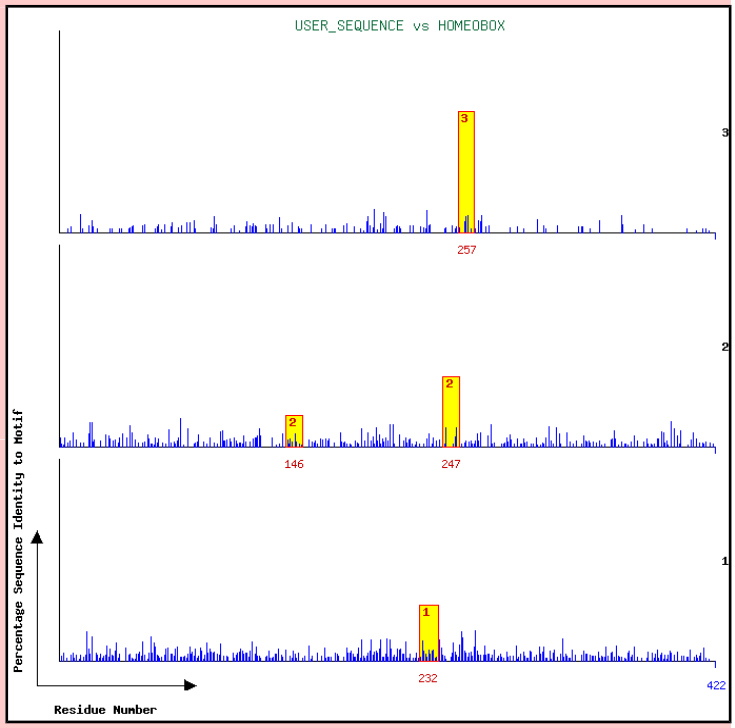


Each motif by itself might not be significant. Together, in a **fingerprint**, they constitute a confident prediction for a **Paired Box domain**.



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 breadth short for the first motif!

From the Top ten scoring fingerprints table, you can see that **fingerPRINTscan** considers the first motif to be missing ("**.II**"). But I 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**!!

DPJ – 2017.10.29