Earlham Institute summer school on bioinformatics

25-29 July 2016

Basic Bioinformatics Sessions

Practical 3: Databases Searching

Wednesday 6 July 2016

Searching for sequence similarities in databases.

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

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

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

Go to the **NCBI** homepage at:

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

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

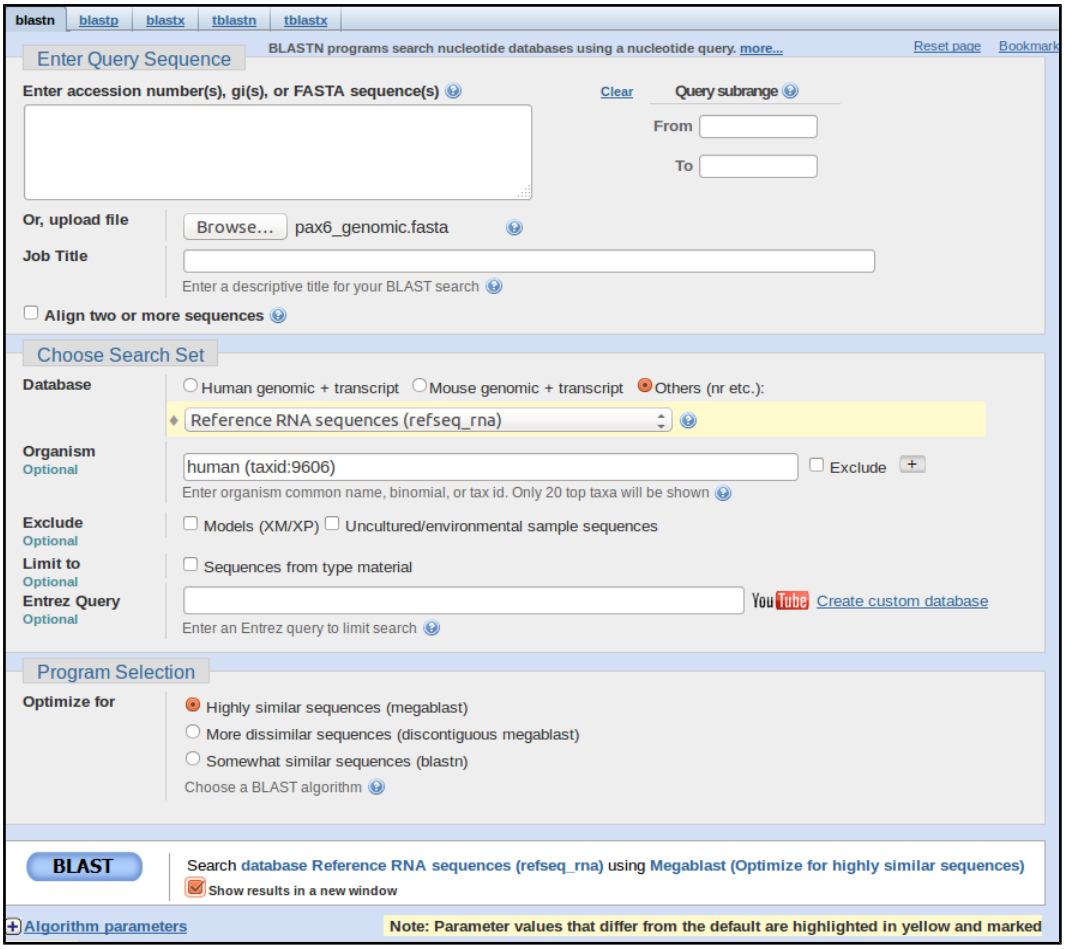
pax6_genomic.fasta.

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

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

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

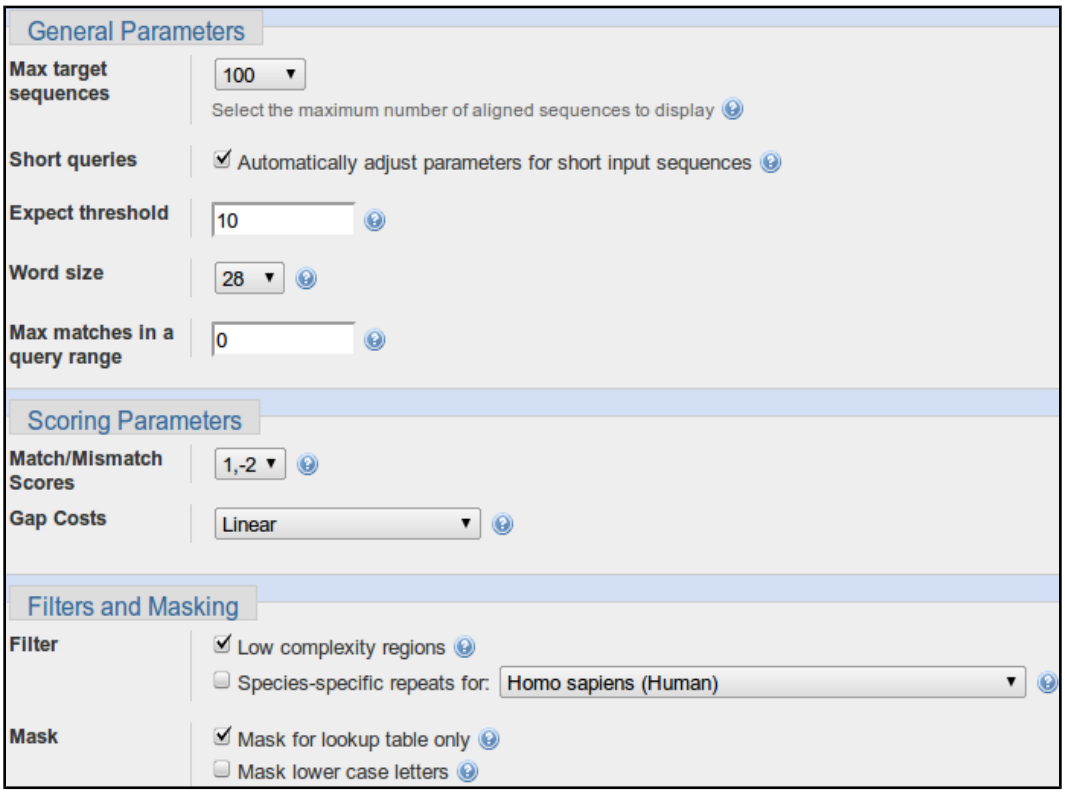
blast is now set to compare the **PAX6** genomic region with all **Human** mRNA sequences in **RefSeq**.



Note that the default **Program Selection** is **Highly similar sequences (megablast[1])**, which seems appropriate here as all the mRNA that correctly match should surely do so almost perfectly.

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

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



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

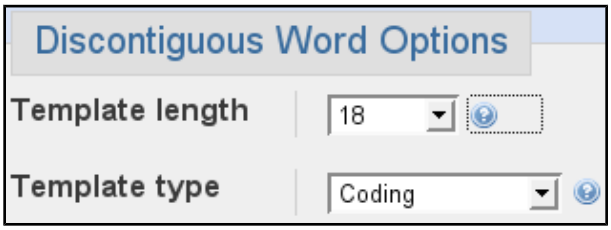
There are differences for **megablast**, where speed is of the essence and sensitivity can be sacrificed.

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

Gapped alignment is time consuming and, by default, considered more crudely by **megablast** than the other two algorithms[2].

Filtering and Masking matches with organism specific repeats and/or low complexity regions takes time, and so only avoiding **Low complexity regions**[3] is on by default for all **Program Selections**.

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



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

_____

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

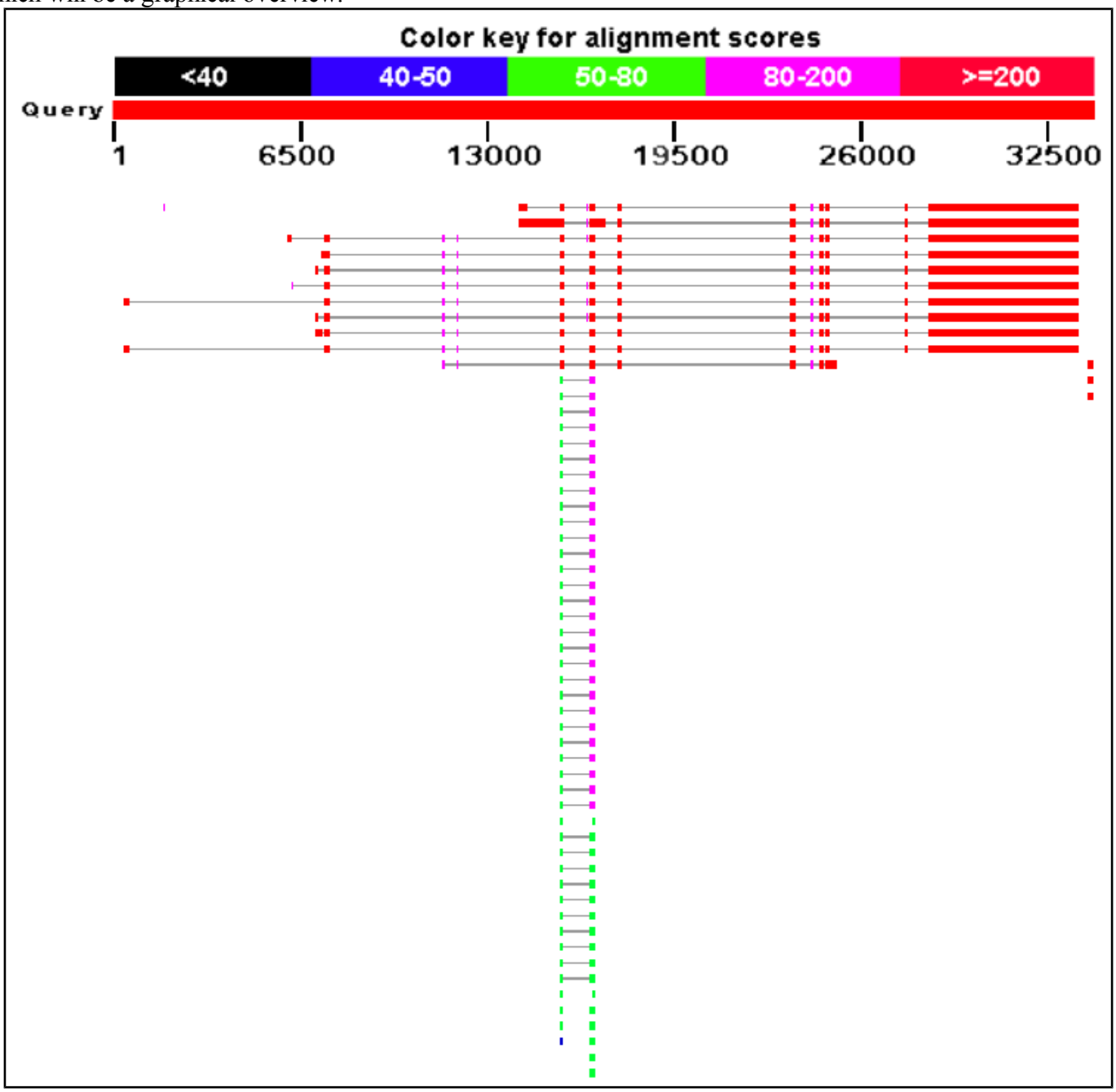
Which parameters would **blast** need to **automatically adjust** to cater for short input sequences (such as primers being tested for uniqueness), and why? _____

_____

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

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

Finally, ensure all the defaults are back in place[4] and the **megablast** is the **Program Selection**, ask **blast** to **Show results in a new window** and then click on the BLAST button. Impressively swiftly, you will have results. At the top of which will be a graphical overview.



This graphic implies that there are **11** full length matches between the genomic sequence and mRNAs in **RefSeq**. The **RefSeq** entries had to be "gapped" in order to compensate for the introns that are represented in the genomic sequence but not in the mRNA sequences. The **red blocks** therefore represent very closely matching (>=200 brownie points) exons, the lines joining the **red blocks** represent introns that have been spliced out. All **11** full length hits match reasonably uniformly except for the first few exons, implying significant variation in the **5′ UTR**.

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

_____

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

_____

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

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

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

In summary, if you hover over the graphical hits, their origin will be displayed above the graphic[5]. The facts are:

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

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

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

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

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

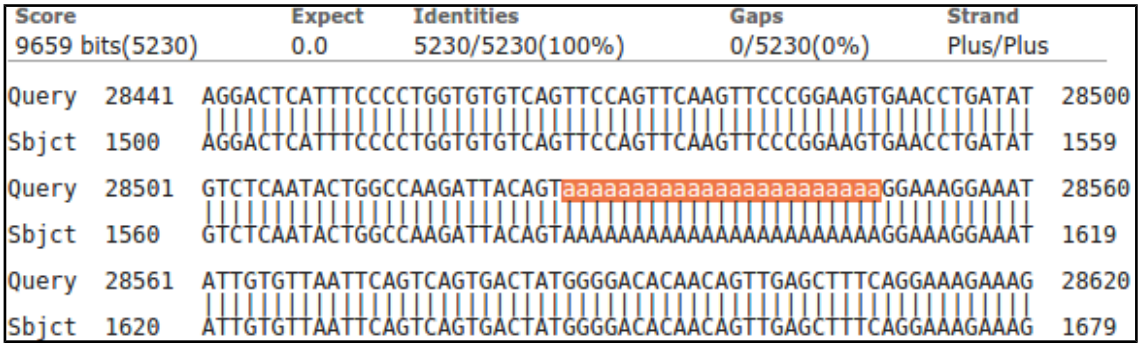
Why were you not surprised to discover **24 PAX6** transcripts in **Refseq** matching this sequence? _____

_____

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

Moving further down the results you will come to the alignments between **pax6-genomic** and the matching database entries. All similarity searches use local alignment strategies[6], so you should not be surprised to see a number of alignments for each "hit" in the list. Here we have a genomic query sequence aligned exclusively with mRNA sequences from **RefSeq**. The expectation is therefore to find an alignments corresponding to exons. The alignments are ordered by quality, though you are provided with a **Sort by:** menu to alter the order to taste[7].

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



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

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

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

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

Notice the lower case string of '**a**'s. The case indicates that they were ignored (**filtered**) as a **Low complexity region** whilst **megablast** was looking for identically matching words that might suggest matching regions[8]. By themselves, the '**a**'s are not sufficient evidence that a biological match exists. Only because the surrounding sequence is compellingly similar, can it be assumed that such a match does exist. The '**a**'s are replaced (lower case to indicate they were filtered) when the final alignment is computed. If you look a little further down the same alignment, you will see several other runs of '**a**'s and '**t**'s for which the same explanation applies.

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

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

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

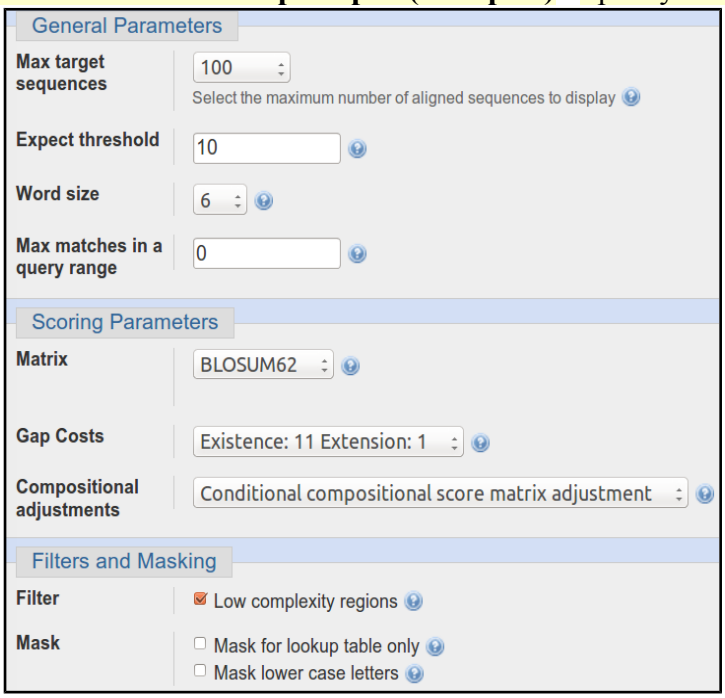
Now use a version of **blast** (called **blastx**) to compare your genomic sequence with a protein database. **blastx** will translate a DNA query sequence in all six reading frames and compare each translation with a protein sequence database. Thus, in a similar fashion to that employed by the **Ensembl**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

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

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

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

Take a look at the **Algorithm parameters**[9].



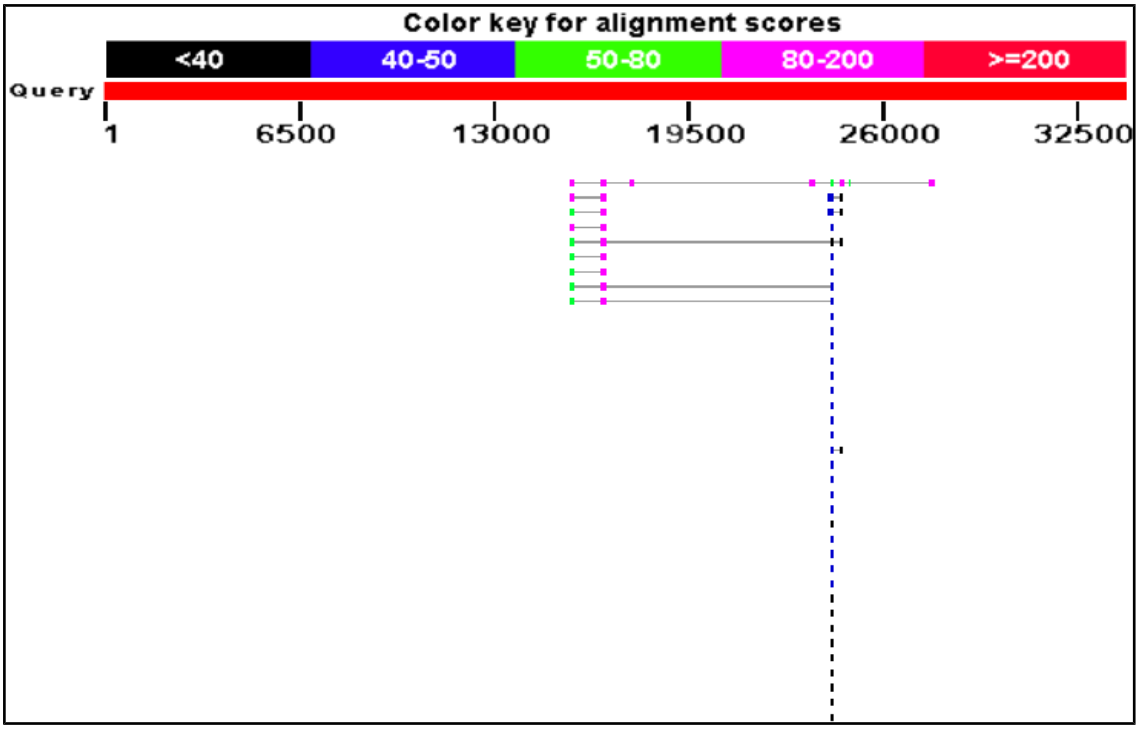
The **Word size** choice is **2, 3** or **6** . The default is **6**. We seek very close matches here, so the largest **Word size** would seem appropriate.

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

Low complexity regions will be filtered by default.

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

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



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

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

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

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

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

Alignments Download GenPept Graphics

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

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

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

_____

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

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

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

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

PSI-BLAST

This program is used to find a comprehensive set of relatives of a protein. First, **BLAST** is used to find closely related proteins. From an alignment of these proteins a general "profile" (a **P**osition **S**pecific **S**coring **M**atrix - **PSSM**) is computed. A **PSSM** is very similar in concept and purpose to an **HMM** profile in that it summarises significant features present in the sequences it represents.

A query against the protein database is then run using the **PSSM**, and a larger more widely associated group of proteins is found. This larger group is used to construct another **PSSM**, and the process is repeated until no more significantly matching new sequences can be detected, or the user tires of the whole process.

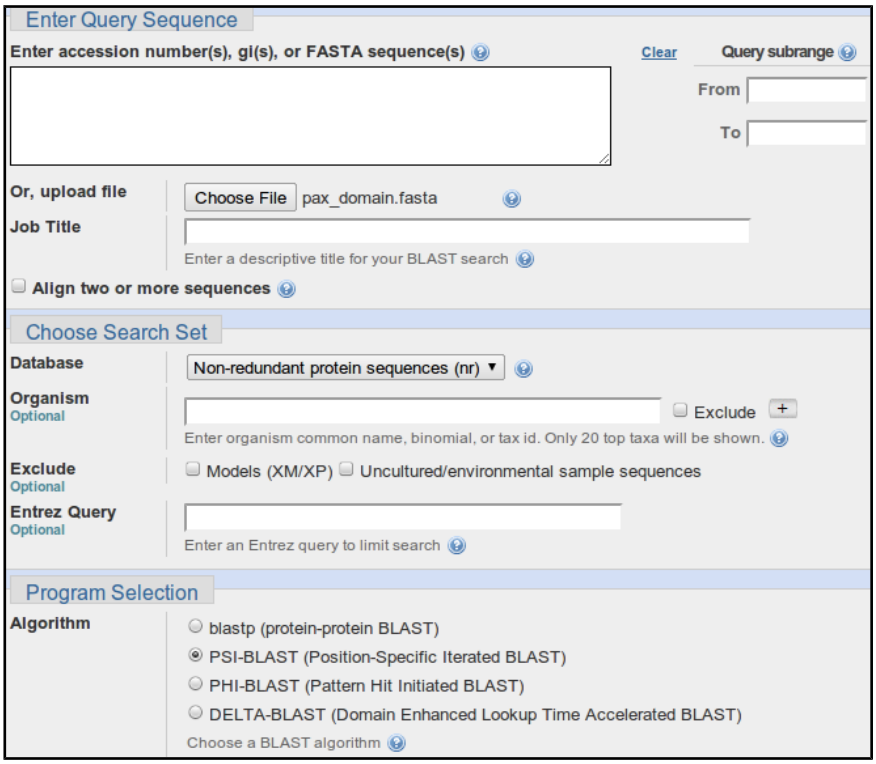
You have used **PSI-BLAST** integrated into **Jpred** already and similar ideas were used to create the **PFAM** alignments. Here we will use **PSI-BLAST** explicitly at the **NCBI** on the **Paired DOMAIN** of the **PAX6** protein that you saved in a file earlier. It should be possible to detect a large family of **PAX** domains and to eventually multiply align them generating something like the **Full** alignment from the **PFAM** database viewed earlier[10].

To investigate **PSI-BLAST** go first to the **NCBI** Home page at:

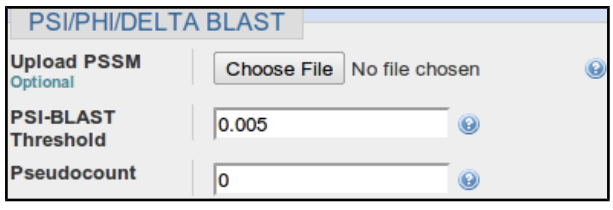
http://www.ncbi.nlm.nih.gov/

Click on the **BLAST** option. Select **protein BLAST** from the **Basic BLAST** section. Upload the **PAX6** paired box domain sequence (stored in the file **pax_domain.fasta**) using the appropriate **Browse** button.

Select **PSI-BLAST** from the **Program Selection** section. Leave all the others options at their default settings, particularly the option to search all the proteins available.

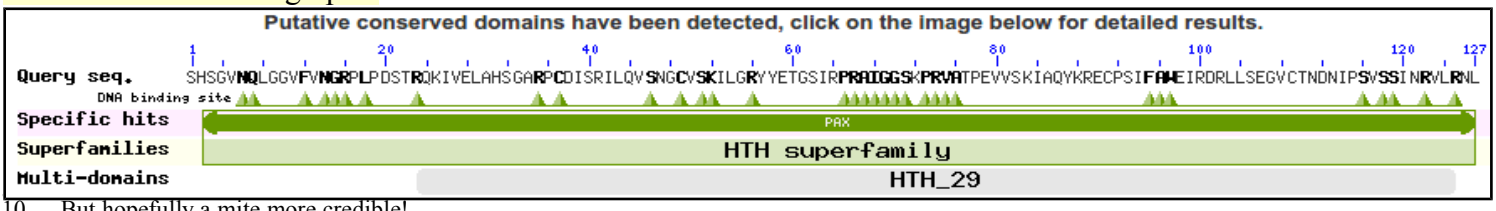


Before you set **PSI-BLAST** going, click on the **Algorithm parameters** link and take a look at the **PSI/PHI/DELTA BLAST** section. Here is offered the option to use a **PSSM** from a previous run **PSI-BLAST**, potentially on a different database (but with the same query sequence). Accept the default that database entries scoring better than an **Expect Threshold** of **0.005** be offered for inclusion into the **PSSM** of each successive **PSI-BLAST** iteration. Remember the buttons.



What do you suppose the choice of **Pseudocount** might influence? _____

Elect to **Show results in a new window** and then click on the BLAST button. After several moments of deep thought, **PSI-BLAST** will come back with its first set of results, at the top of which is a report that (unsurprisingly) matches have been detected between the query sequence and several domain databases. For more detail, click on the **Conserved Domains** graphic.



10 But hopefully a mite more credible!

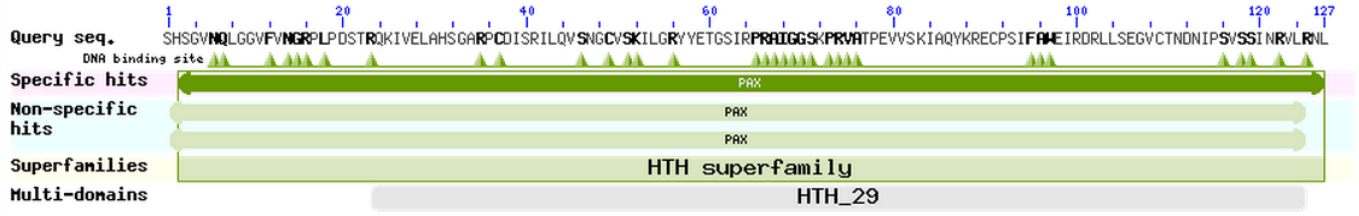
SMART, **Pfam** and the **NCBI Conserved Domains** database hits are reported. None should be a surprise.

Conserved domains on [lcl|Query_2485]

View: Standard Results

Pax-Domain P26367(4-130)

Graphical summary | Zoom to residue level | show extra options »



Search for similar domain architectures | Refine search

List of domain hits

Name	Accession	Description	Interval	E-value
[+] PAX	cd00131	Paired Box domain	2-127	5.03e-80
[+] PAX	smart00351	Paired Box domain;	1-125	2.30e-81
[+] PAX	pfam00292	'Paired box' domain;	1-125	2.38e-81
[+] HTH_29	pfam13551	Winged helix-turn helix; This helix-turn-helix domain is often found in transferases and is ...	23-125	1.15e-04

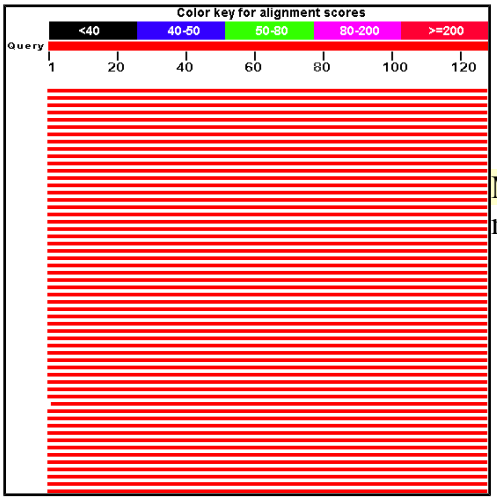
Blast search parameters

Data Source: Live blast search **RID =** XCKC805Y014

User Options: **Database:** CDSEARCH/cdd v3.14 **Low complexity filter:** no **Composition Based Adjustment:** yes **E-value threshold:** 0.01 **Maximum number of hits:** 500

References:

- Marchler-Bauer A et al. (2015), *"CDD: NCBI's conserved domain database."*, **Nucleic Acids Res.43**(D)222-6.
- Marchler-Bauer A et al. (2011), *"CDD: a Conserved Domain Database for the functional annotation of proteins."*, **Nucleic Acids Res.39**(D)225-9.
- Marchler-Bauer A et al. (2009), *"CDD: specific functional annotation with the Conserved Domain Database."*, **Nucleic Acids Res.37**(D)205-10.
- Marchler-Bauer A, Bryant SH (2004), *"CD-Search: protein domain annotations on the fly."*, **Nucleic Acids Res.32**(W)327-331.



Moving back to the main **PSI-BLAST** results, you will see that there are many high quality hits covering the whole length of the query sequence.

The best **500** of these are listed.

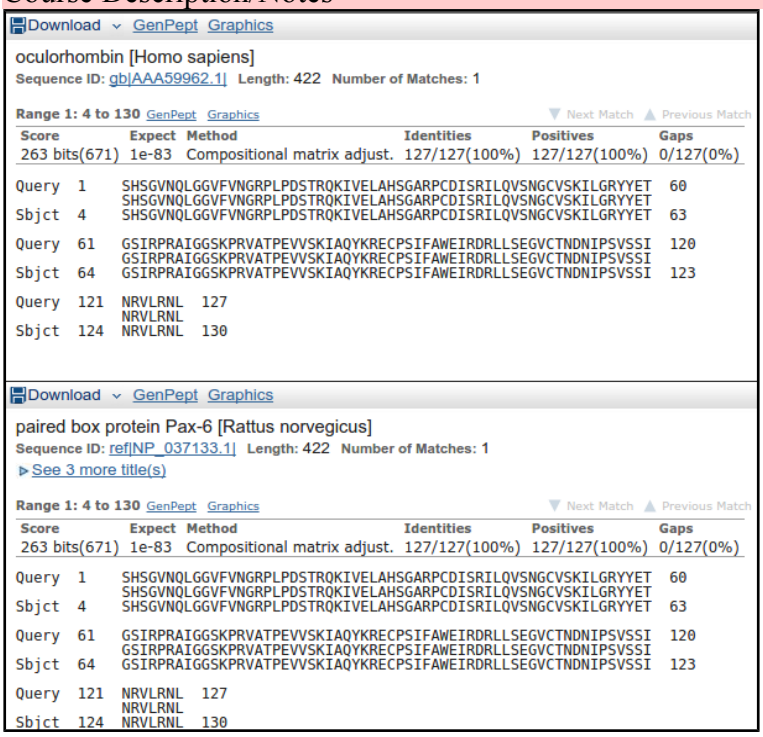
All the listed hits are selected for inclusion into the **PSSM** for the next iteration. Unless you feel strongly about any particular entry, leave them all selected.

Sequences producing significant alignments with E-value BETTER than threshold

Select: All None Selected:0

Alignments | Download | GenPept | Graphics | Distance tree of results | Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession	Select for PSI blast	Used to build PSSM
PREDICTED: paired box protein Pax-6 isoform X4 [Macaca nemestrina]	262	262	100%	1e-83	100%	XP_011722295.1	☑	
PREDICTED: paired box protein Pax-6 isoform X2 [Ursus maritimus]	263	263	100%	1e-83	100%	XP_008685073.1	☑	
oculorhombin [Homo sapiens]	263	263	100%	1e-83	100%	AAA59962.1	☑	
paired box protein Pax-6 [Rattus norvegicus]	263	263	100%	1e-83	100%	NP_037133.1	☑	
PREDICTED: paired box protein Pax-6 isoform X2 [Fukomys damarensis]	263	263	100%	1e-83	100%	XP_010638711.1	☑	
PREDICTED: paired box protein Pax-6 isoform X2 [Cavia porcellus]	263	263	100%	1e-83	100%	XP_003464531.1	☑	
PREDICTED: paired box protein Pax-6 isoform X2 [Aotus nancymaae]	263	263	100%	1e-83	100%	XP_012307699.1	☑	
PREDICTED: paired box protein Pax-6 isoform X2 [Callorhinchus milii]	263	263	100%	1e-83	100%	XP_007885973.1	☑	
PREDICTED: paired box protein Pax-6 isoform X2 [Heterocephalus glaber]	263	263	100%	1e-83	100%	XP_004851665.1	☑	
PREDICTED: paired box protein Pax-6 isoform X2 [Octodon degus]	263	263	100%	1e-83	100%	XP_004638029.1	☑	
PREDICTED: paired box protein Pax-6 [Poecilia reticulata]	261	261	100%	1e-83	98%	XP_008404092.1	☑	


Download GenPept Graphics

oculorhombin [Homo sapiens]

Sequence ID: gb|AAA59962.1| Length: 422 Number of Matches: 1

Range 1: 4 to 130 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
263 bits(671)	1e-83	Compositional matrix adjust.	127/127(100%)	127/127(100%)	0/127(0%)

```
Query  1    SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYET  60
            SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYET
Sbjct  4    SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYET  63

Query  61   GSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSI  120
            GSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSI
Sbjct  64   GSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSI  123

Query  121  NRVLRNL  127
            NRVLRNL
Sbjct  124  NRVLRNL  130
```

Download GenPept Graphics

paired box protein Pax-6 [Rattus norvegicus]

Sequence ID: ref|NP_037133.1| Length: 422 Number of Matches: 1

See 3 more title(s)

Range 1: 4 to 130 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
263 bits(671)	1e-83	Compositional matrix adjust.	127/127(100%)	127/127(100%)	0/127(0%)

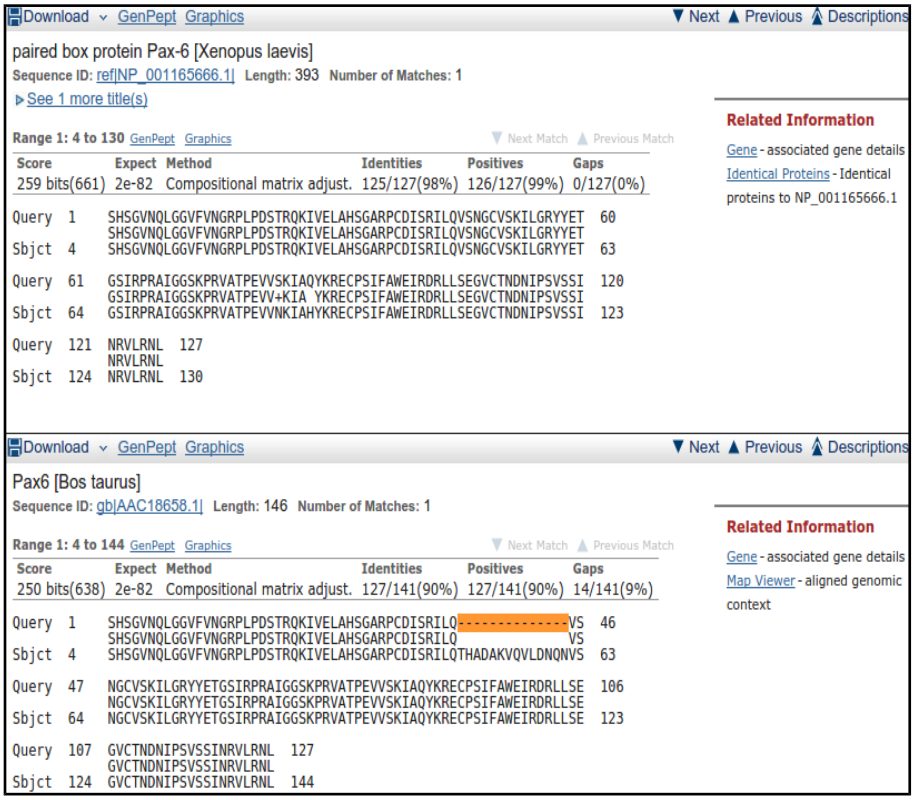
```
Query  1    SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYET  60
            SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYET
Sbjct  4    SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYET  63

Query  61   GSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSI  120
            GSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSI
Sbjct  64   GSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSI  123

Query  121  NRVLRNL  127
            NRVLRNL
Sbjct  124  NRVLRNL  130
```

How might the inclusion of relatively poor quality sequences and the presence of so much duplication have been minimised?_____

_____


Download GenPept Graphics Next Previous Descriptions

paired box protein Pax-6 [Xenopus laevis]

Sequence ID: ref|NP_001165666.1| Length: 393 Number of Matches: 1

See 1 more title(s)

Range 1: 4 to 130 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
259 bits(661)	2e-82	Compositional matrix adjust.	125/127(98%)	126/127(99%)	0/127(0%)

Related Information

Gene - associated gene details

Identical Proteins - Identical proteins to NP_001165666.1

```
Query  1    SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYET  60
            SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYET
Sbjct  4    SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYET  63

Query  61   GSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSI  120
            GSIRPRAIGGSKPRVATPEVV+KIA YKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSI
Sbjct  64   GSIRPRAIGGSKPRVATPEVVNKIAHYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSI  123

Query  121  NRVLRNL  127
            NRVLRNL
Sbjct  124  NRVLRNL  130
```

Move down far enough and you will see less perfect matches, some of which involve proteins with the extra 14 amino acids of **isoform 5a** of **PAX6_HUMAN**.

Download GenPept Graphics Next Previous Descriptions

Pax6 [Bos taurus]

Sequence ID: gb|AAC18658.1| Length: 146 Number of Matches: 1

Range 1: 4 to 144 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
250 bits(638)	2e-82	Compositional matrix adjust.	127/141(90%)	127/141(90%)	14/141(9%)

Related Information

Gene - associated gene details

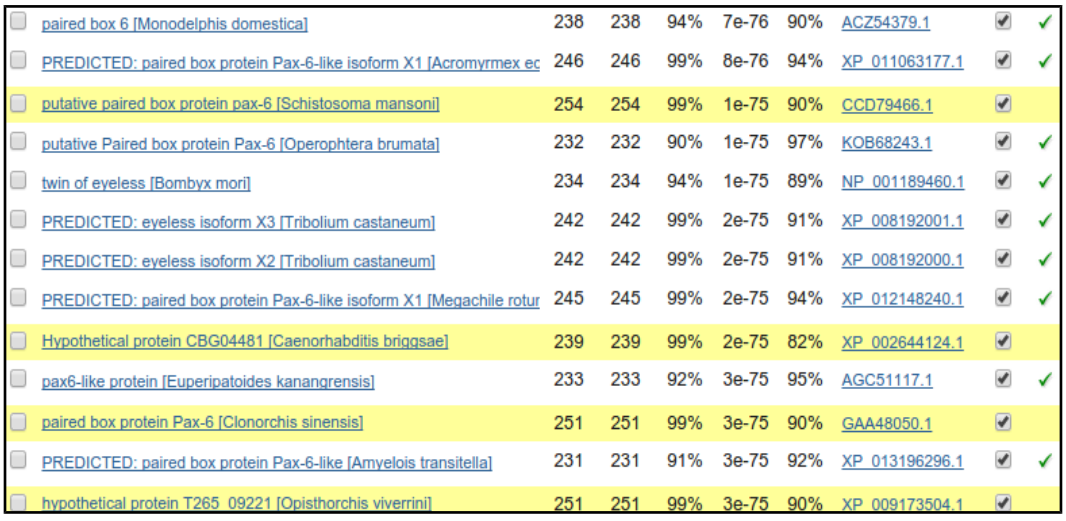
Map Viewer - aligned genomic context

```
Query  1    SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ----------VS  46
            SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ          VS
Sbjct  4    SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNVS  63

Query  47   NGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSE  106
            NGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSE
Sbjct  64   NGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSE  123

Query  107  GVCTNDNIPSVSSINRVLRNL  127
            GVCTNDNIPSVSSINRVLRNL
Sbjct  124  GVCTNDNIPSVSSINRVLRNL  144
```

Having browsed your results sufficiently, click on the Go button to **Run PSI-Blast iteration 2**. It is at the bottom of the hit list.

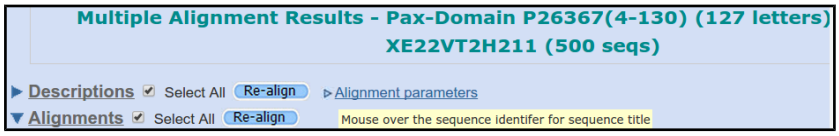


paired box 6 [Monodelphis domestica]	238	238	94%	7e-76	90%	ACZ54379.1	☑	✓
PREDICTED: paired box protein Pax-6-like isoform X1 [Acromyrmex ec	246	246	99%	8e-76	94%	XP_011063177.1	☑	✓
putative paired box protein pax-6 [Schistosoma mansoni]	254	254	99%	1e-75	90%	CCD79466.1	☑	
putative Paired box protein Pax-6 [Operophtera brumata]	232	232	90%	1e-75	97%	KOB68243.1	☑	✓
twin of eyeless [Bombyx mori]	234	234	94%	1e-75	89%	NP_001189460.1	☑	✓
PREDICTED: eyeless isoform X3 [Tribolium castaneum]	242	242	99%	2e-75	91%	XP_008192001.1	☑	✓
PREDICTED: eyeless isoform X2 [Tribolium castaneum]	242	242	99%	2e-75	91%	XP_008192000.1	☑	✓
PREDICTED: paired box protein Pax-6-like isoform X1 [Megachile rotur	245	245	99%	2e-75	94%	XP_012148240.1	☑	✓
Hypothetical protein CBG04481 [Caenorhabditis briggsae]	239	239	99%	2e-75	82%	XP_002644124.1	☑	
pax6-like protein [Euperipatoides kanangrensis]	233	233	92%	3e-75	95%	AGC51117.1	☑	✓
paired box protein Pax-6 [Clonorchis sinensis]	251	251	99%	3e-75	90%	GAA48050.1	☑	
PREDICTED: paired box protein Pax-6-like [Amyelois transitella]	231	231	91%	3e-75	92%	XP_013196296.1	☑	✓
hypothetical protein T265_09221 [Opisthorchis viverrini]	251	251	99%	3e-75	90%	XP_009173504.1	☑	

After a few moments, **PSI-BLAST** will return with the results of searching through the database again using the **PSSM** derived from the hits of the first iteration(☑ed). This time the top of the list will be predominantly filled with hits that have already been incorporated into the **PSI-BLAST PSSM**. However, look far enough down the list and you will find some new ones, highlighted yellow.

Once more, click on the Go button to **Run PSI-Blast iteration 3**. That is probably enough! It took 4 iterations before there were no more new sequences suggested for inclusion into the **PSMM** when I ran this last, so if you really want to take things to their logical conclusion, it should not detain you long.

Next, move to the top of the **Descriptions** list and **Select All**. Click on the **Multiple Alignment** button. You have elected to use the **NCBI** multiple alignment program **Cobalt** to align all the **PAX** domain sequences of your final **PSI-BLAST** iteration that match with an **Expect** score better than **0.001**. In an impressively short time, your alignment will appear.



Move past the long list of proteins that have been aligned (the easiest way is to hide the **Descriptions** view).

At the top of the actual alignment, set **View Format** to **Plain Text** (…. and then hide the **Descriptions** again??), this being the easiest format to understand in a hurry. The alignment will have very ragged ends, but the important region of **120** or so amino acids representing the **PAX** domain is really quite impressive. In particular, the **isoform 5a** insertion is very convincing[11].

XP_003977912	52	TRQKIVELAHSGARPCDISRILQTHDA--VQVLDSEKV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	114
XP_009296159	26	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNENV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	90
XP_003246075	54	TRQKIVELAHSGARPCDISRILQ-------------V----------SNGCVSKILG----RYYETGSIKPRAIGGSK	104
XP_012793883	41	TRQRIIELAHSGARPCDISRILQ-------------V----------SNGCVSKILC----RYYETGSIRPKAIGGSK	91
XP_005991286	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDIQNV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
EFX75780	37	TRQKIVELAHSGARPCDISRILQ-------------V----------SNGCVSKILG----RYYETGSIRPRAIGGSK	87
ABB43131	25	TRQKIVELAHSGARPCDISRILQ-------------V----------SNGCVSKILG----RYYETGSIRPRAIGGSK	75
ETN66652	41	TRQKIVELAHSGARPCDISRILQ-------------V----------SNGCVSKILG----RYYETGSIKPRAIGGSK	91
XP_006128959	56	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	120
XP_010874560	44	TRQKIVELAHSGARPCDISRILQTHDDSKVQVLDNENV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	108
AFJ24746	53	TRQRIVELAHSGARPCDISRILQ-------------V----------SNGCVSKILC----RYYETGSIRPKAIGGSK	103
XP_007885968	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVVDNRKV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
BAA24024	42	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDSQNV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	106
XP_012307695	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
CBY09679	55	TRQKIVELAHSGARPCDISRILQ-------------V----------SNGCVSKILA----RYYETGSIKPRAIGGSK	105
XP_007181079	82	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	146
CAF29075	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDSENV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
XP_004264009	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
XP_009184622	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
AAW24017	55	TRQKIVELAHSGARPCDISRILQ-------------V----------SNGCVSKILA----RYYETGSIKPRAIGGSK	105
XP_008547741	26	TRQKIVELAHSGARPCDISRILQ-------------V----------SNGCVSKILG----RYYETGSIRPRAIGGSK	76
XP_012162452	50	TRQKIVELAHSGARPCDISRILQ-------------V----------SNGCVSKILG----RYYETGSIKPRAIGGSK	100
XP_006975926	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNENV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
KDR14710	21	TRQKIVELAHSGARPCDISRILQ-------------V----------SNGCVSKILG----RYYETGSIRPRAIGGSK	71
XP_005530321	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
ABI98847	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
XP_010794780	44	TRQKIVELAHSGARPCDISRILQTHDE--VQVLDSEKV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	106
NP_001103907	26	TRQKIVELAHSGARPCDISRILQ-------------V----------SNGCVSKILG----RYYETGSIRPRAIGGSK	76
XP_010356630	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
XP_010638709	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
XP_005064878	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNENV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
NP_038655	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNENV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
XP_005401829	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
XP_004638028	25	TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNV----------SNGCVSKILG----RYYETGSIRPRAIGGSK	89
BAM74254	32	TRQRIVELAHSGARPCDISRILQ-------------V----------SNGCVSKILG----RYYETGSIRPRAIGGSK	82

11 Much more so than the **Full** alignment offered by **PFAM**, I would contend. Although, it has to be admitted, the **Pfam** alignment included more sequences and I suspect they would have gone for a less closely homologous set of sequences. Even so … I think the alignment illustrated here is **MUCH** more beautiful!!

Model Answers to Questions in the Instructions Text.

Notes:

For the most part, these "**Model Answers**" just provide the reactions/solutions I hoped you would work out for yourselves. However, sometime I have tried to offer a bit moer back ground and material for thought? Occasionally, I have rambled off into some rather self indulgent investigations that even I would not want to try and justify as pertenent to the objective of these exercises. I like to keep these meanders, as they help and entertain me, but I wish to warn you to only take regard of them if you are feeling particularly strong and have time to burn. Certainly not a good idea to indulge here during a time constrained course event!

Where things have got extreme, I am going to make two versions of the answer. One starting:

Summary:

Which has the answer with only a reasonably digestible volume of deep thought. Read this one.

The other will start:

Full Answer:

Beware of entering here! I do not hold back. Nothing complicated, but it will be long and full of pedantry.

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

From your investigations of **Searching for sequence similarities in databases**

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

Generally, whenever one might suspect the automatic masking algorithms of **blast** might miss a non informative region in a specific query sequence, obviously.

A specific example might be when a query sequence contained a significant informative region that was known to be common amongst the sequences being searched. If this region was left unmasked, **blast** would pick up so many similar matches to this one region that other interesting similarities might be obscured. By manually masking such a region by changing it to lower case, its matches would not be seen by **blast** and matches with other regions of the query sequence should be more apparent.

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

The **word size**: Clearly, if you are trying to find matches for a primer (for example) of around **20** base pairs, it would be pretty silly to use a **word size** of **28** (default for **megablast**). A **word** the same size as the primer would find only exact matches. A **word** of about **7** would allow a couple of mismatches and would probably be most generally appropriate.

The **expect score**: As good chance matches between between a short query sequence and a large database will be abundant, it would not be sensible to choose a demanding (i.e. small) **expect score** to represent the limit of significance. In particular, a primer sized query sequence of around **20** base pairs might easily exactly match more than **10** times (generally the default maximum expect score for a significant match) just by chance. After all, there are only **4** bases, a string of **20** is not that long and the databases can be huge! Typically **blast** chooses very high **expect score** cut off for short query sequences, effectively removing the **expect score** filter altogether.

Earlier versions of **blast** did not automatically adjust these parameters. When a short query sequences were selected, suitable adjustment was left to the user. Without sensible parameter adjustment, results could be greatly confusing. For example, a **21** base pair primer could easily match perfectly more than **10** times against a large DNA sequence database. **blast** is set to ignore matches that are expected to occur more than **10** times by chance. Thus even exact matches with such a small sequences would be ignored! Now automatic parameter adjustment is undertaken by **blast**, the user does not really have to think too hard. However, it does seem to be a good idea to know what **blast** is doing and why.

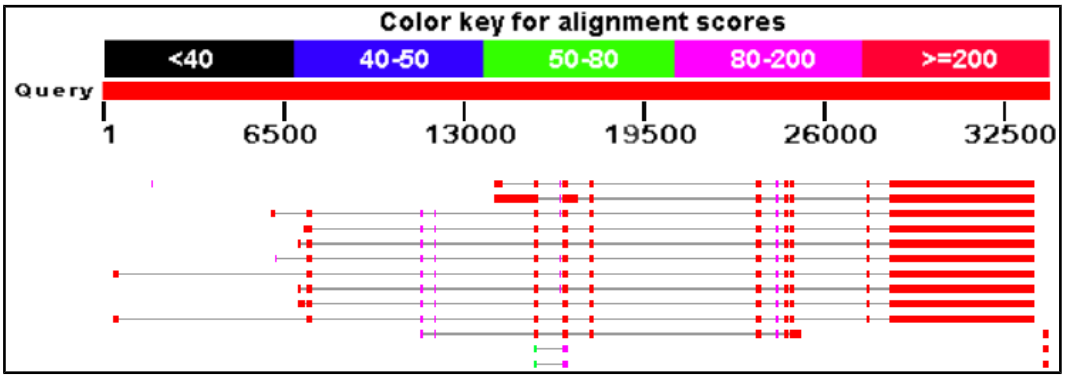
Why do you suppose that a few of the exons do not achieve the maximum score?

Summary:

Each local region of significant alignment between a database entry and a query sequence is scored independently. The scoring method that governs the alignment score colour in this graphic, reflects both the quality of the match ***and*** its length. Unless a particular region is of sufficient length, it cannot achieve the **200 bit** threshold even if the alignment is perfect. Note that is is the shorter regions that fail to reach the >=200 status. All of the illustrated local alignments associated with **PAX6** transcripts are essentially perfect.

Full Answer:

In common with most database searching programs, **blast** compares query sequences with database entries using a local strategy. The overall evaluation of a particular query sequence is taken to be the highest local score.



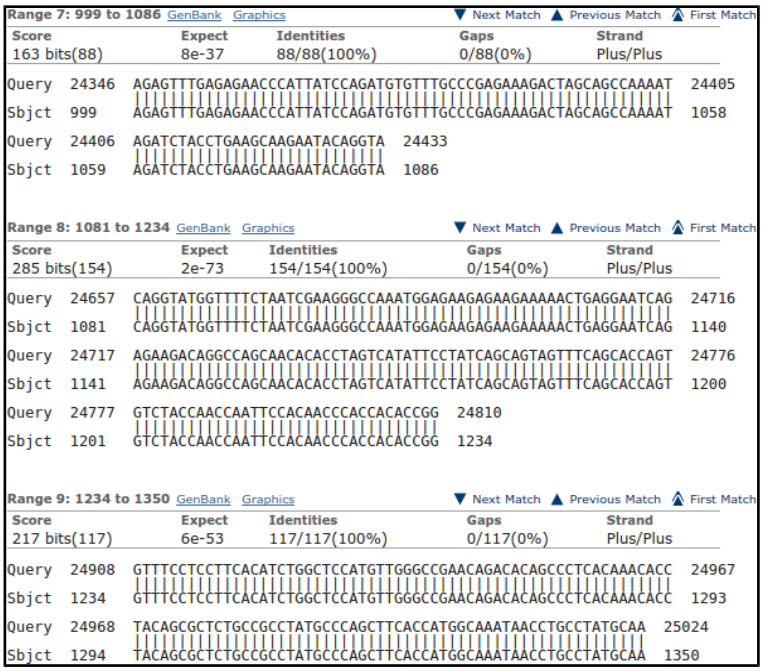
Individual local matches are coloured according to individual quality. In this query, all true matches should be perfect, or very nearly so. Scores might therefore be expected to be maximal (>=200). However, they are not? Some only manage a score in the range 80-200.

The score referenced for this purpose is the **bit score**. For a full, no holds barred definition of this score, try here. I prefer this somewhat gentler version:

> "The **bit score** gives an indication of how good the alignment is; the higher the score, the better the alignment. In general terms, this score is calculated from a formula that takes into account the alignment of similar or identical residues, as well as any gaps introduced to align the sequences. A key element in this calculation is the "substitution matrix ", which assigns a score for aligning any possible pair of residues. The **BLOSUM62** matrix is the default for most **BLAST** programs, the exceptions being **blastn** and **MegaBLAST** (programs that perform **nucleotide–nucleotide** comparisons and hence do not use protein-specific matrices). Bit scores are normalized, which means that the bit scores from different alignments can be compared, even if different scoring matrices have been used."

Still too scary? The important things to note are that:

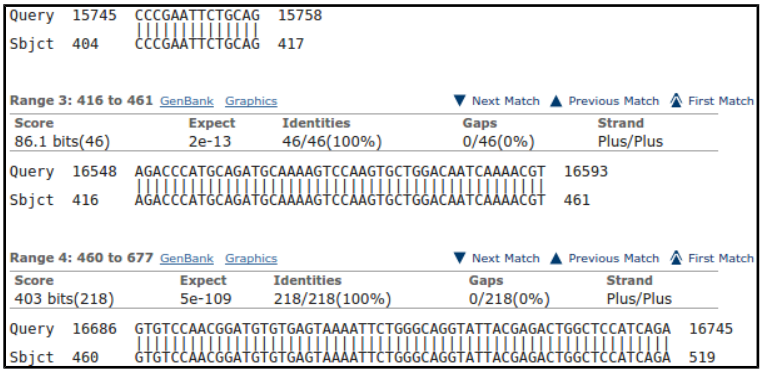
- These scores are based on a simple DNA scoring matrix (**1** for a match, **-2** for a mismatch by default for **megablast**), plus penalties for gaps. So scores will be limited by the length of the alignment, ignoring gaps.
- The scores reflect penalties for **indel**s (**in**sertions or **del**etions).
- The scores are normalised so that they do not depend on the chosen scoring matrix. This allows bits scores from searches using different scoring matrices to be compared.



This being so, **bit scores** will reflect the length of an alignment as well as its quality. If an alignment is very short, it might be perfect but still not achieve a very high value. **bit scores** are designed to reflect significance, not just local quality. A short perfect match clearly can be less significant than a longer less perfect match. That is what you see illustrated here.

You can see evidence of what is occurring in the alignments further down your results. Here is illustrated one of the 80-200 exons that occur in all transcripts at position **24,346**[12]. The match is perfect, but the length of the exon is consistently just to short to get to the heady >=200 level.

Note how imperfectly **blast** finds exon/intron boundaries. If the start of an intron happens to match the start of the next exon, **blast** will included the bases in two alignments[13]. It is not looking for exons and introns as was **spline**, it just mindlessly seeks matches.



For a further example, look at the exon that is found only in the **isoform 5a** transcripts. It is tiny (**42** base pairs) and scores well below >=200 even thought it is a perfect match.

Note that the alignment is **46** base pairs long due to **blast** adding on two bases either side that are actually the highly conserved intron start and end base pairs. As you can see, these extra base pairs occur in the preceding and succeeding alignment also.

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

Oh dear oh dear! Not this again.

Well I refer to the **isoform 5a** exon, of course. The tiny inconsistent one about **9** exons in from the right (when it exists). This will, clearly, only occur in **isoform 5a** transcripts. One day I will tidy these questions up a trifle!

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

Repetitive? True. This question is more "interesting" when the resources you have visited do not agree. That is, most of the time. Whilst the situation is in balance, the answer is:

Because **GeneCards** says there should be **11** quality and a further **13** less supported **PAX6 mRNA** sequences in **RefSeq**. A total of **24 PAX6** implied transcripts in total. In passing, you could have discovered the number of **PAX6 mRNA** sequences in **RefSeq** but asking **RefSeq** directly. Probably a more sensible and certainly a more reliable approach.

12 In order to make this illustration, I needed set **Sort by**: (top of the alignments) to **Query start position**.

13 **6** base pairs (**Sbjct**: **1081-1086**, **CAGGTA**) occur in both the first two matches illustrated. Just **1** base pair is shared between the **2nd** and **3rd** match (**Sbjct**: **1234**, **G**).

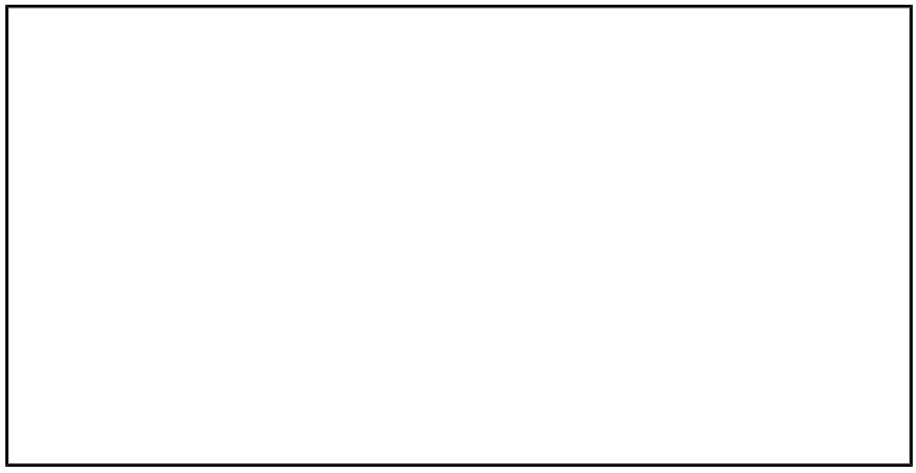
Which of the **Refseq PAX6** transcripts corresponds to **isoform 5a**?

Summary:

As I am sure you are tired of noting by now, all the transcripts with the extra tiny exon around position **1,600** in the genomic sequence are **isoform 5a** transcripts.

Full Answer:

The **isoform 5a** transcripts can be spotted most easily from the graphic. They are the ones with the extra small exon slightly to the left of middle (around base position **1,600**). For example, the **first**, **second** and **third blast** matches displayed. If you hover over these matches with your mouse, you will see that they are t**ranscript variants 1**, **2**, **3**, **6**, **7**, **8**, **11**, **12**, **14**, **18**, **19**, **20**, **21** and **23** (in the vertical order of the graphic).



Stated with the unequalled poetry of **RefSeq Accession Code** and lyrical **Title** Line, that becomes:

TITLE	ACCESSION CODE
Homo sapiens paired box 6 (PAX6), transcript variant 11, mRNA	NM_001310161.1
Homo sapiens paired box 6 (PAX6), transcript variant 10, mRNA	NM_001310160.1
Homo sapiens paired box 6 (PAX6), transcript variant 8, mRNA	NM_001310158.1
Homo sapiens paired box 6 (PAX6), transcript variant 5, mRNA	NM_001258463.1
Homo sapiens paired box 6 (PAX6), transcript variant 4, mRNA	NM_001258462.1
Homo sapiens paired box 6 (PAX6), transcript variant 2, mRNA	NM_001604.5
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X13, mRNA	XM_005252958.3
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X12, mRNA	XM_011520153.1
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X10, mRNA	XM_011520152.1
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X6, mRNA	XM_011520150.1
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X5, mRNA	XM_011520149.1
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X4, mRNA	XM_005252954.3
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X3, mRNA	XM_011520148.1
PREDICTED: Homo sapiens paired box 6 (PAX6), transcript variant X1, mRNA	XM_011520146.1

Yes well, that was fun? When I first wrote this question, there was only **5** or **6 RefSeq** mRNAs! The message of the question was to ensure you could see how to spot the **isoform 5a** transcripts (again!), not to list them! But, never mind, doing so was in fine tune with the ennui of the moment.

Additional Meanderings:

That really should not detain anyone but me? They belong after the consideration of masking the run of As for the **megablast** you ran. I just enjoyed this detour, so I keep it somewhere low profile. The whole journey leads nowhere of any note, so, should you decide to read, expect little!

The **500** base pairs of 3' flanking sequence added on to the **Ensembl** sequence for "good measure", is not part of the alignment (as would be expected). This can be seen easily if you look at the end of the alignment illustrated above, which is the alignment of the last exon of a transcript.

The entire length of this transcript is **6,732** base pairs.

Homo sapiens paired box 6 (PAX6), transcript variant 11, mRNA
Sequence ID: ref|NM_001310161.1| Length: 6732 Number of Matches: 11

The entire length of the genomic query sequence is **34,170** base pairs.

Query ID	lcl|Query_71179
Description	pax6-genomic sequence
Molecule type	nucleic acid
Query Length	34170

The alignment ends at position position **6,729** of the mRNA (**3** from the end) and position **33,670** of the genomic sequence (exactly **500** base pairs from the end).

```
Query  33601  ATTTGACATCCTGGCAAATCACTGTCATTGATTCAATTATTCTAATTCTGAATAAAAGCT  33660
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  6660   ATTTGACATCCTGGCAAATCACTGTCATTGATTCAATTATTCTAATTCTGAATAAAAGCT  6719

Query  33661  GTATACAGTA  33670
              ||||||||||
Sbjct  6720   GTATACAGTA  6729
```

The **3** missing base pairs of the mRNA are all **A**s, due to **polyadenylation**. Position **6,729** being recorded as a **polyA** site by **RefSeq**. A very short **polyA** tail surely? But there is no telling what stage of the mRNA is recorded on **RefSeq**. **Wikipedia** says:

```
polyA_site      2495
                /gene="PAX6"
                /gene_synonym="AN; AN2; D11S812E; FVH1; MGDA; WAGR"
```

```
    6541 aaaaaaatag aataagaaac ctgattttta gtactaatga aatagcgggt gacaaaatag
    6601 ttgtcttttt gattttgatc acaaaaaata aactggtagt gacaggatat gatggagaga
    6661 tttgacatcc tggcaaatca ctgtcattga ttcaattatt ctaattctga ataaaagctg
    6721 tatacagtaa aa
//
```

> "The tail is shortened over time, and, when it is short enough, the mRNA is enzymatically degraded."

Of course, the neatness of this observation does reflect less some profund biological truth than it does that this mRNA just happens to be one that extends furthest to the right in the genome, and there is no chance match between the **polyA** tail and the extra **500** bases of genomic sequence you added on when extracting it from **Ensembl**.

The journey was fun even though the destination was dubious. Much the way of a considerable portion of life in general one might reflect?

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

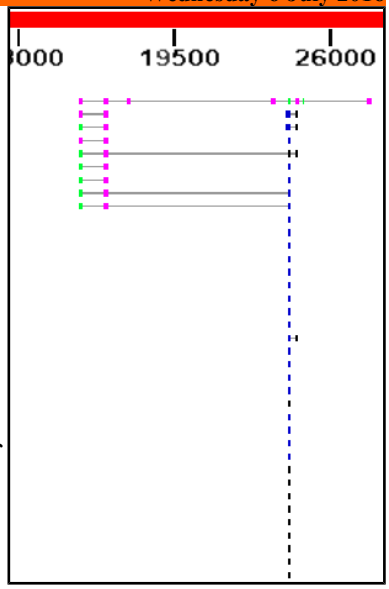
Matches between the regions of genomic DNA encoding **Paired Box** domains.

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

Because that is how many **Paired box** domains are suggested to be in the human genome by counting the number of quality **mRNA** sequences in **RefSeq** claiming to include a **Paired box** coding region. There is **PAX6** plus its **8** paralogues, imaginatively all named:

PAX1, **PAX2**, **PAX3**, **PAX4**, **PAX5**, **PAX6**, **PAX7**, **PAX8** & **PAX9**

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



These are matches between human **mRNA** sequences with the regions of genomic DNA encoding **Homeo Box** domains. As you discovered from **Interpro**, there are many of these.

The thin line joining features implies that those features relate to the same database entry.

Notice that **4** of the **9** proteins matching a **Paired box** genomic region also match a **Homeo box** region. the remaining **5** do not. This implies that **4** of the **9** proteins corresponding to the hits detected here have a **Paired box** domain near the start of the protein and a **Homeo box** domain further along. This is exactly as was suggested by the **PROSITE** annotation you examined.

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

Because they score more highly and so, in the opinion of **blast**, are more worthy. Primarily, they score more highly because they are longer. The list is ranked by **E Value**. Good matches with long sequence are less likely to occur by chance than equally good matches with shorter sequences.

Possibly a more interesting question[14] might have been: "Why are not all the hits which include both domains at the top of the list?". Surely they should be, as they match over a longer proportion of the query sequence and so must, in general at least, be of the greatest significance.

They do not always come at the top of the list because **blast** scores each matching region individually and uses the ranking scores associated with the single region with the highest **E Value** to evaluate the similarity of the entire database entry with the query. This has to be a dubious practice surely? But, it appears to work, so why complain.

To justify this last assertion, Look at your top hit.

Description	Max score	Total score	Query cover	E value	Ident	Accession
RecName: Full=Paired box protein Pax-6; AltName: Full=Aniridia type II protein; AltName: Full=Oculorhombin	160	767	3%	2e-40	97%	P26367.2

E Val = 2e-40, Max score = 160, Total score 767 associated with the whole of **P26367.2**

RecName: Full=Paired box protein Pax-6; AltName: Full=Aniridia type II protein; AltName: Full=Oculorhombin
Sequence ID: sp|P26367.2|PAX6_HUMAN Length: 422 Number of Matches: 8

Range 1: 46 to 123 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
160 bits(406)	2e-40	Compositional matrix adjust.	76/78(97%)	78/78(100%)	0/78(0%)	+3

```
Query  16680  MQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDR  16859
              +QVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDR
Sbjct  46     LQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDR  105

Query  16860  LLSEGVCTNDNIPSVSSL  16913
              LLSEGVCTNDNIPSVSS+
Sbjct  106    LLSEGVCTNDNIPSVSSI  123
```

Range 2: 254 to 305 GenPept Graphics ▼ Next Match ▲ Previous Match ▲ First Match

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

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

Range 3: 312 to 344 GenPept Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
70.5 bits(171)	5e-29	Compositional matrix adjust.	33/33(100%)	33/33(100%)	0/33(0%)	+2

```
Query  24926  GSMLGRTDTALTNTYSALPPMPSFTMANNLPMQ  25024
              GSMLGRTDTALTNTYSALPPMPSFTMANNLPMQ
Sbjct  312    GSMLGRTDTALTNTYSALPPMPSFTMANNLPMQ  344
```

Now look at the first few individual regional alignments for this hit.

As you can see, the **E Value** and **Max score** values used to evaluate the whole protein were computed from just the best (ranked by **E Value**) local alignment! Crude, but never mind.

The **Total score** for the entire protein is the sum (rounded up to the nearest integer) of all the bit scores for all **8** local alignments computed for this protein (I suggest you just trust me on this assertion).

14 That I did not ask, because I only just thought of it.

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

If **Paired boxes** had been more prolific, then the number of **Paired box** matches might have filled the **blast** hit list before the highest scoring **Homeo box** hit was registered.

If **Homeo boxes** were longer, and so justified a better **E value**, then the number of **Homeo box** matches might have filled the **blast** hit list before the highest scoring **Paired box** hit was registered.

Either of these situations would be very unfortunate, but easily avoided by setting the **Max matches in a query range** parameter to something sensible (**50** say). This would ensure that only the top **50** items in the **blast** hit list would be dominated by the strongest hit.

For further discussion of the parameter, see above.

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

blast identifies two non-informative regions. I only discussed the prettiest one above. The region discussed is comprised largely of **Serines**, **Prolines**, **Threonines** & **Isoleucines** the **15** residues between **294-308**.

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

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

The second (to be found much further down your **blast Alignments** output) is comprised entirely of **Arginines**, **Luecines** and **Lysines** and **Glutamines**, the **10** residues between **203 - 212**.

Score	Expect	Method	Identities	Positives	Gaps	Frame
85.9 bits(211)	3e-16	Compositional matrix adjust.	56/66(85%)	58/66(87%)	5/66(7%)	+3

```
Query  23649  YHPILFVP-----DGCQQQEGGGENTNSISSNGEDSDEAQMrlqlkrklqrNRTSFTQEQ  23813
              ++P   VP     DGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQ
Sbjct  162    WYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQ  221

Query  23814  IEALEK  23831
              IEALEK
Sbjct  222    IEALEK  227
```

Uniprotkb also suggests there are two **compositionally biased regions**.

Compositional bias	131 – 209	79	Gln/Gly-rich
Compositional bias	279 – 422	144	Pro/Ser/Thr-rich

Well, hardly an exact match, but there is approximate agreement? One would certainly suppose that **blast** is only willing to mask fairly severe cases of **compositional bias**. It is also probable that **blast** has a rather more mechanistic (i.e. non-biological) interpretation of what **computational bias** is?

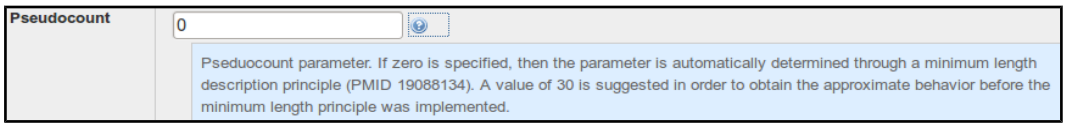
PROSITE also predicts the more obvious region of **computational bias**, rather more generally:

> "An octapeptide and/or a homeodomain can occur C-terminal to the paired domain, as well as a Pro-Ser-Thr-rich C-terminus"

PSI-Blast

What do you suppose the choice of **Pseudocount** might influence?

I clicked with confidences upon the link to the help. It opined as illustrated.



I suppose the next step is to read **PMID 19088134**? There is most certainly no elucidation amongst the strangle of words offered here?

The article **Abstract** says:

"**P**osition **s**pecific **s**core **m**atrices (**PSSM**s) are derived from multiple sequence alignments to aid in the recognition of distant protein sequence relationships. The **PSI-BLAST** protein database search program derives the column scores of its **PSSM**s with the aid of **pseudocounts**, added to the observed amino acid counts in a multiple alignment column. In the absence of theory, the number of **pseudocounts** used has been a completely empirical parameter. This article argues that the minimum description length principle can motivate the choice of this parameter. Specifically, for realistic alignments, the principle supports the practice of using a number of **pseudocounts** essentially independent of alignment size. However, it also implies that more highly conserved columns should use fewer **pseudocounts**, increasing the inter-column contrast of the implied **PSSM**s. A new method for calculating **pseudocounts** that significantly improves **PSI-BLAST**'s; retrieval accuracy is now employed by default."

The article itself, continues in like vein ….. how about we close our eyes and accept the defaults? I would just wonder why the whole thing does not commence with, at least an attempt, to answer the question in the forefront of my inquiry, which is .. "**WHAT, in the current context, IS a pseudocount?**". I do not believe it is as tricky as they appear to wish us to believe. I will try again later, when my view of the world is less storm infested.