ELB17F

Entry Level Bioinformatics

08-12 May 2017

(First 2017 run of this Course)

Basic Bioinformatics Sessions

Practical 6: Multiple Sequence Alignment

Monday 1 May 2017

Multiple Sequence Alignment

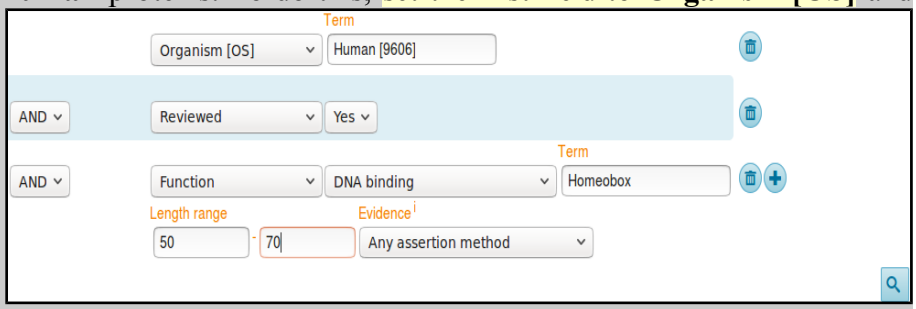
Here we will look at some software tools to align some protein sequences. Before we can do that, we need some sequences to align. I propose we try all the human **homeobox** domains from the well annotated section of **UniprotKB**. Getting the sequences is a trifle clumsy, so concentrate now! There used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

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

http://www.uniprot.org/

Choose the Advanced option of the Search button.

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



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

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

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

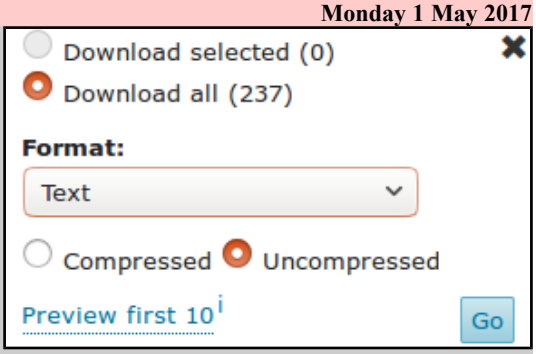
Leave the **Evidence** box as **Any assertion method**, one does not wish to be too fussy! Address the button with authority to get the search going.

BLAST | Align | Download | Add to basket | Columns | 1 to 25 of 237 | Show 25

Entry	Entry name	Protein names	Gene names	Organism	Length
P52952	NKX25_HUMAN	**Homeobox protein Nkx-2.5**	**NKX2-5** CSX, NKX2.5, NKX2E	Homo sapiens (Human)	324
P49639	HXA1_HUMAN	**Homeobox protein Hox-A1**	**HOXA1** HOX1F	Homo sapiens (Human)	335
P26367	PAX6_HUMAN	**Paired box protein Pax-6**	**PAX6** AN2	Homo sapiens (Human)	422
Q99697	PITX2_HUMAN	**Pituitary homeobox 2**	**PITX2** ARP1, RGS, RIEG, RIEG1	Homo sapiens (Human)	317
Q99801	NKX31_HUMAN	**Homeobox protein Nkx-3.1**	**NKX3-1** NKX3.1, NKX3A	Homo sapiens (Human)	234
Q01860	PO5F1_HUMAN	**POU domain, class 5, transcription ...**	**POU5F1** OCT3, OCT4, OTF3	Homo sapiens (Human)	360
Q01826	SATB1_HUMAN	**DNA-binding protein SATB1**	**SATB1**	Homo sapiens (Human)	763
Q15475	SIX1_HUMAN	**Homeobox protein SIX1**	**SIX1**	Homo sapiens (Human)	284
P43699	NKX21_HUMAN	**Homeobox protein Nkx-2.1**	**NKX2-1** NKX2A, TITF1, TTF1	Homo sapiens (Human)	371
P23760	PAX3_HUMAN	**Paired box protein Pax-3**	**PAX3** HUP2	Homo sapiens (Human)	479

A fine miscellany of sequences will assemble upon you screen. Most seem to declare themselves in possession of a **Homeobox** or two (including **PAX6_HUMAN**), so I suggest a declaration of success.

Now save the entire list into a file using the Download button. Set the download to **uncompressed**. Make sure you have **all** sequences selected and that **Text** (i.e. **EMBL** or **SwissProt**) format selected. Press the **Go** button and do whatever it takes to ensure your results end up in a file residing on your **Desktop** called:



human_homeobox_proteins.emb

```
ID   NKX25_HUMAN             Reviewed;         324 AA.
AC   P52952; A8K3K0; B4DNB6; E9PBU6;
DT   01-OCT-1996, integrated into UniProtKB/Swiss-Prot.
DT   01-OCT-1996, sequence version 1.
DT   30-NOV-2016, entry version 177.
DE   RecName: Full=Homeobox protein Nkx-2.5;
DE   AltName: Full=Cardiac-specific homeobox;
DE   AltName: Full=Homeobox protein CSX;
DE   AltName: Full=Homeobox protein NK-2 homolog E;
GN   Name=NKX2-5; Synonyms=CSX, NKX2.5, NKX2E;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC   Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 1).
RC   TISSUE=Heart;
RX   PubMed=8900537;
RA   Turbay D., Wechsler S.B., Blanchard K.M., Izumo S.;
RT   "Molecular cloning, chromosomal mapping, and characterization of the
RT   human cardiac-specific homeobox gene hCsx.";
RL   Mol. Med. 2:86-96(1996).
```

Take a swift look at the file you have just created. Your neat list of **Human Homeobox** sequences will have transformed into a flood of **many SwissProt** format **UniProtKB** entries. Ugly, but what is required.

Search (**Control F**) for the term **DNA_BIND**.

It should occur many times (at least once per sequence) in the **F**eature **T**ables and most often refer to a **Homeobox** region.

In the **DNA_BIND** **F**eature **T**able entries, the position of the **Homeobox**s are recorded and will be used by the next program to isolate the sequence of the **Homeobox**s.

```
FT   CHAIN           1    374       Pre-B-cell leukemia transcription factor
FT                                  4.
FT                                  /FTId=PRO_0000049241.
FT   DNA_BIND      210    272       Homeobox; TALE-type.
FT                                  {ECO:0000255|PROSITE-ProRule:PRU00108}.
FT   VARIANT       169    169       V -> I (in dbSNP:rs8108180).
FT                                  /FTId=VAR_059355.
FT   VARIANT       177    177       M -> V (in dbSNP:rs8108981).
FT                                  /FTId=VAR_059356.
FT   VARIANT       283    283       T -> M (in a colorectal cancer sample;
FT                                  somatic mutation; dbSNP:rs376647012).
FT                                  {ECO:0000269|PubMed:16959974}.
FT                                  /FTId=VAR_036439.
FT   CONFLICT      368    368       I -> T (in Ref. 1; BAG53471).
FT                                  {ECO:0000305}.
SQ   SEQUENCE   374 AA;  40854 MW;  B9CE8BE93D0B7ABC CRC64;
     MAAPPRPAPS PPAPRRLDTS DVLQQIMAIT DQSLDEAQAR KHALNCHRMK PALFSVLCEI
     KEKTVVSIRG IQDEDPPDAQ LLRLDNMLLA EGVCRPEKRG RGGAVARAGT ATPGGCPNDN
     SIEHSDYRAK LSQIRQIYHS ELEKYEQACR EFTTHVTNLL QEQSRMRPVS PKEIERMVGA
     IHGKFSAIQM QLKQSTCEAV MTLRSRLLDA RRKRRNFSKQ ATEVLNEYFY SHLNNPYPSE
     EAKEELARKG GLTISQVSNW FGNKRIRYKK NMGKFQEEAT IYTGKTAVDT TEVGVPGNHA
     SCLSTPSSGS SGPFPLPSAG DAFLTLRTLA SLQPPPGGGC LQSQAQGSWQ GATPQPATAS
     PAGDPGSINS STSN
//
```

Now to extract from the whole protein sequences you have saved in a file, the sequences of just the **Homeobox** domains. One way of doing this (possibly not the best), is to use an **EMBOSS** package program called **extractfeat**. This can be found in many places, including the Bioinformatics server at **Wageningen** in the Netherlands. Go to:

http://emboss.bioinformatics.nl/

EDIT
aligncopy
aligncopypair
biosed
codcopy
cutseq
degapseq
descseq
entret
extractalign
extractfeat

Find the program **extractfeat** (in the **EDIT** section), and set it going.

Use the **Choose File** button to **upload** the **SwissProt** format sequences from **UniProtKB** that you saved in **the file**:

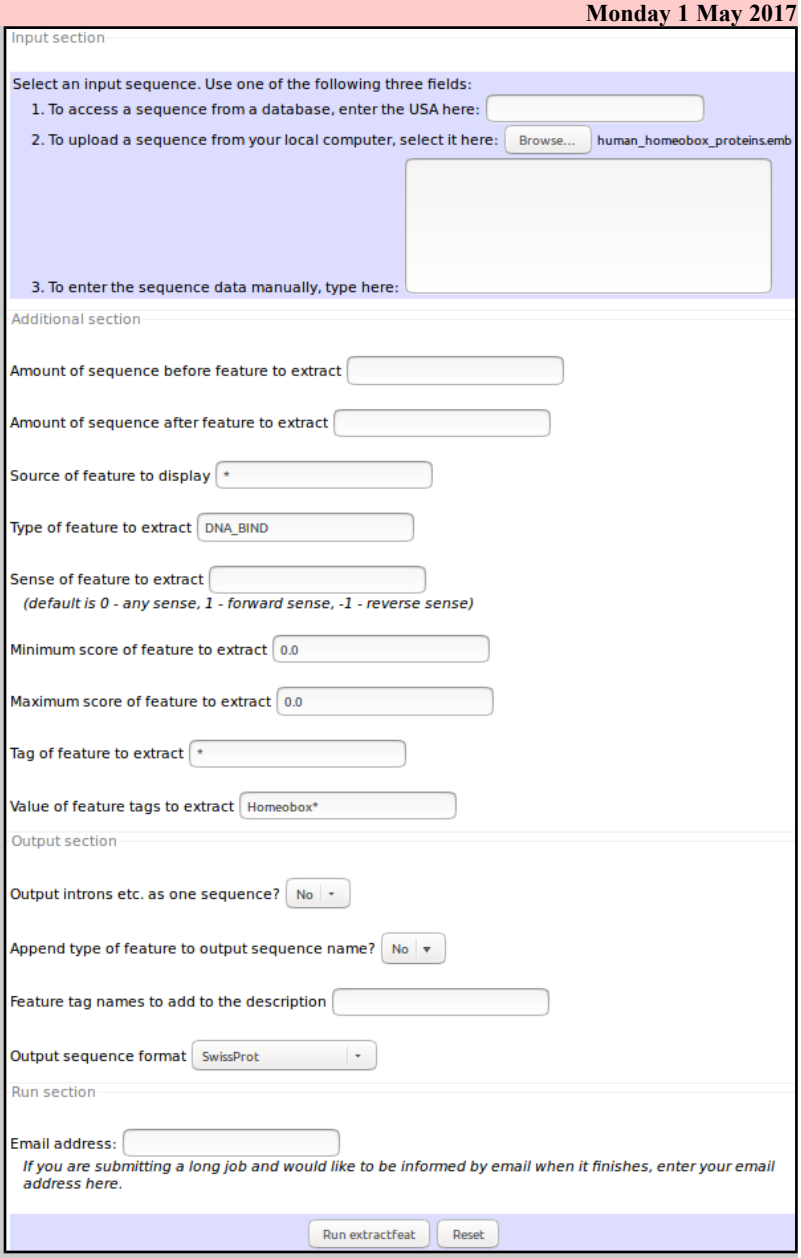
`human_homeobox_proteins.emb`.

Set **Type of feature to extract field** to **DNA_BIND** (Make sure you remove the "*").

Set **Value of feature tags to extract** to **Homeobox*** (Make sure you append the "*" to ensure hits with, for example "**homeoboxes**").

Set the **Output sequence format** to **SwissProt** (**Fasta** would do, but **SwissProt** retains more annotation).

Click on the Run extractfeat button to start **extractfeat** going. Many sequences of **60** amino acids (or so) in length will leap into view.



OUTPUT FILE outseq

```
ID   NKX25_HUMAN_138_197     Reviewed;          60 AA.
DE   [DNA_contact] Homeobox protein Nkx-2.5 (Cardiac-specific homeobox) (Homeobox protein CSX) (Homeobox protein NK-2 homolog E)
SQ   SEQUENCE   60 AA;  7514 MW;  16EE564D071E5E8A CRC64;
     RRKPRVLFSQ AQVYELERRF KQQRYLSAPE RDQLASVLKL TSTQVKIWFQ NRRYKCKRQR
//
ID   HXA1_HUMAN_229_288      Reviewed;          60 AA.
DE   [DNA_contact] Homeobox protein Hox-A1 (Homeobox protein Hox-1F)
SQ   SEQUENCE   60 AA;  7365 MW;  53E2BC59B06F544E CRC64;
     PNAVRTNFTT KQLTELEKEF HFNKYLTRAR RVEIAASLQL NETQVKIWFQ NRRMKQKKRE
//
ID   PAX6_HUMAN_210_269      Reviewed;          60 AA.
DE   [DNA_contact] Paired box protein Pax-6 (Aniridia type II protein) (Oculorhombin)
SQ   SEQUENCE   60 AA;  7447 MW;  075C194DB9F33ED9 CRC64;
     LQRNRTSFTQ EQIEALEKEF ERTHYPDVFA RERLAAKIDL PEARIQVWFS NRRAKWRREE
//
ID   PITX2_HUMAN_85_144      Reviewed;          60 AA.
DE   [DNA_contact] Pituitary homeobox 2 (ALL1-responsive protein ARP1) (Homeobox protein PITX2) (Paired-like homeodomain transcription factor 2) (RIEG bicoid-related homeobox transcription factor) (Solurshin)
SQ   SEQUENCE   60 AA;  7622 MW;  49CF61CFC17E1E0E CRC64;
     QRRQRTHFTS QQLQELEATF QRNRYPDMST REEIAVWTNL TEARVRVWFK NRRAKWRKRE
//
ID   NKX31_HUMAN_124_183     Reviewed;          60 AA.
DE   [DNA_contact] Homeobox protein Nkx-3.1 (Homeobox protein NK-3 homolog A)
SQ   SEQUENCE   60 AA;  7339 MW;  F665B481E2E574BB CRC64;
     QKRSRAAFSH TQVIELERKF SHQKYLSAPE RAHLAKNLKL TETQVKIWFQ NRRYKTKRKQ
//
```

Right click the outseq button and select **Save Link as...** . Do whatever it takes to save all your **Homeobox** domains into a file residing on your **Desktop** called:

`homeobox_human.emb`

Finally, we have some sequences with which to investigate the multiple sequence alignment programs.

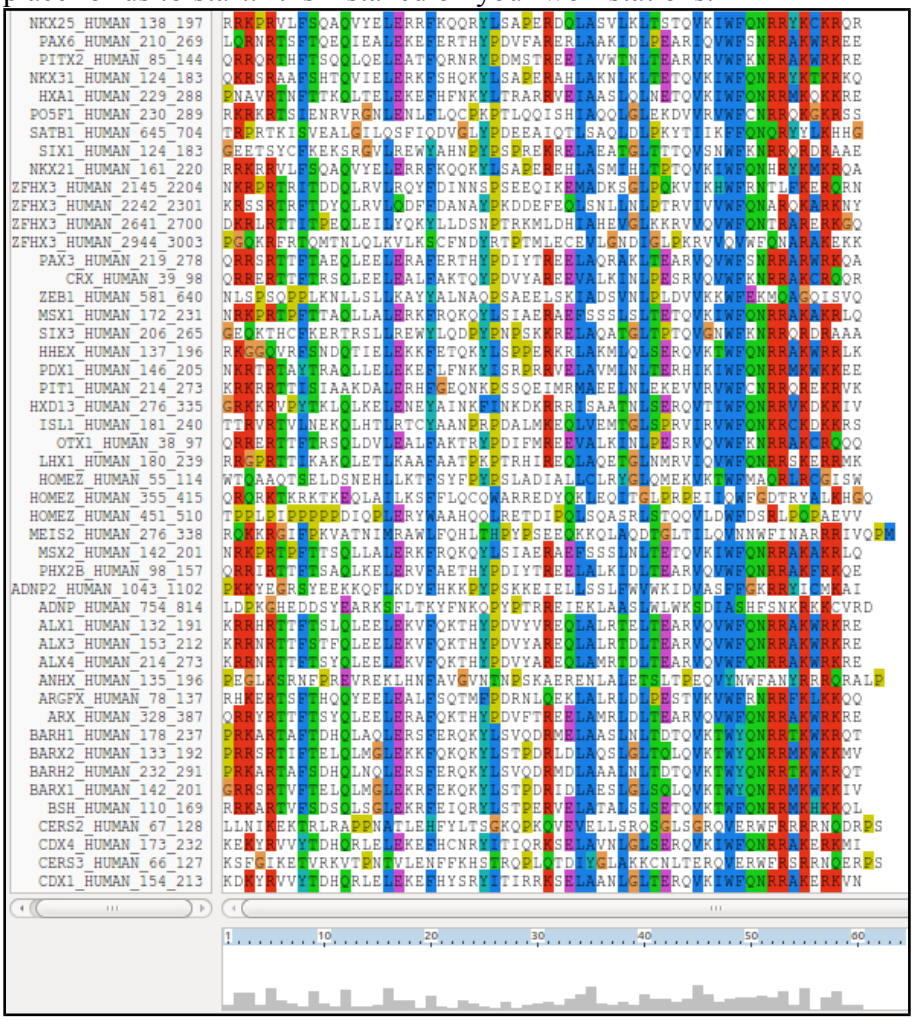
Take a look at the file you have created. You should have many human **homeobox** domains in **SwissProt** format, looking rather as they did in your browser window. Happily **ClustalX**, the first multiple alignment program to be investigated, accepts multiple sequence **SwissProt** format files as input.

ClustalX is a part of the mostly widely known family of **M**ultiple **S**equence **A**lignments (**MSA**) programs, originating in the **1980**s. Until relatively recently, it was the only real option. **ClustalX** still has merit, although it lacks some of the sophistication of more recent programs. **ClustalX** runs on effectively all workstations and has a nice **G**raphical **U**ser **I**nterface (**GUI**). A good place for us to start. It is installed on your workstations.

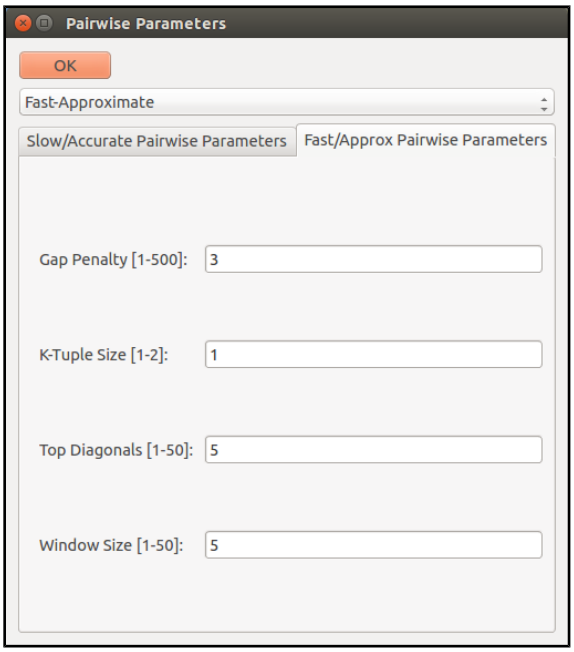
Start up the program **ClustalX**[1]. The **ClustalX G**raphical **U**ser **I**nterface (**GUI**) will regally mount your screen.

Select **Load Sequences** from the **File** pull down menu and load your file of **homeobox** domains (`homeobox_human.emb`)**.**

The sequences will arrange themselves colourfully. Many of the **homeoboxes** are similar enough to look convincing even before alignment. Note the "Manhattan skyline" under the sequences indicating the varying degrees of conservation.



Font: 10 You might like to increase the **Font** size from the minute default setting designed for Hawks and Eagles, to something more comfortable. **24** works tolerably well for me.



From the **Alignment** pull down menu, go to the **Alignment parameters** menu and select **Pairwise Alignment Parameters**. Just for a moment, change the setting from **Slow-Accurate** to **Fast-Approximate**. Bring the corresponding parameters into view by clicking on **Fast/Approx Pairwise Parameters** tab[2].

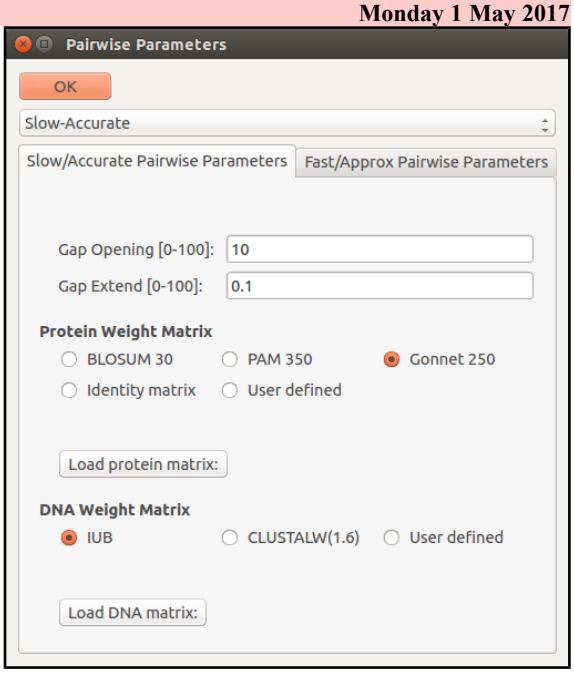
Hopefully, we will have discussed the way **ClustalX** (and similar multiple alignment tools) work. Intuitively, it should not make a lot of difference how the initial pairwise comparison stage is conducted. However, it very often does.

Specifically for this set of proteins, as well as generally, **ClustalX** will give a noticeably better alignment if the initial pairwise alignment stage is done carefully. Accordingly, reverse your whimsical setting change by moving back from **Fast-Approximate** to **Slow-Accurate**.

1 Of course, you could run **Clustal** from websites all over the world if you wished. Specifically, it is available at the Bioinformatics server at **Wageningen**. Try it if you have time. You get the same results but will, sadly, lose the pretty interface.
http://www.bioinformatics.nl/tools/clustalw.html
The **EBI** no longer offer basic **Clustal**.

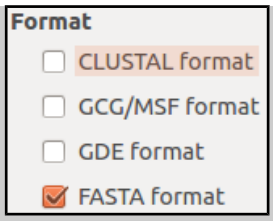
2 The **Fast-Approximate** algorithm is essential that which the database searching program **fasta** employs. Assuming we have discussed how **fasta** (or **blast**) works, it should require little further explanation here.

Click on the **Slow/Accurate Pairwise Parameters** tab for a final look at the default parameters to be used. The **Slow-Accurate** option is essentially a version of **Global Alignment** algorithm we will have discussed previously. Hopefully, all the parameter options will therefore be familiar to you.



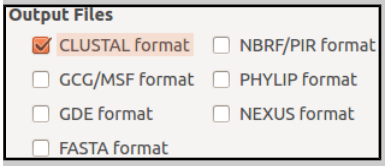
I will assume both sets of parameters at least seem familiar? If not please ask. The default **Slow/Accurate Pairwise Parameters** you now have in view are fine. Click the **OK** button to dismiss the **Pairwise Parameters** window.

Before proceeding, save the **homeobox** sequences in **FASTA** format, which will better suit the other **MSA** programs we will try. Do this by selecting **Save sequences as...** from the **File** pull down menu. Deselect **CLUSTAL format**, select **FASTA format**.



Change the default file output file name to `homeobox_human_full`

Click **OK**. A file called `homeobox_human_full.fasta` will be created. Take a look to check it is as you would expect.



Strangely, saving your sequences in **FASTA** format convinces **clustalx** that it should now output its alignments in **FASTA** format. To prevent this, select **Output Format Options** from the **Alignments** pull down menu. Deselect **FASTA format** and select **CLUSTAL format**. Click **OK**.

From the **Alignment** pull down menu, select **Do Complete Alignment**. Accept the default names for output files and click on the **OK** button. **ClustalX** will start to think deeply and eventually come up with it view of how the **homeobox** domains should be aligned.

Note the display at the bottom of the **ClustalX** window in which the preliminary pairwise comparisons of all sequences is monitored. The scores from these comparisons are used to compute the **Guide Tree**.

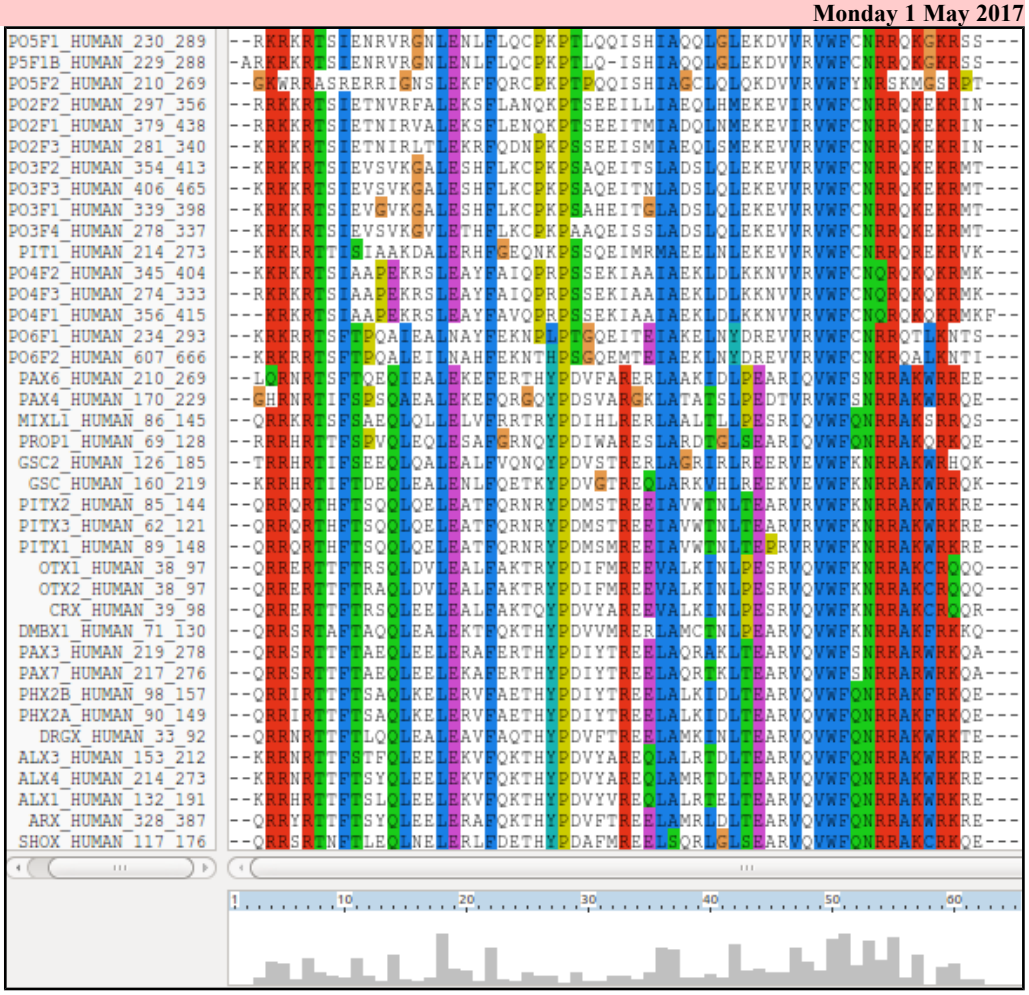
Not a bad first try. From an entirely non scientific, cosmetic, viewpoint, the ragged ends offend a trifle, as does the gap just before position **30**!

In reality, these features might be interesting, but here I go for pretty!

So, just to investigate what is possible, select all the **homeobox** sequences that are causing the gap around position **30** by clicking on their names (quite a lot of them I fear). Hold the **Ctrl** key down to allow multiple selection.

All selected, go to the **Edit** pull down menu and select **Cut Sequences**. Then select **Remove Gap-Only columns** from the **Edit** pull down menu. Nasty gap gone … along with all scientific credibility, but … never mind.

You could recompute the alignment from scratch for the reduced sequence set ending up with the same answer. Just for the sake of it, select **Select All Sequences** from the **Edit** pull down menu. Then select **Remove All gaps** from the **Edit** menu and confirm your intentions. You are now back where you started, but without the sequences that mess up the alignment intolerably!



Save your filtered set of sequences. From the **File** menu select **Save Sequences as...** . Choose **FASTA** format only. This time, create a file with the default name:

`homeobox_human.fasta`

The full original set of sequences was saved in a differently named file, as a precaution. I am convinced the sequences eliminated would not align convincingly with any of the tools we have at hand. Let us lose them! Press the **OK** button.

From the **Alignment** menu, select **Output Format Options** and then select **CLUSTAL format** only.

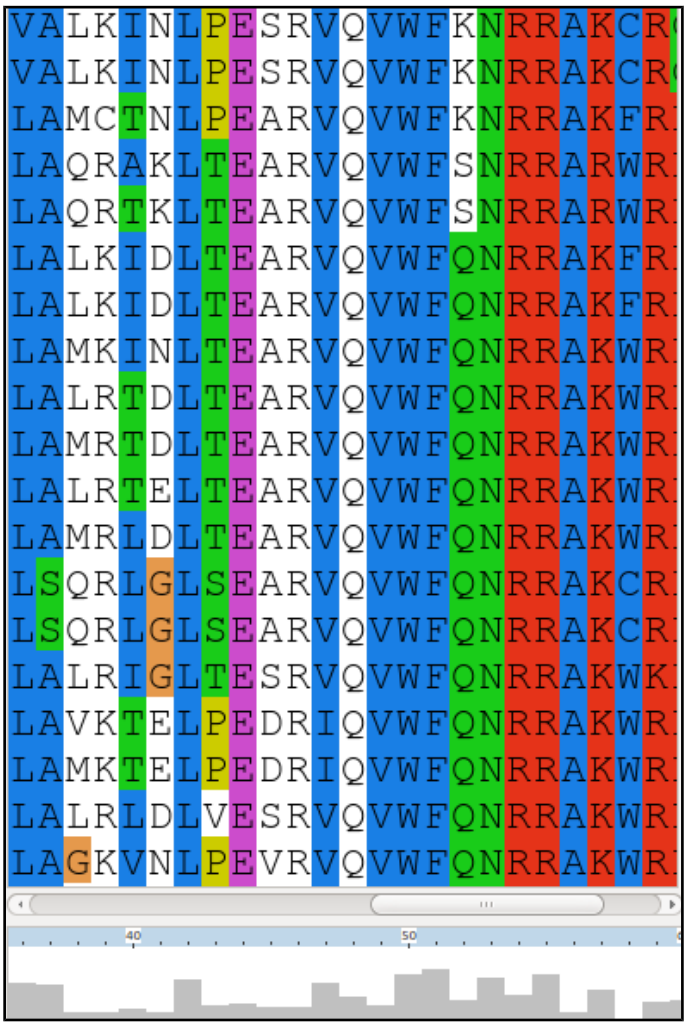
From the **Alignment** menu, select **Do Complete Alignment**. Accept the default names for the output files. This will overwrite your previous efforts, but no matter. More deep thought. Well, I got back to where I was, no gaps around position **30** but still with ragged ends!

It is difficult to prove you have exactly the same alignment as previously as the order of the **MSA** will be different. This order being determined by the pairwise comparison stage of the **ClustalX MSA** computation.

The **Prosite** motif database uses Patterns to represents protein features (in addition to **HMM**s). The pattern for a **homeobox** is the ever memorable:

[LIVMFYG]-[ASLVR]-x(2)-[LIVMSTACN]-x-[LIVM]-{Y}-x(2)-{L}-[LIV]-[RKNQESTAIY]-[LIVFSTNKH]-W-[FYVC]-x-[NDQTAH]-x(5)-[RKNAIMW]

Any speculations as to how this might be interpreted? Hint?

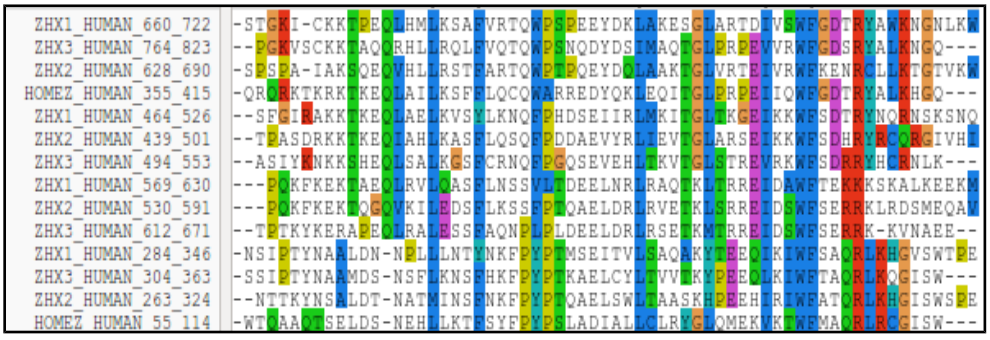


This pattern corresponds to positions **36** to **59** in my alignment. See that the "Manhattan Skyline" is encouraging in the parts of this region that matter.

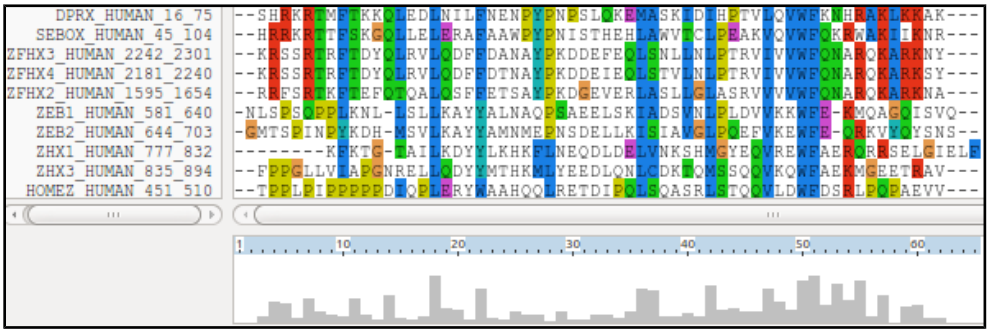
Note that the profile **Tryptophan**, in position **50**, is **very** consistent, but not quite **100%** as suggested by the **Prosite** pattern[3]. The **W** was even conserved in the sequences that were cosmetically removed.

Position **52** is not conserved ("**-x-**") according to the **Prosite** pattern. In the alignment segment offered here, it looks like a pretty consistent **Q**. However, the "**Manhattan skyline**" at this position is quite low, suggesting that the sequences in view might not be typical of the whole alignment set. Which, upon checking …. they are not!

Looking through this alignment, I get the feeling I could design a better, stricter pattern for the region between **36** and **59**. Possibly true, but remember the pattern in **Prosite** aims to represent the conservation of **Homeobox** domains in **ALL** organisms. Here we have only sequences from **Human**.



Of course, things are not quite so convincing throughout. If you look at the top and bottom few sequences, you will see that **ClustalX** had its moments of uncertainty.



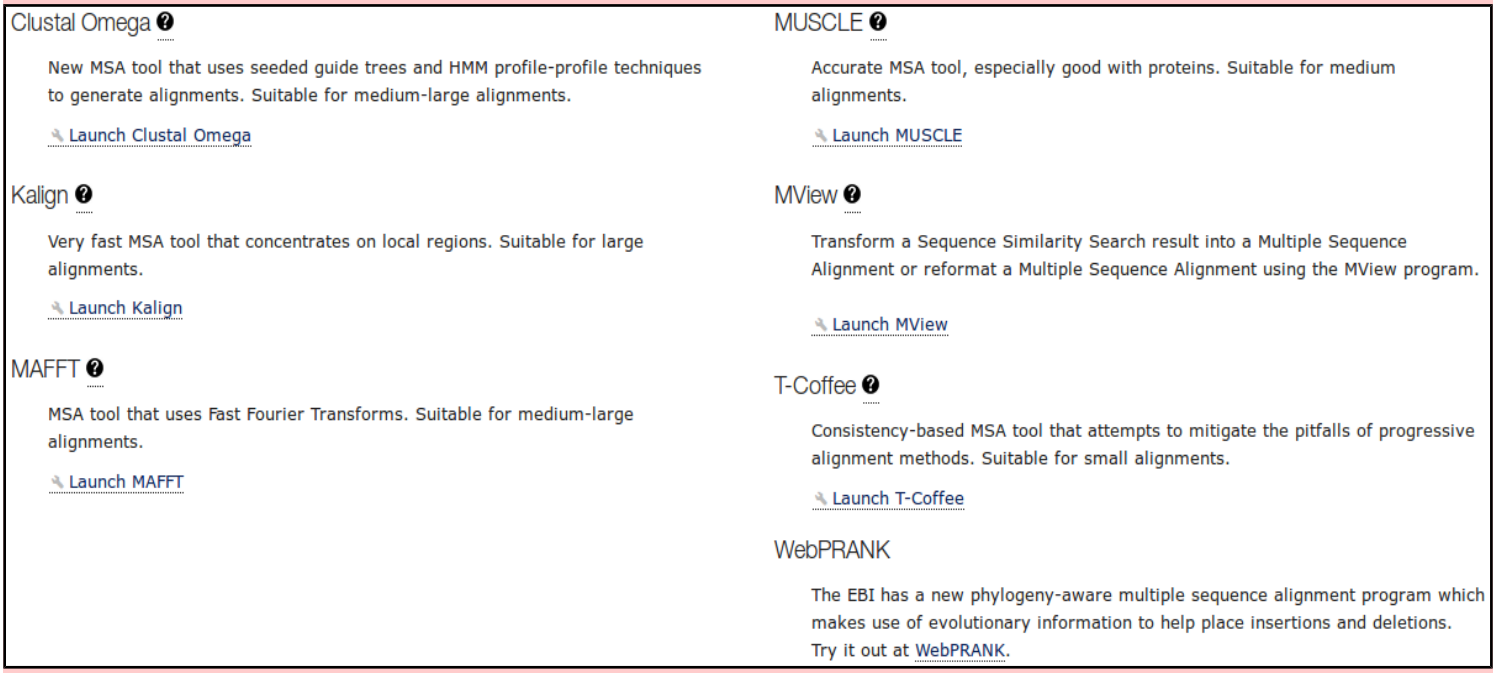
Note, however, the consistent **W** in position **50** despite the surrounding crumble.

3 From the "**Manhattan Skyline**", you can see the conservation is less than **100%**. Less conserved than the **F** that immediately follows in fact? Look at your alignment, the "**Manhattan Skyline**" does not seem to reflect reality? The **W** is **very** well conserved, although the scoring matrices would regard any deviation from **W** as serious? I need to find out more about how the **Skyline** is computed.

Now to show existence of some **msa** program options available on the web. There are many. They are available from a number of server sites. An obvious place to start has to be the **EBI** page dedicated to **MSA**. Go to:

`http://www.ebi.ac.uk/Tools/msa/`

Offered here is a selection of popular, current generation **MSA** tools. Each is accompanied by advice to guide the choice of tool to best fit the circumstances. Each tool is provided with a link to its **Launch** interface. All the **Launch** interfaces are very consistent. Once you have run one of the **MSA** options, you should have no trouble running any of the others.

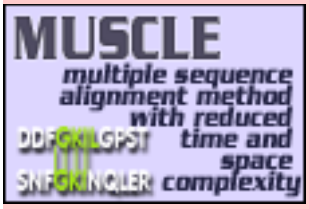


Here I intend to align again the human **homeboxes** with just one of the tools on offer. Then take a quick look at how the machine generated multiple alignment can be manually edited using **Jalview**, a program that is probably installed on your workstation and definitely available as a web service. You might have already used Jalview as an alignment viewer when investigating **Pfam** and/or **Jpred**.

Then I will invite you to try a few of the other options for yourself and see that they do not all produce the same alignment! Differences reflect not only the parameters selected, which we will have discussed, but also the particular objectives of the program selected. For example, a multiple protein sequence alignment optimal for investigating conservation of protein structure might well not be identical to one best representing protein evolution.

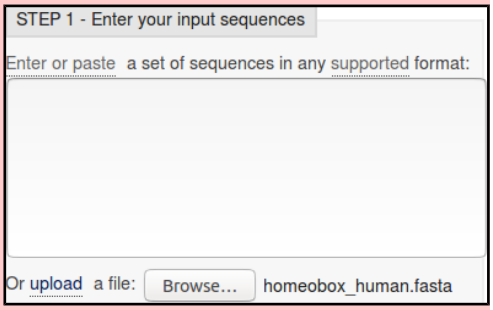
Used to align the **Homeobox** sequences used in this exercise, I do not expect you will see much difference between the outputs of any of these options. They will all work sufficiently on such a simple data set.

The program whose use I choose to describe carefully, leading on to a short **Jalview** exercise is **MUSCLE**. I choose thus as **MUSCLE** is now the first choice of most of the people with whom I work. Also popular are **Clustal Omega**, **MAFFT** and, for **phylogeny**, **WebPRANK**.

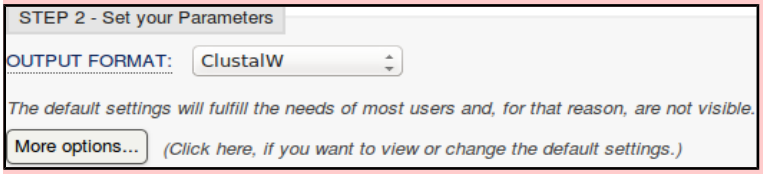


So the plan now is to use **MUSCLE**[4] to align again the **homeobox** sequences previously aligned with **ClustalX**. **MUSCLE** works in a way similar to **clustalX** but it takes rather more care in the generation of the **Guide Tree** used to control the order of pairwise construction of the final multiple alignment[5]. Particularly for more difficult alignments, **MUSCLE** should do a better job than **ClustalX**. The alignment you will generate here will certainly be different. I leave you to judge for yourselves whether it is better.

Start by requesting to Launch MUSCLE.

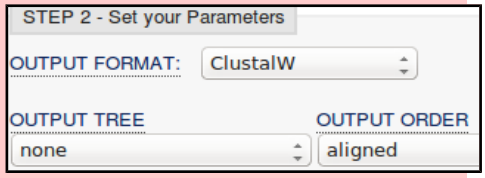


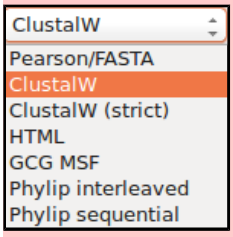
Use the Browse... button to upload the file containing the **FASTA** format **homeobox** sequences, **homeobox_human.fasta**. This file should not included the sequences with a mess around position **30**.



Take a look at the **Set your Parameters** section of the page. I find the claim that "*The default settings will fulfill the needs of most users and, for that reason, are not visible*" a little strange? What about the users who are not in the category "*most*"? I want control over all the programs that their creators deemed sensible to make available[6]?

The default settings behind the More options... button are not those that affect the computation of the **MSA**. I confess myself confused at the lack of any meaningful options to consider? I was expecting at least the **gap open** and **gap extension penalty** options (available elsewhere, including **Wageningen**), plus a way to change the **scoring matrix**. I have inquired why things are as they are (most recently **2016.04.17**). No practical issue here, as I intended to suggest the defaults whatever they were. Look at the range of settings for the **OUTPUT TREE** parameter. **none** is indeed the thinking persons choice, but … one or the other (but not both?) of the **Guide Trees** that **MUSCLE** will compute can be saved if you wish[7]. You may also set the **OUTPUT ORDER** to **aligned** or … **aligned**?





There are a number of **OUTPUT FORMATS** offered. For a quick glance at your results, both **ClustalW** or **HTML** are fine. Here I suggest it would be nice to generate an output that can be downloaded and viewed in **Jalview**[8]. The default **ClustalW** or **Pearson/FASTA** serve for this purpose. As **ClustalW** looks more like an alignment in the web page, I choose **ClustalW**[9].

How do the options for the **OUTPUT TREE** relate to the output files of **ClustalX** and the difference between the way that **ClustalX** and **muscle** work?_____

Comment on how one might choose between the range of options offered for the aligned parameter?_____

4 More available from a variety of websites in addition to the **EBI**, including the Bioinformatics server at **Wageningen**:
`http://www.bioinformatics.nl/tools/muscle.html`

5 As discussed, superficially at least, previously. I hope.

6 I have asked the **EBI** about their policy (the same for all the locally provided **MSA** options). Discussion is ongoing (**2016.04.20**).

7 A useful option if you thought it possible you might want to rerun **MUSCLE** with different parameter setting for the stages after the **Guide Tree(s)** are generated. The same possibilities exist for **ClustalX**. Of course, utterly pointless if it is impossible to control the relevant parameters …. so I really cannot see the point of any of the **More options** section? I am open to elucidation from all/any sources.

8 A widely used **java** alignment editor and viewer.

9 But feel free to try the others. **HTML** is the default at **Wageningen**. The **Phylip** formats are the best if you are going to analyse your output further with the phylogeny programs of the **PHYLIP** package.

After considering these enigmas, or before if you prefer, Click on the Submit button and sit back to admire **muscle** in action.

The alignment that is computed is, superficially at least, similar to that offered by **ClustalX**.

The alignment is irritatingly split into two sections. A nice extra parameter might have been "How wide would you like your alignment to be"? A problem with the format rather than the program, to be fair.

```
ARX_HUMAN_328_387       --QRRYR-TTFTSYQLEELERAFQKTHYPDVFTREELAMRLDLTEARVQVWFQNRRAKWR
ALX1_HUMAN_132_191      --KRRHR-TTFTSLQLEELEKVFQKTHYPDVYVREQLALRTELTEARVQVWFQNRRAKWR
ALX3_HUMAN_153_212      --KRRNR-TTFSTFQLEELEKVFQKTHYPDVYAREQLALRTDLTEARVQVWFQNRRAKWR
ALX4_HUMAN_214_273      --KRRNR-TTFTSYQLEELEKVFQKTHYPDVYAREQLAMRTDLTEARVQVWFQNRRAKWR
ISL1_HUMAN_181_240      --TTRVR-TVLNEKQLHTLRTCYAANPRPDALMKEQLVEMTGLSPRVIRVWFQNKRCKDK
ISL2_HUMAN_191_250      --TTRVR-TVLNEKQLHTLRTCYAANPRPDALMKEQLVEMTGLSPRVIRVWFQNKRCKDK
LHX9_HUMAN_267_326      --TKRMR-TSFKHHQLRTMKSYFAINHNPDAKDLKQLAQKTGLTKRVLQVWFQNARAKFR
LHX2_HUMAN_266_325      --TKRMR-TSFKHHQLRTMKSYFAINHNPDAKDLKQLAQKTGLTKRVLQVWFQNARAKFR
LHX6_HUMAN_219_278      --AKRAR-TSFTAEQLQVMQAQFAQDNNPDAQTLQKLADMTGLSRRVIQVWFQNCRARHK
LHX8_HUMAN_225_284      --AKRAR-TSFTADQLQVMQAQFAQDNNPDAQTLQKLAERTGLSRRVIQVWFQNCRARHK
ZFHX3_HUMAN_2641_2700   --DKRLR-TTITPEQLEILYQKYLLDSNPTRKMLDHIAHEVGLKKRVVQVWFQNTRARER
ZFHX4_HUMAN_2560_2619   --DKRLR-TTITPEQLEILYEKYLLDSNPTRKMLDHIAREVGLKKRVVQVWFQNTRARER
ZFHX2_HUMAN_1857_1916   --DKRLR-TTILPEQLEILYRWYMQDSNPTRKMLDCISEEVGLKKRVVQVWFQNTRARER
ZFHX2_HUMAN_2065_2124   --QRRYR-TQMSSLQLKIMKACYEAYRTPTMQECEVLGEEIGLPKRVIQVWFQNARAKEK
ZFHX3_HUMAN_2944_3003   PGQKRFR-TQMTNLQLKVLKSCFNDYRTPTMLECEVLGNDIGLPKRVVQVWFQNARAKEK
ZFHX4_HUMAN_2884_2943   --HKRFR-TQMSNLQLKVLKACFSDYRTPTMQECEMLGNEIGLPKRVVQVWFQNARAKEK
LMX1A_HUMAN_195_254     --PKRPR-TILTTQQRRAFKASFEVSSKPCRKVRETLAAETGLSVRVVQVWFQNQRAKMK
LMX1B_HUMAN_219_278     --PKRPR-TILTTQQRRAFKASFEVSSKPCRKVRETLAAETGLSVRVVQVWFQNQRAKMK
LHX1_HUMAN_180_239      --RRGPR-TTIKAKQLETLKAAFAATPKPTRHIREQLAQETGLNMRVIQVWFQNRRSKER
LHX5_HUMAN_180_239      --RRGPR-TTIKAKQLETLKAAFAATPKPTRHIREQLAQETGLNMRVIQVWFQNRRSKER
LHX4_HUMAN_157_216      --AKRPR-TTITAKQLETLKNAYKNSPKPARHVREQLSSETGLDMRVVQVWFQNRRAKEK
LHX3_HUMAN_157_216      --AKRPR-TTITAKQLETLKSAYNTSPKPARHVREQLSSETGLDMRVVQVWFQNRRAKEK
                                :  :          .       :  :

HOMEZ_HUMAN_451_510     EVV---
ZHX1_HUMAN_777_832      LGIELF
ZHX3_HUMAN_835_894      RAV---
HOMEZ_HUMAN_55_114      ISW---
ZHX2_HUMAN_263_324      ISWSPE
ZHX3_HUMAN_304_363      ISW---
ZHX1_HUMAN_284_346      VSWTPE
ZEB2_HUMAN_644_703      SNS---
ZEB1_HUMAN_581_640      SVQ---
ZHX1_HUMAN_569_630      LKEEKM
ZHX2_HUMAN_530_591      SMEQAV
ZHX3_HUMAN_612_671      AEE---
ZHX2_HUMAN_439_501      RGIVHI
ZHX3_HUMAN_494_553      NLK---
ZHX1_HUMAN_464_526      NSKSNQ
HOMEZ_HUMAN_355_415     HGQ---
ZHX2_HUMAN_628_690      TGTVKW
```

At the very bottom of the page, **muscle** whines:

PLEASE NOTE: Showing colors on large alignments is slow.

So click the **Show Colors** button at the top of the page and try to live with the pain of such gross Trans-Atlantic inept spelling in a European site!!! Good Grief! They get everywhere!!

Well, an improvement I suppose? Colours are very useful (even slow ones) in the interpretation of alignments. Various colour schemes are used to clarify the message of alignments. Colouring can indicate shared amino acid properties not immediately evident when the letter representations differ.

```
ARX_HUMAN_328_387       --QRRYR-TTFTSYQLEELERAFQKTHYPDVFTREELAMRLDLTEARVQVWFQNRRAKWR
ALX1_HUMAN_132_191      --KRRHR-TTFTSLQLEELEKVFQKTHYPDVYVREQLALRTELTEARVQVWFQNRRAKWR
ALX3_HUMAN_153_212      --KRRNR-TTFSTFQLEELEKVFQKTHYPDVYAREQLALRTDLTEARVQVWFQNRRAKWR
ALX4_HUMAN_214_273      --KRRNR-TTFTSYQLEELEKVFQKTHYPDVYAREQLAMRTDLTEARVQVWFQNRRAKWR
ISL1_HUMAN_181_240      --TTRVR-TVLNEKQLHTLRTCYAANPRPDALMKEQLVEMTGLSPRVIRVWFQNKRCKDK
ISL2_HUMAN_191_250      --TTRVR-TVLNEKQLHTLRTCYAANPRPDALMKEQLVEMTGLSPRVIRVWFQNKRCKDK
LHX9_HUMAN_267_326      --TKRMR-TSFKHHQLRTMKSYFAINHNPDAKDLKQLAQKTGLTKRVLQVWFQNARAKFR
LHX2_HUMAN_266_325      --TKRMR-TSFKHHQLRTMKSYFAINHNPDAKDLKQLAQKTGLTKRVLQVWFQNARAKFR
LHX6_HUMAN_219_278      --AKRAR-TSFTAEQLQVMQAQFAQDNNPDAQTLQKLADMTGLSRRVIQVWFQNCRARHK
LHX8_HUMAN_225_284      --AKRAR-TSFTADQLQVMQAQFAQDNNPDAQTLQKLAERTGLSRRVIQVWFQNCRARHK
ZFHX3_HUMAN_2641_2700   --DKRLR-TTITPEQLEILYQKYLLDSNPTRKMLDHIAHEVGLKKRVVQVWFQNTRARER
ZFHX4_HUMAN_2560_2619   --DKRLR-TTITPEQLEILYEKYLLDSNPTRKMLDHIAREVGLKKRVVQVWFQNTRARER
ZFHX2_HUMAN_1857_1916   --DKRLR-TTILPEQLEILYRWYMQDSNPTRKMLDCISEEVGLKKRVVQVWFQNTRARER
ZFHX2_HUMAN_2065_2124   --QRRYR-TQMSSLQLKIMKACYEAYRTPTMQECEVLGEEIGLPKRVIQVWFQNARAKEK
ZFHX3_HUMAN_2944_3003   PGQKRFR-TQMTNLQLKVLKSCFNDYRTPTMLECEVLGNDIGLPKRVVQVWFQNARAKEK
ZFHX4_HUMAN_2884_2943   --HKRFR-TQMSNLQLKVLKACFSDYRTPTMQECEMLGNEIGLPKRVVQVWFQNARAKEK
LMX1A_HUMAN_195_254     --PKRPR-TILTTQQRRAFKASFEVSSKPCRKVRETLAAETGLSVRVVQVWFQNQRAKMK
LMX1B_HUMAN_219_278     --PKRPR-TILTTQQRRAFKASFEVSSKPCRKVRETLAAETGLSVRVVQVWFQNQRAKMK
LHX1_HUMAN_180_239      --RRGPR-TTIKAKQLETLKAAFAATPKPTRHIREQLAQETGLNMRVIQVWFQNRRSKER
LHX5_HUMAN_180_239      --RRGPR-TTIKAKQLETLKAAFAATPKPTRHIREQLAQETGLNMRVIQVWFQNRRSKER
LHX4_HUMAN_157_216      --AKRPR-TTITAKQLETLKNAYKNSPKPARHVREQLSSETGLDMRVVQVWFQNRRAKEK
LHX3_HUMAN_157_216      --AKRPR-TTITAKQLETLKSAYNTSPKPARHVREQLSSETGLDMRVVQVWFQNRRAKEK
                                :  :          .       :  :

HOMEZ_HUMAN_451_510     EVV---
ZHX1_HUMAN_777_832      LGIELF
ZHX3_HUMAN_835_894      RAV---
HOMEZ_HUMAN_55_114      ISW---
ZHX2_HUMAN_263_324      ISWSPE
ZHX3_HUMAN_304_363      ISW---
ZHX1_HUMAN_284_346      VSWTPE
ZEB2_HUMAN_644_703      SNS---
ZEB1_HUMAN_581_640      SVQ---
ZHX1_HUMAN_569_630      LKEEKM
ZHX2_HUMAN_530_591      SMEQAV
ZHX3_HUMAN_612_671      AEE---
ZHX2_HUMAN_439_501      RGIVHI
ZHX3_HUMAN_494_553      NLK---
ZHX1_HUMAN_464_526      NSKSNQ
HOMEZ_HUMAN_355_415     HGQ---
ZHX2_HUMAN_628_690      TGTVKW
```

But any decoration available here is far short of what can be achieved with **Jalview**, so click on the **Download Alignment File** button to save you alignment in a file on your **Desktop** called:

`homeobox_human_muscle.aln`

Jalview can be easily installed under all commonly used operating systems and run locally. For these exercises, I attempt to use services available freely from the **INTERNET** wherever possible, so let us run **Jalview** from the web here by first going to:

`http://www.jalview.org/`

and selecting the Launch Jalview Desktop link at the top of the page. And agree with all the many questions you will be asked.

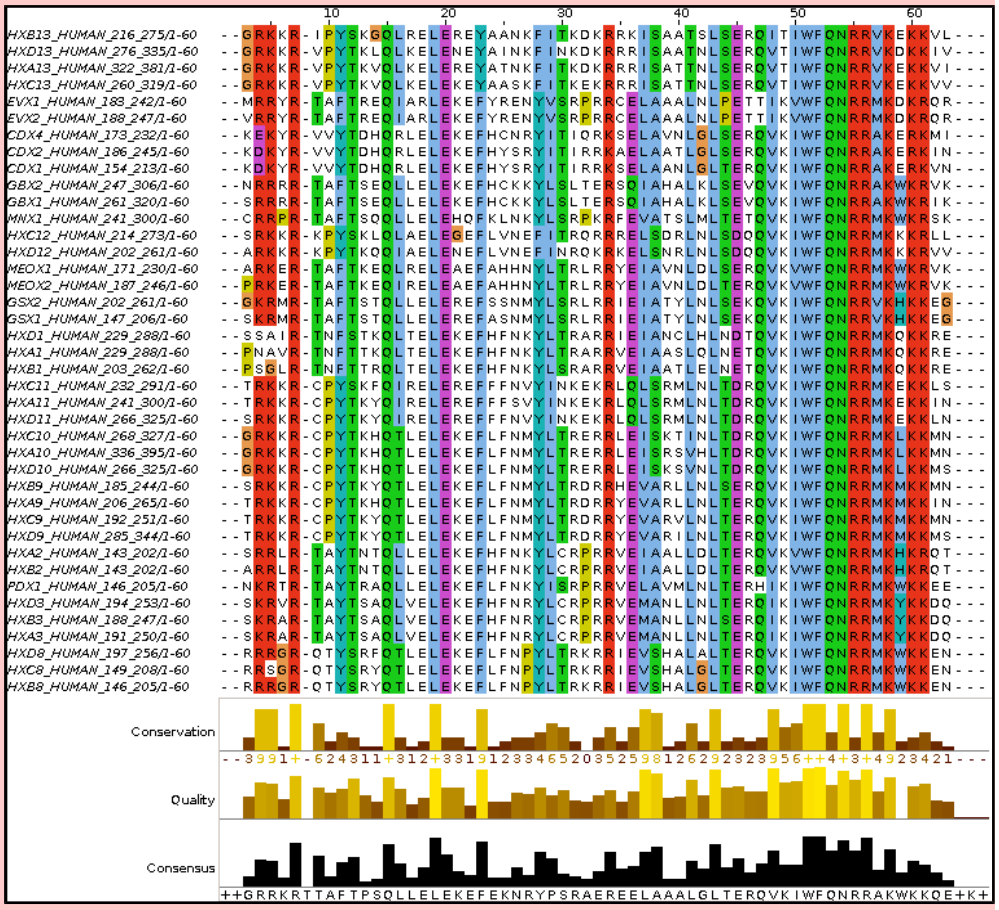
Close down all the example outputs **Jalview** sees fit to show you on start up. From the **File** pull down menu choose **from File** from the **Input Alignment** option. Locate and load the file:

`homeobox_human_muscle.aln`

You might need to adjust the file name filter to included **.aln** files.

The default view is a trifle bland. Try a few of the options from the **Colour** pull down menu.

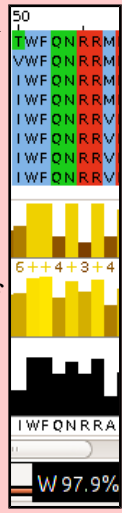
You could try the default colour scheme used by **ClustalX**, for example.



The **MUSCLE** and massaged **ClustalX** alignments now look very similar! In the nicely aligned regions at least.

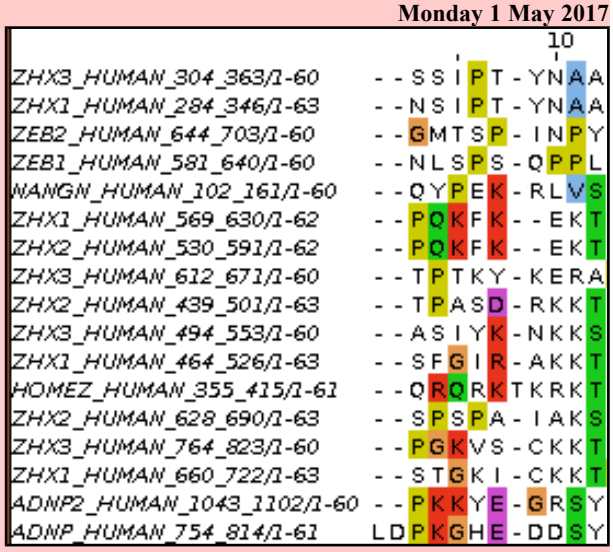
There are many **Jalview** features that merit investigation. Have a look around if you have time. In particular, **Jalview** will compute simple phylogenetic trees for you employing a number of methods (**Calculate Tree** from the **Calculate** pull down menu). Try it, but be aware this is only sensible if you were very sure of your alignment (and have more meaningfully selected sequences maybe?).

Jalview is made by the same group as produce **Jpred** (an extremely effective **Secondary Structure Prediction** system). You could send your alignment for **Secondary Structure Prediction** via the **Web Service** pull down menu, if you wished.



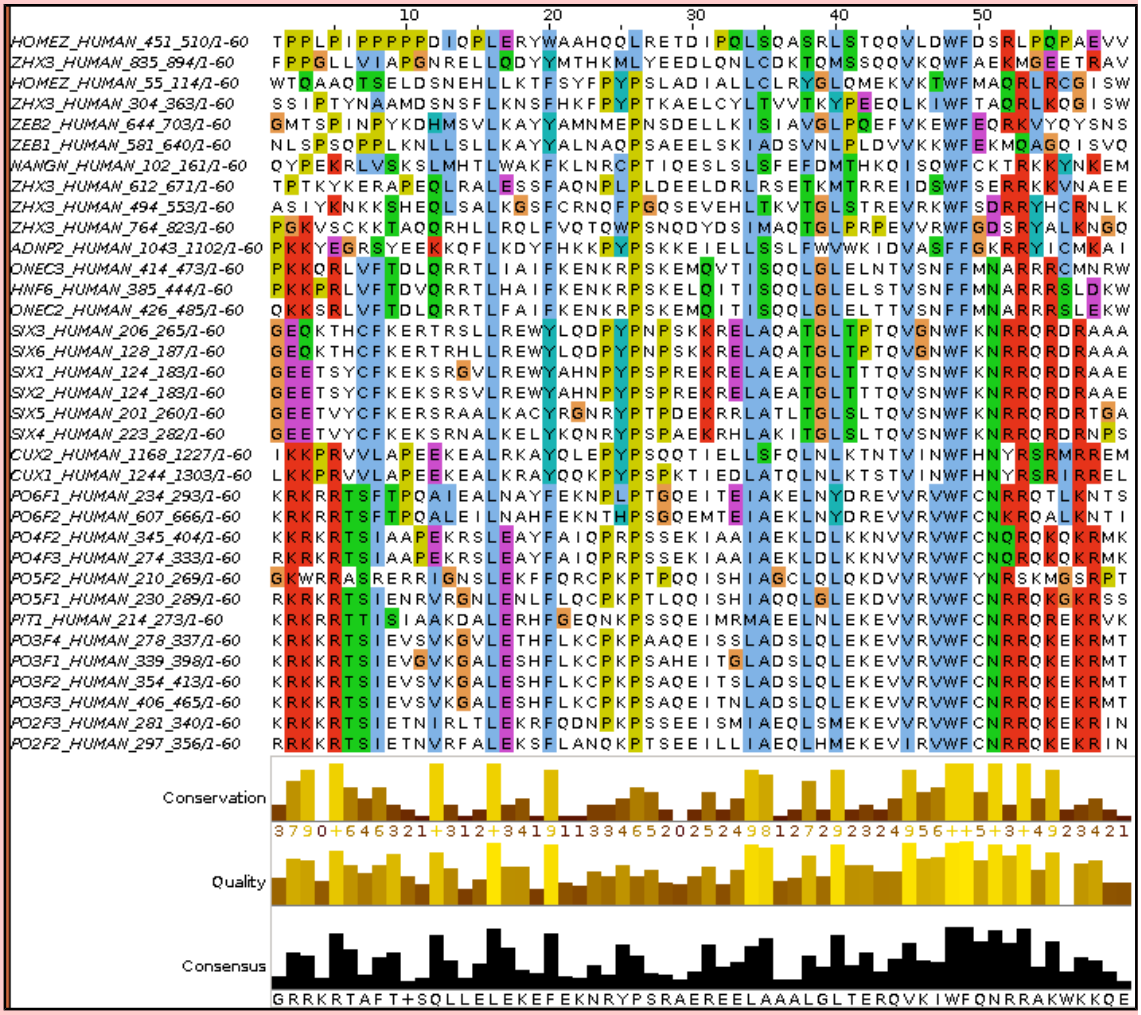
A central purpose of **Jalview** is to allow users to edit alignments as well as just to view them. For example, hold down the **Shift** key, click and hold on any amino acid at the edge of a gap, slide left and right and see that you can introduce and/or alter the position of gaps. It is very important to be able to edit alignments generated by even the best of programs. As I hope has been made clear, the alignment algorithms are crude. If you know something about the sequences you are aligning it is very reasonable to suppose you can improve upon the computer's alignments. **Jalview** tries to make this possibility easy. Look through some of the other **Edit** pull down menu options, maybe to increase the font size in particular!, it does not matter how much you mangle your alignment, you can always make another one.

Finally, take a look at the **Jalview** "**Manhattan Skyline**" for the highly conserved **W** at position **51**. This seems better quality than **clustalX** managed? I am not sure how one can make further comment without knowing what parameters were used. Is there really an improvement? If so, is it due to the improved algorithm or more appropriate choice of parameters? Impossible to discuss further as the parameters used for **MUSCLE** are not revealed.



In my alignment, the **W** at position **51** was at position **50**, according to **clustalx**. This slippage to the right is due to **MUSCLE** introducing an extra gap, inspired by just one sequence at position **8**. Is this sensible? No idea … exactly when it might be good idea to investigate the effect of lighter/heavier gap penalties?

You can also **Select** and **Cut** sequences in a way similar to that you employed with **clustalx**. I could not resist it! I removed all the ugly sequences that caused the gaps at the start and finish of the alignment, ***and*** the sequence that messed up column **8** (just select their names and then select **Cut** or **Delete** from the **Edit** menu). I achieved the gap-free beautiful alignment illustrated.



Of course, **Jalview** does not compute alignments, so once I had removed all the unfortunate proteins, I had to use an **Edit** option to tidy up my meddling. I used **Remove Empty Columns** to get rid of the gap columns at the start of the alignment. The gaps at the end just melted away once the sequences that supported their presence were removed.

Science is easy! Once you remove the need for honesty that is.

If it could be done slightly more meaningfully, I would suggest you might try some of the other **MSA** tools offered by the **EBI**, to investigate the differences in the alignments computed. Any differences might be due to different parameter selection or differences in the algorithms of the tool you select.

For full control, you really need to download the various tools and run them locally. The **EBI** is not the only site that hides significant parameters from their users.

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 further search of the protein database is then run using the **PSSM** as a query, and a larger more widely associated group of proteins is found. This larger group is aligned and 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.

PSI-BLAST is integrated into the **Secondary Structure Prediction** system **Jpred**. Whenever **Jpred** is asked to compute structure form a single protein sequence, it will use **PSI-BLAST** to construct an aligned family of protein sequences to enable an improved prediction. An aligned family of proteins is a much better starting point than any single protein sequence.

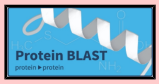
Similar ideas are used by the domain database **PFAM** to create large alignments of domain regions.

Here we will use **PSI-BLAST** directly from 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 alignment from the **PFAM** database.

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

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

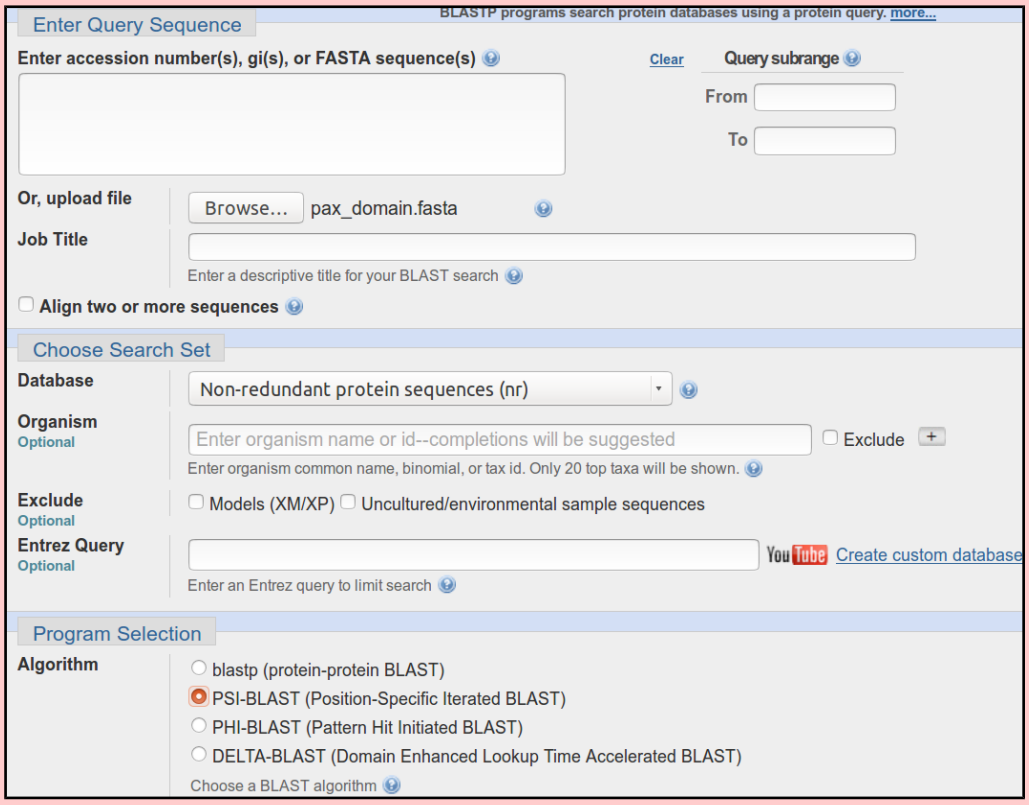
Click on the **BLAST** option from the **Popular Resources** menu.

Select 

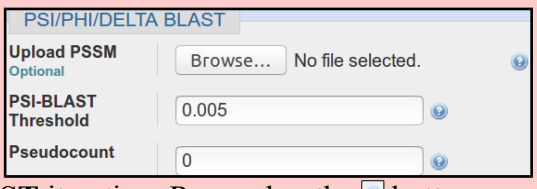
from the **Web BLAST** section.

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.



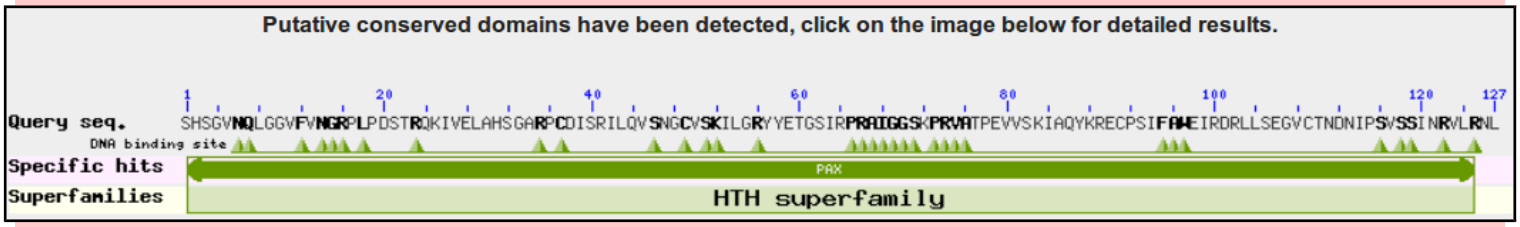
Before you set **PSI-BLAST** going, click on the **Algorithm parameters** link and take a look at the **PSI/PHI/DELTA BLAST** section. Note the option to use a **PSSM** from a previous run of **PSI-BLAST**, potentially on a different database (but with the same query sequence). Accept the default that 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.



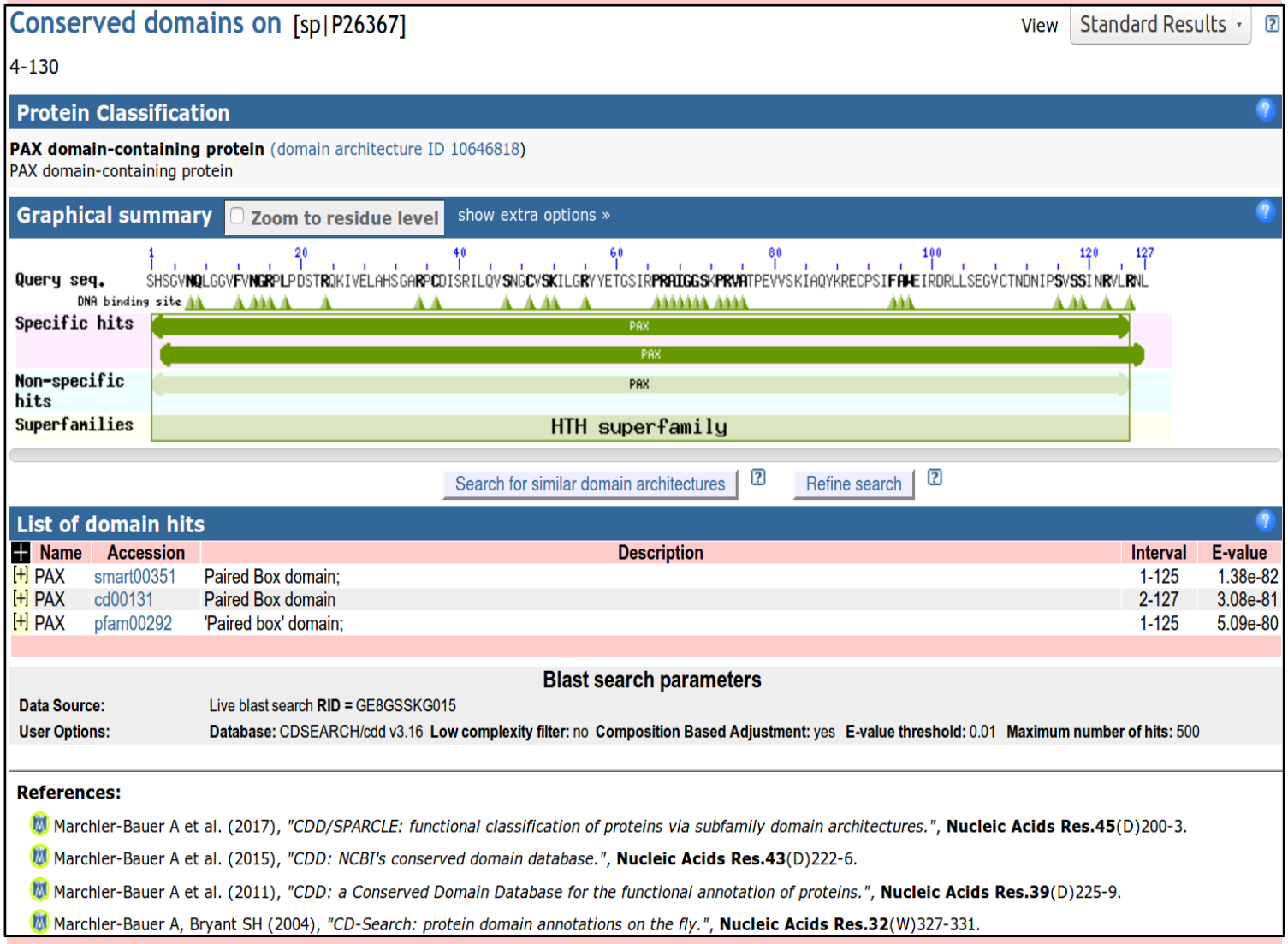
What do you suppose the choice of **Pseudocount** might influence? _____

Elect to **Show results in a new window** and then click on the BLAST button.

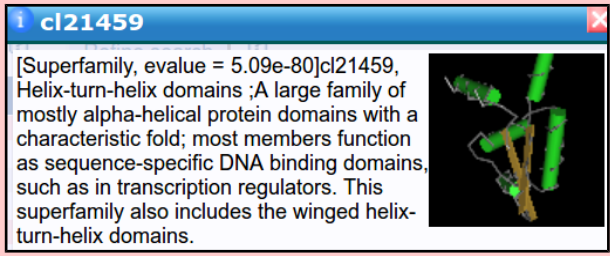
After several moments of deep thought, **PSI-BLAST** will come back with its first set of results, at the top of which is a report that (unsurprisingly) matches have been detected between the query sequence and several domain databases.



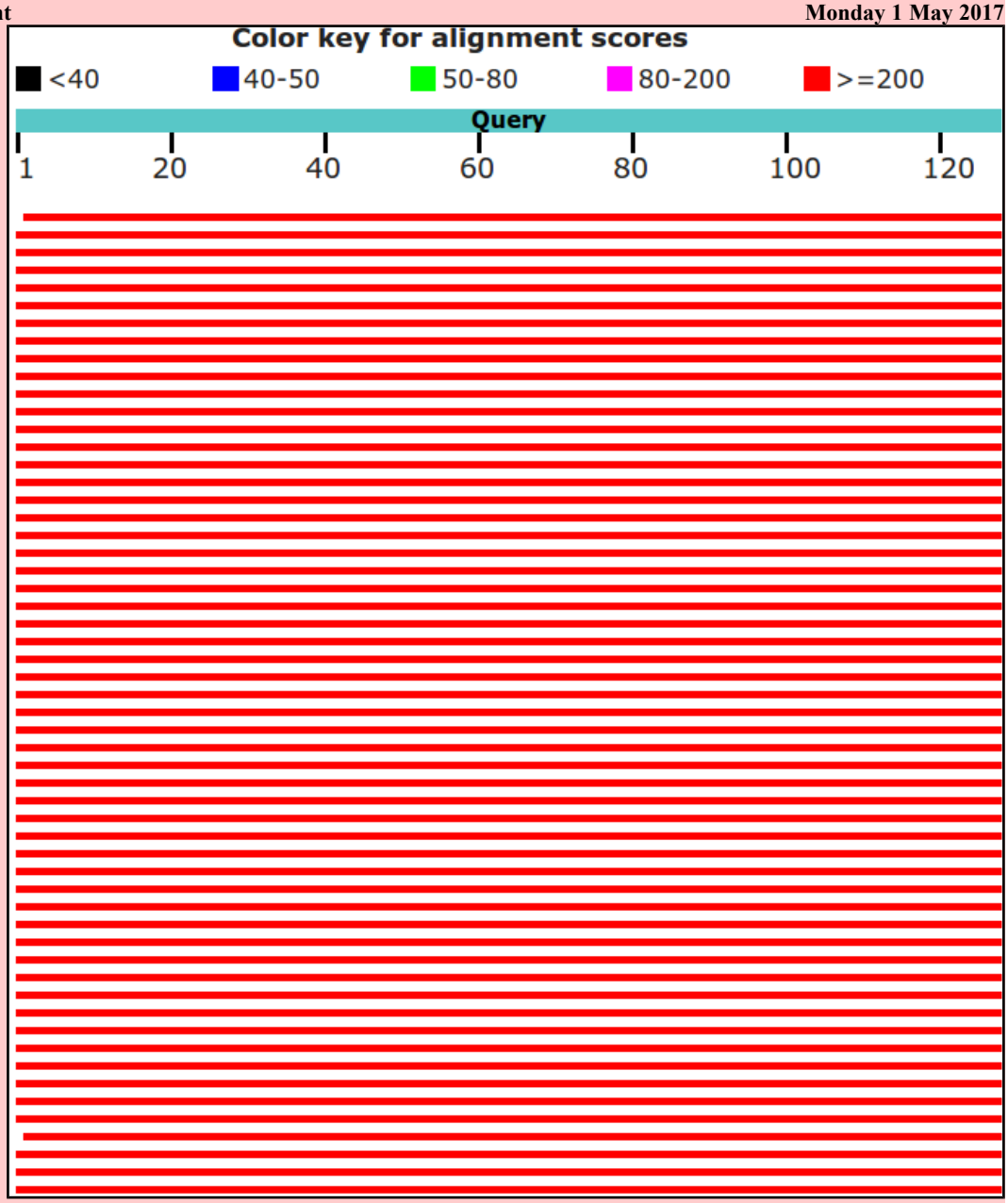
For more detail, click on the **Conserved Domains** graphic.



Hover over the **Specific** / **Non-specific hits** and you will see that **SMART**, **Pfam** and the **NCBI Conserved Domains** database matches for a **PAX** domain are all reported. No surprise here.



There is also a **Superfamilies** (derived from **SCOP** as briefly mentioned previously) hit recognising that a **PAX** domain, in common with many other domains, includes **Helix-Turn-Helices**.



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.

Note the **Accession Codes** that begin **XP_**. As mention previously, these are less well evidenced ptotein sequences from the **NCBI** databases.

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
hypothetical protein A6R68_04829 [Neotoma lepida]	257	257	99%	3e-86	100%	OBS66634.1	☑	
PREDICTED: paired box protein Pax-6 isoform X2 [Felis catus]	258	258	100%	3e-86	100%	XP_019667790.1	☑	
PREDICTED: paired box protein Pax-6 isoform X7 [Protobothrops mucrosquamatus]	262	262	100%	9e-86	100%	XP_015678414.1	☑	
PREDICTED: paired box protein Pax-6 isoform X4 [Papio anubis]	262	262	100%	4e-85	100%	XP_017804011.1	☑	
PREDICTED: paired box protein Pax-6 isoform X4 [Nanorana parkeri]	262	262	100%	5e-85	100%	XP_018423452.1	☑	
PREDICTED: paired box protein Pax-6 isoform X4 [Macaca nemestrina]	262	262	100%	5e-85	100%	XP_011722295.1	☑	
PREDICTED: paired box protein Pax-6 isoform X4 [Macaca mulatta]	262	262	100%	5e-85	100%	XP_014969998.1	☑	
PREDICTED: paired box protein Pax-6 isoform X2 [Acinonyx jubatus]	263	263	100%	5e-85	100%	XP_014922398.1	☑	
PREDICTED: paired box protein Pax-6 isoform X4 [Macaca fascicularis]	262	262	100%	5e-85	100%	XP_015289636.1	☑	
PREDICTED: paired box protein Pax-6 isoform X2 [Ursus maritimus]	263	263	100%	5e-85	100%	XP_008685073.1	☑	
PREDICTED: paired box protein Pax-6 isoform X7 [Pseudopodoces humilis]	262	262	100%	5e-85	100%	XP_014114466.1	☑	
PREDICTED: paired box protein Pax-6 isoform X2 [Manis javanica]	263	263	100%	6e-85	100%	XP_017519499.1	☑	

Download GenPept Graphics

hypothetical protein A6R68_04829, partial [Neotoma lepida]

Sequence ID: OBS66634.1 Length: 170 Number of Matches: 1

Range 1: 15 to 140 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
257 bits(656)	3e-86	Compositional matrix adjust.	126/126(100%)	126/126(100%)	0/126(0%)

```
Query  2    HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG  61
            HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG
Sbjct  15   HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG  74

Query  62   SIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSIN  121
            SIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSIN
Sbjct  75   SIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSIN  134

Query  122  RVLRNL  127
            RVLRNL
Sbjct  135  RVLRNL  140
```

Download GenPept Graphics

PREDICTED: paired box protein Pax-6 isoform X2 [Felis catus]

Sequence ID: XP_019667790.1 Length: 206 Number of Matches: 1

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

Score	Expect	Method	Identities	Positives	Gaps
258 bits(660)	3e-86	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
```

Move down to the **Alignments** section of the results and you will see that many of the top hits match the query exactly.

Note that many of the top hits come from the **GenPept** database (roughly equivalent to the **TrEMBL** section of **UniProtKB**).

How might the inclusion of poor quality and duplicated sequences have been minimised?_____

Download GenPept Graphics

Pax-6 variant form1 [Idiosepius paradoxus]

Sequence ID: BAM74251.1 Length: 408 Number of Matches: 1

Range 1: 12 to 137 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
252 bits(643)	5e-81	Compositional matrix adjust.	121/126(96%)	124/126(98%)	0/126(0%)

```
Query  2    HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG  61
            HSGVNQLGGVFVNGRPLPDSTRQ+IVELAHSGARPCDISRILQVSNGCVSKILGRYYETG
Sbjct  12   HSGVNQLGGVFVNGRPLPDSTRQRIVELAHSGARPCDISRILQVSNGCVSKILGRYYETG  71

Query  62   SIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSIN  121
            SIRPRAIGGSKPRVATPEVV KIAQ+KRECPSIFAWE+RDRLLSEGVCT DNIPSVSSIN
Sbjct  72   SIRPRAIGGSKPRVATPEVVQKIAQFKRECPSIFAWEVRDRLLSEGVCTQDNIPSVSSIN  131

Query  122  RVLRNL  127
            RVLRNL
Sbjct  132  RVLRNL  137
```

Download GenPept Graphics

PREDICTED: LOW QUALITY PROTEIN: paired box protein Pax-6 [Xiphophorus maculatus]

Sequence ID: XP_014326183.1 Length: 437 Number of Matches: 1

Range 1: 4 to 142 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
253 bits(646)	5e-81	Compositional matrix adjust.	125/139(90%)	127/139(91%)	12/139(8%)

```
Query  1    SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ-----------VSNG  48
            SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQ           VSNG
Sbjct  4    SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQTHDEVQVLDSEKVSNG  63

Query  49   CVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGV  108
            CVSKILGRYYETGSIRPRAIGGSKPRVATPEVV+KIAQYKRECPSIFAWEIRDRLLSEG+
Sbjct  64   CVSKILGRYYETGSIRPRAIGGSKPRVATPEVVAKIAQYKRECPSIFAWEIRDRLLSEGI  123

Query  109  CTNDNIPSVSSINRVLRNL  127
            CTNDNIPSVSSINRVLRNL
Sbjct  124  CTNDNIPSVSSINRVLRNL  142
```

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

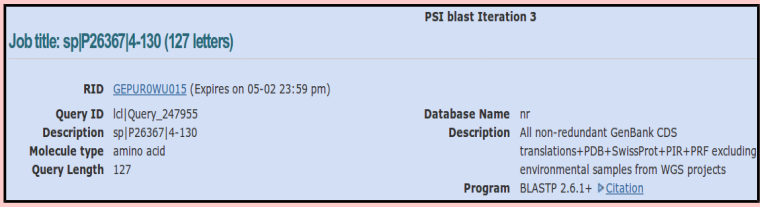
Having browsed your results sufficiently, click on the Go button to **Run PSI-Blast iteration 2**. It is at the bottom of the hit list.

Run PSI-Blast iteration 2 with max 500 Go

PREDICTED: paired box protein Pax-6-like [Notothenia coriiceps]	250	250	99%	3e-80	96%	XP_010768697.1	☑	✓
PREDICTED: paired box protein Pax-6-like isoform X1 [Trichogramma pretiosum]	254	254	99%	3e-80	93%	XP_014229927.1	☑	
Paired box protein Pax-6 [Chelonia mydas]	252	252	99%	3e-80	90%	EMP28095.1	☑	
PREDICTED: paired box protein Pax-6 isoform X2 [Crassostrea gigas]	252	252	99%	3e-80	96%	XP_011433289.1	☑	✓
Paired box protein Pax-6 [Eufriesea mexicana]	249	249	99%	3e-80	94%	OAD62380.1	☑	
XLPAX6 [Xenopus laevis]	250	250	100%	3e-80	89%	AAB07733.1	☑	✓
PREDICTED: paired box protein Pax-6 isoform X5 [Haplochromis burtoni]	250	250	100%	3e-80	90%	XP_014195922.1	☑	✓
PREDICTED: paired box protein Pax-6 isoform X4 [Drosophila suzukii]	247	247	99%	4e-80	92%	XP_016944157.1	☑	
PREDICTED: paired box protein Pax-6-like isoform X2 [Trichogramma pretiosum]	254	254	99%	4e-80	93%	XP_014229933.1	☑	
Paired box protein Pax-6 [Eufriesea mexicana]	254	254	99%	4e-80	95%	OAD53021.1	☑	
pax6 [Leptochiton asellus]	252	252	99%	4e-80	94%	AMB26739.1	☑	✓

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! As dear Eddie oft advised, there are typically but three steps to ultimate fulfilment. Recently I took just **8** iterations before there were no more new sequences suggested for inclusion into the **PSMM**. Recently, I was not so lucky, I got to iteration **21** before I realised that **PSI-Blast** was playing tricks one me! I was oscillating between two minutely different, perfectly acceptable solutions! Having vented my spleen in shame filled fashion I accepted iteration **21**. I advise that you stop here on "*good enough*" iteration **3**, as I will do this time!



Next, move to the just above the **Graphic Summary** and click on the **Multiple alignment** link. You have elected to use the **NCBI** multiple alignment program **Cobalt** to align the **PAX** domain sequences of your final **PSI-BLAST** iteration.

Alignment Parameters	
Gap penalties	-11,-1
End-Gap penalties	-5,-1
CDD Parameters	
Use RPS BLAST	on
Blast E-value	0.003
Find Conserved columns and Recompute	on
Query Clustering Parameters	
Use query clusters	on
Word Size	4
Max cluster distance	0.8
Alphabet	Regular

This can take quite a while. **Cobalt** might even complain wearily and give up occasionally. If it does, tell it not to be silly!! It will get there eventually. When it is done, click on the **Alignment parameters** link at the top of the results.

Cobalt reports the parameters it used to make the alignment. It is possible to recompute the alignment with different parameters by using the **Edit and Resubmit** link at the top of the page and then choosing to set **Advanced parameters**. But, maybe not today?

Recording the parameters chosen for any computation is surely extremely important. How else can published computer generated results be reproducible? Feel free to disagree, but I feel strongly this is a point not sufficiently appreciated by software engineers in this field and often entirely ignored by service providers (e.g. the "*we have chosen the best parameter settings for you and feel you do not need to even know what they are*" approach for the **EBI MSA** options).

```
AAB07733        5   ---HSGVNQLGGVFVNGRPLPDS--TRQKIVELAHSGARPCDISRILQSHADAKVPVLDSQNVSNGCVSKILG----RYY  75
XP_014853331   29   DEGHSGVNQLGGVFVNGRPLPDS--TRQKIVELAHSGARPCDISRILQTHDE--VQVLDSEKVSNGCVSKILG----RYY 100
XP_015229805   29   DEGHSGVNQLGGVFVNGRPLPDS--TRQKIVELAHSGARPCDISRILQTHDE--VQVLDSEKVSNGCVSKILG----RYY 100
XP_015193792    5   ---HSGVNQLGGVFVNGRPLPDS--TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNVSNGCVSKILG----RYY  75
XP_012546782   30   G--HSGVNQLGGVFVGGRPLPDS--TRQKIVELAHSGARPCDISRILQ-------------VSNGCVSKILG----RYY  87
XP_003376863   44   -LGHTGVNQLGGVFVNGRPLPDS--TRQKIIELAHQGARPCDISRILQ-------------VSNGCVSKILC----RYY 102
XP_015364286   72   G--HSGVNQLGGVFVGGRPLPDS--TRQKIVELAHSGARPCDISRILQ-------------VSNGCVSKILG----RYY 129
XP_018423443    5   ---HSGVNQLGGVFVNGRPLPDS--TRQKIVELAHSGARPCDISRILQSHADAKVQVLDSQNVSNGCVSKILG----RYY  75
CAJ40659       42   ---HSGVNQLGGVFVNGRPLPDS--TRQRIVELAHSGARPCDISRILQ-------------VSNGCVSKILG----RYY  98
CAC80515        5   ---HSGVNQLGGVFVNGRPLPDS--TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNENVSNGCVSKILG----RYY  75
XP_015364293   53   G--HSGVNQLGGVFVGGRPLPDS--TRQKIVELAHSGARPCDISRILQ-------------VSNGCVSKILG----RYY 110
XP_003777840    5   ---HSGVNQLGGVFVNGRPLPDS--TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNVSNGCVSKILG----RYY  75
CEH19759        6   ---HSGINQLGGVYVNGRPLPDS--TRQKIVELAHSGARPCDISRILQ-------------VSNGCVSKILG----RYY  62
ACD88758        5   G--HSGVNQLGGVYVNGRPLPDS--TRQKIVELAHSGARPCDISRILQ-------------VSNGCVSKILG----RYY  62
XP_014969997    5   ---HSGVNQLGGVFVNGRPLPDS--TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNVSNGCVSKILG----RYY  75
XP_016844277   33   ---HSGVNQLGGVYVNGRPLPDT--TRQKIVELAHSGARPCDISRILQ-------------VSNGCVSKILG----RYY  89
XP_003246860   72   G--HSGVNQLGGVFVGGRPLPDS--TRQKIVELAHSGARPCDISRILQ-------------VSNGCVSKILG----RYY 129
AAB40616       10   PNGHSGVNQLGGVFVNGRPLPDS--TRQRIVELAHSGARPCDISRILQ-------------VSNGCVSKILG----RYY  69
XP_011722290    5   ---HSGVNQLGGVFVNGRPLPDS--TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNVSNGCVSKILG----RYY  75
CBX88047        5   ---HSGVNQLGGVFVNGRPLPDT--IRQKIVELAHSGARPCDISRILQ-------------VSNGCVSKILG----RYY  61
XP_015289635    5   ---HSGVNQLGGVFVNGRPLPDS--TRQKIVELAHSGARPCDISRILQTHADAKVQVLDNQNVSNGCVSKILG----RYY  75
XP_003246859   53   G--HSGVNQLGGVFVGGRPLPDS--TRQKIVELAHSGARPCDISRILQ-------------VSNGCVSKILG----RYY 110
```

Move past the long list of aligned proteins (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. This might take a while also. I am not sure why? Be patient, it will get therein the end. 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.

DPJ – 2017.05.01

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 **Multiple Sequence Alignment**

How do the options for the **OUTPUT TREE** relate to the output files of **ClustalX** and the difference between the way that **ClustalX** and **muscle** work?

I leave this question here in the hope that one day I will be able to offer a full and sensible answer. First draft answer below.

Essentially, both **ClustalX** and **MUSCLE** work in two stages. First they create **Guide Tree(s)**. Then they create a multiple alignment by pairwise steps ordered by most refined the **Guide Tree**.

ClustalX just computes one based exclusively on the pairwise comparison of its input sequence set.

MUSCLE will create a **Guide Tree** that is the rough equivalent of that computed by **ClustalX**. Then it will offer to refine this **Guide Tree** from computed draft **MSA**s until a user selected maximum number of iterations is met or no further improvement is possible.

ClustalX saves the **Guide Tree** it computes by default. **MUSCLE** offers to save its **Guide Tree** from its first or second refinement iteration.

The purpose of saving the **Guide Tree(s)** to a file is to enable a rerun of the second phase with new parameter settings without having to first recalculate the **Guide Tree**. Of course, as mentioned previously, utterly pointless if there is no way to change the parameters to allow a guide tree to be used as input? but that is the theory.

More investigation by me and expansion of this answer required. Discussion with EBI current (2016.04.20).

Comment on how one might choose between the range of options offered for the aligned parameter?

I cannot … beyond suggesting it simply does not make sense? Going by what is offered at **Wageningen**, the choice should be between **aligned** and **input order**. i.e. the order of the original set of sequences to be aligned or the order after they have all been compared with each other and arranged into a **Guide Tree** … or two.

Currently, the only way of which I am aware to run muscle with full flexibility, is to download it. It is available for **Windows**, **Linux** or **Mac** operating systems but has no pretty **GUI** front end. You have to read the manual carefully and run from the command line.

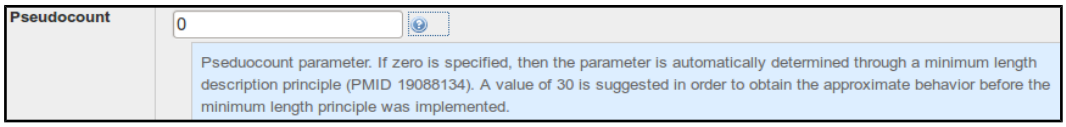
To attempt (with pain) to be fair, one might suggest that web services are for creating draft results primarily. If one wanted to get serious and have full controll over the software and record properly all the settings one has chosen, it would make sense to download the software and run in locally.

That still does not excuse offering selections that only have one option and/or save files that cannot serve any fucntion. I think I give up trying to persuade the **EBI** guys of this and just live with “what is”. So much more restful (2017.05.01).

From your investigations of **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.

In the meantime I will take comfort in the claim that:

"A new method for calculating **pseudocounts** that significantly improves **PSI-BLAST**'s; retrieval accuracy is now employed by default."

Jolly good!

2016.12.04: Aha! **Wikipedia** to the rescues once more. Maybe I will donate again? Wonderful service.

One can forgive the **NCBI** people for not explaining what a **pseudocount** is, as they did not, as I first thought, invent the term. It is an idea/strategy of far wider and general application as wikipedia explains.

My interpretation of this article (feel free to disagree/correct) in the current context is:

PSSMs are computed from the amino acid composition of regions of a protein sequence. Their purpose is to identify other protein regions of the same size that might be homologous. If a given amino acid is not represented at all in the region from which the **PSSM** is computed, the probability of any other region including that missing amino acid will be consider to be **0** (i.e. impossible!) even if the rest of the region matches extremely well.

Generally speaking, that would be a nonsense! Solution? Add a tiny bit (a **pseudocount** even) to all amino acid counts that come to **0**. Then "*impossible*" becomes "*extremely unlikely*", which makes a bit more sense. A trifle more poetry than science here, but I think I follow the logic.

A popular way of implementing pseudocounts is due to **Pierre-Simon Laplace**. A French chap who was pretty famous for having good ideas. His strategy, nattily known as **Laplace's Rule of Succession**, was to add a **psuedocount** of **1** to ***ALL*** the real counts and so pervert the message of the data uniformly. Nice one **Pierre**.

I am not entirely sure why, but this all reminds me of one of the many dubious culinary practices of my dear mother (when not in the kitchen, an unsurpassed example of the human female condition!). To-whit, when confronted with a spice or condiment with which she was unfamiliar, she would avoid the unacceptable **zero condition** by adding a swift **pseudocount** (sometimes **two**!) into whatever she was brewing at the time. The principle being that of "*just in case*" and the avoidance of the horror filled possibilities of "*missing an exciting new flavour*".

She used to protect the family from any ill effects by assiduously, testing the **psuedocount** side effects upon its most dispensable member … the youngest son, say? If he still frisked after a given period, she would let loose the potion upon the rest of the family. Happily, I survive! But repeated **pseudocount** experimentations may well explain much of the condition of what remains.

How might the inclusion of poor quality and duplicated sequences have been minimised?

At the top of your output is recorded some details of the conditions under which you database search was undertaken. This is a very important step towards making your results reproducible. Not sufficient I would opine. Surely the database versions and a complete record of the parameters used by **blast** are required in order to be able to exactly reproduce a search?

Database Name	nr
Description	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Program	BLASTP 2.6.1+ ▷Citation

But at least the version of **blast** and the databases that were searched are recorded. The collection of databases searched is rather optimistically called "**nr**", for non-redundant. A bit of an exaggeration I would think. Surely **PDB** and **SwissProt** overlap a trifle? But let us not be too picky, in fact, surely a noble attempt to remove duplication between these databases has been made, understandably, imperfectly.

The collection of databases that is **nr** includes "*All non-redundant **GenBank CDS** translations*" (aka **GenPept**) which, like it European broad equivalent **TrEMBL**, includes some pretty dubious sequences.

I would think that if one wanted to maximise quality and minimise duplication, it would be best to pick just one good quality database. **SwissProt** is the obvious choice. **blast**, in general, and **PSI-BLAST** in particular, allows such a selection.

However, today the objective is not refinement!!! Bloat is good! More the merrier! Never mind the quality, just admire the volume.

DPJ – 2017.05.01