



# GTPB

The Gulbenkian Training Programme in Bioinformatics  
(Since 1999)

Pedro Fernandes, Organiser



INSTITUTO  
GULBENKIAN  
DE CIÊNCIA

# ELB19<sub>9</sub>F

## Entry Level Bioinformatics

04-08 February 2019

(First 2019 run of this Course)

## Basic Bioinformatics Sessions

### Practical 6: Multiple Sequence Alignment

Friday 8 February 2019

## 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 **Reviewed** (that is, only **SwissProt** entries).

If required, 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**.

Searching in UniProtKB Help

Term

Organism [OS] Human [9606]

AND Reviewed -> Reviewed

AND Function -> DNA binding Homeobox Length range From 50 To 70 Evidence Any assertion method

AND All

Search

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 **homeoboxes** 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 **Search** button with authority to get the search going.

Entry	Entry name	Protein names	Gene names	Organism	Length
P28356	HXD9_HUMAN	Homeobox protein Hox-D9	HOXD9 HOX4C	Homo sapiens (Human)	352
Q92826	HXB13_HUMAN	Homeobox protein Hox-B13	HOXB13	Homo sapiens (Human)	284
P31271	HXA13_HUMAN	Homeobox protein Hox-A13	HOXA13 HOX1J	Homo sapiens (Human)	388
P17482	HXB9_HUMAN	Homeobox protein Hox-B9	HOXB9 HOX2E	Homo sapiens (Human)	250
P31270	HXA11_HUMAN	Homeobox protein Hox-A11	HOXA11 HOX1I	Homo sapiens (Human)	313
P49639	HXA1_HUMAN	Homeobox protein Hox-A1	HOXA1 HOX1F	Homo sapiens (Human)	335
P09629	HXB7_HUMAN	Homeobox protein Hox-B7	HOXB7 HOX2C	Homo sapiens (Human)	217
P31249	HXD3_HUMAN	Homeobox protein Hox-D3	HOXD3 HOX1D, HOX4A	Homo sapiens (Human)	432
P31260	HXA10_HUMAN	Homeobox protein Hox-A10	HOXA10 HOX1H	Homo sapiens (Human)	410
P20719	HXA5_HUMAN	Homeobox protein Hox-A5	HOXA5 HOX1C	Homo sapiens (Human)	270

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**

Download selected (0) ✕  
 Download all (237)  
**Format:**  
  
 Compressed  Uncompressed  
[Preview first 10<sup>i</sup>](#)

```
ID MEOX2_HUMAN Reviewed; 304 AA.
AC P50222; A4D127; B2R8I7; O75263; Q9UPL6;
DT 01-OCT-1996, integrated into UniProtKB/Swiss-Prot.
DT 18-APR-2006, sequence version 2.
DT 25-OCT-2017, entry version 159.
DE RecName: Full=Homeobox protein MOX-2;
DE AltName: Full=Growth arrest-specific homeobox;
DE AltName: Full=Mesenchyme homeobox 2;
GN Name=MEOX2; Synonyms=GAX, MOX2;
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], AND VARIANT HIS-80 DEL.
RC TISSUE=Embryo;
RX PubMed=7607679; DOI=10.1016/0888-7543(95)80174-K;
RA Grigoriou M., Kastriaki M.-C., Modi W., Theodorakis K., Mankoo B.,
RA Pachnis V., Karagogeos D.;
RT "Isolation of the human MOX2 homeobox gene and localization to
RT chromosome 7p22.1-p21.3.";
RL Genomics 26:550-555(1995).
RN [2]
RP NUCLEOTIDE SEQUENCE [MRNA], AND VARIANT 79-HIS-HIS-80 DEL.
RC TISSUE=Heart;
```

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 Feature Tables and most often refer to a **Homeobox** region.

In the **DNA\_BIND** Feature Table 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 1 304 Homeobox protein MOX-2.
FT /FTid=PRO_0000049197.
FT DNA_BIND 187 246 Homeobox. {ECO:0000255|PROSITE-
FT ProRule:PRU00108}.
FT COMPBIAS 42 47 Poly-Ser.
FT COMPBIAS 68 80 Poly-His.
FT COMPBIAS 81 86 Poly-Gln.
FT VARIANT 79 80 Missing. {ECO:0000269|PubMed:7713505}.
FT /FTid=VAR_026040.
FT VARIANT 80 80 Missing. {ECO:0000269|PubMed:12690205,
FT ECO:0000269|PubMed:14702039,
FT ECO:0000269|PubMed:15489334,
FT ECO:0000269|PubMed:7607679}.
FT /FTid=VAR_026041.
FT VARIANT 287 287 I -> L (in dbSNP:rs2237493).
FT /FTid=VAR_049585.
FT MUTAGEN 236 236 Q->E: Abolishes DNA-binding. Does not
FT affect ability to activate expression of
FT CDKN2A. {ECO:0000269|PubMed:22206000}.
FT CONFLICT 58 58 G -> D (in Ref. 2; AAA58497).
FT {ECO:0000305}.
SQ SEQUENCE 304 AA; 33594 MW; 0C008479D6995389 CRC64;
MEHPLFGCLR SPHATAQGLH PFSQSSLALH GRSDHMSYPE LSTSSSSCII AGYPNEEGMF
ASQHRHGHHS HHHHHHHHHH QQQHQALQT NWHLPMSSP PSAARHSLCL QPDSGGPPEL
GSSPVLCSN SSSLGSSTPT GAACAPGDYG RQALSPAEAE KRSGGKRKSD SSDSQEGNYK
SEVNSKPRKE RTAFTKEQIR ELEAEFAHFN YLTRLRRYEI AVNLDLTERQ VKWVFQNRM
KWKRVKGGQQ GAAAREKELV NVKKGTLPLS ELSGIGAATL QQTGDSIANE DSHSDHSSE
//
```

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.

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.

Input section

Select an input sequence. Use one of the following three fields:

- To access a sequence from a database, enter the USA here:
- To upload a sequence from your local computer, select it here:  human\_homeobox\_proteins.emb
- To enter the sequence data manually, type here:

Additional section

Amount of sequence before feature to extract

Amount of sequence after feature to extract

Source of feature to display

Type of feature to extract

Sense of feature to extract   
(default is 0 - any sense, 1 - forward sense, -1 - reverse sense)

Minimum score of feature to extract

Maximum score of feature to extract

Tag of feature to extract

Value of feature tags to extract

Output section

Output introns etc. as one sequence?

Append type of feature to output sequence name?

Feature tag names to add to the description

Output sequence format

Run section

Email address:   
*If you are submitting a long job and would like to be informed by email when it finishes, enter your email address here.*

```

OUTPUT FILE outseq
ID MEOX2_HUMAN_187_246 Reviewed; 60 AA.
DE [DNA_contact] Homeobox protein MOX-2 (Growth arrest-specific homeobox) (Mesenchyme homeobox 2)
SQ SEQUENCE 60 AA; 7615 MW; 7AA1CEC5BB0265F CRC64;
PRKERTAFATK EQIRELEAEF AHHNYLTRLR RYEIAVNLDL TERQVKVWFQ NRRMKWKRVK
//
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;
RRKPRVLF5Q AQVYELERRF KQQRYSAPR RDQLASVLKLL TSTQVKIWFQ NRRYKCKRQR
//
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 PEARIQVWFQ NRRAKWRREE
//
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 KOLTELEKEF HFNKYLTRAR RVEIAASLQL NETQVKIWFQ NRRMKQKKRE
//

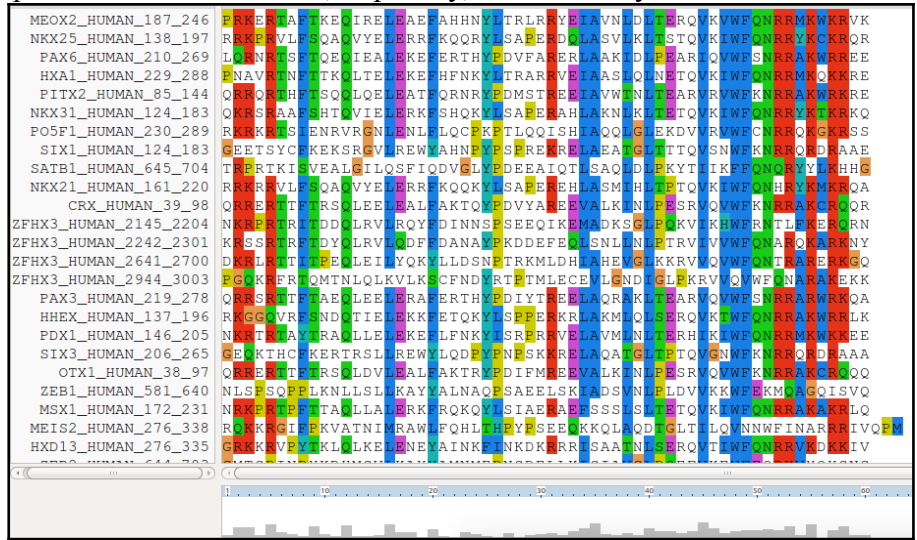
```

**ClustalX** is a part of the mostly widely known family of Multiple Sequence Alignments (MSA) programs, originating in the **1980s**. 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 **Graphical User Interface (GUI)**. A good place for us to start. It is, hopefully, installed on your workstations.

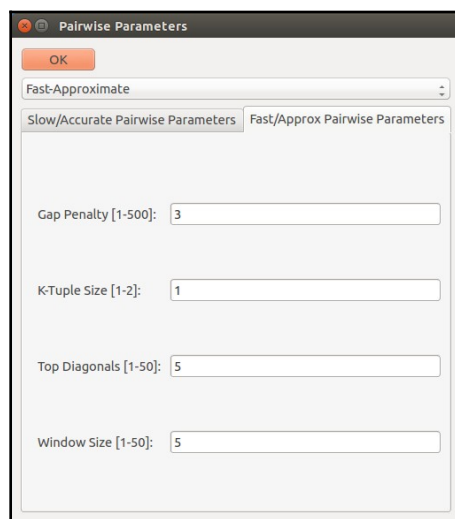
Start up the program **ClustalX**<sup>1</sup>. The **ClustalX Graphical User Interface (GUI)** will regally mount your screen.

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.



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.

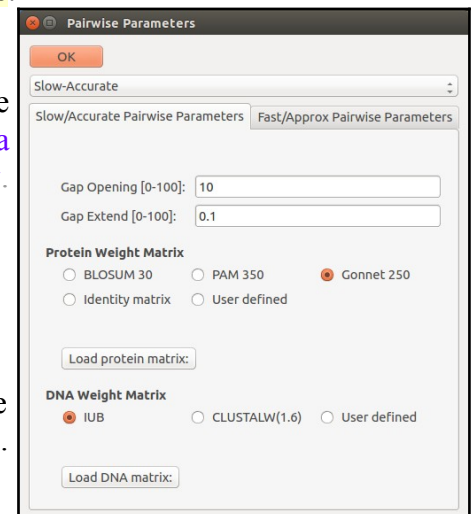


From the **Alignment** pull down menu, go to the **Alignment parameters** menu and select **Pairwise Alignment Parameters**. Just for a moment, change the setting from **Slow-Accurate** to **Fast-Approximate**. Bring the corresponding parameters into view by clicking on **Fast/Approx Pairwise Parameters**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

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**.

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 ring a bell? 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.

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, little further explanation should be required here.

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**.

**Format**

- CLUSTAL format
- GCG/MSF format
- GDE format
- 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.

**Output Files**

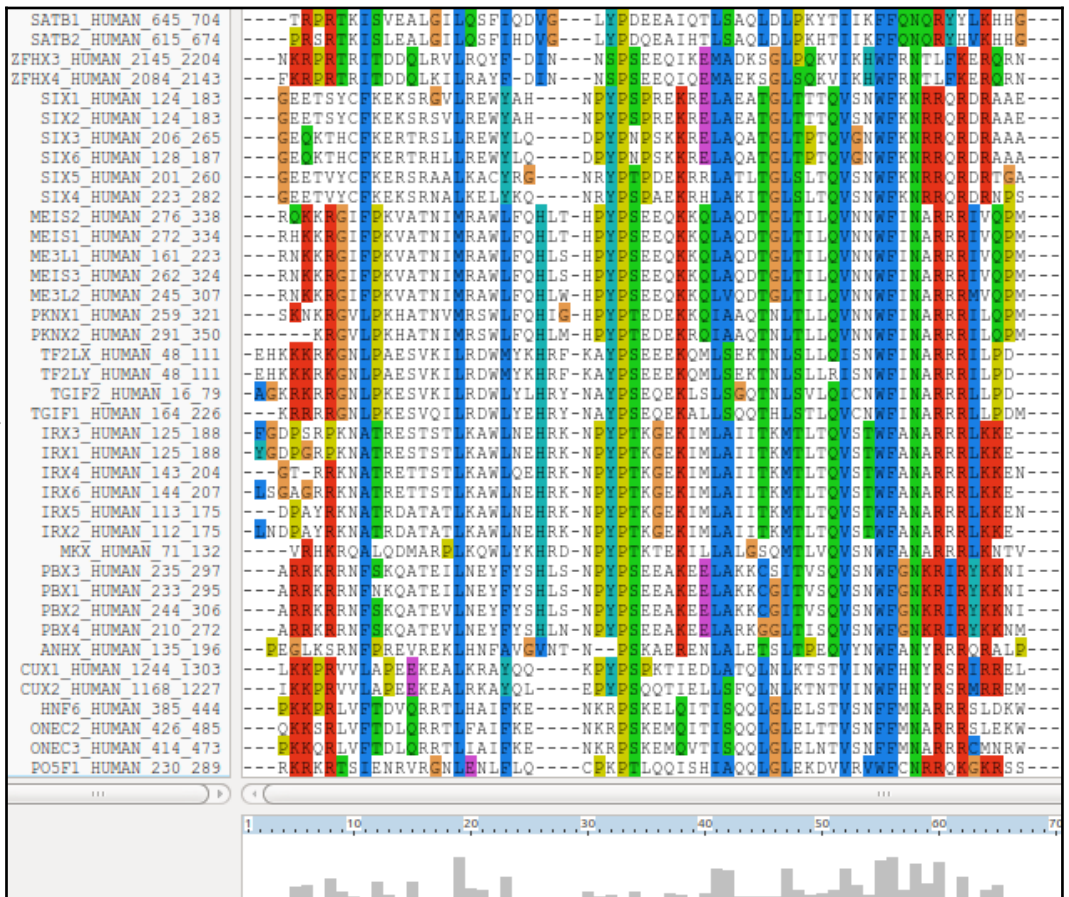
- CLUSTAL format
- NBRF/PIR format
- GCG/MSF format
- PHYLP format
- GDE format
- NEXUS format
- FASTA format

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!

Just to investigate the 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 same alignment from scratch for the reduced sequence set. To justify this assertion, select **Select All Sequences** from the **Edit** 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.



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. Well, I got back to where I was, no gaps around position **30** but still the 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 represent protein features (in addition to **HMMs**). 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? Quick

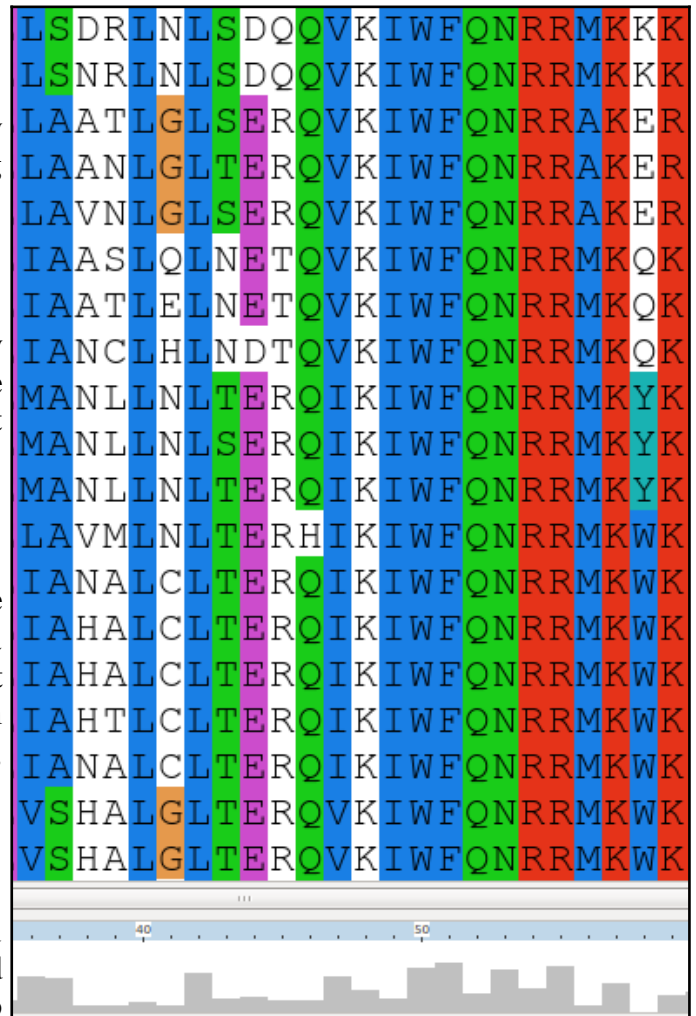
Hint?

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<sup>3</sup>. 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**.



```
ZHX1_HUMAN_660_722 -STGKI-CKKPELHMKSAIVRTQWPSPEEYDKLAKESGLARTDVSWFQDTRYAWKNGNLKW
ZHX3_HUMAN_764_823 --PGKVSCKKTAQRHLRLQLEVQTPWPSNQDYDSIMAOGLPEEVVRFQDSRYALKNGQ---
ZHX2_HUMAN_628_690 -SESPA-IARKSQEQVHLRSTFARTQWPTQOEYDQLAAKGLVRFTEVRFKFNRLGLLTGTVKW
HOMEZ_HUMAN_355_415 -QQRKTKRKRKEQALAIKSPFLQCQWARREYQKLEQIPLRPEIQWFGDTRYALRHGQ---
ZHX2_HUMAN_439_501 --TPASDRKKEQIAHMKASELQSQFDDAEVYRLTEVGLARSEIKKWFSDHRYRCORGIVHI
ZHX1_HUMAN_464_526 --SFGIYAKKKEQIAELKVSYLKNQFPHDSEIIRLMKITGLRGEIKKWFSDTRYNQNSKSNQ
ZHX3_HUMAN_494_553 --ASIIYKNNKSHQALSALKSGFCRNQFPQSEVEHLTKVGLSTREVRKWFSDRRYHCSNLK---
ZHX2_HUMAN_530_591 ---PKFKEKTPQGVKILIDSEFLKSSFPQAEELDRARVEIKLSRREIDSWFSERRKLRSMEQAV
ZHX1_HUMAN_569_630 ---PKFKEKAEQLRVLQASFLNSSVLDEELNRLRAQKLRREIDAWFTEKSKSKALKEEKW
ZHX3_HUMAN_612_671 --TPTKYKERAPEQLRALSSFAQNFPLDEELDRARSEKMRREIDSWFSERRK-KVNAEE--
DNP2_HUMAN_1043_1102 --PKRYEGRYEEKQFLKDYFHKKPYPSKKEIEELSSLEVWVKIDVASFPGKRRYICMKAI---
ADNP_HUMAN_754_814 --DEKGHEDDYEAKKSFPTKYFNKQYPTRRREIEKLAASLWLVKSDSEASHFSKRRKCVRD---
ZHX1_HUMAN_284_346 -NSIPTYNALDN-NFLLLNNTYKFFPYPTMSEITVLSAQAKYEEQIKIWFSAQRLLKHGVSWTPE
ZHX3_HUMAN_304_363 --SSIPTYNAMDS-NSEFLKNSFKHFPYPTKAEELCYLVVVKYPEEQIKLWFTAQRLLKQGISW---
```

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.

```
LHX6_HUMAN_219_278 --AKKARTSFAEQLQVMAQFAQDNNDAQTLQKLADMGLRRVTVVWFQNCRRARHKKHT---
LHX9_HUMAN_267_326 --TKRMRTSKHHQLRTMKSYFAINHNDAKDLKLAQKGLTKRVLQVWFQNCRRARFRNNL---
LHX2_HUMAN_266_325 --TKRMRTSKHHQLRTMKSYFAINHNDAKDLKLAQKGLTKRVLQVWFQNCRRARFRNNL---
DPRX_HUMAN_16_75 --SHRKTMPKQKLEEDNILENENYFNESLKEASKIDHITVQVWFQNCRRARFRNNL---
ZEB1_HUMAN_581_640 -NLSSEPPFKNL-LSLKAYALNAQFAEELSKLADSNLFLDVVKKWFQKMGAGQISVQ---
ZEB2_HUMAN_644_703 -GMTSEINPKYKH-MSVLKAYAMNMEFNSDELLKLSIAVGLPQEFVKEWFQKRVVQYSNS---
ZHX1_HUMAN_777_832 ---KFKTG-FAIKDYLLKHKFLNEQDLDLVNKSHEGVEVREFAERQRSELGTEL
ZHX3_HUMAN_835_894 --FPGLLVAFGNRELLCDYMTHKMLYEEDLQNLCDKQMSQVQVQKQWFAERMGEEETLAV---
HOMEZ_HUMAN_451_510 --TPEPLIPPEPDTQELERYAAHQOQRETDFQLSQASRLTQOVLDFWEDSLPQAEVW---
NANGN_HUMAN_102_161 --QYPERLVSKSLMHTLWAKFKLNRCPFIQESLSLSFEDMHHKIQWFCCKTKKKNKEM---
```

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.

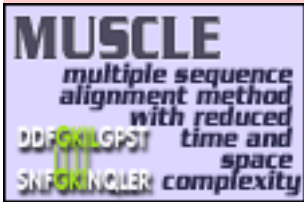
Here I intend to align again the human **homeboxes** with just one of the tools on offer. Then take a quick look at how the machine generated multiple alignment can be manually edited using **Jalview**, a program that is probably installed on your workstation and definitely available as a web service. 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**.

<p><b>Clustal Omega</b> </p> <p>New MSA tool that uses seeded guide trees and HMM profile-profile techniques to generate alignments. Suitable for medium-large alignments.</p> <p><a href="#">Launch Clustal Omega</a></p>
<p><b>Kalign</b> </p> <p>Very fast MSA tool that concentrates on local regions. Suitable for large alignments.</p> <p><a href="#">Launch Kalign</a></p>
<p><b>MAFFT</b> </p> <p>MSA tool that uses Fast Fourier Transforms. Suitable for medium-large alignments.</p> <p><a href="#">Launch MAFFT</a></p>
<p><b>MUSCLE</b> </p> <p>Accurate MSA tool, especially good with proteins. Suitable for medium alignments.</p> <p><a href="#">Launch MUSCLE</a></p>
<p><b>MView</b> </p> <p>Transform a Sequence Similarity Search result into a Multiple Sequence Alignment or reformat a Multiple Sequence Alignment using the MView program.</p> <p><a href="#">Launch MView</a></p>
<p><b>T-Coffee</b> </p> <p>Consistency-based MSA tool that attempts to mitigate the pitfalls of progressive alignment methods. Suitable for small alignments.</p> <p><a href="#">Launch T-Coffee</a></p>
<p><b>WebPRANK</b></p> <p>The EBI has a new phylogeny-aware multiple sequence alignment program which makes use of evolutionary information to help place insertions and deletions.</p> <p>Try it out at <a href="#">WebPRANK</a>.</p>



So the plan now is to use **MUSCLE**<sup>4</sup> 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<sup>5</sup>. 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 include the sequences with a mess around position **30**.

the programs that their creators deemed sensible to make available<sup>6</sup>?

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 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<sup>7</sup>. 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**<sup>8</sup>. The default **ClustalW** or **Pearson/FASTA** serve for this purpose. As **ClustalW** looks more like an alignment in the web page, I choose **ClustalW**<sup>9</sup>.

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?

<sup>4</sup> 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>

<sup>5</sup> As discussed, superficially at least, previously. I hope.

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

<sup>7</sup> 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.

<sup>8</sup> A widely used **java** alignment editor and viewer.

<sup>9</sup> 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- TTFTSLQLEELKVFQKTHYPDVVYREQLALRTELTEARVQVWFQNRRAKWR
ALX4_HUMAN_214_273    --KRRNR- TTFTSYQLEELKVFQKTHYPDVYAREQLAMRDLTEARVQVWFQNRRAKWR
ALX3_HUMAN_153_212    --KRRNR- TTFTFQLEELKVFQKTHYPDVYAREQLALRDLTEARVQVWFQNRRAKWR
ISL1_HUMAN_181_240    --TTRVR- TVLNEKQLHLRTRCYAANPRPDALMKEQLVEMTGLSPRVIRVWFQNKRCCKDK
ISL2_HUMAN_191_250    --TTRVR- TVLNEKQLHLRTRCYAANPRPDALMKEQLVEMTGLSPRVIRVWFQNKRCCKDK
LHX9_HUMAN_267_326    --TKRMR- TSFKHHQLRTMKSIFYAINHNPDADKDLKLAQKTGLTKRVLQVWFQNAKAFR
LHX2_HUMAN_266_325    --TKRMR- TSFKHHQLRTMKSIFYAINHNPDADKDLKLAQKTGLTKRVLQVWFQNAKAFR
LHX8_HUMAN_225_284    --AKRAR- TSFTADQLQVMQAQFAQDNNPDAQTLQKLAERTGLSRRVIQVWFQNCRARHK
LHX6_HUMAN_219_278    --AKRAR- TSFTAEQLQVMQAQFAQDNNPDAQTLQKLAADMTGLSRRVIQVWFQNCRARHK
ZFHX3_HUMAN_2641_2700 --DKRLR- TTITPEQLEILYKQYLSDSNPTRKMLDHIAREVGLKRRVVQVWFQNTRARER
ZFHX4_HUMAN_2560_2619 --DKRLR- TTITPEQLEILYKQYLSDSNPTRKMLDHIAREVGLKRRVVQVWFQNTRARER
ZFHX2_HUMAN_1857_1916 --DKRLR- TTILPEQLEILYRWYMQSDSNPTRKMLDCISEEVGLKRRVVQVWFQNTRARER
ZFHX3_HUMAN_2944_3003 --QRRYR- TQMSSLQKIMKACYEAYRTPMQECEVLGEEIQLPKRVIQVWFQNAKAKEK
ZFHX4_HUMAN_2884_2943 --HKKFR- TQMSNLQKVLKACFSYRTPMQECEMLGNEIQLPKRVIQVWFQNAKAKEK
LHX1A_HUMAN_195_254   --PKRPR- TILTTQRRAFKASFEVSSKPKRVRETAAETGLSRRVVQVWFQNAKAKMK
LHX1B_HUMAN_219_278   --PKRPR- TILTTQRRAFKASFEVSSKPKRVRETAAETGLSRRVVQVWFQNAKAKMK
LHX1_HUMAN_180_239    --RRGPR- TTIKAKQLETLKAAFAATPKPTRHIREQLAQETGLNMRVIQVWFQNRSSKER
LHX5_HUMAN_180_239    --RRGPR- TTIKAKQLETLKAAFAATPKPTRHIREQLAQETGLNMRVIQVWFQNRSSKER
LHX4_HUMAN_157_216    --AKRPR- TTITAKQLETLKNAKYNKSPKPARHVREQLSSETGLDMRVVQVWFQNRRAKEK
LHX3_HUMAN_157_216    --AKRPR- TTITAKQLETLKSAYNTPKPARHVREQLSSETGLDMRVVQVWFQNRRAKEK
      :      :      :      :      :
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---
NANGN_HUMAN_102_161  KEM---
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---
    
```

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.

```

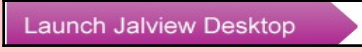
ARX_HUMAN_328_387      --QRRYR- TTFTSYQLEELERAFQKTHYPDVFTREELAMRLDLTEARVQVWFQNRRAKWR
ALX1_HUMAN_132_191    --KRRHR- TTFTSLQLEELKVFQKTHYPDVVYREQLALRTELTEARVQVWFQNRRAKWR
ALX4_HUMAN_214_273    --KRRNR- TTFTSYQLEELKVFQKTHYPDVYAREQLAMRDLTEARVQVWFQNRRAKWR
ALX3_HUMAN_153_212    --KRRNR- TTFTFQLEELKVFQKTHYPDVYAREQLALRDLTEARVQVWFQNRRAKWR
ISL1_HUMAN_181_240    --TTRVR- TVLNEKQLHLRTRCYAANPRPDALMKEQLVEMTGLSPRVIRVWFQNKRCCKDK
ISL2_HUMAN_191_250    --TTRVR- TVLNEKQLHLRTRCYAANPRPDALMKEQLVEMTGLSPRVIRVWFQNKRCCKDK
LHX9_HUMAN_267_326    --TKRMR- TSFKHHQLRTMKSIFYAINHNPDADKDLKLAQKTGLTKRVLQVWFQNAKAFR
LHX2_HUMAN_266_325    --TKRMR- TSFKHHQLRTMKSIFYAINHNPDADKDLKLAQKTGLTKRVLQVWFQNAKAFR
LHX8_HUMAN_225_284    --AKRAR- TSFTADQLQVMQAQFAQDNNPDAQTLQKLAERTGLSRRVIQVWFQNCRARHK
LHX6_HUMAN_219_278    --AKRAR- TSFTAEQLQVMQAQFAQDNNPDAQTLQKLAADMTGLSRRVIQVWFQNCRARHK
ZFHX3_HUMAN_2641_2700 --DKRLR- TTITPEQLEILYKQYLSDSNPTRKMLDHIAREVGLKRRVVQVWFQNTRARER
ZFHX4_HUMAN_2560_2619 --DKRLR- TTITPEQLEILYKQYLSDSNPTRKMLDHIAREVGLKRRVVQVWFQNTRARER
ZFHX2_HUMAN_1857_1916 --DKRLR- TTILPEQLEILYRWYMQSDSNPTRKMLDCISEEVGLKRRVVQVWFQNTRARER
ZFHX3_HUMAN_2944_3003 --QRRYR- TQMSSLQKIMKACYEAYRTPMQECEVLGEEIQLPKRVIQVWFQNAKAKEK
ZFHX4_HUMAN_2884_2943 --HKKFR- TQMSNLQKVLKACFSYRTPMQECEMLGNEIQLPKRVIQVWFQNAKAKEK
LHX1A_HUMAN_195_254   --PKRPR- TILTTQRRAFKASFEVSSKPKRVRETAAETGLSRRVVQVWFQNAKAKMK
LHX1B_HUMAN_219_278   --PKRPR- TILTTQRRAFKASFEVSSKPKRVRETAAETGLSRRVVQVWFQNAKAKMK
LHX1_HUMAN_180_239    --RRGPR- TTIKAKQLETLKAAFAATPKPTRHIREQLAQETGLNMRVIQVWFQNRSSKER
LHX5_HUMAN_180_239    --RRGPR- TTIKAKQLETLKAAFAATPKPTRHIREQLAQETGLNMRVIQVWFQNRSSKER
LHX4_HUMAN_157_216    --AKRPR- TTITAKQLETLKNAKYNKSPKPARHVREQLSSETGLDMRVVQVWFQNRRAKEK
LHX3_HUMAN_157_216    --AKRPR- TTITAKQLETLKSAYNTPKPARHVREQLSSETGLDMRVVQVWFQNRRAKEK
      :      :      :      :      :
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---
NANGN_HUMAN_102_161  KEM---
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---
    
```

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  link at the top of the page. And agree with all the many questions you will be asked.

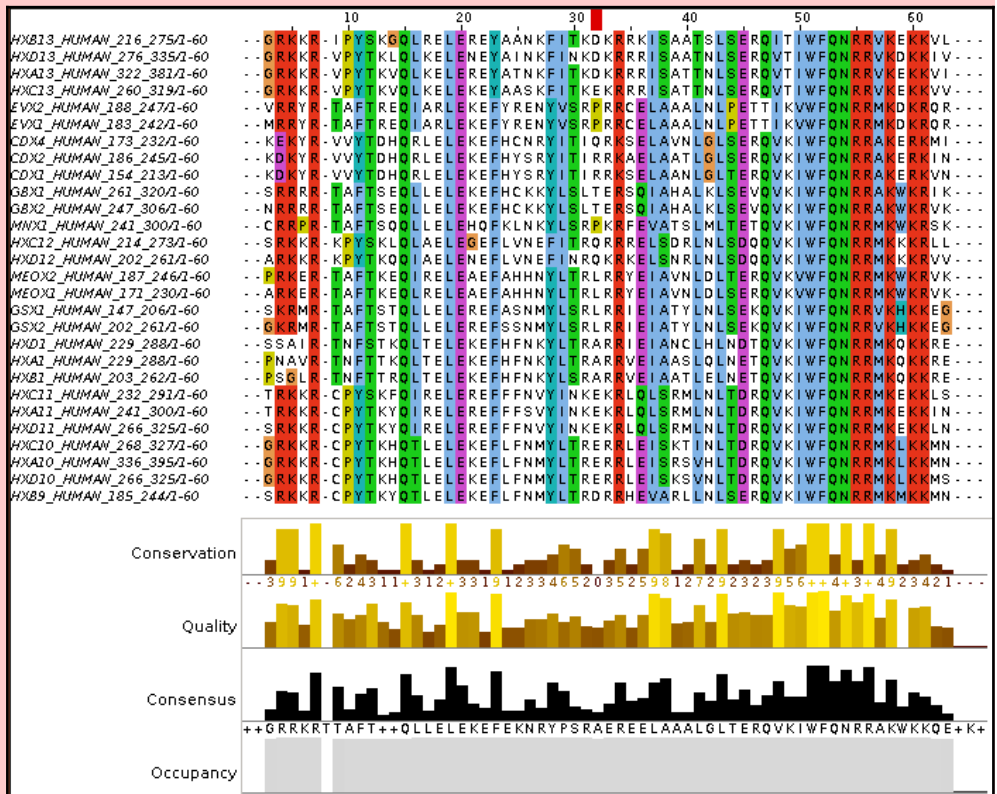
Close down all the example outputs **Jalview** sees fit to show you on start up. From the **File** pull down menu choose from **File** from the **Input Alignment** option. Locate and load the file:

`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.

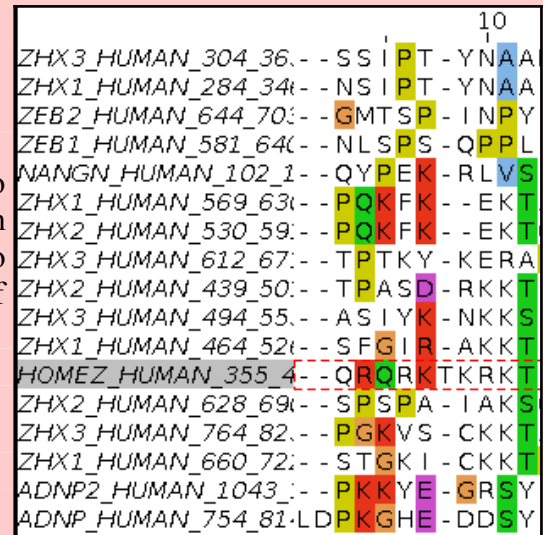
A central purpose of **Jalview** is to allow users to edit alignments as well as just to view them. For example, hold down the **Shift** key, click and hold on any amino acid at the edge of a gap, slide left and right and see that you can introduce and/or alter the position of gaps. It is very important 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

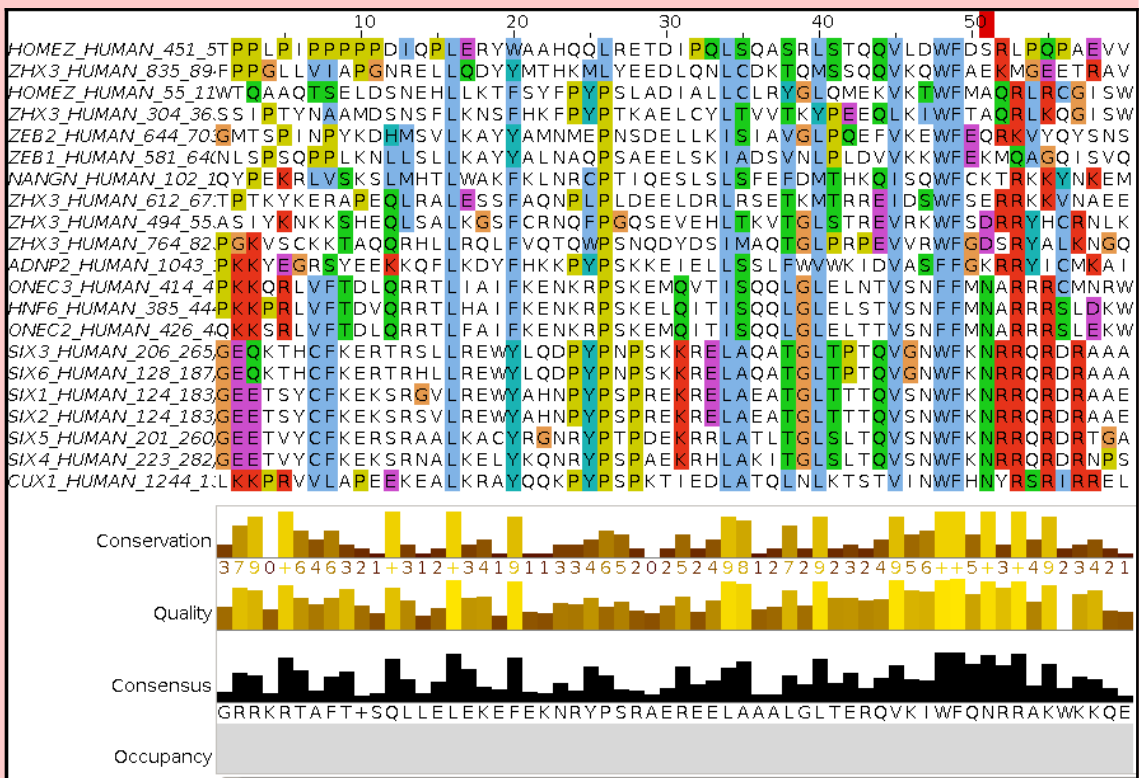


algorithm or more appropriate choice of parameters? Impossible to discuss further as the parameters used for MUSCLE are not revealed.

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. To be fair, one could argue that the web site should only set out to provide draft answers? Maybe the, relatively few users that need/desire full control should expect to download the software, read the manual and do things the hard way?

I am not sure I am sufficiently convinced, particularly when faced with pull down menus with one option and the chance to create data files I cannot use. Make your own mind up.

## 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 pertenant 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 **MSAs** 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 control 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 function. 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).

**DPJ – 2019.02.08**

**Discussion Points and Casual Questions arising from the Instructions Text.****Notes:*****Work in progress I fear.***

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

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

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

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

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



Discussion of the way **ClustalX** (and similar multiple alignment tools) work.

...

Explanation of **clustalX FAST/APPROXIMATE** parameters.

...

Explanation of **clustalX Global Alignment** parameters.

...

The interpretation of the **Homeobox Prosite Pattern?**

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

After reference to the **Quick Hint** mentioned in the text, the boring answer (taking each element in turn, after removing the optional “-” signs) is:

Pattern Position	Pattern Element	Interpretation
1	[LIVMFYG]	Any of the bracketed amino acid codes are acceptable
2	[ASLVR]	Any of the bracketed amino acid codes are acceptable
3	x (2)	Any amino acid is acceptable in the next 2 position
4	[LIVMSTACN]	Any of the bracketed amino acid codes are acceptable
5	x	Any amino acid is acceptable in this position
6	[LIVM]	Any of the bracketed amino acid codes are acceptable
7	{Y}	Any amino acid <i>EXCEPT</i> Y (Tyrosine) is acceptable in this position
8	x (2)	Any amino acid is acceptable in the next 2 position
9	{L}	Any amino acid <i>EXCEPT</i> L (Leucine) is acceptable in this position
10	[LIV]	Any of the bracketed amino acid codes are acceptable
11	[RKNQESTAIY]	Any of the bracketed amino acid codes are acceptable
12	[LIVFSTNKH]	Any of the bracketed amino acid codes are acceptable
13	W	The <i>ONLY</i> acceptable amino acid code in this position is a W (Tryptophan)
14	[FYVC]	Any of the bracketed amino acid codes are acceptable
15	x	Any amino acid is acceptable in this position
16	[NDQTAH]	Any of the bracketed amino acid codes are acceptable
17	x (5)	Any amino acid is acceptable in the next 5 position
18	[RKNAIMW]	Any of the bracketed amino acid codes are acceptable

Note the lack of flexibility of these patterns. An amino acid code is either allowed or not. No reflection of relative frequency of residues in the region of **MSA** from which they are designed (typically by hand).

Note that this particular pattern, though long, is too weak for **Interpro** take take very seriously. As discussed earlier, **Interpro** records a “**Conserved site**” when a match is discovered with this pattern. It is not considered strong enough, by itself, to indicate a **Homeobox** domain.

To examine a few more features of **Prosite**, particularly the very wide degree of relevance to be associated with matches with the patterns, I include a quick exercise to compare all of **Prosite** with the **Human PAX6** protein. In this exercise **protein sequence motifs** and **protein domains** will be sought using just **Prosite** and its associated searching software.

Please do not use class time to go through this. I would hope to discuss the issues briefly anyway. The full instructions are really for people who are going through the exercises by themselves.



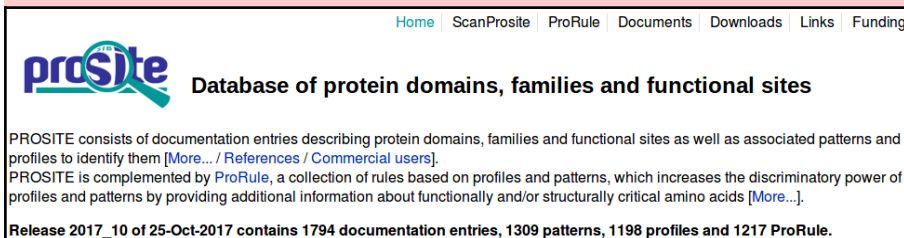
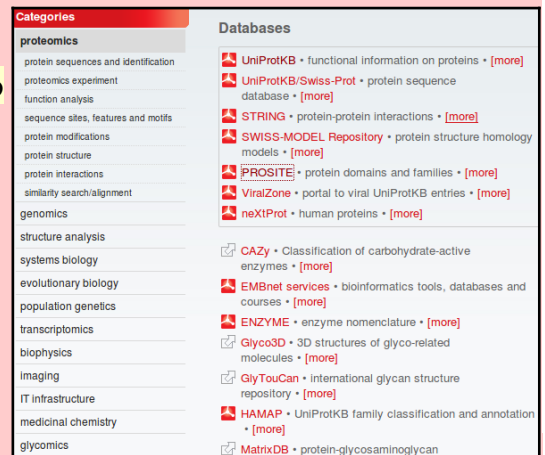
A major database for both motifs and domains is **PROSITE**. Sequence motifs include examples that are extremely simple, and short. These represent such common phenomena possible sites for post-translational modifications (e.g. **glycosylation** or **phosphorylation**). Motifs are generally represented by “Patterns” of characters adhering to some very **trivial rules**.

For a swift experience of using **Prosite**, try the following. **Go to** the **ExPASy**<sup>10</sup> site at:

<http://www.expasy.org>

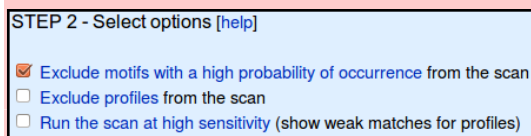
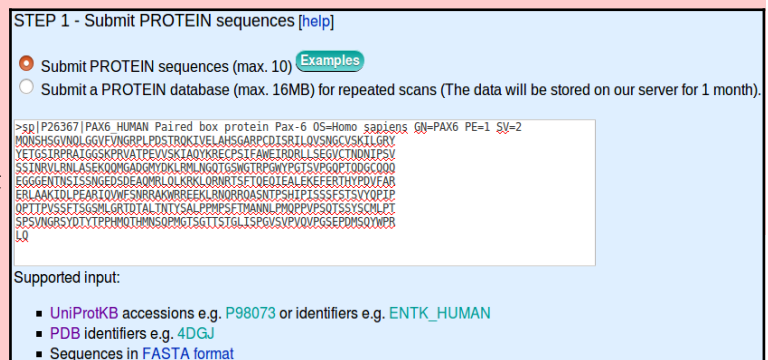
Select **proteomics** from the list of **Categories**.

Select **PROSITE** from the **Databases** section.



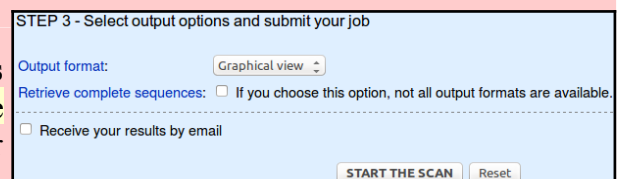
Click on the **ScanProsite** link at the top of your page.

Enter **pax6\_human** in the **STEP 1 - Submit PROTEIN sequences** section.



In the **STEP 2 - Select options** section, ensure that the **Exclude motifs with a high probability of occurrence** box is ticked.

The defaults offered in the **STEP 3 - Select output options and submit your job** section are fine so just click on the **START THE SCAN** button. In but a few moments, your results will burst forth.



<sup>10</sup> **Expasy** is a major site for protein based research in Switzerland. As the all knowing **Wikipedia** puts it: “**ExPASy** is a **bioinformatics** resource portal operated by the **Swiss Institute of Bioinformatics (SIB)** and in particular the **SIB Web Team**. It is an extensible and integrative portal accessing many scientific resources, databases and software tools in different areas of life sciences. Scientists can access a wide range of resources in many different domains, such as **proteomics**, **genomics**, **phylogeny/evolution**, **systems biology**, **population genetics**, and **transcriptomics**.”

Two hits with **PROSITE** profiles suggesting the familiar domains in their familiar places.

**hits by profiles: [2 hits (by 2 distinct profiles) on 1 sequence]**

Upper case represents match positions, lower case insert positions, and the '-' symbol represents deletions relative to the matching profile.

**ruler:** 1 100 200 300 400 500 600 700 800 900 1000

sp-P26367-PAX6\_HUMAN (sp-P26367-PAX6\_HUMAN) (422 aa)

**PS51057 PAIRED\_2 Paired domain profile :**

**4 - 130:** score = 64.941  
 SHSGVNLGGFVNGRPLPDSTRQKIVELAHSGARPCDISRILOVSNQCVSKILGRYYET  
 GSIRPRATGGSKPRVATPEVSKIAQYKRECPISFAMEIRDLLSEGVCTNDNIPSVSSI  
 NRVLRNL

**Predicted feature:**  
 DOMAIN 4 130 Paired [condition: none]

**PS50071 HOMEBOX\_2 'Homeobox' domain profile :**

**208 - 268:** score = 20.164  
 RKLQRNRTSFTQEQIEALEKEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWR  
 E

Two hits with **PROSITE** patterns confirm the same domains by matching highly conserved subregions.

**hits by patterns: [2 hits (by 2 distinct patterns) on 1 sequence]**

**ruler:** 1 100 200 300 400 500 600 700 800 900 1000

sp-P26367-PAX6\_HUMAN (sp-P26367-PAX6\_HUMAN) (422 aa)

**PS00034 PAIRED\_1 Paired domain signature :**

**38 - 54:** [confidence level: (0)] RPCdisrilqvsngCVS

**PS00027 HOMEBOX\_1 'Homeobox' domain signature :**

**243 - 266:** [confidence level: (0)] LAakIdLPeaRIQVWfSNrrakwR

This confirms what has already been discovered more than once, by reading database annotations, by running **Interpro** and by running other individual database search program(s) manually.

Note that **Prosite** is happy to accept the **HOMEBOX** pattern hit as sufficient to predict the presence of a **Homeobox domain**. **Interpro** regards exactly the same evidence to register only a “**Homeobox conserved site**”. I suspect the caution of **Interpro** is justified.

Move back to the search submission. In **Step 2**, deselect **Exclude patterns with a high probability of occurrence**. **START THE SCAN**.

**hits by patterns with a high probability of occurrence or by user-defined patterns: [19 hits (by 5 distinct patterns) on 1 sequence]**

**ruler:** 1 100 200 300 400 500 600 700 800 900 1000

sp-P26367-PAX6\_HUMAN (sp-P26367-PAX6\_HUMAN) (422 aa)

PS00008 MYRISTYL N-myristoylation site :
13 - 18: GVfvNG
36 - 41: GArpCD
110 - 115: GVctND
151 - 156: GQtgSW
154 - 159: GSwgTR
157 - 162: GTrpGW
182 - 187: GGgeNT
183 - 188: GGgeTN
312 - 317: GSmLGR
387 - 392: GTsgTT
390 - 395: GTtsTG

This time you will see many more hits with very short patterns.

Follow the link to the documentation for an **N-myristoylation site (PS00008)**.

See that the pattern is just **6** positions wide. **2** of those positions can be any amino acid. Only one position is fully specified. Not too demanding on the whole. I would expect this to match most proteins of any size and not always because there was an **N-myristoylation site**.

MYRISTYL, PS00008; N-myristoylation site (PATTERN with a high probability of occurrence!)

- Consensus pattern:  
 G-[EDRKHPFYW]-x(2)-[STAGCN]-[P][GisteN-myristoylation site]

The pattern is explained in the database thus.

- The N-terminal residue must be glycine.
- In position 2, uncharged residues are allowed. Charged residues, proline and large hydrophobic residues are not allowed.
- In positions 3 and 4, most, if not all, residues are allowed.
- In position 5, small uncharged residues are allowed (Ala, Ser, Thr, Cys, Asn and Gly). Serine is favored.
- In position 6, proline is not allowed.

The description is not entirely an honest reflection of the information to which the scanning software will respond. The software is given to understand that **ANY** amino acid can occur in positions **3** and **4**. The software has no way to know that “**Serine** is favoured” in position **5**! Maybe you think that my pointing out these transparent truths makes me an intolerable pedant? Well ... so is the computer!

**PROSITE** predicts **11 N-myristoylation** sites in the **Human PAX6** protein. A site every **40** amino acids or so. Without considerable further effort, it is not really possible to suggest how many of these predictions might be “real”. The evidence of this exercise alone is most certainly insufficient. Intuitively, I would expect a large number of false positives from as weakly specified motif as this one. It has been suggested (**May of 2011**) of this **PROSITE** pattern, by researchers looking at more sophisticated detection methods, that:

“**PS00008** of **PROSITE** constructed from a small dataset ... produces a great number of not only false positive but false negative predictions.”

This is good enough to believe the majority of these predictions to be unreliable. It is not good enough for me to hazard a meaningful guess as to how many real sites would be expected in this particular protein.

Consider for a few moments the **Prosite Paired Box** pattern, **R-P-C-x(11)-C-V-S**, specifically its location within the **Paired Box** domain.

At the top of your **ScanProsite Results** page, you will find the **canonical** version of **PAX6 Human** displayed. If you hover over the graphic indicating the position of the **Profile** match for the **Paired Box**, the position of the whole **Paired Box** domain will be highlighted. If you hover over the graphic for the **Pattern** match for **Paired box**, the position of the pattern will be illustrated.

My illustration is of these two views superimposed on each other and prettied up a trifle.

```

M Q N S H S G V N Q L G G V F V N G R P L P D S T R Q K I V E L A H S G A R P C D I S R I L Q V S N G C V S K I L G R Y Y E T G S I
R P R A I G G S K P R V A T P E V V S K I A Q Y K R E C P S I F A W E I R D R L L S E G V C T N D N I P S V S S I N R V L R N L A S
E K Q Q M G A D G M Y D K L R M L N G Q T G S W G T R P G W Y P G T S V P G Q P T Q D G C Q Q Q E G G G E N T N S I S S N G E D S D
E A Q M R L Q L K R K L Q R N R T S F T Q E Q I E A L E K E F E R T H Y P D V F A R E R L A A K I D L P E A R I Q V W F S N R R A K
W R R E E K L R N Q R R Q A S N T P S H I P I S S S F S T S V Y Q P I P Q P T P V S S F T S G S M L G R T D T A L T N T Y S A L P
P M P S F T M A N N L P M Q P P V P S Q T S S Y S C M L P T S P S V N G R S Y D T Y T P P H M Q T H M N S Q P M G T S G T T S T G L
I S P G V S V P V Q V P G S E P D M S Q Y W P R L Q
    
```

The pattern **RPCXXXXXXXXXXCVS** within the entire domain is clear.

If you were to repeat this whole exercise with the **isoform 5a** version of **Human PAX6** (*please do not!*), the equivalent picture would be as illustrated.

```

M Q N S H S G V N Q L G G V F V N G R P L P D S T R Q K I V E L A H S G A R P C D I S R I L Q T H A D A K V Q V L D N Q N V S N G C
V S K I L G R Y Y E T G S I R P R A I G G S K P R V A T P E V V S K I A Q Y K R E C P S I F A W E I R D R L L S E G V C T N D N I P
S V S S I N R V L R N L A S E K Q Q M G A D G M Y D K L R M L N G Q T G S W G T R P G W Y P G T S V P G Q P T Q D G C Q Q Q E G G G
E N T N S I S S N G E D S D E A Q M R L Q L K R K L Q R N R T S F T Q E Q I E A L E K E F E R T H Y P D V F A R E R L A A K I D L P
E A R I Q V W F S N R R A K W R R E E K L R N Q R R Q A S N T P S H I P I S S S F S T S V Y Q P I P Q P T P V S S F T S G S M L G
R T D T A L T N T Y S A L P P M P S F T M A N N L P M Q P P V P S Q T S S Y S C M L P T S P S V N G R S Y D T Y T P P H M Q T H M N
S Q P M G T S G T T S T G L I S P G V S V P V Q V P G S E P D M S Q Y W P R L Q
    
```

The **14** amino acid insertion of **isoform 5a** (**THADAKVQVLDNQN**, corresponding to the entire **3<sup>rd</sup>** coding exon being spliced into the **mRNA**) has landed right in the middle of the pattern! It surely cannot match as intended when used with an **isoform 5a PAX** domain.

PAIRED\_1, PS00034; Paired domain signature (PATTERN)

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

From your **ScanProsite Results** page, follow the link to the documentation for this pattern (**PS00034**). Find and read the description of the pattern where it is claimed that the pattern matches all **58** true **Paired Boxes** in **SwissProt**<sup>11</sup>.

This is bold claim can only be true if *none* the **PAX** domains in **Swissprot** are **isoform 5a** domains. Unsurprisingly, this is the case. All **PAX** proteins are recorded in **Swissprot** in their “**canonical form**”. **Isoform 5a** variants are always only acknowledged in the annotation as “**Features**”. **ScanProsite** is not clever enough to search all variations of a **Swissprot** entry. It just searches the main **canonical** sequence. Yes, it finds all **58 canonical SwissProt PAX** proteins, but it would not find any **isoform 5a PAX** proteins if they were stored as separate entries in **SwissProt** (or input to **ScanProsite** as an independent protein sequence). The **PAX Prosite Pattern** is not as effective as its documentation claims.

<sup>11</sup> This claim is, of course, only possible if all the annotation of **SwissProt** is assumed to be **100%** accurate.

In order to detect just the **PAX isoform 5a**, the pattern would have to be:

**R-P-C-x(25)-C-V-S**

To detect both isoforms, using just one pattern:

**R-P-C-x(11,25)-C-V-S**

would work, but would be insufficiently specific and would generate far too many false positives. These sort of patterns are useful, but only with caution. They are valuable because of their simplicity, but they are very fragile.

In the **Prosite Paired domain documentation page**, just below the **Pattern** description, is the **Profile** description.

<ul style="list-style-type: none"> <li>• Sequences in UniProtKB/Swiss-Prot known to belong to this class: 58             <ul style="list-style-type: none"> <li>◦ detected by PS51057: 58 (true positives)</li> <li>◦ undetected by PS51057: 0 (false negative or 'partial')</li> </ul> </li> <li>• Other sequence(s) in UniProtKB/Swiss-Prot detected by PS51057: NONE.</li> </ul>
---

The claim here is also to find all the **58 PAX** domains in **SwissProt**. This time, with **0** false positives (the **Pattern** had to admit to **7**). A clear but small improvement, but, the real superiority of the **Profile** over the **Pattern** is that it allows enough flexibility to find **Paired** boxes that have the relatively large **14** amino acid **isoform 5a** insertion. The documentation cannot boast that this is true as there are no instances in SwissProt to allow the case to be proven. However, it is true ... because I say so!

This flexibility of the probabilistic approach employed by **pHMMs** was also illustrated when we glanced at **PFAM**. The **PFAM pHMM** for **PAX** was computed from a **5** sequence alignment including no representation of any **isoform 5a** sequence, yet it too will match **isoform 5a PAX domains**.

Comments on **Jalview** as an alignment viewer/editor in various contexts (e.g. **Pfam** and **Jpred**).

**Jalview** has appeared in the exercises twice already. Not however, in particularly high profile sections, so you might have yet to be introduced formally. Here I attempt brief correction of any inappropriate informality.

Alignment algorithms are crude.

## DPJ – 2019.02.08

References for further extension:

[https://en.wikipedia.org/wiki/Multiple\\_sequence\\_alignment](https://en.wikipedia.org/wiki/Multiple_sequence_alignment)