



Earlham
Institute

Earlham Institute summer school on bioinformatics

25-29 July 2016

Basic Bioinformatics Sessions

Practical 4: Multiple Sequence Alignment

Wednesday 6 July 2016

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**.

The screenshot shows the Uniprot search criteria form. It includes fields for Organism [OS] (Human [9606]), AND Reviewed (Yes), AND Function (DNA binding), Term (Homeobox), Length range (50-70), and Evidence (Any assertion method). There are buttons for adding, removing, and searching.

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
P52952	NKX25_HUMAN	Homeobox protein Nkx-2.5	NKX2-5 CSX,NKX2.5,NKX2E	Homo sapiens (Human)	324
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
P49639	HXA1_HUMAN	Homeobox protein Hox-A1	HOXA1 HOX1F	Homo sapiens (Human)	335
Q92786	PROX1_HUMAN	Prospero homeobox protein 1	PROX1	Homo sapiens (Human)	737
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
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 13-APR-2016, entry version 170.
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 **Feature Tables** and most often refer to a **Homeobox** region.

In the **DNA_BIND** **Feature Table** entries, the position of the **Homeoboxes** are recorded and will be used by the next program to isolate the sequence of the **Homeoboxes**.

```
FT CHAIN 1 430 Pre-B-cell leukemia transcription factor
FT 2.
FT /FTId=PRO_0000049237.
FT DNA_BIND 244 306 Homeobox; TALE-type.
FT {ECO:0000255|PROSITE-ProRule:PRU00108}.
FT COMPBIAS 137 145 Poly-Ala.
FT MOD_RES 136 136 Phosphoserine.
FT {ECO:0000269|PubMed:24275569}.
FT MOD_RES 151 151 Phosphoserine.
FT {ECO:0000269|PubMed:18220336,
FT ECO:0000269|PubMed:18669648,
FT ECO:0000269|PubMed:20068231}.
FT MOD_RES 330 330 Phosphoserine.
FT {ECO:0000269|PubMed:18669648,
FT ECO:0000269|PubMed:21406692}.
FT CONFLICT 393 393 M -> I (in Ref. 1; CAA42503).
FT {ECO:0000305}.
SQ SEQUENCE 430 AA; 45881 MW; EF2FFA158C4DAF68 CRC64;
MDERLLGPPP PGGGRGGLGL VSGEPGGPGE PGGGDPGGG SGGVPGGRGK QDIGDILQQI
MTITDQSLDE AQAKKHALNC HRMKPALFSV LCEIKEKTGL SIRSSQEEEP VDPQLMRLDN
MLLAEGVAGP EKGGSAAAAA AAAASGGGV SPDNSIEHSD YRSKLAQIRH IYHSELEKYE
QACNEFTTHV MNLLREQSRT RPVAPKEMER MVSIIHRKFS AIQMOLKQST CEAVMILRSR
FLDARKRRRN FSKQATEVLN EYFYSHLSNP YPSEAKEEL AKKCGITVSQ VSNWFGNKRI
RYKKNIGKFP EEANIYAVKT AVSVTQGGHS RTSSPTPPSS AGSGGSFNLS GSGDMFLGMP
GLNGDSYSAS QVESLRHSMG PGGYGDNLGG GQMYSPPREMR ANGWSQEAVT PSSVTSPTGEG
PGSVHSDTSN
//
```

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

<http://emboss.bioinformatics.nl/>

EDIT
[aligncopy](#)
[aligncopypair](#)
[biosed](#)
[codcopy](#)
[cutseq](#)
[degapseq](#)
[descseq](#)
[entrel](#)
[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.

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_home...oteins.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 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 KQQRYSAPD RDQLASVLKL 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;
LQNRNRSFTQ EQIEALEKEF ERTHYPDVFA RERLAADIDL 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 QQLQLEATF QRRNRYPMST REEIAVWTLN 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 SHQKYLAPD RAHLAKNLKL TETQVKIWFQ NRRYKTKRKQ
//
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
//

```

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.

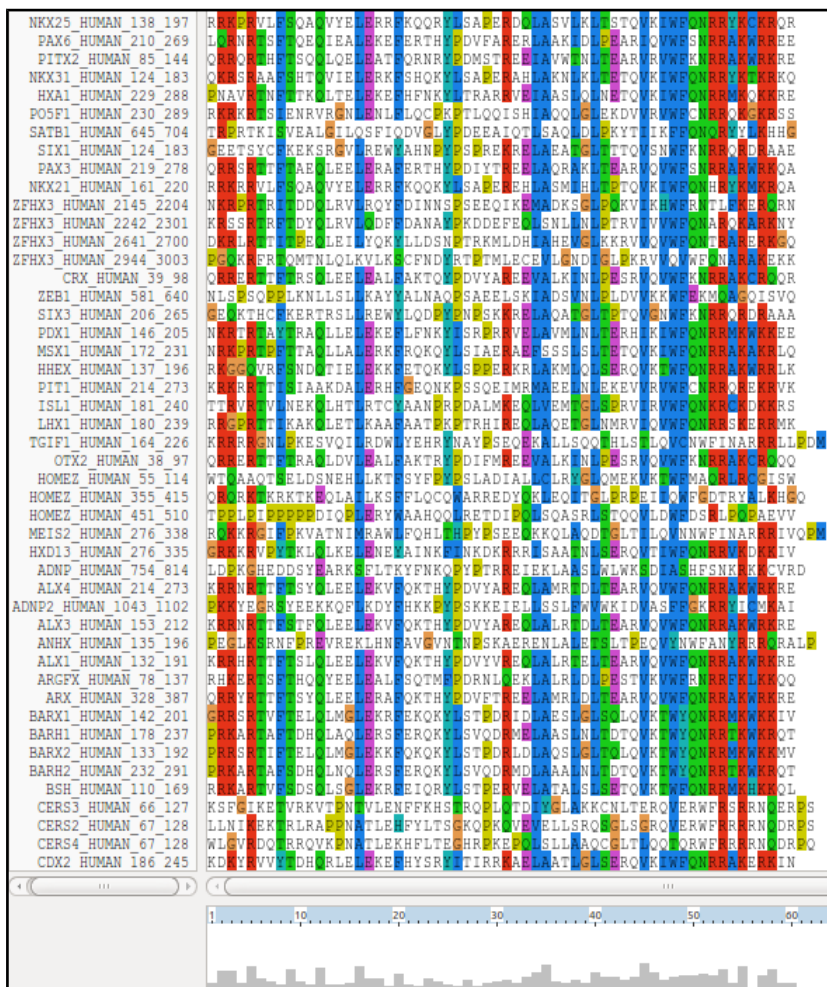
ClustalX is a part of the mostly widely known family of **Multiple Sequence Alignments (MSA)** programs, originating in the **1980s**. Until relatively recently, it was the only real option. **ClustalX** still has merit, although it lacks some of the sophistication of more recent programs. **ClustalX** runs on effectively all workstations and has a nice **Graphical User Interface (GUI)**. A good place for us to start. It is installed on your workstations.

Start up the program **ClustalX**¹. 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.

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** tab².

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 both at the **EBI** and the Bioinformatics server at **Wageningen**. Try it if you have time. You get the same results but will, sadly, lose the pretty interface.

<http://www.ebi.ac.uk/Tools/clustalw2/index.html>

<http://www.bioinformatics.nl/tools/clustalw.html>

² The **Fast-Approximate** algorithm is essential that which the database searching program **fasta** employs. Assuming we have discussed how **fasta** works, it should require no further explanation here.

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

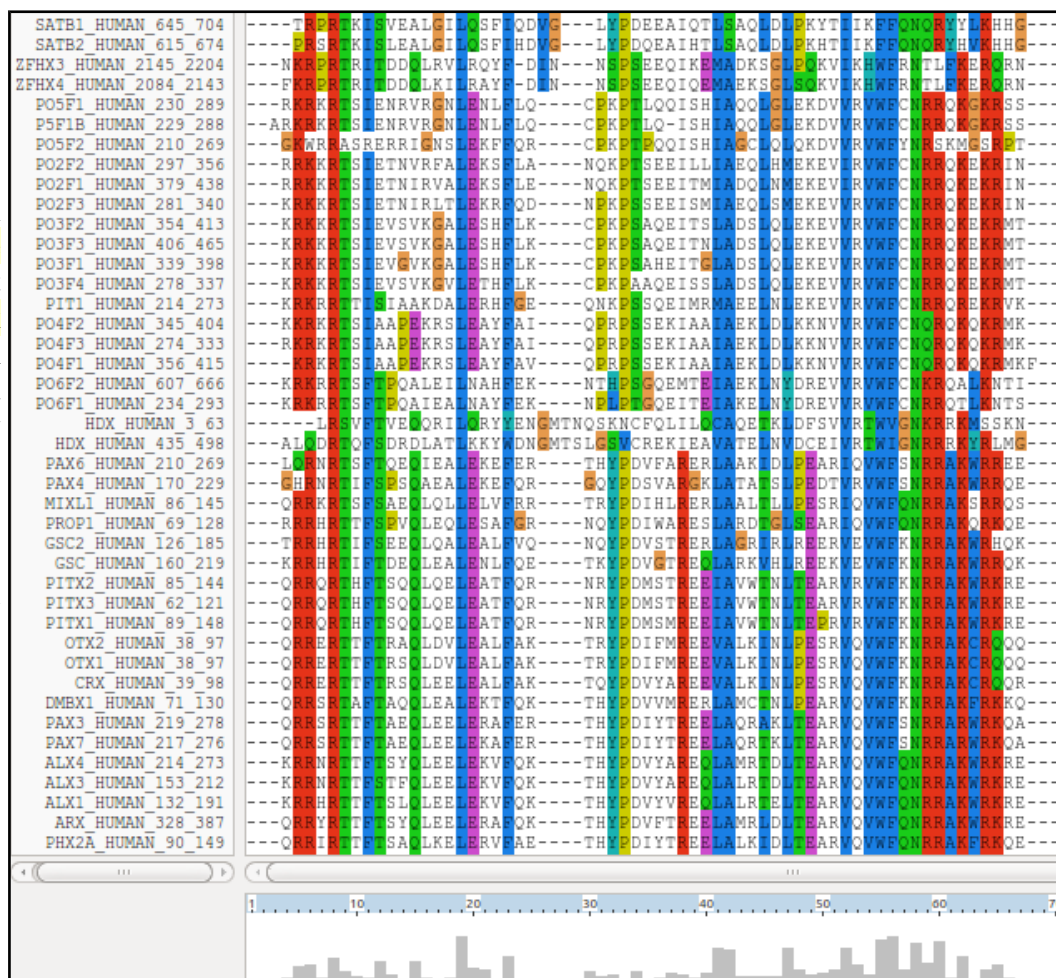
I will assume both sets of parameters at least seem familiar? If not please ask. The default **Slow/Accurate Pairwise Parameters** you now have in view are fine. Click the **OK** button to dismiss the **Pairwise Parameters** window.

Before proceeding, save the **homeobox** sequences in **FASTA** format, which will better suit the other **MSA** programs we will try. Do this by selecting **Save sequences as...** from the **File** pull down menu. Deselect **CLUSTAL** format, select **FASTA** format. Click **OK**. A file called **homeobox_human.fasta** will be created. Take a look to check it is as you would expect.

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

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

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



PO5F1 HUMAN 230 289 --RRKKRRTSLENVRVGNLENLQCCKPPQLQOISHIAQQLGLEKDVVRVWFCRRRQKGRSS--
P5F1B HUMAN 229 288 --ARKKRRTSLENVRVGNLENLQCCKPPQLQ-ISHIAQQLGLEKDVVRVWFCRRRQKGRSS--
PO5F2 HUMAN 210 269 --GKWRKASRRRIINSLEKFFQRCCKPQLQOISHIAQLQCKQKDVVRVWFFYWRKMSRGT--
PO2F2 HUMAN 297 356 --RRKKRTSLETNVRFALESFLANQKPPSEEEILLIAEQLMHEKEVVRVWFCRRRQKGRKN--
PO2F1 HUMAN 379 438 --RRKKRTSETNIRVALESFLDNQKPPSEETIMADQLNWKEVVRVWFCRRRQKGRKN--
PO2F3 HUMAN 281 340 --RRKKRTSETNIRLTLEKRFDNKPPSEEEISMLIAEQLSWEKEVVRVWFCRRRQKGRKN--
PO3F2 HUMAN 354 413 --KKKKRTSEVSVKGALESHFLKCKPPSAQEITSLADSLQLEKEVVRVWFCRRRQKGRMT--
PO3F1 HUMAN 406 465 --KKKKRTSEVSVKGALESHFLKCKPPSAQEITSLADSLQLEKEVVRVWFCRRRQKGRMT--
PO3F3 HUMAN 339 398 --KKKKRTSEVSVKGALESHFLKCKPPSAHEITSLADSLQLEKEVVRVWFCRRRQKGRMT--
PO3F4 HUMAN 278 337 --KKKKRTSEVSVKGVLETHFLKCKPPAAQEISSLADSLQLEKEVVRVWFCRRRQKGRMT--
P1T1 HUMAN 214 273 --KKKKRTTSLIAAKDALERHGEONKPPSQEIMMAEELNLEKEVVRVWFCRRRQKGRVK--
PO4F2 HUMAN 345 404 --KKKKRTSLAAPKRSLEAYIAIQRRPSEKIAAAIEKLDLKKNVVRVWFCRRRQKGRMK--
PO4F3 HUMAN 274 333 --RRKKRTSLAAPKRSLEAYIAIQRRPSEKIAAAIEKLDLKKNVVRVWFCRRRQKGRMK--
PO4F1 HUMAN 356 415 --KKKKRTSLAAPKRSLEAYIAVQRRPSEKIAAAIEKLDLKKNVVRVWFCRRRQKGRMKF--
PO6F2 HUMAN 607 666 --KKKKRTTSPQALEIINAHSEKNTHPSQEMTEIAEKNVNDREVVRVWFCRRRQKGRNTI--
PO6F1 HUMAN 234 293 --KKKKRTTSPQALEIANAYSEKNLPSQETITEIAKELNVDREVVRVWFCRRRQKGRNTS--
PAX6 HUMAN 210 269 --LRNNRTSPQEOLEALEKEKERTGPDVFAELAKTSLPEARTVWFSRRRAKRRREE--
PAX4 HUMAN 170 229 --GRRNRTSPSSQAEALEKEKERTGPDVSVAGKLATSLSPEDTVRWFSRRRAKRRRQE--
MIXL1 HUMAN 86 145 --QRRKRTSFAEDQLQLLELVRRRTRYPDHLEERLALTLLESGRIQVWFQNRRAKRRRQS--
PROPI HUMAN 69 128 --RRRHRTSPSSQLEQLESAGRNQYPDVWAESELRDTELSSEARQVWFQNRRAKRRRQE--
GSC2 HUMAN 126 185 --TRRHRTSPSEEDQLQALEALVVQNYQPDVSTERLAQRRLRERERVEWFKRRRAKRRRQE--
GSC HUMAN 160 219 --KRRHRTSPDEOLEALENLQETKYPDVGTRELRKVHUREEKVEWFKRRRAKRRRQK--
PITX2 HUMAN 85 144 --QRRRTHTPSSQQLQELEATPQRNRYPDMSTREELIAVWNLTEARVQWFKRRRAKRRRRE--
PITX3 HUMAN 62 121 --QRRRTHTPSSQQLQELEATPQRNRYPDMSTREELIAVWNLTEARVQWFKRRRAKRRRRE--
PITX1 HUMAN 89 148 --QRRRTHTPSSQQLQELEATPQRNRYPDMSMREELIAVWNLTEARVQWFKRRRAKRRRRE--
OTX2 HUMAN 38 97 --QRRRTHTPRAOLDVLEALPAKTRYPDIFMREELVALKINLPESRVQWFKRRRAKRRRQK--
OTX1 HUMAN 38 97 --QRRRTHTPERSLDVLEALPAKTRYPDIFMREELVALKINLPESRVQWFKRRRAKRRRQK--
CRX HUMAN 39 98 --QRRRTHTPERSLEELEALPAKQYPDVYAREELVALKINLPESRVQWFKRRRAKRRRQK--
DMBX1 HUMAN 71 130 --QRRSRTFAAQOLEALEKTKTHYPDVVMERLAMCNLEARVQWFKRRRAKRRRQK--
PAX3 HUMAN 219 278 --QRRSRTFAAELEELERAFERTHYPDITYREELAQRAKLTEARVQWFSRRRAKRRRQA--
PAX7 HUMAN 217 276 --QRRSRTFAEOLLEELKAFERTHYPDITYREELARQKLTTEARVQWFSRRRAKRRRQA--
ALX4 HUMAN 214 273 --KRRNRTTPSSYLEELEEKVPQKTHYPDVYAREELAMRLDTEARVQWFSRRRAKRRRRE--
ALX3 HUMAN 153 212 --KRRNRTTPSTFLEELEKVPQKTHYPDVYAREELALRLDTEARVQWFSRRRAKRRRRE--
ALX1 HUMAN 132 191 --KRRHRTTPSSYLEELEKVPQKTHYPDVYAREELALRLDTEARVQWFSRRRAKRRRRE--
ARX HUMAN 328 387 --QRRRTHTPSSYLEELEKAFQKTHYPDVFTREELAMRLDTEARVQWFSRRRAKRRRRE--
PHX2A HUMAN 90 149 --QRRINTTPSAOLKELELVPAETHYPDITYREELALKIDLTEARVQWFSRRRAKRRRQE--
PHX2B HUMAN 98 157 --QRRINTTPSAOLKELELVPAETHYPDITYREELALKIDLTEARVQWFSRRRAKRRRQE--
DRGX HUMAN 33 92 --QRRNRTTPSLQOLEALEAVPAQTHYPDVFTREELAMKINLTEARVQWFSRRRAKRRRTE--
SHOX HUMAN 117 176 --QRRSRTNPLEDLNELERLADETHYPDAFMREELSLQRLGLESEARVQWFSRRRAKRRRQE--
SHOX2 HUMAN 140 199 --QRRSRTNPLEDLNELERLADETHYPDAFMREELSLQRLGLESEARVQWFSRRRAKRRRQE--
OTP HUMAN 104 163 --QRRHRTTPSAOLNELERSAKTHYPDIFMREELALRLIGTESRVQWFSRRRAKRRRRE--
VSX2 HUMAN 148 207 --KRRHRTTPSSYLEELEKANEARYPDVYAREELAMKTELEPEDRIQVWFQNRRAKRRRRE--

1 10 20 30 40 50 60

ZHX3 HUMAN 764 823 -- PCKVSCCK LAQRHLRQLFVQTOWPSNQDYDSIMAGLGRPEVVRVFGDSTRALNKGQ---
 ZHX1 HUMAN 660 722 -- STGKI-CKK EEOHMKSAFVRTOWPSEEDYDKLAKESCHARTDIVFNGDSTRAYKNGNLK--
 ZHX2 HUMAN 628 690 -- SSGA-AIAKQEQVHLRSTFARTQWCKQEQYDLAAKGLVURTIVRWEKENRCLLTGTIVK--
 HOMEZ HUMAN 355 415 -- QRPRKTKRKKKEOLAILKSFSLQCCWARRREDYQKLEQITGLRRELIIONFGDTRVALNHGQ---
 ZHX2 HUMAN 439 501 -- TTRASDRKKKEQIAHRKASTLQSQSPDDAEVYRLTEVFGARSEIKKWFSDHRYECCSGIVH--
 ZHX1 HUMAN 464 526 -- SFGILAKKKKEQIAELKVSILKNQPHDSEIIRLMKLTGLKEIKKESDTRYNQNSKSKNC--
 ZHX3 HUMAN 494 553 -- ASIYHNKKHEOLSAIKKSGFCRNQPPGQSEVEHLKVTGLSTREVRKWFSDARXHNKLNK---
 ZHX2 HUMAN 530 591 --- PCKFKEKQCKWIKLSDSLKSSPQAEELDRVRVEKLSRREIDWFSEARRKLRDSMEQK--
 ZHX1 HUMAN 569 630 --- PCKFKEKAEQVRVLCASFINSSVDEELNRRAQTKLRRREIDAWFATNRKSKALKEEK--
 ZHX3 HUMAN 612 671 --- TTKYKERAELSRALSSFAQNLPLDEELDRLRSEKLRREIDWFSEARRK-KVNAAE---
 ZHX3 HUMAN 304 363 --- SSITYNAAMDSSNFKLNSPHKFPVQKAEELCYLVVVKIKVDEELKIKWETARLQGGISW---
 ZHX1 HUMAN 284 346 --- NSITYNAALDN-NLLNLNKNKFPPTMSEITVLSAQKKYEEELKIKWESAQRLHGVSWTBE--
 ZHX2 HUMAN 263 324 --- NTTKYNALDT-NATMINSNKKFPVQAEELSWLAASKPDEHIFRWFAQRLHGGISWSBE--
 HOMEZ HUMAN 55 114 --- WTALAKSGELDS-NHEHLKTSYGFVPSLADIALLCLRYGQMEKFKWEMARLGGISW---
 ADNP HUMAN 754 814 --- DPYGHEDDSYEARKSFLTKYFNKQVPTREIEKLAASLWKKSDIAPHESKRRKQVVD--
 DNPF2 HUMAN 1043 1102 --- PKHYEGRSYEEKQFLDKYHHKKVPSKKEIELSLGWKKKIDAFFEKRRITLMAAI---
 SIX1 HUMAN 124 183 --- GEETSYCEKKSRSVLREWYAHNPVPSREKRELAEATGLTTTVSNWEKRRQRDRAAE--
 SIX2 HUMAN 124 183 --- GEETSYCEKKSRSVLREWYAHNPVPSREKRELAEATGLTTTVSNWEKRRQRDRAAE--
 SIX3 HUMAN 206 265 --- GEOKTHCKERTRSLREWYLQDPVPSNPKKRELAEATGLTPTGVSNWEKRRQRDRAAE--
 SIX6 HUMAN 128 187 --- GEOKTHCKERTRHLREWYLQDPVPSNPKKRELAEATGLTPTGVSNWEKRRQRDRAAE--
 SIX5 HUMAN 201 260 --- GEETVYCEKERSRAALKACVNRNKPPTDEKRRLATILGLSLTVSNWEKRRQRDRDGA--
 SIX4 HUMAN 223 282 --- GEETVYCEKKSRAALKYKQNRVPSAEKRLAKITGLSLTVSNWEKRRQRDRDGA--
 CUX2 HUMAN 1168 1227 --- IKKPRVVLVAESKEAKRLKAYQLEVPSSQOTIELSPQNLKNTNTVINWFHYRSKRRREM--
 CUX1 HUMAN 1244 1303 --- LKKPRVVLVAESKEAKRLKAYQKKVPSQOTIEDLATQNLKNTSTVINWFHYRSKRRREL--
 HNF6 HUMAN 385 444 --- EKKPRLVFVDVRRTHAIFENKRRPSKELITISQQLGELSTVSNFEMARRSLDKW---
 ONEC2 HUMAN 426 485 --- OKKSRLVFVDLRRTHAIFENKRRPSKEMOITISQQLGELTTVSNFEMARRSLLEKW---
 ONEC3 HUMAN 414 473 --- EKKPRLVFVDLRRTHAIFENKRRPSKEMOITISQQLGELTTVSNFEMARRKLNRRW---
 PO5F1 HUMAN 230 289 --- RKKRRSTENVRGNLENLFLQCCKPTLQISHIAQCLQEKDVRVWFCHRRQKGRSS--
 PO5F1B HUMAN 229 288 --- ARKKRRSTENVRGNLENLFLQCCKPTLQISHIAQCLQEKDVRVWFCHRRQKGRSS--
 PO5F2 HUMAN 210 269 --- CKWRASRERRINSLKFFQRCCKPTLQISHIAQCLQEKDVRVWFVFRSKMSSTRT--
 PO2F2 HUMAN 297 356 --- RKKRRSTENVRFALEKFLANQKPTSEIILLIAEQHNEKEVVRWFCHRRQKKEKIN--
 PO2F1 HUMAN 379 438 --- RKKRRSTETNIRVALRSLLENQKPTSEIITMIAQDLNKEKEVVRWFCHRRQKKEKIN--
 PO2F3 HUMAN 281 340 --- RKKRRSTETNIRLTLEKRFQDNKPTSEESIIMIAEQHNEKEVVRWFCHRRQKKEKIN--
 PO3F2 HUMAN 354 413 --- RKKRRSTEVSVKGALESHFLKCKPSAQEITSLADSLQEKVVRVWFCHRRQKKEKMT--
 PO3F3 HUMAN 406 465 --- RKKRRSTEVSVKGALESHFLKCKPSAQEITSLADSLQEKVVRVWFCHRRQKKEKMT--
 PO3F1 HUMAN 339 398 --- RKKRRSTEVKGALESHFLKCKPSAHEITSLADSLQEKVVRVWFCHRRQKKEKMT--
 PO3F4 HUMAN 278 337 --- RKKRRSTEVSVKGVLETHFLKCKPSAQEISSLADSLQEKVVRVWFCHRRQKKEKMT--
 PIT1 HUMAN 214 273 --- KKKRRSTTAAKDALRRHFGQNKPPSQEIMRMAEELNKEKEVVRVWFCHRRQKKEKRVK--
 PO4F2 HUMAN 345 404 --- KKKRRSTIAAPKKRSLAYFAIQRPSSKIAAIAEKLDLKKNNVVRWFCHRRQKKEKMK--
 PO4F3 HUMAN 274 333 --- RKKRRSTIAAPKKRSLAYFAIQRPSSKIAAIAEKLDLKKNNVVRWFCHRRQKKEKMK--
 PO4F1 HUMAN 356 415 --- KKKRRSTIAAPKKRSLAYFAVQRPSSKIAAIAEKLDLKKNNVVRWFCHRRQKKEKMF--
 PO6F2 HUMAN 607 666 --- KKKRRSTIAAPKQAEILNAHFEKNTSPSQEMTIAEKLNLDREVVRVWFCHRRQKKEKNTI--

ZHX3 HUMAN 764 823 -- PCKVSCCK LAQRHLRQLFVQTOWPSNQDYDSIMAGLGRPEVVRVFGDSTRALNKGQ---
 ZHX1 HUMAN 660 722 -- STGKI-CKK EEOHMKSAFVRTOWPSEEDYDKLAKESCHARTDIVFNGDSTRAYKNGNLK--
 ZHX2 HUMAN 628 690 -- SSGA-AIAKQEQVHLRSTFARTQWCKQEQYDLAAKGLVURTIVRWEKENRCLLTGTIVK--
 HOMEZ HUMAN 355 415 -- QRPRKTKRKKKEOLAILKSFSLQCCWARRREDYQKLEQITGLRRELIIONFGDTRVALNHGQ---
 ZHX2 HUMAN 439 501 -- TTRASDRKKKEQIAHRKASTLQSQSPDDAEVYRLTEVFGARSEIKKWFSDHRYECCSGIVH--
 ZHX1 HUMAN 464 526 -- SFGILAKKKKEQIAELKVSYLKNQPHDSEIIRLMKLTGLKEIKKESDTRYNQNSKSKNC--
 ZHX3 HUMAN 494 553 -- ASIYHNKKHEOLSAIKKSGFCRNQPPGQSEVEHLKVTGLSTREVRKWFSDARXHNKLNK---
 ZHX2 HUMAN 530 591 --- PCKFKEKQCKWIKLSDSLKSSPQAEELDRVRVEKLSRREIDWFSEARRKLRDSMEQK--
 ZHX1 HUMAN 569 630 --- PCKFKEKAEQGRVLCASFINSSVLTDEELNRRAQTKLTREIDAWFATRNKSKSKALKEEK--
 ZHX3 HUMAN 612 671 --- TTKYKERAELSRALSSFAQNLPLDEELDRLRSEKKTREIDWFSEARRK-KVNAAE---
 ZHX3 HUMAN 304 363 --- SSITYNAAMDSSNFKLNSPHKFPVQKAEELCYLVVVKIKVEELKIKWETARLQGGISW---
 ZHX1 HUMAN 284 346 --- NSITYNAALDN-NLLNLNKNKFPPTMSEITVLSAQKKYEEELKIKWESAQRLHGVSWTBE--
 ZHX2 HUMAN 263 324 --- NTTKYNALDT-NATMINSNKKFPVQAEELSWLAASKPHEIHWFAFARLHGGISWSBE--
 HOMEZ HUMAN 55 114 --- WTALAKSGELDS-NHEHLKTSYGFVPSLADIALLCLRYGQMEKFKWEMARLGGISW---
 ADNP HUMAN 754 814 -- DPYGHEDDSYEARKSFELTKYFNKQVPTREIEKLAASLWKKSDIAPHESKRRKGVVD--
 DNPF2 HUMAN 1043 1102 -- PKHYEGRSYEEKQFLRDYHKKVPSPSKIEELSLGWKKKIDAFFEKRRITLMAAI---
 SIX1 HUMAN 124 183 -- GEETSYCEKKSRSVLREWYAHNPVPSREKRELAELATGLTTTVSNWEKRRQRDRAAE--
 SIX2 HUMAN 124 183 -- GEETSYCEKKSRSVLREWYAHNPVPSREKRELAELATGLTTTVSNWEKRRQRDRAAE--
 SIX3 HUMAN 206 265 -- GEKTHCKSKERTRSLREWYLQDPVPSNPKKRELAQATGLTPTGVSNWEKRRQRDRAAE--
 SIX6 HUMAN 128 187 -- GEKTHCKSKERTRHLREWYLQDPVPSNPKKRELAQATGLTPTGVSNWEKRRQRDRAAE--
 SIX5 HUMAN 201 260 -- GEETVYCEKERSRAALKACVGRNKPPTDEKRRLATILGLSLTVSNWEKRRQRDRDGA--
 SIX4 HUMAN 223 282 -- GEETVYCEKKSRAALKYKQNRVPSAEKRLHAKITGLSLTVSNWEKRRQRDRDGA--
 CUX2 HUMAN 1168 1227 -- IKKPRVVLVAESKEAKRLKAYQLEVPSSQOTIELSPQNLKTKNTVINWFHYRSKRRREM--
 CUX1 HUMAN 1244 1303 -- LKKPRVVLVAESKEAKRLKAYQKKVPSKQTIEDLATQNLKTKSTVINWFHYRSKRRREL--
 HNF6 HUMAN 385 444 -- EKKPRLVFVDVVRTLHAIFENKRRPSKELITISQQLGELSTVSNFEMARRSLDKW---
 ONEC2 HUMAN 426 485 -- QKKSRLVFVDLORRTLFAIFENKRRPSKEMOITISQQLGELTTVSNFEMARRSLLEKW---
 ONEC3 HUMAN 414 473 -- EKKPRLVFVDLORRTLFAIFENKRRPSKEMOITISQQLGELTTVSNFEMARRKLNWRW---
 PO5F1 HUMAN 230 289 -- RKKRRSTENVRGNLENLFLQCCKPTLQ-QISHIAQCLQEKDVRVWFCHRRQKGRSS--
 PO5F1B HUMAN 229 288 -- ARKKRRSTENVRGNLENLFLQCCKPTLQ-QISHIAQCLQEKDVRVWFCHRRQKGRSS--
 PO5F2 HUMAN 210 269 -- CKWRASRRRIINSLKFFQRCQKPTLQ-QISHIAQCLQEKDVRVWFVFRSKMSSTRT--
 PO2F2 HUMAN 297 356 -- RKKRRSTENVRFALEKFLANQKPTSEIILLIAEQHNEKEVVRWFCHRRQKKEKIN--
 PO2F1 HUMAN 379 438 -- RKKRRSTETNIRVALRSLLENQKPTSEIITMIAQDLNKEKEVVRWFCHRRQKKEKIN--
 PO2F3 HUMAN 281 340 -- RKKRRSTETNIRLTLEKRFQDNKPTSEESIIMIAEQHNEKEVVRWFCHRRQKKEKIN--
 PO3F2 HUMAN 354 413 -- RKKRRSTEVSVKGALESHFLKCKPSAQEITSLADSLQEKVVRVWFCHRRQKKEKMT--
 PO3F3 HUMAN 406 465 -- RKKRRSTEVSVKGALESHFLKCKPSAQEITSLADSLQEKVVRVWFCHRRQKKEKMT--
 PO3F1 HUMAN 339 398 -- RKKRRSTEVKGALESHFLKCKPSAHEITSLADSLQEKVVRVWFCHRRQKKEKMT--
 PO3F4 HUMAN 278 337 -- RKKRRSTEVSVKGVLETHFLKCKPSAQEISSLADSLQEKVVRVWFCHRRQKKEKMT--
 PIT1 HUMAN 214 273 -- KKKRRSTTAAKDALRRHFGQNKPPSQEIMRMAEELNKEKEVVRVWFCHRRQKKEKRVK--
 PO4F2 HUMAN 345 404 -- RKKRRSTLAAPKKRSLAYFAIQRPSSKIAAIAEKLDLKKNNVVRWFCHRRQKKEKMK--
 PO4F3 HUMAN 274 333 -- RKKRRSTLAAPKKRSLAYFAIQRPSSKIAAIAEKLDLKKNNVVRWFCHRRQKKEKMK--
 PO4F1 HUMAN 356 415 -- KKKRRSTLAAPKKRSLAYFAVQRPSSKIAAIAEKLDLKKNNVVRWFCHRRQKKEKMF--
 PO6F2 HUMAN 607 666 -- RKKRRSTLAAPKEIINAHPEKNTSPSQEMTIAEKLNLDREVVRVWFCHRRQKKEKNTI--

1 10 20 30 40 50 60

ZHX3 HUMAN 764 823 -- PCKVSCCK LAQRHLRQLFVQTOWPSNQDYDSIMAGLGRPEVVRVFGDSTRALNKGQ---
 ZHX1 HUMAN 660 722 -- STGKI-CKK EEOHMKSAFVRTOWPSEEDYDKLAKESCHARTDIVFNGDSTRAYKNGNLK--
 ZHX2 HUMAN 628 690 -- SSGA-AIAKQEQVHLRSTFARTQWCKQEQYDLAAKGLVURTIVRWEKENRCLLTGTIVK--
 HOMEZ HUMAN 355 415 -- QRPRKTKRKKKEOLAILKSFSLQCCWARRREDYQKLEQITGLRRELIIONFGDTRVALNHGQ---
 ZHX2 HUMAN 439 501 -- TTRASDRKKKEQIAHRKASTLQSQSPDDAEVYRLTEVFGARSEIKKWFSDHRYECCSGIVH--
 ZHX1 HUMAN 464 526 -- SFGILAKKKKEQIAELKVSYLKNQPHDSEIIRLMKLTGLKEIKKESDTRYNQNSKSKNC--
 ZHX3 HUMAN 494 553 -- ASIYHNKKHEOLSAIKKSGFCRNQPPGQSEVEHLKVTGLSTREVRKWFSDARXHNKLNK---
 ZHX2 HUMAN 530 591 --- PCKFKEKQCKWIKLSDSLKSSPQAEELDRVRVEKLSRREIDWFSEARRKLRDSMEQK--
 ZHX1 HUMAN 569 630 --- PCKFKEKAEQGRVLCASFINSSVLTDEELNRRAQTKLTREIDAWFATRNKSKSKALKEEK--
 ZHX3 HUMAN 612 671 --- TTKYKERAELSRALSSFAQNLPLDEELDRLRSEKKTREIDWFSEARRK-KVNAAE---
 ZHX3 HUMAN 304 363 --- SSITYNAAMDSSNFKLNSPHKFPVQKAEELCYLVVVKIKVEELKIKWETARLQGGISW---
 ZHX1 HUMAN 284 346 --- NSITYNAALDN-NLLNLNKNKFPPTMSEITVLSAQKKYEEELKIKWESAQRLHGVSWTBE--
 ZHX2 HUMAN 263 324 --- NTTKYNALDT-NATMINSNKKFPVQAEELSWLAASKPHEIHWFAFARLHGGISWSBE--
 HOMEZ HUMAN 55 114 --- WTALAKSGELDS-NHEHLKTSYGFVPSLADIALLCLRYGQMEKFKWEMARLGGISW---
 ADNP HUMAN 754 814 -- DPYGHEDDSYEARKSFELTKYFNKQVPTREIEKLAASLWKKSDIAPHESKRRKGVVD--
 DNPF2 HUMAN 1043 1102 -- PKHYEGRSYEEKQFLRDYHKKVPSPSKIEELSLGWKKKIDAFFEKRRITLMAAI---
 SIX1 HUMAN 124 183 -- GEETSYCEKKSRSVLREWYAHNPVPSREKRELAELATGLTTTVSNWEKRRQRDRAAE--
 SIX2 HUMAN 124 183 -- GEETSYCEKKSRSVLREWYAHNPVPSREKRELAELATGLTTTVSNWEKRRQRDRAAE--
 SIX3 HUMAN 206 265 -- GEKTHCKSKERTRSLREWYLQDPVPSNPKKRELAQATGLTPTGVSNWEKRRQRDRAAE--
 SIX6 HUMAN 128 187 -- GEKTHCKSKERTRHLREWYLQDPVPSNPKKRELAQATGLTPTGVSNWEKRRQRDRAAE--
 SIX5 HUMAN 201 260 -- GEETVYCEKERSRAALKACVGRNKPPTDEKRRLATILGLSLTVSNWEKRRQRDRDGA--
 SIX4 HUMAN 223 282 -- GEETVYCEKKSRAALKYKQNRVPSAEKRLHAKITGLSLTVSNWEKRRQRDRDGA--
 CUX2 HUMAN 1168 1227 -- IKKPRVVLVAESKEAKRLKAYQLEVPSSQOTIELSPQNLKTKNTVINWFHYRSKRRREM--
 CUX1 HUMAN 1244 1303 -- LKKPRVVLVAESKEAKRLKAYQKKVPSKQTIEDLATQNLKTKSTVINWFHYRSKRRREL--
 HNF6 HUMAN 385 444 -- EKKPRLVFVDVVRTLHAIFENKRRPSKELITISQQLGELSTVSNFEMARRSLDKW---
 ONEC2 HUMAN 426 485 -- QKKSRLVFVDLORRTLFAIFENKRRPSKEMOITISQQLGELTTVSNFEMARRSLLEKW---
 ONEC3 HUMAN 414 473 -- EKKPRLVFVDLORRTLFAIFENKRRPSKEMOITISQQLGELTTVSNFEMARRKLNWRW---
 PO5F1 HUMAN 230 289 -- RKKRRSTENVRGNLENLFLQCCKPTLQ-QISHIAQCLQEKDVRVWFCHRRQKGRSS--
 PO5F1B HUMAN 229 288 -- ARKKRRSTENVRGNLENLFLQCCKPTLQ-QISHIAQCLQEKDVRVWFCHRRQKGRSS--
 PO5F2 HUMAN 210 269 -- CKWRASRRRIINSLKFFQRCQKPTLQ-QISHIAQCLQEKDVRVWFVFRSKMSSTRT--
 PO2F2 HUMAN 297 356 -- RKKRRSTENVRFALEKFLANQKPTSEIILLIAEQHNEKEVVRWFCHRRQKKEKIN--
 PO2F1 HUMAN 379 438 -- RKKRRSTETNIRVALRSLLENQKPTSEIITMIAQDLNKEKEVVRWFCHRRQKKEKIN--
 PO2F3 HUMAN 281 340 -- RKKRRSTETNIRLTLEKRFQDNKPTSEESIIMIAEQHNEKEVVRWFCHRRQKKEKIN--
 PO3F2 HUMAN 354 413 -- RKKRRSTEVSVKGALESHFLKCKPSAQEITSLADSLQEKVVRVWFCHRRQKKEKMT--
 PO3F3 HUMAN 406 465 -- RKKRRSTEVSVKGALESHFLKCKPSAQEITSLADSLQEKVVRVWFCHRRQKKEKMT--
 PO3F1 HUMAN 339 398 -- RKKRRSTEVKGALESHFLKCKPSAHEITSLADSLQEKVVRVWFCHRRQKKEKMT--
 PO3F4 HUMAN 278 337 -- RKKRRSTEVSVKGVLETHFLKCKPSAQEISSLADSLQEKVVRVWFCHRRQKKEKMT--
 PIT1 HUMAN 214 273 -- KKKRRSTTAAKDALRRHFGQNKPPSQEIMRMAEELNKEKEVVRVWFCHRRQKKEKRVK--
 PO4F2 HUMAN 345 404 -- KKKRRSTLAAPKKRSLAYFAIQRPSSKIAAIAEKLDLKKNNVRVWFCHRRQKKEKMK--
 PO4F3 HUMAN 274 333 -- RKKRRSTLAAPKKRSLAYFAIQRPSSKIAAIAEKLDLKKNNVRVWFCHRRQKKEKMK--
 PO4F1 HUMAN 356 415 -- KKKRRSTLAAPKKRSLAYFAVQRPSSKIAAIAEKLDLKKNNVRVWFCHRRQKKEKMF--
 PO6F2 HUMAN 607 666 -- KKKRRSTLAAPKEIINAHFEKNTSPSQEMTIAEKLNLDREVVRVWFCHRRQKKEKNTI--

ZHX3 HUMAN 764 823 -- PCKVSCCK LAQRHLRQLFVQTOWPSNQDYDSIMAGLGRPEVVRVFGDSTRALNKGQ---
 ZHX1 HUMAN 660 722 -- STGKI-CKK EEOHMKSAFVRTOWPSEEDYDKLAKESCHARTDIVFNGDSTRAYKNGNLK--
 ZHX2 HUMAN 628 690 -- SSGA-AIAKQEQVHLRSTFARTQWCKQEQYDLAAKGLVURTIVRWEKENRCLLTGTIVK--
 HOMEZ HUMAN 355 415 -- QRPRKTKRKKKEOLAILKSFSLQCCWARRREDYQKLEQITGLRRELIIONFGDTRVALNHGQ---
 ZHX2 HUMAN 439 501 -- TTRASDRKKKEQIAHRKASTLQSQSPDDAEVYRLTEVFGARSEIKKWFSDHRYECCSGIVH--
 ZHX1 HUMAN 464 526 -- SFGILAKKKKEQIAELKVSILKNQPHDSEIIRLMKLTGLKEIKKESDTRYNQNSKSKNC--
 ZHX3 HUMAN 494 553 -- ASIYHNKKHEOLSAIKKSGFCRNQPPGQSEVEHLKVTGLSTREVRKWFSDARXHNKLNK---
 ZHX2 HUMAN 530 591 --- PCKFKEKQCKWIKLSDSLKSSPQAEELDRVRVEKLSRREIDWFSEARRKLRDSMEQK--
 ZHX1 HUMAN 569 630 --- PCKFKEKAEQVRVLCASFINSSVDEELNRRAQTKLRRREIDAWFATNRKSKALKEEK--
 ZHX3 HUMAN 612 671 --- TTKYKERAELSRALSSFAQNLPLDEELDRLRSEKLRREIDWFSEARRK-KVNAAE---
 ZHX3 HUMAN 304 363 --- SSITYNAAMDSSNFKLNSPHKFPVQKAEELCYLVVVKIKVDEELKIKWETARLQGGISW---
 ZHX1 HUMAN 284 346 --- NSITYNAALDN-NLLNLNKNKFPPTMSEITVLSAQKKYEEELKIKWESAQRLHGVSWTBE--
 ZHX2 HUMAN 263 324 --- NTTKYNALDT-NATMINSNKKFPVQAEELSWLAASKPDEHIFRWFAQRLHGGISWSBE--
 HOMEZ HUMAN 55 114 --- WTALAKSGELDS-NHEHLKTSYGFVPSLADIALLCLRYGQMEKFKWEMARLGGISW---
 ADNP HUMAN 754 814 --- DPYGHEDDSYEARKSFLTKYFNKQVPTREIEKLAASLWKKSDIAPHESKRRKQVVD--
 DNPF2 HUMAN 1043 1102 --- PKHYEGRSYEEKQFLDKYHHKKVPSKKEIELSLGWKKKIDAFFEKRRITLMAAI---
 SIX1 HUMAN 124 183 --- GEETSYCEKKSRSVLREWYAHNPVPSREKRELAEATGLTTTVSNWEKRRQRDRAAE--
 SIX2 HUMAN 124 183 --- GEETSYCEKKSRSVLREWYAHNPVPSREKRELAEATGLTTTVSNWEKRRQRDRAAE--
 SIX3 HUMAN 206 265 --- GEOKTHCKERTRSLREWYLQDPVPSNPKKRELAEATGLTPTGVSNWEKRRQRDRAAE--
 SIX6 HUMAN 128 187 --- GEOKTHCKERTRHLREWYLQDPVPSNPKKRELAEATGLTPTGVSNWEKRRQRDRAAE--
 SIX5 HUMAN 201 260 --- GEETVYCEKERSRAALKACVNRNKPPTDEKRRLATILGLSLTVSNWEKRRQRDRDGA--
 SIX4 HUMAN 223 282 --- GEETVYCEKKSRAALKYKQNRVPSAEKRLAKITGLSLTVSNWEKRRQRDRDGA--
 CUX2 HUMAN 1168 1227 --- IKKPRVVLVAESKEAKRLKAYQLEVPSSQOTIELSPQNLKNTNTVINWFHYRSKRRREM--
 CUX1 HUMAN 1244 1303 --- LKKPRVVLVAESKEAKRLKAYQKKVPSQOTIEDLATQNLKNTSTVINWFHYRSKRRREL--
 HNF6 HUMAN 385 444 --- EKKPRLVFVDVRRTHAIFENKRRPSKELITISQQLGELSTVSNFEMARRSLDKW---
 ONEC2 HUMAN 426 485 --- OKKSRLVFVDLRRTHAIFENKRRPSKEMOITISQQLGELTTVSNFEMARRSLLEKW---
 ONEC3 HUMAN 414 473 --- EKKPRLVFVDLRRTHAIFENKRRPSKEMOITISQQLGELTTVSNFEMARRKLNRRW---
 PO5F1 HUMAN 230 289 --- RKKRRSTENVRGNLENLFLQCCKPTLQISHIAQCLQEKDVRVWFCHRRQKGRSS--
 PO5F1B HUMAN 229 288 --- ARKKRRSTENVRGNLENLFLQCCKPTLQISHIAQCLQEKDVRVWFCHRRQKGRSS--
 PO5F2 HUMAN 210 269 --- CKWRASRERRINSLKFFQRCCKPTLQISHIAQCLQEKDVRVWFVFRSKMSSTRT--
 PO2F2 HUMAN 297 356 --- RKKRRSTENVRFALEKFLANQKPTSEIILLIAEQHNEKEVVRWFCHRRQKKEKIN--
 PO2F1 HUMAN 379 438 --- RKKRRSTETNIRVALRSLLENQKPTSEIITMIAQDLNKEKEVVRWFCHRRQKKEKIN--
 PO2F3 HUMAN 281 340 --- RKKRRSTETNIRLTLEKRFQDNKPTSEESIIMIAEQHNEKEVVRWFCHRRQKKEKIN--
 PO3F2 HUMAN 354 413 --- RKKRRSTEVSVKGALESHFLKCKPSAQEITSLADSLQEKVVRVWFCHRRQKKEKMT--
 PO3F3 HUMAN 406 465 --- RKKRRSTEVSVKGALESHFLKCKPSAQEITSLADSLQEKVVRVWFCHRRQKKEKMT--
 PO3F1 HUMAN 339 398 --- RKKRRSTEVKGALESHFLKCKPSAHEITSLADSLQEKVVRVWFCHRRQKKEKMT--
 PO3F4 HUMAN 278 337 --- RKKRRSTEVSVKGVLETHFLKCKPSAQEISSLADSLQEKVVRVWFCHRRQKKEKMT--
 PIT1 HUMAN 214 273 --- KKKRRSTTAAKDALRRHFGQNKPPSQEIMRMAEELNKEKEVVRVWFCHRRQKKEKRVK--
 PO4F2 HUMAN 345 404 --- KKKRRSTLAAPKKRSLAYFAIQRPSSKIAAIAEKLDLKKNNVRVWFCHRRQKKEKMK--
 PO4F3 HUMAN 274 333 --- RKKRRSTLAAPKKRSLAYFAIQRPSSKIAAIAEKLDLKKNNVRVWFCHRRQKKEKMK--
 PO4F1 HUMAN 356 415 --- KKKRRSTLAAPKKRSLAYFAVQRPSSKIAAIAEKLDLKKNNVRVWFCHRRQKKEKMF--
 PO6F2 HUMAN 607 666 --- KKKRRSTLAAPKEIINAHPEKNTSPSQEMTIAEKLNLDREVVRVWFCHRRQKKEKNTI--

ZHX3 HUMAN 764 823 -- PCKVSCCK LAQRHLRQLFVQTOWPSNQDYDSIMAGLGRPEVVRVFGDSTRALNKGQ---
 ZHX1 HUMAN 660 722 -- STGKI-CKK EEOHMKSAFVRTOWPSEEDYDKLAKESCHARTDIVFNGDSTRAYKNGNLK--
 ZHX2 HUMAN 628 690 -- SPSGA-TAKG QEOVHLRSTFARTOWPSEEDYDLAAKGLVURTIVRWEKENRCLLTGTGVK--
 HOMEZ HUMAN 355 415 -- QRPRKTKRKKKEOLAILKSFSLQCCWARRREDYQKLEQITGLRRELIIONFGDTRVALNHGQ---
 ZHX2 HUMAN 439 501 -- TTRASDRKKKEOAHKRASFLQSQFDDAEVYRLTEVGLARSEIKKWFSDHRYECCGSGIVH--
 ZHX1 HUMAN 464 526 -- SPGLRAKKKEOAEALKVSFLKNQFPHDSEIIRLMKLTGLKEIKKESDTRYNQNSKSKNC--
 ZHX3 HUMAN 494 553 -- ASIYHNKKHEOISALKKGSFCRNQFPGQSEVEHLKVTGLSTREVRKWFSDARXHNKLNK---
 ZHX2 HUMAN 530 591 --- PCKFKEKQCKWIKLSDSLKSSPQAEELDRVRVEKLSRREIDWFSEARRKLRDSMEQK--
 ZHX1 HUMAN 569 630 --- PCKFKEKAEQGRVLCASFINSSVDEELNRLRAQTKLRRREIDAWFATRNKSKALKKEEK--
 ZHX3 HUMAN 612 671 --- TTKYKERAEEERALSFAQNLPLDEELDRLRSEKRRRREIDWFSEARRK-KVNAAE---
 ZHX3 HUMAN 304 363 --- SSITYNAAMDSSNFKLNSPHKFPVQKAEALCYLVVVKVKEEELKIKWETARLQGGISW---
 ZHX1 HUMAN 284 346 --- NSITYNAALDN-NLLNLNKNKFPPTMSEITVLSAQKKYEEELKIKWESAQRHAGVSWTEB--
 ZHX2 HUMAN 263 324 --- NTTKYNALDT-NATMINSNKKFPVQAEALSWLAASKPHEHRAWFATRLHGGISWSBEE--
 HOMEZ HUMAN 55 114 --- WTALASGSELDSENEHLKTSYGFVPSLADIALLCLRYGQMEKVFQWEMARLGGISW---
 ADNP HUMAN 754 814 --- DPYGHEDDSYEARKSFETKYFNKQVPTREIEKLAASLWKKSDIAPHESKRRKQVVD---
 DNPF2 HUMAN 1043 1102 --- PKHYEGRSYEEKQFLDYHKKFVPSKKEIELSLGWKKKIDAFFEKRRITLMAAI---
 SIX1 HUMAN 124 183 --- GEETSYCEKKSRSVLREWYAHNPVPSREKRELAEATGLTTTVSNWEKRRQRDRAAE---
 SIX2 HUMAN 124 183 --- GEETSYCEKKSRSVLREWYAHNPVPSREKRELAEATGLTTTVSNWEKRRQRDRAAE---
 SIX3 HUMAN 206 265 --- GEOKTHCKERTRSLREWYLQDPVPSNPKKRELAEATGLTPTGVSNWEKRRQRDRAAE---
 SIX6 HUMAN 128 187 --- GEOKTHCKERTRHLREWYLQDPVPSNPKKRELAEATGLTPTGVSNWEKRRQRDRAAE---
 SIX5 HUMAN 201 260 --- GEETVYCEKERSRAALKACVNRNKPTEDEKRLATILGLSLTVSNWEKRRQRDRDGA---
 SIX4 HUMAN 223 282 --- GEETVYCEKKSRAALKYKQNRVPSAEKRLAKITGLSLTVSNWEKRRQRDRDGA---
 CUX2 HUMAN 1168 1227 --- IKKPRVVLVAESKEAKRLAYQLEVPSSQOTIELSPQNLKTNITVINWFHYRSKRRREM---
 CUX1 HUMAN 1244 1303 --- LKKPRVVLVAESKEAKRLAYQKKVPSQOTIEDLATQNLKTNITVINWFHYRSKRRREL---
 HNF6 HUMAN 385 444 --- EKKPRLVFVDVRRTHAIFENKRRPSKELITISQQLGELSTVSNFEMARRSLDKW---
 ONEC2 HUMAN 426 485 --- OKKSRLVFVDLRRTHAIFENKRRPSKEMOITISQQLGELTTVSNFEMARRSLLEKW---
 ONEC3 HUMAN 414 473 --- EKKPRLVFVDLRRTHAIFENKRRPSKEMOITISQQLGELTTVSNFEMARRKLNRRW---
 PO5F1 HUMAN 230 289 --- RKKRRSTENVRGNLENLFLQCCKPTLQISHIAQCLQEKDVRVWFCHRRQKGRSS---
 PO5F1B HUMAN 229 288 --- ARKKRRSTENVRGNLENLFLQCCKPTLQISHIAQCLQEKDVRVWFCHRRQKGRSS---
 PO5F2 HUMAN 210 269 --- CKWRASRERRINSLKFFQRCQKPTLQISHIAQCLQEKDVRVWFVFRSKMSSTRT---
 PO2F2 HUMAN 297 356 --- RKKRRSTENVRFALEKFLANQKPTSEIILLIAEQHNEKEVVRWFCHRRQKKEKIN---
 PO2F1 HUMAN 379 438 --- RKKRRSTETNIRVALRSLLENQKPTSEIITMIAQDLNKEKEVVRWFCHRRQKKEKIN---
 PO2F3 HUMAN 281 340 --- RKKRRSTETNIRLTLEKRFQDNKPTSEESIIMIAEQHNEKEVVRWFCHRRQKKEKIN---
 PO3F2 HUMAN 354 413 --- RKKRRSTEVSVKGALESHFLKCKPSAQEITSLADSLQKEKEVVRWFCHRRQKKEKMT---
 PO3F3 HUMAN 406 465 --- RKKRRSTEVSVKGALESHFLKCKPSAQEITSLADSLQKEKEVVRWFCHRRQKKEKMT---
 PO3F1 HUMAN 339 398 --- RKKRRSTEVKGALESHFLKCKPSAHEITSLADSLQKEKEVVRWFCHRRQKKEKMT---
 PO3F4 HUMAN 278 337 --- RKKRRSTEVSVKGVLETHFLKCKPSAQEISSLADSLQKEKEVVRWFCHRRQKKEKMT---
 PIT1 HUMAN 214 273 --- KKKRRSTTAAKDALRRHFGQNKPSSEQIMRMAEELNKEKEVVRWFCHRRQKKEKRVK---
 PO4F2 HUMAN 345 404 --- KKKRRSTLAAPKKRSLAYFAIQRPSEKIAAIAEKLDLKKNNVVRWFCHRRQKKEKMK---
 PO4F3 HUMAN 274 333 --- RKKRRSTLAAPKKRSLAYFAIQRPSEKIAAIAEKLDLKKNNVVRWFCHRRQKKEKMK---
 PO4F1 HUMAN 356 415 --- KKKRRSTLAAPKKRSLAYFAVQRPSEKIAAIAEKLDLKKNNVVRWFCHRRQKKEKMF---
 PO6F2 HUMAN 607 666 --- KKKRRSTLAAPKEIINAHFEKNTSPSQEMTIAEKLNLDREVVRWFCHRRQKKEKNTI---

07:09:47 PM

You will recall from earlier that the **Prosite** pattern for a **homeobox** is the ever memorable:

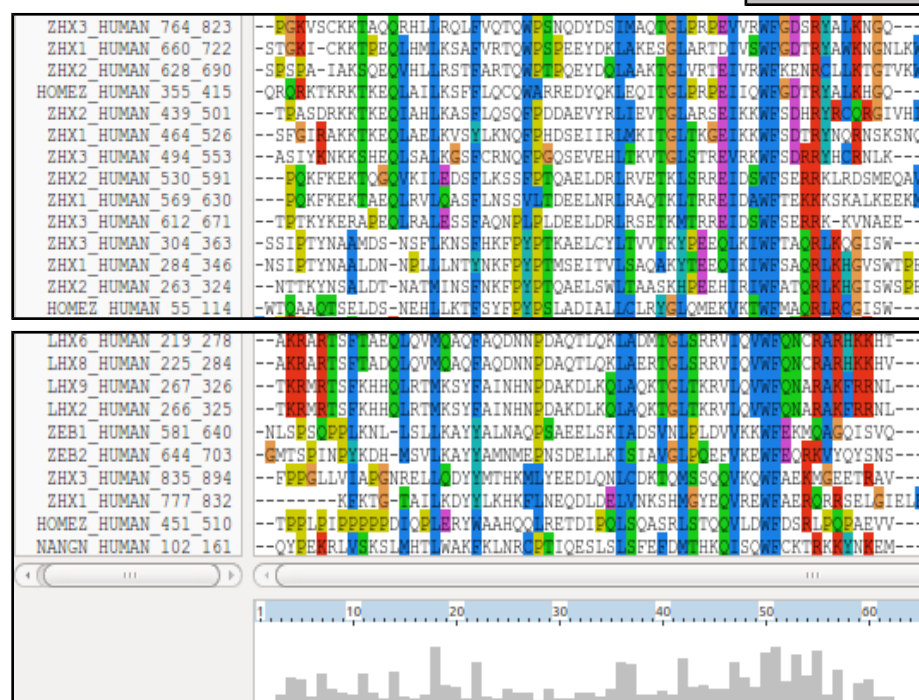
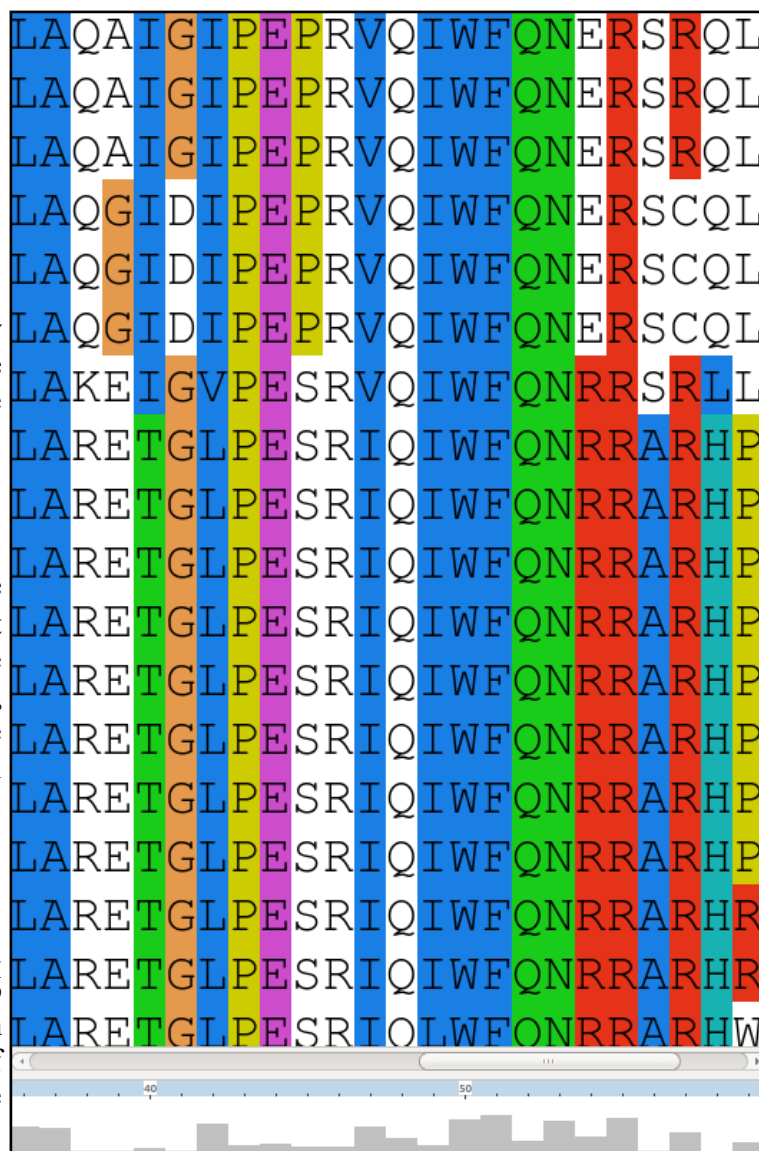
[LIVMFYG] - [ASLVR] -x (2) - [LIVMSTACN] -x- [LIVM] - {Y} -x (2) - {L} - [LIV] - [RKNQESTAIY] - [LIVFSTNKH] -W- [FYVC] -x- [NDQTAH] -x (5) - [RKNAIMW]

This corresponds to positions **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** pattern³. 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 very low, suggesting that the sequences in view might not be typical of the whole alignment set. Which, upon checking they are not!

Looking through this alignment, I get the feeling I could design a better, stricter pattern for the region between **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.

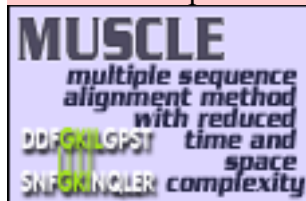
Clustal Omega ? New MSA tool that uses seeded guide trees and HMM profile-profile techniques to generate alignments. Suitable for medium-large alignments. Launch Clustal Omega	MUSCLE ? Accurate MSA tool, especially good with proteins. Suitable for medium alignments. Launch MUSCLE
Kalign ? Very fast MSA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

Here I intend to align again the human **homeboxes** with just one of the tools on offer. Then take a quick look at how the machine generated multiple alignment can be manually edited using **Jalview**, a program that is installed on your workstation and that you have already used as an alignment viewer when investigating **Pfam** and also **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 alignment⁵. Particularly for more difficult alignments, **MUSCLE** should do a better job than **clustalX**. The alignment you will generate here will certainly be different. I leave you to judge for yourselves whether it is better.

Start by requesting to [Launch MUSCLE](#).

Use the **Choose File** button to upload the file containing the **FASTA** format **homeobox** sequences, **homeobox_human.fasta**. This file should no longer included the sequences with a mess around position **30**.

STEP 1 - Enter your input sequences

Enter or paste a set of sequences in any supported format:

Or upload a file: **Choose File** homeobox_human.fasta

STEP 2 - Set your Parameters

OUTPUT FORMAT: **ClustalW**

The default settings will fulfill the needs of most users and, for that reason, are not visible.

More options... (Click here, if you want to view or change the default settings.)

over the programs I use that their creators deemed it sensible to make available⁶?

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 all the control

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 wish⁷. You may also set the **OUTPUT ORDER** to **aligned** or ... **aligned**?

STEP 2 - Set your Parameters

OUTPUT FORMAT: **ClustalW**

OUTPUT TREE: **none**

OUTPUT ORDER: **aligned**

ClustalW

Pearson/FASTA

ClustalW

ClustalW (strict)

HTML

GCG MSF

Phylip interleaved

Phylip sequential

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 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 The java alignment editor and viewer you used to look at the **Pfam** and **Jpred** alignments earlier.

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.

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.

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

```

ARX_HUMAN_328_387      --QRRYR--TTFTSYQLEELERAFQKTHYPDVFTREELAMRLDLTEARVQVWFQNRRAKWR
ALX1_HUMAN_132_191     --KRRHR--TTFTSLQLEELKVFQKTHYPDVVREQLALRTELTEARVQVWFQNRRAKWR
ALX4_HUMAN_214_273     --KRRNR--TTFTSYQLEELKVFQKTHYPDVVAREQLAMRTDLTEARVQVWFQNRRAKWR
ALX3_HUMAN_153_212     --KRRNR--TTFSTFQLEELKVFQKTHYPDVVAREQLALRTDLTEARVQVWFQNRRAKWR
ISL1_HUMAN_181_240     --TTRVR--TVLNEKQLHLTLRTCYAANPRPDALMKEQLVEMTGLSPRVIRVWFQNKRCCKDK
ISL2_HUMAN_191_250     --TTRVR--TVLNEKQLHLTLRTCYAANPRPDALMKEQLVEMTGLSPRVIRVWFQNKRCCKDK
LHX9_HUMAN_267_326     --TKRMR--TSFKHHQLRTMKSIFYAINHNPDADKDLQAQKTGLTKRVLQVWFQNARAKFR
LHX2_HUMAN_266_325     --TKRMR--TSFKHHQLRTMKSIFYAINHNPDADKDLQAQKTGLTKRVLQVWFQNARAKFR
LHX6_HUMAN_219_278     --AKRAR--TSFTAELQVMQAQFAQDNNPDAQTLQKLADMTGLSRRVIQVWFQNCRAHK
LHX8_HUMAN_225_284     --AKRAR--TSFTADQLQVMQAQFAQDNNPDAQTLQKLAERTGLSRRVIQVWFQNCRAHK
ZFHX3_HUMAN_2641_2700  --DKRLR--TTITPEQLEILYQKYLSDSNPTRKMLDHIAREVGLKRRVVQVWFQNTRARER
ZFHX4_HUMAN_2560_2619  --DKRLR--TTITPEQLEILYQKYLSDSNPTRKMLDHIAREVGLKRRVVQVWFQNTRARER
ZFHX2_HUMAN_1857_1916  --DKRLR--TTILPEQLEILYRWYMQDSNPTKMLDCISEEVGLKRRVVQVWFQNTRARER
ZFHX2_HUMAN_2065_2124  --QRRYR--TQMSSLQKIMKACYEAYRTPTMQECEVLGEEIGLPKRVIQVWFQNARAKEK
ZFHX3_HUMAN_2944_3003  --PGQKFRF--TQMTNLQKVLKSCFNDYRTPMTLECEVLGNDIGLPKRVIQVWFQNARAKEK
ZFHX4_HUMAN_2884_2943  --HKRFR--TQMSNLQKVLKACFSYDRTPTMQECEMLGNEIGLPKRVIQVWFQNARAKEK
LMX1A_HUMAN_195_254    --PKRPR--TILTTQORAFKASFEVSSKPCRKVRETAAETGLSVRVVQVWFQNRRAKMK
LMX1B_HUMAN_219_278    --PKRPR--TILTTQORAFKASFEVSSKPCRKVRETAAETGLSVRVVQVWFQNRRAKMK
LHX1_HUMAN_180_239     --RRGPR--TTIKAKQLETLKAAFAATPKPTRHIREQLAQETGLNMRVIQVWFQNRSSKER
LHX5_HUMAN_180_239     --RRGPR--TTIKAKQLETLKAAFAATPKPTRHIREQLAQETGLNMRVIQVWFQNRSSKER
LHX3_HUMAN_157_216     --AKRPR--TTITAKQLETLKSAYNTSPKPARHVREQLSSETGLDMRVVQVWFQNRRAKEK
LHX4_HUMAN_157_216     --AKRPR--TTITAKQLETLKSAYNTSPKPARHVREQLSSETGLDMRVVQVWFQNRRAKEK
:      :      :      :      :
HOMEZ_HUMAN_451_510    EVV---
ZHX3_HUMAN_835_894     RAV---
ZHX1_HUMAN_777_832     LGIELF
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---

```

```

ARX_HUMAN_328_387      --QRRYR--TTFTSYQLEELERAFQKTHYPDVFTREELAMRLDLTEARVQVWFQNRRAKWR
ALX1_HUMAN_132_191     --KRRHR--TTFTSLQLEELKVFQKTHYPDVVREQLALRTELTEARVQVWFQNRRAKWR
ALX4_HUMAN_214_273     --KRRNR--TTFTSYQLEELKVFQKTHYPDVVAREQLAMRTDLTEARVQVWFQNRRAKWR
ALX3_HUMAN_153_212     --KRRNR--TTFSTFQLEELKVFQKTHYPDVVAREQLALRTDLTEARVQVWFQNRRAKWR
ISL1_HUMAN_181_240     --TTRVR--TVLNEKQLHLTLRTCYAANPRPDALMKEQLVEMTGLSPRVIRVWFQNKRCCKDK
ISL2_HUMAN_191_250     --TTRVR--TVLNEKQLHLTLRTCYAANPRPDALMKEQLVEMTGLSPRVIRVWFQNKRCCKDK
LHX9_HUMAN_267_326     --TKRMR--TSFKHHQLRTMKSIFYAINHNPDADKDLQAQKTGLTKRVLQVWFQNARAKFR
LHX2_HUMAN_266_325     --TKRMR--TSFKHHQLRTMKSIFYAINHNPDADKDLQAQKTGLTKRVLQVWFQNARAKFR
LHX6_HUMAN_219_278     --AKRAR--TSFTAELQVMQAQFAQDNNPDAQTLQKLADMTGLSRRVIQVWFQNCRAHK
LHX8_HUMAN_225_284     --AKRAR--TSFTADQLQVMQAQFAQDNNPDAQTLQKLAERTGLSRRVIQVWFQNCRAHK
ZFHX3_HUMAN_2641_2700  --DKRLR--TTITPEQLEILYQKYLSDSNPTRKMLDHIAREVGLKRRVVQVWFQNTRARER
ZFHX4_HUMAN_2560_2619  --DKRLR--TTITPEQLEILYQKYLSDSNPTRKMLDHIAREVGLKRRVVQVWFQNTRARER
ZFHX2_HUMAN_1857_1916  --DKRLR--TTILPEQLEILYRWYMQDSNPTKMLDCISEEVGLKRRVVQVWFQNTRARER
ZFHX2_HUMAN_2065_2124  --QRRYR--TQMSSLQKIMKACYEAYRTPTMQECEVLGEEIGLPKRVIQVWFQNARAKEK
ZFHX3_HUMAN_2944_3003  --PGQKFRF--TQMTNLQKVLKSCFNDYRTPMTLECEVLGNDIGLPKRVIQVWFQNARAKEK
ZFHX4_HUMAN_2884_2943  --HKRFR--TQMSNLQKVLKACFSYDRTPTMQECEMLGNEIGLPKRVIQVWFQNARAKEK
LMX1A_HUMAN_195_254    --PKRPR--TILTTQORAFKASFEVSSKPCRKVRETAAETGLSVRVVQVWFQNRRAKMK
LMX1B_HUMAN_219_278    --PKRPR--TILTTQORAFKASFEVSSKPCRKVRETAAETGLSVRVVQVWFQNRRAKMK
LHX1_HUMAN_180_239     --RRGPR--TTIKAKQLETLKAAFAATPKPTRHIREQLAQETGLNMRVIQVWFQNRSSKER
LHX5_HUMAN_180_239     --RRGPR--TTIKAKQLETLKAAFAATPKPTRHIREQLAQETGLNMRVIQVWFQNRSSKER
LHX3_HUMAN_157_216     --AKRPR--TTITAKQLETLKSAYNTSPKPARHVREQLSSETGLDMRVVQVWFQNRRAKEK
LHX4_HUMAN_157_216     --AKRPR--TTITAKQLETLKSAYNTSPKPARHVREQLSSETGLDMRVVQVWFQNRRAKEK
:      :      :      :      :
HOMEZ_HUMAN_451_510    EVV---
ZHX3_HUMAN_835_894     RAV---
ZHX1_HUMAN_777_832     LGIELF
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---

```


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/>

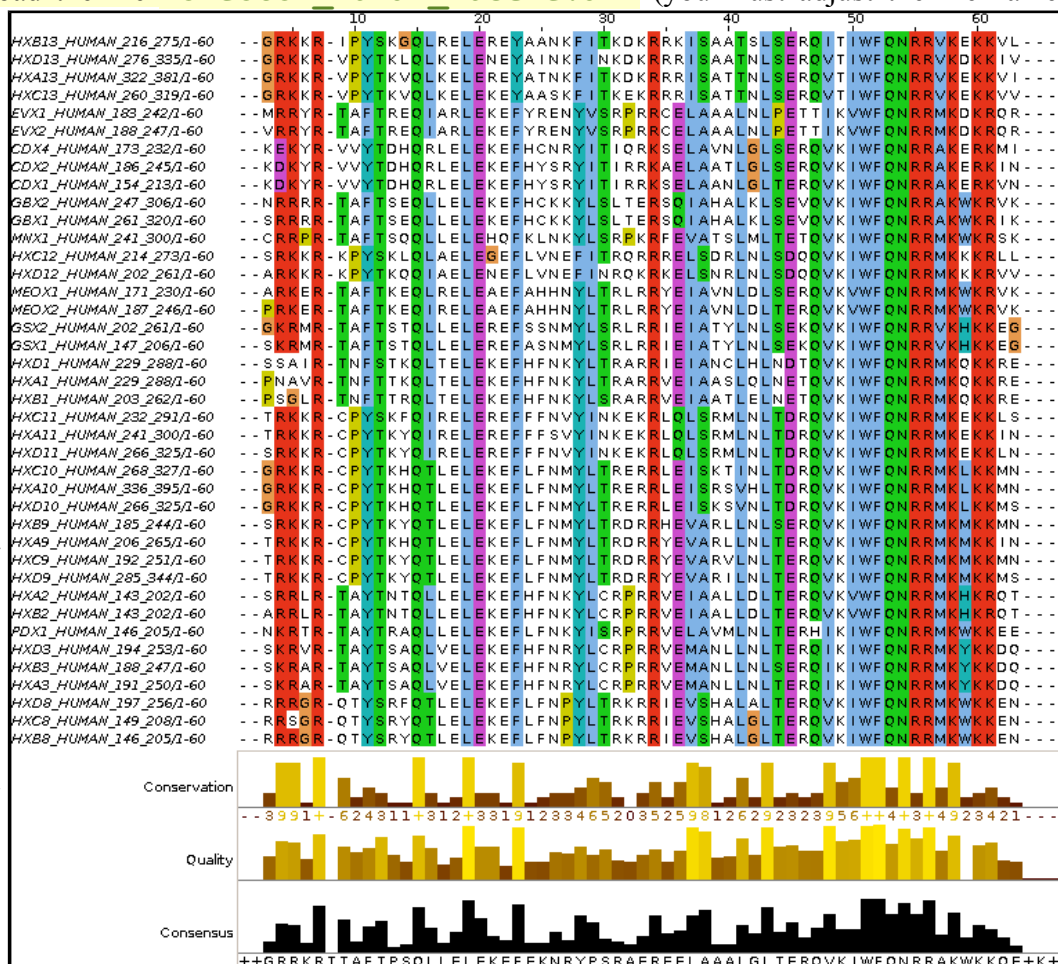
and selecting the **Launch Jalview Desktop** link at the top of the page. Close down all the example outputs

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 must 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.

Now the **MUSCLE** and massaged **ClustalX** alignments look even more similar! In the nicely aligned regions at least.

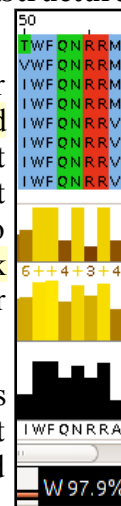


There are many **Jalview** features that merit investigation. Have a look around if you have time. In particular, **Jalview** will compute simple phylogenetic trees for you employing a number of methods (**Calculate Tree** from the **Calculate** pull down menu). Try it, but be aware this is only sensible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

Jalview is made by the same group as produce **JPred**. You could send your alignment for **Secondary Structure Prediction** via the **Web Service** pull down menu, if you wished.

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

Finally, take a look at the **Jalview** "Manhattan Skyline" for the highly conserved **W** at position 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?



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 gap penalties? Not possible, but fear not children!! The sages of the **EBI** assure us that they have selected the correct setting for our sequences that they have never seen??? Sing praises and have faith brothers and sisters for we are in the safe hands of the blessed!!!

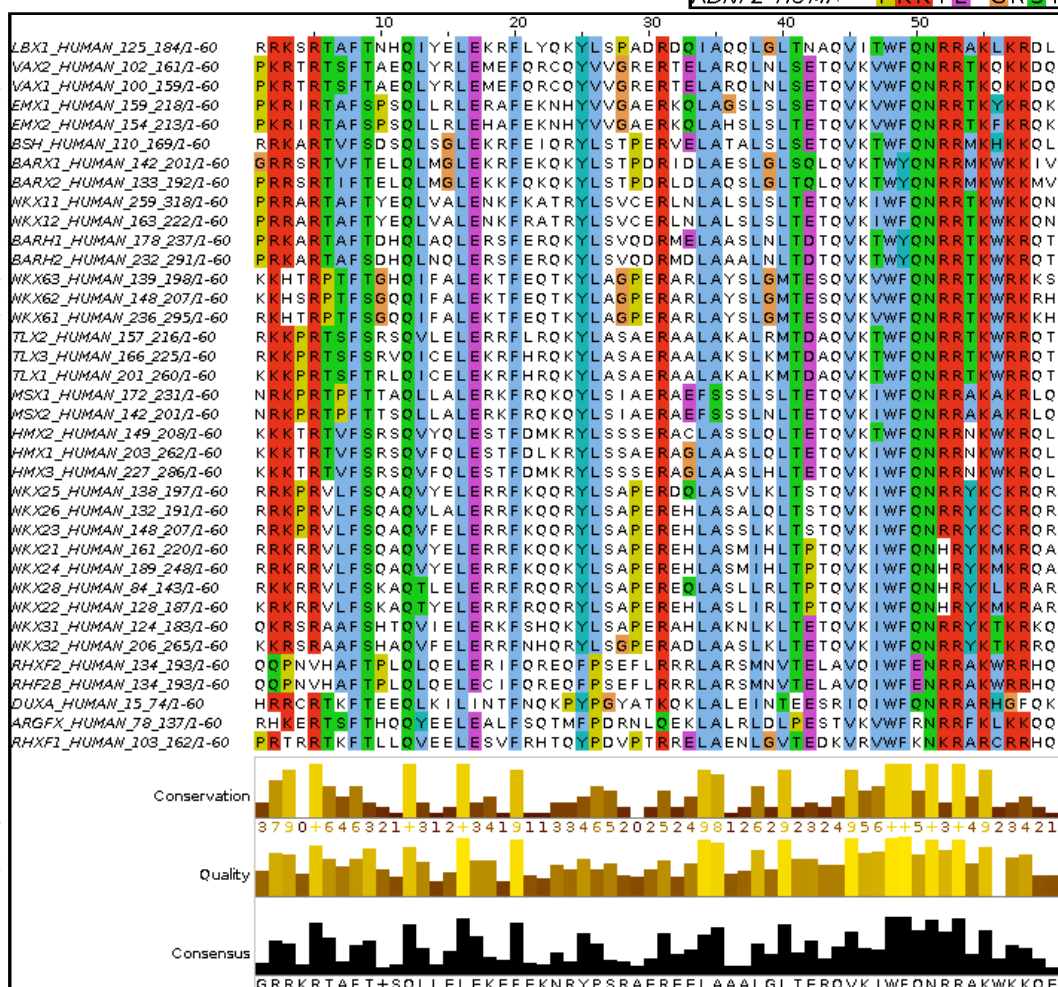
```

10
ZHX3_HUMAN - - S S I P T - Y N A A
ZHX1_HUMAN - - N S I P T - Y N A A
ZEB2_HUMAN - - G M T S P - I N P Y
ZEB1_HUMAN - - N L S P S - Q P P L
NANGN_HUM - - Q Y P E K - R L V S
ZHX1_HUMAN - - P Q K F K - - E K T
ZHX2_HUMAN - - P Q K F K - - E K T
ZHX3_HUMAN - - T P T K Y - K E R A
ZHX2_HUMAN - - T P A S D - R K K T
ZHX3_HUMAN - - A S I Y K - N K K S
ZHX1_HUMAN - - S F G I R - A K K T
HOMEZ_HUM - - Q R Q R K T K R K T
ZHX2_HUMAN - - S P S P A - I A K S
ZHX3_HUMAN - - P G K V S - C K K T
ZHX1_HUMAN - - S T G K I - C K K T
ADNP_HUMAN - P K G H E - D D S Y
ADNP2_HUMA - P K K Y E - G R S Y

```

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

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



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

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 alignment. Whilst the **EBI** refuse to share with their users how they are running the programs, however, it is hardly worth the effort. Maybe I look for a more obliging service for next time?

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 **Position Specific Scoring Matrix - PSSM**) is computed. A **PSSM** is very similar in concept and purpose to an **HMM** profile in that it summarises significant features present in the sequences it represents.

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

You have used **PSI-BLAST** integrated into **Jpred** already and similar ideas were used to create the **PFAM** alignments. Here we will use **PSI-BLAST** explicitly at the **NCBI** on the **Paired DOMAIN** of the **PAX6** protein that you saved in a file earlier. It should be possible to detect a large family of **PAX** domains and to eventually multiply align them generating something like the **Full** alignment from the **PFAM** database viewed earlier¹⁰.

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

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

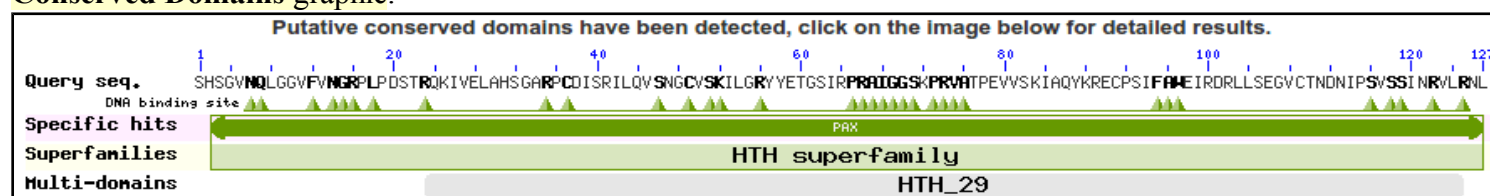
Click on the **BLAST** option. Select **protein BLAST** from the **Basic BLAST** section. Upload the **PAX6** paired box domain sequence (stored in the file **pax_domain.fasta**) using the appropriate **Browse** button.

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

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

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

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



¹⁰ But hopefully a mite more credible!

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

Conserved domains on [lcl|Query_2485] View **Standard Results** ▾ ?

Pax-Domain P26367(4-130)

Graphical summary ☐ Zoom to residue level show extra options »

Query seq. SHSGVNLGGVFVNGRPLDSTRQKIVELAHSGARPCDISRIQLQVSNCGYSKILGRYYETGSIKPPRAIGGSKPRVATPEVWSKIAQYKRECPISIFAMIEIRDRLSEGVCCTNDNIPSSVSSINRVLRL

DNA binding site

Specific hits PAX

Non-specific hits PAX

Superfamilies HTH superfamily

Multi-domains HTH_29

[Search for similar domain architectures](#) ? [Refine search](#) ?

List of domain hits

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

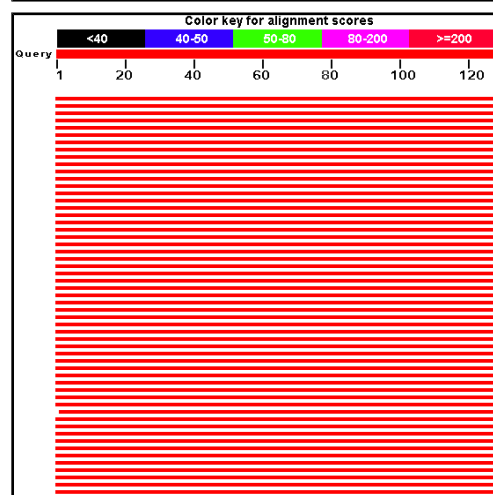
Blast search parameters

Data Source: Live blast search RID = XCKC805Y014

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

References:

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



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

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

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

Sequences producing significant alignments with E-value BETTER than threshold

Select: **All** None Selected: 0

[Alignments](#) [Download](#) [GenPept](#) [Graphics](#) [Distance tree of results](#) [Multiple alignment](#) [Settings](#)

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

Download ▾ GenPept Graphics

oculorhombin [Homo sapiens]
Sequence ID: [gb|AAA59962.1|](#) Length: 422 Number of Matches: 1

Range 1: 4 to 130 [GenPept](#) [Graphics](#) ▾ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
263 bits(671)	1e-83	Compositional matrix adjust.	127/127(100%)	127/127(100%)	0/127(0%)
Query 1	SHSGVNQLGGVFVNGRPLDPSTRQKIVELAHSGARPCDISRILQVSNCGVSKILGRYYET	60			
Sbjct 4	SHSGVNQLGGVFVNGRPLDPSTRQKIVELAHSGARPCDISRILQVSNCGVSKILGRYYET	63			
Query 61	GSIRPRAIGGSKPRVATPEVVS KIAQYKRECP SIFAW EIRDRL SEGVC TNDNIP SVSSI	120			
Sbjct 64	GSIRPRAIGGSKPRVATPEVVS KIAQYKRECP SIFAW EIRDRL SEGVC TNDNIP SVSSI	123			
Query 121	NRVLRNL 127				
Sbjct 124	NRVLRNL 130				

Download ▾ GenPept Graphics

paired box protein Pax-6 [Rattus norvegicus]
Sequence ID: [ref|NP_037133.1|](#) Length: 422 Number of Matches: 1
[► See 3 more title\(s\)](#)

Range 1: 4 to 130 [GenPept](#) [Graphics](#) ▾ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
263 bits(671)	1e-83	Compositional matrix adjust.	127/127(100%)	127/127(100%)	0/127(0%)
Query 1	SHSGVNQLGGVFVNGRPLDPSTRQKIVELAHSGARPCDISRILQVSNCGVSKILGRYYET	60			
Sbjct 4	SHSGVNQLGGVFVNGRPLDPSTRQKIVELAHSGARPCDISRILQVSNCGVSKILGRYYET	63			
Query 61	GSIRPRAIGGSKPRVATPEVVS KIAQYKRECP SIFAW EIRDRL SEGVC TNDNIP SVSSI	120			
Sbjct 64	GSIRPRAIGGSKPRVATPEVVS KIAQYKRECP SIFAW EIRDRL SEGVC TNDNIP SVSSI	123			
Query 121	NRVLRNL 127				
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 relatively poor quality sequences and the presence of so much duplication have been minimised?

Download ▾ GenPept Graphics

paired box protein Pax-6 [Xenopus laevis]
Sequence ID: [ref|NP_001165666.1|](#) Length: 393 Number of Matches: 1
[► See 1 more title\(s\)](#)

Range 1: 4 to 130 [GenPept](#) [Graphics](#) ▾ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
259 bits(661)	2e-82	Compositional matrix adjust.	125/127(98%)	126/127(99%)	0/127(0%)
Query 1	SHSGVNQLGGVFVNGRPLDPSTRQKIVELAHSGARPCDISRILQVSNCGVSKILGRYYET	60			
Sbjct 4	SHSGVNQLGGVFVNGRPLDPSTRQKIVELAHSGARPCDISRILQVSNCGVSKILGRYYET	63			
Query 61	GSIRPRAIGGSKPRVATPEVVS KIAQYKRECP SIFAW EIRDRL SEGVC TNDNIP SVSSI	120			
Sbjct 64	GSIRPRAIGGSKPRVATPEV+KIA YKRECP SIFAW EIRDRL SEGVC TNDNIP SVSSI	123			
Query 121	NRVLRNL 127				
Sbjct 124	NRVLRNL 130				

Download ▾ GenPept Graphics

Pax6 [Bos taurus]
Sequence ID: [gb|AAC18658.1|](#) Length: 146 Number of Matches: 1

Range 1: 4 to 144 [GenPept](#) [Graphics](#) ▾ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
250 bits(638)	2e-82	Compositional matrix adjust.	127/141(90%)	127/141(90%)	14/141(9%)
Query 1	SHSGVNQLGGVFVNGRPLDPSTRQKIVELAHSGARPCDISRILQ-----VS	46			
Sbjct 4	SHSGVNQLGGVFVNGRPLDPSTRQKIVELAHSGARPCDISRILQ-----VS	63			
Query 47	NGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVS KIAQYKRECP SIFAW EIRDRL LSE	106			
Sbjct 64	NGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVS KIAQYKRECP SIFAW EIRDRL LSE	123			
Query 107	GVCTNDNIPSVSSINRVLRNL 127				
Sbjct 124	GVCTNDNIPSVSSINRVLRNL 144				

Download ▾ GenPept Graphics

Related Information

[Gene](#) - associated gene details
[Identical Proteins](#) - Identical proteins to NP_001165666.1

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.

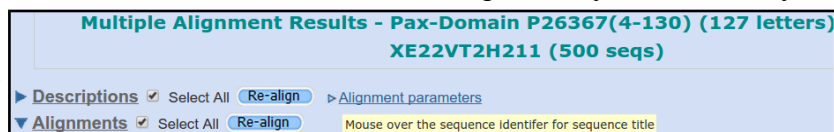
<input type="checkbox"/>	paired box 6 [Monodelphis domestica]	238	238	94%	7e-76	90%	ACZ54379.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/>	PREDICTED: paired box protein Pax-6-like isoform X1 [Acromyrmex ec	246	246	99%	8e-76	94%	XP_011063177.1	<input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

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

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

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

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



At the top of the actual alignment, set **View Format to Plain Text** (... and then hide the **Descriptions** again??), this being the easiest format to understand in a hurry. The alignment will have very ragged ends, but the important region of **120** or so amino acids representing the **PAX** domain is really quite impressive. In particular, the **isoform 5a** insertion is very convincing¹¹.

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

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

Model Answers to Questions in the Instructions Text.

Notes:

For the most part, these “**Model Answers**” just provide the reactions/solutions I hoped you would work out for yourselves. However, sometime I have tried to offer a bit more 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 pertinent to the objective of these exercises. I like to keep these meanders, as they help and entertain me, but I wish to warn you to only take regard of them if you are feeling particularly strong and have time to burn. Certainly not a good idea to indulge here during a time constrained course event!

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

Summary:

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

The other will start:

Full Answer:

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

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

From your investigations of Multiple Sequence AlignmentClustalx

How might you have saved the need to recompute the alignment by selecting an **Edit** option other than **Remove All Gaps**?

Pretty sure using **Remove Gap-Only columns** rather than **Remove All Gaps** would have got to the second alignment in one step. I struggle to see how this could fail, but cannot convince myself it would be absolutely certain to work. Maybe some of the sequences that were removed would subtly alter the alignment calculations? Doubtful, given how crude the whole thing is. Anyway, it would have been good enough for me in “Real Life” for this data. I did not suggest this approach as we had to make a **Fasta** format file with the entire sequence set to be aligned by the alternative software as a by product of working with **clustalx**. This would not be easily possible if the **Remove Gap-Only columns** short cut was taken.

Muscle

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

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

Essentially, both **ClustalX** and **MUSCLE** work in two stages. First they create **Guide Tree(s)** (**ClustalX** just one, **MUSCLE** creates 2). Then from the **Guide Tree(s)** they compute the final **MSA**. 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 rerun the first **Guide Tree** generation stage.

Of course, as mentioned above, utterly pointless if there is no way to change the parameters? 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.

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.

Pseudocount	<input type="text" value="0"/>	
<small>Pseudocount parameter. If zero is specified, then the parameter is automatically determined through a minimum length description principle (PMID 19088134). A value of 30 is suggested in order to obtain the approximate behavior before the minimum length principle was implemented.</small>		

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:

“Position specific score matrices (**PSSMs**) 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 **PSSMs** 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 **PSSMs**. 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.

DPJ – 2015.09.14