

ELB18F

Entry Level Bioinformatics

19-23 February 2018

(First 2018 run of this Course)

Basic Bioinformatics Sessions

Practical 6: Multiple Sequence Alignment

Tuesday 20 February 2018

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.o	rg/	
Choose the Advanced Toption of the Search button.		
First specify that you are only interested in Human	n proteins. To do this, set the first field to	Organism [OS] and
Term to Human [9606].	Term Organism [OS] Human [9606]	۵
Set the second field selector to Reviewed	Perviewed v Vac v	
and the corresponding Term to Yes (that is,	Term	
choose to find only SwissProt entries).	Function V DNA binding V Homeobox	
Click on the the button to request a further	Length range Evidence ¹ 50 70 Any assertion method V	
field selection option. Set the new field to		Q
Function. Set the type of Function to DNA bindin	g. Set the Term selection to Homeobox.	

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

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

8	BLAST	Align 土 Downloa	d 🗎	Add to basket 🛛 🖊 Columns >		1 to 25 of 237	▶ Show 25 ▼
	Entry 🖨	Entry name 🗘		Protein names 🗘 🛛 🔊	Gene names 🖨	Organism 🗘	Length 🗘 🖊
	P50222	MEOX2_HUMAN	☆	Homeobox protein MOX-2	MEOX2 GAX, MOX2	Homo sapiens (Human)	304
	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
	P49639	HXA1_HUMAN	¢	Homeobox protein Hox-A1	HOXA1 HOX1F	Homo sapiens (Human)	335
	Q99697	PITX2_HUMAN	☆	Pituitary homeobox 2	PITX2 ARP1, RGS, RIEG, RIEG1	Homo sapiens (Human)	317
	Q99801	NKX31_HUMAN	¢	Homeobox protein Nkx-3.1	NKX3-1 NKX3.1, NKX3A	Homo sapiens (Human)	234
	Q01860	PO5F1_HUMAN	☆	POU domain, class 5, transcription	POU5F1 OCT3, OCT4, OTF3	Homo sapiens (Human)	360
	Q15475	SIX1_HUMAN	¢	Homeobox protein SIX1	SIX1	Homo sapiens (Human)	284
	Q01826	SATB1_HUMAN	☆	DNA-binding protein SATB1	SATB1	Homo sapiens (Human)	763
	P43699	NKX21_HUMAN	☆	Homeobox protein Nkx-2.1	NKX2-1 NKX2A, TITF1, TTF1	Homo sapiens (Human)	371

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.

Practical 6: Multiple Sequence Alignment					Tuesday 20 February 2018		
Now save the entire list into a file using the	Download selected (0)						
download to uncompressed . Make sure you have	all	sequenc	es sele	ected	O Download all (237)		
and that Text (i.e. EMBL or SwissProt) format	Format						
button and do whatever it takes to ensure your r	esu	lts end u	ip in a	a file			
residing on your Desktop called:	Text ~						
human_homeobox_protei	ins	s.emb					
ID MEOX2 HUMAN Reviewed; 304 AA.		Preview first 10 ¹ Go					
AC P50222; A4D127; B2R817; 075263; Q9UPL6;							
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;							
GN Name=MEOX2; Synonyms=GAX, MOX2;			Take	a swi	ift look at the file you have just created.		
OS Homo sapiens (Human). OC Eukarvota: Metazoa: Chordata: Craniata: Vertebrata: Eut	مام	stomi ·	Your	neat 1	list of Human Homeobox sequences will		
OC Mammalia; Eutheria; Euarchontoglires; Primates; Haplorr	hini	;	have	transf	formed into a flood of many SwissProt		
OC Catarrhini; Hominidae; Homo. OX NCBI TaxID=9606;			form	at II	iProtKR entries Usly but what is		
RN [1]			requir	n UI rad	In rotado entries. Ogiy, out what is		
RC TISSUE=Embryo;			requi	icu.			
RX PubMed=7607679; DOI=10.1016/0888-7543(95)80174-K; BA Grigoriou M., Kastrinaki MC., Modi W., Theodorakis K.	. Ma	nkoo B.					
RA Pachnis V., Karagogeos D.;							
RT "Isolation of the human MOX2 homeobox gene and localiza" RT chromosome 7p22.1-p21.3.";	tion	to					
RL Genomics 26:550-555(1995).							
RP NUCLEOTIDE SEQUENCE [MRNA], AND VARIANT 79-HIS-HIS-80 D	EL.						
RC TISSUE=Heart;							
	FT	CHAIN	1	304	Homeobox protein MOX-2.		
	FT FT	DNA_BIND	187	246	/FTId=PRO_0000049197. Homeobox. {ECO:0000255 PROSITE-		
Search (Control F) for the term DNA_BIND.	FT FT	COMPREAS	42	47	ProRule: PRU00108}.		
	FT	COMPBIAS	68	80	Poly-His.		
	FT FT	COMPBIAS VARIANT	81 79	86 80	Poly-Gln. Missing. {ECO:0000269 PubMed:7713505}.		
	FT	177 D T 7 NM	80	90	/FTId=VAR_026040.		
	FT	VARIANI	80	80	ECO: 0000269 PubMed: 14702039,		
It should occur many times (at least once per	FT FT				ECO:0000269 PubMed:15489334, ECO:0000269 PubMed:7607679}.		
sequence) in the Feature Tables and most often	FT		007	0.07	/fTId=VAR_026041.		
refer to a Homeobox region.	FT	VARIANT	287	287	/FTId=VAR_049585.		
, in the second s	TT FT	MUTAGEN	236	236	Q->E: Abolishes DNA-binding. Does not		
	FT				CDKN2A. {ECO:0000269 PubMed:22206000}.		
	FT FT	CONFLICT	58	58	G -> D (in Ref. 2; AAA58497). {ECO:0000305}.		
	SQ	SEQUENCE	304 AZ	A; 335	94 MW; 0C008479D6995389 CRC64; FSOSSIALH CPSDHMSYDE LSTSSSSCIL ACYDNEFCME		
In the DNA BIND Feature Table entries, the		ASQHHRGHHI	н ннннн	инннн о	QQQHQALQT NWHLPQMSSP PSAARHSLCL QPDSGGPPEL		
position of the Homeoboxs are recorded and will		GSSPPVLCSI SEVNSKPRK	N SSSLGS RTAFT	SSTPT G EQIR E	AACAPGDYG RQALSPAEAE KRSGGKRKSD SSDSQEGNYK LEAEFAHHN YLTRLRRYEI AVNLDLTERO VKVWFONRRM		
be used by the next program to isolate the		KWKRVK	2 GAAARE	CKELV N	VKKGTLLPS ELSGIGAATL QQTGDSIANE DSHDSDHSSE		
sequence of the Homeobox s.	//	HAHL					
Now to extract from the whole protein sequences	VO	u have c	aved i	n a fi	le the sequences of just the Homeobox		
domains One way of doing this (nossibly not the b	yU est)	is to us	e an E		SS nackage program called extractfoot		
This can be found in many places including the Dist	ust)	, is to us		ot W/	aganingan in the Notherlands Co ter		
This can be found in many places, including the Biol	1110	matics s	erver		igennigen in the Netherlands. Go to:		

http://emboss.bioinformatics.nl/

EDIT
aligneopy
angheopypan
Diosed
Find the program extractleat (in the EDIT section), and set it going.
degapsed
entret
extractangn
extractient

Practical 6: Multiple Sequence Alignment	Tuesday 20 February 2018
	Input section
	Select an input sequence. Use one of the following three fields:
Use the Choose File button to upload the SwissProt	1. To access a sequence from a database, enter the USA here:
format sequences from UniProtKB that you saved in	2. To upload a sequence from your local computer, select it here: Browse human_homeobox_proteins.emb
the file:	
human_homeobox_proteins.emb.	
	3. To enter the sequence data manually, type here:
	Additional section
Set Type of feature to extract field to DNA BIND	Amount of sequence before feature to extract
(Make sure you remove the "*")	
(interior you relie to the).	Amount of sequence after feature to extract
	Source of feature to display (*
	Type of feature to extract DNA_BIND
	Sense of feature to extract
Set Value of feature tags to extract to Homeobox*	(default is 0 - any sense, 1 - forward sense, -1 - reverse sense)
(Make sure you append the "*" to ensure hits with.	Minimum score of feature to extract
for example " Homeoboxes ").	
· · · · · · · · · · · · · · · · · · ·	Maximum score of feature to extract 0.0
	Tag of feature to extract *
	Value of feature tags to extract Homeobox*
	Output section
Set the Output sequence format to SwissProt	Output introns etc. as one sequence? No -
(Fasta would do, but SwissProt retains more	
annotation).	Append type of feature to output sequence name? No 🔻
	Feature tag names to add to the description
	Output sequence format SwissProt -
	Run section
Citi 1 di Dup extractfeat	Email address:
Click on the Lead active button to start extractfeat	If you are submitting a long job and would like to be informed by email when it finishes, enter your email address here.
going. Many sequences of 60 amino acids (or so) in	
length will leap into view.	Run extractfeat Reset

Right click the outsed 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

DUTPL	IT 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; 7AA1CEC5BBC0265F CRC64;
	PRKERTAFTK EQIRELEAEF AHHNYLTRLR RYEIAVNLDL TERQVKVWFQ NRRMKWKRVK
11	
ID	NKX25_HUMAN_138_197 Reviewed; 60 AA.
DE	[DNA_contact] Homeobox protein Nkx-2.5 (Cardiac-specific homeobox) (Homeobox protein CSX) (Homeobox protein NK-2 homolog E)
SQ	SEQUENCE 60 AA; 7514 MW; 16EE564D071E5E8A CRC64;
	RRKPRVLFSQ AQVYELERRF KQQRYLSAPE RDQLASVLKL TSTQVKIWFQ NRRYKCKRQR
11	
ID	PAX6_HUMAN_210_269 Reviewed; 60 AA.
DE	[DNA_contact] Paired box protein Pax-6 (Aniridia type II protein) (Oculorhombin)
SQ	SEQUENCE 60 AA; 7447 MW; 075C194DB9F33ED9 CRC64;
	LQRNRTSFTQ EQIEALEKEF ERTHYPDVFA RERLAAKIDL PEARIQVWFS NRRAKWRREE
11	
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;
	PNAVRINFTT KQLTELEKEF HFNKYLTRAR RVEIAASLQL NETQVKIWFQ NRRMKQKKRE
11	
aro	m to be investigated accents multiple sequence Swigs Prot format files

ClustalX, the first multiple alignment program to be investigated, accepts multiple sequence SwissProt format files as input.

Tuesday 20 February 2018

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 ClustalX¹. 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.



Font: 10 Eagles, to something more comfortable. 24 works tolerably well for me.

👂 🗐 🛛 Pairwise Paramel	ters	
ОК		
Fast-Approximate		:
Slow/Accurate Pairwise	Parameters	Fast/Approx Pairwise Parameters
Gap Penalty [1-500]:	3	
K-Tuple Size [1-2]·	1	
it represente (r ep		
Top Diagonals [1-50]:	5	
Window Size [1-50]:	5	

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 dignment tools) work. Intuitively, it should not make a lot of difference how the nitial 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.

Slow-Accurate	*
Slow/Accurate Pairwise Paramet	ers Fast/Approx Pairwise Parameters
Gap Opening [0-100]: 10	
Gap Extend [0-100]: 0.1	
Protein Weight Matrix	
O BLOSUM 30 O PA	M 350 💿 Gonnet 250
○ Identity matrix ○ Us	er defined
Load protein matrix:	
DNA Weight Matrix	
● IUB ○ CL	USTALW(1.6) O User defined
Load DNA matrix:	

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.

Basic Bioinformatics - A Practical User Introduction

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.

Practical 6: Multiple Sequence Alignment	Tuesday 20 February 2018
Before proceeding, save the homeobox sequences in FASTA format, which will better suit the	Format
other MSA programs we will try. Do this by selecting Save sequences as from the File pull	CLUSTAL format
down menu. Deselect CLUSTAL format, select FASTA format.	GCG/MSF format
Change the default file output file name to homeobox_human_full	GDE format
Click OK. A file called homeobox human full.fasta will be created. Take a look to	👿 FASTA format
check it is as you would expect.	

Output Files		
SCLUSTAL format	NBRF/PIR format	Strangely, saving your sequences in FASTA format convinces clustalx that it should
GCG/MSF format	PHYLIP format	now output its alignments in FASTA format. To prevent this, select Output Forma
GDE format	NEXUS format	Options from the Alignments pull down menu. Deselect FASTA format and select
FASTA format		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**!

1	SATB1 HUMAN 645 704	TRPRTK	ISVEALGIL(QSFIQDVG	LYP	DEEAIQT	SAQLDI	PKYTIIK	FONORYL	KHHG
•	SATB2 HUMAN 615 674	PRSRTK	ISLEALGIL(OSFIHDVG	L <mark>Y</mark> P	DQEAIHTL	SAQLDI	PKHTIIK	FONORYHV	KHHG
e	ZFHX3_HUMAN_2145_2204	NKRPRTR	ITDDQLRVL	RQY F -D I N	(N <mark>S P</mark>	SEEQIKEN	ADKSGI	POKVIKHV	FRNTLFKE	RQRN
t	ZFHX4_HUMAN_2084_2143	FKRPRTR	ITDDOLKIL	RAY 🖬 – D 🔳 N	(N <mark>SP</mark>	SEEQIQEM	AEKSGI	SOKVIKH	FRNTLFKE	R <mark>O</mark> RN
ι	SIX1_HUMAN_124_183	GEETSYC	FKEKSR <mark>G</mark> VL	REW <mark>Y</mark> AH	• – – N <mark>P Y P</mark>	SPREKREL	AEATGI	TTTQVSN	FKNRRQRD	RAAE
ζ	SIX2_HUMAN_124_183	GEETSYC	FKEKSRSVL	REW <mark>Y</mark> AH	N <mark>P Y</mark> P	SPREKREL	AEATGI	TTTQVSN	FKNRRQRD	RAAE
	SIX3_HUMAN_206_265	GEQKTHC	FKERTRSLL	REW <mark>Y</mark> LQ	·D <mark>PY</mark> P	NPSKKREL	AQATGI	TPTQVGN	FKNRRQRD	RAAA
I	SIX6_HUMAN_128_187	GEOKTHC	FKERTRHLL	REW <mark>Y</mark> LQ	·D <mark>PY</mark> P	NPSKKREI	AQATGI	TPTQVGN	FKNRRQRD	RAAA
1	SIX5_HUMAN_201_260	GEETVYC	FKERSRAAL	KAC <mark>Y</mark> R <mark>G</mark>	·NR <mark>YP</mark>	TPDEKRRI	ATLTGI	SLTQVSN	FKNRRQRD	RTGA
1	SIX4_HUMAN_223_282	GEETVYC	FKEKSRNAL	KEL <mark>Y</mark> KQ	NR MP	SPAEKRHI	AKITGI	SLTQVSN	FKNRRQRD	RNPS
7	MEIS2_HUMAN_276_338	R <mark>QK</mark> K R GI	F <mark>P</mark> KVATNI MH	RAWLFQHI	'I – H <mark>B A b</mark>	SEEQKKOL	AQDTGI	TILQUNN	FINARRI	VQPM
V	MEIS1_HUMAN_272_334	RHKKRGI	F <mark>P</mark> KVATNI MB	RAW l FQ h I	JT-H <mark>PY</mark> P	SEEQKKOL	AQDTGI	TILQVNN	FINARRI	V <mark>QP</mark> M
c	ME3L1_HUMAN_161_223	RNKKRGI	F <mark>P</mark> KVATNI MB	RAWLFQHI	JS-H <mark>PYP</mark>	SEEQKKOL	AQDTGI	TILQVNN	FINARRI	V <mark>QP</mark> M
5	MEIS3_HUMAN_262_324	RNKKRGI	F <mark>P</mark> KVATNI M	RAWLFQHI	JS-H <mark>PY</mark> P	SEEQKKOL	AQDTGI	TILQVNN	FINARRI	VQPM
	ME3L2_HUMAN_245_307	RNKKRGI	F P K V A T N I M H	RAWLFQHI	'M-H DAD	SEEQKKQL	VQDTGI	TILQUNN	FINARRA	VQPM
	PKNX1_HUMAN_259_321	SKNKRGV	L <mark>P</mark> KHATNV M H	RSWLFQHI	G-HPYP	TEDEKKQI	AAQTNI	TLLQVNN	FINARRII	L <mark>QP</mark> M
	PKNX2_HUMAN_291_350	KRGV	L <mark>P</mark> KHATNI M H	RSWLFQHI	N-H <mark>BAb</mark>	TEDEKROI	AAQINI	TLLQUNN	FINARREI	L <mark>QP</mark> M
	TF2LX_HUMAN_48_111	-EHKKKRKGN	L <mark>P</mark> AESVKI L	RDWMYKHF	(F-KAYP	SEEEKQML	SEKTNI	SLLQISN	FINARREI	LPD
	TF2LY_HUMAN_48_111	-EHKKKRKGN	L <mark>P</mark> AESVKI LE	RDWMYKHF	(F-KAYP	SEEEKQMI	SEKINI	SLLRISN	FINARREI	LPD
	TGIF2_HUMAN_16_79	- AGKRKRRGN	L <mark>P</mark> KESVKI L	RDWLYLHF	(Y-NAYP	SEQEKLSI	SGQTNI	SVLQICN	FINARREL	LPD
f	TGIF1_HUMAN_164_226	KRRRGN	PKESVQIL	RDWLYEHF	Y-NAMP	SEQERAL	SQQTHI	STLOVCN	FINARREL	LPDM
-	IRX3_HUMAN_125_188	- FGDPSRPKN	ATRESTSTL	KAWLNEHF	K-NPMP	TKGEKIMI	AIITKM	TLTOVST	FANARREL	KKE
1	IRX1_HUMAN_125_188	- MGDPGRPKN	ATRESTST	KAWLNEHE	(K-NPYP	TKGEKIMI	AIITKM	TLTOVST	FANARREL	KKE
_	IRX4_HUMAN_143_204	GT-RRKN	ATRETTSTL	KAWLQEHF	K-NPMP	TKGEKIMI	AIITKM	TLTOVST	FANARREL	KKEN
e	IRX6_HUMAN_144_207	- SGAGRRKN	ATRETTSTL	KAWLNEHF	K-NPMP	TKGEKIMI	AIITKM	TLTOVST	FANARREL	KKE
	IRX5_HUMAN_113_175	DPAYRKN	ATRDATATL	KAWLNEHF	(K-NPYP	TKGEKIMI	AIITKM	TLTOVST	FANARREL	KKEN
S	IRX2_HUMAN_112_175	- NDPAYRKN	ATRDATATL	KAWLNEHF	(K-NPYP	TKGEKIMI	AIIKM	TLTOVST	FANARREL	KKE
n	MKX_HUMAN_71_132	VRHKRQ	ALQDMARPL	KQWLIKHE	(D-NPTP	TKTERIL	ALGSQN	TLVQVSN	FANARREL	NTV
1	PBX3_HUMAN_235_297	ARRKERN	SKQATEILI	NEIFISHI	IS-NALA	SEEAREEL	AKKUSI	IVSOVSN	FGNKKIRI	KKNI
)	PBX1_HUMAN_233_295	ARRARK	NKQATEILI	NEIBISHI	12-NALA	SEEAREEL	AKKCG	IVSQVSN	FGNKKIRI	KKNI
-	PBA2_HUMAN_244_306	ARRARK	SKQATEVI	NEIPISHI	72 - N FIF	SEEANEEL	ANNOGI	TROUCH	FGNKKIKI	
	PBA4 HUMAN 210 272	AKKAKKK	DRUBER	NEIPISHI		SEEANE SI	AKNGGI	TISU SN		
	CUVI UMAN 1244 1303	PLGLKSKN	I A DE DE E E E E E	KD3 MOO-		SCHEREN D	ATOINI	KTCTUIN	CHNINARV.	D D D I
	CUX1_HOMAN_1244_1303	INKPRVV		DVANOT		SOOTIPI	C DOL NI	KTNTUIN.		D D D M
	HNE6 HIMAN 385 444	DKK DR LU	TOVORRTI	HYIRKE	NKRD	SKELOIT		FLSTUSNE	EMNARRES	
	ONEC2 HUMAN 426 485	OKKSRLV	TDLORRT	FAIRKE	NKRD	SKEMOIT	SOOLCI	FLTTUSNE	EMNARRES!	LEKW
	ONEC3 HUMAN 414 473	DKKORLV	TDLORRTL	TATEKE	NKRD	SKEMOUT	SOOLCI	FLNTUSNE	EMNARR.	INRW
1	PO5F1 HUMAN 230 289		ENRVRONT	RNI. RLO		LOOISH	ACOLGI	EKDVURUS	FCNRROKG	KRSS
	POSPI HORAN 250 205	C.C.		1111112	U . IL	144101		DIED I I I I I I		
,)))	CIC								
S		1		20	30	40		50	60	
۰I									1	
1		1	. .							_
			and the state		and the second	a second				

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!

05F1 HUMAN 230 289	R	KRK	TS	ENRVR	GN	LENL	LOC	ΧP	LOOI	ΙSΗ	IAOO	LG	EKDV	R	IWEC	NRRO	KGKI	RSS
5F1B HUMAN 229 288	-AR	KRK	TS	ENRVR	GN	ENL	LOCE	ΚP	TLO-I	ΙSΗ	IAOO	LG	EKDV	R	WEC	NRRO	KGKI	RSS
05F2 HUMAN 210 269	G	KWR	AS	RERRIG	NS	EKF	ORCE	KP	TP001	ISH	LACC	LO	OKDV	R	WEY	NRSK	MGS	RPT
02F2 HUMAN 297 356	R	RKK	TS	ETNVR	FA	EKS	LANC	KP	SEEL	LL	LARO	H	EKEV	R	WEC	NRRC	KEK	RIN
02F1 HUMAN 379 438	R	RKK	S	ETNIR	VA	EKS	LENC	KP	SEEL	ITΜ	LADO	N	EKEV	R	WEC	NRRO	KEK	RIN
02F3 HUMAN 281 340	K	RKK	S	ETNIR	LT	RKR	ODN	KP	SEE	LSM	TARO	S	RKEV		WEC	NRRO	K E K	T N
03F2 HUMAN 354 413	K	RKK	S	EVSVK	GA	E SH	LKCP	KP	SAOR	ITS	LADS	1.0	RKRV		WEC	NRRO	KEK	RMT
03F3 HUMAN 406 465	K	RKK	S	EVSVK	GA	E S H	LKCP	K D	SAORI	TTN	LADS	0	EKEV		WEC	NRRO	KEK	RMT
03F1 HUMAN 339 398	K	RKK	S	EVGUK	GA	E S H	LKCP	K D	SAHEI	TTG	LADS	L O	EKEV		WEC	NRRO	KEKI	RMT
03F4 HUMAN 278 337	K	RKK	S	EVSVK	GV	RTH	LKC	KP	AAOEI	LSS	LADS	ō	RKEV	R	WEC	NRRO	KEK	RMT
PTT1 HUMAN 214 273	K	RKR	T	STAAK	DA	RRH	GEON	IK P	SOF	IMR	MARE	N	EKEV		WEC	NRRO	REKI	VK
04F2 HUMAN 345 404	K	KRK	S	AADEK	RS	RAY	ATOP	RP	SEKI	TAA	TAEK	D	KKNV		WEC	NORO	KOK	RMK
04F3 HUMAN 274 333	R	KRK	S	AADEK	RS	RAY	ATOP	RD	SEKI	TAA	TAEK	D	KKNV		WEC	NORC	KOK	RMK
04F1_HUMAN_356_415		KRK	S	AADEK	RS	RAY	AVOD	RD	SEKI	TAA	TAEK	D	KKNV		WEC	NORC	KOK	RMKF
06F1 HUMAN 234 293	K	RKR	S	FTDOAT	FA	NAY	EKND	D.D	TGORI	TTE	TAKE	N	DREV	URI	WEC	NRRO	TIK	NTS
06F2 HUMAN 607 666	K	RKR	S	FTPOAL	ΕT	NAH	EKNT	HD	SCOR	MT F	TAEK	N	DREV		WEC	NKRO	ALK	NTT
PAX6_HUMAN_210_269	T.		S	FTOROT	EA	RKE	ERTH	YD	DVFA	ER	LAAK	D	DRAR		WES	NRRA	KKR	REE
DAX4 HUMAN 170 229		HEN	Т	FSPSOA	EA	FKE	ORCO	Y D	DSVA	CK	LATA	S	DRDT		WES	NRRA	RER	ROE
MIXI. HUMAN 86 145	0	RRK	ŝ	FSAROL	OT.	RT.V	RRTR	V D	DIHL	FR	LAAL	T.	DESR		WEO	NRRA	KSRI	05
DRODI HUMAN 69 128	R	RRH	T	RSDUAT	RO	RSA	GRNC	V D	DIWA	RS		i a	SFAR		JWEO	NRRA	RORI	08
CSC2 HUMAN 126 185	T	RRH	÷,	FSFFOI	0.2	RAT.	VONC	V D	DUST	FR	LACR	R	REFR		JWF K	NRRA	RER	HOK
GSC HUMAN 160 219	K	RRH	T T	FTDEOL	EA	FNT.	OFTR	Y D	DVGT	FO	LARK	U H	REEK	URI	WEK	NRRA	KKR	OK
DTTX2 HIMAN 85 144	0	RROI	н Н	FTSOOL	OR	RAT	ORNR	V D	DMST	FF	TAVW	N	TRAR		WEK	NRRA	RER	RR
DTTY3 HUMAN 62 121	ŏ	RROE	н Н	FTSOOT	OF	RAT	ORNR	V D	DMST	TTT	TAVW	N	TRAR		WEK	NRRA	RER	RF
DTTY1 HUMAN 89 148	ŏ	RROE	н Н	FTSOOT	OF	RAT	ORNR	V D	DMSM		TAVW	N	TEDR		JWFK	NRRA	RER	RR
OTVI HUMAN 38 97	ŏ		T	TREAT	DV.	RAL.	AKTR	V D	DIEM		UALK	M	DEGR		INE K	NRRZ	RCR.	00
OTY2 HUMAN 38 97	ŏ		.	FTRAOT	DV	PAL.	AKTR	V D	DIFM		UALK	M	DECR		INE K	NPPZ	N C D	00
CDV UIIMAN 39 99	ŏ		.	PERCOT		PAL	AKTO	V D	DUVA		UALK	M	DECE		IN P.K	NDDB	N C D	0P
DMBX1 HUMAN 71 130		RRGE	a a	ETAOOI	R A	RKT	OKTH		DUUM	E R R	LAMC	M	DRAR		INE K	NRRZ	KER	KO
DAY3 HUMAN 219 278	ŏ	RRCE		TAPOT	ER.	PRA	FRTH	V D	DIVT		LAOR	K	TEAR		INE'S	NRRZ	RERI	03
DAX7 HUMAN 217 276	ŏ		.	TAROL	R R	PKA	FRTH	V D	DIVT		LAOR	K	TRAR			NDDZ	DUDI	03
DHY2B HIMAN 98 157	ŏ	RRTE	- ÷	TO 6 2 T	KE	RRV	AFTH	V D	DIVT		LALK	D	TRAR			NRRZ	RERI	OF
DHY2A HUMAN 90 149	ŏ	RRTE	- ÷	ETS AOT	KE	PRU	AFTH	V D	DIVT		LALK	E D	TRAR		JWE O	NRRA	KERI	08
DRCY HUMAN 33 92	ŏ	RRME	.	TLOOL	R B	PAU	AOTH	V D	DVFT		LAMK	N	TRAR			NRRS	K R	TT
ALV3 HUMAN 153 212	K		.	C S T F O I	D D	PKU	OKTH	V D	DUVA		LALR	D	TPAR			NDDZ	N N D	
ALXA HUMAN 214 273	K		.			DYU	OKTH	V D	DUVA		LAMP	D D	TPAD			NDDA	V V D	
AT V1 UIMAN 132 191	V		.	CT CT OT		DVU	OFTH	VD.	DUVU					ň		NDDX	P P D	
ARY HUMAN 328 387	0		.	TOYOT	R R	PRA	OKTH	V D	DVFT		LAMR	D	TRAR			NPPZ	N N R	RR
SHOX HUMAN 117 176	ŏ	RRSE	n n	TLEOL	NF	R R L	DETH	V D	DAFM			c	SPAR			NRRA	RCRI	OF
SHOK HOMAN III/ I/U			14	C T D C V L	14.12		DEIN		DAFE		- MIX			M	inc v			ΥY.
	C		-		-			-				_	10	_				
	1			10		20			30			0			50		(0
																1 B.		
		1 I	ь.			L. I			-			.		h.,			1.0	
					-													

Tuesday 20 February 2018



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.

Tuesday 20 February 2018

The **Prosite** motif database uses **Patterns** to represents 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** 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 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**.

LS	DR	L	Ν	L	S	D	Q	Q	V	K	Ι	W	F	Q	Ν	R	R	М	K	K	K
LS	NR	L	Ν	L	S	D	Q	Q	V	K	Ι	W	F	Q	Ν	R	R	М	K	K	K
LΑ	ΑT	L	G	L	S	Е	R	Q	V	K	Ι	W	F	Q	Ν	R	R	A	K	Е	R
LΑ	AN	L	G	L	Т	Е	R	Q	V	K	Ι	W	F	Q	Ν	R	R	A	K	Е	R
LΑ	VN	L	G	L	S	Е	R	Q	V	K	Ι	W	F	Q	Ν	R	R	A	K	Ε	R
ΙA	AS	L	Q	L	Ν	Е	Т	Q	V	K	Ι	W	F	Q	Ν	R	R	М	K	Q	K
ΙA	ΑT	L	Ε	L	Ν	Е	Т	Q	V	K	Ι	W	F	Q	Ν	R	R	М	K	Q	K
ΙA	NC	L	Η	L	Ν	D	Т	Q	V	K	Ι	W	F	Q	Ν	R	R	М	K	Q	K
MA	ΝL	L	Ν	L	Т	Е	R	Q	Ι	K	Ι	W	F	Q	Ν	R	R	М	K	Y	K
MA	ΝL	L	Ν	L	S	Е	R	Q	Ι	K	Ι	W	F	Q	Ν	R	R	М	K	Y	K
MA	ΝL	L	Ν	L	Т	Е	R	Q	Ι	K	Ι	W	F	Q	Ν	R	R	М	K	Y	K
LΑ	VM	L	Ν	L	Т	Е	R	Η	Ι	K	Ι	W	F	Q	Ν	R	R	М	K	W	K
ΙA	ΝA	L	С	L	Т	Е	R	Q	Ι	K	Ι	W	F	Q	Ν	R	R	М	K	W	K
ΙA	ΗA	L	С	L	Т	Ε	R	Q	Ι	K	Ι	W	F	Q	Ν	R	R	М	K	W	K
ΙA	ΗA	L	С	L	Т	Ε	R	Q	Ι	K	Ι	W	F	Q	Ν	R	R	М	K	W	K
ΙA	ΗT	L	С	L	Т	Е	R	Q	Ι	K	Ι	W	F	Q	Ν	R	R	М	K	W	K
ΙA	NA	L	С	L	Т	Е	R	Q	Ι	K	Ι	W	F	Q	Ν	R	R	М	K	W	K
VS	ΗA	L	G	L	Т	Е	R	Q	V	K	Ι	W	F	Q	Ν	R	R	М	K	W	K
VS	ΗA	L	G	L	Т	Ε	R	Q	V	K	Ι	W	F	Q	Ν	R	R	М	K	W	K
		40										50									
				_					_												

	GNLKW	YAWKN	ΤR	FGD	SW	IV.	RTD	LAF	SG	AKE	DK	PEES	2WPS	VRT	KSA	LHM	TPEC	I-CKF	-STGK	ZHX1 HUMAN 660 722
	GQ	YALKN	SR	FGD	RW	VVI	RPE	L <mark>P</mark> F	TG	MAQ	DS	NQDY	2WPS	VQT	RQL	RHL	TAQ	VSCKF	<mark>PG</mark> K	ZHX3 HUMAN 764 823
~	GTVK <mark>W</mark>	CL <mark>LK</mark> T	ΝR	FKE	RW	IVI	RΤΕ	LVF	ΤG	AAK	DQ	PQE 1	DWP1	ART	RST	VHL	SQE <mark>(</mark>	A-IAF	-S <mark>P</mark> SF	ZHX2 HUMAN 628 690
Ο	GQ	YALKH	ΤR	FGD	QW	II	RPE	L <mark>P</mark> F	TG	EQI	QK	REDY	2 <mark>WA</mark> F	LQC	KSF	LAI	TKE	KTKRF	-QR <mark>Q</mark> F	HOMEZ HUMAN 355 415
	GIVH <mark>I</mark>	Y <mark>rcq</mark> r	ΗR	FSD	K W	IKI	RSΕ	LAF	ΤG	IEV	ΥR	DAE	2 <mark>F P</mark> E	LQS	KAS	IAH	TKE	SDRKK	– – T <mark>P</mark> A	ZHX2 HUMAN 439 501
C	SKSNQ	YNQRN	ΤR	FSD	K W	IKI	(<mark>G</mark> E	L T P	TG	MKI	IR	DSEI	<u>P</u> FPH	LKN	KVS	LAE	TKE	IRAKF	SFG	ZHX1 HUMAN 464 526
+h	LK	Y H C R N	RR	FSD	KW	V RJ	[R <mark>E</mark>	L <mark>S</mark> I	TG	TKV	ΕH	QSEV	P P C	CRN	KGSI	LSA	SHE	Y <mark>K</mark> NKF	ASI	ZHX3_HUMAN_494_553
u	MEQA <mark>V</mark>	KLRDS	RR	FSE	SW	ID.	rr <mark>e</mark>	LSF	ΤK	RVE	DR	QAEI	F <mark>P</mark> I	LKS	EDSI	VKI	TQG	KFKEÞ	<mark>PC</mark>	ZHX2 HUMAN 530 591
11 7	K E E K <mark>M</mark>	KSKAL	KK	FΤΕ	AW	ΙD	RR <mark>E</mark>	L T F	ΤK	RAQ	NR	DEEI	SVL <mark>I</mark>	LNS	QAS	LRV	TAE	KFKEF	<mark>P</mark> C	ZHX1_HUMAN_569_630
w	AEE	K-KVN	RR	FSE	SW	ID.	RR <mark>E</mark>	MT F	TK	RSE	DR	DEEI	PLPI	AQN	ESSI	LRA	APE	KYKEF	– – T <mark>P</mark> T	ZHX3 HUMAN 612 671
of	AI	YICM <mark>K</mark>	RR	FGK	SF	VA.	ΚID	V ₩ F	FW	SL	ΕL	KKE I	PY PS	HKK	KDY	KQF	YEE	YE <mark>G</mark> R <mark>S</mark>	–– <mark>P</mark> KK	DNP2_HUMAN_1043_1102
01	D	K <mark>KC</mark> VR	KR	FSN	SH	ΙA	ζSD	LWF	LWJ	AAS	ΕK	RRE I	ΡΥ <mark>ΡΊ</mark>	NKQ	TKY	KSF	YEA	HEDDS	LDPKG	ADNP_HUMAN_754_814
	SWT <mark>P</mark> E	L <mark>K</mark> HGV	QR	FSA	ΙW	ΙK	Ε <mark>Ο</mark>	ΥTΕ	AK.	SAQ	TV	MSE1	PY PI	NKF	LNT	NPL	LDN-	TYNA <mark>7</mark>	-NSI <mark>F</mark>	ZHX1 HUMAN 284 346
	SW	LKQGI	QR	FΤΑ	ΙW	LK	ΕQ	ΥPΕ	ΤK	TVV	СҮ	KAEI	PY PI	HKF	KNS	NSF	MDS-	TYNA	-SSI <mark>F</mark>	ZHX3 HUMAN 304 363

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 AKRARTSFTAEQLQVMQAQFAQDNNPDAQTLQKLADMTGLSRRVIQVWFONCRARHKKH	T
LHX9 ⁻ HUMAN ⁻ 267 ⁻ 326T <mark>KR</mark> M <mark>RT</mark> S ^F KHH <mark>QL</mark> RTMKSYFAINHN ^P DAKDLK <mark>QLA</mark> QK <mark>TGLT</mark> KRVLQ <mark>VWFQNARAKFRR</mark> N	L
LHX2_HUMAN_266_325T <mark>KRMRT</mark> SFKHHQLRTMKSYFAIN <u>HNP</u> DAKDLKQLAQKTGLTKRVLQVWFQNARAKFRRN	L
DPRX_HUMAN_16_75SHRKRTMFTKKOLEDLNILENENPYPNPSLOKEMASKIDIHPTVLOVWFKNHRAKLKKA	к Г
ZEB1_HUMAN_581_640 -NLS <mark>PSQPPL</mark> KNL-LSLLKAYYALNAQP <mark>S</mark> AEELSK <mark>IA</mark> DS <mark>VNLPL</mark> DV <mark>V</mark> KKWFEKMQAGQISV	Q
ZEB2_HUMAN_644_703 -GMTSPINPYKDH-MSVLKAYYAMNMEPNSDELLK <mark>IS</mark> IA <mark>VGLPQ</mark> EFVKEWFEQ RK VYQYSN	s p
ZHX1_HUMAN_777_832KEKTG-TAILKDYYLKHKFLNEQDLDELVNKSHMGYEQVREWFAERORESLG	IELF
ZHX3_HUMAN_835_894F <mark>PPG</mark> LLV I APGNRELLQDYYMTHKMLYEEDLONLCDK T QM <mark>S</mark> SQ QV KQ WF AE K MGEET R A	v C
HOMEZ_HUMAN_451_510 TPPLPIPPPPDIQPLERYWAAHQQLRETDIPQLSQASRLSTQQVLDWFDSELPQPAEV	V
NANGN HUMAN 102 161 QYPEKRLVSKSLMHTLWAKEKLNRCPTIQESLSLSFEDDMTHKQISQMECKTEKKYNKE	M

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.

Tuesday 20 February 2018

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/

Offerend have in a subschiere of manufacture	
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	Clustal Omega New MSA tool that uses seeded guide trees and HMM profile-profile techniques to generate alignments. Suitable for medium-large alignments.
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.	Kalign Very fast MSA tool that concentrates on local regions. Suitable for large alignments.
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	MAFFT MSA tool that uses Fast Fourier Transforms. Suitable for medium-large alignments.
workstation and definitely available as a web service. You might have already used Jalview as an alignment viewer when investigating Pfam and/or Jpred .	MUSCLE Accurate MSA tool, especially good with proteins. Suitable for medium alignments.
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	MView Transform a Sequence Similarity Search result into a Multiple Sequence Alignment or reformat a Multiple Sequence Alignment using the MView 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.	T-Coffee O Consistency-based MSA tool that attempts to mitigate the pitfalls of progressive alignment methods. Suitable for small alignments.
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.	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.

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**.

	ce Alignment	Tuesday 20 February 2018
MUSCLE multiple sequence alignment method with reduced DUTATION time and space SUPACION complexity	So the plan now is to use MUSCLE ⁴ to align ag aligned with ClustalX . MUSCLE works in a way more care in the generation of the Guide Tree construction of the final multiple alignment ⁵ . Pa MUSCLE should do a better job than ClustalX . certainly be different. I leave you to judge for yours	gain the homeobox sequences previously ay similar to clustalX but it takes rather e used to control the order of pairwise articularly for more difficult alignments, The alignment you will generate here will selves whether it is better.
		STEP 1 - Enter your input sequences
Start by requesting to	Launch MUSCLE	Enter or paste a set of sequences in any supported format:
Use the Browse butte homeobox sequences, I	on to upload the file containing the FASTA for homeobox_human.fasta . This file should not includes around position 30 .	nat ded
the sequences with a me		
the sequences with a me		Or upload a file: Browse homeobox_human.fasta
the sequences with a me	Take a look at	Or upload a file: Browse homeobox_human.fasta

the programs that their creators deemed sensible to make available⁶?

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

0750 0 0 1 0	
STEP 2 - Set your F	arameters
OUTPUT FORMAT:	ClustalW ^
OUTPUT TREE	OUTPUT ORDER
CON OF THEE	CONT OT ONDER
none	1 aligned
	*)[

(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?

ClustalW Pearson/FASTA 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 Jalview⁸. 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?

Basic Bioinformatics - A Practical User Introduction

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

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

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

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.

A widely used java alignment editor and viewer.

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.

Practical 6: Multiple Sequence Alignment		Tuesday 20 February 2018
	ARX HUMAN 328 387	QRRYR-TTFTSYQLEELERAFQKTHYPDVFTREELAMRLDLTEARVQVWFQNRRAKWR
	ALXI HUMAN 132 191	KRRHR-TTFTSLQLEELEKVFQKTHYPDVYVREQLALRTELTEARVQVWFQNRRAKWR
	ALX4 HUMAN 214 273	KRRNR-TTFTSYQLEELEKVFQKTHYPDVYAREQLAMRTDLTEARVQVWFQNRRAKWR
	ALX3_HUMAN_153_212	KRRNR-TTFSTFQLEELEKVFQKTHYPDVYAREQLALRTDLTEARVQVWFQNRRAKWR
	ISL1_HUMAN_181_240	TTRVR - TVLNEKQLHTLRTCYAANPRPDALMKEQLVEMTGLSPRVIRVWFQNKRCKDK
After considering these enigmas, or before if	ISL2_HUMAN_191_250	TTRVR - TVLNEKQLHTLRTCYAANPRPDALMKEQLVEMTGLSPRVIRVWFQNKRCKDK
	LHX9_HUMAN_267_326	TKRMR - TSFKHHQLRTMKSYFAINHNPDAKDLKQLAQKTGLTKRVLQVWFQNARAKFR
you prefer Click on the Submit button and	LHX2_HUMAN_266_325	TKRMR - TSFKHHQLRTMKSYFAINHNPDAKDLKQLAQKTGLTKRVLQVWFQNARAKFR
	LHX8_HUMAN_225_284	AKRAR - TSFTADQLQVMQAQFAQDNNPDAQTLQKLAERTGLSRRVIQVWFQNCRARHK
sit back to admire muscle in action.	LHX6_HUMAN_219_278	AKRAR - TSFTAEQLQVMQAQFAQDNNPDAQTLQKLADMTGLSRRVIQVWFQNCRARHK
	ZFHX3_HUMAN_2641_2700	DKRLR- I I I I PEQLEI LYQKYLLDSNP I RKMLDH I AHEVGLKKRVVQVWFQN I RARER
	ZFHX4_HUMAN_2560_2619	UKKLK-IIIIPEULEILYEKYLLUSNPIKKMLDHIAKEVGLKKKVVUVWFUNIKAKEK
	ZFHX2_HUMAN_1857_1910	UKKLK-ITILPEQLEILTKWIMQUSNPTKKMLUCISEEVGLKKKVVQVWFQNTKAKEK
	ZFHX2_HUMAN_2005_2124	UKKTK-TUMSSLULKIMKAUTEATKTPTMUECEVLUEETULPKKVTUVWFUNAKAKEK
	ZENXA UIMAN 2004 2043	
The alignment that is computed is	LMX1A HUMAN 195 254	- PKRPR-TTI TTOORRAEKASEEVSSKPCRKVRETI AAETGI SVRVVOVWEONORAKMK
The diffinitent that is computed is,	LMX1B HUMAN 219 278	PKRPR - TTI TTOORRAFKASEEVSSKPCRKVRETI AAETGI SVRVVOVWEONORAKMK
superficially at least, similar to that offered	LHX1 HUMAN 180 239	RRGPR - TTTKAKOI FTI KAAFAATPKPTRHTREOI AOFTGI NMRVTOVWFONRRSKER
hy ClustelV	LHX5 HUMAN 180 239	RRGPR - TTIKAKOLETLKAAFAATPKPTRHIREOLAOETGLNMRVIOVWFONRRSKER
by Clustala.	LHX4 HUMAN 157 216	AKRPR-TTITAKOLETLKNAYKNSPKPARHVREOLSSETGLDMRVVOVWFONRRAKEK
	LHX3_HUMAN_157_216	AKRPR-TTITAKQLETLKSAYNTSPKPARHVREQLSSETGLDMRVVQVWFQNRRAKEK
		510/
	HOMEZ_HUMAN_451_510	EVV
The alignment is invitatingly galit into two	ZHX1_HUMAN_777_832	LGIELF
The angliment is initiatingly split into two	ZHX3_HUMAN_835_894	KAV
sections. A nice extra parameter might have	HUMEZ_HUMAN_263_324	
	ZHX2_HUMAN_30/_363	TSW
been How wide would you like your	ZHX1_HUMAN_284_346	VSWTPE
alignment to be"? A problem with the format	ZEB2 HUMAN 644 703	SNS
anguinent to be ? A problem with the format	ZEB1 HUMAN 581 640	SV0
rather than the program, to be fair.	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:

So	click th	ne Show	Colors	s button	at the	top
of	the pag	e and the	ry to liv	ve with	the pai	n of
suc	h gross	Trans-	Atlantic	inept s	pelling	in a
Eur	ropean	site!!!	Good	Grief!	They	get
eve	rywher	e!!				

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.

--TTRVR-TVLNEKQLHTLRTCYAANPRPDALMKEQLVEMTGLSPRVIRVWFQNKRCKD --TTRVR-TVLNEKQLHTLRTCYAANPRPDALMKEQLVEMTGLSPRVIRVWFQNKRCKD ISL1_HUMAN_181_240 ISL2_HUMAN_191_250 LHX9_HUMAN_267_326 LHX2_HUMAN_266_325 --TKRMR-TSFKHHQLRTMKSYFAINHNPDAKDLKQLAQKTGLTKRVLQVWFQNARAKF --TKRMR-TSFKHHQLRTMKSYFAINHNPDAKDLKQLAQKTGLTKRVLQVWFQNARAKF LHX8_HUMAN_225_284 --AKRAR-TSFTADQLQVMQAQFAQDNNPDAQTLQKLAERTGLSRRVIQVWFQNCRARH LHX6_HUMAN_219_278 --AKRAR-TSFTAEQLQVMQAQFAQDNNPDAQTLQKLADMTGLSRRVIQVWFQNCRARH ZEHX3 HUMAN 2641 2700 -- DKRI R-TTTTPEOLETI YOKYI I DSNPTRKMI DHTAHEVGI KK ZFHX4 HUMAN 2560 2619 --DKRLR-TTITPEQLEILYEKYLLDSNPTRKMLDHIAREVGL ZFHX2_HUMAN_1857_1916 --DKRLR-TTILPEQLEILYRWYMQDSNPTRKMLDCISEEVGLI ZFHX2 HUMAN 2065 2124 - - QRRYR-TQMSSLQLKIMKACYEAYRTPTMQECEVLGEEIGLPH ZEHX3_HUMAN_2944_3003 PGOKRER - TOMTNLOL KVLKSCENDYRTPTMLECEVLGNDTGLPKRVVOVWEONARAKE ZFHX4 HUMAN 2884 2943 --HKRFR-TQMSNLQLKVLKACFSDYRTPTMQECEMLGNEIGLPKRVVQVWFQNARAKE LMX1A_HUMAN_195_254 --PKRPR-TILTTQQRRAFKASFEVSSKPCRKVRETLAAETGLSVRVVQVWFQNQRAKM IMX18_HUMAN_219_278 --PKRPR-TILTTQQRRAFKASFEVSSKPCRKVRETLAAETGLSVRVVQVWFQNQRAKM --RRGPR-TTIKAKQLETLKAAFAATPKPTRHIREQLAQETGLNMRVIQVWFQNRRSKE LHX1 HUMAN 180 239 LHX5_HUMAN_180_239 LHX4_HUMAN_157_216 --RRGPR-TTIKAKQLETLKAAFAATPKPTRHIREQLAQETGLNMRVIQVWFQNRRSKE --AKRPR-TTITAKQLETLKNAYKNSPKPARHVREQLSSETGLDMRVVQVWFQNRRAKE --AKRPR-TTITAKQLETLKSAYNTSPKPARHVREQLSSETGLDMRVVQVWFQNRRAKE LHX3 HUMAN 157 216 HOMEZ_HUMAN_451_510 EVV ---ZHX1_HUMAN_777_832 ZHX3_HUMAN_835_894 LGIELF RAV ---ISW--HOMEZ HUMAN 55 114 ZHX2_HUMAN_263_324 ISWSPE ZHX3_HUMAN_304_363 TSW--ZHX1_HUMAN_284_346 ZEB2_HUMAN_644_703 VSWTPF SNS - - -ZEB1 HUMAN 581 640 SVQ---NANGN HUMAN 102 161 ZHX1 HUMAN 569 630 KEM---LKEEKM ZHX2 HUMAN 530 591 SMEQAV ZHX3_HUMAN_612_671 AEE - -7HX2 HUMAN 439 501 RGIVHI ZHX3_HUMAN_494_553 NLK ZHX1_HUMAN_464_526 NSKSNQ

HGQ-

PLEASE NOTE: Showing colors on large alignments is slow.

- - QRRYR-TTFTSYQLEELERAFQKTHYPDVFTREELAMRLDLTEARVQVWFQNRRAKW - - KRRHR-TTFTSLQLEELEKVFQKTHYPDVYVREQLALRTELTEARVQVWFQNRRAKW

--KRRNR-TTFTSYQLEELEKVFQKTHYPDVYAREQLAMRTDLTEARVQVWFQNRRAKW

--KRRNR-TTFSTFQLEELEKVFQKTHYPDVYAREQLALRTDLTEARVQVWFQNRRAKW

ARX HUMAN 328 387

ALXI_HUMAN_132_191 ALX4_HUMAN_214_273

ALX3 HUMAN 153 212

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:

HOMEZ HUMAN 355 415

homeobox human muscle.aln

RVVOVWFONTRARF

VVQVWFQNTRARE

VVQVWFQNTRARE

(RVTOVWEONARAKE

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:



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.



Tuesday 20 February 2018

10

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?

ŀ	<i>ZHX3_HUMAN_304_36</i> SSI <mark>P</mark> T-Y	ΝA	A
ŀ	<i>ΖΗΧ1_ΗUMAN_284_34</i> (<u>N</u> SI <mark>P</mark> T-Υ	ΝA	А
ŀ	ZEB2_HUMAN_644_703 <mark>G</mark> MT S <mark>P</mark> - T	Ν <mark>Ρ</mark>	Υ
	<i>ZEB1_HUMAN_581_64</i> (NLS <mark>P</mark> S-Q	ΡP	L
,	NANGN_HUMAN_102_1QY <mark>P</mark> E <mark>K</mark> -R	LV	S
	ZHX1_HUMAN_569_63(<mark>PQK</mark> F <mark>K</mark>	ΕK	Т
Ľ	ZHX2_HUMAN_530_59: PQK FK	ΕK	Т
'	<i>ZHX3_HUMAN_612_67</i> :Т <mark>Р</mark> ТКҮ-К	ΕR	A
Ì	<i>ZHX2_HUMAN_439_50</i> :T <mark>P</mark> AS <mark>D</mark> -R	КΚ	Т
	<i>ZHX 3_HUMAN_</i> 494_55、 A S I Y <mark>K</mark> - N	КΚ	s
ļ	ZHX1_HUMAN_464_52(S F <mark>G</mark> I <mark>R</mark> - A	КΚ	Т
	HOMEZ_HUMAN_355_4 Q <mark>RQ</mark> RKTK	RК	Ŧ
	ZHX2_HUMAN_628_69(S <mark>P</mark> S <mark>P</mark> A-I	ĀΚ	s
	ZHX3_HUMAN_764_82、 <mark>PGK</mark> VS-C	КΚ	Т
ŀ	<i>ZHX1_HUMAN_660_72:</i> ST <mark>G</mark> KI-С	КK	Т
ł	ADNP2_HUMAN_1043_: <mark>PKK</mark> Y <mark>E</mark> - G	RS	Υ
ł	ADNP_HUMAN_754_81.LD <mark>PKG</mark> H <mark>E</mark> -D	DS	Υ

You can also Select and **Cut** sequences in a way similar to that you employed with clustalx. I could not resist it! removed all the ugly sequences that caused the ZHA3_HUMAN_434_33. gaps at the start and finish ADNP2_HUMAN_1043_ of the alignment, and the HNF6_HUMAN sequence that messed up ONEC2_HUMAN_4. column 8 (just select their SIX6 HUMAN names and then select **Cut** or Delete from the Edit SXX5 HUMAN 201 260GE menu). I achieved the gapfree beautiful alignment illustrated.



30

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.

10

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 epect 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 menues with one option and the chance to create data files I cannot use. Make your own mind up.

DPJ - 2017.12.23

Model Answers to Questions in the Instructions Text.

Notes:

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

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

Summary:

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

The other will start:

Full Answer:

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

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

From your investigations of Multiple Sequence Alignment

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

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

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

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

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

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

The purpose of saving the **Guide Tree**(\mathbf{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 – 2017.12.23

Discussion Points and Casual Questions arising from the Instructions Text.

Notes:

Work in progress I fear.

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

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

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

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

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

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	Pattern	Interpretation
Position	Element	
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 EXCEPT 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 <u>ONLY</u> 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.

			l I	Categories	Databases
				proteomics	Databases
				protein sequences and identification	UniProtKB • functional information on proteins • [more]
	For a swift experience of using Prosite , try the fo	ollowing. G	o to	proteomics experiment function analysis	UniProtKB/Swiss-Prot • protein sequence database • [more]
	the ExPASy ¹⁰ site at:	_		sequence sites, features and motifs	STRING • protein-protein interactions • [more]
	the EATTAby Site ut.			protein modifications	SWISS-MODEL Repository • protein structure homology
				protein structure	models * [more] PROSITE * protein domains and families * [more]
				similarity search/alignment	ViralZone • portal to viral UniProtKB entries • [more]
	httn•//www.exnasy.org			genomics	neXtProt • human proteins • [more]
	http://www.capasy.org			structure analysis	
				systems biology	CAZy • Classification of carbohydrate-active enzymes • [more]
				evolutionary biology	EMBnet services • bioinformatics tools, databases and
	Select proteomics from the list of Categories			population genetics	courses • [more]
	erect proteonnes nom me not of eutegones.			transcriptomics	Givenable 3 and the structures of diven-related
				biophysics	molecules • [more]
				imaging	GlyTouCan • international glycan structure
	Select PROSITE from the Databases section			IT infrastructure	HAMAP • UniProtKB family classification and annotation
				medicinal chemistry	• [more]
			L	glycomics	MatrixDB • protein-glycosaminoglycan
_				-	
	Home ScanProsite ProRule Documents	Downloads Links	Funding	9	
	Detabase of protein domains, families and func	tional sites			
	Database of protein domains, families and func	lional siles		Click on the S	ScanProsite link at the top
DRO	NTE consists of documentation entries describing protein domains, families and functional sites as	well as associated as	tterne end	of your page	1
profile	st consists of documentation entries describing protein domains, families and functional sites as as to identify them [More/ Beferences / Commercial users]	well as associated pa	illerns and	of your page.	
PRO	SITE is complemented by ProRule, a collection of rules based on profiles and patterns, which increa	ases the discriminator	y power of		
profile	es and patterns by providing additional information about functionally and/or structurally critical amin	no acids [More].			
Relea	ase 2017 10 of 25-Oct-2017 contains 1794 documentation entries, 1309 patterns, 1198 profile	s and 1217 ProBule.			
	,,,,,,,			-	
		EP 1 - Submit PRO	TEIN Sec	uences [neip]	
	l c	Submit PROTEIN se	quences (i	max 10) Examples	
		Submit a PROTEIN	database (max. 16MB) for repeated scan	s (The data will be stored on our server for 1 month).
				· ·	
	≥s M0	p P26367 PAX6_HUMAN Pai NSHSGVNOLGGVFVNGRPLPDST	red box pro	itein Pax-6 OS=Homo <u>sapiens</u> GN ARPCDISRILOVSNGCVSKILGRY	=PAX6 PE=1 SV=2
-		TGSIRPRAIGGSKPRVATPEVVS	KIA0YKRECP	IFAWEIRDRLLSEGVCTNDNIPSV GTRPGWYPGTSVPG0PT0DGC000	
Ent	er pax6_human in the SIEP I - Submit	GGENTNSISSNGEDSDEAQMRLQ	LKRKLORNRT	FTOEOIEALEKEFERTHYPDVFAR	
PR	OTEIN sequences section	TPVSSFTSGSMLGRTDTALTNT	YSALPPMPSF	MANNLPMOPPVPSOTSSYSCMLPT	
		SVNGRSTRITTEPHNUTHUNSUP	19615611516	15PGV5VPVQVPG5EPUM5Q1WP8	
	Su	pported input:			
		 UniProtKB accessi 	ons e.g. PS	8073 or identifiers e.g. ENTK	HUMAN
		 PDB identifiers e.g. 	4DGJ		_
		Sequences in FAS	TA format		
STE	P 2 - Select options [help]				
0.1					an arrest that the Earlands
	Evolute motifs with a high probability of occurrence from the scan	P Z - Sele	ect of	btions section,	ensure that the Exclude
	Evolute profiles from the scan	a high pro	obabi	ility of occurre	ence box is ticked.
	Pup the scan at high constituity (show weak matches for profiles)				
	tur ne soar at high sensitivity (show weak matches for promes)				
		s	STEP 3 - S	elect output options and su	bmit your job
The	e defaults offered in the STEP 3 - Select output	t options of	Dutput form	at: Graphical v	view 🗘
and	submit your job section are fine so just clic	k on the	Retrieve con	nplete sequences: 🗆 If you c	hoose this option, not all output formats are available.
	A DELETITE COANT 1 W	K OII UIC	Receivo	your results by email	
ST	AKI THE SCAN button. In but a few mome	nts, your	neceive	Joan roound by crindil	
res	ults will burst forth.				START THE SCAN Reset

"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."

¹⁰ Expasy is a major site for protein based research in Switzerland. As the all knowing Wikipedia puts it:

Discussion Points

Tuesday 20 February 2018

	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-PAX PATRED_2 totteor (422 aa) 6_HUMAN)									
Two hits with PROSITE profiles	PS51057 PAIRED_2 Paired domain profile :									
suggesting the familiar domains in their familiar places.	4 - 130: score = 64.941 SHSGVN0LGGVFVNGRPLPDSTR0KTVELAHSGARPCDISRIL0VSNGCVSKILGRYYET GSIRPRAIGGSKPRVATPEVVSKLAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSI NRVLRNL Predicted feature:									
	DOMAIN 4 130 Paired [condition: none]									
	PS50071 HOMEOBOX_2 'Homeobox' domain profile :									
	208 - 268: score = 20.164 RKLORNRTSFTQEQIEALEKEFERTHYPDVFARERLAAKIDLPEARIOVWFSNRRAKWRR E									
	hits by patterns: [2 hits (by 2 distinct patterns) on 1 sequence]									

Two hits with **PROSITE** patterns confirm the same domains by hits by patterns: [2 hits (by 2 distinct patterns) on 1 sequence] matching highly conserved subregions.

This confirms what has already been discovered more than once, by PAX6-HUMAN reading database annotations, by running Interpro and by running other individual database search program(s) manually.

PS00027 HOMEOBOX_1 'Homeobox' domain signature : Note that **Prosite** is happy to accept the HOMEOBOX pattern hit as 243-266:

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.

PS00034 PAIRED 1 Paired domain signature

100

ruler:

sp-P26367-

6 HUMAN)

38 - 54:

(sp-P26367-PAX

200

[confidence level: (0)] RPCdisrilqvsngCVS

[confidence level: (0)] LAakIdLPeaRIQVWFsNrrakwR

300

400

600

500

(422 aa)

700

800

900

1000

Move back to the search submission. In Step 2, deselect Exclude patterns with a high probability occurrence. START of 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:	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$
sp-P26367- PAX6_HUMAN (sp-P26367-PAX 6_HUMAN)	

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

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

G-{EDRKHPFYW}-x(2)-[STAGCN]-{P}[GistheN-myristoylationsite]

proteins of any size and not always because there was an N-myristoylation site.

PS00008 MYRISTYL N-my

GVfvNG

GArpCD

GVctND

GQtqSW

GSwgTR

GTroGW

GGgeNT

GGenTN

GSmlGR

GTsaTT

GTtsTG

13 - 18:

36 - 41:

110 - 115:

151 - 156:

154 - 159:

157 - 162:

182 - 187:

183 - 188:

312 - 317:

387 - 392-

390 - 395

Dis	cussion Points		Tuesday 20 February 2018
	The pattern	is I	The N-terminal residue must be glycine.
	avalained in	tho	• In position 2, uncharged residues are allowed. Charged residues, proline and large hydrophobic residues are not allowed.
	explained in	ule	In positions 3 and 4, most, if not all, residues are allowed.
	database thus.	•	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.

The pattern **RPC**xxxxxxxxxx**CVS** within the entire domain is clear.

equivalent picture would be illustrated.

2NSHSGVNQLGGVFVNGKFLFDSIKQKIVELAHSGARFCDISKILQVSNGCVSKILGKIIEIGSI
PRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLAS
KQQMGADGMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTQDGCQQQEGGGENTNSISSNGEDSD
AQMRLQLKRKLQRNRTSFTQEQIEALEKEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAK
RREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTPVSSFTSGSMLGRTDTALTNTYSALP
MPSFTMANNLPMQPPVPSQTSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMNSQPMGTSGTTSTGL
SPGVSVPVQVPGSEPDMSQYWPRLQ
<u>Q</u> N <mark>SHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGA</mark> RPCDISRILQTHADAKVQVLDNQNVSNGC

If you were to repeat this whole ^{MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARECTISKINGTHADAKVVVDDNQNVNGE exercise with the **isoform 5a** version of **Human PAX6** (*please do not*!), the ^{MQNSHSGVNQLGGVFVNGRPLDSTRQKIVELAHSGARECTISKINGTHADAKVVVDDNQNVNGE ENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEQIEALEKEFERTHYPDVFARERLAAKIDLP EARIQVWFSNRRAKWRREEKLRNQRRQASNTPSHIPISSSFSTSVQPIPQPTTPVSSFTSGSMLG}} as RTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQTSSYSCMLPTSPSVNGRSYDTYTPPHMQTHMN SQPMGTSGTTSTGLISPGVSVPVQVPGSEPDMSQYWPRLQ

The 14 amino acid insertion of isoform 5a (THADAKVQVLDNQN, corresponding to the entire 3rd 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)

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 SwissProt¹¹.

- 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:

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 assemble and 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.

Discussion Points

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.

Sequences in UniProtKB/Swiss-Prot known to belong to this class: 58

 detected by PS51057: 58 (true positives)
 undetected by PS51057: 0 (false negative or 'partial')

 Other sequence(s) in UniProtKB/Swiss-Prot detected by PS51057: NONE.

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 its 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 - 2017.12.23

References for further extension:

https://en.wikipedia.org/wiki/Multiple_sequence_alignment