IB16S

Introductory Bioinformatics

12-16 December 2016

(Second 2016 run of this Course)

Basic Bioinformatics Sessions

Practical 5: Secondary Structure Prediction

Friday 9 December 2016

Computational Protein Analysis

In this section, the plan is to look exclusively at the **protein** of **PAX6**. The object is to use various software items to confirm what has already discovered from the various web resources. Often the software you will use will be exactly that which was used to determine the pre-computed results you browsed previously.

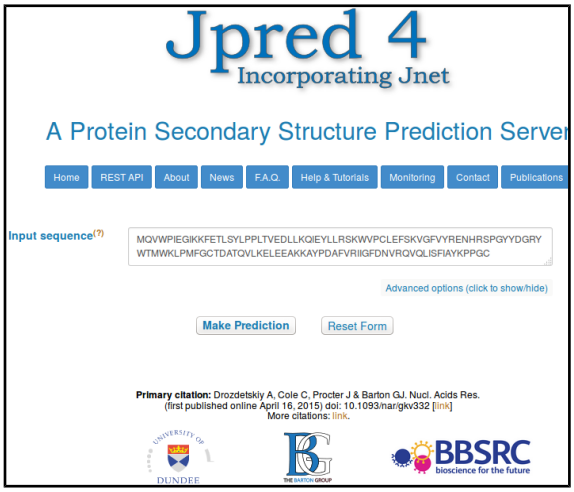
Predicting Protein Secondary Structure.

Beta strand	6 – 8	3
Beta strand	14 – 16	3
Helix	23 – 34	12
Helix	39 – 46	8
Helix	50 – 63	14
Beta strand	77 – 79	3
Helix	81 – 93	13
Helix	99 – 108	10
Turn	114 – 116	3
Helix	120 – 133	14
Helix	219 – 229	11
Helix	237 – 246	10
Helix	251 – 275	25

A first step is to look at to look at ways to predict the protein secondary structure of the **PAX6** protein. Evidence from various sources suggests that the **PAX6** protein has **9** helices arranged in triplets, plus a few beta strands.

To save time, I show here the relevant section from the **Uniprot Feature Table**. The helical triplets are involved in binding. **2** triplets are to be found in the paired box region, the other in the homeobox a little further along. Here we will try one of the most sophisticated methods available, to predict, essentially from from the primary sequence, what we already know. If you really wish, I also offer a supplementary exercise based around one of the earlier prediction methods, still used, but although faster, significantly less accurate than more modern methods.

Early Secondary Structure Prediction Methods - GOR



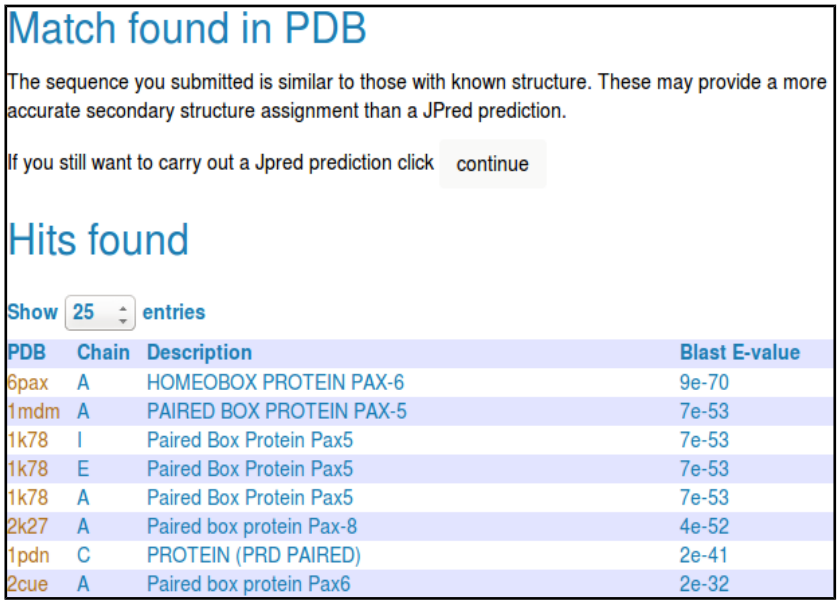
The service considered by many to offer the most effective method of predicting secondary structure is called **Jpred**. This is developed by the Barton group now located at Dundee University. Over **80%** accuracy is claimed for **Jpred** predictions. Due to the inherent imprecision in defining the end positions of secondary structure elements, **80%** is pretty much as good as is practically possible.

Go to the **Barton Group** web site at:

http://www.compbio.dundee.ac.uk

and follow the link to the Jpred 4 Incorporating Jnet server. Copy and paste the **PAX6** protein (from the file **pax6_human.fasta**) into the appropriate text box. Click on **Make Prediction**.

With alacrity, **JPred** will report several hits with proteins of known **3D** structure (using **blast** against a database of proteins of known **3D** structure). Links are offered to a number of entries in the **PDB** structure database. At least **2** of the **PDB** entries listed should be familiar.



Match found in PDB

The sequence you submitted is similar to those with known structure. These may provide a more accurate secondary structure assignment than a JPred prediction.

If you still want to carry out a Jpred prediction click continue

Hits found

Show 25 entries

PDB	Chain	Description	Blast E-value
6pax	A	HOMEOBOX PROTEIN PAX-6	9e-70
1mdm	A	PAIRED BOX PROTEIN PAX-5	7e-53
1k78	I	Paired Box Protein Pax5	7e-53
1k78	E	Paired Box Protein Pax5	7e-53
1k78	A	Paired Box Protein Pax5	7e-53
2k27	A	Paired box protein Pax-8	4e-52
1pdn	C	PROTEIN (PRD PAIRED)	2e-41
2cue	A	Paired box protein Pax6	2e-32

Jpred offers the suggestion that it really does not make sense to continue. After all, if the **3D** structure is effectively known, why predict (guess?) the **2D** structure? The answer to this challenge being a petulant "**Because we want to!**"

Click purposefully on the **Continue** button. **JPred**, with a small sigh of exasperation, will submit your job and tells you how busy it is. **Jpred** typically takes a while as it has much to consider[1].

Jpred will use **PSI-Blast** to align your sequence with all sequences deemed to be homologous, from a particularly appropriate database. **Jpred** then makes its structure predictions based on an aligned "family" of proteins, rather than just one individual sequence. Intuitively at least, this has to be a fine idea. A **M**ultiple **S**equence **A**lignment (**MSA**) of related proteins will typically represent far more evidence for prediction than any single protein.

1 If the wait becomes unbearable, consider opening a new window/tab and moving on to the next section, returning to **Jpred** later.

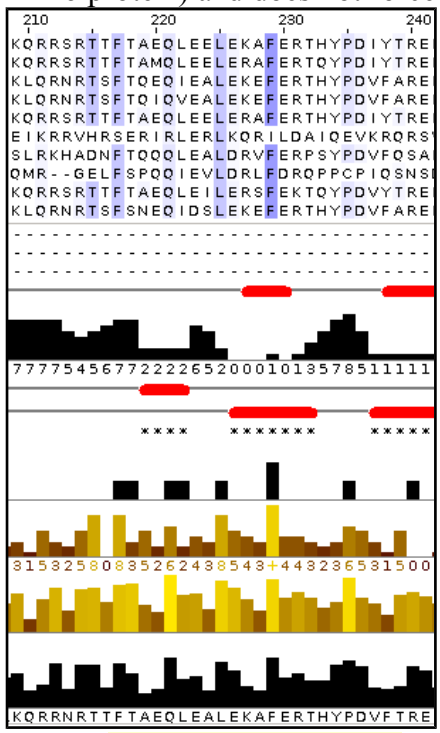
JPred presents the results of running two secondary structure predictions, using the program **JNET**, based on two different representations of the alignment (**HMM** and **PSSM**, similar ideas that will be discussed at some point). Predicted helices are represented as red blocks, predicted beta sheets as green arrows. A consensus prediction is presented (**jnetpred**) as is an indication of prediction confidence (**JNETCONF**). Algorithms are also run to predict **coiled coils** (**Lupas**, with window sizes **21**, **14**, **28**). The first view of the results offered is a graphical overview aligned with your original single sequence.

The annotation bars below the alignment are as follows:

- Lupas_21, Lupas_14, Lupas_28
  Coiled-coil predictions for the sequence. These are binary predictions for each location.
- JNETSOL25,JNETSOL5,JNETSOL0
  Solvent accessibility predictions - binary predictions of 25%, 5% or 0% solvent accessibility.
- JNetPRED
  The consensus prediction - helices are marked as red tubes, and sheets as dark green arrows.
- JNetCONF
  The confidence estimate for the prediction. High values mean high confidence. prediction - helices are marked as red tubes, and sheets as dark green arrows.
- JNetALIGN
  Alignment based prediction - helices are marked as red tubes, and sheets as dark green arrows.
- JNetHMM
  HMM profile based prediction - helices are marked as red tubes, and sheets as dark green arrows.
- jpred
  Jpred prediction - helices are marked as red tubes, and sheets as dark green arrows.
- JNETPSSM
  PSSM based prediction - helices are marked as red tubes, and sheets as dark green arrows.
- JNETFREQ
  Amino Acid frequency based prediction - helices are marked as red tubes, and sheets as dark green arrows.
- JNETJURY
  A '' in this annotation indicates that the JNETJURY was invoked to rationalise significantly different primary predictions.*

The full key to all the abbreviations used (and more information about **JNet**) can be displayed by clicking on the **details on acronyms used** link.

For a fuller view, elect to **View results in Jalview**[2]. You will arrive at a page inviting you to select from various viewing options. The options are explained clearly, but to save you time reading and pain deciding, I suggest you go for **Option 1** for the clearest view. This option does not confuse the picture by gapping your query sequence (and thus making it more difficult to associate structure predictions with regions of the **PAX6** protein) and does not force you to look at the entire, huge, **MSA** generated by **PSI-Blast**.

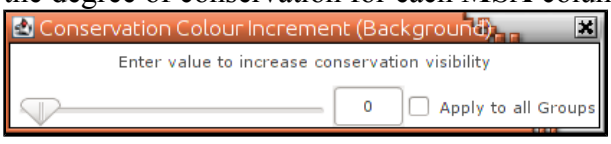


Jalview presents something very similar to the original view of the **Jpred** results. This time though, the most significant part of the **PSI-Blast MSA** from which the predictions were computed is displayed, if rather blandly.

To highlight the conserved regions of the alignment, some colour is required. **Jalview**, offers a number of colouring strategies. I refer you to the **Help** for the full story. Here I will choose what I think is a revealing option with minimal explanation[3].

From the **Jalview Colour** pull down menu, select **BLOSUM62 Score**, to suggest that the inclination to colour any amino acid of the **MSA** be determined from its **BLOSUM 62** score with the corresponding **Consensus** sequence residue and the quality of that consensus sequence residue. A few of the most highly conserved **MSA** positions around the homeobox region will now be coloured a watery blue, but otherwise, the display will remain unimpressed.

In order to remove unwanted subtlety, from the **Jalview Colour** pull down menu, select **By Conservation**, thus electing for the colour intensity to be reflected by the degree of conservation for each **MSA** column. A jolly little slider bar will leap forward. Slide the bar to and fro to get a feel for what it is doing. Terminate your oscillations with the minimum value selected, thus demanding that any slight odour of conservation should elicit a maximal colour burst! Clearly appropriate. You can ignore the reference to **Groups** as you have not specified any, so the entire **MSA** is regarded as a single **Group** at this point.

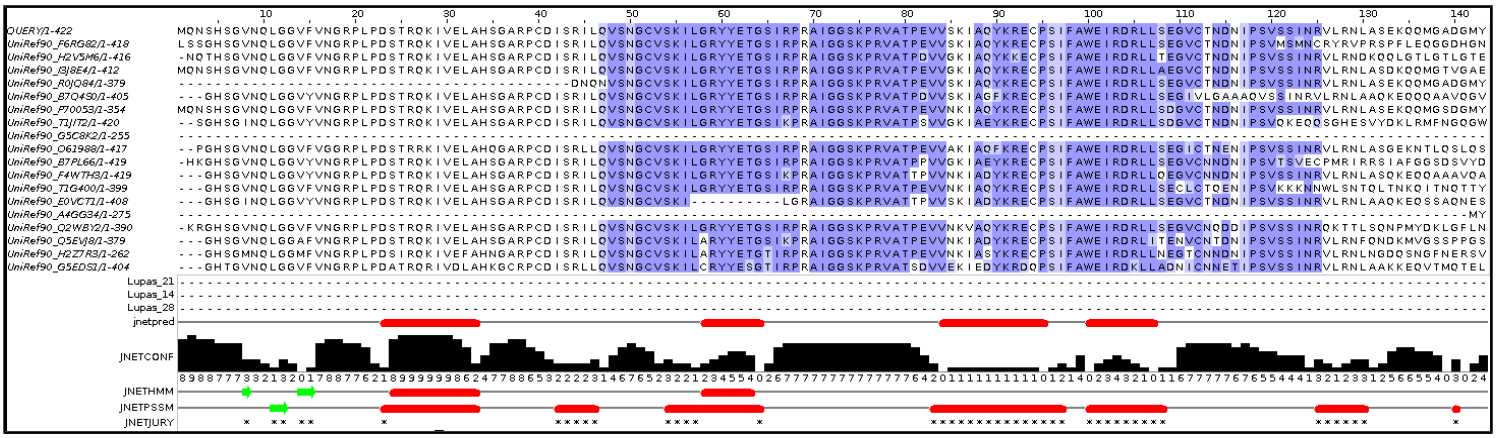


Now, all the regions regarded as even vaguely conserved glow enthusiastically blue. All regions, in this instance, being everywhere a conservation score of "**0**" or more is recorded. That is, everywhere except where the conservation is represented by a "**-**" by the conservation histogram.

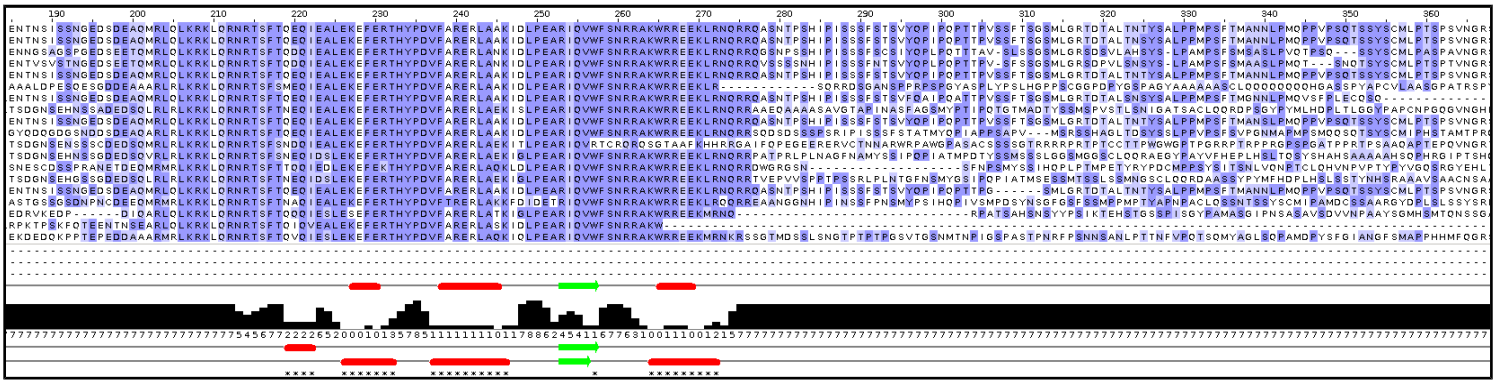
Slide along the entire width of the **MSA** and you should clearly see that, roughly, the **Paired Box domain**, **Homeobox domain** and the **compositionally biased C-terminus** are exclusively highlighted.

2 Should that not work, try **Full HTML**.
3 I have made some notes on my choice, but they should not really detain you at his point. If you insist, they are here.

Here I have included the **Jalview** version of the MSA and structure predictions around the **PAX** region



and those around the **Homeobox**, including some of the **C-terminus compositionally biased region**.



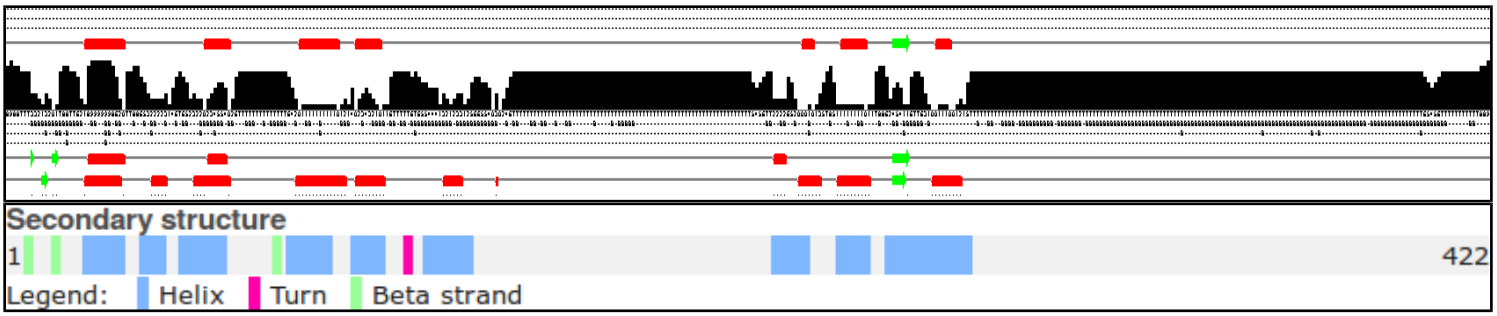
Note that, even though **JNET** has produced a reasonable secondary structure prediction for the start of the **PAX** region, **Jalview** does not consider this region to be sufficiently conserved to colour? Why this might be so will become apparent when you consider the quality of this prediction overall (in a couple of Questions time).

What protein database has **Jpred** chosen to search for protein sequences for the alignment upon which its predictions will be based? _____

Why do you suppose this database was used in preference to, say **UniprotKB**?_____

_____

Also, I have lined up the entire prediction with the **Uniprot** Feature Table graphic.



It would appear the helices predicted least confidently by **Jpred** are the same ones with which **GOR IV** (investigated in a supplementary exercise) had problems.

How would you rate the **Jpred** prediction overall?_____

_____

Protein Tertiary Structure

Protein Data Bank (PDB)

The **P**rotein **D**ata **B**ank (**PDB**) archive is the major repository of information about the 3D structures of biological molecules, including proteins and nucleic acids. Structures in the archive range from tiny proteins and bits of DNA to complex molecular machines like the ribosome.

In 1998, the **R**esearch **C**ollaboratory for **S**tructural **B**ioinformatics (**RCSB**) became responsible for the management of the **PDB**.

In 2003, the **wwPDB** formed to maintain a single **PDB** archive of macromolecular structural data that is freely and publicly available to the global community. It consists of organizations that act as deposition, data processing and distribution centres for **PDB** data.

PDBe is the European resource for the collection, organisation and dissemination of data on biological macromolecular structures. In collaboration with the other Worldwide Protein Data Bank (**wwPDB**) and **EMDataBank** partners, they work to collate, maintain and provide access to the global repositories of macromolecular structure data (the Protein Data Bank (**PDB**) and Electron Microscopy Data Bank (**EMDB**)).

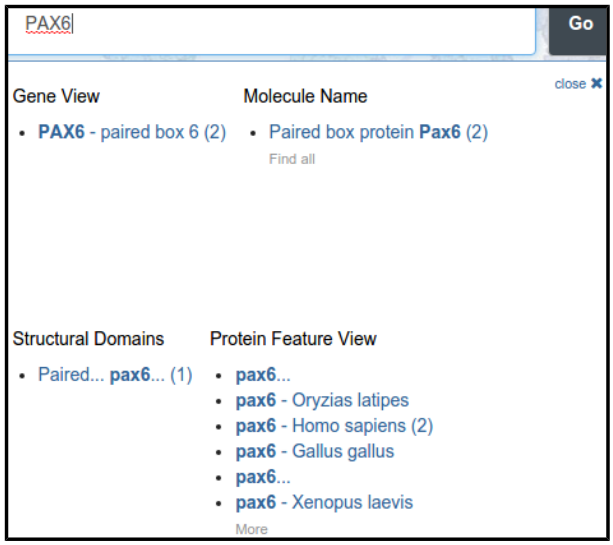
In the course of the exercises undertaken to this point, you will have already had a look at the 3D structures for the 2 major domains of the human **PAX6** protein. You might have taken a more direct route to these structures by asking for them directly from the **RCSB PDB** site as follows.

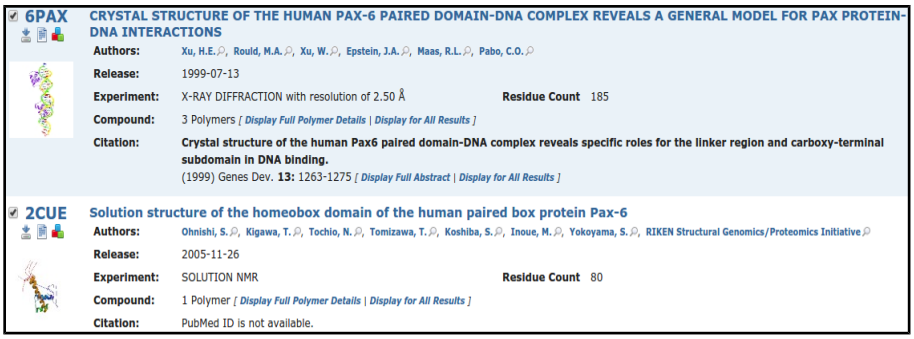
Go to:

http://www.rcsb.org

Enter **PAX6** in the **Search** box and click on the **Go** button.

Click on the link under the **Molecule Name** title..





The two **PDB** structure hits will, hopefully, be familiar. Links are provided with each hit to view the structure with **Jmol** (a java based structure viewer), view the textual **PDB** entry and download the **PDB** entry to a file.

Take a look at the **Jmol** view of the **6PAX PDB** entry. This you have seen this previously, but now I suggest a very quick visualisation of the main mutation that causes aniridia occurs in the **PAX** protein. The idea is to locate and highlight the **Alanine** that mutates to a **Proline** in an aniridia sufferer. As you have discovered, this is the residue **33** in the canonical protein, as recorded by **UniProtKB**. It is residue **30** in the protein as visualised here, the difference being explained by **post translational modification** which, in this instance, removes the first three amino acids.

Instructions for using **Jmol** can be found in many places. For a **Quick Guide**, you might try:

http://blc.arizona.edu/courses/mcb184/graphics/JmolQuickReferenceSheet.pdf

One place for the full manual is:

http://jmol.sourceforge.net/docs/JmolUserGuide/

Please note **Jmol** is not the only structure visualisation option available to you, nor is it the most sophisticated. It is just the one used by **PDB**. Here we look at just the minimum of **Jmol** skills to see what is required. First notice you can zoom in and out with the wheel of your mouse. You can also rotate the image in all directions using your left hand mouse button. Use these two tricks as needed.

To proceed any further, you really need a console window into which you can type commands. To get a console, choose **Console**. From the right hand mouse button pull down menu.

In the lower section of the console, select the **30th** amino acid with the command:

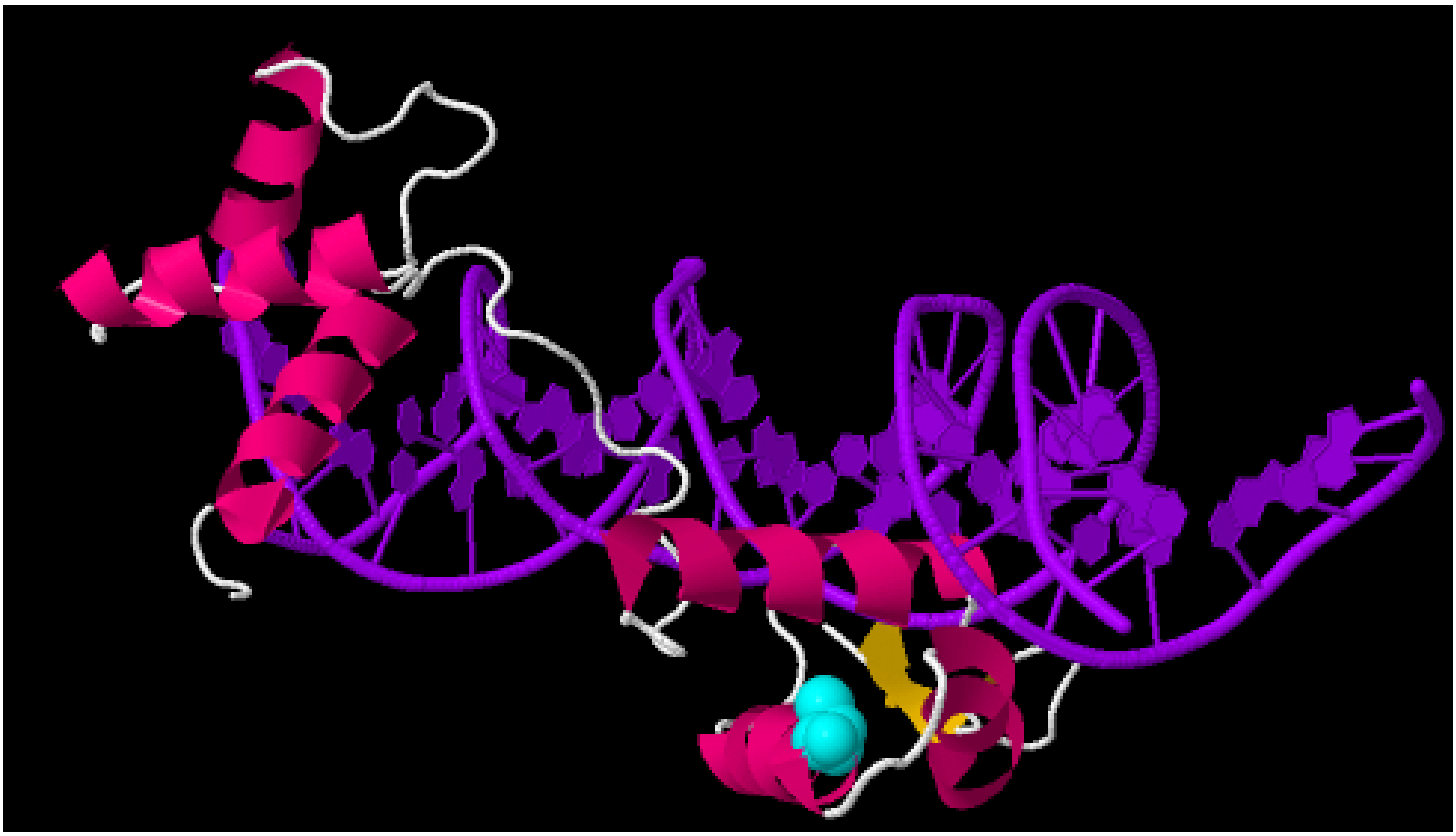
select 30

Then to make the selected residue more evident, type in the two commands:

spacefill

color cyan

and then manipulate the structure until the selected amino acid can be best observed.



DPJ 2016.12.06

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 **Protein Secondary Structure Prediction with Jpred**

Some Notes on colouring the MSA generated by Jpred:

(Click here to return to the Instructions.)

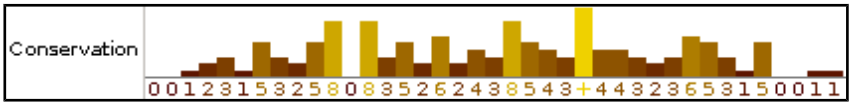
Mostly to remind me why and how I decided to colour the **MSA** as I did. My objective was purely to make obvious where the family of proteins were meaningfully similar. If you are happy that this objective was achieved, it is probably best to read no further.

I discovered most of what follows by Selecting the **Help** (easiest way is to press **F1** key, otherwise there is a pull down option at the top of the display, choose **Documentation** option) and searching for "**conservation**". From the list of hits, I first selected "**Alignment Conservation Annotation**". There it says:

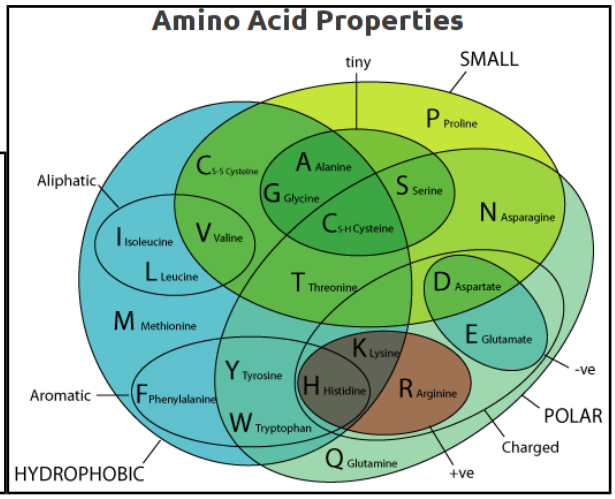
> "**Conservation** is measured as a numerical index reflecting the **conservation** of physico-chemical properties in the alignment: Identities score highest, and the next most conserved group contain substitutions to amino acids lying in the same physico-chemical class.
>
> **Conservation** is visualised on the alignment or a sequence group as a histogram giving the score for each column. Conserved columns are indicated by '*' (score of **11** with default amino acid property grouping), and columns with mutations where all properties are conserved are marked with a '+' (score of **10**, indicating all properties are conserved).
>
> Mousing over a conservation histogram reveals a tooltip which contains a series of symbols corresponding to the physico-chemical properties that are conserved amongst the amino acids observed at each position. In these tooltips, the presence of ! implies that the lack of a particular physico-chemical property is conserved (e.g. !proline)."

I think to understand the detail of the scoring, one would have to read the paper quoted in the **Help**. I think I will leave that until another day! For now, I just make a few notes.

- The numbers under the histogram columns appear to represent simply the number of phsico-chemical properties considered to be conserved. At least, this is consistently true for this example, shown by hovering the mouse over the histogram columns.

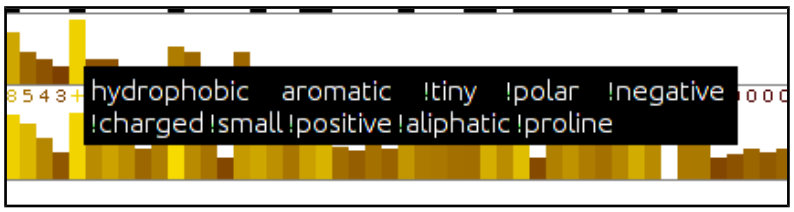


- **Jalview** admits to exactly **10** phsico-chemical properties that must be one of "**Not conserved**", "**positively conserved**" or "**negatively conserved**".




```
ILVCAGMFYWHKREQDNSTPBZX-
XXXXXXXXXXX·····X···XX Hydrophobic
·····XXXXXXXXXX·XXXXX Polar
··XXXX·····XXXXX··XX Small
·····X··XX Proline
····XX·····X····XX Tiny
XXX·····XX Aliphatic
·····XXXX·····XX Aromatic
·····XXX·····XX Positive
·····X·X·····XX Negative
·····XXXX·X·····XX Charged
```

- The column achieving a "+" has all **10** conserved phsico-chemical properties either **positively** or **negatively conserved**. It is a highly, but not completely **conserved** "**F**". This would appear to agree with the **Help**? There is no example of a **100%** conserved column in this example. If there was, I would expect it would be represented by a "*" representing a score of **11**.

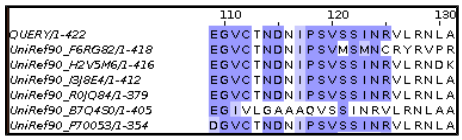


- Conservation of any given property does not have to be **100%** and gaps are tolerated. Reasonable as to be too exacting would eliminating. I expect the details are explained in the original paper. I justify this statement, unnecessarily, by claiming there are both gaps and a **Proline** in the column represented by a "+".
- I am still uncertain about the difference between a "**0**" column and a "–" column? I decide to believe they are both columns were there is no measurable conservation, but "**0**" columns are in regions where they are surrounded by significant conservation? One day, I will read the paper.
- By observation, it can be seen that "conservation" is measured relative to the consensus sequence rather than the query sequence. This seems a reasonable choice to me.

Well that was fun? Now I write some instructions to turn the nasty bland alignment into one that glows blue.

Click here to return to the Instructions.

What protein database has **Jpred** chosen to search for protein sequences for the alignment upon which its predictions will be based?



The database **Jpred** instructed **PSI-blast** to use to seek proteins homologous to the **PAX6** query can be determined by looking at the sequence identifiers displayed down the left hand side of the alignment in **Jalview**. The identifiers are constructed from the name of the database and the entry identifier separated by an underline character. So the database is the **UniRef90** cluster database built from the **UniProtKB** database.

The **UniRef** cluster databases comprise entries that are not individual protein sequences, but cluster of similar sequences. In the case of the **UniRef90** database, each entry includes all sequences **90%** identical to a given seed sequence. A representative sequence is elected as the only one of the cluster to be considered by such as **PSI-blast**, but clearly, a hit with any representative sequence implies significant similarity with all the sequences of its cluster.

I offer a supplementary exercise to investigate these cluster databases for those to whom they might be of particular interest.

Why do you suppose this database was used in preference to, say **UniprotKB**?

The reason **Jpred** runs **PSI-blast** is to identify sequences representing as wide a family of proteins as possible, to which a **Query** sequence belongs. For the purpose of structure prediction, there is little value in this collection including many sequences that are essentially identical. A wide variety of sequences, as long as they still are likely to be homologous, is of far greater value than a huge number of sequences. Using a **UniRef** database allows that only the **Representative** sequence of each cluster of very similar sequences will be recognised and aligned by **PSI-blast**. This allows the **PSI-blast MSA** to include an extensive range of variation without being bloated by sequences too similar to be individually interesting.

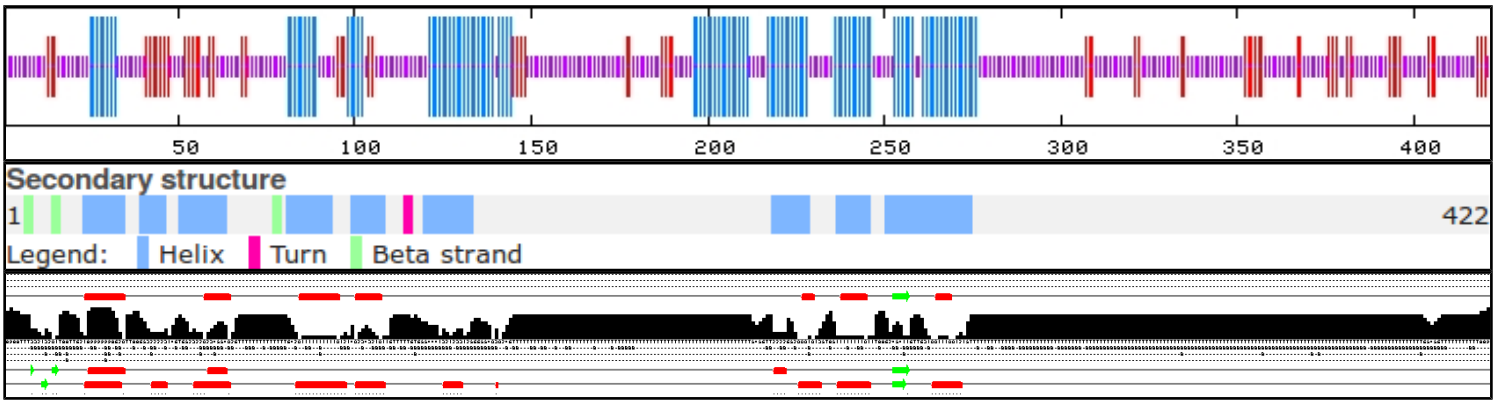
How would you rate the **Jpred** prediction overall?

Well, frankly, not as wonderful as I was expecting. Better than **GOR IV** (investigated in a supplementary exercise), but still leaves room for improvement? **jnetpred** (essentially the answer) is reasonable. It misses a couple of helices including one that **GOR IV** also overlooks. However, it has considerably less false positive prediction tendencies than **GOR IV**. The **JNETHMM** predictions are particularly poor, saved by the much more accurate deliberations of **JNETPSSM**.

JNETHMM is a prediction computed from the **H**idden **M**arkov **M**odel (**HMM**) representation of the final **PSI-blast MSA**.

JNETPSSM is a prediction computed from the **P**osition **S**pecific **S**coring **M**atrix (**PSSM**) representation of the final **PSI-blast MSA**. **PSI-blast** uses **PSSM**s of the **MSA** of each iteration of its search as a **Query** for the next iteration.

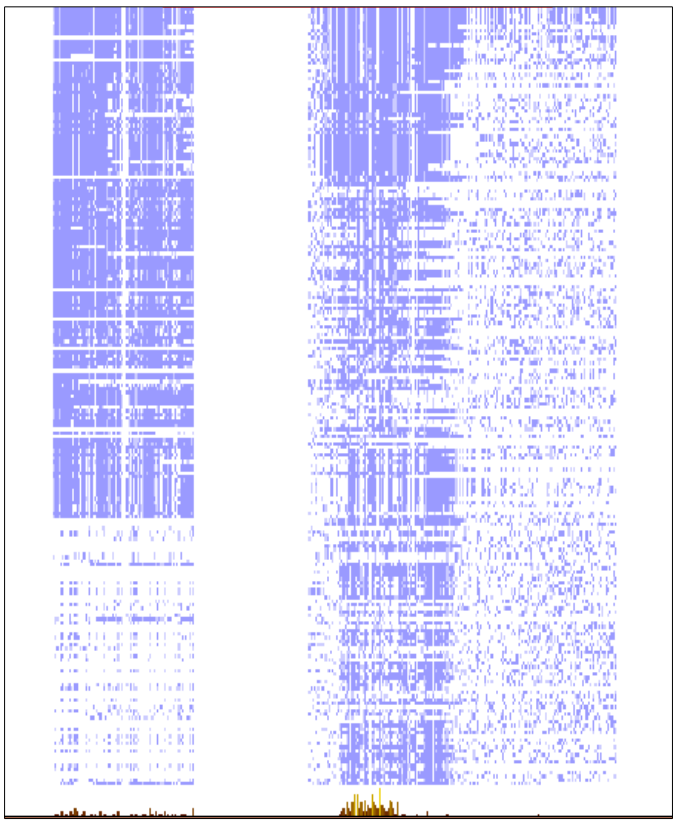
The **jnetpred** prediction is effectively the consensus of the predictions of **JNETHMM** and **JNETPSSM**.



Here I have aligned the **GOR IV** and **Jpred predictions** with the secondary structure as recorded by **UniProtKB**.

So, can the prediction be improved? **Jpred** is better than this result suggests!

On reflection, maybe just throwing in the entire sequence of **PAX6_HUMAN** and hoping for the best was a little crude? Our protein has two major domains whose secondary structure one might expect to be conserved. **PSI-blast** will gather together a mountain of sequences that have one, or the other, or both of the domains and try to align them as if they were homologous over their entire length (a **global alignment**). ***BUT***, they are not all globally homologous! This means that the alignment of both the domains regions will be polluted by sequence that represent proteins that do not include that domain. This must substantially reduce the quality of the prediction?



This phenomena can be illustrated by choosing to view the **Jalview Overview Window** (available from the **View** pull down menu).

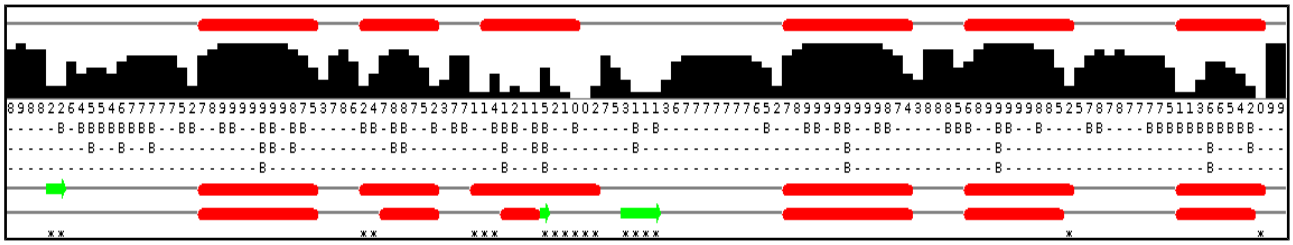
The wider column of blueness at the start of the alignment represents the **paired box** domains. The picture suggests about one third of the aligned sequences do not have a **paired box** domain, but those sequences will have unrelated sequence in that region that will reduce the degree to which the alignment represents the properties of a **paired box** and so also the likelihood of a sensible structure prediction[4].

The problem for the more common **homebox** domain looks less severe, however, the alignment clearly includes many sequences that do not look to have a **homeobox** domain.

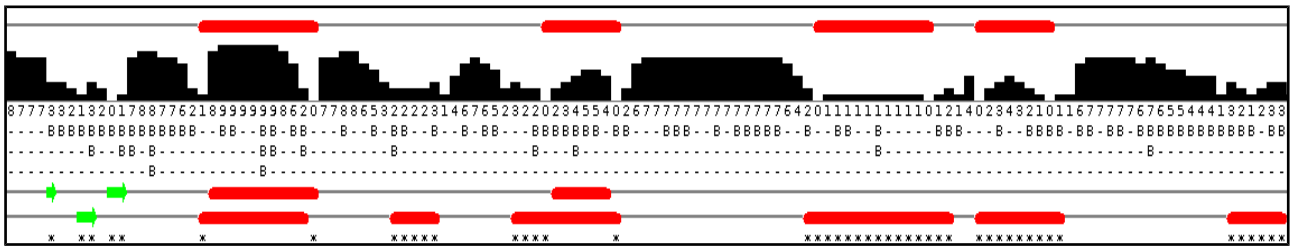
So, what to do? I suggest the two domains might be investigated separately? Why not run **Jpred** twice, once with just the **PAX6_HUMAN** paired box region and then again with just the **homebox** region.

I have done this for you and will now show you the results, however, should you wish to try it yourself, you already have the isolated sequence of both domains saved in local files. The sequence of the **paired box** region should be in a file called **pax_domain.fasta**. The **homeobox** sequence should be in a file called **homeobox_domain.fasta**. Run **Jpred** again with each sequence and you should get results very similar to mine.

First the new **paired box** prediction (top) compared to the original (bottom).



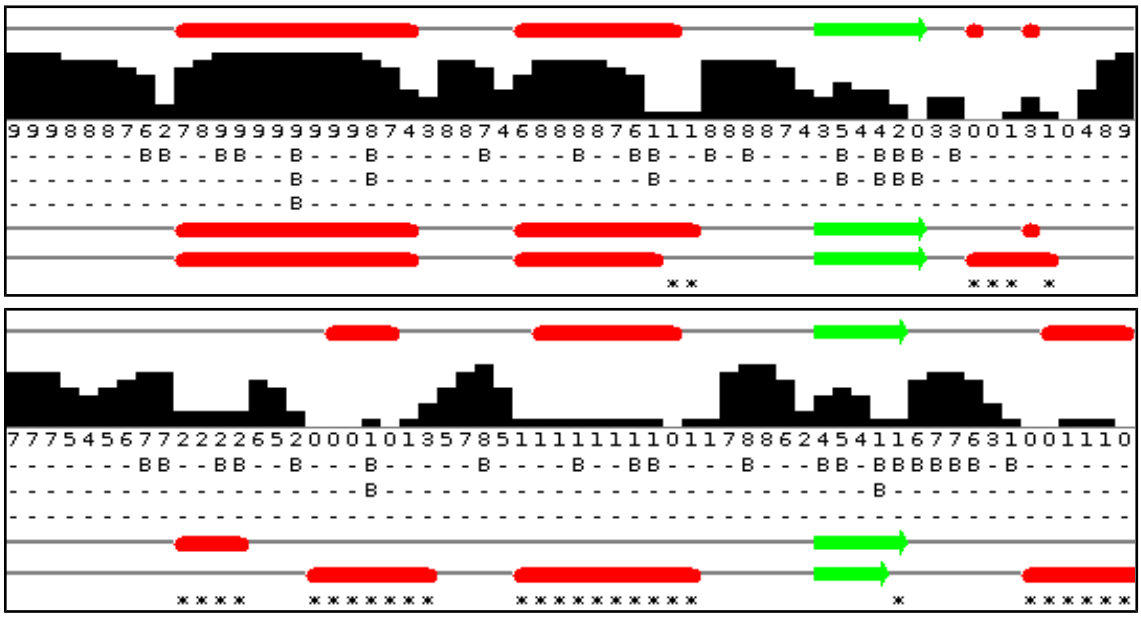
Massively improved I would suggest. All helices present and accurately placed. The **JPREDHMM** prediction, in



particular, is very much improved. The Beta Sheet predictions seem weak? It finds only one (accurately) of the three that **UniProtKB** suggests to be present. I wonder why, but the helices for the paired box domain specific prediction are excellent.

4 This could be why, as noted in the instructions, the start of the **PAX** region was considered insignificantly conserved by **Jalview**.

And so to the **homeobox** specific results. Once more, the new **homeobox** prediction (top) compared to the original (bottom).



As the **homeobox**s are significantly more numerous than the **paired box**s, less interference from sequences not including a **homeobox** might have been expected. I imagined the improvement in prediction would be minimal. However, it is very much better! All three helices are predicted in the correct positions, although **Jpred** appears to be a little reluctant about the third helix? There is a rather strong beta sheet prediction that is unsupported by **UniProtKB**. There is no reason to suppose that **UniProtKB** is **100%** correct, of course, but nothing I can find suggests that a beta sheet should appear in the middle of a homeobox. An enigma for another day.

So I conclude that this sort of protein analysis requires a little bit more than just throwing an entire sequence at a dumb program and assuming something marvellous will occur. In this case, considering the regions of the protein that are expected to be homologus separately is a very logical thing to do (and entirely obvious, retrospectively at least). Geoff Barton, whose group is responsible for **Jpred** agrees. He says[5]:

> " … Always split proteins into domains when searching. … "

So for both domains the prediction of the helices is far more accurate when each domain is considered separately. However, it is not just the red bars indicating the position of the helical predictions that should be noted. Look also at the confidence histogram. It indicates clearly that with more specific data to work on, better predictions can be made with much improved confidence (i.e. likelihood of being correct!).

DPJ – 2016.12.09

5 As does the **Jpred Help**.