Jalview/JPred4 training course Cambridge Day 3 programme and practical

Protein Secondary Structure Prediction and Sub-Family Analysis

Geoff Barton

email: gjbarton@dundee.ac.uk
web: www.compbio.dundee.ac.uk
twitter: @gjbarton

Contents

Introduction

In this course you'll learn some background to protein structure prediction, why it is useful and its limits. You will then do some hands-on predictions with the JPred4 server for sequences that give good and not so good predictions.

You will also work with these predictions in Jalview to explore the effect of editing the multiple alignment that JPred uses for the prediction in different ways. Finally, you will learn more about the protein sub-family analysis, a useful technique for identifying functionally important amino acids from a multiple sequence alignment.

The example files for this course (including this document) are all in:

http://www.jalview.org/tutorial/training-materials/2015/Cambridge/May/day3/

Further hyperlinks are provided below for information that is on the JPred website.

Programme and practicals

Lecture: "Protein Structure Prediction".

- Briefly revises important concepts about protein structure
- Briefly explains the limits of methods to determine three-dimensional structure experimentally.
- Briefly summarises different approaches to predict the tertiary structure of a protein from its amino acid sequence.
- Explains principles of protein secondary structure prediction:
    - Features in protein sequences that are diagnostic of different secondary structures
    - Example of a "Blind" prediction that worked
    - Example of problems that can arise
    -

Lecture: "JPred and JNet: Protein Secondary Structure Prediction"

- Explains what JPred and JNet are
- Describes how they work
- Shows how their accuracy has improved since 1999

Practical 1: "Predicting Secondary Structure with the JPred Server"

The JPred web server can be found on: www.compbio.dundee.ac.uk/jpred.

I'm going to go through the example here for you on the screen, then we will move on to using Jalview to view and interact with predictions made on the website.

1. Go to the JPred address above and you should see something that looks like Figure 1.

2. Now Click on the button marked "Help and Tutorials".

3. Click on the JPred Tutorial link on "making a prediction from a single sequence". You can get there from this link as well: http://bit.ly/1RkVxCz

4. Skip to the "Detailed step-by-step guide" which starts on Page 6. You can follow the detailed instructions in that guide to see how to run JPred with a single sequence and discover the different types of results you get back. Please try the following example:

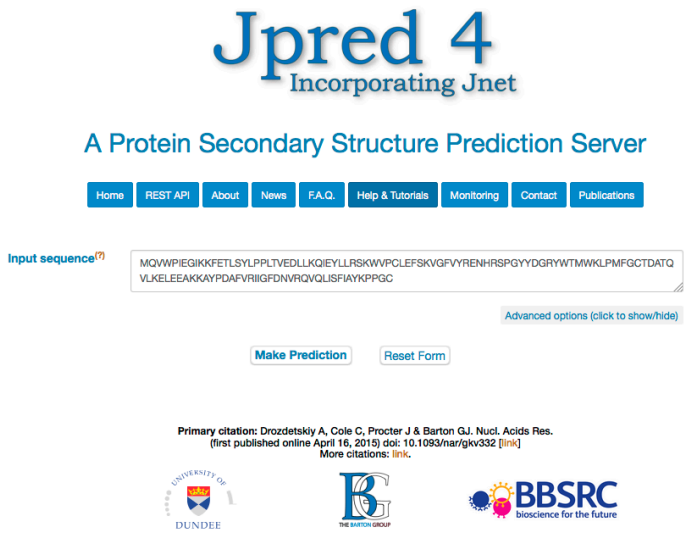


Figure 1 JPred webserver home page

5. Try pasting in the following sequence:

GGIQVNGPRLESLVLTYVNAISSGDLPCMENAVLALAQIENSAAVQKAIAHYEQQMGQKVQLPTESLQELLD
LHRDSEREAIEVFIRSSFKDVDHLFQKELAAQLEKKRDDFCKQNQEASSDRCSGLLQVIFSPLEEEVKAGIY
SKPGGYRLFVQKLQDLKKKYYEEPRKGIQAEEILQTYLKSKESMTDAILQTDQTLTEKEKEIEVERVKAESA
QASAKMLHEMQRKNEQMMEQKERSYQEHLKQLTEKMENDRVQLLKEQERTLALKLQEQEQLLKEGFQKESRI
MKNEIQDLQTKM

Then click: "Make Prediction". JPred will tell you there is a Match found in PDB and list the proteins that match. What are they?

If you click "Continue", the server will run JPred anyway. This example has been stored on the server so will come back very quickly and you will see something like the image in Figure 2.

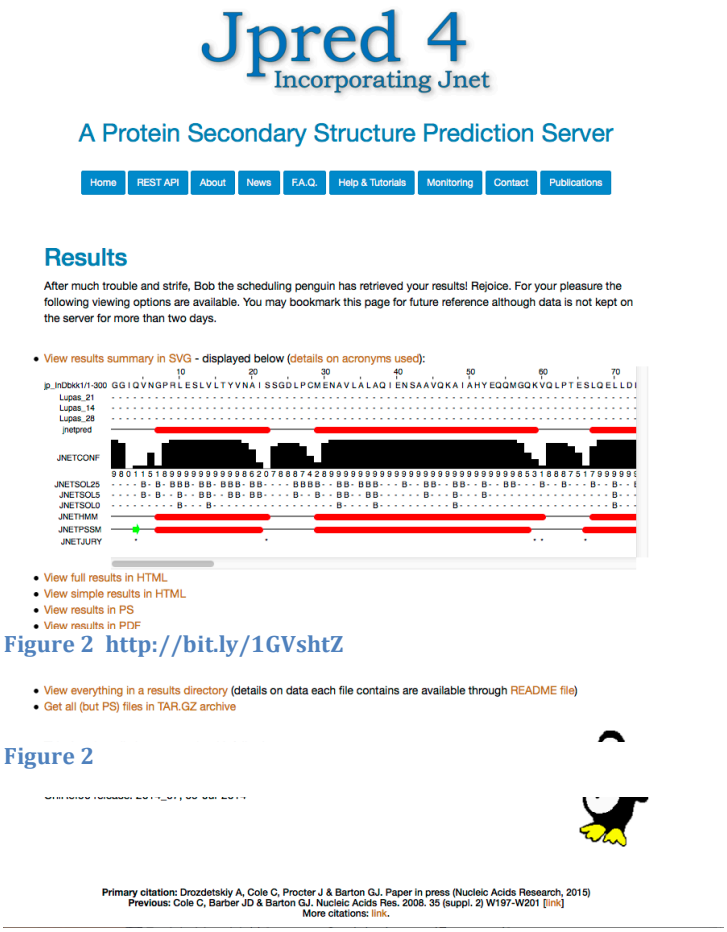


Figure 2 http://bit.ly/1GVshtZ

Figure 2

The central panel shows a summary of the results in a scrollable window. It contains:

1. the query sequence
2. LUPAS_21 and so on: predictions of coiled-coils
3. jnetpred: the summary consensus JNet prediction
4. JNETCONF: a histogram showing the confidence in the prediction
5. Numbers between 0 and 9 that also illustrate confidence (9 is most confident)
6. Three lines that represent solvent accessibility predictions. JNETSOL25 has a "B" where JNet predicts the position to be <25% solvent accessible. Similarly for JNETSOL5 and JNETSOL0.
7. JNETHMM is the prediction made by the HMMER profile network in JNet.
8. JNETPSSM is the prediction made by the PSIBLAST profile network in JNet.
9. JNETJURY shows positions where the two disagree. This is used to help make the consensus prediction.

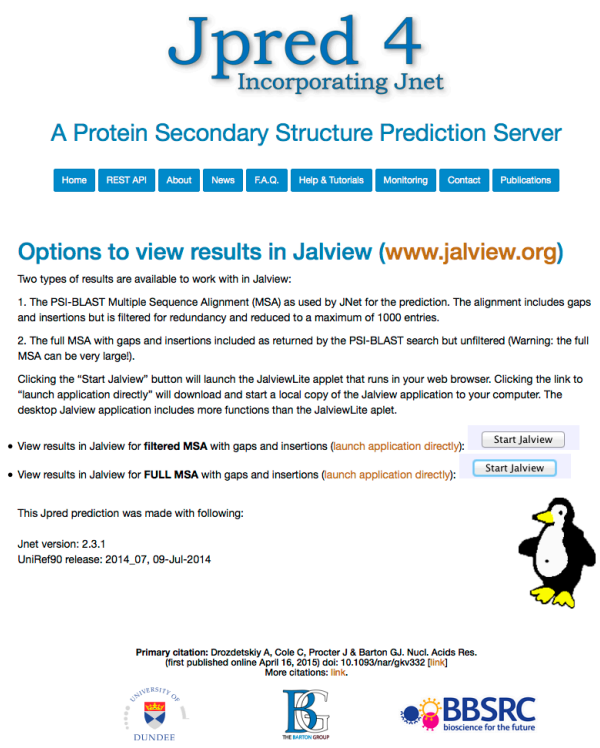
After the scrollable display there are links to further options to view the prediction data. For now, **DON'T click on the "View results in Jalview" option,** we'll do that in the next exercise.

- View FULL results in HTML: shows a representation of the multiple sequence alignment used by JPred to make the prediction as well as the predictions at the bottom.
  - **Question: What is unusual about the multiple sequence alignment?**
- View SIMPLE results in HTML: shows just the query sequence and the JNet summary prediction.
- View results in PS: provides a PostScript file that can be read into illustrator or printed.
- View results in PDF: provides the same output as the PS option, but in PDF format. **Click on this one to see!**
  - **Question: How useful is this representation of the alignment and prediction when there are so many sequences?**
- There are two options to view the full multiple sequence alignment with gaps and insertions. These are not very useful on their own, so skip this bit.

- View everything in a results directory: Provides the complete set of files produced by JPred so that you can choose what to look at or download.

Practical 2: "Viewing and interacting with the JPred prediction in Jalview"

1. Now click on the "View results in Jalview" option. This brings up a new page with some further options as shown in Figure 3 below. The "Start Jalview" button will open the JalviewLite applet and run it in the web page.



However, we are going to launch the Full Jalview Application by clicking on: "Launch application directly" next to the "Filtered MSA" row.

This will download and start up Jalview with the MSA and secondary structure prediction loaded. **Question: What do you notice about the alignment view?**

2. This protein is predicted to be mostly helical. Look at the "Conservation" histogram under the alignment. **Question: Can you see any patterns of conservation characteristic of an α-helix?** Try hiding all but the first 20 sequences. Does this help?

3. We'll now associate a PDB structure with this alignment and prediction.

Right click on the identifier for the first sequence, then select: "View Structure". In the query box, type: 1f5n and click "View". A structure viewer (JMol or Chimera) will open with a copy of the structure loaded in. Now on the "Colour" menu, select "Colour by Annotation". This opens a dialog box with "Conservation" highlighted. Select "jnetpred" from the drop-down menu to colour the PDB structure according to the JNet secondary structure prediction.

4. Use JMol or Chimera to examine the structure in detail.
Questions: Does the prediction of helix look OK? Are there any regions that are incorrectly predicted, if so, what is the most serious error? Where are the positions that appear to be conserved in a helical pattern?

Practical 3: Experimenting with prediction from alignment

We'll now see if there is any effect if you change the sequences used to make the prediction.

1. Select the first 20 sequences (It does not matter if it is exactly 20, fewer will do.) in the alignment by clicking on the identifiers and dragging down. 2. Go to the web services menu ->Secondary structure prediction-> JNet
3. Jalview will open a new window to show progress of the job and when finished, it will open a new window with the sub-alignment and prediction in it.
4. Load the PDB structure 1f5n again and the prediction on the structure in the same way as in Practical 2.

Question: Is the prediction different? Is it better or worse? Why?

In particular, look at the short β-strand at the beginning of the sequence.

Question: How do your different alignments affect prediction? Why do you think you see the β-strand in one prediction and not in the other?

Practical 4: Experiment with a sequence that works less well…

1. Make a prediction from the sequence in file: **1vyia.fasta**. You'll need to select "Advanced Options" to do this and upload the file from the Day3 directory.

2. View the prediction and alignment in Jalview. Colour the structure by the secondary structure prediction – the structure file is **1vyi**.

3. Question: What does the prediction get wrong? Why do you think it makes mistakes for this protein?

Also try this protein – **WARNING this may break Jalview since it is a large alignment:**

IKSALLVLEDGTQFHGRAIGATGSAVGEVVFNTSMTGYQEILTDPSYSRQIVTLTYPHIGNVGTNDADEESSQV
HAQGLVIRDLPLIASNFRNTEDLSSYLKRHNIVAIADIDTRKLTRLLREKGAQNGCIIAGDNPDAALALEKARAF
PG

Use structure: 1a9x . *This may break Jalview if your computer has low memory.* In order to see Chain B of the molecule, switch off the display of all other chains. In JMol this can be done from the "View" drop down menu.

Practical 5: Exploring protein structure predictions used to train and test JPred

In this exercise you will explore over a thousand different protein structural domains whose three-dimensional structure *has* already been determined by X-ray crystallography. You will look at some of their alignments and some of the predictions. Don't worry you won't need to look at all of them!

If you follow this link:

http://www.compbio.dundee.ac.uk/jpred4/jnet231retrTable_blind.shtml

It will take you to a table containing all the 150 protein domains in the JPred4 Blind Test. There are a lot of columns in this table, they are all listed here, but the most important ones are highlighted in BLUE.:

1. Domain ID. This looks like: d**1bco**a1. The part in bold is the PDB ID code, the "a" is the chain.
2. Score: the percentage accuracy of the prediction.
3. Length: the length of the sequence.
4. Num.Seq.Full: the number of sequences in the alignment returned by PSI-BLAST.
5. Num.Char.Full: the number of characters in the multiple alignment including gaps.
6. Num.Seq.Filt: the number of sequences in the filtered alignment
7. Num.Char.Filt: the number of characters in the filtered multiple alignment.
8. Timing (min): the time JPred took to run this prediction
9. Protein: The protein domain name according to SCOPe
10. Class: the SCOPe domain structural class
11. Fold: the SCOPe domain fold
12. Superfamily: the SCOPe superfamily

13. Family: the SCOPe
14. Species
15. SCOPeURL: repeats the Domain ID as a hyperlink to the SCOPe page for this domain.
16. Sequence with DSSP details and JNet prediction: The JNet prediction if clicked will take you to the JNet results page for Jalview.

You can sort the table for the numerical columns by clicking on the column header. Try sorting by score, then looking down the set of predictions in the last column of the table.

Question: If you sort the table by "Score", which structural class of protein has the best score? Which do less well? Why do you think this is?

You should explore some of the predictions by clicking on them and then loading the structure and colouring by prediction as you did in Practical 2 above.

If you are feeling adventurous, then explore the following table as well:
http://www.compbio.dundee.ac.uk/jpred4/jnet231retrTable_train.shtml

This is the full set of domains used to train JNet.

Question: Again, which class of protein domain does well and which does badly in the predictions?

Question: Are there any "odd" domains with unusual secondary structure composition?

Other features of the JPred webserver

In these practicals we have only used the webserver to predict from a single sequence. However, the server also supports uploading of a multiple alignment (MSA) and allows for batch submission of many sequences. For help with how to use these functions, see the tutorials that are listed on the website.

Lecture: Protein Sub-Family analysis

- Summarises principles of sub-family analysis
- Shows example of the Annexins and prediction of charge-pair
- Discusses tree-determinants

Practical 6: Identify "interesting" positions in an Annexin alignment

In this practical you will repeat the analysis described in the lecture, but using Jalview to find the salt-bridge pair in this sequence family.

1. Read the file "lall_ra_mult.blc" into Jalview either through the File menu or by dragging and dropping. Hint, this multiple alignment file is in BLC file format.
2. Calculate a tree using the "Average Distance Using PAM250" method and look at the resulting tree. You should see two outlier sequences at the top of the tree.
3. Identify the outliers and "hide" them from the multiple alignment, then remove empty columns using the option under the Edit menu.
4. Select all visible sequences and recalculate the tree in the same way.

5. On the new tree you should be able to see four clear groups. Cut the tree by clicking on the tree at a position that will divide the groups into four.
6. The groups will now be colour coded on both the multiple alignment and the tree.
7. Under Calculate->Sort sort the alignment according to the tree.
8. Now colour the alignment using ClustalX colouring and click "Colour by conservation".
9. In the conservation colour increment box, tick "Apply to all groups", then move the slider to the right. This has the effect of highlighting the most highly conserved positions in each group.
10. Inspect the most heavily coloured positions to see if you can see pairs that are highly conserved at a position in two groups, but are different amino acid types. You should be able to see a position where Arg (R) is conserved in one group, but Glu (E) is conserved in another.
11. Moving the conservation slider to the left a bit will reveal more subtle conservation.

The program AMAS which can be run online at www.compbio.dundee.ac.uk/amas provides more sophisticated automated methods to do sub-group analysis, but it is not interactive.

Practical 7: Repeat the same analysis but with a much larger contemporary alignment

This is probably one you will have to try at home. The point of the exercise is to illustrate the challenges of working with hundreds of sequences. Select one of the sequences in the Annexing alignment you have been working with.

1. Run the JNet secondary structure prediction program on it from within Jalview, or alternatively copy-paste the sequence into the JPred website.
2. The job will take a few minutes to run and will return a sequence alignment with over 600 sequences in it. If it is too slow, just load in the file "big_annexins.fasta" and work from that.
3. Try calculating a tree for this alignment and then use the tree to identify outliers you can hide to give you an alignment with four groups.

This is hard to do but the Jalview team are actively researching methods to make it easier to work with large alignments for this type of analysis!