Multiple Sequence Alignment using the Jalview alignment workhorse and viewer

by Toby Gibson

In these learning by doing exercises, we will introduce and work with Jalview, the Java Alignment Viewer. Jalview is powerful visualisation software that can allow alignments to be generated, manipulated, edited and annotated. It interfaces remotely with tools such as multiple sequence alignment programs and secondary structure predictors. We will visualise alignments of modular proteins with Jalview, discussing sequence features such as folded protein domains, short functional peptide motifs and natively disordered polypeptide. It is important to be aware of these structure-function modules that underlie cell regulation and they will reappear regularly during the course.

The Jalview developers provide extensive online help. The developers have also prepared training videos for YouTube. You can access these at the Jalview Youtube channel.

Installing Jalview

Go to the Jalview home page. Find and click on the download button. Download and install Jalview v.2.11 for your operating system. It should just work, but some computer setups have problems with Java. Ask a trainer to help if you are stuck – if all fails in the time available, you will have to work with your partner.

We will do the exercises together and I will project on the screen.

PART 1. USING JALVIEW WITH EPSINS

Epsins are important proteins for receptor mediated endocytosis. And they illustrate protein modularity very well.

- Start Jalview
  On first use, Jalview shows off what it can do. This can be turned off in Tools>Preferences.
- Load a set of sequences by cut and paste into Jalview using:
- File->Input Alignment-> from Text Box
- Get the sequences from this page of Epsin Sequences in FASTA format
- Examine the sequences. Mouse over the names. What kind of organisms are represented?
- Remotely align sequences via the Web Service->Alignment-> ClustalO (I usually use Clustal Omega) and run with defaults
- Use the colour menu to colour the alignment in various ways.
- Examine the alignment, identify possible regions of misalignment, and try correcting these by moving bits of sequence as described in the Jalview documentation; remember to Select->Deselect All if you are unable to make the edits you want
- Get a remote secondary structure prediction by the JNet server using the Web Service link
- Select the Epn1_Human sequence. Get a remote natively disordered structure prediction from one of the Protein Disorder Web Service links
- Look at the conservation of some short linear motifs: Select and copy over the motif text pattern shown in bold and paste into the Select>Find box, then clicking “Find All”

- NPF motif RegExp is **NPF**
- DPW motif RegExp is **DP[FW]**
- Clathrin box RegExp is **L[IVLMF].[IVLMF][DE]**

(These motifs have entries in the ELM Server, DPW, NPF, and clathrin boxes which we will introduce tomorrow.)

- Follow this up by creating and naming a New Feature from the pattern matches
- Sort the order of sequences in the alignment, clustering by pairwise identity Calculate->Sort->By Pairwise Identity
- Save the annotated alignment data in a file in Jalview format – this allows you to examine in future these and other featuresannotations you may add to your alignment File->Save As->FileFormat Jalview (.jar)

QUESTIONS:

1. What is the basis for the "ClustalX" colouring scheme provided by Jalview?
   a. which residues are assigned similar colours?
   b. which residues are assigned different colours?
   c. in which situations are residues left uncoloured?
   d. are there residues that are always coloured? Why is that useful?
   e. try some other colouring schemes.
2. Why are amino acid residues conserved in evolution?
3. Are matches to the linear motif regular expressions more likely to be conserved in regions known/predicted to be globular or in IUP regions? Does the JNet 2D structure predictor suggest large numbers of alpha helices and beta strands throughout the alignment?
4. What are the characteristics of regions of the MSA that you expect to be well aligned? Consider:
   a. residue identity/property conservation
   b. number and size of gaps
5. Would you expect the affinity of a domain – linear motif interaction to be higher or lower than one between two domains? Why? What assumptions are you making about the nature of these different kinds of interactions?

PART 2. USING JALVIEW WITH P53

p53 is a "master regulator" of apoptosis and the "guardian of the genome". (I don't like the term master regulator…) We will now make a p53 alignment using this linked set of unaligned p53 sequences. Search in the sequences and make new features with different colours and names for the following motifs: - Cyclin RxL motif – Long CDK site - SUMO modification site. Paste their regular expressions into the Select>Find box

- Cyclin RxL is

(.|([KRH].{0,3}))[^EDWNSG][^D][RK][^D]L.{0,1}[FLMP].{0,3}[EDST]

- Long Cdk is **...([ST])P..[RK]**
- SUMO site is **[VILMAFP](K).E**

QUESTIONS:

1. Some of the sequences included in the MSA are shorter than others. Why might this be? Do all sequences begin with a Met residue? If you want a high quality alignment, will you keep these sequences or discard them?
2. Do all p53 sequences have cyclin box candidates? Are they all in the same place in the sequences?

3. Do all p53 sequences have CDK sites? CDK sites require cyclin boxes to function: Is there any correlation between the presence or absence of cyclin and CDK sites?
4. Do all p53 sequences have SUMO sites? Can they all be aligned? If not, is there an evolutionary process that can account for their change in position?

PART 3. USING JALVIEW WITH XYXIN, A FOCAL ADHESION PROTEIN THAT INDUCES ACTIN FILAMENTS

The EVH1 domain in proteins like VASP binds to a Proline-rich motif. This interaction enables VASP to induce actin filament formation. Zyxin is a VASP-interacting protein, enabling acting filament induction at focal adhesions.

Make an alignment in Jalview from this set of unaligned Zyxin sequences. Then find all examples of this EVH1-binding motif **([FYWL]P.PP)|([FYWL]PP[ALIVTFY]P)**

QUESTIONS:

1. Is any part of Zyxin folded?
2. Might there be Zinc Fingers in Zyxin? What do we need to look for?
3. How many EVH1 motifs are there in human Zyxin?
4. How many EVH1 motifs are there in zebra fish (look for *_DANRE*) Zyxin?
5. Why do you think there are multiple EVH1 motifs in Zyxin?
6. Make a tree from the sequence divergence under the Calculate menu. Find both the PIG branches. What are the nearest relatives in the tree? Do they make sense? Using the control key on the name of the mystery relative, we can retrieve the UniProt entry to determine the species – is it an ungulate?

PUBLICATIONS

- *Jalview Version 2–a multiple sequence alignment editor and analysis workbench.* Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Bioinformatics. 2009 May 1;25(9):1189-91. PMID: 19151095
- *Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.* Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. Mol Syst Biol. 2011 Oct 11;7:539. PMID: 21988835