MSA & Jalview

Toby Gibson

MSA & JALVIEW

In these 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. These structure-function modules will reappear regularly during the course.

The Jalview developers have prepared training videos for YouTube. You can access these at the Jalview Youtube channel.

We want to get Jalview installed on your computers before starting.

PART 1. USING JALVIEW WITH EPSINS

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

- 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
- Remotely align sequences via the Web Service->Alignment-> options (Choose a service like 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 motifs from the ELM Server, DPW, NPF, and clathrin boxes] using JalView Select->Find...-> and copying over the appropriate motif text pattern, and then clicking Find All*]. Select and copy over the motif text pattern shown in bold, and then clicking Find All

  1. - NPF motif RegExp is **NPF**
  1. - DPW motif RegExp is **DP[FW]**

     - Clathrin box RegExp is **L[IVLMF].[IVLMF][DE]**
  1.

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

QUESTIONS:

1. What is the basis for the "ClustalX" colouring scheme provided by Jalview?
   - which residues are assigned similar colours?
   - which residues are assigned different colours?
   - in which situations are residues left uncoloured?
   - are there residues that are always coloured? Why is that useful?
   - try some other colouring schemes.
2. 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?
3. What are the characteristics of regions of the MSA that you expect to be well aligned? Consider:
   - residue identity/property conservation
   - number and size of gaps
4. 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 ELM motifs: - Cyclin box – Long CDK site - SUMO modification site - SUMO reversed site - MDM2-interaction motif.

- Cyclin box motif is **(.|([KRH].{0,3}))[ˆEDWNSG][ˆD][RK][ˆD]L.{0,1}[FLMP].{0,3}[EDST]**
- Long CDK site motif is **...([ST])P..[RK]**
- Sumo modification site is **[VILMAFP](K).E**
- MDM2 degron site is **F[ˆP]{3}W[ˆP]{2,3}[VIL]**

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 MDM2 interaction motifs? Are they all the same length? How can an alpha helical sequence vary in length?
3. Do all p53 sequences have cyclin box candidates? Are they all in the same place in the sequences?
4. 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?
5. 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?
6. Do all P53 sequences have reverted SUMO sites? Can they all be aligned?
   - If some SUMO sites cannot be aligned, is there an evolutionary process that can account for their change in position?

PART 3. USING JALVIEW WITH Tir PROTEIN ISOLATES FROM PATHOGENIC E. COLI

Tir proteins are secreted by pathogenic E. coli. They attach to targeted mammalian cells and both the N- and C- termini enter through the membrane, taking over the local cell regulation and, with other inserted proteins, induce the actin pedestal. The central portion of Tir remains extracellular and is bound by the bacterium. Many Tir isolates have been sequenced and are in Uniprot. Load by cut and paste this already aligned set of Tir proteins into Jalview. Use the p53 motifs above to find the Cyclin and CDK motif entries and use the regular expressions to create new features in all sequences.

- Do all sequences have both motifs?
- Are they all alignable, or can they move around?
- Some are juxtaposed – can they both be functional at the same time?

Note that as far as we know in creating this exercise, these motifs have not been studied, but there is some evidence that cell cycle is disrupted by pathogenic E. coli (e.g. PMID: 11598051).

Now put an SH2-binding motif **Y..[IVLM]** regular expression into the alignment and make new features

- Do the sequences have matches to SH2 motifs?
- Do you think the Tir proteins are phosphorylated by Tyrosine Kinases?

Proteins that are natively disordered, and contain linear motifs to control cell regulation, are known to be secreted by pathogens into the cells that they take over.

Now find the PRMT1 Arginine methylase motif **GGRGG** - Do you think Tir is a substrate?

Now find the **NPY** motif which binds the I-BAR domain and is essential for pedestal formation. This motif is well described in bacteria but not yet in a human host cell protein (PMID:21893288).

Tir has a lot of known motifs that interact with host proteins. However there is still a lot of conserved sequence, suggesting that Tir will make more interactions than have yet been described.

PART 4. USING JALVIEW WITH CagA PROTEIN ISOLATES FROM PATHOGENIC Helicobacter

CagA effector proteins are secreted by pathogenic Helicobacter directly into the cytosol. These large proteins modulate the actin cytoskeleton and the overall status of the cell. Load by cut and paste this already aligned set of CagA proteins into Jalview. The EPIYA motif regular expression from ELM is **EP[IL]Y[TAG]** – use it to search the alignment, making a new feature.

- Do the sequences have one EPIYA motif or do they have more?
- Do they all have the same number?
- What is the most EPIYA motifs in one protein?
- Do any of the EPIYA motifs match to typical **Y..[IVLM]** SH2 motifs?
- Do you think the CagA proteins are phosphorylated by Tyrosine Kinases?

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