EXPLORING SHORT LINEAR MOTIFS IN PATHOGENS

Manjeet Kumar, Toby Gibson & Holger Dinkel

PART 1: USING JALVIEW WITH Tir PROTEIN ISOLATES FROM PATHOGENIC *E. COLI*

Tir proteins are secreted by pathogenic EHEC and EPEC *E. coli* Strains. 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.

Cyclin box motif - **(. | ([KRH].{0,3}))[^EDWNSG][^D][RK][^D]L.{0,1}[FLMP].{0,3}[EDST]**
Long CDK site motif - **...([ST])P..[RK]**

Use the 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?

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 with no known function, suggesting that Tir will make more interactions than have yet been described.

See which of these motifs have been annotated in ELM by querying with UniProt:TIR_ECO27 – leave cell compartment and species blank; a later exercise shows how to set up pathogen queries correctly.

PART 2: 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 the already aligned set of CagA proteins in 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?

PART 3:

A. SEARCHING ELM (http://elm.eu.org) FOR BACTERIAL EFFECTOR PROTEINS

ELM contains information for Eukaryotic Linear Motifs although we do capture non-Eukaryotic motifs in cases where motif mimicry might be involved. To explore a bacterial protein in ELM, we will use the IDP-rich TarP effector from *Chlamydophila caviae* for which the natural host is guinea pig (*Cavia porcellus*). The UniProt accession for this protein is Q824H6_CHLCV; search ELM using this accession.

- How may motifs you retrieve for the protein?
- TarP is extracellular for the bacterium, think if the default search is enough in this case?
- What should be the correct compartment and taxonomic filter for the host?
- Change the compartment and taxonomic filter accordingly and rerun the ELM search.
- How many VBS motifs do you on the TarP protein and what is their biological relevance for pathogen?

Other interesting bacterial proteins to check for motif mimicry are *Chlamydomonas* **Y005_CHLTR and Cholera C5IZN1_VIBCL**

B. EXPLORING ELM (http://elm.eu.org) FOR VIRAL PROTEINS - E1A adenoviral Protein

Adenoviruses are non-enveloped dsDNA viruses. Human adenoviruses are responsible for respiratory diseases, croup, and bronchitis outbreaks and gastroenteritis in children. The adenovirus E1A protein is unique to the Mastadenovirus genus. All members of the Mastadenovirus genus infect mammals. E1A plays a role in viral genome replication by driving entry of quiescent cells into the cell cycle. Stimulation of progression from G1 to S phase allows the virus to efficiently use the cellular DNA replicating machinery to achieve viral genome replication.

1. Search in ELM E1A_ADE05. Remember to define cellular compartments and taxonomic context appropriately for the host.

a) What can you say about the structure of the protein?

b) How many annotated instances are?

c) How many annotated instances belong to cellular targets? How many are related?

d) How many linear motifs for kinases are annotated and how many are predicted?

2. Search in ELM E1A_ADE02. Remember to define cellular compartments and taxonomic context.

a) What can you say about the structure of the protein? Is this different from E1A_ADE05?

b) How many annotated instances are? Are those different from E1A_ADE05?

c) How many annotated instances belong to cellular targets? How many are related?

d) How many instances are assigned by homology?

e) How many linear motifs for kinases are annotated and how many are predicted?

3. If you have to test which kinase phosphorylates E1A, which of all the predictions would you test?

4. Search in ELM E1A_ADECR.

a) Which is the taxonomic context?

b) How many instances are annotated? Why do you think is that?

c) What can you say about the structure of the protein? What can you say in general about E1A proteins?

Other interesting proteins to check for motif mimicry in ELM are Vaccina A36_VACCW and Ebola NCAP_EBOZM

Useful resources:

1. **http://slim.icr.ac.uk/articles/**
2. **http://slim.ucd.ie/slimsearch/**

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
- ELM-the eukaryotic linear motif resource in 2020. Kumar M, Gouw M, Michael S, Sámano-Sánchez H, Pancsa R, Glavina J, Diakogianni A, Valverde JA, Bukirova D, Čalyševa J, Palopoli N, Davey NE, Chemes LB, Gibson TJ.Nucleic Acids Res. 2019 Nov 4. pii: gkz1030. doi: 10.1093/nar/gkz1030.