Short Linear Motifs and the Eukaryotic Linear motif resource

Toby Gibson & Holger Dinkel with help from Juliana Glavina and Hugo Samano

[**Resources**]

UniProt http://www.uniprot.org/

ELM http://elm.eu.org

SlimSearch http://slim.ucd.ie/slimsearch/

PeCan https://pecan.stjude.cloud/proteinpaint/

ProViz http://proviz.ucd.ie

[ELM exercises]

Objective: Get familiar with the ELM (Eukaryotic Linear Motif) prediction tool.

1. Search in ELM by copy/pasting the following sequence and using the following parameters:

> P12931

MGSNKSKPKDASQRRRSLEPAENVHGAGGGAFPASQTPSKPASADGHRGPSAAFAPAAAE
PKLFGGFNSSDTVTSPQRAGPLAGGVTTFVALYDYESRTETDLSFKKGERLQIVNNTEGD
WWLAHSLSTGQTGYIPSNYVAPSDSIQAEEWYFGKITRRESERLLLNAENPRGTFLVRES
ETTKGAYCLSVSDFDNAKGLNVKHYKIRKLDSGGFYITSRTQFNSLQQLVAYYSKHADGL
CHRLTTVCPTSKPQTQGLAKDAWEIPRESLRLEVKLGQGCFGEVWMGTWNGTTRVAIKTL
KPGTMSPEAFLQEAQVMKKLRHEKLVQLYAVVSEEPIYIVTEYMSKGSLLDFLKGETGKY
LRLPQLVDMAAQIASGMAYVERMNYVHRDLRAANILVGENLVCKVADFGLARLIEDNEYT
ARQGAKFPIKWTAPEAALYGRFTIKSDVWSFGILLTELTTKGRVPYPGMVNREVLDQVER
GYRMPCPPECPESLHDLMCQCWRKEPEERPTFEYLQAFLEDYFTSTEPQYQPGENL

- Cell Compartment: **Not specified**
- Motif Probability Cutoff: **100**
- Context information: **(leave blank)**

1. Pay attention to many instances you find
2. What can you say about the structure of the protein?
    a. Do you find any domains?
    b. Do you find any disordered regions?

2. Repeat the previous search (again accession P12931) using these parameters:

- Cell Compartment: **cytosol**
- Motif Probability Cutoff: **0.01**
- Context information: **Homo sapiens**

1. How many instances (roughly) do you find now?

2. How many of the instances are 'annotated'?
3. Do the structural predictors/filters (SMART, GlobPlot, IUPRED, Secondary Structure) agree in terms of which regions are structured/disordered?
4. Compare the location of the annotated instances with structural information at hand (IUPRED, Secondary Structure).

3. Submit the sequence of Paxillin (P49023) to ELM, using default parameters.

1. Compare the results with a search for the same sequence when using the cellular compartment 'plasma membrane'

4. Search protein SRC_MOUSE (P05480) for ELMs.

1. Do you find "annotated instances"?
2. If not, what's the closest to an 'annotated instance' that you can find? Investigate where this information might come from.

5. Submit the entry name 'P53_HUMAN'

1. Do the cell compartments make sense?
2. How many degrons are there in p53?
3. Is there a CDK site in p53? Is there a Cyclin Box in p53?

6. (Optional) Search ELM using the protein name 'MDM4_HUMAN' and look for the 'USP binding motif' DOC_USP7_MATH_1.

1. How many such motif instances are found in this protein sequence?

7. (Optional) Repeat this exercise with protein 'AMPH_HUMAN' and ELM class 'LIG_Clathr_ClatBox_1'

1. Try to assess the biological relevance of each of these instances.
2. Is the annotation for the biological relevance in accordance with the globular structure?

8. (Optional) Get all annotated instances for "Homo sapiens" that contain the search term "cilium"

(Hint: Use url http://elm.eu.org/elms/browse_instances.html).

1. How many are there?
2. Which experimental evidence is annotated and how reliable is this evidence?
3. Try to get these instances TSV-file (tab-separated values)

E1A adenoviral Protein

Objective: Apply the ELM (Eukaryotic Linear Motif) prediction tool to a viral protein.

Background Information: Adenoviruses are non-enveloped DNAds virus. 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 infects 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.

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 phosphorylation sites are annotated in Phospho.ELM?

e) 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 phosphorylation sites are annotated in Phospho.ELM?

f) 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?

Helicobacter pylori CagA

Objective: Use ELM to predict Eukaryotic Linear Motifs in bacterial proteins.

Background Information: *H. pylori* infection causes gastritis, peptide ulcer or gastric cancer. There is a stronger probability to develop gastric cancer if an East Asian strain (like F32) is responsible for the infection compared to a Western strain (like NCTC 11637). East Asian and Western strains differ in the number and sequence context of the EPIYA motifs. (Higashi, H., et al., 2002; Jones, K.R., et al., 2009)

1. Paste in ELM prediction server the following sequences of CagA from a Western and an East Asian strain. Specify 'Cytosol' cell compartment, '*Homo sapiens*' and a Motif probability cutoff of 0.001.

> NCTC11637_CagA

MTNETIDQQPQTEAAFNPQQFINNLQVAFLKVDNAVASYDPDQKPIVDKNDRDNRQAFDGISQLREEYSNKAIKNPTKKN
QYFSDFINKSNDLINKDNLIDIGSSIKSFQKFGTQRYRIFTSWVSHQNDPSKINTRSIRNFMENIIQPPIPDDKEKAEFL
KSAKQSFAGIIIGNQIRTDQKFMGVFDEFLKERQEAEKNGEPTGGDWLDIFLSFVFNKEQSSDVKEAINQEPVPHVQPDI
ATTTTHIQGLPPESRDLLDERGNFSKFTLGDMEMLDVEGVADIDPNYKFNQLLIHNNALSSVLMGSHNGIEPEKVSLLYA
GNGGFGAKHDWNATVGYKNQQGDNVATLINVHMKNGSGLVIAGGEKGINNPSFCLYKEDQLTGSQRALSQEEIRNKIDFM
EFLAQNNAKLDNLSEKEKEKFQNEIEDFQKDSKAYLDALGNDRIAFVSKKDPKHSALITEFGKGDLSYTLKDYGKKADRA
LDREKNVTLQGNLKHDSVMFVNYSNFKYTNASKSPDKGVGVTNGVSHLDAGFSKVAVFNLPDLNNLAITSFVRRNLENKL
VTEGLSLQEANKLIKDFLSSNKELVGKALNFNKAVADAKNTGNYDEVKKAQKDLEKSLRKREHLEKEVEKKLESKSGNKN
KMEAKAQANSQKDKIFALINKEANRDARAIAYSQNLKGIKRELSDKLEKINKDLKDFSKSFDEFKNGKNKDFSKAEETLK
ALKGSVKDLGINPEWISKVENLNAALNEFKNGKNKDFSKVTQAKSDLENSVKDVIVNQKITDKVDNLNQAVSMAKATGDF

SRVEQALADLKNFSKEQLAQQTQKNESFNVGKKSEIYQSVKNGVNGTLVGNGLSGIEATALAKNFSDIKKELNEKFKNFN
NNNNNGLENEPIYAKVNKKKTGQVASPEEPIYAQVAKKVNAKIDRLNQAASGLGGVGQAGFPLKRHDKVDDLSKVGRSVS
PEPIYATIDDLGGPFPLKRHDKVDDLSKVGRSVSPEPIYATIDDLGGPFPLKRHDKVDDLSKVGRSVSPEPIYATIDDLG
GPFPLKRHDKVDDLSKVGLSRNQELAQKIDNLSQAVSEAKAGFFSNLEQTIDKLKDSTKYNSVNLWVESAKKVPASLSAK
LDNYATNSHTRINSNIQNGAINEKATGMLTQKNPEWLKLVNDKIVAHNVGSVPLSEYDKIGFNQKNMKDYSDSFKFSTKL
NNAVKDVKSSFTQFLANAFSTGYYSLARENAEHGIKNVNTKGGFQKS

> F32_CagA

MTNETIDQTTTPDQTGFVPQRFINNLQVAFIKVDNAVASFDPDQKPIVDKNDKDNRQAYEKISQLREEYANKAIKNPAKK
NQYFSDFINKSNDLINKDNLIAVDSSVESFRKFGDQRYQIFTSWVSLQKDPSKINTQQIRNFMENVIKPPISDDKEKAEF
LRSAKQSFAGIIIGNQIRSDEKFMGVFDESLKARQEAEKNAEPAGGDWLDIFLSFVFNKKQSSDLKETLNQEPRPDFEQN
LATTTTDIQGLPPEARDLLDERGNFFKFTLGDVEMLDVEGVADKDPNYKFNQLLIHNNALSSMLMGSHSNIEPEKVSLLY
GDNGGPEARHDWNATVGYKNQQGNNVATLINAHLNNGSGLIIAGNEDGIKNPSFYLYKEDQLTGLKQALSQEEIQNKVDF
MEFLAQNNAKLDNLSEKEKEKFQTEIENFQKDRKAYLDALGNDHIAFVSKKDPKHLALVTEFGNGELSYTLKDYGKKQDK
ALDGETKTTLQGSLKYDGVMFVNYSNFKYTNASKSPNKGLGTTNGVSHLEANFSKVAVFNLPNLNNLAITNYIRRDLEDK
LWAKGLSPQEANKLIKDFLNSNKEMVGKVSNFNKAVAEAKNTGNYDEVKKAQKDLEKSLRKREHLEKEVAKKLESRNDNK
NRMEAKAQANSQKDKIFALISQEASKEARVATFDPYLKGVRSELSDKLENINKNLKDFGKSFDELKSGKNNDFSKAEETL
KALKDSVKDLGINPEWISKIENLNAALNDFKNGKNKDFSKVTQAKSDLENSIKDVIINQKITDKVDNLNQAVSEIKLTGD
FSKVEQALAELKNLSLDLGKNSDLQKSVKNGVNGTLVSNGLSKTEATTLTKNFSDIRKELNEKLFGNSNNNNNGLKNNTE
PIYAQVNKKKTGQATSPEEPIYAQVAKKVSAKIDQLNEATSAINRKIDRINKIASAGKGVGGFSGAGRSASPEPIYATID
FDEANQAGFPLRRSAAVNDLSKVGLSREQELTRRIGDLSQAVSEAKTGHFGNLEQKIDELKDSTKKNALKLWVESAKQVP
TSLQAKLDNYATNSHTRINSNVQSGTINEKATGMLTQKNPEWLKLVNDKIVAHNVGSAPLSAYDKIGFNQKNMKDYSDSF
KFSTKLNNAVKDIKSSFVQFLTNTFSTGSYSLMKANVEHGVKNTNTKGGFQKS

1. What are the differences in EPIYA motif predictions? Is the 'Assigned by homology' indicator showing any difference?

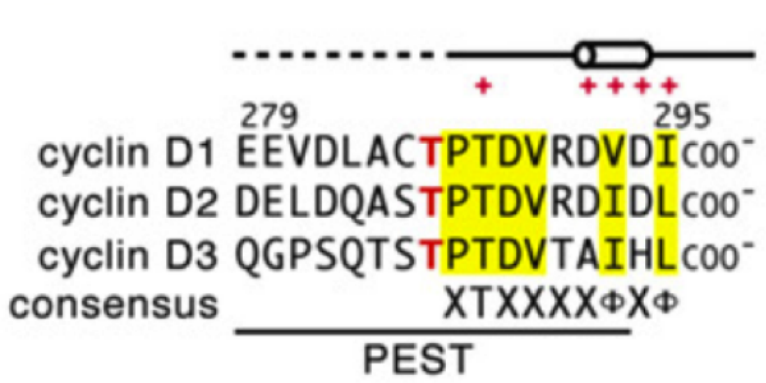
[PeCan exercise]

- Go to elm.eu.org and enter ETV1_HUMAN as search term; you should find a single annotated true positive instance of a degron. Where/What ELM class is it? Note down it's amino acid position.

Now go to PeCan and enter ETV1 as search term. Do you fidn any pediatric mutations? How about Cosmic? Select only "Fusion transcript" mutations; can you find any near the degron motif? Which tissue types are these mutations found in? Are there differences in fusions and frameshifts in tissue types?

- Go to elm.eu.org and enter NOTC1_HUMAN as search term. Find the annotated FBW7 degron (TP) and note down its position.

Now go to PeCan and enter notch1 as search term. Select only "Frameshift" and "nonsense" mutations; if possible load COSMIC data as well; analyse the area around the degron; additionally, by adding "missense" is there another hotspot?



- CCND2 possesses a degron in the C-Terminus, however the entry has not yet been annotated in elm.eu.org. Fortunately, a new structure came out recently: PMID 29279382; open CCND2 in PeCan and look at the differences in pediatric & Cosmic data

(graphics from: "

Structural basis of the phosphorylation-independentrecognition of cyclin D1 by the SCF^{FBXO31} ubiquitin ligase" *Li* PNAS 2018)

[ProViz exercise]

ProViz aggregates and displays useful information from many resources where relevant to linear motif discovery.

Go to the **ProViz server** http://proviz.ucd.ie

Put p53 into the **Search for a Protein** field

Explore the results, then check out these operations and questions:

- Are there any methyllysine modifications?
- What is a cumulative switch?
- Are there any motifs in p53 than can antagonise MDM2 ubiquitination?
- Are there any splice variants in the DNA-binding domain? Do any motifs get removed in splice variants? Do you think p53 researchers study all the splice variants?
- What do the long structure modules correspond to?
- Can tetrameric p53 be exported from the nucleus?
- Click on the structure element of a short segment of p53. It goes to the PDB website (RCSB). Display the structure in the browser and adjust the viewing controls.
- What is the difference between mutagenesis sites and sequence variants?
- Which residue has the most sequence variants listed?
- Are there any sequence variants in the MDM2 degron (towards the N-terminus)?

[References:]

Alexander et al. Sci. Sig 2011 “Spatial exclusivity combined with positive and negative selection of phosphorylation motifs is the basis for context-dependent mitotic signaling” [URL]

Davey NE, Travé G and Gibson TJ (2011), *"How viruses hijack cell regulation"*, Trends Biochem Sci., Mar, 2011. Vol. 36, pp. 159-169. [DOI] [URL]

Davey NE, Van Roey K, Weatheritt RJ, Toedt G, Uyar B, Altenberg B, Budd A, Diella F, Dinkel H and Gibson TJ (2012), *"Attributes of short linear motifs"*, Mol Biosyst., Jan, 2012. Vol. 8, pp. 268-281. [DOI] [URL]

Dinkel H, Van Roey K, Michael S, Davey NE, Weatheritt RJ, Born D, Speck T, Krüger D, Grebnev G, Kuban M, Strumillo M, Uyar B, Budd A, Altenberg B, Seiler M, Chemes LB, Glavina J, Sánchez IE, Diella F, Gibson TJ. (2015), "*The eukaryotic linear motif resource ELM: 10 years and counting.*" Nucleic Acids Res., Nov, 2013. [DOI] [URL]

Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, Gibson TJ and Diella F (2011), *"Phospho.ELM: a database of phosphorylation sites--update 2011."*, Nucleic Acids Res., Jan, 2011. Vol. 39 (Database issue), pp. D261-D267. [DOI] [URL]

Dyson HJ and Wright PE (2005), *"Intrinsically unstructured proteins and their functions"*, Nat Rev Mol Cell Biol., Mar, 2005. Vol. 6, pp. 197-208. [DOI] [URL]

Van Roey K, Orchard S, Kerrien S, Dumousseau M, Ricard-Blum S, Hermjakob H and Gibson TJ (2013), *"Capturing cooperative interactions with the PSI-MI format"*, Database (Oxford). Vol. 2013, pp. bat066. [DOI] [URL]

Van Roey K, Dinkel H, Weatheritt RJ, Gibson TJ, Davey NE. (2013) "*The switches.ELM resource: a compendium of conditional regulatory interaction interfaces.*" Sci Signal. 2013 Apr 2;6(269):rs7. [DOI] [URL]

Gibson TJ, Dinkel H, Van Roey K, Diella F (2015) "Experimental detection of short regulatory motifs in eukaryotic proteins: tips for good practice as well as for bad", Cell Communication & Signalling [URL]

Mészáros B, Kumar M, Gibson TJ, Uyar B, Dosztányi Z. (2017) "Degrons in cancer." Sci Signal. 2017 Mar 14 [URL]