Mapping protein-protein interactions with ELM and 3DID

Marina Vallejo Vallés
Computational and Synthetic Biology Group
Group Leader: Dr. Jae-Seong Yang
15/07/2020

ABOUT ELM

"Eukaryotic Linear Motif". Online resource that contains experimentally validated SLiMs.

From this database we obtain a .tsv file with 338 interaction domains annotated.

These domains stablish interactions with linear-motifs of other proteins.

Each one has information about:

ELM ID / INTERACTION DOMAIN ID (Pfam) / INTERACTION DOMAIN NAME / INTERACTION DOMAIN DESCRIPTION

Our interest is to finally determine which of the protein-protein interactions are domain-linear motif (DLI).

This kind of interactions are in part explained by 3DID. So once we've done the mapping with ELM it will remain to do the mapping with 3DID DMI data.

DLI = ELM + 3DID DMI

SLiMs

"Short Linear Motifs".

Compact protein interaction sites that are composed of short parts of adjacent aminoacids.

They are very present in the intrinsically altered regions of eukaryotic proteomes and are very important for cell signaling and protein regulation. Some of them mediate interactions between proteins as they provide a binding site for peptide-binding domains. These interactions are DLI type.

Intrinsically disordered region covers the 30% of human proteome.

The estimation is that human proteome holds more tan 100000 SLiMs. Only a small part has been discovered. For the large-scale identification we need experimental methods.

The SLiM-based interactions are not easily detected due to their nature. But they are of great interest as many diseases such as cancer alter their function. Also, pathogens use to mimic them in order to hijack hosts system.

DLI vs DDI

Physical characteristics of PPI are determined by interface structure. There are two main groups:

- DDI: Domain/Domain Interactions. Typical case of big/strong interfaces between two globular domains. Tend to connect proteins within the same biological modules. Stronger fold affinities than DLI.

- DLI: Domain/Linear-Motif Interactions. Common in small and weak interfaces between short peptides. Enriched in protein interactions connecting different functional groups.

The PPI classification allows to find a relationship between interaction strength and modular architecture of protein interaction network.

DLI vs DDI

- Domain: globular structures of long peptides with defined binding or catalytic activities. These structures tend to conserve.

- Linear-Motif: short peptides with specific sequence patterns that bind to domains. They used to be created by few substitutions in short peptides.

There has been an expansion of these interactions in complex organism and has contributed to an increase in network modularity.

Biological modules are protein groups with close functional relationship.

The number of DLI interactions is higher in metazoan species. There's an increment in the number of linear-motifs and peptide-binding domains when the complexity of the species is higher. This fact contributes to the modular architecture of metazoan PPI networks.

DLI vs DDI

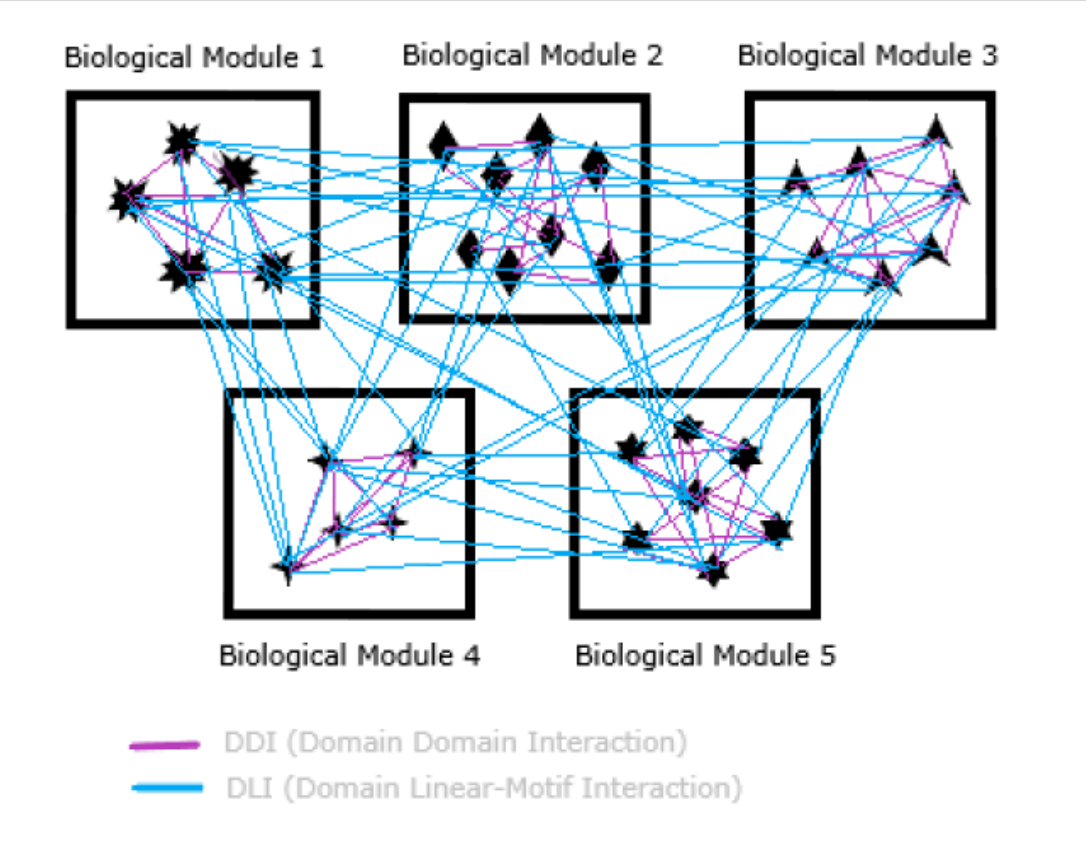


Figure 1. Protein Protein Interaction (PPI) Network.

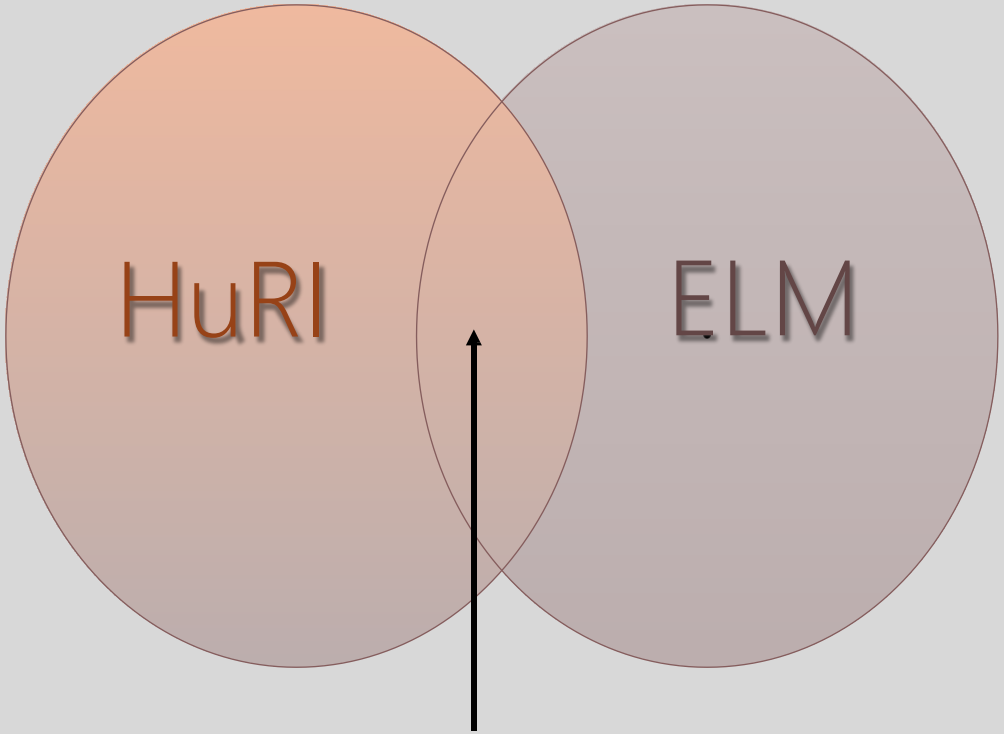
MAPPING WITH ELM, MAIN OBJECTIVE

Determine which protein-protein interactions.

can be explained by the Eukaryotic Linear Motive resource.

These interactions will be Domain-Linear Motif type (DLI).

We would like to know how many Domains found with HMM are likely to have a DLI interaction.



INPUTS

The compared files are :

- HuRI.fa.pfam converted to a .csv file. It is available in HuRI folder in Github.

- ELM interaction domains downloaded from http://elm.eu.org/interactiondomains and with some modifications. This .csv file has the column correspondant to Interaction Domain Name duplicated, in order to proceed with the comparison later. Maybe some code modifications can be done so it is not necessary.

To find the common items in both files I used the Domain Name as I had some problems with the Pfam code, as one file has got decimal points in it and I wasn't able to remove them.

PYTHON CODE

The analysis was performed with **Pandas** because it has great tools to work with large dataframes.

In order to work with **Pandas** I used **Jupyter Notebook**. The conversion of the code file was done with the proper commands in the cmd, as the code file was in **.ipynb** and I wanted the **.py**

Full code is available in Github : https://github.com/lionking0000/YangLabIntern/tree/master/Y2H with the name **huri_elm.py**

```
df_merge = pd.merge(df, df2, how='inner')   #to obtain common values between both files
```

```
df_merge
```

	<hmm name>	<seq id>	<alignment start>	<alignment end>	<envelope start>	<envelope end>	<hmm acc>	<hmm name>.1	<type>	<hmm start>	<hmm end	<hmm length>	sc
0	RRM_1	ENST00000428680.6	130	187	111	187	PF00076.23	RRM_1	Domain	17	70	70	
1	RRM_1	ENST00000428680.6	130	187	111	187	PF00076.23	RRM_1	Domain	17	70	70	

Figure 2. Screenshot of one part of the huri_elm.py code.

OUTPUT

As an output we obtain the file **“common_huri_elm.csv”**. Which contains extended information about each match.

This file shows 12411 matches with ELM, and for each entry contains complete information.

	<hmm name>
0	RRM_1
1	RRM_1
2	RRM_1
3	RRM_1
4	RRM_1
...	...
12407	STT3
12408	STT3
12409	Focal_AT
12410	ALIX_LYPXL_bnd
12411	ALIX_LYPXL_bnd

12412 rows × 1 columns

Figure 3. Output displayed, number of rows are equal to matches.

	A	B	C	D	E	F	G	H	I	J
1		<hmm name>	<seq id>	<alignment start>	<alignment end>	<envelope start>	<envelope end>	<hmm acc>	<hmm name>.1	<type>
2	0	RRM_1	ENST00000428680.6	130	187	111	187	PF00076.23	RRM_1	Domain
3	1	RRM_1	ENST00000428680.6	130	187	111	187	PF00076.23	RRM_1	Domain
4	2	RRM_1	ENST00000373993.5	58	124	58	127	PF00076.23	RRM_1	Domain
5	3	RRM_1	ENST00000373993.5	58	124	58	127	PF00076.23	RRM_1	Domain

Figure 4. Partial output generated and available in common_huri_elm.csv.

RESULTS DISCUSSION

The 12411 obtained matches are the number of all possible interactions combinations.

Some examples:

1) (*Figure 5*). Protein O60551 is likely to have a DLI interaction type with another protein, as the Domain NMT_C presented a match with the ELM file. This interaction is more probable than NMT interacting (as it wasn't found).

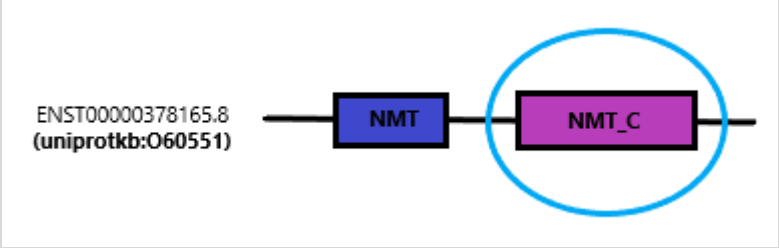


Figure 5. Example nº1.

RESULTS DISCUSSION

2) Protein SUMO-conjugating enzyme UBC9 (uniprotkb:B0QYN7) there were found 2 possible DLI interactions with linear motifs : MOD_SUMO_for_1 and MOD_SUMO_rev_2.

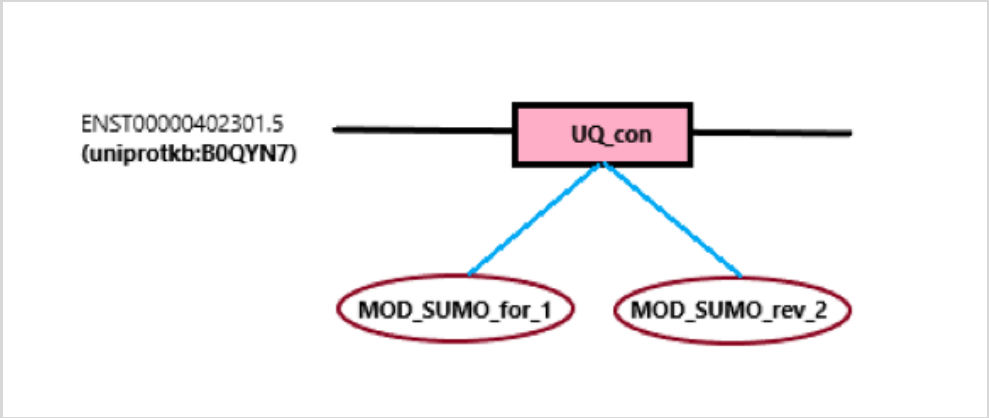


Figure 6. Example nº2.

NEXT TASKS

1. ~~Do the same process with DMI and DDI.~~
2. ~~Obtain a unique file with all the results and a better display.~~
3. ~~Clean code.~~
4. ~~Map with protein file available in Github.~~

REFERENCES AND INTERESTING BIBLIOGRAPHY

1. Kumar, M. *et al.* ELM-the eukaryotic linear motif resource in 2020. *Nucleic Acids Res.* **48**, D296–D306 (2020).

2. Luck, K. *et al.* Putting into practice domain-linear motif interaction predictions for exploration of protein networks. *PLoS One* **6**, e25376 (2011).

3. Davey, N. E. *et al.* Discovery of short linear motif-mediated interactions through phage display of intrinsically disordered regions of the human proteome. *FEBS J.* **284**, 485–498 (2017).

4. Kim, I., Lee, H., Han, S. K. & Kim, S. Linear Motif-Mediated Interactions Have Contributed to the Evolution of Modularity in Complex Protein Interaction Networks. *PLoS Comput. Biol.* **10**, e1003881 (2014).