Mapping protein-protein interactions with ELM

First steps

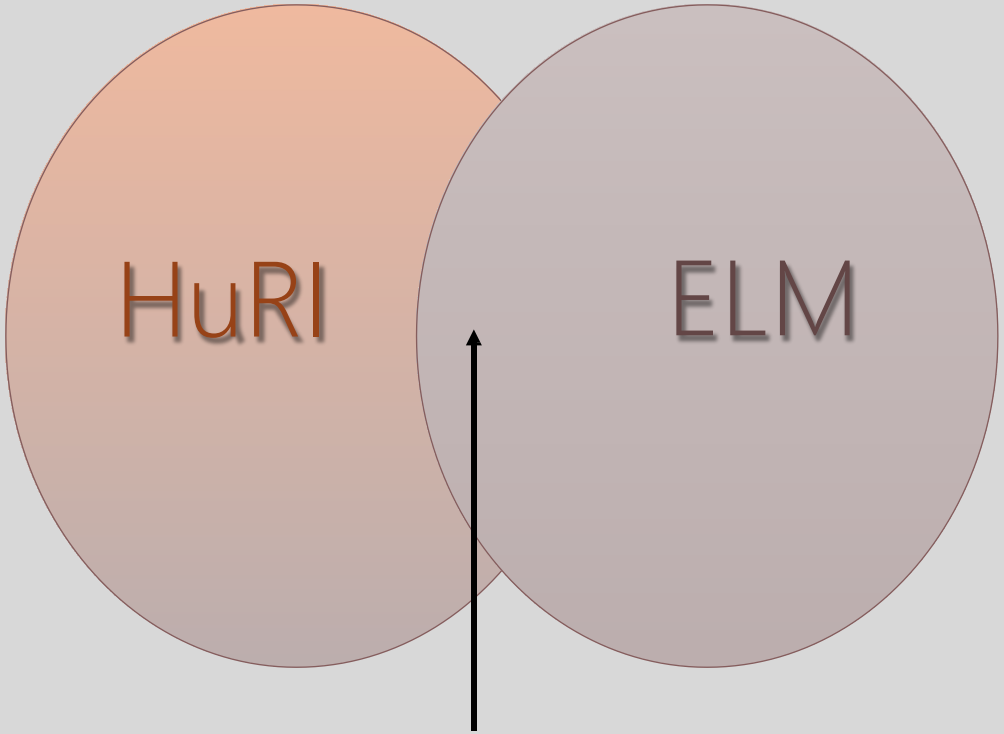
Marina Vallejo Vallés
Computational and Synthetic Biology Group
Group Leader: Dr.Jae-Seong Yang
09/07/2020

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 1. 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. ~~This are the number of domains that are likely to have a DLI interaction.~~
Number of possible combinations

~~Also the result is close to the total value of 15797 HuRI's domain.~~

	<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 2. 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 3. Partial output generated and available in common_huri_elm.csv.

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.