Mapping protein file with PPI

Marina Vallejo Vallés
Computational and Synthetic Biology Group
Group Leader: Dr. Jae-Seong Yang
20/07/2020

Steps

Clean the input files

- In order to have only the necessary information and eliminate duplicated interactions.
- Files: all_interactions1.csv ; all_interactions2.csv ; proteins.csv

Run Python codes

- So we can proceed with the mapping.
- Files: map_proteins1.py ; map_proteins2.py

Obtain output files

- Files: mapping1.csv ; mapping2.csv

Why two “all_interactions.csv” files?

- Notice that in the previous slide we have as input :

  all_interactions1.csv ; all_interactions2.csv

- In the file “proteins.csv”, we have two interacting proteins with their ensembl codes (so they can be identified).
- This interacting proteins are distributed in two columns (A, B). And their respectives ensembl codes ENST, ENSP and ENSG are in C/D/E for protein 1 and in F/G/H for protein 2.
- For the mapping, first we’ll search for ENST similarities for the column C, with the file “all_interactions1.csv” and code “map_proteins1.py”. And we obtain the file “mapping1.csv”.
- Then we do the same process but searching similarities for column F, with the file “all_interactions2.csv” and code “map_proteins2.py”. And we obtain the file “mapping2.csv”.

Why two “all_interactions.csv” files?

- We have to explore this two columns because maybe the ENST we are looking for its just available in one of them.
- The ENST is a unique ID for each human transcript and contains a 11 digit number. In the protein file this ENST is related with protein it encodes.
- With this methodology we’ll obtain a large amount of duplicates as in many cases the ENST is available in both columns and linked to the same interacting protein we found before.

Input files

	A	B	C	D	E	F	G	H	I
1	PROTEIN 1 (uniprotkb)	PROTEIN 2 (uniprotkb)	ENST 1	ENSP 1	ENSG 1	ENST 2	ENSP 2	ENSG 2	
2	X6RM59	Q9UHD9	ENST00000610140.5	ENSP00000476480.1	ENSG00000122643.18	ENST00000338222.6	ENSP00000345195.5	ENSG00000188021.8	
3	X6RM59	Q6ICB0	ENST00000610140.5	ENSP00000476480.1	ENSG00000122643.18	ENST00000263256.6	ENSP00000263256.6	ENSG00000100418.7	
4	X6RLT1	Q9H3P2-1	ENST00000460601.5	ENSP00000436783.2	ENSG00000101158.13	ENST00000411638.6	ENSP00000399165.1	ENSG00000185049.14	
5	X6RLT1	G8JLG2	ENST00000460601.5	ENSP00000436783.2	ENSG00000101158.13	ENST00000376288.2	ENSP00000365465.2	ENSG00000204539.3	

Figure 1. Screenshot of the file "proteins.csv". For each row we have the two interacting proteins and the respectives ensembl id. ENST/ENSP/ENSG 1 are from PROTEIN 1 and ENST/ENSP/ENSG 2 are from PROTEIN 2.

	A	B	C	D	E
1	ENST 1	INTERACTING DOMAIN	DOMAIN	LINEAR MOTIF	INTERACTION TYPE
2	ENST00000428680.6	zf-RING_4	UQ_con		DDI
3	ENST00000428680.6	RRM_1	2OG-FeII_Oxy_2		DDI
4	ENST00000428680.6	RRM_1	Bud13		DDI
5	ENST00000428680.6	RRM_1	CPSF_A		DDI

Figure 2. Screenshot of the file "all_interactions1.csv". Is the output from the mapping with ELM/3DID databases, we can see the domains/ linear motifs interacting for each ENST. Here the ENST has been named ENST 1 so we can search similarities with columna ENST 1 in file "proteins.csv".

	A	B	C	D	E
1	ENST 2	INTERACTING DOMAIN	DOMAIN	LINEAR MOTIF	INTERACTION TYPE
2	ENST00000428680.6	zf-RING_4	UQ_con		DDI

Figure 3. Screenshot of the file "all_interactions2.csv". Is the duplicated file of "all_interactions1.csv" with a small change: Column A now is ENST2 so we can search similarities with columna ENST 2 in file "proteins.csv".

The original file "all_interactions" was obtained in the previous mapping of HuRI with ELM and 3DID databases.

Some information was deleted as there were too many details for each entry. This way we can obtain a precise output with just the information we need.

Our goal is to obtain a file with the two interacting proteins and which interactions are likely to happen.

Input files

- For the "proteins.csv" file some modifications were made. There were a large amount of duplicated interactions.
- After some data treatment this duplicated interactions were deleted.
- For example, the interaction between proteins X6RM59 and Q9UHD9 just appears one time in all the file.
- Notice that proteins are coded with their respectives uniprotkb codes.

Output files

- Once we run the codes we obtain two output files.
- For each pair of interacting proteins we have displayed all the likely interaction types detailed with Domain/Domain or Domain/Linear-Motif.

Next tasks

- Obtain a unique output file.
- Do some modifications to improve the output file.
- Proceed with the Statistical Analysis.