Mapping protein file with PPI

Marina Vallejo Vallés
Computational and Synthetic Biology Group
Group Leader: Dr. Jae-Seong Yang
23/07/2020

Steps

Clean the input files

- In order to have only the necessary information and eliminate duplicated interactions.
- Files: all_interactions.csv; proteins1.csv; proteins2.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?

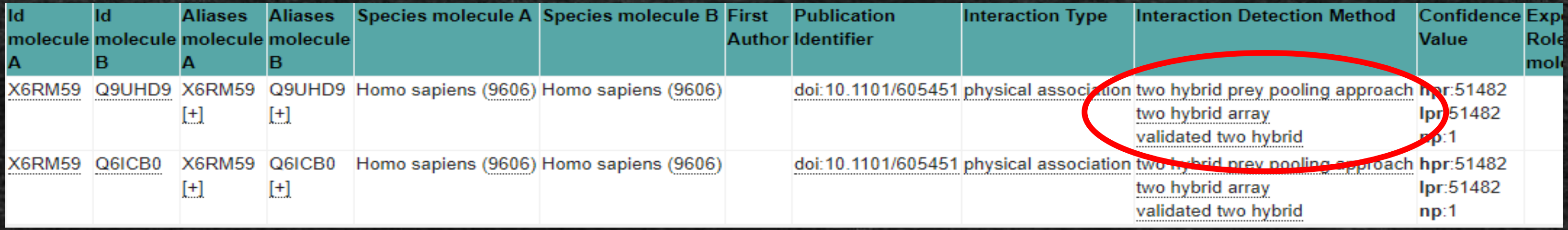
- Notice that in the previous slide we have as input :

  proteins1.csv ; proteins2.csv

- In the files “proteinsX.csv”, we have two interacting proteins with their ensembl codes (so they can be identified).
- For the mapping, first we’ll search for ENST similarities for the column C, with the file “’proteins1.csv” and code “map_proteins1.py”. And we obtain the file “mapping1.csv”.
- Then we do the same process but searching similarities with the file “proteins2.csv” and code “map_proteins2.py”. And we obtain the file “mapping2.csv”.
- The ENST is a unique ID for each human transcript and contains an 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.
- Notice that proteins are coded with their respectives uniprotkb codes.

Input files

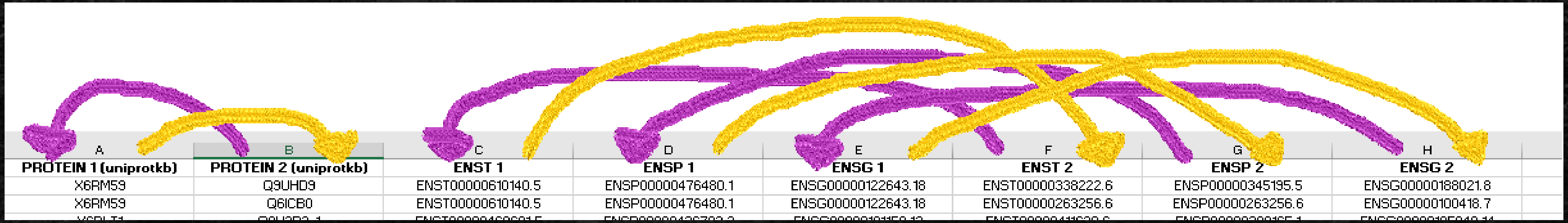
- For the "proteins.csv" file some modifications were made. There were a large amount of duplicated interactions. Some data treatment is done and then these duplicated interactions are deleted.



Id molecule A	Id molecule B	Aliases molecule A	Aliases molecule B	Species molecule A	Species molecule B	First Author	Publication Identifier	Interaction Type	Interaction Detection Method	Confidence Value	Exp Role mol
X6RM59	Q9UHD9	X6RM59 [+]	Q9UHD9 [+]	Homo sapiens (9606)	Homo sapiens (9606)		doi:10.1101/605451	physical association	two hybrid prey pooling approach two hybrid array validated two hybrid	hpr:51482 lpr:51482 np:1	
X6RM59	Q6ICB0	X6RM59 [+]	Q6ICB0 [+]	Homo sapiens (9606)	Homo sapiens (9606)		doi:10.1101/605451	physical association	two hybrid prey pooling approach two hybrid array validated two hybrid	hpr:51482 lpr:51482 np:1	

Figure 1. There are duplicates in the file because the interaction was detected by more than one method.

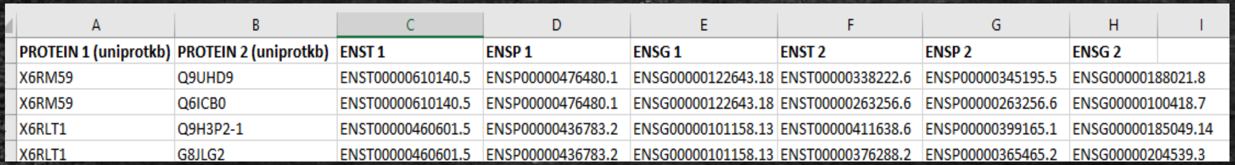
- This interacting proteins are distributed in two columns (A, B). And their respective ensembl codes ENST, ENSP and ENSG are in C/D/E for protein 1 and in F/G/H for protein 2. Maybe an specific protein is only available as protein 2 but we have found interesting information about how it can interact, that's why I fliped columns and created "proteins2.csv", where protein 2 takes the role of protein 1 and viceversa.



A	B	C	D	E	F	G	H
PROTEIN 1 (uniprotkb)	PROTEIN 2 (uniprotkb)	ENST 1	ENSP 1	ENSG 1	ENST 2	ENSP 2	ENSG 2
X6RM59	Q9UHD9	ENST00000610140.5	ENSP00000476480.1	ENSG00000122643.18	ENST00000338222.6	ENSP00000345195.5	ENSG00000188021.8
X6RM59	Q6ICB0	ENST00000610140.5	ENSP00000476480.1	ENSG00000122643.18	ENST00000263256.6	ENSP00000263256.6	ENSG00000100418.7

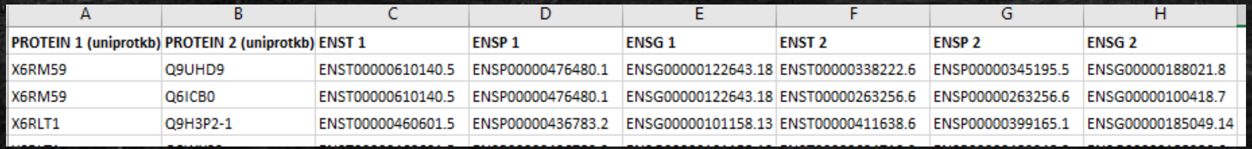
Figure 2. Flip data in file in order to have protein 2 as protein 1 and viceversa.

Input files



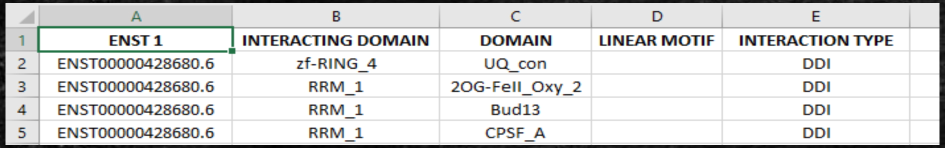
PROTEIN 1 (uniprotkb)	PROTEIN 2 (uniprotkb)	ENST 1	ENSP 1	ENSG 1	ENST 2	ENSP 2	ENSG 2
X6RM59	Q9UHD9	ENST00000610140.5	ENSP00000476480.1	ENSG00000122643.18	ENST00000338222.6	ENSP00000345195.5	ENSG00000188021.8
X6RM59	Q6ICB0	ENST00000610140.5	ENSP00000476480.1	ENSG00000122643.18	ENST00000263256.6	ENSP00000263256.6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

Figure 3. Screenshot of the file "proteins1.csv". For each row we have the two interacting proteins and the respective ensembl id. ENST/ENSP/ENSG 1 are from PROTEIN 1 and ENST/ENSP/ENSG 2 are from PROTEIN 2.



PROTEIN 1 (uniprotkb)	PROTEIN 2 (uniprotkb)	ENST 1	ENSP 1	ENSG 1	ENST 2	ENSP 2	ENSG 2
X6RM59	Q9UHD9	ENST00000610140.5	ENSP00000476480.1	ENSG00000122643.18	ENST00000338222.6	ENSP00000345195.5	ENSG00000188021.8
X6RM59	Q6ICB0	ENST00000610140.5	ENSP00000476480.1	ENSG00000122643.18	ENST00000263256.6	ENSP00000263256.6	ENSG00000100418.7
X6RLT1	Q9H3P2-1	ENST00000460601.5	ENSP00000436783.2	ENSG00000101158.13	ENST00000411638.6	ENSP00000399165.1	ENSG00000185049.14

Figure 4. Screenshot of the file "proteins2.csv". Here protein 1 is protein 2 from the original file and viceversa.



	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 3. Screenshot of the file "all_interactions.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 column ENST 1 in file "proteinsX.csv".

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

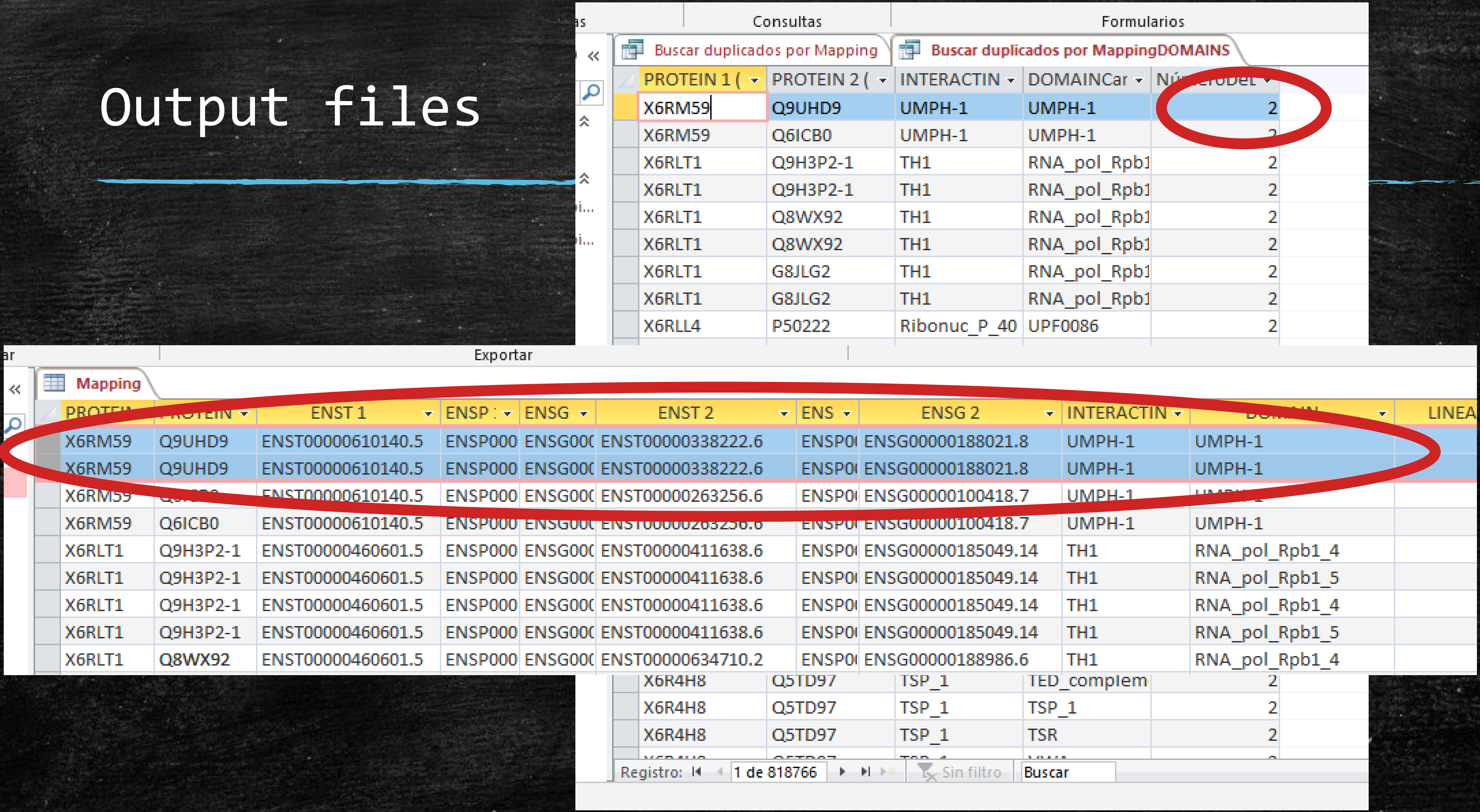
Some information was deleted as there were too many details for each entry. This way we can obtain a precise output with just the information we need.

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

Output files

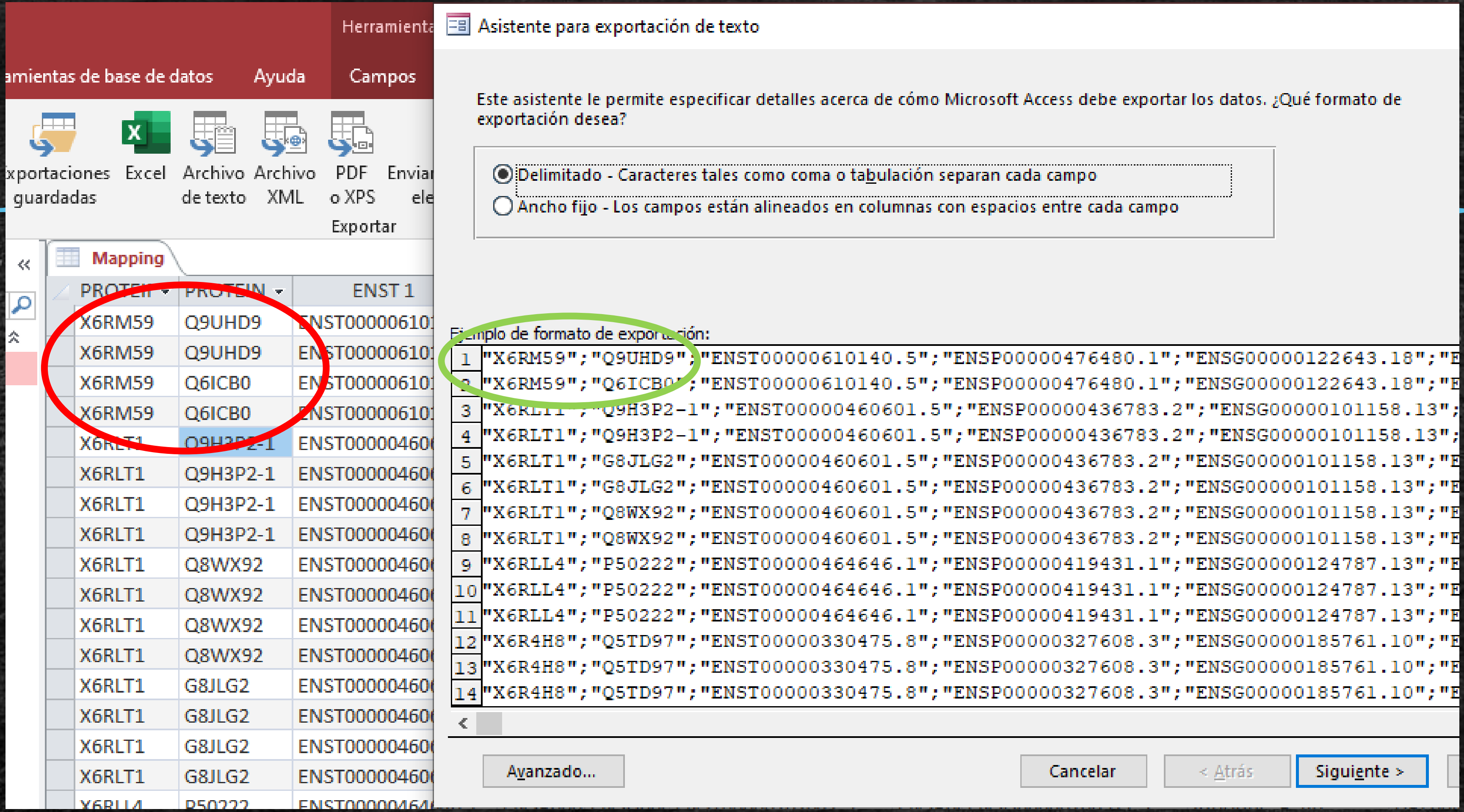
- Once we have run the codes we obtain two output files. We convert these two files in a single one that has all the information.
- For each pair of interacting proteins we have displayed all the likely interaction types detailed with Domain/Domain or Domain/Linear-Motif.

Output files


Consultas
Formularios
Buscar duplicados por Mapping
Buscar duplicados por MappingDOMAINS

PROTEIN 1 (	PROTEIN 2 (	INTERACTIN	DOMAINCar	NúmeroDe
X6RM59	Q9UHD9	UMPH-1	UMPH-1	2
X6RM59	Q6ICB0	UMPH-1	UMPH-1	2
X6RLT1	Q9H3P2-1	TH1	RNA_pol_Rpb1	2
X6RLT1	Q9H3P2-1	TH1	RNA_pol_Rpb1	2
X6RLT1	Q8WX92	TH1	RNA_pol_Rpb1	2
X6RLT1	Q8WX92	TH1	RNA_pol_Rpb1	2
X6RLT1	G8JLG2	TH1	RNA_pol_Rpb1	2
X6RLT1	G8JLG2	TH1	RNA_pol_Rpb1	2
X6RLL4	P50222	Ribonuc_P_40	UPF0086	2
X6R4H8	Q5TD97	TSP_1	TED_complem	2
X6R4H8	Q5TD97	TSP_1	TSP_1	2
X6R4H8	Q5TD97	TSP_1	TSR	2

PROTEIN	PROTEIN	ENST 1	ENSP :	ENSG	ENST 2	ENS	ENSG 2	INTERACTIN	DOMAIN	LINEA
X6RM59	Q9UHD9	ENST00000610140.5	ENSP000	ENSG00(	ENST00000338222.6	ENSP0(	ENSG00000188021.8	UMPH-1	UMPH-1	
X6RM59	Q9UHD9	ENST00000610140.5	ENSP000	ENSG00(	ENST00000338222.6	ENSP0(	ENSG00000188021.8	UMPH-1	UMPH-1	
X6RM59	Q6ICB0	ENST00000610140.5	ENSP000	ENSG00(	ENST00000263256.6	ENSP0(	ENSG00000100418.7	UMPH-1	UMPH-1	
X6RM59	Q6ICB0	ENST00000610140.5	ENSP000	ENSG00(	ENST00000263256.6	ENSP0(	ENSG00000100418.7	UMPH-1	UMPH-1	
X6RLT1	Q9H3P2-1	ENST00000460601.5	ENSP000	ENSG00(	ENST00000411638.6	ENSP0(	ENSG00000185049.14	TH1	RNA_pol_Rpb1_4	
X6RLT1	Q9H3P2-1	ENST00000460601.5	ENSP000	ENSG00(	ENST00000411638.6	ENSP0(	ENSG00000185049.14	TH1	RNA_pol_Rpb1_5	
X6RLT1	Q9H3P2-1	ENST00000460601.5	ENSP000	ENSG00(	ENST00000411638.6	ENSP0(	ENSG00000185049.14	TH1	RNA_pol_Rpb1_4	
X6RLT1	Q9H3P2-1	ENST00000460601.5	ENSP000	ENSG00(	ENST00000411638.6	ENSP0(	ENSG00000185049.14	TH1	RNA_pol_Rpb1_5	
X6RLT1	Q8WX92	ENST00000460601.5	ENSP000	ENSG00(	ENST00000634710.2	ENSP0(	ENSG00000188986.6	TH1	RNA_pol_Rpb1_4	



PROTEI	PROTEIN	ENST 1
X6RM59	Q9UHD9	ENST00000610
X6RM59	Q9UHD9	ENST00000610
X6RM59	Q6ICB0	ENST00000610
X6RM59	Q6ICB0	ENST00000610
X6RLT1	Q9H3P2-1	ENST00000460
X6RLT1	Q9H3P2-1	ENST00000460
X6RLT1	Q9H3P2-1	ENST00000460
X6RLT1	Q9H3P2-1	ENST00000460
X6RLT1	Q8WX92	ENST00000460
X6RLT1	Q8WX92	ENST00000460
X6RLT1	Q8WX92	ENST00000460
X6RLT1	Q8WX92	ENST00000460
X6RLT1	G8JLG2	ENST00000460
X6RLT1	G8JLG2	ENST00000460
X6RLT1	G8JLG2	ENST00000460
X6RLT1	G8JLG2	ENST00000460
X6RLL4	P50222	ENST00000464

Asistente para exportación de texto

Este asistente le permite especificar detalles acerca de cómo Microsoft Access debe exportar los datos. ¿Qué formato de exportación desea?

● Delimitado - Caracteres tales como coma o tabulación separan cada campo

○ Ancho fijo - Los campos están alineados en columnas con espacios entre cada campo

Ejemplo de formato de exportación:

```
"X6RM59";"Q9UHD9";"ENST00000610140.5";"ENSP00000476480.1";"ENSG00000122643.18";"E
"X6RM59";"Q6ICB0";"ENST00000610140.5";"ENSP00000476480.1";"ENSG00000122643.18";"E
"X6RLT1";"Q9H3P2-1";"ENST00000460601.5";"ENSP00000436783.2";"ENSG00000101158.13";
"X6RLT1";"Q9H3P2-1";"ENST00000460601.5";"ENSP00000436783.2";"ENSG00000101158.13";
"X6RLT1";"G8JLG2";"ENST00000460601.5";"ENSP00000436783.2";"ENSG00000101158.13";"E
"X6RLT1";"G8JLG2";"ENST00000460601.5";"ENSP00000436783.2";"ENSG00000101158.13";"E
"X6RLT1";"Q8WX92";"ENST00000460601.5";"ENSP00000436783.2";"ENSG00000101158.13";"E
"X6RLT1";"Q8WX92";"ENST00000460601.5";"ENSP00000436783.2";"ENSG00000101158.13";"E
"X6RLL4";"P50222";"ENST00000464646.1";"ENSP00000419431.1";"ENSG00000124787.13";"E
"X6RLL4";"P50222";"ENST00000464646.1";"ENSP00000419431.1";"ENSG00000124787.13";"E
"X6RLL4";"P50222";"ENST00000464646.1";"ENSP00000419431.1";"ENSG00000124787.13";"E
"X6R4H8";"Q5TD97";"ENST00000330475.8";"ENSP00000327608.3";"ENSG00000185761.10";"E
"X6R4H8";"Q5TD97";"ENST00000330475.8";"ENSP00000327608.3";"ENSG00000185761.10";"E
"X6R4H8";"Q5TD97";"ENST00000330475.8";"ENSP00000327608.3";"ENSG00000185761.10";"E
```

When I was exporting the data from Access to .txt in order to delete the duplicates… it just imported the value once!!! It didn't import the duplicates!!

Next tasks

- ~~Obtain a unique output file.~~
- ~~Eliminate duplicates.~~
- Proceed with the Statistical Analysis.
- PPI network with Cytoscape.