Package appinetwork

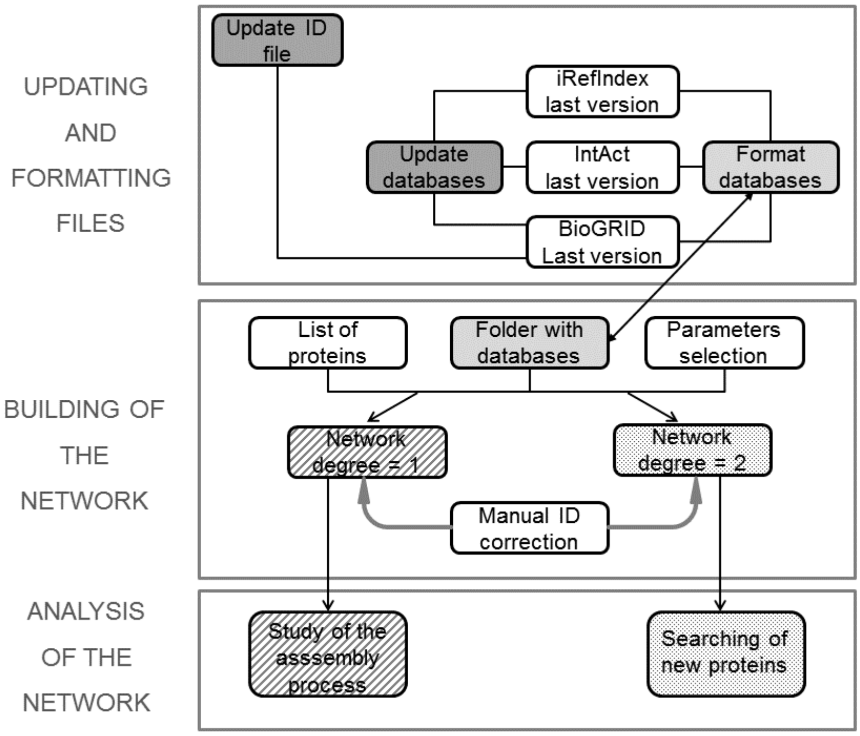
General description

The APPInetwork package is implemented in R, python and C, it is composed of five independent and complementary modules allowing the construction and analysis of protein-protein interaction networks (PPI) (see the workflow below). It was developed for the 8 most studied species but it works for any species. Examples and datasets are provided at each step.

This package allows to format data coming from 4 different websites (databases). The first one is Uniprot (https://www.uniprot.org/). Uniprot is the Universal Protein resource, a central repository of protein data combining information coming from different databases. It a general website, gathering all information around sequences and names. In particular, this website keeps record of all names and IDs for each gene. The user can download data from the UniProt website (following the instruction below) and format this data using the R package. The package will create a ID correspondence file that will be used for following formatting and analysis.

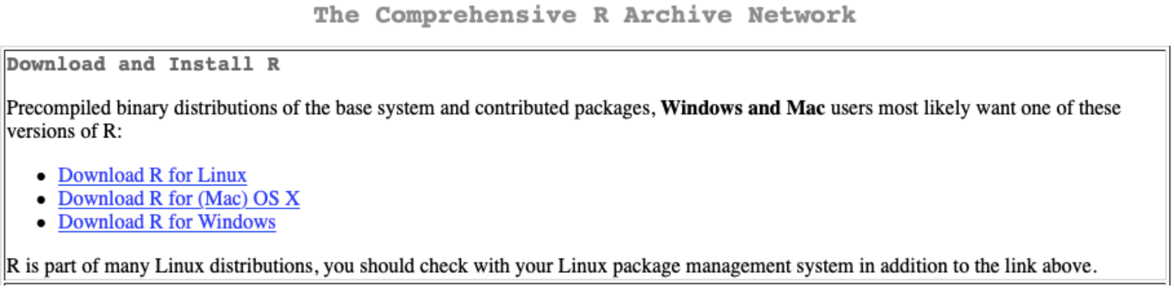
The package enables the formatting of PPI interactions data coming from 3 different databases, Intact, Biogrid and IrefIndex. The 3 databases gather information around PPI interactions. Each database has to be downloaded from their respective website. Each database has its own format and ID numbers. The ID correspondence is useful to find links between databases.

Workflow



Download the package in practice

The appinetwork package is a R package. It can be downloaded from github, directly using the R console. To download and install R, users can go to the following website https://cran.r-project.org and choose the operating system.



The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Once the R software is installed, users have acces to the R console and can first download all required tools by typing the following code in the R console, line by line, as displayed below :

```
install.packages("devtools")
library("devtools")
devtools::install_github("melinagallopin/appinetwork")
```

```
install.packages( pkgs = c("R.methodsS3","rPython","stringr"), dependencies =
TRUE)
```

When downloading all the required tools above, users are required to select the R CRAN repository (by clicking on one city name), and choose to download all required libraries.

If the download fails, users can type the following lines and try again to download

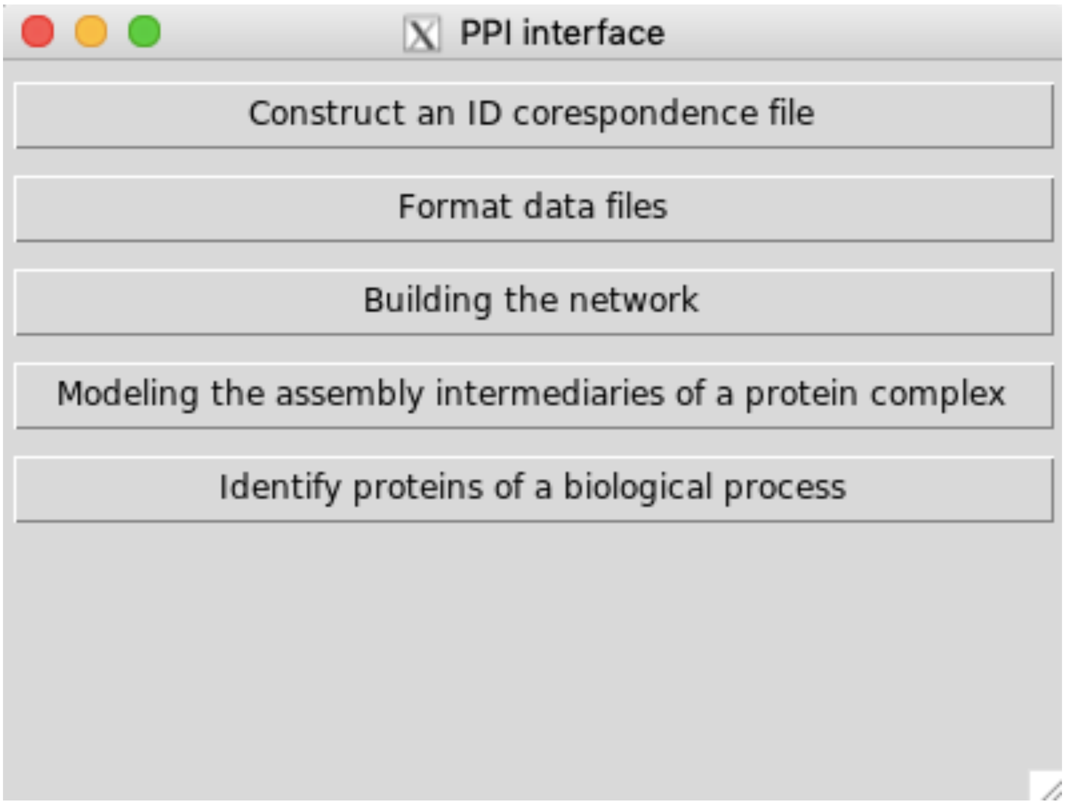
```
install.packages("gWidgets2")
install.packages("digest")
install.packages("gWidgets2tcltk")
install.packages("memoise")
install.packages("tcltk")
install.packages("ape")
```

Once all packages have been downloaded, the appinetwork package can be used by typing the following two lines :

```
library(appinetwork)
interface()
```

Note that only the two previous lines have to be typed every time the user wants to use the appinetwork package.

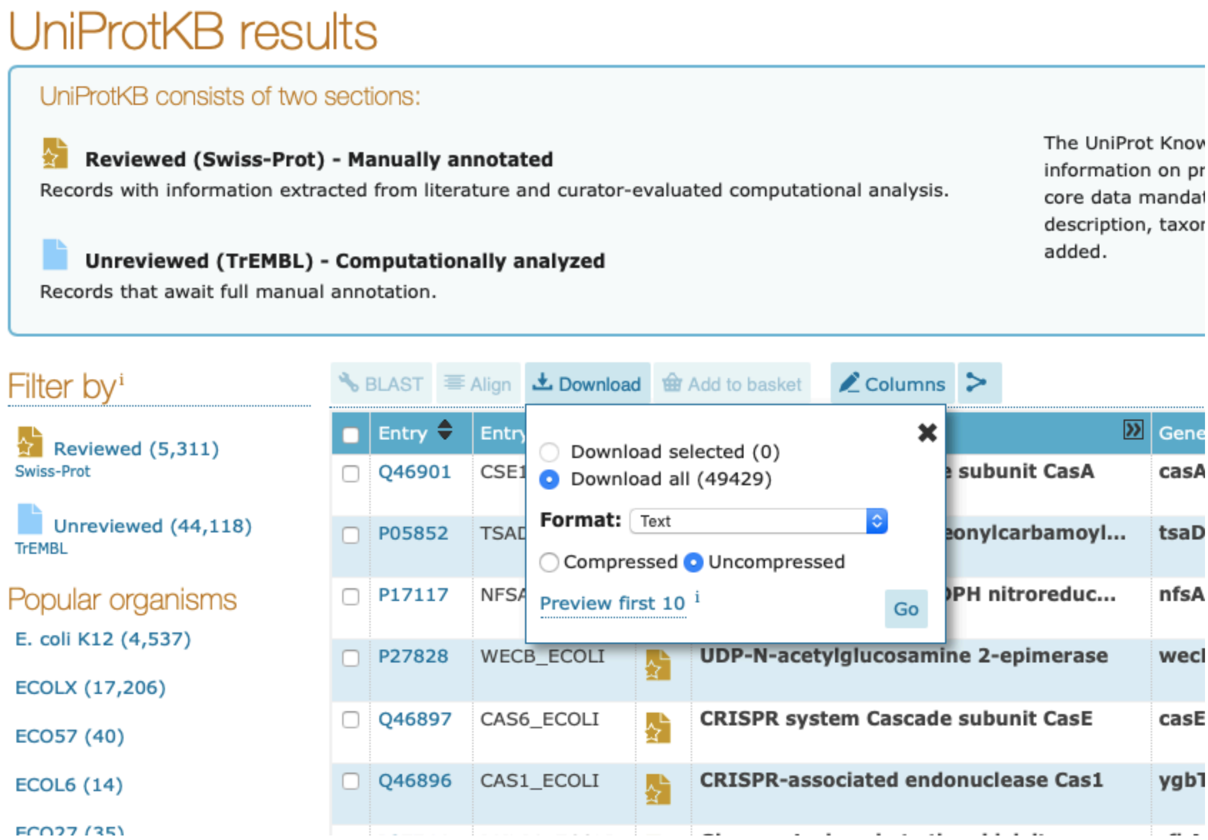
The user sees the following window :



PPI interface
Construct an ID corespondence file
Format data files
Building the network
Modeling the assembly intermediaries of a protein complex
Identify proteins of a biological process

Download the UniProt file

The first step is to constuct the ID correspondence file. This step requires the UniProt file of the organism to study. It can be downloaded from the following website (https://www.uniprot.org/uniprot/), search for the organism. The name of the organism should appear in a left column, under “popular organisms”. The user clicks on the name of the organism (with the number of proteins named in parenthesis). Then, the user clicks on “download”, select the format to “Text” and “uncompressed”.



UniProtKB results

UniProtKB consists of two sections:

Reviewed (Swiss-Prot) - Manually annotated
Records with information extracted from literature and curator-evaluated computational analysis.

Unreviewed (TrEMBL) - Computationally analyzed
Records that await full manual annotation.

The UniProt Know
information on pr
core data manda
description, taxo
added.

Filter by

Reviewed (5,311)
Swiss-Prot

Unreviewed (44,118)
TrEMBL

Popular organisms
E. coli K12 (4,537)
ECOLX (17,206)
ECO57 (40)
ECOL6 (14)
ECO27 (35)

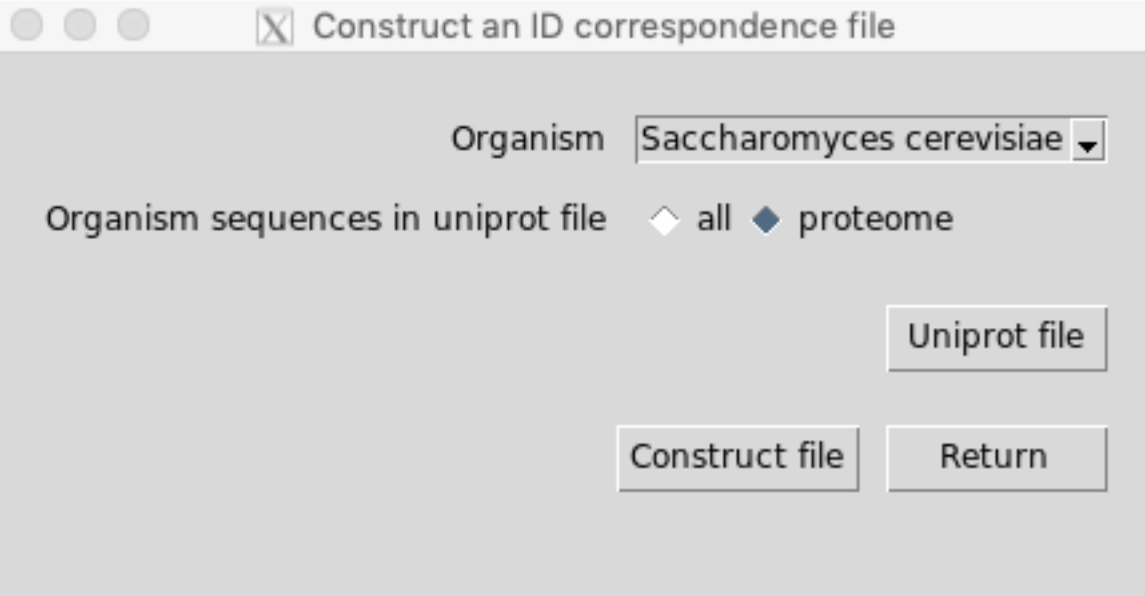
○ Download selected (0)
● Download all (49429)
Format: Text
○ Compressed ● Uncompressed
Preview first 10
Go

☐	Entry	Entry			Gene
☐	Q46901	CSE1		subunit CasA	casA
☐	P05852	TSAD		eonylcarbamoyl...	tsaD
☐	P17117	NFSA		PH nitroreduc...	nfsA
☐	P27828	WECB_ECOLI		UDP-N-acetylglucosamine 2-epimerase	wecB
☐	Q46897	CAS6_ECOLI		CRISPR system Cascade subunit CasE	casE
☐	Q46896	CAS1_ECOLI		CRISPR-associated endonuclease Cas1	ygbT

If the user wants to study specific part of the genome, speciftic chromosome can be selected using the following instructions : go to https://www.uniprot.org/proteomes, search for the organism, and choose the chromosomes to select, then click on « View all proteins » and download the data (using « Text » format and « uncompressed »).

Construct the ID correspondence file

Once the file is downloaded, click on the button « Construct an ID correspondence file» (see window named « appinetwork interface »). The « construct an IC correspondence file » appears. The user selects the organism, click on « proteome », click on « Uniprot file » to select the uniprot file, and click on «Construct file ».



Construct an ID correspondence file

Organism Saccharomyces cerevisiae

Organism sequences in uniprot file ◇ all ◆ proteome

The resulting formatted file is named "Thesaurus_Saccharomyces-cerevisiae_proteome-sequences.txt". This file can be found at the directory printed in the last line appearing in the R console. The name of the directory containing the file can be display in the R console typing :

```
getwd()
```

The "Thesaurus_nameoftheorganism_proteome-sequences.txt" contains 8 columns ID correspondence file including Uniprot-ID, Biogrid-ID, Gene Name, Ref-Seq NP number, Protein-Name, Gene-ID, Old Uniprot-ID and number of Isoformes.

To test this functionality, the user can download the following file "Scerevisiae-Uniprot_for-test.txt" available here https://github.com/melinagallopin/data/blob/master/Scerevisiae-Uniprot_for-test.txt. The resulting file is stored in organism repository in "Thesaurus_Saccharomyces-cerevisiae_proteome-sequences.txt" (path written in the R console) and one example is available here https://github.com/melinagallopin/data/blob/master/Thesaurus_Saccharomyces-cerevisiae_proteome-sequences.txt

Construct the Input list

The user choses the subset of proteins he wants to work on. For example, a list of proteins of a complex or a list of proteins involved in a specific biological process. The input list file should contain 5 columns: the Name of the proteins to study, their UniProtID that will be used to search interactions and UniProtName, the alias and Systematic Name.

Example for one protein of *S. cerevisiae*

Name	UniProt ID	UniProt Name	alias	systematic Name
COR1	P07256	QCR1_YEAST	QCR1	YBL045C

The input list name will be used to give the name of the directories and the files created during the progress of the script. The name should be as short and representative as possible. One example of this file available here https://github.com/melinagallopin/data/blob/master/input_list.txt.

Choosing the PPI database

The user can choose the PPI database he wants to use (personal database) or public databases. The package enables the automatic formatting of the three public databases (iRefIndex, IntAct or Biogrid). If the user wants to user other database, he needs to format the database manually. iRefIndex provides protein-protein interactions available in primary databases of the 8 most studied organisms, but the update is not done often. For those reasons, the package offer the possibility to update and format the BioGRID and IntAct bases. BioGRID is monthly updated and contains physical and genetic interactions. IntAct is a PPI interaction database and provides confidence scores about the interaction.

Formatting the databases

The formatted databases should have the following 15 columns format.

The files should contain UniProt identifiants of each protein (named *uid* and *alias*), the name of the method that enables the identification of the interaction, the author of the publication, the PubMed ID, the name of the taxon that should be identical in two columns (otherwise the interaction is discarded), the interaction type (physical or genetic), the name of the databases were the data were stored with the associated confidence's score, the number of participants that should be equal to two, finally the GeneName corresponding to each protein.

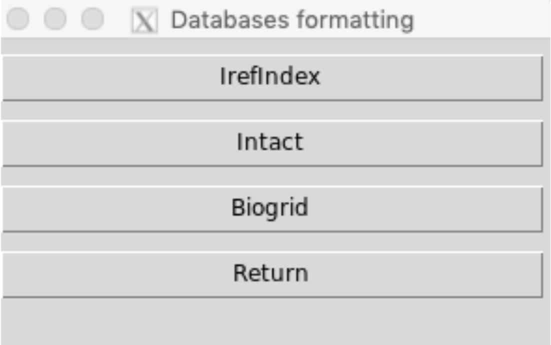
Format of the Databases: 15 columns; n rows
Columns name
Column number: 1 (uidA)
Column number: 2 (uidB)
Column number: 3 (aliasA)
Column number: 4 (aliasB)
Column number: 5 (Method)
Column number: 6 (pmids)
Column number: 7 (author)
Column number: 8 (taxa)
Column number: 9 (taxb)
Column number: 10 (interactionType)
Column number: 11 (sourcedb)
Column number: 12 (confidence)
Column number: 13 (numParticipants)
Column number: 14 (GeneNameA)
Column number: 15 (GeneNameB)

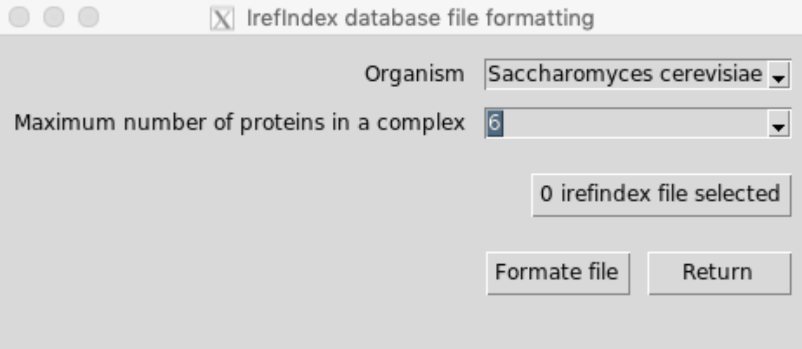
Formatting the iRefIndex database

The iRefIndex database (http://irefindex.org/) can be downloaded at the following link http://irefindex.org/wiki/index.php?title=README_MITAB2.6_for_iRefIndex

The organism ID are : Caenorhabditis elegans : 6239, Drosophila melanogaster : 7227, Escherichia coli : 562, Homo sapiens : 9606, Mus musculus : 10090, Rattus norvegicus : 10116, Saccharomyces cerevisiae : 559292.

Complexes are included in the iRefIndex file, it is possible to list the interactions between all its subunits, but to avoid too much noise in the data the user must set a threshold which depends of the size of the complex to study (example threshold 6 for a complex made of 6 subunits). The file is split in primary bases and interactions involving organism other than the suited organism are eliminated (taxon Id should be identical).





Organism: Saccharomyces cerevisiae

Maximum number of proteins in a complex: 6

The user choses the irefindex file and click on formate file.

The formatting program splits the iRefIndex secondary database into the primary databases used to construct the secondary databases (see iRefIndex website for more information).

To perfom the test on small files, the user can download the following file “Scerevisiae_IrefIndex_for-tests.txt” avalaible here: https://github.com/melinagallopin/data/blob/master/Scerevisiae_IrefIndex_for-tests.txt

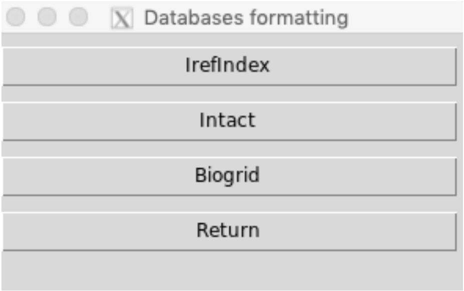
The result files are stored in the folder named “Databases” (the path to the corresponding directory is written by the program in the R console). One example of resulting file from the formatting step is available here https://github.com/melinagallopin/data/blob/master/Saccharomyces-cerevisiae_irefindex_bind_threshold_6.txt .

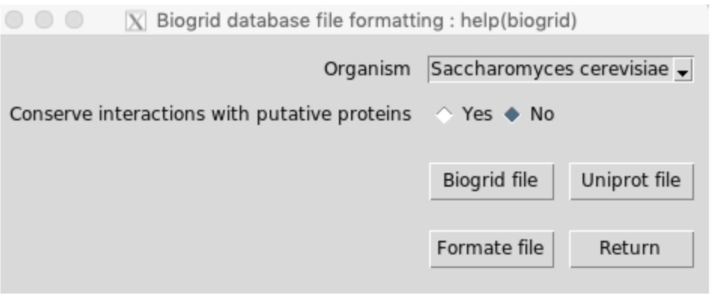
Formatting the BioGRID database

The user goes to BioGRID (https://thebiogrid.org/), then click on “Latest Downloads”, then click on the folder “Current-Release” and and select “BIOGRID-ORGANISM-*****-

tab2.zip”. The BIOGRID repository numbers (*****) change every month. The user uncompress the zip file and has access to the database containing one tab2.txt file for each organism. The user chooses the .tab2.txt file corresponding to the organism of his choice.

The user open the interface (typing “interface()” in the R console, if necessary), click on “format data files”, click on Biogrid and choose the organism, choose to conserve or not interactions with putative proteins, then select the Biogrid database corresponding to the organism (.tab2.txt file) and select the Uniprot file downloaded previously (see **Download the UniProt file** section**).**





Organism Saccharomyces cerevisiae

Conserve interactions with putative proteins ○ Yes ● No

To perfom the test on small files, the user can download the following file

“Scerevisiae_BIOGRID_for-tests.txt” available here

https://github.com/melinagallopin/data/blob/master/Scerevisiae-BIOGRID_for-test.txt,

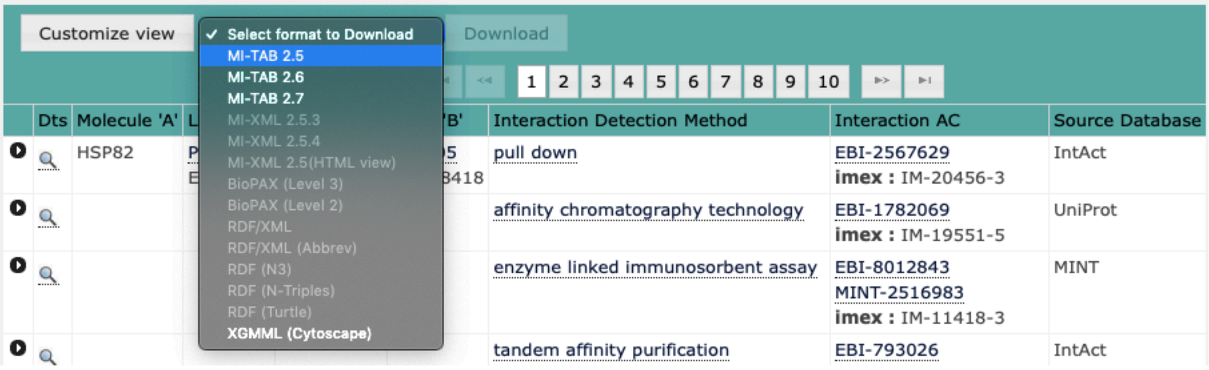
Download the Uniprot “Scerevisiae-Uniprot_for-test.txt” file

https://github.com/melinagallopin/data/blob/master/Scerevisiae-Uniprot_for-test.txt

The resulting file is “Saccharomyces-cerevisiae _biogrid.txt“ available here https://github.com/melinagallopin/data/blob/master/Saccharomyces-cerevisiae%20_biogrid.txt .

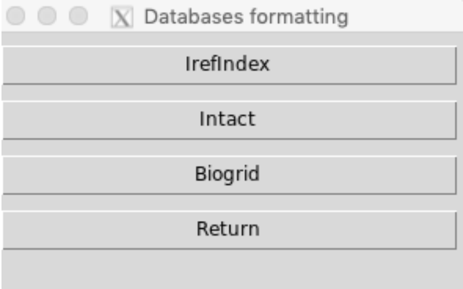
Formatting the intact database

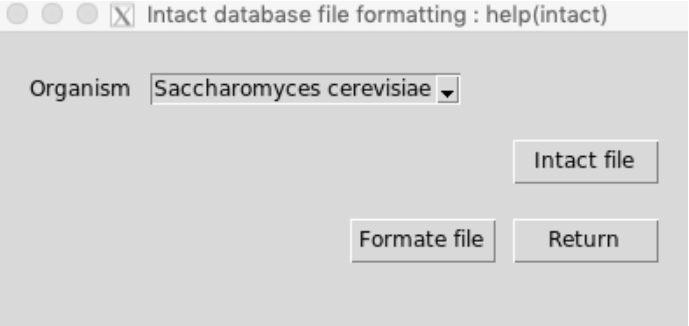
To download the IntAct database to format, the user goes to the website IntAct (https://www.ebi.ac.uk/intact/), search for the organism (example, “saccharomyces cerevisiae”) and select “format to download” and choose MI-TAB 2.5 file. The result is a formatted base named: saccharomy.txt



Dts	Molecule 'A'	L	'B'	Interaction Detection Method	Interaction AC	Source Database
	HSP82	P E	5 3418	pull down	EBI-2567629 **imex :** IM-20456-3	IntAct
				affinity chromatography technology	EBI-1782069 **imex :** IM-19551-5	UniProt
				enzyme linked immunosorbent assay	EBI-8012843 MINT-2516983 **imex :** IM-11418-3	MINT
				tandem affinity purification	EBI-793026	IntAct

The user open the interface (typing “interface()” in the R console, if necessary), click on “format data files”, click on IntAct and choose the organism, then select the IntAct database.





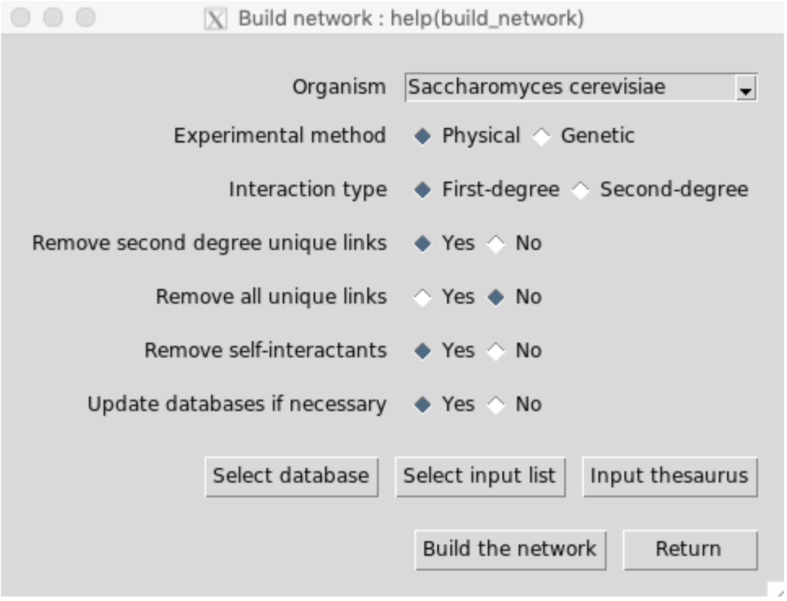
Organism: Saccharomyces cerevisiae

To perfom the test on small files, the user can download the following file “Scerevisiae_intact_for-tests.txt” available here https://github.com/melinagallopin/data/blob/master/Scerevisiae-intact_for-test.txt

The resulting file is “Saccharomyces-cerevisiae_intact.txt“ available here https://github.com/melinagallopin/data/blob/master/Saccharomyces-cerevisiae_intact.txt

Build a network

To build a network starting from a list of proteins, the user selects the “Input list” (see section **Construct the Input list** for details). The user needs to select options that depends on the kind of analysis the user needs to do.



Organism: Saccharomyces cerevisiae

Experimental method ◆ Physical ◇ Genetic

Interaction type ◆ First-degree ◇ Second-degree

Remove second degree unique links ◆ Yes ◇ No

Remove all unique links ◇ Yes ◆ No

Remove self-interactants ◆ Yes ◇ No

Update databases if necessary ◆ Yes ◇ No

Options

To build the network the user can choose some options. The script allows:

- to look for interactions coming from physical or genetic experiments.

- to build network of degree one or two [1].

  To study assembly process a degree 1 network should be used.

  For other studies network of degree 2 can be used [2].

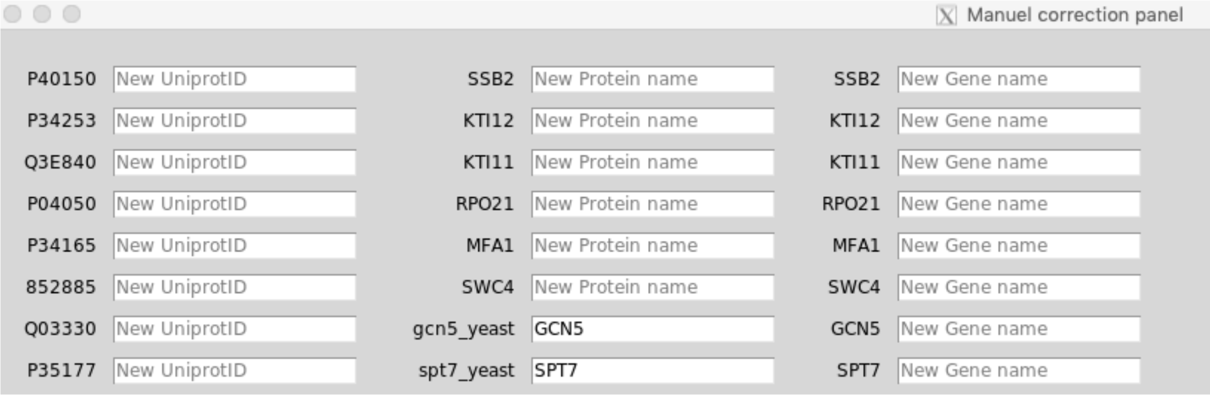
- to remove unique links and/or self-interacts.

To study assembly process keep the unique links and remove self-interactions.

To look for interactions that were not mentioned previously, keep all the unique links.

- the user can update the databases to use them later. The IDs can change from time to time, the correction of the database is time consuming so it can be useful to keep an updated version.

Once all options are selected, the user select the database, the input list and the thesaurus file (ID correspondence file) and click on build the network. Sometimes, there are discrepancies between gene ID names. The user is allowed to correct manually the name using the “manual correction panel”.



Manuel correction panel

P40150	New UniprotID	SSB2	New Protein name	SSB2	New Gene name
P34253	New UniprotID	KTI12	New Protein name	KTI12	New Gene name
Q3E840	New UniprotID	KTI11	New Protein name	KTI11	New Gene name
P04050	New UniprotID	RPO21	New Protein name	RPO21	New Gene name
P34165	New UniprotID	MFA1	New Protein name	MFA1	New Gene name
852885	New UniprotID	SWC4	New Protein name	SWC4	New Gene name
Q03330	New UniprotID	gcn5_yeast	GCN5	GCN5	New Gene name
P35177	New UniprotID	spt7_yeast	SPT7	SPT7	New Gene name

Once the correction is done, click on “correct network”.

To perfom the test on small files, the user can download the following databases “Saccharomyces-cerevisiae%20_biogrid.txt” and “Saccharomyces-cerevisiae_intact.txt”.

https://github.com/melinagallopin/data/blob/master/Saccharomyces-cerevisiae%20_biogrid.txt

https://github.com/melinagallopin/data/blob/master/Saccharomyces-cerevisiae_intact.txt

The user uploads the input list

https://github.com/melinagallopin/data/blob/master/input_list.txt.

The user uploads the ID correspondence file

https://github.com/melinagallopin/data/blob/master/Thesaurus_Saccharomyces-cerevisiae_proteome-sequences.txt

Then, click on build network. Eventually, manual correction of the names is required.

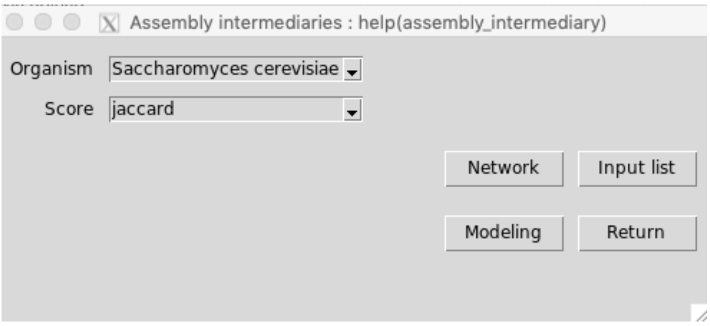
The two files we obtain are available here

https://github.com/melinagallopin/data/blob/master/physical_Network_Saccharomyces-cerevisiae_first-degree_interactions.txt

https://github.com/melinagallopin/data/blob/master/physical_Network_Saccharomyces-cerevisiae_second-degree_interactions.txt

Study of assembly process

This analysis is performed on the "first-degree interactions" network. To study an assembly process, 6 computing methods are proposed (modified Jaccard, dice, chi2, lidell, ms, zscore) [2]. Results are .text files and .jpeg image.



Assembly intermediaries : help(assembly_intermediary)

Organism: Saccharomyces cerevisiae

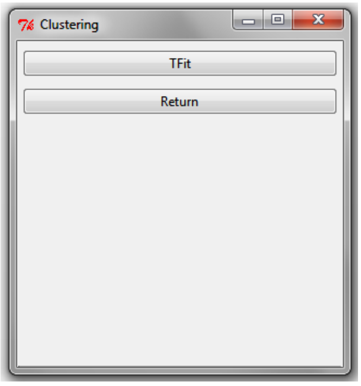
Score: jaccard

To perfom the test on small files, the user can download the following network https://github.com/melinagallopin/data/blob/master/physical_Network_Saccharomyces-cerevisiae_first-degree_interactions.txt and the following Input list https://github.com/melinagallopin/data/blob/master/input_list.txt.

The analysis can be found in the directory named “Analysis” and “subdirectory Assembly_intermediaries”.

Clustering the second-degree network

To visualize and analyze the clusters one method TFit is implemented [3].



To perfom the test on small files, the user can download the following network https://github.com/melinagallopin/data/blob/master/physical_Network_Saccharomyces-cerevisiae_second-degree_interactions.txt

The analysis can be found in the directory named “Analysis” and “TFit”.

[1] Glatigny A *et al.* (2011) An *in silico* approach combined with in vivo experiments enables the identification of a new protein whose overexpression can compensate for specific respiratory defects in *Saccharomyces cerevisiae*. *BMC Syst Biol.* 5, 173-185.

[2] Glatigny A *et al*. (2017). Development of an *in silico* method for the identification of subcomplexes involved in the biogenesis of multiprotein complexes in *Saccharomyces cerevisiae*. *BMC Syst Biol.* 11, 67-79.

[3] Gambette, P , and Guénoche, A. (2011) Bootstrap clustering for graph partitioning. *RAIRO - Operations Research - Recherche Opérationnelle* 45, 339-352.