Data Mining tools for Biomedical Discovery: a framework application in a Biotech Pharma Industry.

A case report of a working group

Gualtieri F., Colace F., Garofalo P., Rovati LC.

Rottapharm Biotech, Monza (Italy)

EAHIL 2016

Biblioteca Virtual
del Sistema Sanitario Público de Andalucía

Knowledge, Research, Innovation ... eHealth

15th EAHIL 2016 Conference 11-13 june, Seville, Spain

Data Mining – Introduction and Aims

In the era of *big data*, Text Mining (TM) represents an essential tool to automate the process of the information retrieval, extraction, interpretation and analysis.

This study provides a practical example of using the TM tool in drug discovery and hypothesis generation.

We compared **Knime** and **R**, two web based open source and free tools, as useful methods in improving the workflow of information search in the Rheumatology field.

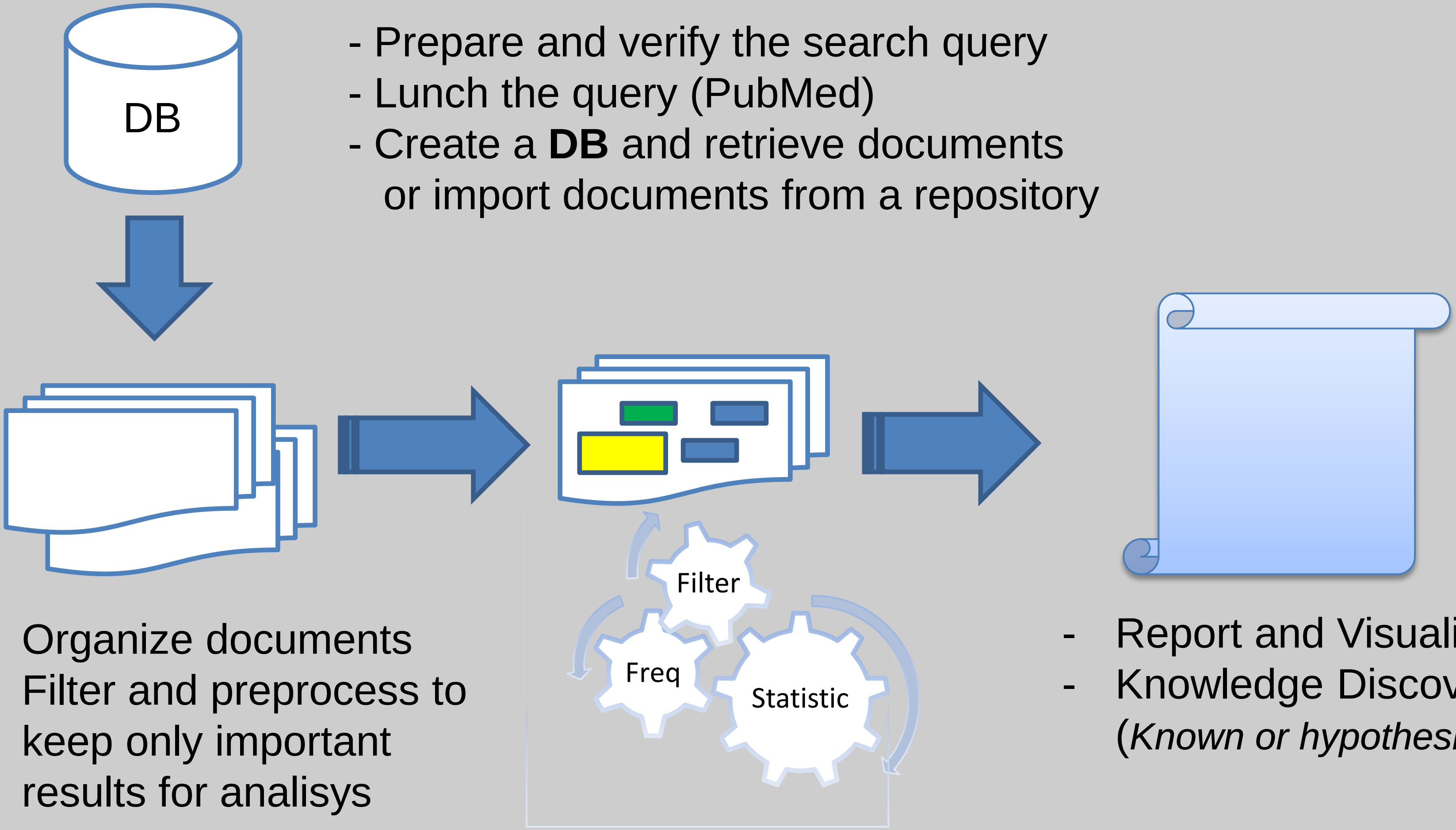
Data Mining - Query

We performed the following search query:

Osteoarthritis, Knee [MeSH*] AND *Humans* [MeSH] AND *Pain

in both systems, to search for new valuable target(s) in Knee OA.

General Text Mining Plan and Workflow



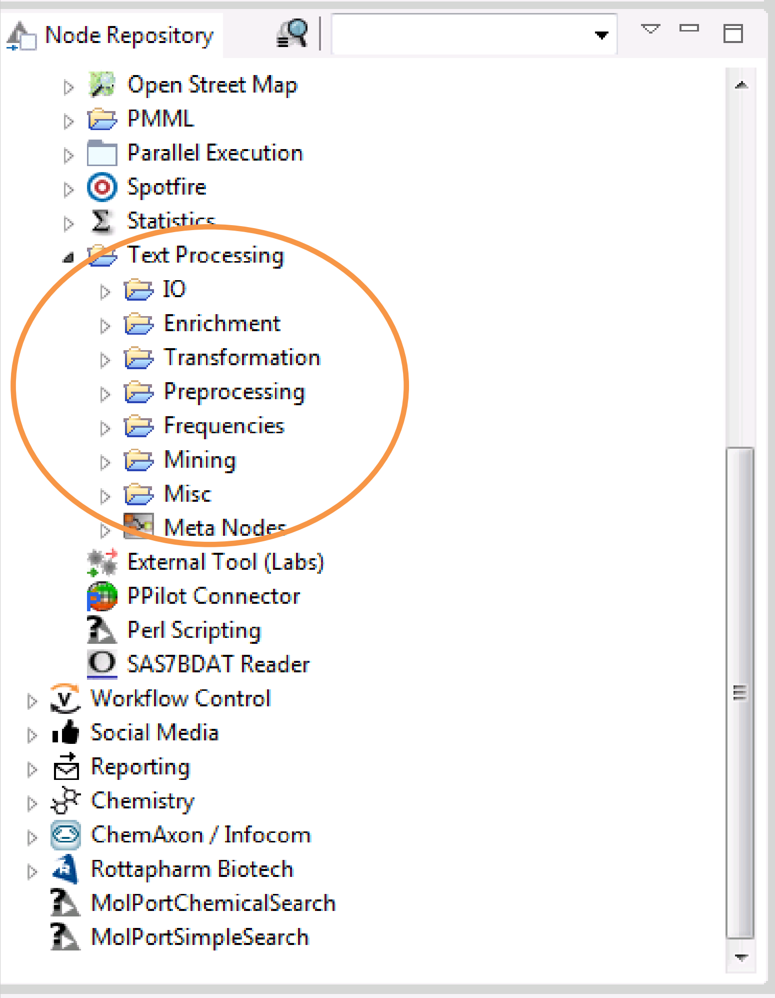
- Prepare and verify the search query
- Lunch the query (PubMed)
- Create a **DB** and retrieve documents or import documents from a repository

- Organize documents
- Filter and preprocess to keep only important results for analisys

- Tag terms(Abner), filter,
- analize for occurence and classify documents or extract relevant terms

- Report and Visualization
- Knowledge Discovery (*Known or hypothesis*)

Knime plugin

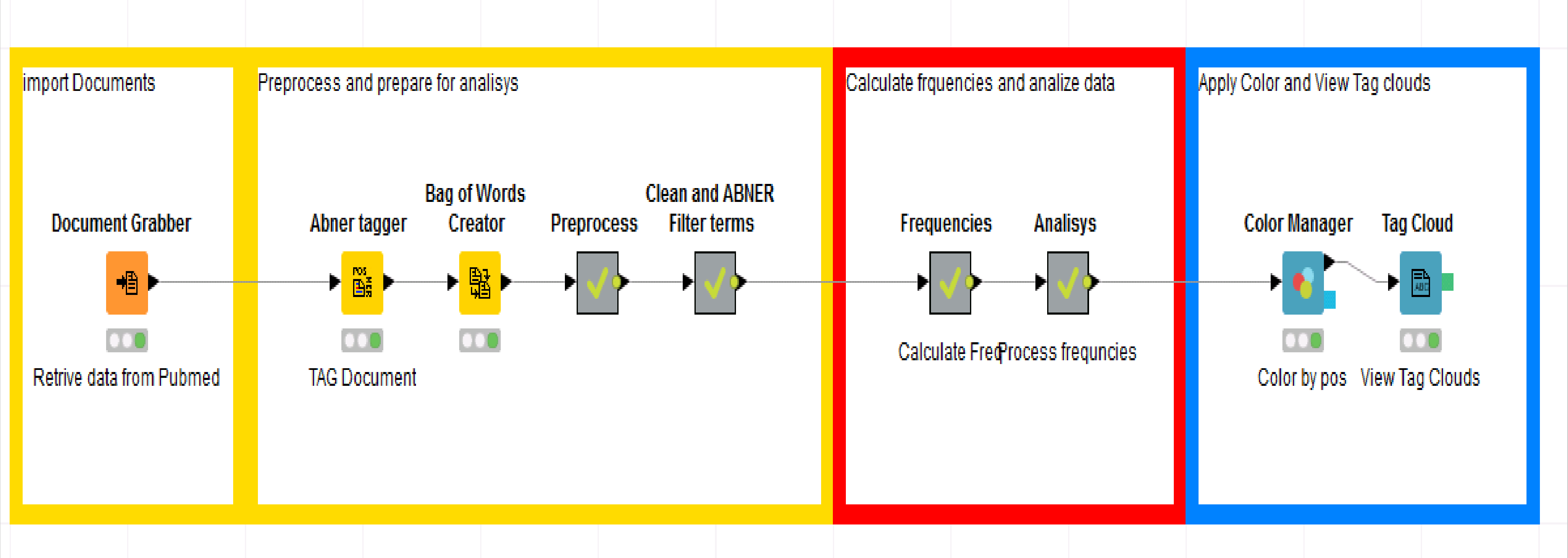


Node Repository

- Open Street Map
- PMML
- Parallel Execution
- Spotfire
- Statistics
- Text Processing
  - IO
  - Enrichment
  - Transformation
  - Preprocessing
  - Frequencies
  - Mining
  - Misc
  - Meta Nodes
- External Tool (Labs)
- PPilot Connector
- Perl Scripting
- SAS7BDAT Reader
- Workflow Control
- Social Media
- Reporting
- Chemistry
- ChemAxon / Infocom
- Rottapharm Biotech
- MolPortChemicalSearch
- MolPortSimpleSearch

- **IO** = Several Nodes are available for Query/Read documents
- **Enrichment** = assign Tags
- **Preprocess** = clean the «bag of words»
- **Frequencies** = Statistic analysis of frequency or co-occurence of Terms
- **Mining** = extract important terms to report

Case 1: Knime Abner tagger node



import Documents

Preprocess and prepare for analisys

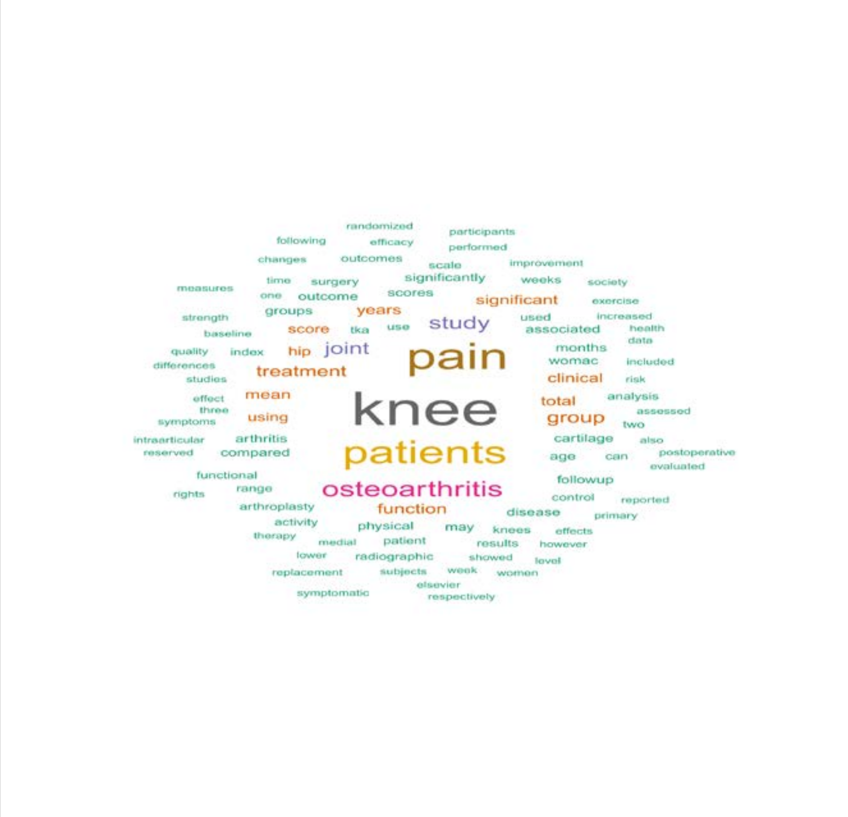
Calculate frquencies and analize data

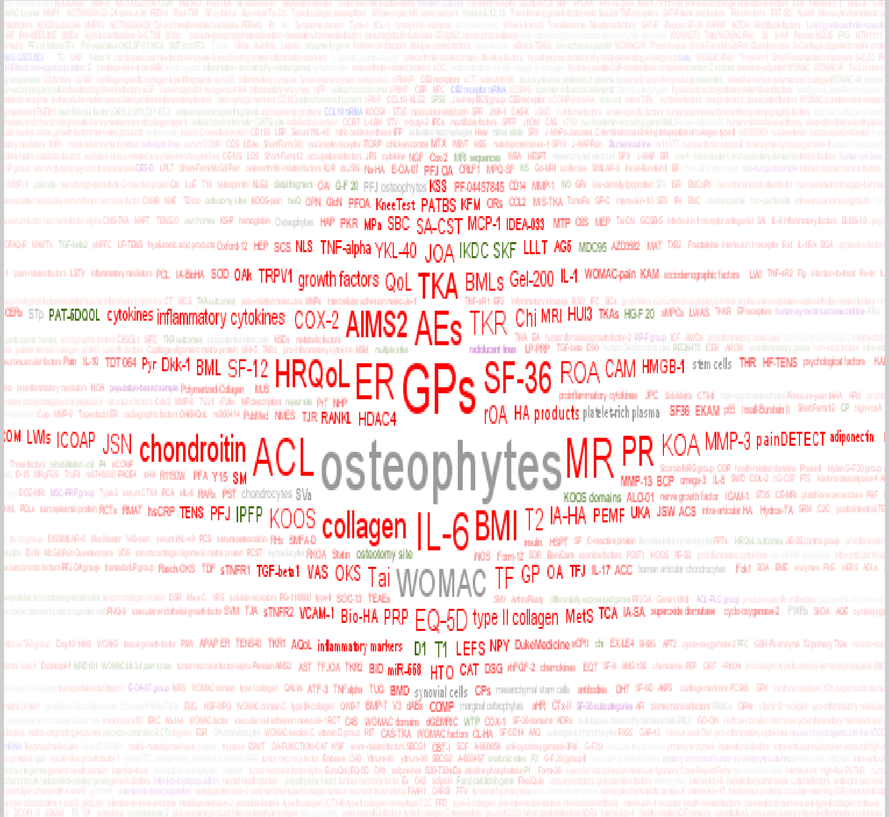
Apply Color and View Tag clouds

Case 1: Word Cloud Visualization

R: Word-cloud visualization

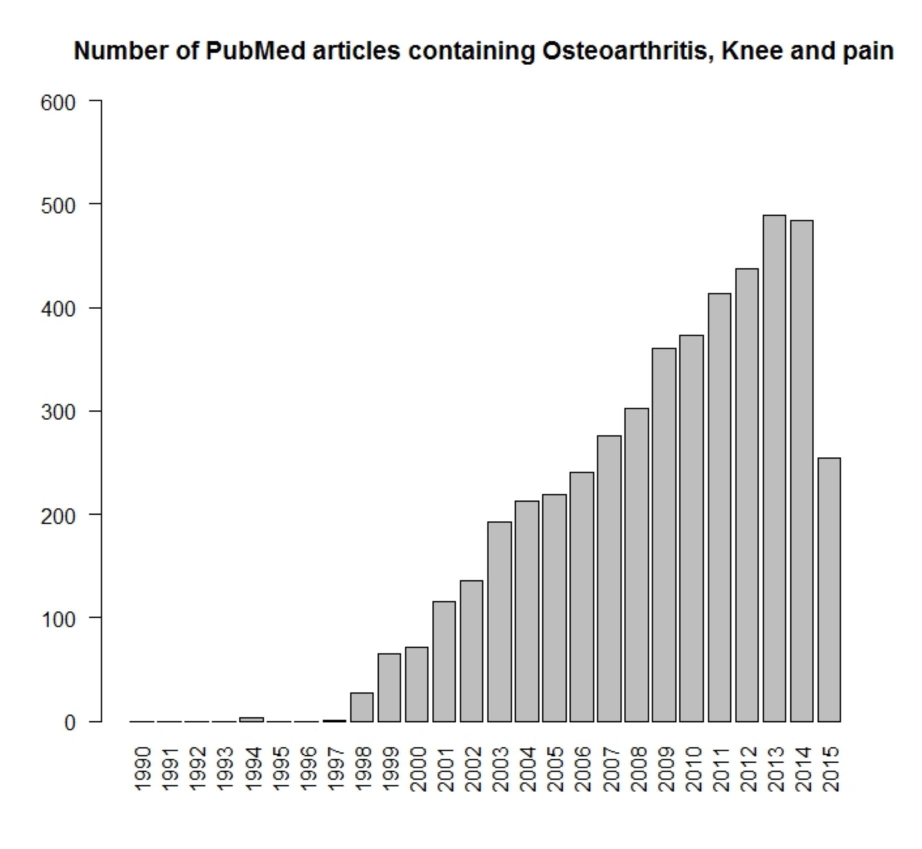
Knime: Word-cloud visualization (abner tagger)



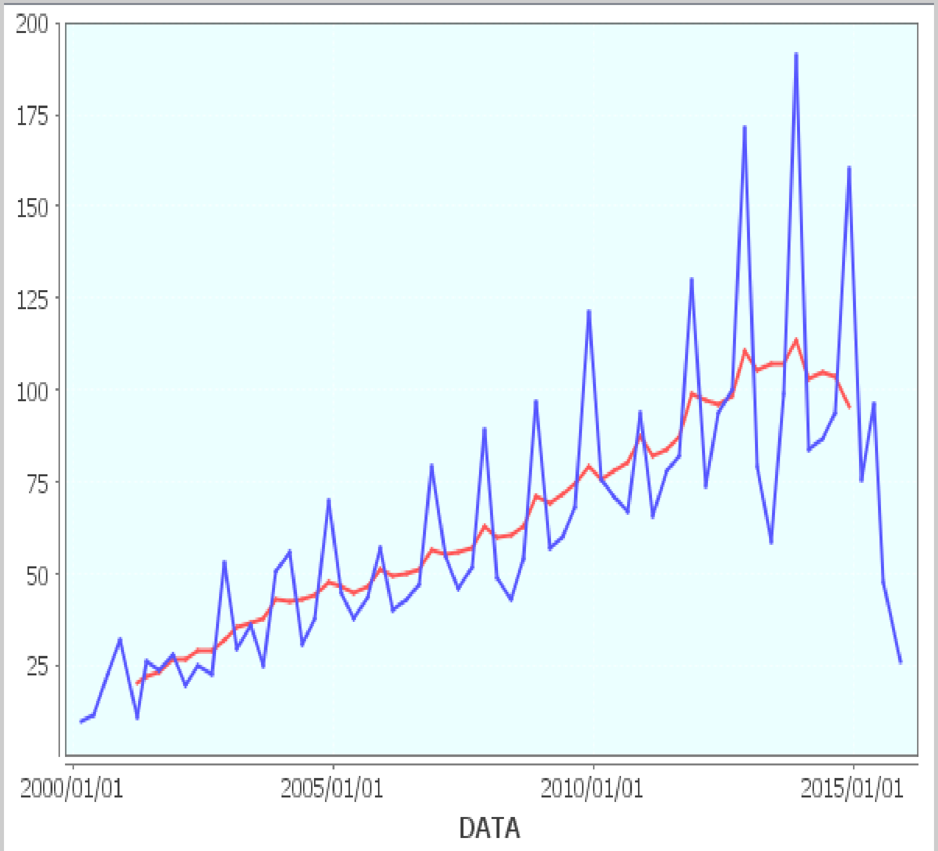


Case 2(a): temporal tendency in query

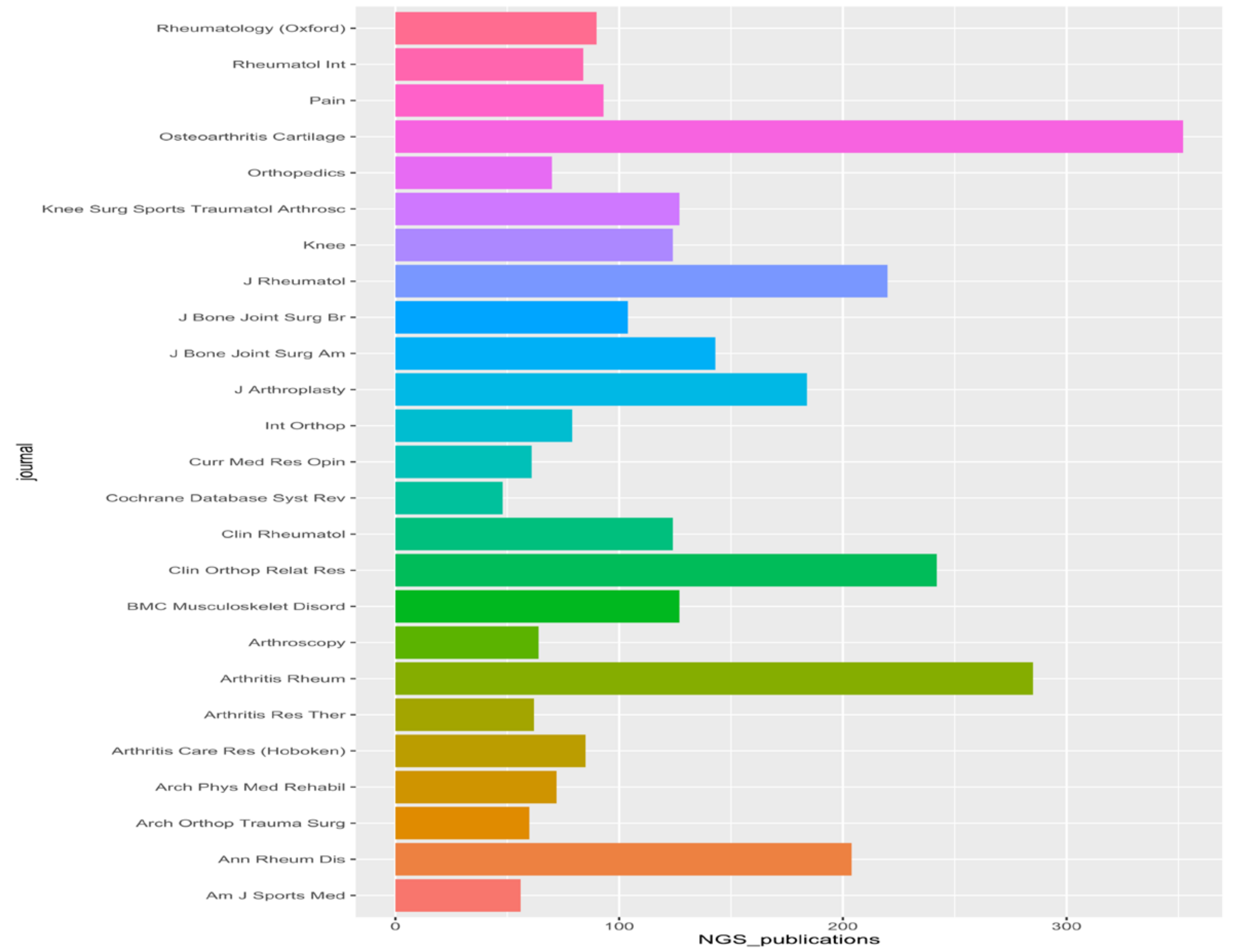
R:



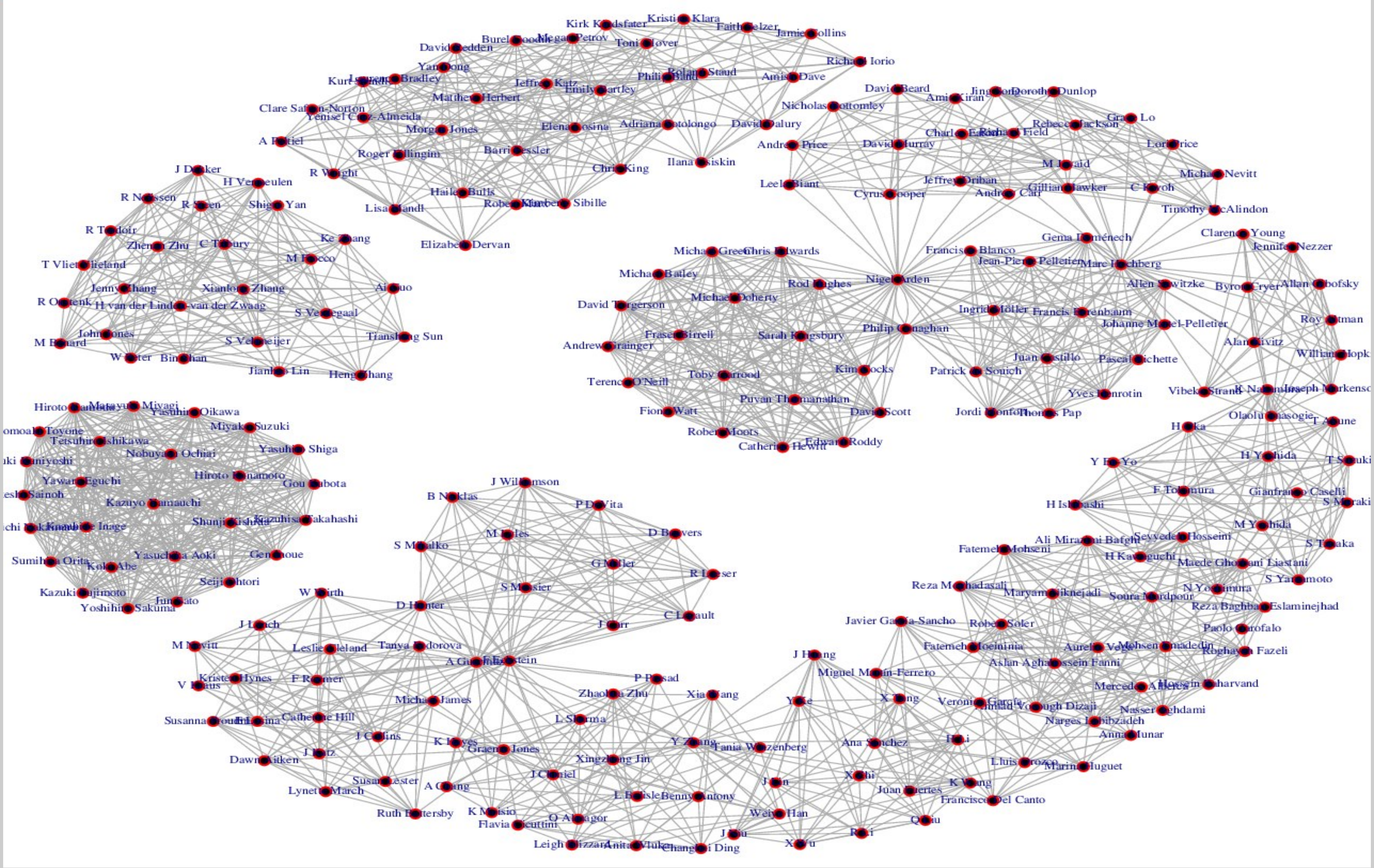
Knime



Case 2(a1): R - number of publications per journal

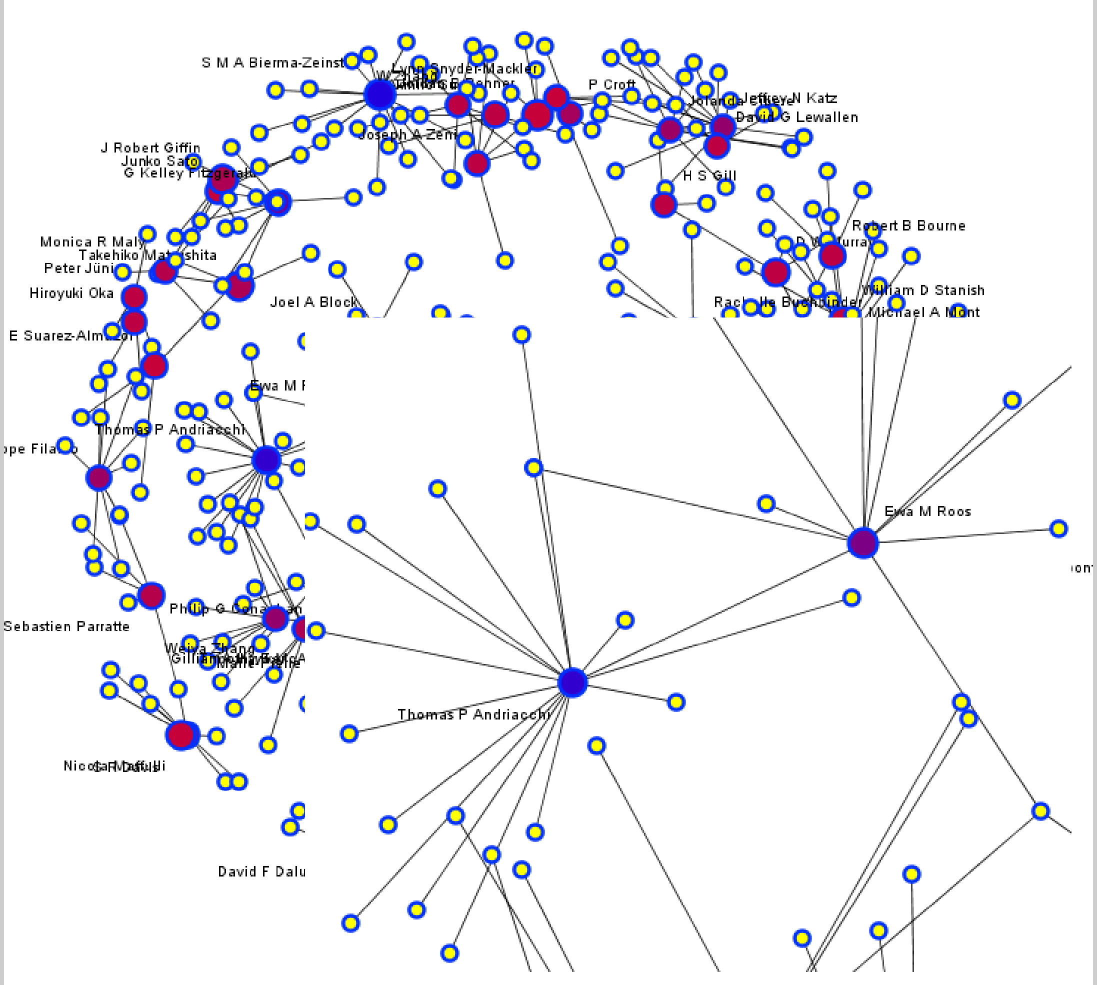


Case 2(b): R - Author network



Case 2(b): Author network

Authors' network related to No. of papers on Knee OA



Legend Color

● = High n° publications

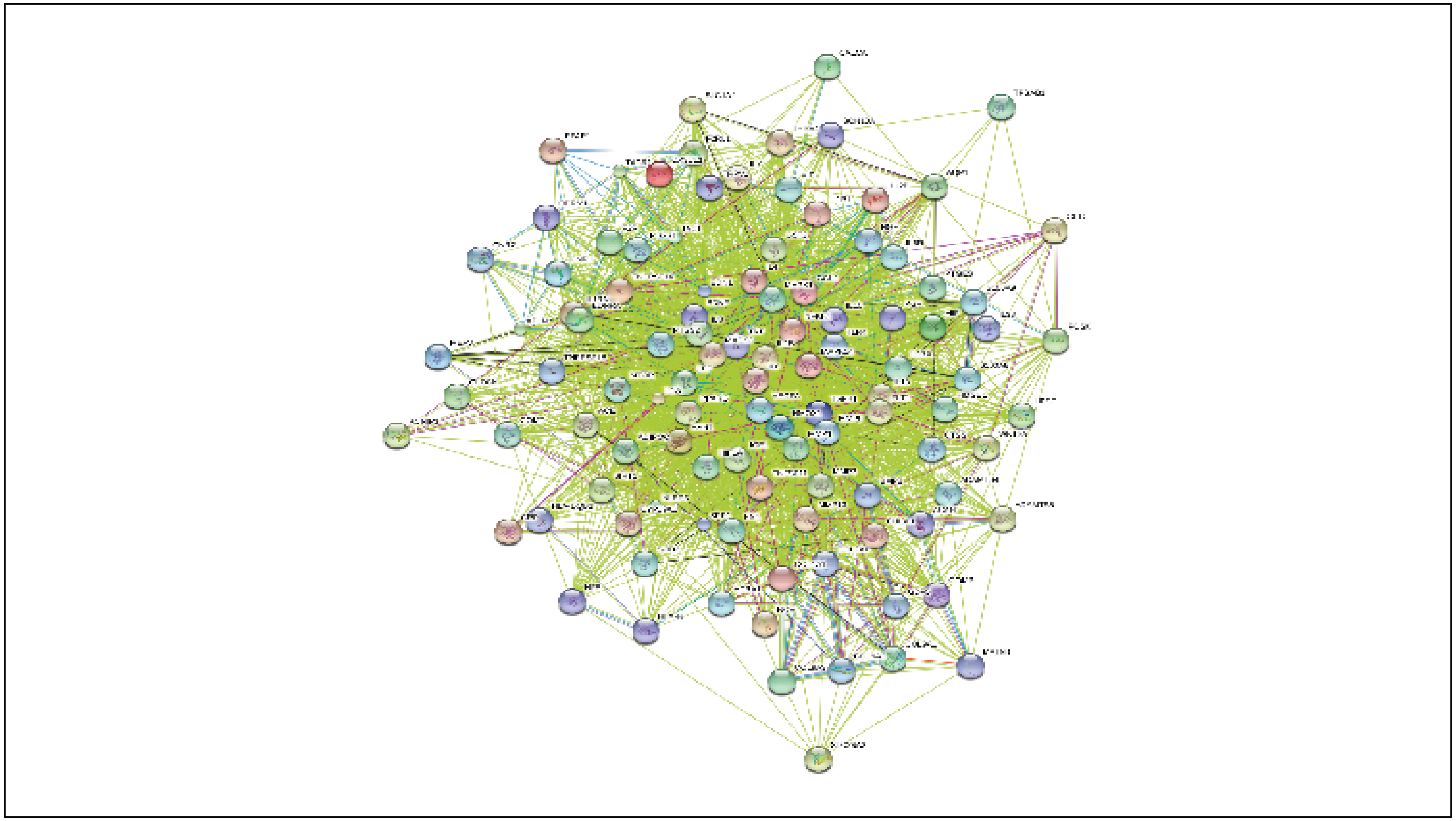
● = Medium n° publications

Case 4:
target discovery PP-interaction network

- Workflow can extract Protein-gene and Protein-Protein Interaction from literature.
- based on a custom dictionary of terms of interaction.
  - It is important to find out and extract the right Interactions
- The most «interesting» protein lists are sent to iHOP to investigate and compare the results

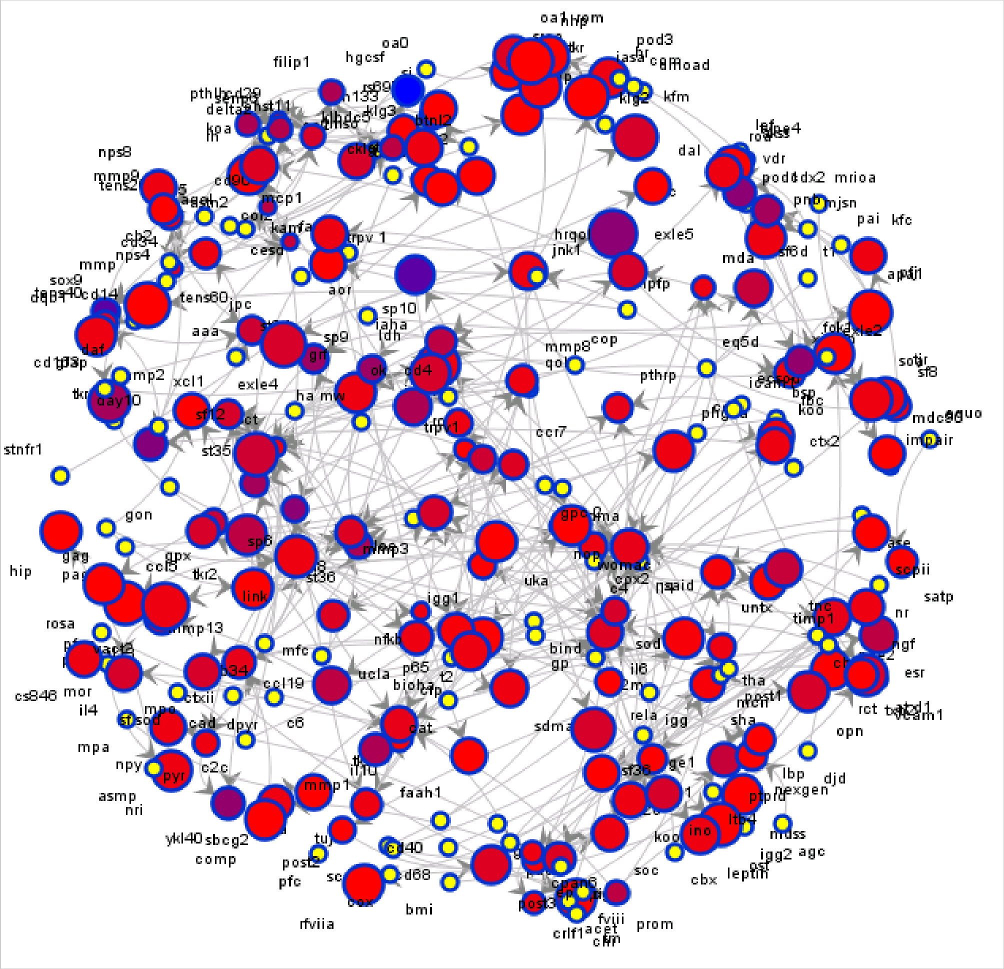
Case 4: target discovery PP-Interaction network by R

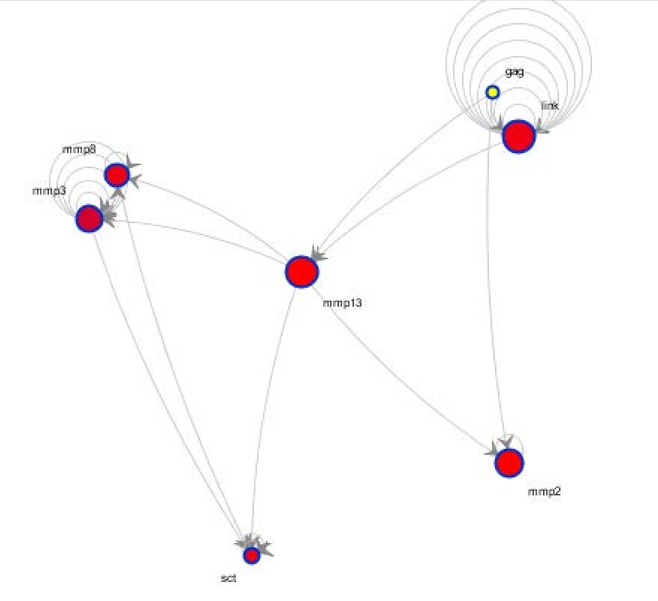
proteins: 200
interactions: 2492
expected interactions: 1021 (p−value: 0)



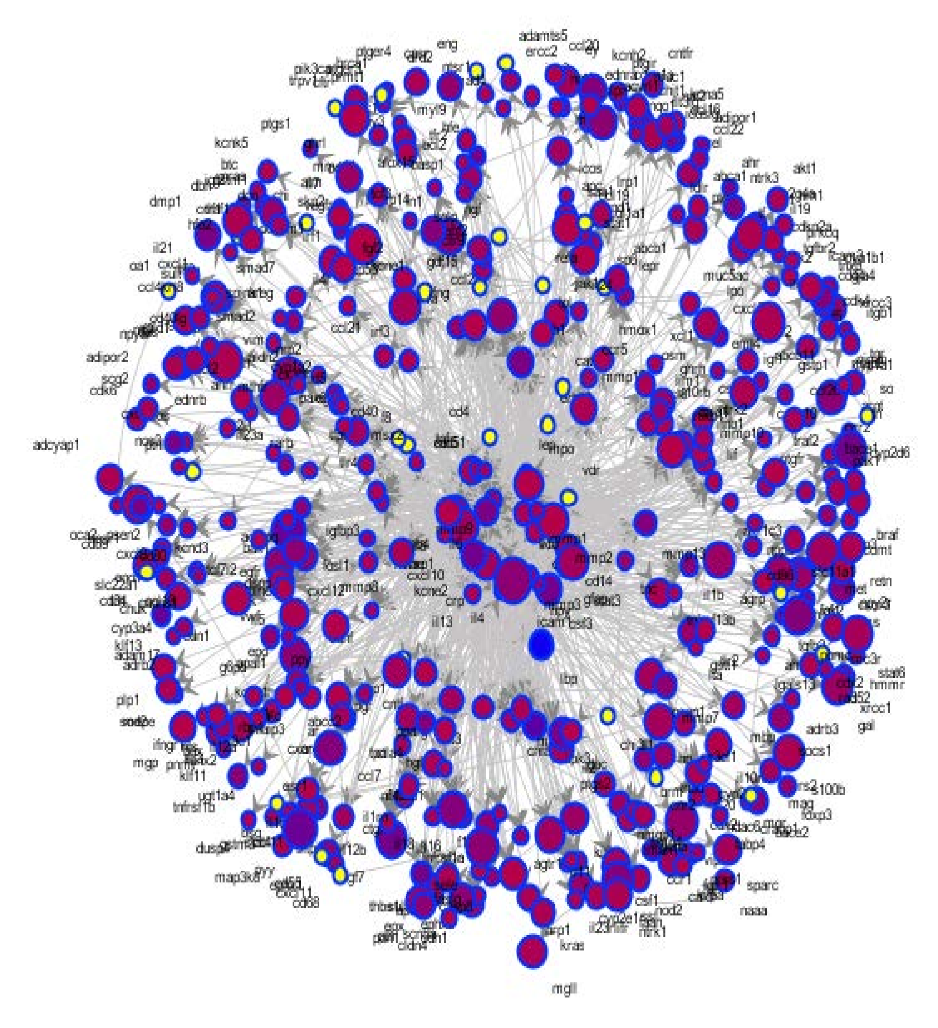
http://string−db.org/10/p/8344194184

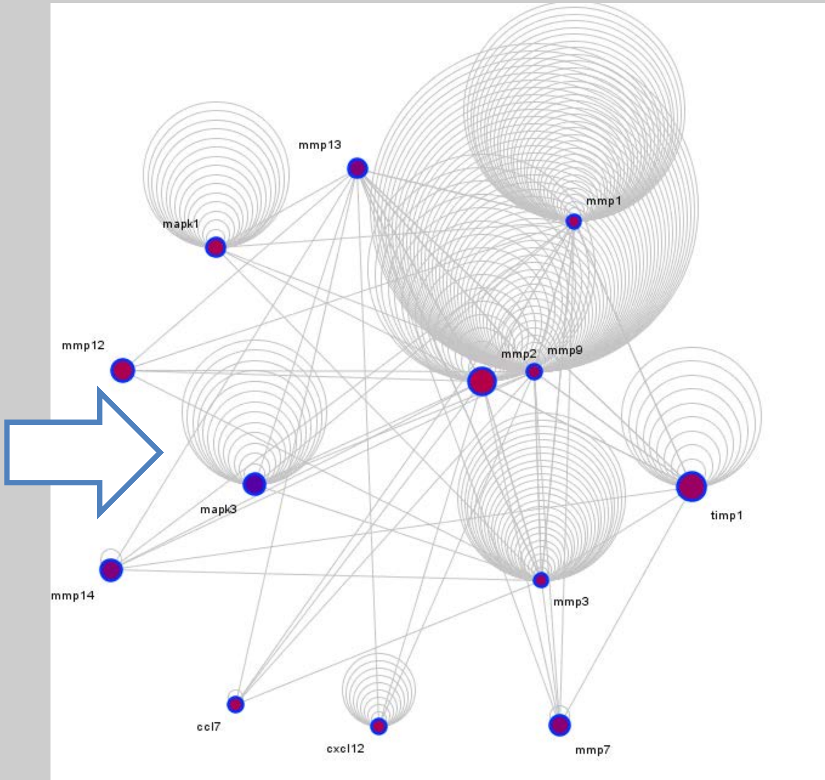
Case 4:
target discovery PP-Interaction network (knime)





Case 4:
target discovery PP-Interaction network (iHOP)





Conclusions

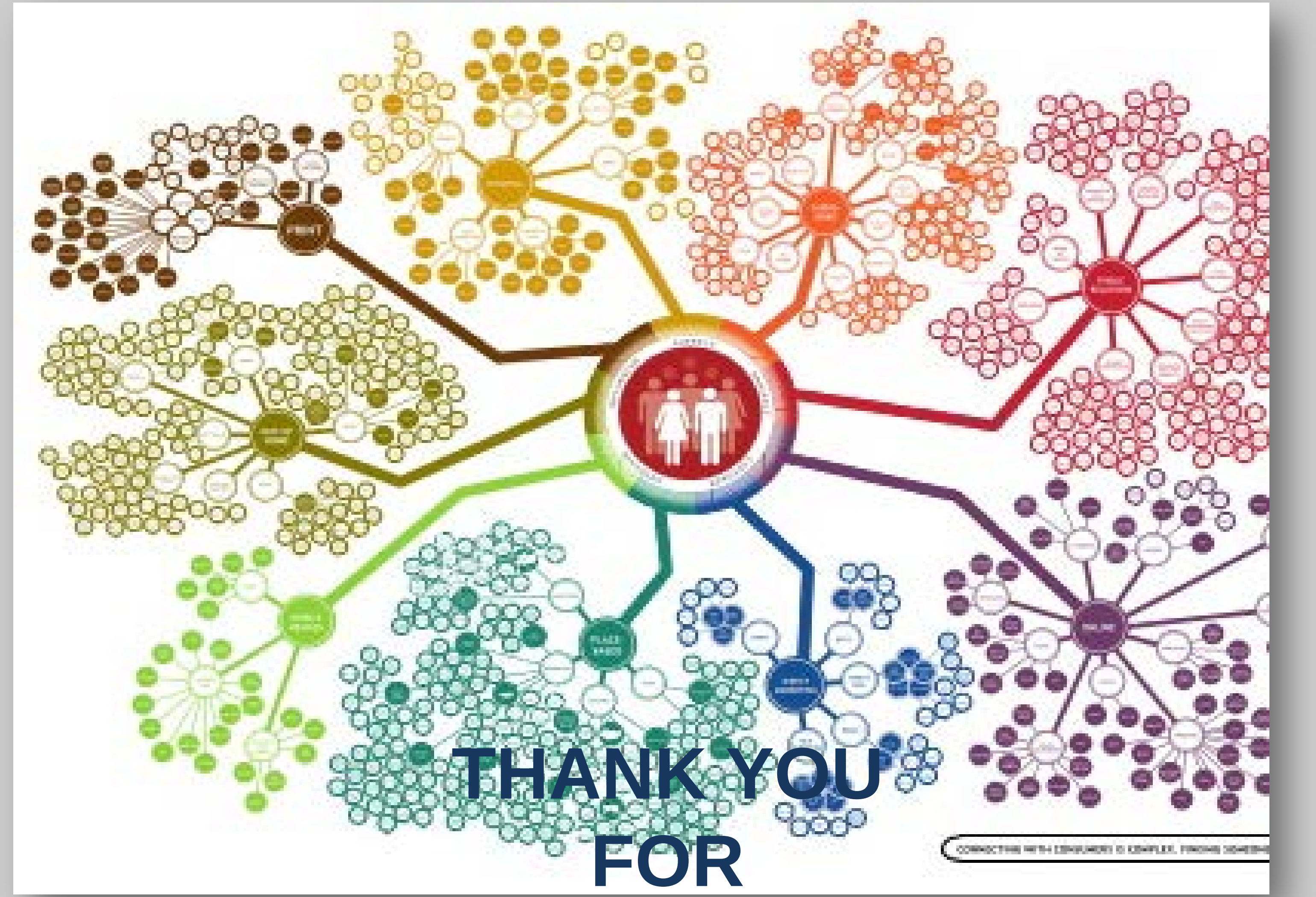
- TM brings knowledge from literature to identify known and new relationship between genes, pathways, drugs and diseases.
- TM is integration of different sources even in human ware
- A collaborative team with Biologists, Chemists, Statisticians, Librarians and Computer Technicians will be a new challenge in drug discovery.
- Biomedical Librarians can play a critical role in TM process, providing researchers access to taxonomy and vocabulary use or creation, and sharing content platform.

Rottapharm Biotech TM Team

Paolo Garofalo, Biologist
Fabrizio Colace, Medicinal Chemist
Francesca Gualtieri, Biomedical Librarian

Lucio Rovati, Medical Doctor

We would like to thank Fabrizio Arosio, computer Technician, for his helpful work.



THANK YOU FOR YOUR ATTENTION

info @: francesca.gualtieri@rottapharmbiotech.com