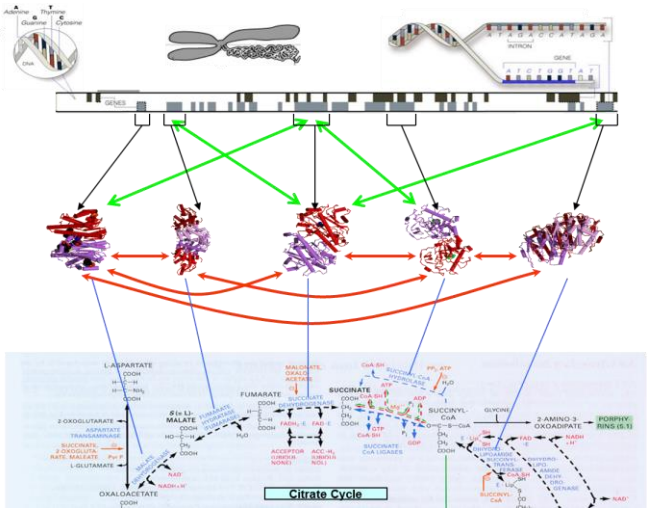
Function Prediction in Protein-Protein-Interaction Networks

E.M. Bakker| LIACS
and Hossein Rahmani| Maastricht University

11-12 2018

Universiteit Leiden
The Netherlands

Bio-Map



GENOME

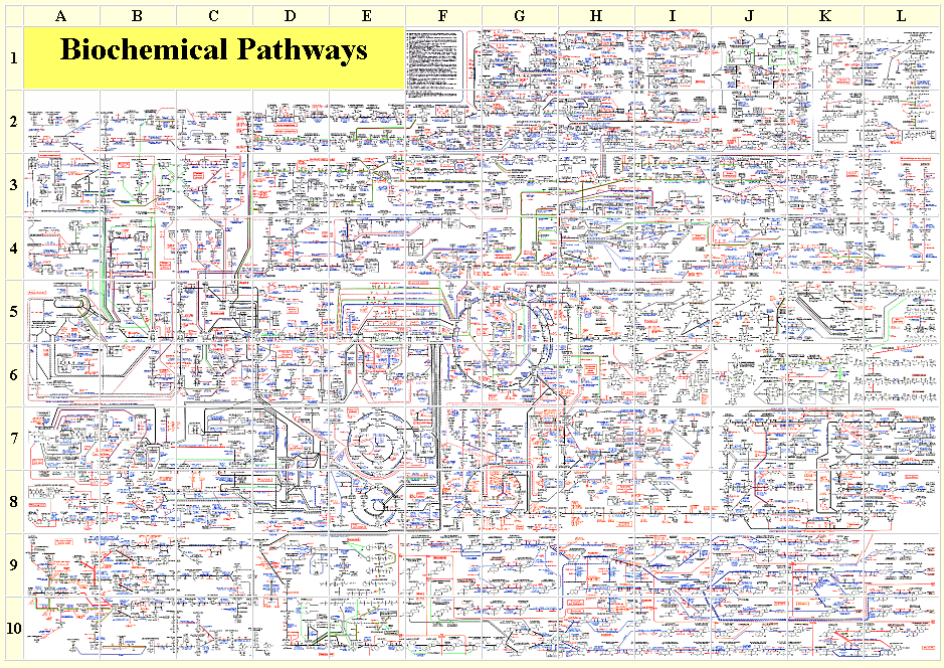
protein-gene interactions

PROTEOME

protein-protein interactions

METABOLISM

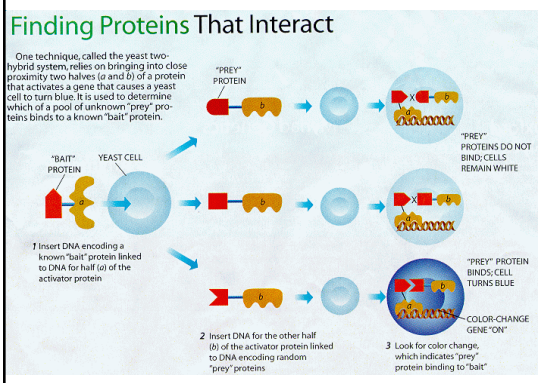
Bio-chemical reactions

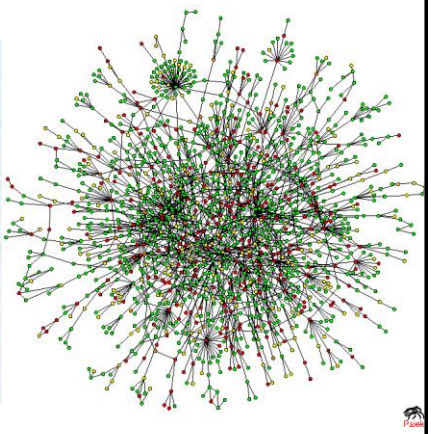


Protein Interaction Map: Yeast Protein Network

Nodes: proteins

Links: physical interactions (binding)

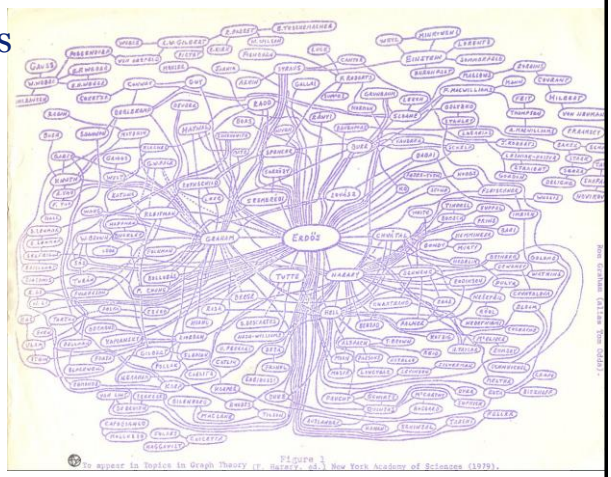




P. Uetz, *et al.*, *Nature* **403**, 623-7 (2000)

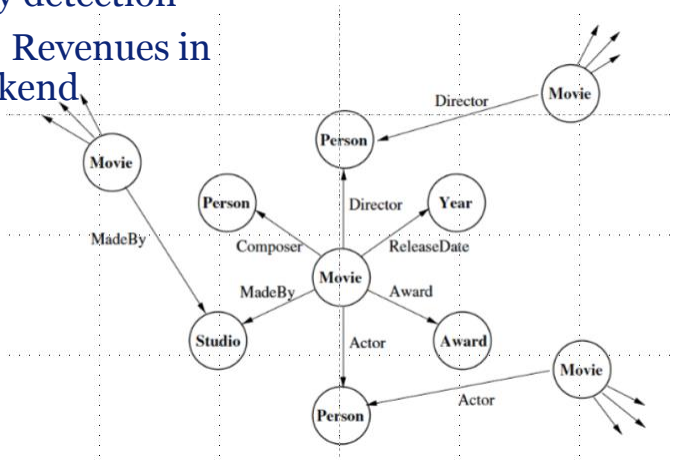
Graph Mining Domains

- Scientific Cooperation
- Movie Databases
- Web Data
- Social Networks
- Bioinformatics
- Etc.



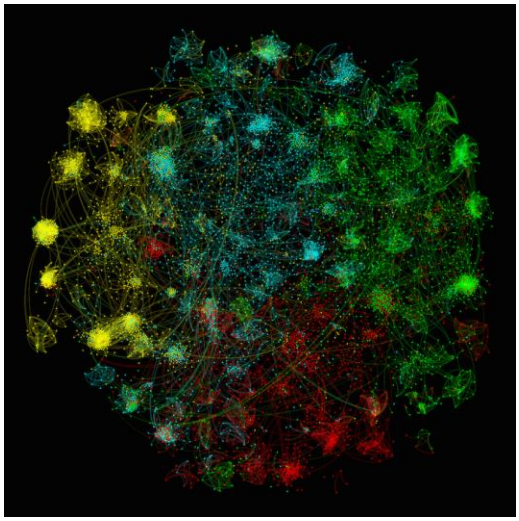
Internet Movie Databases

- Movie Recommendation
- Community detection
- Prediction: Revenues in opening weekend
- Etc.



Case Studies Similarity Network

- Similarity between texts
- Community detection
  - biomedical sciences
  - physical sciences and engineering
  - social sciences
  - arts and humanities
  - red in yellow: case studies on public outreach in science.

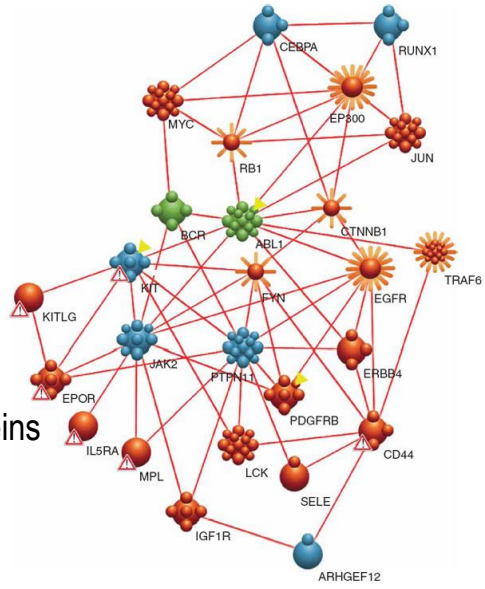


http://dx.doi.org/10.6084/m9.figshare.1476881 (2016)

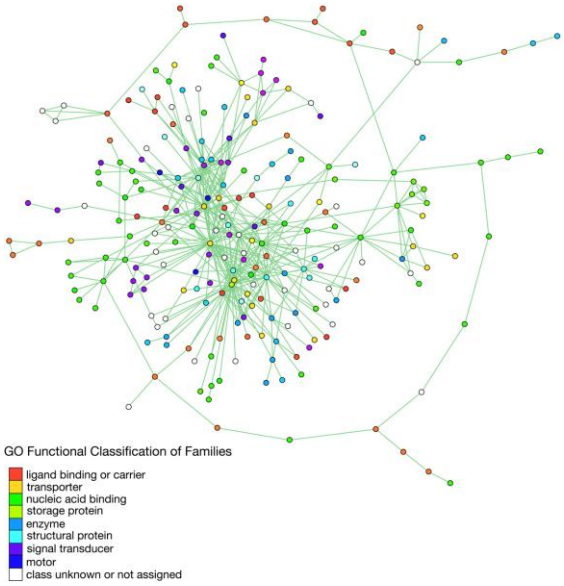
Protein-Protein Interaction (PPI) Network

- Nodes: Proteins
- Edges: Interactions

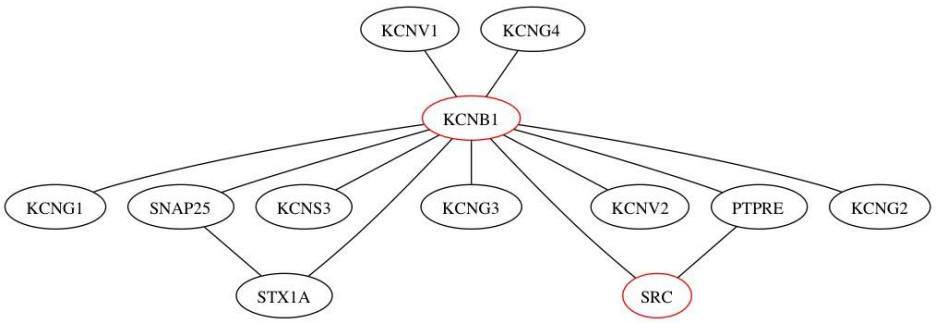
- Important Problems
  - Function Prediction
    - Similar Proteins
    - Cancer-Related Proteins
    - Disease association
    - Unknown Proteins
    - …



Function Prediction in PPI Networks



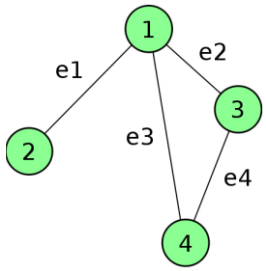
Predicting Cancer-Related Proteins in PPI Networks



Graphs: some definitions

- Graph: G = (V ,E, μ, v)
  - V : finite set of nodes.
  - $E \subseteq V \times$ V denotes a set of edges.
  - $\mu : V \rightarrow L_V$ denotes a node labeling function.
  - $v : E \rightarrow L_E$ denotes an edge labeling function.
- Let $G_1 = (V_1, E_1, \mu 1, v_1)$ and $G_2 = (V_2, E_2, \mu_2, v_2)$
- Graph G_1 is a subgraph of G_2, written $G_1 \subseteq G_2$, if:
  - $V_1 \subseteq V_2$
  - $E_1 \subseteq E_2$
  - $\mu_1(u) = \mu_2(u)$ for all u $\in V_1$.
  - $v_1(u, v) = v_2(u, v)$ for all (u, v) $\in E_1$.

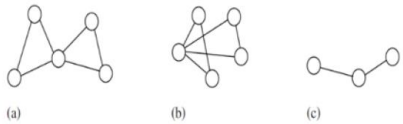
  (i.e., labels of nodes and edges remain identical)



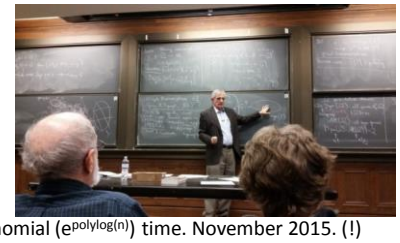
Graphs: some definitions

- Let $G_1 = (V_1, E_1, \mu_1, v_1)$ and $G_2 = (V_2, E_2, \mu_2, v_2)$,
- A graph isomorphism between G_1 and G_2 is a bijective function $f : V_1 \rightarrow V_2$ satisfying
  - $\mu_1(u) = \mu_2(f(u))$ for all nodes $u \in V_1$.
  - For every edge $e_1 = (u, v) \in E_1$, there exists an edge $e_2 = (f(u), f(v)) \in E_2$ such that $v_1(e_1) = v_2(e_2)$.
  - Conclusion: Isomorphic graphs are identical in terms of structure and labels.

Isomorphic graphs are identical in terms of structure and labels.



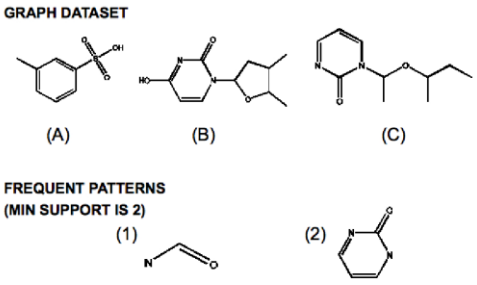
Graph (b) is isomorphic to (a) and (c) is isomorphic to a subgraph of (a)



László Babai, Graph Isomorphism is in Quasipolynomial ($e^{polylog(n)}$) time. November 2015. (!)
January 4, 2017 posting: quasipolynomial claim withdrawn
January 9, 2017 update: quasipolynomial claim restored
Degree of polyn. in exponent = 3 *Helfgott, Harald (January 16, 2017) not peer-reviewed (?).*

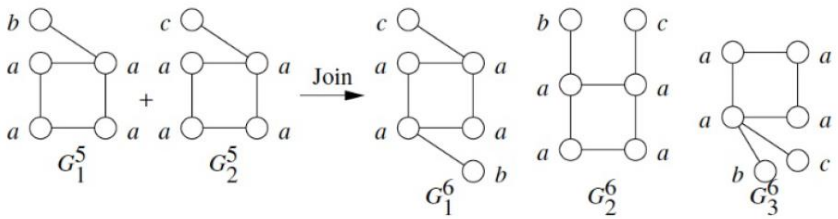
Frequent Subgraphs

- Frequent subgraphs:
  - support (subgraph) ≥ minimum support
- Usage:
  - Graph Classification
  - Graph Clustering
  - Graph Indexing
- Detection Algorithms:
  - Apriori-Based Approach
  - Pattern Growth Approach



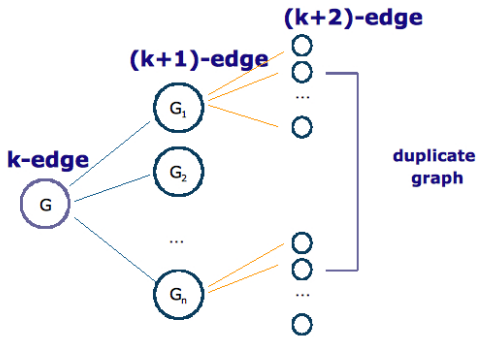
Frequent Subgraphs: Apriori-Algorithm

```
1: S_{k+1} ← ∅;
2: for each frequent g_i ∈ S_k do
3:   for each frequent g_j ∈ S_k do
4:     for each size (k+1) graph g formed by the merge of
       g_i and g_j do
5:       if g is frequent in D and g ∉ S_{k+1} then
6:         insert g to S_{k+1};
7: if s_{k+1} ≠ ∅ then
8:   call Apriori(D, min_support, S_{k+1});
9: return;
```



Frequent Subgraphs: FP-Growth

- A graph *G* is extended by adding new edges *e*
- Adding edge *e* may or may not introduce a new node to *G*

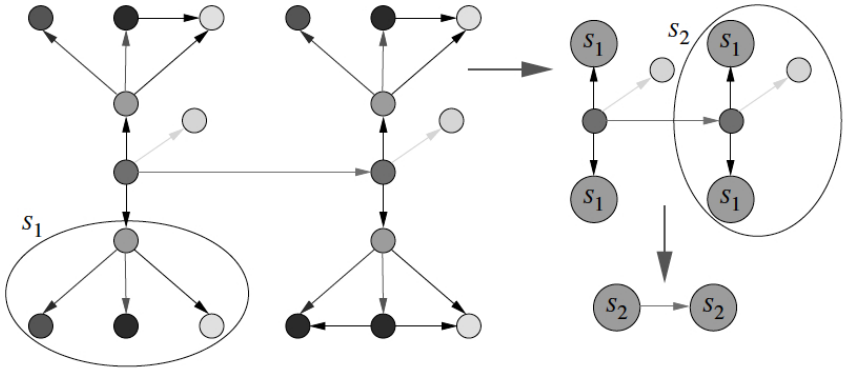


Frequent Subgraphs: Algorithms

- Apriori-based approach
  - AGM/AcGM: Inokuchi, et al. (PKDD'00)
  - FSG: Kuramochi and Karypis (ICDM'01)
  - PATH#: Vanetik and Gudes (ICDM'02, ICDM'04)
  - FFSM: Huan, et al. (ICDM'03)
- Pattern growth approach
  - MoFa, Borgelt and Berthold (ICDM'02)
  - gSpan: Yan and Han (ICDM'02)
  - Gaston: Nijssen and Kok (KDD'04)

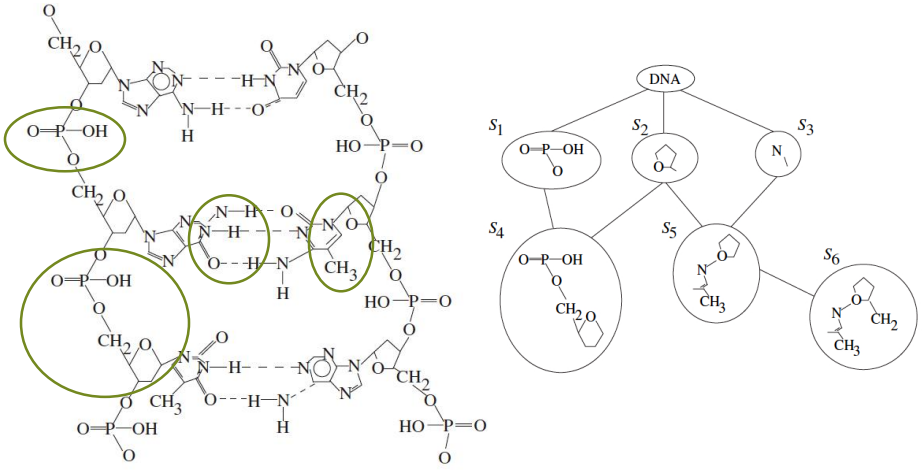
T. Ramraj et al. Frequent Subgraph Mining Algorithms – A Survey. Procedia Computer Science 47, pp 197 – 214, 2015.

Frequent Subgraphs: Compression



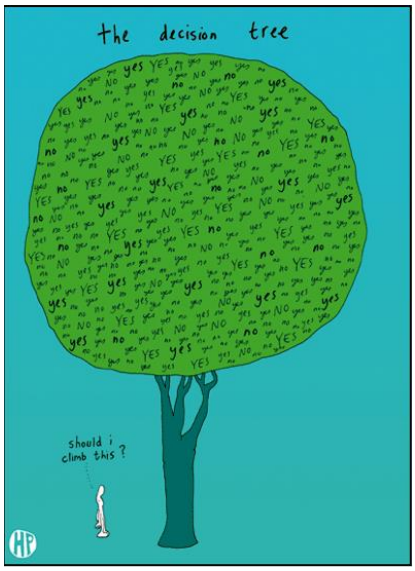
Graphs may be very big and consisting of a network of very similar subgraphs.
Graph compression helps to improve the complexity of the problem.

Frequent Subgraphs: Compression



Graphs may be very big and consisting of a network of very similar subgraphs.
Graph compression helps to improve the complexity of the problem.

Graph Based Decision Trees



Branches: attribute values
Leafs: classes

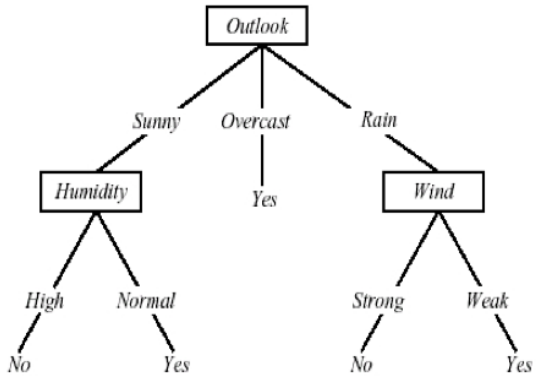
Graph Based Decision Trees

When to play tennis?

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

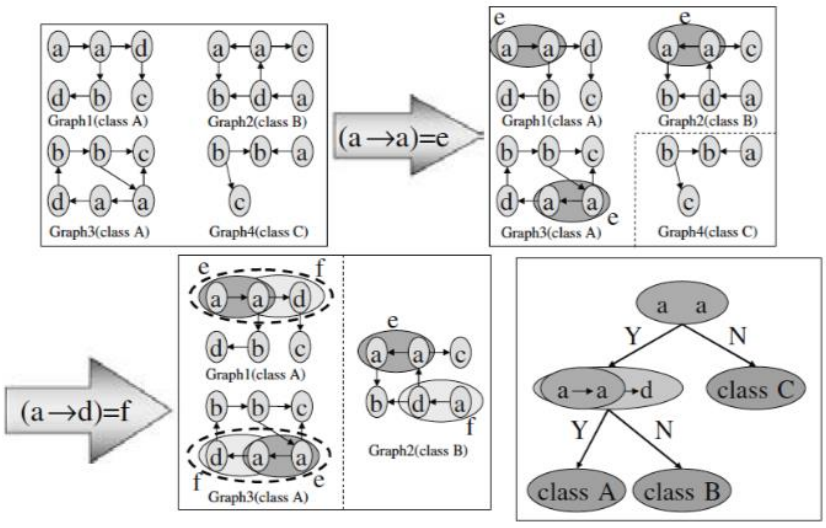
Graph Based Decision Trees

When to play tennis?

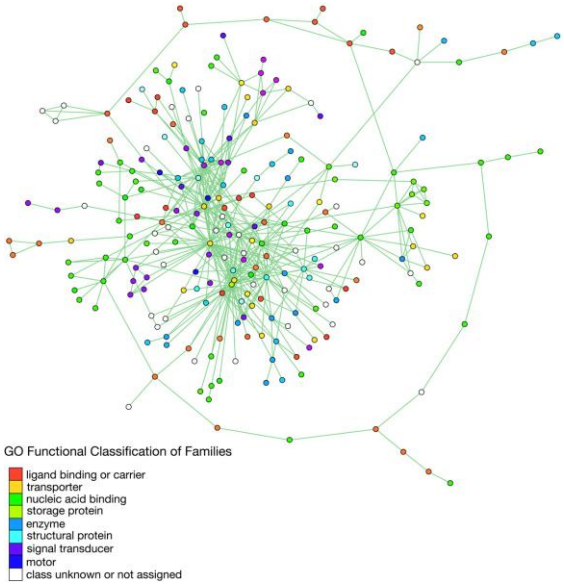


Graph Based Decision Trees

Which graph belongs to which class?

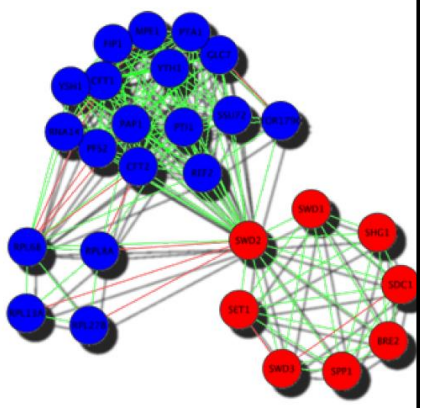


Function Prediction in PPI Networks



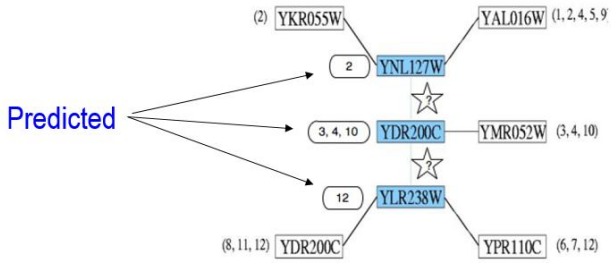
Function Prediction in PPI Networks: Similarity Based

- Assumption:
  - Interacting proteins have similar functions
- Optimization criteria:
  - Minimizing the number of interacting of proteins with no common function
- Majority Rule
- Functional Clustering



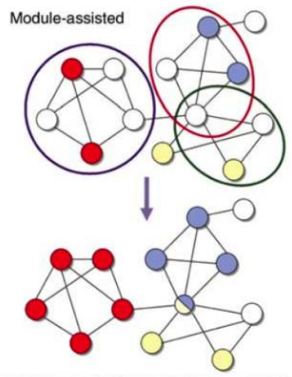
Function Prediction in PPI Networks: Majority Rule

- Predicted function: Most common function(s) among classified partners
- Problem: Links unclassified-unclassified proteins completely neglected (Coverage problem.)



Function Prediction in PPI Networks: Functional Clustering

- Cluster the PPI network
- Predict the function of unclassified protein based on the cluster they belong to



Function Prediction in PPI Networks: Collaboration Based

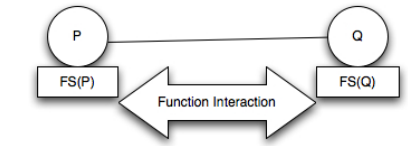
- Main Idea: A biological process is the aggregation of each individual protein's functions
- Assumption: Topologically close proteins tend to have *collaborative functions*
- Collaborative functions: Pairs of functions that frequently interface with each other in different interacting proteins
- A Reinforcement Based Function Predictor (RL)
- SOM Based Function Predictor
- protein *p*:

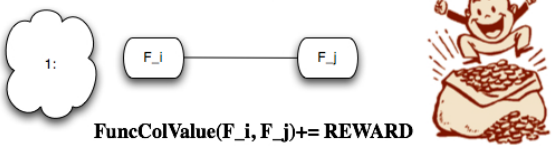
$$\begin{cases} Function\ \ Set: \ \ FS_p; FS_p(f_i) \ (= 1, \text{ if } f_i \text{ occurs in } N_p, \text{ 0 otherwise}) \\ Neighborhood\ \ Function\ \ Vector: NB_p; NB_p(f_j) \end{cases}$$


Hossein Rahmani

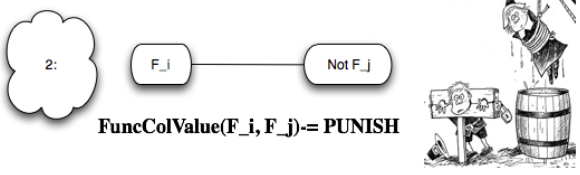
where $NB_p(f_j)$= number of times function f_j occurs in N_p (Neighborhood of p)

A Reinforcement Based Function Predictor





FuncColValue(F_i, F_j)+= REWARD



FuncColValue(F_i, F_j)-= PUNISH

A Reinforcement Based Function Predictor

- Prediction Time:
  - Select candidate functions
  - Rank candidate functions based on how well they collaborate with the neighborhood of unclassified protein p
  - Formula (1) assigns a collaboration score to each candidate function f_c:

$$Score(f_c) = \sum_{\forall f_j \in F} NB_p(f_j) * FuncColVal(f_j, f_c) \quad (1)$$

Similarity vs Collaboration

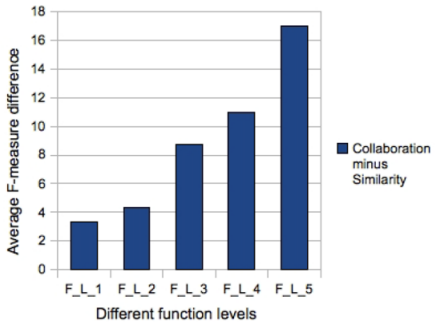
- Three Yeast Datasets: Krogan, VonMering and DIP-Core
- Five different function levels
  - 11.02.01 (rRNA synthesis) Vs 11.02.03 (mRNA synthesis)

01 METABOLISM
01.01 amino acid metabolism
01.01.03 assimilation of ammonia, metabolism
01.01.03.01 metabolism of glutamine
01.01.03.01.01 biosynthesis of glutamine
01.01.03.01.02 degradation of glutamine
01.01.03.02 metabolism of glutamate
01.01.03.02.01 biosynthesis of glutamate

Similarity vs Collaboration

- In all three datasets, collaboration methods predict functions more accurately than similarity based methods
- More detailed functions level → More difference in performance

0 - 100



References

- Diane J. Cook and Lawrence B. Holder. "Mining Graph Data". John Wiley & Sons, 2006.
- Hossein Rahmani, Hendrik Blockeel and Andreas Bender, "Collaboration-based function prediction in protein-proein interaction networks", Proceedings of the Fourth International Workshop on Machine Learning in Systems Biology, October 2010.

Kohler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. The Amerian Journal of Human Genetics. 2008; 82(4):949±58. https://doi.org/10.1016/j.ajhg.2008.02.013 PMID: 18371930

Exam

Date: Monday 7-1 2019

Time:14.00 - 17.00

Place: Room F104 (Van Steenis Building)

Do not forget your student-card.

- Please note, it is an open book exam, you can take with you your book, and printed course notes (slides). **No electronic equipment is allowed though.**
- Materials to be studied:
  - All contents covered and discussed during lectures (see links to all slides in the schedule).
- References:
  - For background information on the study material see references mentioned in the slides, and Chapters 2 - 7 of the book J. Han et al. Data Mining Concepts and Techniques.
- Exam example questions: 2014, **2015, 2016, 2017**

Snellius Building
Niels Bohrweg 1
Leiden

Van Steenis Building
Einsteinweg 2
Leiden

Data Mining Assignment 2

Databases and Data Mining 2018
Data Mining Assignment 2

Topology based Protein-Protein Interaction Network Analysis for Prioritization of Candidate Genes associated with Menière's Disease.

4. Lin Li, YanShu Wang, Lifeng An, XiangYin Kong, Tao Huang, *A network-based method using a random walk with restart algorithm and screening tests to identify novel genes associated with Menière's disease.* PLOSL ONE, August 7, 2017

- https://string-db.org/cgi/download.pl?sessionId=2FWULbtZb2d4&species_text=Homo+sapiens
- Select

INTERACTION DATA

File	Description	Access
9606.protein.links.v10.5.txt.gz (**65.9 Mb**)	protein network data (scored links between proteins)	

Due Monday 17-1 2019