Analysis of bio-molecular networks through RANKS (RAnking of Nodes with Kernelized Score Funcions)

Giorgio Valentini

Computer Science Department

UNIVERSITÀ DEGLI STUDI DI MILANO

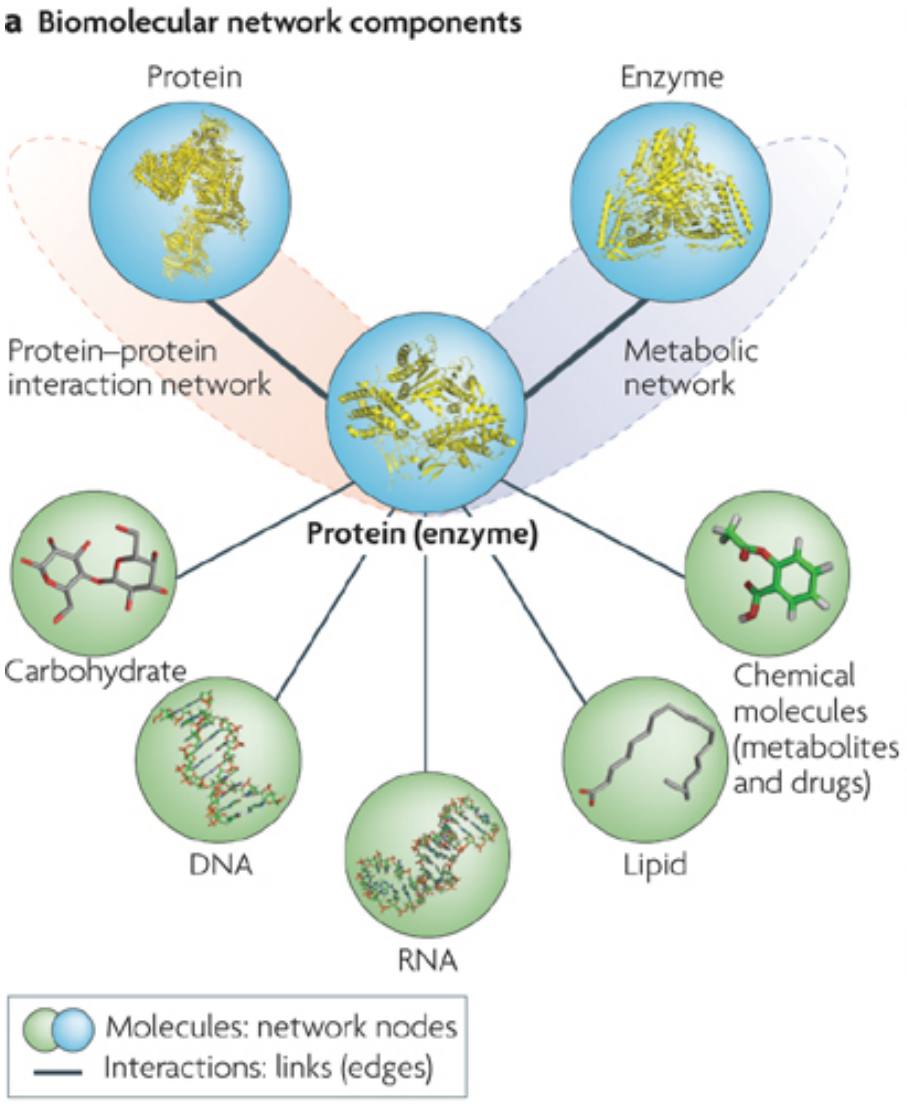
Anacleto Lab

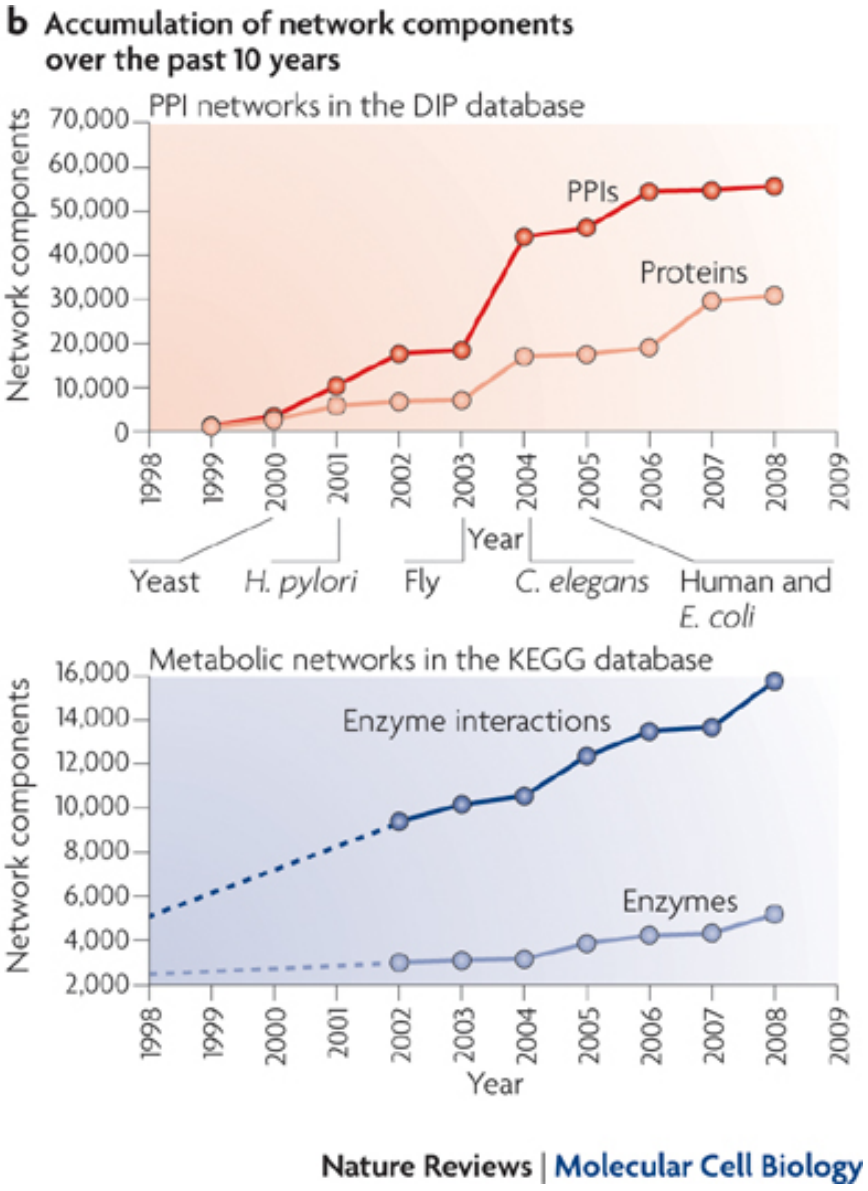
Computational Biology and Bioinformatics

Outline

- Relevant problems in molecular biology and medicine can be modeled through graphs
- The node labeling and ranking problem in complex biological networks
- Merging local and global learning strategies: the kernelized score functions algorithmic scheme

Biomolecular networks

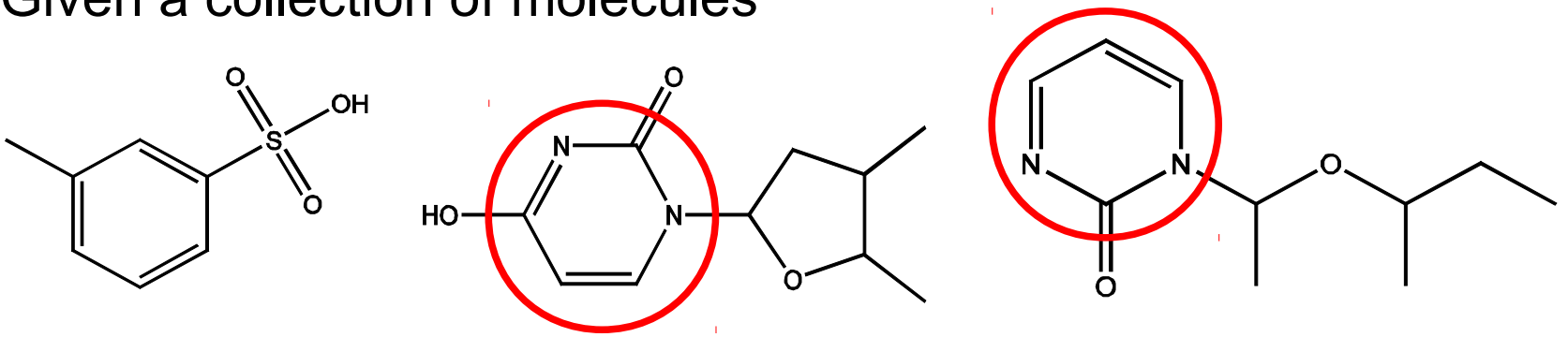




Nature Reviews | Molecular Cell Biology

Drug repositioning

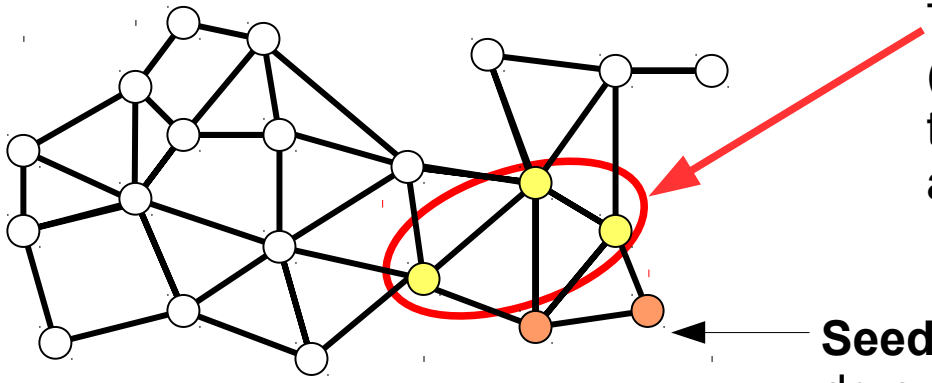
Given a collection of molecules



(A) (B) (C)

Find a meaningful way to express a similarity between them (i.e. binary profiles indicating the presence/absence of substructures used as proxy for the computation of a global similarity score between each pair of molecules).

Nodes: drugs
Edges: similarity between drugs



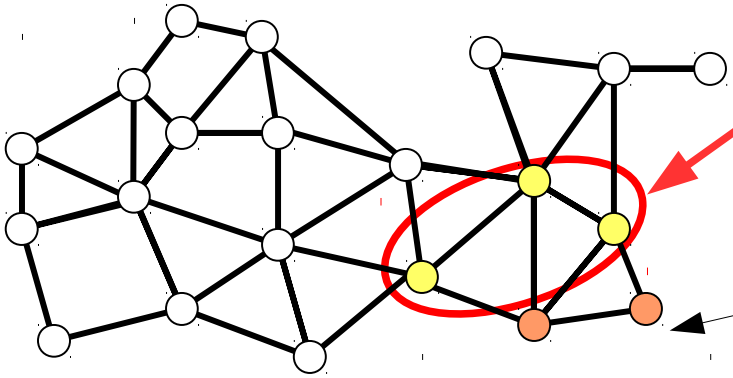
The **most similar** nodes (drugs) are candidates for the development of novel anticonvulsant drugs

Seed node, a **marketed** drug (i.e. anticonvulsant)

Automated Function Prediction (AFP)

Given a collection of proteins.

Find a meaningful way to express a similarity between them (i.e. binary profiles indicating the presence/absence of protein domains, 3D structure signatures, presence/absence of catalytic groups used as proxy for the computation of a global similarity score between each pair of ptoreins).

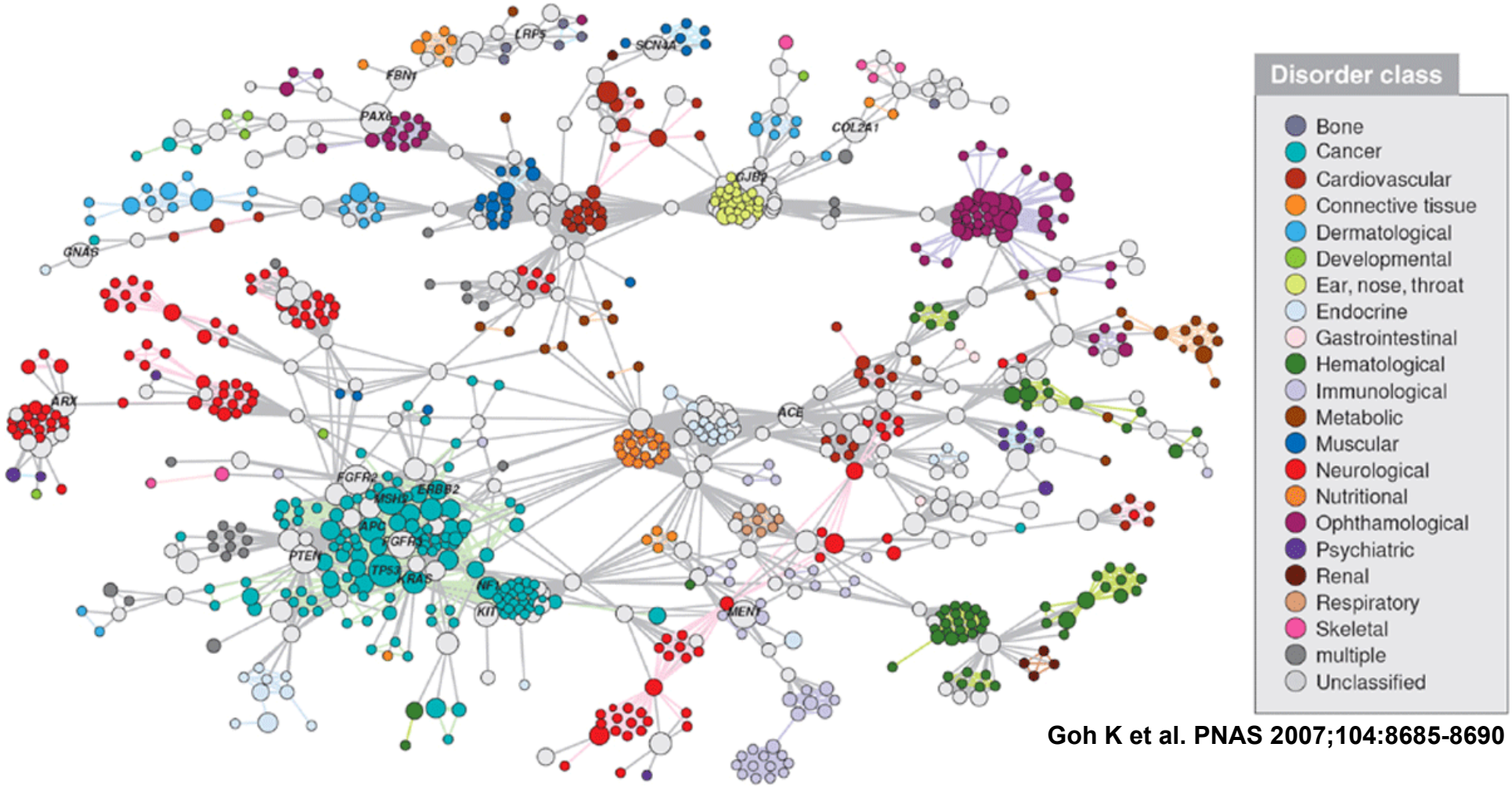


The **most similar** nodes (proteins) are candidates for the association to the functional term associated to the seeds

Seed node, associated to a **functional vocabulary term** (i.e. Gene Ontology)

Disease gene networks

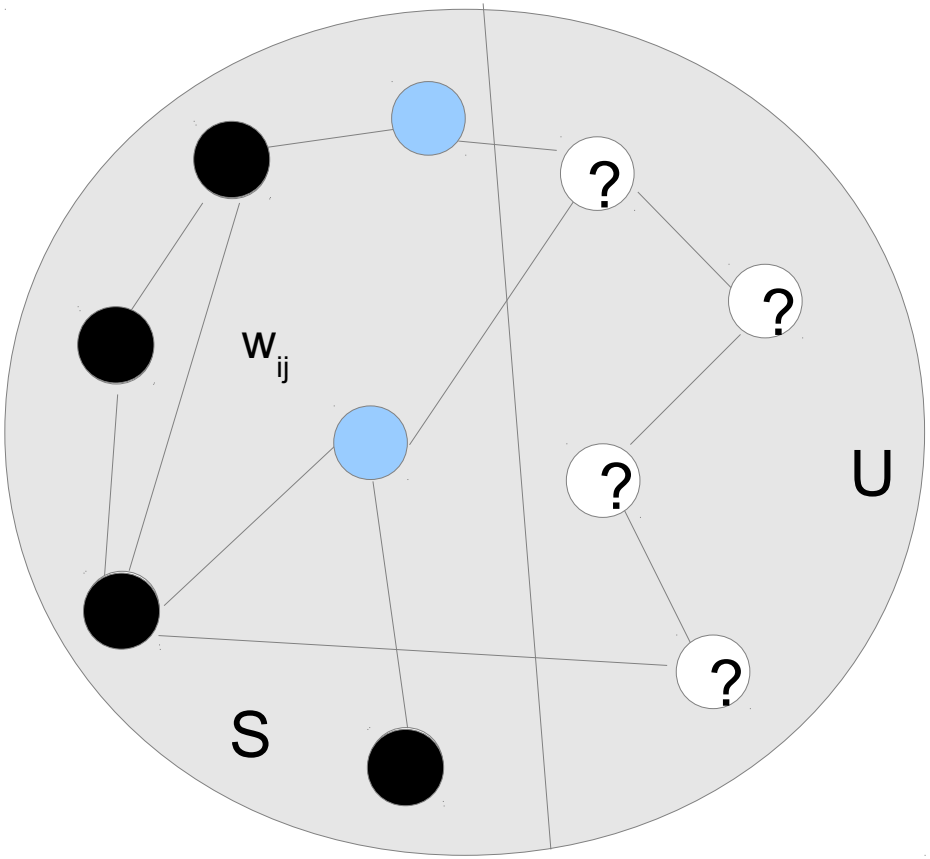
Given a collection of genes. Build a network whose nodes (genes) are connected only if they are involved into disorders of the same class.



Goh K et al. PNAS 2007;104:8685-8690

Graph Semi-Supervised Learning (GSSL) problem

$G = \langle V, E \rangle$



V : proteins,genes,drugs,...

E : functional similarities/relationships

W : similarity matrix

S : labeled nodes

U : unlabeled nodes

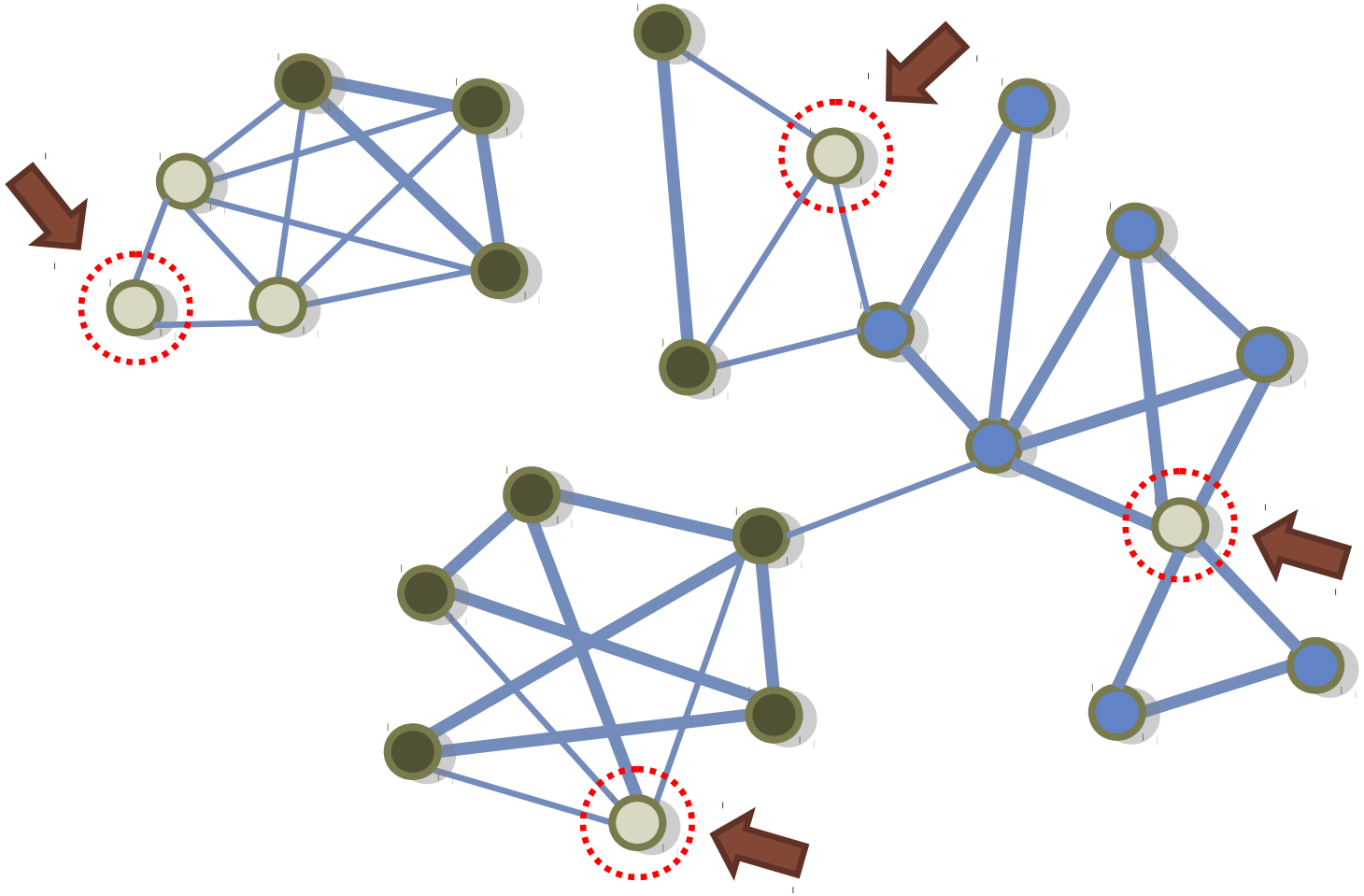
GOAL: predict labels for unlabeled nodes (*labeling problem*) or rank nodes with respect to the class to be predicted (*ranking problem*)

Node labeling/ranking methods in computational biology

- Guilt by association (*Marcotte* et al., 1999, *Oliver* et al. 2000)
- Evaluation of functional flow in graphs (*Vazquez* et al. 2003)
- Hopfield network-based methods (*Karaoz* et al. 2004, *Bertoni et al*. 2011)
- Local learning and weighed integration (*Chua* et al 2007)
- Label propagation based on Markov fields (*Deng* et al. 2004)
- Kernel methods for semi-supervised learning and integration of networks (*Tsuda* et al. 2005, *Borgwardt et al.* 2011)
- Label propagation based on Gaussian random fields and ridge regression (*Mostafavi* et al. 2008)
- Random walk-based algorithms (*Kohler et al*., 2008, *Bogdanov* and *Singh*, 2010)
- ...

Local learning strategy:

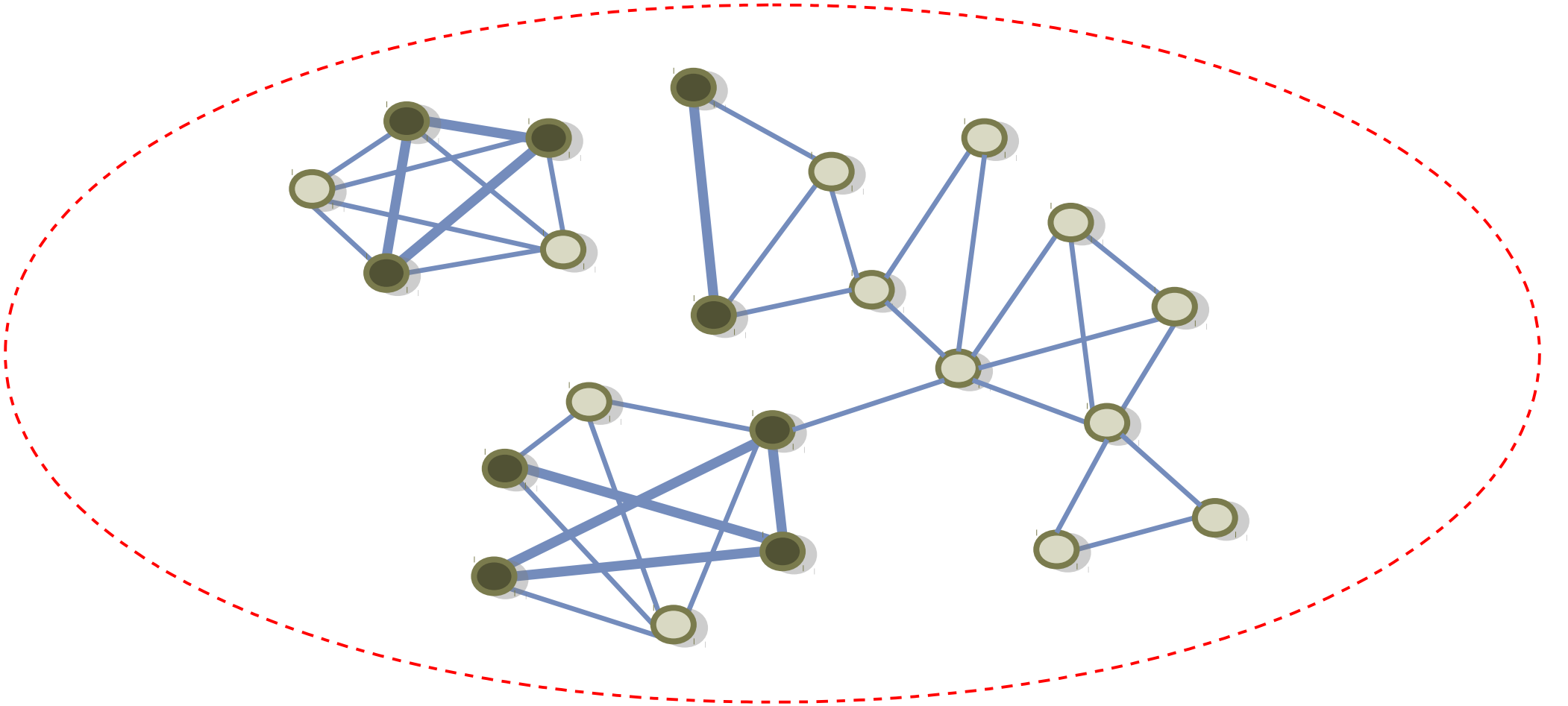
Guilt-by-association (*Marcotte* et al., 1999, *Oliver* et al. 2000)



Global learning strategy:

Exploitation of the overall network topology

(*Karaoz et al. 2004, Bengio et al. 2008, Borgwardt et al. 2011*)

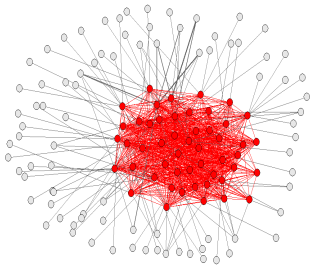


Kernelized score functions: putting together local and global learning strategies (Valentini *et al. 2016*)

Global learning

Kernel

Any kernel. E.g.:
- *Linear kernel*
- *Gaussian kernel*
- *Graph kernels*



Score function

Node ranking

Local learning

Average score : $S_{AV}(v,V_C)=\frac{1}{|V_C|}\sum_{x\in V_C} K(v,x)$

kNN score : $S_{kNN}(v,V_C)=\sum_{x\in kNN(v)} K(v,x)$

NN score : $S_{NN}(v,V_C)=max_{x\in V_C} K(v,x)$

Example of a kernel well-suited to capture the topology of the graph: the Random Walk Kernel (Smola and Kondor, 2003)

$$L = D - W \quad d_{ii} = \sum_j w_{ij}$$

Normalized graph Laplacian

$$\tilde{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}} =$$
$$D^{-\frac{1}{2}} D D^{-\frac{1}{2}} - D^{-\frac{1}{2}} W D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$$

$$K_{rw} = aI - \tilde{L} = aI - I + D^{-\frac{1}{2}} W D^{-\frac{1}{2}} =$$
$$(a-1)I + D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$$

1 - step RW kernel

$$K_{rw}^{q} = (aI - \tilde{L})^{q}$$

q - step RW kernel

Derivation of kernelized score functions

$$\phi : X \rightarrow \mathcal{H} \qquad D_{AV}(i, V_C) = \left|\left| \phi(\boldsymbol{x}_i) - \frac{1}{|V_C|} \sum_{j \in V_C} \phi(\boldsymbol{x}_j) \right|\right|^2$$

$$D_{AV}(i, V_C) = < \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_i) > - \frac{2}{|V_C|} \sum_{j \in V_C} < \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) > + \frac{1}{|V_C|^2} \sum_{k \in V_C} \sum_{j \in V_C} < \phi(\boldsymbol{x}_k), \phi(\boldsymbol{x}_j) >$$

$$Sim_{AV}(i, V_C) = -K(\boldsymbol{x}_i, \boldsymbol{x}_i) + \frac{2}{|V_C|} \sum_{j \in V_C} K(\boldsymbol{x}_i, \boldsymbol{x}_j - \frac{1}{|V_C|^2} \sum_{k \in V_C} \sum_{j \in V_C} K(\boldsymbol{x}_k, \boldsymbol{x}_j)$$

$$S_{AV}(i, V_C) = -K(\boldsymbol{x}_i, \boldsymbol{x}_i) + \frac{2}{|V_C|} \sum_{j \in V_C} K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

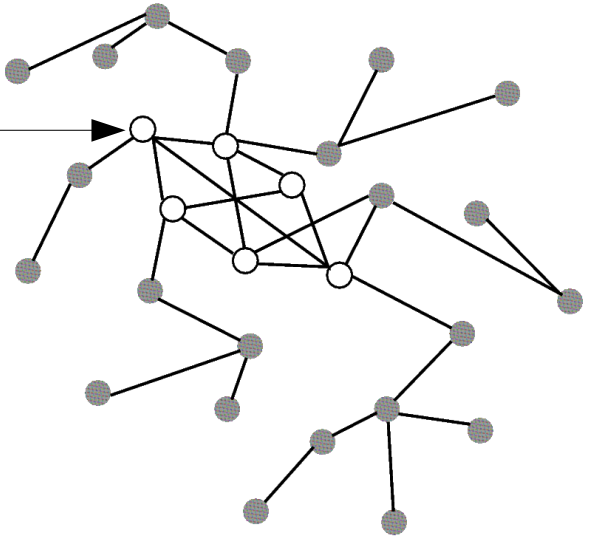
Score functions are used to rank nodes in a undirected graph

A modular approach:

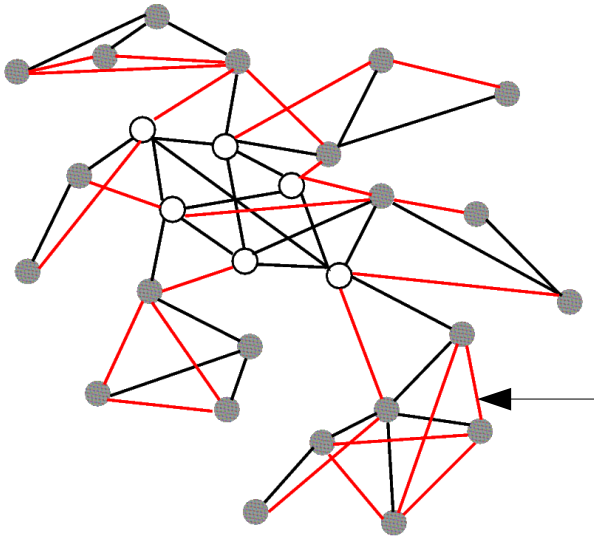
1. Select a distance - score function
2. Select a suitable kernel

Kernelized score functions: a picture of the ranking method

A positive node

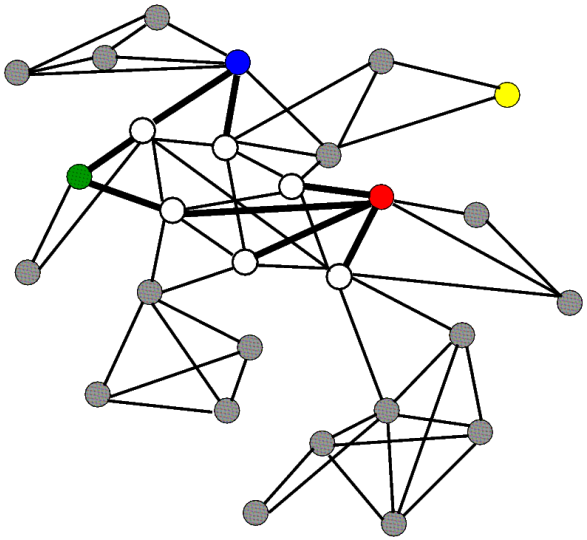


Original network



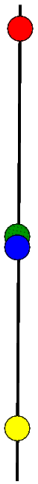
Random walk kernel

Augmented connectivity



Scoring of unlabeled nodes

high rank



connected with 4 positives

connected with 2 positives

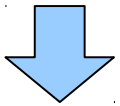
. . .

low rank

Kernelized score functions : a drug repositioning case study

M. Re, and G. Valentini, Network-based Drug Ranking and Repositioning with respect to DrugBank Therapeutic Categories, IEEE ACM Transactions on Computational Biology and Bioinformatics 10(6), pp. 1359-1371, Nov-Dec 2013

- A network G=(V,E) connecting a large set of drugs: { Nodes → drugs; Edges → similarities }
- A subset $V_C \subset V$ of drugs belonging to a given therapeutic category *C*

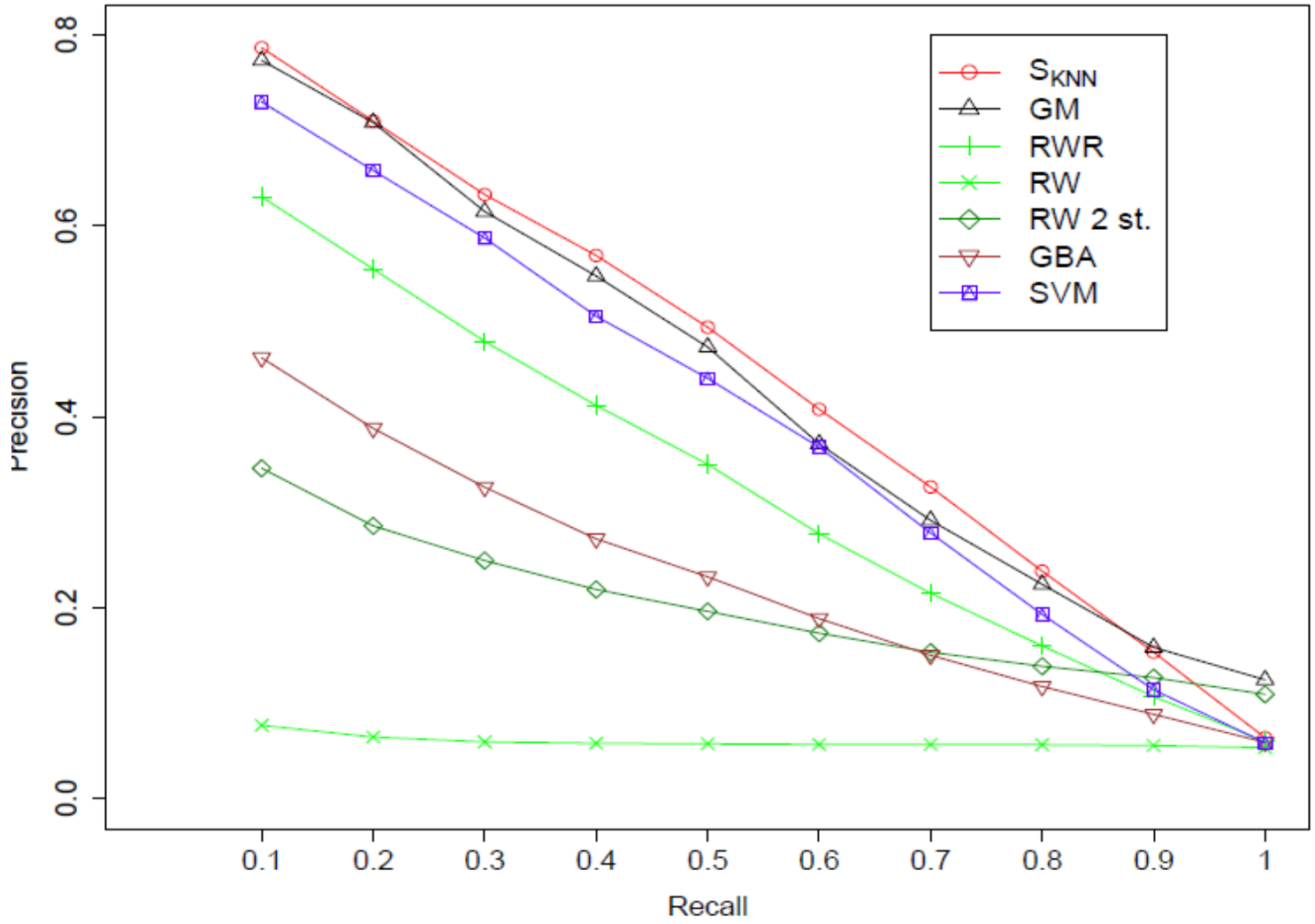


Rank drugs $v \in V$ w.r.t. to a given therapeutic category *C*

Many strategies for drugs networks construction: pairwise chemical similarity, bipartite network projection (projection in drug space of drug-target networks : drugs connected if they target the same protein/s).

Kern. score functions : a gene function prediction case study

M. Re, M. Mesiti, and G. Valentini, "A Fast Ranking Algorithm for Predicting Gene Functions in Biomolecular Networks," IEEE ACM Transactions on Computational Biology and Bioinformatics, vol. 9, no. 6, pp. 1812–1818, 2012.



Kern. score functions : a gene disease prioritization case study

G. Valentini, A. Paccanaro, H. Caniza, A. Romero, M. Re, An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods, Artificial Intelligence in Medicine 61 (2) (2014)

Goals:

- An extensive analysis of gene-disease associations, considering a large set of diseases (708 MeSH diseases)
- Finding novel gene-disease associations for unannotated genes
- Analysis of the impact of network integration on gene prioritization

Kern. score functions : a gene disease prioritization case study

G. Valentini, A. Paccanaro, H. Caniza, A. Romero, M. Re, An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods, Artificial Intelligence in Medicine 61 (2) (2014)

Goals:

- An extensive analysis of gene-disease associations, considering a large set of diseases (708 MeSH diseases)
- Finding novel gene-disease associations for unannotated genes
- Analysis of the impact of network integration on gene prioritization

Summary

- Semi-supervised graph-based methods are widely applied in several relevant problems in computational biology and medicine
- Kernelized score functions is a flexible algorithmic framework that can be applied in a broad range of interesting bioinformatics problems
- Kernelized score functions and the others state-of-the-art semi-supervised learning methods for biological network analysis are affected by serious scalability problems on big networks
- RANKS software library is avaialble as an R package from CRAN: `https://cran.r-project.org/web/packages/RANKS`

References:

- G. Valentini, G. Armano, M. Frasca, J. Lin, M. Mesiti and M. Re RANKS: a flexible tool for node label ranking and classification in biological networks, *Bioinformatics*, 32(18), 2016.
- M. Mesiti, M. Re, G. Valentini Think globally and solve locally: secondary memory-based network learning for automated multi-species function prediction, *GigaScience*, 3:5, 2014
- G. Valentini, A. Paccanaro, H. Caniza, A. Romero, M. Re, An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods, *Artificial Intelligence in Medicine*, Volume 61, Issue 2, pages 63-78, June 2014
- M. Re, and G. Valentini, Network-based Drug Ranking and Repositioning with respect to DrugBank Therapeutic Categories, *IEEE ACM Transactions on Computational Biology and Bioinformatics* 10(6), pp. 1359-1371, Nov-Dec 2013
- M. Frasca, A. Bertoni, M. Re, and G. Valentini, A neural network algorithm for semi-supervised node label learning from unbalanced data, *Neural Networks* 43, pp.84-98, July 2013
- M. Re, M. Mesiti and G. Valentini, A Fast Ranking Algorithm for Predicting Gene Functions in Biomolecular Networks, *IEEE ACM Transactions on Computational Biology and Bioinformatics* 9(6) pp. 1812-1818, 2012