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

Giorgio Valentini
DI – Dipartimento Informatica
Università degli Studi di Milano
• Relevant problems in molecular biology and medicine can be modeled through graphs

• Local and global semi-supervised learning strategies to learn node labels in graphs

• Merging local and global learning strategies: the kernelized score functions algorithmic scheme (RANKS)
Biomolecular networks

Analysis of bio-molecular networks through semi-supervised graph-based learning methods

G. Valentini
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 proteins).

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
Node labeling and ranking

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 label prediction methods

Node labeling/ranking methods in computational biology

- Local learning and weighed integration (*Chua* et al 2007)
- Label propagation based on Markov fields (*Deng* et al. 2004)
- 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 \textit{(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)
Globlal learning strategies minimize a quadratic cost function

A. Consistency with the initial labeling:

\[ \sum_{i=1}^{l} (\hat{y}_i - y_i)^2 = \| \hat{Y}_l - Y_l \|^2. \]

B. Consistency with the geometry of the data (internal consistency):

\[
\frac{1}{2} \sum_{i,j=1}^{n} W_{ij}(\hat{y}_i - \hat{y}_j)^2 = \frac{1}{2} \left( 2 \sum_{i=1}^{n} \hat{y}_i^2 \sum_{j=1}^{n} W_{ij} - 2 \sum_{i,j=1}^{n} W_{ij} \hat{y}_i \hat{y}_j \right) \\
= \hat{Y}^T (D - W) \hat{Y} \\
= \hat{Y}^T L \hat{Y}
\]

A + B + regularizations

\[
C(\hat{Y}) = \| \hat{Y}_l - Y_l \|^2 + \mu \hat{Y}^T L \hat{Y} + \mu \epsilon \| \hat{Y} \|^2
\]
Kernelized score functions: putting together local and global learning strategies (Valentini et al. 2016)

Random walk kernel

Kernel

Score function

Node ranking

Global learning

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

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

Analysis of bio-molecular networks through semi-supervised graph-based learning methods

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

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

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

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

\[
K_{rw}^q = (aI - \tilde{L})^q
\]
Derivation of kernelized score functions

\[ \phi : X \rightarrow \mathcal{H} \]

\[ D_{AV}(i, V_C) = \left\| \phi(x_i) - \frac{1}{|V_C|} \sum_{j \in V_C} \phi(x_j) \right\|^2 \]

\[ D_{AV}(i, V_C) = \langle \phi(x_i), \phi(x_i) \rangle - \frac{2}{|V_C|} \sum_{j \in V_C} \langle \phi(x_i), \phi(x_j) \rangle + \frac{1}{|V_C|^2} \sum_{k \in V_C} \sum_{j \in V_C} \langle \phi(x_k), \phi(x_j) \rangle \]

\[ \text{Sim}_{AV}(i, V_C) = -K(x_i, x_i) + \frac{2}{|V_C|} \sum_{j \in V_C} K(x_i, x_j) - \frac{1}{|V_C|^2} \sum_{k \in V_C} \sum_{j \in V_C} K(x_k, x_j) \]

\[ S_{AV}(i, V_C) = -K(x_i, x_i) + \frac{2}{|V_C|} \sum_{j \in V_C} K(x_i, x_j) \]

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 reppositioning 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 $\rightarrow$ drugs
  - Edges $\rightarrow$ 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).
1253 FDA approved drugs
51 DrugBank therapeutic classes
3 pharmacological networks:
- $N_{\text{structSim}}$: pairwise chemical similarity ($Tanimoto$ coefficients)
- $N_{\text{drugTarget}}$: projection from drug-target interactions (from DrugBank 3.0)
- $N_{\text{drugChem}}$: projection from chemical interactions (from STITCH 2.0)

Kernelized score functions

Network construction by bipartite network projection

Bipartite network (e.g. drug-target, drug-drug interaction) ➔ One-mode drug network
Kernelized score functions: experiments

High coverage
100% ........................................................       50%

\(N_{\text{structSim}}\) \(\rightarrow\) \(W_1\) (1253 nodes, 13010 edges)

\(N_{\text{structSim}} + N_{\text{drugTarget}} \rightarrow W_2\) (1253, 43827)

\(N_{\text{structSim}} + N_{\text{drugTarget}} + N_{\text{drugChem}} \rightarrow W_3\) (1253, 96711)

NB: networks integration increase the connectivity!
A view of the integrated pharmacological network
**Kernelized score functions: results (AUC)**

*Kernelized score functions* with random walk kernels compared with *Random Walk* (RW) and *Random Walk with Restart* (RWR) algorithms:

- 5-fold CV
- Results averaged across 51 DrugBank therapeutic classes having more than 15 drugs:

<table>
<thead>
<tr>
<th>Methods</th>
<th>AUC</th>
<th>P40R</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$W_1$</td>
<td>$W_2$</td>
</tr>
<tr>
<td>$S_{AV}$ 3 steps</td>
<td>0.8332</td>
<td>0.9233</td>
</tr>
<tr>
<td>$S_{kNN}$ 2 steps $k=31$</td>
<td><strong>0.8373</strong></td>
<td><strong>0.9261</strong></td>
</tr>
<tr>
<td>$S_{NN}$ 3 steps</td>
<td>0.8271</td>
<td><strong>0.9067</strong></td>
</tr>
<tr>
<td>RW $\theta = 0.6$</td>
<td>0.8078</td>
<td>0.9203</td>
</tr>
<tr>
<td>RW 1 step</td>
<td>0.8175</td>
<td>0.9201</td>
</tr>
<tr>
<td>GBA</td>
<td>0.8027</td>
<td>0.9028</td>
</tr>
<tr>
<td>RW</td>
<td>0.6846</td>
<td>0.5780</td>
</tr>
</tbody>
</table>

- $W_1 \rightarrow W_2 \rightarrow W_3$: AUC increments are statistically significant (Wilcoxon rank sum test, $\alpha=0.01$)
- $S_{AV}$ and $S_{kNN}$ significantly better than the other methods (Wilcoxon rank sum test, $\alpha=0.01$)
**Kernelized score functions:** Exploring deeply the integrated pharmacological space yields better results

Counts of the "wins" across the 1254 therapeutic classes for the average score with 1, 2, 3, 5 and 10 steps random walk kernels
Kern. score functions: a gene function prediction case study

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
Analysis of the impact of network integration on gene prioritization

But also proper pre-processing and normalization of the networks is fundamental...
### Analysis of the impact of network integration on gene prioritization

<table>
<thead>
<tr>
<th>Network</th>
<th>Description</th>
<th>Type</th>
<th>Nodes</th>
<th>Edges</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>finet</td>
<td>Obtained from multiple sources of evidence</td>
<td>Binary</td>
<td>8449</td>
<td>271466</td>
<td>0.0038</td>
</tr>
<tr>
<td>hmnet</td>
<td>Obtained from multiple sources of evidence</td>
<td>Binary</td>
<td>8449</td>
<td>502222</td>
<td>0.0070</td>
</tr>
<tr>
<td>cmnet</td>
<td>Network projections from cancer modules</td>
<td>Binary</td>
<td>8449</td>
<td>3414722</td>
<td>0.0478</td>
</tr>
<tr>
<td>gcnet</td>
<td>Network projections from CTD</td>
<td>Binary</td>
<td>7649</td>
<td>1421298</td>
<td>0.0242</td>
</tr>
<tr>
<td>bgnet</td>
<td>Network projections from BioGRID</td>
<td>Binary</td>
<td>8449</td>
<td>120169</td>
<td>0.0016</td>
</tr>
<tr>
<td>dbnet</td>
<td>Direct relationships obtained from BioGRID</td>
<td>Binary</td>
<td>8449</td>
<td>3023084</td>
<td>0.0423</td>
</tr>
<tr>
<td>bpnet</td>
<td>Semantic similarity network from GO BP</td>
<td>Real valued</td>
<td>6923</td>
<td>44506147</td>
<td>0.9286</td>
</tr>
<tr>
<td>mfnet</td>
<td>Semantic similarity network from GO MF</td>
<td>Real valued</td>
<td>6145</td>
<td>26611887</td>
<td>0.7047</td>
</tr>
<tr>
<td>ccnet</td>
<td>Semantic similarity network from GO CC</td>
<td>Real valued</td>
<td>6693</td>
<td>39652637</td>
<td>0.8851</td>
</tr>
</tbody>
</table>

**Diagram:** Box plots showing the distribution of gene prioritization scores across different network types and integration strategies.
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 available as a R package from CRAN: https://cran.r-project.org/web/packages/RANKS
References:

- 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