Network-based gene-disease prioritization using ProphNet

Víctor Martínez Gómez
Carlos Cano Gutierrez
Armando Blanco Morón

Department of Computer Science and Artificial Intelligence
University of Granada
Spain

November 2012

Prioritization

Prioritization aims to identify the most promising biological entities among a larger pool of candidates through integrative computational analysis of genomic data. These methods are based on guilt-by-association hypothesis.

Guilt-by-association hypothesis

Biological entities showing a similar behaviour or sharing interactions/relations are more likely to belong to the same biological process, to be functionally related or to share molecular basis.

Prioritization strategies

Several prioritization strategies have been proposed:

- **Filtering**
- **Text-mining**
- **Profiling**
- **Network-based**

Network-based methods

Advantages

- Better average performance than other approaches under similar conditions.
- Biological information is usually easy to representate as networks.

Disadvantages

- Difficulties integrating an arbitrary number of data sources since network-based methods usually are limited to 2 networks.
- Very specific methods that do not allow to perform different types of queries.

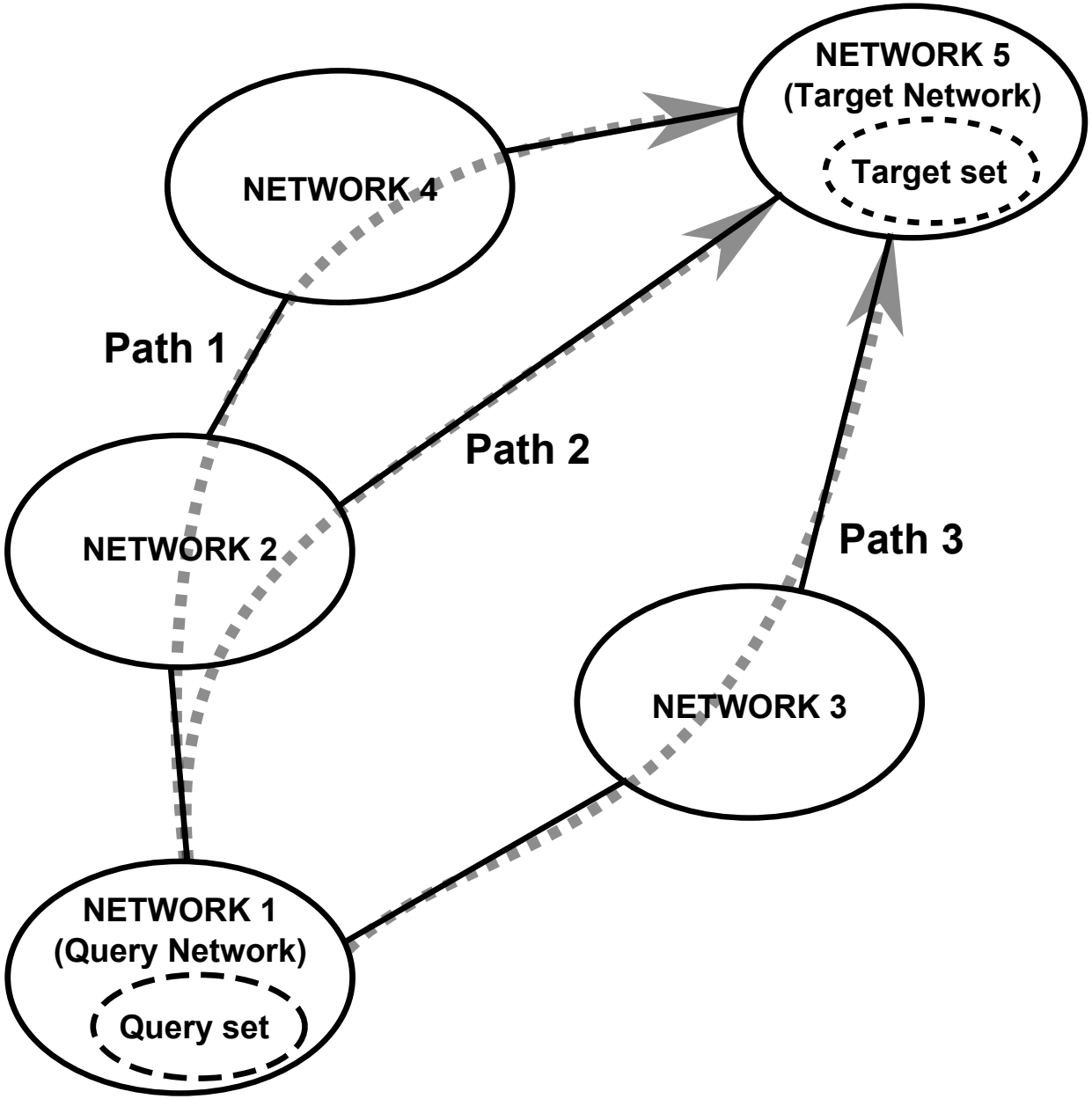
ProphNet

- **ProphNet** is a **network-based** prioritization method.
- Allows **integration** of an **arbitrary** number of networks.
- **Outperforms** recently proposed methods.
- **Flexibility in queries** allowing any prioritization task (e.g., genes-diseases or protein domains-genes).

Data representation

- **Data of a particular type** is represented by a **network**.
- **Nodes** represent **biological entities** (e.g., genes or diseases).
- **Arcs** represent **interactions or relations**.
- Networks are interconnected by other networks (e.g., gene-disease or domain-gene network) forming the **Global Graph**.

Definitions

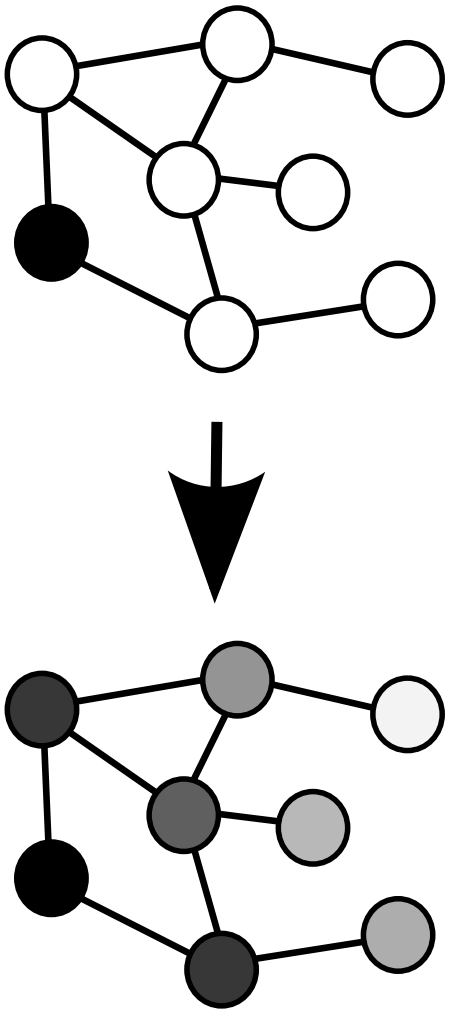


More details about data representation

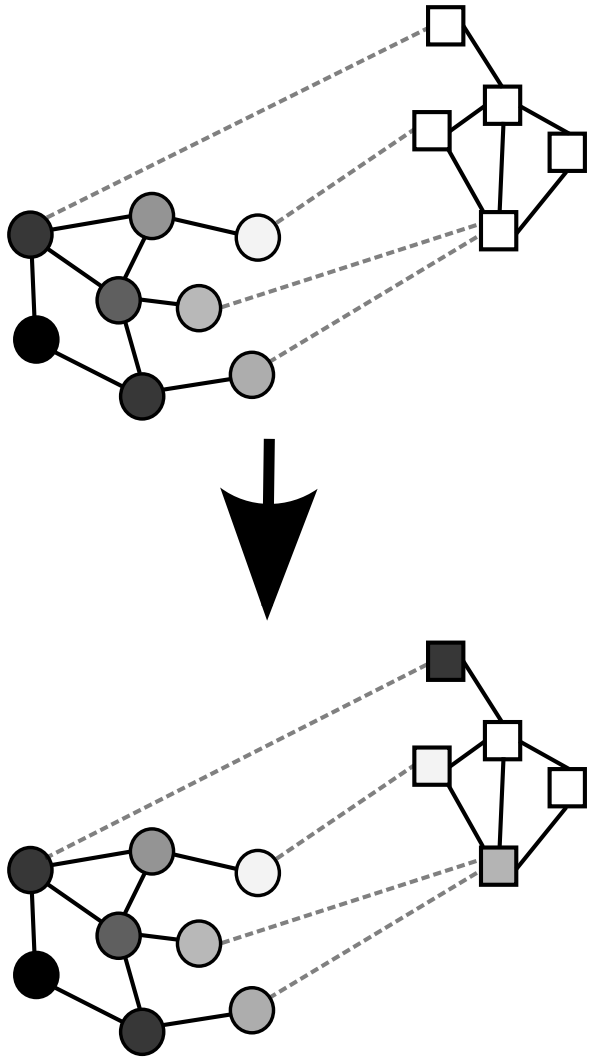
- Nodes have **variable values** which will be updated based on the **degree of relation** with the query or target sets.
- Query and Target set nodes are initially assigned with a value. The rest are assigned to zero.
- Arcs have **constant weights** representing the **strength** of the interaction or the relation.
- Networks are represented as **adjacency matrices**.
- A **normalization step** is performed in order to decrease the influence of each **node degree**.

Propagation operations

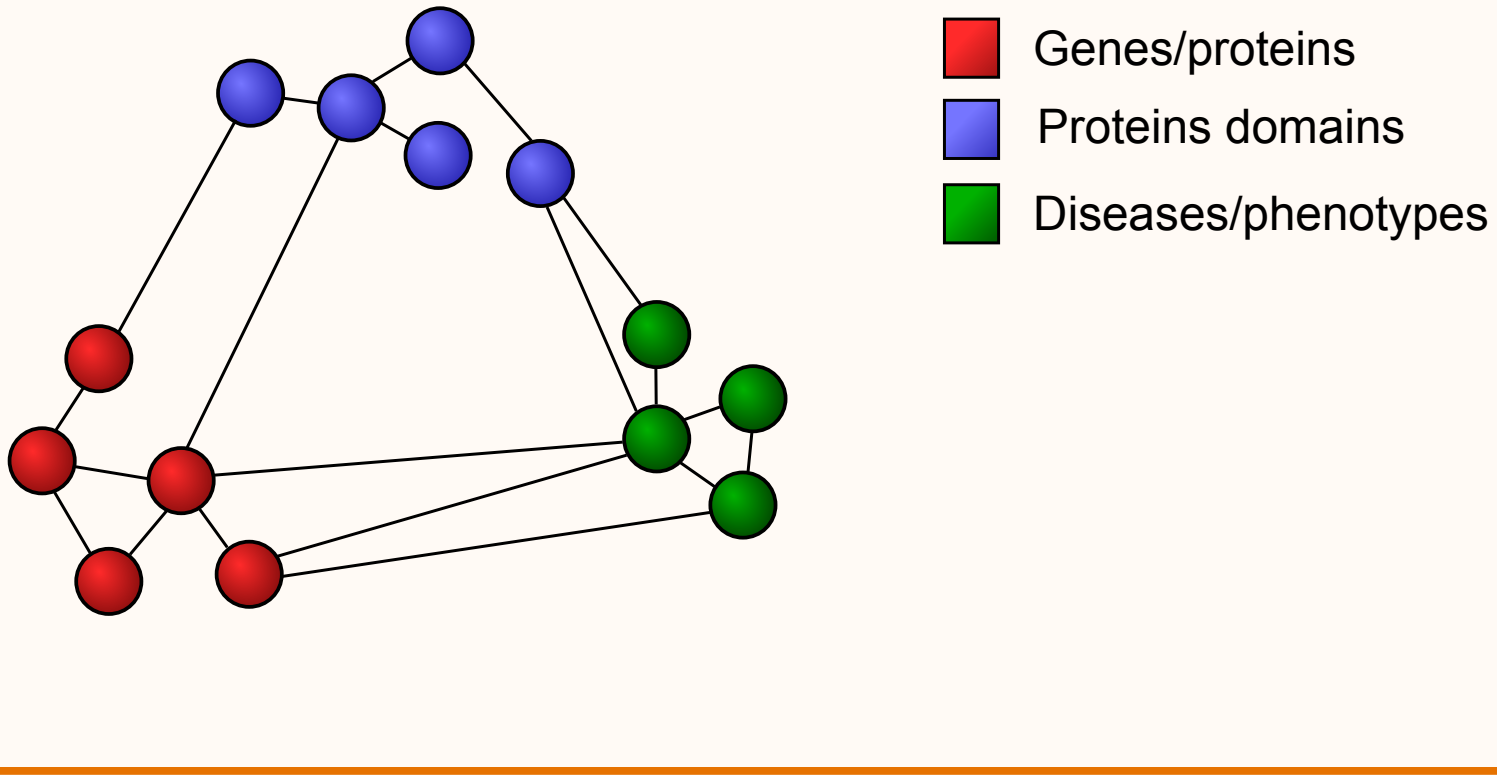
Propagation inside network



Propagation to the next network



Used Global graph



Algorithm

ProphNet algorithm pseudocode

1: Query set values are propagated inside Query Network.
2: paths $\leftarrow$ Compute all paths from Query to Target Net.
3: **for each** node e **in** Target Network **do**
4: Set node e as Target Set and propagate values inside Target Network
5: **for each** path **in** paths **do**
6: **for each** step **in** path **do**
7: Propagate from current network to next network
8: Propagate inside next network
9: **end for**
10: **end for**
11: $S_e \leftarrow$ Correlate paths with Target Network values
12: **end for**
13: Sort S_x decrementally to obtain prioritized list

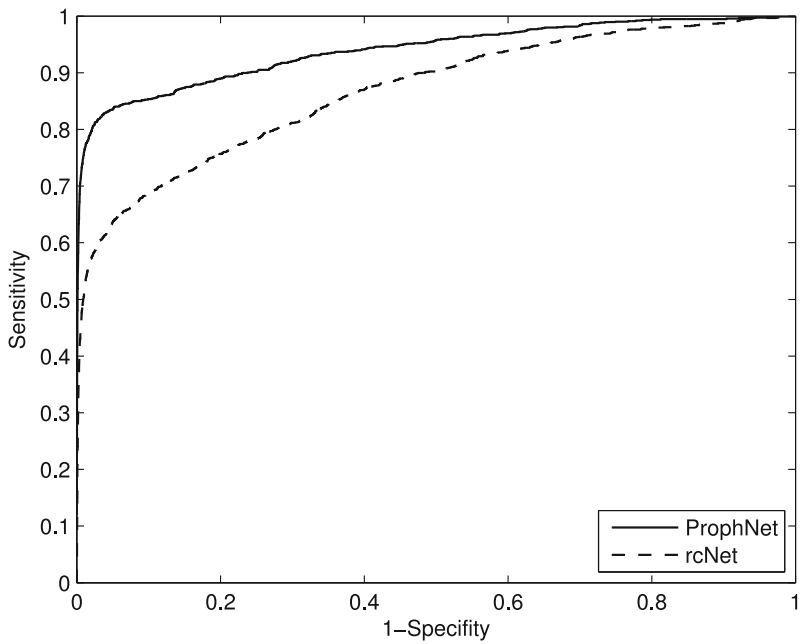
Validation tests

- ProphNet has been applied to **obtain a prioritized list** of genes for some diseases. Top ranked genes were related with query diseases.
- **Three test** have been performed to **validate** ProphNet.
- Against **rcNet**: **LOO gene-disease** prioritization and **new associations** prioritization.
- Against **domainRBF**: **LOO domain-disease**.
- **Leave-one-out (LOO) test:** Remove known A-B association, prioritize using A as Query Set and measure where B is ranked.

Tests against rcNet

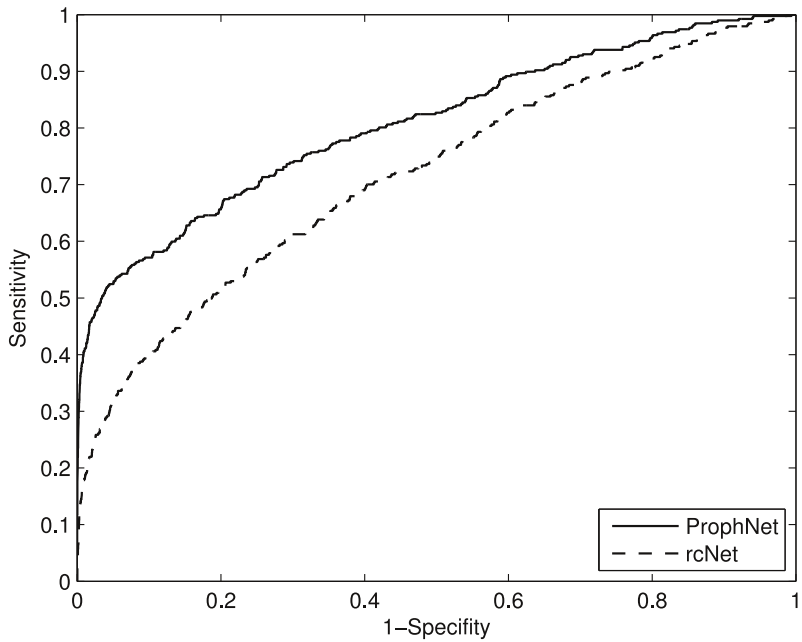
LOO gene-disease validation

16 % AUC gain



New gene-disease relation validation

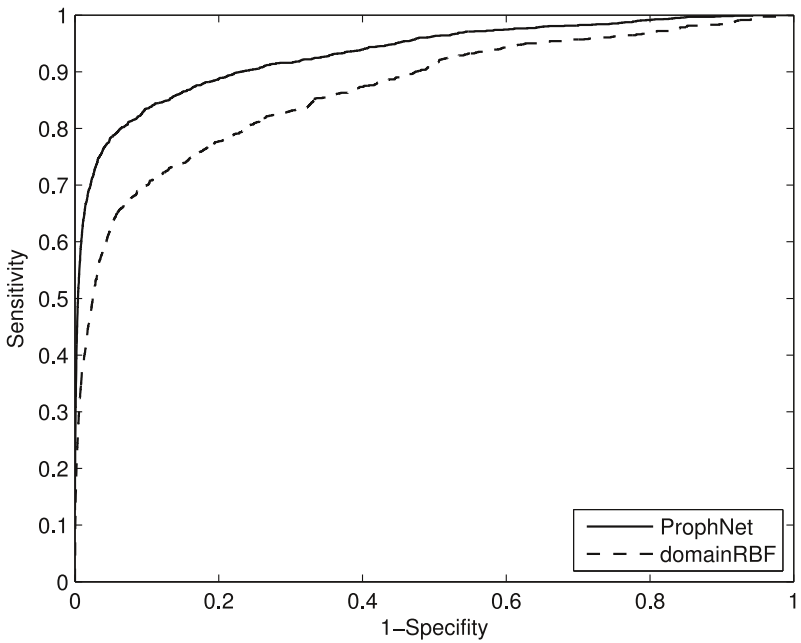
13 % AUC gain



Tests against domainRBF

LOO protein domain-disease validation

8 % AUC gain



Conclusions

- We have proposed a method that **overcomes some limitations** in network-based methods.
- Data integration allows **better performance** than some state-of-the-art methods.
- **Integration** of other type of data has shown an **increase** in ProphNet performance (e.g., drugs or pathways).

Thanks for your attention!

Víctor Martínez Gómez

University of Granada

fvictor@correo.ugr.es