A Machine Learning approach to generate Gene Interaction Networks

Nicola Lazzarini

Supervisors:

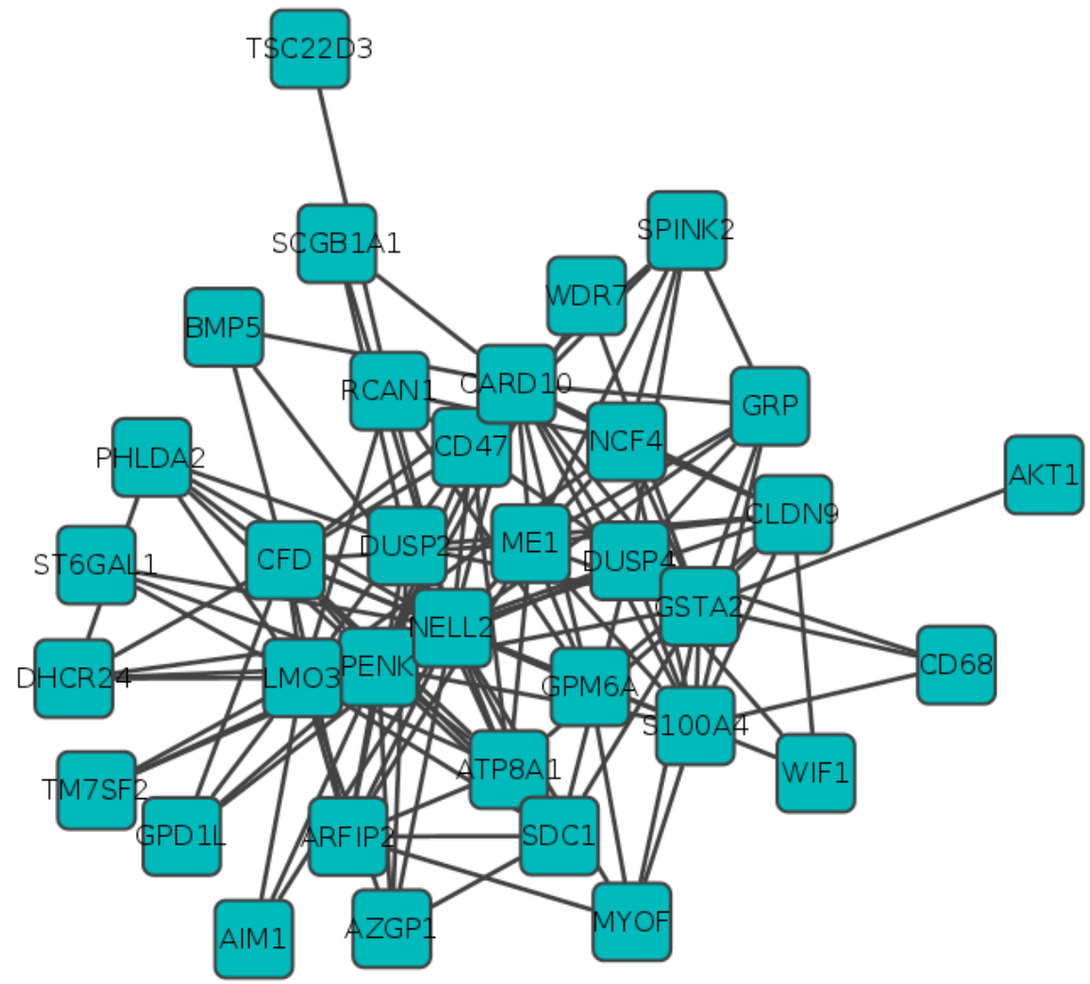
Dr. Jaume Bacardit
Prof. Natalio Krasnogor

Outline

- Gene Interaction Networks
- Co-Prediction
- General Pipeline
- Co-Prediction evaluation
- Co-Expression comparison
- Results
- Conclusions

Gene Interaction Networks

A graph where nodes represents genes and edges indicate a relationship among them.



There exist many methods to infer GIN: Correlation based, Bayesian Networks, ODEs, etc.

ML Approach

Machine Learning approach:

Use the dataset to learn a model (e.g classification) from the existing samples, and afterwards infer gene interactions from the structure of the model

ML techniques used to infer GINs:

- Model Trees

  Inferring gene regression networks with model trees (Nepomuceno-Chamorro et. al)

- Rule Based Systems

  An analysis pipeline with statistical and visualization-guided knowledge discovery for Michigan-style learning classifier systems (Urbanowicz et al.)

- Association Rules

  Inferring gene-gene associations from Quantitative Association Rules (Martinez-Ballestreros et. al)

<u>Our approach</u>: use a rule based classification algorithm to define GINs

Network Generation

BioHEL[1] is a rule based classification algorithm:

rule1: **if** AttributeX > value1 **AND** AttributeY < value2 **then** **Cancer**					
rule2: **if** AttributeX < value3 **AND** AttributeZ < value4 **then** **Normal**					
...	...	...	...	...	...
rule n:	...	...	...	...	...

Core assumption:

Genes present within the same rules, due to their co-operation in classification, might be functionally related

List of edges:

Attribute1	Attribute2	score
Attribute1	Attribute4	score
Attribute1	Attribute5	score
...	....	....

List of nodes:

Attribute1	score
Attribute2	score
Attribute3	score
....	....

1: *Improving the scalability of rule-based evolutionary learning* (Bacardit et al.)

What is Co-Prediction

This way of inferring GINs is termed: **Co-Prediction**
Interaction between genes that act together for class prediction

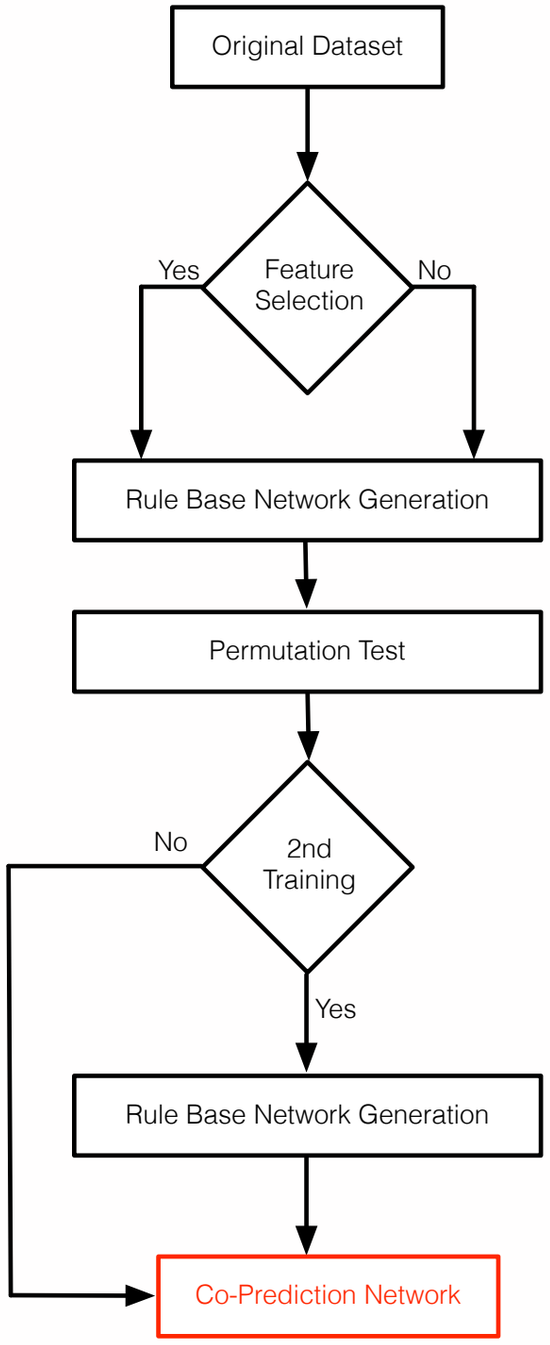
Already used in specific studies:

- *Reduced Neonatal Mortality in Meishan Piglets: A Role for Hepatic Fatty Acids?* (H.P. Fainberg et al.)
- *Contact map prediction using a large-scale ensemble of rule sets and the fusion of multiple predicted structural features* (Bacardit et al.)
- *Functional Network Construction in Arabidopsis Using Rule-Based Machine Learning on Large-Scale Data Sets* (Bassell et al.)

We want to generalise this analysis:

- Define a general pipeline to create Co-Prediction networks
- Test on a set of data (8 datasets)
- Compare with other GINs inferring methods (Co-Expression)

General Pipeline



There are different ways of create the Co-Prediction Network

2 main **switches** allow different variants of the inferring process

<u>4 different variants/configurations</u>

Feature Selection

Remove irrelevant features that might penalize the classification

Selected method: SVM-RFE (Recursive Feature Elimination)

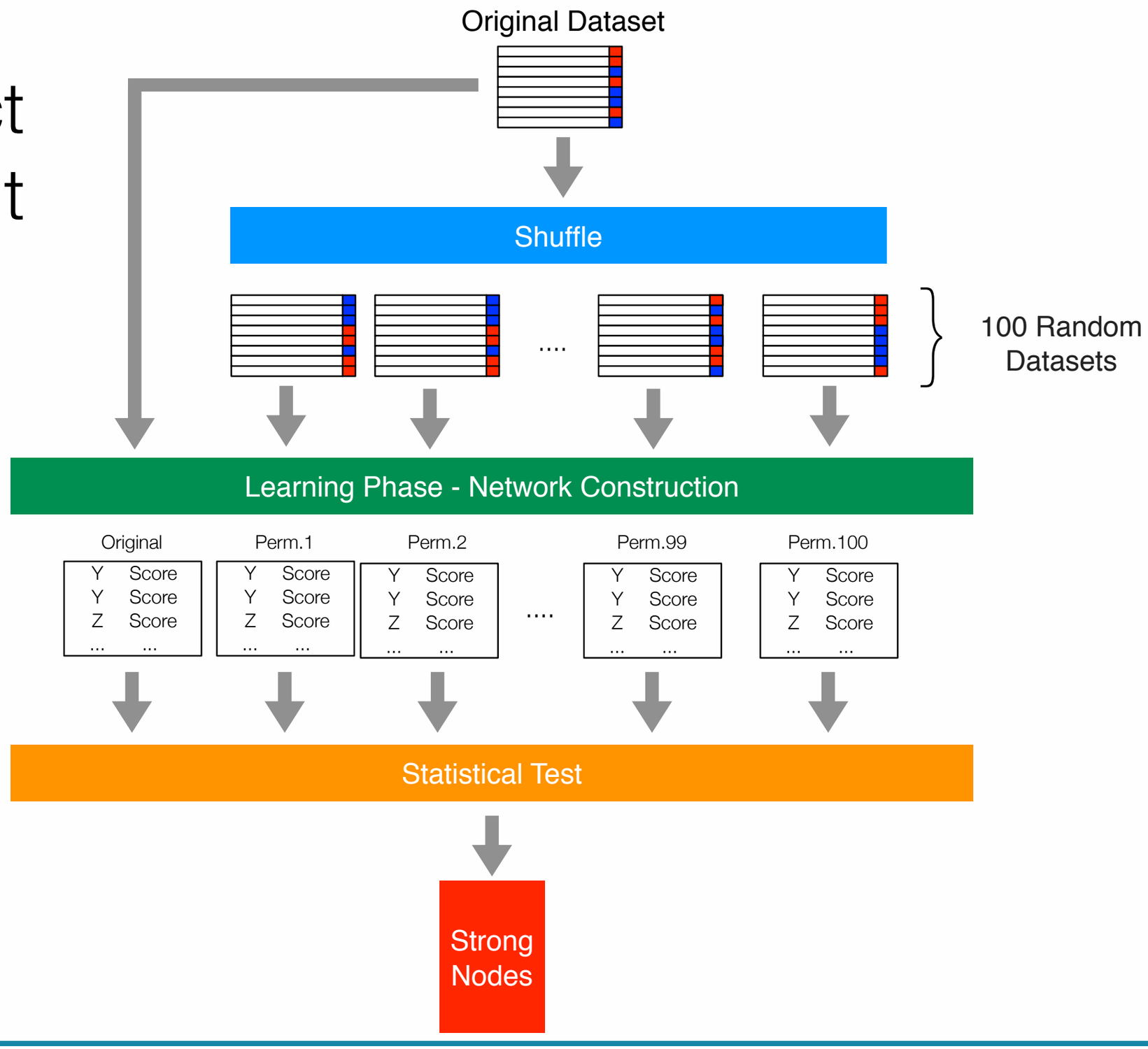
Gene Selection for Cancer Classification using Support Vector Machines (Guyon et al.)

Removes features based on SVM model

Only 10% of features are selected from the original set

Permutation Test

We want to select only the most **important** nodes



Permutation Test

Two ways to use strong nodes:

1) Filter the list of edges: keep only the interactions where at least one node is “strong” (consider their neighbours)

2) Further Feature Selection method: consider the “strong” nodes (and their neighbours) for a second rule based network inferring process

Configurations

Final 4 Configurations used:

Configuration	Description
Conf1	Feature Selection + 1 Training Phase
Conf3	Original Dataset + 1 Training Phase
Conf5	Feature Selection + 2 Training Phases
Conf7	Original Dataset + 2 Training Phases

Conf2, Conf4, Conf6, Conf8 discarded by previous tests

Datasets

Cohort of 8 human cancer related datasets

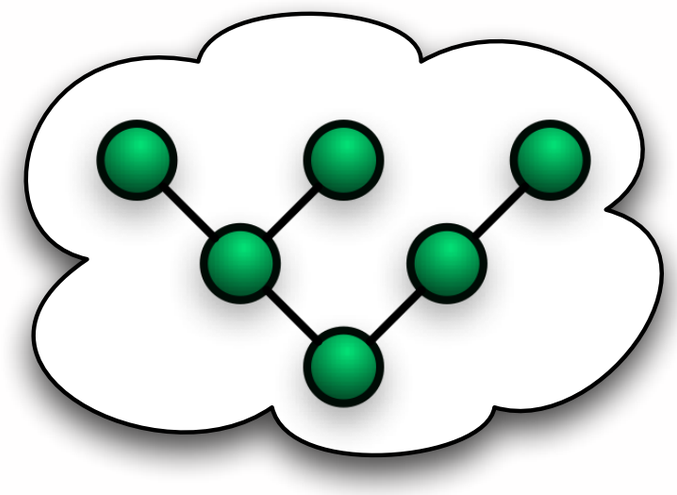
Name	Attributes	Samples
Leukemia	7129	72
Lung Harvard	12534	181
Lung Michigan	7129	96
CNS	7129	60
dlbcl	2647	77
GSE2191	12625	54
GSE3726	22283	52
ProstateML	12600	136

Topological Analysis

Topology: layout of a network, is the arrangement of its various elements

Five different parameters:

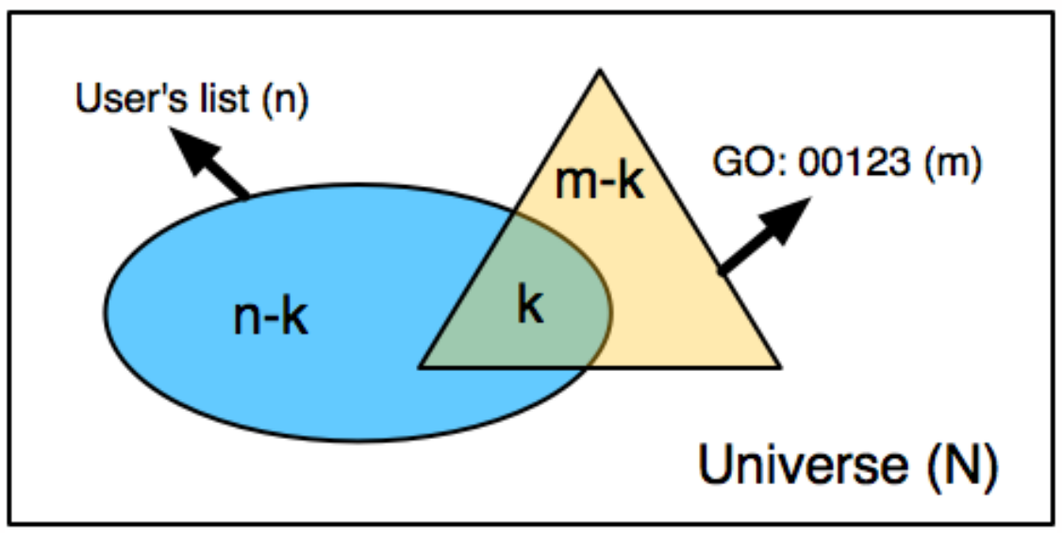
- Clustering Coefficients (Triangle & Square)
- Density
- Diameter
- Shortest Path Length



Enrichment Analysis

Biological terms overrepresented in a gene set

Selected categories: Gene Ontology (GO), KEGG pathways and PubMed



	User's list	Universe
GO: 000123	k	m
not GO: 000123	n-k	N-m

$$P(X=k)=\frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

Enrichment Analysis done using DAVID web tool

Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources (Huang et al.)

Co-Expression Comparison

Pearson Correlation Coefficient (PCC)

$$r_{x,y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \times \sqrt{\sum_i (y_i - \bar{y})^2}}$$

Ranges from -1 to +1

Measures how two genes tend to respond in the same direction across different samples

Co-Expression Comparison

Given a Co-Prediction network with ***N*** nodes and ***E*** edges we build 2 type of Co-Expression networks:

- **SE type:** with *E* edges
- **SN type:** with *N* nodes

Considering the edges with highest PCC

This is done for every Configuration

Topology Analysis

Dataset: Leukemia

Type: Co-Prediction

Configuration: Conf1

Nodes: 421

Edges: 1529

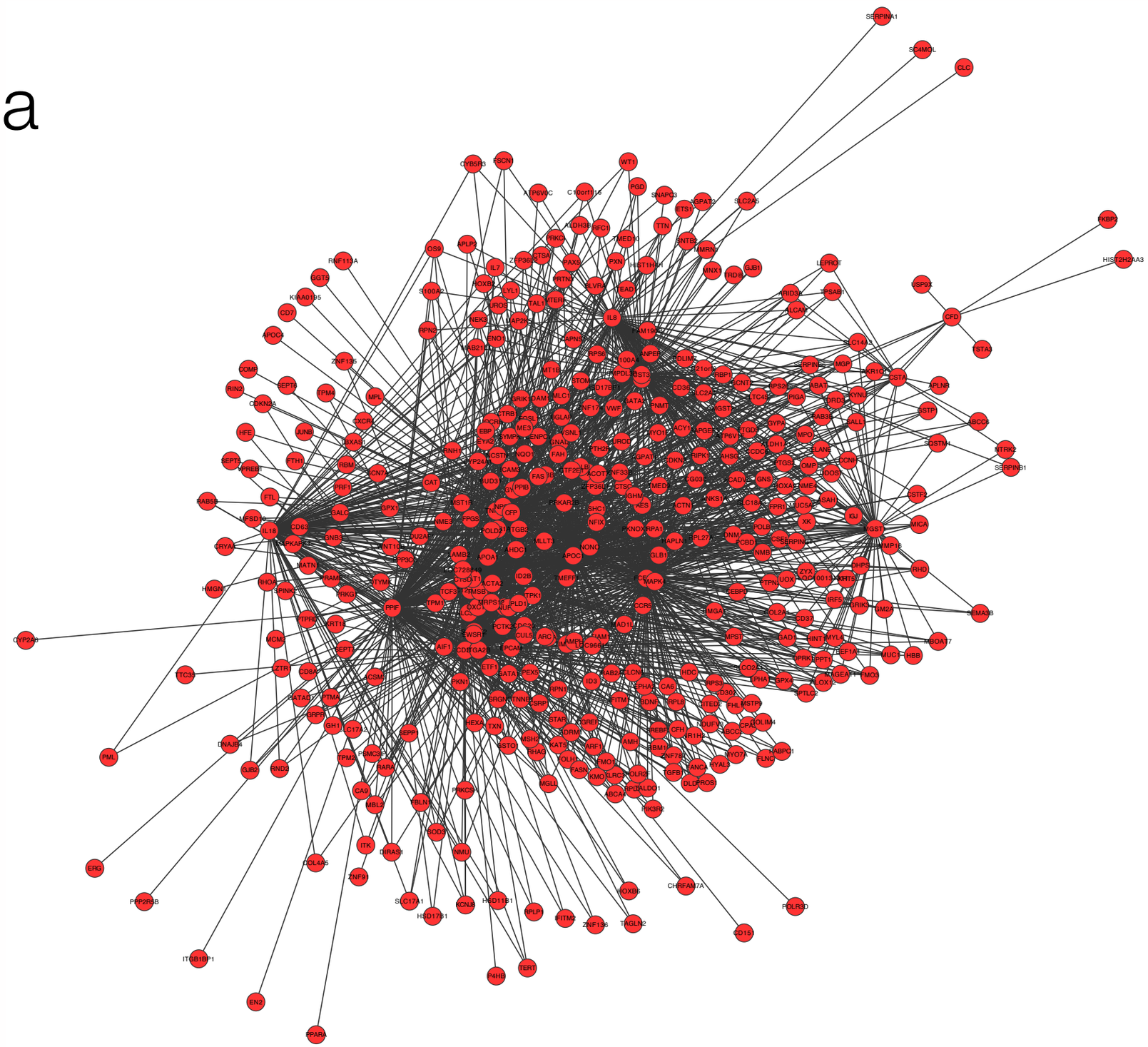
Triangle Coeff: 0.712

Square Coeff: 0.013

Density: 0.017

Diameter: 5

Avg. SPL: 2.284



1 Connected Component

Topology Analysis

Dataset: Leukemia

Type: Co-Expression

Configuration: Conf1_SE

Nodes: 683

Edges: 1529

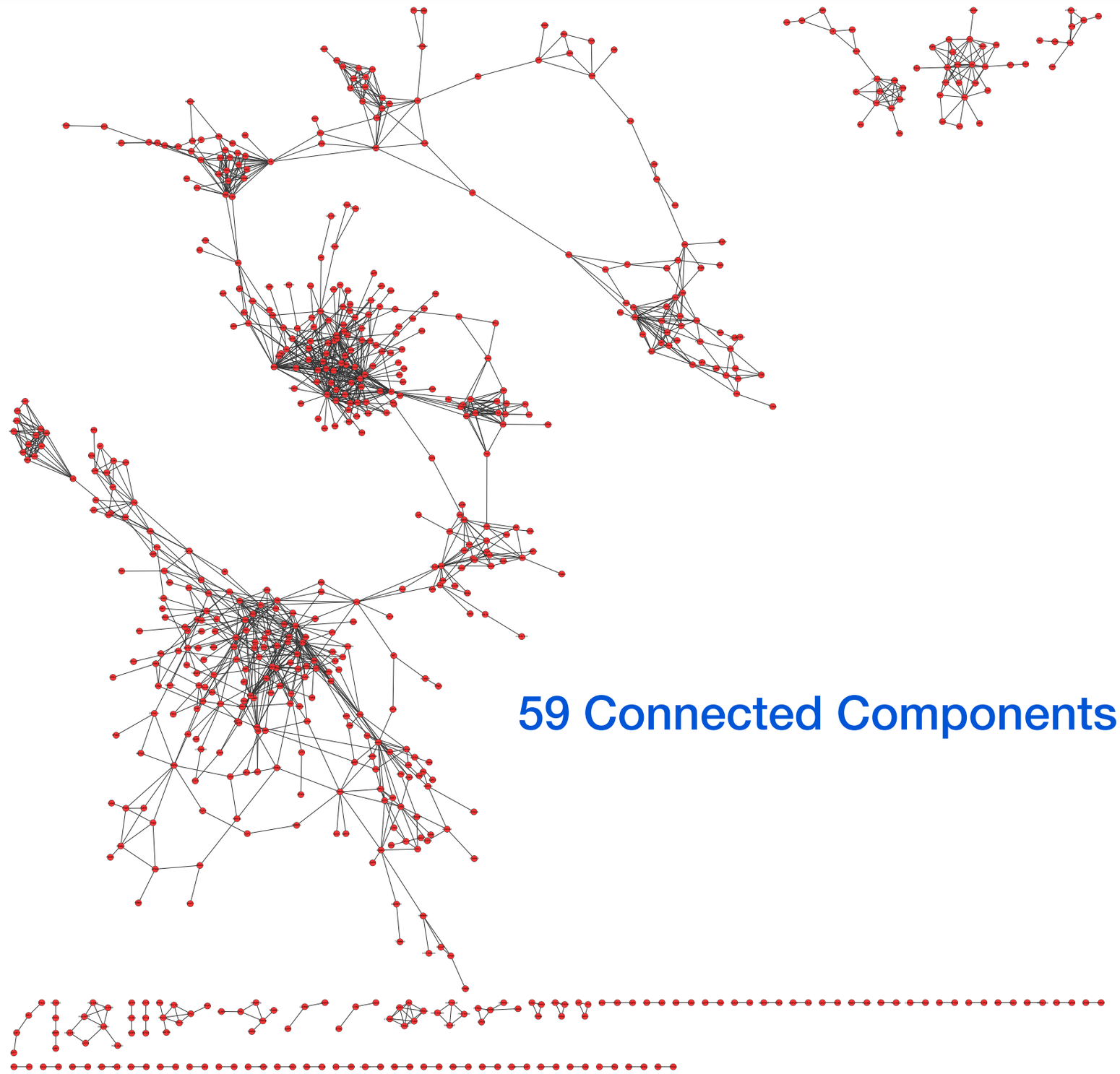
Triangle Coeff: 0.333

Square Coeff: 0.14

Density: 0.007

Diameter: 24

Avg. SPL: 8.856



59 Connected Components

Topology Analysis

Dataset: Leukemia

Type: Co-Expression

Configuration: Conf1_SN

Nodes: 422

Edges: 680

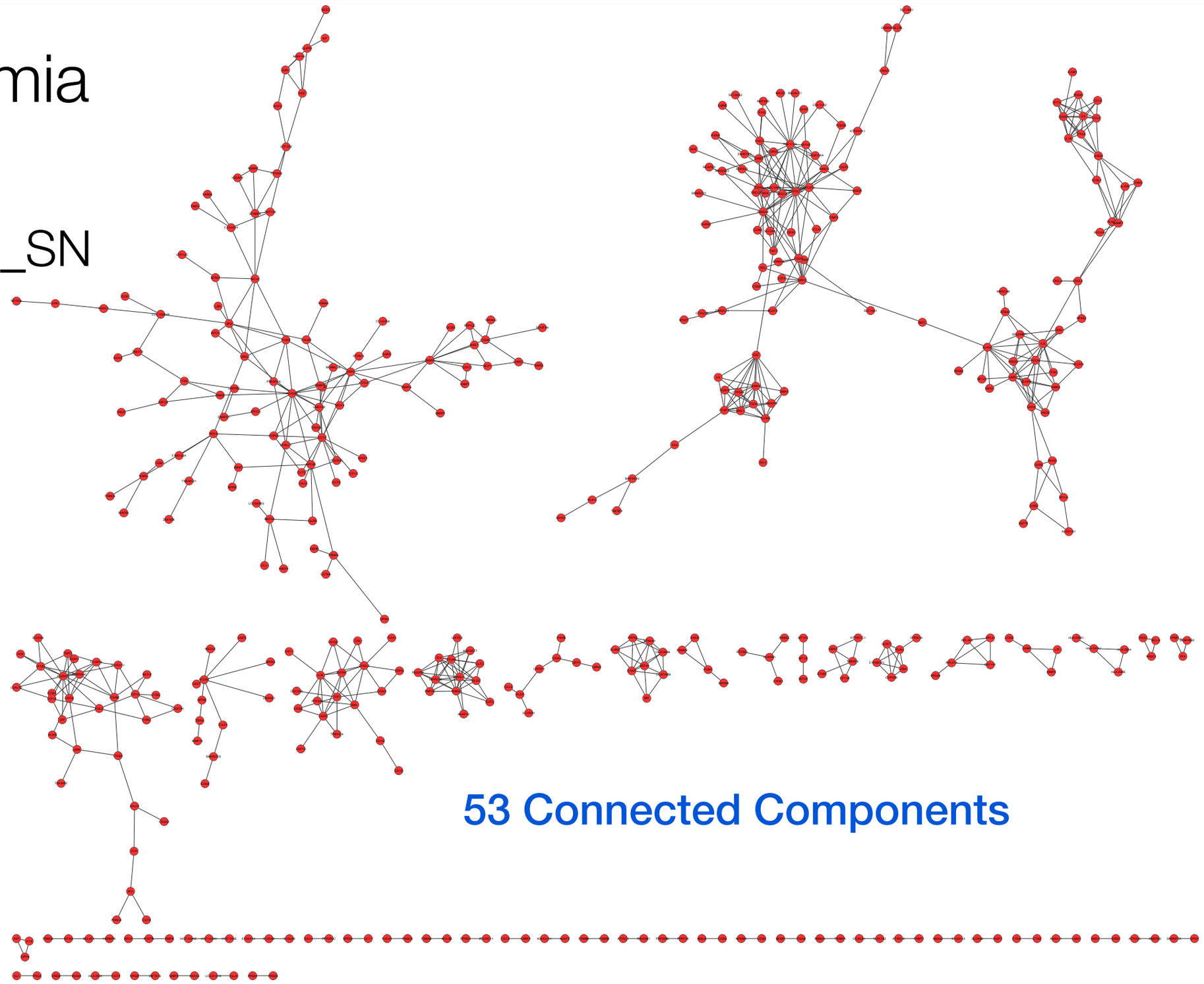
Triangle Coeff: 0.323

Square Coeff: 0.199

Density: 0.008

Diameter: 16

Avg. SPL: 5.644

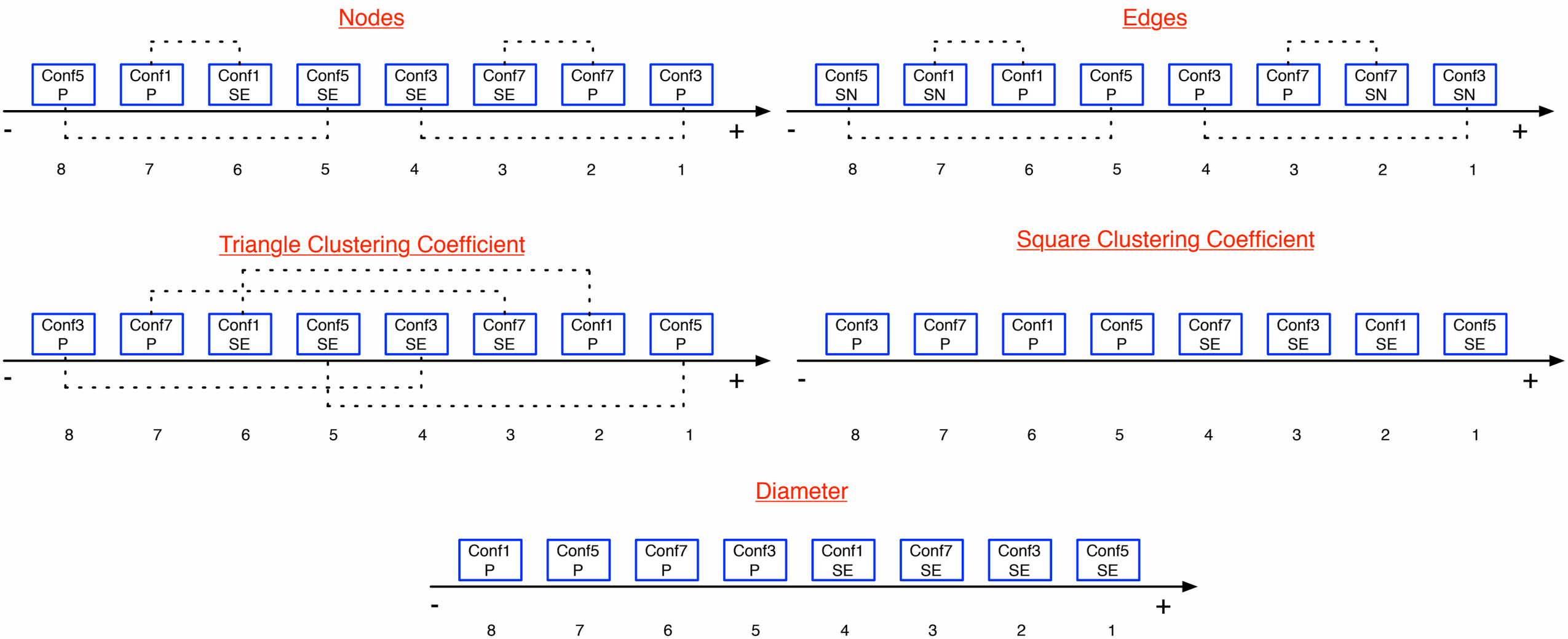


53 Connected Components

Topology Analysis

Co-Prediction vs. Co-Expression

Configuration	Description
Conf1	F.S + 1 Train.
Conf3	Orig + 1 Train
Conf5	F.S + 2 Train.
Conf7	Orig + 2 Train.



Co-Prediction configurations diversity

Overlap of Enriched terms among different configurations

$$O(Conf1,Conf2) = \frac{Commons}{Commons + Unique1 + Unique2}$$

GO	Conf1	Conf3	Conf5	Conf7
Conf1	-	0.166	0.658	0.190
Conf3		-	0.156	0.571
Conf5			-	0.179
Conf7				-

KEGG	Conf1	Conf3	Conf5	Conf7
Conf1	-	0.238	0.690	0.124
Conf3		-	0.127	0.508
Conf5			-	0.139
Conf7				-

Co-Prediction configurations diversity

PubMed	Conf1	Conf3	Conf5	Conf7
Conf1	-	0.065	0.720	0.080
Conf3		-	0.062	0.423
Conf5			-	0.077
Conf7				-

Generally small overlap

Similar configurations (1,5) and (3,7) share more

Variants on the inferring process lead to different networks

Co-Prediction vs. Co-Expression biological knowledge

Enrichment Score to have unbiased results

$$ES = \frac{\text{\# of Enriched Terms}}{\text{\# of Nodes}}$$

Average Ranking across 8 datasets

Co-Expression SE type

	Conf1_P	Conf1_ESE	Conf3_P	Conf3_ESE	Conf5_P	Conf5_ESE	Conf7_P	Conf7_ESE
GO	3.000 (2.5)	3.000 (2.5)	6.562 (8)	5.250 (6)	**2.250 (1)**	4.625 (4)	5.062 (5)	6.250 (7)
KEGG	4.437 (6)	**3.125 (1)**	6.687 (8)	4.312 (5)	3.625 (2)	4.125 (3)	4.250 (4)	5.437 (7)
PubMed	3.750 (4)	**3.375 (1)**	5.750 (7)	3.500 (2)	3.625 (3)	4.375 (5)	6.250 (8)	5.375 (6)

Co-Expression SN type

	Conf1_P	Conf1_ESN	Conf3_P	Conf3_ESN	Conf5_P	Conf5_ESN	Conf7_P	Conf7_ESN
GO	3.00 (3)	2.750 (2)	6.187 (7)	7.625 (8)	**2.000 (1)**	3.375 (4)	4.943 (5)	6.125 (6)
KEGG	4.500 (5)	3.625 (3.5)	6.562 (8)	5.937 (7)	3.625 (3.5)	**3.250 (1)**	3.562 (2)	4.937 (6)
PubMed	4.000 (4)	**3.125 (1)**	5.875 (8)	5.000 (5)	3.375 (2)	3.562 (3)	5.75 (7)	5.125 (6)

Overlap of Enriched terms: Co-Prediction vs. Co-Expression

Co-Expression SE type

	Conf1	Conf3	Conf5	Conf7
GO	0.155	0.325	0.166	0.294
KEGG	0.068	0.256	0.092	0.215
PubMed	0.045	0.120	0.043	0.147

Co-Expression SN type

	Conf1	Conf3	Conf5	Conf7
GO	0.149	0.469	0.144	0.371
KEGG	0.049	0.421	0.068	0.300
PubMed	0.048	0.270	0.047	0.193

Statistical test for performance evaluation

Is one method better than the other?

Consider each configuration as independent algorithm

Friedman test

Null-hypothesis: the performances of the algorithms are equivalent

Post-Hoc with Nemenyi test

Compare each algorithm vs. each other to check if one is better

Enrichment Score used as metric

Significance level α=0.05

Statistical test for performance evaluation

GO Terms

- Null-hypothesis rejected with p-value $5.322e^{-07}$ for SN type

Conf5_P better than: Conf3_P, Conf7_SN, Conf3_SN
Conf1_P better than: Conf3_SN
Conf1_SN and Conf5_SN better than: Conf3_SN

- Null-hypothesis rejected with p-value 0.001012 for SE type

Conf5_P better than: Conf3_P, Conf7_SE

KEGG and PubMed Terms

Null-hypothesis is not rejected

Literature Mining

Manual literature mining for “important” nodes:

- Top Hubs (highest degree)
- Central Nodes (highest betweenness centrality)

Those important nodes appear to play an important role on the disease studied in the microarray

Example: GSE3726 (Colon vs. Breast Cancer)

6 out 7 “important” genes involved in one of those 2 cancers.

Conclusions

We presented a ML approach to define GINs called Co-Prediction.

- It has general pipeline to infer GINs with 4 different variants
- It identifies interactions in a different way compared with statistical methods as Co-Expression
- Compared to Co-Expression networks, Co-Prediction GINs:
  - ✦Encode the same amount of biological knowledge
  - ✦Different in:
    - Topology
    - Enriched Terms
- It creates biologically relevant networks

Thank you!