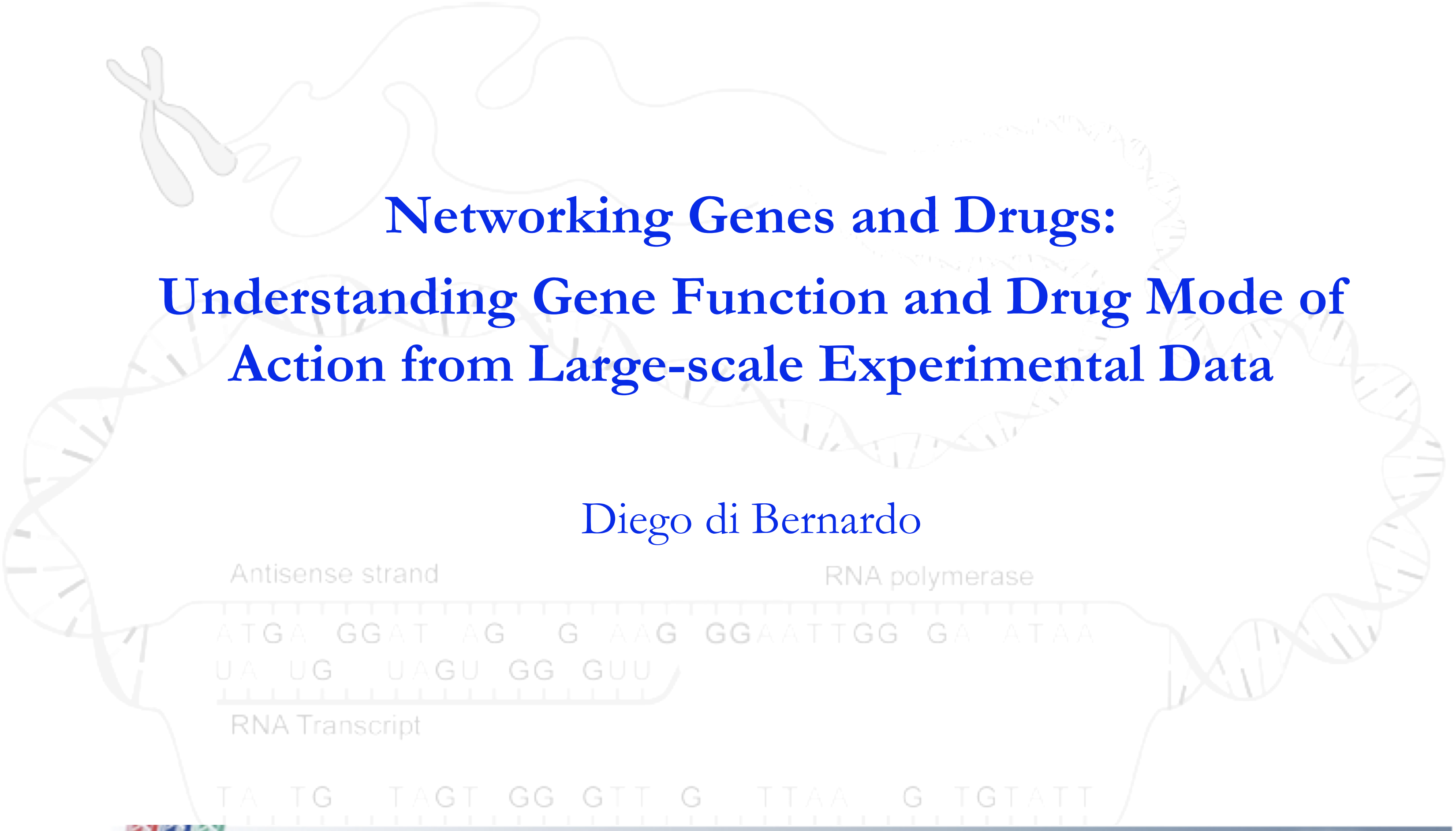
Networking Genes and Drugs:

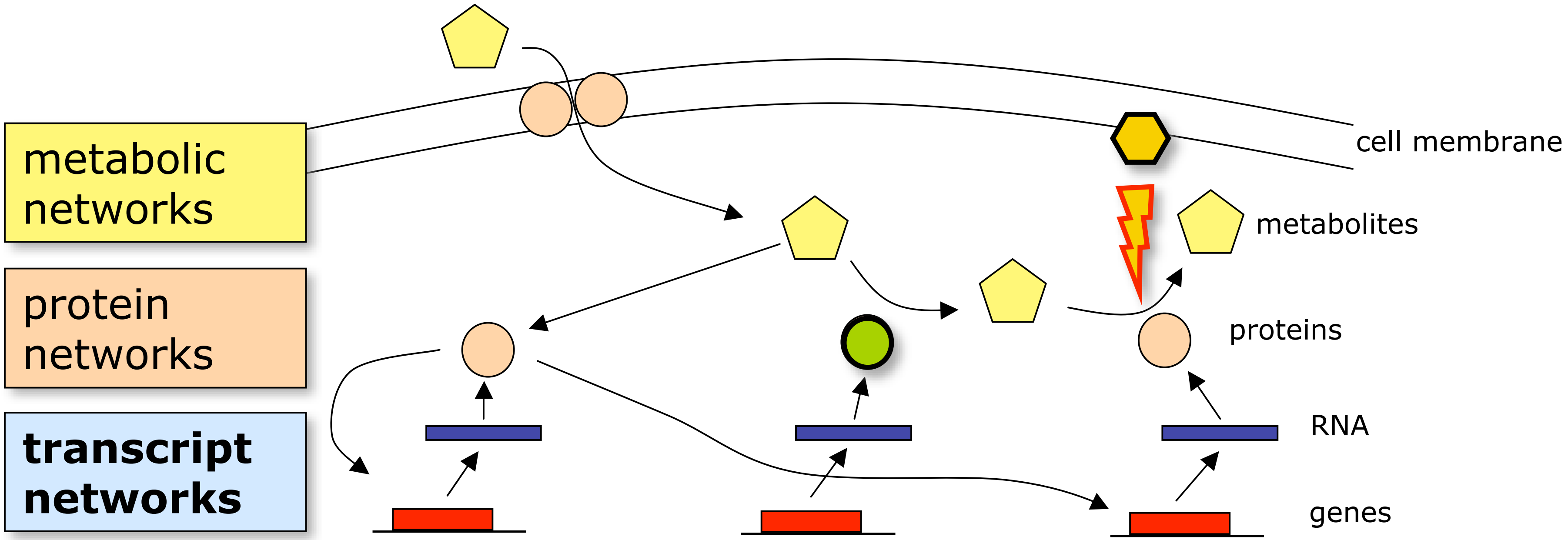
Understanding Gene Function and Drug Mode of Action from Large-scale Experimental Data

Diego di Bernardo


Antisense strand
RNA polymerase
ATGA GGAT AG G AAG GGAATTGG GA ATAA
UA UG UAGU GG GUU
RNA Transcript
TA TG TAGT GG GTT G TTAA G TGTATT

Telethon Institute of Genetics and Medicine

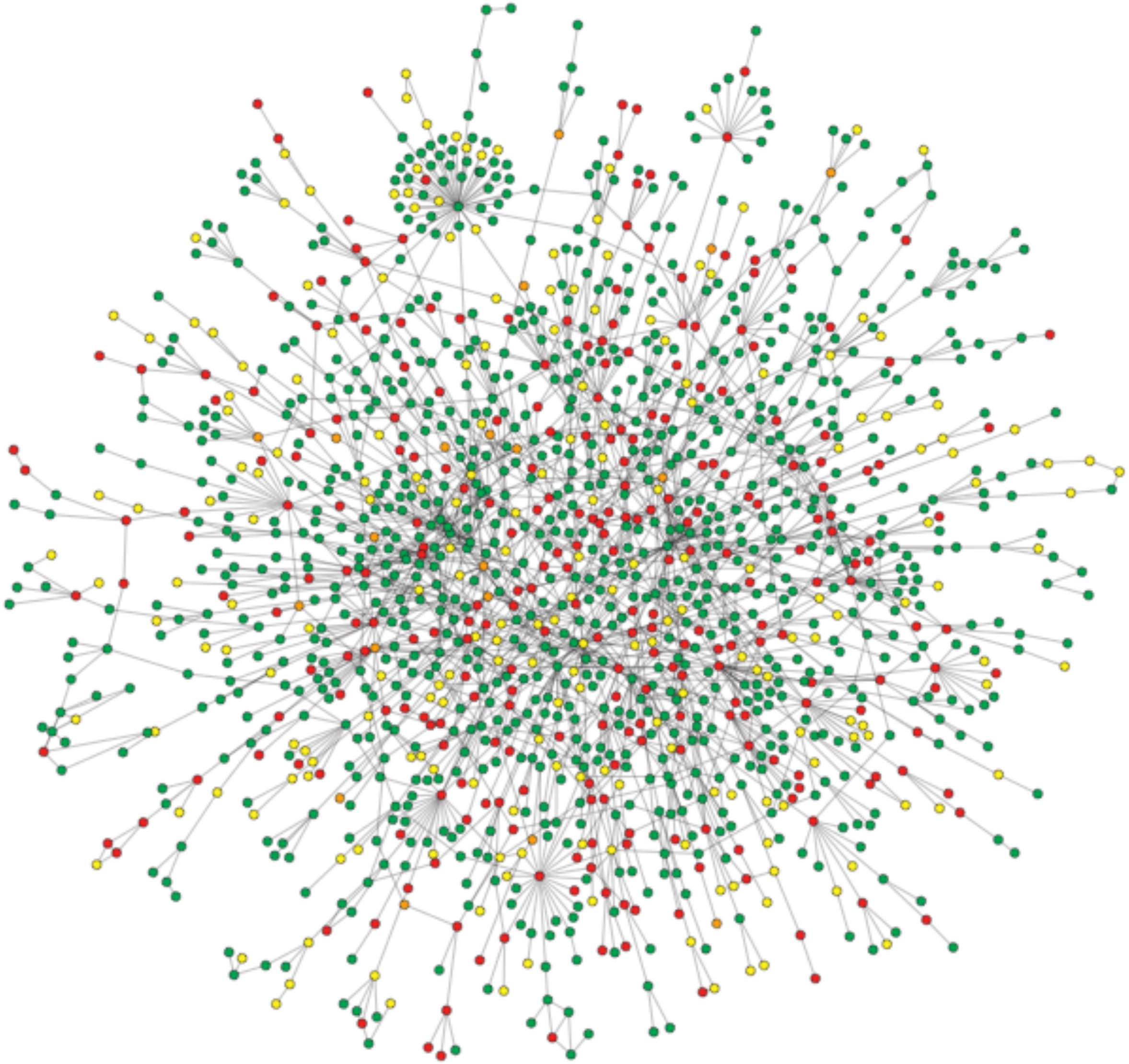
The problems we (and everybody else) are tackling:



● What is the role of my gene?

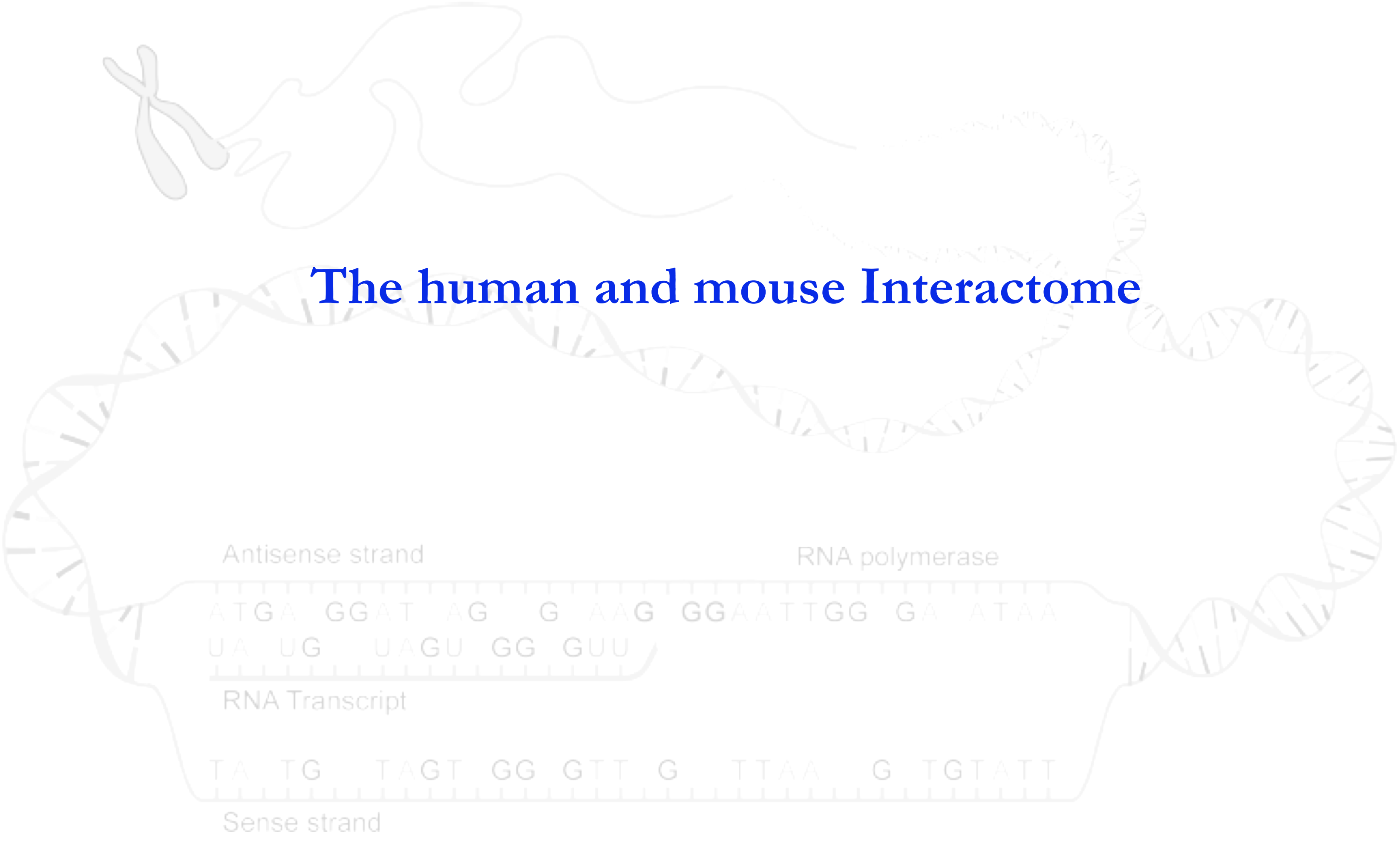
⬢ Which small molecule (drug) can modify the pathway of interest?

A 'simple' protein-protein interaction network (yeast S. cerevisiae)



Part I: Understanding Gene Function

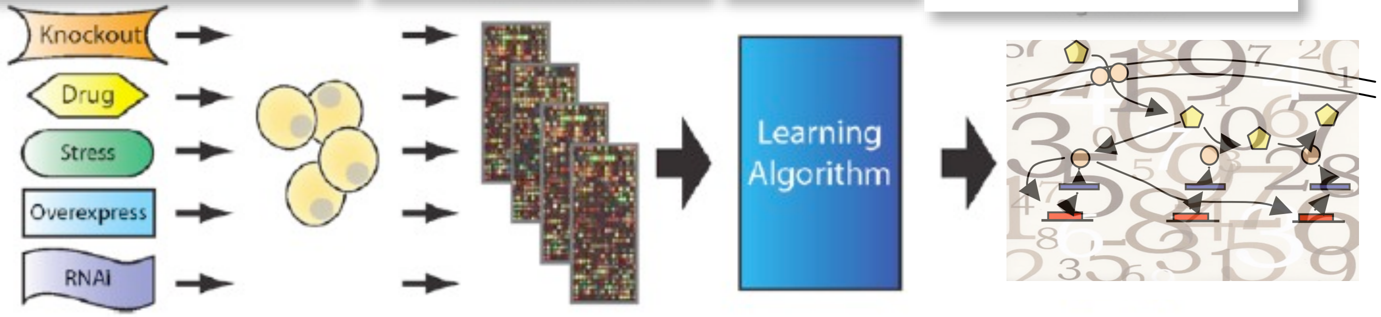
The human and mouse Interactome



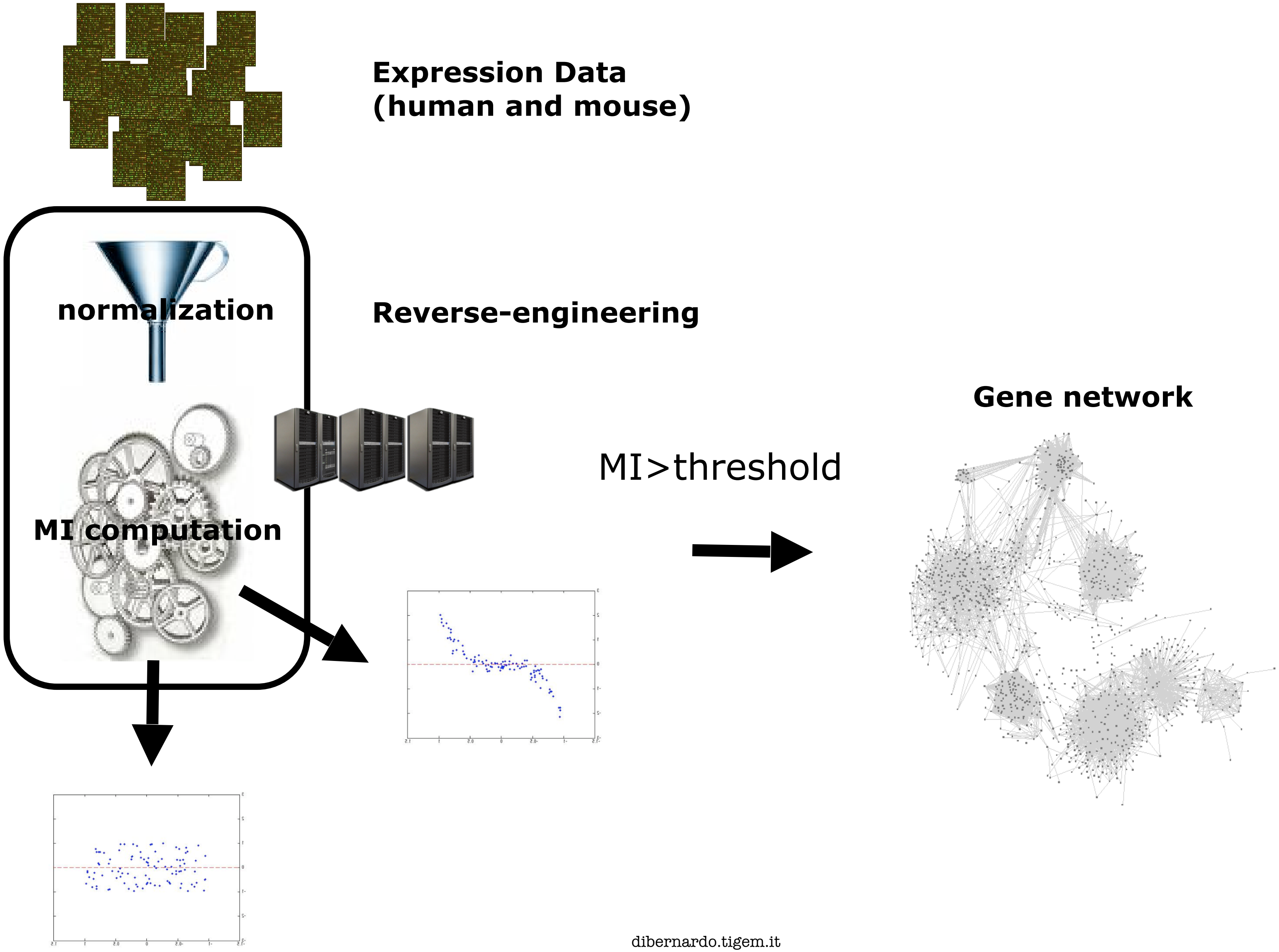
Belcastro V et al, UNPUBLISHED (confidential)

Reverse-engineering: from data to model

In vivo perturbations | Measure response | Learn | Computer model



Reverse engineering human and mouse gene networks:



Expression Data
(human and mouse)

normalization

Reverse-engineering

MI computation

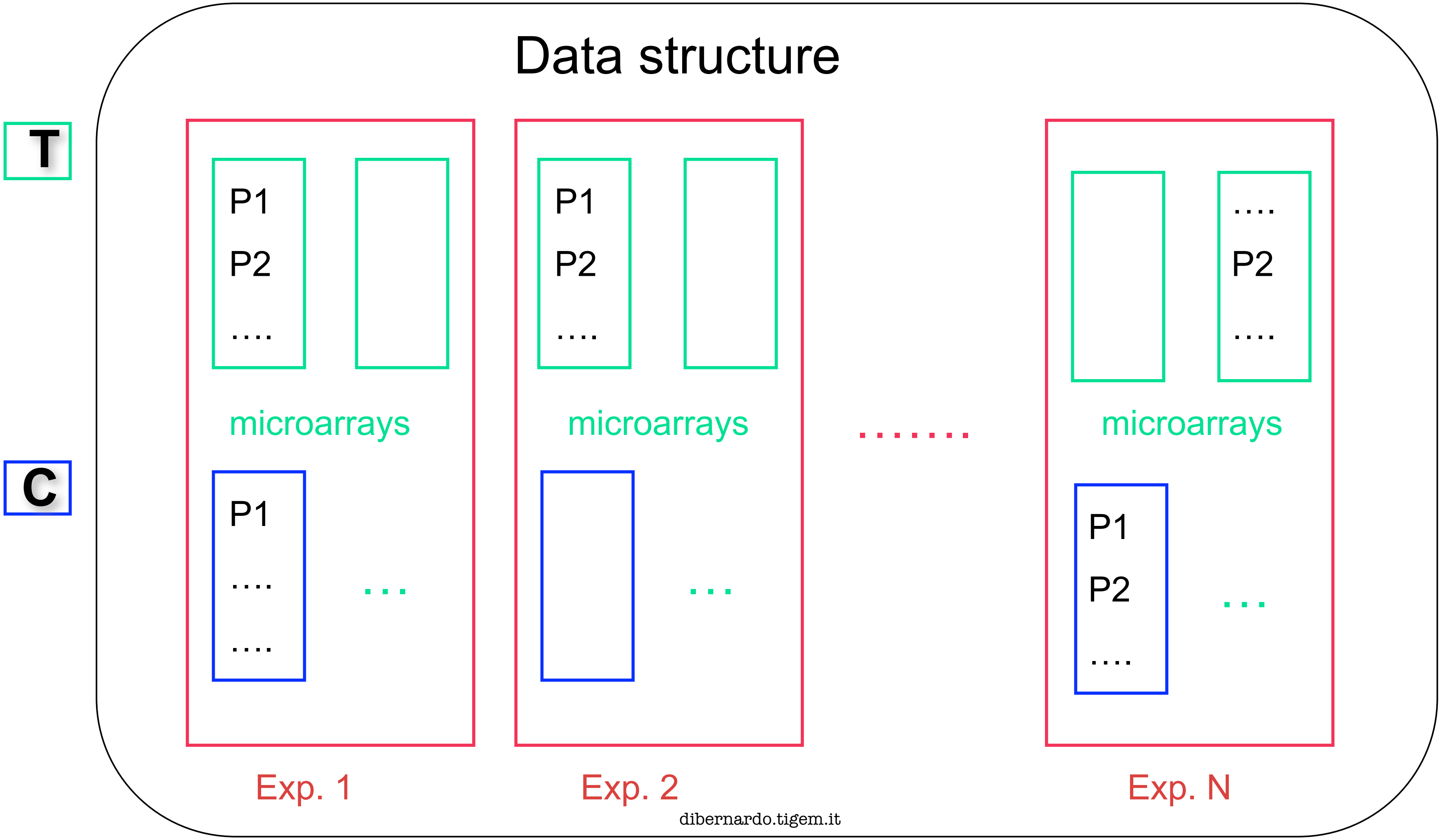
MI>threshold

Gene network

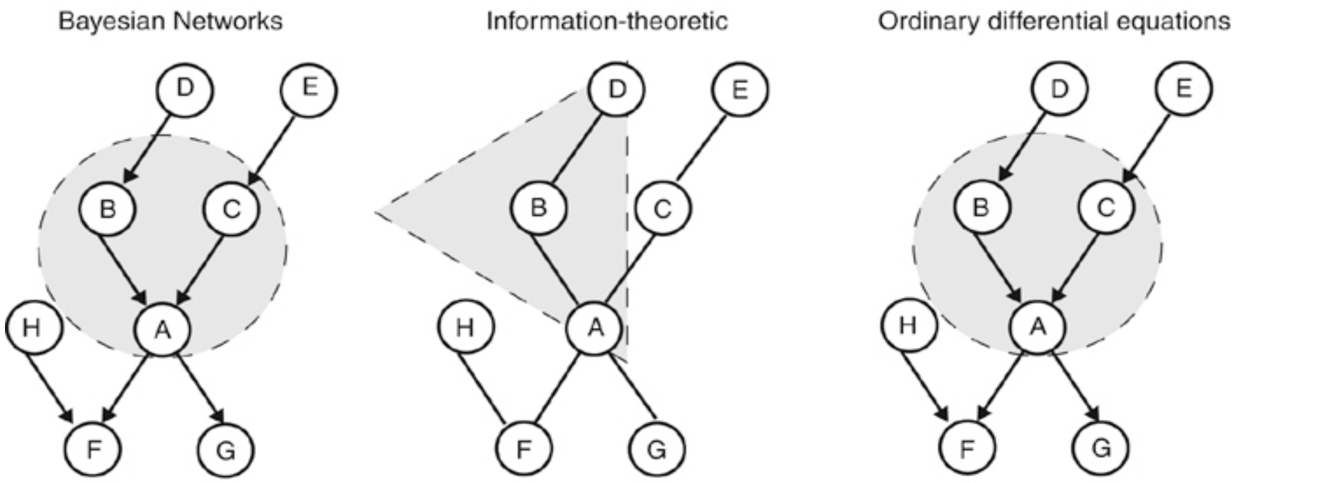
Expression Data:

HUMAN (HG-U133A)
702 experiments (**20255** hyb.)
22283 probesets (P)
14340 genes

MOUSE (Mouse430_2)
797 experiments (**8895** hyb.)
45101 probesets (P)
28219 genes



Reverse-engineering:



Bayesian Networks

Information-theoretic

Ordinary differential equations

$P(A/B,C,D,E)=P(A/B,C)$

$MI(A,H)=0$
$MI(A,B)>0$
$0<MI(A,D)\leq\min\{MI(A,B),\ MI(B,D)\}$

$dA/dt=\theta_1A+\theta_2B+\theta_3C$

or more generally:

$dA/dt=f(A,B,C,\theta)$

BANJO

(Hartemink, A. Nature Biotechnology, 2005.)

ARACNE

(Basso et al., Nature Genetics, 2006)

NIR and TSNI *(Gardner, et al, Science, 2003; Bansal et al, Bioinformatics, 2006; Della Gatta et al, Genome Research, 2008)*

DYNAMIC AND

STEADY-STATE (n-way)

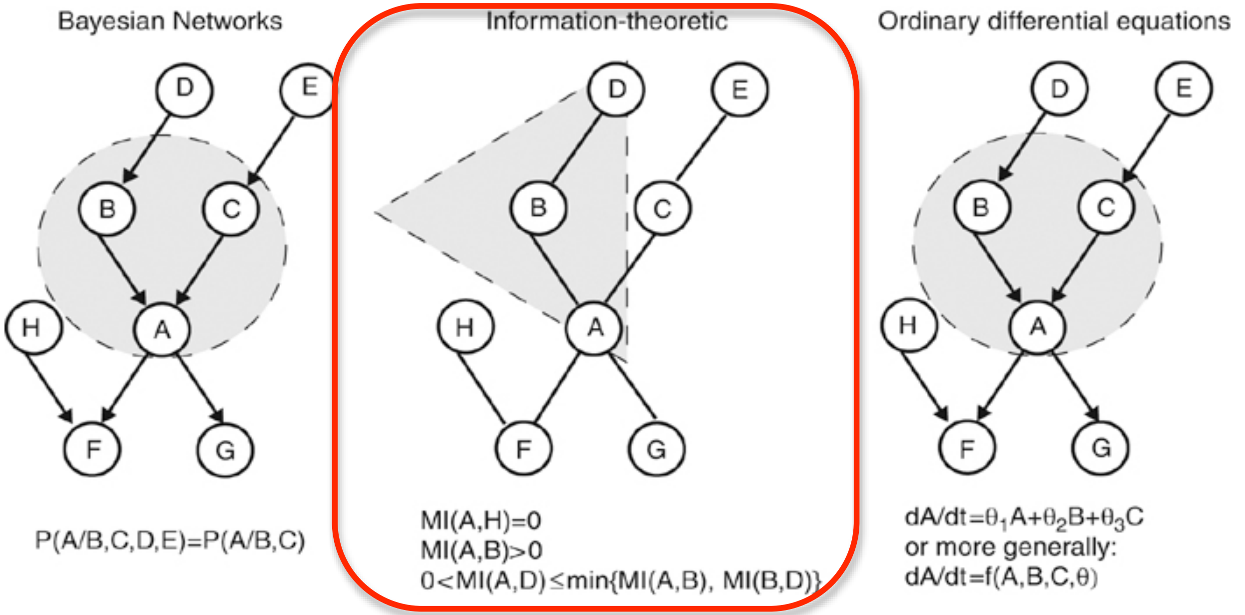
STEADY-STATE

(2-way)

DYNAMIC AND

STEADY-STATE (n-way)

Reverse-engineering:



$P(A/B,C,D,E)=P(A/B,C)$

$MI(A,H)=0$
$MI(A,B)>0$
$0<MI(A,D)\leq\min\{MI(A,B), MI(B,D)\}$

$dA/dt=\theta_1 A+\theta_2 B+\theta_3 C$
or more generally:
$dA/dt=f(A,B,C,\theta)$

BANJO

(Hartemink, A. Nature Biotechnology, 2005.)

DYNAMIC AND

STEADY-STATE (n-way)

ARACNE

(Basso et al., Nature Genetics, 2006)

STEADY-STATE

(2-way)

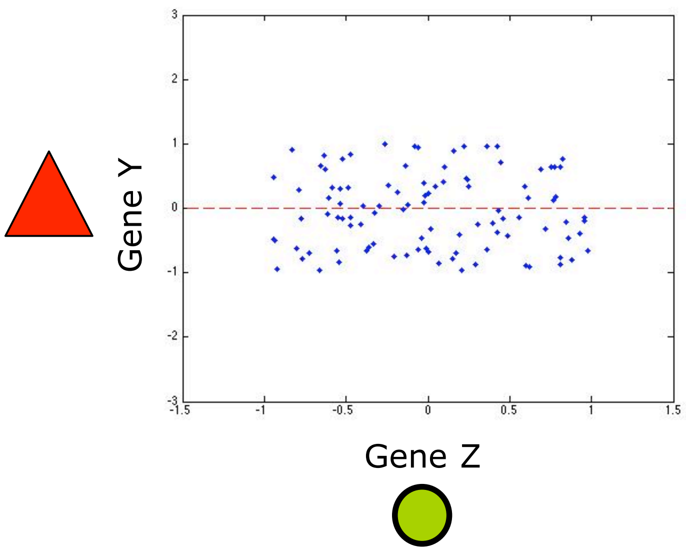
NIR and TSNI *(Gardner, et al, Science, 2003; Bansal et al, Bioinformatics, 2006; Della Gatta et al, Genome Research, 2008)*

DYNAMIC AND

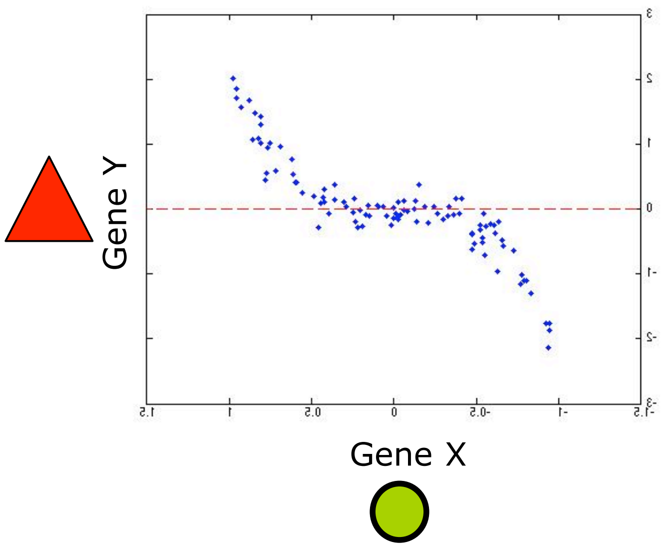
STEADY-STATE (n-way)

Mutual Information:

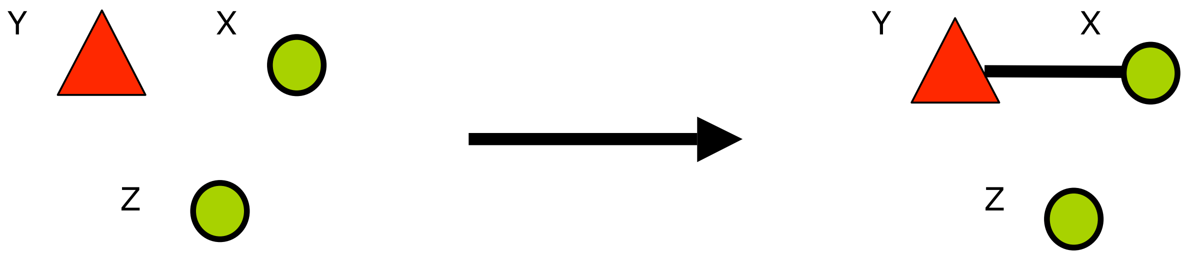
Independent genes



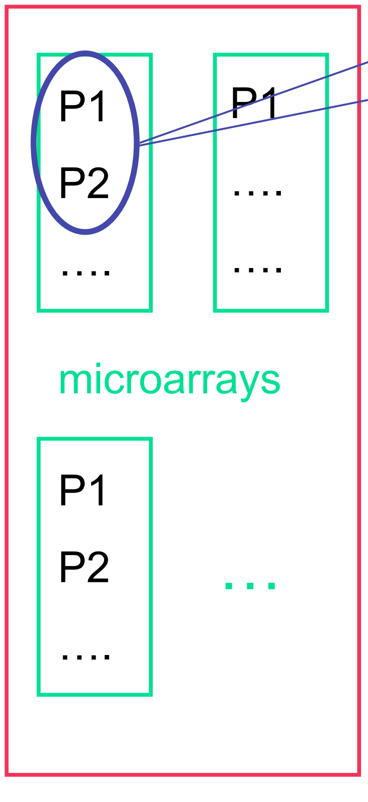
Genes that regulate each other



$$I(X,Y) = \sum_{\substack{x \in \{1..r\} \\ y \in \{1..s\}}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$



Computation of Mutual Information:



Exp. k

| -------------- Differential expression values (T/C)------ |

P1	2.3	2	-1	0.7	2.6	-0.7	-0.7	-0.3	…
P2	5	8.2	1	-0.7	-0.2	0.3	2.3	-0.7	…

▼ Discretization

•(Liu *et al.*,2002) *Data Mining and Knowledge Discovery*, **6**, 393-423.

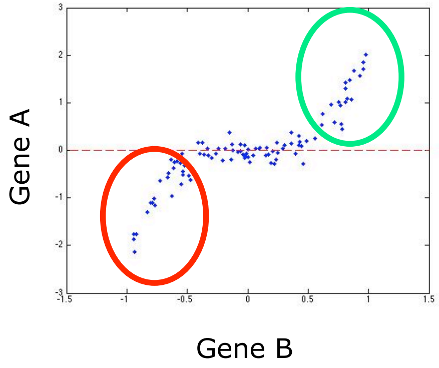
P1	+	+	-	=	+	-	-	-	…
P2	+	+	=	-	-	=	+	-	…

▼

P1\P2	-	=	+
-	1	1	1
=	1	0	1
+	1	0	2

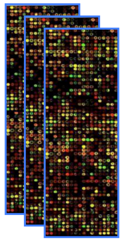
Matrix of Occurrences

(3 bins)



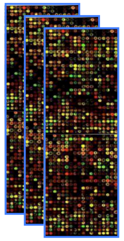
How to obtain one huge datasets? (Dataset merging)

Exp. #1



P1\P2	-	=	+
-	1	2	1
=	1	0	1
+	1	0	1

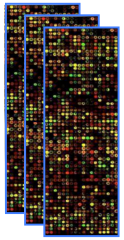
Exp. #2



P1\P2	-	=	+
-	1	2	0
=	2	3	1
+	1	1	0

...

Exp. #N



P1\P2	-	=	+
-	0	2	1
=	1	0	3
+	1	2	1

P1\P2	-	=	+
-	10	12	10
=	8	10	11
+	6	16	7

/ #common microarrays

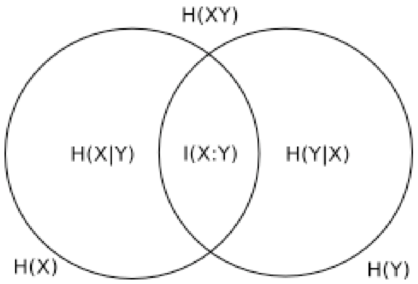
P1\P2	-	=	+
-	10/90	12/90	14/90
=	8/90	20/90	21/90
+	6/90	18/90	7/90

frequencies

Frequentist apporach to MI:

P1\P2	-	=	+
-	10/90	12/90	14/90
=	8/90	20/90	21/90
+	6/90	18/90	7/90

Mutual Information (MI) is the amount of information two random variables share.



MI can be used to measure how dependent two probes are.

$$I(X,Y) = \sum_{\substack{x\in\{1..r\}\\ y\in\{1..s\}}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

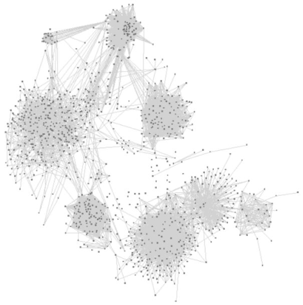
$$\hat{I}(X,Y) = \sum_{\substack{x\in\{1..r\}\\ y\in\{1..s\}}} f(x,y)\log\frac{f(x,y)}{f(x)f(y)}$$

where

$$f(z) = \frac{n_z}{n}$$

$$f(x,y) = \frac{n_{xy}}{n}$$

Network statistics and properties



HUMAN
20255 experiments
22283 probes

Threshold **0.04**
4.817.629 edges

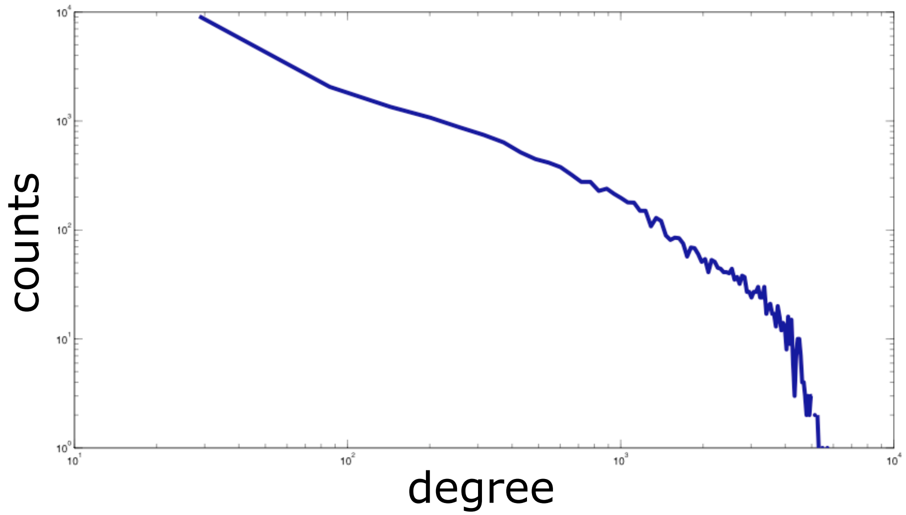
MOUSE (Mouse430_2)
8895 experiments
45101 probes

Threshold **0.025**
14.461.095 edges

Mouse+**H**uman
10415 genes

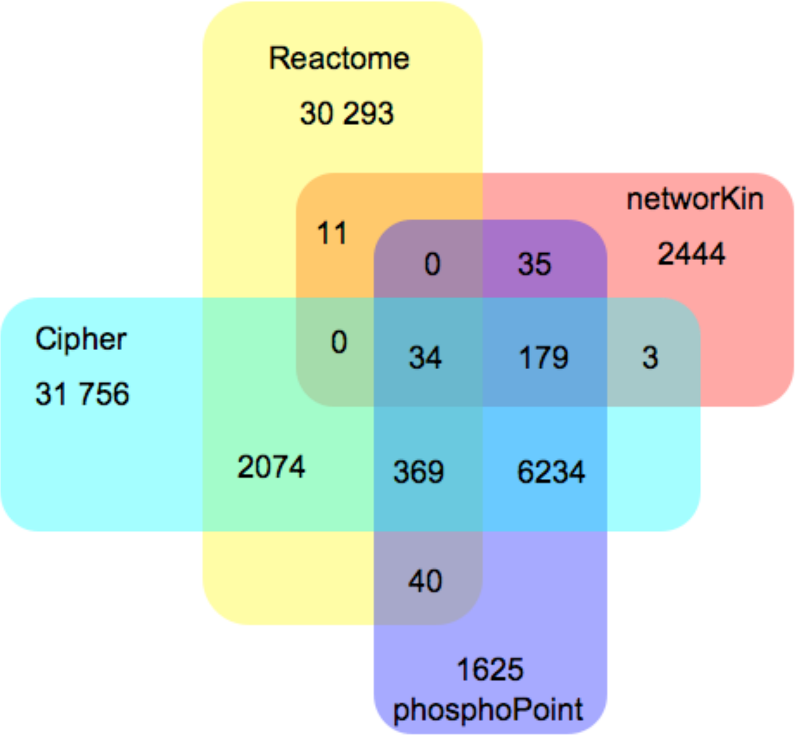
Threshold **H 0.04, M 0.025**
3.283.347 edges

- 20123 of the human genes belong to the same component.

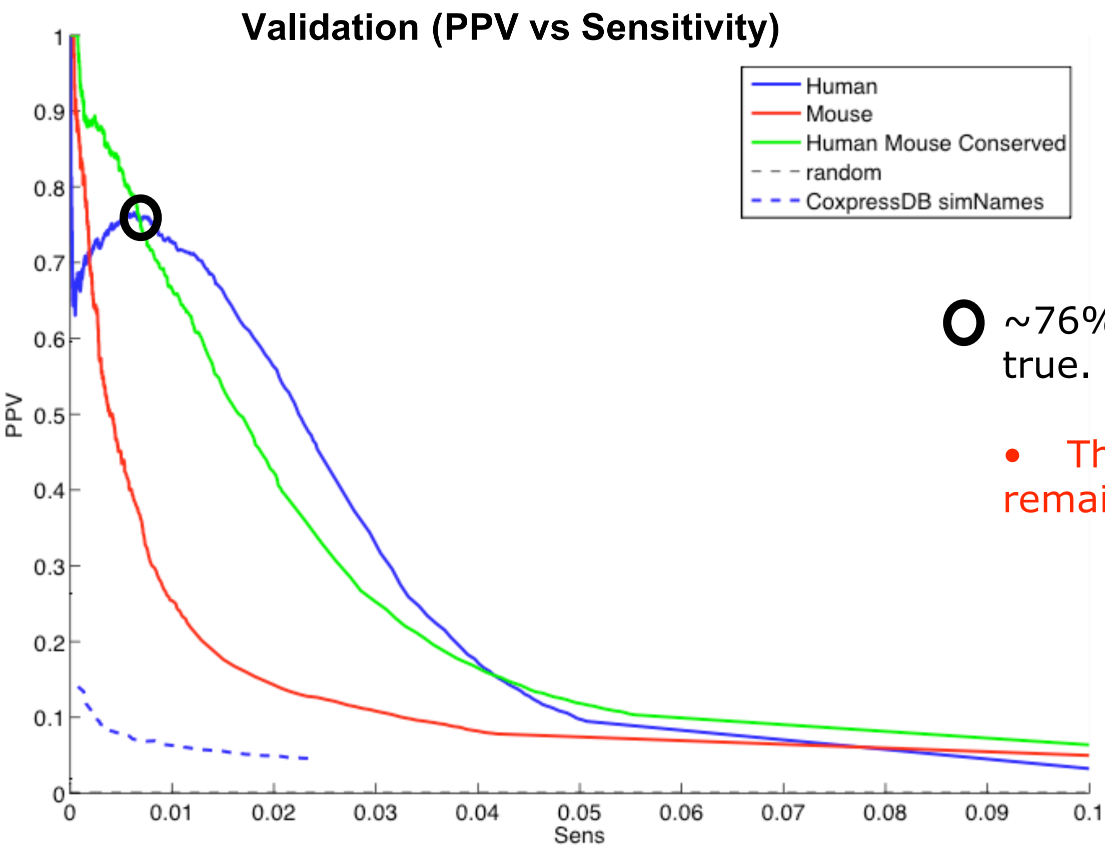


Interactome validation on experimentally verified interactions

- The Golden standard is a collection of experimentally validate edges for a total of **105.688** edges from a wide renge of publicly available databases:



Results --> Validation of gene-to-gene interactions

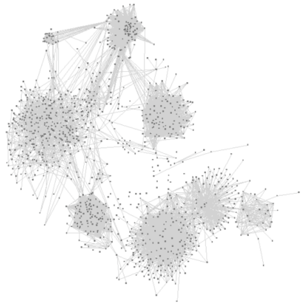


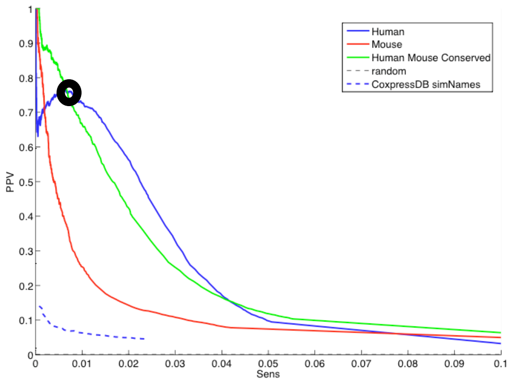
○ ~76% of the predicted edges were true.

- This doesn’t mean that the remaining ~24% are not correct.

- **COXPRESdb:** a database of coexpressed gene networks in mammals
Nucleic Acids Research, 2008, Vol. 36, Database issue D77-82

Results --> Validation of gene-to-gene interactions



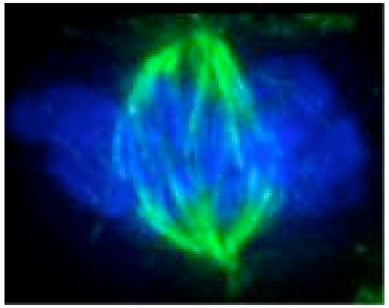


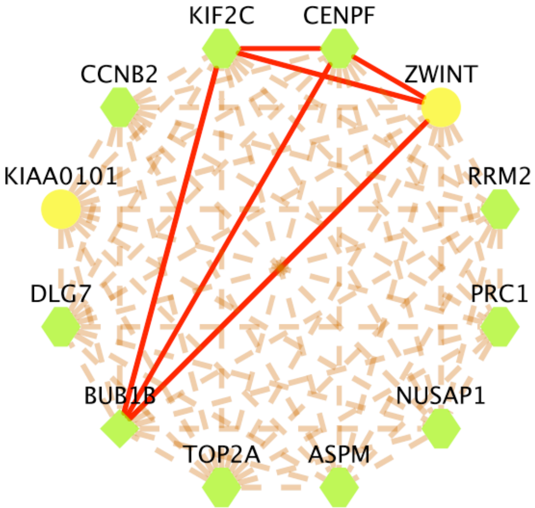
~76% of the edges predicted are true,

- This doesn’t mean that the remaining ~24% are not correct.

Genes involved into the spindle check point. We are currently experimentally validating these interactions via Y2H:

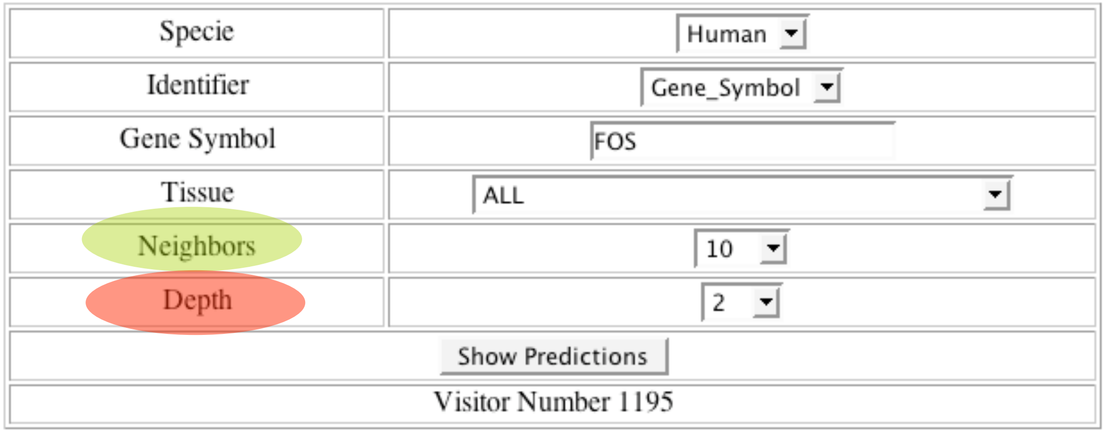
Cell division





Netview: Online visualization tool

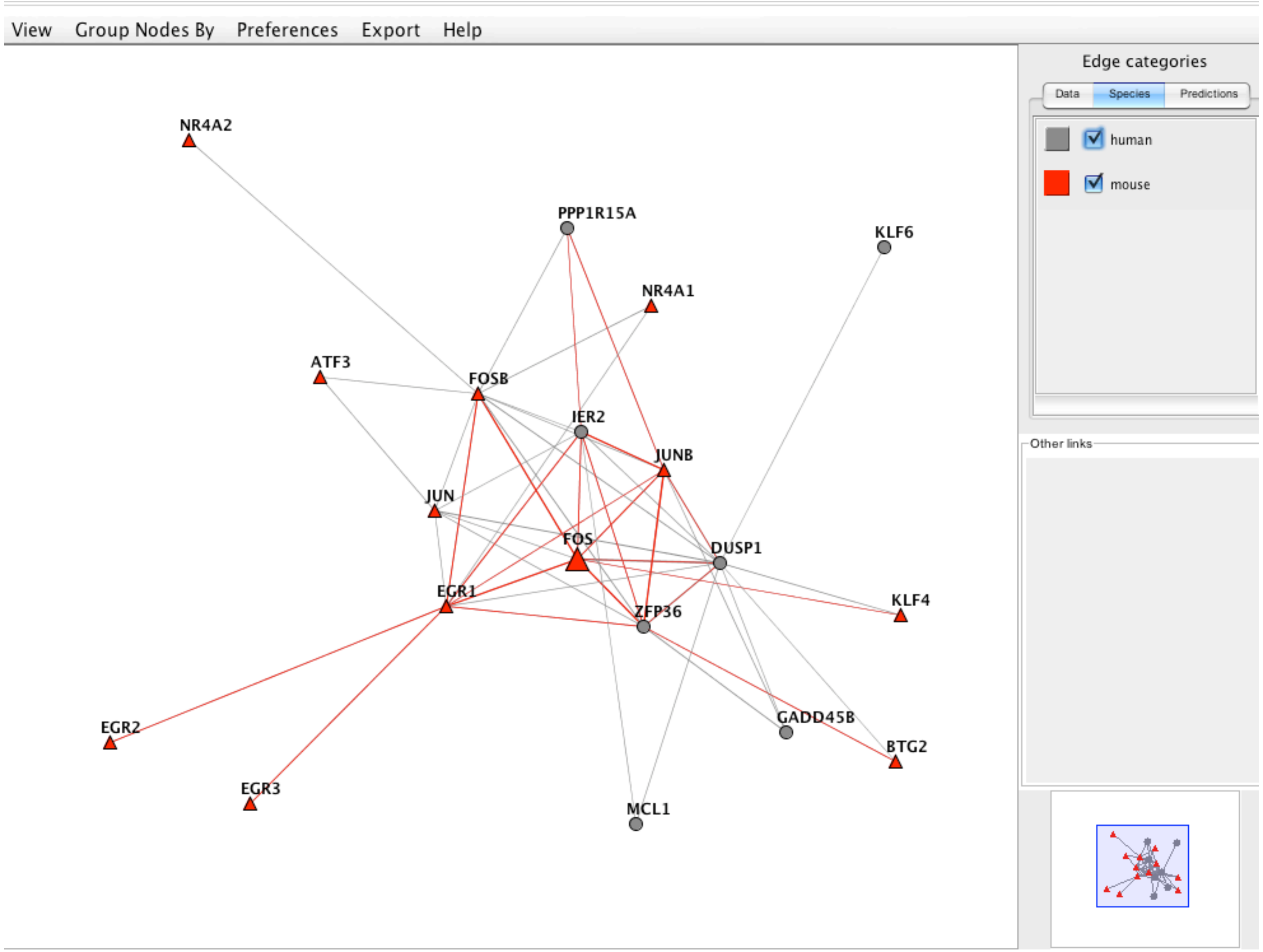
Query the database with a Gene Symbol



Specie	Human
Identifier	Gene_Symbol
Gene Symbol	FOS
Tissue	ALL
Neighbors	10
Depth	2
Show Predictions	
Visitor Number 1195	

Neighbors: # of nodes directly connected to the queried node.
Depth: # of network levels to explore (root is the queried node).

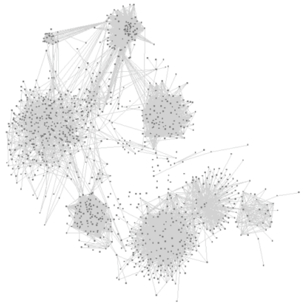
Netview: Online visualization tool

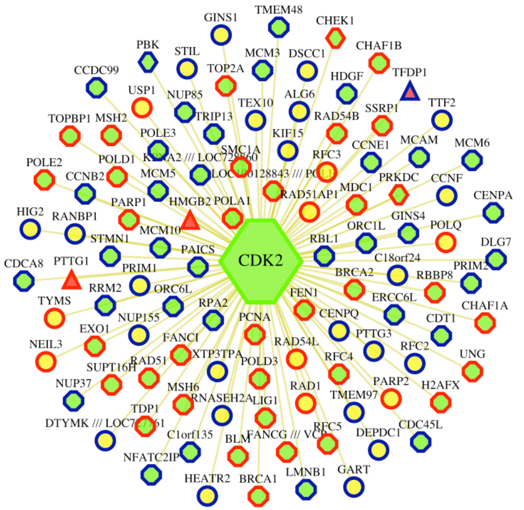


- **jSquid**: a Java applet for graphical on-line network exploration Bioinformatics 2008 24(12):1467-1468.

dibernardo.tigem.it

Using the network to understand gene function



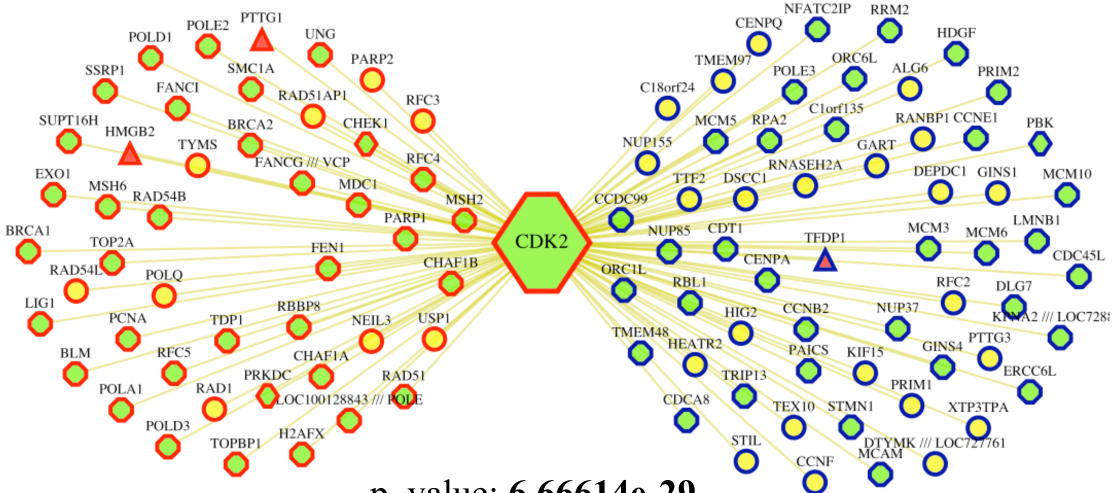


1. Neighbors selection.

2. Neighbors Enrichment analysis via hypergeometric distribution.

Cell Cycle *gene ontology*

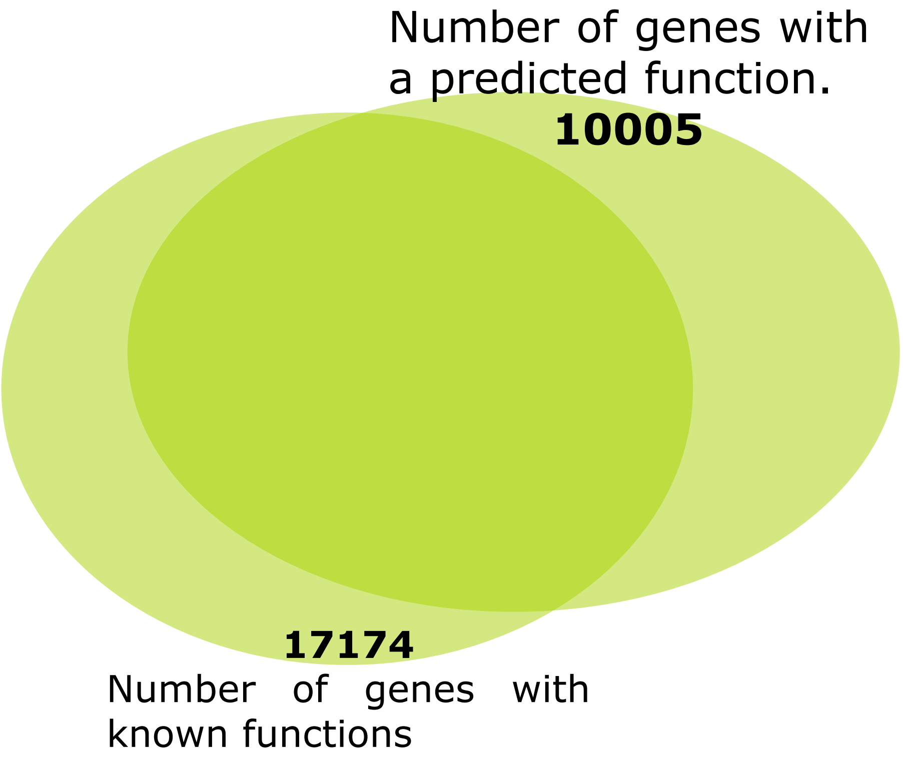
other gene ontology



p_value: **6.66614e-29**

3. Gene function prediction.

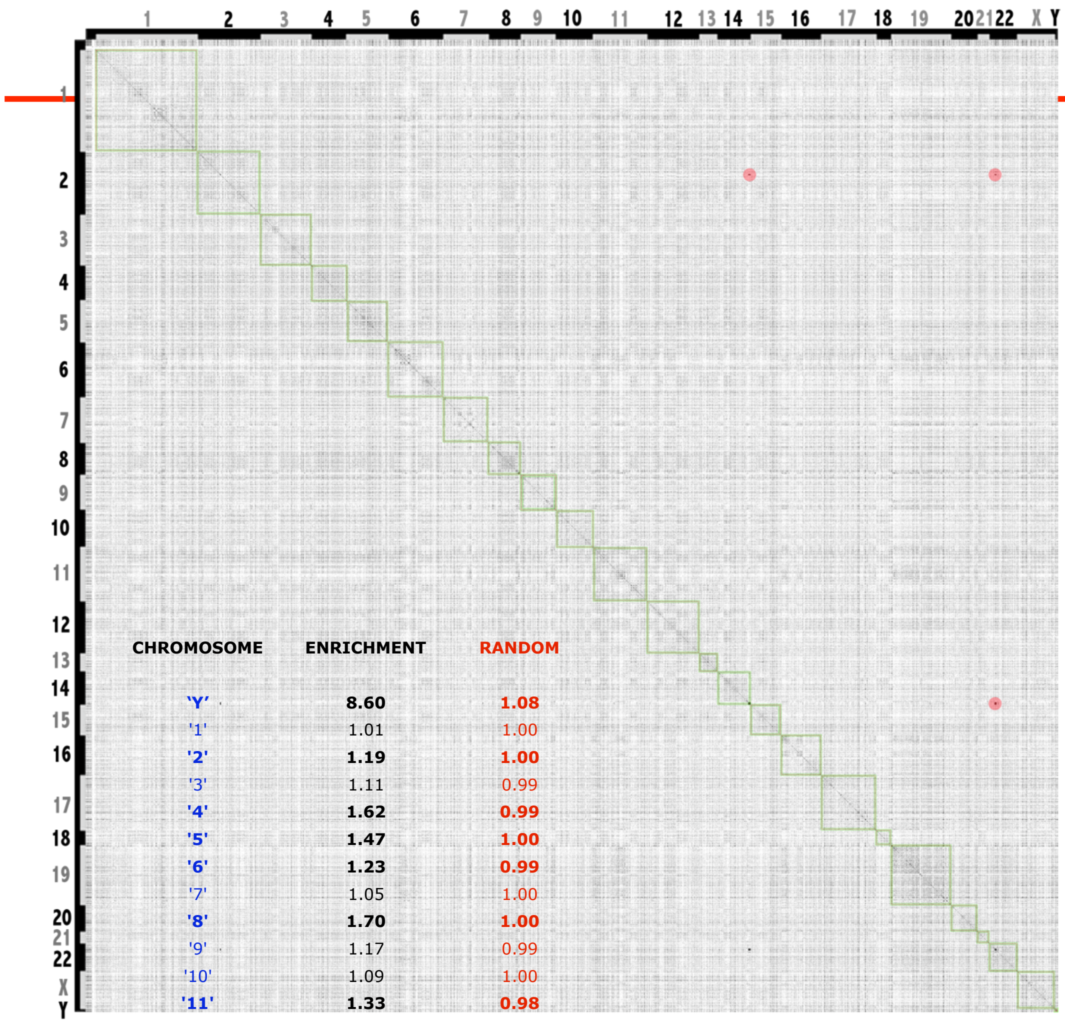
Using the network to understand gene function: validation



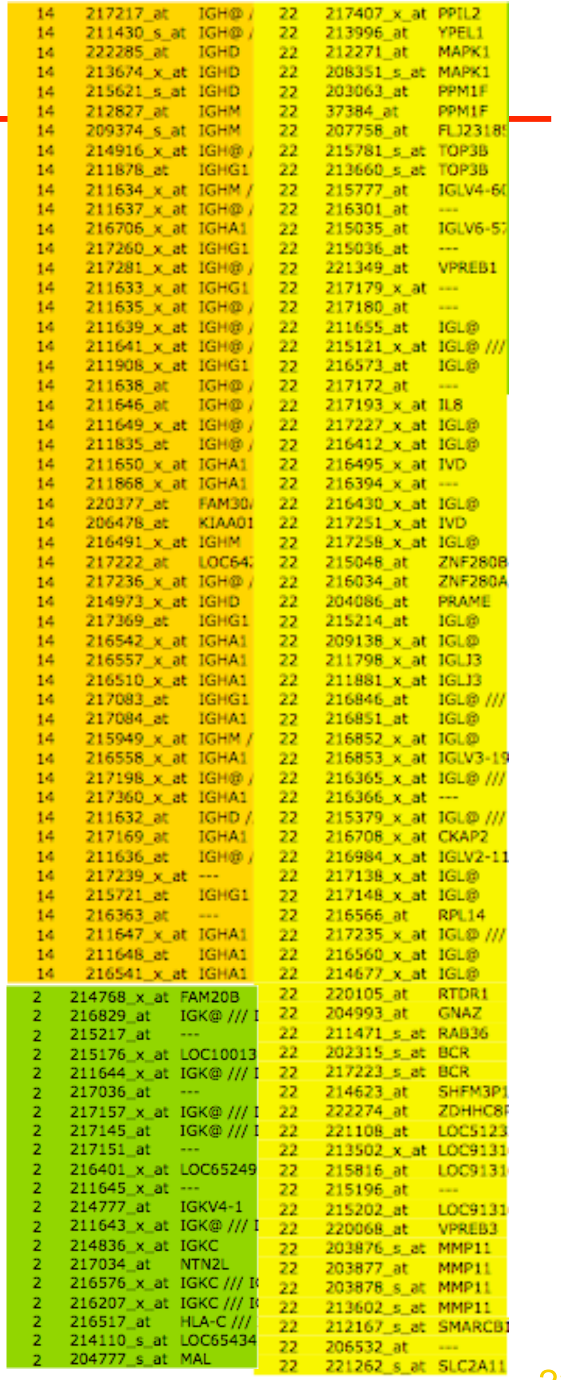
- 58% of the genes were correctly assigned to a gene function.

- This doesn’t mean that the remaining 42% are not properly assigned to a gene function.

- We are now validating experimentally gene function for 8 genes predicted to be localised in mitochondria



CHROMOSOME	ENRICHMENT	RANDOM
'Y'	8.60	1.08
'1'	1.01	1.00
'2'	1.19	1.00
'3'	1.11	0.99
'4'	1.62	0.99
'5'	1.47	1.00
'6'	1.23	0.99
'7'	1.05	1.00
'8'	1.70	1.00
'9'	1.17	0.99
'10'	1.09	1.00
'11'	1.33	0.98
'12'	1.08	1.01
'13'	1.69	1.03
'14'	1.25	1.02
'15'	1.35	0.98
'16'	1.04	1.00
'17'	1.18	1.00
'18'	1.15	0.97
'19'	1.76	0.99
'20'	1.26	1.01
'21'	1.07	0.96
'22'	1.59	1.01
'X'	1.38	0.99



14	217217_at	IGH@ /
14	211430_s_at	IGH@ /
14	222285_at	IGHD
14	213674_x_at	IGHD
14	215621_s_at	IGHD
14	212827_at	IGHM
14	209374_s_at	IGHM
14	214916_x_at	IGH@ /
14	211878_at	IGHG1
14	211634_x_at	IGHM /
14	211637_x_at	IGH@ /
14	216706_x_at	IGHA1
14	217260_x_at	IGHG1
14	217281_x_at	IGH@ /
14	211633_x_at	IGHG1
14	211635_x_at	IGH@ /
14	211639_x_at	IGH@ /
14	211641_x_at	IGH@ /
14	211908_x_at	IGHG1
14	211638_at	IGH@ /
14	211646_at	IGH@ /
14	211649_x_at	IGH@ /
14	211835_at	IGH@ /
14	211650_x_at	IGHA1
14	211868_x_at	IGHA1
14	220377_at	FAM30/
14	206478_at	KIAA01
14	216491_x_at	IGHM
14	217222_at	LOC64:
14	217236_x_at	IGH@ /
14	214973_x_at	IGHD
14	217369_at	IGHG1
14	216542_x_at	IGHA1
14	216557_x_at	IGHA1
14	216510_x_at	IGHA1
14	217083_at	IGHG1
14	217084_at	IGHA1
14	215949_x_at	IGHM /
14	216558_x_at	IGHA1
14	217198_x_at	IGH@ /
14	217360_x_at	IGHA1
14	211632_at	IGHD /
14	217169_at	IGHA1
14	211636_at	IGH@ /
14	217239_x_at	---
14	215721_at	IGHG1
14	216363_at	---
14	211647_x_at	IGHA1
14	211648_at	IGHA1
14	216541_x_at	IGHA1

2	214768_x_at	FAM20B
2	216829_at	IGK@ /// I
2	215217_at	---
2	215176_x_at	LOC10013
2	211644_x_at	IGK@ /// I
2	217036_at	---
2	217157_x_at	IGK@ /// I
2	217145_at	IGK@ /// I
2	217151_at	---
2	216401_x_at	LOC65249
2	211645_x_at	---
2	214777_at	IGKV4-1
2	211643_x_at	IGK@ /// I
2	214836_x_at	IGKC
2	217034_at	NTN2L
2	216576_x_at	IGKC /// I
2	216207_x_at	IGKC /// I
2	216517_at	HLA-C ///
2	214110_s_at	LOC65434
2	204777_s_at	MAL

22	217407_x_at	PPIL2
22	213996_at	YPEL1
22	212271_at	MAPK1
22	208351_s_at	MAPK1
22	203063_at	PPM1F
22	37384_at	PPM1F
22	207758_at	FLJ2318
22	215781_s_at	TOP3B
22	213660_s_at	TOP3B
22	215777_at	IGLV4-6
22	216301_at	---
22	215035_at	IGLV6-5
22	215036_at	---
22	221349_at	VPREB1
22	217179_x_at	---
22	217180_at	---
22	211655_at	IGL@
22	215121_x_at	IGL@ ///
22	216573_at	IGL@
22	217172_at	---
22	217193_x_at	IL8
22	217227_x_at	IGL@
22	216412_x_at	IGL@
22	216495_x_at	IVD
22	216394_x_at	---
22	216430_x_at	IGL@
22	217251_x_at	IVD
22	217258_x_at	IGL@
22	215048_at	ZNF280B
22	216034_at	ZNF280A
22	204086_at	PRAME
22	215214_at	IGL@
22	209138_x_at	IGL@
22	211798_x_at	IGLJ3
22	211881_x_at	IGLJ3
22	216846_at	IGL@ ///
22	216851_at	IGL@
22	216852_x_at	IGL@
22	216853_x_at	IGLV3-19
22	216365_x_at	IGL@ ///
22	216366_x_at	---
22	215379_x_at	IGL@ ///
22	216708_x_at	CKAP2
22	216984_x_at	IGLV2-11
22	217138_x_at	IGL@
22	217148_x_at	IGL@
22	216566_at	RPL14
22	217235_x_at	IGL@ ///
22	216560_x_at	IGL@
22	214677_x_at	IGL@
22	220105_at	RTDR1
22	204993_at	GNAZ
22	211471_s_at	RAB36
22	202315_s_at	BCR
22	217223_s_at	BCR
22	214623_at	SHFM3P1
22	222274_at	ZDHHC8P
22	221108_at	LOC5123
22	213502_x_at	LOC9131
22	215816_at	LOC9131
22	215196_at	---
22	215202_at	LOC9131
22	220068_at	VPREB3
22	203876_s_at	MMP11
22	203877_at	MMP11
22	203878_s_at	MMP11
22	213602_s_at	MMP11
22	212167_s_at	SMARCB1
22	206532_at	---
22	221262_s_at	SLC2A11

Conclusion of Part I:

- Using expression data from a wide variety of tissues and cell lines enables the identification of functional modules within the cell regulatory network

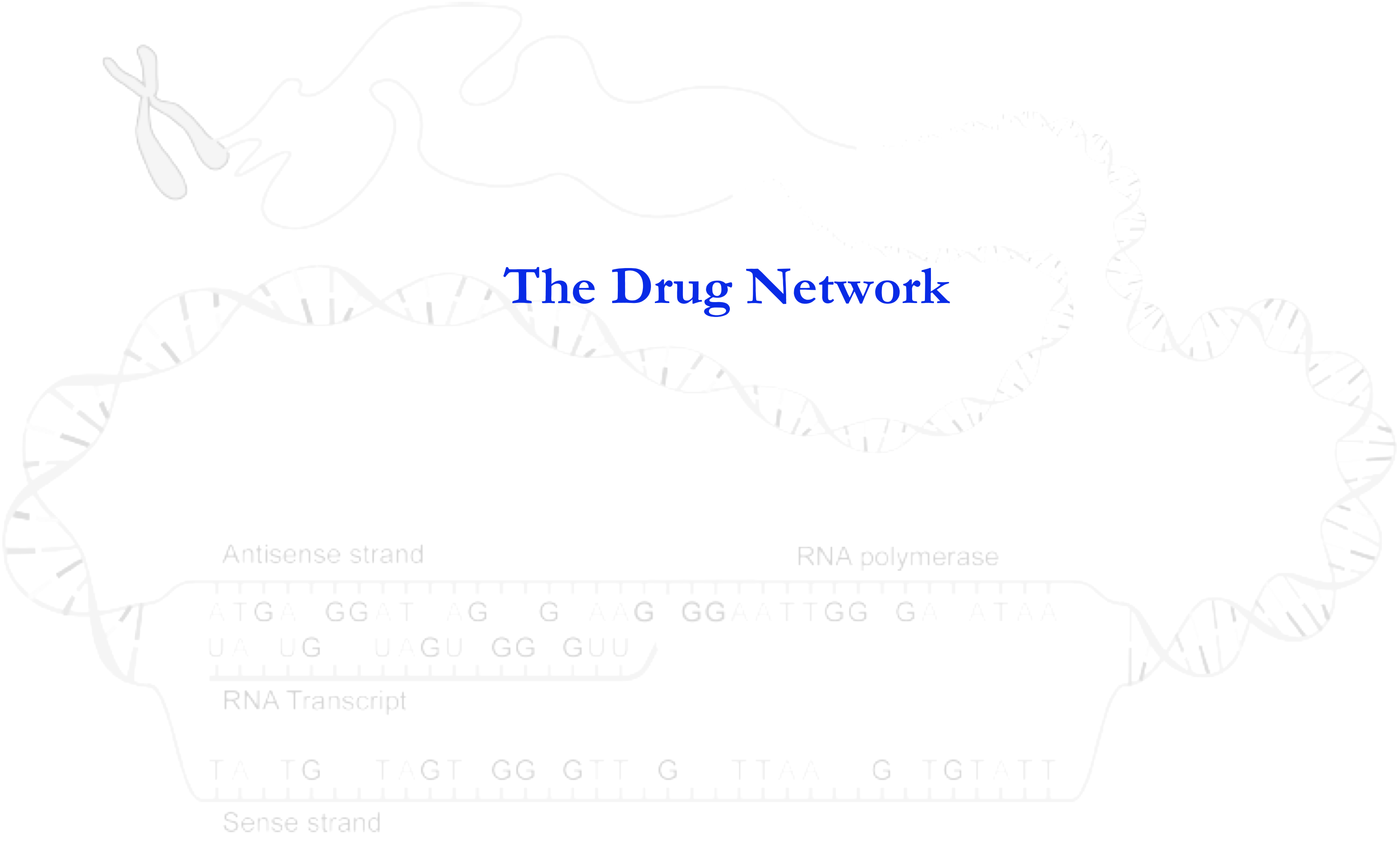
- It is possible to predict functional and physical interactors of a gene using co-expression networks

- It is possible to predict the function of a gene from its interactors (i.e. co-expressed genes)

- We are now looking at how this global interactome network can be useful in interpreting gene expression data and to understand the global organisation of the cell regulatory network

Part II: Understanding Drug Mode of Action

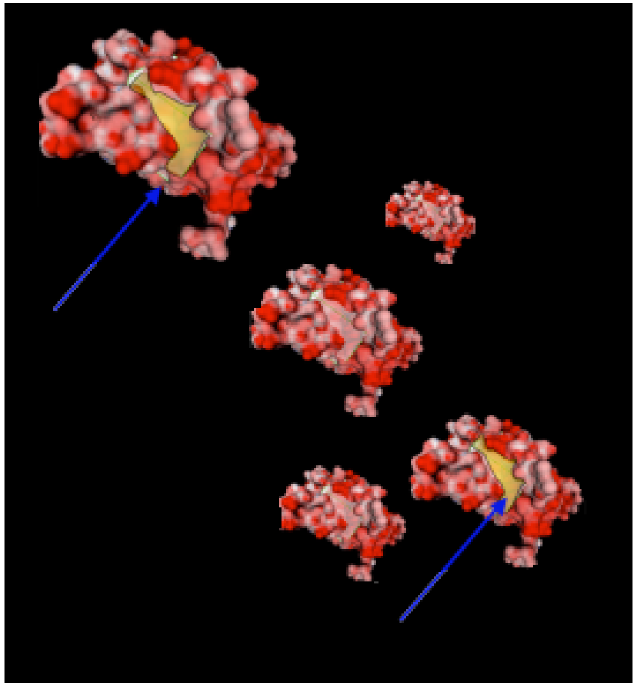
The Drug Network



Iorio F et al, UNPUBLISHED (confidential)

Drug Discovery Problem

We want to investigate the mode of action of a novel drug...

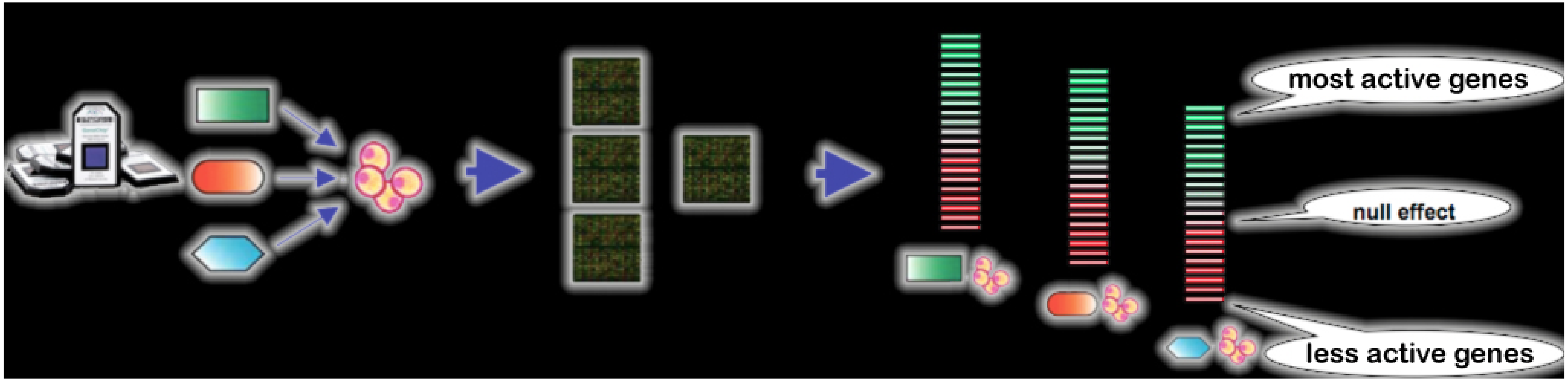


Therapeutic Target

Off Target

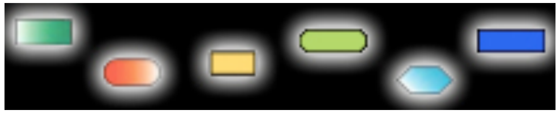
Drug Molecule

The Connectivity Map DataSet (microarrays):

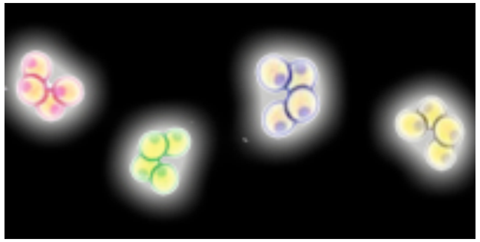


Differential Expression Profiles
as Ranked Lists of Genes

small molecules: 1309 perturbagens tested
(FDA approved and nondrug bioactive compounds)



cell lines:
MCF7 (human epitelial breast cancer)
PC3 (human epitelial prostate cancer)
HL60 (human leukemia)
SKMEL5 (human melanoma)
ssMCF7 (MCF7 grown in a different veichle)



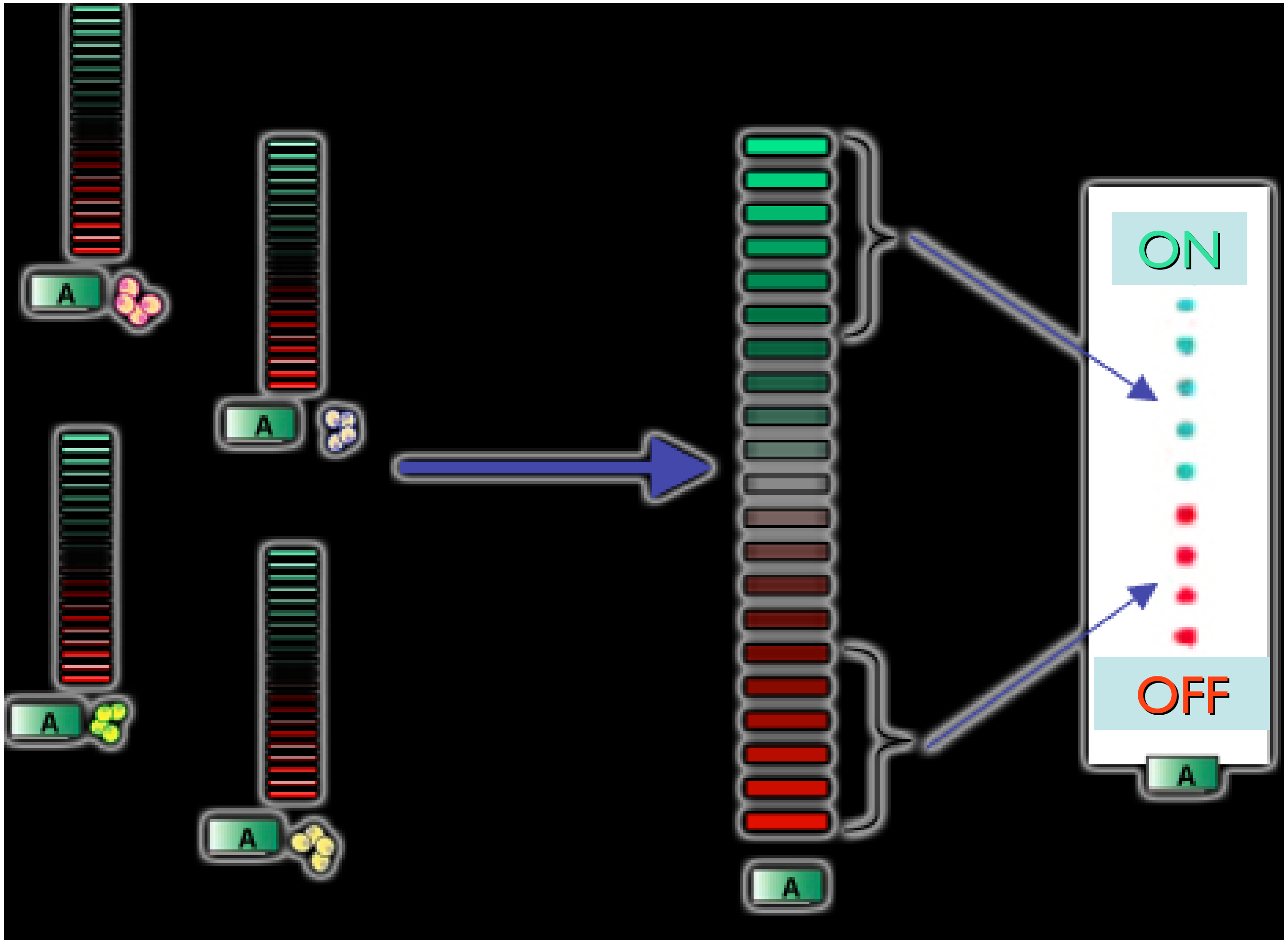
Concentration and treatment
10mM (when the optimal concentration is unknown) x 6h

Negative Control
cells in the same plate and treated with vehicle alone (medium, DMSO…)

[Lamb et Al, Science 2006]

General Cellular Response to a Drug:

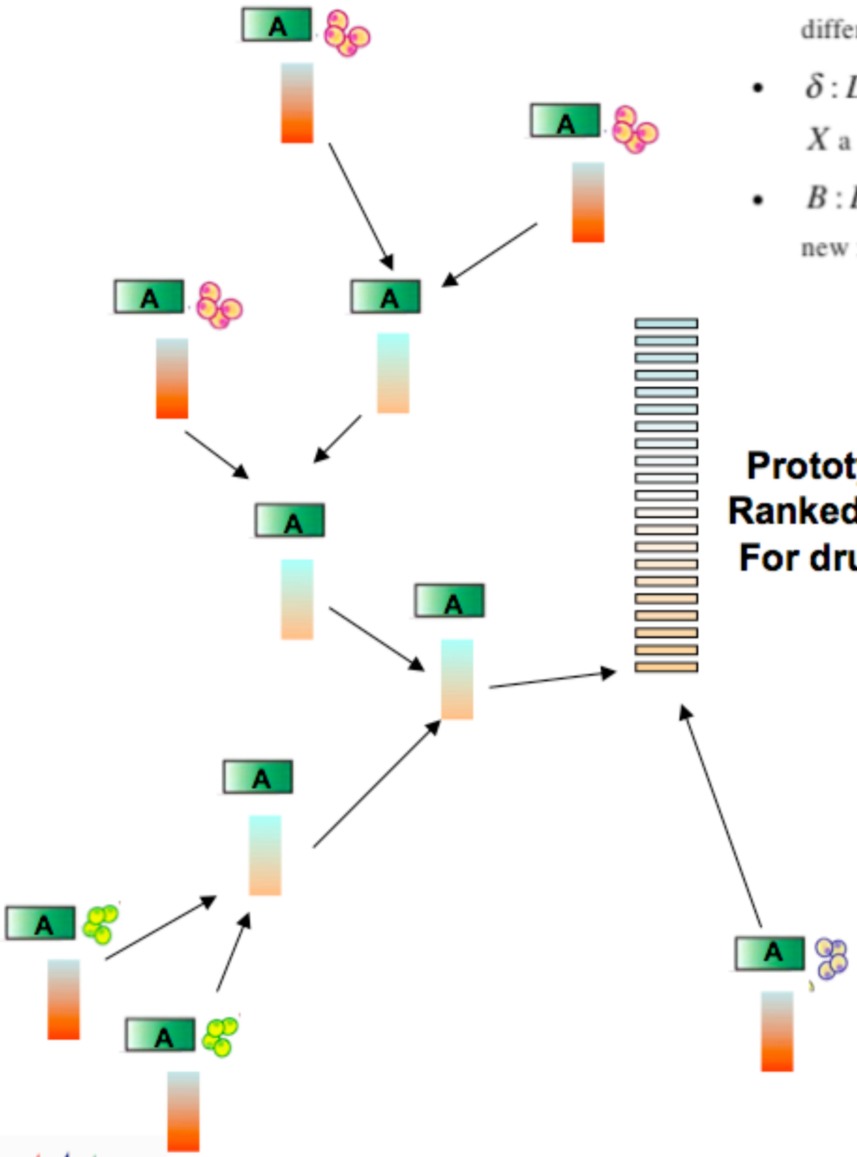
Using a novel rank aggregation method (next slide)



Prototype
Ranked List (PRL)
for Drug A

[Iorio et Al, Journal of Computational Biology 2009]

The Kru-Bor Merging Method



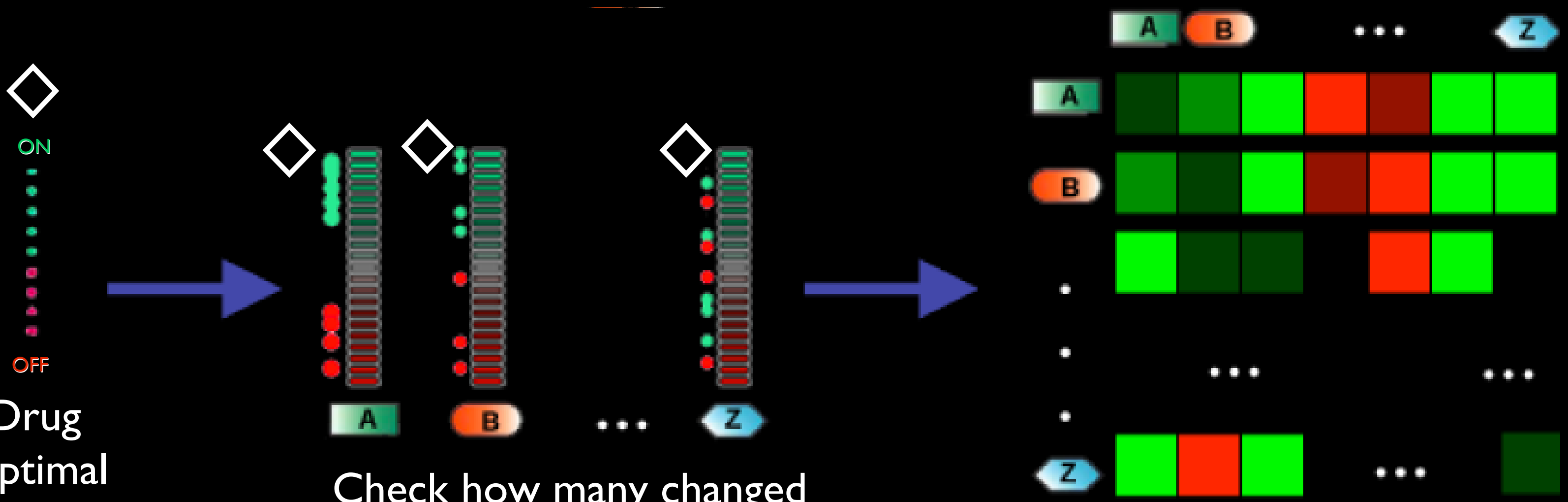
Prototype Ranked List For drug A

- D : The set of all the possible pemutations of microarray probes;
- X : A set of ranked lists of probes computed by sorting, in decreasing order, the genome-wide differential expression profiles (GEP) obtained by treating cell lines with the same drug;
- $\delta : D^2 \rightarrow N$: The *Spearman's Foot-Rule* distance associating to each pair of ranked lists in X a natural number quantifying the similarity between them;
- $B : D^2 \rightarrow D$: The *Borda Merging Function*, associating to each pair of ranked lists in X , a new ranked lists obtained by merging them with the *Borda Merging Method*;

```
$n = |X|$
while $n > 1$
    find $i, j : \delta(x_i, x_j) = \min_{p,q=1,\ldots,n;\ p \neq q} \delta(x_p, x_q)$
    $y = B(x_i, x_j)$
    $X = (X / \{x_i, x_j\}) \cup \{y\}$
    $n = |X|$
end
```

The Drug Distance Matrix

For each drug ◇



Drug Optimal Signature

Check how many changed genes are in common

Compute similarity by using Enrichment Scores

Computation of the drug distance:

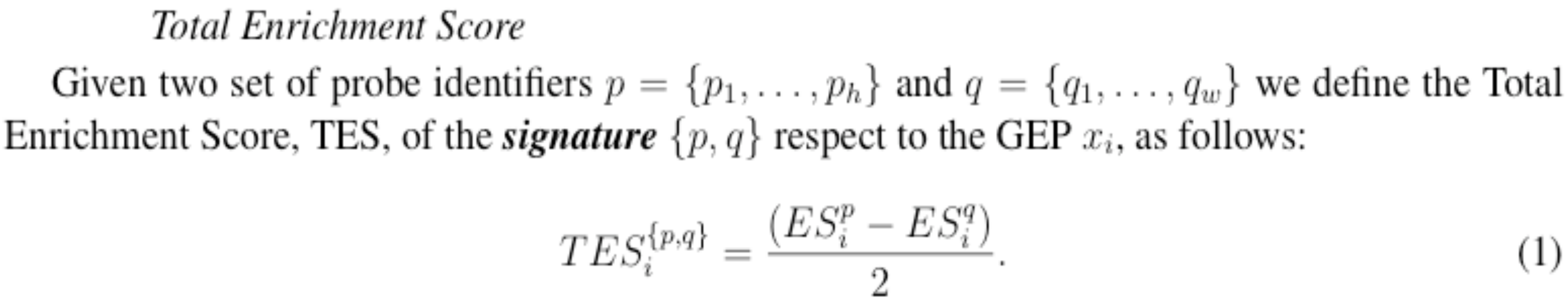
- Given a set of N_H probes in S and a ranked list of N probes:

- The Enrichment Score of S on the list is defined as:
  - $\max_i |P_{hit}-P_{miss}|$
  - where:

$$P_{\text{hit}}(S, i) = \sum_{\substack{g_j \in S \\ j \leq i}} \frac{1}{N_H}$$

$$P_{\text{miss}}(S, i) = \sum_{\substack{g_j \notin S \\ j \leq i}} \frac{1}{(N - N_H)}.$$

Computation of the distance:



Total Enrichment Score

Given two set of probe identifiers $p = \{p_1, \ldots, p_h\}$ and $q = \{q_1, \ldots, q_w\}$ we define the Total Enrichment Score, TES, of the ***signature*** $\{p, q\}$ respect to the GEP x_i, as follows:

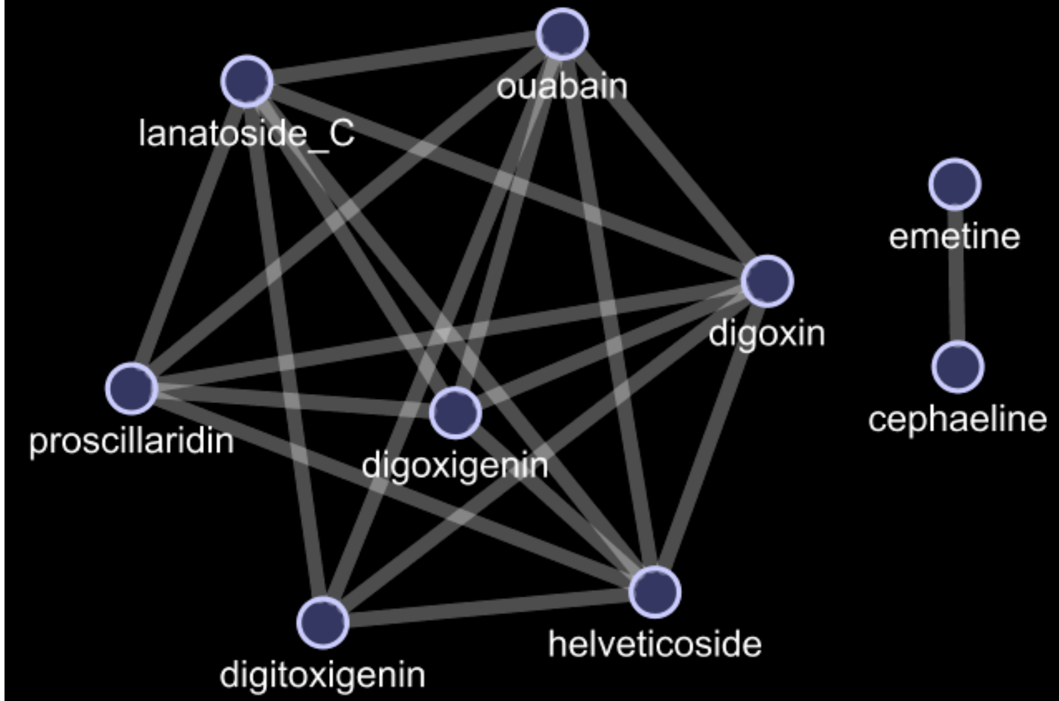
$$TES_i^{\{p,q\}} = \frac{(ES_i^p - ES_i^q)}{2}. \qquad (1)$$

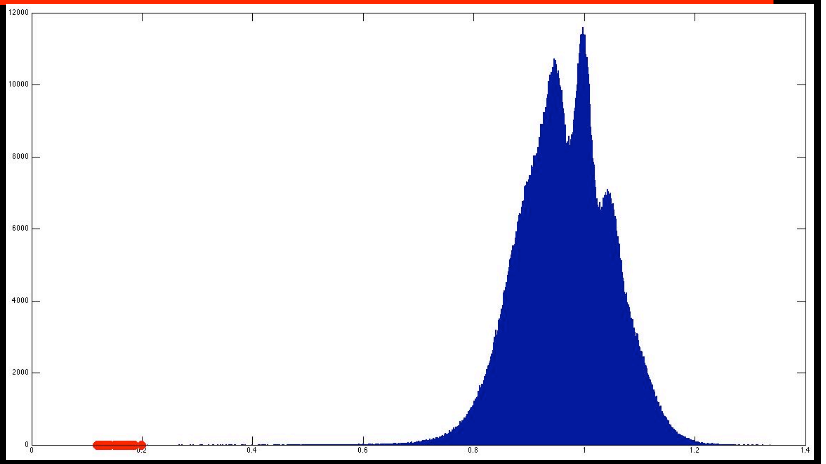
We then define as a distance between two compounds **i** and **j** the following quantity:

$$\frac{1}{2}(TES_i^{\{p_j, q_j\}} + TES_j^{\{p_i, q_i\}})$$

The Drug Network is obtained by setting a threshold:

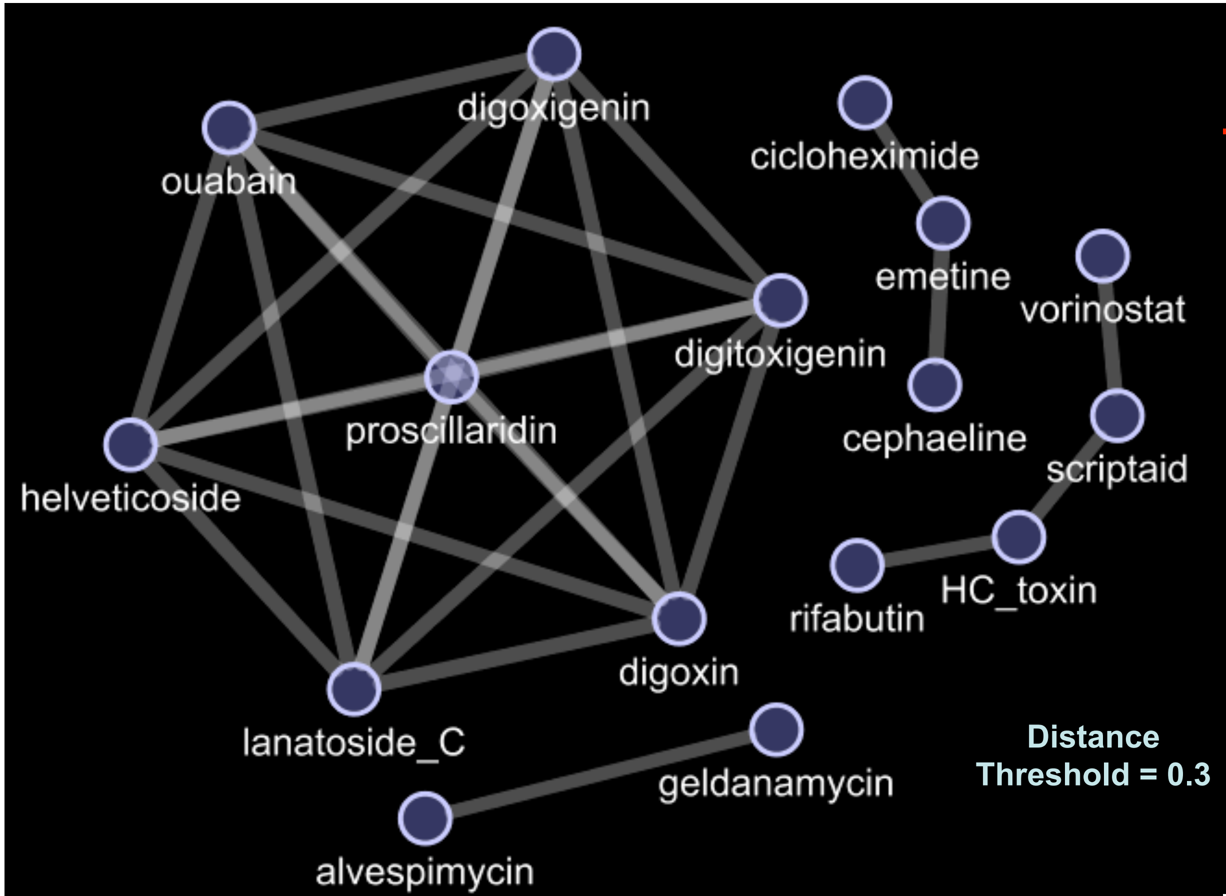
Distance Threshold = 0.2



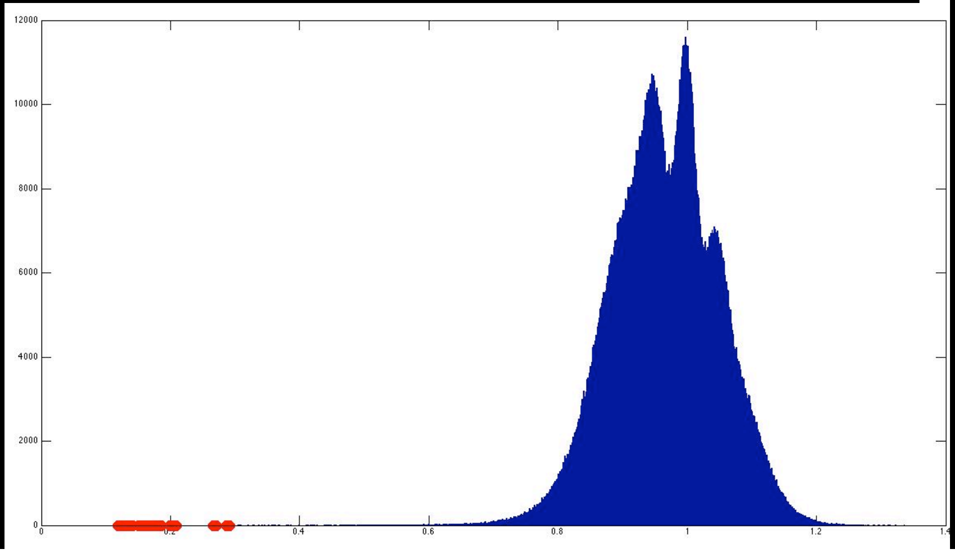


Compound	Informations
Cephaeline	Ipecac (a plant) Alkaoids - Protein Synthesis inhibition
Emetine	
Digoxigenin	Steroids found in some species of digitalis (purpurea or lanata), a plants. Used to treat Cardiac Diseases.
Digoxin	
Digitoxigenin	
Helveticoside	Cardiac Glycoside
Ouabain	Endogenous hormone found in the ripe seeds of the african plant Strophanthus. It blocks the sodium pump and it is used to cure human heart failure, angina pectoris and Myocardial infarction.
Proscillaridin	Cardiotonic Glycoside isolated from Scilla maritima var. Alba (a plant)
Lanatoside C	Cardiac Glycoside

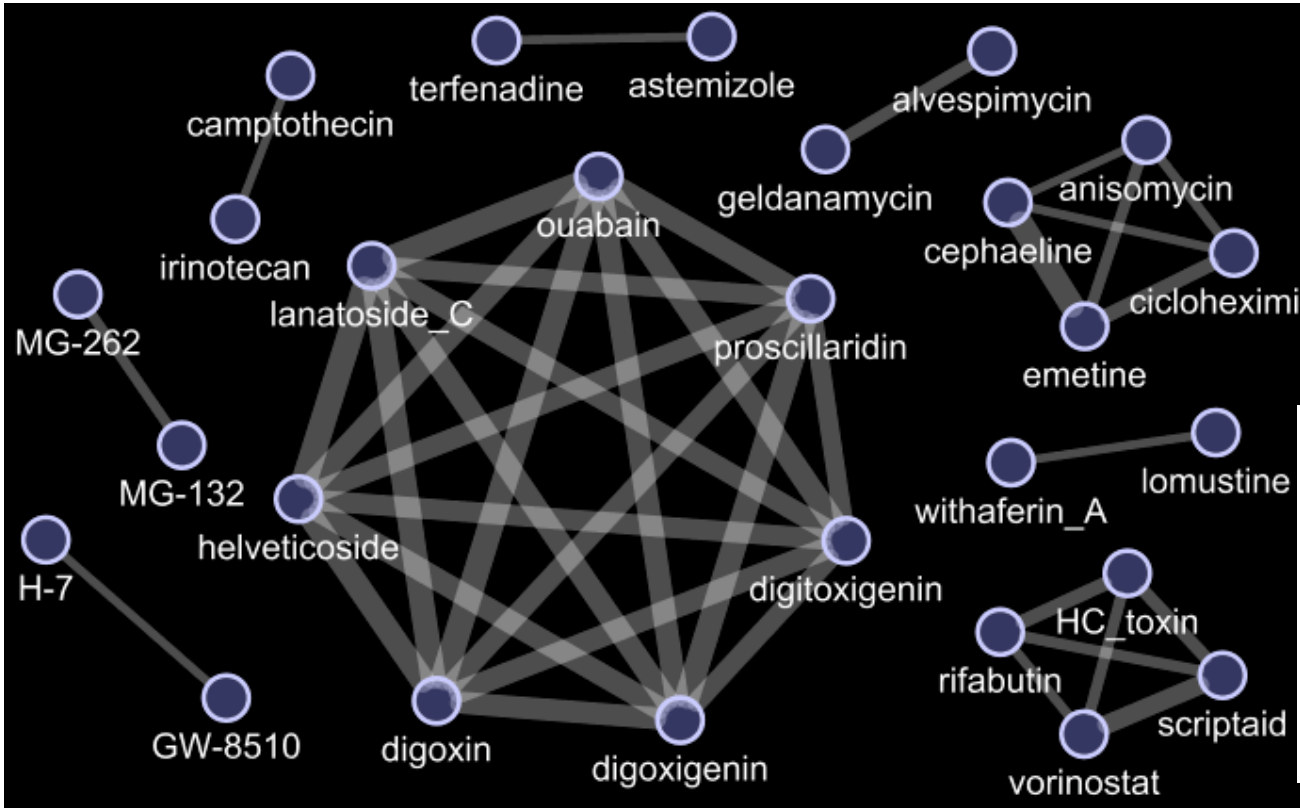
tigem TELETHON INSTITUTE OF GENETICS AND MEDICINE



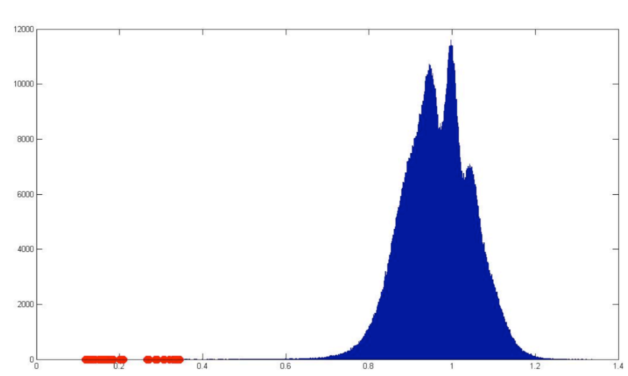
Distance Threshold = 0.3



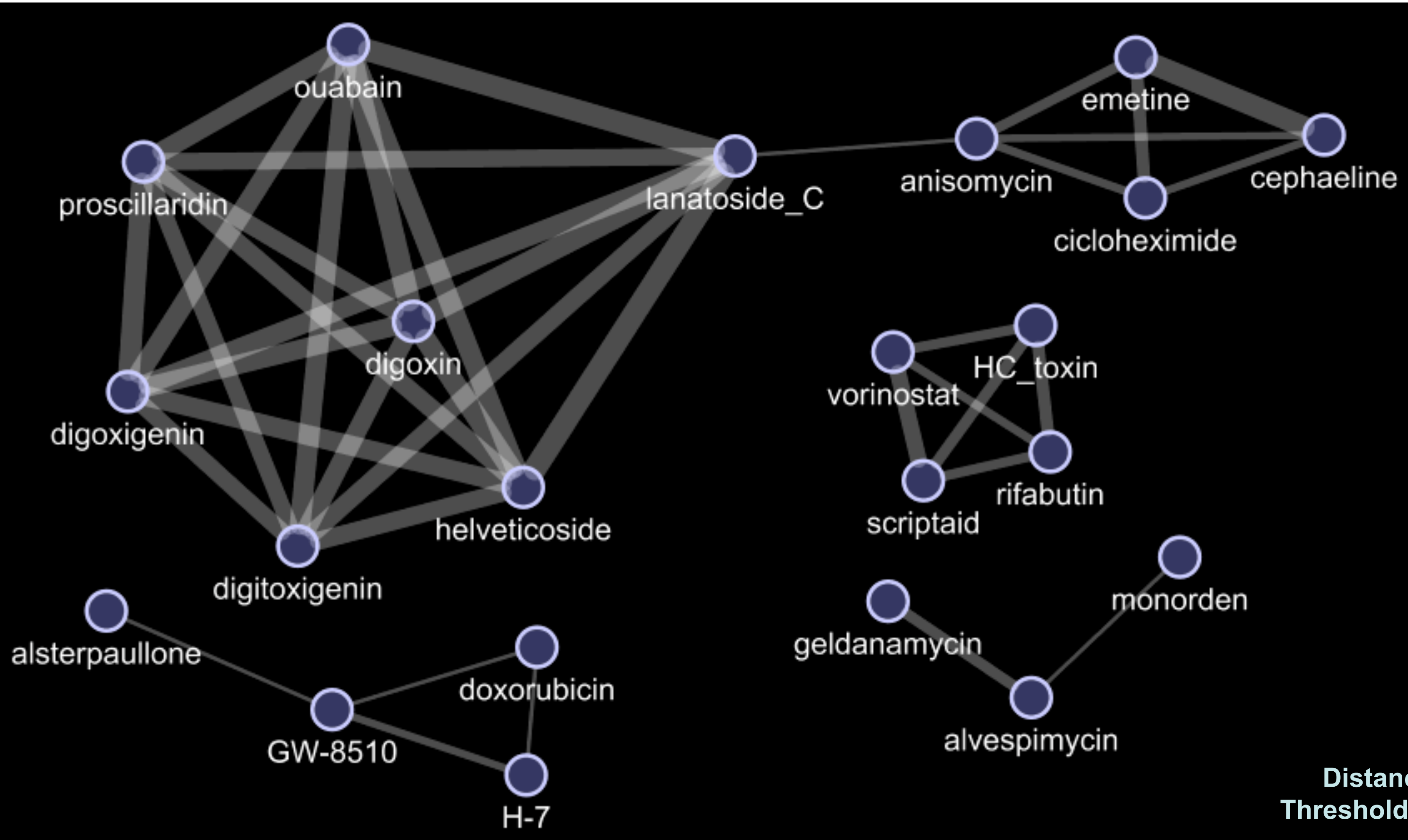
Compound	Informations
Cicloheximide	Antibiotic substance isolated from streptomycin-producing strains of Streptomyces griseus. It acts by inhibiting elongation during protein synthesis.
Geldanamycin	a benzoquinone ansamycin antibiotic that binds to Hsp90 (Heat Shock Protein 90) and alters its function.
Alvespimycin	Hsp90 inhibitor that has demonstrated the potential to disrupt the activity of multiple oncogenes and cell signaling pathways implicated in tumor growth, including HER2, a key signaling pathway in breast cancer.
vorinostat	or suberoylanilide hydroxamic acid (SAHA) is a member of a larger class of compounds that inhibit histone deacetylases (HDAC).
scriptaid	A novel histone deacetylase inhibitor
HC Toxin	Inhibition of Maize Histone Deacetylases by HC Toxin, the Host-Selective Toxin of Cochliobolus carbonum
Rifabutin	Rifabutin is a bactericidal antibiotic drug primarily used in the treatment of tuberculosis. The drug is a semi-synthetic derivative of rifamycin S. Its effect is based on blocking the DNA-dependent RNA-polymerase of the bacteria.



Distance Threshold = 0.35



Compound	Informations	Compound	Informations
anisomycin	also known as flagecidin is an antibiotic produced by Streptomyces griseolus which inhibits protein synthesis.	MG-262	a proteasome inhibitor
captothecin	a cytotoxic quinoline alkaloid which inhibits the DNA enzyme topoisomerase I (topo I).	MG-132	a proteasome inhibitor
irinotecan	a chemotherapy agent that is a topoisomerase 1 inhibitor. Chemically, it is a semisynthetic analogue of the natural alkaloid camptothecin.	Lomustine	is an alkylating nitrosourea compound used in chemotherapy. It is in the same family as streptozotocin
astemizole	a second generation antihistamine drug which has a long duration of action.	Withaferin A	an anticancer
terfenadine	an antihistamine formerly used for the treatment of allergic conditions.	H-7	Protein Kinase C Inhibitor
		GW-8510	Cyclin Dependant Kinase (CDK) inhibitor . (CDK2 in.)



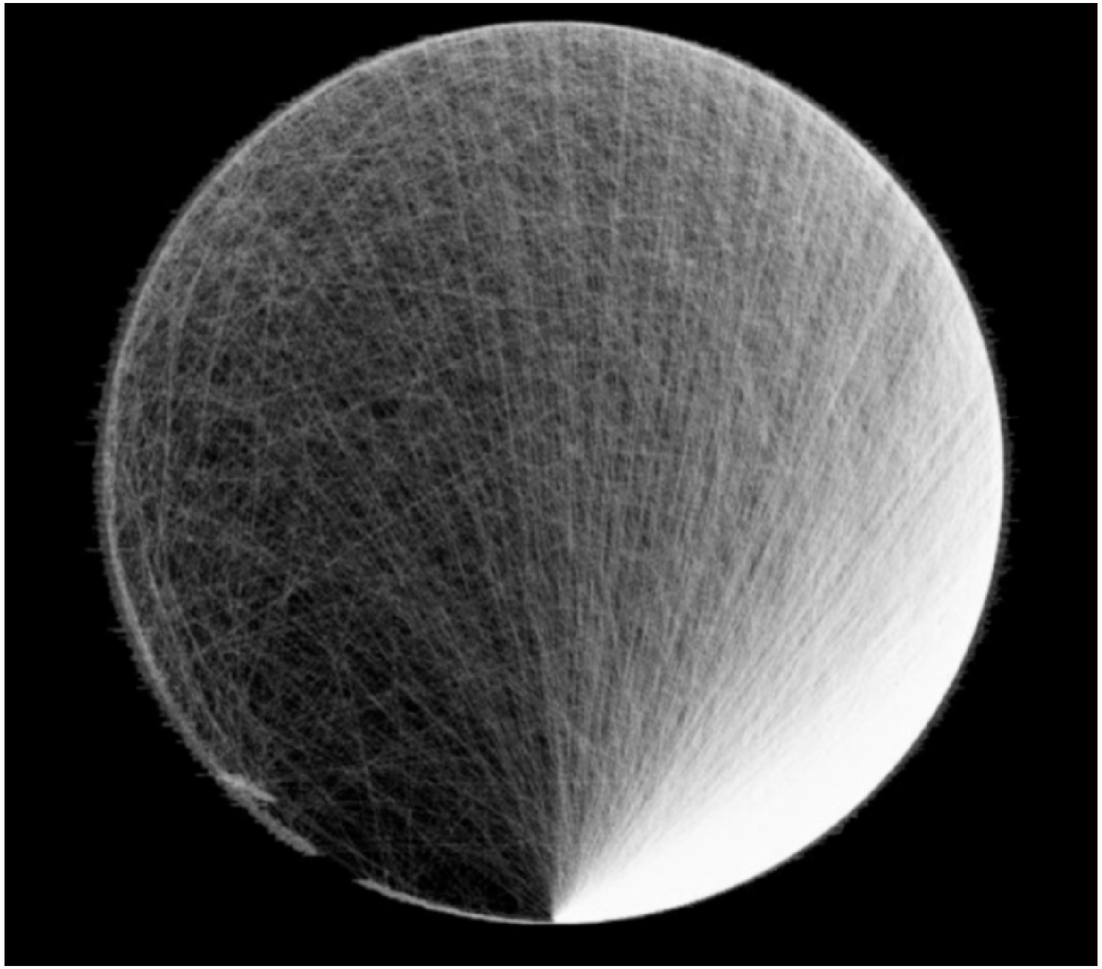
Distance Threshold = 0.4

Secondary Similarity	Informations
Lanatoside C, anisomycin	Caspase-3 inhibitors

Compound	Informations
monorden	(radicicol) antifugal metabolites. It inhibits the Hsp90 Chaperone
alsterpaullone	CDKs inhibitor
doxorubicin	a drug widely used in cancer chemotherapy. It is an anthracycline antibiotic and structurally closely related to daunomycin. Used in combination with CDKs inhibitors

The Drug network

There is an edge connecting two drugs if their distance is below a fixed threshold



Distance Threshold = 0.8049

Statistics

number of connected vertices = 1302

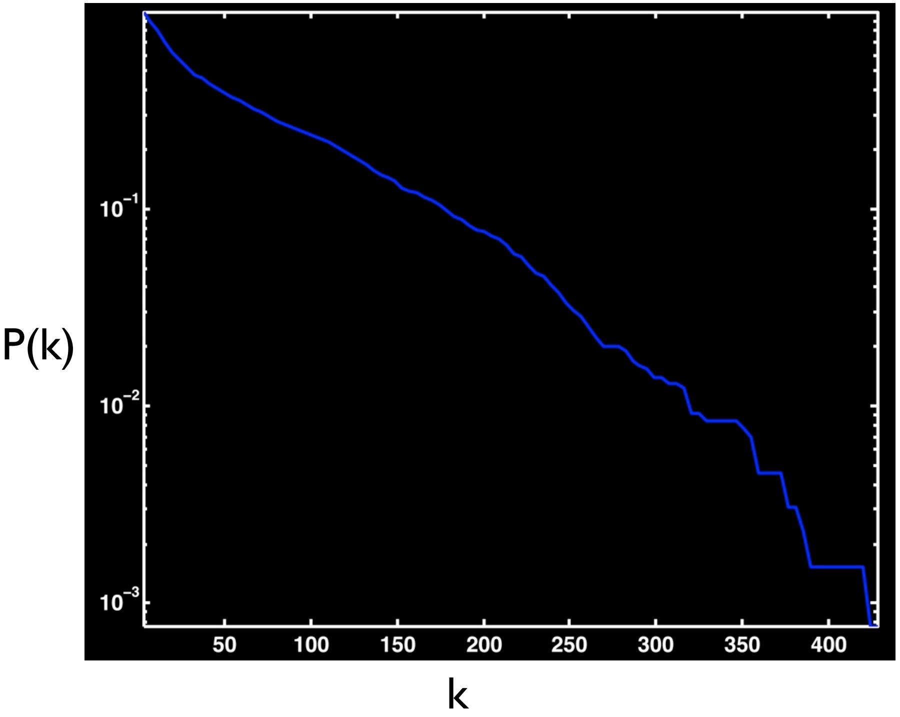
number of edges = 41047 (~ 5% of a fully connected network with the same number of nodes)

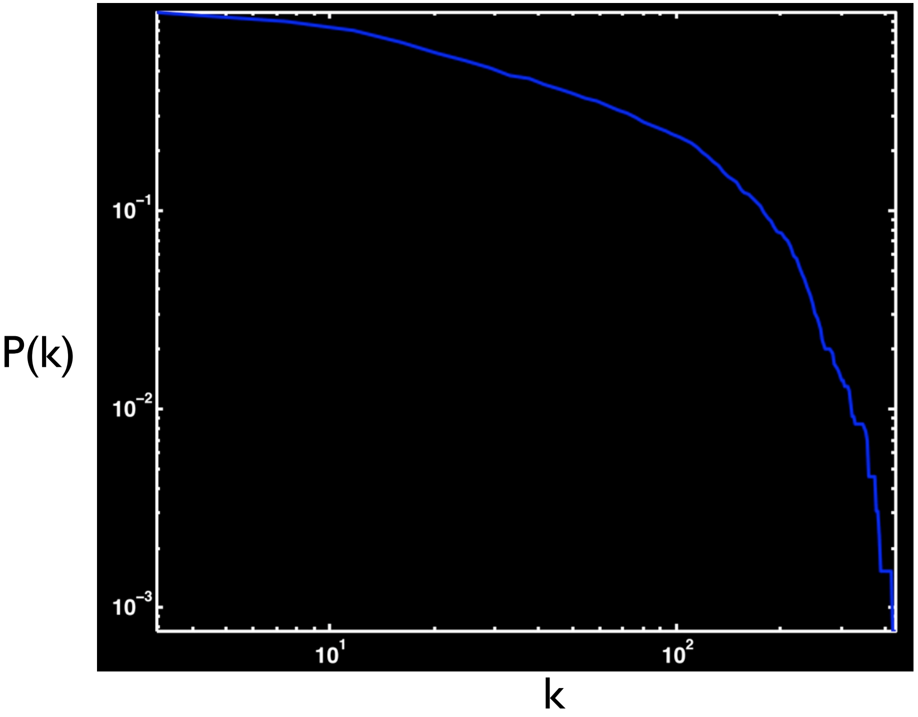
Avg. Shortest Path length = 2.5

Avg. Local Clustering Coefficient = 0.44

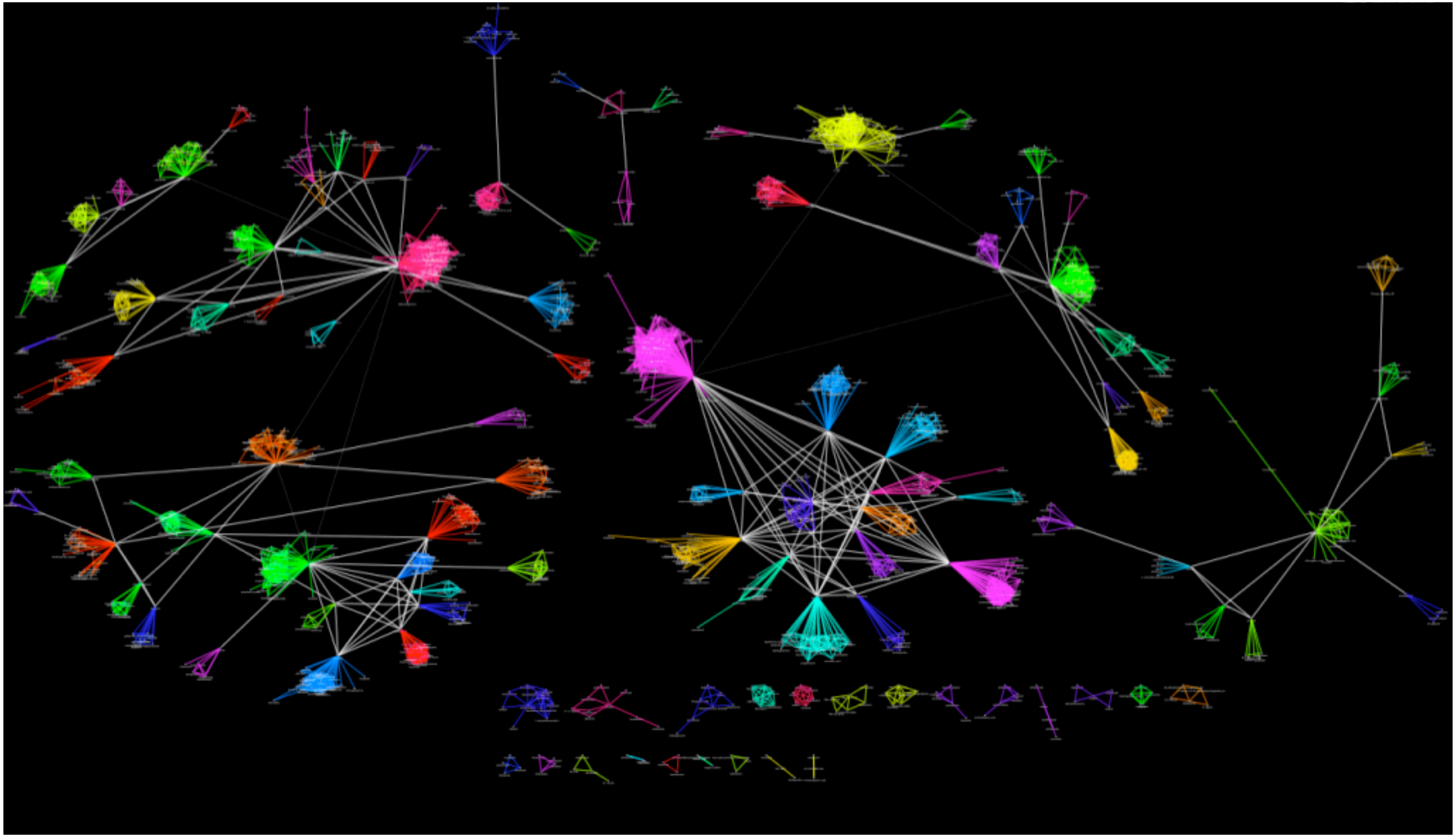
Maximum Shortest Path = 7

Node Degree Empirical cdf



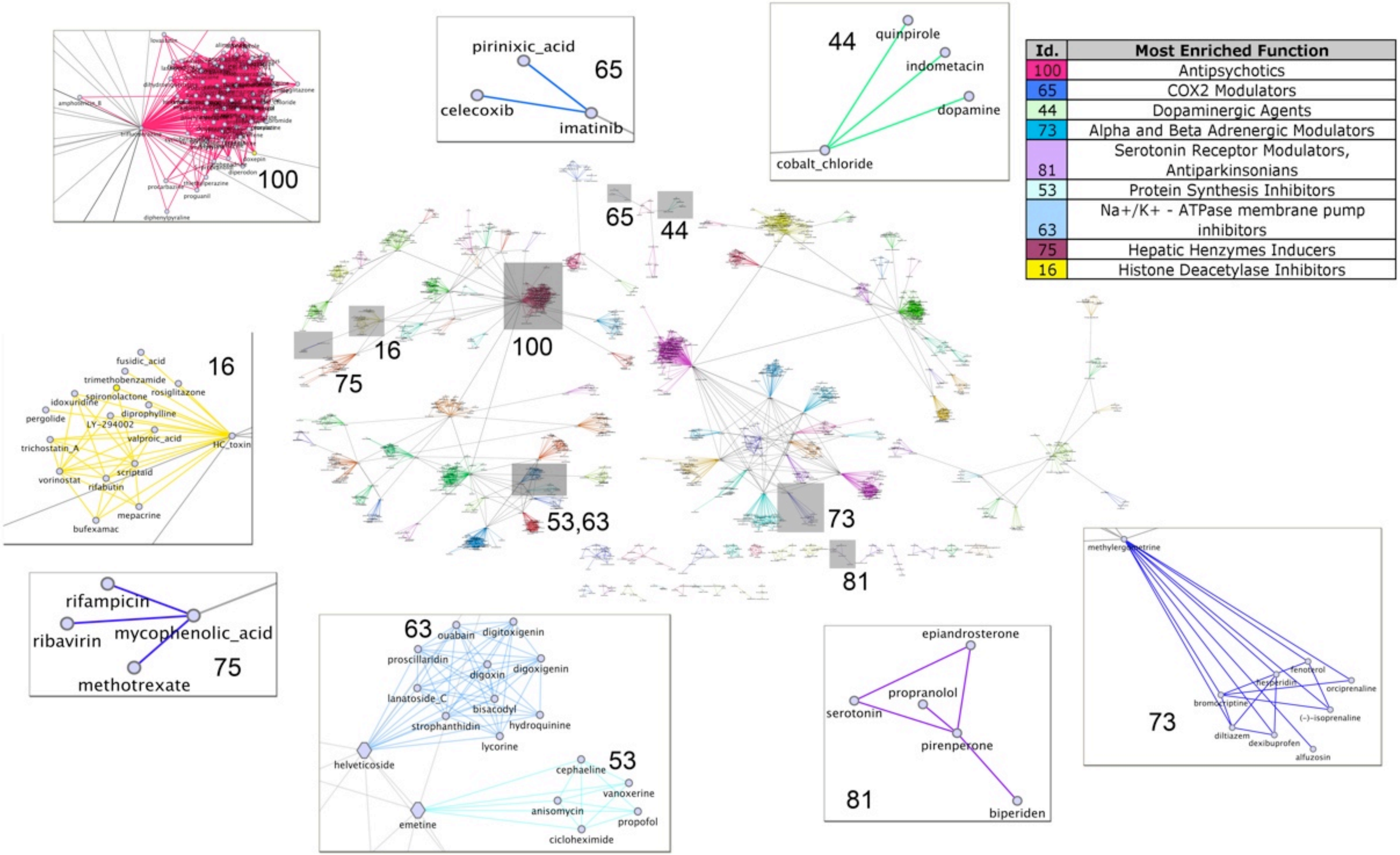


Community Identification by Hierarchical Clustering by Message Passing



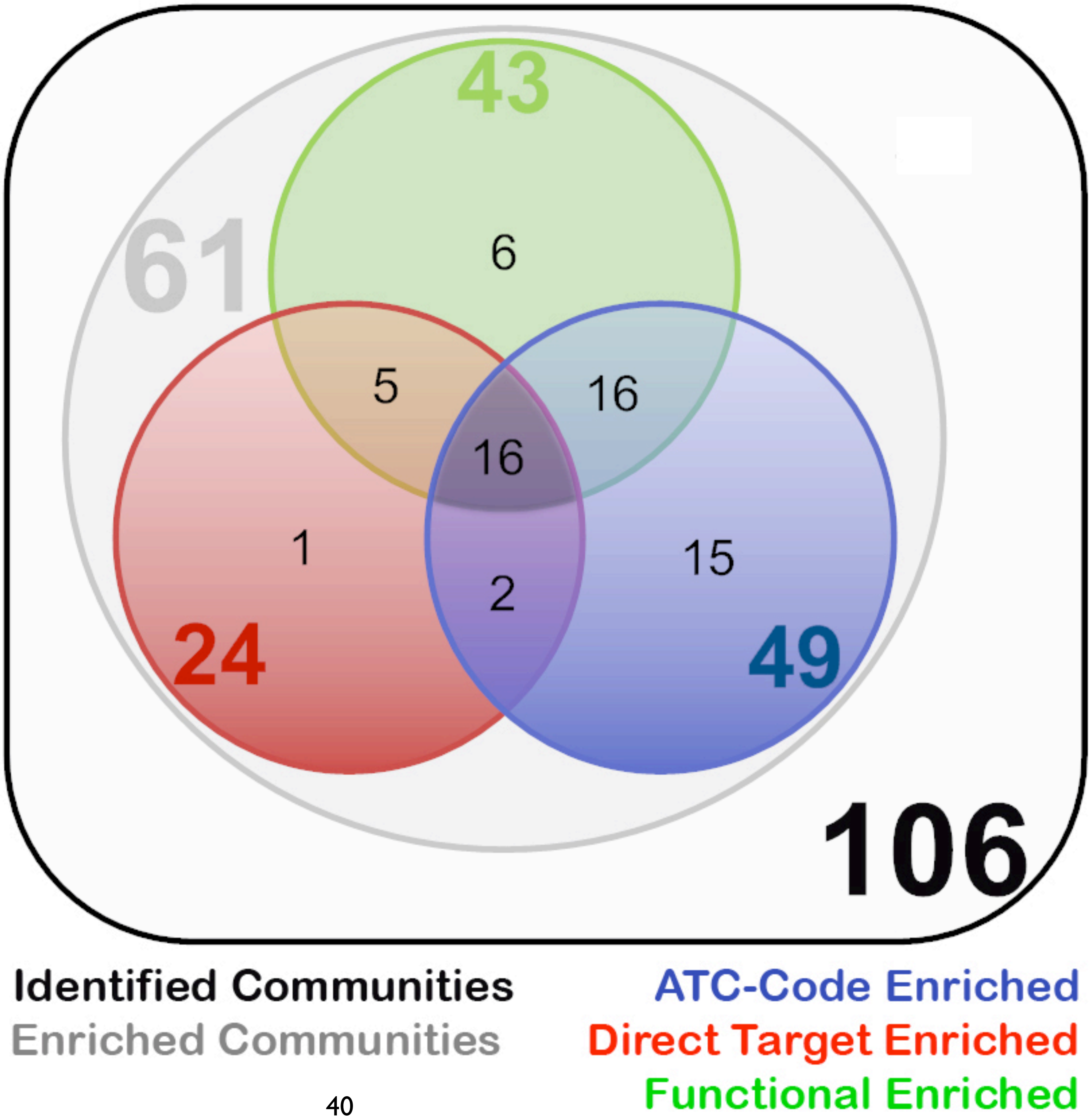
Frey et al., Science 2007

Community Validation

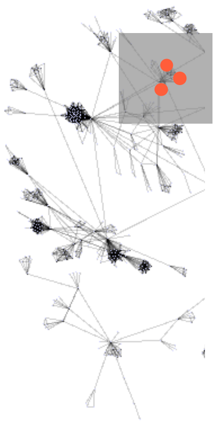


Id.	Most Enriched Function
100	Antipsychotics
65	COX2 Modulators
44	Dopaminergic Agents
73	Alpha and Beta Adrenergic Modulators
81	Serotonin Receptor Modulators, Antiparkinsonians
53	Protein Synthesis Inhibitors
63	Na+/K+ - ATPase membrane pump inhibitors
75	Hepatic Henzymes Inducers
16	Histone Deacetylase Inhibitors

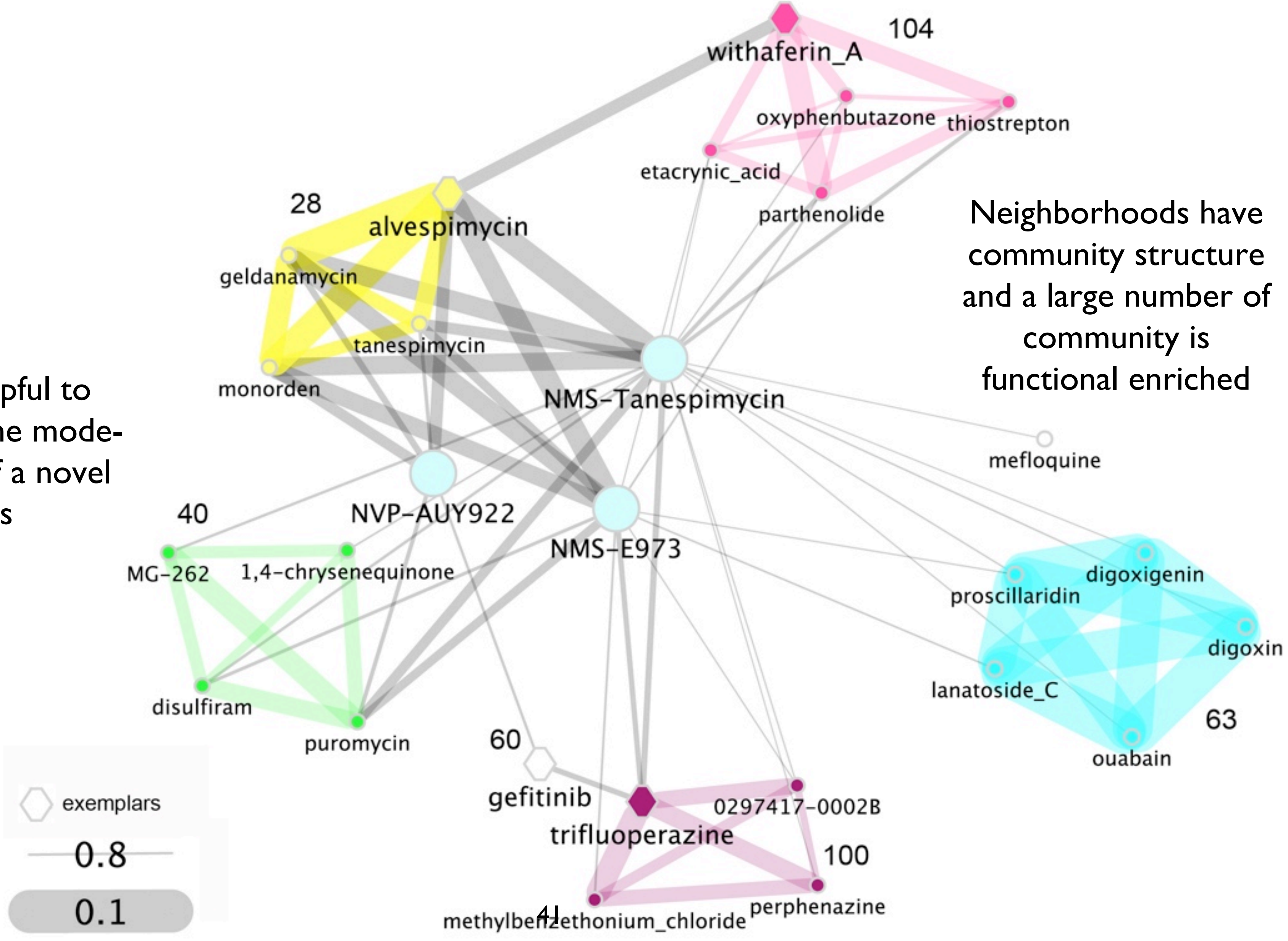
Community Enrichment Analysis



Experimental validation with three compounds:

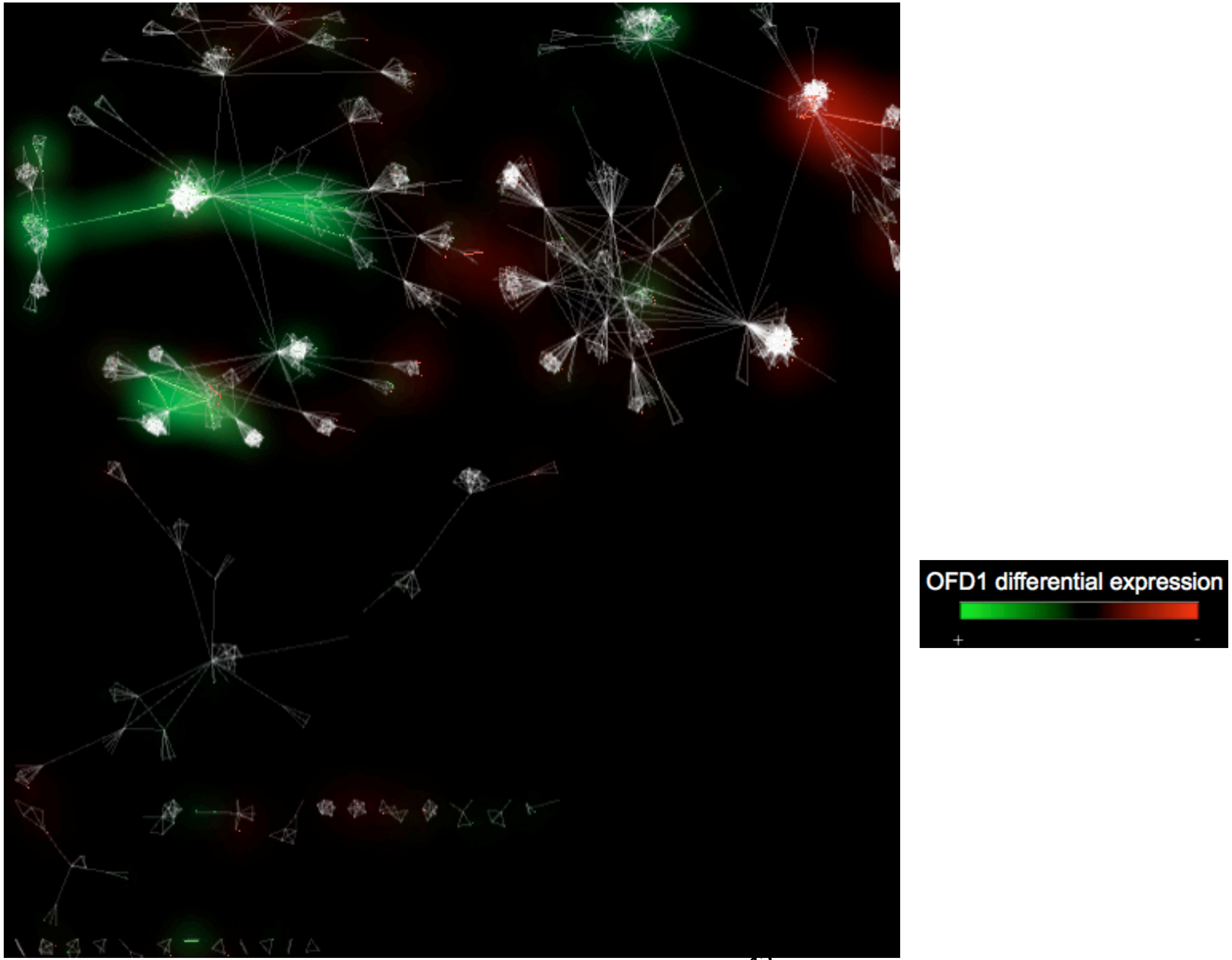


This is helpful to “recover” the mode-of-action of a novel drugs



Neighborhoods have community structure and a large number of community is functional enriched

Mapping gene changes due to drug treatment:



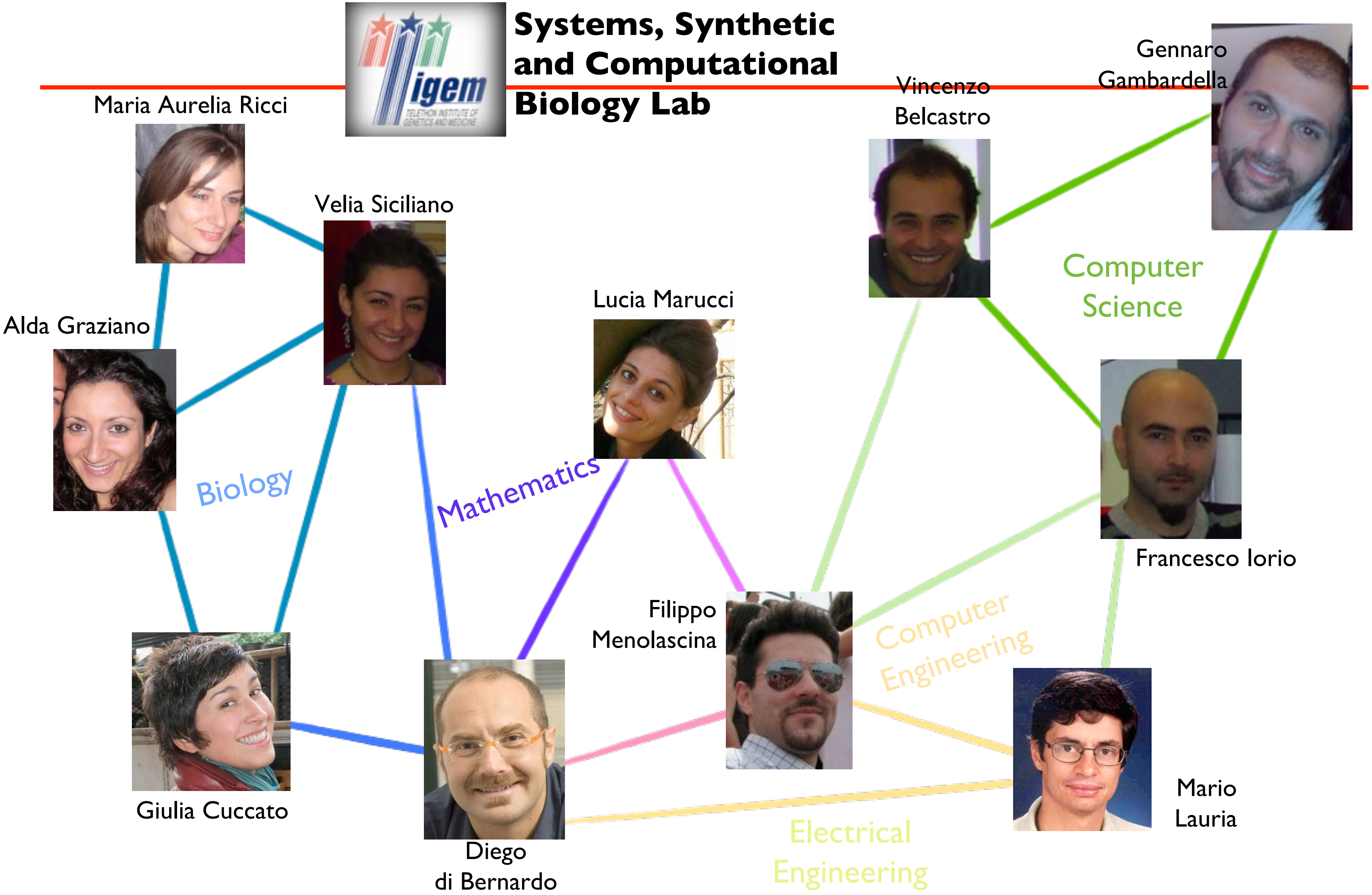
Part II: Conclusions

Drug Network = A novel, efficient tool to study drugs and their mode of action by gene expression profiling

- Performance assessment showed that 91% of tested compounds were correctly classified

- The modular structure of the sub-network that surrounds a new drug elucidates the MOA of the drug

Systems, Synthetic and Computational Biology Lab



Maria Aurelia Ricci

Velia Siciliano

Alda Graziano

Lucia Marucci

Vincenzo Belcastro

Gennaro Gambardella

Francesco Iorio

Filippo Menolascina

Giulia Cuccato

Diego di Bernardo

Mario Lauria

Roberta Bosotti, Antonella Isacchi
Nerviano Medical Sciences (Milano) - Italy