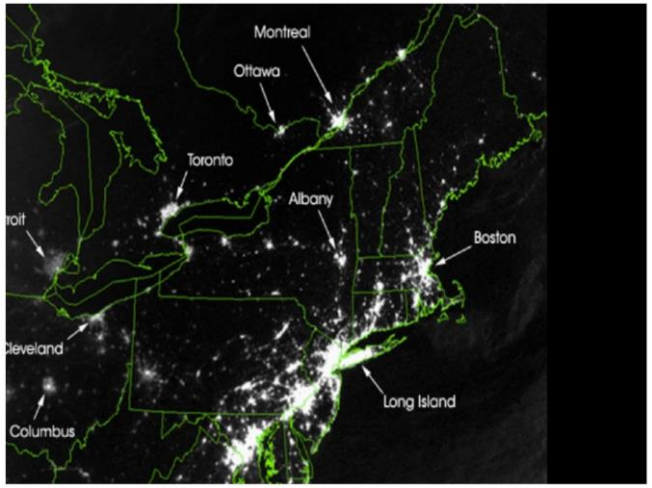
Introduction: networks and complex systems

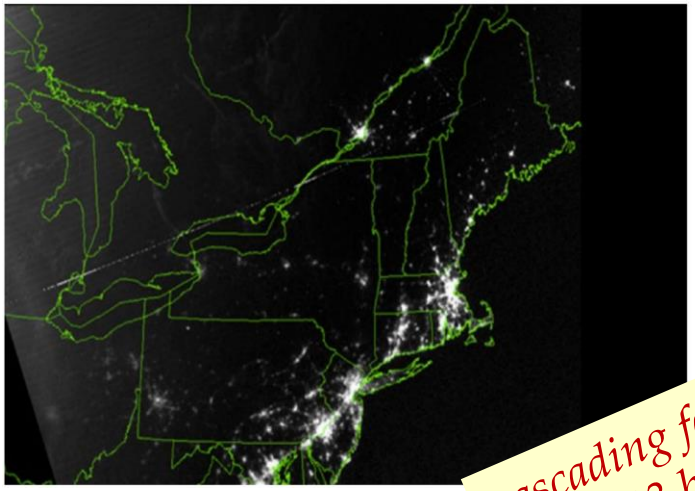
Alberto Paccanaro
EMAp – FGV

www.paccanarolab.org

Some material and images are from (or adapted from):
A. Barabási, and M. Pósfai. Network science, Cambridge University Press, 2016

Interconnectivity – advantages and vulnerabilities





cascading failure -- the 2003 blackout

We need to:

a) Understand the structure of the network

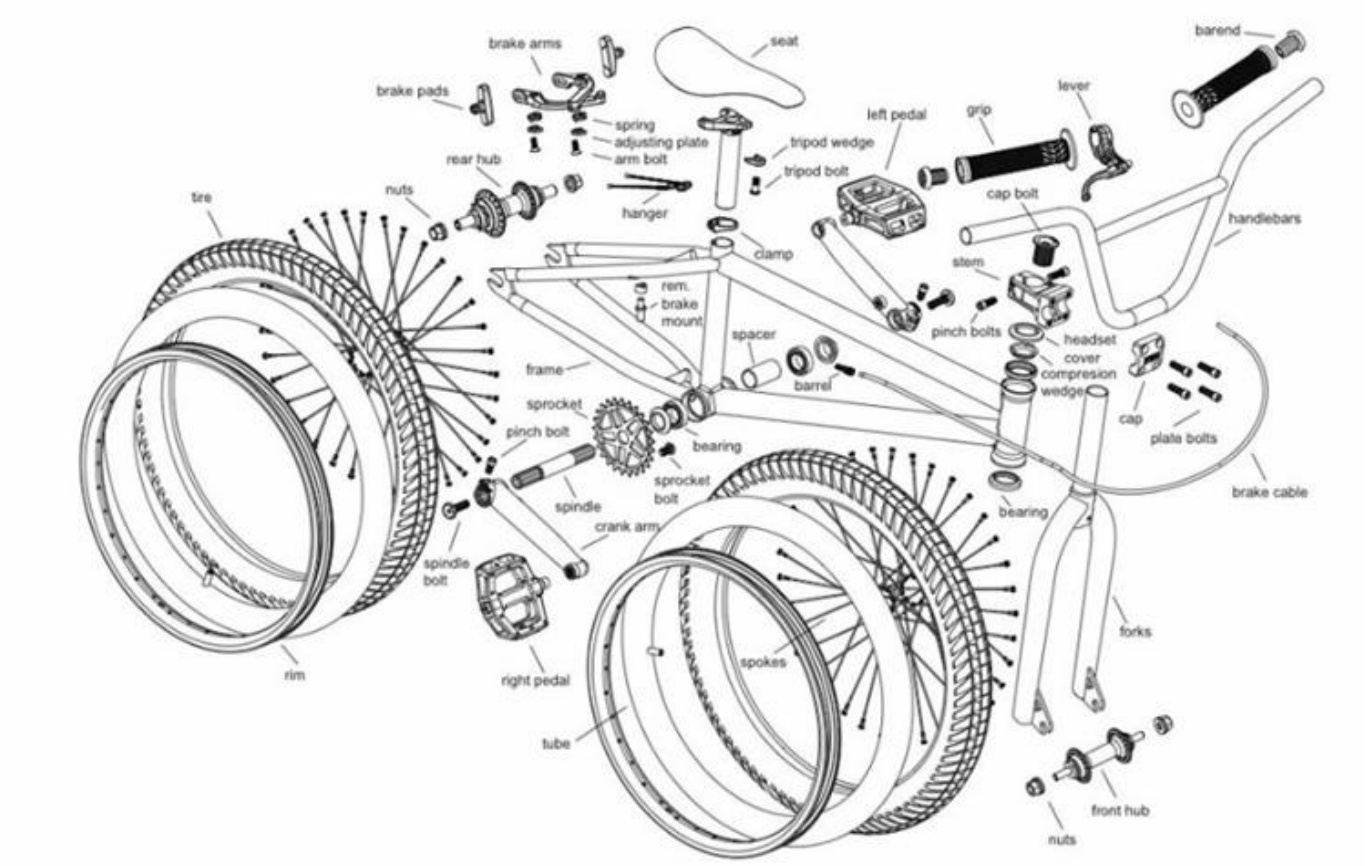
b) Be able to model dynamical processes on this networks

Interconnectivity induces non-locality

Complex systems

- Social networks
- Trade networks
- Communication infrastructure
- The brain
- Cell

It is difficult/impossible to derive the behaviour of a complex system from knowledge of the system's components.



Complex systems

- Social networks
- Trade networks
- Communication infrastructure
- The brain
- Cell

It is difficult/impossible to derive the behaviour of a complex system from knowledge of the system's components.

Despite the diversity of complex systems, their network structure and the evolution is driven by **common organizing principles**.

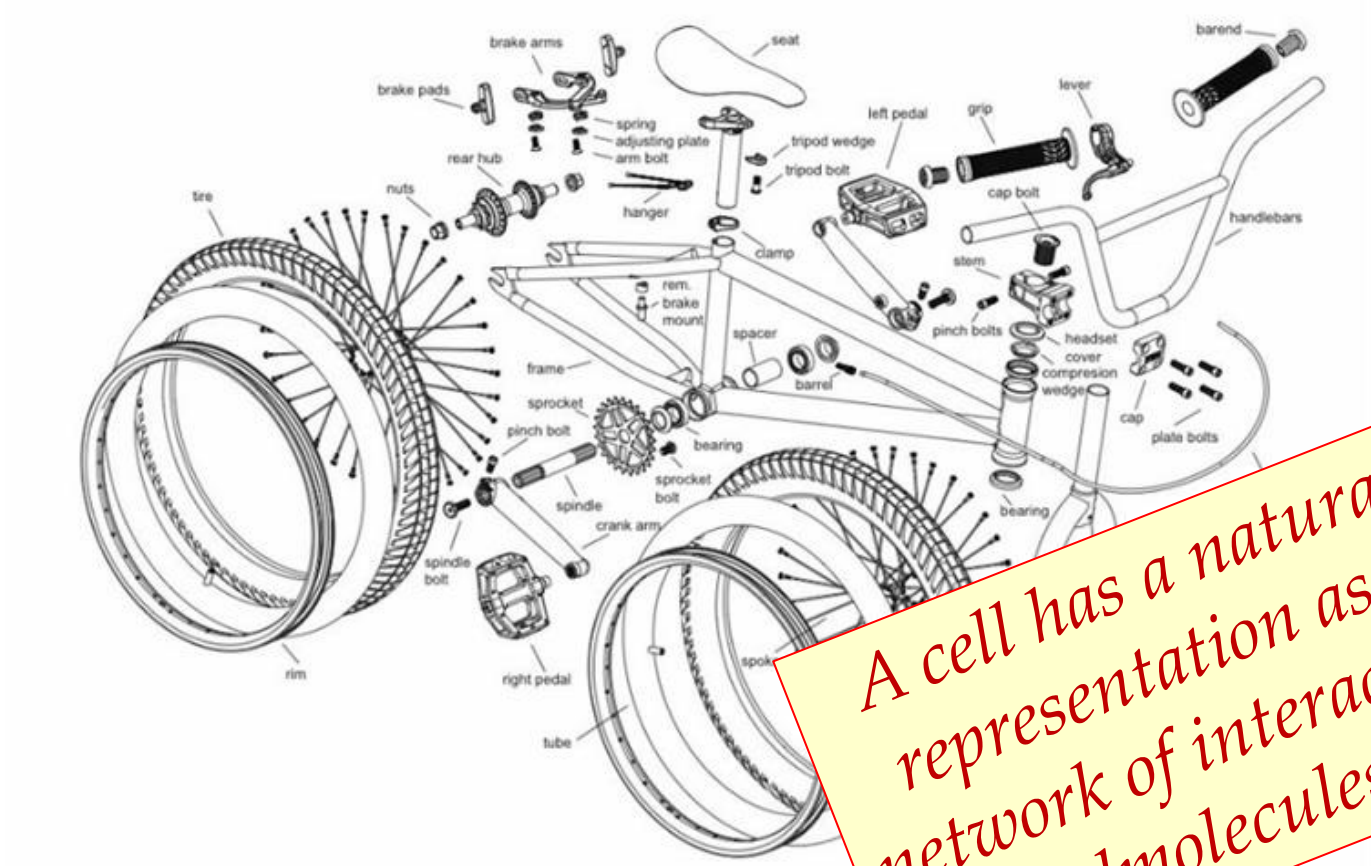
➔ we can use a common set of mathematical tools to explore these systems.

Chapter 1 discussed networks in Networks in different areas of science. Read it by yourself.

What do we do in my lab?
We focus on Molecular Biology, Medicine, Pharmacology.

www.paccanarolab.org

We need to analyse the cell at systems level



A cell has a natural representation as a network of interacting biolmolecules !

We need to analyse the cell at systems level

Large scale experiments interrogate the cell at the system level

Machine Learning:

- *Detect patterns in large amounts of very noisy data*
- *Integrate diverse sets of data from different sources*

In my lab, we develop Machine Learning methods for answering questions in Biology, Medicine, Pharmacology

- At the heart of our research is the **question, not the methodology – different areas of ML**
- **Diverse problems**
- Collaborate with **experimentalists and clinicians**
- We implement **software tools**
- **Explainable models**

Most of what we do can be phrased in terms of networks!

Problem	Type of Data	ML approach	Reference
De-noising of proteomics data	protein-protein interactions, mass spectrometry data (AP/MS)	Information diffusion over PPI networks	Havugimana et al, *Cell*, 2012
Quantifying the functional similarity between genes	protein seq., transcriptomics, proteomics	Random Walks over ontology structures (DAGs)	Caniza et al, *Bioinf.*,'14; Yang, Nepusz, Paccanaro *Bioinf*, 2012
Detection of protein complexes from protein interaction data	protein-protein interaction data (AP/MS and Y2H)	Overlapping clustering of large scale weighted graphs	Nepusz, Yu, Paccanaro *Nature Methods*, 2012
Selecting transcriptomics experiments for a given functional category	Transcriptomics, functional genomics	Supervised learning	Bhat, Yang, Paccanaro *PLoS ONE*, 2017
Protein function prediction	Any –omics network data	Semi-supervised learning	Torres, Romero, Yang, Paccanaro, *Nature Machine Intelligence*, 2021
Denoising of Hi-C data	Hi-C contact maps (networks)	Network modularity in random graphs	Ye, Paccanaro et al, *BMC Comp Biol, 2021*

PROBLEM	TYPE OF DATA	TYPE OF ML APPROACH	REFERENCE
Prediction of patients phenotype/outcome	transcriptomics, protein interaction network	Semi-supervised learning, feature selection	Gliozzo et al, *Nature Scient. Reports, 2020*
Prediction of drug cocktails against Chagas disease	genomics, transcriptomics, metabolic pathways	Supervised learning	Jimenez et al, *in preparation*
A measure of distance between hereditary diseases	disease phenotype description (text)	Random Walks over ontology structures (DAGs)	Caniza, Romero, Paccanaro, *Nature Scient. Reports*, 2015
Prediction of disease genes for uncharted diseases	protein interaction networks, disease genes	Semi-supervised learning	Caceres, Paccanaro, *PLoS Comp. Biol.*, 2019
Predicting the frequency of drug side effects	Drug side effects	Collaborative filtering (matrix factorization)	Galeano, Paccanaro *Nature Communications*, 2020
Drug Repurposing for COVID-19	protein interaction networks, transcriptomics, drug targets	matrix factorization and graph kernels	Santos, Torres, Galeano et al, *Cell Patterns, 2021*

ROADMAP for today

- **Networks in Systems Biology**

  Lab projects:

  - *Protein Function Prediction*
  - *Protein Complex prediction*

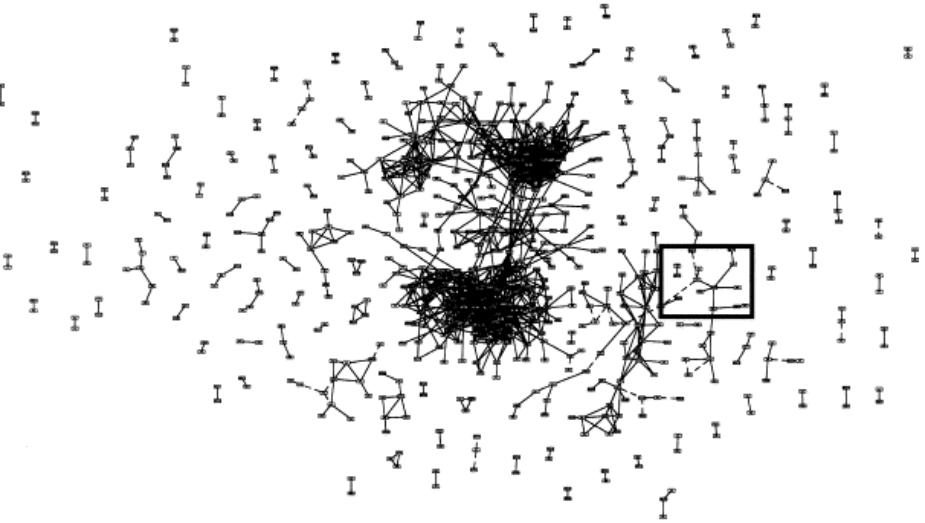
- **Networks in Medicine & Pharmacology**

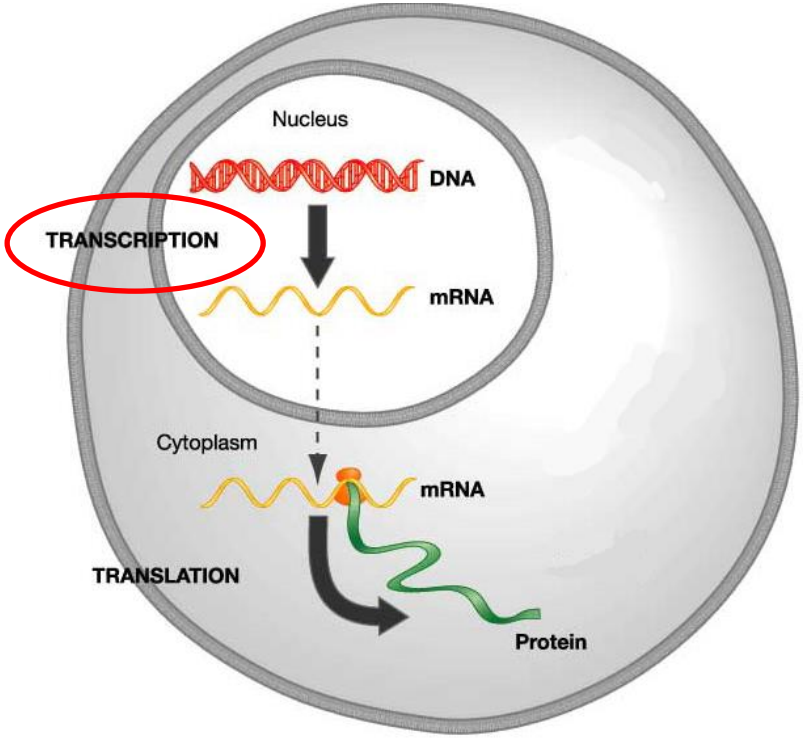
  Lab Projects:

  - *Disease gene prediction*
  - *Predicting drugs for repurposing for COVID-19*

Networks in Systems Biology (biological networks)

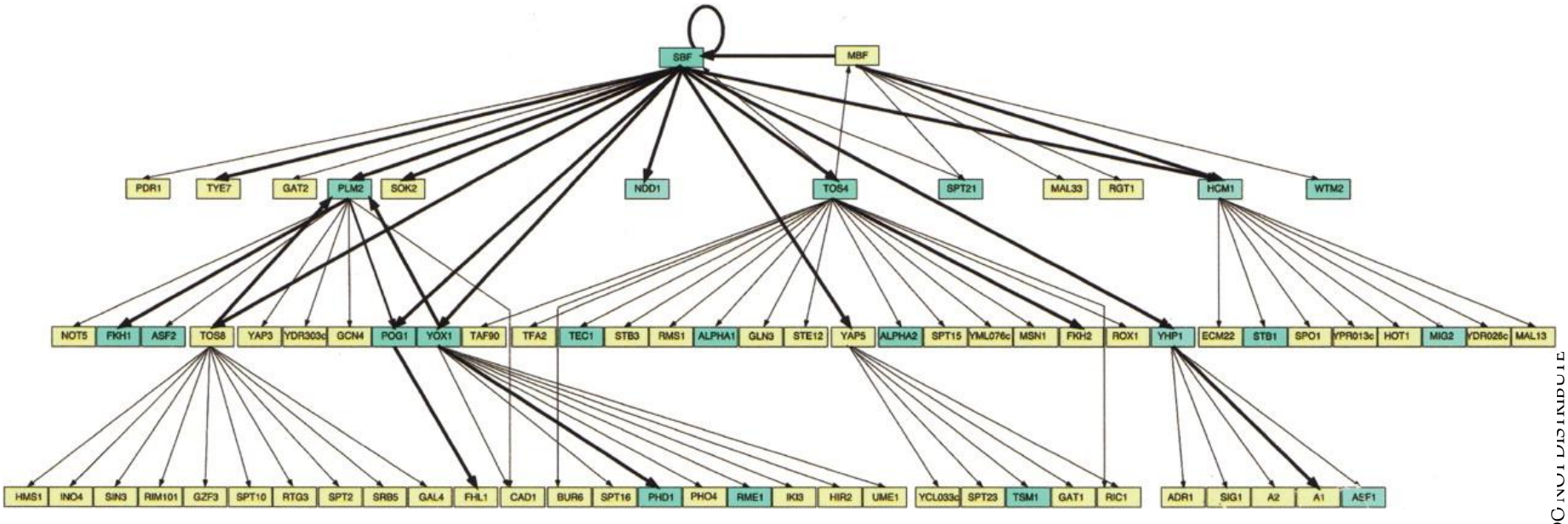
Expression networks





[Qian, et al, J. Mol. Bio., 314:1053-1066]

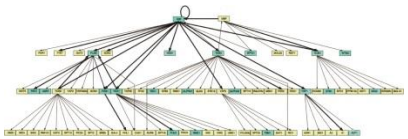
Regulatory networks



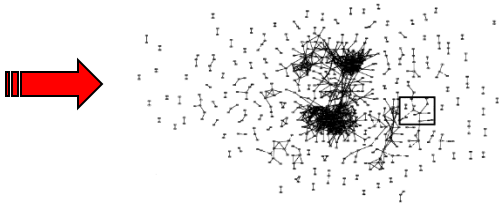
[Horak, et al, Genes & Development, 16:3017-3033]

Nodes are either proteins or a putative DNA regulatory element and directed edges represent:

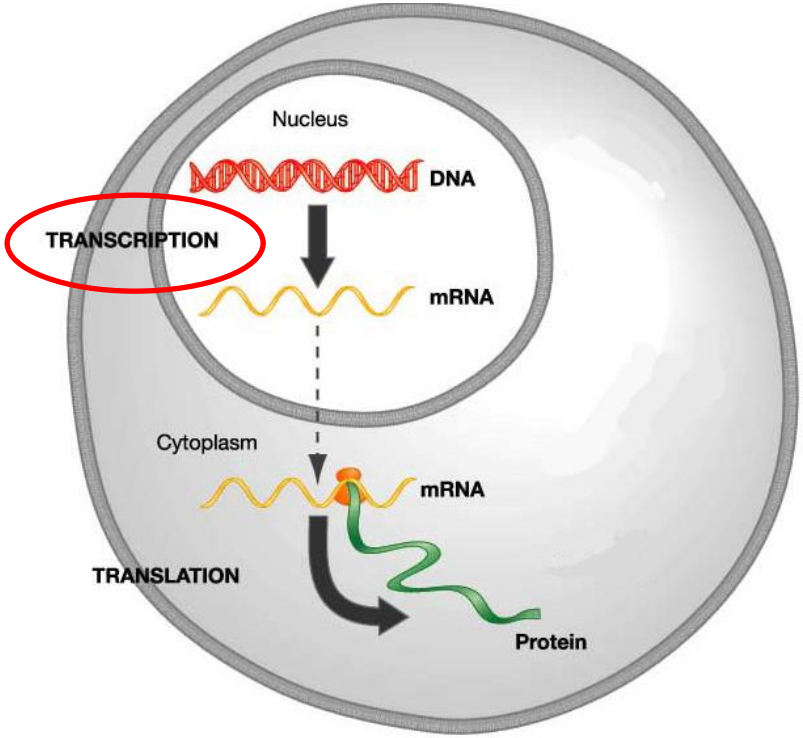
1. **Regulatory relationships**
2. **Post-translational modifications**



Regulatory networks



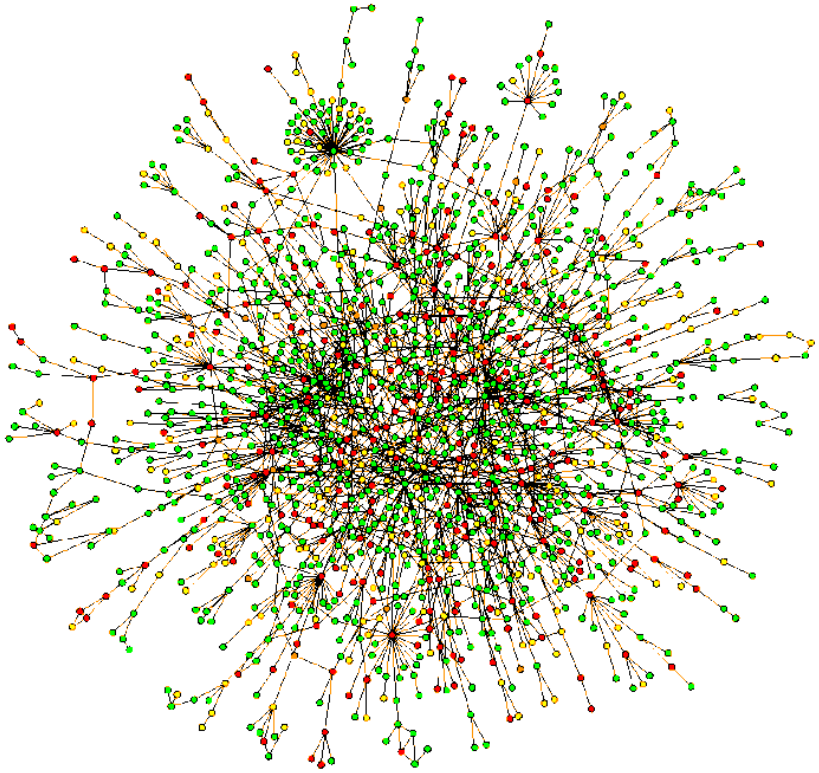
Expression networks

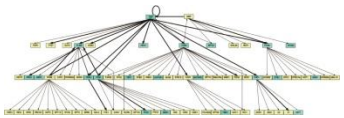


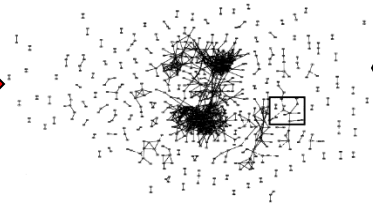
Protein-Protein Interaction Networks

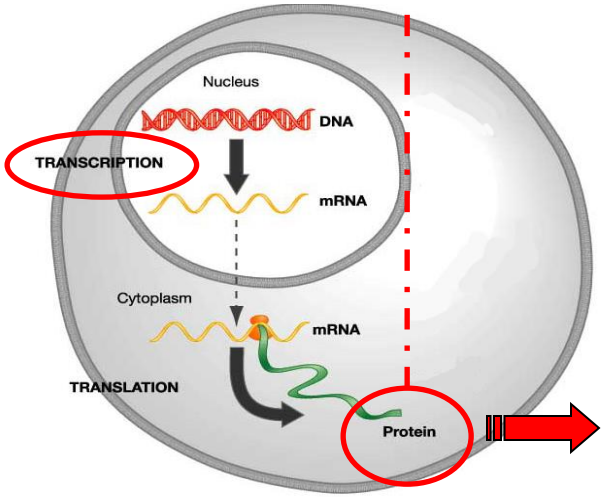
Nodes represent proteins and edges represent a physical interaction between two proteins.

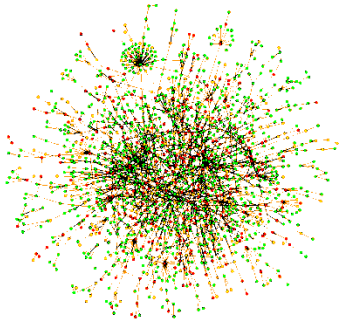
Edges are non-directed











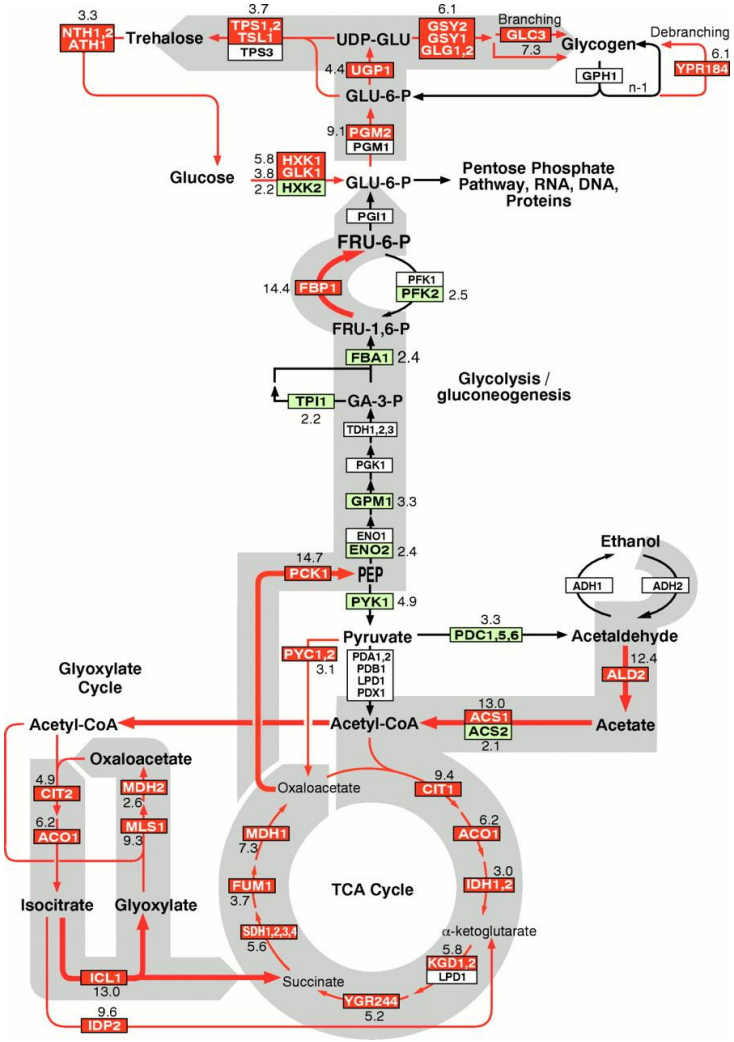
Regulatory networks

Expression networks

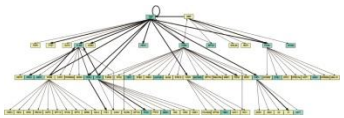
Interaction networks

Metabolic networks

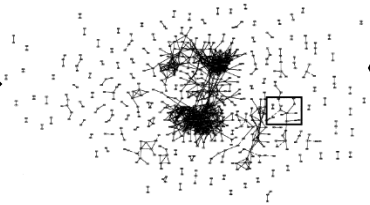
Metabolic network maps attempt to comprehensively describe all possible biochemical reactions for a particular cell or organism



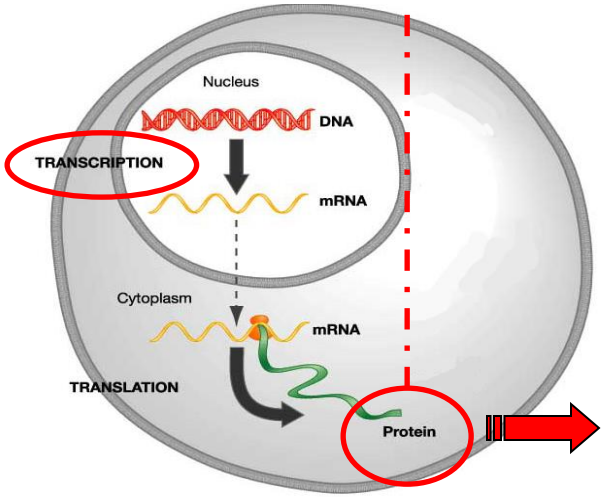
[DeRisi, Iyer, and Brown, Science, 278:680-686]



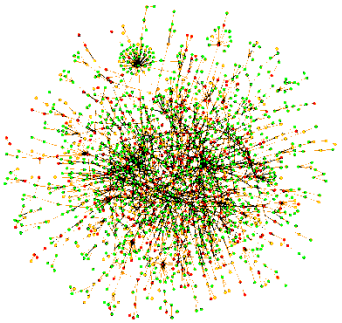
Regulatory networks

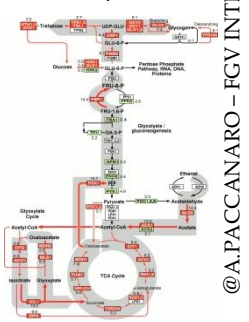


Expression networks



Interaction networks

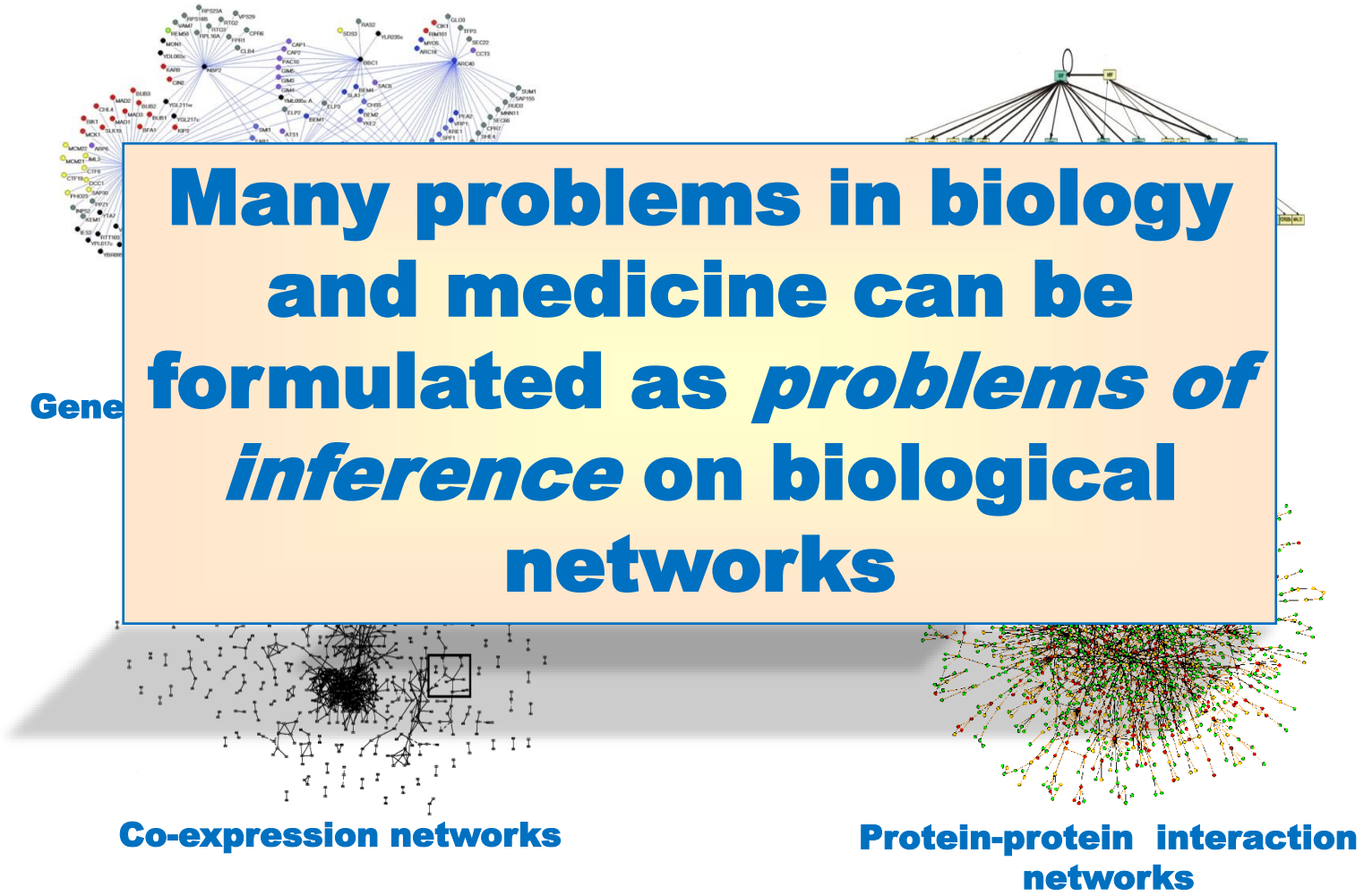




Metabolic networks

Biological networks

Cell as webs of interactions between biomolecules
Experimental data have a natural representation as networks



[Horak, Genes & Dev.; DeRisi, Science; Qian, J. Mol. Bio; Jeong, Nature,;Tong, Science; Goh, PNAS]

When I look at biological networks in terms of principles from network science, what do I see?

The organizing principles of Biological Networks

The organizing principles of Biological Networks

- Modules: high degree of clustering, implying the existence of topological modules – highly interlinked regions.

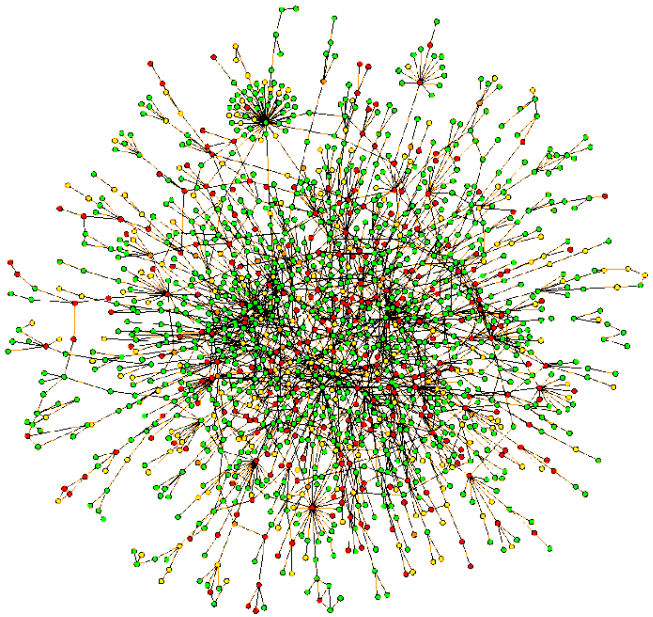
- Degree distribution: the degree distribution $P(k) \sim k^{-\gamma}$ – it does not follow a Poisson distribution…

- Hubs: few highly connected hubs hold the whole network together.

In protein interaction networks:

- hub proteins tend to be encoded by essential genes
- genes encoding hubs are older and evolve more slowly

- *'party' hubs*: interact with most of their partners simultaneously
- *'date' hubs*: bind different partners at different locations and times



- Small world phenomena: relatively short paths between any pair of nodes.

- Motifs: Some subgraphs in biological networks appear more (or less) frequently than expected

In regulatory networks:

- bottlenecks (nodes with high betweeness centrality) tend to correlate with essentiality.

Two projects in Systems Biology from the lab

1. Protein Function prediction

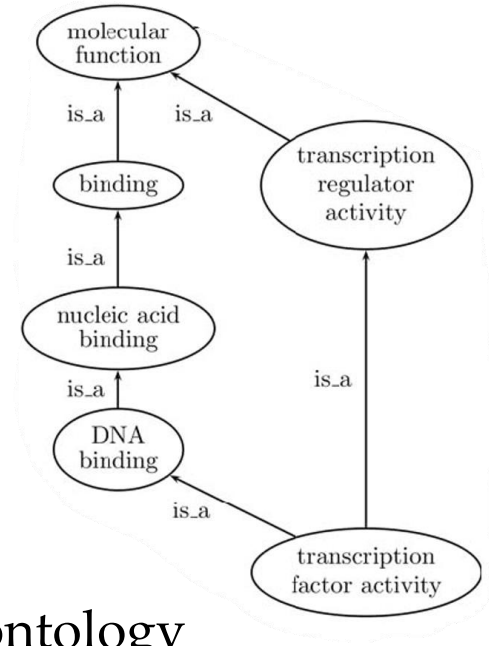
What do the proteins do?

Mateo Torres

One of the "Holy Grails" of computational biology

A rough sketch of the problem:

- *Classification problem: ~ 40,000 classes*
- *Bacteria: ~ up to 5,000 proteins*
- *Eukaryotes: ~ up to 25,000 proteins*
- *Multi-label*
- *Classes have relations*

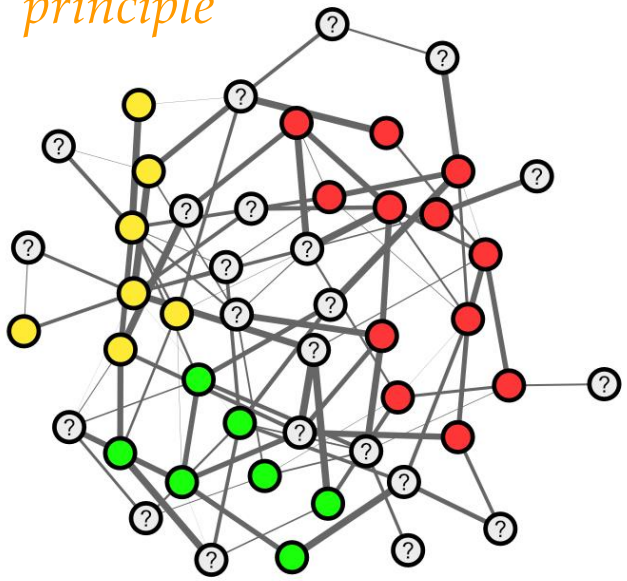


Gene ontology

Available data for PFP

"Unary" information
Similarity with known proteins
Motifs
Structure
…

*The "**guilt by association**" principle*

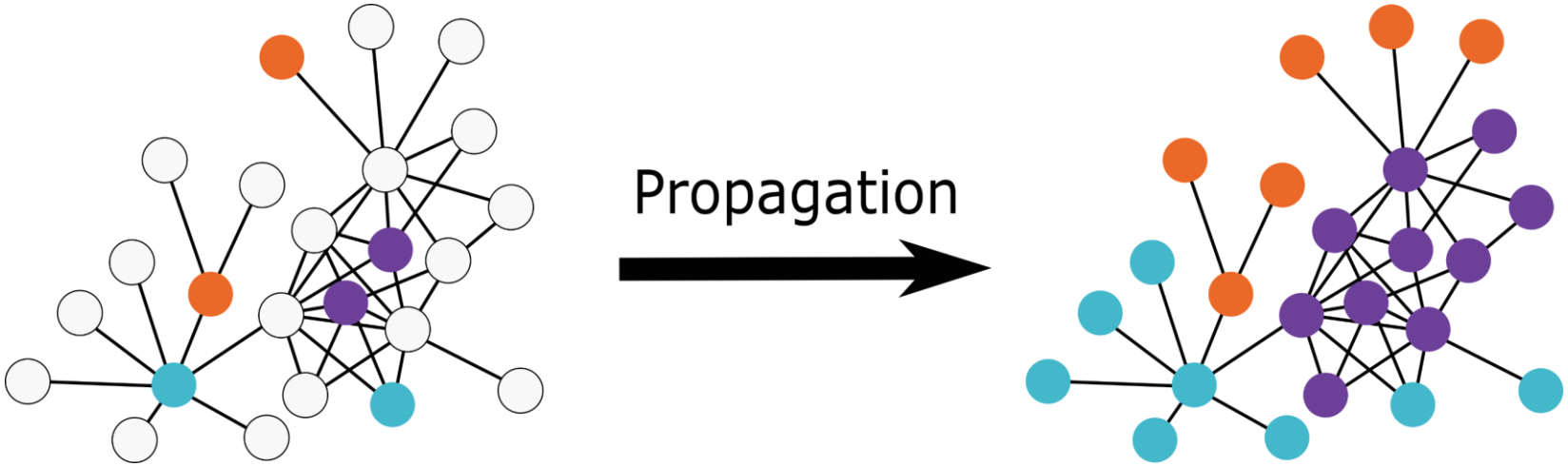


- *Proteins that are "close" in the graphs tend to have similar function*
- *Proteins form "functional **communities**"*

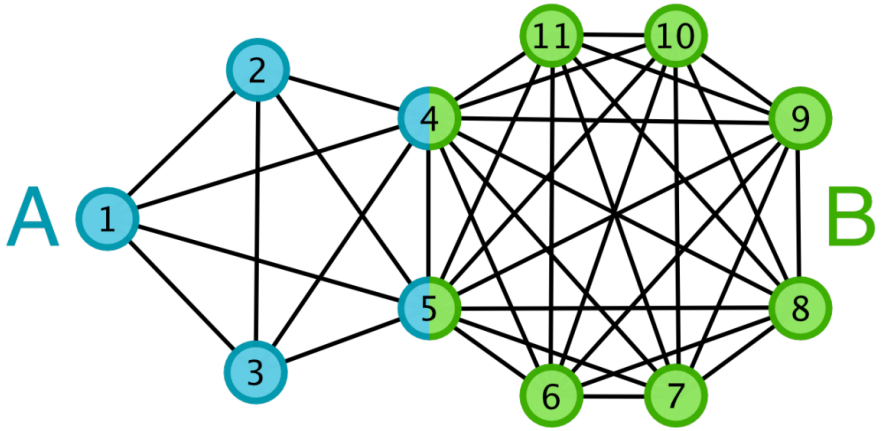
Thus, the problem can be framed as a *semi-supervised learning* problem

Protein function prediction in a semi-supervised setting

To predict function, we propagate the initial labelling, from *unary information*, onto the network built from *binary information*.



A new semi-supervised learning algorithm, that exploits the community structure



Previous algorithms did not exploit the ***community*** *structure*

The initial assignment should be kept (fitting constraint)

Neighbours should have similar label (smoothness constraint)

$$Q(F) = \sum_{i=1}^{n} (F_i - Y_i)^2 + \frac{\lambda}{2} \sum_{i=1}^{n} \frac{1}{d_i} \sum_{j=1}^{n} J_{ij} W_{ij} (F_i - F_j)^2$$

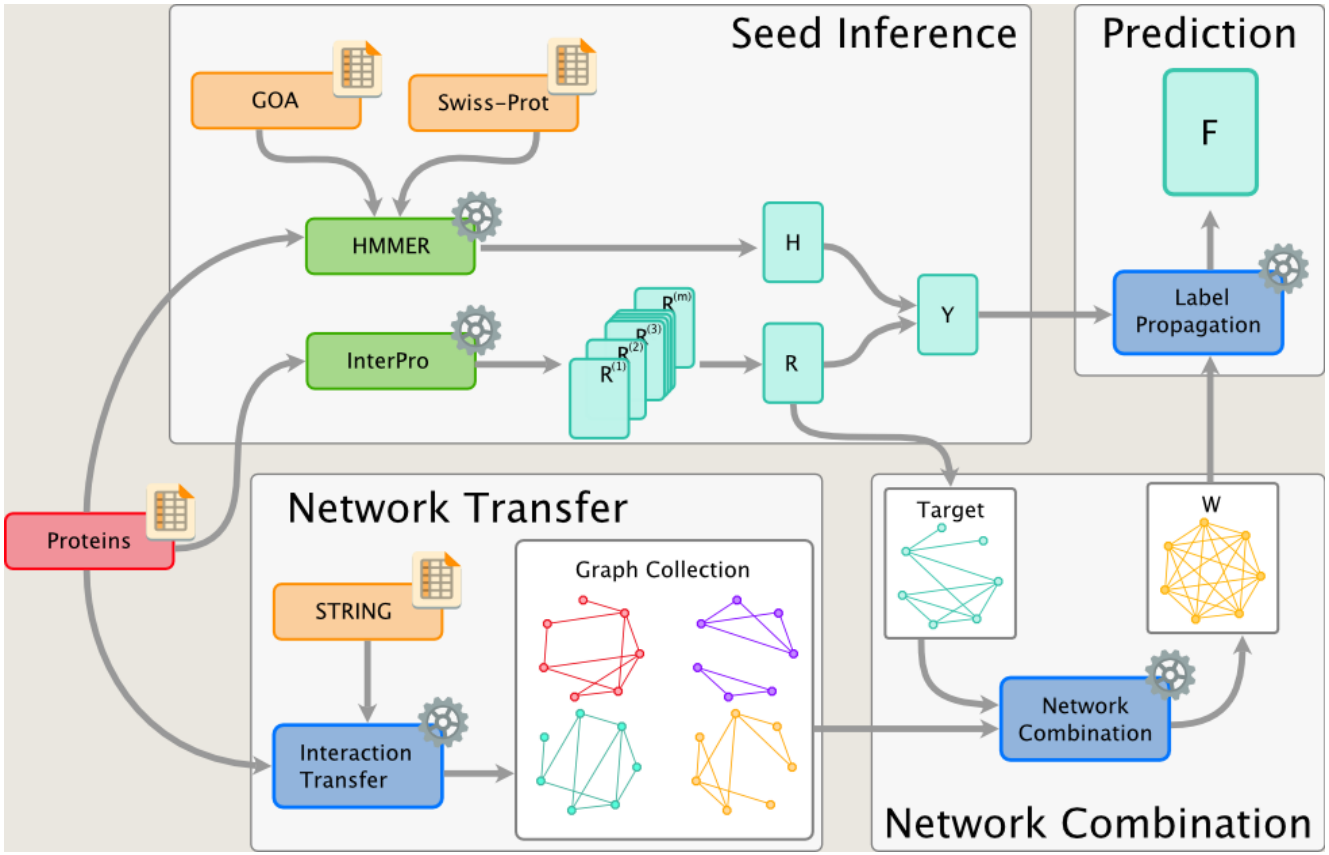
$\boldsymbol{F}$ assignment vector
Y_i known labels
W graph adj matrix
d_i degree of node i

$$J_{ij} = \frac{\sum_k W_{ik} W_{jk}}{\sum_k W_{ik} + \sum_k W_{jk} - \sum_k W_{ik} W_{jk}}$$

$$F^* = \arg\min_F Q(F)$$

S2f – Sequence to Function

Torres, Yang, Romero, Paccanaro,
Nature Machine Intelligence, 2021



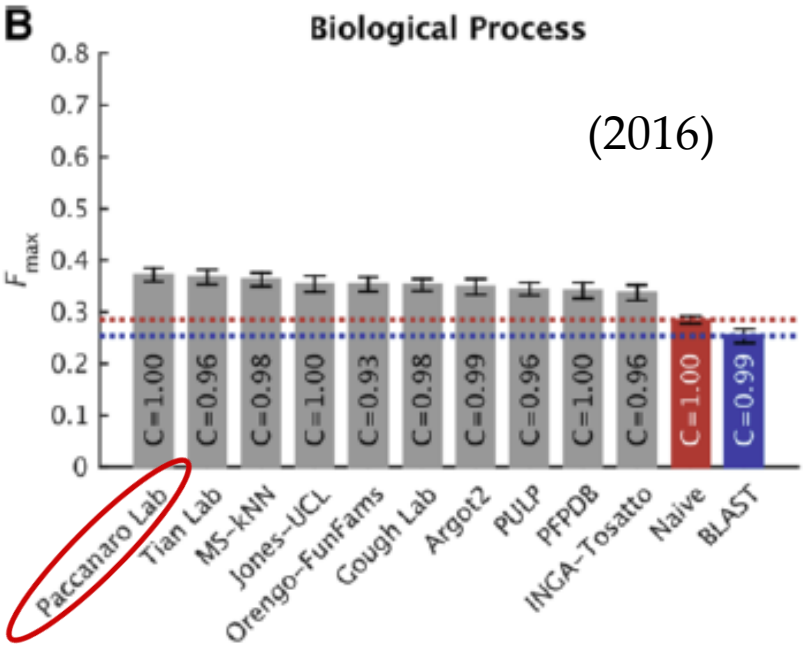
The best system for predicting function in microbes

Genome Biology

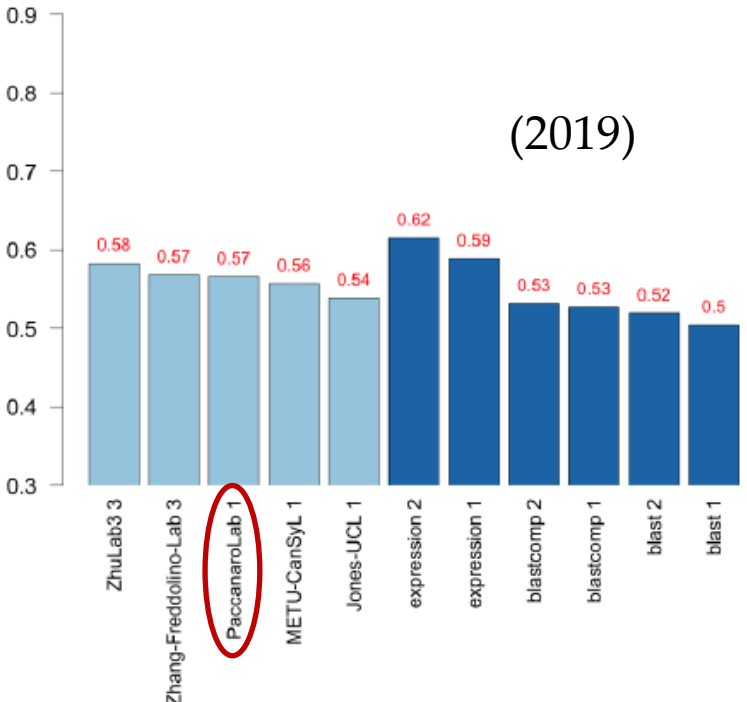
nature methods

Top performer at CAFA competitions

Explainable !



(2016)



(2019)

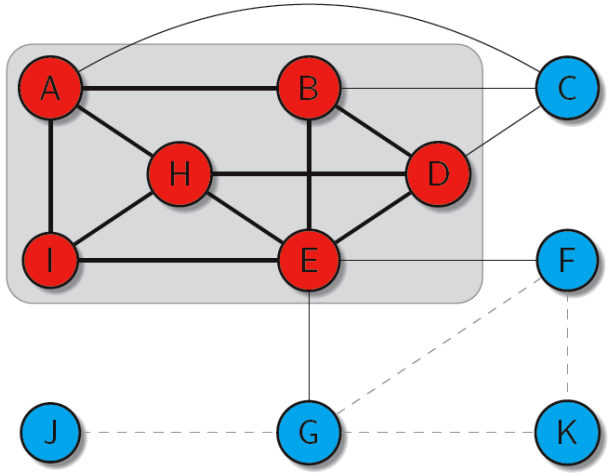
2. Protein complex detection

Tamas Nepusz

We developed an overlapping clustering algorithm for large weighted networks
(ClusterONE: Clustering with Overlapping Neighbour Expansion)

A rough sketch of the problem:

- *Large network (up to 20,000 nodes)*
- *Links are weighted*
- *We need to identify areas that are more tightly connected*
- *A node can belong to more than one group*



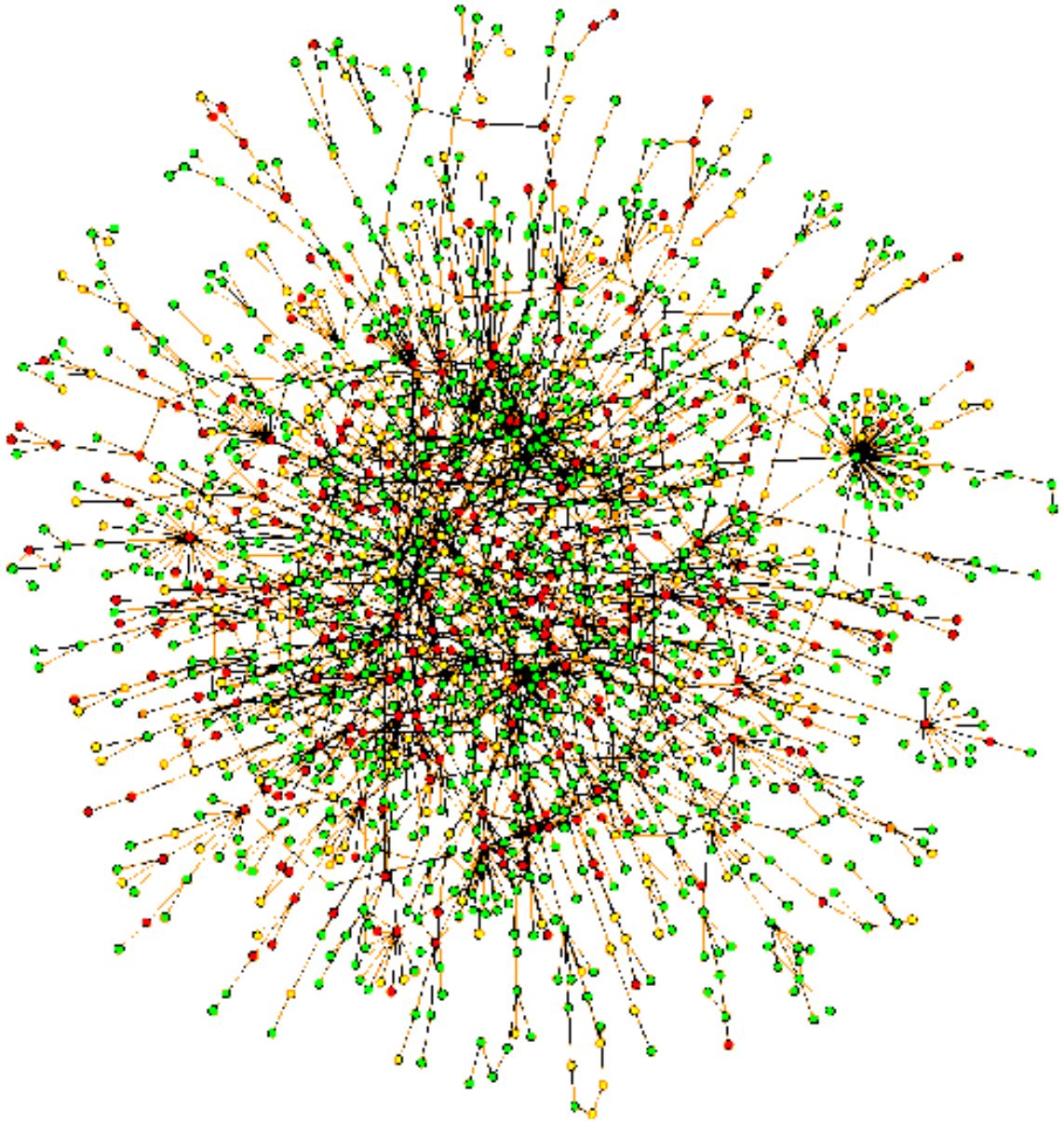
Two structural properties:

a. **contain many reliable links**

b. be **well-separated** from the rest

Cohesiveness

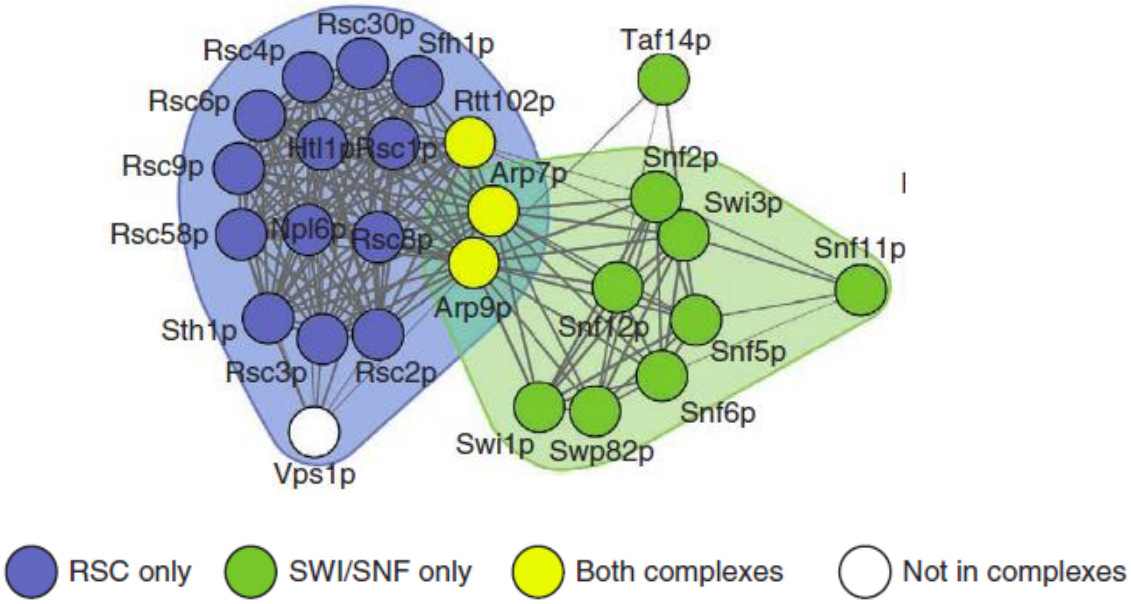
$$f(V) = \frac{w^{in}(V)}{w^{in}(V) + w^{bound}(V) + p|V|}$$



nature methods

One of the most used method for protein complex detection

Explainable !

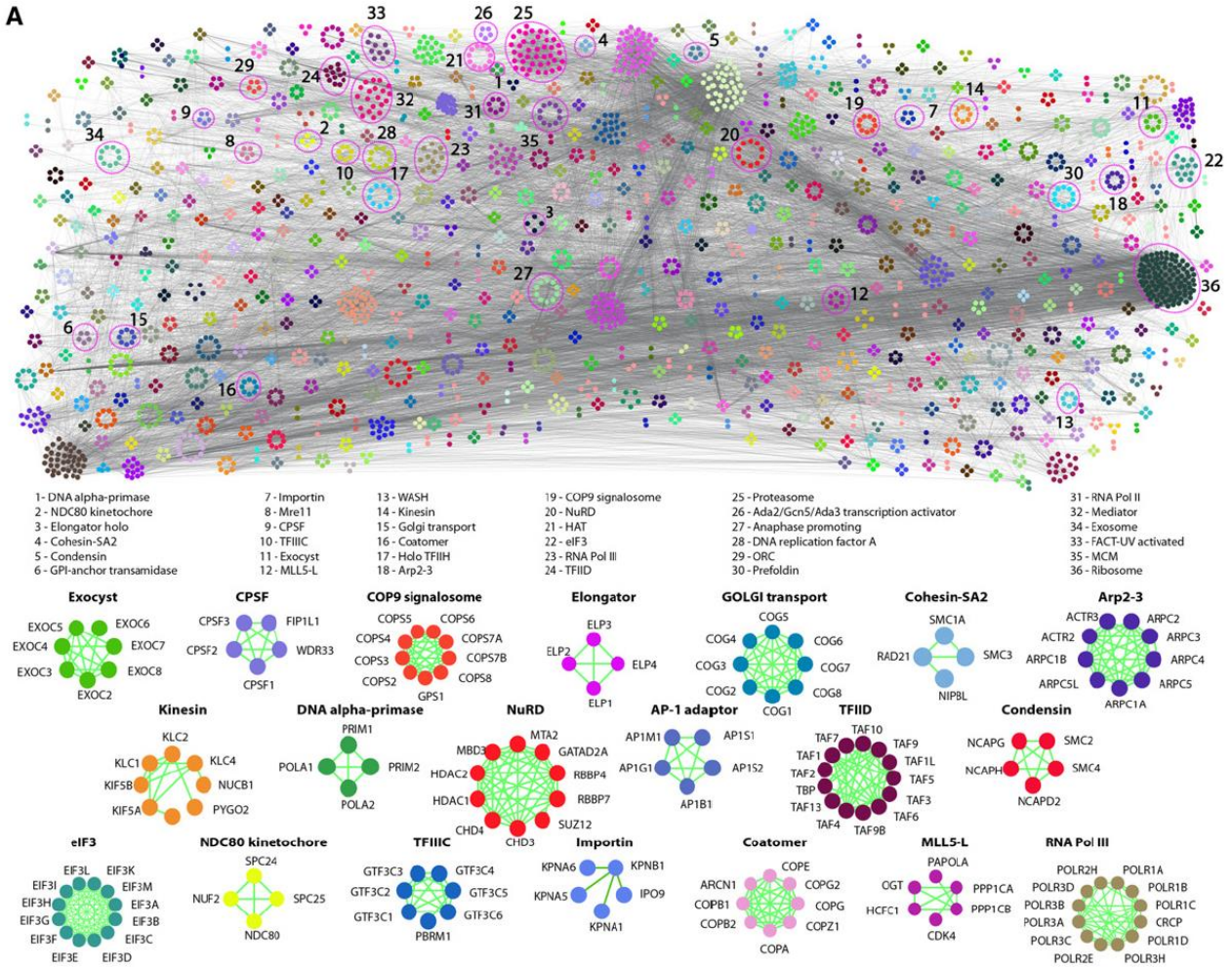


scales up to graphs containing millions of vertices and edges

Over 20,000 downloads !

A Census of Human Soluble Protein Complexes

Paccanaro (RHUL), Marcotte (UTAustin), Emili (Utoronto)



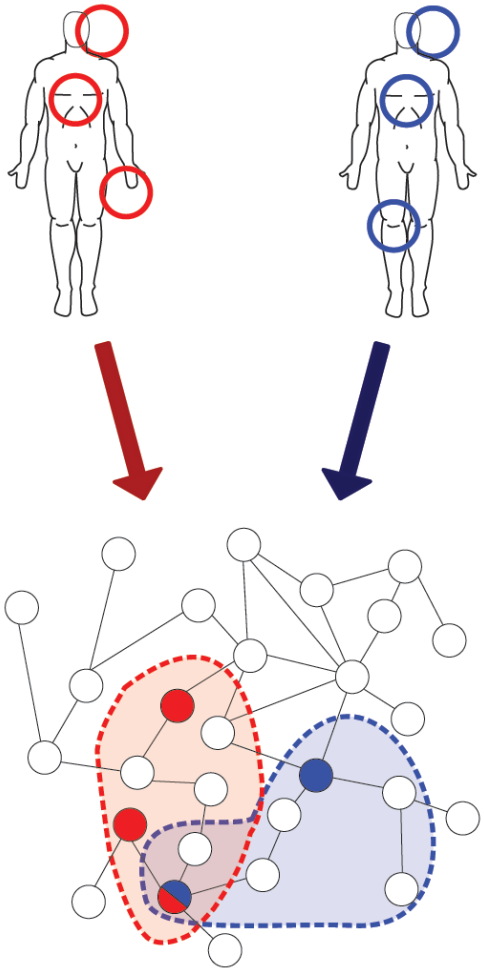
Largest set of human protein complexes to date

Networks in Medicine

When I map our current knowledge of human disease onto the human biological networks, and I analyze it in terms of principles from network science, what do I see?

The principles of Network Medicine

The principles of Network Medicine



Disease genes tend to avoid hubs and segregate at the functional periphery of the interactome.

Genes associated with a specific disease tend to cluster in the same neighbourhood – the disease module

The disease modules of diseases that are phenotypically similar tend to be located in closeby regions of the interactome.

A drug, binding to a target, also causes a perturbation on the human network that will propagate.

***Network Medicine*: diseases as perturbations of molecular networks**

Two projects in Network Medicine from the lab

3. Disease gene prediction

Predicting causative genes for hereditary diseases

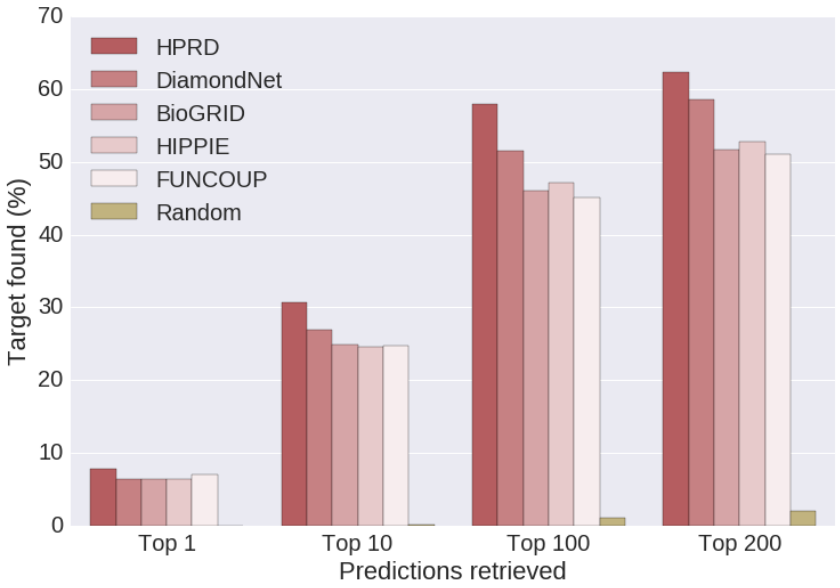
PLoS COMPUTATIONAL BIOLOGY
An official journal of the International Society for Computational Biology
www.ploscompbiol.org

Our method:

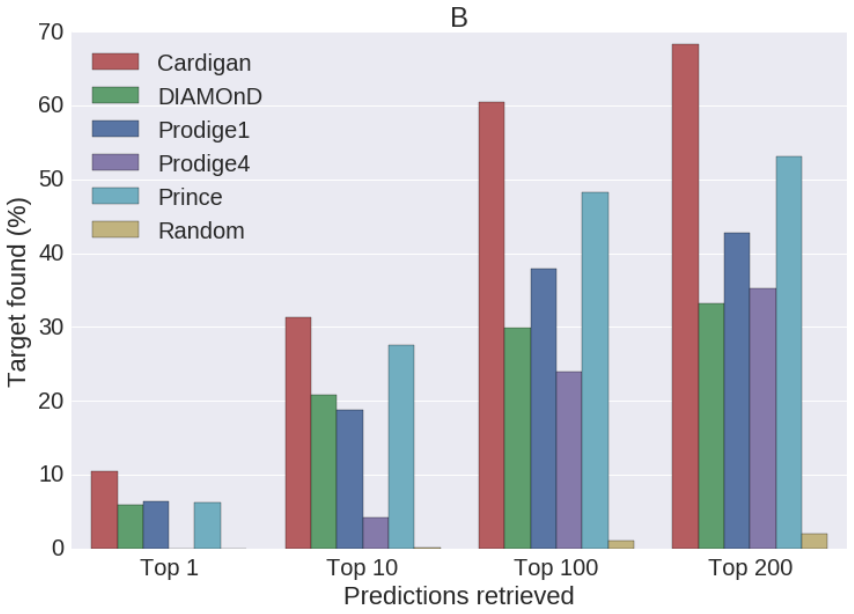
- *text mining of scientific papers*
- *semi-supervised learning algorithm*

Explainable !

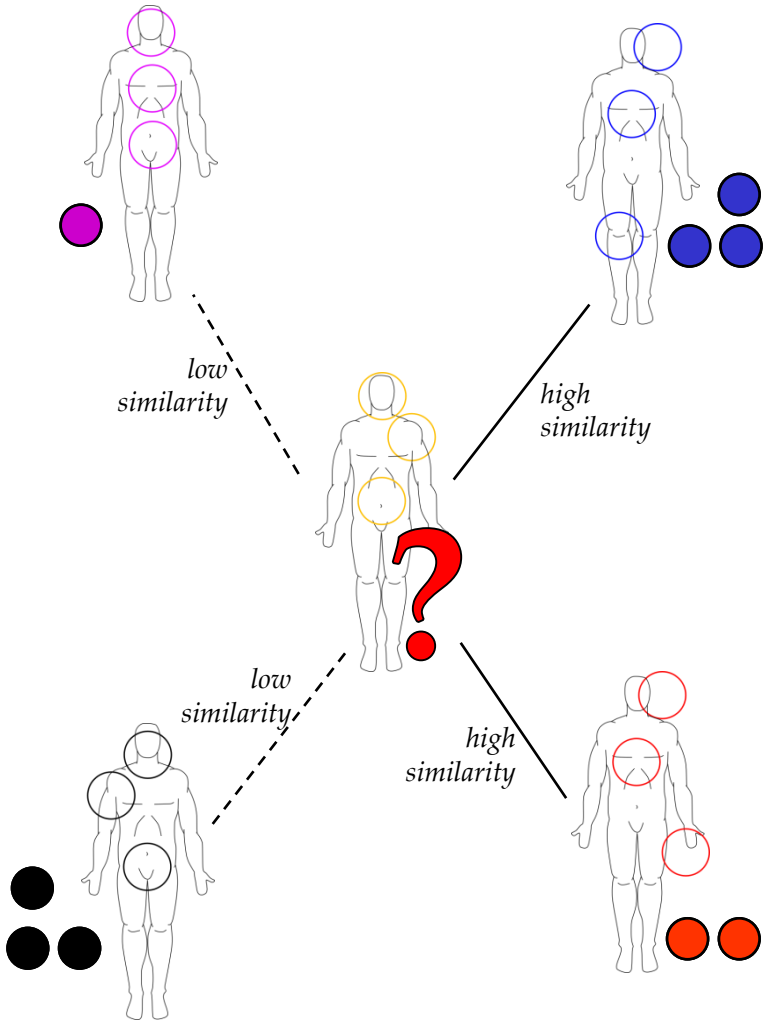
The first method that can predict disease genes for uncharacterized diseases

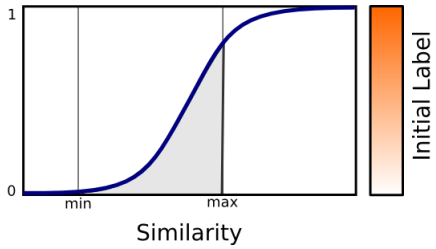


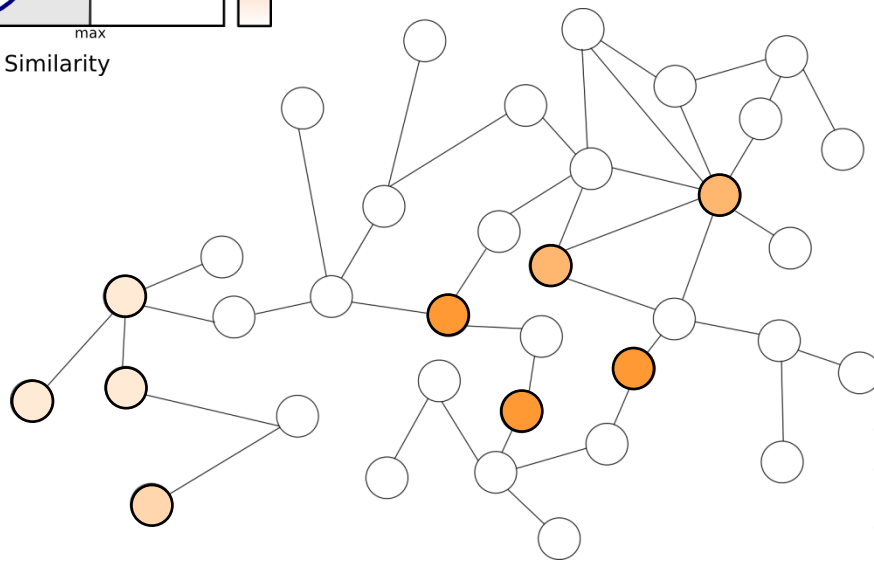
uncharacterized diseases



characterized diseases



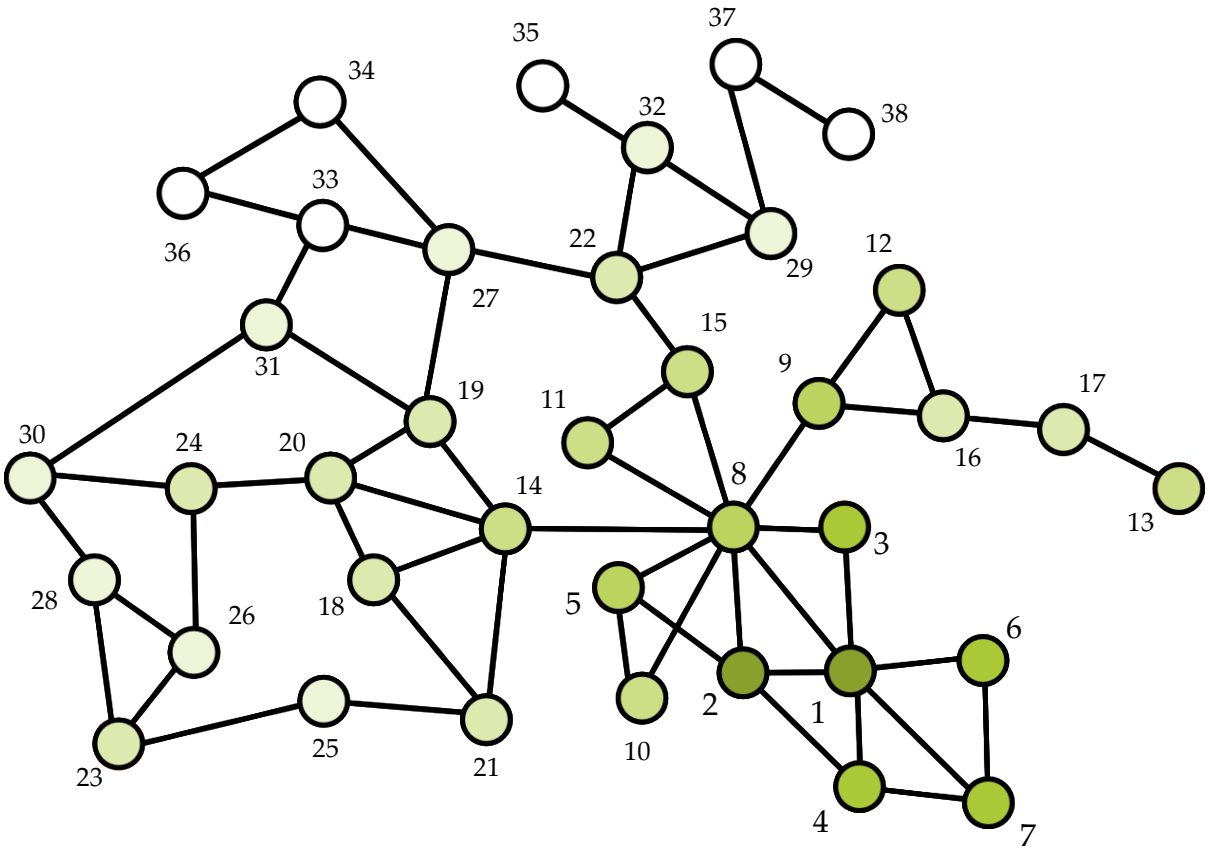




1. Calculate the similarity between our uncharted disease and each charted disease
2. Place known genes in the interactome.
3. Learn a *similarity-to-label* mapping
4. Assign a *"soft"* label to the disease genes
5. Diffuse the soft labels

Diffusing soft labels (semi-supervised learning)

For a given disease, the soft label of each gene is related to the ***probability for that gene to be a disease gene for that disease****.*



4. Drug repurposing for COVID-19

Predicting which drugs could be re-used against the disease

Conacyt project
Luca Cernuzzi
Maria del Mar Sanchez
Aldo Galeano
Rafael Adorno

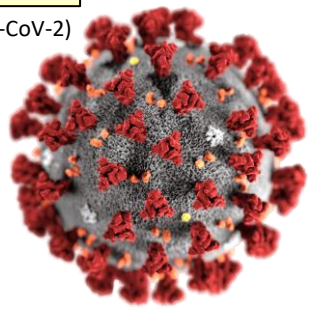
Mateo Torres

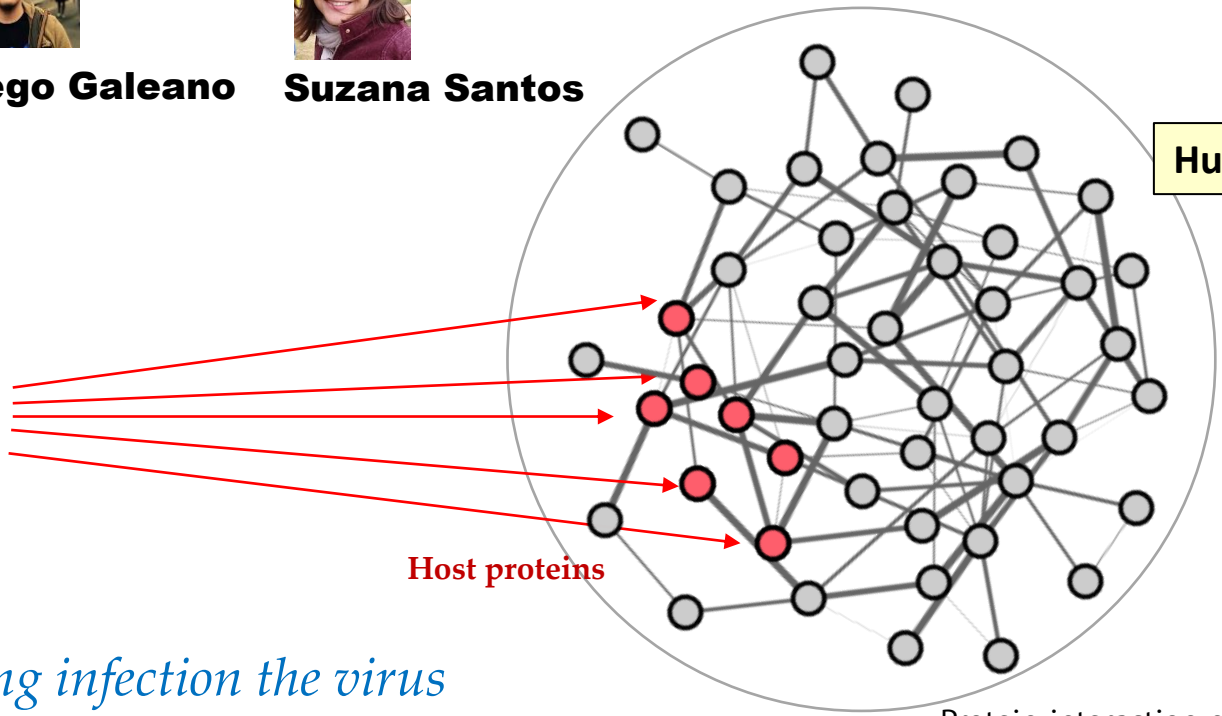
Diego Galeano

Suzana Santos

Virus

(SARS-CoV-2)





During infection the virus interacts with this network

1. Systems Pharmacology

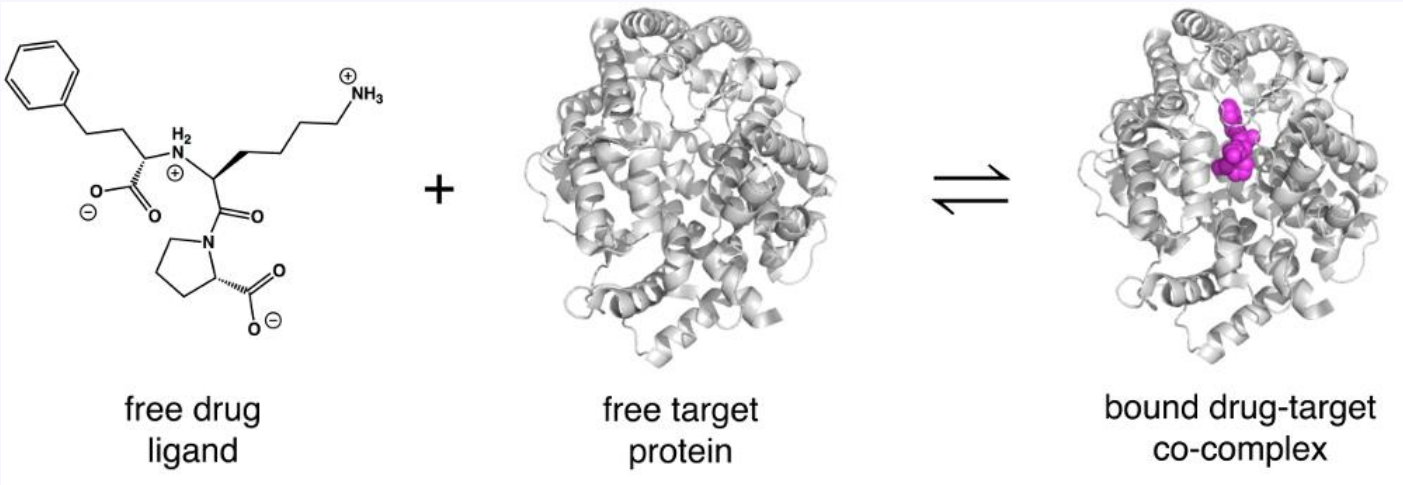
predict drugs (broad-spectrum antivirals) effective against the virus (SARS-CoV-2)

2. Network Medicine

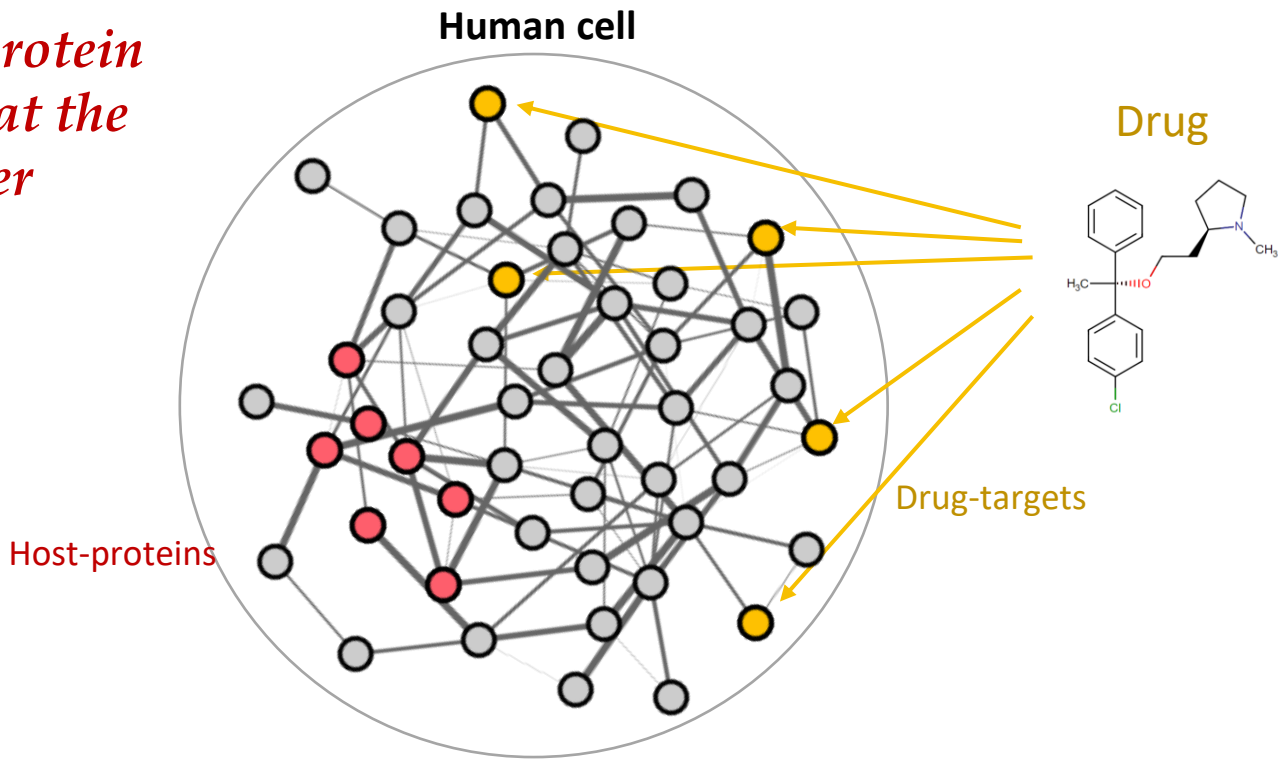
perturb the host protein subnetwork so that the virus can no longer interact (~infect)

How drugs work

Drugs bind to proteins (drug targets)



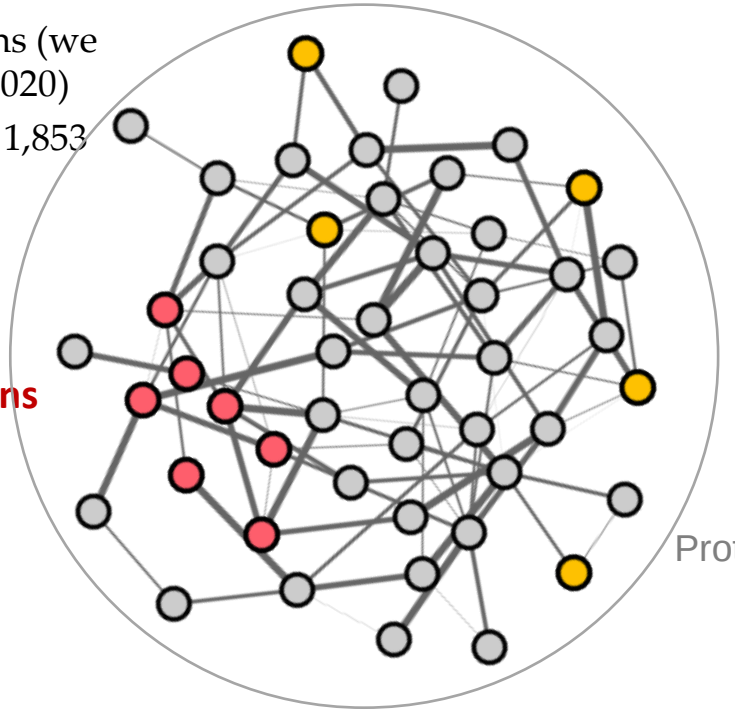
perturb the host protein subnetwork so that the virus can no longer interact (~infect)



1. *A drug binding to a protein causes a perturbation to the network*
2. *Perturbations propagate in the network*

The idea: rank the drugs based on how much they perturb the host protein subnetwork

- List of SARS-CoV-2 host proteins (we have 332, Krogan Lab, Nature 2020)
- List of drug targets (Drugbank: 1,853 drugs; 2,083 targets)



Calculate distances on the network (graph kernels)

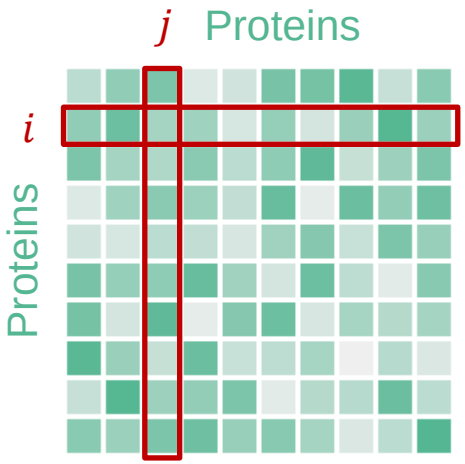
Using Graph Kernels

Normalised Laplacian $\tilde{L} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$

Kernels:

$K = (I + \sigma^2 \tilde{L})^{-1}$	Regularised Laplacian
$K = \exp(-\sigma^2/2\tilde{L})$	Diffusion process
$K = (aI - \tilde{L})^p$ with $a \geq 2$	p-Step random walk
$K = ((1-\alpha)I + \alpha\tilde{L})^{-1}$	Consistency method

Matrix representation of K



1. Raw kernel-based scores
2. Weighting the host proteins (using gene expr)
3. Including predicted drug targets (using our **DTP method**)

$score = v_i^T K h$
v_i^T: drug target vector
h: host-protein vector

Looking at the 20 top predicted drugs

Cannabidiol inhibits SARS-CoV-2 replication through induction of the host ER stress and innate immune responses (Science 2022)

Rank	Drug name (ID)	Main ATC Category	Additional curated evidence for COVID-19
1	Fostamatinib (DB12010)	Blood and blood forming organs (B)	several clinical trials (NCT04579393, NCT04581954, NCT04629703), *in silico* evidence[59]
2	NADH (DB00157)		in Clinical trials (NCT04604704), *in silico* evidence[60–62]
3	Copper (DB09130)		Combinatorial therapy[63], *in silico* evidence[64]
4	Cannabidiol (DB09061)	Nervous System (N)	In Clinical Trials (NCT04467918)
5	Glutathione (DB00143)	Various (V)	In Clinical Trials CTRI/2021/01/030793
6	Doxorubicin (DB00997)	Antineoplastic and immunomodulating agents (L)	*in silico* evidence[65]
7	Flavin adenine dinucleotide (DB03147)		*in silico* evidence[64]
8	Verapamil (DB00661)	Cardiovascular System (C)	multiple clinical trials (NCT04351763, NCT04330300, NCT04467931)
9	Zinc (DB01593)	Cardiovascular System (C)	included in more than 60 clinical trials, *in silico* evidence[64]
10	Zinc acetate (DB14487)	Cardiovascular System (C), Alimentary tract and metabolism (A)	included in more than 60 clinical trials, *in silico* evidence[64]
11	Zinc chloride (DB14533)	Blood and blood forming organs, Cardiovascular System (B)	included in more than 60 clinical trials, *in silico* evidence[64]
12	Moexipril (DB00691)	Cardiovascular System (C)	in clinical trials (NCT04467931)
13	Conjugated estrogens (DB00286)	Genito-urinary system and sex hormones (G)	NA
14	Clozapine (DB00363)	Nervous System (N)	NA
15	Rifampicin (DB01045)	Antiinfectives for systemic use (J)	*in silico* evidence[66]
16	Amitriptyline (DB00321)	Nervous System (N)	*in vitro* evidence[38]
17	Phenobarbital (DB01174)	Nervous System (N)	NA
18	Desipramine (DB01151)	Nervous System (N)	NA
19	Progesterone (DB00396)	Genito-urinary system and sex hormones (G)	in several clinical trials (NCT04365127, NCT04539626, NCT04865029 - combinatorial therapy)
20	Ethanol (DB00898)	Dermatologicals (D), Various (V)	in more than 100 clinical trials.

Table S6. Kernel-based top-20 predicted drugs. For each drug we show whether there is evidence from other *in silico* approaches, *in vitro* experiments, or clinical trials.

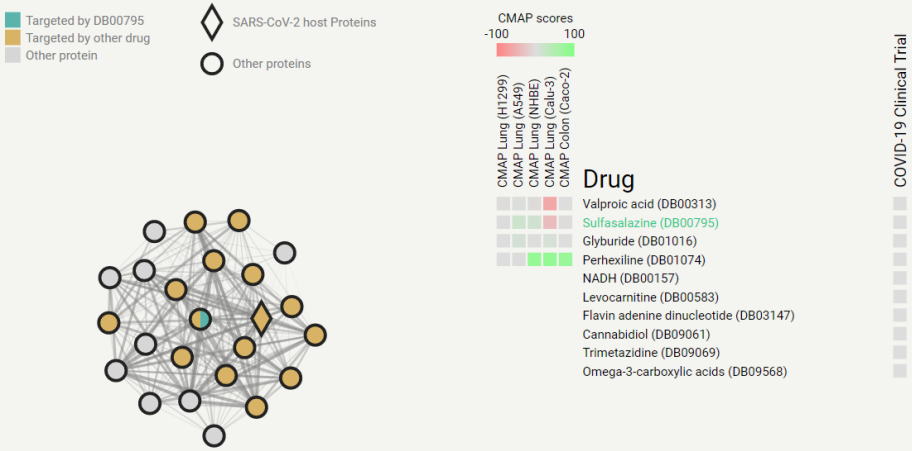
CoREx

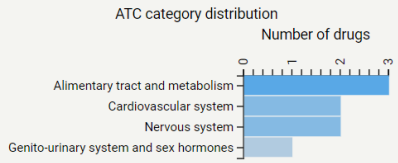
the COVID-19 Repositioning Explorer

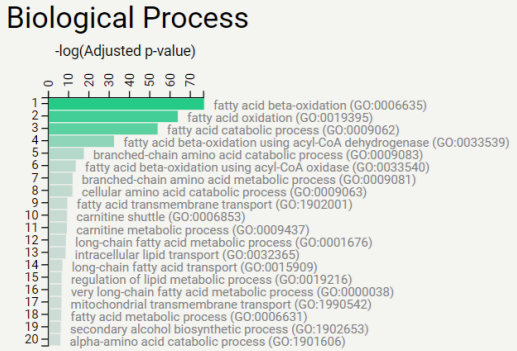
An online exploration tool for experimentalists

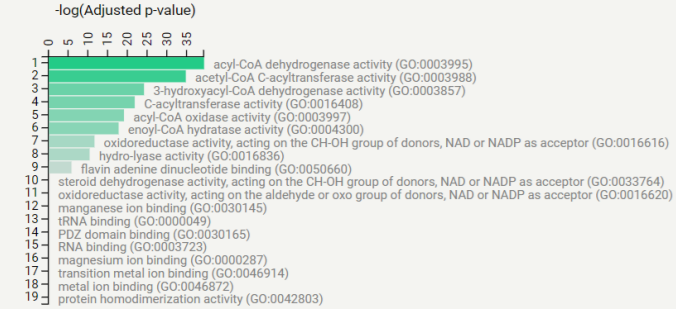
1. *Examine predictions in the context of the interactome*
2. *Analysis of functional modules*

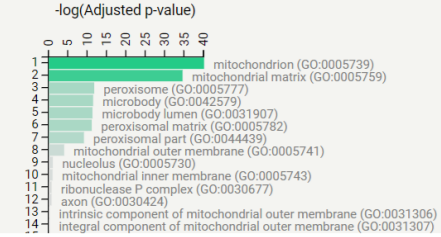
Drug: DB00795 (Sulfasalazine)
Network: S2F











www.paccanarolab.org/corex

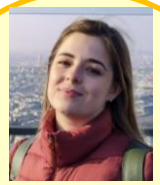
Ongoing projects

- Drug repurposing for Chagas disease
- Drug repurposing for viruses
- Phenotype prediction in cardiovascular disease
- Prediction of phenotype/outcome in cancer patients
- Drug target prediction
- Drug-Drug interaction prediction

Acknowledgements


Horacio Caniza


Santiago Noto


Bruna Fistarol


Phil Ovington


Haixuan Yang


Aldo Galeano

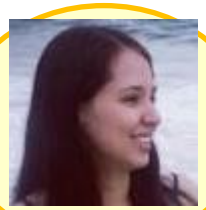

Mateo Torres


Tamas Nepusz


Diego Galeano


Suzana Santos


Ruben Jimenez


Maria del Mar Sanchez

Ongoing collaborations:
Mark Gerstein – Yale University
Haiyuan Yu – Cornell University
Andrew Emili – Un of Toronto
Edward Marcotte – Un Texas, Austin
Laszlo Bogre – Royal Holloway, UoL
Paul Matthews – Imperial College
Michael Bronstein – Imperial College
Giorgio Valentini – Un of Milan
Raghava Velagaleti – Boston Un.
Celeste Vega – Un of Asuncion

BBSRC bioscience for the future

EPSRC Engineering and Physical Sciences Research Council

FUNDAÇÃO GETULIO VARGAS

MARIE CURIE ACTIONS

CONACYT

NSF

MRC Medical Research Council

THE ROYAL SOCIETY

http://www.paccanarolab.org