Interactome based analyzes for translational medicine

Emre Güney, PhD

Hospital del Mar Research Institute (IMIM)
& Pompeu Fabra University(UPF)

MSc on Omics Data Analysis - Bioinformatics Applications

January 29th, 2019

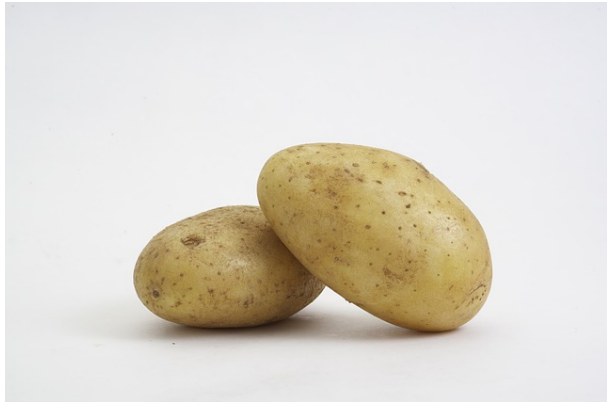
RESEARCH PROGRAMME ON BIOMEDICAL INFORMATICS

UNIVERSITAT POMPEU FABRA

Proteins – Why do I care?

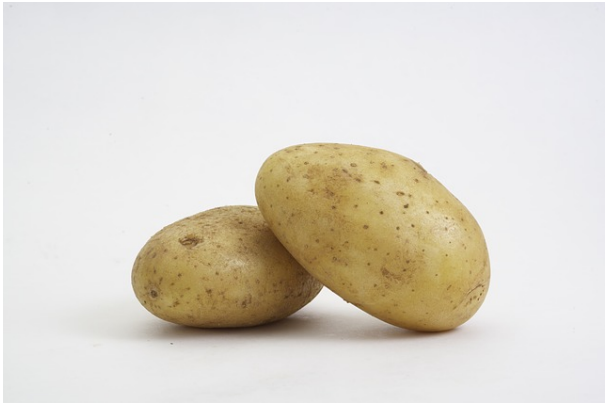
- Tortilla de patatas
- Patatas bravas
- Pulpo a Gallega

Proteins – Why do I care?

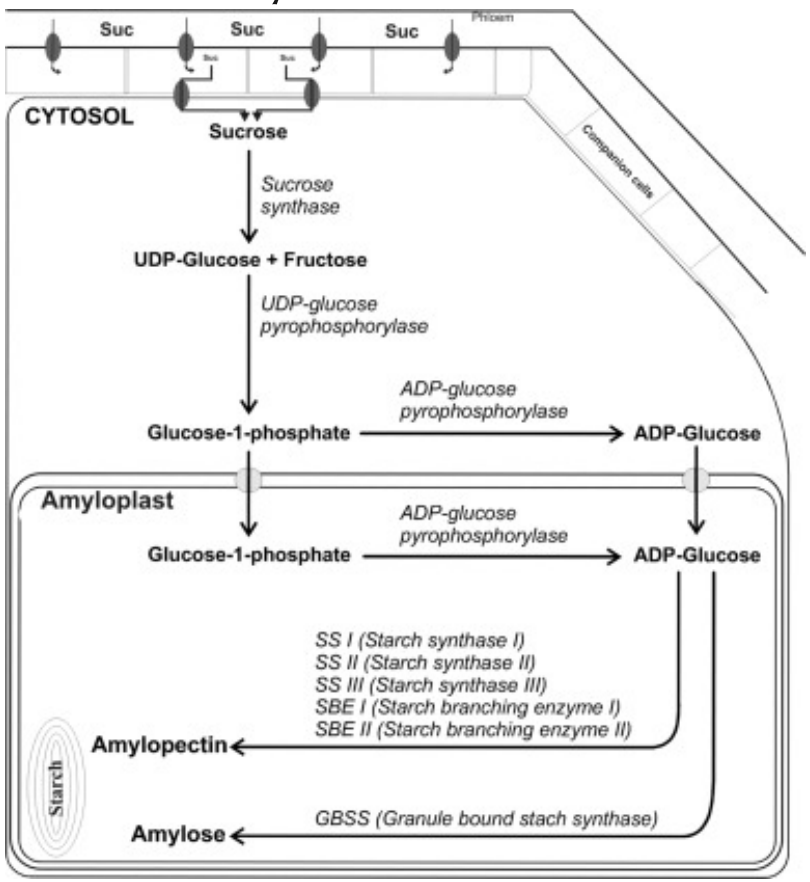


Images from pixabay.com

Proteins – Why do I care?

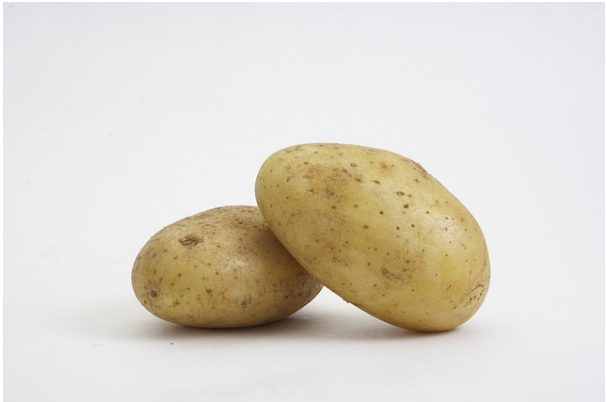


starch biosynthesis

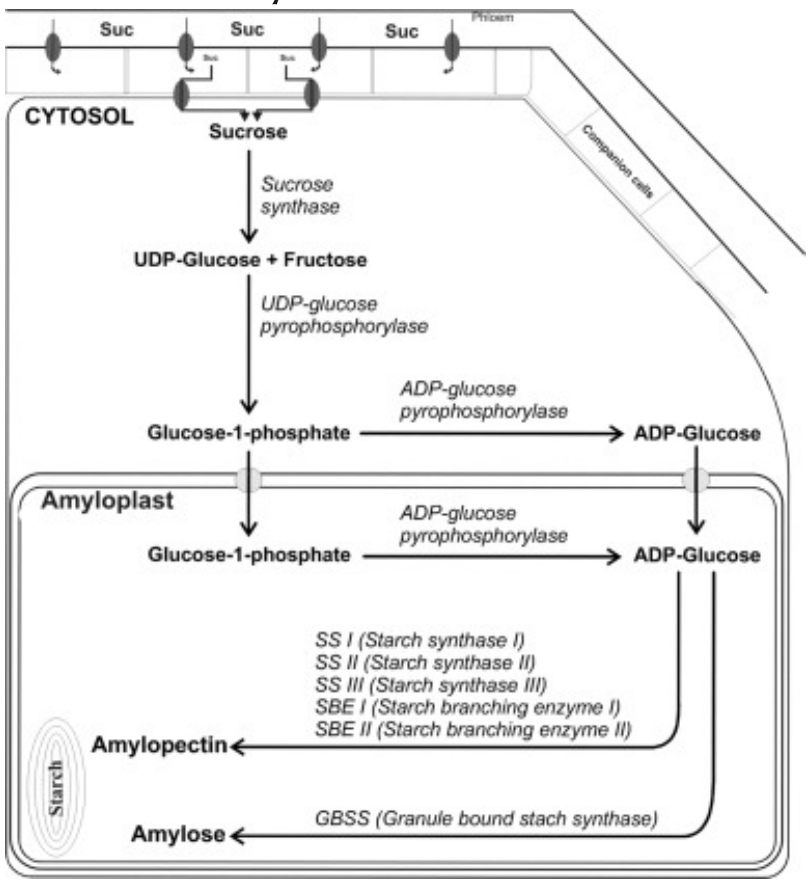


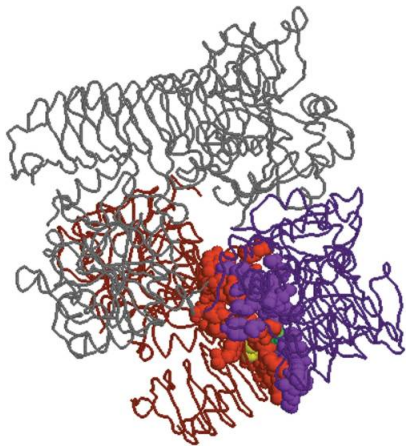
Images from pixabay.com | *Nazarian-Firouzabadi and Visser, 2017, Biochem & Biophy Rep*

Proteins – Why do I care?



starch biosynthesis

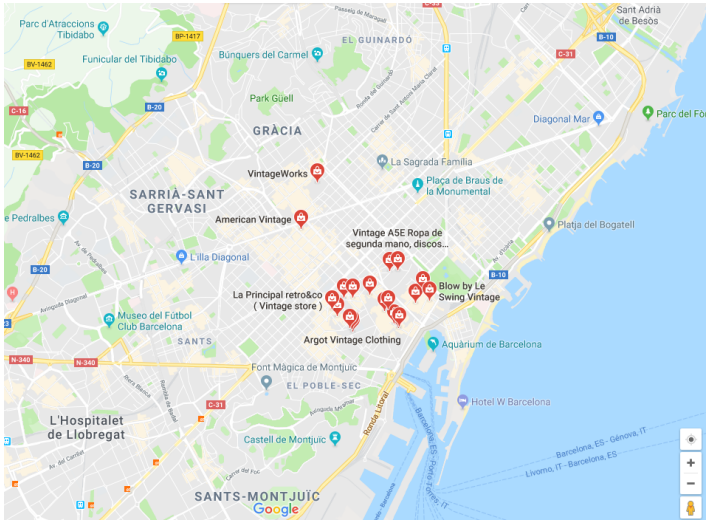


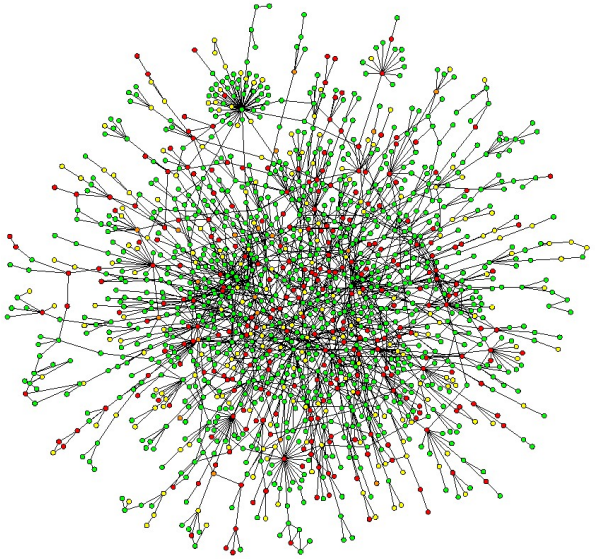


ADP-glucose pyrophosphorylase

Images from pixabay.com | *Nazarian-Firouzabadi and Visser, 2017, Biochem & Biophy Rep* | *Guney et al., 2008, NAR*

Interactome – The map of protein interactions in the cell

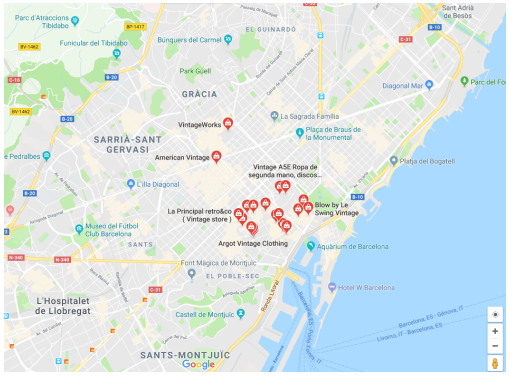


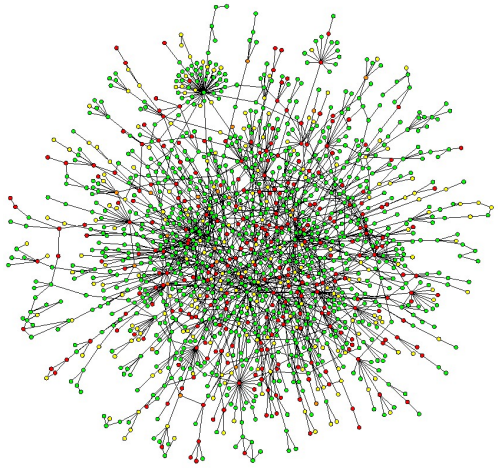


(Barabási and Oltvai, 2004, Nat Rev Genet)

- Proteins “talk to each other” by physically interacting with each other
- These interactions are essential for performing biological processes
- The network of interactions between proteins: Interactome

How to generate such a map for the cell?





Protein interaction data is spread across various repositories

DIP

Reactome

Human Protein Reference Database

IntAct

STRING

MIPS

MINT

BioGrid

KEGG

Hands On! – Interactions of AGPS1

Finding interactions of AGPS1

AGPS1 is the large subunit of ADP-glucose pyrophosphorylase in potato

- Go to `www.uniprot.org`
- Search for “AGPS1” and “potato”
- Select the AGPS1 protein
- Check “Interaction” field

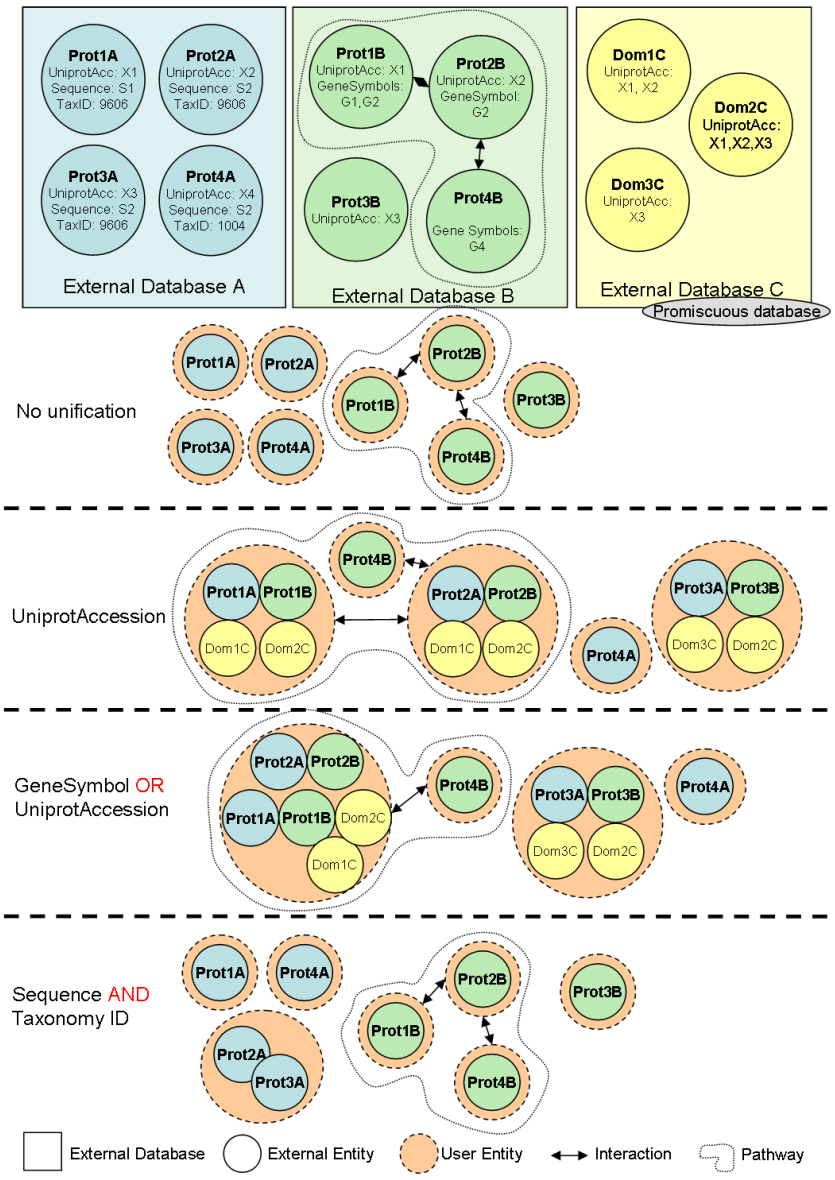
Hands On! – Interactions of AGPS1

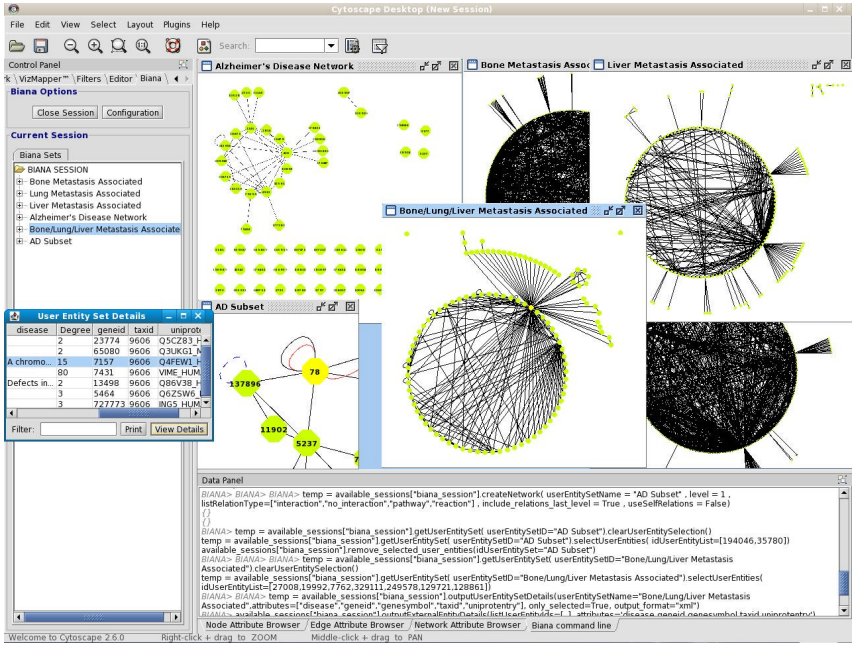
Finding interactions of AGPS1

AGPS1 is the large subunit of ADP-glucose pyrophosphorylase in potato

- Which are the databases that provide interaction information?
- With which other protein AGPS1 interacts according to IntAct database?
- What are the interaction detection methods for this interaction?
- What is the ENTREZ gene id of the interacting protein?

Integrating protein interaction data





BIANA: Biological Interaction data integration and Network Analysis framework

(Garcia-Garcia et al., 2010, BMC Bioinformatics)

Available interactomes across species

Number of proteins and their interactions (in thousands, rounded to the closest 1K)

Species	Proteins (SwissProt)	BIANA	IID*	CPDB†	HIPPIE‡
Human	20	299	335	397	274
Mouse	17	64	37	23	-
Rat	8	5	6	-	-

Data on human protein interactions is ∼**10-50x** of what is available for mouse and rat

* Integrated Interactions Database *(Kotlyar et al., 2015, NAR)*

† ConsensusPathDB *(Kamburov et al., 2011, NAR)*

‡ Human Integrated Protein–Protein Interaction rEference *(Alanis-Lobato et al., 2016, NAR)*

Hands On! – PPICoin

Generating PPICoin

A new pseudo-crypto currency based on protein interactions

Rules for block generation and transaction verification

- Generate a new coin using a **seed gene (X)** as the starting block
- Add a gene **Y interacting with X** to the block, where Y interacts with at least one other gene Z
- If there are multiple candidate proteins Y, add the one which has the **smallest ENTREZ gene id**
- Write that protein Y on a paper (along with your name) and pass it to the person on your left
- When you receive the paper, verify the correctedness of previous transaction and repeat the three steps above to generate the next block

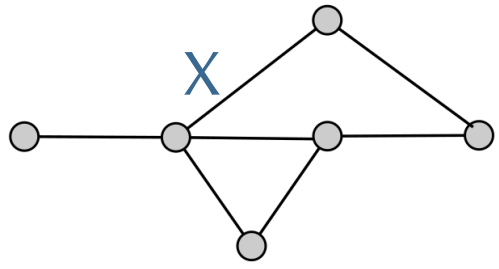
Hands On! – PPICoin

PPICoin

X	Y	Z	...
Emre	Alice	Bob	...

- $X, Y \in V$
- $Candidates(Y) = \{y : (X, y) \in E, k_y > 1\}$
- $Y = argmin_{Geneid}(Candidates(Y))$

V: proteins in the interactome
E: interactions between proteins
k: node degree (number of the node's neighbors)



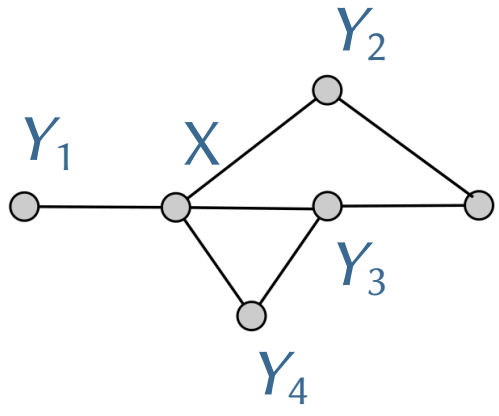
Hands On! – PPICoin

PPICoin

X	Y	Z	...
Emre	Alice	Bob	...

- $X, Y \in V$
- $Candidates(Y) = \{y : (X, y) \in E, k_y > 1\}$
- $Y = argmin_{Geneid}(Candidates(Y))$

V: proteins in the interactome
E: interactions between proteins
k: node degree (number of the node's neighbors)



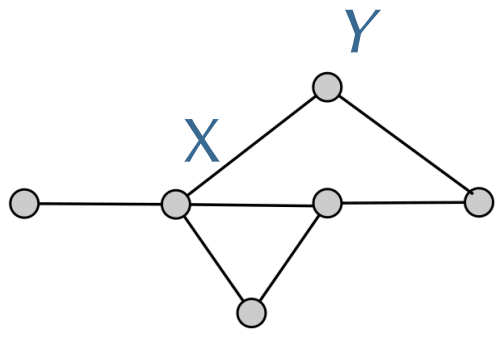
Hands On! – PPICoin

PPICoin

X	Y	Z	...
Emre	Alice	Bob	...

- $X, Y \in V$
- $Candidates(Y) = \{y : (X, y) \in E, k_y > 1\}$
- $Y = argmin_{Geneid}(Candidates(Y))$

V: proteins in the interactome
E: interactions between proteins
k: node degree (number of the node's neighbors)



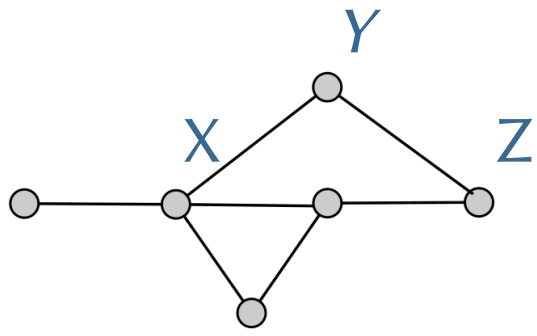
Hands On! – PPICoin

PPICoin

X	Y	Z	...
Emre	Alice	Bob	...

- $X, Y \in V$
- $Candidates(Y) = \{y : (X, y) \in E, k_y > 1\}$
- $Y = argmin_{Geneid}(Candidates(Y))$

V: proteins in the interactome
E: interactions between proteins
k: node degree (number of the node's neighbors)



Hands On! – PPICoin

Generating PPICoin

Verify the coin using the gene associated to actin-accumulation myopathy as seed

Resource	URL
Disease-gene information	
DisGeNET	`www.disgenet.org`
Interactome generation	
BIANA (web server)	`http://sbi.imim.es/BIANA.php`
BIANA generated files	`www.emreguney.net/doc/ms_omics.zip`
Interactome visualization	
Cytoscape	`www.cytoscape.org`

Hands On! – PPICoin

Generating PPICoin

Verify the coin using the gene associated to actin-accumulation myopathy as seed

- Use the gene associated to actin-accumulation myopathy as the seed gene (X)
- Add a gene **Y interacting with X** to the block, where Y interacts at least with another gene
- Write your name and Y on the paper and pass it to the left

Hands On! – PPICoin

Generating PPICoin

Verify the coin using the gene associated to actin-accumulation myopathy as seed

- Use the gene associated to actin-accumulation myopathy as the seed gene (X)
  - Go to `www.disgenet.org`
  - Browse for "CURATED" associations
  - Filter by disease name "actin-accumulation myopathy"
- Add a gene **Y interacting with X** to the block, where Y interacts at least with another gene

- Write your name and Y on the paper and pass it to the left

Hands On! – PPICoin

Generating PPICoin

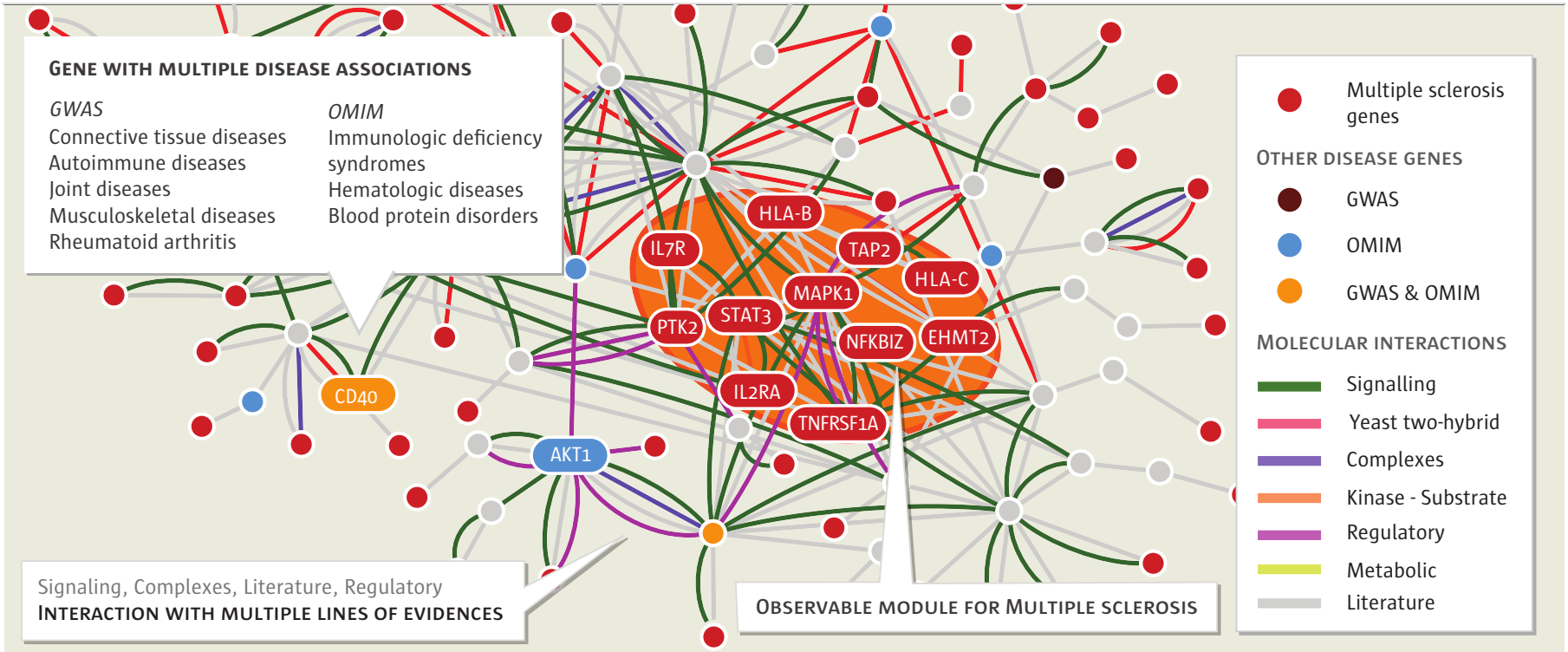
Verify the coin using the gene associated to actin-accumulation myopathy as seed

- Use the gene associated to actin-accumulation myopathy as the seed gene (X)

- Add a gene **Y interacting with X** to the block, where Y interacts at least with another gene
  - Download and unzip the BIANA generated interactome files
  - Import “subnetwork.sif” as the network file and “9606.txt” as the node table file in Cytoscape
  - Locate & select the seed gene in the interactome (using the search box in Cytoscape)
  - Go to “Select::Nodes::First neighbors” to find the interacting genes
- Write your name and Y on the paper and pass it to the left

Network analysis

Graphs provide a systems-level representation of the cellular network such as the interactome



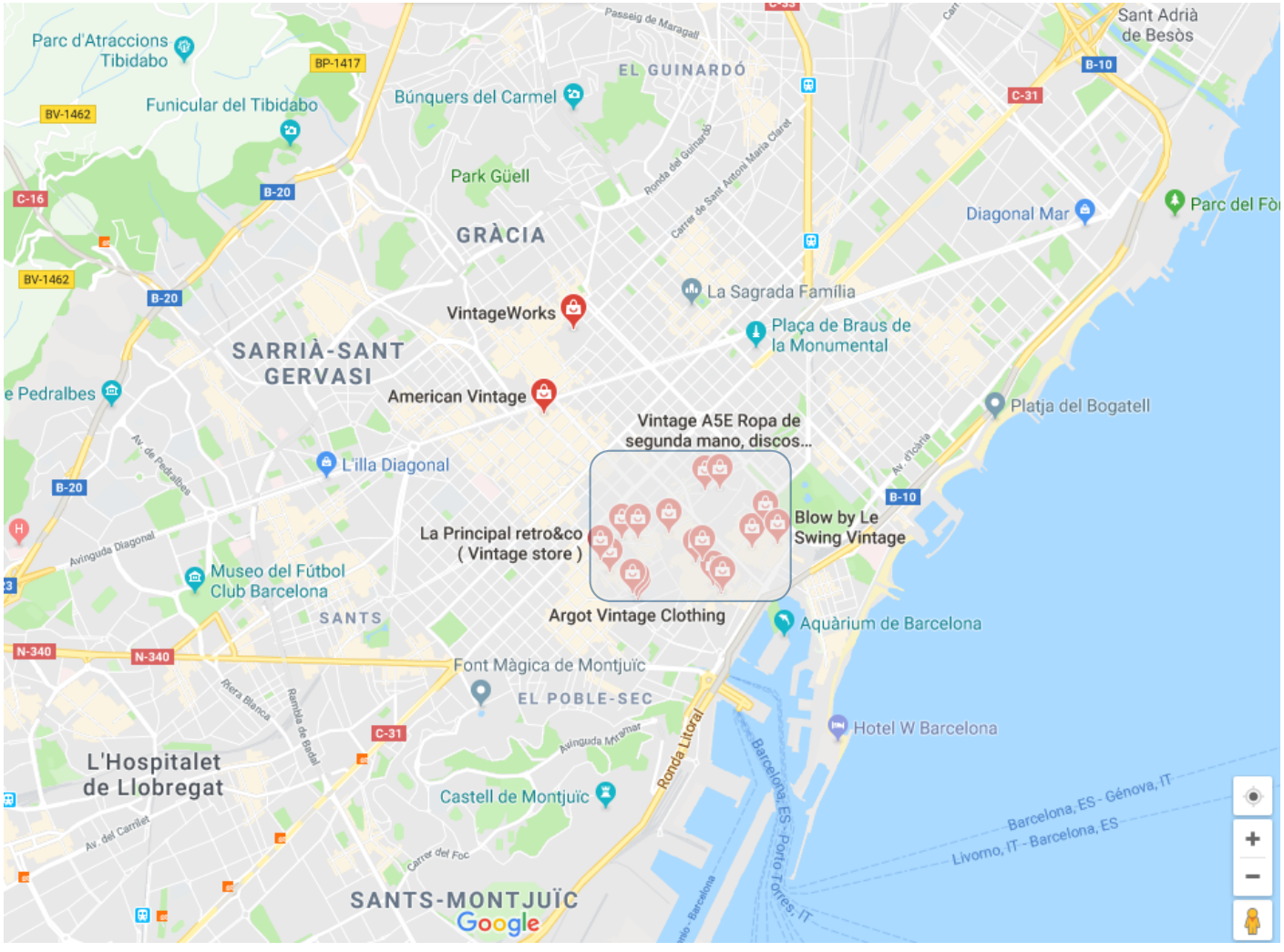
(Menche et al., 2015, Science)

Network analysis is used to extract biologically meaningful information from interactomes

Network analysis – Clustering vs guilt-by-association

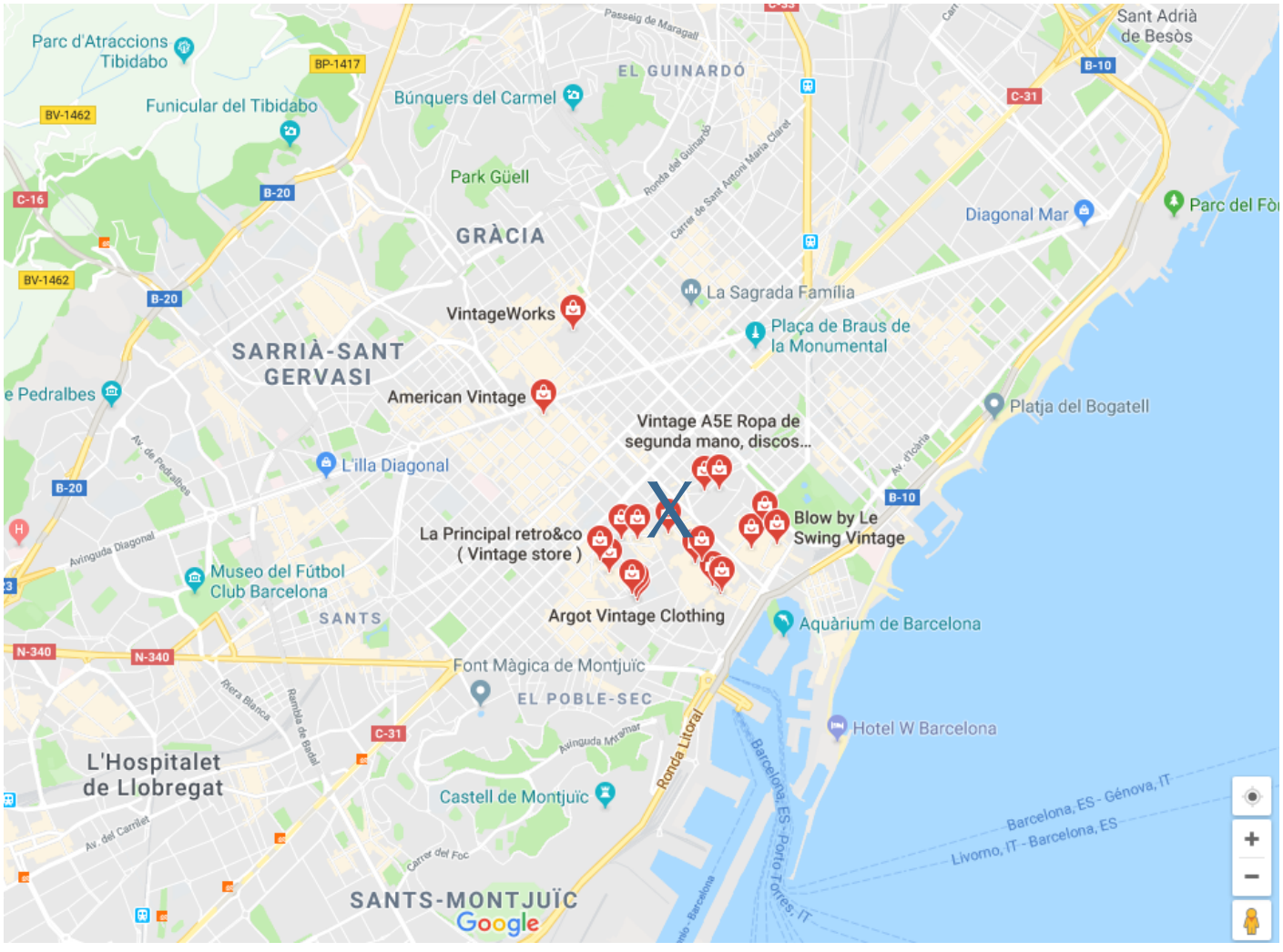
Vintage shops in Barcelona

Network analysis – Clustering vs guilt-by-association



clustering (Vintage shops are located in the center)

Network analysis – Clustering vs guilt-by-association



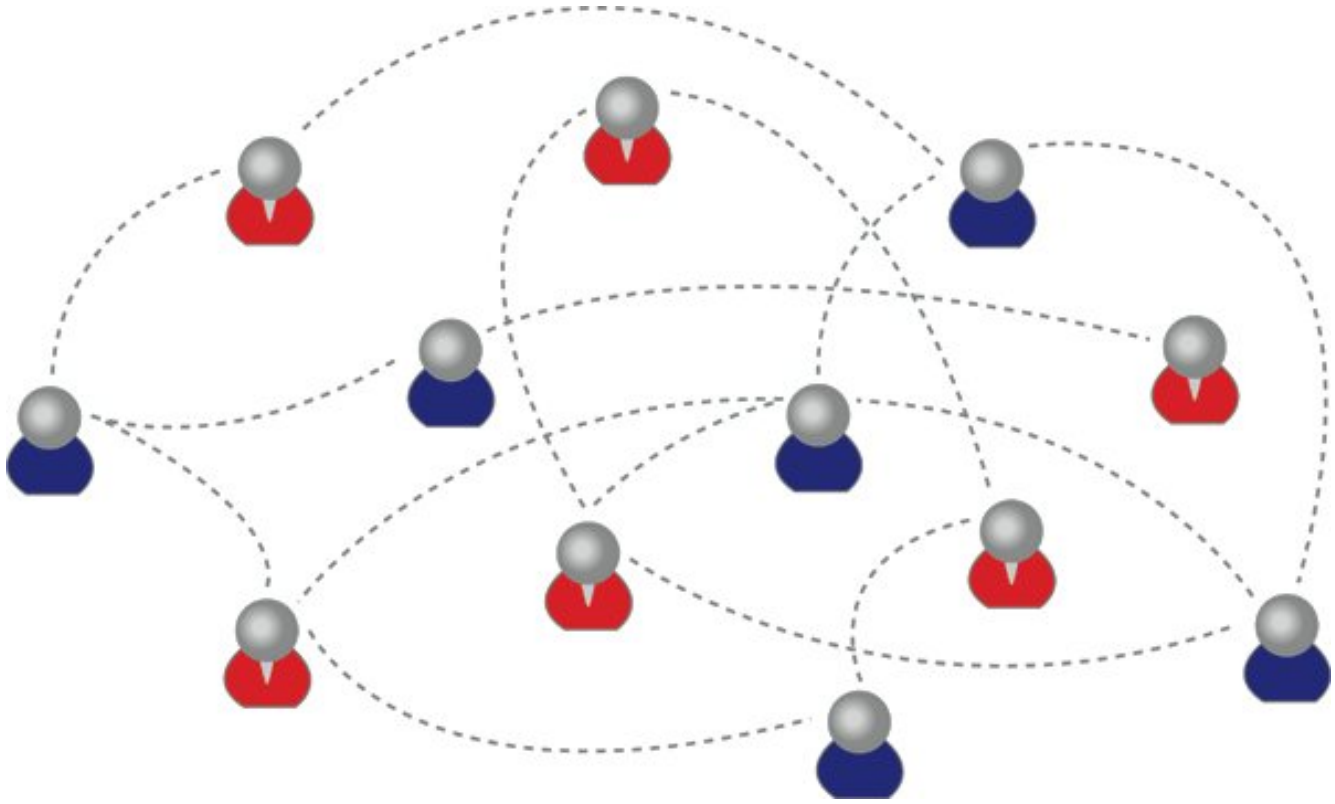
guilt-by-association (X is a vintage shop)

Guilt-by-association works in practice due to the clustering phenomenon

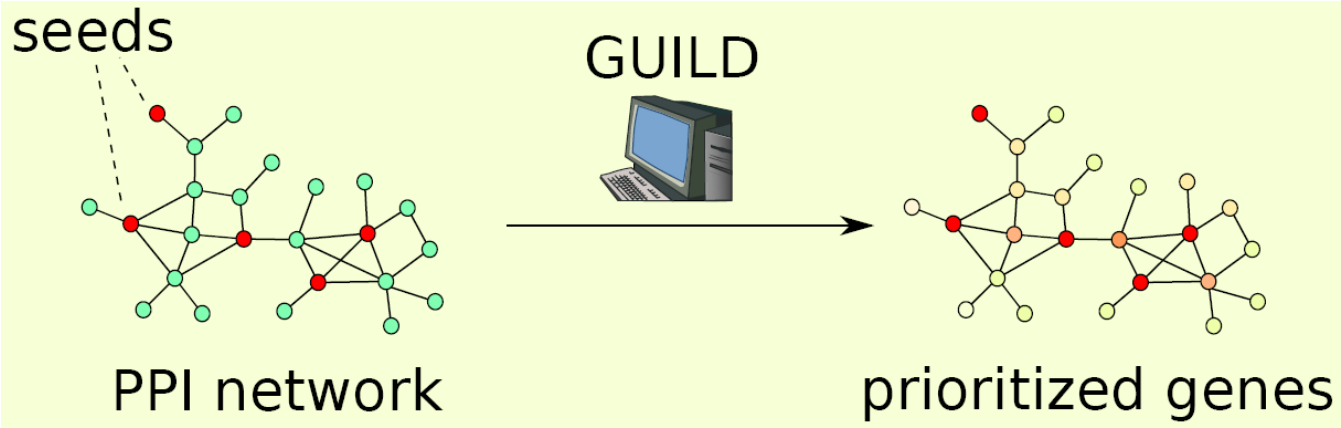
Percentage of your Facebook friends from the same country

(Ugander et al., 2011, arXiv:1111.4503)

~85%



GUILD framework



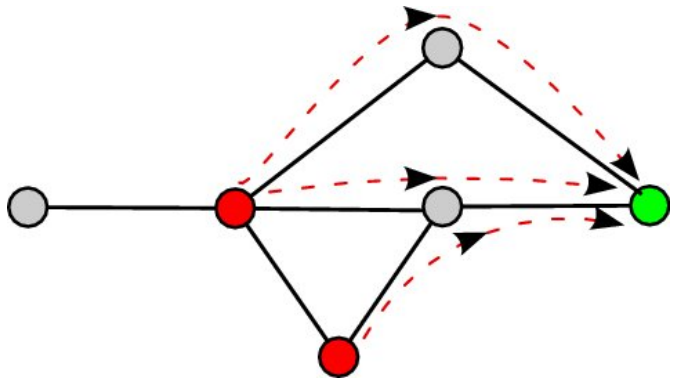
- NetScore
- NetZcore
- NetShort
- NetCombo

- Functional Flow *(Nabieva et al., 2005)*
- Random walk with restart *(Kohler et al., 2008)*
- PageRank with priors *(Chen et al., 2009)*
- Network propagation *(Vanunu et al., 2011)*

Guney and Oliva, 2012, PLoS ONE

GUILD framework – NetScore

- considers multiple shortest paths in-between the two nodes

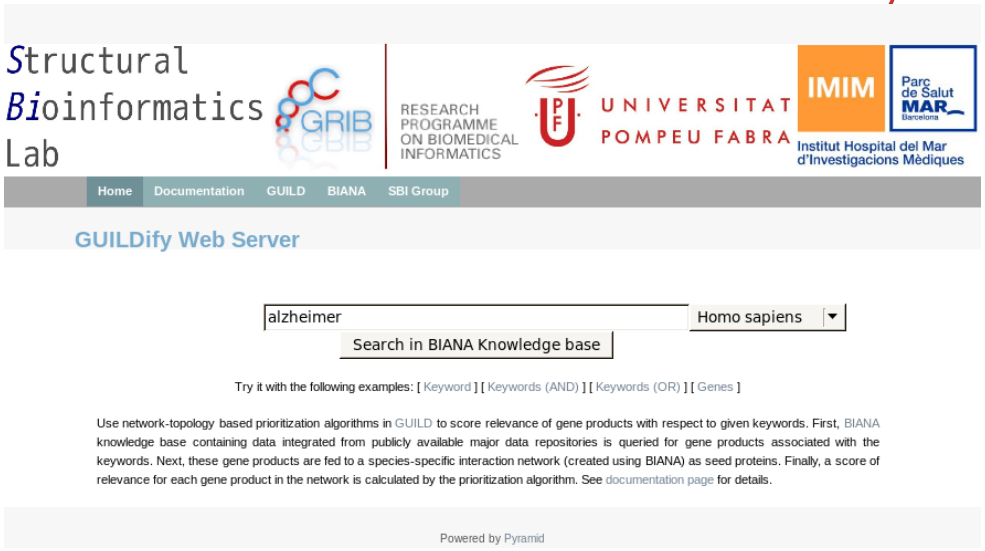


$$message = \left\{ \begin{array}{l} source \\ timestamp \\ path\_weight \end{array} \right\}$$

u
m_u
⋮
m_v
⋮

Guney and Oliva, 2012, PLoS ONE

GUILD framework – GUILDify



- free text search on UniProt, OMIM, DisGeNET databases and GO annotations for a given species
- matching genes are retrieved and used as seeds
- the PPI network is compiled from DIP, HPRD, IntAct, MINT, MPact, BioGRID, BIND
- tissue-specific PPI networks for human using HPA and Tissues databases

Guney and Oliva, 2014, Bioinf

GUILD framework – GUILDify

GUILDify Web Server - Matching Proteins

35 BIANA entries are found for the query ***alzheimer*** in *Homo sapiens*. 2 of these entries have no interactions (not in the network).

GUILDify! using selected entries below [Options]

Options

☑ NetScore (Repetition: 3 ▾ Iteration: 2 ▾)

☐ NetZcore (Iteration: 5 ▾)

☐ NetShort

[Hide]

BIANA entries in the network

[Select All / None]

Keep	Gene ID	UniProt ID	Gene Symbol	Description
☑	2	P01023	A2M	{alzheimer disease, susceptibility to}, 104300 (3) omim
☑	26574	Q9NY61	AATF	May function as a general inhibitor of the histone deacetylase HDAC1. Binding to the pocket region of RB1 may displace HDAC1 from RB1/E2F complexes, leading to activation of E2F target genes and cell cycle progression. Conversely, displacement of HDAC1 from SP1 bound to the CDKN1A promoter leads to increased expression of this CDK inhibitor and blocks cell cycle progression. Also antagonizes PAWR mediated induction of aberrant amyloid peptide production in alzheimer disease (presenile and senile dementia), although the molecular basis for this phenomenon has not been described to date. swissprot
☑	1636	P12821	ACE	{alzheimer disease, susceptibility to}, 104300 (3) omim
☑	9507	O75173	ADAMTS4	Cleaves aggrecan, a cartilage proteoglycan, and may be involved in its turnover. May play an important role in the destruction of aggrecan in arthritic diseases. Could also be a critical factor in the exacerbation of neurodegeneration in alzheimer disease. Cleaves aggrecan at the '392-Glu-|-Ala-393' site. swissprot
☑	323	Q92870	APBB2	{alzheimer disease, late-onset}, 104300 (3) omim
☑	348	P02649	APOE	alzheimer disease swissprot Defects in APOE are a cause of hyperlipoproteinemia type 3 (HLPP3) [MIM:107741]; also known as familial dysbetalipoproteinemia. Individuals with HLPP3 are clinically characterized by xanthomas, yellowish lipid deposits in the palmar crease, or less specific on tendons and on elbows. The disorder rarely manifests before the third decade in men. In women, it is usually expressed only after the menopause. The vast majority of the patients are homozygous for APOE*2 alleles. More severe cases of HLPP3 have also been observed in individuals heterozygous for rare APOE variants. The influence of APOE on lipid levels is often suggested to have major implications for the risk of coronary artery disease (CAD). Individuals carrying the common APOE*4 variant are at higher risk of CAD. Genetic variations in APOE are associated with alzheimer disease type 2 (AD2) [MIM:104310]. It is a late-onset neurodegenerative disorder characterized by progressive dementia, loss of cognitive abilities, and deposition of fibrillar amyloid proteins as intraneuronal neurofibrillary tangles, extracellular amyloid plaques and vascular amyloid deposits. The major constituent of these plaques is the neurotoxic amyloid-beta-APP 40-42 peptide (s), derived proteolytically from the transmembrane precursor protein APP by sequential secretase processing. The cytotoxic C-terminal fragments (CTFs) and the caspase-cleaved products such as C31 derived from APP, are also implicated in neuronal death. Note=The APOE*4 allele is genetically associated with the common late onset familial and sporadic forms of alzheimer disease. Risk for AD increased from 20% to 90% and mean age at onset decreased from 84 to 68 years with increasing number of APOE*4 alleles in 42 families with late onset AD. Thus APOE*4 gene dose is a major risk factor for late onset AD and, in these families, homozygosity for APOE*4 was virtually sufficient to cause AD by age 80. The mechanism by which APOE*4 participates in pathogenesis is not known. Defects in

GUILD framework – GUILDify

GUILDify Web Server - Results

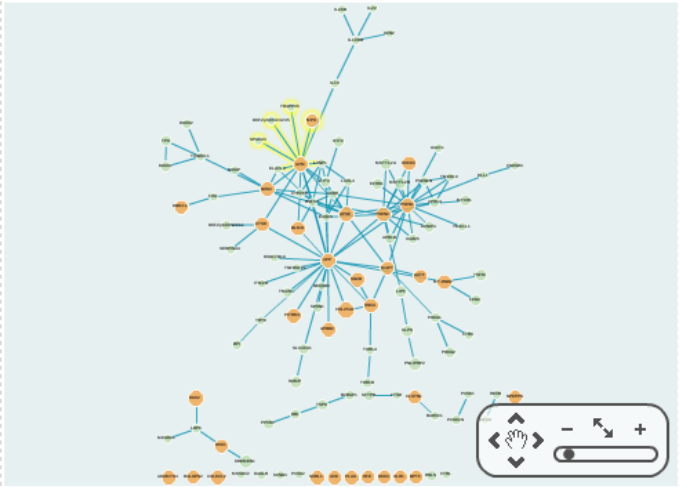
Calculated scores for proteins in the BIANA ***Homo sapiens*** interactome:

[Download all scores] [Download seed proteins] [Download interactome]

Proteins **1** - **20** of 11238 ||<< << >> >>||

Rank	Gene ID	UniProt ID	Gene Symbol	GUILD Score
1	84570	Q9BXS0	seed COL25A1	1.0000
2	27123	Q9UBU2	seed DKK2	0.9398
3	348	P02649	seed APOE	0.8481
4	6620	Q16143	seed SNCB	0.8435
5	10531	Q5JRX3	seed PITRM1	0.8373
6	5664	P49810	seed PSEN2	0.8123
7	5663	P49768	seed PSEN1	0.7905
8	6653	Q92673	seed SORL1	0.7861
9	1795	Q8IZD9	seed DOCK3	0.7858
10	1636	P12821	seed ACE	0.7848
11	5328	P00749	seed PLAU	0.7824
12	351	P05067	seed APP	0.7774
13	323	Q92870	seed APBB2	0.7756
14	9507	O75173	seed ADAMTS4	0.7754
15	4846	P29474	seed NOS3	0.7717
16	-	Q8IVG9	seed MT-RNR2	0.7705
17	2	P01023	seed A2M	0.7599
18	642	Q13867	seed BLMH	0.7556
19	81035	Q5KU26	seed COLEC12	0.7523
20	6622	P37840	seed SNCA	0.7518

Visualization options: Top 1% Top 5% Top 10%



Selected 4 nodes:

Rank	Gene ID	UniProt ID	Gene Symbol	GUILD Score	Type
96	124912	Q8IXA5	SPACA3	0.227186	non_seed
97	126003	Q8IUR0	TRAPPC5	0.227186	non_seed
78	1562	P33260	DKFZp686I24235	0.274228	non_seed
24	4353	P05164	MPO	0.721588	seed

Hands On! – Guilt-by-association analysis using the interactome

Identifying commonalities across diseases

Investigate the common biological processes involved in between Alzheimer's disease and type 2 diabetes

Resource	URL
Interactome-based analysis	
GUILDify web server	`http://aleph.upf.edu/guildify2`

Hands On! – Guilt-by-association analysis using the interactome

Identifying commonalities across diseases

Investigate the common biological processes involved in between Alzheimer's disease and type 2 diabetes

Resource	URL
Interactome-based analysis	
GUILDify web server	`http://aleph.upf.edu/guildify2`

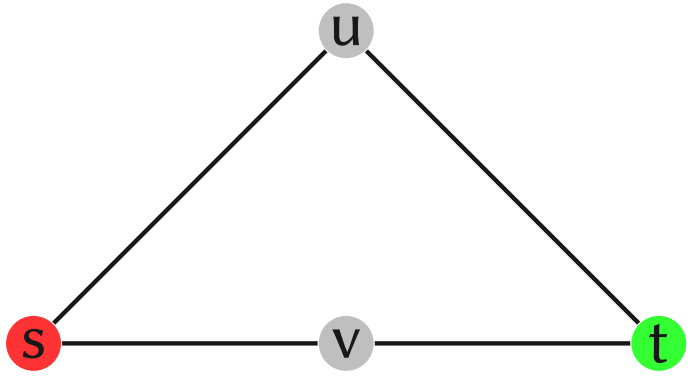
- Go to GUILDify web server
- Query for "Alzheimer disease" and run prioritization
- Query "type 2 diabetes" and run prioritization
- Check the overlap between the two results

Concluding remarks

- Biological processes involve coordinated activity of proteins that physically interact with each other
- Interactome, the network of interactions between proteins, provides a framework for understanding and characterizing biological processes
- Network medicine is an emerging field that aims to use interactome-based analyzes to identify genes associated to diseases and develop novel therapeutics
- Several tools available for the generation, visualization and analysis of the interactome

Appendix

NetScore

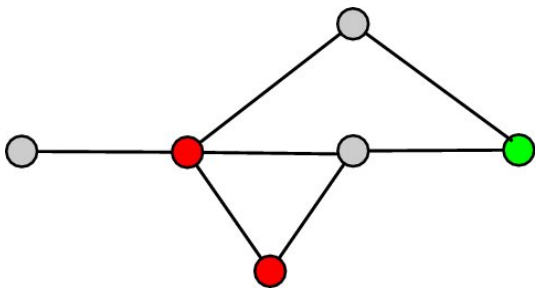


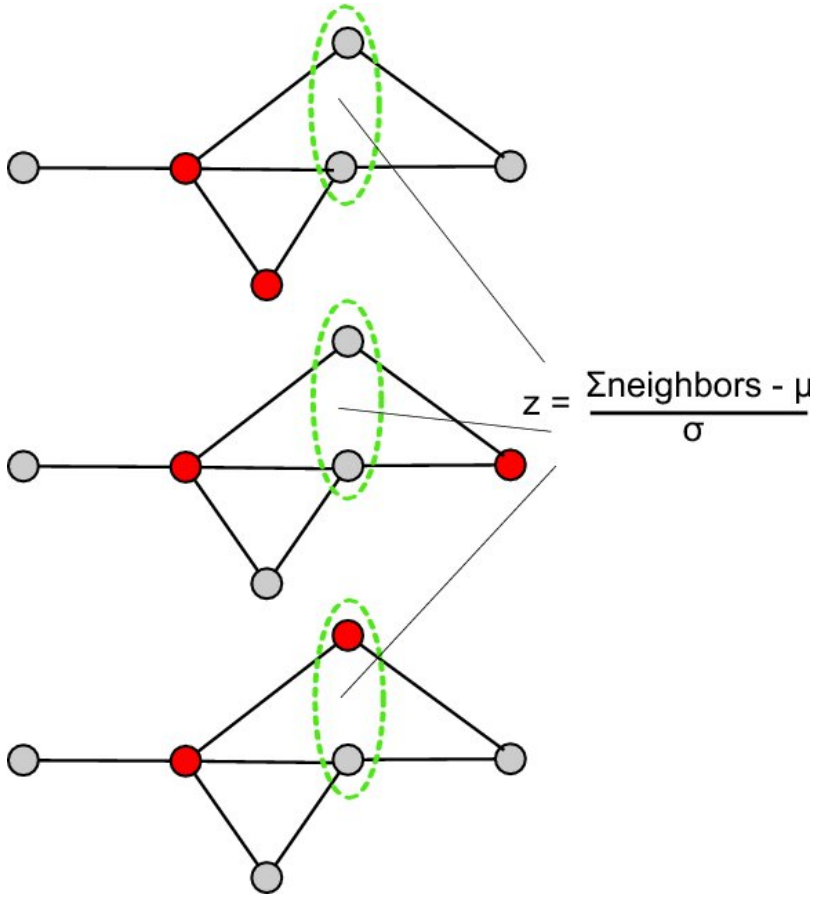

```
Input: G = ( V, E) graph with score property score:V → [0, 1],
       weight property weight: E → [0, 1], nRepetition, nIteration.
Output: G = ( V, E) graph with score property score′:V → [0, 1].
for i=1 to nRepetition do
   /* Initialize message arrays                                    */
   foreach u ∈ V do
      m.source ← u // source node id
      m.timestamp ← 0 // the iteration when recieved
      m.path_weight ← 1 // weights of the traveled path
      messages(u) ← { m }
   end
   for j=1 to nIteration do
      /* Digest messages                                          */
      foreach u ∈ V do
         foreach v ∈ {(u, v) ∈ E } do
            foreach m ∈ messages(v) do
               /* Do not accept messages from the same node
                  recieved during previous iterations            */
               if AcceptMessage(messages(u), m) then
                  m.timestamp ← j
                  m.path_weight ← m.path_weight ∗ weight(u, v)
                  messages(u) ← messages(u) ∪ m
               end
            end
         end
      end
      /* Update node scores                                       */
      foreach u ∈ V do
         foreach m ∈ messages(u) do
            score(u) ← score(u) + m.path_weight ∗
            score(m.source)
         end
         score(u) ← score(u) / || messages(u) ||
      end
   end
end
```

	s	u	v	t
initially	m_s^0	m_u^0	m_v^0	m_t^0
iteration 1	m_u^1	m_s^1	m_s^1	m_u^1
iteration 2	m_v^1	m_t^1	m_t^1	m_v^1
	$2m_t^2$	$2m_v^2$	$2m_u^2$	$2m_s^2$

NetZcore

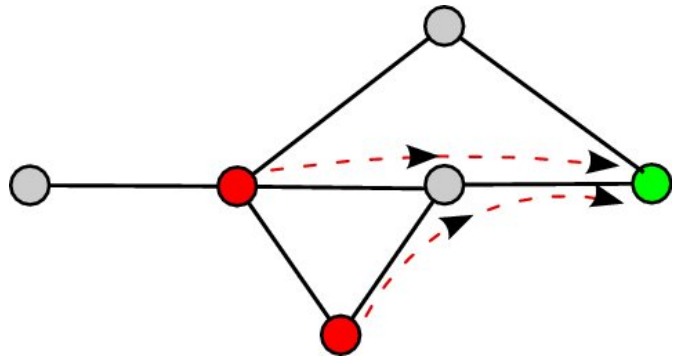
- checks the “significance” of the neighborhood configuration

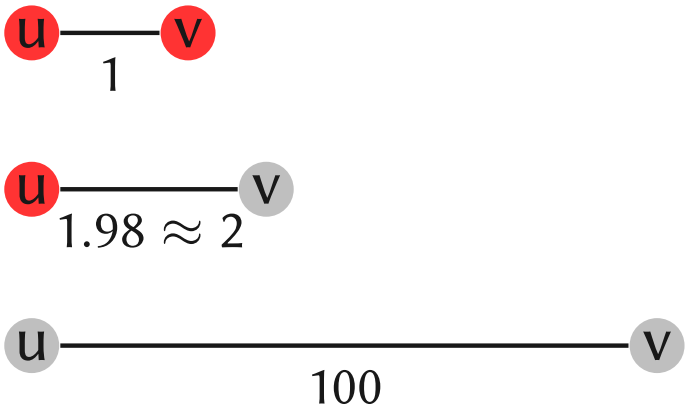




NetShort

- incorporates “disease-relevance” of a path
- the more seeds the path has, the shorter it is

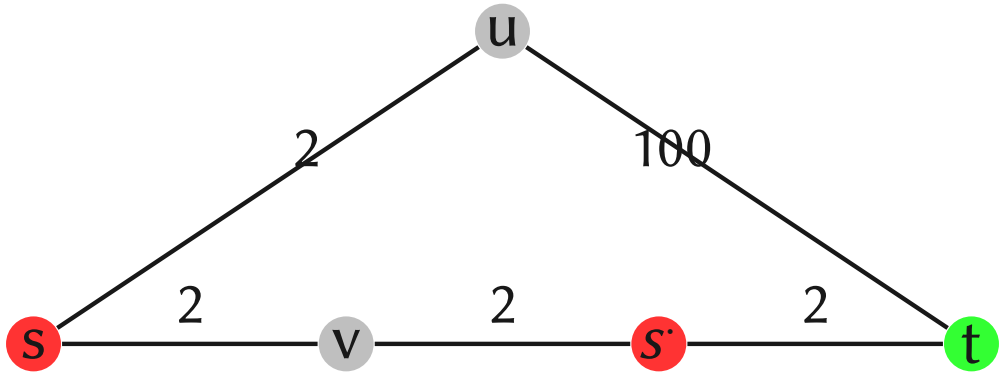


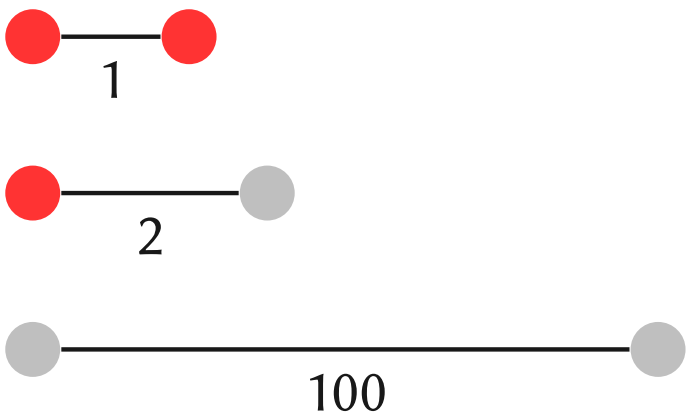


$$score(u) = \begin{cases} 1, & \text{if } u \text{ is seed} \\ 0.01, & \text{otherwise} \end{cases}$$

$$weight(u, v) = \frac{1}{(score(u)+score(v))/2}$$

NetShort





$$path\_length(s,t)_{\{s,u,t\}} = 2 + 100 = 102$$

$$path\_length(s,t)_{\{s,v,s',t\}} = 2 + 2 + 2 = 6$$

NetCombo

- combines the scores of the three methods
- calculates a z-score for each node using the score distribution of each method
- averages the z-score from all three algorithms for each node

$$score_u^{NetCombo} = \frac{1}{3} * \sum_{method \in \{NetScore, NetZcore, NetCombo\}} \frac{score_u^{method} - \mu^{method}}{\sigma^{method}}$$

GUILD accuracy

Data Set	Metric	NetScore	NetZcore	NetShort	NetCombo	Func. Flow	PageRank	Random Walk	Network Prop.
OMIM	AUC	67.49	62.99	65.63	**72.09**	58.55	57.03	55.36	65.97
	Sens.	20.69	19.62	15.41	21.46	22.31	10.76	14.64	**23.24**
Goh	AUC	**67.32**	61.45	55.36	67.08	54.78	52.39	49.35	54.74
	Sens.	**11.61**	11.05	4.88	11.34	6.22	4.00	5.69	8.66
Chen	AUC	75.92	72.80	63.11	**78.41**	63.56	65.30	61.78	69.07
	Sens.	**18.89**	12.84	9.06	17.51	12.43	6.00	9.64	15.30

Improvement over existing methods is significant

- for NetCombo in all three datasets
- for NetScore in Goh and Chen datasets

Several applications of GUILD

- implicating genes in AD *(Guney and Oliva, 2012, PLoS ONE)*
- extending apoptosis pathway *(Planas et al., 2012, OMICS: A Jour Int Bio)*
- prioritizing genes in bone metastatic breast cancer *(Santana-Codina et al., 2013, Mol & Cel Prot)*
- identifying subnetworks driving brain and lung metastasis in breast cancer *(Engin et al., 2013, PLoS ONE)*
- analysis of functional diversity of disease genes *(Guney and Oliva, 2014, PLoS ONE)*
- drug repurposing in brain metastatic breast cancer *(Martinez-Aranda et al., 2015, Oncotarget)*
- prioritization of regulatory modules in insulin secretion *(Hänzelmann et al., 2015, Islets)*