Bioinformatics — Lecture 7
Network Biology
(EG Ch. 13, MM. Ch. 10)

Krzysztof Bartoszek

Linköping University

krzysztof.bartoszek@liu.se

10 XII 2024 (R35)

Today

Introduction

Random graphs

Examples of Biological Networks

Inferring Networks

Additional reading

D R. Durret. Random Graph Dynamics, Cambridge, 2007. Cambridge University Press

L A.M. Lesk. Introduction to Bioinformatics, Oxford, 2014. Oxford University Press

R S. Raychaudhuri. Computational Text Analysis for Functional Genomics and Bioinformatics, 2006, Oxford University Press

Graphs

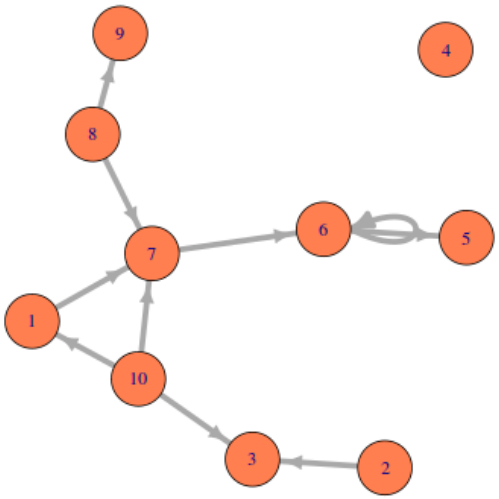
$$\mathcal{G} = (V_{\mathcal{G}}, E_{\mathcal{G}})$$

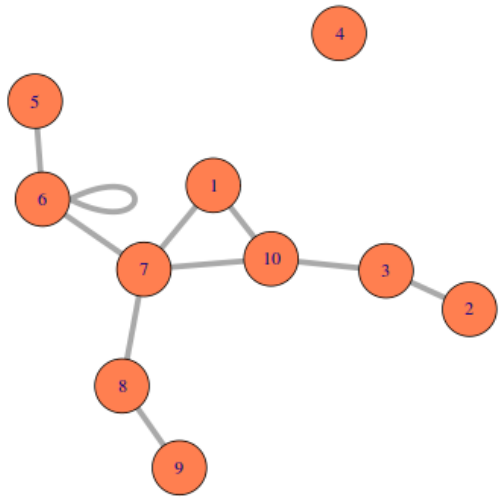
$V_{\mathcal{G}}$: set of *vertices*

$E_{\mathcal{G}}$: set or *edges*, unordered or ordered pairs of vertices

Undirected or directed graphs

Representation: adjacency matrix , adjacency list, edge matrix

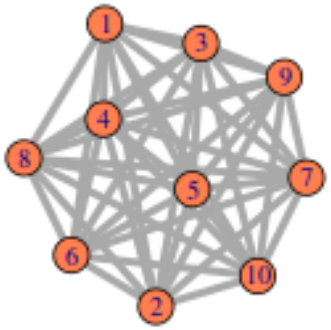




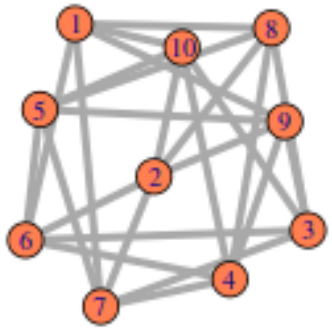
CRAN: `igraph`; Biconductor: `Rgraphviz`

Examples of graphs

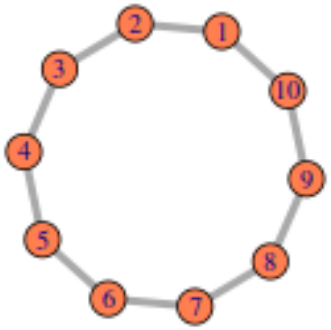
Full graph



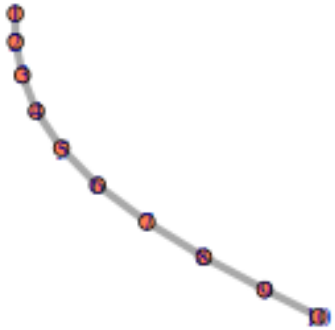
Full bipartite graph



Ring graph (cycle)



Path



code: `732A51_BioinformaticsHT2024_Lecture07codeSlide05graphs.R`

Analyzing graphs

degree of node: number of neighbouring nodes (*indegree*, *outdegree*) maximum degree, minimum degree
diameter of graph: length of longest shortest path
Component: *induced subgraph* where all nodes are connected by a *path*

Some properties:
Shortest Path Length (Dijkstra's algorithm $\in O(|V_{\mathcal{G}}|^2 \log |V_{\mathcal{G}}|)$)
Mean Path Length
Square of Adjacency matrix: number of paths of length 2

Decomposition: Split of graphs into subgraphs such that each edge is in exactly one
Path Decomposition
Cycle Decomposition

Clustering coefficient

M: adjacency matrix, n_k: number of nodes of degree k

$$C_i = \frac{\sum_{j,k} \mathbf{M}_{ij}\mathbf{M}_{jk}\mathbf{M}_{ki}}{\text{degree}(i)(\text{degree}(i)-1)}$$

$$\mathbf{C}(\mathbf{k}) = \frac{1}{n_k} \sum_{i:\text{degree}(i)=k} C_i$$

Counting triangles: number of edges between nodes compared to the maximal possible number of them
Module in network: clusters where density of edges is higher than expected

Betweenness

$$b(v) = \sum_{V_{\mathcal{G}} \ni i,j \neq v} \frac{\#\{\text{shortest paths from } i \text{ to } j \text{ passing through } v\}}{\#\{\text{shortest paths from } i \text{ to } j\}}$$

Centrality of a node v: number of shortest paths between two vertices passing through v compared to the number of shortest paths between these nodes

Bottleneck, Hub: nodes with largest values of betweenness, controlling connectivity of graph

Erdős–Rényi graph (D Ch2)

$$V_{\mathcal{G}} = \{1, \ldots, n\} \qquad \forall_{x,y \in V_{\mathcal{G}}} P(\{x, y\} \in E_{\mathcal{G}}) = p$$

Take $p = \lambda/n$
$\lambda < 1$: largest component $\in O(\log n)$
$\lambda > 1$: a giant component $\sim g(\lambda)n$
$\lambda < \log n$: probabilty of connected graph $\to 0$
$\lambda \in (\frac{1}{2}\log n, \log n)$: with probability $\to 1$, giant component and isolated vertices
$\lambda > \log n$: probabilty of connected graph $\to 1$

Degree distribution: $p_k \sim$ Poisson

Barabási–Albert Model (D Ch4)

1. start with a small number of vertices m_0
($m_0 = 2$ with m parallel edges)
2a. at each step add a vertex connecting it with m vertices
choose those vertices with probabilty $\deg(i)/\sum \deg(i)$
(preferential attachement, *one edge at a time with update degrees*)
2b. choose vertices with probability $\sim \deg(i)^{\gamma}$

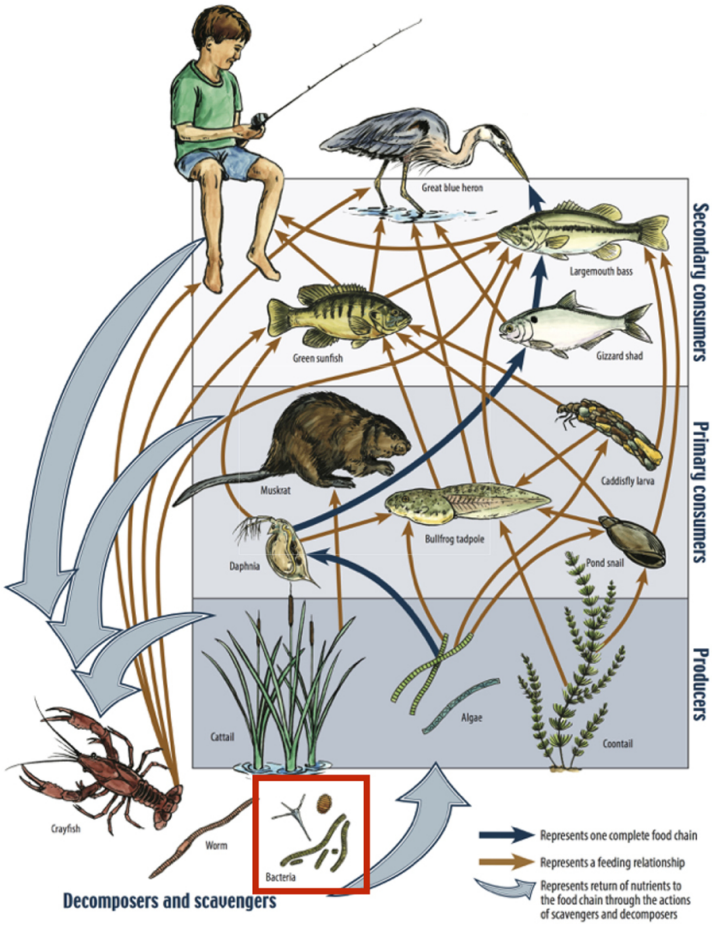
Degree distribution: power law $p_k \sim Ck^{-\beta}$
Scale free network (properties same with network growth)

resiliant to random damage
sensitive to targeted attacks (hubs)

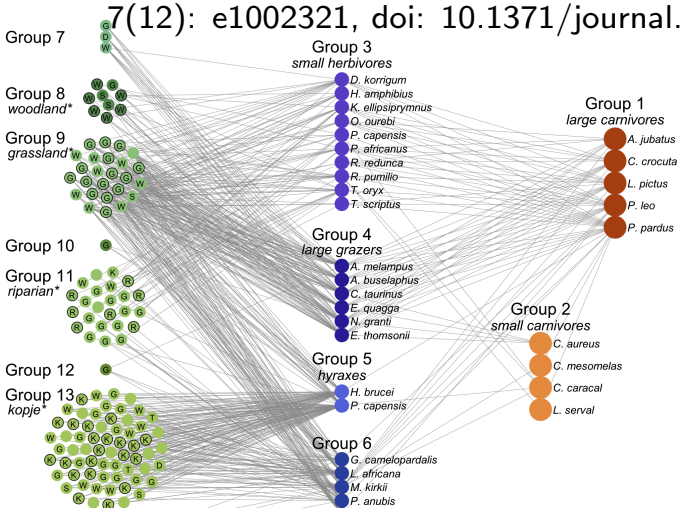
gene duplication: proteins with most links more likely to gain new

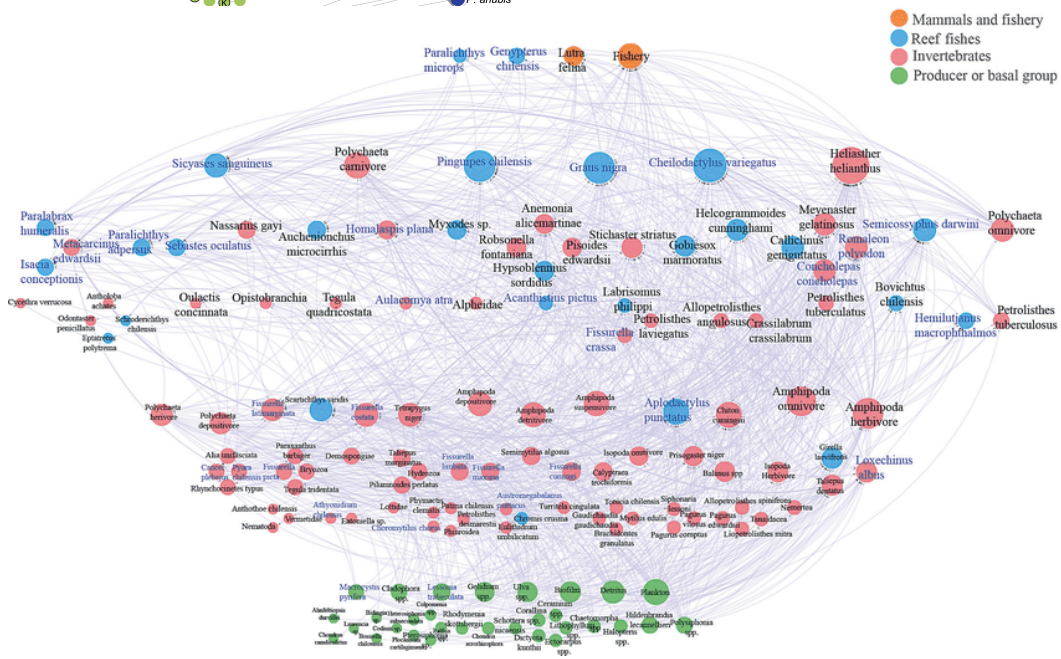
E.B. Baskerville et al. (2011), Spatial Guilds in the Serengeti Food Web Revealed by a Bayesian Group Model. PLoS Comput Biol 7(12): e1002321, doi: 10.1371/journal.pcbi.1002321, CC Lic.

Food web (trophic levels)



Wikimedia Commons, K. Schulz, M.W. Smit, L. Herfort, H.M. Simon, image (amended) provided courtesy of the Missouri Department of Conservation, doi:10.3389/frym.2018.00004, CC BY-SA 4.0



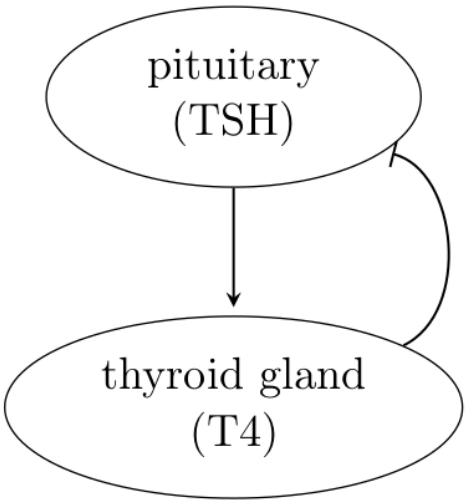


A. Pérez-Matus et al. (2017), Temperate rocky subtidal reef community reveals human impacts across the entire food web, Mar. Ecol. Prog. Ser. 567:1–16, doi: 10.3354/meps12057, CC BY-SA 4.0

Systems Biology

Modelling complex biological systems with many interacting components.
What model fits biological networks?

TSH: thyroid stimulating hormone; T4: thyroxine hormon

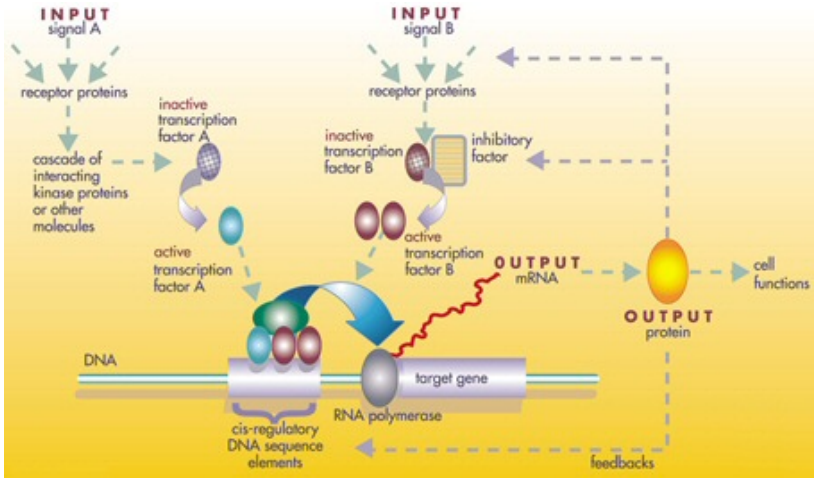


$$TRH'(s) = c - gTRH(s)$$
$$T4'(s) = aTRH(s) - bT4(s)$$

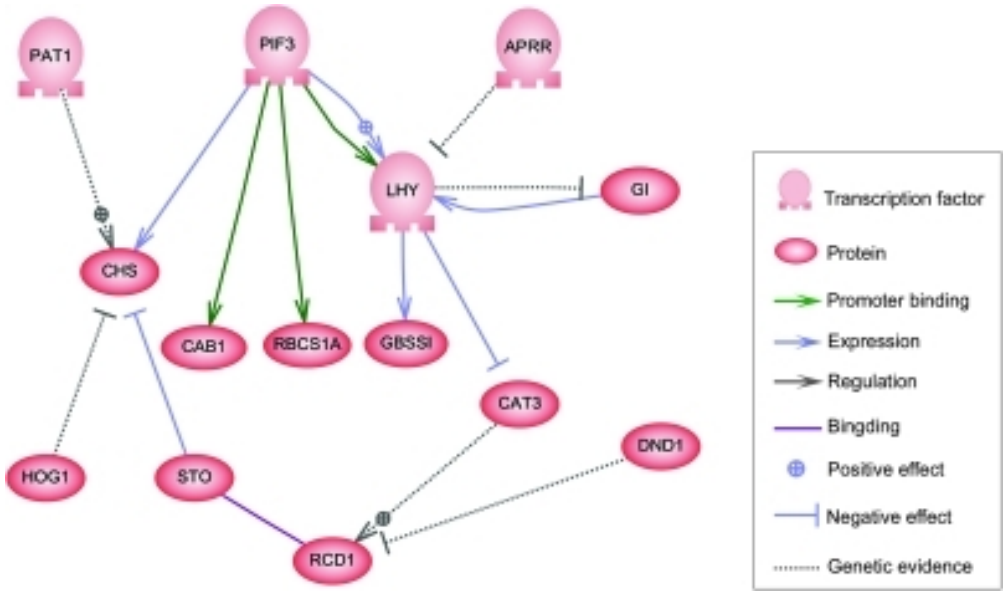
$$TRH'(s) = c - h \min\left\{T4(s - \tau_{TRH}), \frac{c}{h}\right\} - gTRH(s)$$
$$T4'(s) = aTRH(s - \tau_{T4}) - bT4(s)$$

A. Bartłomiejczyk, B. Jackowska–Zduniak, 2018. Dynamics of simplified HPT model in relation to 24h TSH profiles. Mathematica Applicanda 46, 16-24 doi:10.14708/ma.v46i1.6389 (reproduced with Authors' permission).

Gene regulatory networks (L Ch9)



source `http://genomics.energy.gov` (public domain)

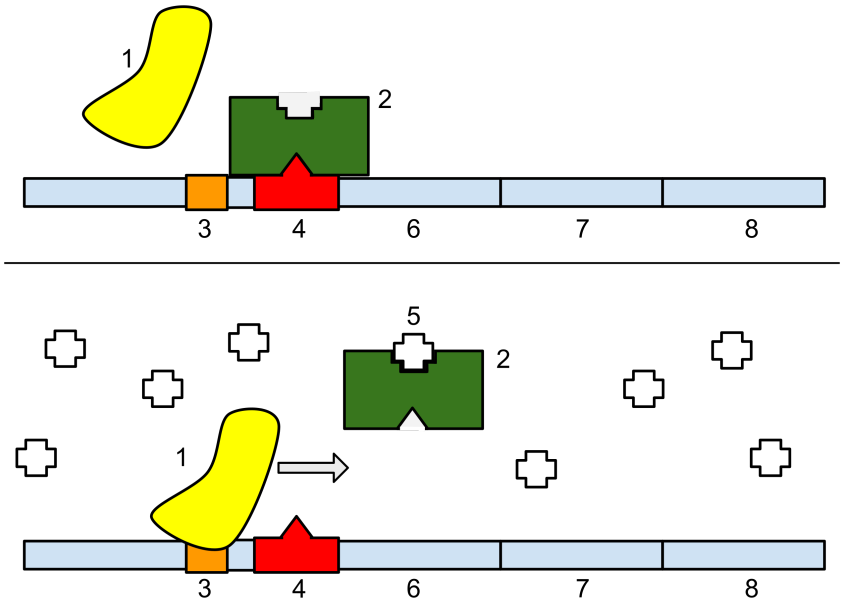


source `http://openi.nlm.nih.gov/detailedresult.php?img=2993235_mplantssq046f07_4c&req=4` (public domain)

`https://en.wikipedia.org/wiki/Gene_regulatory_network`

"collection of molecular regulators that interact with each other and with other substances in the cell to govern the gene expression levels of mRNA and proteins which, in turn, determine the function of the cell. ... The regulator can be DNA, RNA, protein and complexes of these. The interaction can be direct or indirect (through transcribed RNA or translated protein)."

Lactose operon



1: RNA polymerase

2: Repressor

3: Promoter

4: Operator

5: Lactose

6–8: operon structural genes

https://en.wikipedia.org/wiki/Lac_operon
by T A RAJU (own work), CC BY-SA 3.0

The operon is needed for the transport and metabolism of lactose in *Escherichia coli* and some other bacteria. While glucose is preferred, tha lac operon allows lactose to be digested, when glucose is lacking. First gene regulatory mechanism to be understood.

Conversion of glyceraldehyde–3–phosphate is in **both** paths to pyruvate.

Metabolic network (L Ch8)

Metabolite: molecule undergoing transformation

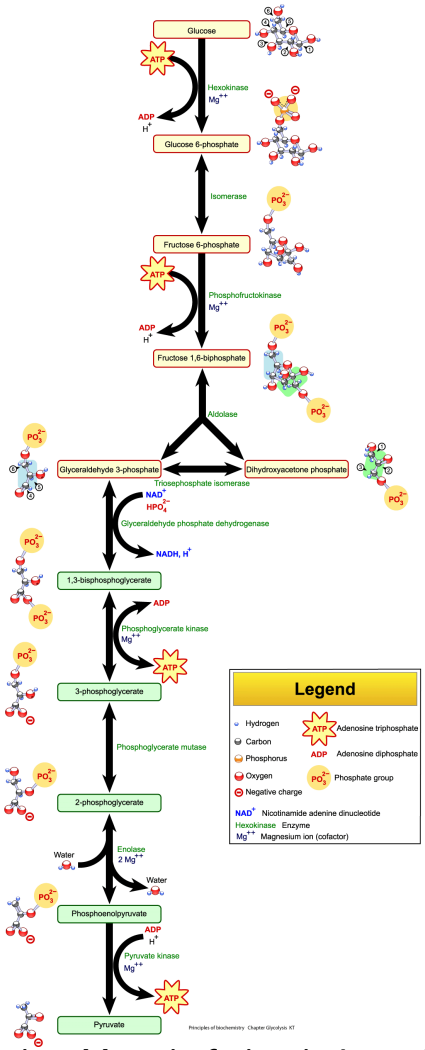
*Metabolic pathway***s**: a network representing transformations

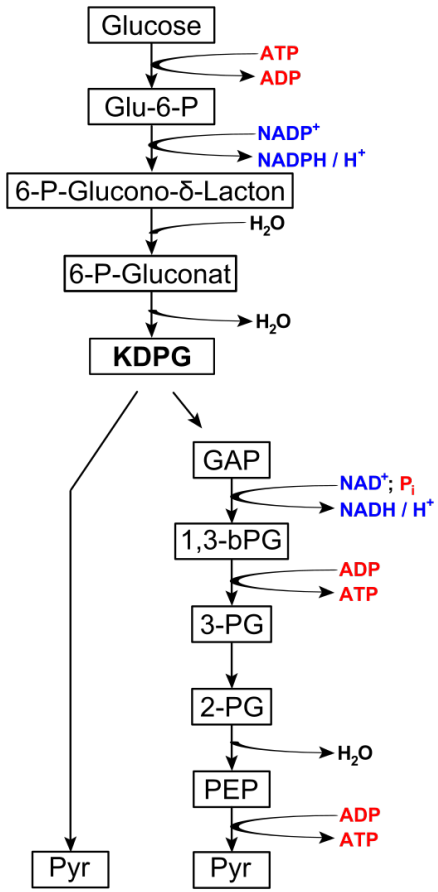
Nodes: metabolites

edges: reactions

directed edges: irreversible reactions

catalyzer: edge label

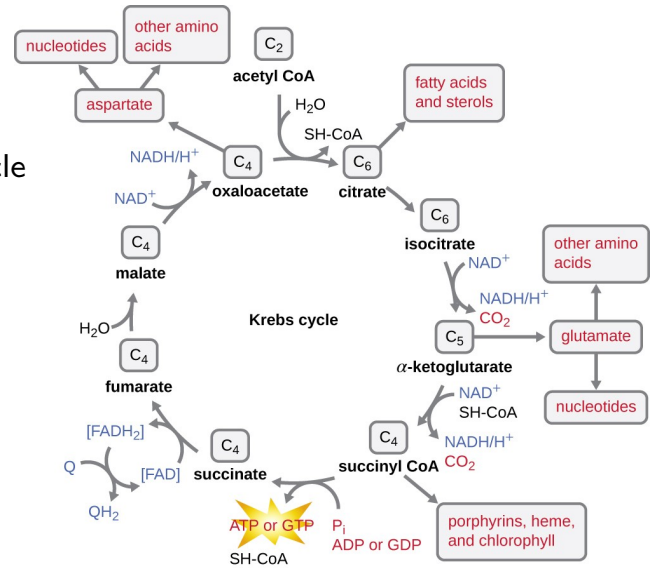




Entner–Doudoroff pathway, Wikimedia Commons, by Yikrazuul, public domain

Embden–Meyerhof glycolytic pathway , Wikimedia Commons, by Narayanese, WikiUserPedia, YassineMrabet, TotoBaggins, CC BY-SA 3.0,

citric acid cycle



N. Parker, M. Schneegurt, A.-H.T. Tu, P. Lister, B.M. Forster, OpenStax Microbiology, 2016, CC BY-SA 4.0. Access for free at `https://openstax.org/books/microbiology/pages/1-introduction`

Robustness (L p323)

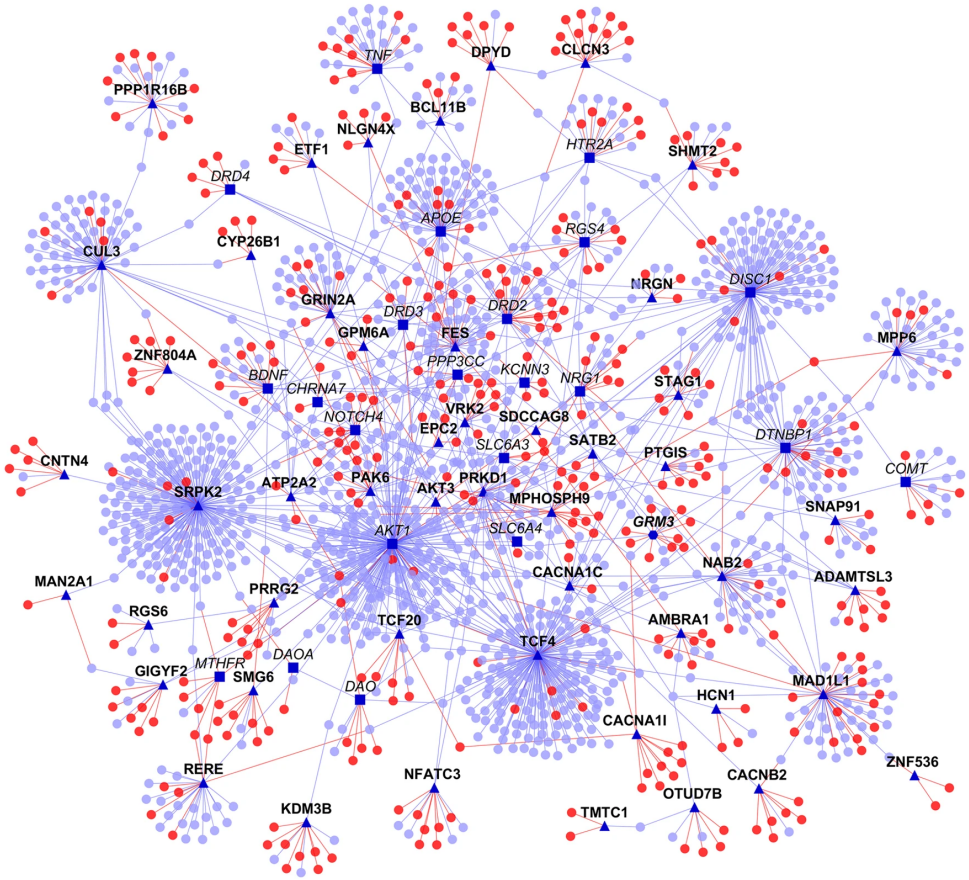
substitutional redundancy: two proteins capable of same job, can be due to gene duplication (similar sequence, similar protein, similar function)

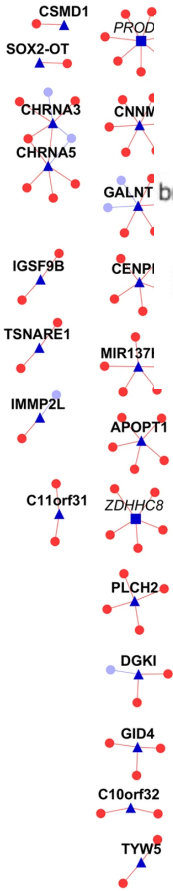
functional redundancy: one protein essential, the other sometimes activated for same job
NOT robust to loss of first protein

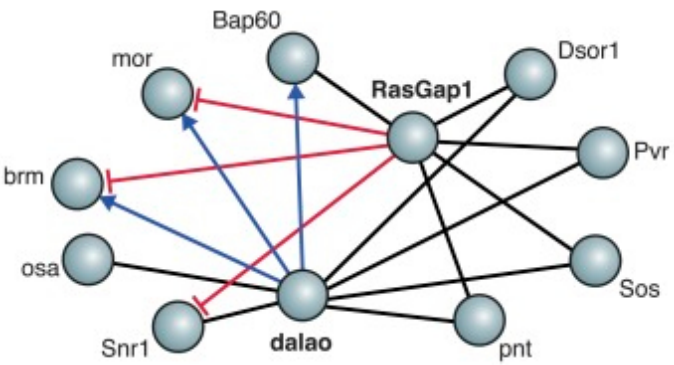
distributed redundancy: equivalent effects, different routes

Protein interaction network

https://en.wikipedia.org/wiki/Protein-protein_interaction





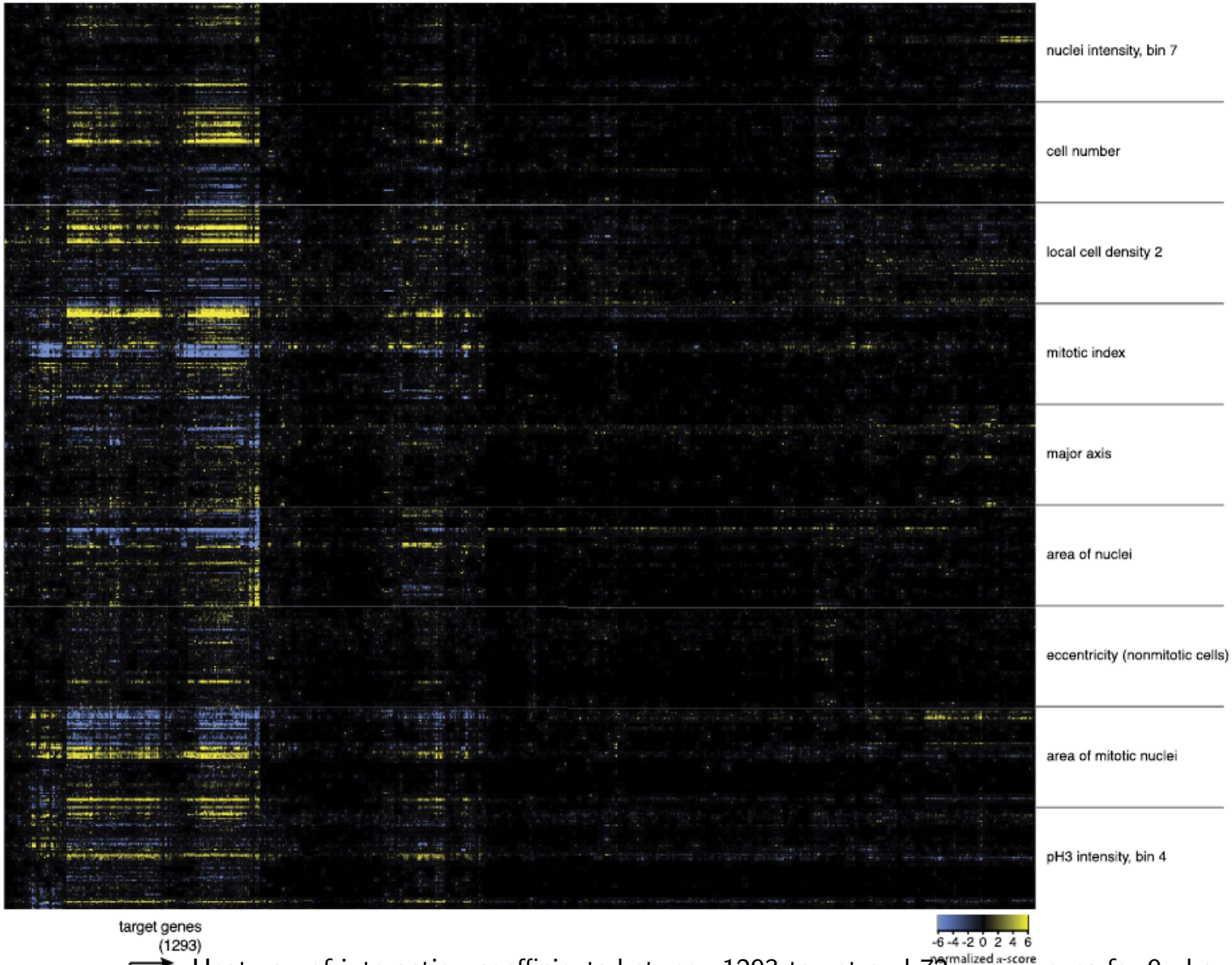


Protein–protein interactions are represented as a signed graph with different types of edges, corresponding to different interaction types.

B. Fischer, T. Sandmann, T. Horn, M. Billmann, V. Chaudhary, W. Huber, M. Boutros, 2015. A map of directional genetic interactions in a metazoan cell. eLife 4:e05464. doi:10.7554/eLife.05464, CC BY-SA 4.0

Schizophrenia interactome, genes: nodes, edges: protein–protein interactions

Fig. 2 in M. Ganapathiraju et al. (2016), Schizophrenia interactome with 504 novel protein-protein interactions. npj Schizophr 2:16012, doi: 10.1038/npjschz.2016.12, CC BY-SA 4.0

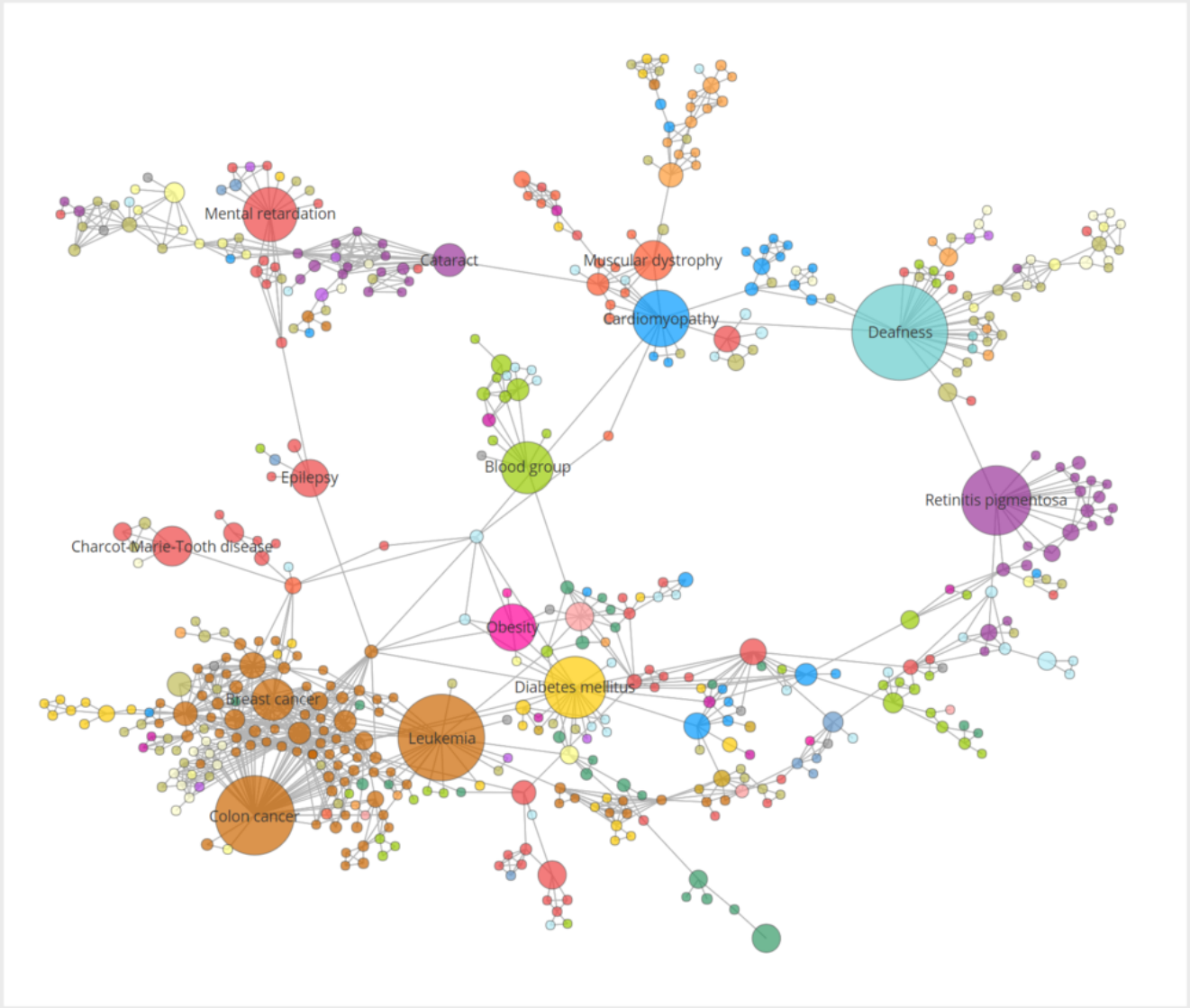


Heatmap of interaction coefficients between 1293 target and 72 query genes for 9 phenotypic features. Blue: negative interactions, yellow: positive.

Fig. 3 (Supp. 1) of B. Fischer et al. (2015) doi:10.7554/eLife.05464, CC BY-SA 4.0

Genotype disease network

https://en.wikipedia.org/wiki/Human_disease_network.png



graphic by Empetrisor
Wikiemedia Commons, CC BY-SA 4.0
Two diseases are connected if they share a genetic component.

Pathways

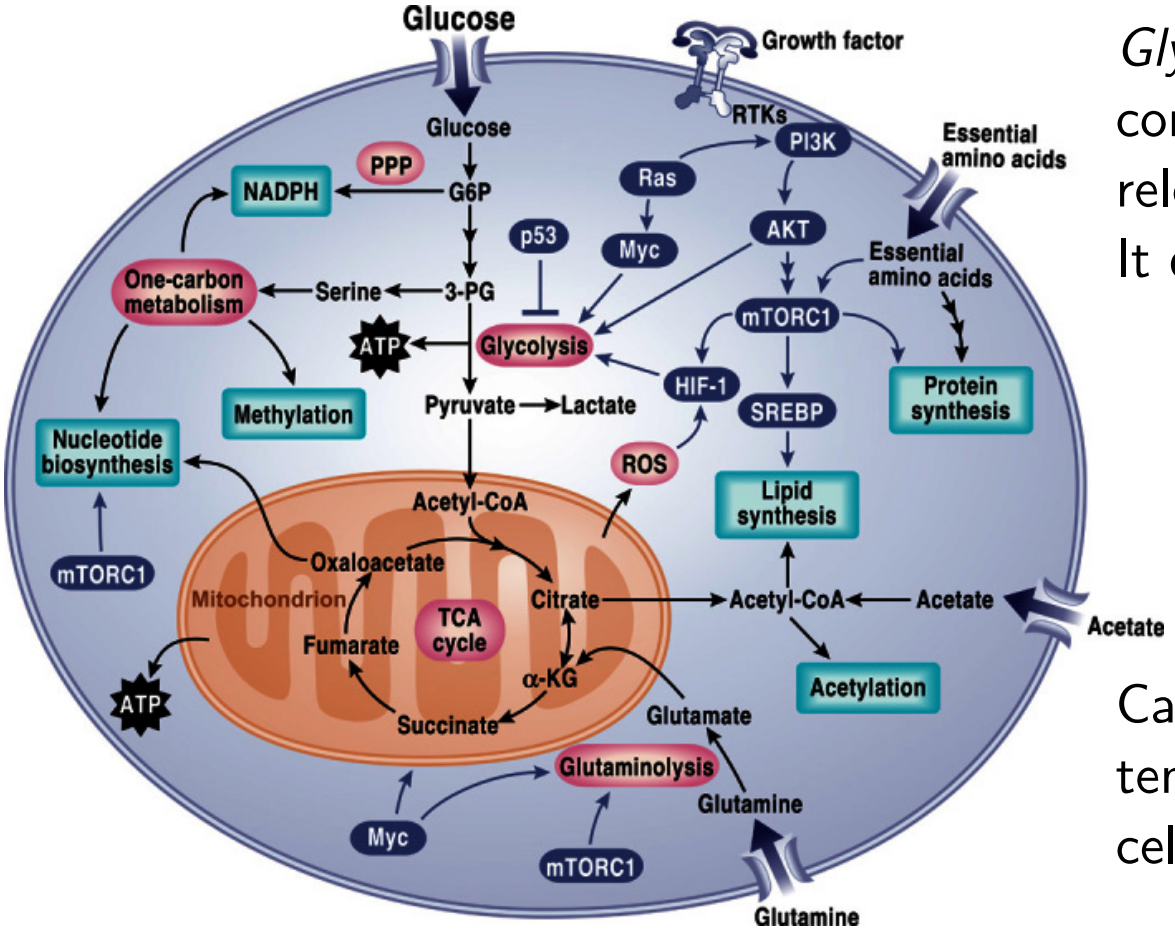
Path redundancy

Cancer : altered pathways

Hub–knockouts: p53 hub protein, inactivated by mutation in multiple human (breast) cancers; apoptosis signalling pathway disrupted

Drug resistance: drug targets a given pathway, but other pathways can also result in disease

p53: "Guardian of the Genome"; prevents cancer formation.



Glycolysis: metabolic pathway that converts glucose, into pyruvic acid, releasing free energy that forms ATP. It does not require oxygen.

Cancer cells can perform glycolysis ten times faster than noncancerous cells.

Fig. 1. of R.J. DeBerardinis, N. S. Chandel, 2016. Fundamentals of cancer metabolism. Sci Adv. 2(5): e1600200 doi:10.1126/sciadv.1600200, CC BY-NC 4.0

Tumour cells are often, at genesis, starved of oxygen. Glycolysis may provide them with ATP.

Pathways

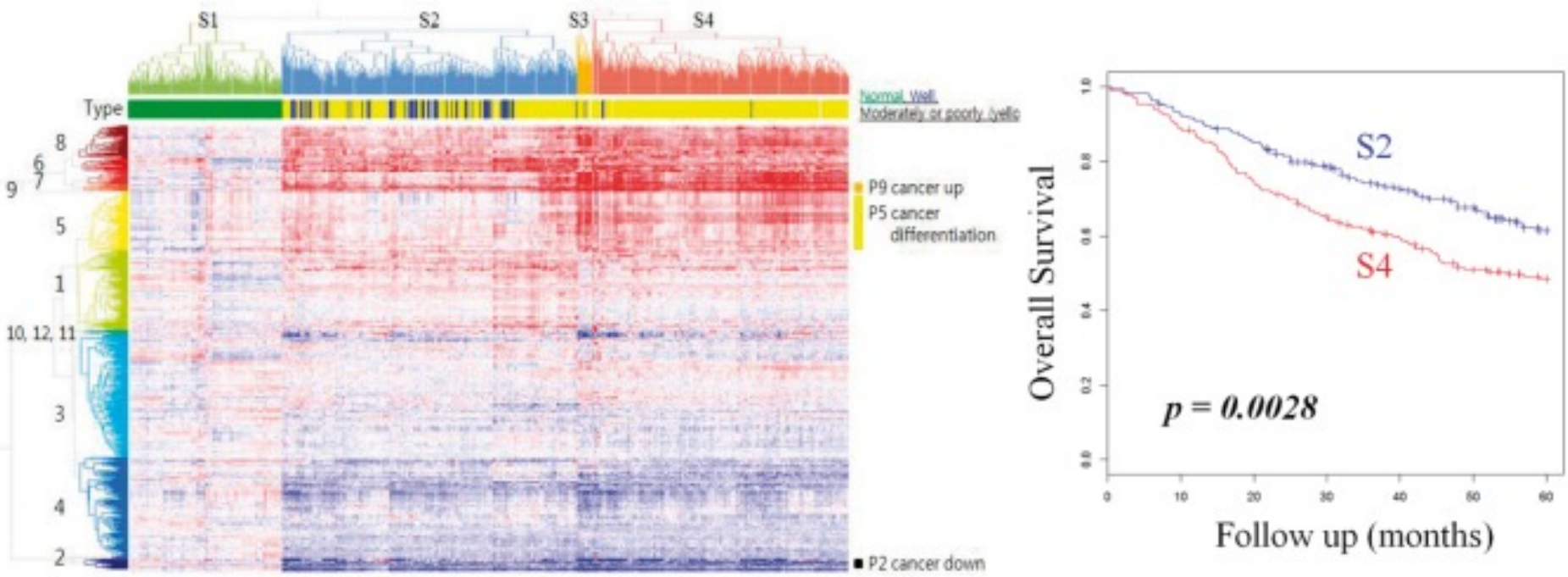


Fig. 3 of TJ. Ahn, E. Lee, N. Huh, T. Park, 2014. Personalized identification of altered pathways in cancer using accumulated normal tissue data, Bioinformatics, 30(17), i422-i429, doi:10.1093/bioinformatics/btu449, CC BY-NC 4.0

S1: normal samples, S2-S4 tumours

rows: pathways, columns: samples

red:up–regulated, blue: down–regulated

Gene enrichment analysis

"Determination of statistical significance for functional category enrichment." *Hypergeometric distribution* as null model for p–value of observing a given number of genes from some functional category inside a cluster of genes.

Probablity to observe at least k genes (ORFs in original) from a functional category within a cluster of size n (genes)

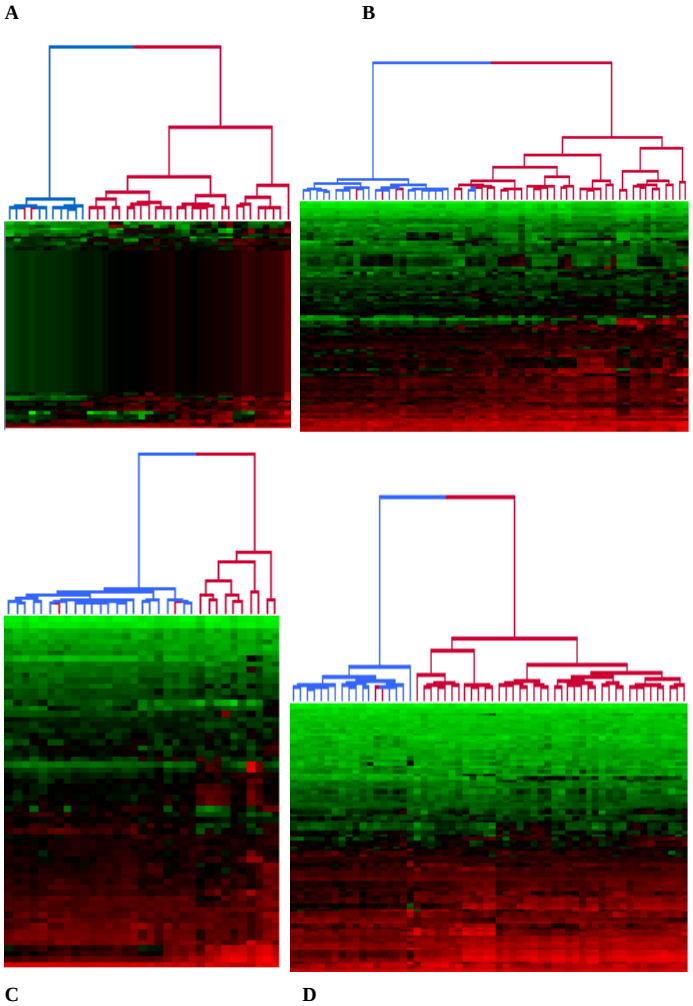
$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{f}{i}\binom{g-f}{n-i}}{\binom{g}{f}}$$

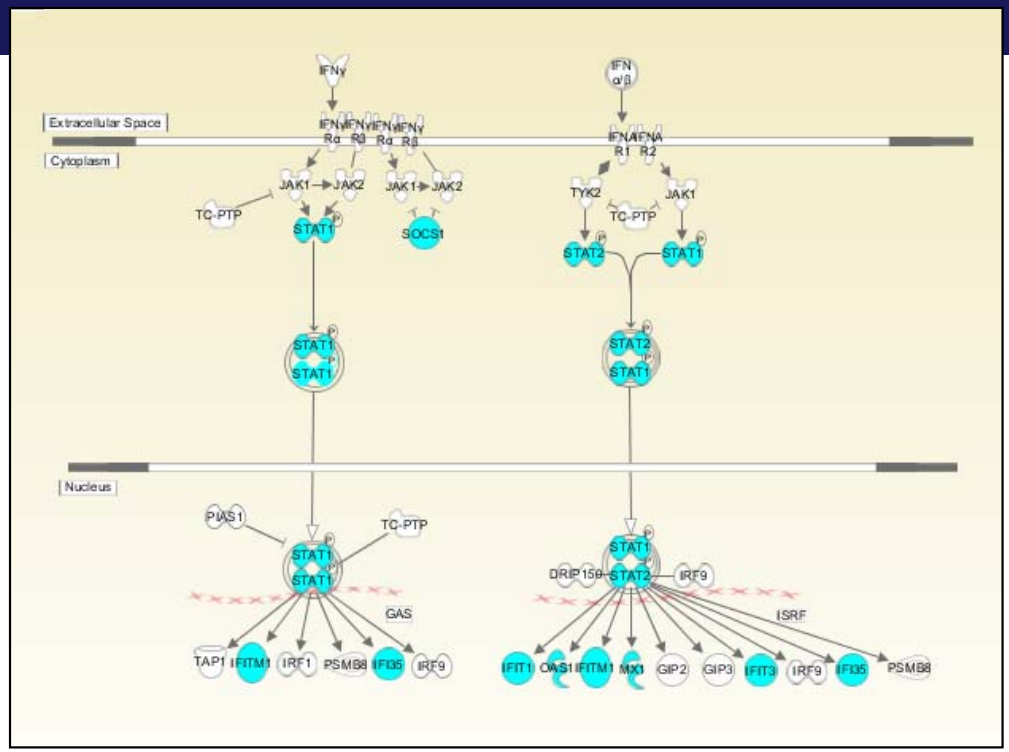
f: total number of genes from *the* functional category

g: total number of genes in the genome

S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, G. M. Church, 1999.
Systematic determination of genetic network architecture Nature Genetics, 22, 281-285

Inferring networks





"Hierarchical clustering analysis for all scenarios. Blue branches indicate healthy samples, and red branches indicate SLE patients."
Scenarios correspond to choice of test data.

Systemic lupus erythematosus: autoimmune disease, immune system mistakenly attacks healthy cells.

Figs. 3, 5 of D. Arasappan et al. (2011)Meta-analysis of microarray data using a pathway-based approach identifies a 37–gene expression signature for systemic lupus erythematosus in human peripheral blood mononuclear cells. BMC Med 9:65, doi: 10.1186/1741-7015-9-65, CC BY 2.0

Correlation based inference

Calculate correlation (Pearson, Spearman) between pairs of genes

interaction is *bidirectional* (who is regulating?)

cannot distinguish *direct* from *indirect* regulation

Gene set enrichment analysis (GESA)

Input: gene list ordered by expression levels,
expressions correspond to *phenotype*
gene set (pathway)

1. Calculate the *enrichment score* (ES) as the amount of overrepresentation of the genes from the set at the bottom or top of the list ordered by expression levels.
2. Find statistical significance of *ES* by a permutation test (assign random group to gene).
3. Correct for multiple testing (if analyzing multiple gene sets). Normalize ES values and calculate FDR.

`https://en.wikipedia.org/wiki/Gene_set_enrichment_analysis#Methods_of_GSEA`

A. Subramanian et al., 2005. Gene set enrichment analysis: A knowledge–based approach for interpreting genome–wide expression profiles. PNAS 102:43, 15545–15550., grouping called phenotype

Partial correlations

$\mathcal{P}$ set of p measured variables (gene expression levels) in multiple observations

$\vec{r}_x$: residuals of regressing x on $\mathcal{P} \setminus \{x, y\}$
$\vec{r}_y$: residuals of regressing y on $\mathcal{P} \setminus \{x, y\}$

partial correlation between x and y: Pearson correlation between $\vec{r}_x$ and $\vec{r}_y$

GeneNet algorithm (Graphical Gaussian models)

1. Estimate correlation and partial correlations between every pair of genes
2. Multiple testing correction of all significance of all estimates from 1.
3. Edges are only between those genes that have both correlation and partial correlations significant (undirected graph).

GeneNet R package `cran.r-project.org/web/packages/GeneNet`; `pypi.org/project/pygenenet/`

J. Schäfer, R. Opgen–Rhein, and K. Strimmer, 2006. Reverse engineering genetic networks using the GeneNet package. R News 6/5:50-53.

J. Schäfer and K. Strimmer, 2005. An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics 21:754-764 `doi:10.1093/bioinformatics/bti062`

Small sample estimate of partial correlations (more genes than samples)

B. Efron, 2004. Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. J. Am. Stat. Assoc. 99(465):96-104.

Causality (directed graph)

Standardized partial variance:

$$SPV_x = \mathrm{Var}[\vec{r}_x]/\mathrm{Var}[\vec{x}]$$

$$R_x^2 = 1 - SPV_x$$

$$\gamma_y^x = \log(SPV_x) - \log(SPV_y)$$

if $\gamma_y^x > 0$ *significantly*, then x has larger SPV_x and so smaller R_x^2: *less explained by other data*, edge would point $x \to y$ in the directed graph

Other Examples of Networks

Deep Learning for Network Biology:
`http://snap.stanford.edu/deepnetbio-ismb/`

RRosseta and Visunet:
`https://github.com/mategarb/R.ROSETTA`
`https://github.com/komorowskilab/VisuNet`

Questions?