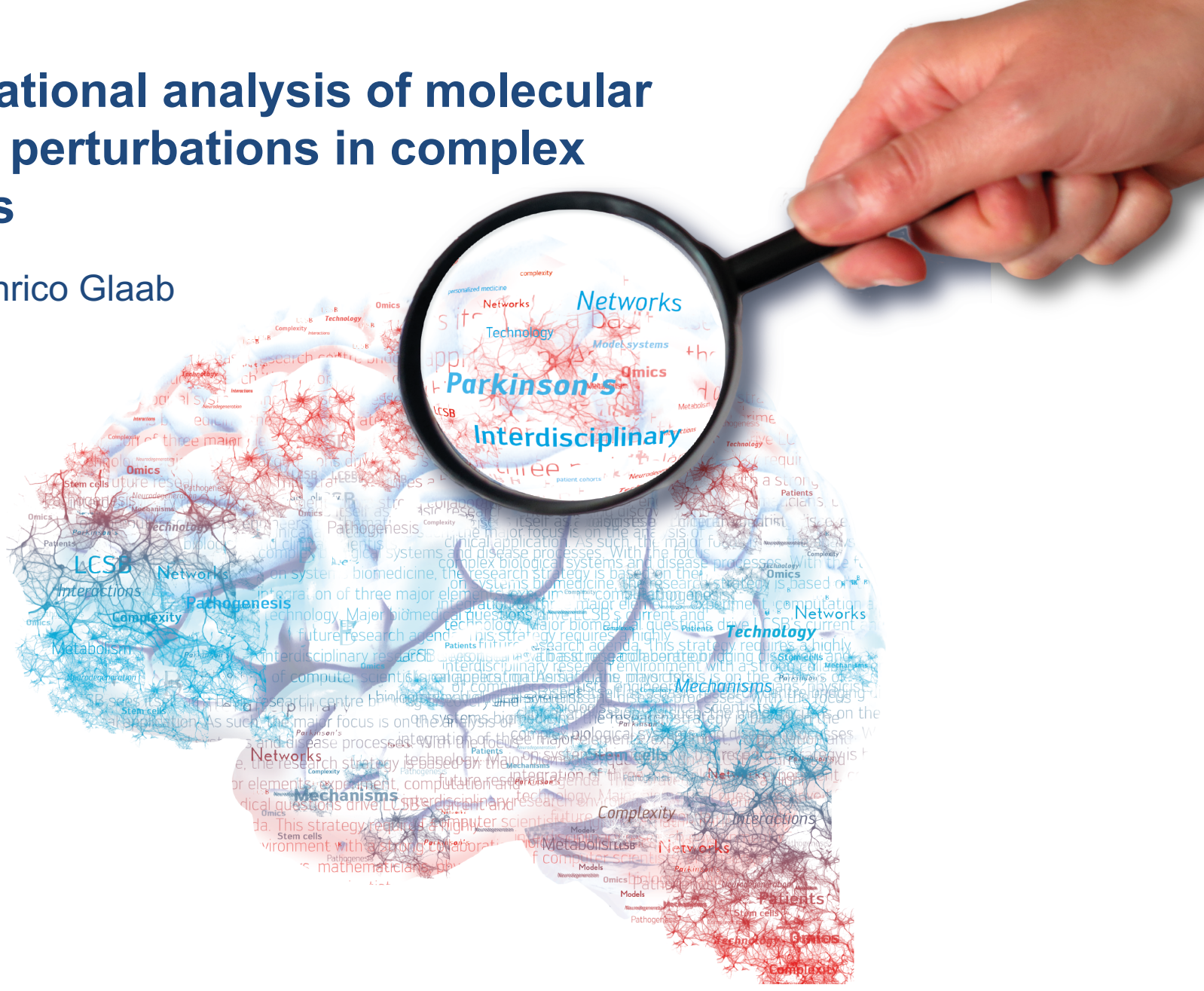
Computational analysis of molecular network perturbations in complex diseases

Speaker: Enrico Glaab

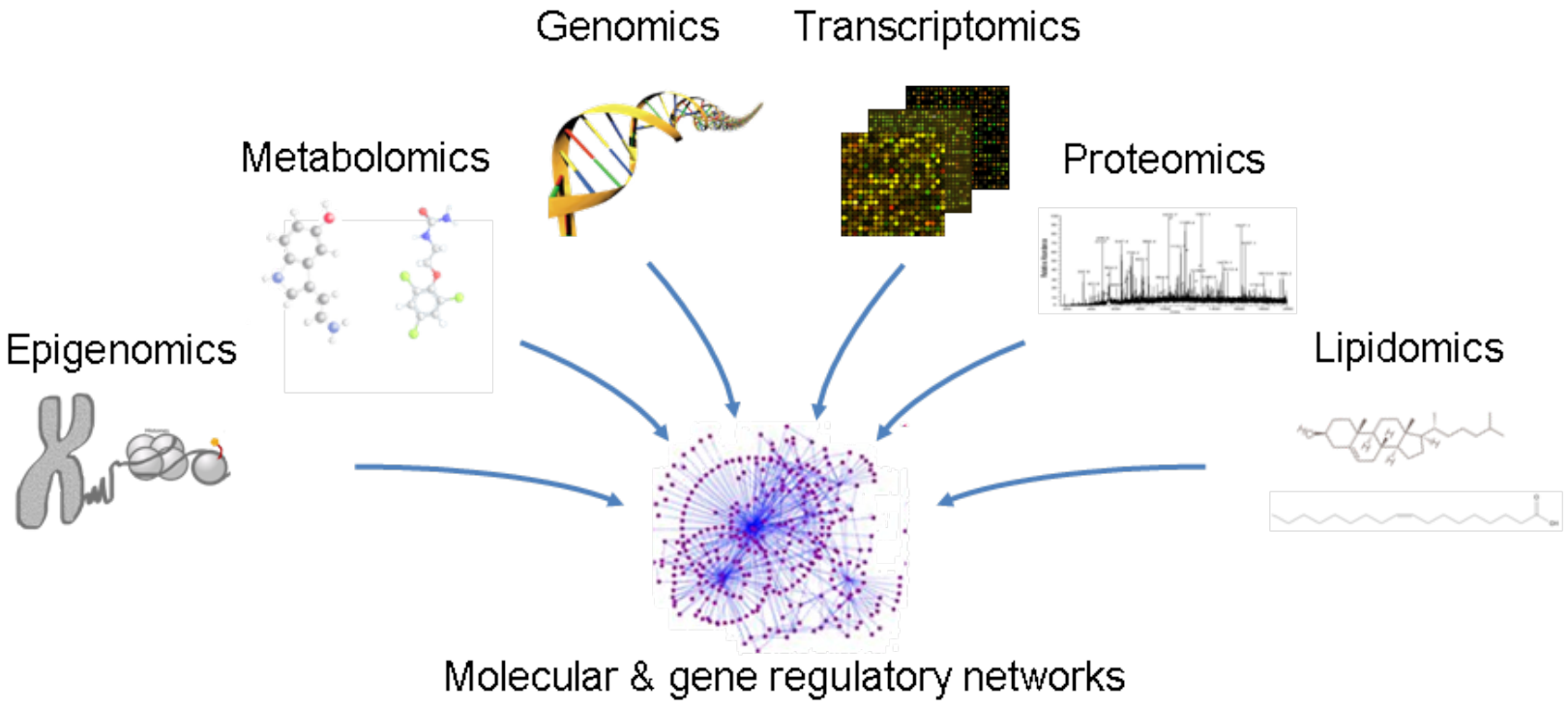


Luxembourg Centre for Systems Biomedicine (LCSB)

Luxembourg Centre
for Systems Biomedicine

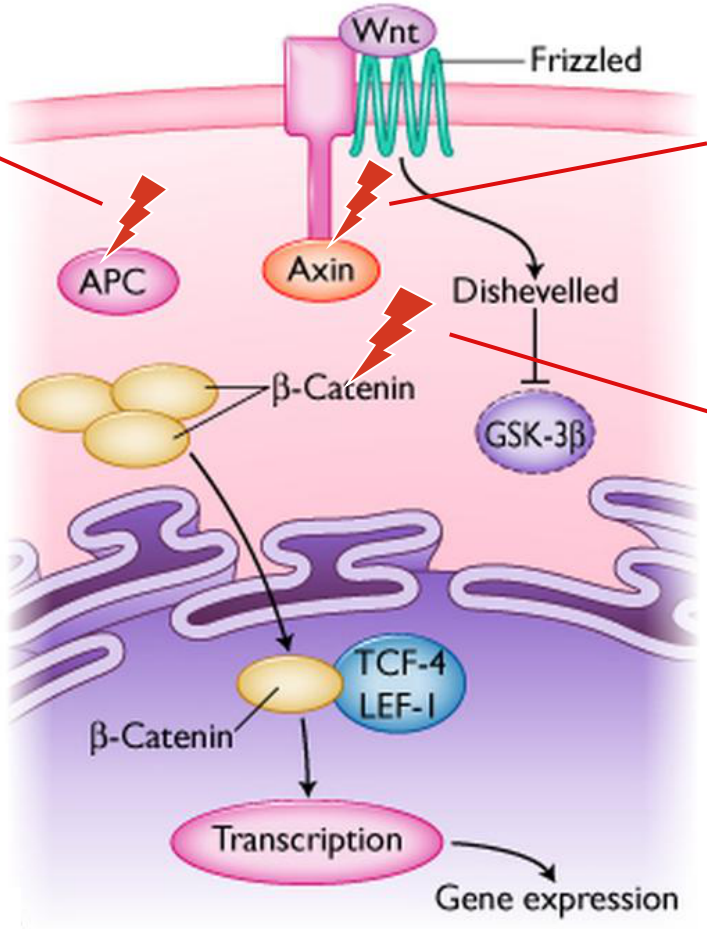
Group focus

Main goal: Interpret molecular changes in complex disorders by integrating diverse data types using network analyses



Motivation: Diseases as network perturbations

APC mutations
→ Colorectal cancer



AXIN1/2 mutations
→ Colorectal cancer

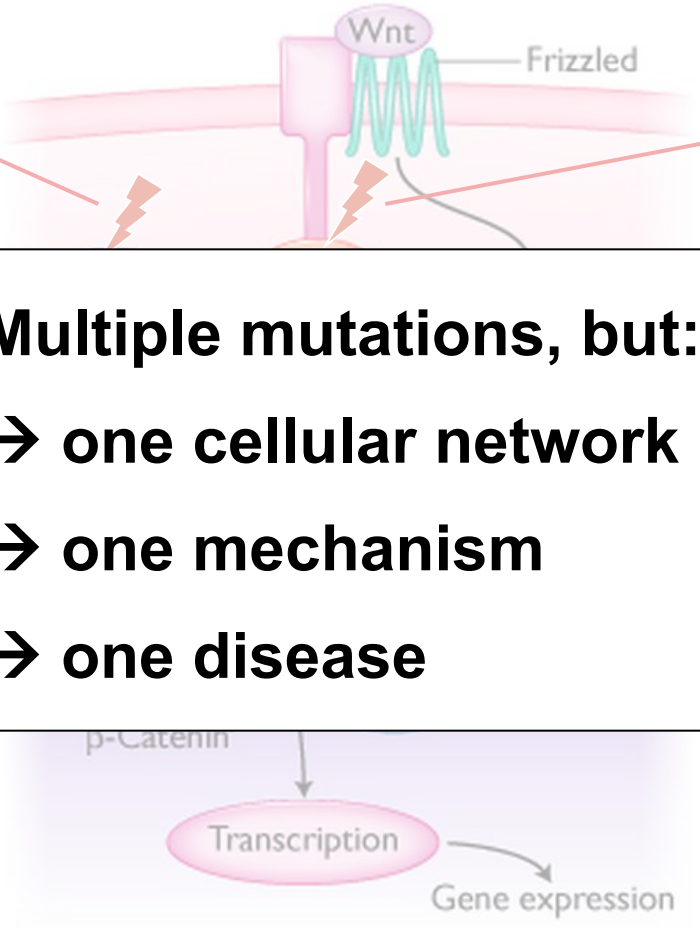
ß-Catenin mutations
→ Colorectal cancer

Wnt/β-catenin signaling pathway
(= affected by disease-related mutations)

Motivation: Diseases as network perturbations

APC mutations
→ Colorectal cancer

AXIN1/2 mutations
→ Colorectal cancer



Multiple mutations, but:

→ one cellular network

→ one mechanism

→ one disease

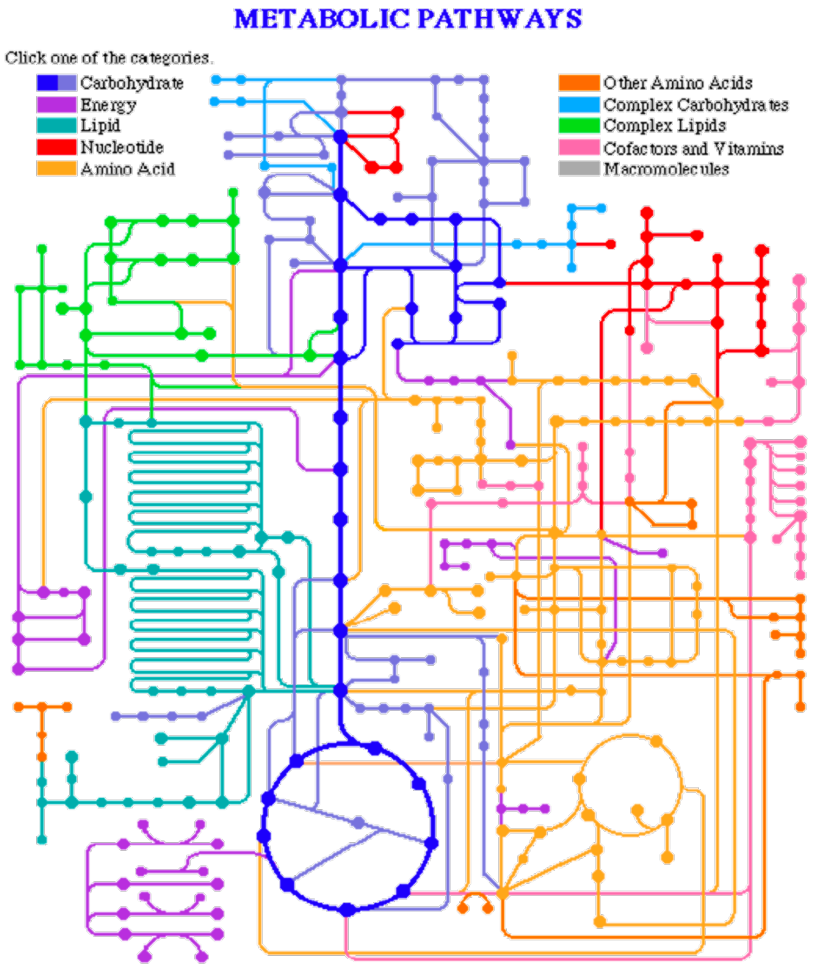
-Catenin mutations
→ Colorectal cancer

Wnt/β-catenin signaling pathway
(= affected by disease-related mutations)

Classical scientific approach: Reductionist Method

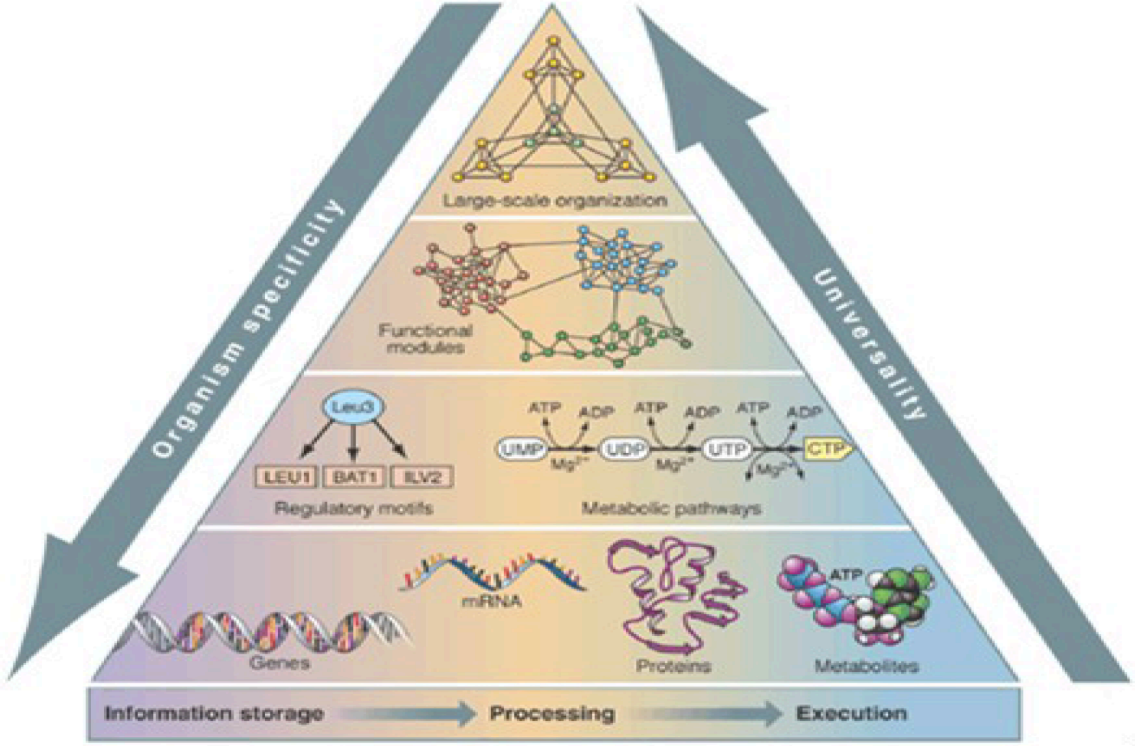
Reductionist Method:

- Hypotheses are specific and of narrow scope (local hypotheses)
- Understanding of an overall biological system (ecosystem, organism, cell) is supposed to be achieved by combining local insights
- however, the combinatorial nature of many biological systems challenges this method



Computational Systems Biology & Network Medicine

Systems biology: The study of an organism, viewed as an integrated and interacting network of genes, proteins and biochemical reactions.



Source: Oltvai & Barabasi, 2002

Computational Systems Biology (2)

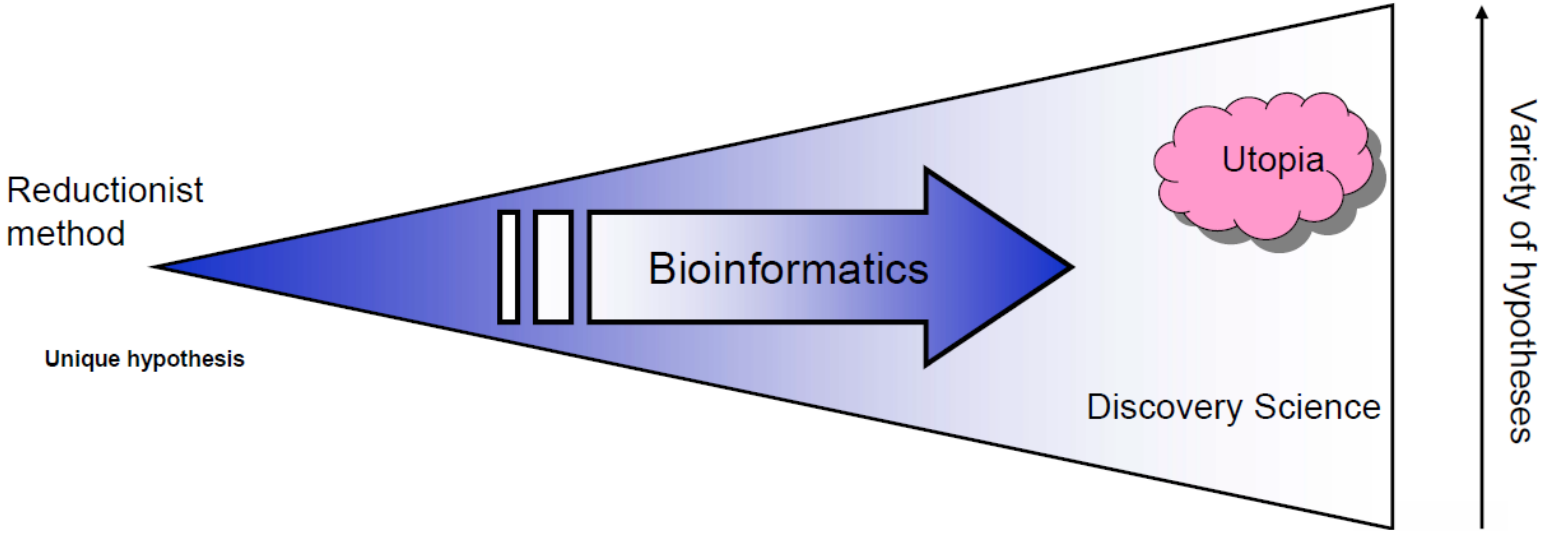
Two main driving forces:

- New high-throughput experimental profiling approaches enable organism-wide data collection:
  - Genomics → whole-genome sequencing
  - Transcriptomics → DNA chips / RNAseq
  - Interactomics → High-throughput Y2H screens

- Modern computers enable system-wide bioinformatics analyses for generating new valid or plausible hypotheses, e.g.:
  - which (combinations of) gene variations cause a disease?
  - which drugs inhibit the activity of target proteins most effectively?

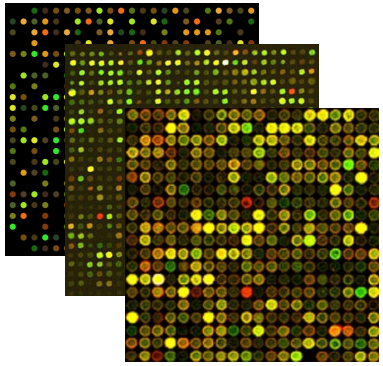
Computational Systems Biology (3)

- Bioinformatics is expected to drive progress in Systems Biology
- Some expectations may be too optimistic, but:
  - Bioinformatics can provide useful hypotheses for subsequent targeted experimental testing
  - Bioinformatics can help to select the most promising hypotheses from a larger set of plausible hypotheses

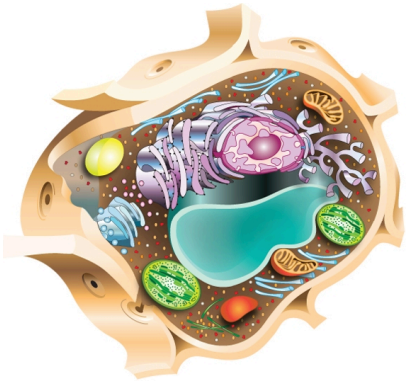


Challenge: From big data to biological function

Big data → **Statistics** → **Function**



Gene 1	4.3
Gene 2	–2.4
Gene 3	–3.1
Gene 4	1.8
...	...



Bottleneck:

Which changes are causal / secondary?

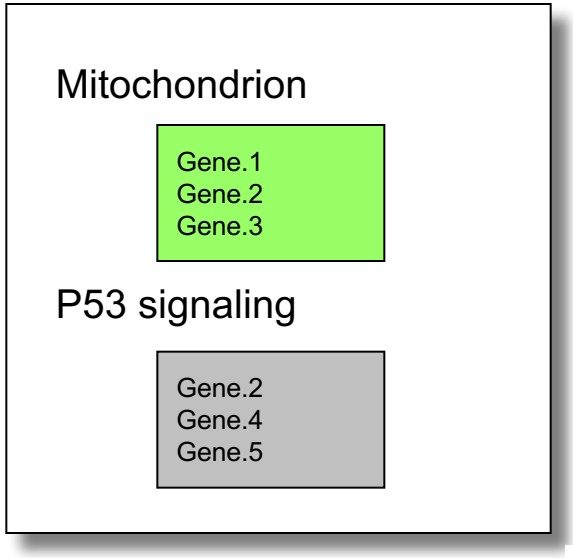
Which changes are correlative / predictive?

Which confounders modulate the system?

Which changes are disease-relevant / actionable?

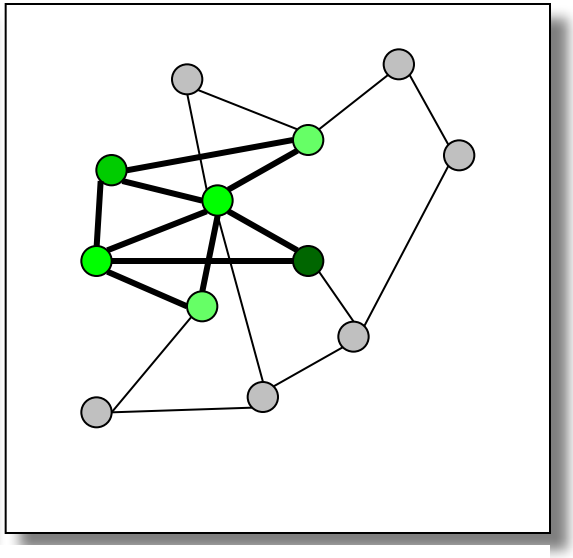
Representing and modeling cellular processes

GENE SETS



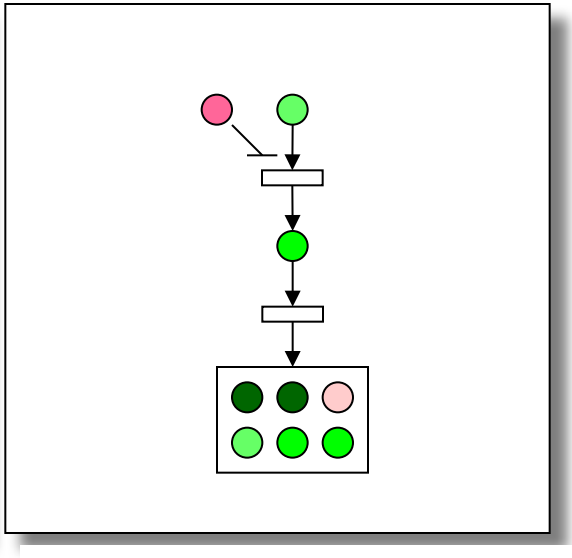
→ pure statistical scoring of enriched expression changes

NETWORKS



→ scoring of topological + expression criteria

PATHWAYS



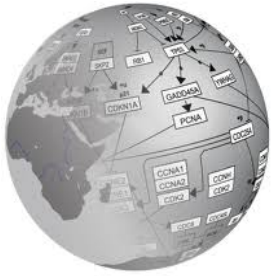
→ scoring of topology + expression changes + consistency criteria

Gene set / pathway resources

- Many public databases on functional gene sets and pathways available
- Both generic, multi-organism pathway collections covered and specialized collections (e.g. disease pathways: PD-Map, AlzMap)
- A total of over 10,000 public pathways available for the human species

BioCarta

KEGG Kyoto Encyclopedia of Genes and Genomes



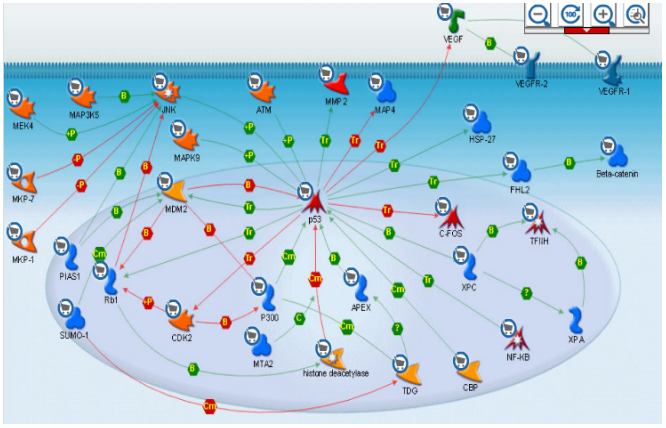
WikiPathways

GO the Gene Ontology

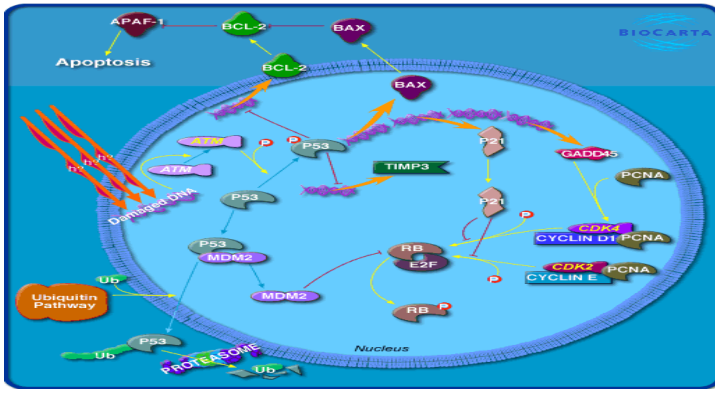
Reactome

Limitations of pathway databases

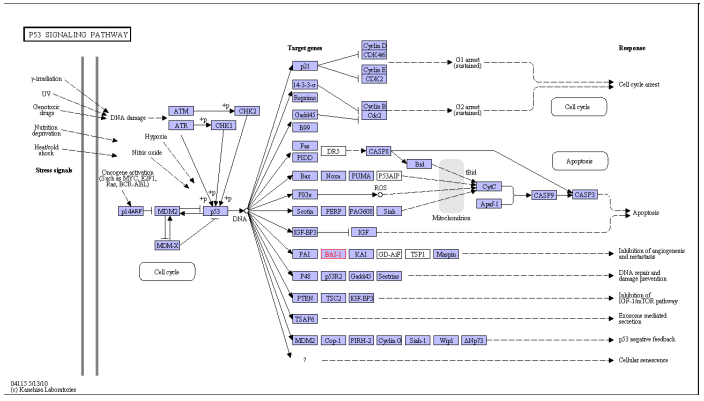
- manual curation → subjective decisions on pathway members & boundaries
- false-positive and false-negatives among molecular interactions
- database inconsistencies, e.g. “p53 signaling“:



Invitrogen iPath (p53 signaling)



BioCarta (p53 signaling)



KEGG (p53 signaling)

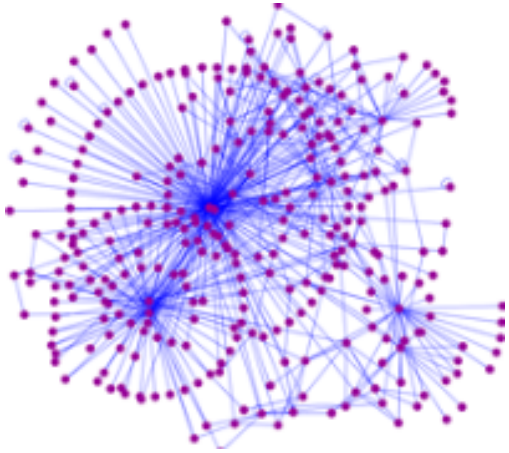
Improving pathway definitions using networks

- **Questions**: Can we make pathway definitions more objective? Can we improve existing pathways according to quantitative criteria?

- **Strategy**: Use genome-scale networks to redefine pathways:

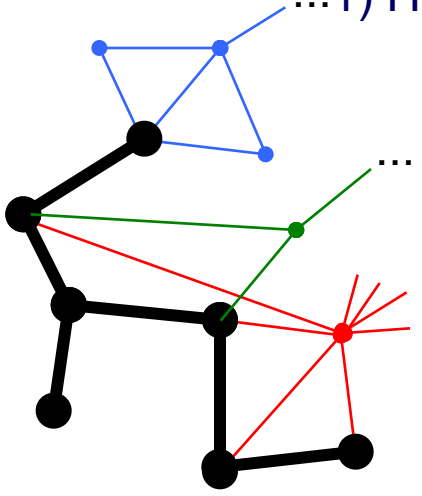
  - protein-protein interactions
  - genetic interactions
  - gene co-expression relations

  → large-scale, higher coverage, less biased
  → can also reveal communication between pathways (“cross-talk“)



Network-based pathway extension

Idea: Extend pathways by adding genes according to graph-theoretic criteria:



...1) High density of connections (“triangle-links“)

...2) High specificity of connections (more in than out)

...3) High coverage of connected pathway members

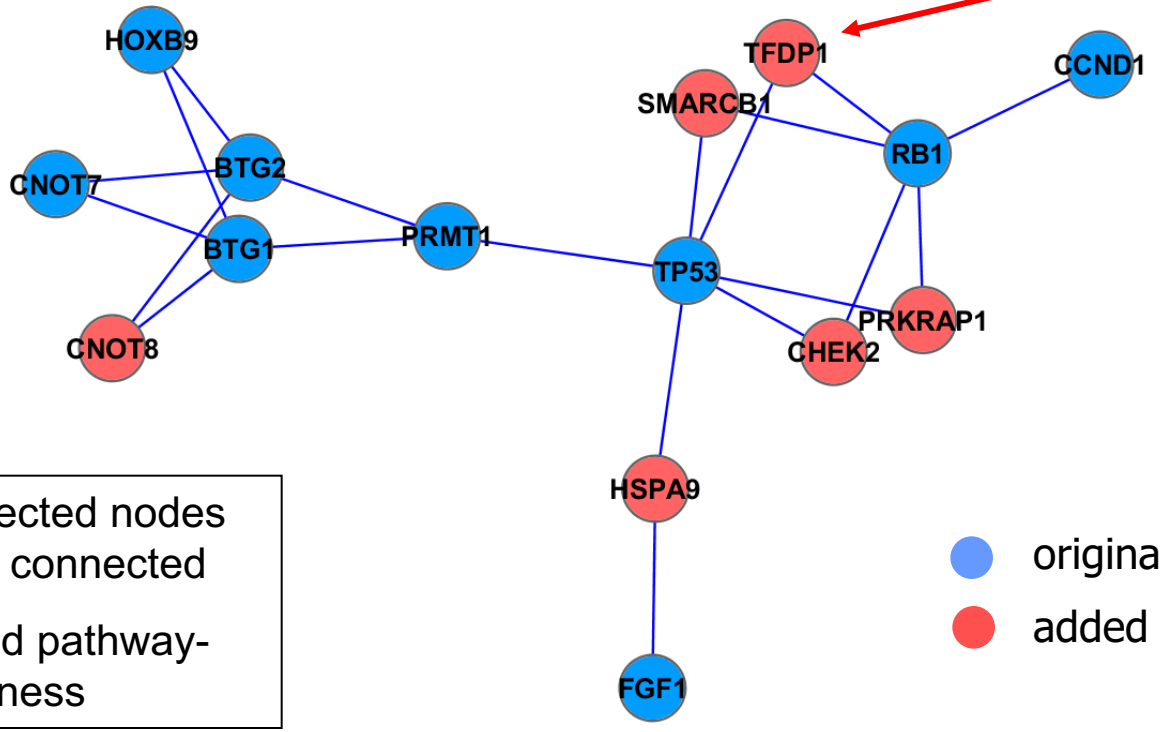
black = pathway members

red **blue** **green** = new candidate pathway members

Automated pathway extension: Example

Known cancer pathway: "*BTG family proteins and cell cycle regulation*" (BioCarta)

Mutations linked to colorectal cancer



→ Disconnected nodes become connected

→ increased pathway-compactness

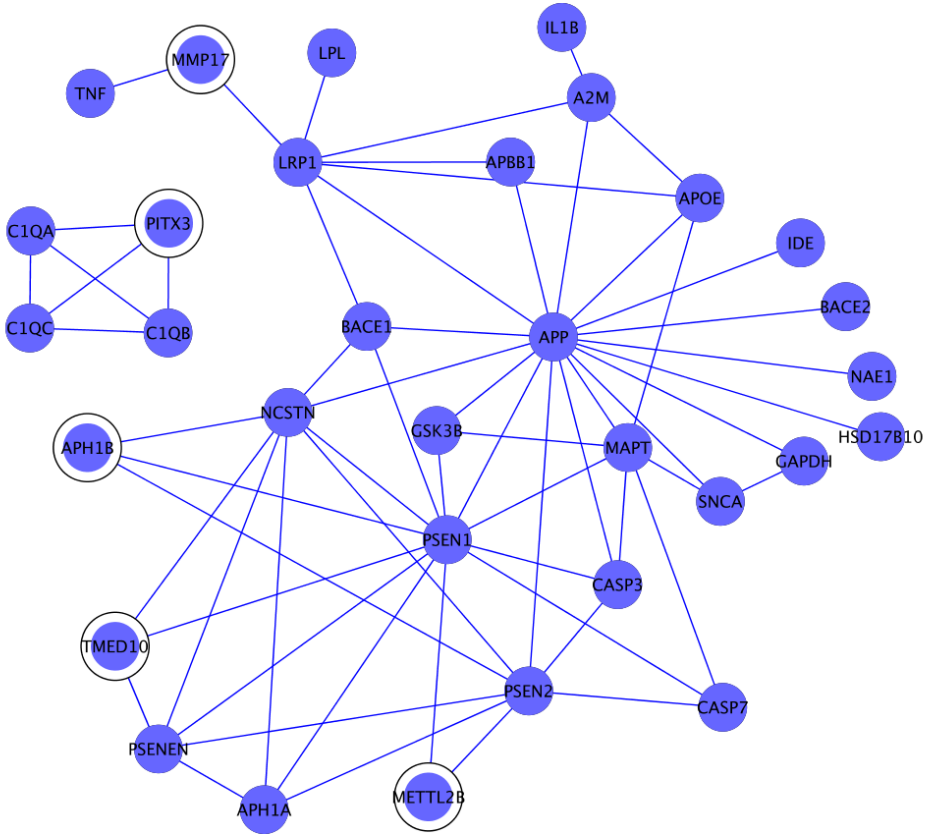
original pathway

added genes

Biological applications (1): Alzheimer‘s disease

- More than 20 proteins annotated in our molecular network
- 5 proteins added by the extension process (circled)
- 3 known to be associated with the disease
- 2 novel candidates: METTL2B, TMED10*

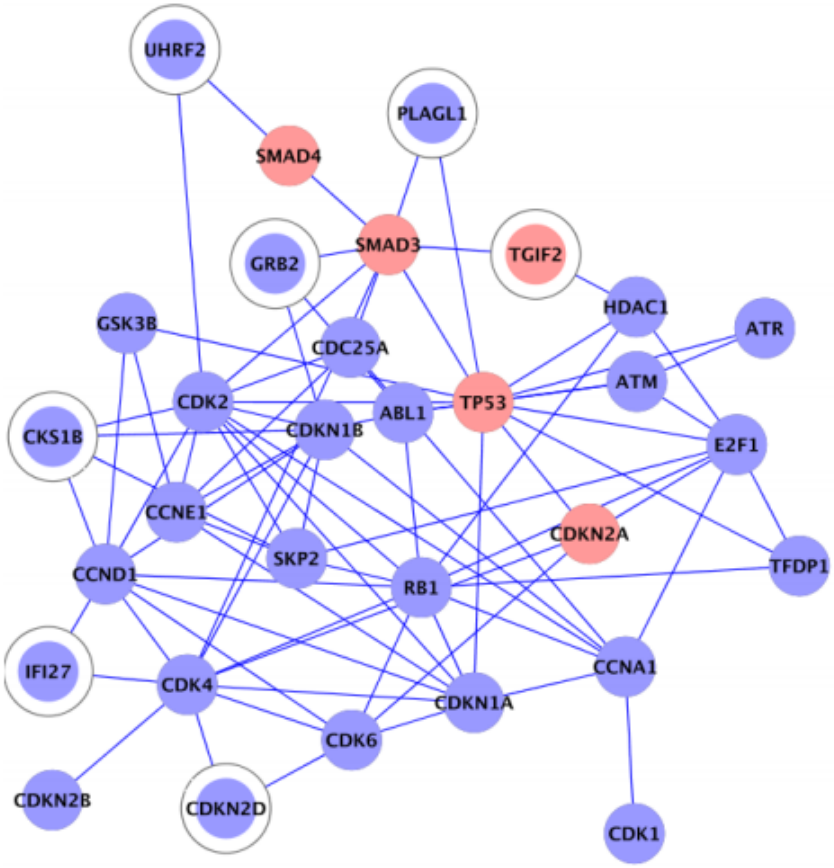
(*later confirmed: Shin et al., Autophagy, 2018)



KEGG Alzheimer disease pathway mapped on human protein interaction network

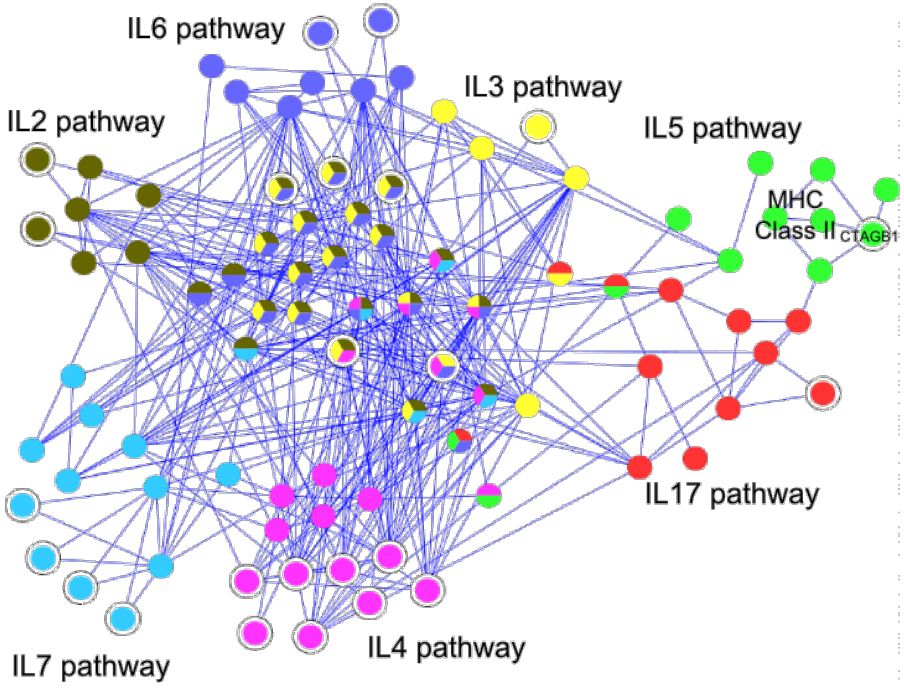
Biological applications (2): Pancreatic cancer

- "Cell cycle G1/S check point process" - extension procedure adds 7 proteins
- 6 of the added proteins are involved in cell cycle regulation
- the 7th (TGIF2) is known to be mutated in pancreatic cancer
- points to functional role of added proteins



Biological applications (3): Interleukin signaling

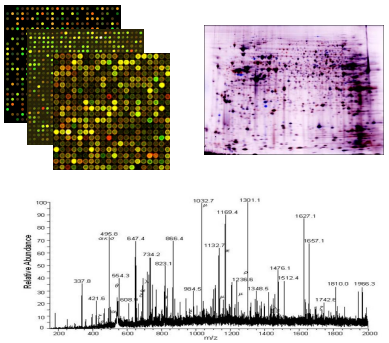
- Complex system of intracellular signaling cascades
- New putative pathway regulators identified
- New “cross-talk proteins” identified (associated with multiple pathways)

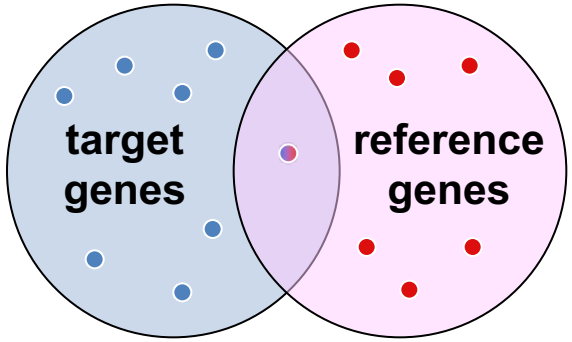


Two functions: pathway-regulation & pathway-communication?

Scoring of omics/pathway associations (EnrichNet)

Experimentally derived genes (**target genes**)





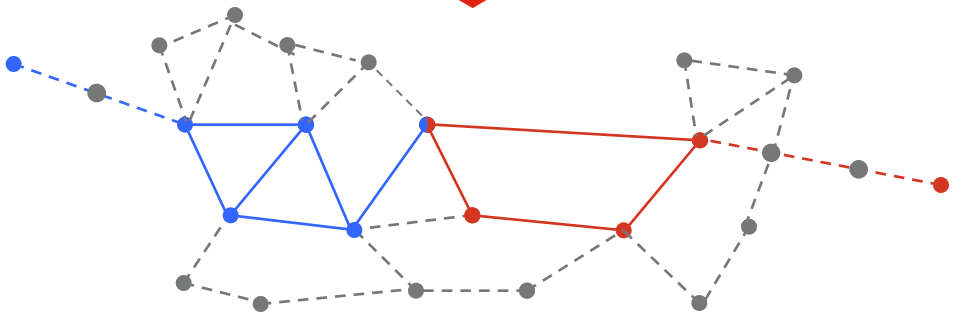
Pathway genes (**reference genes**)

KEGG Kyoto Encyclopedia of Genes and Genomes

WIKIPATHWAYS

GO the Gene Ontology

Reactome



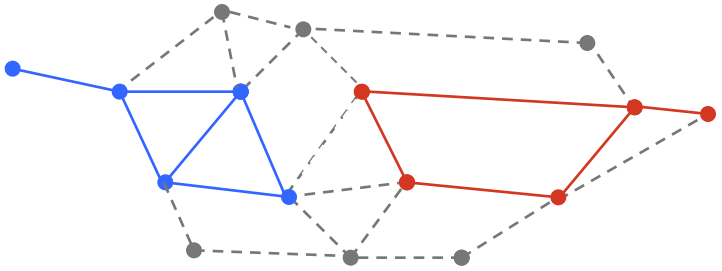
target node reference node other nodes

Network association scoring (EnrichNet)

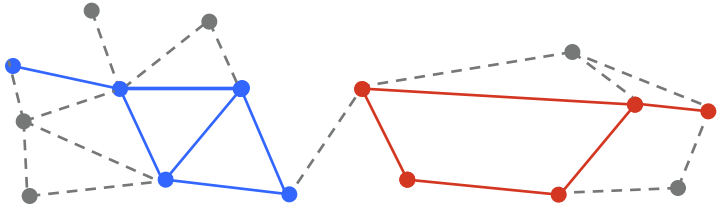
Scoring criteria:

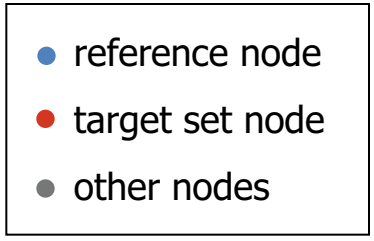
- distances between target and reference genes in network
- multiplicity of interactions between target and reference genes
- density of interactions between target and reference genes (compared to rest of the network)

Example 1: dense inter-connections



Example 2: sparse inter-connections





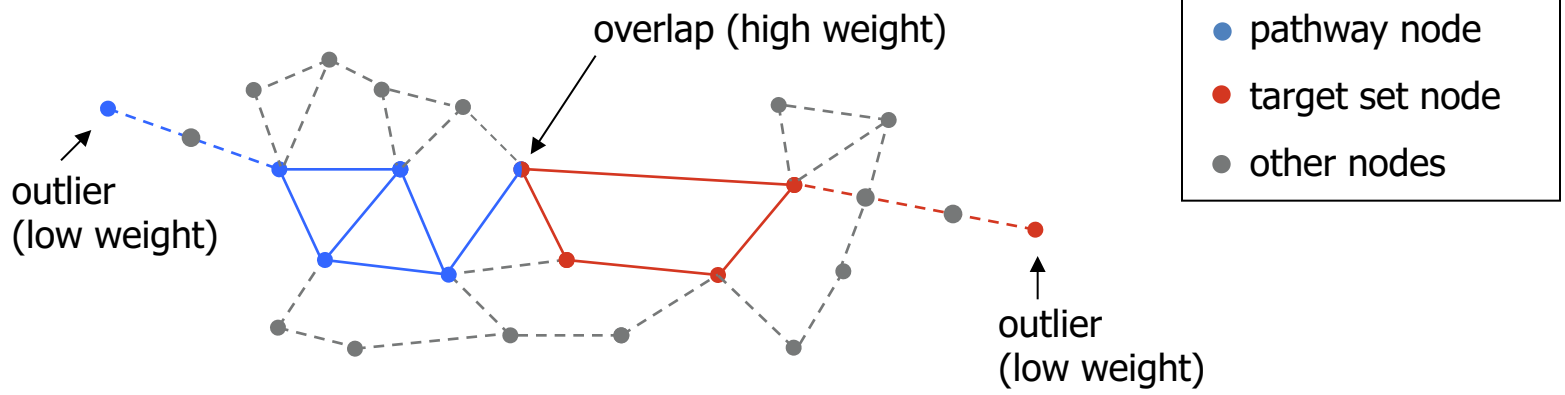
Network association scoring (EnrichNet)

Handling of overlapping genes and long distance outliers:

→ overlapping nodes and small distance node pairs → heigher weight

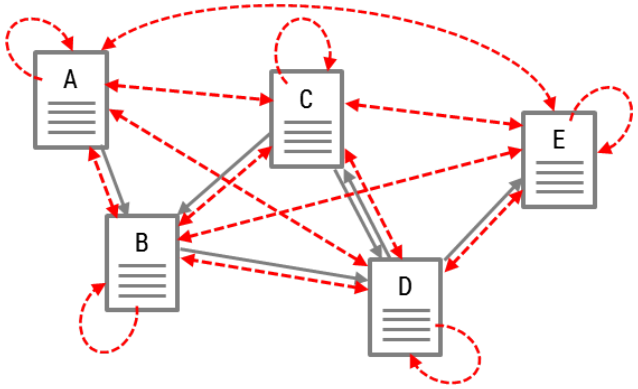
→ outlier nodes / large distance node pairs → lower weight

Example:

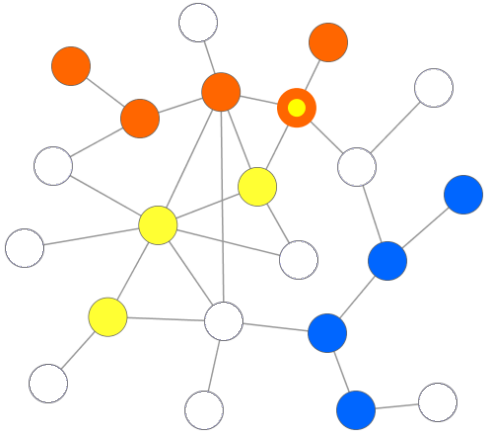


Network association scoring (EnrichNet)

Algorithm: Google‘s “Personalized Page Rank“



Transfer approach to molecular networks



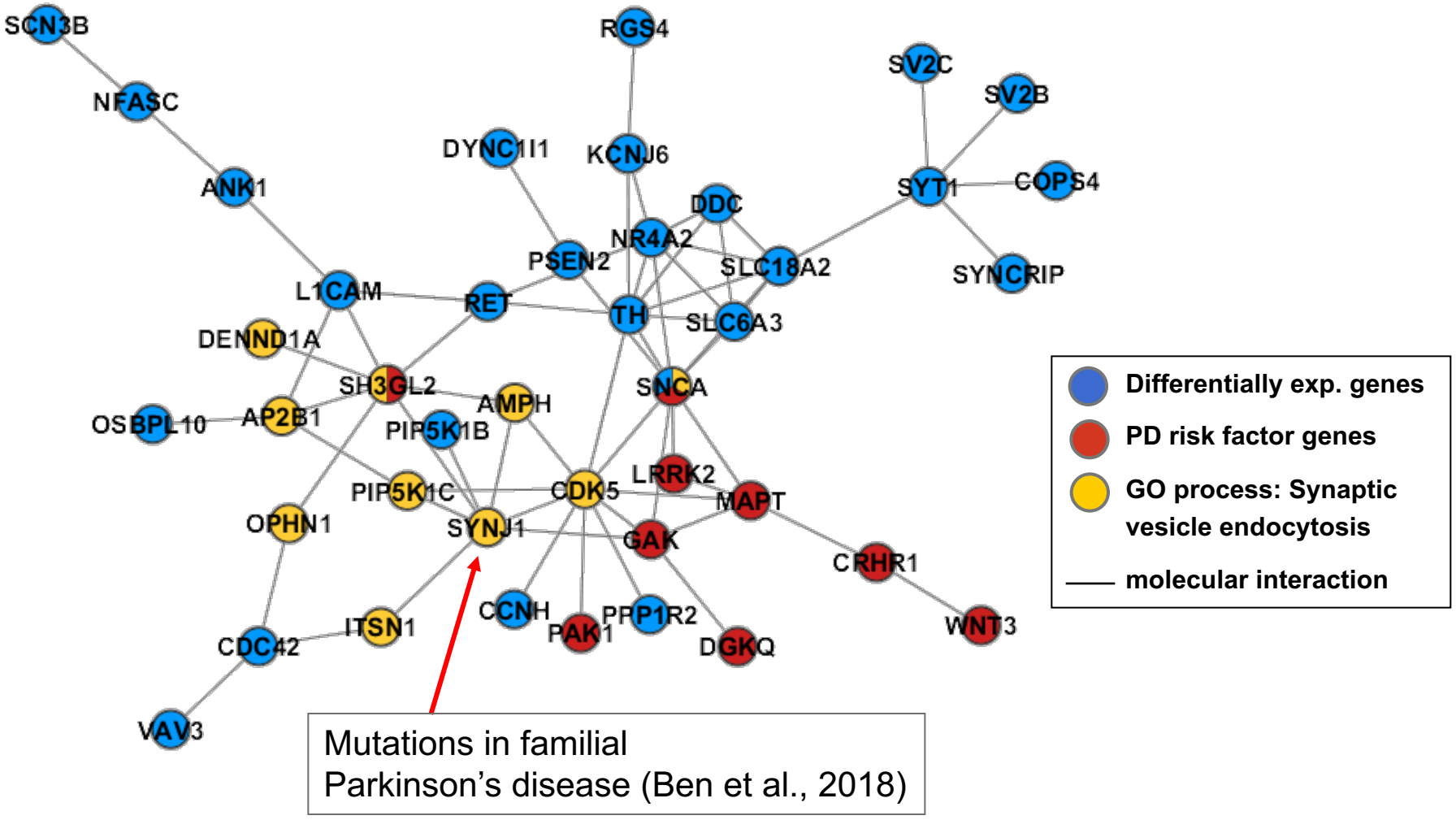
web-page
Real hyperlink
virtuel hyperlink

target genes
target/pathway overlap
pathway 1
pathway 2
interaction

Output: Relevance scores for each web-page (in relation to other web-pages)

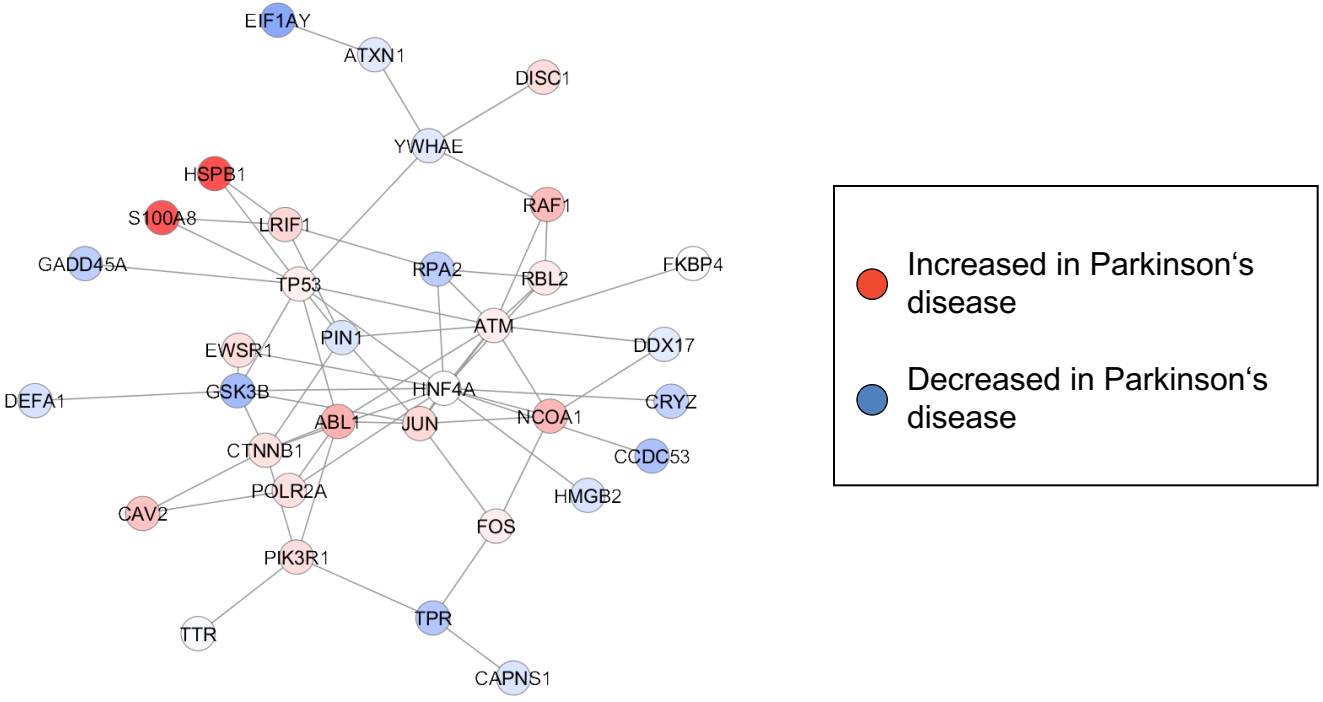
Output: Relevance scores for each pathway (in relation to a target set of genes)

Example Result: Parkinson‘s disease



Pathway-independent network analysis

Motivation: Disease perturbations may cluster in network regions outside of known pathways. Finding these clusters may lead to more robust biomarker models.



Question: How can we find clustered gene/protein groups efficiently, accounting for their diagnostic predictivity and connectedness in the network?

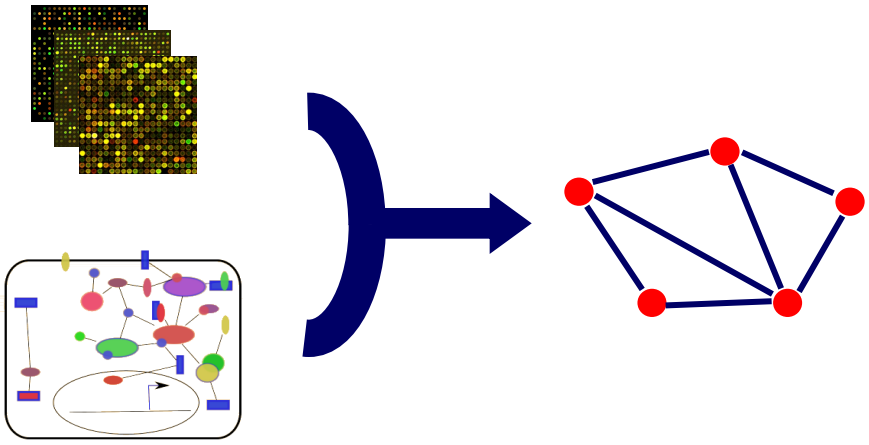
Network analysis software (GenePEN)

Input:

- Omics dataset (table with rows = genes/biomolecules, columns = samples)
- Class labels (e.g. “patient” or “control”)
- Table of interactions between the biomolecules (e.g. protein-protein interactions)

Output:

- A subset of discriminative biomolecules (rows) representing a connected component in the network that provides a predictive signature to classify new samples



Network analysis approach (GenePEN)

Idea: Find genes maximizing two quantities:

- the diagnostic prediction accuracy of their omics biomarker signature
- the connectedness of the selected genes in the network

→ formulate a corresponding scoring function (details not shown):

$$\min_{w} loss(w) + \lambda \cdot penalty(w)$$

loss-function (minimize error) trade-off parameter penalty-function (network grouping)

→ Minimize the function to find a good gene selection

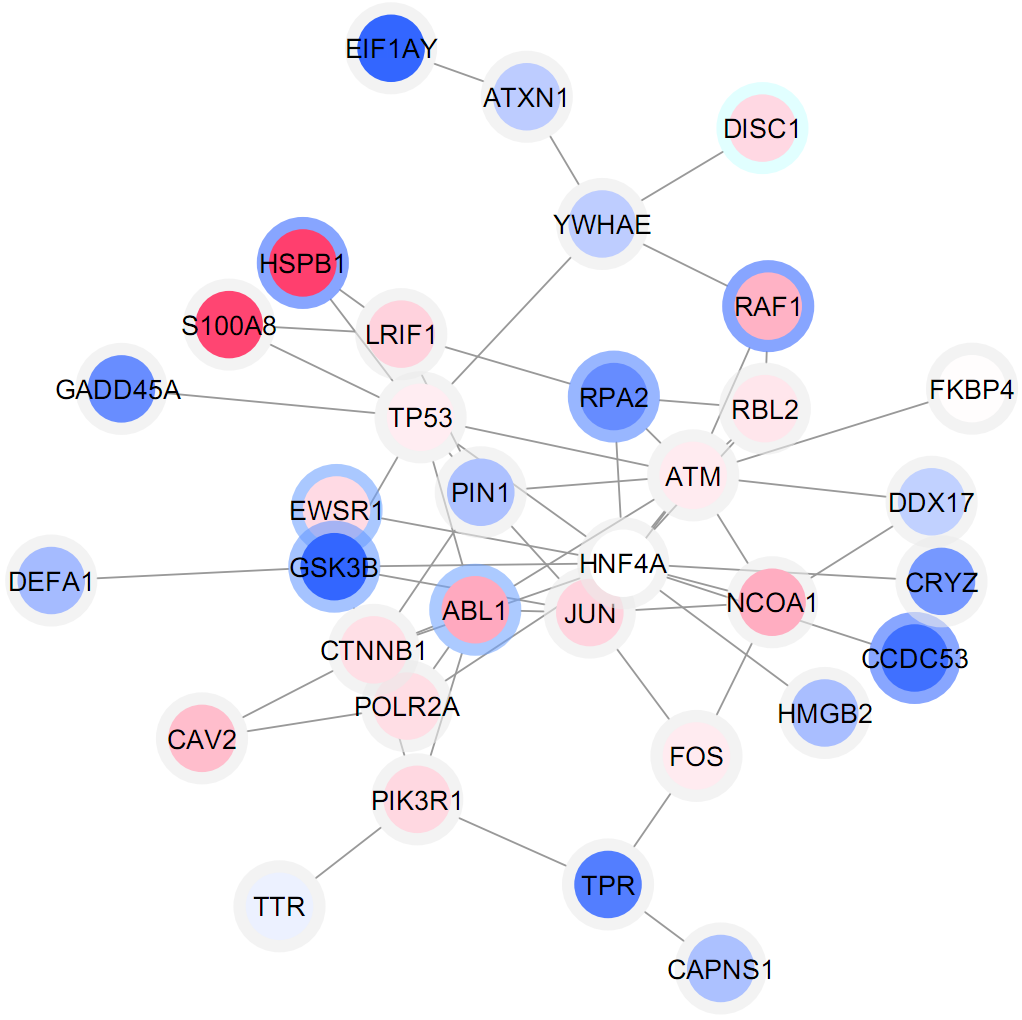
Application to Parkinson‘s disease (GenePEN)

Network alteration in Parkinson’s disease:

- red = over-expressed in PD
  blue = under-expressed in PD
  node borders = significance of alteration (from gray to blue with increasing significance)

- significant genes are over-represented in the sub-network (p = 0.01)

- GSK3B, the top significant gene in the sub-network, contains polymorphisms associated with Parkinson’s disease



Conclusion & Summary

- Why study diseases using network analysis?
  → to identify common mechanisms and combinatorial changes

- Three approaches presented:
  1) Automated network extension of disease pathways
  2) Scoring disease/pathway associations using network information
  3) Pathway-independent network analysis using machine learning

- **Future**: Time series data analysis of causal network perturbation

References

1. E. Glaab, A. Baudot, N. Krasnogor, A. Valencia. Extending pathways and processes using molecular interaction networks to analyse cancer genome data, BMC Bioinformatics, 11(1):597, 2010
2. E. Glaab, A. Baudot, N. Krasnogor, R. Schneider, A. Valencia. EnrichNet: network-based gene set enrichment analysis, Bioinformatics, 28(18):i451-i457, 2012
3. N. Vlassis, E. Glaab, *GenePEN: analysis of network activity alterations in complex diseases via the pairwise elastic net*, Statistical Applications in Genetics and Molecular Biology (2015), 14(2), 221
4. E. Glaab, J.P. Trezzi, A. Greuel, C. Jäger, Z. Hodak, A. Drzezga, L. Timmermann, M. Tittgemeyer, N. J. Diederich, C. Eggers, Integrative analysis of blood metabolomics and PET brain neuroimaging data for Parkinson's disease, Neurobiology of Disease (2019), Vol. 124, No. 1, pp. 555
5. E. Glaab, *Using prior knowledge from cellular pathways and molecular networks for diagnostic specimen classification*, Briefings in Bioinformatics (2015), 17(3), pp. 440
6. E. Glaab, R. Schneider, *Comparative pathway and network analysis of brain transcriptome changes during adult aging and in Parkinson's disease*, Neurobiology of Disease (2015), 74, 1-13
7. Z. Zhang, P. P. Jung, V. Grouès, P. May, C. Linster, E. Glaab, *Web-based QTL linkage analysis and bulk segregant analysis of yeast sequencing data*, GigaScience (2019), 8(6), 1-18
8. S. Köglsberger, M. L. Cordero-Maldonado, P. Antony, J. I. Forster, P. Garcia, M. Buttini, A. Crawford, E. Glaab, *Gender-specific expression of ubiquitin-specific peptidase 9 modulates tau expression and phosphorylation: possible implications for tauopathies*, Molecular Neurobiology (2017), 54(10), pp. 7979
9. L. Grandbarbe, S. Gabel, E. Koncina, G. Dorban, T. Heurtaux, C. Birck, E. Glaab, A. Michelucci, P. Heuschling, *Inflammation promotes a conversion of astrocytes into neural progenitor cells via NF-kB activation*, Molecular Neurobiology (2016), Vol. 53, No. 8, 5041-5055
10. S. Kleiderman, J. Sá, A. Teixeira, C. Brito, S. Gutbier, L. Evje, M. Hadera, E. Glaab, M. Henry, S. Agapios, P. Alves, U. Sonnewald, M. Leist, *Functional and phenotypic differences of pure populations of stem cell-derived astrocytes and neuronal precursor cells*, Glia (2016), Vol. 64, No. 5, 695-715
11. E. Glaab, R. Schneider, *RepExplore: Addressing technical replicate variance in proteomics and metabolomics data analysis*, Bioinformatics (2015), 31(13), pp. 2235
12. E. Glaab, *Building a virtual ligand screening pipeline using free software: a survey*, Briefings in Bioinformatics (2015), 17(2), pp. 352
13. E. Glaab, R. Schneider, *PathVar: analysis of gene and protein expression variance in cellular pathways using microarray data*, Bioinformatics, 28(3):446-447, 2012
14. E. Glaab, J. Bacardit, J. M. Garibaldi, N. Krasnogor, *Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data*, PLoS ONE, 7(7):e39932, 2012
15. E. Glaab, A. Baudot, N. Krasnogor, A. Valencia. *TopoGSA: network topological gene set analysis*, Bioinformatics, 26(9):1271-1272, 2010
16. E. Glaab, J. M. Garibaldi and N. Krasnogor. *ArrayMining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization*, BMC Bioinformatics,10:358, 2009
17. E. Glaab, J. M. Garibaldi, N. Krasnogor. *Learning pathway-based decision rules to classify microarray cancer samples*, German Conference on Bioinformatics 2010, Lecture Notes in Informatics (LNI), 173, 123-134
18. E. Glaab, J. M. Garibaldi and N. Krasnogor. *VRMLGen: An R-package for 3D Data Visualization on the Web*, Journal of Statistical Software, 36(8),1-18,2010