Computational analysis of molecular network perturbations in complex diseases

Speaker: Enrico Glaab
Luxembourg Centre for Systems Biomedicine (LCSB)
**Main goal:** Interpret molecular changes in complex disorders by integrating diverse data types using **network analyses**
Motivation: Diseases as network perturbations

APC mutations $\rightarrow$ Colorectal cancer

AXIN1/2 mutations $\rightarrow$ Colorectal cancer

β-Catenin mutations $\rightarrow$ Colorectal cancer

Wnt/β-catenin signaling pathway

( $\rightarrow$ affected by disease-related mutations)
Motivation: Diseases as network perturbations

- APC mutations $\rightarrow$ Colorectal cancer
- AXIN1/2 mutations $\rightarrow$ Colorectal cancer
- Catenin mutations $\rightarrow$ Colorectal cancer

Multiple mutations, but:
- one cellular network
- one mechanism
- one disease
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
**Systems biology:** The study of an organism, viewed as an integrated and interacting network of genes, proteins and biochemical reactions.

Source: Oltvai & Barabasi, 2002
Two main driving forces:

- New high-throughput experimental profiling approaches enable organism-wide data collection:
  - Genomics $\rightarrow$ whole-genome sequencing
  - Transcriptomics $\rightarrow$ DNA chips / RNAseq
  - Interactomics $\rightarrow$ 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?
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

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**

- Mitochondrion
  - Gene.1
  - Gene.2
  - Gene.3

- P53 signaling
  - Gene.2
  - Gene.4
  - Gene.5

→ 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
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“)
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
Automated pathway extension: Example

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

→ Disconnected nodes become connected
→ increased pathway-compactness

Mutations linked to colorectal cancer

- 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)

- Target genes
- Reference genes
- Other nodes

[Diagram showing the overlap between experimentally derived genes and pathway genes, with nodes and connections highlighted for target and reference 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 interconnections

**Example 2:**
sparse interconnections

- reference node
- target set node
- other nodes
Network association scoring (EnrichNet)

Handling of overlapping genes and long distance outliers:

→ overlapping nodes and small distance node pairs → higher weight
→ outlier nodes / large distance node pairs → lower weight

Example:
Network association scoring (EnrichNet)

Algorithm: Google’s “Personalized Page Rank“

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

Transfer approach to molecular networks

Output: Relevance scores for each pathway (in relation to a target set of genes)
Example Result: Parkinson’s disease

Mutations in familial Parkinson’s disease (Ben et al., 2018)
**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):

\[
\min_w \text{loss}(w) + \lambda \cdot \text{penalty}(w)
\]

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

→ Minimize the function to find a good gene selection
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

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