Data-Driven Medicine

Nataša Pržulj, PhD, MAE
ICREA Research Professor
Barcelona Supercomputing Center
Medicine: complex world of inter-connected entities

1. Motivation

2. New Methods – Examples: mine inter-connected data

   i. **Single type of omics data**: Molecular networks → function, disease

   ii. **Multiple layers of heterogeneous data**:
       - Patient-centered data integration → Precision medicine
       - Disease re-classification
       - Gene Ontology reconstruction
       - Network alignment

3. Vision
Overview

Medicine: complex world of inter-connected entities

1. Motivation

2. New Methods – Examples: mine inter-connected data
   i. Single type of omics data: Molecular networks $\rightarrow$ function, disease
   ii. Multiple layers of heterogeneous data:
      • Patient-centered data integration $\rightarrow$ Precision medicine
      • Disease re-classification
      • Gene Ontology reconstruction
      • Network alignment

3. Vision
1. Motivation
Medicine: complex world of inter-connected entities

Technological advances →
astounding harvest of various molecular and clinical data

**Integrative methods for analyzing big data in precision medicine**

*Vladimir Gligorijević, Noël Malod-Dognin and Nataša Pržulj*
1. Motivation
Medicine: complex world of inter-connected entities

Data growth:

- Guided by empirical reductionism:
  - Striving to dissect a biological entity into its constituent parts
  - To better understand it

- However, knowing parts is not enough:
  - 1859 — Darwin\(^1\) saw biology as a “tangled bank” with all its aspects interconnected
  - 1855 — Virchow\(^2\): all diseases involve changes in normal cells

---

Data growth about a cell:

- **Hit the wall of bio-complexity**

- **Cells:**
  - are not just loosely coupled arrangements of quasi-independent molecules
  - highly intricately and precisely integrated **networks of entities** and **interactions** within the cells and with the environment
  - **Data types complement each other**
  - **Seek joint modeling and mining**

---


Time to:

- Replace the mostly reductionist molecular perspective that dominated the 20th century
- New and holistic view of the living world
- Required to explain biological and medical phenomena
- Biology’s innate complexity

1. Motivation
Medicine: complex world of inter-connected entities

Requires:

- Establishing a perspective and framework not only for one problem, but for biology and medicine in general

A foremost challenge:

- How to re-synthesize biology
- Put the elements back into their complex, dynamic environments
- Connect them all within a unified framework
- Reformulate biological paradigms within the non-linear world

1. Motivation

Medicine: complex world of inter-connected entities

Vision:

- Bridge this gap by developing a mathematically principled framework for integration of networked data
- Marry biomedical problems and data with algorithms from:
  - ML, such as NMTF
  - Mathematical non-linear optimization
  - Network science
  - Algebraic topology…
  - High-performance computing
- Propose modelling & computational advances that will link the medicine’s:
  - reductionist past with its holistic future
- Enable
  - displacement of the dominant molecular representation of biology
  - by a new, integrative paradigm that is deeper, more comprehensive and inspiring

€2M ERC Consolidator Grant for 2018-2023
Title: “Integrated Connectedness for a New Representation of Biology”
1. Motivation

Medicine: complex world of inter-connected entities

Computational challenges

- Need new tools to mine complex data systems
- Why?
  - Analysing sequences: “computationally easy” → still lacking
  - Analysing interconnected heterogeneous data: “computationally hard”
- Sophisticated methods carefully tuned to extract new knowledge from particular data
1. Motivation
Medicine: complex world of inter-connected entities

Computational challenges
Medicine: complex world of inter-connected entities

1. Motivation

2. New Methods – Examples: mine inter-connected data
   
   i. **Single layer of omics data**: Molecular networks $\rightarrow$ function, disease

   ii. **Multiple layers of heterogeneous data**:
       
       • Patient-centered data integration $\rightarrow$ Precision medicine
       • Disease re-classification
       • Gene Ontology reconstruction
       • Network alignment

3. Vision
2. Novel Methods
Mine the Medical World of Inter-Connected Entities

i. Molecular Networks

- The number of nodes
- The number of links
- Links of each node: degree
- Distribution of links (degrees)
2. Novel Methods
Mine the Medical World of Inter-Connected Entities

i. Molecular Networks

- The number of nodes
- The number of links
- Links of each node: degree
- Distribution of links (degrees)
2. Novel Methods
Mine the Medical World of Inter-Connected Entities

Graphlets “Legos of Networks”
2. Novel Methods
Mine the Medical World of Inter-Connected Entities

ERC StG: 278212 (2012-2017): "Biological Network Topology Complements Genome as a Source of Biological Information"

Graphlets
"Legos of Networks"

2. Novel Methods
Mine the Medical World of Inter-Connected Entities
2. Novel Methods
Mine the Medical World of Inter-Connected Entities

90% similar wiring – significantly enriched:

→ Biological function
→ Protein complexes
→ Sub-cellular localization
→ Tissue expression
→ Disease

2. Novel Methods
Mine the Medical World of Inter-Connected Entities

Cancer research:
→ New proteins for melanin production
→ Same cancer type: more similar wiring
→ Far away in the network

2. Novel Methods
Mine the Medical World of Inter-Connected Entities

2. Novel Methods
Mine the Medical World of Inter-Connected Entities

Network Alignment

GrAAL:
267 nodes
900 edges

Isorank:
116 nodes
261 edges

GraAL:
267 nodes
900 edges

MI-GraAL:
1,858 nodes
3,467 edges

L-GraAL:
5,726 nodes
16,084 edges

Yeast:
98% proteins
21% interactions

T. Milenkovic, W.L. Wong, W. Hayes, & N. Pržulj, *Cancer Informatics*, 9:121-37, June 30, 2010 (Highly visible)
2. Novel Methods
Mine the Medical World of Inter-Connected Entities

Alignment of PPI Networks — **Ulign**

- Many methods
- All heuristic
- No gold standard
- **Questions:**
  - Which aligner for which data?
  - Which scoring scheme for evaluation?
  - Coverage: biological and topological?
  - Contribution of topology vs sequence?

- Map biologically and topologically **different** network regions
- Each covers only about 50% of the proteins of the larger network
- **Together** — map entire networks → **Ulign**
  - Biologically coherent

- The most topologically coherent — using topology only
- The most biologically coherent — using sequence only

→ **Combine** topology and sequence information

Why?
- Existing annotations ill-suited?
- Methodological limitations?

2. Novel Methods
Mine the Medical World of Inter-Connected Entities

✓ The best performing
✓ Robust
✓ …

☑ PPI networks are geometric


…

➢ Directed Networks
➢ Track dynamics

2. Novel Methods
Mine the Medical World of Inter-Connected Entities

Graphlets
“Legos of Networks”

Network analytics in the age of Big Data
How can we holistically mine big data?

By Nataša Pržulj and Noël Malod-Dognin

We live in a complex world of interconnected entities. In all areas of human endeavor, from biology to medicine, economics, and climate science, we are flooded with large-scale data sets. They describe intricate real-world systems from different and complementary viewpoints, with entities being modeled as nodes and their connections as edges, comprising large networks. This is

Network structures
The four networks shown have exactly the same size (the same number of nodes and edges), and each node within each network has the same degree (the number of interactions with other nodes), but each network can be of very different structure.

Four triangles

Three squares

into RNAs and translated into proteins, which adopt various three-dimensional structures to carry out particular cellular functions. Molecular interactions are captured by different high-throughput biotechnologies and modeled with different types of networks. Individual analyses of molecular networks have revealed that molecules involved in similar functions tend to group together in a network and are similarly wired (12), leading to better understanding of gene functions (6) and molecular organization of the cell (7) and to im-
2. Novel Methods
Mine the Medical World of Inter-Connected Entities

A global genetic interaction network maps a wiring diagram of cellular function

Network analytics in the age of Big Data
How can we holistically mine big data?

By Nataša Pržulj and Noël Malod-Dognin

Network structures
The four networks shown have exactly the same size (the same number of nodes and edges), and each node within each network has the same degree (the number of interactions with other nodes), but each network can be of very different structure.

Four triangles
Three squares

into RNAs and translated into proteins, and various three-dimensional molecular interactions are carried by high-throughput biotechnologies and modeled with different types of networks. Individual analyses of molecular networks have revealed that molecules involved in similar functions tend to group together in a network and are similarly wired (12), leading to better understanding of gene functions (6) and molecular organization of the cell (7) and to im-

8 JULY 2016 • VOL 353 ISSUE 6295
2. Novel Methods
Mine the Medical World of Inter-Connected Entities

Systematic protein–protein interaction mapping for clinically relevant human GPCRs

Kate Sokolina, Saranya Kittankom, Jamie Snider, Max Kotlyar, Pascal Maurice, Jorge Candia, Abla Benleulmi-Chaoucha, Kenjiro Tadagaki, Atsuro Oishi, Victoria Wong, Ramy H Malty, Viktor Deineko, Hiroyuki Aoki, Shahreen Amin, Zhong Yao, Xavier Morató, David Otasek, Hiroyuki Kobayashi, Javier Menendez, Daniel Auerbach, Stephane Angers, Natasa Pržulić, Michel Bouvier, Mohan Babu, Francisco Ciruela, Ralf Jockers, Igor Jurisica, & Igor Staglar

✓ “Spine” of the network
  ➢ “Dominating set” heuristic
✓ Functionally and topologically separates the cell
✓ Predict new GPCRs:
  ➢ e.g., chromosome 20 open reading frame 39 (TMEM90B)
2. Novel Methods
Mine the Medical World of Inter-Connected Entities

Graphlets
“Legos of Networks”
Overview

Medicine: complex world of inter-connected entities

1. Motivation

2. New Methods – Examples: mine inter-connected data
   i. **Single layer of omics data**: Molecular networks $\rightarrow$ function, disease
   ii. **Multiple layers of heterogeneous data:**
       - Patient-centered data integration $\rightarrow$ Precision medicine
       - Disease re-classification
       - Gene Ontology reconstruction
       - Network alignment

3. Vision
2. Novel Methods
Mine the Medical World of Inter-Connected Entities

Multi-disciplinary, data-fusion methodology

Motivation:
- Captures all systems-level
- Captures how data relate
- Mechanistic explanations

\[ \min \{ \sum_{1 \leq i \leq j \leq p} \| W_{ij} \cdot (D_{ij} - G_i S_{ij} G_j^T) \|^2 + \alpha \| S_{ij} \|^2 + \alpha_i \text{tr}(G_i^T L_i G_i) + \alpha_j \text{tr}(G_j^T L_j G_j) \} : G_i, S_{ij} \geq 0 \]

\( \alpha \| S_{ij} \|^2 \) maintain sparsity of \( S_{ij} \), \( \alpha_i \text{tr}(G_i^T L_i G_i) \) and \( \alpha_j \text{tr}(G_j^T L_j G_j) \) adding prior knowledge (penalties), \( G_i, S_{ij} \geq 0 \) is needed for cluster interpretation
Multi-disciplinary, data-fusion methodology

2. Novel Methods
Mine the Medical World of Inter-Connected Entities

Motivation:

- Captures all systems-level
- Captures how data relate
- Mechanistic explanations

\[
\min \left\{ \sum_{1 \leq i \leq j \leq p} \left( ||W_{ij} \circ (D_{ij} - G_i S_{ij} G_j^T)||^2 + \alpha ||S_{ij}||^2 + \alpha_i \operatorname{tr}(G_i^T L_i G_i) + \alpha_j \operatorname{tr}(G_j^T L_j G_j) \right) : G_i, S_{ij} \geq 0 \right\}
\]

\[\alpha ||S_{ij}||^2 \text{ maintain sparsity of } S_{ij}, \alpha_i \operatorname{tr}(G_i^T L_i G_i) \text{ and } \alpha_j \operatorname{tr}(G_j^T L_j G_j) \text{ adding prior knowledge (penalties), } G_i, S_{ij} \geq 0 \text{ is needed for cluster interpretation} \]
2. Novel Methods
Mine the Medical World of Inter-Connected Entities

Patient-Specific Data Fusion → Personalized Treatment

Co-clustering: patients, genes and drugs

Data:

- TCGA
- BioGRID, KEGG: PPI, GI, MI
- DrugBank: DTI
- DrugBank: SMILES

353 serous ovarian cancer patients from TCGA:
1. Patient stratification
2. Driver gene prediction
3. Drug repurposing
2. Novel Methods
Mine the Medical World of Inter-Connected Entities

Patient-Specific Data Fusion → Personalized Treatment

Co-clustering: patients, genes and drugs

Data:
- TCGA
- BioGRID, KEGG: PPI, GI, MI
- DrugBank: DTI
- DrugBank: SMILES

\[
\begin{align*}
R_{12} &\approx \mathbf{G}_1 \mathbf{H}_{12} \mathbf{G}_2^T \\
R_{23} &\approx \mathbf{G}_2 \mathbf{H}_{23} \mathbf{G}_3^T \\
k_1 \ll n_1 &\text{ - patient clusters} \\
k_2 \ll n_2 &\text{ - gene clusters} \\
k_3 \ll n_3 &\text{ - drug clusters} \\
\mathbf{G}_1, \mathbf{G}_2 \text{ and } \mathbf{G}_3 &\text{ are cluster indicator matrices}
\end{align*}
\]

Ovarian cancer patients:
1. Patient stratification → \( \hat{\mathbf{C}}_1 \)
2. Driver gene prediction → \( \hat{\mathbf{C}}_2 \)
3. Drug repurposing → \( \hat{\mathbf{R}}_{23} \)

\[
\min_{\mathbf{G}_i \geq 0, 1 \leq i \leq 3} \min_{\mathbf{G}_i \geq 0, 1 \leq i \leq 3} \left[ \| \mathbf{R}_{12} - \mathbf{G}_1 \mathbf{H}_{12} \mathbf{G}_2^T \|_F^2 + \| \mathbf{R}_{23} - \mathbf{G}_2 \mathbf{H}_{23} \mathbf{G}_3^T \|_F^2 + \right.
\]
\[
tr(\mathbf{G}_2^T \mathbf{L}_2 \mathbf{G}_2) + tr(\mathbf{G}_3^T \mathbf{L}_3 \mathbf{G}_3)
\]

V. Gligorijevic, N. Malod-Dognin and N. Przulj, Patient-specific data fusion for cancer stratification and personalized treatment, PSB, 2016
Some results:

Kaplan-Meier survival curves for 3 patient groups found by GNMTF (log-rank p-val = 5.3 x 10^{-3})

\[
\begin{align*}
R_{12} & \approx G_1 H_{12} G_2^T \\
R_{23} & \approx G_2 H_{23} G_3^T \\
\end{align*}
\]

\(k_1 \ll n_1\) - patient clusters
\(k_2 \ll n_2\) - gene clusters
\(k_3 \ll n_3\) - drug clusters

\(G_1, G_2\) and \(G_3\) are cluster indicator matrices

Ovarian cancer patients:

1. Patient stratification \(\hat{C}_1\)
2. Driver gene prediction \(\hat{C}_2\)
3. Drug repurposing \(\hat{R}_{23}\)

\[
\begin{align*}
J = \min_{G_i \geq 0, 1 \leq i \leq 3} \min_{G_i \geq 0, 1 \leq i \leq 3} \left[ \| R_{12} - G_1 H_{12} G_2^T \|_F^2 + \| R_{23} - G_2 H_{23} G_3^T \|_F^2 + tr(G_2^T L_2 G_2) + tr(G_3^T L_3 G_3) \right]
\end{align*}
\]

V. Gligorijevic, N. Malod-Dognin and N. Przulj, Patient-specific data fusion for cancer stratification and personalized treatment, *PSB*, 2016
Some results: ~40% of our 809 predicted driver genes in CCGD, Census, or IntOGen

<table>
<thead>
<tr>
<th>New driver</th>
<th>Known drivers</th>
<th>Score</th>
<th>DB</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADAM32</td>
<td>BMPR2</td>
<td>1.000</td>
<td>–</td>
</tr>
<tr>
<td>REG1P</td>
<td>CLASP2</td>
<td>1.000</td>
<td>–</td>
</tr>
<tr>
<td>PCDHA2</td>
<td>CHD4</td>
<td>1.000</td>
<td>–</td>
</tr>
<tr>
<td>NCR1</td>
<td>BMPR2</td>
<td>1.000</td>
<td>–</td>
</tr>
<tr>
<td>USPL1</td>
<td>CLASP2</td>
<td>1.000</td>
<td>–</td>
</tr>
<tr>
<td>GDPD3</td>
<td>DDX5</td>
<td>1.000</td>
<td>–</td>
</tr>
<tr>
<td>LECT1</td>
<td>CLASP2</td>
<td>1.000</td>
<td>CCGD</td>
</tr>
<tr>
<td>IL25</td>
<td>CDK12, CCAR1</td>
<td>0.975</td>
<td>–</td>
</tr>
<tr>
<td>BAK1</td>
<td>ATRX, TFDP1, NDRG1</td>
<td>0.967</td>
<td>–</td>
</tr>
<tr>
<td>MOGAT2</td>
<td>ATRX, TFDP1, NDRG1</td>
<td>0.967</td>
<td>–</td>
</tr>
<tr>
<td>CHAF1A</td>
<td>ATRX, TFDP1, NDRG1</td>
<td>0.967</td>
<td>CCGD</td>
</tr>
<tr>
<td>PITX2</td>
<td>ATRX, TFDP1, NDRG1</td>
<td>0.967</td>
<td>–</td>
</tr>
<tr>
<td>SIN3B</td>
<td>ATRX, TFDP1, NDRG1</td>
<td>0.967</td>
<td>–</td>
</tr>
<tr>
<td>RPL30</td>
<td>ATRX, TFDP1, NDRG1</td>
<td>0.967</td>
<td>–</td>
</tr>
<tr>
<td>GRWD1</td>
<td>ATRX, TFDP1, NDRG1</td>
<td>0.967</td>
<td>–</td>
</tr>
<tr>
<td>SNAI1</td>
<td>ATRX, TFDP1, NDRG1</td>
<td>0.967</td>
<td>CCGD</td>
</tr>
<tr>
<td>RBMXP4</td>
<td>ATRX, TFDP1, NDRG1</td>
<td>0.967</td>
<td>–</td>
</tr>
<tr>
<td>CPNE7</td>
<td>ATRX, TFDP1, NDRG1</td>
<td>0.967</td>
<td>–</td>
</tr>
<tr>
<td>HIPK3</td>
<td>ATRX, TFDP1, NDRG1</td>
<td>0.967</td>
<td>CCGD</td>
</tr>
<tr>
<td>EPOR</td>
<td>ATRX, TFDP1, NDRG1</td>
<td>0.967</td>
<td>CCGD</td>
</tr>
</tbody>
</table>

\[
\min_{G_i \geq 0, 1 \leq i \leq 3} \min_{G_i \geq 0, 1 \leq i \leq 3} \left[ \| R_{12} - G_1 H_{12} G_2^T \|_F^2 + \| R_{23} - G_2 H_{23} G_3^T \|_F^2 + \right. \\
\left. tr(G_2^T L_2 G_2) + tr(G_3^T L_3 G_3) \right]
\]

Ovarian cancer patients:
1. Patient stratification → \( \hat{C}_1 \)
2. Driver gene prediction → \( \hat{C}_2 \)
3. Drug repurposing → \( \hat{R}_{23} \)

V. Gligorijevic, N. Malod-Dognin and N. Przulj, Patient-specific data fusion for cancer stratification and personalized treatment, PSB, 2016
Some results: 5-fold cross validation, average AUC: ROC and PR

Ovarian cancer patients:
1. Patient stratification → $\hat{C}_1$
2. Driver gene prediction → $\hat{C}_2$
3. Drug repurposing → $\hat{R}_{23}$

V. Gligorijevic, N. Malod-Dognin and N. Przulj, Patient-specific data fusion for cancer stratification and personalized treatment, PSB, 2016
Some results: 37% of our ~225K predicted DTIs confirmed in MATADOR or CTD

<table>
<thead>
<tr>
<th>Gene</th>
<th>Drug</th>
<th>Score</th>
<th>Clusters</th>
<th>DB</th>
</tr>
</thead>
<tbody>
<tr>
<td>KIT</td>
<td>ATP</td>
<td>0.873</td>
<td>1, 2, 3</td>
<td></td>
</tr>
<tr>
<td>GABRQ</td>
<td>Adinazolam</td>
<td>0.808</td>
<td>1 M</td>
<td></td>
</tr>
<tr>
<td>GABRQ</td>
<td>Fludiazepam</td>
<td>0.808</td>
<td>1 M</td>
<td></td>
</tr>
<tr>
<td>GABRQ</td>
<td>Cinolazepam</td>
<td>0.809</td>
<td>1 M</td>
<td></td>
</tr>
<tr>
<td>GABRQ</td>
<td>Cloxiazepam</td>
<td>0.809</td>
<td>1 M</td>
<td></td>
</tr>
<tr>
<td>HTR2A</td>
<td>Dopamine</td>
<td>0.809</td>
<td>1, 3 C, M</td>
<td></td>
</tr>
<tr>
<td>GRIN3A</td>
<td>Pethidine</td>
<td>0.801</td>
<td>1, 2</td>
<td></td>
</tr>
<tr>
<td>CACNA2D1</td>
<td>Verapamil</td>
<td>0.761</td>
<td>1, 3 M</td>
<td></td>
</tr>
<tr>
<td>PDGFRB</td>
<td>ATP</td>
<td>0.724</td>
<td>1, 2</td>
<td></td>
</tr>
<tr>
<td>KDR</td>
<td>ATP</td>
<td>0.724</td>
<td>1, 3 C</td>
<td></td>
</tr>
<tr>
<td>HTR1A</td>
<td>Mirtazapine</td>
<td>0.720</td>
<td>1, 2 C, M</td>
<td></td>
</tr>
<tr>
<td>GABRA6</td>
<td>Adinazolam</td>
<td>0.688</td>
<td>1 M</td>
<td></td>
</tr>
<tr>
<td>GABRA6</td>
<td>Fludiazepam</td>
<td>0.688</td>
<td>1 M</td>
<td></td>
</tr>
<tr>
<td>GABRA6</td>
<td>Cinolazepam</td>
<td>0.688</td>
<td>1 M</td>
<td></td>
</tr>
<tr>
<td>GABRA6</td>
<td>Cloxiazepam</td>
<td>0.688</td>
<td>1 M</td>
<td></td>
</tr>
<tr>
<td>GABRA4</td>
<td>Adinazolam</td>
<td>0.687</td>
<td>1, 3 M</td>
<td></td>
</tr>
<tr>
<td>GABRA4</td>
<td>Fludiazepam</td>
<td>0.687</td>
<td>1, 3 M</td>
<td></td>
</tr>
<tr>
<td>GABRA4</td>
<td>Cinolazepam</td>
<td>0.687</td>
<td>1, 3 M</td>
<td></td>
</tr>
<tr>
<td>GABRA4</td>
<td>Cloxiazepam</td>
<td>0.687</td>
<td>1, 3 M</td>
<td></td>
</tr>
<tr>
<td>CACNA1D</td>
<td>Magnesium Sulfate</td>
<td>0.676</td>
<td>1, 2, 3 M</td>
<td></td>
</tr>
</tbody>
</table>

Ovarian cancer patients:
1. Patient stratification $\rightarrow \hat{C}_1$
2. Driver gene prediction $\rightarrow \hat{C}_2$
3. Drug repurposing $\rightarrow \hat{R}_{23}$

V. Gligorijevic, N. Malod-Dognin and N. Przulj, Patient-specific data fusion for cancer stratification and personalized treatment, *PSB*, 2016
2. Novel Methods
Mine the Medical World of Inter-Connected Entities

Patient-Specific Data Fusion → Personalized Treatment

- DNA elements
  - SMPs, Expressions, Methylation, Copy numbers
  - PPIs, GIs, eQTLs

- Clinical profile similarities
  - Patients
  - Targets

- Chemical similarities
  - Drugs
  - Treatments, Adverse effects
2. Novel Methods
Mine the Medical World of Inter-Connected Entities

Patient-Specific Data Fusion → Personalized Treatment

- DNA elements
- Germline mutations, Expressions, Methylation
- Targets
- PPIs, GIs, eQTLs
- Clinical profile similarities
- Treatments, Adverse effects
- Vaccine compounds
- Chemical similarities

- Systems vaccinology
  - With Dr. Nuria Izquierdo, IGTP IrsiCaixa, Badalona
  - Scientific Advisory Board of the Helmholtz Centre for Infection Research (HZI / Braunschweig, Germany)
2. Novel Methods
Mine the Medical World of Inter-Connected Entities

Patient-Specific Data Fusion → Personalized Treatment

Obstacles:
1. Different NP-hard continuous optimization problem:
   • propose objective function,
   • optimization solver — prove convergence and correctness
2. Optimization is slow → HPC
Disease Classification from Systems-Level Molecular Data

Method

Some Results:

→ 14 disease-disease associations currently not present in DO:
  - evidence for their relationships through comorbidity data and literature curation

→ GI the most important predictor of a link between diseases, despite small

→ Omission of any one of the included data sources reduces prediction quality
  - Importance of systems-level data fusion

→ DO ∩ disease class → 80% DO from only network data

4 Objects: Genes, GO terms, DO terms, Drugs
Constraints: $\Theta_i$ (network topology, ontology relations)
Relation matrices: $R_{ij}$

2. Novel Methods
Mine the Medical World of Inter-Connected Entities

Disease Classification from Systems-Level Molecular Data

- Co-clustering GO terms, DO terms, Genes and Drugs under pairwise constraints:

  
  Minimizing Frobenious distance between $R_{ij}$ and $G_iS_{ij}G_j^T$, for all relation matrices:

  
  - $i = \{\text{Genes}\}$, $j = \{\text{DO terms, GO terms, Drugs}\}$
  - $G_i$ is a cluster indicator matrix for data type $i$ (genes, DO terms, GO terms and Drugs)

  
  with additional penalty terms:

  
  $\min J = \min G \geq 0 \left[ ||R - GS G^T||^2_F + \sum_{i=1}^5 tr(G^T\Theta^{(i)}G) \right]$

  
  - Interested in $G_2$ (DO terms)
    - used for cluster assignment and inferring new disease associations from clusters

Optimization problem which minimizes \( \| R_{12} - G_1 S_{12} G_2^T \|_F^2 \)
under the guidance of pairwise constraints
(connectivity and GDV similarity) between genes in networks:

\[
\min_{G_1 \geq 0, G_2 \geq 0} J = \min_{G_1 \geq 0, G_2 \geq 0} \left( \| R_{12} - G_1 S_{12} G_2^T \|_F^2 + \sum_{i=1}^{4} tr(G_1^T L_1^{(i)} G_1) + \sum_{i=1}^{4} tr(G_1^T \Lambda_1^{(i)} G_1) + tr(G_2^T L_2 G_2) \right)
\]

→ using topology of molecular networks as constraints (penalty terms) in this optimization problem:
→ \( L^{(i)}_1 \) is Laplacian of adjacency matrix of a molecular network \( i=1,2,3,4 \):
\( L^{(i)}_1 = D^i - A^i \), where \( D^i \) is diagonal matrix of degrees (row summation of \( A^i \)), \( A^i \) is adjacency matrix

→ \( \Lambda^{(i)}_1 \) are Laplacians of GDV similarity matrices over all genes for each molecular network \( i \):
\( \Lambda^{(i)}_1 = D^i - \sigma^{(i)} \), where \( D^i \) is diagonal matrix of row summation of \( \sigma^{(i)} \), \( \sigma^{(i)} \) is binary GDV similarity matrix (containing only significantly similar gene/protein pairs)

→ \( L_2 \) is Laplacian of Gene Ontology graph

Outperform Dutkowski et al. [2013]
96% of GO reconstructed!
Correct assignment of GO terms to genes (3-fold cross-validation, AUC=0.874 ± 0.002)
Graphlets improve results
Validated biologically by Bonne’s yeast Genetic Interaction profile data, Science, 2016

V. Gligorijevic, V. Janjic and N. Przulj, Integration of molecular network data reconstructs GO, Bioinformatics, Vol. 30 ECCB 2014, i594-i600 (14% acceptance rate), 2014
2. Novel Methods
Mine the Medical World of Inter-Connected Entities

Multiple Network Alignment: **Fuse**

We use a block-based representation of relation (R) and Laplacian (L) matrices and matrix factors (S and G) for our 5 PPI networks as follows:

\[
R = \begin{bmatrix}
0 & R_{12} & \ldots & R_{15} \\
R_{12} & 0 & \ldots & R_{25} \\
\vdots & \vdots & \ddots & \vdots \\
R_{15} & R_{25} & \ldots & 0
\end{bmatrix},
L = \begin{bmatrix}
L_1 & 0 & \ldots & 0 \\
0 & L_2 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & L_5
\end{bmatrix};
\]

\[
S = \begin{bmatrix}
0 & S_{12} & \ldots & S_{15} \\
S_{12}^T & 0 & \ldots & S_{25} \\
\vdots & \vdots & \ddots & \vdots \\
S_{15}^T & S_{25}^T & \ldots & 0
\end{bmatrix},
G = \begin{bmatrix}
G_1 & 0 & \ldots & 0 \\
0 & G_2 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & G_5
\end{bmatrix};
\]

To simultaneously factorize all relation matrices, \( R_{ij} \approx G_i S_{ij} G_j^T \), \( 0 \leq i, j \leq 5 \), under the constraints of PPI networks, we minimize the following objective function:

\[
\min_{G \geq 0} J = \left\| R - G S G^T \right\|^2_F + \gamma Tr(G^T L G)
\]

where \( Tr \) denotes the trace of a matrix and \( \gamma \) is a regularization parameter which balances the influence of network topologies in reconstruction of the relation matrix. The second term of equation 2 is the penalization term.

---

2. Novel Methods
Mine the Medical World of Inter-Connected Entities

Multiple Network Alignment: **Fuse**

---

Fig. 2. Functional consistency of NMTF associations. For both NMTF associations and sequence similarity of protein pairs, we plot the cumulative number of protein pairs with both proteins annotated (x-axis) against the percentages of them sharing GO terms (y-axis). Biological process (BP) and molecular function (MF) annotations are considered separately.

Medicine: complex world of inter-connected entities

1. Motivation

2. New Methods – Examples: mine inter-connected data

   i. **Single layer of omics data**: Molecular networks $\rightarrow$ function, disease

   ii. **Multiple layers of heterogeneous data**:  
       • Patient-centered data integration $\rightarrow$ Precision medicine  
       • Disease re-classification  
       • Gene Ontology reconstruction  
       • Network alignment

3. Vision
3. Vision

Biomedical Data: complex system of heterogeneous interacting entities

- Large
- Heterogeneous
- Highly dimensional
- Growing Complexity
- Noisy
- Dynamic
- Different time and space scales

- World Economic Forum in Davos 2016:
  - “Big data” potential to transform medicine
  - Make it more effective due to increased life expectancy and exposure to environmental risks

- Nature Insight and Outlook of 2015 and 2016

- I was awarded 2014 BCS Roger Needham Award in recognition of “the potential my work and research have to revolutionize health and pharmaceutics”
3. Vision

Biomedical Data: complex system of heterogeneous interacting entities

- Large
- Heterogeneous
- Highly dimensional
- Growing Complexity
- Noisy
- Dynamic
- Different time and space scales

- Each type: limited, but complementary information
- Seek principled, joint organization and mining within the same framework

- World Economic Forum in Davos 2016:
  - “Big data” potential to transform medicine
  - Make it more effective due to increased life expectancy and exposure to environmental risks

- Nature Insight and Outlook of 2015 and 2016

- I was awarded 2014 BCS Roger Needham Award in recognition of “the potential my work and research have to revolutionize health and pharmaceutics”

€2M ERC Consolidator Grant for 2018-2023
Title: “Integrated Connectedness for a New Representation of Biology”
3. Vision

Holistically Mine All Available Data

→ Paradigm shifts

1. Conceptual  
2. Methodological

3. Vision

Holistically Mine All Available Data

→ Paradigm shifts

1. Conceptual

Do not analyze single data type in isolation of others (e.g., sequence align.)

- Analyze all types of data within a single framework
- New, bottom-up, data-driven biological concepts
  - Elucidate that a cell may be governed by yet undiscovered principles of life
  - Point to ways to re-think biology and approaches to medicine

3. Vision

Holistically Mine All Available Data

→ Paradigm shifts

1. Conceptual

Do not analyze single data type in isolation of others (e.g., sequence align.)

- Introduce a concept of an “Integrated Cell (iCell)”

3. Vision

Holistically Mine All Available Data

→ Paradigm shifts

2. Methodological

3. Vision

Holistically Mine All Available Data

→ Paradigm shifts

2. Methodological

- Mathematical formalisms
  - Capture multi-scale organization
  - Dynamics, stochasticity of the data,…
  - E.g., multiplex networks, hypergraphs, simplicial complexes …

- Algorithms to compute and extract information from those formalisms

3. Vision

Holistically Mine All Available Data

→ Paradigm shifts

2. Methodological

- Mathematical formalisms
  - Capture multi-scale organization
  - Dynamics, stochasticity of the data, …

  - E.g., multiplex networks, hypergraphs, simplicial complexes …

- Algorithms to compute and extract information from those formalisms

How: e.g.
- Utilize dependencies in local network topology (orbits) — data set dependent
- Uncover latent low-dimensional structure of data
- Exploit structure for developing efficient toolsets for particular data

3. Vision

Holistically Mine All Available Data

→ Paradigm shifts

2. Methodological

• Mathematical formalisms
  • Capture multi-scale organization
  • Dynamics, stochasticity of the data, …

  • E.g., multiplex networks, hypergraphs, simplicial complexes …

• Algorithms to compute and extract information from those formalisms

Computational issues remain to be addressed, arising from intractability:
• large sizes, complexity, heterogeneity, noisiness, and
• different time and space scales of the data

“Embedded” data scientists: problem-specific heuristic methods, HPC

3. Vision

Holistically Mine All Available Data

→ Paradigm shifts

Guided by Needs of Biomedical Collaborators and Industry

E.g.:

• Cancer
• Rare genetic diseases
• Viral medicines
• JnJ
• GSK
• Medium, start-ups, …
Acknowledgements

➢ **Funding:**

European Research Council  
NSERC  
Ontario  
gsk  
IBM  
Google

BSC  
The Farr Institute of Health Informatics Research  
**the PROSTATE project**  
ARRS SLOVENIAN RESEARCH AGENCY

➢ **Group members (present and past):**

1. Dr. Noel Malod-Dogning  
2. Dr. Julia Petschnigg  
3. Dr. Chhedi Gupta  
4. Sam Windels  
5. Thomas Gaudelet  
6. Dr. Omer Yaveroglu  
7. Prof. Tijana Milenković  
8. Dr. Oleksii Kuchaiev  
9. Dr. Vesna Memišević  
10. Dr. Vladimir Gligorijevic

➢ **Collaborators:**

Robin Ketteler, Harry Hemmingway,  
Igor Stagljar, Charles Boone, ...
Acknowledgements

Funding:

- ERC Consolidator Grant:
  - Post-Doc positions
  - PhD student positions

- JnJ:
  - Post-Doc position
Thank you

Comments and Questions