Uncovering Proteins Functions Through Multi-Layer Tissue Networks

Marinka Zitnik
marinka@cs.stanford.edu
Joint work with Jure Leskovec
Why tissues?

A unified view of cellular functions across human tissues is essential for understanding biology, interpreting genetic variation, and developing therapeutic strategies

[Greene et al. 2015, Yeger & Sharan 2015, GTEx and others]
What Does My Protein Do?

Goal: Given a set of proteins and possible functions, predict each protein’s association with each function

Proteins $\times$ (Functions, Tissues) $\rightarrow$ [0,1]

$WNT1 \times \text{(Midbrain development, Substantia nigra)} \rightarrow 0.9$

$RPT6 \times \text{(Angiogenesis, Blood)} \rightarrow 0.05$

Marinka Zitnik, Stanford, ISMB/ECCB 2017
Existing Research

- Guilty by association: protein’s function is determined based on who it interacts with [Zuberi et al. 2013, Radivojac et al. 2013, Kramer et al. 2014, Yu et al. 2015] and many others
  - No tissue-specificity

- Protein functions are assumed constant across organs and tissues:
  - Functions in heart are the same as in skin

Lack of methods for predicting protein functions in different biological contexts
Challenges

- Tissues have inherently multiscale, hierarchical organization
- Tissues are related to each other:
  - Proteins in biologically similar tissues have similar functions [Greene et al. 2015, ENCODE 2016]
  - Proteins are missing in some tissues
- Interaction networks are tissue-specific
- Many tissues have no annotations
Multi-label node classification: **midbrain development**, **angiogenesis**, etc.
Machine Learning Lifecycle

- Machine learning lifecycle: This feature, that feature
- Every single time!

Raw Networks → Node and edge profiles → Learning Algorithm → Prediction Model

Downstream task: Protein function prediction

Feature engineering: Automatically learn the features
Feature Learning in Multi-Layer Graphs

OhmNet: Unsupervised feature learning for multi-layer networks

\[ u \rightarrow \mathbb{R}^d \]

Vectors, node embeddings

Layer \( i \)

Layer \( k \)

Layer \( l \)

Scale “3”

Scale “2”

Scale “1”

Marinka Zitnik, Stanford, ISMB/ECCB 2017
Features in Multi-Layer Tissue Network

- **Given**: Layers $\{G_i\}_i$, hierarchy $\mathcal{M}$
  - Layers $\{G_i\}_{i=1..T}$ are in leaves of $\mathcal{M}$
- **Goal**: Learn functions: $f_i : V_i \rightarrow \mathbb{R}^d$
- **Multi-scale model**:
  - Learn node embeddings at each possible scale
  - Layers $i, j, k, l$
  - Scales “3”, “2”, “1”
OhmNet Learning Approach

OhmNet has two components:

1. **Single-layer objectives**
   Nodes with similar network neighborhoods in each layer are embedded close together

2. **Hierarchical dependency objectives**
   Nodes in nearby network layers in the hierarchy share similar features
Single-Layer Objectives

- **Intuition:** For each layer, embed nodes to $d$ dimensions by preserving their similarity.
- Two nodes are similar if their neighborhoods are similar.
- For node $u$ in layer $i$ we define nearby nodes as nodes in $G_i$ visited by random walks starting at $u$. 

![Diagram showing single-layer objectives with nodes and random walk paths]
Dependencies Between Network Layers

- **Intuition:** Proteins in biologically similar tissues share similar features
- **Use tissue hierarchy to recursively regularize features at** $i$ **to be similar to features in** $i$’s parent

**OhmNet** generates multi-scale node embeddings

OhmNet generates multi-scale node embeddings
Data: 107 Tissue Layers

- **Layers** are PPI nets:
  - Nodes: proteins
  - Edges: tissue-specific PPIs

- **Node labels**:
  - “Cortex development” in renal cortex tissue
  - “Artery morphogenesis” in artery tissue
Experimental Setup

- Protein function prediction is a multi-label node classification task
- Every node (protein) is assigned one or more labels (functions)

**Setup:**
- Learn **OhmNet embeddings** for multi-layer tissue network
- Train a classifier for each function based on a fraction of proteins and all their functions
- Predict functions for new proteins
Tissue-Specific Protein Functions

OhmNet

Protein function prediction methods

Mono-layer network embeddings

Tensor decompositions

>10% improvement over function prediction methods

>18% improvement over non-hierarchical versions of the dataset

>15% improvement over matrix-based methods
Case Study: 9 Brain Tissues

9 brain tissue PPI networks in a two-level hierarchy

Marinka Zitnik, Stanford, ISMB/ECCB 2017
Multi-Scale Node Embeddings

Brainstem
- Cerebellum
- Medulla oblongata
- Substantia nigra

Brain
- Frontal lobe
- Temporal lobe
- Pons
- Parietal lobe
- Occipital lobe
- Midbrain
Annotating Proteins in a New Tissue

- Transfer protein functions to an unannotated tissue
- Task: Predict functions in target tissue without access to any annotation/label in that tissue

<table>
<thead>
<tr>
<th>Target tissue</th>
<th>Tissue-specific (OhmNet)</th>
<th>Tissue non-specific</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Placenta</td>
<td>0.758</td>
<td>0.684</td>
<td>11%</td>
</tr>
<tr>
<td>Spleen</td>
<td>0.779</td>
<td>0.712</td>
<td>10%</td>
</tr>
<tr>
<td>Liver</td>
<td>0.741</td>
<td>0.553</td>
<td>34%</td>
</tr>
<tr>
<td>Forebrain</td>
<td>0.755</td>
<td>0.632</td>
<td>20%</td>
</tr>
<tr>
<td>Blood plasma</td>
<td>0.703</td>
<td>0.540</td>
<td>40%</td>
</tr>
<tr>
<td>Smooth muscle</td>
<td>0.729</td>
<td>0.583</td>
<td>25%</td>
</tr>
<tr>
<td>Average</td>
<td>0.746</td>
<td>0.617</td>
<td>21%</td>
</tr>
</tbody>
</table>

Reported are AUROC values (see paper for other metrics)
Conclusions

- Unsupervised feature learning for multi-layer networks
- Learned embeddings can be used for any downstream prediction task: node classification, node clustering, link prediction
- OhmNet predicts protein functions across biological contexts

A shift from flat networks to large multiscale systems in biology
Poster A-294

PREICTING MULTICELLULAR FUNCTIONS THROUGH MULTI-LAYER TISSUE NETWORKS

Marinka Zitnik, Stanford, ISMB/ECCB 2017