CS 5854: Projects

T. M. Murali

February 20, 2020
Student Presentations

- Start on Tuesday, February 25.
- Each presentation is 75 minutes long (including 20–30 minutes of questions).
- Prepare reading notes (Google doc).
  - Send me draft of your notes about 10-14 days before your presentation.
  - Share final version of notes with the class about 7 days before your presentation.
- Any questions about the papers I have assigned?
Suggestions on Reading and Presenting Papers

- Be sceptical/critical: even papers in Nature, Science, or PNAS have errors or invalid thinking.
Suggestions on Reading and Presenting Papers

- Be sceptical/critical: even papers in Nature, Science, or PNAS have errors or invalid thinking.

- Algorithmic/computational papers:
  - Are the biological assumptions valid?
  - Is the algorithm good and computational efficient? Can you improve the technique?
  - Can you mathematically describe the output of the algorithm?
  - Don’t have to give all details. You can just present the essential ideas.
Suggestions on Reading and Presenting Papers

- Be sceptical/critical: even papers in Nature, Science, or PNAS have errors or invalid thinking.

- Algorithmic/computational papers:
  - Are the biological assumptions valid?
  - Is the algorithm good and computational efficient? Can you improve the technique?
  - Can you mathematically describe the output of the algorithm?
  - Don’t have to give all details. You can just present the essential ideas.

- Read supplementary information. Often has details about the assumptions, the techniques, and the results.
**Class Projects that Resulted in Papers**

List of Projects

1. Develop BEELINE 2.0
2. Predict cell types
3. Analyze PanCancer data
4. Predict HPO annotations
Overview

1. BEELINE 2.0
2. Cell Type Prediction
3. Analyse PanCancer data
4. Predict HPO Annotations
1 BEELINE 2.0

- **Goal:** Improve usefulness of BEELINE for experimental scRNA-seq datasets.

  - **High priority**
    - Implement continuous integration.
    - Add GRN inference methods and test them.
    - Develop alternative gene selection strategies.
    - Implement additional evaluation measures developed in GRN inference papers.

  - **Medium priority**
    - Add experimental scRNA-seq datasets.
    - Find better ground truth datasets. Automate selection of cell-type.

  - **Low priority**
    - Try imputation of missing data first.
    - Add denoising methods, e.g., molecular cross validation paper.
    - For real datasets, run parameter search on each type of network when using that type as ground truth.
Overview

1. BEELINE 2.0
2. Cell Type Prediction
3. Analyse PanCancer data
4. Predict HPO Annotations
2 Cell Type Prediction

- A comparison of automatic cell identification methods for single-cell RNA sequencing data
- OnClass uses a 2-layer perceptron in combination with the Cell Ontology.
2 Cell Type Prediction

- A comparison of automatic cell identification methods for single-cell RNA sequencing data
- OnClass uses a 2-layer perceptron in combination with the Cell Ontology.
- Use network-based algorithms for predicting cell types.
  - Two networks: one is among cells and the other is the Cell Ontology.
  - Evaluate network propagation algorithms that respect the ontology structure.
  - Alternative is to develop improved deep learning methods.
Overview

1. BEELINE 2.0
2. Cell Type Prediction
3. Analyse PanCancer data
4. Predict HPO Annotations
3 Prediction of tumour origin

- A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns
- Use random forests and a simple deep learning architecture.
3 Prediction of tumour origin

- A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns
- Use random forests and a simple deep learning architecture.
3 Prediction of tumour origin

- A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns
- Use random forests and a simple deep learning architecture.
3 Prediction of tumour origin

- A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns
- Use random forests and a simple deep learning architecture.
3 Prediction of tumour origin

- A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns
- Use random forests and a simple deep learning architecture.
- Goal: improve performance for poorly-predicted cancers. Try multi-task learning?
Overview

1. BEELINE 2.0
2. Cell Type Prediction
3. Analyse PanCancer data
4. Predict HPO Annotations
Goal: Predict gene annotations to HPO terms.

Develop BEELINE-like framework:

- Multiple algorithms.
- Multiple networks.
- Different evaluation criteria.

Benchmarking network propagation methods for disease gene identification
A functional linkage network (FLN) is a graph where each node corresponds to a gene and each edge connects two genes that may be associated with the same phenotype.

FLN construction can use diverse sources of data: gene expression, protein interactions, genomic context, text mining, etc.

An edge may not indicate which phenotype the connected genes share.
We adopt a term-by-term approach.

For each HPO term, add context to FLN by labelling nodes.

Each node \( v \) has an associated state \( s(v) \):

- \( s(v) = 1 \): gene \( v \) is annotated with HPO term \( f \). **Positives**
- \( s(v) = -1 \): gene \( v \) is annotated with another HPO term \( f' \). **Negatives**
- \( s(v) = 0 \): otherwise. **Unknowns**
Algorithm: SinkSource

- Compute state $s(v)$ between 0 and 1 for each unknown example $v$. 
Algorithm: SinkSource

- Compute state $s(\nu)$ between 0 and 1 for each unknown example $\nu$. 
Algorithm: SinkSource

- Compute state $s(v)$ between 0 and 1 for each unknown example $v$.
- Compute voltage at each node by solving linear system of equations:

$$s(v) = \sum_{u \text{ is a neighbour}} \alpha w(u, v)s(u) + f(v)$$

$u$ is unknown
Hardware Support for Projects

- Research virtual machines maintained by the Department of Computer Science.
- 40-processor, 20-node cluster in the Department of Computer Science dedicated to bioinformatics (baobab.cs.vt.edu).
- Obtain accounts on bioinformatics.cs.vt.edu from Rob Hunter (rhunter at vt dot edu).
Software Support for Projects

- My software page:  
  http://bioinformatics.cs.vt.edu/~murali/software
- My Github page:  http://github.com/Murali-group
- FastSinkSource:  https://github.com/jlaw9/FastSinkSource
- Biorithm software suite (C++):  http://bioinformatics.cs.vt.edu/~murali/software/biorithm-docs
Ground Rules for Projects

- Send me project choices by Thursday, February 20.
- 1 hour meetings with each group every 2 weeks or 4 weeks.
- Maintain Google docs describing your project and your progress.
- Project descriptions (motivation, background, related and previous research, approach, data, any preliminary results) due on Tuesday, March 17.
- Final project presentations possibly on Tuesday, May 5. Also possible to set aside a special day just for final presentations.
- Final project reports due on 5pm, Friday, May 8: 11pt font, 10 pages (not counting references), formatted like a journal paper.
List of Projects

1. Develop BEELINE 2.0
2. Predict cell types
3. Analyze PanCancer data
4. Predict HPO annotations