CS 5854: Projects

T. M. Murali

January 26, 28, February 2, 2016
Goals of the Course

- Emphasise a data-driven approach to systems biology.
- Integrate massive quantities of different types of data
- Stress methods that can prioritise experiments.
- Learn techniques from clustering, data mining, and graph theory and apply them to solve specific biological questions.
Student Presentations

- Start on Tuesday, February 9.
- Each presentation is 75 minutes long (including 20–30 minutes of questions).
- Forms groups of $2 \leq k \leq 3$ students.
- A group of $k$ students will make $k$ presentations.
- Each group of $k$ students should select one–two papers, one for each class they will present.
- Single student can lead/give each individual presentation, but read, understand, and discuss papers as a group.
- Use the tag 2016-spring-csb-papers on CiteULike.
- Read (quickly) the papers before you select them!
- Use the current schedule only as a guideline for how long I think each paper will take to present. It is not the final schedule!
- Send me the names of the groups and your top four paper choices by Tuesday, January 26.
- I will post the schedule by Thursday, January 28.
Class Projects that Resulted in Papers


List of Projects

1. Develop PathLinker 2.0
2. Predict signal transduction pathways from gene expression data.
3. Compute chemical response networks
4. Analyze PanCancer data
5. Predict orientation of interactions in interaction networks.
6. Develop XTalk 2.0, discover links between insulin signaling and inflammation in diabetes
7. Use NLP to build gold standard for XTalk
8. Compute feedback loops in yeast regulatory networks
Overview

Focus of the Course

Projects

Develop PathLinker 2.0

Predict signal transduction pathways from gene expression data

Compute chemical response networks

Analyze PanCancer data

Predict orientation of interactions in networks

Develop XTalk 2.0

Use NLP to build gold standard for XTalk

Compute feedback loops in yeast regulatory networks

Support
Wnt Signaling in a Pathway Database

T. M. Murali January 26, 28, February 2, 2016 CS 5854: Projects
Signaling Pathways as Directed Graphs

Bidirected Physical Interactions

Directed Regulatory Interactions

T. M. Murali January 26, 28, February 2, 2016 CS 5854: Projects
Signaling Pathways as Directed Graphs

Receptors

Transcription Factors and Transcriptional Regulators
Reconstructing Signaling Pathways

Human protein-protein interactome
Reconstructing Signaling Pathways

Curated pathway is a subnetwork of the interactome
Reconstructing Signaling Pathways

**Question:** Can we reconstruct the curated pathway given only receptors and transcriptional regulators?
Reconstructing Signaling Pathways

Proposed pathway reconstruction
Evaluating Reconstructed Pathways

Curated Pathway
Evaluating Reconstructed Pathways

Curated Pathway and Proposed Reconstruction
Evaluating Reconstructed Pathways

Curated Pathway and Proposed Reconstruction
- Developed **PathLinker** to reconstruct *proteins and interactions*
- **Systematically evaluated** PathLinker and other algorithms on human signaling pathways from the NetPath and KEGG databases
▶ Developed PathLinker to reconstruct proteins and interactions
▶ Systematically evaluated PathLinker and other algorithms on human signaling pathways from the NetPath and KEGG databases
How Does PathLinker work?

Wnt network in GraphSpace

T. M. Murali January 26, 28, February 2, 2016 CS 5854: Projects
PathLinker 2.0

- Current algorithm does not learn from the structure of the pathway.
PathLinker 2.0

- Current algorithm does not learn from the structure of the pathway.
- Goal: develop a (machine-learning) algorithm that can use information on the edges in a pathway to predict new edges.
PathLinker 2.0

- Current algorithm does not learn from the structure of the pathway.
- Goal: develop a (machine-learning) algorithm that can use information on the edges in a pathway to predict new edges.
PathLinker 2.0

- Current algorithm does not learn from the structure of the pathway.
- Goal: develop a (machine-learning) algorithm that can use information on the edges in a pathway to predict new edges.
- Use cross-validation to test performance: develop meaningful ways of deleting edges.
Evaluating Reconstructions with Cross Validation

Curated pathway
Evaluating Reconstructions with Cross Validation

Edges removed for cross validation
Evaluating Reconstructions with Cross Validation

Proposed reconstruction
Evaluating Reconstructions with Cross Validation

Curated pathway and proposed reconstruction
Evaluating Reconstructs with Cross Validation

Cross validation edges and proposed reconstruction
Evaluating Reconstructions with Cross Validation

Precision and recall
Project Details

- PathLinker code on GitHub
- Update network dataset used in PathLinker paper.
- Develop machine learning method that can utilize partial information about edges in a pathway.
- Find several meaningful alternatives methods to compare your algorithm with.
- Create computational analyses that are different from the PathLinker paper.
- Do literature analysis of predicted interactions.
Overview

Focus of the Course

Projects

Develop PathLinker 2.0

Predict signal transduction pathways from gene expression data

Compute chemical response networks

Analyze PanCancer data

Predict orientation of interactions in networks

Develop XTalk 2.0

Use NLP to build gold standard for XTalk

Compute feedback loops in yeast regulatory networks

Support
Dissect Cellular Responses to Signals

Collaboration with Padma Rajagopalan (Chem Eng)

LSECs and Kupffer cells

Endothelial/Kupffer cells

Hepatocytes

Collagen

 PEM
Dissect Cellular Responses to Signals

Collaboration with Padma Rajagopalan (Chem Eng)
Dissect Cellular Responses to Signals

Collaboration with Padma Rajagopalan (Chem Eng)
Project Details

- **Papers:**
- **Use liver gene expression data from Padma Rajagopalan’s group.**
- **Find other appropriate datasets (by the middle of February):**
  1. Gene expression measurements after a signal.
     - ideally for signals on liver tissue.
     - ideally, accompanied by a proteomic dataset.
  2. Prefer human or mammalian datasets.
  3. Create a high-quality dataset of molecular interactions in appropriate organism.
  4. If proteomic data is available, use it to measure performance.
- Use software for existing algorithms (PathLinker, eQED, PCSF, etc).
- Perform literature based validation of predicted pathways.
Overview

Focus of the Course

Projects

Develop PathLinker 2.0

Predict signal transduction pathways from gene expression data

**Compute chemical response networks**

Analyze PanCancer data

Predict orientation of interactions in networks

Develop XTalk 2.0

Use NLP to build gold standard for XTalk

Compute feedback loops in yeast regulatory networks

Support
ToxCast → Chemical Response Networks

- Efforts such as the EPA’s ToxCast seek to quickly and efficiently screen thousands of chemicals for potential human and environmental effects.
- **Information is partial:** for each chemical, these efforts study a fixed set of proteins and pathways.
- **Assays are independent:** Do not consider the complex network of interactions among assayed proteins and pathways.
ToxCast $\rightarrow$ Chemical Response Networks

- Efforts such as the EPA’s ToxCast seek to quickly and efficiently screen thousands of chemicals for potential human and environmental effects.
- **Information is partial:** for each chemical, these efforts study a fixed set of proteins and pathways.
- **Assays are independent:** Do not consider the complex network of interactions among assayed proteins and pathways.
- We seek to connect the responding proteins in the context of the underlying network of regulatory and physical interactions.
- **Toxicant response network:** network of regulatory, signaling and physical interactions that connects proteins that are perturbed as a result of toxicant exposure.
ToxCast → Chemical Response Networks

- Efforts such as the EPA’s ToxCast seek to quickly and efficiently screen thousands of chemicals for potential human and environmental effects.

- **Information is partial:** for each chemical, these efforts study a fixed set of proteins and pathways.

- **Assays are independent:** Do not consider the complex network of interactions among assayed proteins and pathways.

- We seek to connect the responding proteins in the context of the underlying network of regulatory and physical interactions.

- **Toxicant response network:** network of regulatory, signaling and physical interactions that connects proteins that are perturbed as a result of toxicant exposure.

- Toxicant response networks may reveal
  - important intermediate proteins that have not have been tested and
  - physiological processes that have not been previously implicated in connection with the chemical.
Bisphenol-A Response Network

- Activation of AP1
- PYK2 Pathway
- Akt Pathway
- RXR/VDR Pathway

Physical Interaction
Regulatory Interaction
Responsive Protein
Non-responsive Protein

ALK1 Pathway
BMP Pathway
RXR/VDR Pathway
PYK2 Pathway
Akt Pathway
RXR/VDR Pathway

Down-regulation of TGFβ Receptor Signaling

Pathway Probability
- ALK1 Pathway (PID) 0.998
- BMP Pathway (PID) 0.996
- RXR/VDR Pathway (PID) 0.981
- AKT Pathway (Biocarta) 0.881
- P38γδ Pathway (PID) 0.808
- Activation of AP1 Family of TFs (Reactome) 0.711
- Downregulation of TGFβ Receptor Signaling (Reactome) 0.669
- PYK2 Pathway (Biocarta) 0.654

T. M. Murali January 26, 28, February 2, 2016 CS 5854: Projects
# Bisphenol-A Response Network

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Posterior Probability</th>
<th>Literature Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALK1 Pathway (PID)</td>
<td>0.998</td>
<td>✓</td>
</tr>
<tr>
<td>BMP Pathway (PID)</td>
<td>0.996</td>
<td>✓</td>
</tr>
<tr>
<td>RXR VDR Pathway (PID)</td>
<td>0.981</td>
<td>✓</td>
</tr>
<tr>
<td>AKT Pathway (Biocarta)</td>
<td>0.881</td>
<td>✓</td>
</tr>
<tr>
<td>P38 γδ Pathway (PID)</td>
<td>0.808</td>
<td>✓</td>
</tr>
<tr>
<td>Activation of AP1 Family of TFs (Reactome)</td>
<td>0.711</td>
<td>✓</td>
</tr>
<tr>
<td>Downregulation of TGFβR Signaling (Reactome)</td>
<td>0.669</td>
<td>✓</td>
</tr>
<tr>
<td>PYK2 Pathway (Biocarta)</td>
<td>0.654</td>
<td>✓</td>
</tr>
</tbody>
</table>
Project Details

- Papers:

- Find other papers on ToxCast and Tox21 datasets and understand how these data are accessible through the EPA and NIEHS websites.
- Apply PathLinker to each chemical to compute its “response network.”
- Integrate relevant gene expression dataset for each chemical, if available.
- Demonstrate increased knowledge in response networks when compared to the original ToxCast data.
Overview

Focus of the Course

Projects

Develop PathLinker 2.0

Predict signal transduction pathways from gene expression data

Compute chemical response networks

Analyze PanCancer data

Predict orientation of interactions in networks

Develop XTalk 2.0

Use NLP to build gold standard for XTalk

Compute feedback loops in yeast regulatory networks

Support
**PanCancer + PathLinker**

- Use PathLinker to analyze PanCancer data.
- Papers:
  - Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE)
  - Pathway and network analysis of cancer genomes (Review)
- Most of these papers analyze overlap of the genes in the computed networks with known physiological processes.
- Can we use PathLinker to discover mechanisms by which genomic events control transcriptional changes?
Overview

Focus of the Course

Projects

Develop PathLinker 2.0
Predict signal transduction pathways from gene expression data
Compute chemical response networks
Analyze PanCancer data

Predict orientation of interactions in networks

Develop XTalk 2.0
Use NLP to build gold standard for XTalk
Compute feedback loops in yeast regulatory networks

Support
T. M. Murali January 26, 28, February 2, 2016 CS 5854: Projects
Wnt Signaling in a Pathway Database

- **Ligand**: Wnt
- **Receptor**: Frizzled
- **Protein**: LRP6, LRP5, ROR1, RYK, DVL1, GSK3B, AXIN1, CTNNB1, MAPK8, MAPK9, PRKCA, PRKCB, PRKCG
- **mRNA**: DVL3, RAC1, PIP5K1B, PI4K2A

**Protein-protein interaction**: Green lines
**Auto catalysis**: Blue lines
**Phosphorylation**: Red lines
**Induced catalysis**: Turquoise lines
**Induction**: Orange lines
**Transport**: Purple lines

**Cytoplasm**: CY
**Nucleus**: NU

**Netslim**: www.netpath.org/netslim

T. M. Murali January 26, 28, February 2, 2016 CS 5854: Projects
Orienting Interactions

- Interactions in pathways are often directed because they correspond to regulation.
- Interactions in PPI networks are undirected because they are physical.
- Human protein interactome:
  - 12K nodes and 110K edges
  - 87K physical interactions
    - BIND, DIP, InnateDB, IntAct, MINT, MatrixDB, Reactome, NetPath, KEGG, SPIKE
  - 33K signaling interactions
    - NetPath, KEGG, SPIKE
Orienting Interactions

- Interactions in pathways are often directed because they are correspond to regulation.
- Interactions in PPI networks are undirected because they are physical.
- Human protein interactome:
  - 12K nodes and 110K edges
  - 87K **physical interactions**
    - BIND, DIP, InnateDB, IntAct, MINT, MatrixDB, Reactome, NetPath, KEGG, SPIKE
  - 33K **signaling interactions**
    - NetPath, KEGG, SPIKE
- Develop methods to assign direction to edges in PPI networks and quantitatively evaluate your results.
Approach for Orienting Interactions

Human protein-protein interactome
Approach for Orienting Interactions

Pathway is a subnetwork of the interactome
Approach for Orienting Interactions

Question: Can we assign edge orientations given only receptors and transcriptional regulators?
Approach for Orienting Interactions

Proposed orientations
Project Details

- **Papers:**

- **Software:** *PathLinker* (Murali), *OrientEdges* (Gitter), *Shortest* (Sharan)
Project Details

- **Papers:**

- **Software:** PathLinker (Murali), OrientEdges (Gitter), Shortest (Sharan)

- Implement OrientEdges and Shortest yourself in Python within the PathLinker package.

- Evaluate by comparing to gold standard dataset of interactions with experimentally known orientation.
Overview

Focus of the Course

Projects

Develop PathLinker 2.0

Predict signal transduction pathways from gene expression data

Compute chemical response networks

Analyze PanCancer data

Predict orientation of interactions in networks

Develop XTalk 2.0

Use NLP to build gold standard for XTalk

Compute feedback loops in yeast regulatory networks

Support
Signaling pathway crosstalk

Hippo Receptor

NF2 → Mst1/2 → Lats1/2 → YAP/TAZ

General Assumption: pathways that share nodes crosstalk
Signaling pathway crosstalk

General Assumption: pathways that share nodes crosstalk.

T. M. Murali January 26, 28, February 2, 2016 CS 5854: Projects
Signaling pathway crosstalk

Hippo Receptor

TGFBR1/2

SMAD2/3

NF2

Mst1/2

Lats1/2

YAP

TAZ

SMAD4

E2F4/5

Cell

General Assumption: pathways that share nodes crosstalk
General Assumption: pathways that share nodes crosstalk
Importance of Crosstalk

- Crosstalk is a naturally occurring phenomenon
Importance of Crosstalk

- Crosstalk is a naturally occurring phenomenon
- Crosstalk has been implicated in many diseases, including cancer, host pathogen defense, and regulation of cell survival
Importance of Crosstalk

- Crosstalk is a naturally occurring phenomenon
- Crosstalk has been implicated in many diseases, including cancer, host pathogen defense, and regulation of cell survival
- Surprisingly, no computational methods exist to predict which pairs of pathways may crosstalk
Importance of Crosstalk

- Crosstalk is a naturally occurring phenomenon
- Crosstalk has been implicated in many diseases, including cancer, host pathogen defense, and regulation of cell survival
- Surprisingly, no computational methods exist to predict which pairs of pathways may crosstalk

**XTalk**: a method to identify which pathway pairs may crosstalk.
**XTalk: average length of the $k$ shortest paths**

**Inputs:**

- **A** Pathway A
- **B** Pathway B
- **$R_A$** Receptors in A
- **$T_B$** TFs in B
- **$k$** number of paths
- **G** background network

**Notation:**

- $d(r, t, k)$ the length of the $k$th shortest path from $r$ to $t$
- $\chi(A, B, k)$ the XTalk statistic

$$\chi(A, B, k) = \frac{k}{|T_B|} \sum_{t \in T_B} \sum_{l=1}^{k} d(R_A, t, l),$$
**XTalk: average length of the $k$ shortest paths**

**Inputs:**

- $A$ Pathway A
- $B$ Pathway B
- $R_A$ Receptors in A
- $T_B$ TFs in B
- $k$ number of paths
- $G$ background network

**Notation:**

- $d(r, t, k)$ the length of the $k^{th}$ shortest path from $r$ to $t$
- $\chi(A, B, k)$ the $\text{XTalk}$ statistic

$$\chi(A, B, k) = \frac{1}{k|T_B|} \sum_{t \in T_B} \sum_{l=1}^{k} d(R_A, t, l),$$
Performance of XTalk

AUC: 0.65

AUC: 0.58

AUC: 0.50

XTalk  BPLN  NIC

AUC: 0.65
AUC: 0.58
AUC: 0.50
**XTalk network from Hippo to TGF-β Pathway**

http://graphspace.org/graphs/ategge@vt.edu/kegg-hippo-tgfbeta
**XTalk 2.0**

- Simple: improve the AUC of XTalk!
- Just like PathLinker, XTalk does not learn anything from the structure of the pathways.

**Papers:**

- XTalk: a path-based approach for identifying crosstalk between signaling pathways, Tegge, Sharp, and Murali, Bioinformatics, 32, 242–251
Simple: improve the AUC of XTalk!

Just like PathLinker, XTalk does not learn anything from the structure of the pathways.

Can we use machine learning techniques to improve XTalk?

XTalk can use edges from any pathway in its crosstalk networks. Can we constrain the types of labels that can appear in a path?
XTalk 2.0

- Simple: improve the AUC of XTalk!
- Just like PathLinker, XTalk does not learn anything from the structure of the pathways.
- Can we use machine learning techniques to improve XTalk?
- XTalk can use edges from any pathway in its crosstalk networks. Can we constrain the types of labels that can appear in a path?

Papers:

- XTalk: a path-based approach for identifying crosstalk between signaling pathways, Tegge, Sharp, and Murali, Bioinformatics, 32, 242–251
Overview

Focus of the Course

Projects

Develop PathLinker 2.0

Predict signal transduction pathways from gene expression data

Compute chemical response networks

Analyze PanCancer data

Predict orientation of interactions in networks

Develop $\textit{XTalk}$ 2.0

\textbf{Use NLP to build gold standard for $\textit{XTalk}$}

Compute feedback loops in yeast regulatory networks

Support
Selected 17 signaling pathways in KEGG.

For each pair of pathways,

- Performed PubMed query and
- Read papers for evidence of crosstalk
- Recorded sentence providing evidence, if any.
XTalk Gold Standard

- Selected 17 signaling pathways in KEGG.
- For each pair of pathways,
  - Performed PubMed query and
  - Read papers for evidence of crosstalk
  - Recorded sentence providing evidence, if any.
- Goal: use natural language processing techniques to train system that reads articles or their abstracts to automatically determine whether a pair of pathways crosstalk.
- Requires some experience with NLP tools.

Crosstalk database
Overview

Focus of the Course

Projects

Develop PathLinker 2.0

Predict signal transduction pathways from gene expression data

Compute chemical response networks

Analyze PanCancer data

Predict orientation of interactions in networks

Develop XTalk 2.0

Use NLP to build gold standard for XTalk

Compute feedback loops in yeast regulatory networks

Support
Focus Projects PL 2.0 Expression to Signals ToxCast PanCancer Orient XTalk 2.0 NLP Feedback Support

GraphSpace: Graphs Groups Help

Graph Details
Export
Owner
Sharing
Layouts
Save Layout
Auto Manual
My Layouts 1 Shared 0
Public 1
layout1

T. M. Murali January 26, 28, February 2, 2016 CS 5854: Projects
Focus Projects PL 2.0 Expression to Signals ToxCast PanCancer Orient XTalk 2.0 NLP Feedback Support

T. M. Murali January 26, 28, February 2, 2016 CS 5854: Projects
Feedback Loops

- These computed networks contain no loops or feedback cycles!
- Feedback cycles are fundamental components of regulatory circuits.
Feedback Loops

- These computed networks contain no loops or feedback cycles!
- Feedback cycles are fundamental components of regulatory circuits.
- How difficult is to compute all the cycles of a length three in a graph?
Feedback Loops

- These computed networks contain no loops or feedback cycles!
- Feedback cycles are fundamental components of regulatory circuits.
- How difficult is to compute all the cycles of a length three in a graph? Length four, five, six?
Feedback Loops

- These computed networks contain no loops or feedback cycles!
- Feedback cycles are fundamental components of regulatory circuits.
- How difficult is to compute all the cycles of a length three in a graph? Length four, five, six? The large number of possible cycles may make cost prohibitive.
- Computing all cycles also may not be very interesting.
- Goal: Given a pair of query nodes $s$ and $t$, compute the $k$ shortest cycles containing $s$ and $t$. 
Feedback Loops

- These computed networks contain no loops or feedback cycles!
- Feedback cycles are fundamental components of regulatory circuits.
- How difficult is to compute all the cycles of a length three in a graph? Length four, five, six? The large number of possible cycles may make cost prohibitive.
- Computing all cycles also may not be very interesting.
- Goal: Given a pair of query nodes \( s \) and \( t \), compute the \( k \) shortest cycles containing \( s \) and \( t \).
- Develop algorithm.
- Apply it to yeast or human regulatory and signaling networks.
- Develop methods to evaluate results.
- Use literature to provide additional support.
Overview

Focus of the Course

Projects

Develop PathLinker 2.0

Predict signal transduction pathways from gene expression data

Compute chemical response networks

Analyze PanCancer data

Predict orientation of interactions in networks

Develop XTALK 2.0

Use NLP to build gold standard for XTALK

Compute feedback loops in yeast regulatory networks

Support
Hardware Support for Projects

- 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

▶ Biorithm software suite:  http://bioinformatics.cs.vt.edu/~murali/software/biorithm-docs

▶ PathLinker, PCSF, ResponseNet, etc.: Code or pipelines implemented in Python and other languages.
Ground Rules for Projects

▶ 1 hour meetings with each group every 2 weeks or 4 weeks.
▶ Maintain Google docs describing your project and your progress.
▶ Mid-term project review presentations on Tuesday, March 15 in class.
▶ Project descriptions (motivation, background, related and previous research, approach, data, any preliminary results) due on Tuesday, March 15.
▶ Final project presentations possibly on Tuesday, May 3 and Wednesday, May 4. Also possible to set aside a special day just for final presentations.
▶ Final project reports due on 5pm, Friday, May 6.
List of Projects

1. Develop PathLinker 2.0
2. Predict signal transduction pathways from gene expression data.
3. Compute chemical response networks
4. Analyze PanCancer data
5. Predict orientation of interactions in interaction networks.
6. Develop \textit{XTalk} 2.0, discover links between insulin signaling and inflammation in diabetes
7. Use NLP to build gold standard for \textit{XTalk}
8. Compute feedback loops in yeast regulatory networks