A Convex Optimization Approach for Identification of Human Tissue-specific Interactomes

Shahin Mohammadi and Ananth Grama

Department of Computer Science
Purdue University

July, 2016

Outline

1. **Background**
   - Problem statement
   - Previous work
   - Example
2. **Activity Propagation (ActPro)**
   - Standardizing gene expression profiles
   - Computing functional activity of genes
   - Updating global interactome
3. **Results**
   - Network statistics
   - Predicting known biology
   - Identifying disease-related pathways

Thanks

Travel funding to ISMB/ECCB 2016 was generously provided by ISCB

Thank you!

Motivation

Global human interactome is a superset of all possible physical interactions that can take places in the cell. It does not provide any information as to which one of these interactions do take place in a given tissue/cell-type context.

Outline

1. **Background**
   - Problem statement
   - Previous work
   - Example
2. **Activity Propagation (ActPro)**
   - Standardizing gene expression profiles
   - Computing functional activity of genes
   - Updating global interactome
3. **Results**
   - Network statistics
   - Predicting known biology
   - Identifying disease-related pathways

What we have – What we want

Available data sources:

1. A global interactome, which contains the set of *possible* interacting pairs.
2. A tissue-specific measurement of gene/protein activity within each tissue/cell type.

Problem

How can we optimally utilize transcriptional activity of gene products to construct the most informative tissue-specific sub-network?

Outline

1. **Background**
   - Problem statement
   - Previous work
   - Example
2. **Activity Propagation (ActPro)**
   - Standardizing gene expression profiles
   - Computing functional activity of genes
   - Updating global interactome
3. **Results**
   - Network statistics
   - Predicting known biology
   - Identifying disease-related pathways

Definition

1. Node Removal (NR): Tissue-specific network is generated by removing from the global network proteins that are not expressed in the relevant tissue.
2. Edge Reweight (ERW): It modify the edge weights to reflect the probability that the corresponding interactions take place in the specific tissue.

$$w'_{ij} = w_{ij} * \alpha^k$$

where $0 \leq \alpha \leq 1$ is the re-weighting factor and $k \in \{0, 1, 2\}$ is the number of end-points for the protein-protein interaction that are expressed in the tissue of interest.

Problems with Previous Methods

- Rely on an ad hoc threshold for identifying expressed genes.
- Utilize only local topology around each edge, specifically its end-points, to decide about existence/probability of an interaction.

Outline

1. **Background**
   - Problem statement
   - Previous work
   - Example
2. **Activity Propagation (ActPro)**
   - Standardizing gene expression profiles
   - Computing functional activity of genes
   - Updating global interactome
3. **Results**
   - Network statistics
   - Predicting known biology
   - Identifying disease-related pathways

Example subnetwork

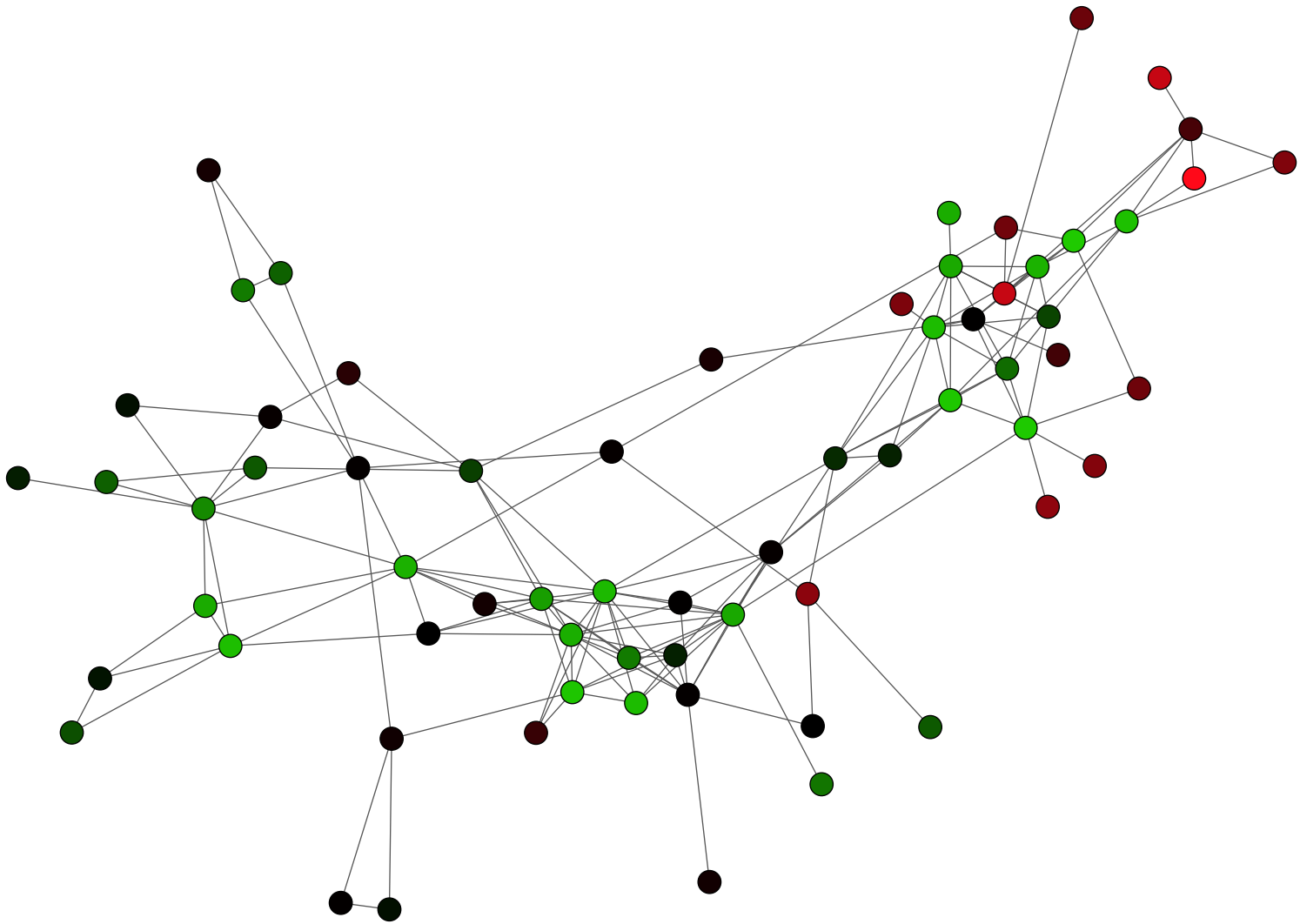


Figure : Example of an upregulated pathway in blood cells– Antigen processing and presentation

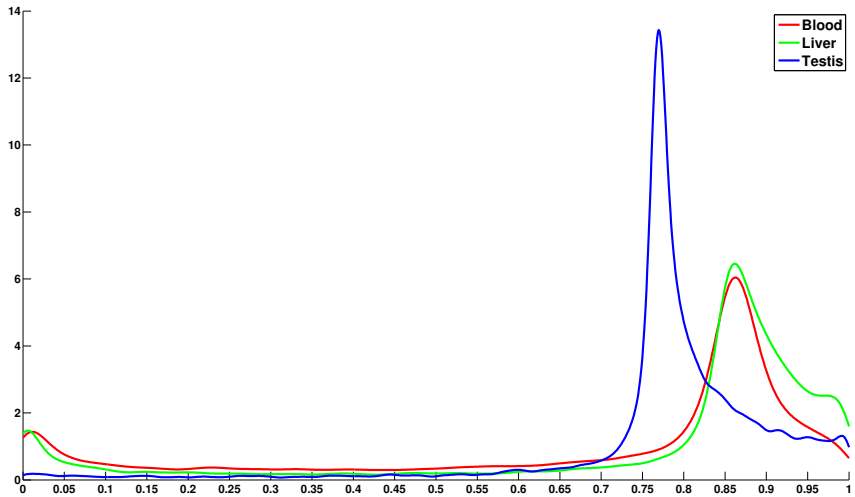
Outline

1. Background
   - Problem statement
   - Previous work
   - Example
2. **Activity Propagation (ActPro)**
   - Standardizing gene expression profiles
   - Computing functional activity of genes
   - Updating global interactome
3. Results
   - Network statistics
   - Predicting known biology
   - Identifying disease-related pathways

Universal exPression Codes (UPC)

Computing transcriptional activity of genes

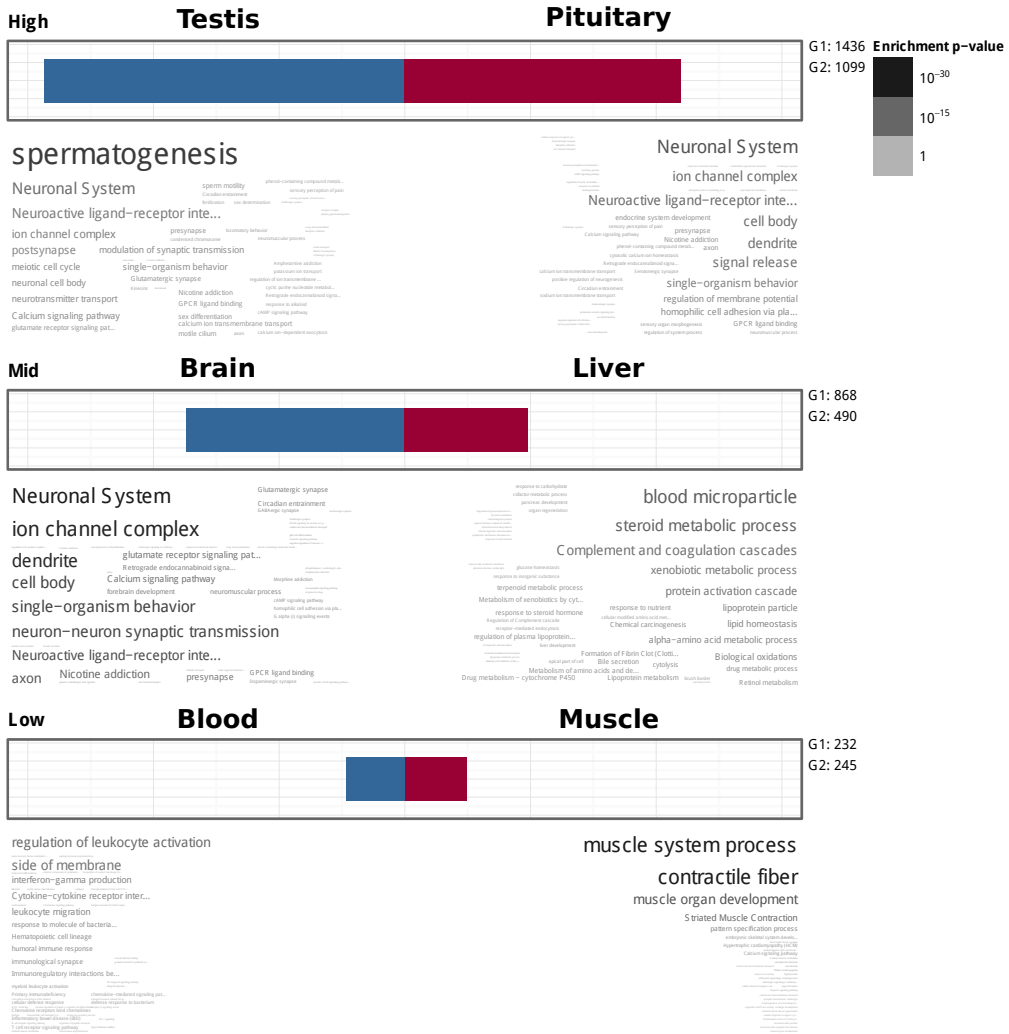
- Processes each sample individually
- Corrects for platform-specific background noise
- Uses a mixture model to estimate whether a gene transcriptional activity
- Has been demonstrated that, for tissue samples profiled using both microarrays and RNA-Seq, UPC values can be highly concordant



Distribution of transcriptional activities in three tissues with low, medium, and high number of expressed genes

Universal exPression Codes (UPC)

Validating tissue-specific markers



GO enrichment for tissues with high, medium, and low number of markers

Outline

1. Background
   - Problem statement
   - Previous work
   - Example
2. **Activity Propagation (ActPro)**
   - Standardizing gene expression profiles
   - Computing functional activity of genes
   - Updating global interactome
3. Results
   - Network statistics
   - Predicting known biology
   - Identifying disease-related pathways

Minimal number of changes that smooths transcriptional activities over adjacent nodes in the network:

$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmin}} \left\{ (1-\alpha)\mathbf{x}\mathbf{L}\mathbf{x} + \alpha \parallel \mathbf{x} - \mathbf{z} \parallel_1 \right\}$$

$$\text{Subject to: } \begin{cases} \mathbf{1}^T\mathbf{x} = 1 \\ 0 \leq \mathbf{x} \end{cases}$$

- Vector $\mathbf{z}$ initial value of transcriptional activities estimated by UPC
- Matrix $\mathbf{L}$ is the *Laplacian* matrix, defined as $\mathbf{A} - \mathbf{D}$, where d_{ii} is the weighted degree of i^{th} vertex in the global interactome.
- Parameter α controls the relative importance of regularization

- The first term defines a *diffusion kernel* that propagates activity of genes through network links.
- We can expand it as $\sum_{i,j} w_{i,j}(x_i - x_j)^2$, which is the accumulated difference of values between adjacent nodes scaled by the weight of the edge connecting them.
- The Laplacian operator $\mathbf{L}$ acts on a given function defined over vertices of a graph, such as $\mathbf{x}$, and computes the smoothness of $\mathbf{x}$ over adjacent vertices.
- It can be also computed as $\| \mathbf{Bx} \|_2^2$, where $\mathbf{B}$ is the incident matrix of the graph

- The second term is a *regularizer* which penalizes changes or deviations
- We can expand it as $\sum_i |x_i - z_i|$, where x_i and z_i are the (inferred) functional and the transcriptional activity of gene i, respectively.
- It enforces sparsity over the vector of differences between *transcriptional* and *functional* activities.

Outline

1. Background
   - Problem statement
   - Previous work
   - Example
2. **Activity Propagation (ActPro)**
   - Standardizing gene expression profiles
   - Computing functional activity of genes
   - Updating global interactome
3. Results
   - Network statistics
   - Predicting known biology
   - Identifying disease-related pathways

Updating edges

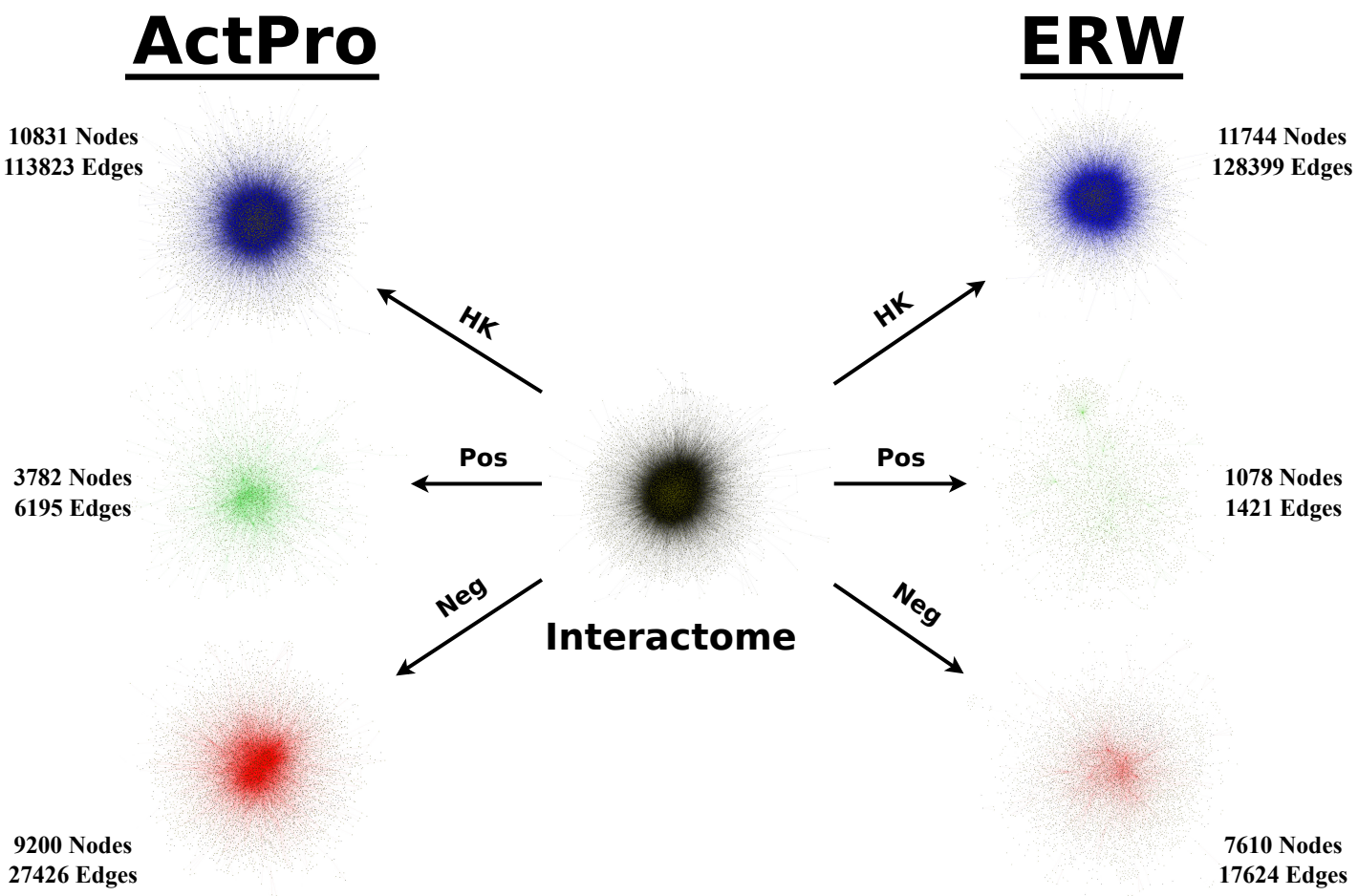
$$\hat{\mathbf{A}} = \mathbf{diag}(\mathbf{x}^*) * \mathbf{A} * \mathbf{diag}(\mathbf{x}^*)$$

- $\mathbf{x}^*$ is the solution of optimization problem
- It represents functional activity of genes
- Functional activities are inferred from the global network context
- We update each edge according to the functional activity of its end-points

Outline

1. Background
   - Problem statement
   - Previous work
   - Example
2. Activity Propagation (ActPro)
   - Standardizing gene expression profiles
   - Computing functional activity of genes
   - Updating global interactome
3. Results
   - Network statistics
   - Predicting known biology
   - Identifying disease-related pathways

Decomposition of global interactome



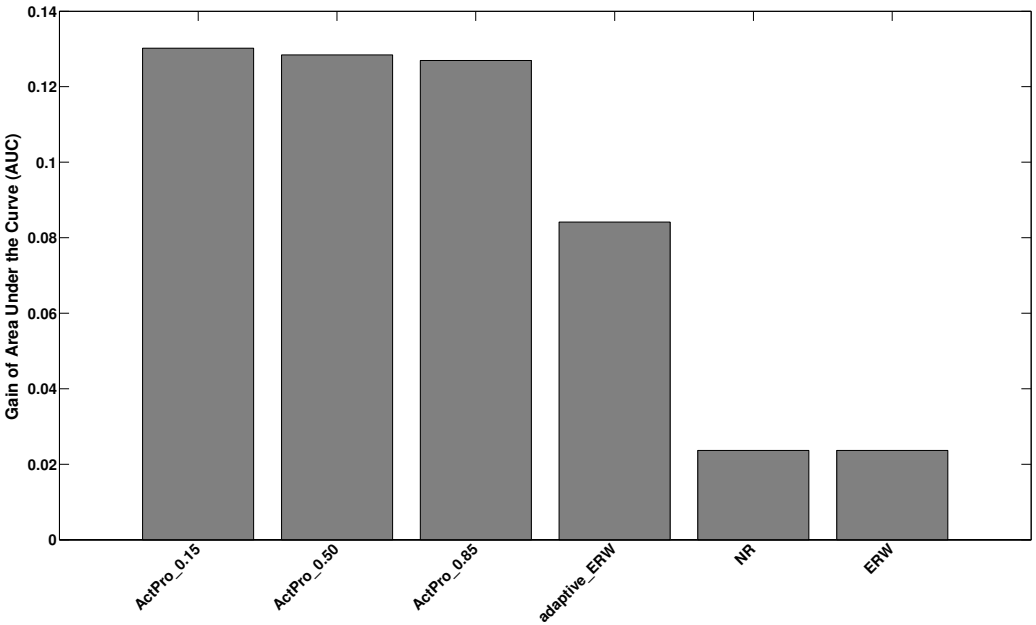
Brain-specific network using ERW and ActPro ($\alpha = 0.5$) methods

Outline

1. Background
   - Problem statement
   - Previous work
   - Example
2. Activity Propagation (ActPro)
   - Standardizing gene expression profiles
   - Computing functional activity of genes
   - Updating global interactome
3. Results
   - Network statistics
   - Predicting known biology
   - Identifying disease-related pathways

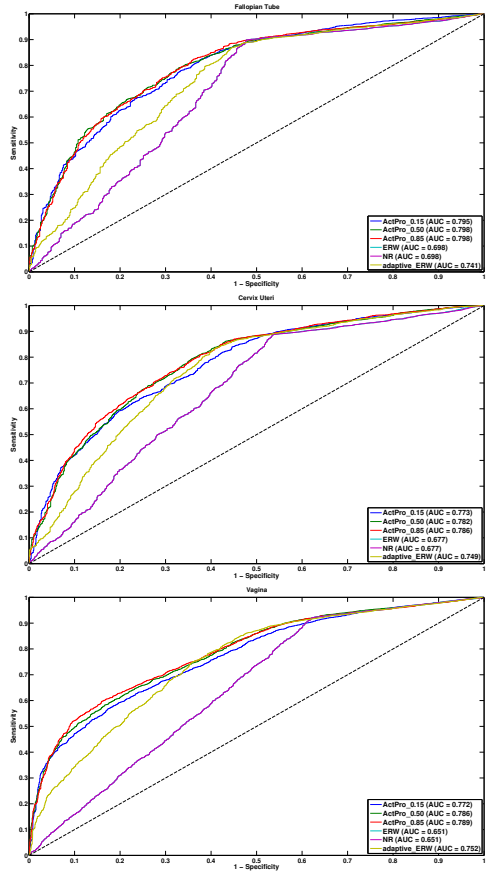
Results

Predicting tissue-specific interactions in known functional pathways



Average performance of different methods in predicting differential interactions

Edge Set Enrichment Analysis (ESEA)



Tissues with the highest gain of AUC for predicting tissue-specific pathway edges

Results

Compactness of disease-related genes

	global	ActPro_0.15	ActPro_0.50	ActPro_0.85	ERW	NR
Alzheimer's disease	**4.12E-3**	6.96E-3	5.98E-3	5.44E-3	5.32E-3	9.60E-2
breast carcinoma	1.83E-3	1.11E-3	8.40E-4	**8.30E-4**	4.09E-3	8.15E-2
chronic lymphocytic leukemia	8.20E-4	7.40E-4	**4.80E-4**	5.10E-4	8.50E-4	2.94E-2
coronary artery disease	3.95E-1	1.58E-1	1.09E-1	1.03E-1	1.33E-1	**1.93E-2**
Crohn's disease	2.56E-2	1.93E-2	1.50E-2	**1.44E-2**	8.54E-2	4.14E-1
metabolic syndrome X	1.11E-2	1.09E-2	**1.07E-2**	1.12E-2	1.02E-1	7.39E-1
Parkinson's disease	1.59E-2	1.25E-2	9.89E-3	**9.50E-3**	1.34E-2	9.62E-2
primary biliary cirrhosis	**7.20E-4**	1.32E-3	3.16E-3	3.40E-3	2.80E-2	6.86E-1
psoriasis	**2.10E-4**	1.10E-3	1.16E-3	9.50E-4	4.67E-3	3.24E-1
rheumatoid arthritis	1.70E-2	**9.28E-3**	1.06E-2	1.10E-2	6.39E-2	3.61E-1
systemic lupus erythematosus	4.98E-2	1.19E-2	7.56E-3	7.22E-3	2.55E-3	**1.60E-4**
type 1 diabetes mellitus	2.64E-2	3.01E-2	**2.38E-2**	2.40E-2	2.64E-1	9.39E-1
type 2 diabetes mellitus	1.57E-3	2.90E-4	2.40E-4	**1.80E-4**	5.60E-4	7.90E-3
vitiligo	**1.17E-3**	2.13E-3	3.04E-3	3.54E-3	1.84E-2	5.69E-1
schizophrenia	3.47E-1	2.13E-1	1.93E-1	1.84E-1	1.40E-1	**4.10E-2**
combined	1.53E-13	1.24E-17	6.62E-19	**3.70E-19**	9.03E-14	2.43E-03

- Symmetric random-walk as a measure of distance
- Empirical p-value for each tissue
- p-value combination using Edgington method

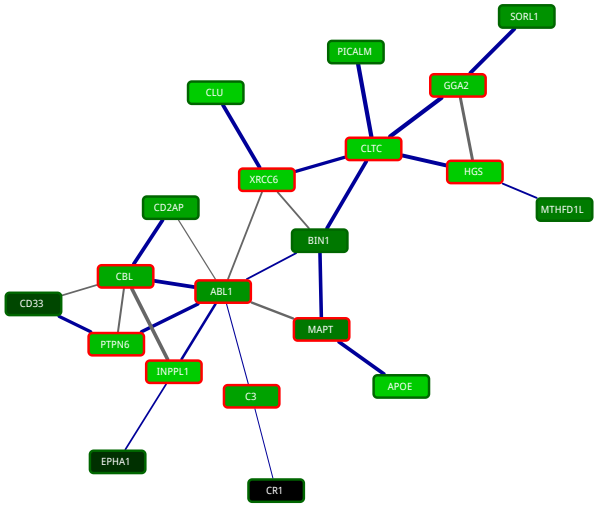
$\Rightarrow$ *ActPro* yields more significant compactness for known disease genes

Outline

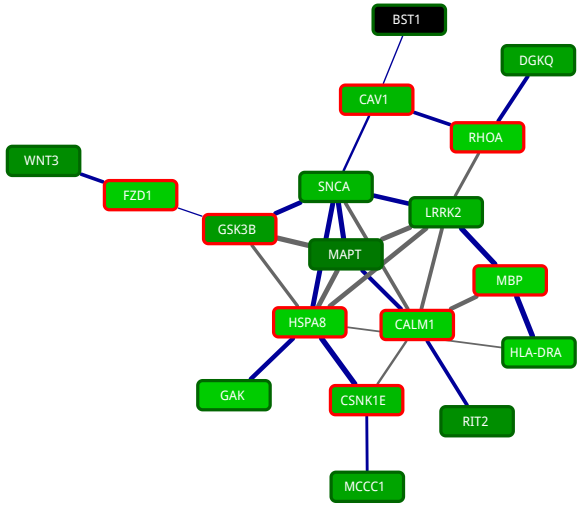
1. Background
   - Problem statement
   - Previous work
   - Example
2. Activity Propagation (ActPro)
   - Standardizing gene expression profiles
   - Computing functional activity of genes
   - Updating global interactome
3. Results
   - Network statistics
   - Predicting known biology
   - Identifying disease-related pathways

Results

Identifying novel disease-related pathways



Alzheimer's Disease



Parkinson's Disease

- Prize Collecting Steiner Tree (PCST)

$$\underset{<v,e>\in T}{\operatorname{argmin}} \left\{ \sum_{e} c_e - \lambda \sum_{v} b_v \right\}$$

- Solved using known message-passing algorithm

Questions

PS: I am defending this summer and will be in the job marker for PostDocs.