

# A SYSTEMS APPROACH TO DISSECTING THE TISSUE-SPECIFIC ARCHITECTURE OF CELLULAR NETWORKS

Ananth Grama

Center for Science of Information  
Purdue University

IUPUI 2014

Work with S. Mohammadi (Purdue), S. Subramaniam (UCSD),  
and G. Kollias (IBM)

# OUTLINE

## 1 PART 1: MOTIVATION – CONSTRUCTING AGING PATHWAYS IN YEAST

- Overview
- Materials and Methods
  - Datasets
  - Tracing Information Flow
- Results and Discussion

## 2 PART 2: TISSUE SPECIFIC NETWORKS AND COMPARATIVE ANALYSIS

- Motivation
- Materials and methods
  - Datasets
  - Algorithmic contributions
- Results and discussion
  - Core alignment graph of housekeeping genes
  - Computing similarity of human tissues with yeast

# OUTLINE

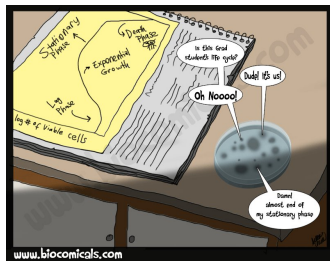
## 1 PART 1: MOTIVATION – CONSTRUCTING AGING PATHWAYS IN YEAST

- Overview
- Materials and Methods
  - Datasets
  - Tracing Information Flow
- Results and Discussion

## 2 PART 2: TISSUE SPECIFIC NETWORKS AND COMPARATIVE ANALYSIS

- Motivation
- Materials and methods
  - Datasets
  - Algorithmic contributions
- Results and discussion
  - Core alignment graph of housekeeping genes
  - Computing similarity of human tissues with yeast

# YEAST AGING



Courtesy of Alper Uzan, PhD.

- Yeast as a model organism for aging research:
  - ✓ Rapid growth
  - ✓ Ease of manipulation
- **Replicative life-span (RLS):** the number of buds a mother cell can produce before senescence occurs
- **Chronological life-span (CLS):** duration of viability after entering the stationary-phase

# OUTLINE

## 1 PART 1: MOTIVATION – CONSTRUCTING AGING PATHWAYS IN YEAST

- Overview
- **Materials and Methods**
  - Datasets
  - Tracing Information Flow
- Results and Discussion

## 2 PART 2: TISSUE SPECIFIC NETWORKS AND COMPARATIVE ANALYSIS

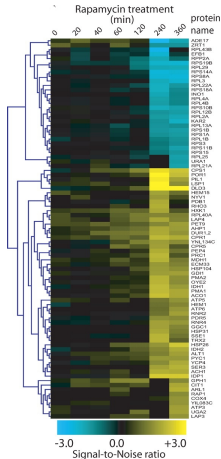
- Motivation
- Materials and methods
  - Datasets
  - Algorithmic contributions
- Results and discussion
  - Core alignment graph of housekeeping genes
  - Computing similarity of human tissues with yeast



# TRANSCRIPTIONAL REGULATORY NETWORK (TRN) OF YEAST

- Directed graph
- Downloaded from the Yeast Search for Transcriptional Regulators And Consensus Tracking (YEASTRACT)
- Consists of 48,082 interactions between 183 transcription factors (TF) and 6,403 target genes (TG).

# RAPAMYCIN-TREATMENT DATASET



Adopted from Fournier et al., 2010

- **Rapamycin:** A lipophilic macrolide that directly binds to and inhibits TOR *in vivo*
- Temporal analysis of gene expression changes for 6,000 ORFs in Baker's yeast, *Saccharomyces cerevisiae*, over 6h of rapamycin treatment.
- 366 repressed and 291 induced genes at a minimum threshold of 2-fold change.



# DIRECTIONAL INFORMATION FLOW

## RANDOM WALK

### DEFINITION

**Random walk** on a graph  $G$ , initiated from vertex  $v$ , is the sequence of transitions among vertices, starting from  $v$ . At each step, the random walker randomly chooses the next vertex from among the neighbors of the current node.

It is a Markov chain with the transition matrix  $P$ , where  $p_{ij} = \text{Prob}(S_{n+1} = v_i | S_n = v_j)$  and random variable  $S_n$  represents the state of the random walk at the time step  $n$ .

# DIRECTIONAL INFORMATION FLOW

## RANDOM WALK WITH RESTART

### DEFINITION

**Random walk with restart (RWR)** is a modified Markov chain in which, at each step, a random walker has the choice of either continuing along its path, with probability  $\alpha$ , or jump (teleport) back to the initial vertex, with probability  $1 - \alpha$ .

The transition matrix of the modified chain,  $M$ , can be computed as  $M = \alpha P + (1 - \alpha) \mathbf{e}_v \mathbf{1}^T$ , where  $\mathbf{e}_v$  is a stochastic vector of size  $n$  having zeros everywhere, except at index  $v$ , and  $\mathbf{1}$  is a vector of all ones.

# DIRECTIONAL INFORMATION FLOW

## STATIONARY DISTRIBUTION

The portion of time spent on each node in an infinite random walk with restart initiated at node  $v$ , with parameter  $\alpha$ .

### DEFINITION

**Stationary distribution** of the modified chain

$$\begin{aligned}\pi_v(\alpha) &= M\pi_v(\alpha) \\ &= (\alpha P + (1 - \alpha)\mathbf{e}_v\mathbf{1}^T)\pi_v(\alpha)\end{aligned}$$

Enforcing a unit norm on the dominant eigenvector to ensure its stochastic property,  $\|\pi_v(\alpha)\|_1 = \mathbf{1}^T \pi_v = 1$ , we will have:

# DIRECTIONAL INFORMATION FLOW

## STATIONARY DISTRIBUTION—CONTINUE

### DEFINITION

**Iterative form** of the information flow process:

$$\pi_v(\alpha) = \alpha P \pi_v(\alpha) + (1 - \alpha) \mathbf{e}_v,$$

### DEFINITION

**Explicit (direct) formulation** of the information flow process:

$$\pi_v(\alpha) = \underbrace{(1 - \alpha)(I - \alpha P)^{-1}}_Q \mathbf{e}_v,$$

# DIRECTIONAL INFORMATION FLOW

## INTERPRETATION

### DEFINITION

Expansion using the Neumann series:

$$\pi_v(\alpha) = (1 - \alpha) \sum_{i=0}^{\infty} (\alpha P)^i \mathbf{e}_v$$

Thus,  $\pi_v(\alpha)$  is a function of:

- Distance to source node ( $v$ )
- Multiplicity of paths

## SIDEBAR: FUNCTIONAL PAGERANK (PR)

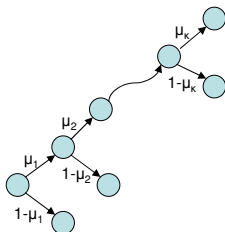
### Computing PageRank (PR)

- PageRank as a *random surfer process*: Start surfing from a random node and keep following links with probability  $\mu$  restarting with probability  $1 - \mu$ ; the node for restarting will be selected based on a personalization vector  $v$ . The ranking value  $x_i$  of a node  $i$  is the probability of visiting this node during surfing.
- PR can also be cast in power series representation as  $x = (1 - \mu) \sum_{j=0}^k \mu^j S^j v$ ;  $S$  encodes column-stochastic adjacencies.

### Functional rankings

- A general method to assign ranking values to graph nodes as  $x = \sum_{j=0}^k \zeta_j S^j v$ . PR is a functional ranking,  $\zeta_j = (1 - \mu)\mu^j$ .
- Terms attenuated by outdegrees in  $S$  and damping coefficient  $\zeta_j$ .

# FUNCTIONAL RANKINGS THROUGH MULTIDAMPING [KOLLIAS, GALLOPOULOS, AG, TKDE'13]



## COMPUTING $\mu_j$ IN MULTIDAMPING

Simulate a functional ranking by random surfers following emanating links with probability  $\mu_j$  at step  $j$  given by :

$$\mu_j = 1 - \frac{1}{1 + \frac{\rho_{k-j+1}}{1 - \mu_{j-1}}}, j = 1, \dots, k,$$

where  $\mu_0 = 0$  and  $\rho_{k-j+1} = \frac{\zeta_{k-j+1}}{\zeta_{k-j}}$

## Examples

*LinearRank (LR)*  $x^{\text{LR}} = \sum_{j=0}^k \frac{2(k+1-j)}{(k+1)(k+2)} S^j v : \mu_j = \frac{j}{j+2}, j = 1, \dots, k.$

*TotalRank (TR)*  $x^{\text{TR}} = \sum_{j=0}^{\infty} \frac{1}{(j+1)(j+2)} S^j v : \mu_j = \frac{k-j+1}{k-j+2}, j = 1, \dots, k.$

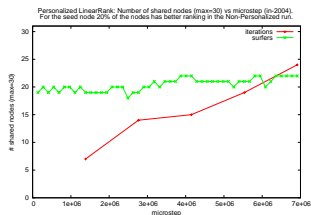
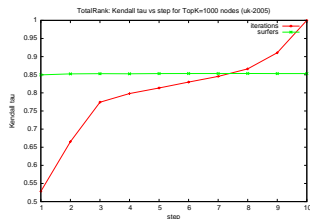
# MULTIDAMPING AND COMPUTATIONAL COST

## Advantages of multidamping

- Interpretability and Design!
- Reduced computational cost in *approximating* functional rankings using the Monte Carlo approach. A random surfer terminates with probability  $1 - \mu_j$  at step  $j$ .
- Inherently parallel and synchronization free computation.



# MULTIDAMPING PERFORMANCE



**Approximate ranking:** Run  $n$  surfers to completion for graph size  $n$ . How well does the computed ranking capture the “reference” ordering for  $\text{top-}k$  nodes, compared to standard iterations of equivalent computational cost/number of operations? [Left]

**Approximate personalized ranking:** Run less than  $n$  surfers to completion (each called a microstep, x-axis), from a selected node (personalized). How well can we capture the “reference”  $\text{top-}k$  nodes, i.e., how many of them are shared (y-axis), compared to the

# OUTLINE

## 1 PART 1: MOTIVATION – CONSTRUCTING AGING PATHWAYS IN YEAST

- Overview
- Materials and Methods
  - Datasets
  - Tracing Information Flow

## ● Results and Discussion

## 2 PART 2: TISSUE SPECIFIC NETWORKS AND COMPARATIVE ANALYSIS

- Motivation
- Materials and methods
  - Datasets
  - Algorithmic contributions
- Results and discussion
  - Core alignment graph of housekeeping genes
  - Computing similarity of human tissues with yeast

## EXPERIMENTAL SETTINGS

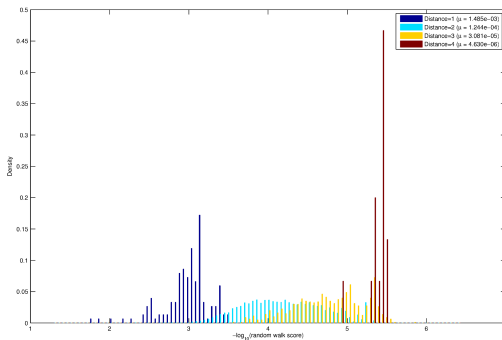
We set the **preference vector** as:

$$e_S(i) = \begin{cases} \frac{1}{|S|} & \text{if } v_i \in S, \\ 0 & \text{O.W.} \end{cases}$$

for  $S$  being the subset of vertices in the yeast interactome corresponding to members of the TORC1 protein complex. The diameter of the network is computed to be 6 and  **$\alpha$  parameter** is set to  $\frac{d}{d+1} = \frac{6}{7} \sim 0.85$  accordingly to give all nodes a fair chance of being visited.

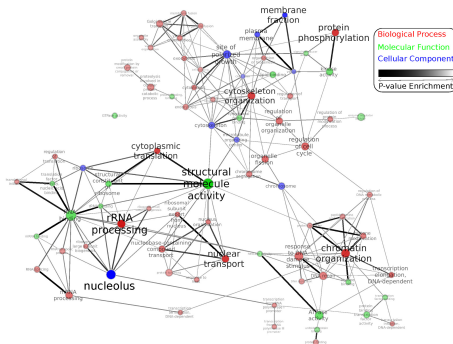
## DISTRIBUTION OF INFORMATION FLOW SCORES

Distribution of information flow scores across nodes with similar distance from members of TORC1 are color coded accordingly. The  $\mu$  parameter is the average of information flow scores for nodes under each distribution.



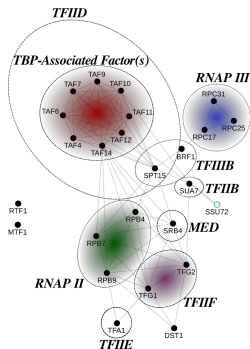
# ENRICHMENT MAP OF YEAST GOSLIM TERMS

Enriched terms are identified by mHG p-value, computed for the ranked-list of genes based on their information flow scores. Each node represents a significant GO term and edges represent the overlap between genesets of GO terms. Terms in different branches of GO are color-coded with red, green, and blue. Color intensity of each node represents the significance of its p-value, while the node size illustrates the size of its geneset. Thickness of edges is related to the extent of overlap among genesets.



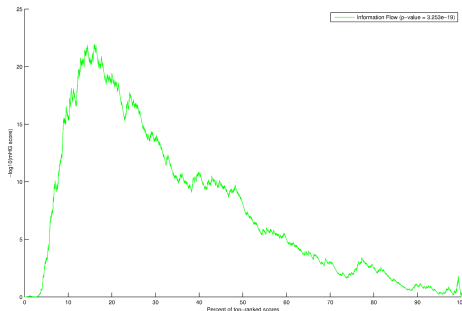
# TOR-DEPENDENT CONTROL OF TRANSCRIPTION INITIATION

Induced subgraph in the yeast interactome, constructed from the top-ranked genes in the information flow analysis that are annotated with the transcription initiation GO term. Different functional subunits are marked and color-coded appropriately.



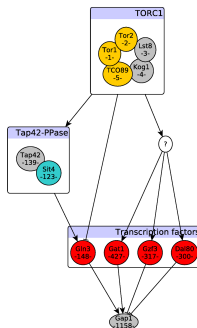
# ENRICHMENT PLOT FOR RAPAMYCIN-TREATMENT DATASET

Enrichment score as a function of the score percentage. Computations are based on the set of differentially expressed genes in response to Rapamycin treatment. The peak of plot occurs at around top 15% of scores, resulting in the minimum hypergeometric (mHG) score of  $\sim 1e - 22$ . The exact p-value for this score is computed, using dynamic programming, to be  $3.3e - 19$ .



# TORC1-DEPENDENT REGULATION OF GAP1

The schematic diagram is based on literature evidence for the interactions. Each node in the signaling pathway is annotated with the rank of its information flow score from TORC1. Ranking of nodes based on their information flow scores respect our prior knowledge on the structure of this pathway.





# OUTLINE

## 1 PART 1: MOTIVATION – CONSTRUCTING AGING PATHWAYS IN YEAST

- Overview
- Materials and Methods
  - Datasets
  - Tracing Information Flow
- Results and Discussion

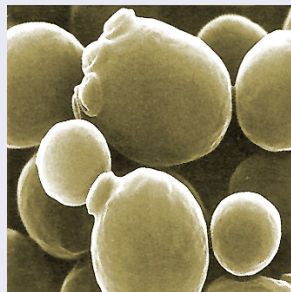
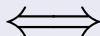
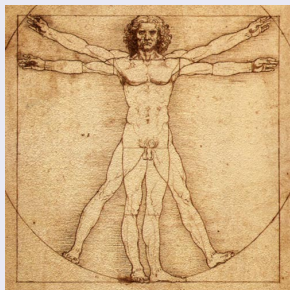
## 2 PART 2: TISSUE SPECIFIC NETWORKS AND COMPARATIVE ANALYSIS

- Motivation
- Materials and methods
  - Datasets
  - Algorithmic contributions
- Results and discussion
  - Core alignment graph of housekeeping genes
  - Computing similarity of human tissues with yeast

# COMPARATIVE NETWORK ANALYSIS

## TRADITIONAL APPROACH

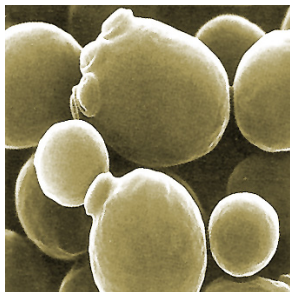
### KEY CHALLENGE



To project functional pathways from a well-studied organism, such as yeast, back to a higher-order organism, such as humans.

# YEAST AS A KEY MODEL ORGANISM

## SIMPLE YET POWERFUL



"... yeast has graduated from a position as the premier model for eukaryotic cell biology to become the pioneer organism that facilitated the establishment of entirely new fields of study called *functional genomics* and *systems biology*." – D. Botstein and G. Fink (2011).

# YEAST AS A KEY MODEL ORGANISM

## WHY YEAST?

- Rapid growth and ease of manipulation
- Mature genetic and molecular toolbox, including deletion mutants, over-expression libraries, and green fluorescent protein (GFP)-tagged yeast strains
- Multitude of high-throughput datasets, ranging from genetic arrays, transcriptome, proteome, and metabolome profiles
- Saccharomyces Genome Database (SGD)



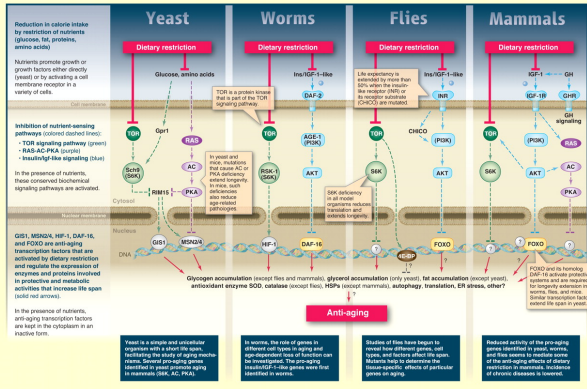
## CONSERVED PATHWAYS BETWEEN YEAST AND HIGHER-ORDER ORGANISMS

Many of the underlying functionalities and associated machineries are shared with higher eukaryotes:

- Cell cycle
- Programmed cell death
- Protein folding, quality control, and degradation
- Signaling pathways, such as MAPK, TOR, and insulin/IGF-I
- Aging and CR-mediated pathways
  - **Chronological:** amount of time cells survive in post-mitotic state
  - **Replicative:** number of times a cell can divide before senescence occurs.

# CONSERVED PATHWAYS CONTINUED

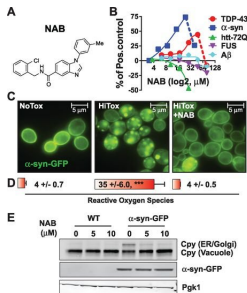
## Conserved Nutrient Signaling Pathways Regulating Longevity



Fontana *et al*, Science (2010)

# YEAST AS A MODEL ORGANISM FOR HUMAN DISEASE

## RECENT SUCCESS STORIES

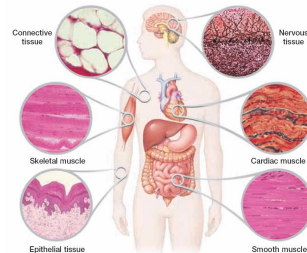


Adopted from D. Tardiff *et al.* (2013)

- Heterologous expression of disease gene(s)
- Yeast as an unbiased phenotypic screen
- N-aryl benzimidazole (NAB) strongly protects cells from  $\alpha$ -synuclein toxicity in the humanized yeast model
- Validated this discovery using iPS cell from Parkinson's patients with  $\alpha$ -Syn mutation

## PROBLEM STATEMENT

For which tissues  
is yeast a good  
model organism?



What are the  
shared/ missing  
functional  
components in  
yeast, compared  
to human tissues?

Different human tissues, while inheriting a similar genetic code, exhibit unique anatomical and physiological properties.



# OUTLINE

## 1 PART 1: MOTIVATION – CONSTRUCTING AGING PATHWAYS IN YEAST

- Overview
- Materials and Methods
  - Datasets
  - Tracing Information Flow
- Results and Discussion

## 2 PART 2: TISSUE SPECIFIC NETWORKS AND COMPARATIVE ANALYSIS

- Motivation
- Materials and methods
  - Datasets
  - Algorithmic contributions
- Results and discussion
  - Core alignment graph of housekeeping genes
  - Computing similarity of human tissues with yeast

# TISSUE-SPECIFIC GENE EXPRESSION

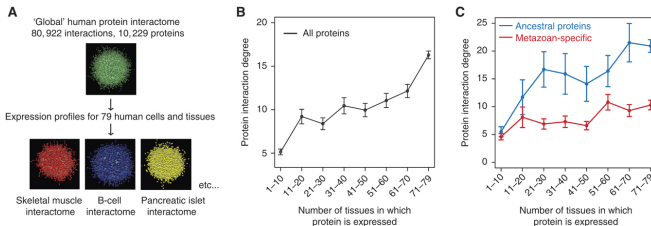
The GNF Gene Atlas dataset:



- 79 different tissues
- 44,775 human transcripts
- Platforms:
  1. Affymetrix HG-U133A.
  2. Custom GNF1H array.

## TISSUE-SPECIFIC INTERACTOMES

- Vertex-induced subgraphs of the global human interactome
- Based on the GNF Gene Atlas dataset
  - ⇒ A gene is considered as present in a tissue, if its normalized expression level is  $> 200$  (average difference between match-mismatch pairs).



Adopted from Bossi et al., 2009

## SEQUENCE SIMILARITY OF PROTEIN PAIRS

- Protein sequences are downloaded from Ensembl database, release 69.
- Reference genomes:
  - ▷ **Human:** GRCh37
  - ▷ **Yeast:** EF4
- Number of protein sequences:
  - ▷ **Human:** 101,075
  - ▷ **Yeast:** 6,692
- Low-complexity regions are masked using **pseg**
- Smith-Waterman algorithm is used to compute local sequence alignments.

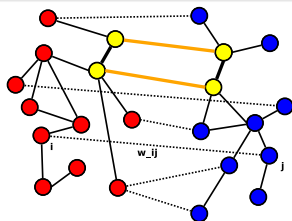
# SPARSE NETWORK ALIGNMENT

**Integer Quadratic Program**— Approximated using Belief Propagation:

$$\begin{aligned} \max_{\mathbf{x}} \quad & (\alpha \mathbf{w}^T \mathbf{x} + \frac{\beta}{2} \mathbf{x}^T \mathbf{S} \mathbf{x}) \\ \text{Subject to:} \quad & \begin{cases} \mathbf{C} \mathbf{x} \leq \mathbf{1}_{n_G * n_H} & \text{Matching constraints;} \\ x_{ij'} \in \{0, 1\}, & \text{Integer constraint.} \end{cases} \end{aligned}$$

- $\mathbf{x}$ : Matching vector
- $L$ : Bipartite graph of similarities between pair of proteins in input networks
- $\mathbf{w}$ : Edge-weights in the graph  $L$  (based on sequence similarities)
- $\mathbf{S}$ : Matrix encoding conserved edges in the product graph ( $G \otimes H$ )
- $\mathbf{C}$ : Incidence matrix of graph  $L$

## SIDEBAR: NETWORK ALIGNMENT



- **Node similarity:** Two nodes are similar if they are linked by other similar node pairs. By pairing similar nodes, the two graphs become *aligned*.

- Let  $\tilde{A}$  and  $\tilde{B}$  be the normalized adjacency matrices of the graphs (normalized by columns),  $H_{ij}$  be the independently known similarity scores (preferences matrix) of nodes  $i \in V_B$  and  $j \in V_A$ , and  $\mu$  be the fractional contribution of topological similarity.
- To compute  $X$ , IsoRank iterates:

$$X \leftarrow \mu \tilde{B} X \tilde{A}^T + (1 - \mu) H$$

# NETWORK SIMILARITY DECOMPOSITION (NSD) [KOLLIAS, MOHAMMADI, AG, TKDE'12]

## Network Similarity Decomposition (NSD)

- In  $n$  steps of we reach

$$X^{(n)} = (1 - \mu) \sum_{k=0}^{n-1} \mu^k \tilde{B}^k H (\tilde{A}^T)^k + \mu^n \tilde{B}^n H (\tilde{A}^T)^n$$

- Assume that  $H = uv^T$  (1 component). Two phases for  $X$ :

1.  $u^{(k)} = \tilde{B}^k u$  and  $v^{(k)} = \tilde{A}^k v$  (*preprocess/compute iterates*)
2.  $X^{(n)} = (1 - \mu) \sum_{k=0}^{n-1} \mu^k u^{(k)} v^{(k)T} + \mu^n u^{(n)} v^{(n)T}$  (*construct  $X$* )

This idea extends to  $s$  components,  $H \sim \sum_{i=1}^s w_i z_i^T$ .

- NSD computes matrix-vector iterates and builds  $X$  as a sum of outer products; these are much cheaper than triple matrix products.

We can then apply Primal-Dual or Greedy Matching (1/2 approximation) to extract the actual node pairs.

# NSD: PERFORMANCE [KOLLIAS, MADAN, MOHAMMADI, AG, BMC RN'12]

| Species           | Nodes | Edges |
|-------------------|-------|-------|
| celeg (worm)      | 2805  | 4572  |
| dmela (fly)       | 7518  | 25830 |
| ecoli (bacterium) | 1821  | 6849  |
| hpylo (bacterium) | 706   | 1414  |
| hsapi (human)     | 9633  | 36386 |
| mmusc (mouse)     | 290   | 254   |
| scere (yeast)     | 5499  | 31898 |

| Species pair | NSD (secs)  | PDM (secs) | GM (secs) | IsoRank (secs) |
|--------------|-------------|------------|-----------|----------------|
| celeg-dmela  | <b>3.15</b> | 152.12     | 7.29      | 783.48         |
| celeg-hsapi  | <b>3.28</b> | 163.05     | 9.54      | 1209.28        |
| celeg-scere  | <b>1.97</b> | 127.70     | 4.16      | 949.58         |
| dmela-ecoli  | <b>1.86</b> | 86.80      | 4.78      | 807.93         |
| dmela-hsapi  | <b>8.61</b> | 590.16     | 28.10     | 7840.00        |
| dmela-scere  | <b>4.79</b> | 182.91     | 12.97     | 4905.00        |
| ecoli-hsapi  | <b>2.41</b> | 79.23      | 4.76      | 2029.56        |
| ecoli-scere  | <b>1.49</b> | 69.88      | 2.60      | 1264.24        |
| hsapi-scere  | <b>6.09</b> | 181.17     | 15.56     | 6714.00        |

- We compute similarity matrices  $X$  for various pairs of species using Protein-Protein Interaction (PPI) networks.  $\mu = 0.80$ , uniform initial conditions (outer product of suitably normalized 1's for each pair), 20 iterations, one component.
- We then extract node matches using PDM and GM.
- *Three orders of magnitude speedup* from NSD-based approaches compared to IsoRank.



## NSD: PARALLELIZATION [KKG JPDC'14 (TO APPEAR)]

**Parallelization:** NSD has been ported to parallel and distributed platforms.

- We have aligned up to million-node graph instances using over 3K cores.
- We process graph pairs of over a billion nodes and twenty billion edges each (!), on MapReduce-based distributed platforms.

# RANDOM MODEL FOR TISSUE-SPECIFIC NETWORKS

## DEFINITION

- **Global human interactome:** All potential interactions between human proteins, represented by graph  $G = (V_G, E_G)$
- **Tissue-specific network(s):** Vertex-induced subgraph(s) of the Global human interactome, represented by  $G_T = (V_T, E_T)$  with  $n_T = |V_T|$ ,  $V_T \subset V_G$ , and  $E_T \subset E_G$
- **Universal genes:** Ubiquitously expressed subset of human genes corresponding to housekeeping functions, represented by  $V_U \subset V_G$ , and  $n_U = |V_U|$
- **Random tissue-specific network(s):** Vertex-induced subgraphs of  $G$ , constructed from  $V_{\mathcal{R}} = V_U \cup V_S$ , with  $V_S$  being random set of vertices of size  $n_T - n_U$  selected from  $V_G \setminus V_U$

## SIGNIFICANCE OF NETWORK ALIGNMENT(S)

### DEFINITION

- **Original alignment:**  $\mathcal{W} = \mathbf{w}^T \mathbf{x}$ ,  $\mathcal{O} = \frac{1}{2} \mathbf{x}^T \mathbf{S} \mathbf{x}$
- **Monte-Carlo simulation:** Let  $\mathcal{W}_{\mathcal{R}}$  and  $\mathcal{O}_{\mathcal{R}}$  be the random vectors representing the weight and overlap of aligning  $k_{\mathcal{R}}$  random tissue-specific networks with yeast
- **Positive/Negative cases:**  $k_P$  is the number of random cases with both  $\mathcal{W}_{\mathcal{R}} \leq \mathcal{W}$  and  $\mathcal{O}_{\mathcal{R}} \leq \mathcal{O}$ .  $k_N$  is defined as the size of complement set.
- **p-value** bounds:

$$\delta_{\mathcal{R}} = \frac{k_P}{k_{\mathcal{R}}} \leq \text{alignment p-value} \leq 1 - \frac{k_N}{k_{\mathcal{R}}} = \Delta_{\mathcal{R}}$$

- **Alignment p-value:**

$$p - \text{value} = \text{Prob}(\alpha * \mathcal{O} + \beta * \mathcal{W} \leq \mathcal{O} \mathcal{W}_{\mathcal{R}})$$

# PARTITIONING HUMAN GENES BASED ON THEIR EXPRESSION SELECTIVITY

## DEFINITION

**Selectivity  $p$ -value**– Given a cluster of homogenous tissues:

$$\begin{aligned} p\text{-value}(X = c_n) &= \text{Prob}(c_n \leq X) \\ &= \text{HGT}(c_n | N, n, c_N) \\ &= \sum_{x=c_n}^{\min(c_N, n)} \frac{C(c_N, x) C(N - c_N, n - x)}{C(N, n)} \end{aligned}$$

$N$ : total number of tissues,  $n$ : number of tissues in the cluster,  $c_N$ : number of tissues in which a given gene is expressed,  $c_n$ : number of tissue in the cluster that the given gene is expressed.

# HUMAN-SPECIFIC OR CONSERVED?

## DEFINITION

Classification of human tissue-selective genes:

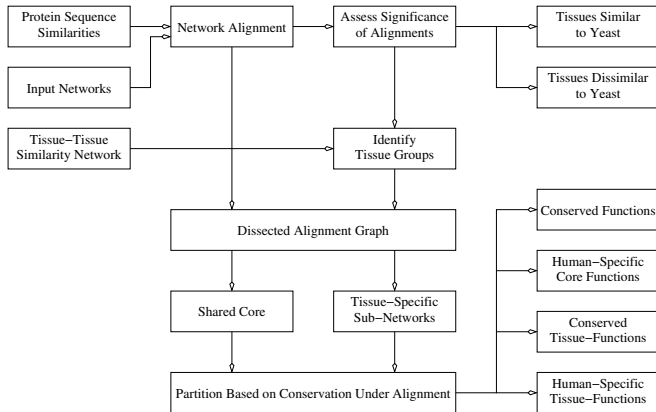
- **Conserved:** Subset of tissue-selective genes that are consistently aligned in the "majority" of aligned tissues in the given group
- **Human-specific:** Subset of tissue-selective genes that are consistently unaligned in the "majority" of tissues in the given group
- **Unclassified:** None of the above

## DEFINITION

Majority voting:

- **Alignment consistency table:** Yeast partner of each tissue-selective gene in the given cluster of tissues
- **Consensus rate:** Minimum percentage of tissues (columns) in each row of the alignment consistency table that have to agree to make a decision about conserved/human-specificity

# SUMMARY



Input

Processing

Output

# OUTLINE

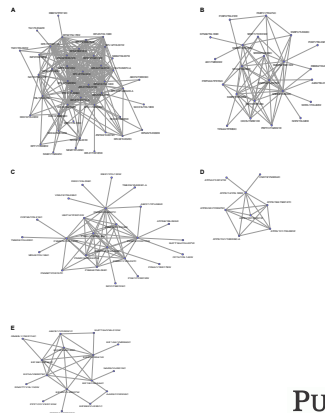
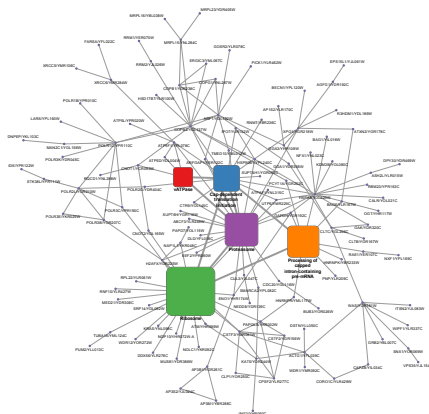
## 1 PART 1: MOTIVATION – CONSTRUCTING AGING PATHWAYS IN YEAST

- Overview
- Materials and Methods
  - Datasets
  - Tracing Information Flow
- Results and Discussion

## 2 PART 2: TISSUE SPECIFIC NETWORKS AND COMPARATIVE ANALYSIS

- Motivation
- Materials and methods
  - Datasets
  - Algorithmic contributions
- Results and discussion
  - Core alignment graph of housekeeping genes
  - Computing similarity of human tissues with yeast

# CORE GENES– THE MOST CONSERVED SUBSET OF HOUSEKEEPING GENES





# FUNCTIONAL ENRICHMENT OF HK GENES

## CORE SUBSET

- Ribosome biogenesis
- Translation
- Protein targeting
- RNA splicing
- mRNA surveillance

# FUNCTIONAL ENRICHMENT OF HK GENES

## HUMAN-SPECIFIC SUBSET

- Anatomical structure development
- Paracrine signaling
- NADH dehydrogenase (mitochondrial Complex I)

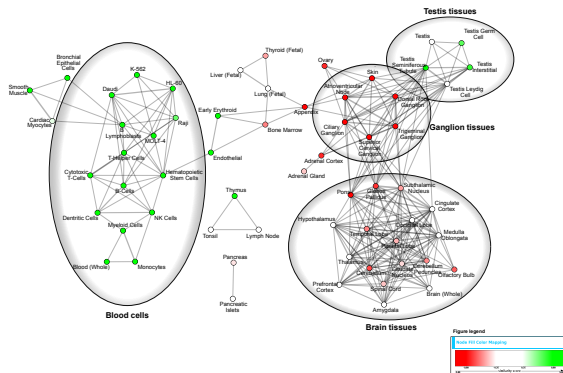
# THE MOST SIMILAR TISSUES TO YEAST

| Name                       | pval lower bound | overall pval | pval upper bound | confidence |
|----------------------------|------------------|--------------|------------------|------------|
| Myeloid Cells              | < 1.00e-04       | < 1.00e-04   | < 1.00e-04       | 1          |
| Monocytes                  | < 1.00e-04       | < 1.00e-04   | < 1.00e-04       | 1          |
| Dendritic Cells            | < 1.00e-04       | < 1.00e-04   | < 1.00e-04       | 1          |
| NK Cells                   | < 1.00e-04       | < 1.00e-04   | < 1.00e-04       | 1          |
| T-Helper Cells             | < 1.00e-04       | < 1.00e-04   | < 1.00e-04       | 1          |
| Cytotoxic T-Cells          | < 1.00e-04       | < 1.00e-04   | < 1.00e-04       | 1          |
| B-Cells                    | < 1.00e-04       | < 1.00e-04   | < 1.00e-04       | 1          |
| Endothelial                | < 1.00e-04       | < 1.00e-04   | < 1.00e-04       | 1          |
| Hematopoietic Stem Cells   | < 1.00e-04       | < 1.00e-04   | < 1.00e-04       | 1          |
| MOLT-4                     | < 1.00e-04       | < 1.00e-04   | < 1.00e-04       | 1          |
| B Lymphoblasts             | < 1.00e-04       | < 1.00e-04   | < 1.00e-04       | 1          |
| HL-60                      | < 1.00e-04       | < 1.00e-04   | < 1.00e-04       | 1          |
| K-562                      | < 1.00e-04       | < 1.00e-04   | < 1.00e-04       | 1          |
| Early Erythroid            | < 1.00e-04       | < 1.00e-04   | < 1.00e-04       | 1          |
| Bronchial Epithelial Cells | < 1.00e-04       | < 1.00e-04   | 0.0002           | 0.9998     |
| Colorectal Adenocarcinoma  | < 1.00e-04       | < 1.00e-04   | 0.0004           | 0.9996     |
| Daudi                      | < 1.00e-04       | < 1.00e-04   | 0.0009           | 0.9991     |
| Testis Seminiferous Tubule | < 1.00e-04       | < 1.00e-04   | 0.0012           | 0.9988     |
| Smooth Muscle              | < 1.00e-04       | < 1.00e-04   | 0.0016           | 0.9984     |
| Blood (Whole)              | < 1.00e-04       | < 1.00e-04   | 0.0053           | 0.9947     |
| Thymus                     | < 1.00e-04       | 0.0001       | 0.0062           | 0.9938     |
| Testis Interstitial        | < 1.00e-04       | 0.0004       | 0.0086           | 0.9914     |

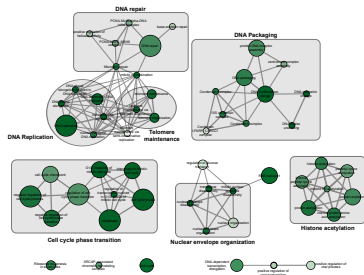
# THE LEAST SIMILAR TISSUES TO YEAST

| Name                       | pval lower bound | overall pval | pval upper bound | confidence |
|----------------------------|------------------|--------------|------------------|------------|
| Trigeminal Ganglion        | 0.9947           | 0.9994       | 1                | 0.9947     |
| Superior Cervical Ganglion | 0.9847           | 0.9991       | 1                | 0.9847     |
| Ciliary Ganglion           | 0.9407           | 0.9813       | 0.9964           | 0.9443     |
| Atrioventricular Node      | 0.8746           | 0.9792       | 0.9921           | 0.8825     |
| Skin                       | 0.8355           | 0.9297       | 0.9809           | 0.8546     |
| Heart                      | 0.7934           | 0.9585       | 0.9815           | 0.8119     |
| Appendix                   | 0.7596           | 0.9371       | 0.973            | 0.7866     |
| Dorsal Root Ganglion       | 0.7065           | 0.933        | 0.9717           | 0.7348     |
| Skeletal Muscle            | 0.3994           | 0.5902       | 0.7866           | 0.6128     |
| Uterus Corpus              | 0.233            | 0.7736       | 0.8769           | 0.3561     |
| Lung                       | 0.0771           | 0.3853       | 0.5544           | 0.5227     |
| Pons                       | 0.0674           | 0.5201       | 0.6983           | 0.3691     |
| Salivary Gland             | 0.0639           | 0.3449       | 0.5173           | 0.5466     |
| Liver                      | 0.0600           | 0.6857       | 0.8519           | 0.2081     |
| Ovary                      | 0.0388           | 0.2735       | 0.4481           | 0.5907     |
| Trachea                    | 0.0259           | 0.2376       | 0.4146           | 0.6113     |
| Globus Pallidus            | 0.0206           | 0.2471       | 0.4336           | 0.587      |
| Cerebellum                 | 0.0127           | 0.1950       | 0.3783           | 0.6344     |

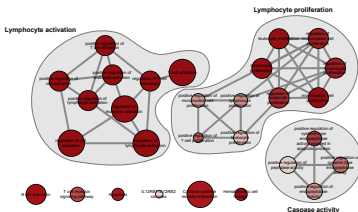
# TISSUE-TISSUE SIMILARITY NETWORK



# BLOOD CELLS



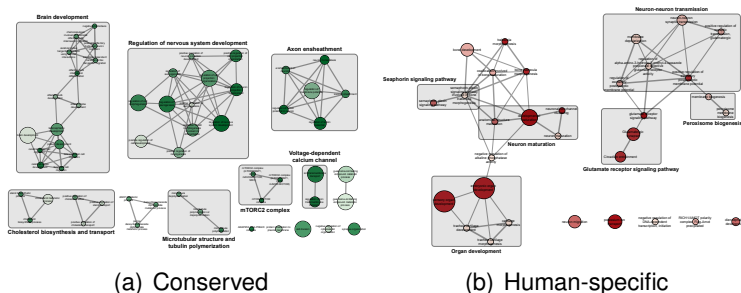
(a) Conserved



(b) Human-specific

**FIGURE :** Enrichment map of unique blood-selective functions.

# BRAIN TISSUES



**FIGURE : Enrichment map of unique brain-selective functions.**

## ENRICHED DISEASE CLASSES

|               | Conserved genes |                  | Human-specific genes |                  |
|---------------|-----------------|------------------|----------------------|------------------|
|               | Disease class   | <i>p</i> -value  | Disease class        | <i>p</i> -value  |
| Blood cells   | Cancer          | $2.85 * 10^{-3}$ | Immune               | $1.88 * 10^{-9}$ |
|               |                 |                  | Infection            | $1.00 * 10^{-2}$ |
| Brain tissues | Psych           | $3.59 * 10^{-4}$ | Psych                | $5.70 * 10^{-8}$ |
|               | Chemdependency  | $2.60 * 10^{-3}$ | Neurological         | $2.97 * 10^{-2}$ |
|               | Pharmacogenomic | $9.74 * 10^{-2}$ |                      |                  |



# COMPARATIVE ANALYSIS OF BRAIN-SPECIFIC PATHOLOGIES

| Disorder  | Conserved genes | Human-specific genes |
|---|-----------------|----------------------|
| schizophrenia   | 0.008573        | 8.4905E-06           |
| autism  | 0.048288        | 0.00077448           |
| dementia  | 0.0014356       | -                    |
| schizophrenia; schizoaffective disorder; bipolar disorder   | -               | 0.0021433            |
| myocardial infarct; cholesterol, HDL; triglycerides; atherosclerosis, coronary; macular degeneration; colorectal cancer | 0.0051617       | -                    |
| epilepsy  | 0.071562        | 0.0064716            |
| seizures  | -               | 0.020381             |
| bipolar disorder  | 0.048288        | 0.022016             |
| attention deficit disorder conduct disorder oppositional defiant disorder   | 0.032444        | 0.023865             |

## CONCLUDING REMARKS

- Tissue-specific networks and interactions emerging as important datasets for understanding and mitigating pathologies.
- This work represents among the first computational investigations into tissue-specific networks.
- Our results reveal a significantly refined understanding of tissue-specific processes and their functions.
- Tissue-specific networks show significant enrichment of a number of important diseases, identifying a number of drug targets.
- Comparative analysis with yeast sheds light on humanized yeast models and their use in identifying targets.

## ACKNOWLEDGEMENTS.

- Thanks to the National Science Foundation, the Center for Science of Information, and the Department of Energy.
- Thank you!