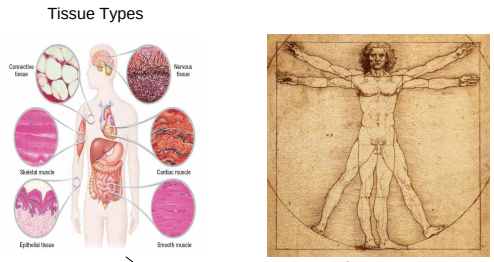
A Systems Study of Aging and Age-Related Pathologies

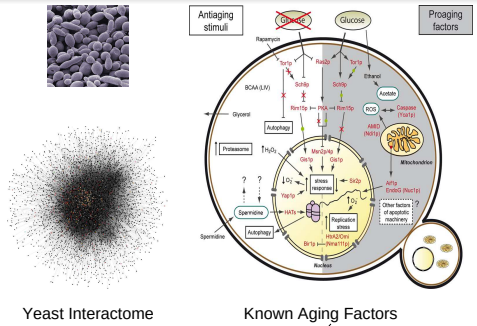
Ananth Grama

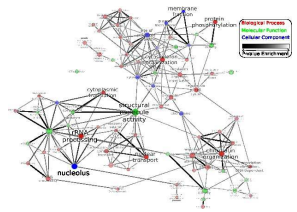
Center for Science of Information
Purdue University

EAC, Chicago 2013

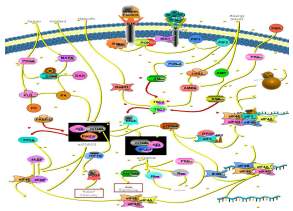
PURDUE UNIVERSITY

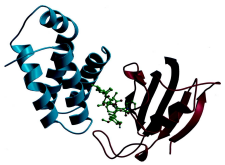






Network Alignment





Predicting New Targets

Novel Therepies, Interventions and Markers for Degenerative Diseases

PURDUE UNIVERSITY

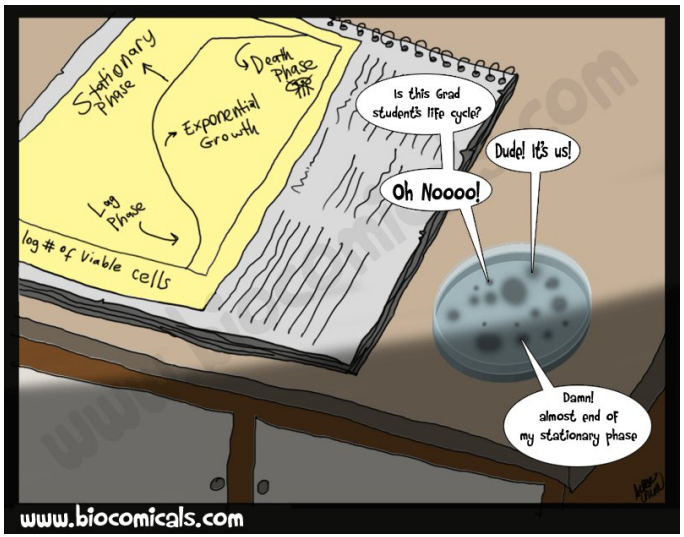
Outline

PURDUE UNIVERSITY

Outline

PURDUE UNIVERSITY

Yeast Aging



Courtesy of Alper Uzan, PhD.

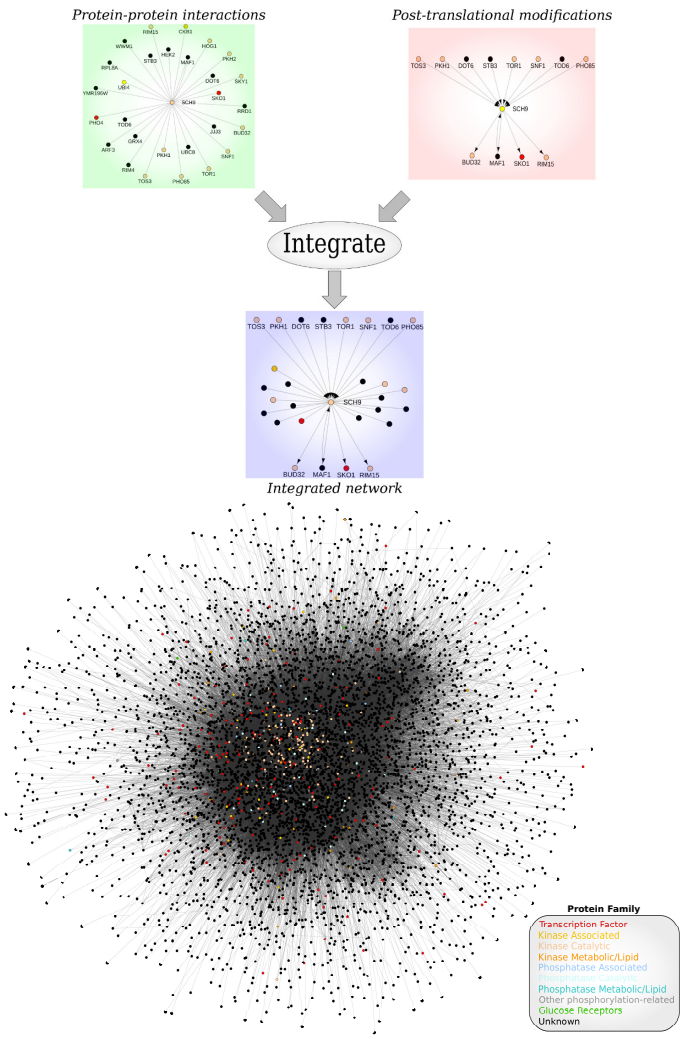
- Yeast as a model organism for aging research:
  - ✓ Rapid growth
  - ✓ Ease of manipulation
- Replicative life-span (RLS): the number of buds a mother cell can produce before senescence occurs
- Chronological life-span (CLS): duration of viability after entering the stationary-phase

PURDUE UNIVERSITY

Outline

PURDUE UNIVERSITY

YEAST INTERACTOME



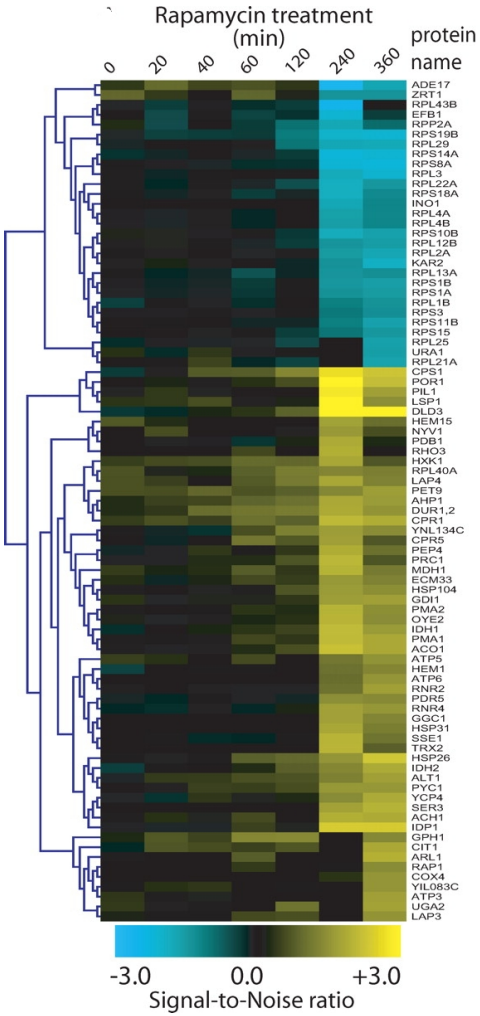
- *Mixed network:* Contains both directed (biochemical activities) and undirected (protein-protein interactions) edges
- 103,619 (63,395 non-redundant) physical interactions among 5,691 proteins.
- 5,791 (5,443 non-redundant) biochemical activities (mostly phosphorylation events) among 2,002 kinase-substrate pairs.

PURDUE UNIVERSITY

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.

PURDUE UNIVERSITY

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 $p_{ij} = Prob(S_{n+1} = v_i | S_n = v_j)$ and random variable S_n represents the state of the random walk at the time step n.

PURDUE UNIVERSITY

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 α, or jump (teleport) back to the initial vertex, with probability $1 - \alpha$.

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

PURDUE UNIVERSITY

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 α.

Definition

Stationary distribution of the modified chain

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

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

PURDUE UNIVERSITY

Directional information flow

Stationary distribution–continue

Definition

Iterative form of the information flow process:

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

Definition

Explicit (direct) formulation of the information flow process:

$$\pi_v(\alpha) = \underbrace{(1-\alpha)(I-\alpha P)^{-1}}_{Q}\boldsymbol{e}_v,$$

PURDUE UNIVERSITY

Directional information flow

Interpretation

Definition

Expansion using the Neumann series:

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

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

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

PURDUE UNIVERSITY

SIDEBAR: FUNCTIONAL PAGERANK (PR)

Computing PageRank (PR)

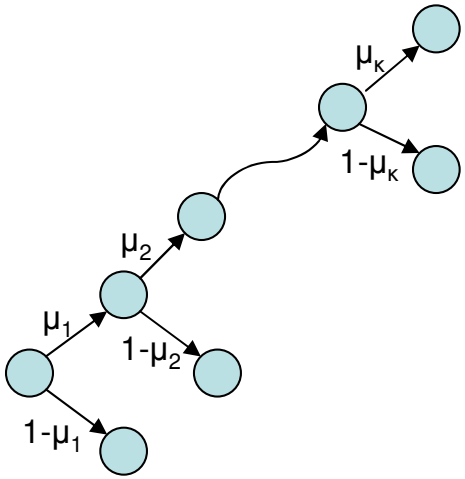
- PageRank as a *random surfer process*: Start surfing from a random node and keep following links with probability μ restarting with probability $1 - \mu$; the node for restarting will be selected based on a personalization vector v. The ranking value x_i of a node i is the probability of visiting this 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 coefficients ζ_j.

PURDUE UNIVERSITY

Functional Rankings Through Multidamping [Kollias, Gallopoulos, AG, TKDE'13]



Computing μ_j in Multidamping

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

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

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

Examples

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

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

PURDUE UNIVERSITY

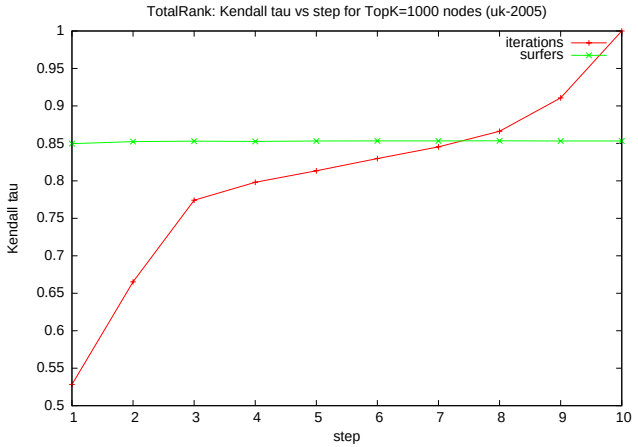
Multidamping and Computational Cost

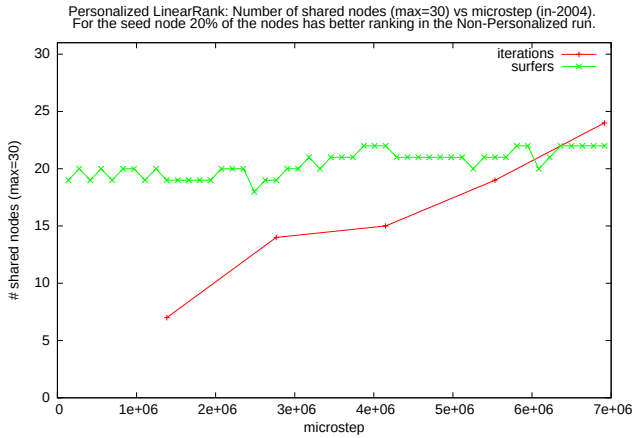
Advantages of multidamping

- 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.

PURDUE UNIVERSITY

MULTIDAMPING PERFORMANCE





Approximate ranking: Run n surfers to completion for graph size n. How well does the computed ranking capture the "reference" ordering for `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" `top-`k nodes, i.e., how many of them are shared (y-axis), compared to the simple approach? *[Right]*

PURDUE UNIVERSITY

Outline

PURDUE UNIVERSITY

Experimental settings

We set the preference vector as:

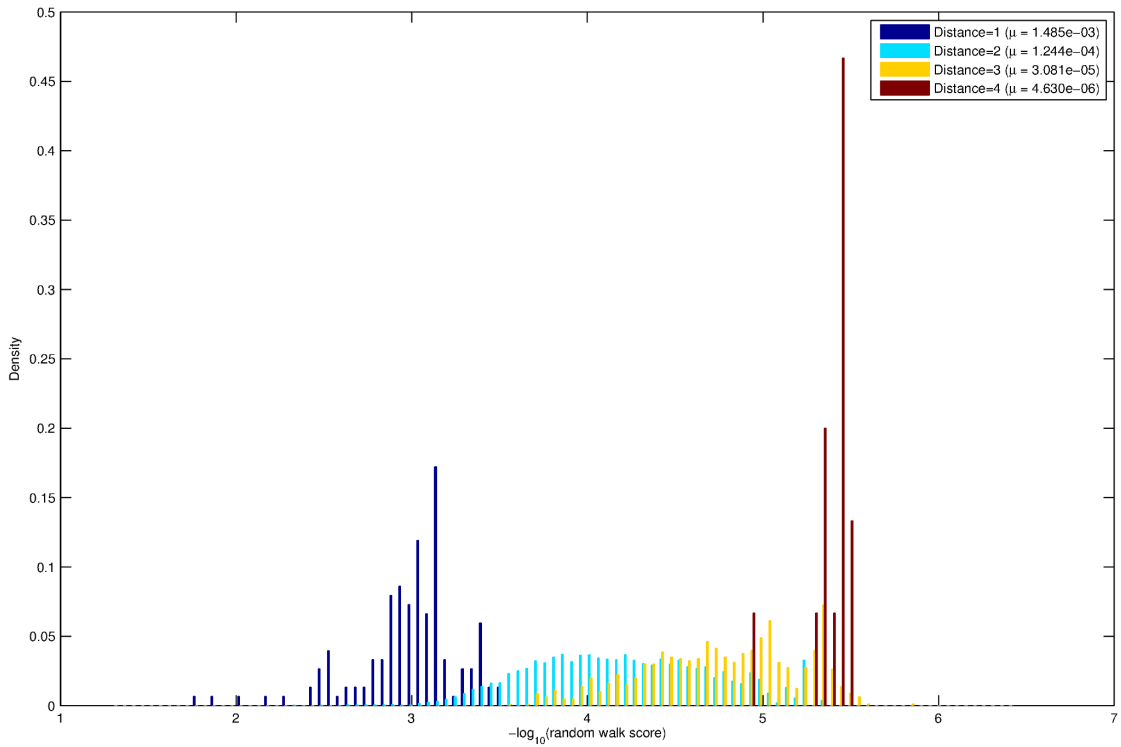
$$e_{\mathcal{S}}(i) = \begin{cases} \frac{1}{|\mathcal{S}|} & \text{if } v_i \in \mathcal{S}, \\ 0 & \text{O.W.} \end{cases}$$

for $\mathcal{S}$ 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 α 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.

PURDUE UNIVERSITY

DISTRIBUTION OF INFORMATION FLOW SCORES

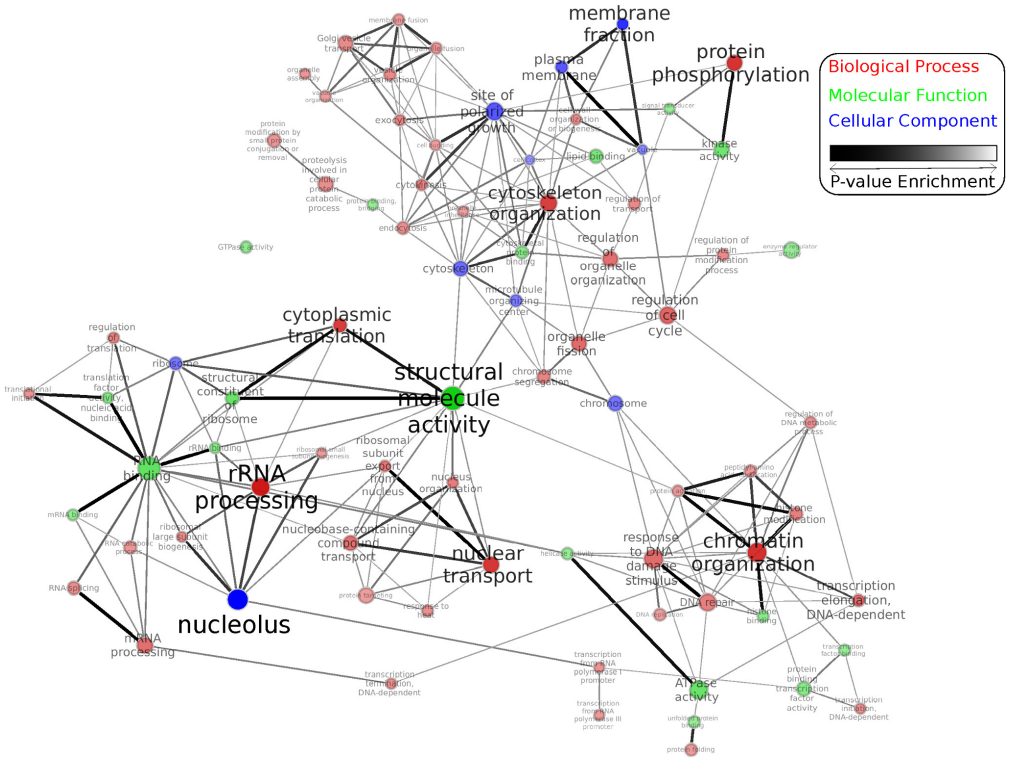
Distribution of information flow scores across nodes with similar distance from members of TORC1 are color coded accordingly. The μ parameter is the average of information flow scores for nodes under each distribution.



PURDUE UNIVERSITY

Enrichment map of yeast GOslim terms

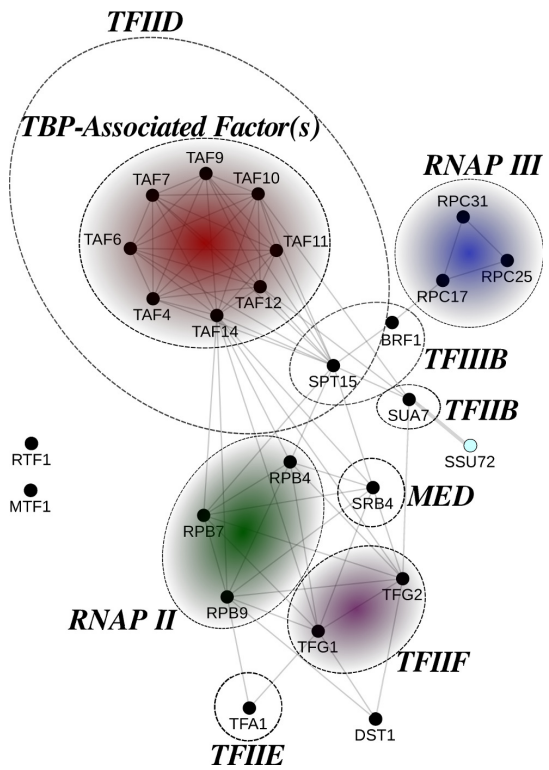
Enriched terms are identified by mHG p-value, computed for the ranked-list of genes based on their information flow scores. Each node represents a significant GO term and edges represent the overlap between genesets of GO terms. Terms in different branches of GO are color-coded with red, green, and blue. 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.



PURDUE UNIVERSITY

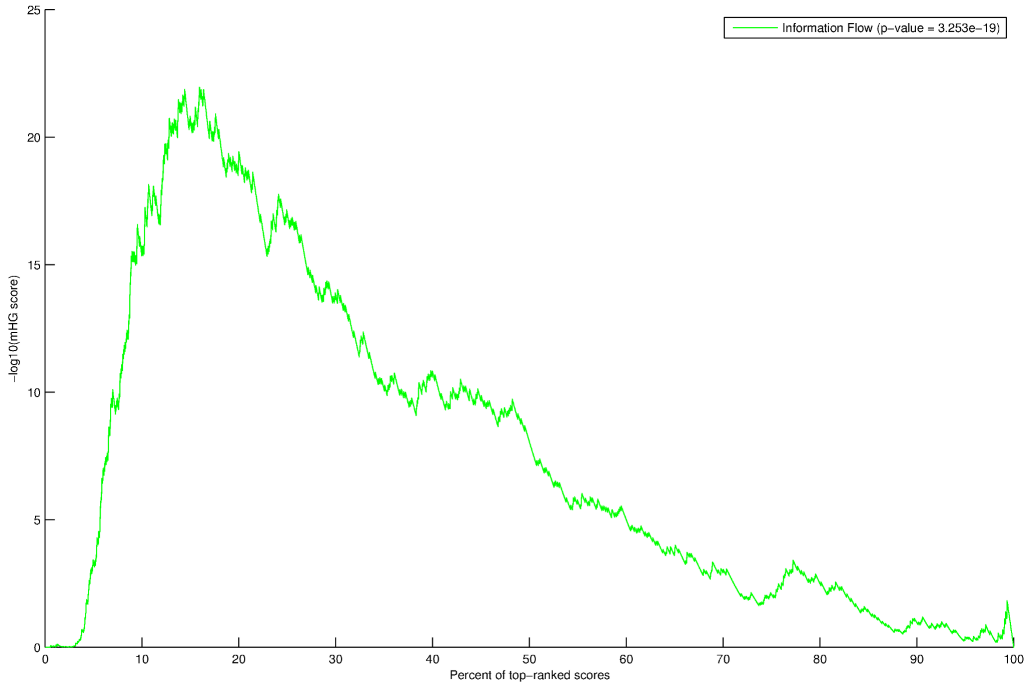
TOR-DEPENDENT CONTROL OF TRANSCRIPTION INITIATION

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

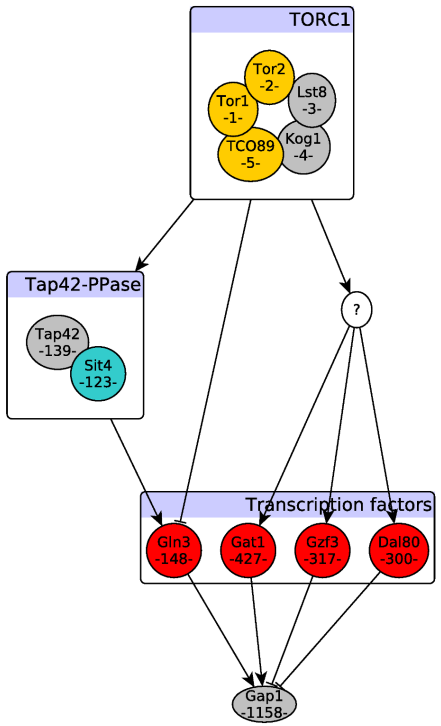
Enrichment score as a function of the score percentage. Computations are based on the set of differentially expressed genes in response to Rapamycin treatment. The peak of plot occurs at around top 15% of scores, resulting in the minimum hypergeometric (mHG) score of $\sim 1e-22$. The exact p-value for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



PURDUE UNIVERSITY

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.



PURDUE UNIVERSITY

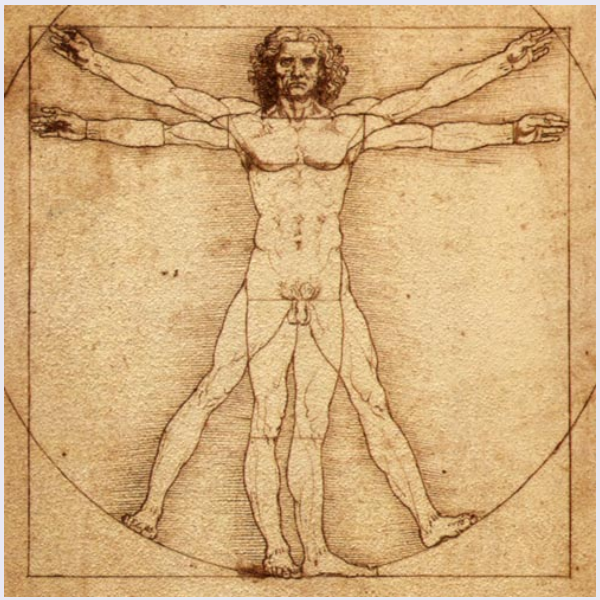
Outline

PURDUE UNIVERSITY

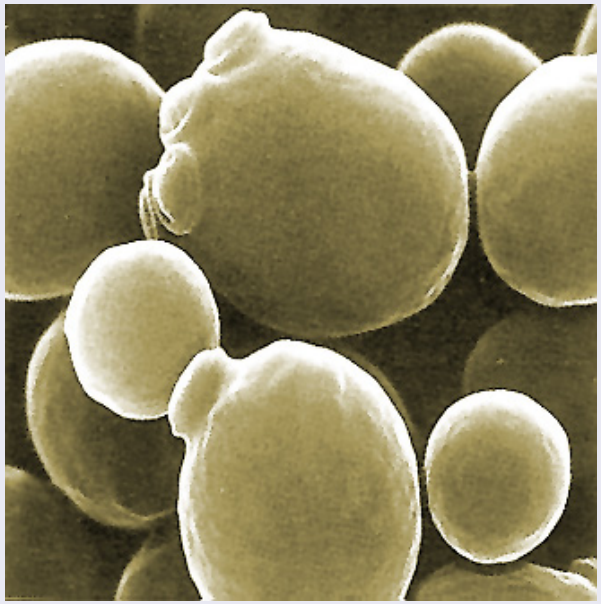
Comparative Network Analysis

Traditional approach

Basic Idea



$\Longleftrightarrow$



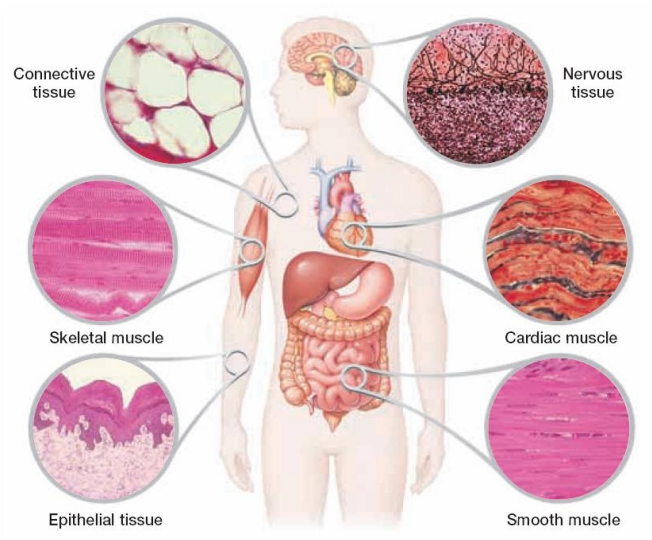
To project functional pathways from a well-studied organism, such as yeast, back to a higher-order organism, such as humans.

PURDUE UNIVERSITY

New outlook

Tissue-specific analysis

Different human cell types inherit a similar genetic code, but exhibit unique characteristics and functions.



- How does regulation of different genes contribute to functional differences in human tissues?
- How does a uni-cellular organism, such as yeast, contributes to the biological understanding of higher-order, multi-cellular organisms, such as humans?

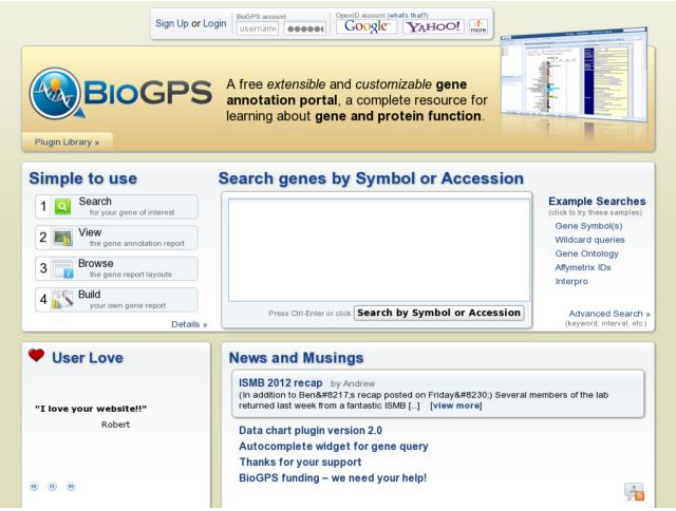
PURDUE UNIVERSITY

Outline

PURDUE UNIVERSITY

TISSUE-SPECIFIC GENE EXPRESSION

The GNF Gene Atlas dataset:



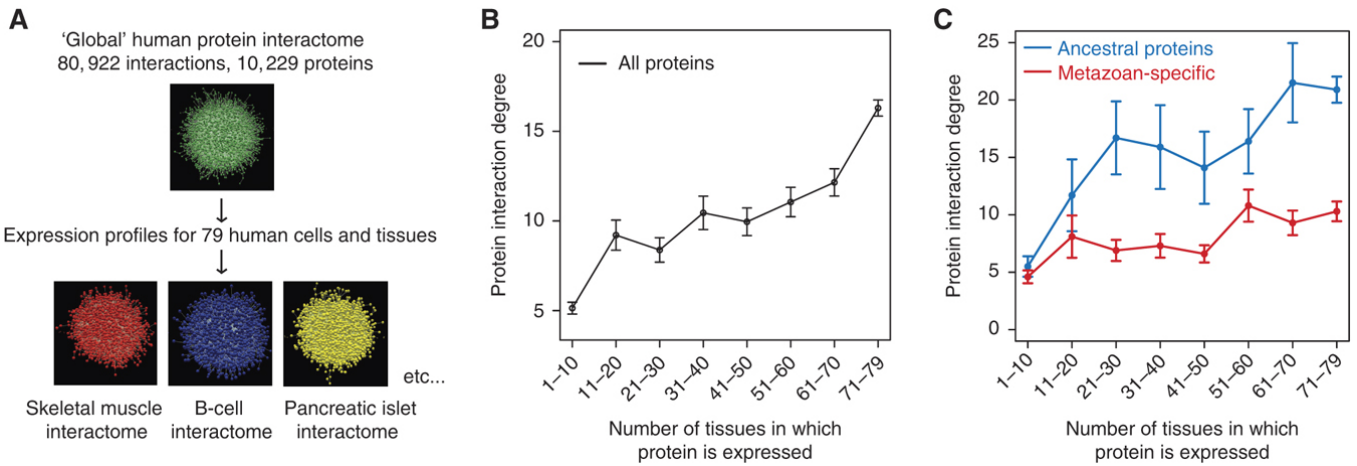
- 79 different tissues
- 44,775 human transcripts
- Platforms:
  1. Affymetrix HG-U133A.
  2. Custom GNF1H array.

PURDUE UNIVERSITY

TISSUE-SPECIFIC INTERACTOMES

- Vertex-induced sub-graphs of the human interactome
- Based on the GNF Gene Atlas dataset

  $\Rightarrow$ 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

PURDUE UNIVERSITY

SEQUENCE SIMILARITY OF PROTEIN PAIRS

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

PURDUE UNIVERSITY

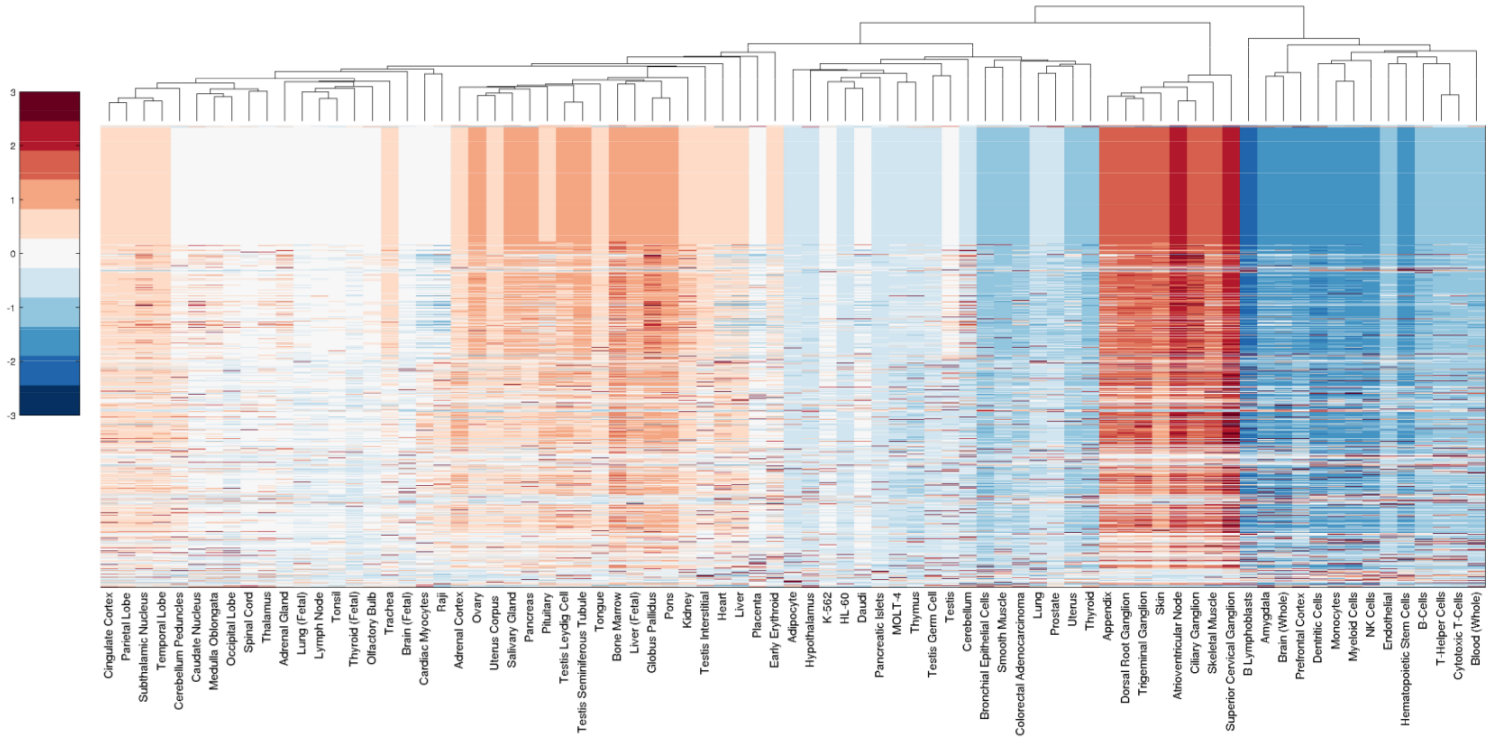
Outline

PURDUE UNIVERSITY

Gene expression signature

A unique identifier for similar tissues

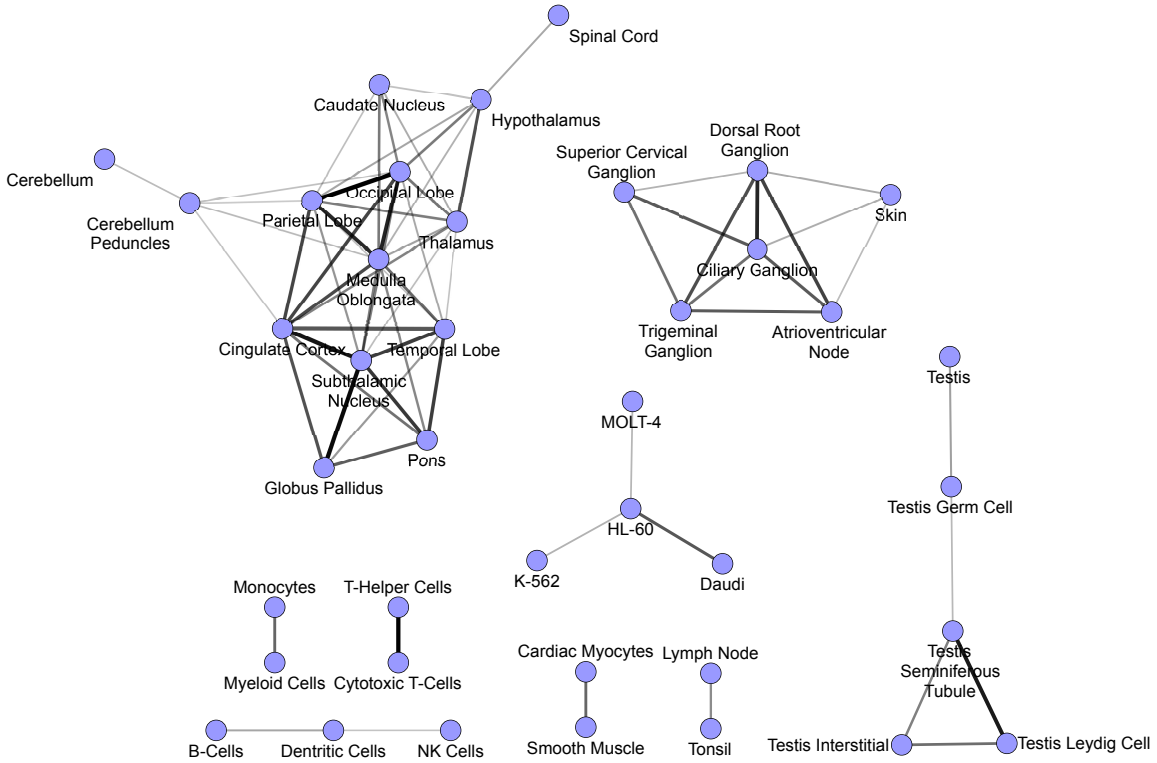
- Each row corresponds to a unique ORF.
- Each col corresponds to a unique tissue.
- Coloring is based on the normalized expression of each gene across different tissues.



PURDUE UNIVERSITY

TISSUE-TISSUE SIMILARITY NETWORK (TTSN)

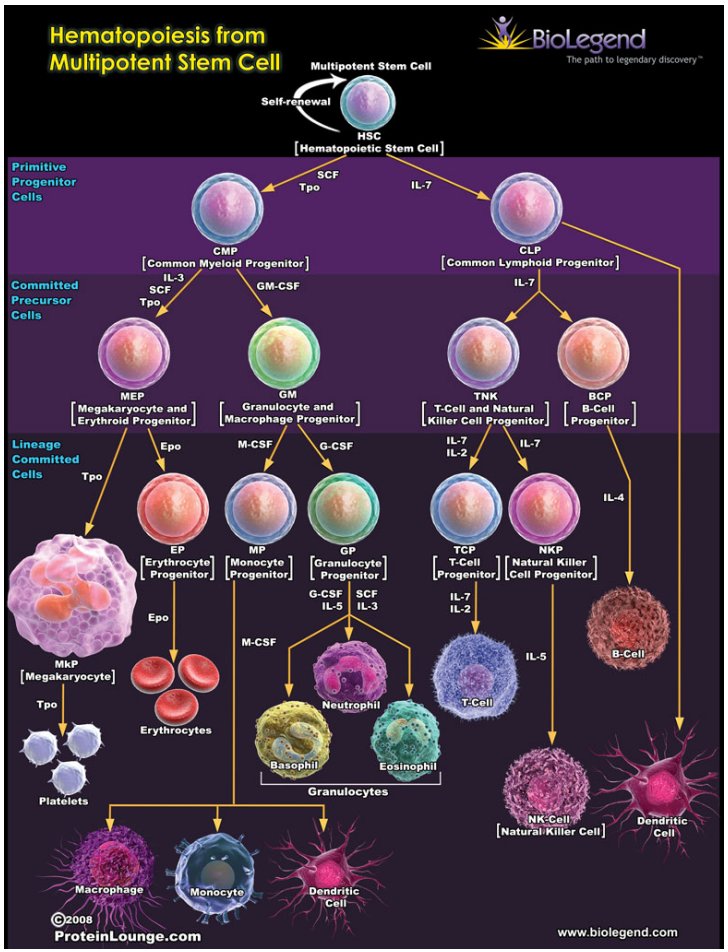
- First of its kind
- Based on similarity of expression signatures
- Differentiates between similar and dissimilar tissues



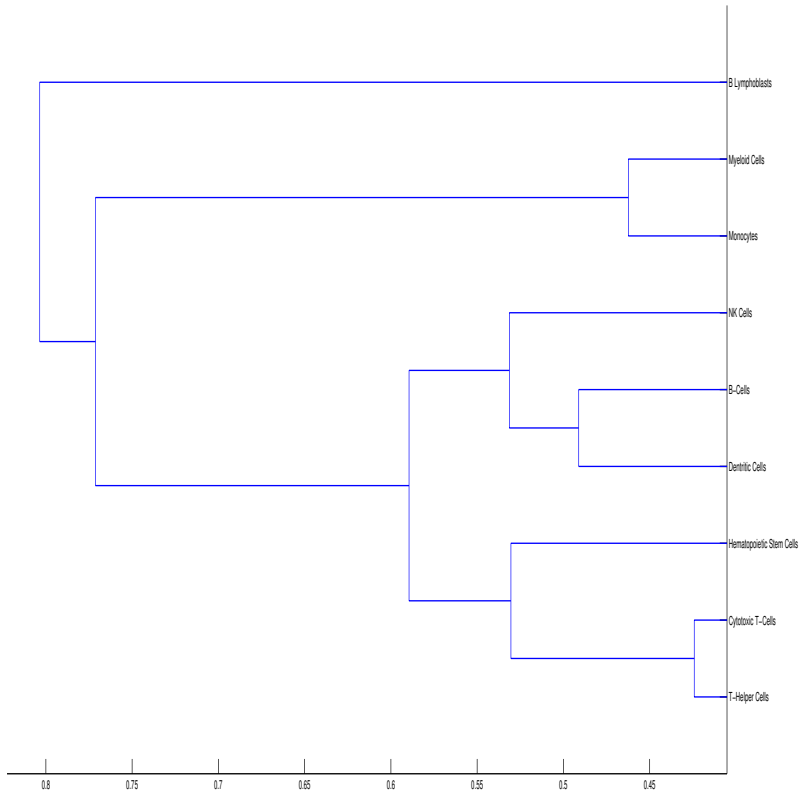
TTSN with stringent threshold (1.96 < Z-score)

PURDUE UNIVERSITY

RECONSTRUCTING THE DIFFERENTIATION TREE OF IMMUNE CELLS



Adopted from BioLegend



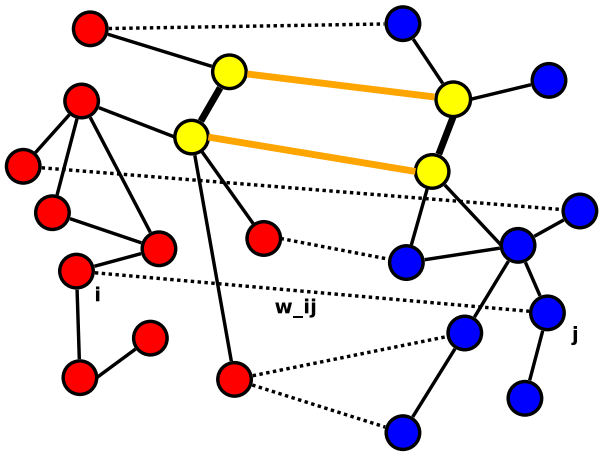
Similarity of different tissues to yeast

Summary of methods

- Align each tissue-specific network to yeast using *Belief Propagation (BP)* algorithm.
- For each alignment, compute:
  - Number of conserved edges.
  - Sequence similarity of the aligned proteins.
- Create and ensemble of random, *pseudo tissues* seeded around housekeeping proteins (proteins that are expressed in all tissues).
- Align random tissues with yeast and compute p-values of original alignment.

PURDUE UNIVERSITY

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 μ be the fractional contribution of topological similarity.
- To compute X, IsoRank iterates:

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

PURDUE UNIVERSITY

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.

PURDUE UNIVERSITY

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	**3.15**	152.12	7.29	783.48
celeg-hsapi	**3.28**	163.05	9.54	1209.28
celeg-scere	**1.97**	127.70	4.16	949.58
dmela-ecoli	**1.86**	86.80	4.78	807.93
dmela-hsapi	**8.61**	590.16	28.10	7840.00
dmela-scere	**4.79**	182.91	12.97	4905.00
ecoli-hsapi	**2.41**	79.23	4.76	2029.56
ecoli-scere	**1.49**	69.88	2.60	1264.24
hsapi-scere	**6.09**	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.

PURDUE UNIVERSITY

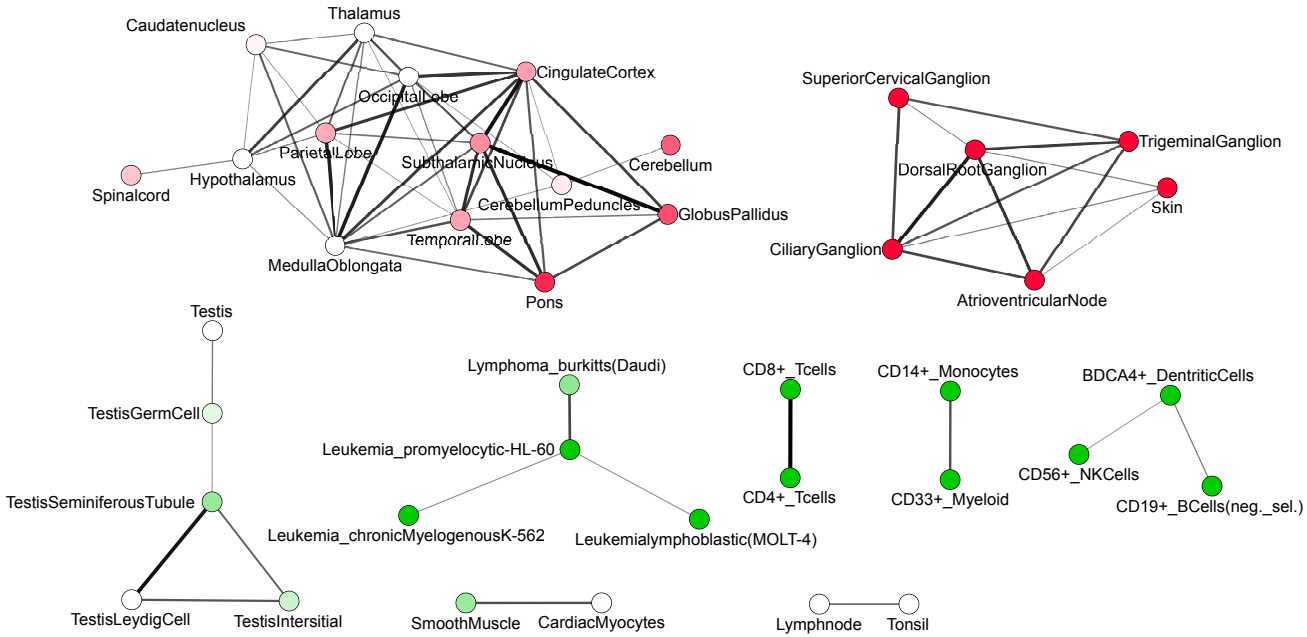
NSD: Parallelization [KKG JPDC'13, Submitted, KMSAG ParCo'13 Submitted]

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.

SIMILARITY OF DIFFERENT TISSUES TO YEAST

PRELIMINARY P-VALUE RESULTS

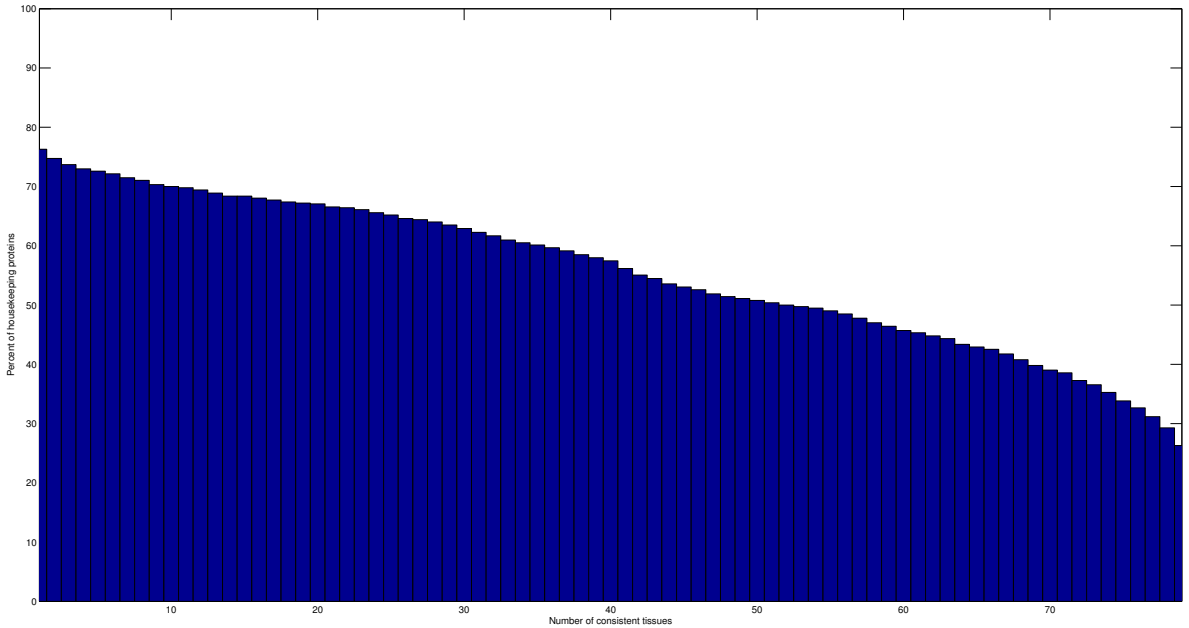


- Green nodes show tissues with significant similarity to yeast, while red nodes show dissimilar tissues.
- Similar tissues tend to have consistent p-values.

PURDUE UNIVERSITY

SIMILARITY OF DIFFERENT TISSUES TO YEAST

ALIGNMENT CONSISTENCY OF HOUSEKEEPING PROTEINS



- Approximately 75% of housekeeping proteins are aligned with yeast proteins.
- Among aligned HK proteins, 25% of them consistently aligned across all 79 tissues.

PURDUE UNIVERSITY

SUMMARY

- Aging is the primary risk factor for a number of human diseases.
- Emerging evidence supports the hypothesis that large classes of age-related pathologies share their underlying biology
- Constructing a comprehensive map of aging pathways is a critical step towards deciphering key lifespan mediators, their crosstalk, and systems-level organization.
- Tissue-specificity analysis is needed to precisely model aging in both proliferating and post-mitotic cells.

- Future works:
  - Construction of comprehensive, yet reliable tissue-specific networks by integrating various available high-throughput datasets.
  - Devising a *biased* information flow method for targeting specific subsets of TOR effectors.

PURDUE UNIVERSITY

For Further Reading I

SC. Johnson et al.
mTOR is a key modulator of ageing and age-related disease
Nature, 493(7432):338–45, 2013.

RM. Naylor et al.
Senescent cells: a novel therapeutic target for aging and age-related diseases
linical pharmacology and therapeutics, 93:105–116, 2013.

J Campisi et al.
Cellular senescence: a link between cancer and age-related degenerative disease?
Seminars in cancer biology, 21: 354–359, 2012.

PURDUE UNIVERSITY