Pairwise Local Alignment of Protein Interaction Networks Based on Models of Evolution

Mehmet Koyutürk, Ananth Grama, Wojciech Szpankowski

Dept of. Computer Sciences
Purdue University

Work supported by
NIH & NSF

Outline

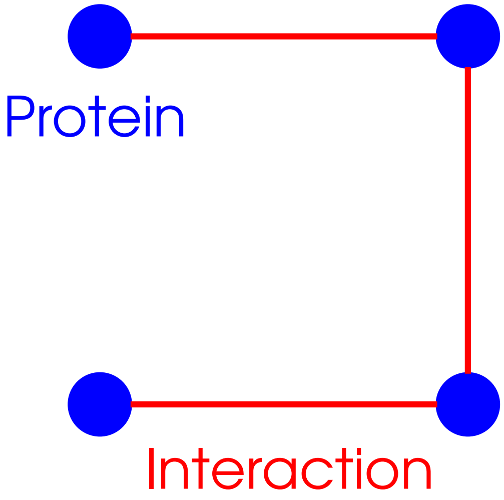
- Background
  - Protein Interaction Networks
  - Modularity of Function & Evolution
  - Theoretical Models on Evolution of PPI networks
- Model & Algorithms
  - Pairwise Local Alignment of PPI Networks: Match, Mismatch, Duplication
  - Alignment Graph & Maximum-Weight Subgraph Problem
  - Implementation, Parameters, Extensions
- Results
  - Alignment of Human-Mouse and Yeast-Fly PPI Networks
- Conclusion
  - Related Work

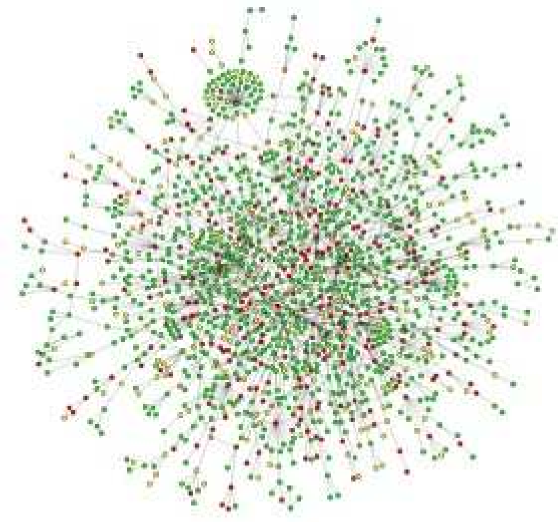
Outline

- Background
  - Protein Interaction Networks
  - Modularity of Function & Evolution
  - Theoretical Models on Evolution of PPI networks
- Model & Algorithms
  - Pairwise Local Alignment of PPI Networks: Match, Mismatch, Duplication
  - Alignment Graph & Maximum-Weight Subgraph Problem
  - Implementation, Parameters, Extensions
- Results
  - Alignment of Human-Mouse and Yeast-Fly PPI Networks
- Conclusion
  - Related Work

Protein-Protein Interaction (PPI) Networks

- Proteins interact with each other to perform cellular functions
  - Signaling, transport, cell cycling, protein modification...
- Interacting proteins can be discovered experimentally by high-throughput screening
  - Two-hybrid (Ito et al., *PNAS*, 2001)
  - Mass spectrometry (Ho et al., *Nature*, 2002)
  - Tandem affinity purification (TAP) (Gavin et al., *Nature*, 2002)



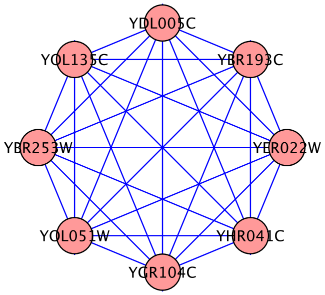


S. Cerevisiae protein interaction network

Source: (Jeong et al., *Nature*, 2001)

Modularity of Protein Interactions

- Functional modules
  - “a spatially or chemically isolated set of functionally associated components that accomplishes a discrete biological process” (Pereira-Leal & Teichmann, *Genome Research*, 2005)
  - e.g., protein complexes
  - Proteins in a functional module densely interact with each other



RNA polymerase II transcription mediator activity in yeast

- Modular evolution
  - Proteins that are part of dense topological motifs show higher degree of conservation (Wuchty et al., *Nature Genetics*, 2003)
  - Proteins that interact with each other follow similar evolutionary trajectories (Pellegrini et al., *PNAS*, 1999)
  - Selective pressure on function ⇒ modular conservation

Comparative Analysis of PPI Networks

- Pairwise local alignment of PPI networks
  - Find groups of proteins with highly conserved interactions
  - Conservation of interactions suggests conservation of function
  - Find subgraphs that are highly conserved in terms of interactions

- What do we gain from comparative analysis of protein interactions?
  - Identification of orthologous modules & interactions
  - Detailed understanding of functional conservation and divergence at a modular level
  - Functional annotation of modules, interactions, and proteins

Evolution of PPI Networks

- PPI networks can be modeled by power-law graphs
  - Relative frequency of proteins that interact with k proteins is proportional to $k^{-\gamma}$
  - γ is a network-specific parameter
  - Most proteins interact with a single protein, while there are only a few hubs $\Rightarrow$ robustness to random attacks
- Network growth model based on preferential attachment (Barábasi & Albert, *Science*, 1999)
  - When a new protein u_{n+1} is added to the network, its probability of interacting with protein u_i, is proportional to the number of interactions of u_i, *i.e.*, $P(u_i u_{n+1} \in E_{n+1}) \propto d(u_i)^{\beta}$ for $1 \leq i \leq n$
  - Why do proteins choose to interact with well-connected proteins?
  - Selective pressure on maintaining connectivity of strongly connected proteins (Eisenberg & Levanon, *Phys. Rev. Let.*, 2003)

Theoretical Models on Evolution of Interactions

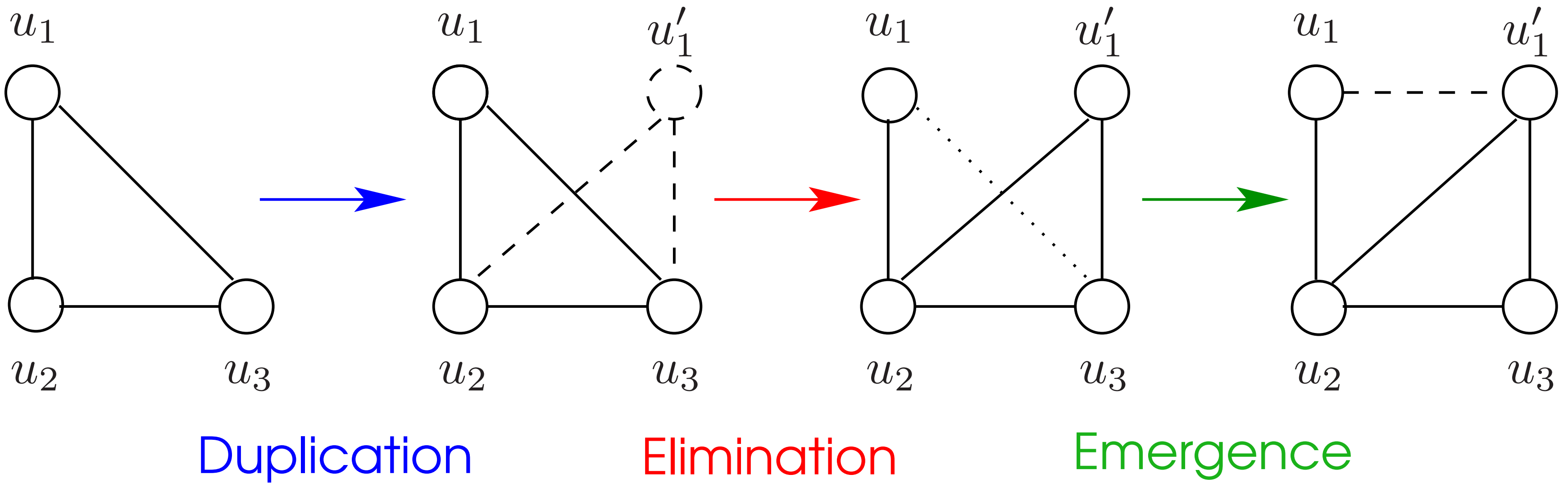
- Duplication/divergence models incorporate evolution of sequences, interactions, and function (Wagner, *Proc. R. Soc. Lond.*, 2003), (Pastor-Sotorras et al., *J Theo. Bio.*, 2003), (Vázquez et al., *ComPlexUs*, 2003)

- Gene duplication
  - Interactions of duplicated proteins are also duplicated
  - Provides redundancy, relaxing pressure

- Protein Divergence
  - Loss of interactions are likely to be tolerated for duplicated proteins
  - Duplicated proteins rapidly diverge by losing (and sometimes gaining) interactions through sequence mutations

- Theoretically shown to generate power-law graphs (Chung et al., *J Comp. Bio.*, 2003)

Duplication/Divergence Models



- Duplication/divergence models form the theoretical basis for comparative analysis of interactions
  - They provide us with a simplified basis for solving a very hard problem
  - Discovered alignments can be annotated through gene duplications and elimination/emergence of interactions

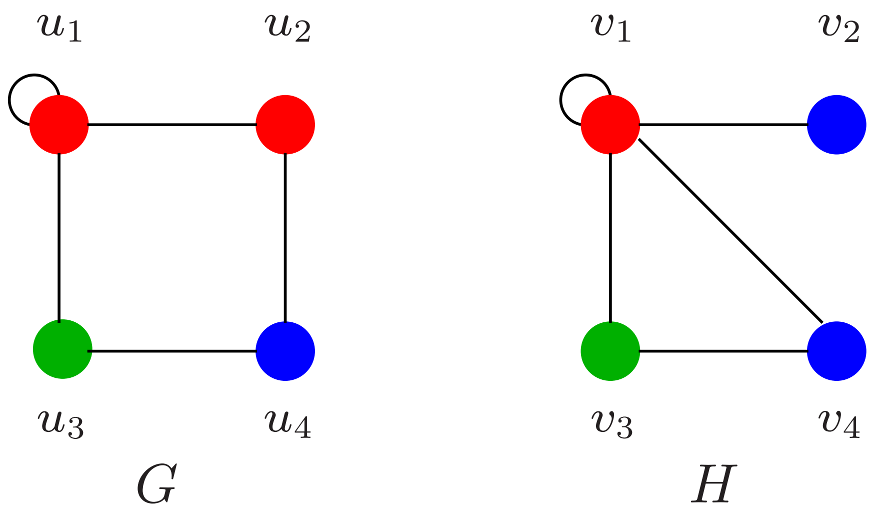
Outline

- Background
  - Protein Interaction Networks
  - Modularity of Function & Evolution
  - Theoretical Models on Evolution of PPI networks
- Model & Algorithms
  - Pairwise Local Alignment of PPI Networks: Match, Mismatch, Duplication
  - Alignment Graph & Maximum-Weight Subgraph Problem
  - Implementation, Parameters, Extensions
- Results
  - Alignment of Human-Mouse and Yeast-Fly PPI Networks
- Conclusion
  - Related Work

Modeling PPI Networks

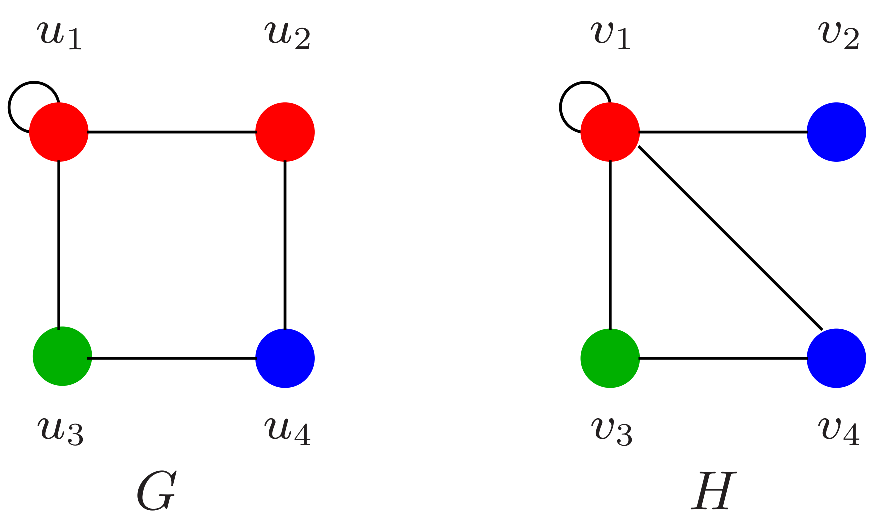
- PPI networks $G(U, E)$ and $H(V, F)$ that belong to two different organisms
  - U and V are sets of proteins (nodes) in each organism
  - E and F are sets of interactions (undirected edges) in each organism

- Sparse similarity function $S(u, v)$ for all $u, v \in U \cup V$
  - $S(u, v)$ is a function of sequence similarity (homology)
  - If $S(u, v) > 0$, u and v are potentially orthologous

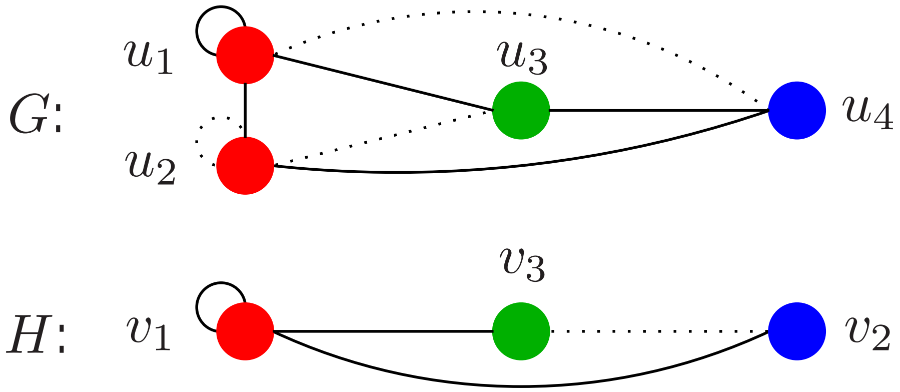


Identical color $\Rightarrow S > 0$

Local Alignment



PPI Networks



Alignment induced by protein subset pair

$\{\{u_1, u_2, u_3, u_4\}, \{v_1, v_2, v_3\}\}$

Match, Mismatch, Duplication

- Alignment induced by protein subset pair $\mathcal{P} = \{\tilde{U} \in U, \tilde{V} \in V\}$: $\mathcal{A}(G, H, S, \mathcal{P}) = \{\mathcal{M}, \mathcal{N}, \mathcal{D}\}$
  - A match $M \in \mathcal{M}$ corresponds to two pairs of homolog proteins from each protein subset such that both pairs interact in both PPI networks. A match is associated with score μ.

    $$\mathcal{M} = \{u, u' \in \tilde{U}, v, v' \in \tilde{V} : S(u, v) > 0, S(u', v') > 0, uu' \in E, vv' \in F\}$$

  - A mismatch $N \in \mathcal{N}$ corresponds to two pairs of homolog proteins from each protein subset such that only one of the pairs is interacting. A mismatch is associated with penalty ν.

    $$\begin{aligned} \mathcal{N} &= \{u, u' \in \tilde{U}, v, v' \in \tilde{V} : S(u, v) > 0, S(u', v') > 0, uu' \in E, vv' \notin F\} \\ &\cup \{u, u' \in \tilde{U}, v, v' \in \tilde{V} : S(u, v) > 0, S(u', v') > 0, uu' \notin E, vv' \in F\} \end{aligned}$$

  - A duplication $D \in D$ corresponds to a pair of homolog proteins that are in the same protein subset. A duplication is associated with score δ.

    $$\mathcal{D} = \{u, u' \in \tilde{U} : S(u, u') > 0\} \cup \{v, v' \in \tilde{V} : S(v, v') > 0\}$$

Match, Mismatch, Duplication: Interpretation

- Match corresponds to a conserved interaction
  - Rewarded for functional conservation after speciation
- Mismatch corresponds to interactions that have been eliminated/ have emerged after speciation
  - May correspond to experimental error/incomplete data
  - Penalized for functional divergence after speciation
- Duplication corresponds to a gene duplication
  - May have happened before (out-paralog) or after speciation (in-paralog)
  - The protein pair may correspond to orthologs or distant paralogs
  - Orthologs are likely to be part of the same functional module, while paralogs may drop from or become part of different modules (Wagner, *Mol. Bio. Evol.*, 2001)
  - Scored to account for trade-off between functional divergence and functional conservation after speciation

Pairwise Local Alignment of PPI networks: Formulation

- For an alignment $\mathcal{A}$ induced by protein subset pair $\mathcal{P}$, we define alignment score

$$\sigma(\mathcal{A}(P)) = \sum_{M \in \mathcal{M}} \mu(M) + \sum_{N \in \mathcal{N}} \nu(N) + \sum_{D \in \mathcal{D}} \delta(D)$$

  - a measure of homology between protein sets from each organism, assessing the likelihood of these sets being a conserved functional module

- Problem: Find all protein subset pairs with unusually high (statistically significant) alignment score
  - high scoring subgraph pair

- A graph equivalent to local sequence alignment
  - Match, mismatch, gap $\longrightarrow$ Match, mismatch, duplication
  - Sequence homology $\Rightarrow$ ortholog molecules
    Subgraph homology $\Rightarrow$ ortholog modules

Weighted Alignment Graph

- Given PPI networks G and H, we construct weighted alignment graph $\mathbf{G}(\mathbf{V},\mathbf{E})$

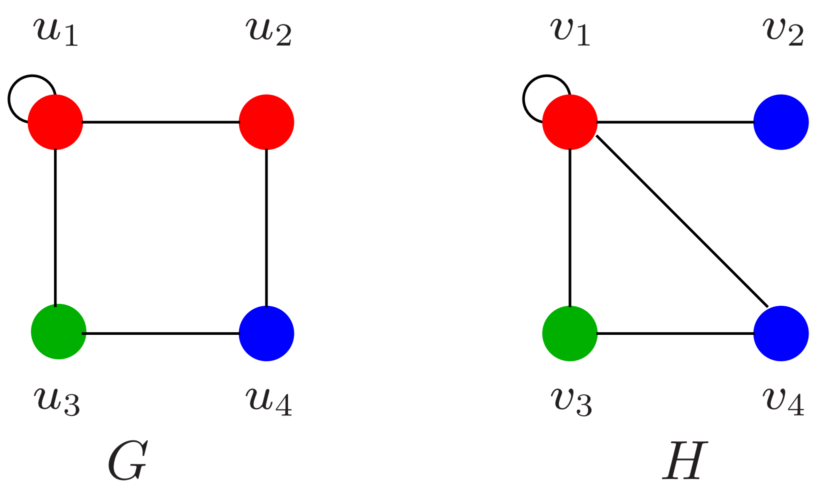
- $\mathbf{V}$ consists all pairs of homolog proteins $\mathbf{v} = \{u \in U, v \in V\}$

- An edge $\mathbf{vv}' = \{uv\}\{u'v'\}$ in $\mathbf{E}$ is assigned weight

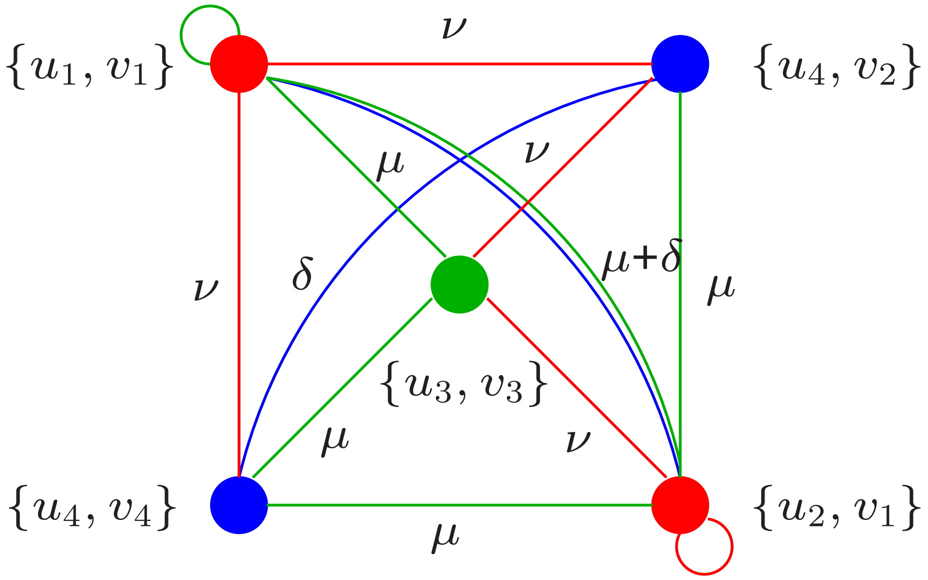
$$w(\mathbf{vv}') = \mu(uu', vv') + \nu(uu', vv') + \delta(u, u') + \delta(v, v')$$

- An edge is called a
  - match edge if $uu' \in E$ and $vv' \in E$, with weight $w(\mathbf{vv}') = \mu(uv, u'v')$
  - mismatch edge if $uu' \in E$ and $vv' \notin E$ or vice versa, with weight $w(\mathbf{vv}') = \nu(uv, u'v')$
  - duplication edge if $S(u, u') > 0$ or $S(v, v') > 0$, with weight $w(\mathbf{vv}') = \delta(u, u')$ or $w(\mathbf{vv}') = \delta(v, v')$

A Sample Alignment Graph



PPI Networks



Weighted alignment graph

Maximum Weight Induced Subgraph Problem

- Definition: (MAWISH)
  - Given graph $\mathbf{G}(\mathbf{V}, \mathbf{E}, w)$ and a constant ϵ, find $\tilde{\mathbf{V}} \in \mathbf{V}$ such that

$$W(\tilde{\mathbf{V}}) = \sum_{\mathbf{v}, \mathbf{u} \in \tilde{\mathbf{V}}} w(\mathbf{vu}) \geq \epsilon.$$

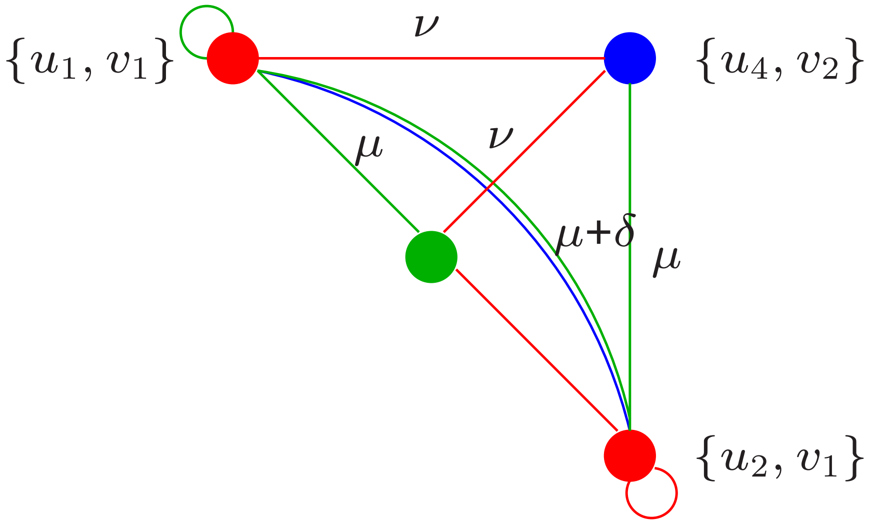
  - NP-complete

- Theorem: (MAWISH $\equiv$ Pairwise alignment)
  - If $\tilde{\mathbf{V}}$ is a solution for the MAWISH problem on $\mathbf{G}(\mathbf{V}, \mathbf{E}, w)$, then $\mathbf{P} = \{\tilde{U}, \tilde{V}\}$ induces an alignment $\mathcal{A}(\mathbf{P})$ with $\sigma(\mathcal{A}) = W(\tilde{\mathbf{V}})$, where

$$\tilde{U} = \{u \in U : \exists v \in V \text{ s.t. } \{u, v\} \in \tilde{\mathbf{V}}\}$$
$$\tilde{V} = \{v \in V : \exists u \in U \text{ s.t. } \{u, v\} \in \tilde{\mathbf{V}}\}$$

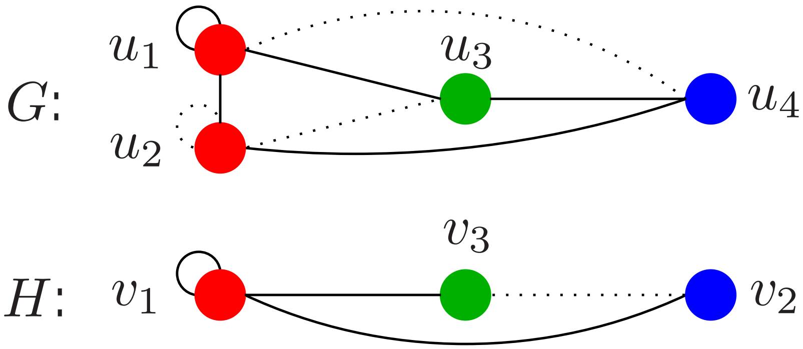
  - We are looking for locally optimal solutions

MaWISh $\equiv$ Pairwise Alignment



Subgraph induced by vertex set

$$\tilde{\mathbf{V}} = \{\{u_1, v_1\}, \{u_2, v_1\}, \{u_3, v_3\}, \{u_4, v_2\}\}$$



Alignment induced by protein subset pair

$$\{\{u_1, u_2, u_3, u_4\}, \{v_1, v_2, v_3\}\}$$

A Greedy Algorithm for MAWISH

- Greedy graph growing
  - Start with a heavy node, put it in $\tilde{\mathbf{V}}$
  - Choose $\mathbf{v}$ that is most heavily connected to $\tilde{\mathbf{V}}$ and put it in $\tilde{\mathbf{V}}$ until no $\mathbf{v}$ is positively connected to $\tilde{\mathbf{V}}$
  - If total weight of the subgraph induced by $\tilde{\mathbf{V}}$ is statistically significant, return $\tilde{\mathbf{V}}$
  - Linear time

- For all (possibly overlapping) local alignments
  - mark nodes of discovered subgraph and run the greedy algorithm again by choosing only unmarked nodes as seed

- $O(|E| + |F|)$ time algorithm if number of homologs is bounded for each protein

Assessing Protein Homology

- Similarity score $S(u, v)$ reflects our confidence in two proteins being orthologous

- BLAST E-value
  - $S(u, v) = log_{10} \frac{p(u,v)}{p_{random}}$
  - $p(u, v)$ is the probability of true homology between u and v, given BLAST E-value (Kelley et al., *PNAS*, 2004)

- Ortholog clustering
  - INPARANOID: Discovers ortholog groups of proteins among two species, while distinguishing in-paralogs and out-paralogs (Remm et al., *J Mol. Bio.*, 2001)
  - $S(u, v) = c(u)c(v)$, where $0 \leq c(u) \leq 1$ is the confidence of the INPARANOID algorithm in assigning u to its corresponding cluster
  - Filters out redundant homologies
  - Does not leave room for network alignment to identify distant paralogs

Scoring Matches, Mismatches and Duplications

- Match score
  - Two interactions are orthologous only if both interacting partners are orthologous
  - $\mu(uu', vv') = \bar{\mu} \min\{S(u,v), S(u',v')\}$

- Mismatch penalty
  - $\nu(uu', vv') = -\bar{\nu} \min\{S(u,v), S(u',v')\}$

- Duplication score
  - Reward orthologs, penalize distant paralogs
  - $\delta(u, u') = \bar{\delta}(S(u, u') - \bar{d})$
  - $S(u, v) \geq \bar{d} \Rightarrow u$ and v are orthologs

- $\bar{\mu}$, $\bar{\nu}$, and $\bar{\delta}$ are relative weights for match, mismatch, and duplication, respectively

Accounting for Experimental Error

- PPI networks are incomplete
  - PPI networks obtained from high-throughput screening are prone to errors in terms of both false negatives and positives
  - PPI data come from different sources

- Several methods have been developed to combine data and account for experimental error
  - These methods assess the likelihood of an interaction between two proteins (Jansen et al., *Science*, 2003)

- The proposed framework can be easily extended to align weighted PPI networks
  - $\mu(uu', vv') = \bar{\mu} S(uu', vv') \varpi_{uu'} \varpi_{vv'}$
  - $\nu(uu', vv') = -\bar{\nu} S(uu', vv') (\varpi_{uu'}(1 - \varpi_{vv'}) + (1 - \varpi_{uu'}) \varpi_{vv'})$
  - Here, $\varpi_{uu'}$ denotes the likelihood of an interaction between u and u'

Tuning Model Components and Parameters

- Shortest-path mismatch model
  - Proteins that are linked by a short alternative path are more likely to tolerate losing their interaction
  - Penalize mismatches based on the distance between proteins

$$\nu(uu', vv') = \bar{\nu} S(uu', vv')(\bar{\Delta} - \max\{\Delta(u, u'), \Delta(v, v')\}),$$

- Linear duplication model
  - Alignment graph model enforces each duplicate pair in alignment to be scored ⇒ $\binom{n}{2}$ for n duplicates (quadratic duplication model)
  - In the evolutionary process, each paralog is the result of a single duplication
  - Score only $n-1$ duplications for n duplicates

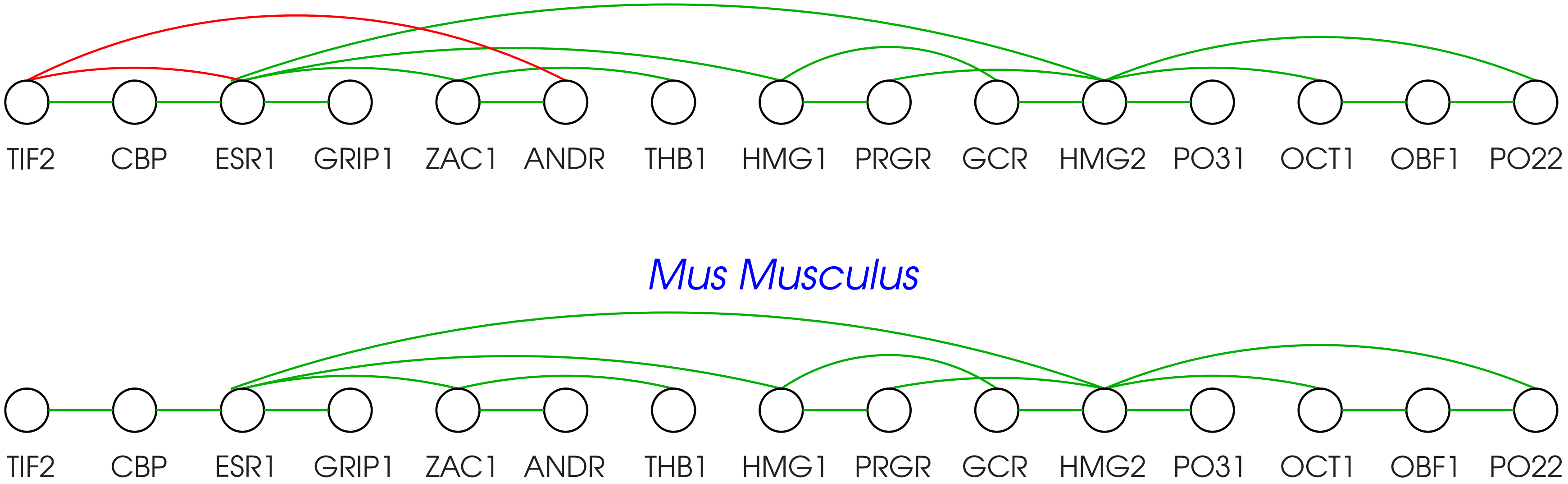
Outline

- Background
  - Protein Interaction Networks
  - Modularity of Function & Evolution
  - Theoretical Models on Evolution of PPI networks
- Model & Algorithms
  - Pairwise Local Alignment of PPI Networks: Match, Mismatch, Duplication
  - Alignment Graph & Maximum-Weight Subgraph Problem
  - Implementation, Parameters, Extensions
- Results
  - Alignment of Human-Mouse and Yeast-Fly PPI Networks
- Conclusion
  - Related Work

Experimental Results

- Interaction data is obtained from DIP
- Homo Sapiens vs Mus Musculus
  - H. Sapiens: 1065 proteins, 1369 interactions
  - M. Musculus: 329 proteins, 286 interactions
  - Alignment graph consists of 273 nodes and 1233 edges
  - 305 matches, 205 mismatches in human, 149 mismatches in mouse
  - 536 duplications in human 384 duplications in mouse
  - Trying alternate settings for relative weights, we identify 54 non-redundant alignments, 15 of which contain at least 3 proteins
- Saccharomyces Cerevisiae vs Drosophila Melanogaster
  - S.Cerevisiae: 4773 proteins, 15481 interactions
  - D. Melanogaster: 7068 proteins, 20988 interactions
  - Alignment graph consists of 1901 nodes and 15811 edges
  - 232 matches, 9278 mismatches in yeast, 2689 mismatches in fly
  - 1862 duplications in yeast, 3050 duplications in fly
  - 62 alignments, 18 contain at least three proteins

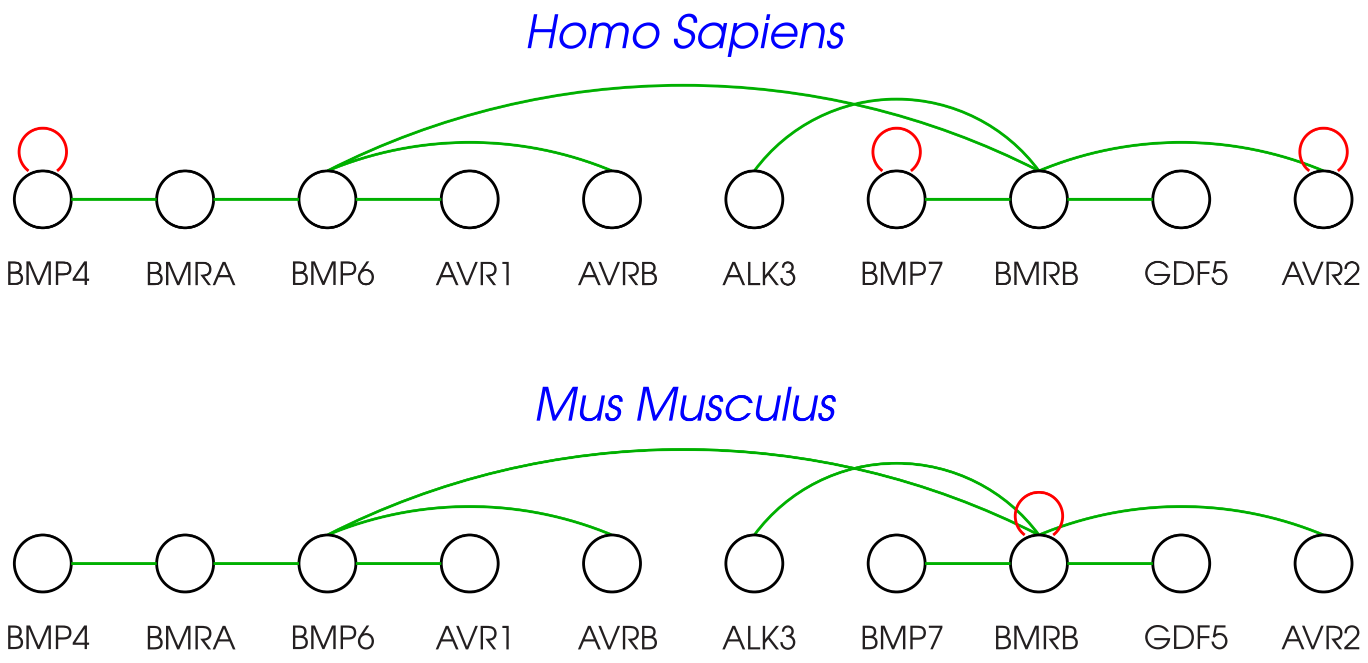
Alignment of Human and Mouse PPI Networks


Homo Sapiens

Mus Musculus

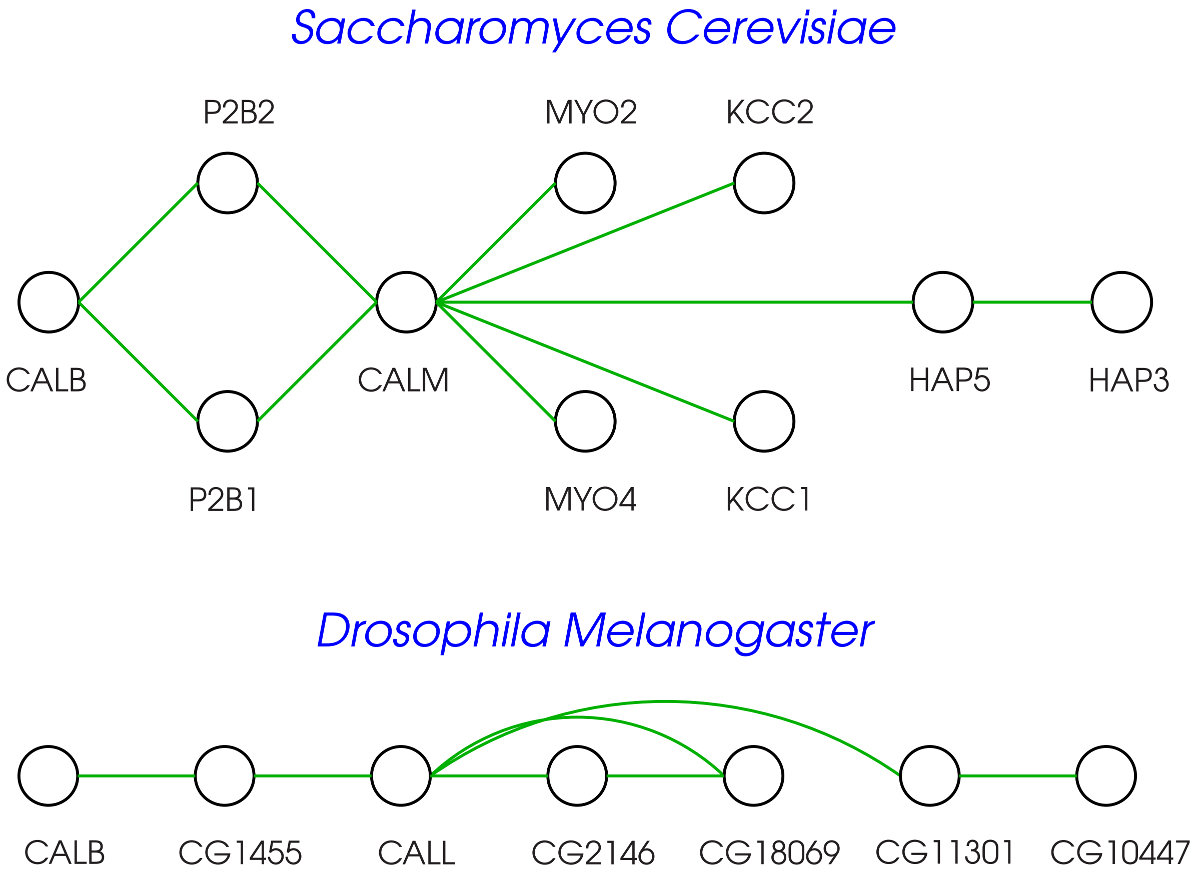
A conserved subnet that is part of
DNA-dependent transcription regulation

Alignment of Human and Mouse PPI Networks



A conserved subnet that is part of
transforming growth factor beta receptor signaling pathway

Alignment of Yeast and Fly PPI Networks

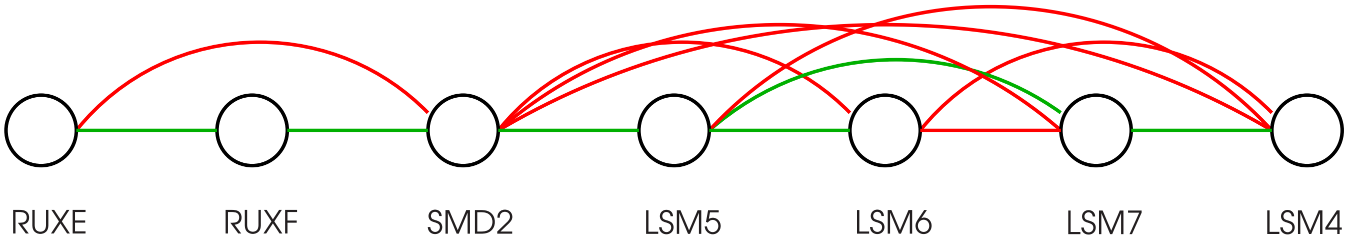


CALM (Calmodulin) mediates the control of protein kinases and phosphatases via Ca(2+) in yeast

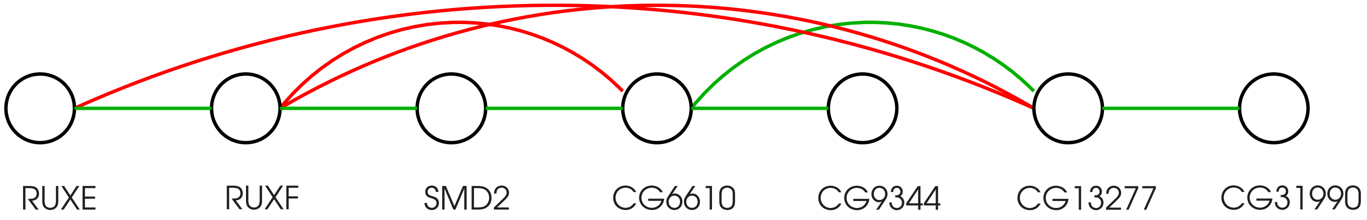
CALL (Androcam) may be involved in calcium-mediated signal transduction in fly

Alignment of Yeast and Fly PPI Networks

Saccharomyces Cerevisiae



Drosophila Melanogaster



A conserved pathway that is part of
Nuclear mRNA splicing

Outline

- Background
  - Protein Interaction Networks
  - Modularity of Function & Evolution
  - Theoretical Models on Evolution of PPI networks
- Model & Algorithms
  - Pairwise Local Alignment of PPI Networks: Match, Mismatch, Duplication
  - Alignment Graph & Maximum-Weight Subgraph Problem
  - Implementation, Parameters, Statistical Significance, Extensions
- Results
  - Alignment of Human-Mouse and Yeast-Fly PPI Networks
- Conclusion
  - Related Work

Related Work

- PathBLAST: Identification of conserved pathways within bacteria & yeast through PPI network alignment (Kelley et al., *PNAS*, 2003)
  - Gaps and mismatches to account for evolutionary variations and experimental error
- Complex identification by comparative analysis of yeast & bacterial PPI networks (Sharan et al., *RECOMB*, 2004)
  - PPI networks are joined into an orthology graph based on probabilistic model
  - Edge weights are assigned based on likelihood
  - Superposing networks to identify complexes vs comparing networks to understand conservation/divergence
- Extension to multiple PPI networks
  - Conserved patterns of protein interaction in multiple species (Sharan et al., *PNAS*, 2005)
  - Graph mining based on contraction of orthologs (Koyutürk et al., *ISMB*, 2004)

Thanks...

- Shankar Subramaniam (UCSD)
- Yohan Kim (UCSD)
- Umut Topkara (Purdue)
- Anonymous RECOMB reviewers
- NIH & NSF

Statistical Significance

- Refence model
  - PPI networks & protein sequences that belong to different species are independent from each other
  - Interactions are generated randomly from a distribution characterized by a given degree sequence, independently from each other
  - Sequences are generated by a memoryless source
- Parameter estimation
  - Probability of interaction
    $q_{uu'} = \frac{d(u)d(u')}{|E|}$ for $u, u' \in U$, $q_{vv'} = \frac{d(v)d(v')}{|F|}$ for $v, v' \in V$, where $d(u)$ is the degree of u
  - Probability of homology between-species
    $p = \frac{\sum_{u \in U, v \in V} S(u,v)}{|U||V|}$
  - Probability of homology within-species
    $p_U = \frac{\sum_{u \in U, u' \in U} S(u,u')}{|U|^2}$, $p_V = \frac{\sum_{v \in V, v' \in V} S(v,v')}{|V|^2}$

Statistical Significance

- Expected value of the score of an alignment (weight of corresponding induced subgraph)

  $$E[W(\tilde{\mathbf{V}})] = \sum_{\mathbf{v},\mathbf{u}\in\tilde{\mathbf{V}}} E[w(\mathbf{vu})],$$

  where

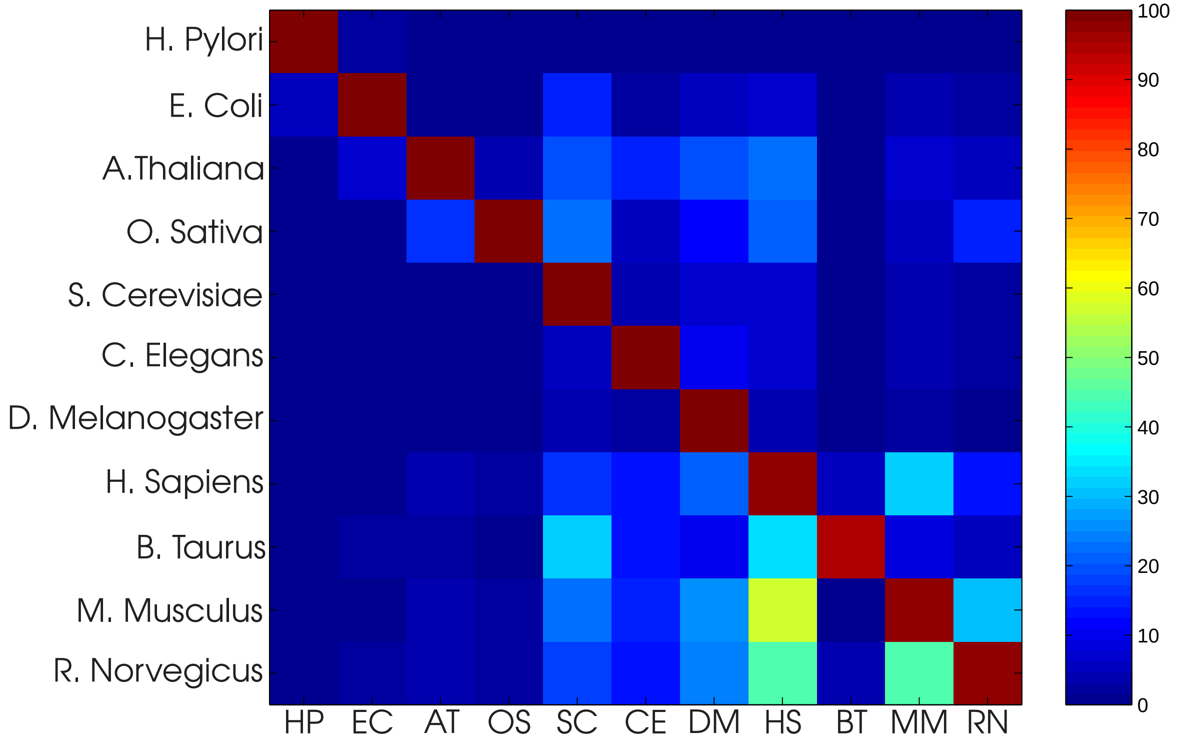
  $$\begin{aligned} E[w(\mathbf{vu})] = \quad & \bar{\mu}p^2 q_{uu'}q_{vv'} - \bar{\nu}p^2(q_{uu'}(1-q_{vv'}) + (1-q_{uu'})q_{vv'}) \\ & -\bar{\delta}(p_U(1-p_U) + p_V(1-p_V)) \end{aligned}$$

- Based on the independence assumption, variance of subgraph weight can be estimated similarly

- Normal approximation allows us to compute z-score

Conservation of Interactions

- Percentage of interactions that have orthologs in the respective species
  - Data from BIND & DIP



Penalties must be relaxed while analyzing distant species

Ongoing Work

- Adjustment of scores & penalties based on experimental analysis & probabilistic models

- Comprehensive alignment of PPI networks obtained by combining different sources
  - Comparison with existing approaches
  - Annotation of discovered alignments

- Statistical significance
  - Probabilistic analysis of density & conservation in power-law graphs

- Web server for PPI network alignment