20.453J / 2.771J / HST.958J Biomedical Information Technology
Fall 2008

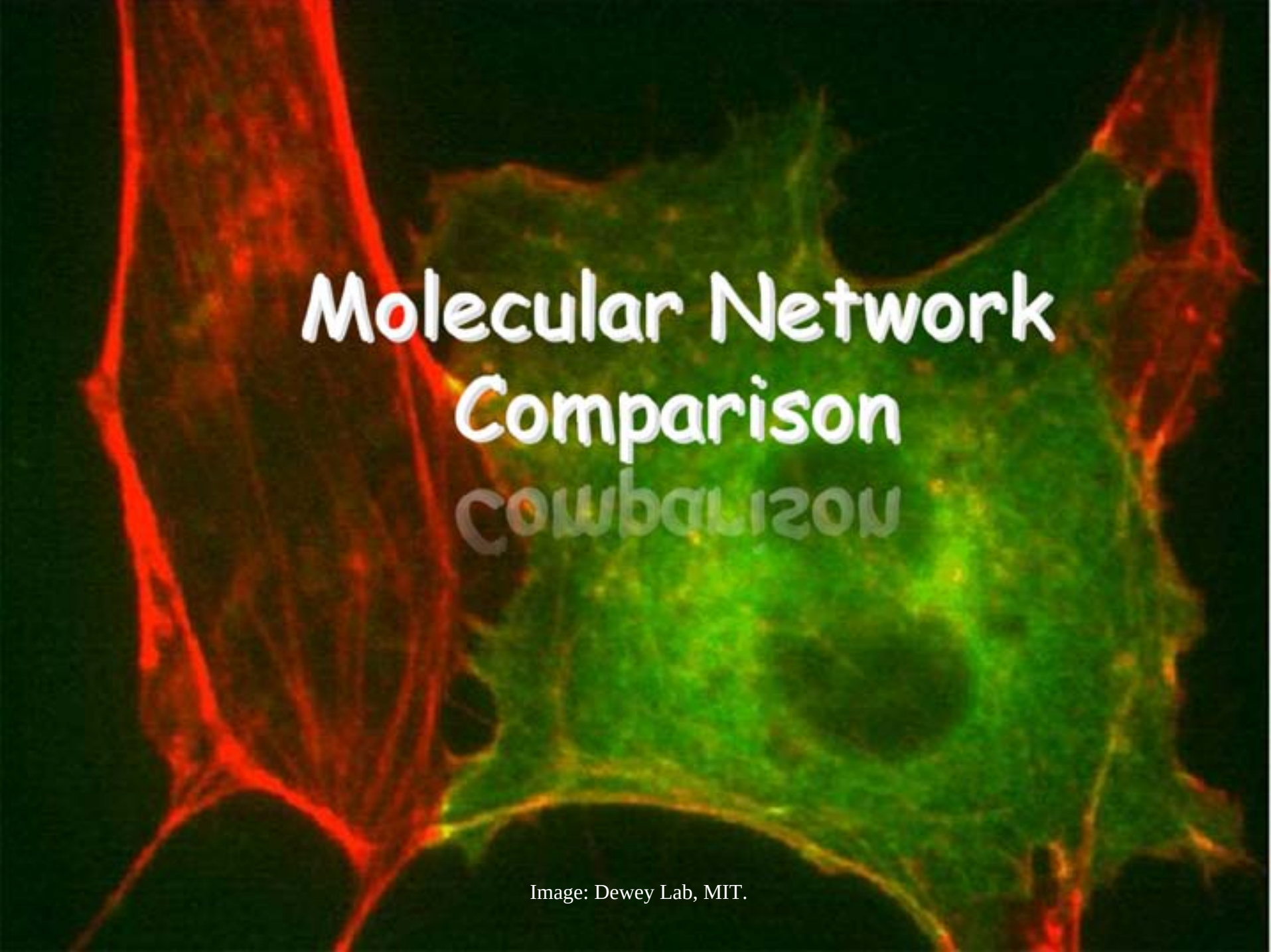
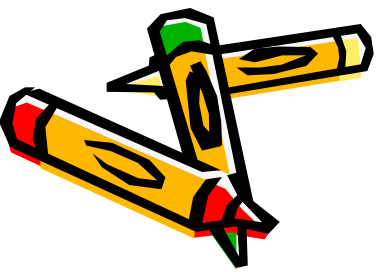


Image: Dewey Lab, MIT.

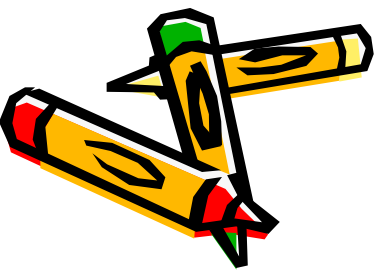
Nothing in Biology Makes Sense Except in the Light of Evolution

Theodosius Dobzhansky (1900-1975)

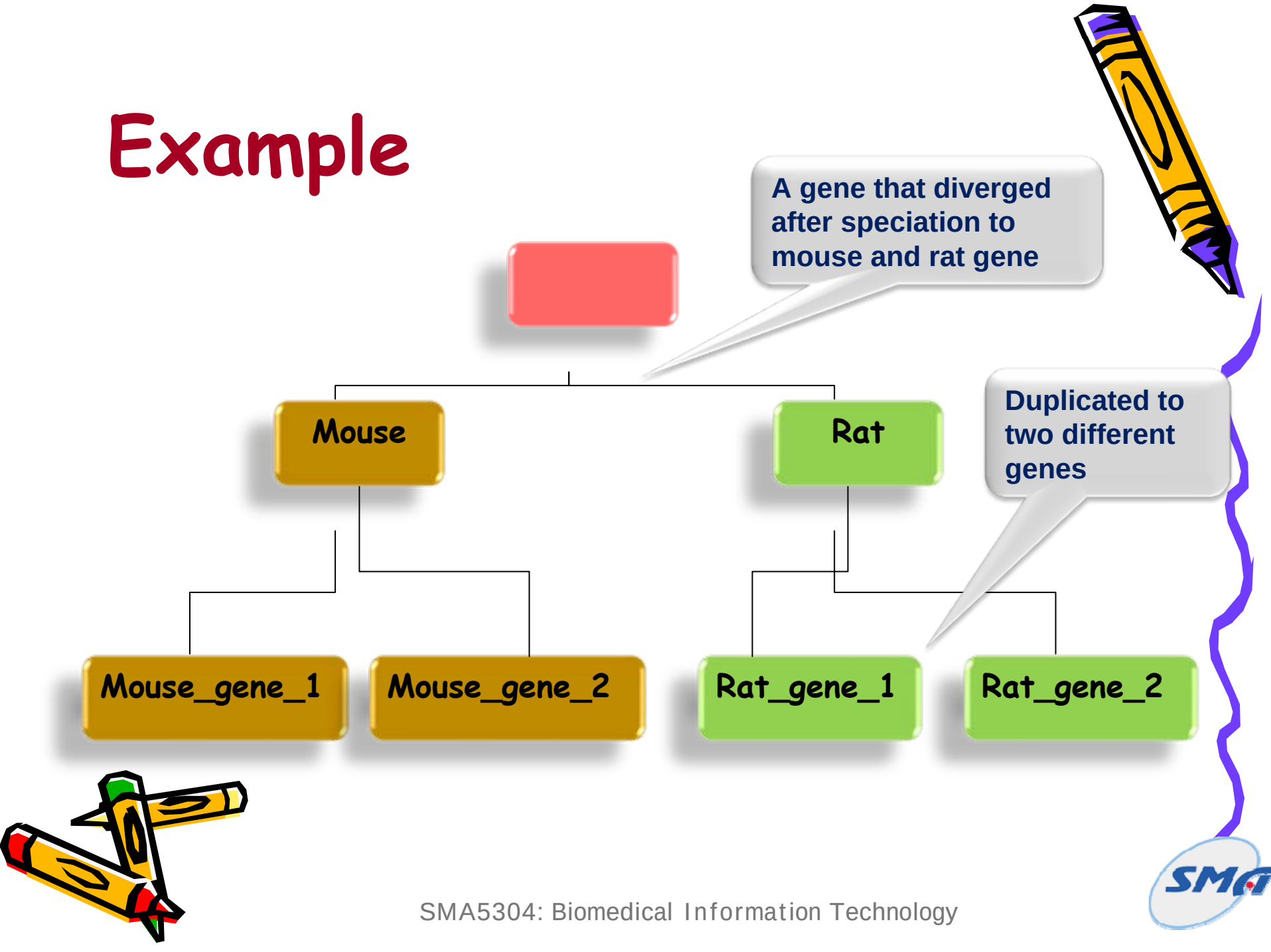


Evolution in Biology

- Mutations and rearrangement in genomic DNA
  - Changes in protein structures, abundances, and modification states
- Variation at the protein level
  - Impact proteins interaction with one another, with DNA, and with small molecules
  - Affects signaling, regulatory, and metabolic networks
- Changes in network organization
  - Affects cellular function, tissue-level responses, behavior and morphology of whole organisms

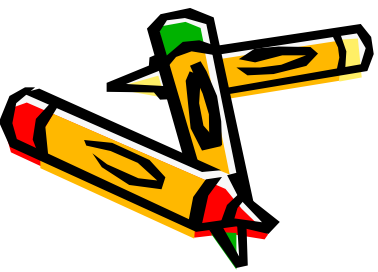


Example


A gene that diverged after speciation to mouse and rat gene
Mouse
Rat
Duplicated to two different genes
Mouse_gene_1
Mouse_gene_2
Rat_gene_1
Rat_gene_2

Metrics for Evolutionary Change

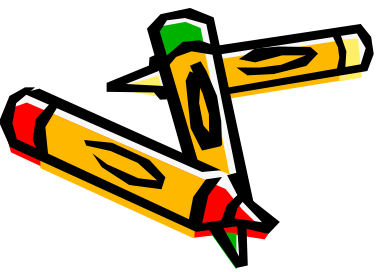
- Gene and protein sequences
- Why?
  - Fundamental level of biological variation
  - Readily available through automated sequence technology
- Network of protein interactions as metric?



Protein-Protein Interaction

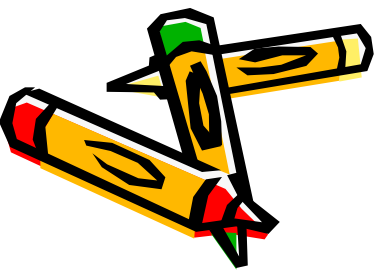
- Often, proteins interact with other proteins to perform their functions
- Backbone of molecular activity within the cell
  - Cell can be understood as a complex network of interacting proteins

Image removed due to copyright restrictions.
See: "MAPK/ERK in Growth and Differentiation."
http://focosi.altervista.org/mapkmap2.html.



Why?

- Increased complexity (function) is not explained simply by variations in gene (or protein) count
- Exponential growth in data
  - Mass spectrometry
  - Genome-wide-chromatin immunoprecipation
  - Yeast two-hybrid assays
  - Combinatorial reverse genetic screens
  - Literature mining



Questions in evolutionary and comparative biology

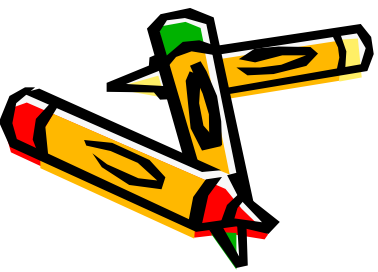
Given that protein sequences and structures are conserved in and among species, are networks of protein interactions conserved as well?

Is there some minimal set of interaction pathways required for all species?

Can we measure evolutionary distance at the level of network connectivity rather than at the level of DNA or protein sequence?

Molecular Network Comparison

Process of contrasting two or more interaction networks, representing different species, conditions, interaction types or time points.



Answer To

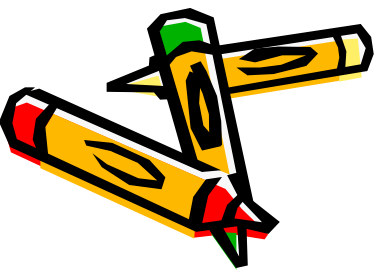
Why proteins, protein interactions and groups of interactions are likely to have equivalent functions across species?

Based on these similarities, can we predict new functional information about proteins and interactions that are poorly characterized?

What do these relationships tell us about the evolution of proteins, networks and whole species?

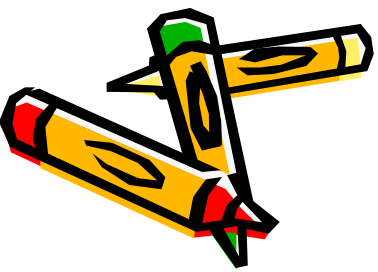
Answer To

Given that systematic screens for protein interactions may report large numbers of false-positive measurements, which interactions represent true binding events?

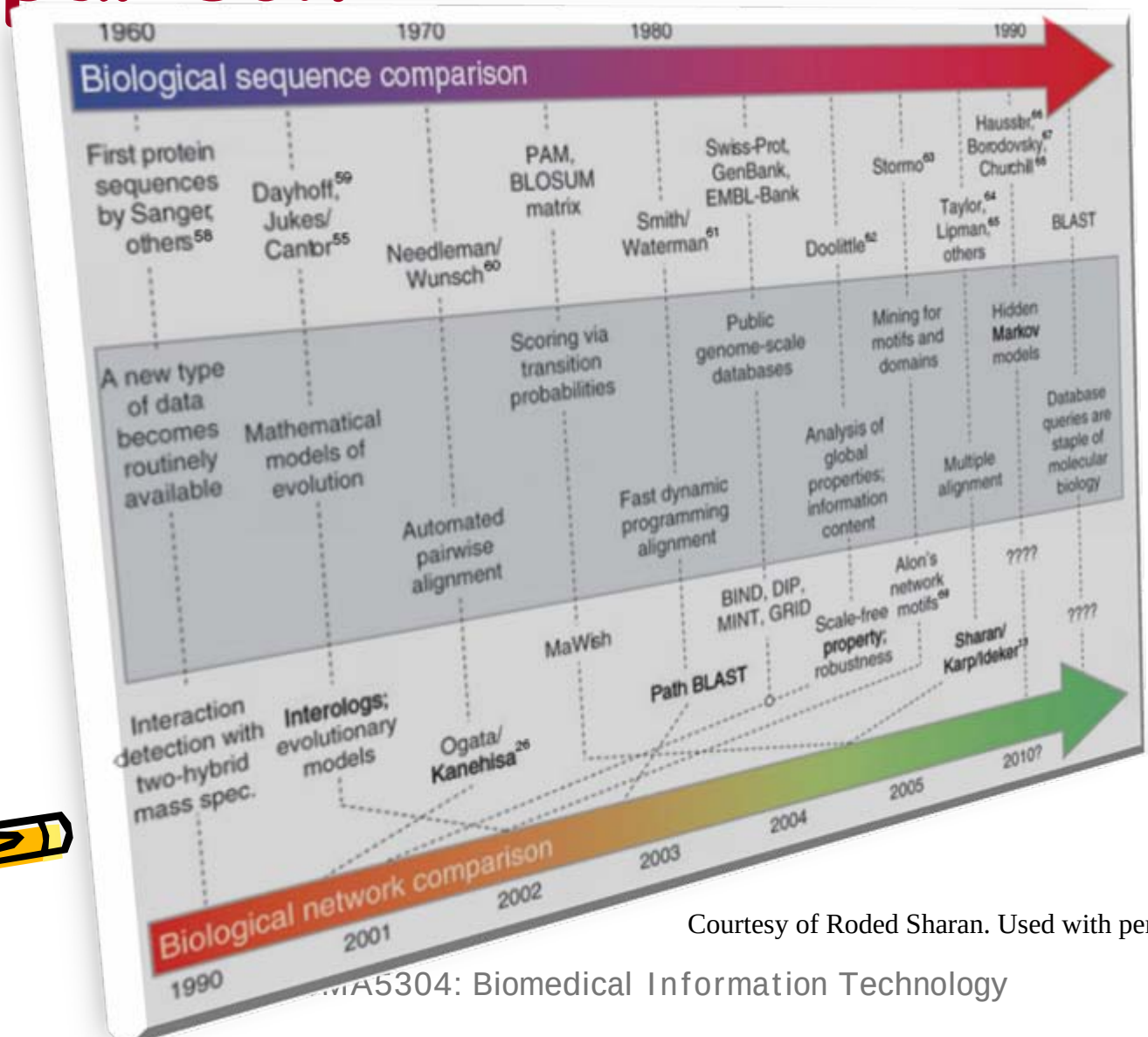


Types of Network Comparison

- Network alignment
  - process of globally comparing two networks, identifying regions of similarity and dissimilarity
- Network integration
  - process of combining several networks, encompassing interactions of different types over the same set of elements, to study their interrelations.
- Network querying
  - a given network is searched for subnetworks that are similar to a subnetwork query of interest



Network & Sequence Comparison



Courtesy of Roded Sharan. Used with permission.

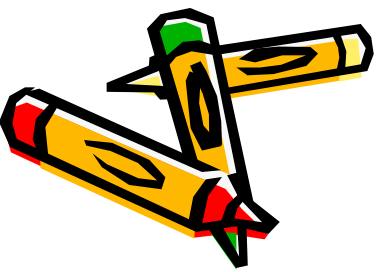
Network & Sequence Comparison

Sequence Comparison

- Sequence alignment methods were proposed long before large sequence databases were widely available.

Network Comparison

- Large network and interaction databases have been available from the late 1990s onwards, three to four years before the first network comparisons were performed.



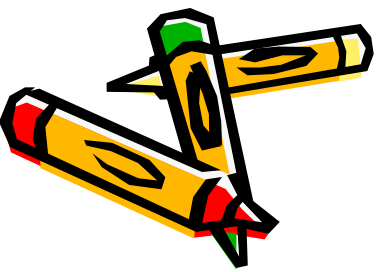
Network & Sequence Comparison

Sequence Comparison

- Computational searches for motifs and systematic characterization of global properties arose relatively late in the history of sequence analysis.
- Local sequence alignment can be solved efficiently

Network Comparison

- Occurred early in the field of network comparison.
- Analogous problem of identifying conserved protein modules is computationally hard



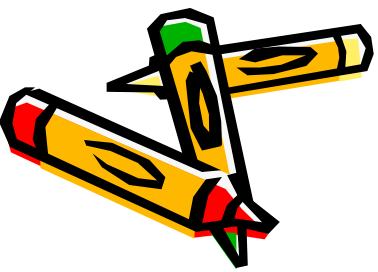
Network & Sequence Comparison

Sequence Comparison

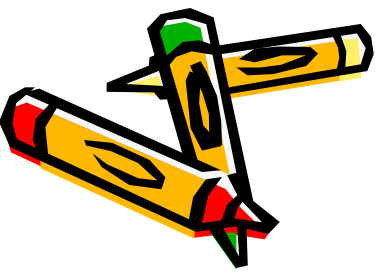
- Integrating biological sequences data types (nucleotides or amino acids) has not posed a major problem.

Network Comparison

- Integrating different data types of molecular networks is a challenging problem

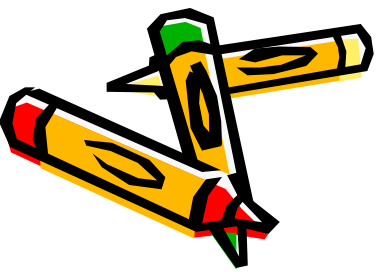


Network Alignment



Protein-Protein Interaction Network (PPI)

An undirected graph G(V, E) where V denotes the set of proteins and (u, v) $\in$ E denotes the interaction between proteins u and v



Network Alignment Problem

Given k different PPI networks belonging to different species, we wish to find conserved subnetworks within these networks

- The problem of finding conserved subnetworks in a group of networks is NP-Hard
- We can reduce it to subgraph-isomorphism

- Solution
  - A heuristic approach is required

Usefulness

- Functional Annotation
  - Assign roles to unknown proteins
- Annotation transfer
  - Assigns to a protein of unknown function the annotation of a protein to which it is aligned
- Landmark extension
  - If a protein of unknown function appears as part of an alignment together with a “landmark” protein of known function, we can label the protein with a similar annotation

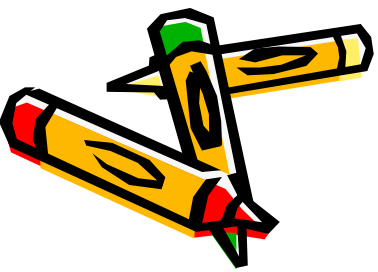
Usefulness

- Compute functional orthologs
  - Proteins which perform the same function across species
- Separating true protein-protein interactions from false positives
- Organizing large-scale interaction data into models of cellular signaling and regulatory machinery
- Identify novel modules by detecting unusual conserved subnetworks

Core Problems

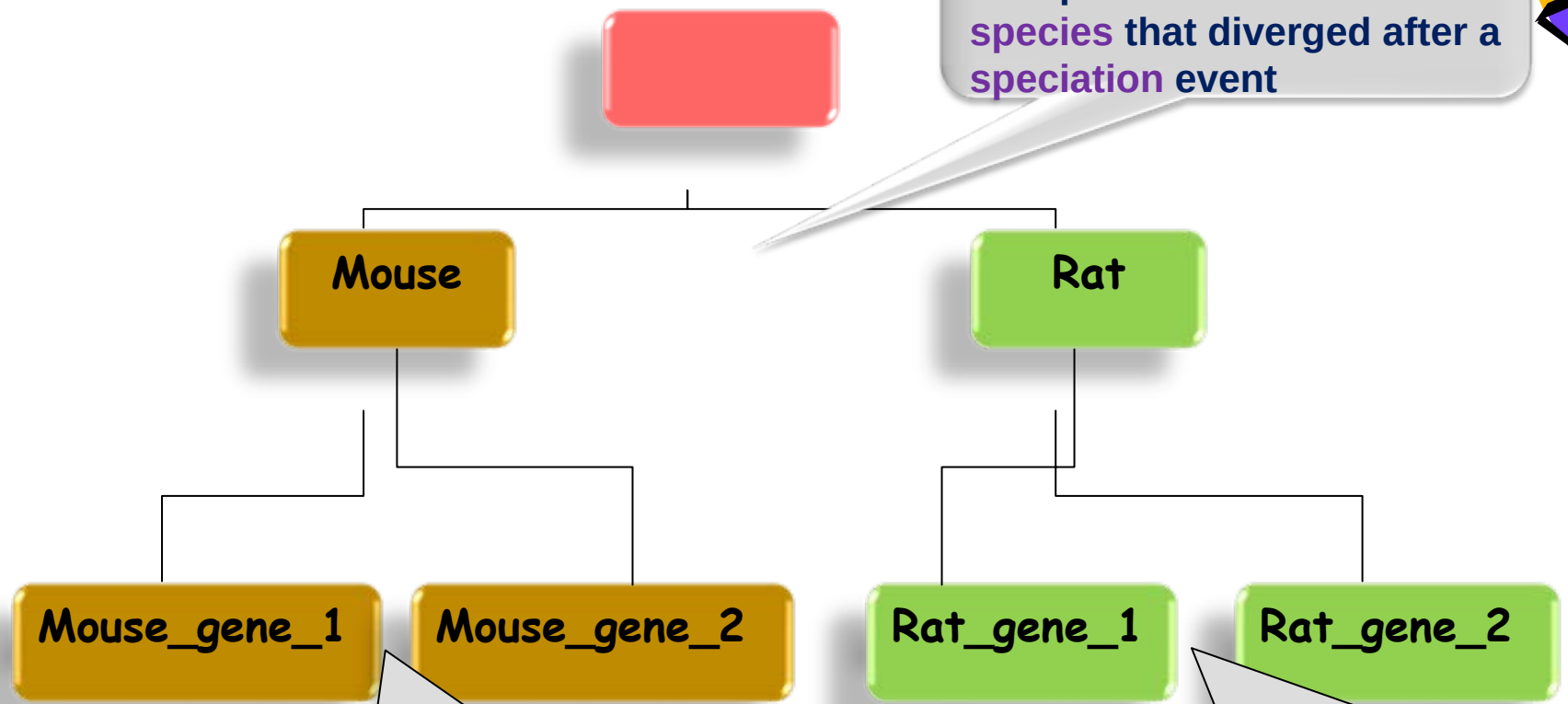
A scoring framework that captures the knowledge about module evolution

Identify high scoring alignments (conserved functional modules) from among exponentially large set of possible alignments



Protein Similarity

Orthologous proteins – Two proteins from different species that diverged after a speciation event



Paralogous proteins – Two proteins from the same species that diverged after a duplication event

Homologous proteins – Two proteins that have common ancestry

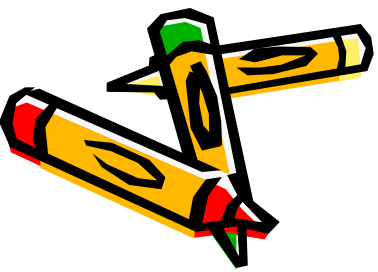
- Detected by sequence similarity
- Proteins can be from same or different species

Overview

- B. P. Kelley et al. **Pathblast: a tool for alignment of protein interaction networks.** *PNAS.*, 2003.
- M. Chen, R. Hofestaedt. **PathAligner: Metabolic Pathway Retrieval and Alignment.** *Applied Bioinformatics*, 3(4), 241–252, 2004.
- R. Sharan, S. Suthram, R. M. Kelley et al. **Conserved patterns of protein interaction in multiple species.** *PNAS*, 102, pp. 1974-1979, 2005.
- R. Y. Pinter, O. Rokhlenko, E. Yeger-Lotem, and M. Ziv-Ukelson. **Alignment of metabolic pathways.** *Bioinformatics*, 21(16):3401–8, 2005.
- M Koyuturk, Y Kim et al. **Pairwise Alignment of Protein Interaction Networks**. *Journal of Comp Biology,* 13(2), 2006.
- J. Flannick, A. Novak, B.S. Srinivasan et al. **Græmlin: General and robust alignment of multiple large interaction networks.** *Genome Research*, 2006.
- Dutkowsky, J., Tiuryn, J.: **Identication of functional modules from conserved ancestral protein-protein interactions.** Bioinformatics, 2007.
- N Przulj. **Biological Network Comparison with Graphlet Degree Distribution.** Bioinformatics, 23, e177-e183, 2007.
- M. Kalaev, V. Bafna, R. Sharan. **Fast and Accurate Alignment of Multiple Protein Networks.** *In Proc. of ACM RECOMB*, 2008.

PathBlast

- Implements a scoring function and search algorithm to find high probability pathway alignments between two protein interaction networks

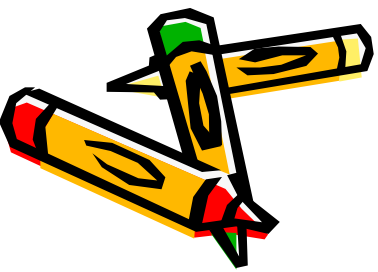


Pathway Alignment

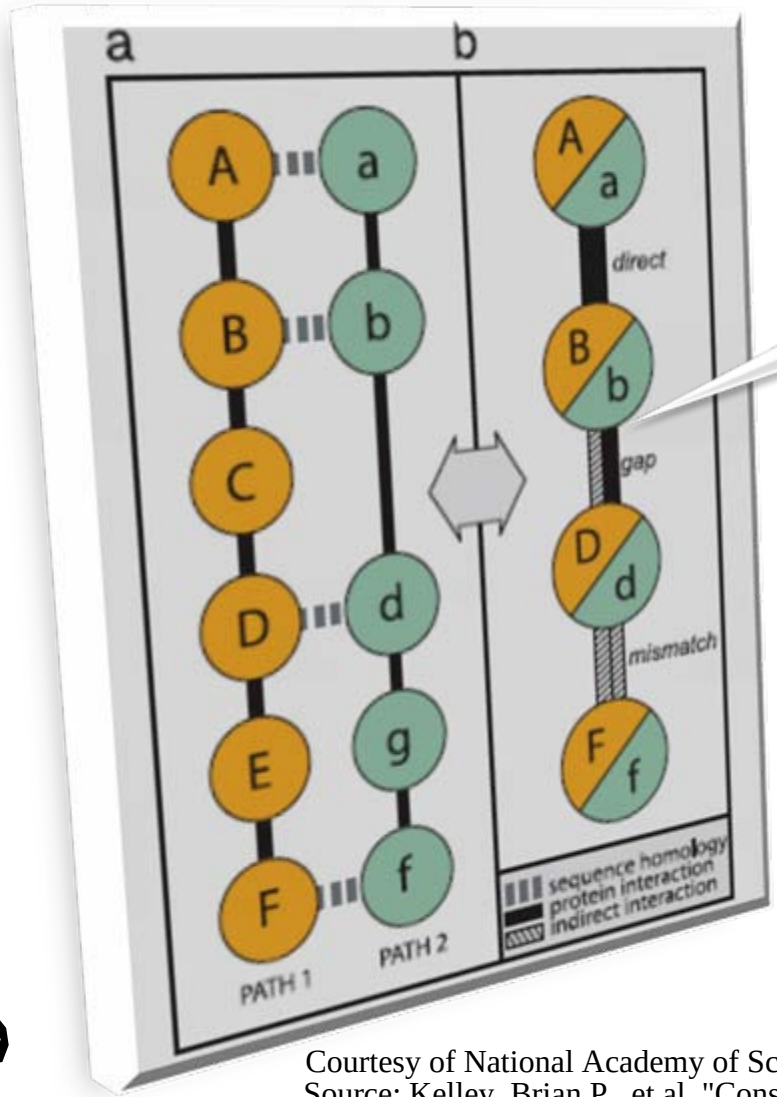
- Consists of two paths one from each network (N1 and N2)
- Proteins in the first path pair with putative homologs occurring in the same order in the second path
  - A homologous protein pair may not occur more than once per pathway alignment
- May include nonhomologous proteins
  - Using gaps and mismatches

Gaps and Mismatches

- Why?
  - Overcome noisy PPI data as well as evolutionary variations
- Gap
  - Occurs when a protein interaction in one path skips over a protein in the other
- Mismatch
  - Occurs when two proteins at the same position in the alignment do not share sequence homology
- Neither gaps nor mismatches may occur consecutively



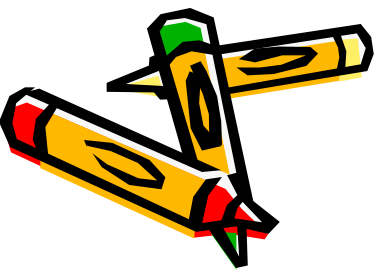
Global Alignment Graph



A path through this represents a conserved pathway between the two networks

Similarity Between Proteins

- In order to build the global alignment graph, similarity between proteins need to be measured
  - Using BLAST
- Quantifies the similarity
- Assigns it a p-value
  - probability of observing such similarity at random
- E-value or Expectation value
  - number of different sequence pairs with score equivalent or better than this hit's score that are expected to result by a random search
  - Unalignable proteins are assigned max E-value of 5



Log Probability Score

Probability of true homology within the protein pair represented by v in P

Probability that the protein-protein interaction represented by e is real (not false positive error)

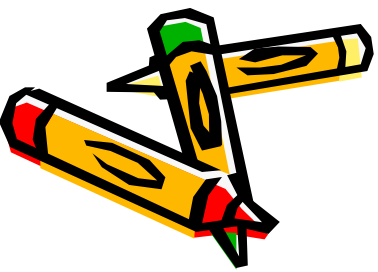
$$S(P) = \sum_{v \in P} \log_{10} \frac{p(v)}{p_{random}} + \sum_{e \in P} \log_{10} \frac{q(e)}{q_{random}},$$

Expected values of p(v) over all vertices and edges in G

Expected values of q(e) over all vertices and edges in G

Alignment Algorithm

- Identify the highest-scoring pathway alignment *P* of* fixed length *L (L vertices and L -1 edges).*
- *If G is directed and acyclic*,
  - linear time (in the number of edges)
  - using dynamic programming



Alignment Algorithm

- The highest-scoring path of length *l = 2…L ending in vertex v will have score*

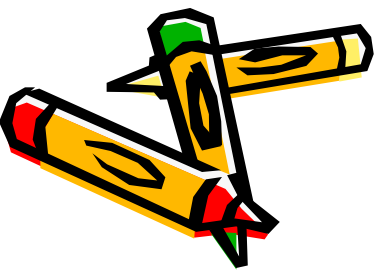
$$S(v,l) = \underset{u \in parents(v)}{\arg\max} \left[S(u,l-1) + \log \frac{p(v)}{p_{random}} + \log \frac{q(e_{u \to v})}{q_{random}} \right]$$

$$S(v,1) = \log \frac{p(v)}{p_{random}}$$

Base Case

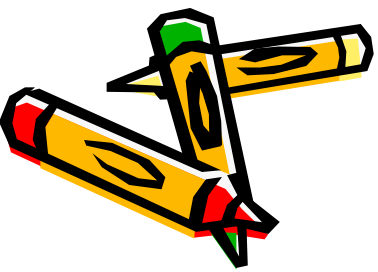
Alignment Algorithm

- G is generally not acyclic
  - First construct a sufficient number (5*L!*) *of* directed acyclic subgraphs
  - Use the dynamic programming method to compute the highest-scoring paths for each.
  - Can be done in linear time



Experiments

- Yeast (S. cerevisiae) vs. Bacteria (H. pylori)
  - Orthologous pathways between the networks of two species.
- Yeast vs. Yeast
  - Paralogous pathways within the network of a single species, by aligning the yeast PPI network versus itself.
- Yeast vs. Yeast
  - Interrogating the protein network with pathway queries, by aligning the yeast PPI network versus simple pathways.



Comparison between Yeast and Bacteria GAGs to the corresponding randomized networks

Graph size and best pathway-alignment scores were significantly larger for the real aligned networks

	Vertices (homologs)	Edges				CPU, min	Score	
		Total	Direct	Gap	Mismatch		Best*	Best 50†
Yeast vs. *H. pylori* ($E_{cutoff} = 10^{-2}$)	829	2,036	7	260	1,769	0.38	8.1	7.5
Random: mean ± SD		509.0 ± 128.0	2.5 ± 1.9	68.8 ± 23.8	437.7 ± 110.3	0.4 ± 0.02	6.1 ± 0.8	4.8 ± 0.7
Yeast vs. yeast ($E_{cutoff} = 10^{-10}$)	5,593	1,389	1,389	N/A	N/A	7.08	11.9	11.0
Random: mean ± SD		62.3 ± 29.4	62.3 ± 29.4	N/A	N/A	6.9 ± 0.2	−4.1 ± 9.5	−15.3 ± 6.5

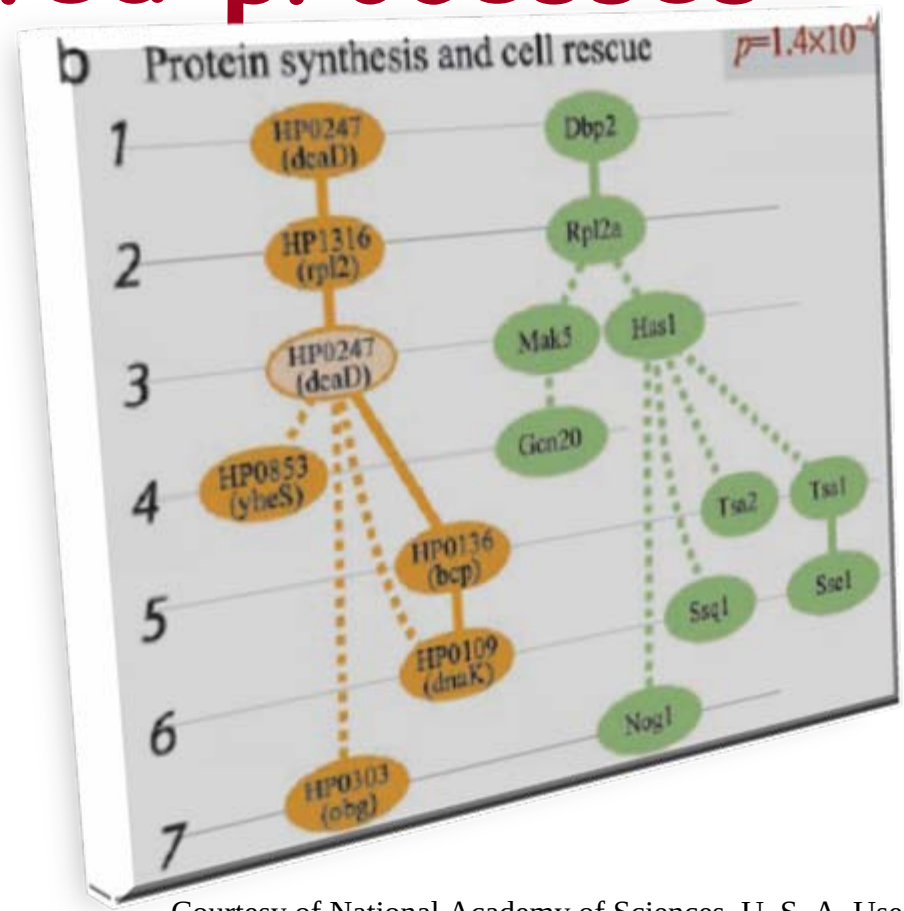
Both species indeed share conserved interaction pathways

Comparison between Yeast and Bacteria *GAGs* to the corresponding randomized networks

- Direct interaction was rare
- Gaps and Mismatch allowed to find much larger regions that were conserved

Vertices (homologs)	Edges			
	Total	Direct	Gap	Mismatch
829	2,036	7	260	1,769
5,593	509.0 ± 128.0	2.5 ± 1.9	68.8 ± 23.8	437.7 ± 110.3
	1,389	1,389	N/A	N/A
	62.3 ± 29.4	62.3 ± 29.4	N/A	N/A

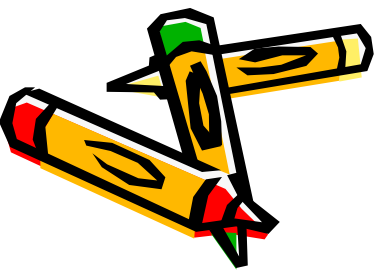
Relation between seemingly unrelated processes



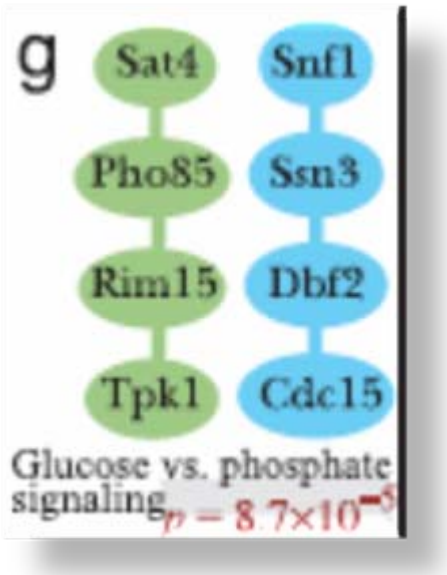
Courtesy of National Academy of Sciences, U. S. A. Used with permission. Source: Kelley, Brian P., et al. "Conserved Pathways Within Bacteria and Yeast as Revealed by Global Protein Network alignment." *PNAS* 100, no. 20 (September 20, 2003): 11304-11309. Copyright (c) 2003 National Academy of Sciences, U.S.A.

Yeast vs Yeast

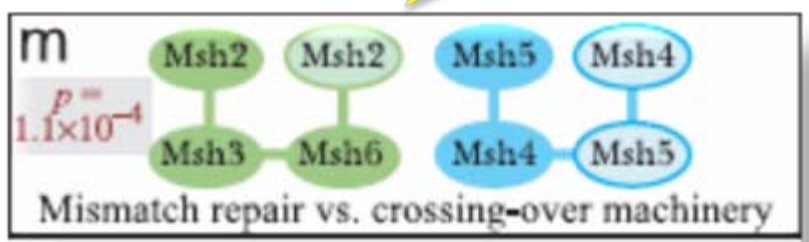
- Search for paralogous pathways
- Constructing a GAG merging the yeast protein interaction network with an identical copy of itself
  - Only direct edge permitted
  - Proteins were not allowed to pair with themselves or their network neighbors
  - Obtain 300 highest scoring pathway alignments of length 4 (level of significance p <= 0.0001)



Paralogous Pathways

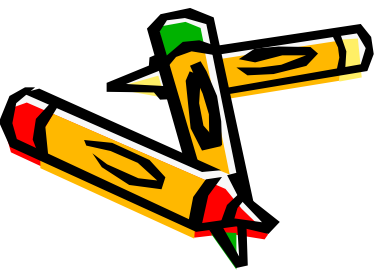


- **Both have DNA binding activity, they act in two distinct processes**
- **Msh2/3/6 is involved in mismatch repair during meiosis and vegetative growth**
- **Msh4/5 facilitates crossing over during homologous recombination and is specific to meiosis**



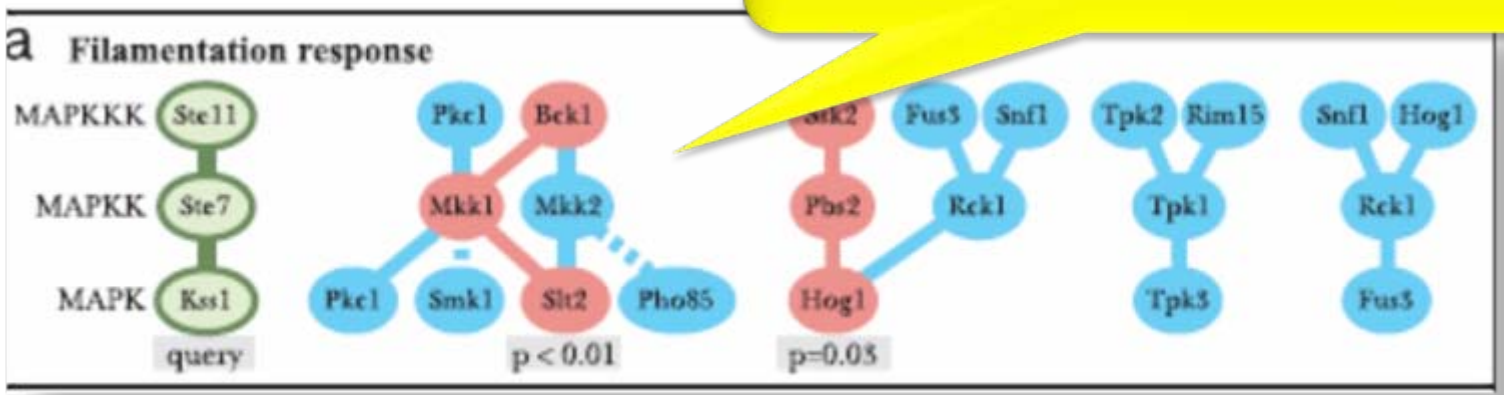
Pathway Queries

- Query a single protein network with specific pathways of interest
  - Similar to using BLAST to query a sequence database with a short nucleotide or amino acid sequence query
- Query Yeast network with a MAPK pathway associated with filamentation response

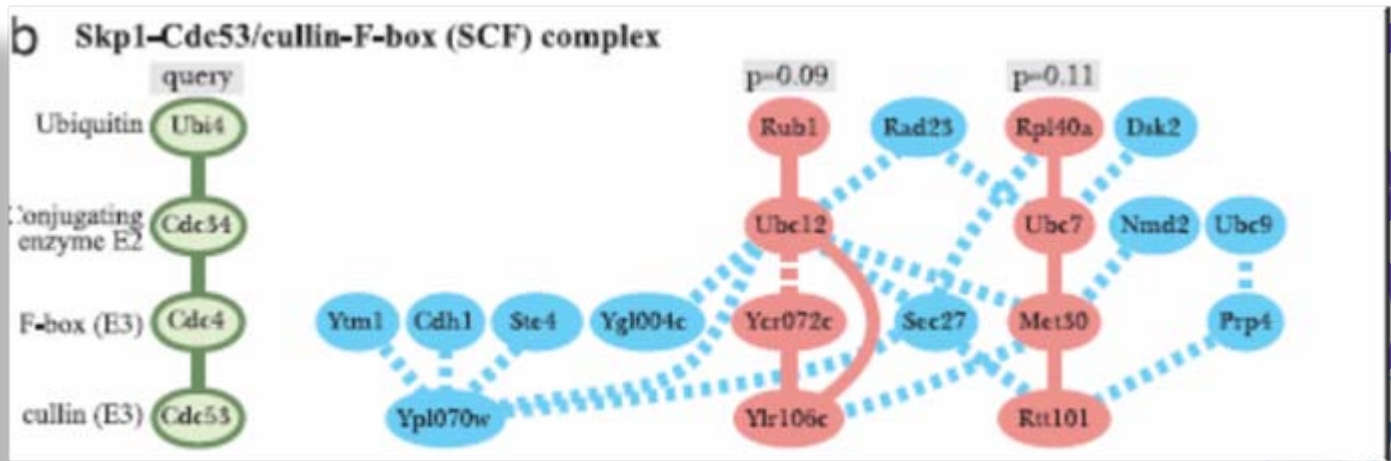


Querying Yeast Network

- Two other well known pathways
- low and high-osmolarity response pathways



Courtesy of National Academy of Sciences, U. S. A. Used with permission.
Source: Kelley, Brian P., et al. "Conserved Pathways Within Bacteria and Yeast as Revealed by Global Protein Network Alignment." *PNAS* 100, no. 20 (September 20, 2003): 11304-11309. Copyright (c) 2003 National Academy of Sciences, U.S.A.



Query: Ste20-Act1-Myo1

Alignment 1 6.835

Query	Match	Function
YHR023W (MYO1)	YKL129C	myosin I
YFL039C (ACT1)	YJR065C	actin-related gene
YKL007C (STE20)	YDR523C	dispensable for mitosis, involved in middle/late stage of meiosis, required for spore wall formation

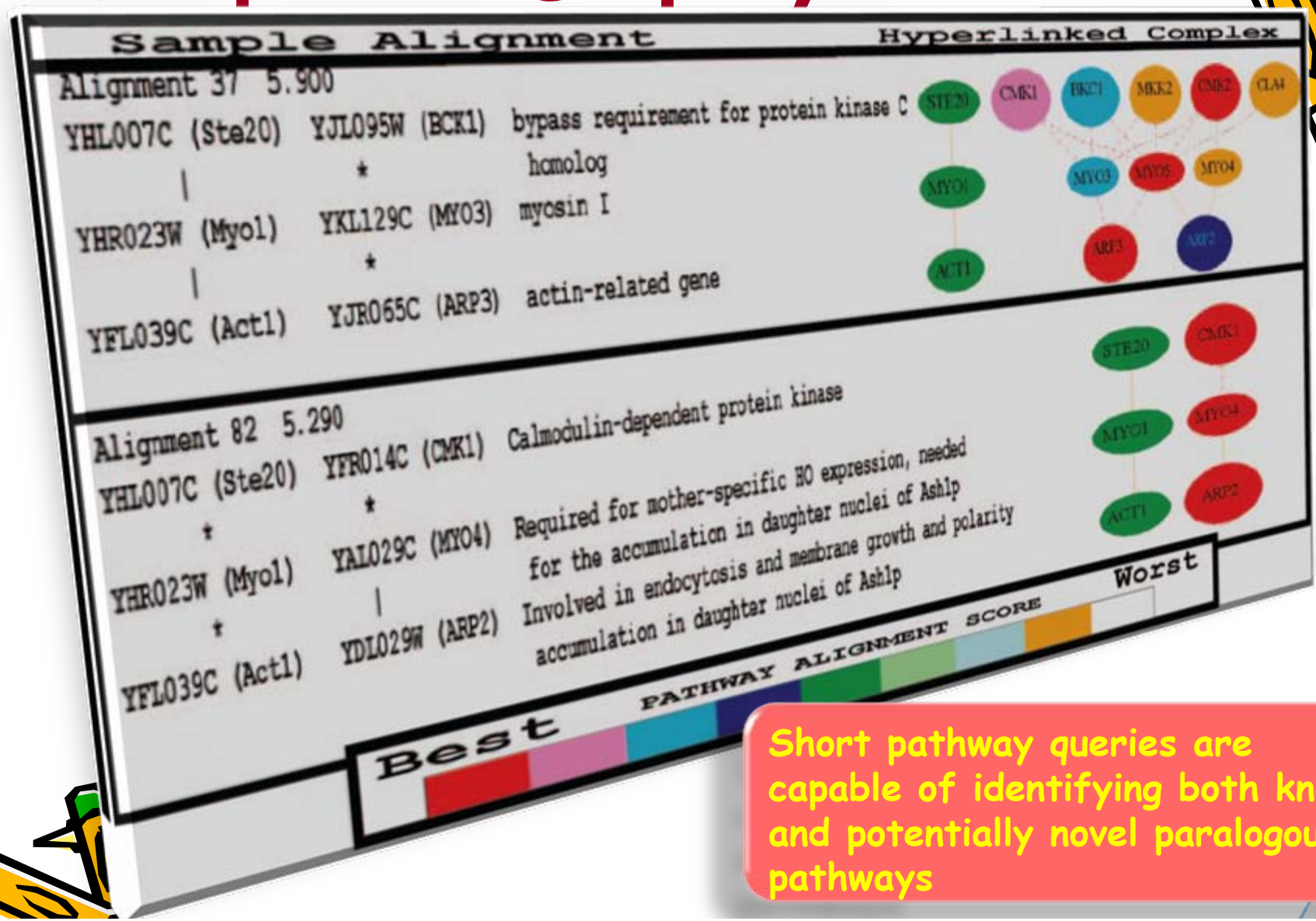
Alignment 2 6.835

Query	Match	Function
YKL007C (STE20)	YDR523C	dispensable for mitosis, involved in middle/late stage of meiosis, required for spore wall formation
YFL039C (ACT1)	YJR065C	actin-related gene
YHR023W (MYO1)	YKL129C	myosin I

Alignment 3 6.768

Query	Match	Function
YHR023W (MYO1)	YKL129C	myosin I
YFL039C (ACT1)	YJR065C	actin-related gene
YKL007C (STE20)	YPL140C	Member of MAP kinase pathway involving PKC1, BCK1, and SLT2. Shows functional redundancy with MKK1

Graphical Display

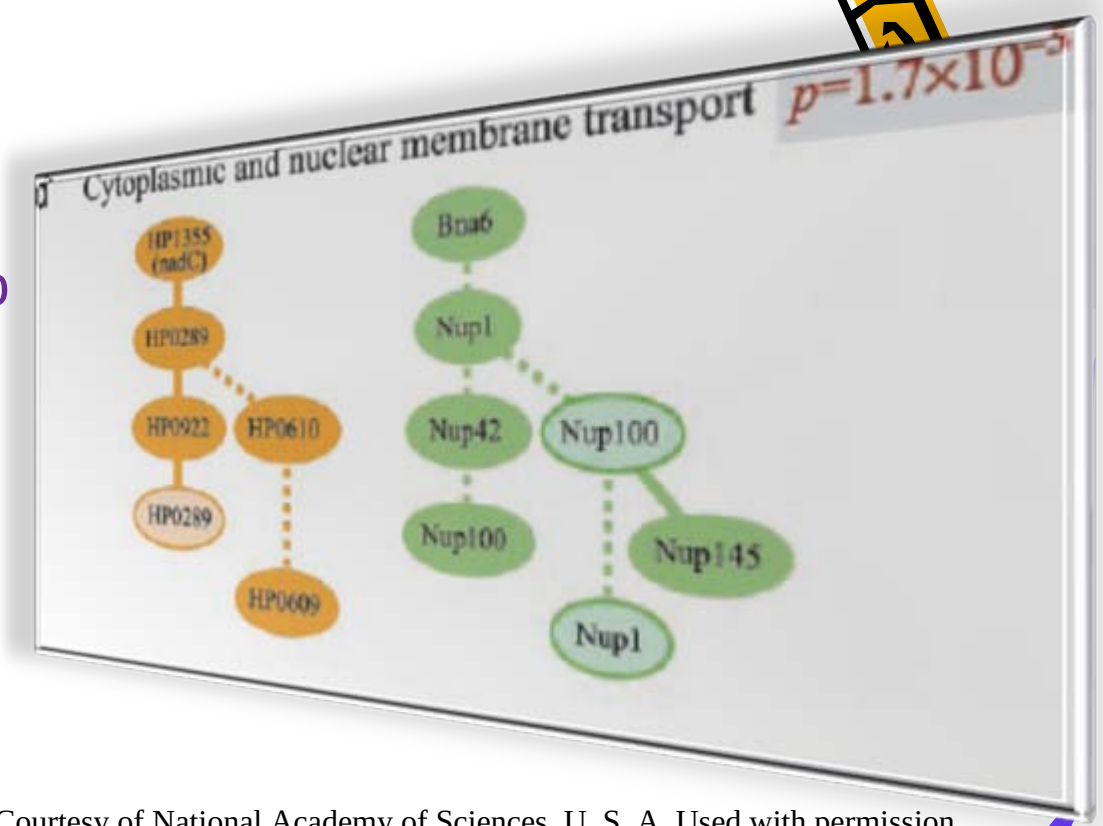


Short pathway queries are capable of identifying both known and potentially novel paralogous pathways

Image courtesy of Brian P. Kelley. Source: Kelley, B. P., et al "PathBLAST: A Tool for Alignment of Protein Interaction Networks." *Nucleic Acids Res* 32, web server issue, (July 1, 2004): W83-W88. © the authors, published by Oxford University Press under open access.

Summary

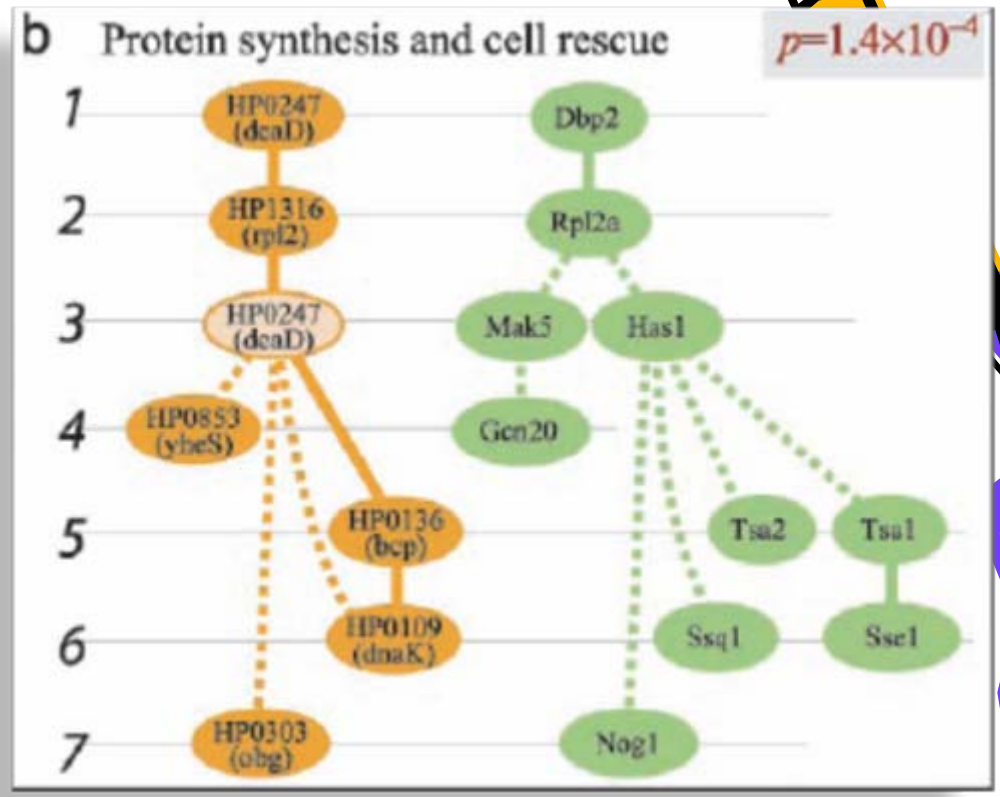
- Pathways from a well studied network is used to shed light on their aligned counterparts from a less well characterized ones
- HP0609 is adjacent to HP0610 and HP0289, which localize to the bacterial outer membrane, and opposite yeast Nup1, which localizes to the nuclear pore.



HP0609 is also membrane-specific and that the bacterial pathway shares homology with the yeast nuclear pore complex

Summary

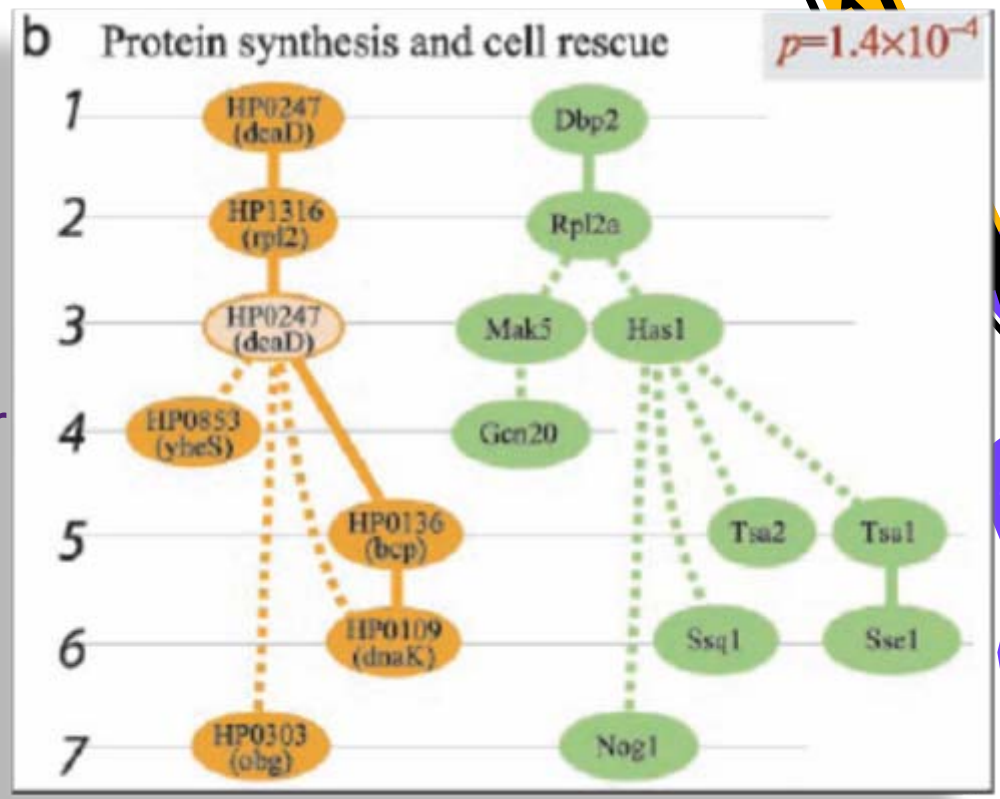
- Single pathways in bacteria frequently correspond to multiple pathways in yeast
- Yeast has undergone or more whole-genome duplications relative to bacteria

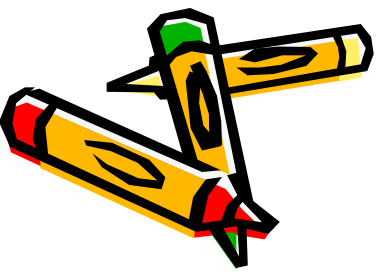


A single bacterial RNA helicase (deaD) occupies the same pathway position, and perhaps functional role, as three different helicases in yeast (Dbp2, Mak5, and Has1).

Summary

- Proteins within high-scoring pathway alignments did not necessarily pair with their best sequence matches in other pathway
- Bcp and Tsa1 are functional orthologs despite weak sequence similarity

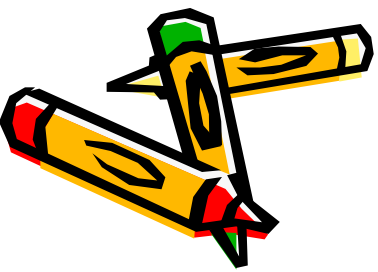




Best match for bcp is Dot5 and not Tsa1
Best match for Tsa1 is TsaA

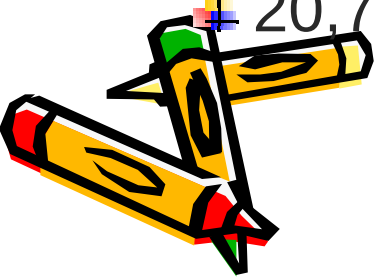
Limitations

- Proteins may occur more than once in an identified matched pathway
  - biologically implausible
- The algorithm provides limited support for identifying non-exact pathway matches
  - supporting no more than a single consecutive deletion of proteins from the query pathway
  - no more than a single consecutive insertion of proteins to the matched pathway
- The running time of the algorithm involves a factorial function of the pathway length



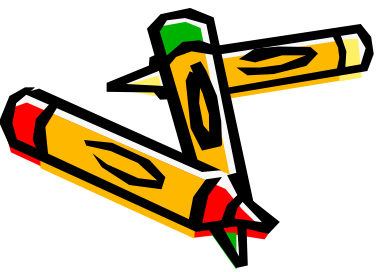
Multiple Alignment (Sharan et al, 2005)

- Three-way alignment of the protein–protein interaction networks
  - *Caenorhabditis elegans*
  - *Drosophila melanogaster*
  - *Saccharomyces cerevisiae.*
- Protein interaction data were obtained from the Database of Interacting Proteins
  - 14,319 interactions among 4,389 proteins in yeast
  - 3,926 interactions among 2,718 proteins in worm
  - 20,720 interactions among 7,038 proteins in fly

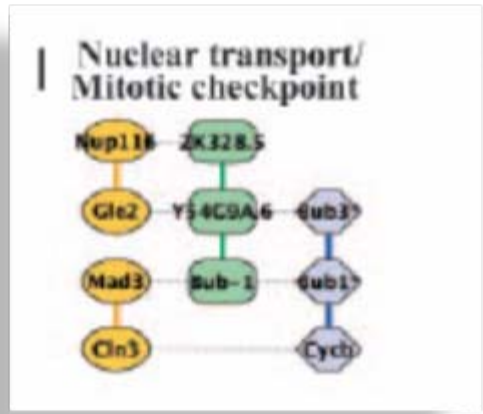


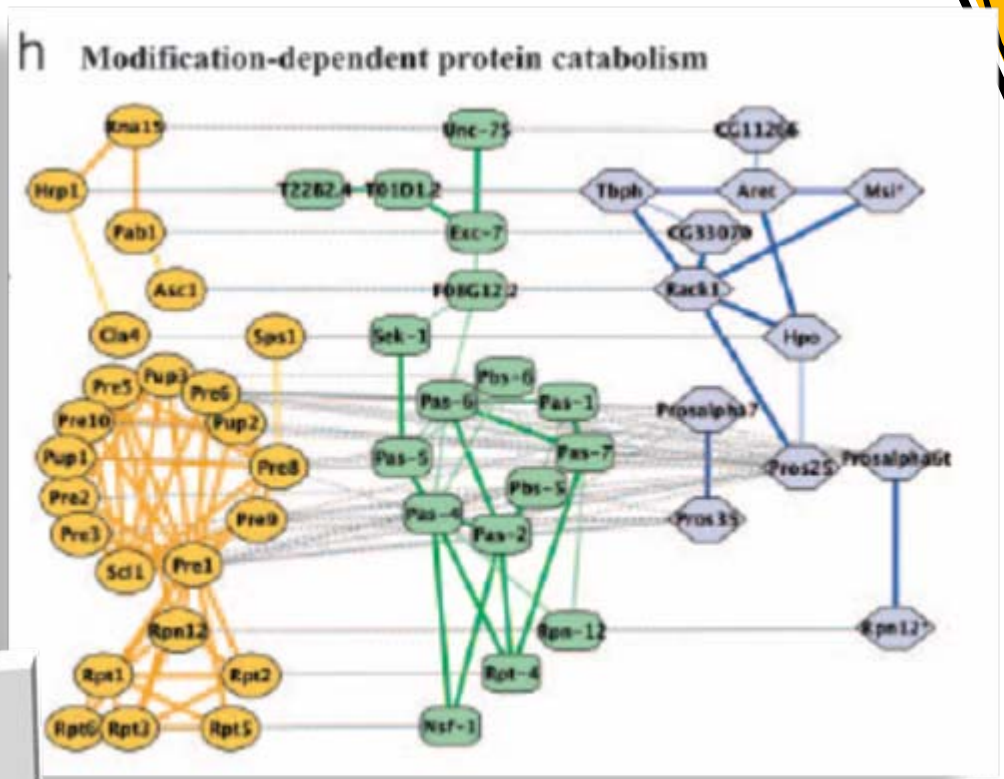
Experimental Results

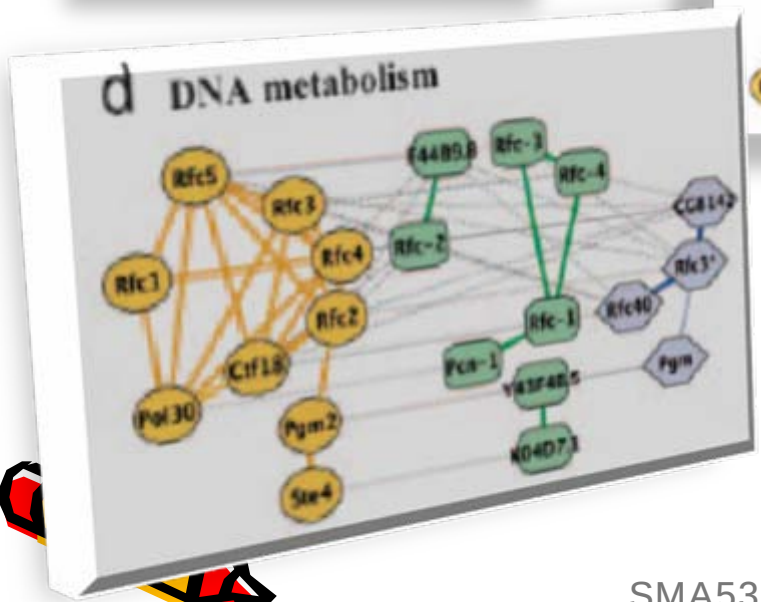
- Protein sequences obtained from
  - Saccharomyces Genome Database
  - WormBase
  - FlyBase
- Combined with the protein interaction data to generate a network alignment
  - 9,011 protein similarity groups
  - 49,688 conserved interactions for the three networks
- 183 protein clusters and 240 paths conserved at a significance level of $P < 0.01$



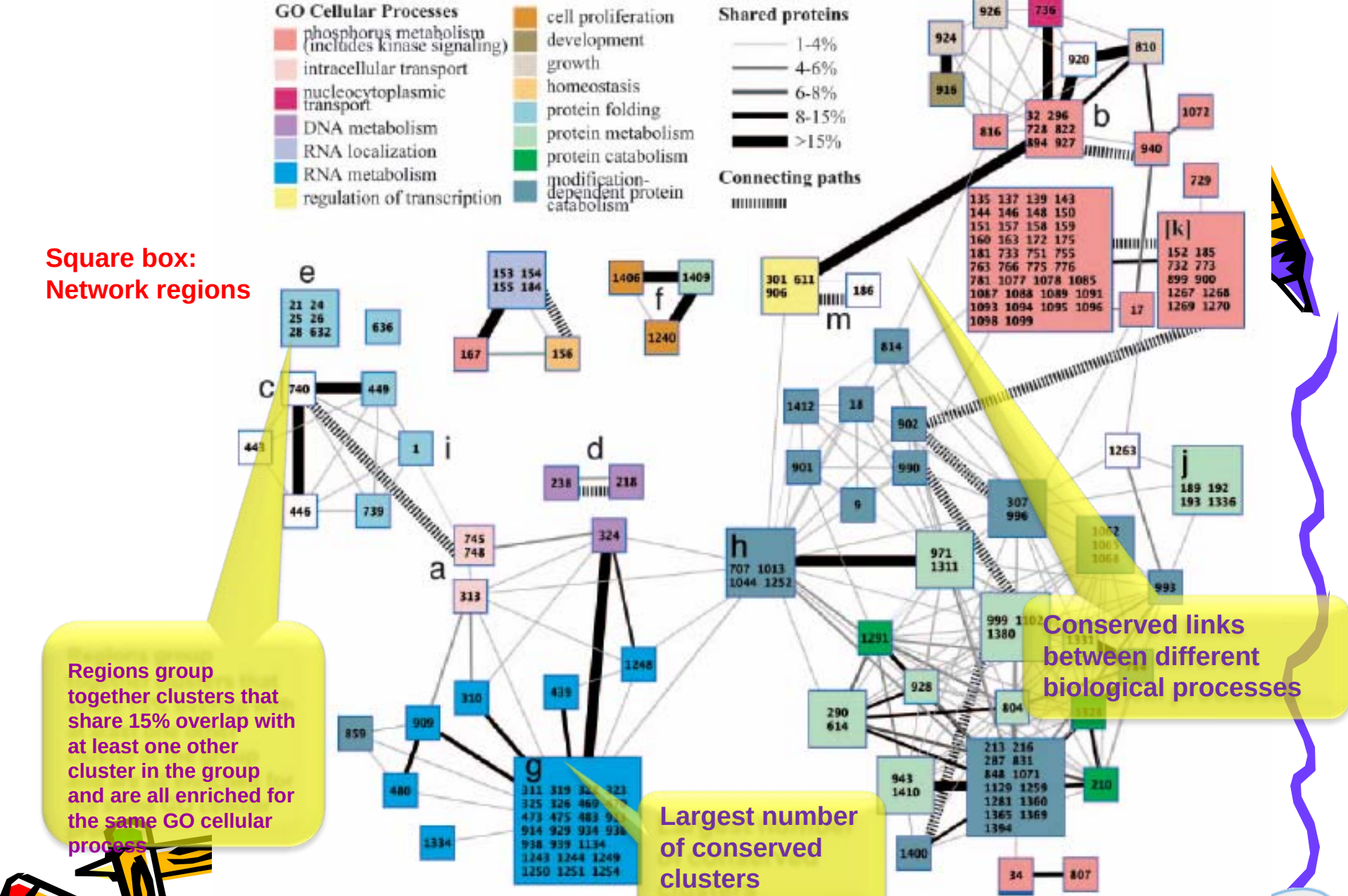
Conserved Network Regions





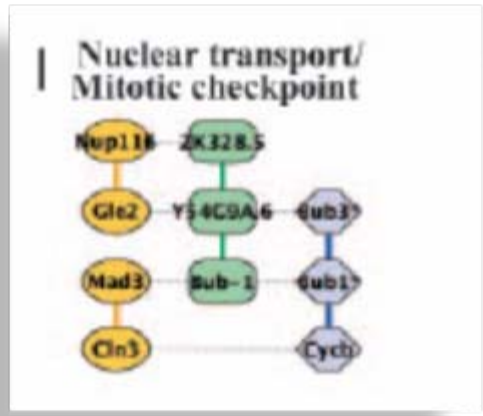


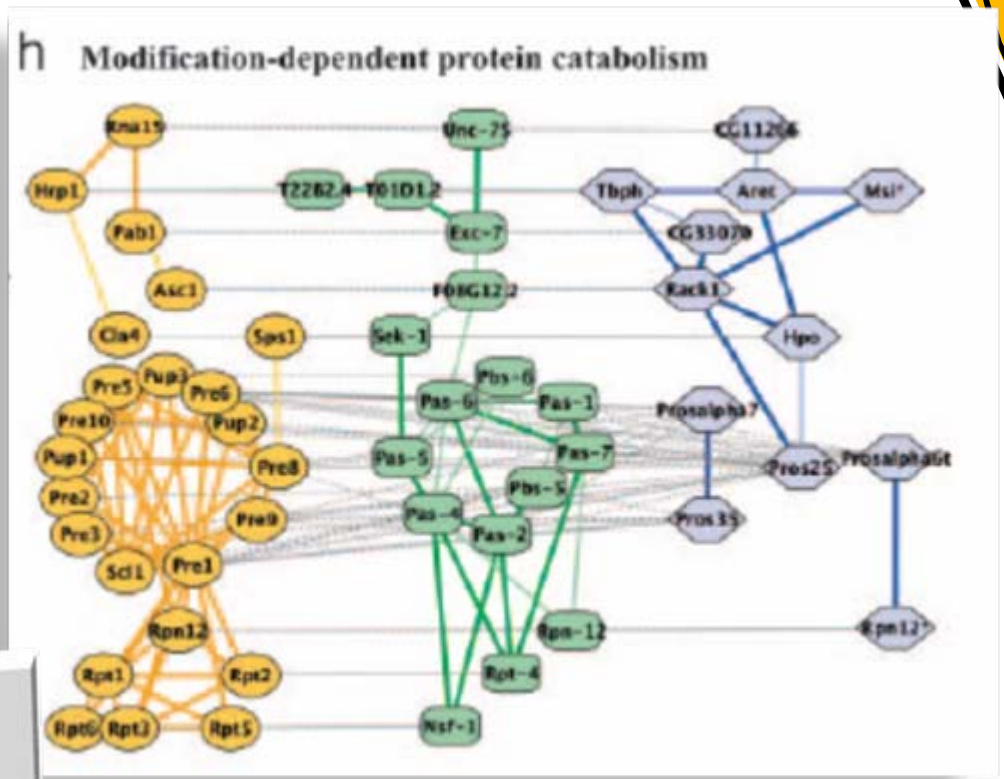
Orange oval - Yeast
Green rectangle - Worm
Blue hexagon - Fly

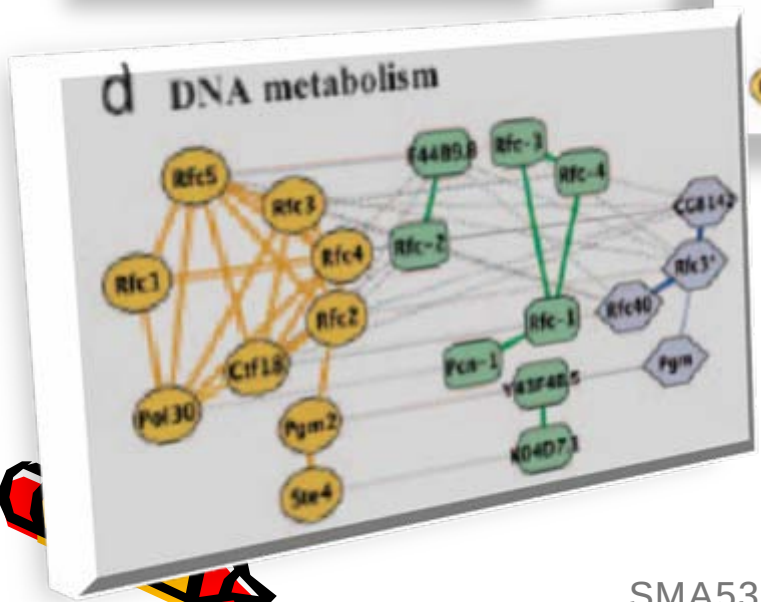


Courtesy of National Academy of Sciences, U. S. A. Used with permission.
Source: Sharan, R., et al. " Conserved Patterns of Protein Interaction in Multiple Species." *PNAS* 102, no. 6 (February 8, 2005): 1974-1979. Copyright (c) 2005 National Academy of Sciences, U.S.A.

Conserved Network Regions







Orange oval - Yeast
Green rectangle - Worm
Blue hexagon - Fly

Prediction of Protein Functions

Better than sequence-based methods (accuracy is 37-53%)

Species	No. correct	No. of predictions	Success rate, %
Yeast	114	198	58
Worm	57	95	60
Fly	115	184	63

For each species, the number of correct predictions, the total number of predictions, and the success rate in 10-fold cross-validation are listed.

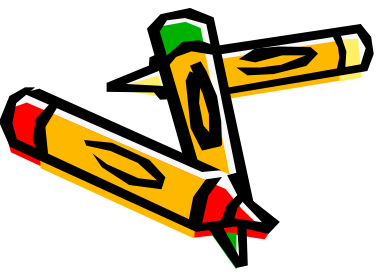
Two-hybrid tests of predicted interactions yielded a success rate in the range of 40-52%

Prediction of Proteins Interaction

Species	Sensitivity, %	Specificity, %	*P* value	Strategy
Yeast	50	77	1.1e-25	*i*
Worm	43	82	1e-13	*i*
Fly	23	84	5.3e-5	*i*
Yeast	9	99	1.2e-6	*i* + *ii*
Worm	10	100	6e-4	*i* + *ii*
Fly	0.4	100	0.5	i + ii

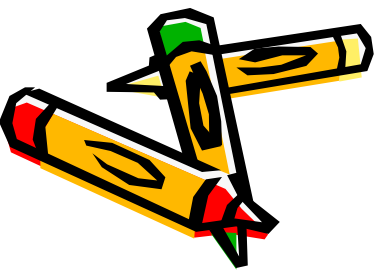
Prediction accuracy was highly significant

Network Querying



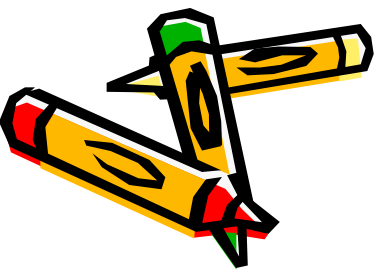
Overview

- T. Shlomi, D. Segal, E. Ruppin, and R. Sharan. **QPath: A Method for Querying Pathways in a Protein-Protein Interaction Network.** *BMC Bioinformatics*, 7(199), 2006.
- B. Dost, T. Shlomi, N. Gupta et al. **QNet: A Tool for Querying Protein Interaction Networks.** *In Proc. of ACM RECOMB*, 2007.
- Y. Tian, R. C. McEachin, C. Santos et al. **SAGA: A subgraph matching tool for biological graphs.** *Bioinformatics*, 2006.



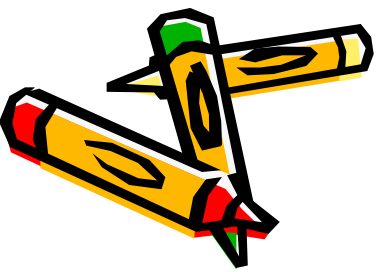
QPath

- Goal
  - Querying linear pathways within a given network
- QPath
  - Searches for matching pathways composed of distinct proteins that are similar to the query proteins in their sequence and interaction patterns



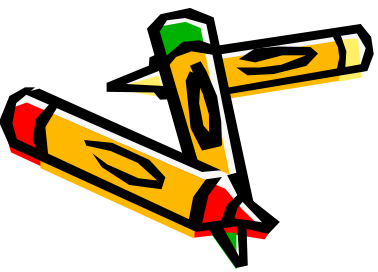
Path Query Problem

- Input
  - a target network, represented as an undirected weighted graph $G(V,E)$ with a weight function on the edges $w : E \times E \rightarrow R$
  - a path query $Q = (q_1,\ldots,q_k)$.
  - Additionally, a scoring function $H : Q \times V$ is given.



Path Query Problem

- Output
  - A set of best matching pathways P = $(p_1,\ldots,p_k)$ in G, where a good match is measured in two respects
- Each node in the matched pathway and its corresponding node in the query are similar with respect to the given scoring function H.
- The reliability of edges in the matched pathway is high



Example of Alignment

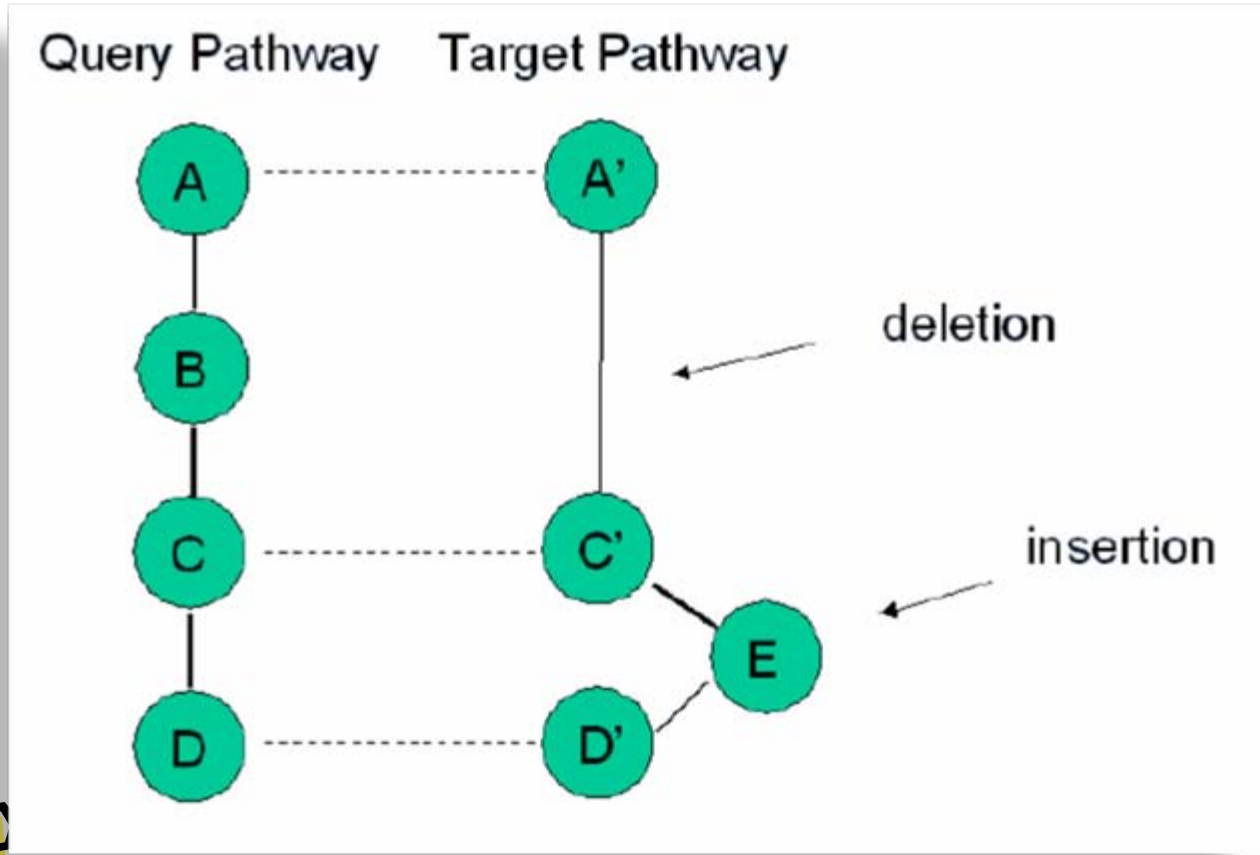


Image source: Figure 1b in Shlomi, T., et al. "QPath: A Method for Querying Pathways in a Protein-protein Interaction Network." *BMC Bioinformatics* 7 (2006): 199.

Evaluation of Pathway Queries

- Queried the yeast network with the yeast filamentous growth MAPK cascade.

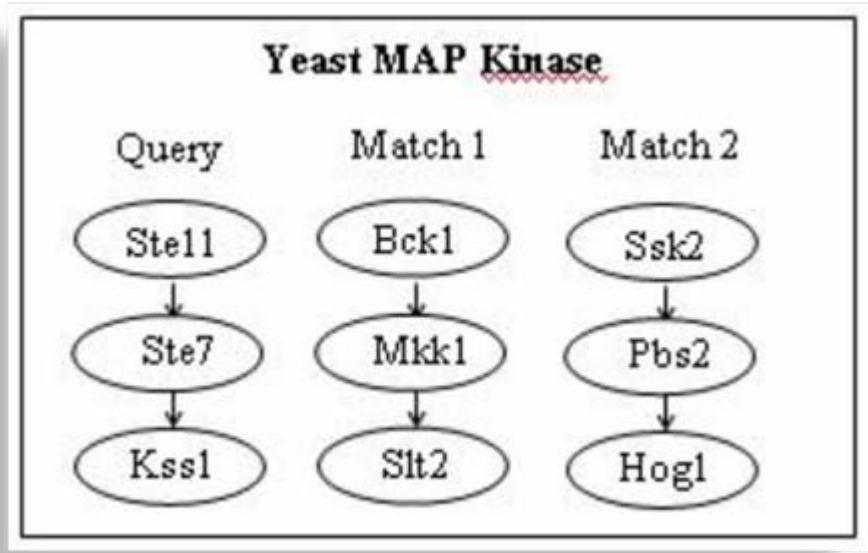
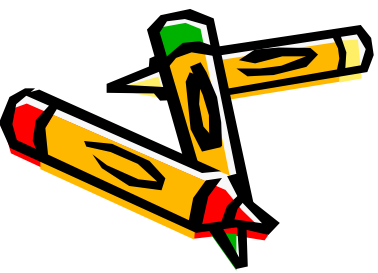


Image source: Figure 6 (supplemental material) in Shlomi, T., et al. "QPath: A Method for Querying Pathways in a Protein-protein Interaction Network." *BMC Bioinformatics* 7 (2006): 199.

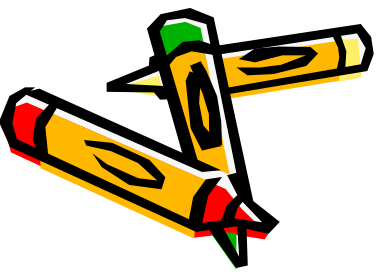
Evaluation of QPath

- Modified QPath algorithm is used to search the network for pathways that have high interaction scores
  - Limited to pathways consisting of 6 proteins
  - Allow for (up to 3) insertions and deletions.
- Identified a set of 271 non-redundant pathways whose scores exceeded those of 99% of randomly chosen pathways



Quality Assessment of the Pathways

- Functional enrichment
  - Representing the tendency of the pathway's proteins to have coherent GO functions
- Expression coherency
  - measuring the similarity in expression profiles of the pathway's coding genes across different experimental conditions

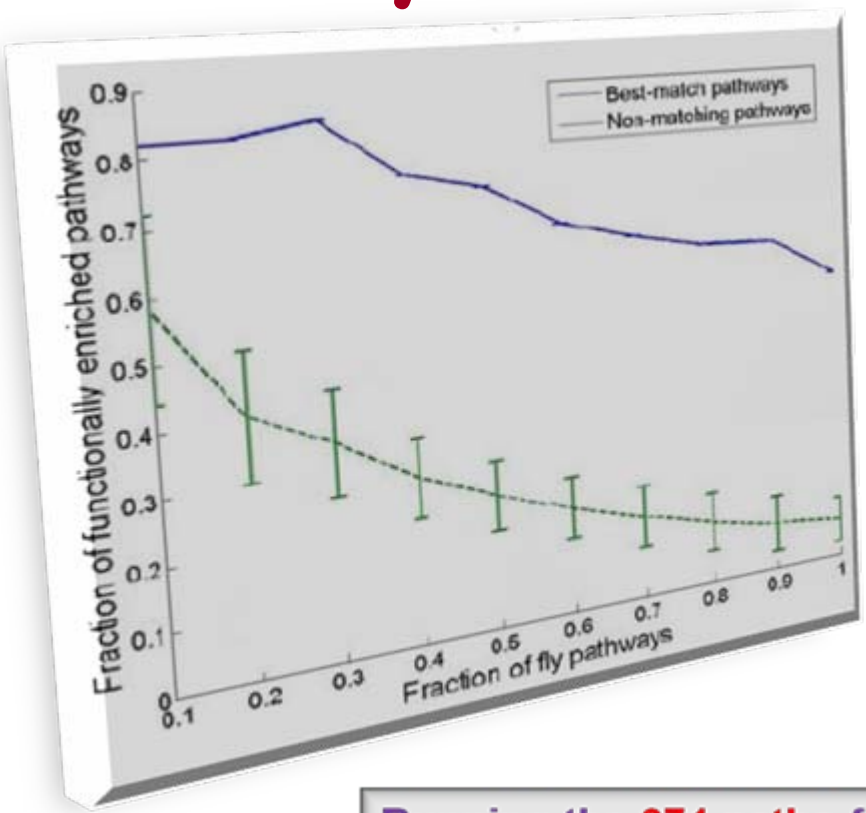


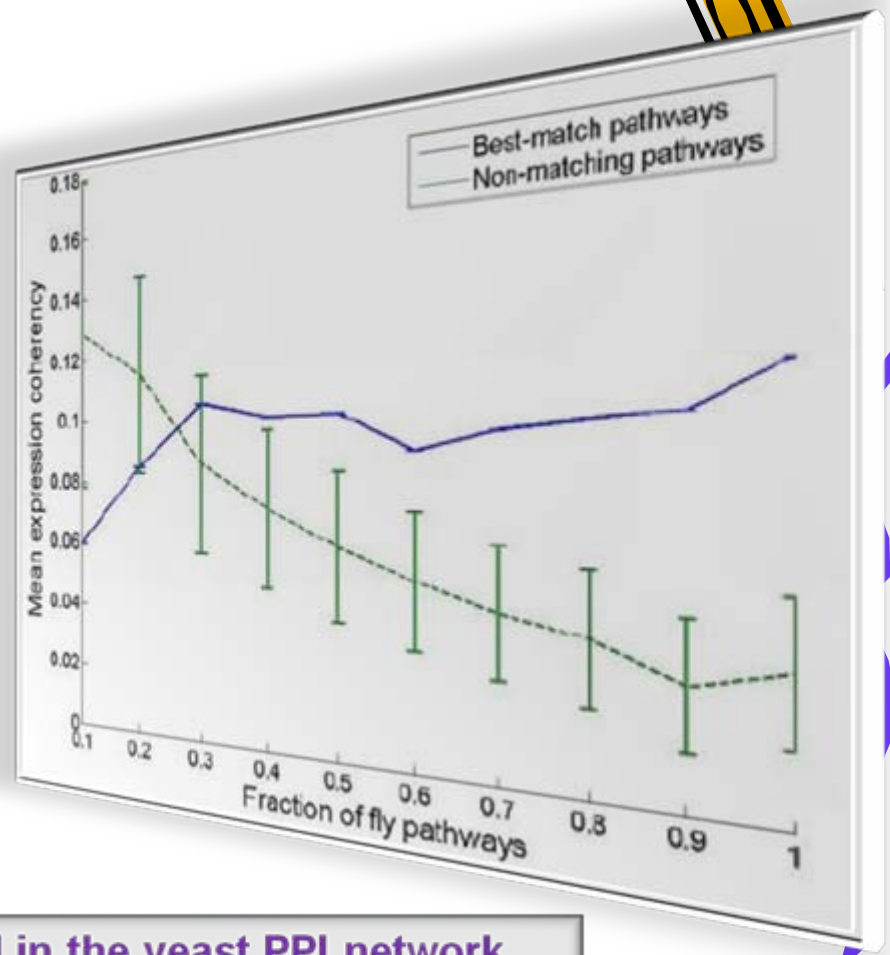
Functional Significance

	High interaction score pathways			Random pathways	
	Number of pathways	Functional enrichment	Expression coherency (p-value)	Functional enrichment	Expression coherency (p-value)
Yeast	271	80%	< 1e-300	17%	4.0e-4
Fly	193	20%	0.024	0%	> 0.05

Source: Table 1 in Shlomi, T., et al. "QPath: A Method for Querying Pathways in a Protein-protein Interaction Network." *BMC Bioinformatics* 7 (2006): 199.

Best Matching Pathways in Fly



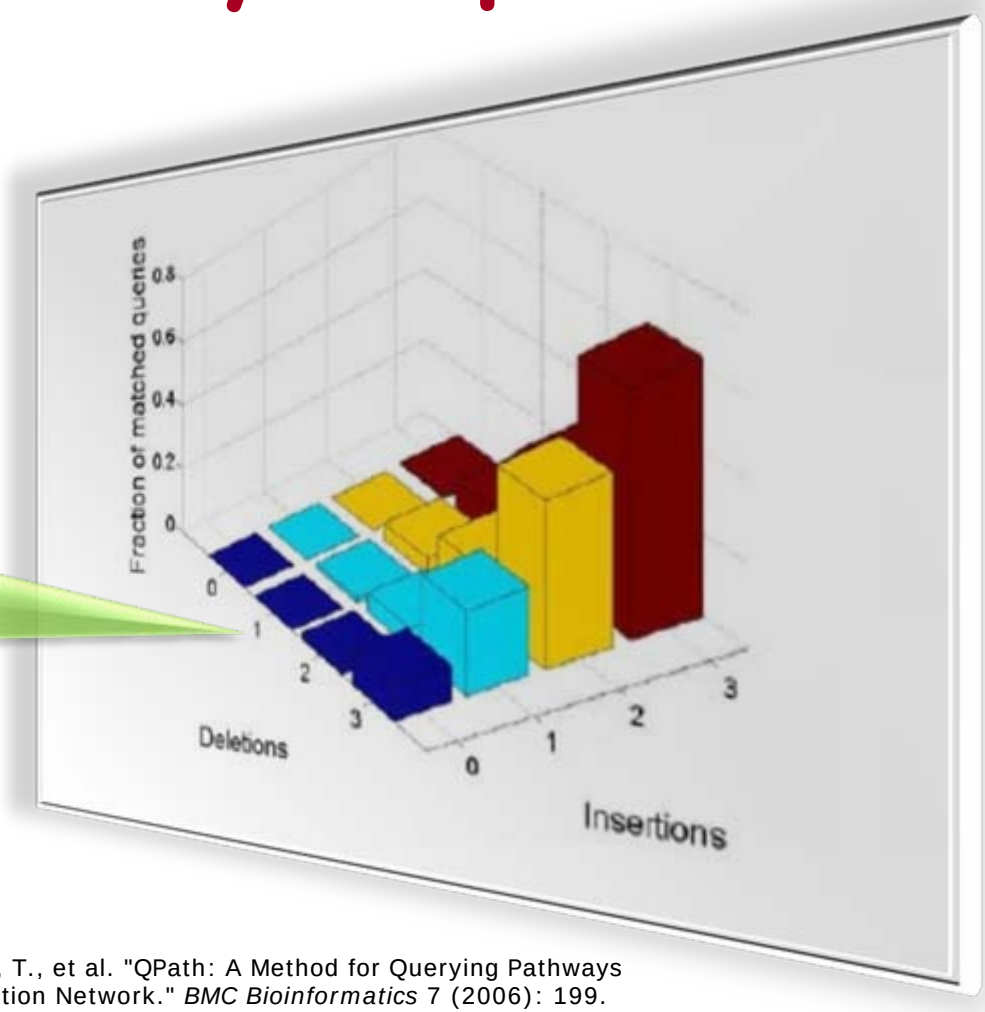


Running the 271 paths found in the yeast PPI network as queries on the fly network discovered that 63% of them had a match in the fly network

Source: Figure 3 in Shlomi, T., et al. "QPath: A Method for Querying Pathways in a Protein-protein Interaction Network." *BMC Bioinformatics* 7 (2006): 199.

Is the Insertion and Deletion Flexibility Really Required?

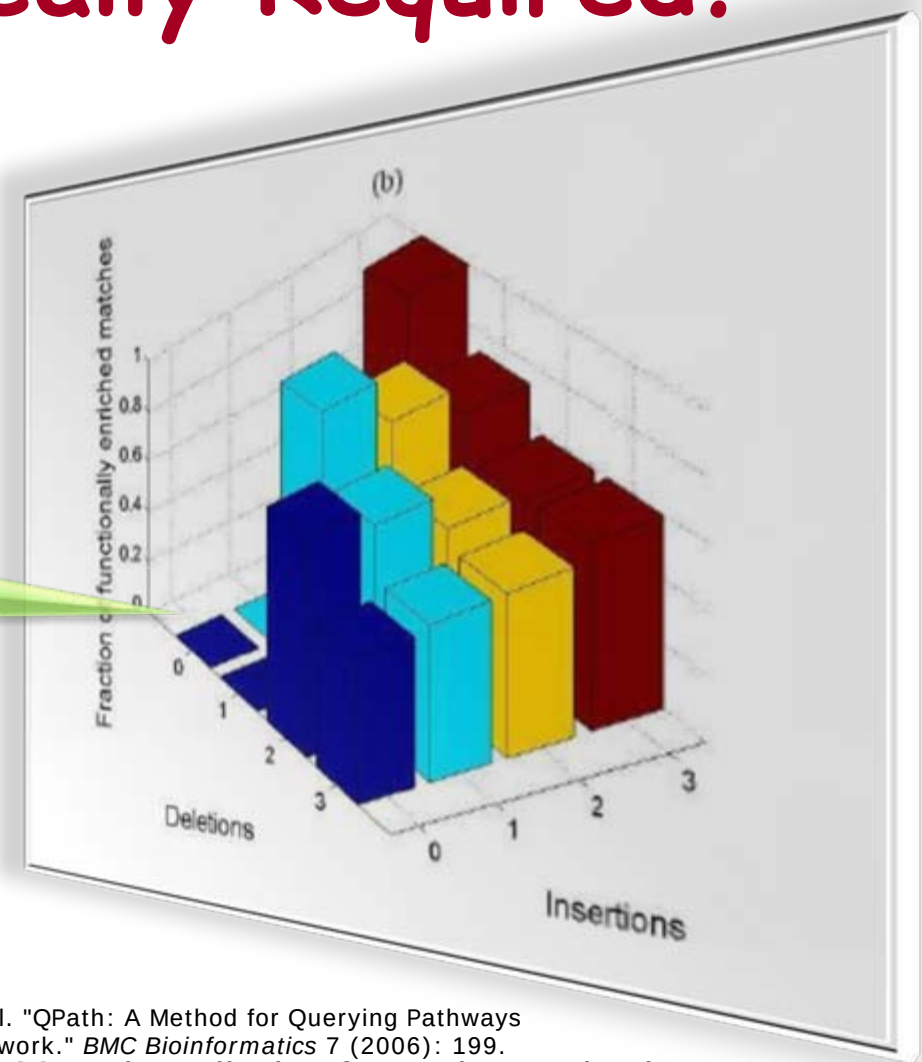
Most conserved paths between the yeast and the fly required more than one insertion and deletion



Source: Figure 2 in Shlomi, T., et al. "QPath: A Method for Querying Pathways in a Protein-protein Interaction Network." *BMC Bioinformatics* 7 (2006): 199.

Is the Insertion and Deletion Flexibility Really Required?

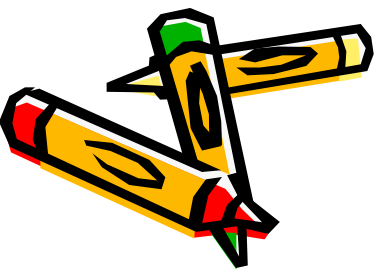
Functionally enriched paths are strongly depended on Ins/Del



Source: Figure 2 in Shlomi, T., et al. "QPath: A Method for Querying Pathways in a Protein-protein Interaction Network." *BMC Bioinformatics* 7 (2006): 199.

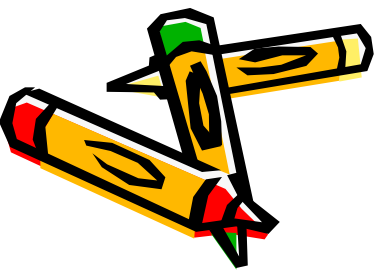
Functional Conservation

- For 64% of the conserved paths, the matched paths in the fly network conserved one or more functions of the yeast query pathways



Conclusions

- Very young field!
- Advanced computational methodology
  - Scaling multiple network alignment
- Association of network features with diseases



References Used

- T. Shlomi, D. Segal, E. Ruppin, and R. Sharan. **QPath: A Method for Querying Pathways in a Protein-Protein Interaction Network.** *BMC Bioinformatics*, 7(199), 2006.
- B. P. Kelley et al. **Pathblast: a tool for alignment of protein interaction networks.** *PNAS.*, 2003.
- R. Sharan, S. Suthram, R. M. Kelley et al. **Conserved patterns of protein interaction in multiple species.** *PNAS*, 102, pp. 1974-1979, 2005.
- Roded Sharan, Trey Ideker. **Modeling Cellular Machinery Through Biological Network Comparison**. Nature Biotechnology, 24(4), April 2006.