Systems Biology Group

# Theoretical properties of cross species interactions conservation

Guy Zinman

Carnegie Mellon University

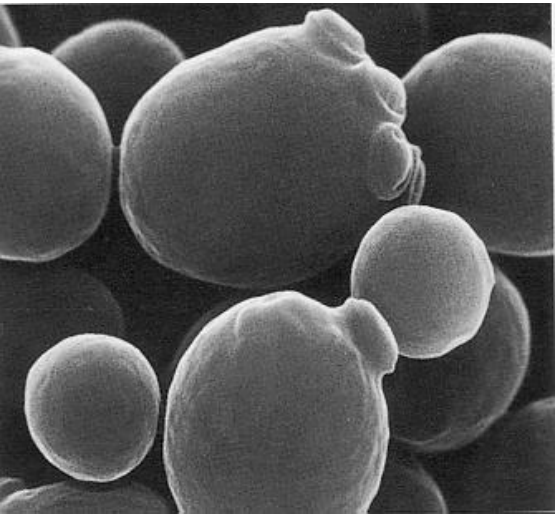
## “Nothing in Biology Makes Sense Except in the Light of Evolution”

**- Theodosius Dobzhansky, 1973**

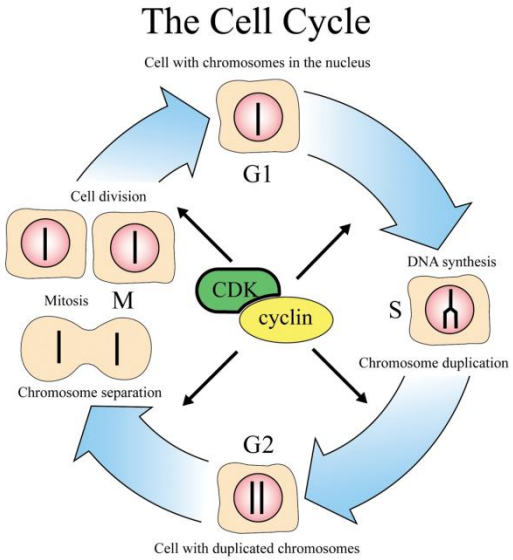
“The diversity and the unity of life are equally striking and meaningful aspects of the living world.”

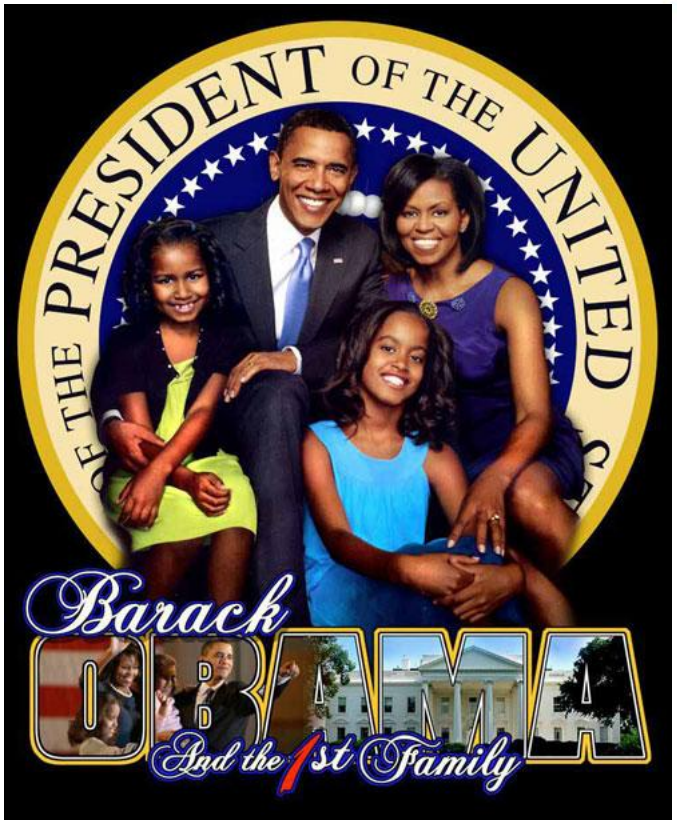
# Cell cycle conservation as example

From the unicellar Baker's yeast → To the 1st family:



**Most of the cell cycle components (phases progression, DNA replication mechanism, sister chromatid cohesion and separation through the ubiquitin system are pretty much similar**

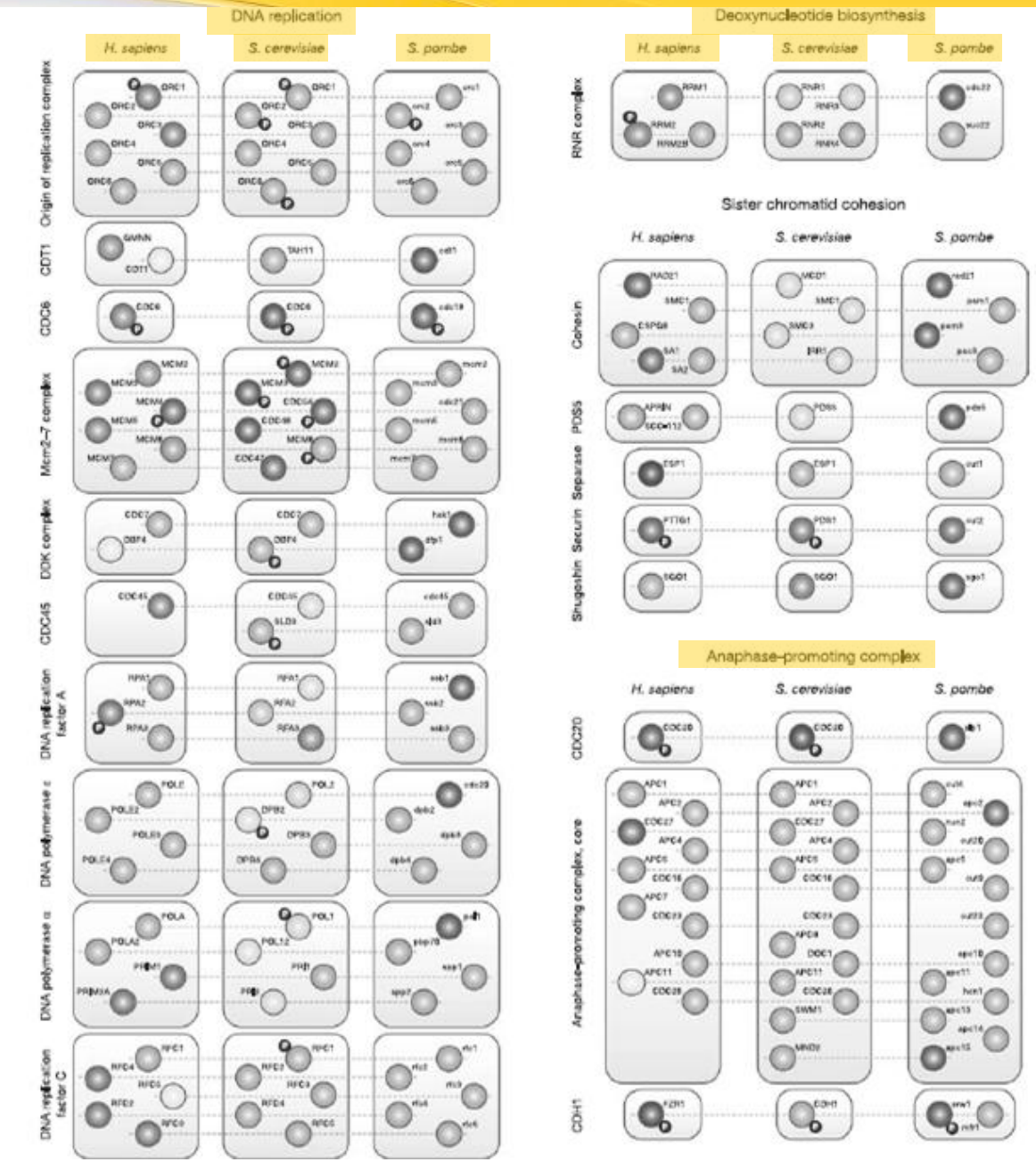



You can tell that he just got elected….

# Cell cycle complexes

Most complexes that are part of the cell cycle including DNA replication, deoxynucleotide biosynthesis, and Anaphase-promoting utilize similar genes in human and in yeast.




**Co-evolution of transcriptional and post-translational cell cycle regulation**
Jensen *et al*. Nature, 2006

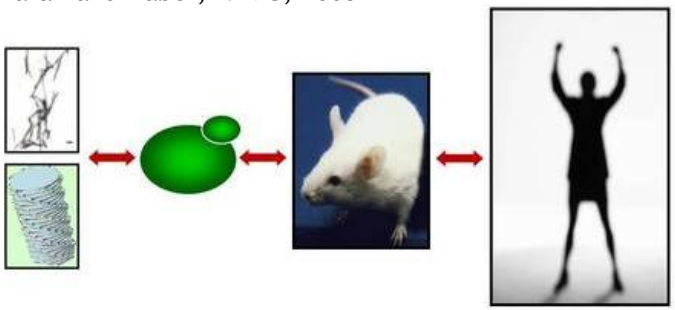
# Cross species analysis in drug discovery

| Model Organism | Common Name | Research Applications |
|---|---|---|
| *Saccharomyces cerevisiae* | Yeast | Cell processes e.g. mitosis and diseases (e.g. cancer) |
| *Drosophila melanogaster* | Fruit fly | A wide variety of studies ranging from early gene mapping to mutant screens to identify genes related to specific biological functions |
| *Caenorhabditis elegans* | Nematode | Development of simple nervous systems and the aging process |
| *Danio rerio* | Zebra fish | Mapping and identifying genes involved in organ development |
| *Mus musculus* | House mouse | Used to study genetic principles and human disease |
| *Rattus norvegicus* | Brown rat | Used to study genetic principles and human disease |

**The use of animal models in studying genetic disease: Transgenesis and induced mutation**
Simmons, Nature Education, 2008.

It is thought that 60 to 80 percent of disease-causing genes in humans have orthologs in the fly genome.

**Yeast, Flies, Worms, and Fish in the Study of Human Disease**
Hariharan and Haber, NEMG, 2003



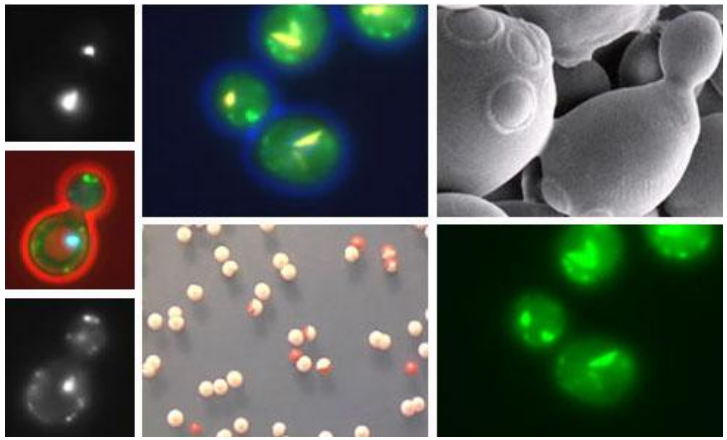
**Major use of model organisms in drug discovery:**

1. Discovering new genes.
2. Identifying genes causing human disease.
3. Defining cellular pathways.
4. Finding pathway perturbations leading to diseases.

# Examples of the use of model organisms to study human diseases

## Neurodegenerative diseases

**Yeast as a model for studying human neurodegenerative disorders**
Miller-Fleming *et al*. Biotechnology journal, 2008.



***Drosophila* as a model for Human Neurodegenerative disease**
Bilen and Bonini. Annual Rev. Genetics, 2005

FLY MODELS OF HUMAN POLYGLUTAMINE DISEASES

FLY MODELS OF PARKINSON DISEASE

FLY MODELS OF NONCODING TRINUCLEOTIDE REPEAT DISEASES

FLY MODELS OF ALZHEIMER AND RELATED DISEASES

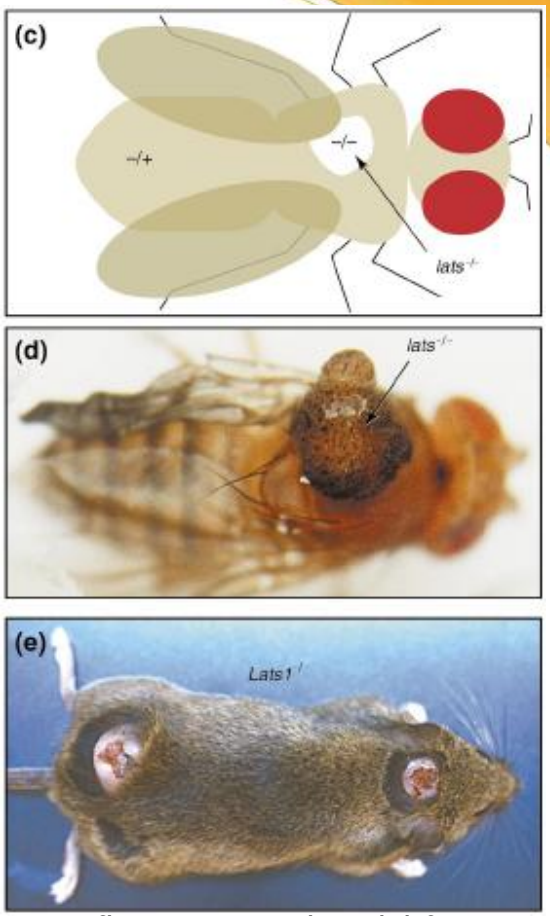
**Human genetic diseases: a cross-talk between man and yeast**
Foury, Gene, 1997

Lists 105 yeast homologues of human disease-associated genes including Immunodeficiency, Anemia, Autism, Diabetes and Insulin resistance, Cataracts and Glaucoma, and Brain tumors.

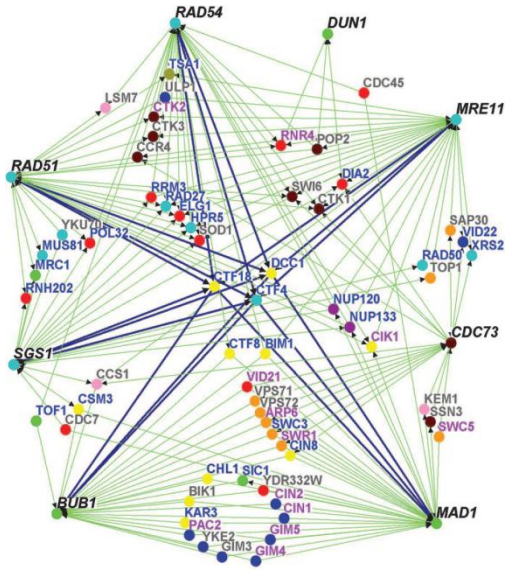
# Examples of the use of model organisms to study human diseases

## Cancer

***Drosophila* in cancer research**
Potter *et al*. Trends genetics, 2000




Mosaic flies are a good model for patients with cancer predisposition syndromes such as those heterozygous for mutated tumor suppressor genes



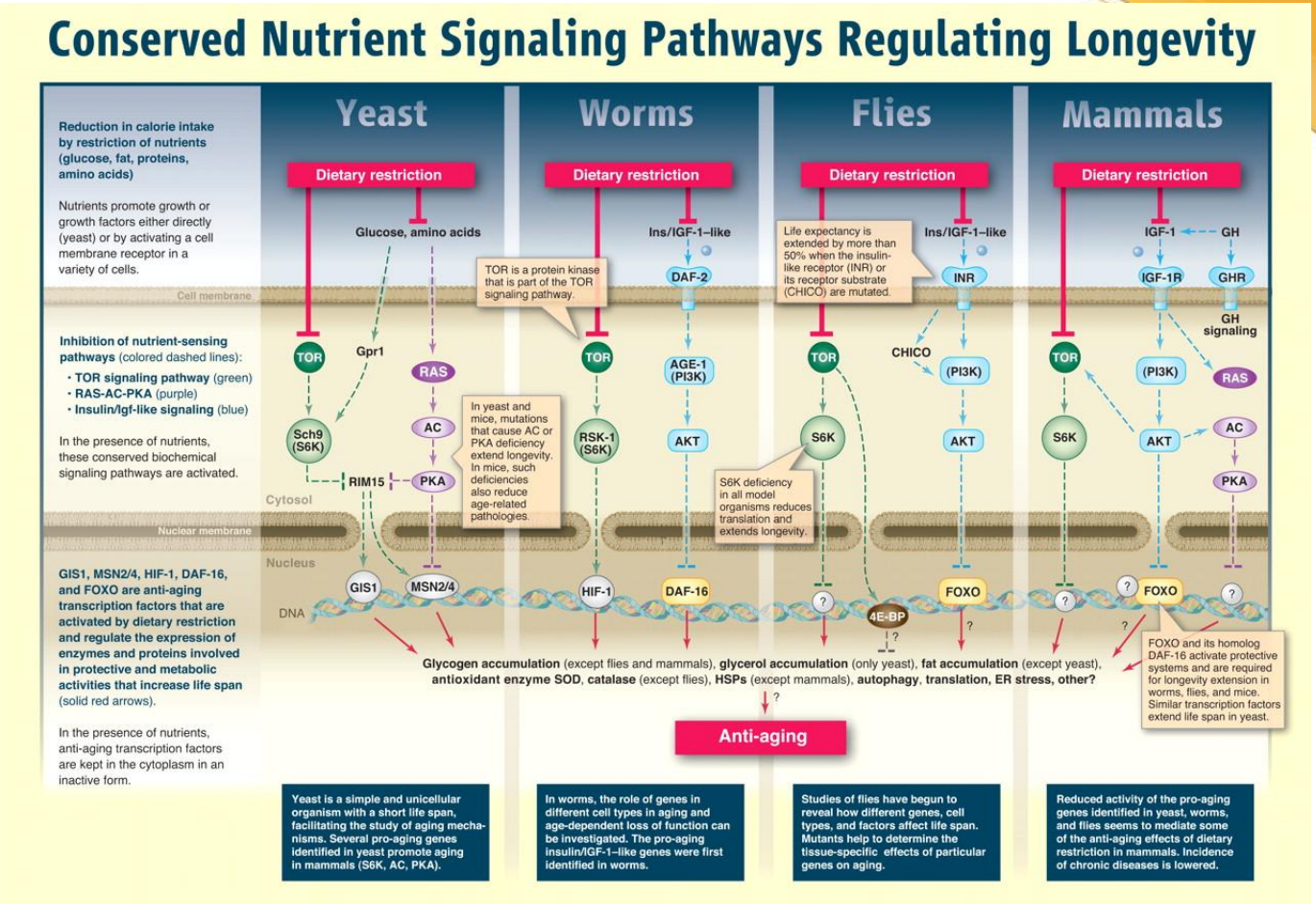

Common synthetic lethal interactions among yeast CIN genes that have human homologs mutated in cancer.

**Systematic genome instability screens in yeast and their potential relevance to cancer**
Yuen *et al*. PNAS, 2007

**Many Yeast CIN Genes Are Conserved.** Current understanding of mechanisms that contribute to genome stability has been largely fueled by work from model systems. This approach has been informative for human biology because of remarkable functional conservation within the chromosome cycle.

# Examples of the use of model organisms to study human diseases

## Aging

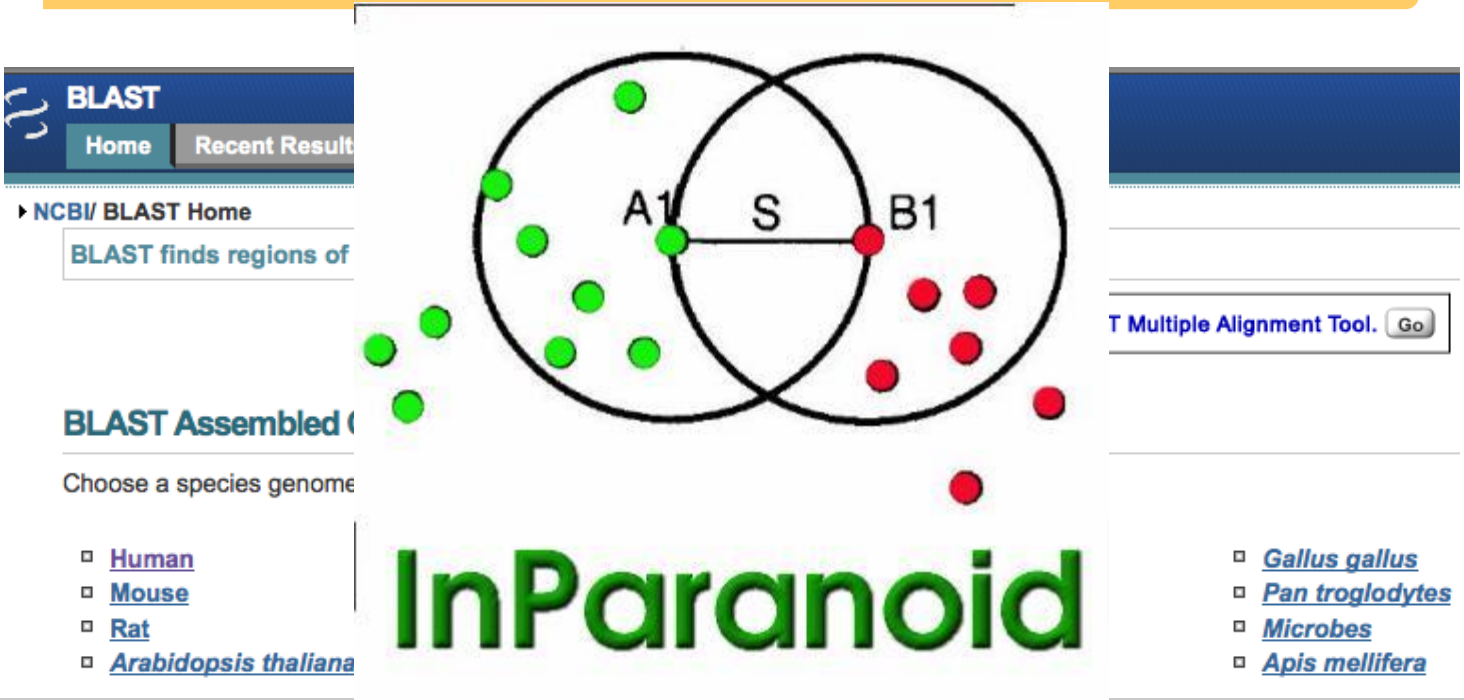



**Extending Healthy Life Span--From Yeast to Humans**
Luigi Fontana, *et al. Science 328, 321 (2010);*

# Cross species analysis through sequence similarity

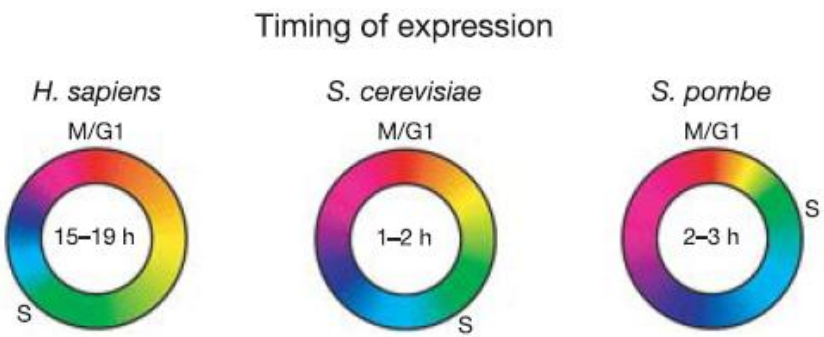
- Many of the above studies were made possible with the development of BLAST that allowed easy identification of orthologs and their control regions.
  - Orthologs – divergent copies of a single gene that were separated by a speciation event.

**Key assumption: Sequence similarity implies functional similarity**

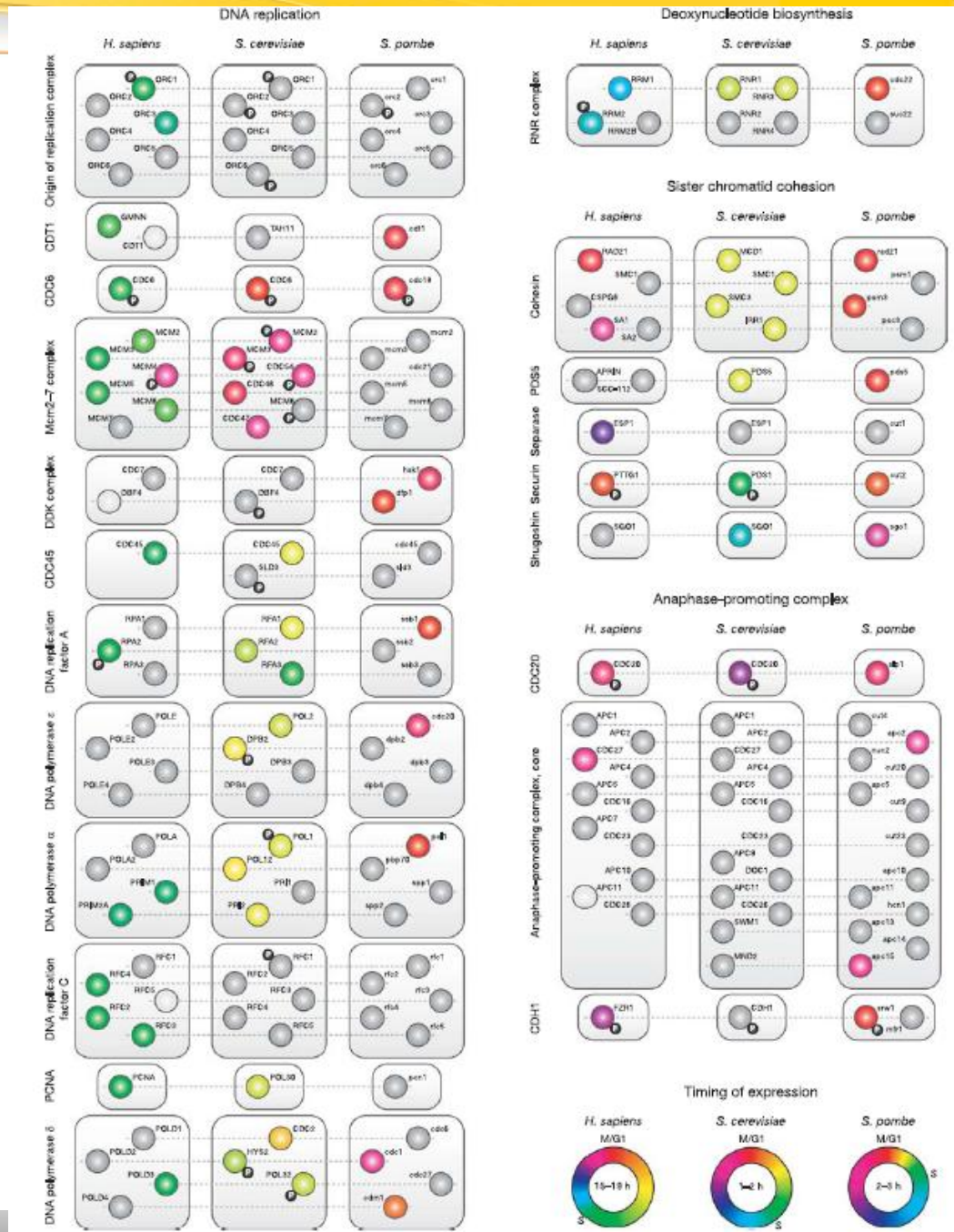


# Sequence information tells only part of the story

- In-time expression of the cell cycle proteins shows significant changes in their dynamics.

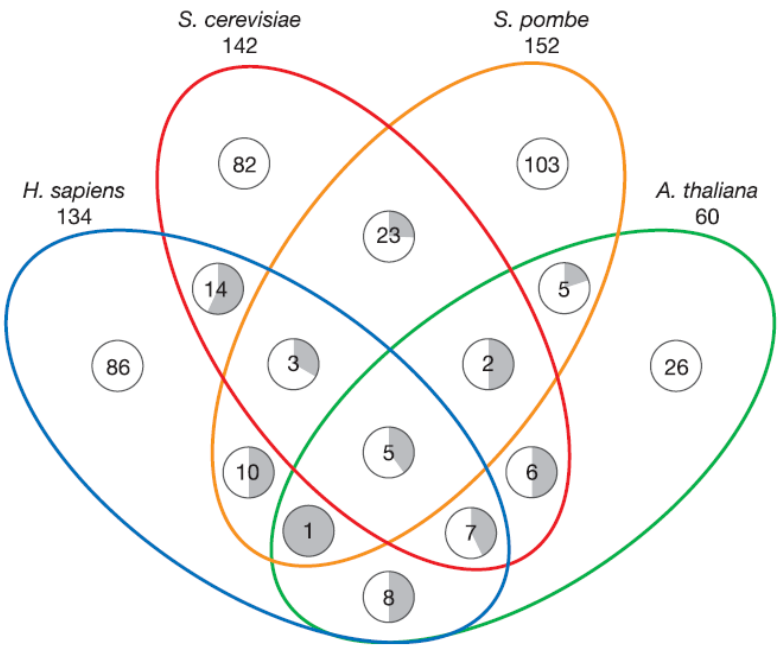



**Co-evolution of transcriptional and post-translational cell cycle regulation**
Jensen *et al*. Nature, 2006

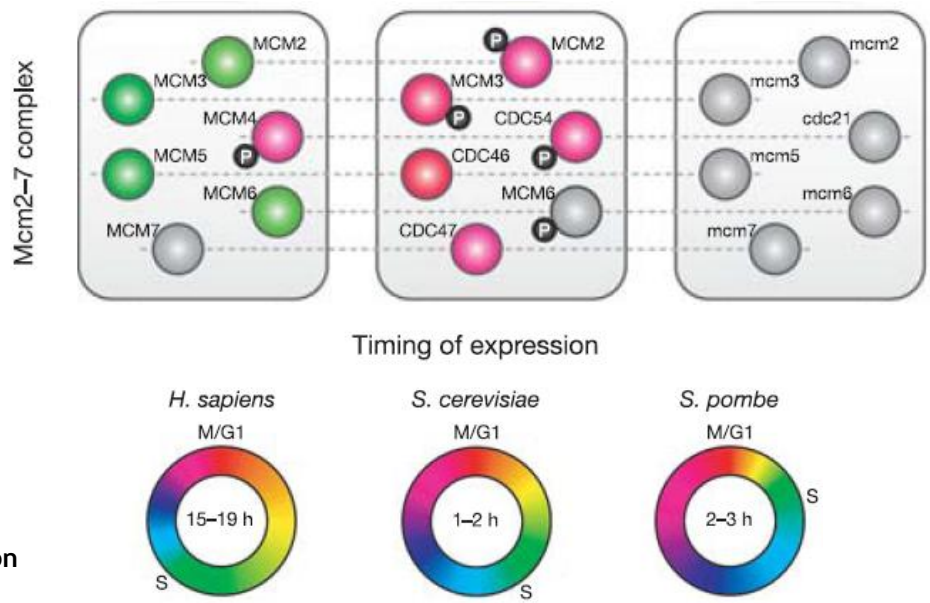


# Sequence information tells only part of the story

- Periodic expression is poorly conserved
- Only 5 / 381 in all four species




- Post translational regulation co-evolved independently
- Many different solutions have evolved for assembling the same molecular machines at the right time during the cell cycle




**Co-evolution of transcriptional and post-translational cell cycle regulation**
Jensen *et al*. Nature, 2006

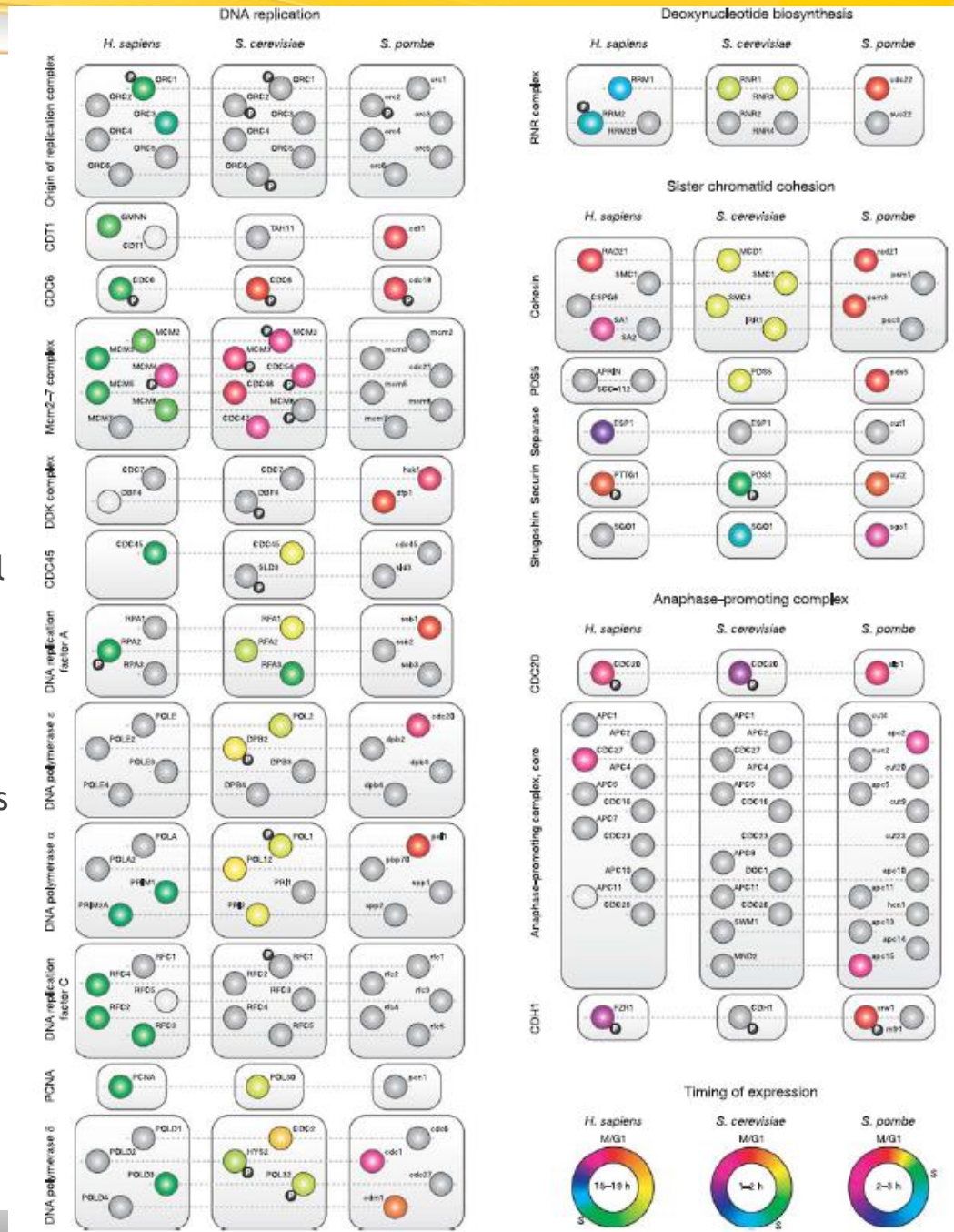
# Sequence information tells only part of the story

**This example demonstrates:**

1. Dynamic measurements are needed for understanding condition specific responses (driving factors still need to be elucidated).

2. Global functional similarity vs. local functional differences.

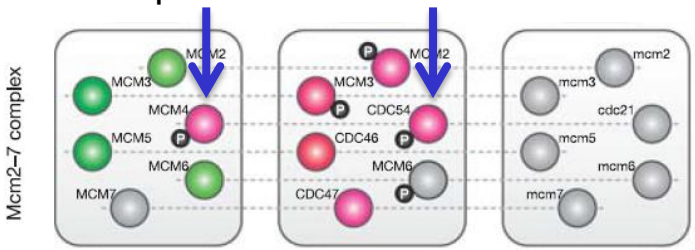
3. Quantifying the differences is challenging:
   - Different biological settings.
   - Requires adjustments due to cross species analysis.
   - Variability in experimental settings.
   - Dynamic measurements are noisy.
   - Orthology assignment.
   - Differences in coverage and quality of measurements.

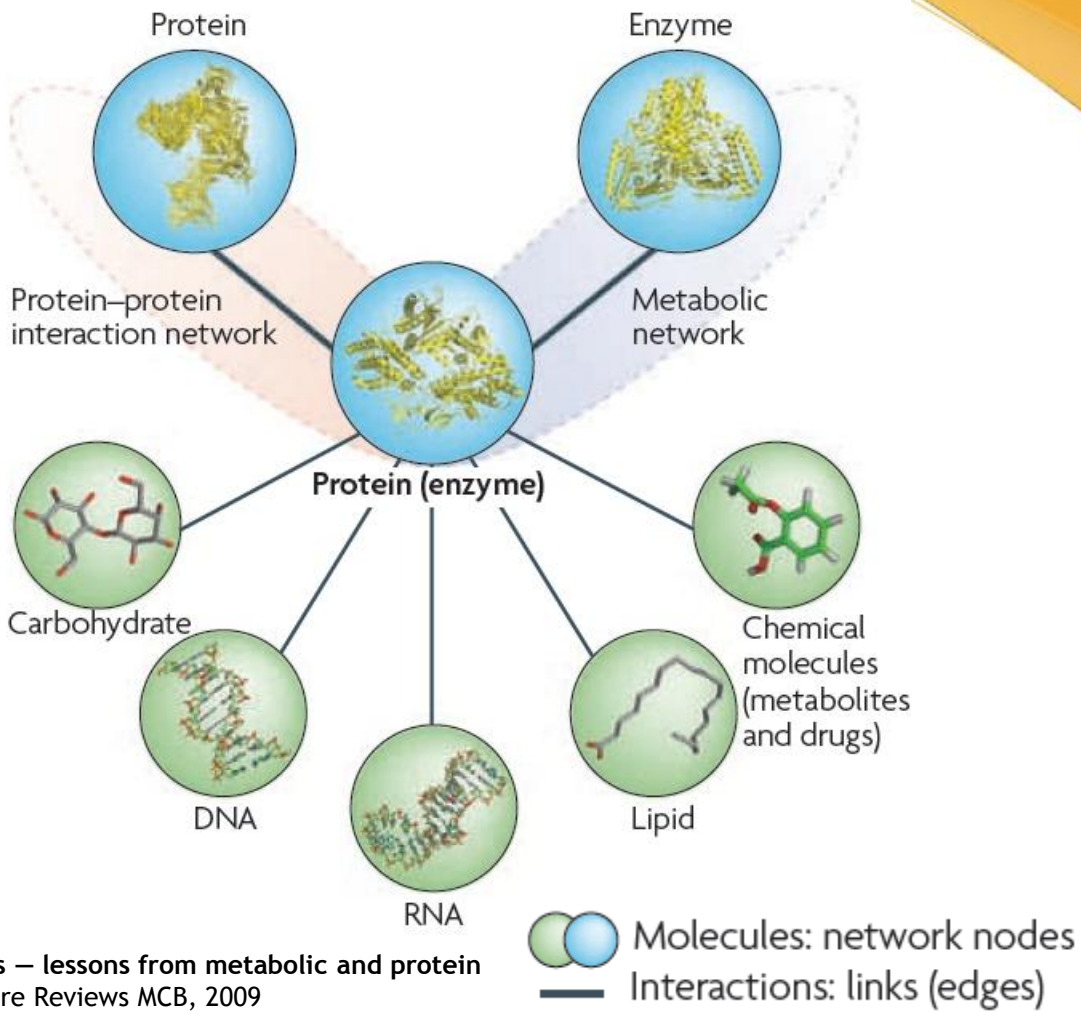


# From individual proteins to protein-interactions

- **Proteins operate in groups and interact one with another**
- **"Tell me who your friends are and I will tell you who you are"**

Co-expression as an interaction

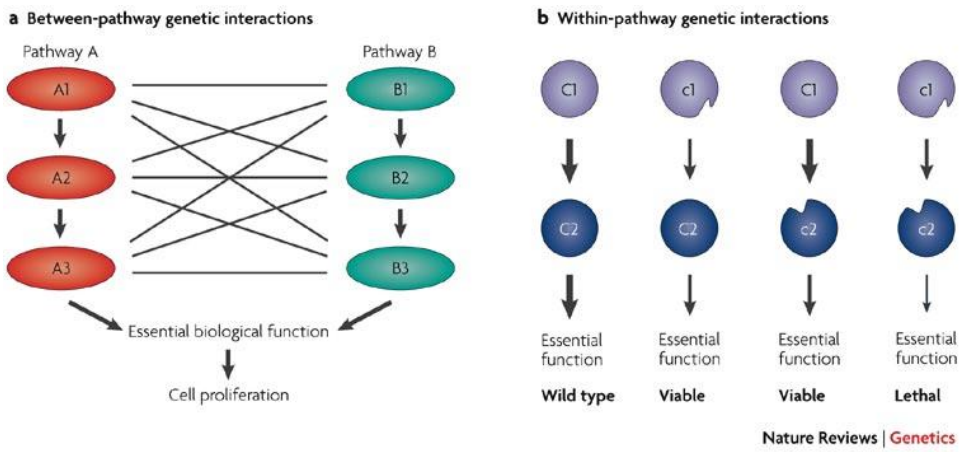



**Evolution of biomolecular networks — lessons from metabolic and protein interactions** Yamada and Bork. Nature Reviews MCB, 2009

# Types of high throughput interactions

- In recent years, new experimental methods were developed that produce high throughput data for more data types and in more species.
- These include co-expression data, protein-protein interactions (PPI), genetic interactions (GI), protein-DNA interactions, mircoRNA-target regulation, kinase-substrate phosporylation, metabolic reactions, and others.




**Negative GI -** the double mutant has a less severe phenotype than either single mutant (left).

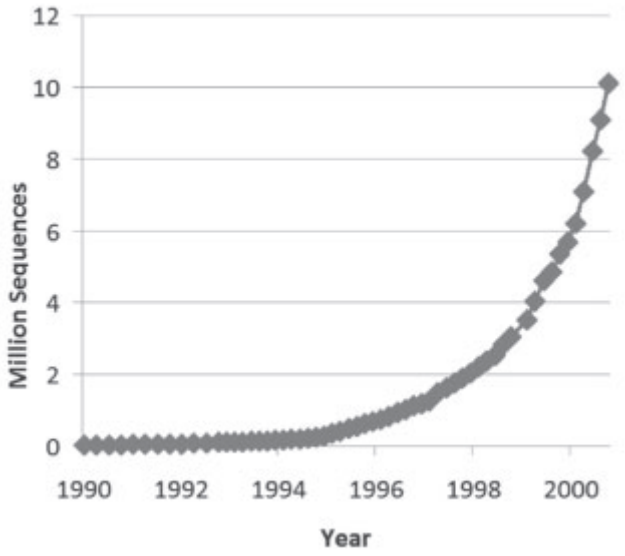
**Positive GI** - the double mutant has a more severe phenotype than one predicted by the additive effects of the single mutants (right).

**Exploring genetic interactions and networks with yeast**
Boone *et al*. Nature Rev. genetics, 2007

# HT datasets grow exponentially

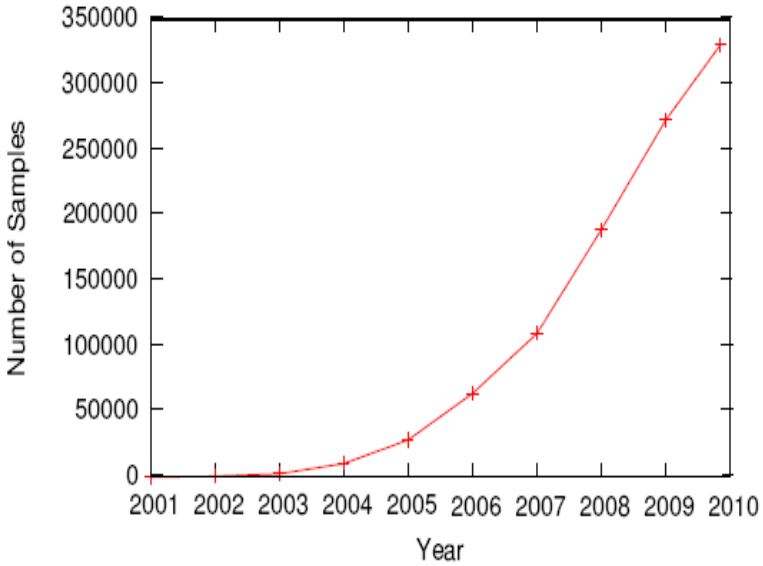
- While the data for some data-types is still limited for many species, the amounts of data grow exponentially.

## Growth of GenBank (sequence)




**Cross species analysis of microarray expression data**
Lu *et al*. Bioinformatics, 2009

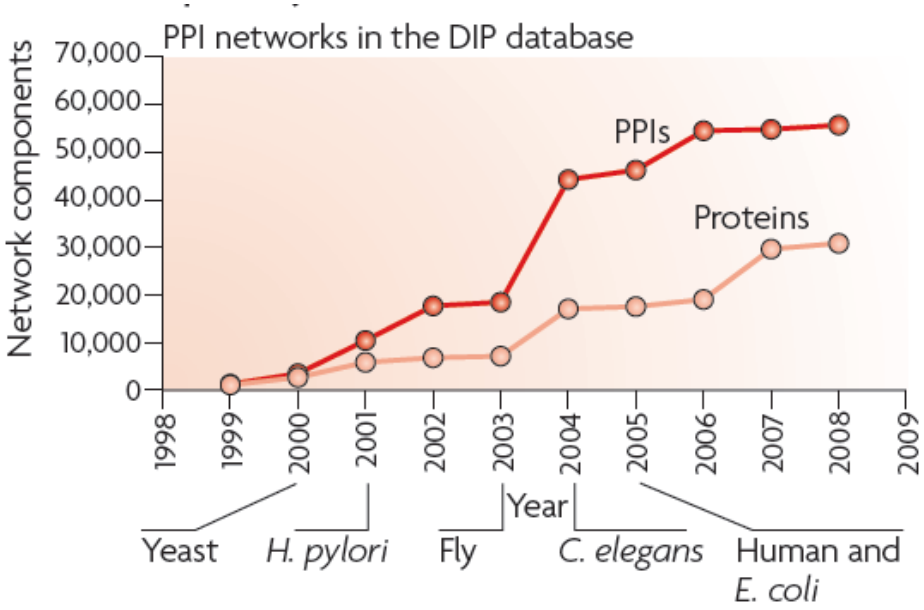
## Growth of GEO (expression)

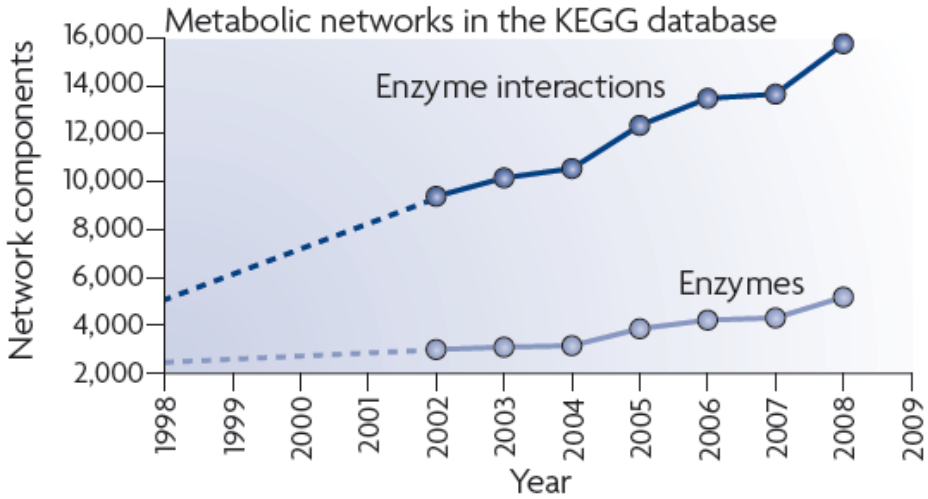



**Cross species queries of large gene expression databases**
Le *et al*. Bioinformatics, 2010

# Accumulation of network components

- **Number of physical interactions and enzymatic reactions is growing; available for more species**

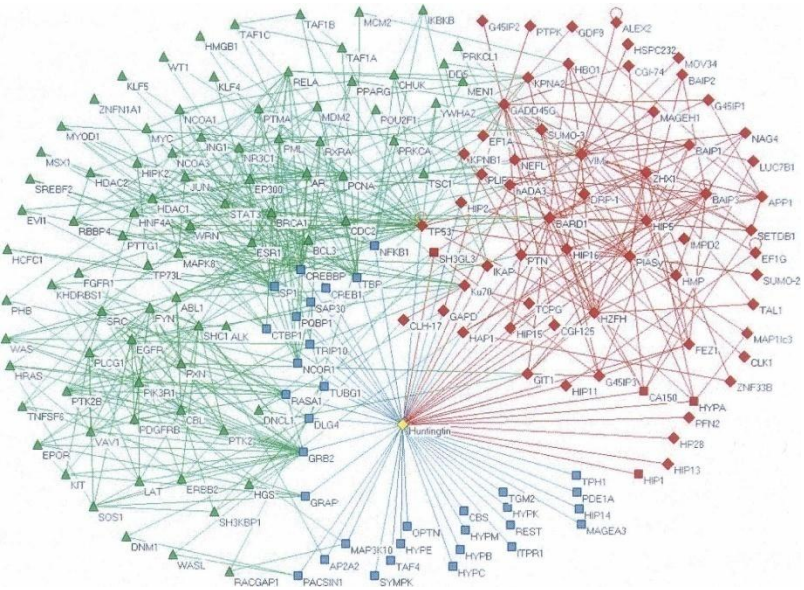



**Evolution of biomolecular networks — lessons from metabolic and protein interactions** Yamada and Bork. Nature Reviews MCB, 2009

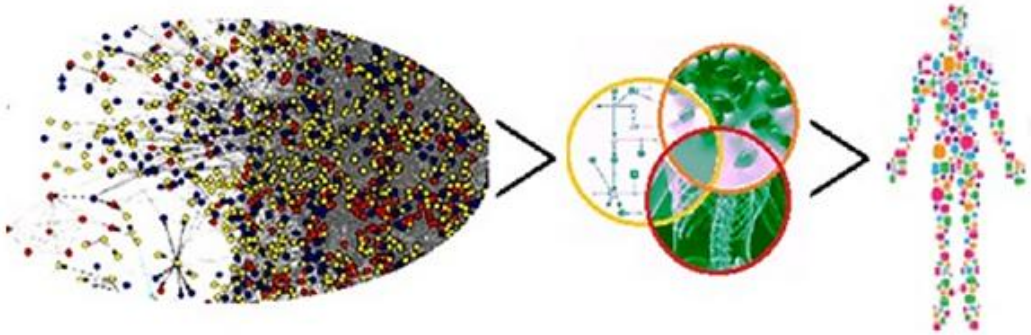
# Applications to medical treatments

- Microarray-based genetics tests were approved by the FDA.
- Aggregation to networks to understand the molecular basis of diseases.




**Protein interaction network for Huntigton disease**

**Protein networks in disease**
Ideker and Sharan Genome Res., 2008



**Protein Interaction Networks**
**Pathways, Networks**

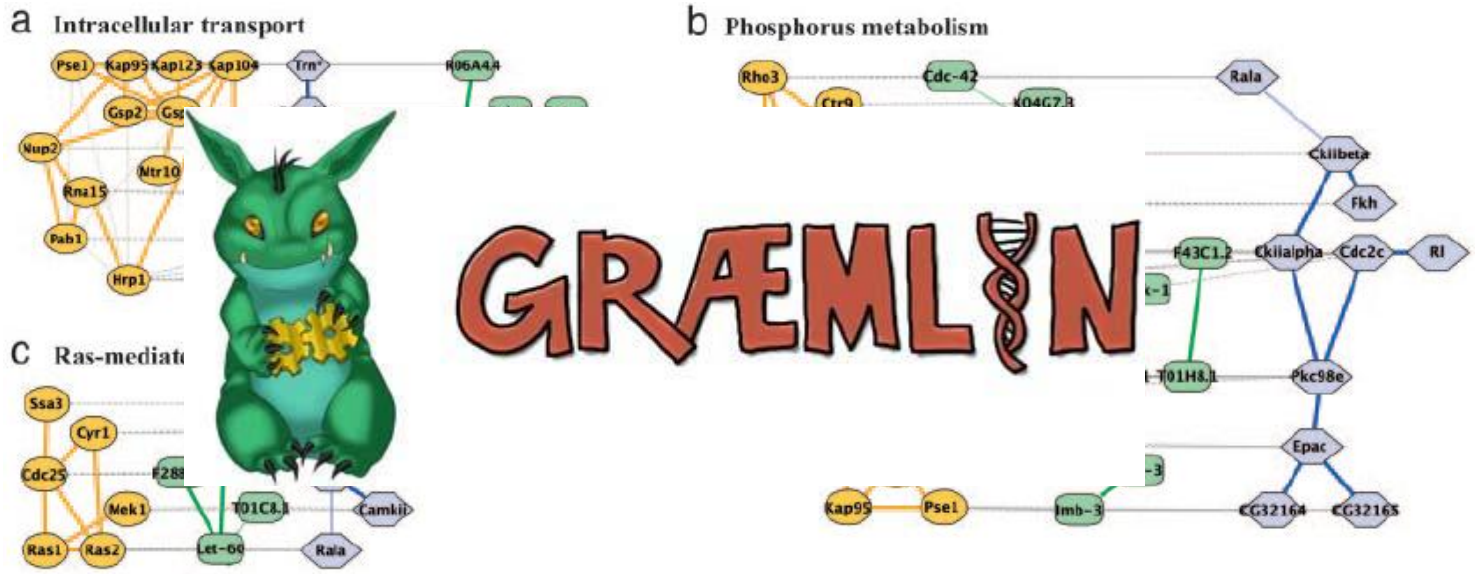
**Physiology & Pathology, Genomics**
**Proteomics, Metabolism, Toxicology**
**Pharmacogenomics, Chemistry…**

**Cause-effect relationships in medicine: a protein network perspective**
Fliri *et al*. Trends in Pharm. Sci., 2010

# Cross species studies on interaction data

As a result, in recent years there is a variety of studies on cross species analysis of HT datasets.

**Proteins interaction networks alignment**




**Conserved patterns of protein interactions in multiple species**
Sharan *et al*. PNAS, 2004

Are interactions conserved?

# Conservation of HT datasets

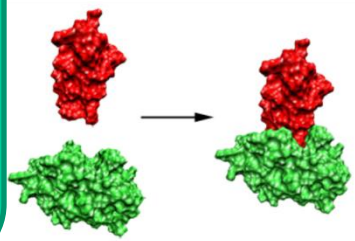
Protein- DNA interactions

Weakly conserved (11% between human and mouse, 20% between 3 yeasts with 85% sequence identify)

Sequence

Note: the analysis on the right was done only for orthologs for which the sequence similarity is close to 100%.

Protein- protein interactions

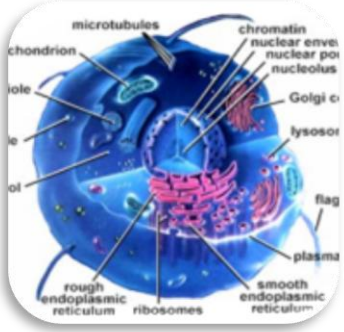


15-22% conservation between the two yeasts.

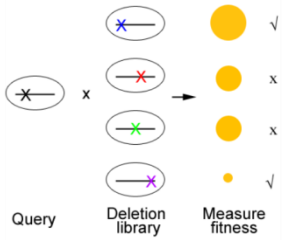
.

Functional annotation

Relatively well conserved (roughly 50% between budding and fission yeast)

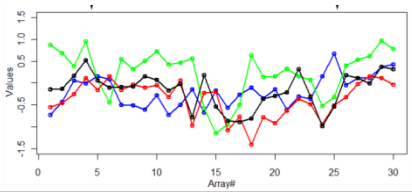



Genetic interactions (positive & negative)



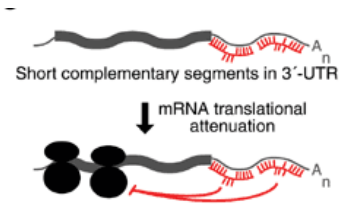

3-17% conservation between the two yeasts.

Expression




Correlations of 0.17-0.37 found for human and mouse

MicroRNA



# System level / gene level discrepancies

Functional conservation

**On the system level:**

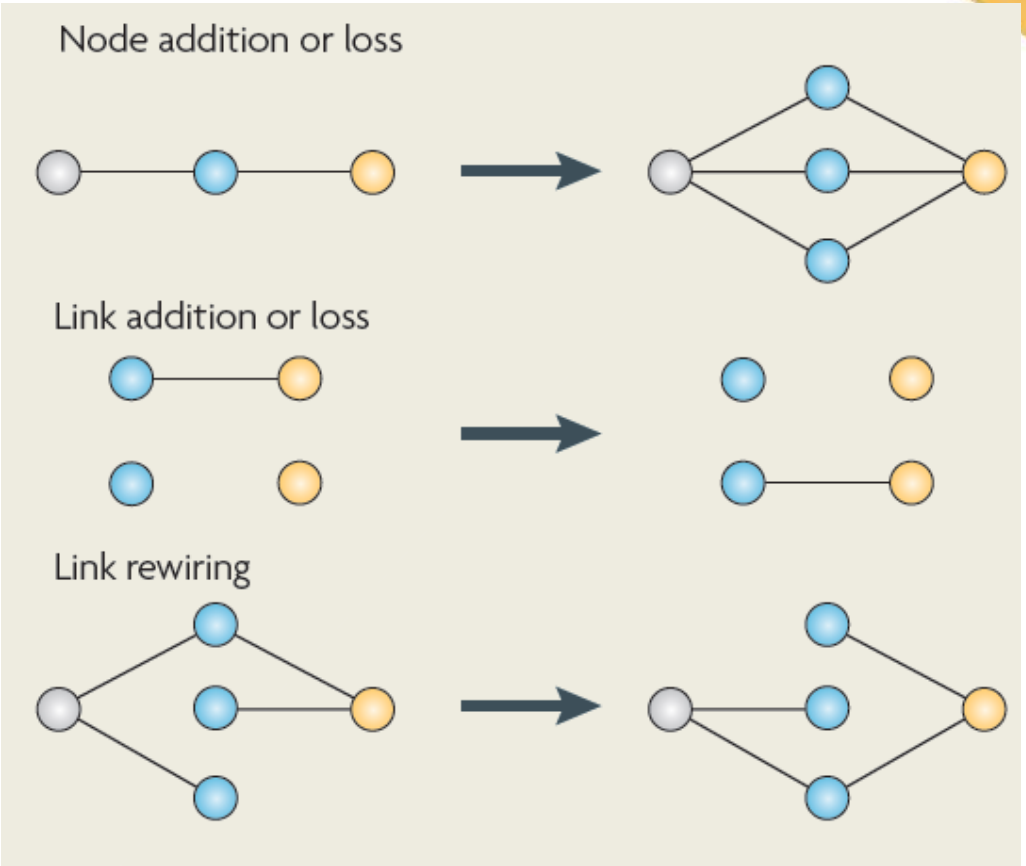
Biological processes are well conserved. (e.g. cell cycle, metabolism). (Wilkins *et al.*, 2001).

**On gene / protein level:**

Interaction data is far less conserved. (Roguev *et al.*, 2008, Tiscler *et al.*, 2008, Odom *et al.*, 2007, Lu *et al.*, 2009).

# Basic network evolution

- **Node addition / loss:**
  - Gene duplication / deletion
  - Horizontal transfer
- **Link addition / loss:**
  - Point mutations
  - Domain accretion or loss
  - Alternative splicing
  - Insertion or deletions in genes or regulatory regions
- **Link rewiring:**
  - Secondary effects of the above




**Evolution of biomolecular networks — lessons from metabolic and protein interactions** Yamada and Bork. Nature Reviews MCB, 2009

# Accuracy issues in comparative network analysis(1)

- **High node coverage, low interaction coverage**

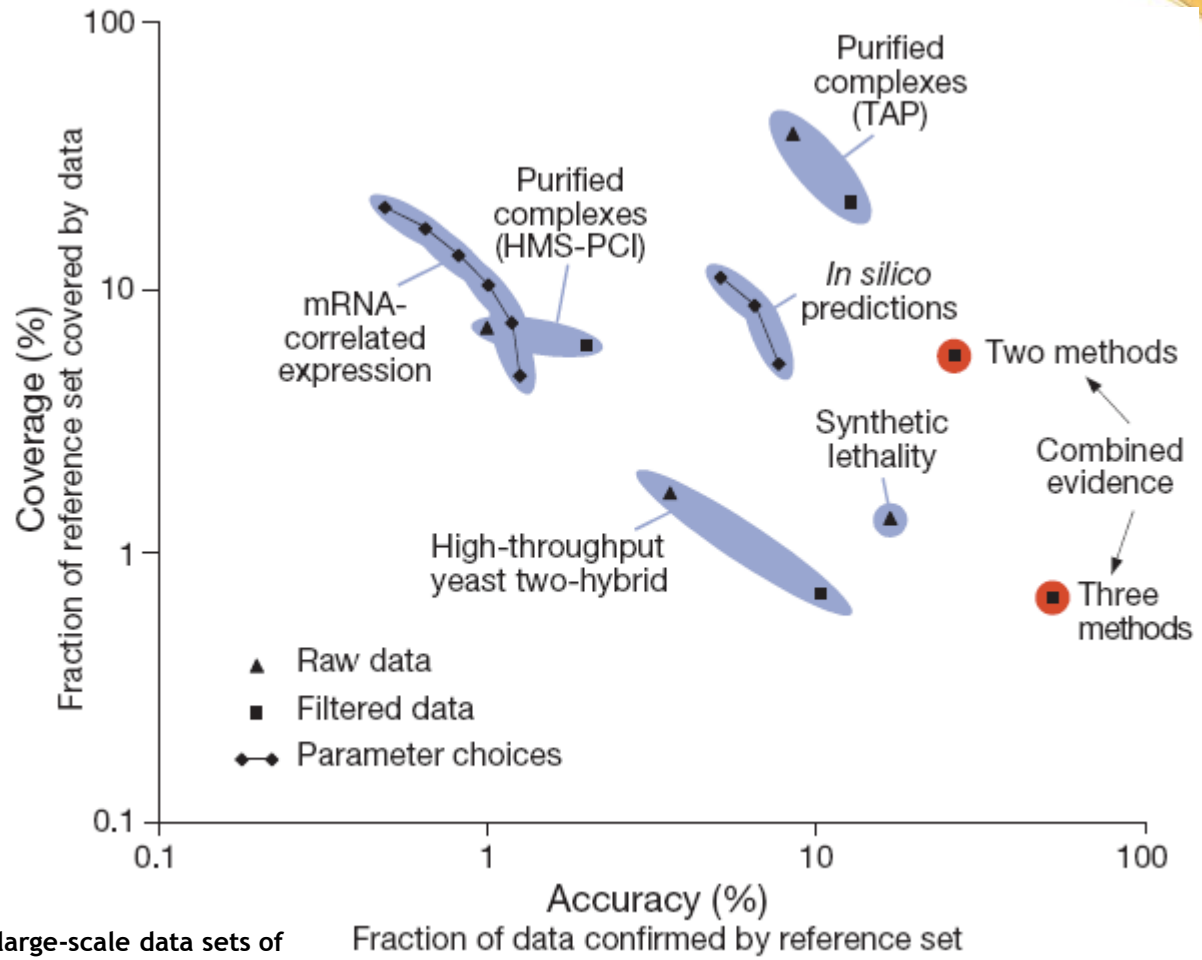
| PPI | | | |
|---|---|---|---|
| DIP (Salwinski *et al* 2004) | 4975/17612 | 1819/2171 | 1882/7002 |
| STRING (Jensen *et al* 2009) | 5764/55860 | 11214/91401 | 2842/14619 |
| Y2H | 817/692 (Uetz *et al* 2000) | 1549/2754 (Rual etal 2005) | |
| | 797/841 (Itoh *et al* 2001) | 1705/3186 (Stelzl et al 2005) | |
| PCA | 1124/2270 (tarassov *et al* 2008) | | |
| TAP-MS | 2762/6942 (Gavin *et al* 2002) | 2235/6463 (Ewing *et al* 2007) | 1209/4871 (Butland *et al* 2005 ) |
| | 2708/7123 (krogan *et al* 2006) | | 2457/8664 (Arifuzzaman *et al* 2006 ) |

proteins / interactions

**Evolution of biomolecular networks — lessons from metabolic and protein interactions** Yamada and Bork. Nature Reviews MCB, 2009

# Accuracy issues in comparative network analysis(2)

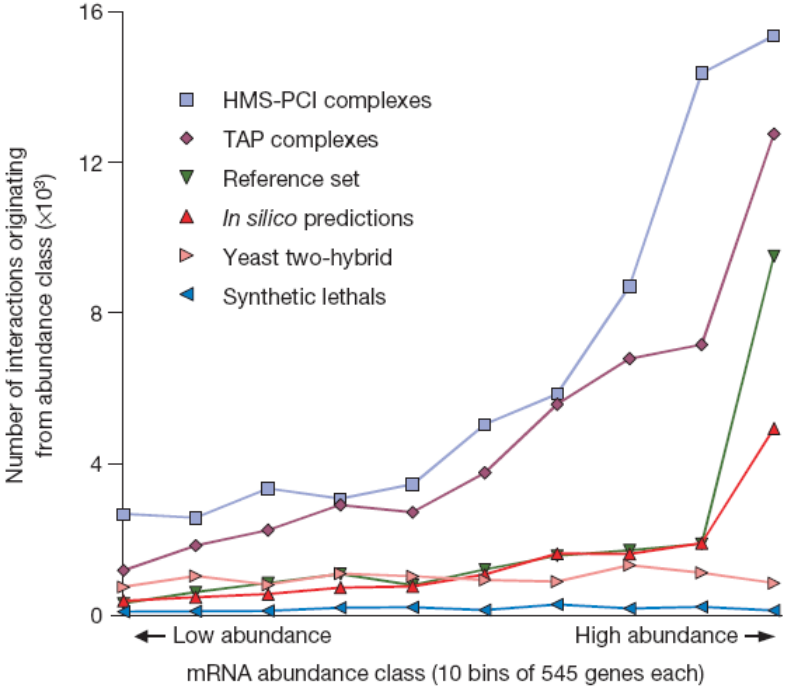
- **Accuracy**
- Different PPI screens are not the same.

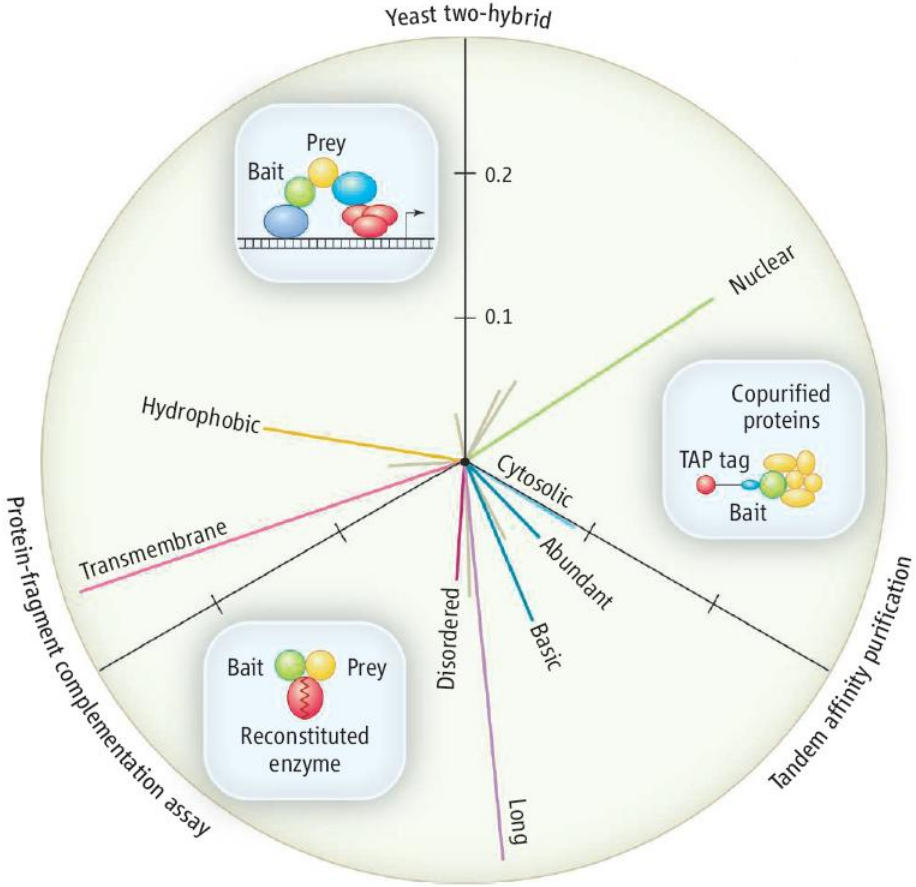



**Von Mering *et al*. Comparative assessment of large-scale data sets of protein-protein interactions.** Nature, 2002

# Accuracy issues in comparative network analysis(3)

- **Different experimental screens create different coverage bias**



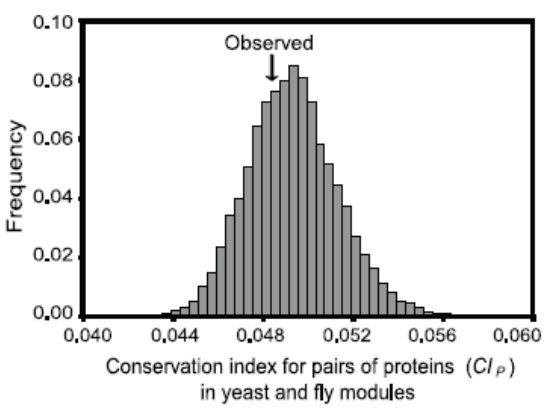




**Von Mering *et al*. Comparative assessment of large-scale data sets of protein-protein interactions.** Nature, 2002

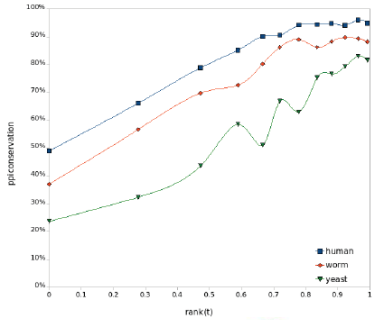
**Jensen, L.J. & Bork, P. Biochemistry. Not comparable, but complementary.** Science, 2008

# Previous observations on interactions conservation (1)

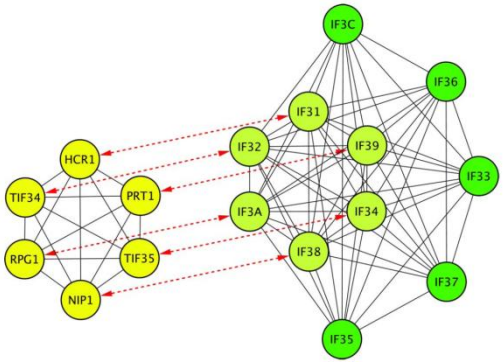
- **Wang and Zhang, 2007 - failed to observe any evolutionary conservation among yeast, fly, and nematode PPI modules**
  - For 27 Giant modules based only on Y2H




- **Fox *et al.*, 2009 – interactions within hubs are more conserved.**

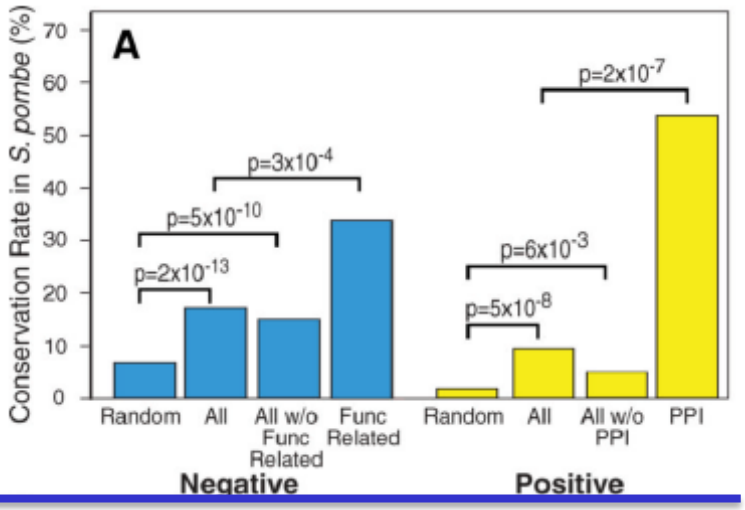



- **Van Dam and Snel, 2008 – interactions within complexes are more conserved.**

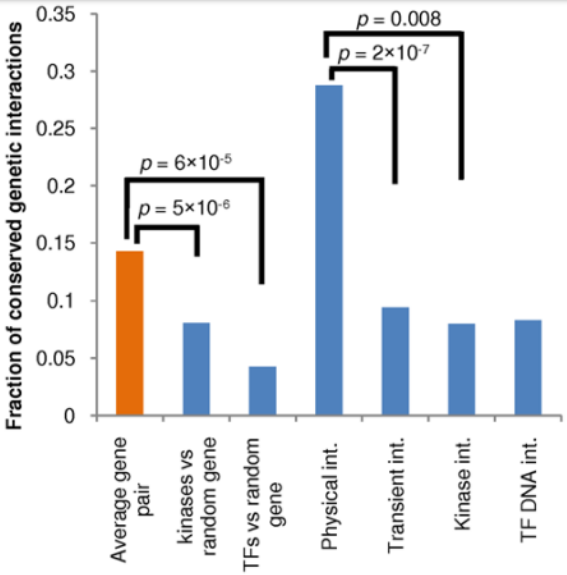


# Previous observations on interactions conservation (2)

- **Roguev *et al.*, 2008 - PPI pose constraints on functional divergence in evolution**

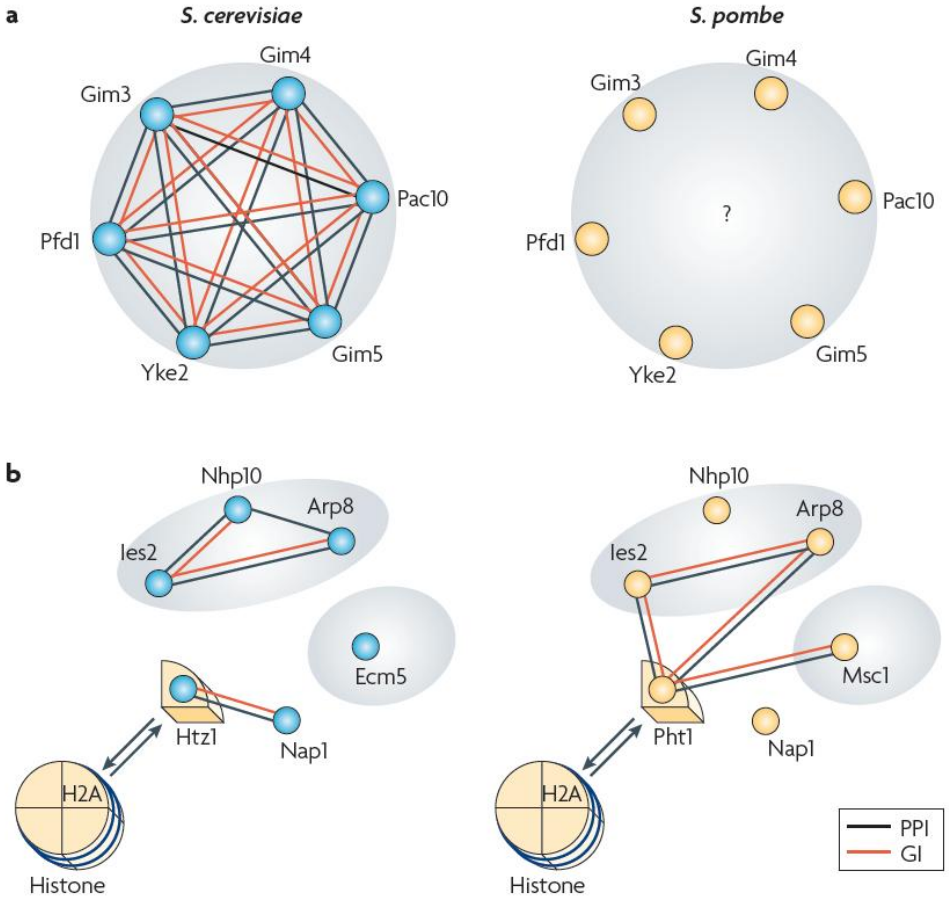



- **Beltrao *et al.*, 2009 – transient interactions (like kinases) are less conserved**



# No negative data

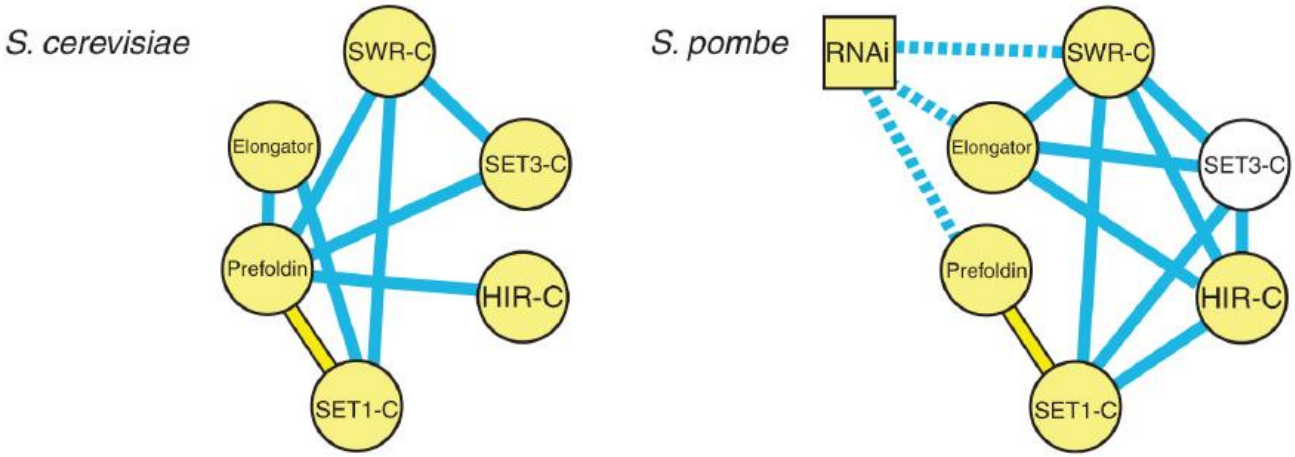
- **Requiring more sources for an edge is good, but might be too strict**




**Evolution of biomolecular networks — lessons from metabolic and protein interactions** Yamada and Bork. Nature Reviews MCB, 2009

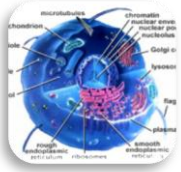
# New biological mechanisms

- **Divergence can be interesting:**
  - **Differences can indicate that new mechanisms have evolved (RNAi)**
  - Roguev *et al.*, 2008



# Hypothesis

Functional conservation



System level

Intermediate “meta-gene” level?

Gene/protein level

**Is there an intermediate level that retains more of the interaction data?**

Zinman *et al. BMC Systems Biology* 2011, **5**:134
http://www.biomedcentral.com/1752-0509/5/134

BMC
Systems Biology

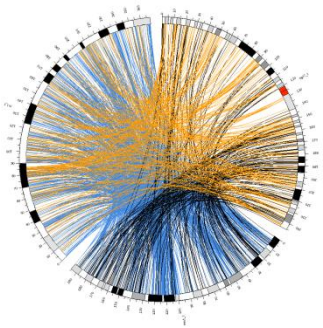
RESEARCH ARTICLE Open Access

# Biological interaction networks are conserved at the module level

Guy E Zinman[1†], Shan Zhong[1†] and Ziv Bar-Joseph[1,2*]

# Hypothesis - bridging the gap

**Two possible reasons for the 'meta gene' model:**



- Most interactions are spurious - only core network interactions are conserved
  - Many participants are playing supporting but not central roles.

- Conservation of networks is on a different level than individual interactions
  - A different substructure is conserved that is not at the individual interaction level.

# Comparing interaction data across species

- **We collected data from four different species:**
  *S. cerevisiae, S. pombe, C. elegans, D. melanogaster*

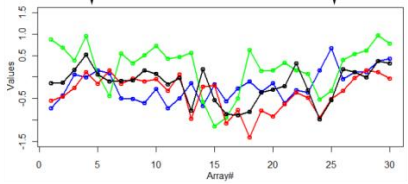
- **Datasets**

## Co-expression

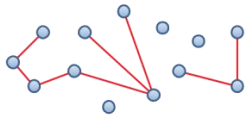
Stanford Microarray Database (cerev, elegans and fly)

NCBI GEO (pombe)

Combination and Filtering




Pairwise Spearman correlation coefficient and apply cutoff

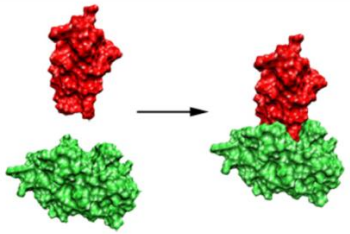


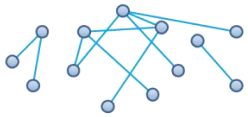
**Coexp net**

## PPI

BioGRID

Protein-protein interactions



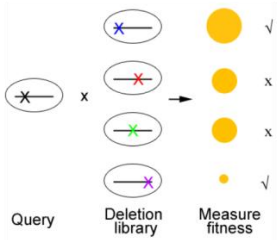


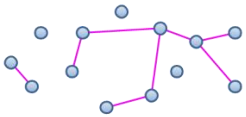
**Ppi net**

## GI

BioGRID

Genetic interactions



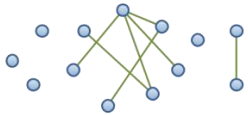



**Gi net**

## paralogs

Protein sequences pairwise BLASTP

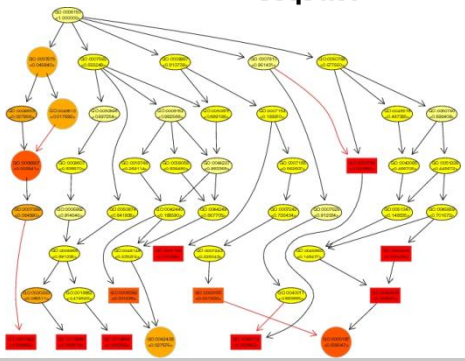
Apply cutoff



**Seqs net**

## GO annotation

(Used for evaluation)



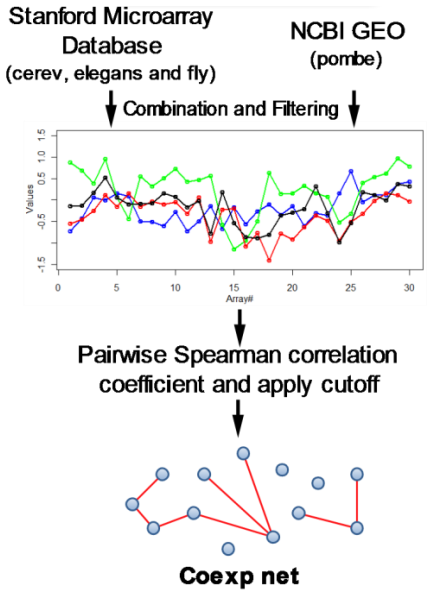
Orthology information was calculated using reciprocal best-hit

# Co-expression network construction

- **Not a straight-forward task:**
  - Arrays have different formats (one channel/two channels)
  - Different chips (different probe names)
  - Multiple probes per array / missing values
  - Different experimental settings (Normalization issues)

- **How to construct the network?**
  - How to associate genes?

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$



# Log likelihood scores for edges

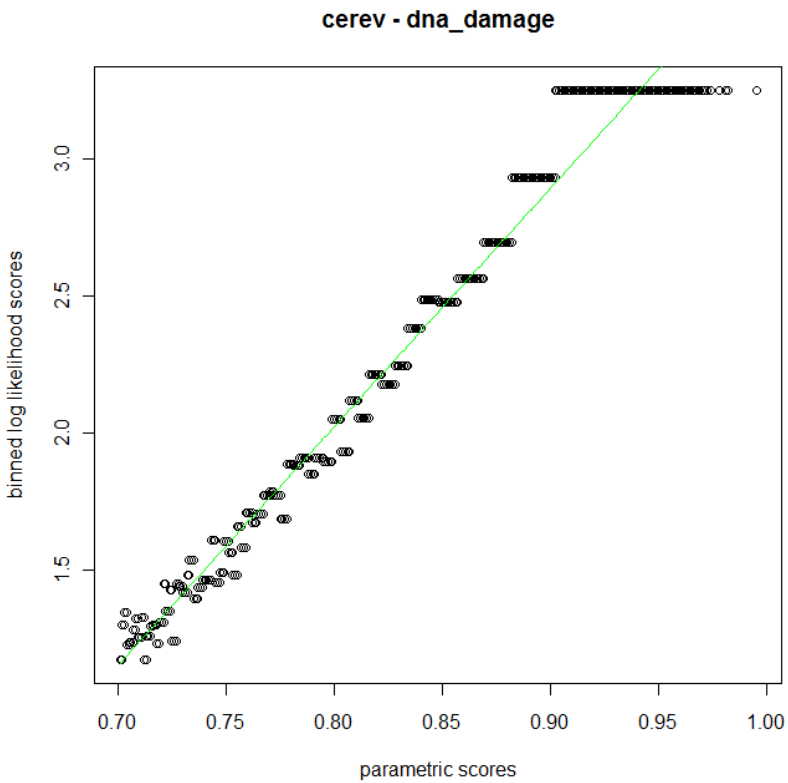
- Quantifies the confidence we have in each interaction.

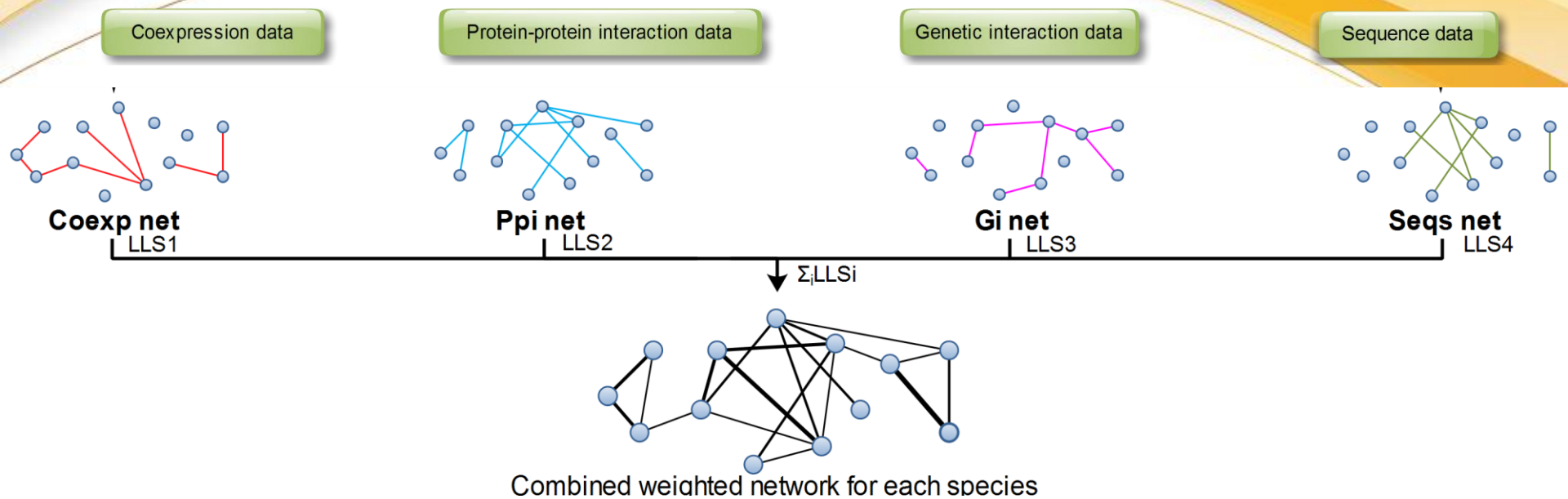
$$LLS = \ln\left(\frac{p(L|E)/\,p(\neg L|E)}{p(L)/p(\neg L)}\right)$$

$p(L|E) \quad p(\neg L|E)$

are the frequencies of linkages (L) observed in a given experiment (E) between annotated genes operating in the same pathway and in different pathways.

Insuk *et al.*, 2004



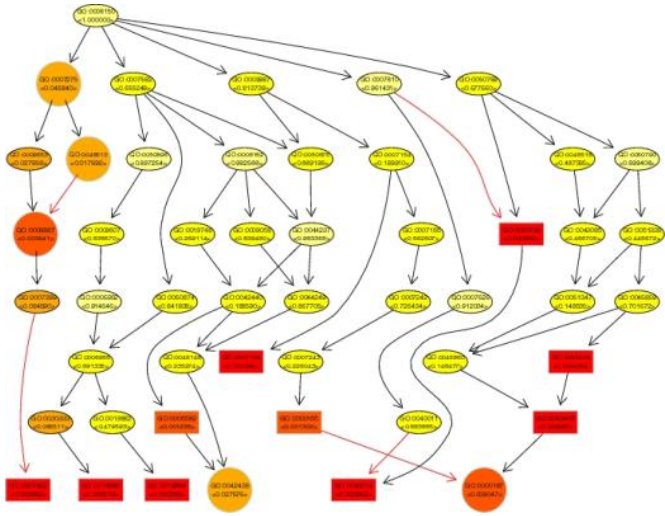



Interactions with more sources have a higher score

# GO network construction

- **How to construct the network - Semantic similarity score** (Wang *et al.*, 2007)
  - The idea is that the more parents are shared for each two terms, the more significant is their relationship.
    - Score representing the contribution of a parent *t* to semantics of term A:

$$S_A(t) = \begin{cases} 1 & t = A \\ \max\limits_{t':childrenof(t)} (w \times S_A(t')) & t \neq A \end{cases}$$



    - Semantic similarity of each pair of GO terms

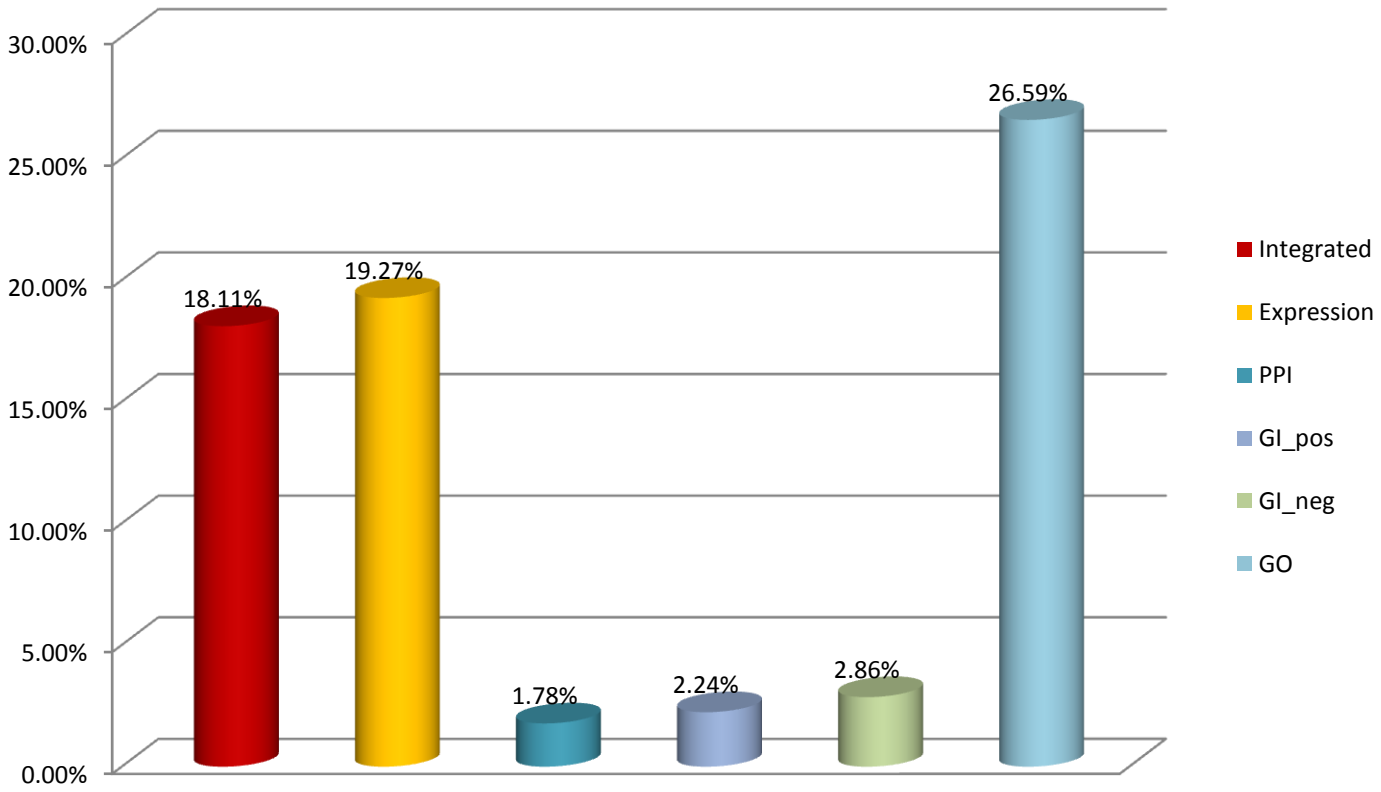
$$S_{GO}(A,B) = \frac{\sum\limits_{t \in T_A \cap T_B} \left(S_A(t) + S_B(t)\right)}{\sum\limits_{t \in T_A} S_A(t) + \sum\limits_{t \in T_B} S_B(t)}$$

    - And finally, semantic similarity of a pair of genes

$$Sim(G1,G2) = \frac{\sum\limits_{1 \le i \le m} \max\limits_{1 \le j \le n}\left(S_{GO}(go_{1i}, go_{2j})\right) + \sum\limits_{1 \le j \le n} \max\limits_{1 \le i \le m}\left(S_{GO}(go_{1i}, go_{2j})\right)}{m+n}$$

# Basic interaction data conservation rates

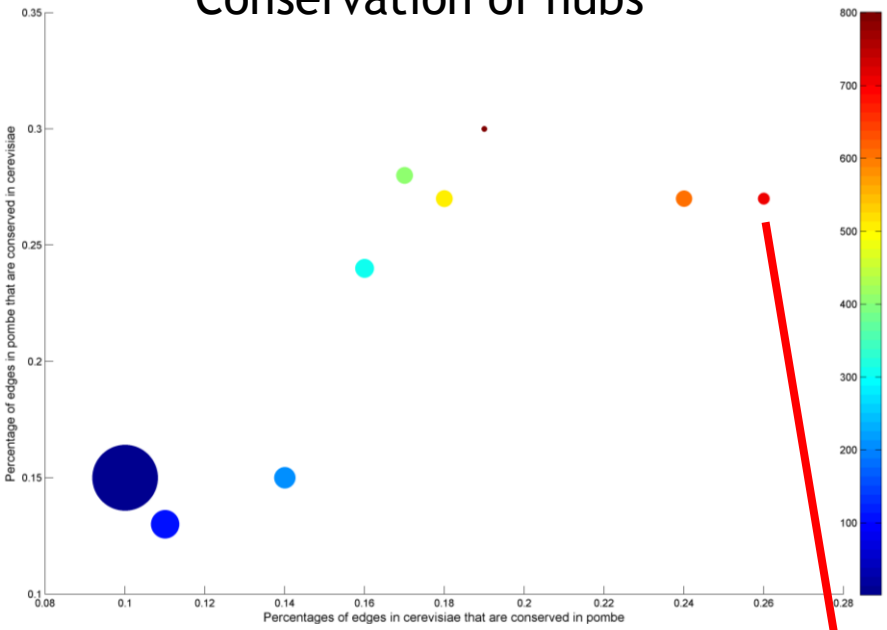
## Example for *S. cerevisiae* wrt. *S. pombe*



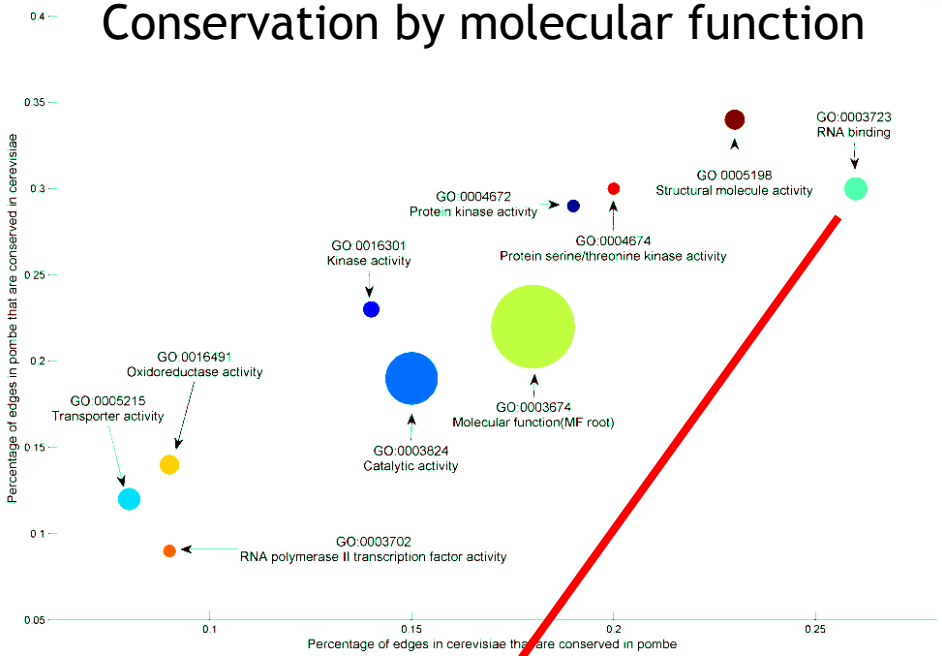
# All the previous explanations are limited

## Example for *S. cerevisiae* wrt. *S. pombe*

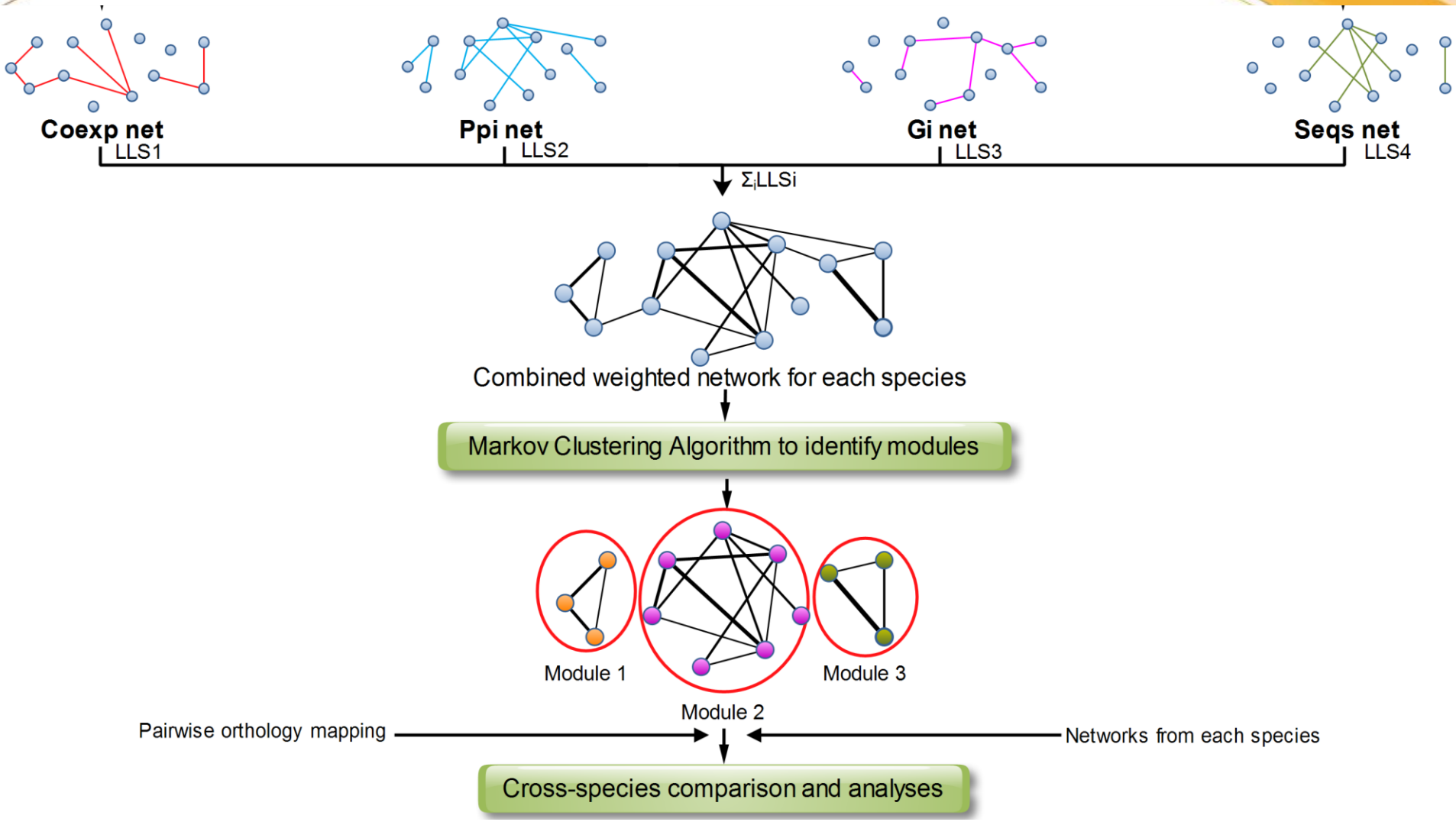
Conservation of hubs




Conservation by molecular function




| | Basic | Hubs | Complexes | Molecular Function |
|---|---|---|---|---|
| Integrated | **18.11%** | **26%** | **26%/35%** | **26%** |

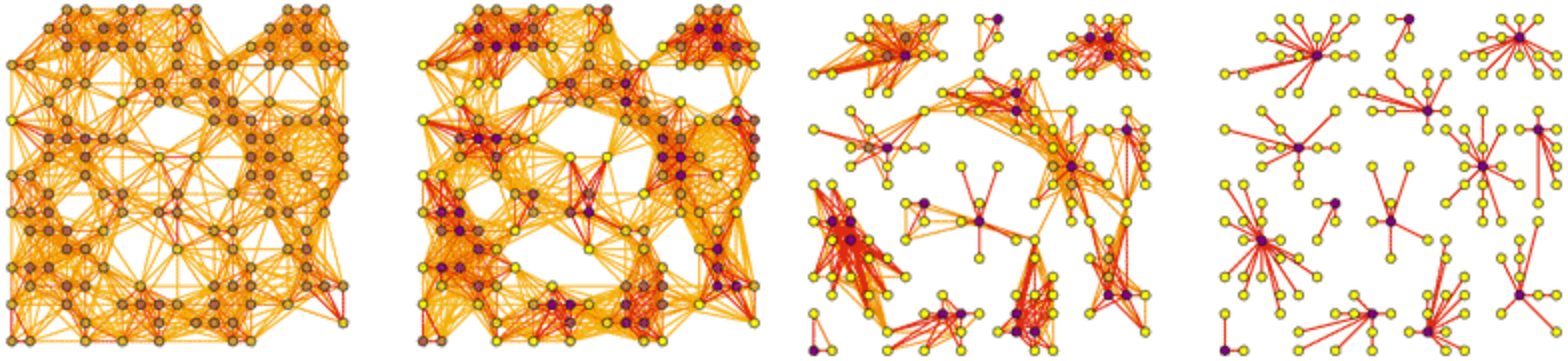

Coexp net
LLS1
Ppi net
LLS2
Gi net
LLS3
Seqs net
LLS4
$\Sigma_i$LLSi
Combined weighted network for each species
Markov Clustering Algorithm to identify modules
Module 1
Module 2
Module 3
Pairwise orthology mapping
Networks from each species
Cross-species comparison and analyses

# Markov CLustering algorithm (MCL)

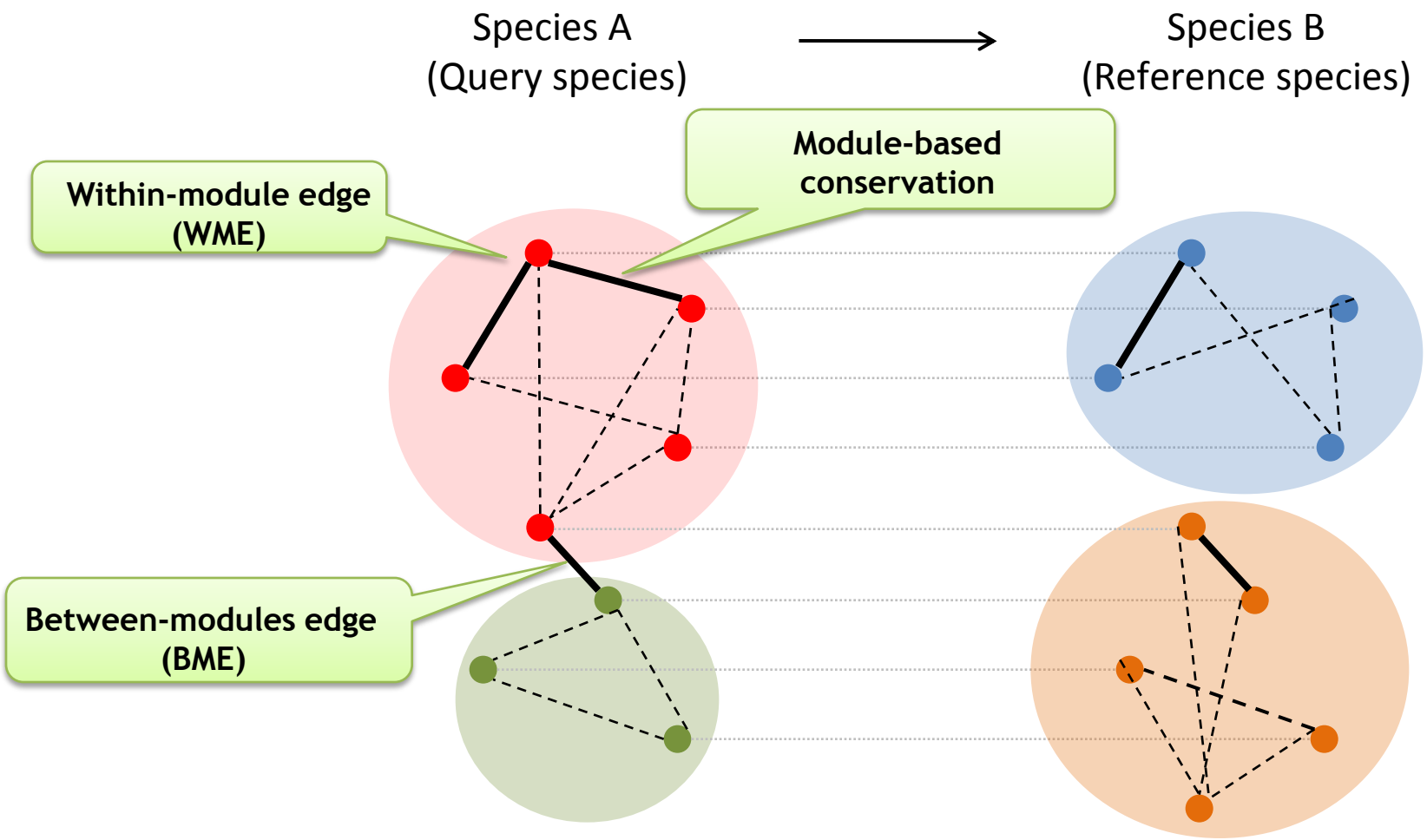
- **MCL partitions a graph via simulation of random walks.**
  - Random walks on a graph are likely to get stuck within dense subgraphs rather than shuttle between dense subgraphs via sparse connections (van Dongen, 2000).

- **The method effectively places each node into exactly one cluster.**

- **MCL proved to be robust to random edge addition and removal (Brohee and van Helden, 2006).**

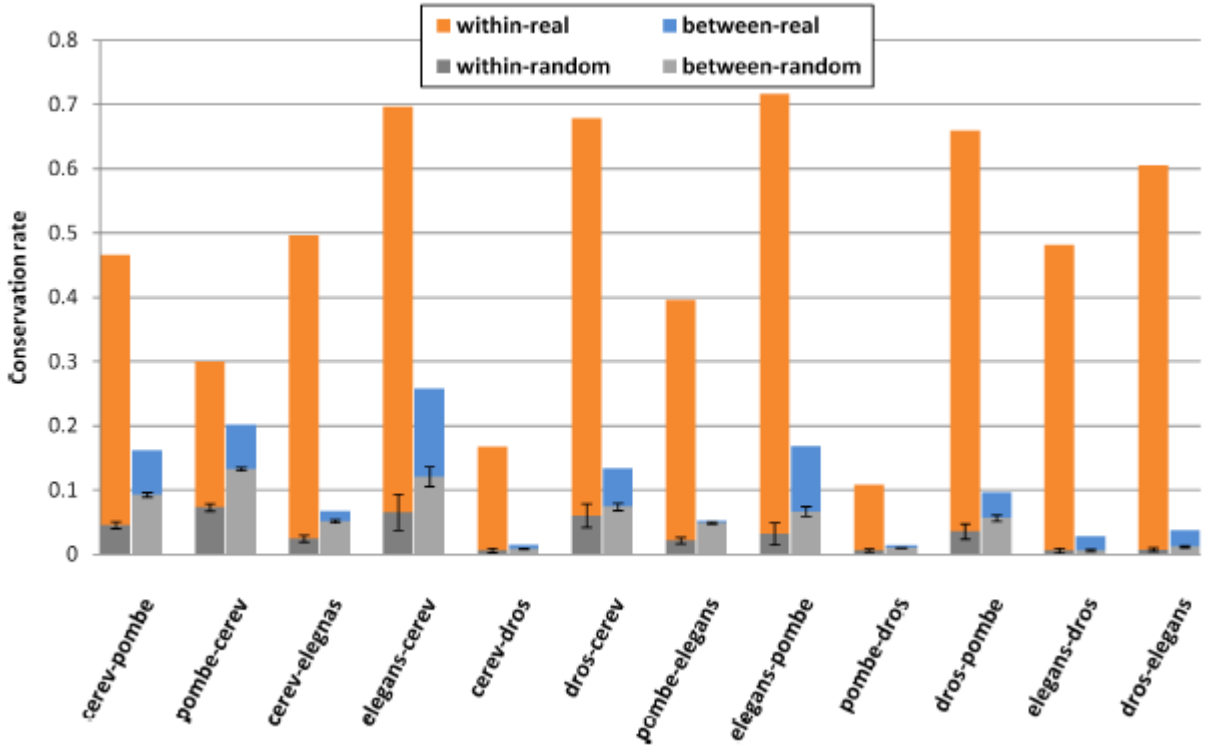


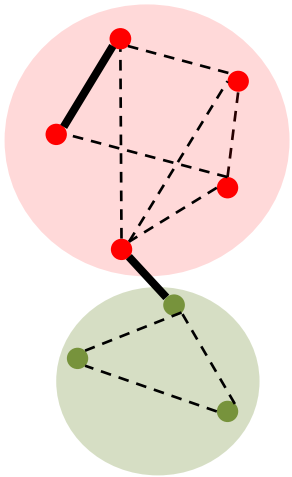
# Are module interactions more conserved?



# WME conservation vs. BME conservation

- **Within-module edges are much more conserved than between-modules edges.**

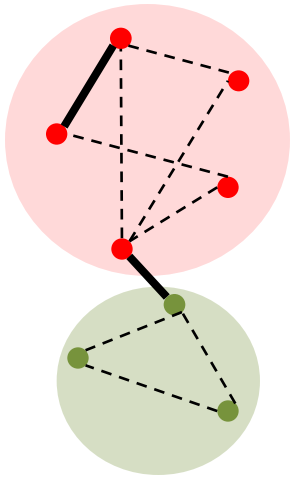


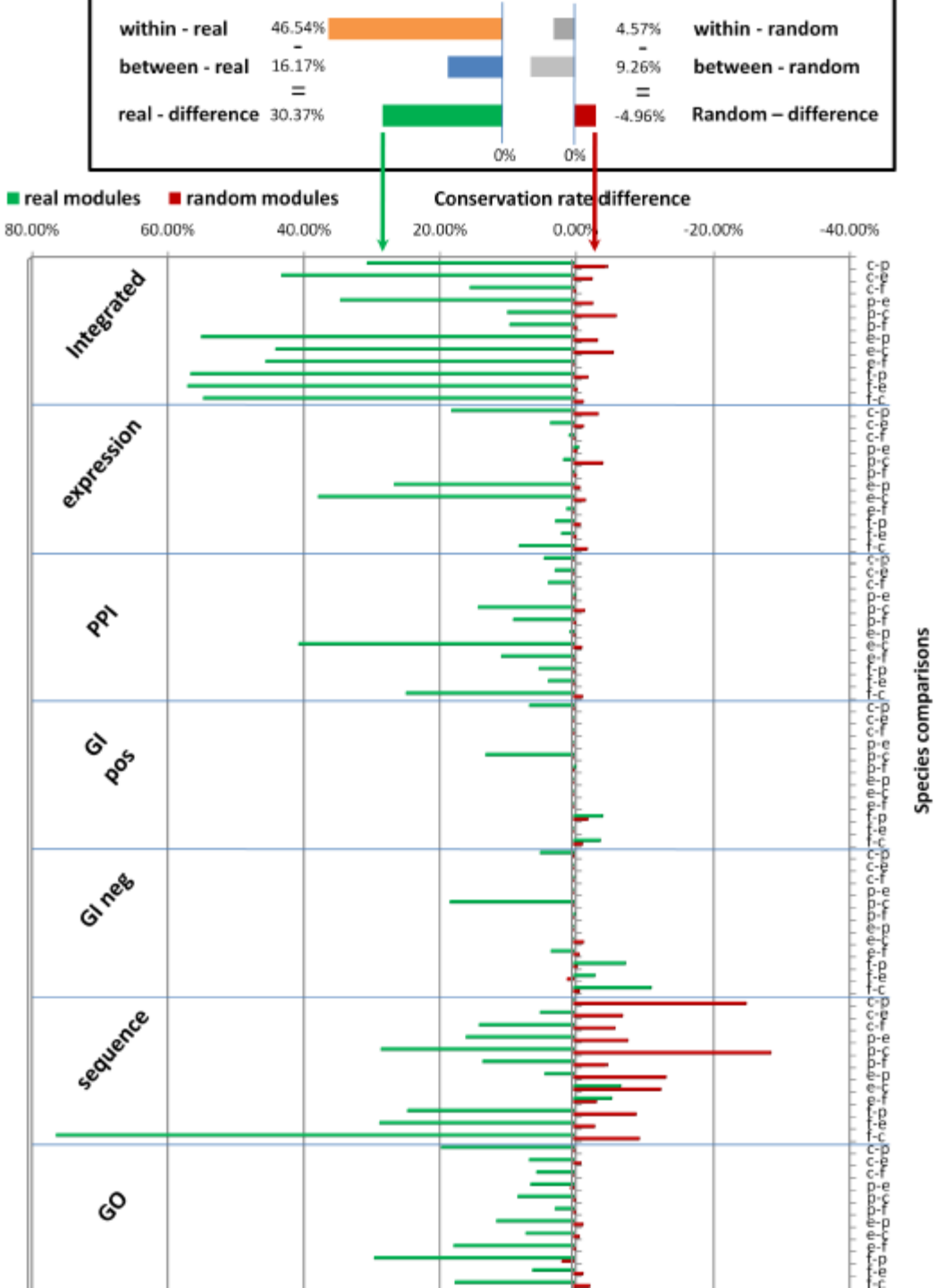



| Baseline | Previous explanations | | | Module based explanations | | |
|---|---|---|---|---|---|---|
| | Hubs | Complexes | Molecular function | WMI | WMI -no hubs | WMI ext. |
| 18.11% | 26% | 26%/35% | 26% | 46.54% | 42.87% | 49.66% |

# Within-modules vs. between modules edge conservation

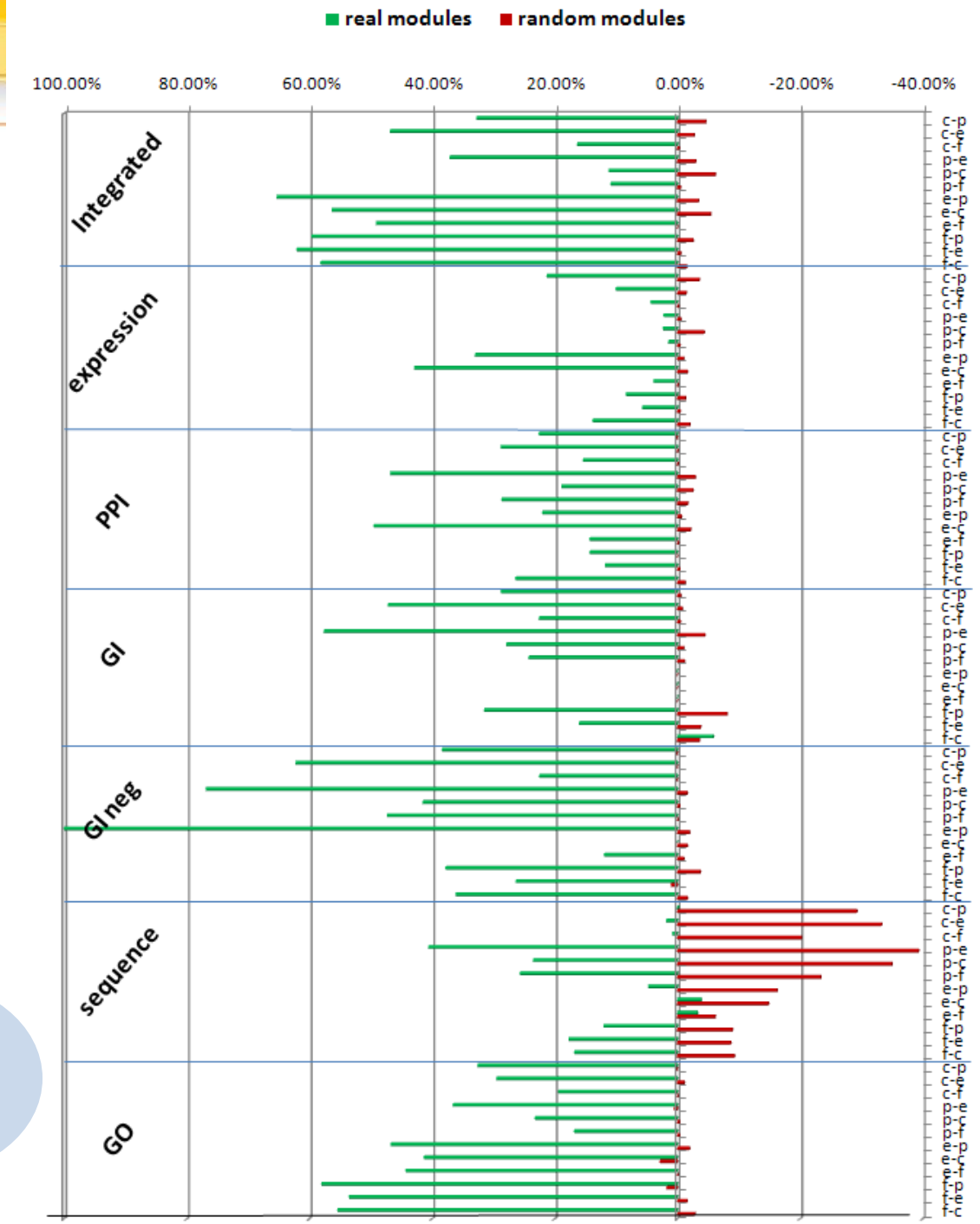
- Trend is consistent across almost all data types and almost all cross species comparisons.
- This trend is not seen in random modules.





# Module-based conservation

- Even higher conservation rates
- Trend is consistent across almost all data types and almost all cross species comparisons.
- This trend is not seen in random modules.

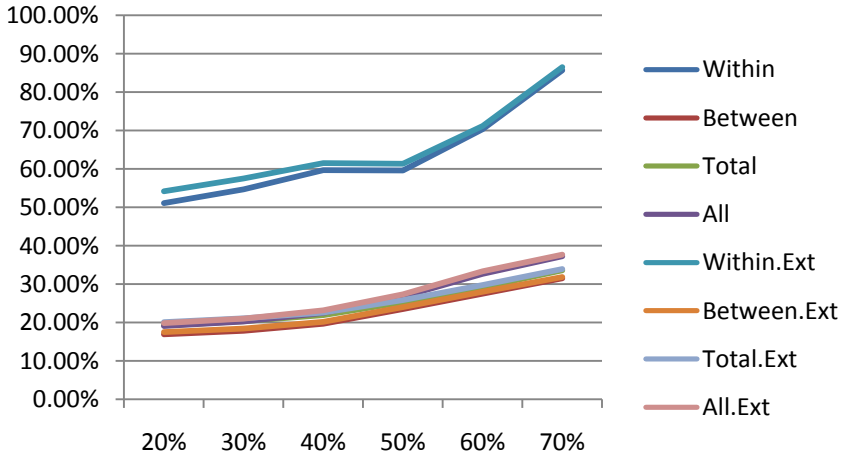


# Modules & results are robust

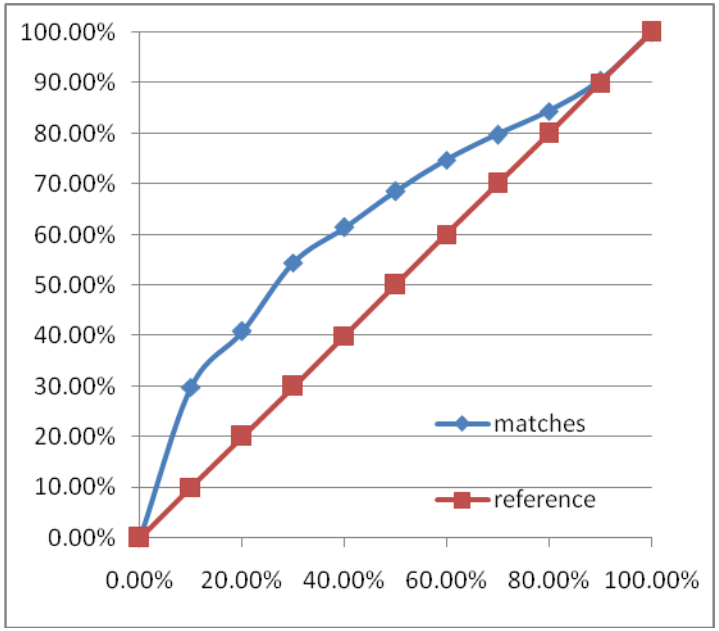
*S. cerevisiae* vs. *S. pombe*

Did not change the trend:

- Other clustering algorithms (SPICi)
- Other randomization methods (nodes)
- Removing sequence networks
- M:N Inparanoid mapping
- Stricter orthology relations (by varying the minimum sequence similarity on orthology relations)




- Insufficient data coverage - randomly removing interactions from the *S. cerevisiae* network, retained many modules.

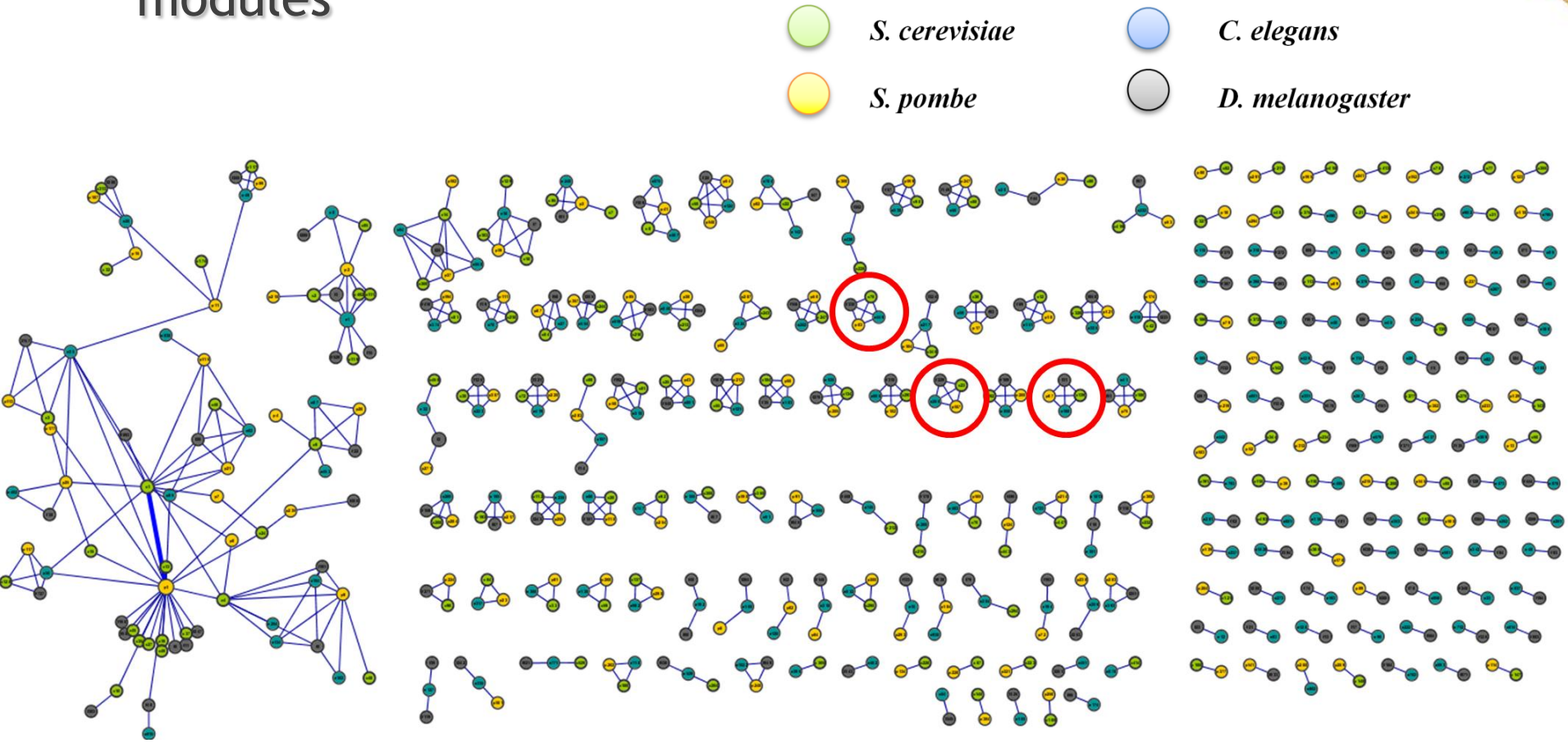


# Matching between modules

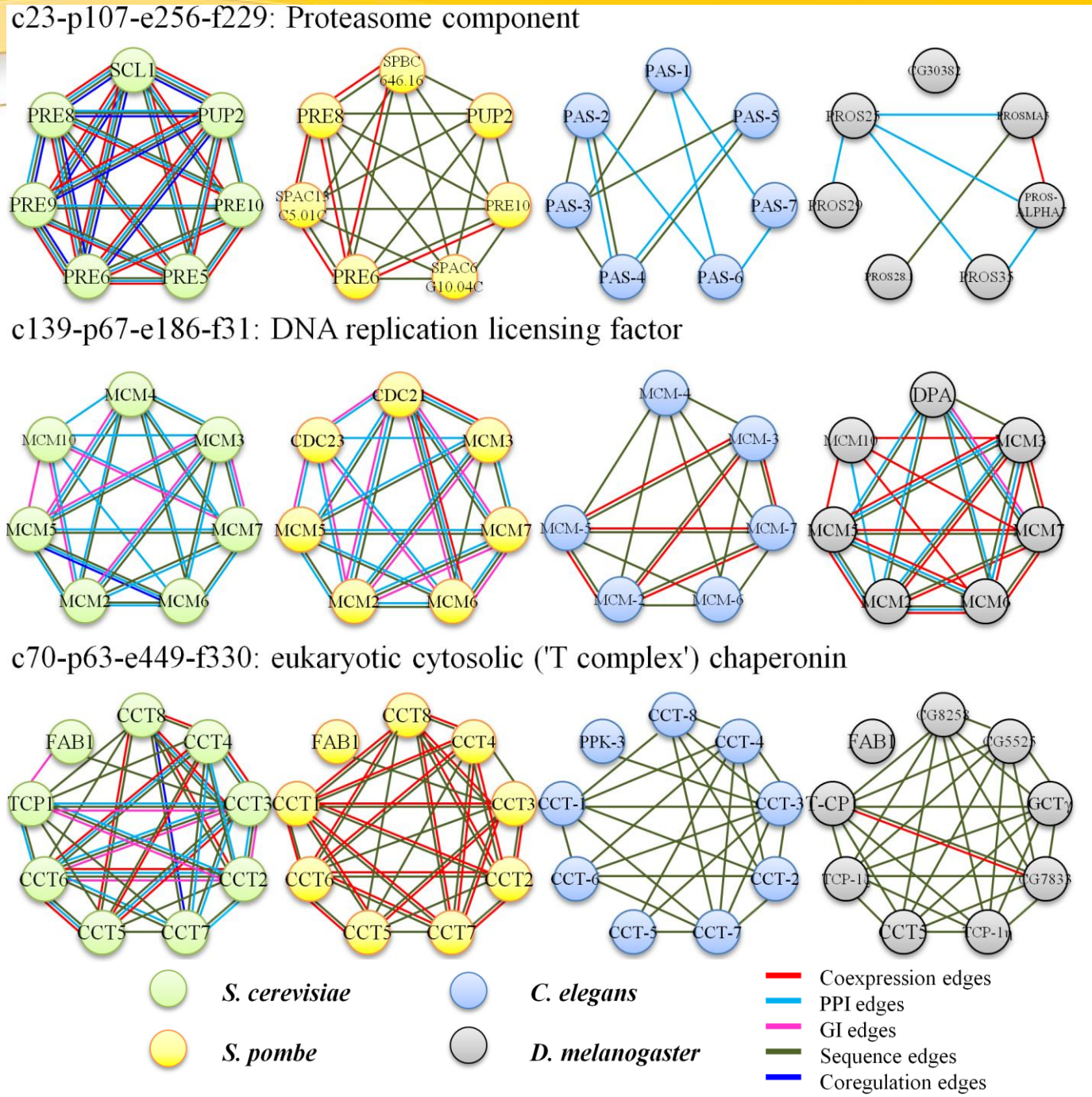
**Based on a reciprocal conditional hyper-geometric test**

*p*(*M* orthologs | *N* nodes) * *p*(*m* matches | *M* orthologs)

Bonferroni corrected cutoffs on all pairwise reciprocal matches



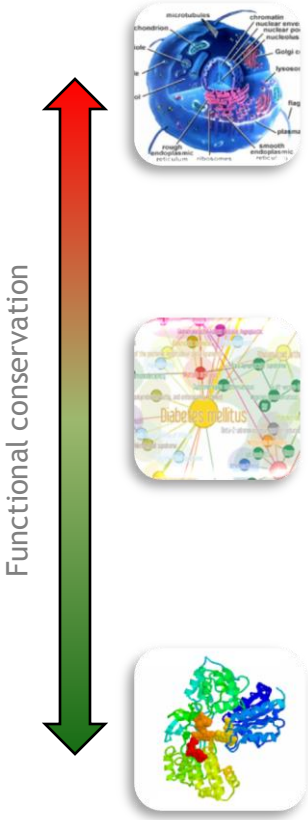
## Examples for matched modules and their interactions



# Conclusions for this part

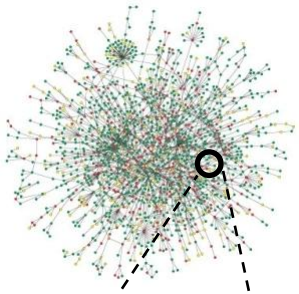
- There is a big discrepancy between system level conservation and conservation of individual edges

- Intermediate ‘meta gene’ modules were found to have a higher rate of conservation:
  - Within-module edges are more conserved than the general case.
  - Higher conservation rate is seen for module-based conservation, that is significantly different from random.



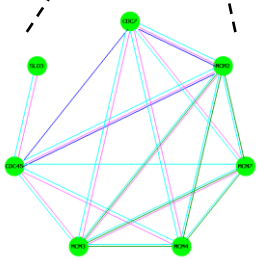
# Analogy to sequence similarity

Genome



Network

Gene

Module

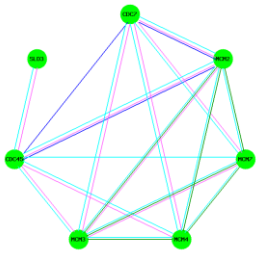
Coding sequence more conserved

module interactions more conserved

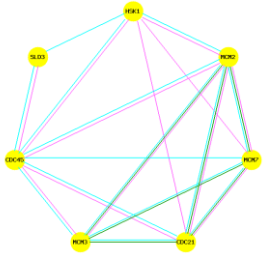
Gene product

Species1
…AGGCTCGAA**CGG**ACCGGGTAC…
Species2
…AGGCTCGAA**CGA**ACCGGGTAC…
Protein
…ArgLeuGlu**Arg**ThrGlyTyr…

Species1



Species2



Module function

DNA replication initiation

Next talk:

# Tools for comparing expression data across species