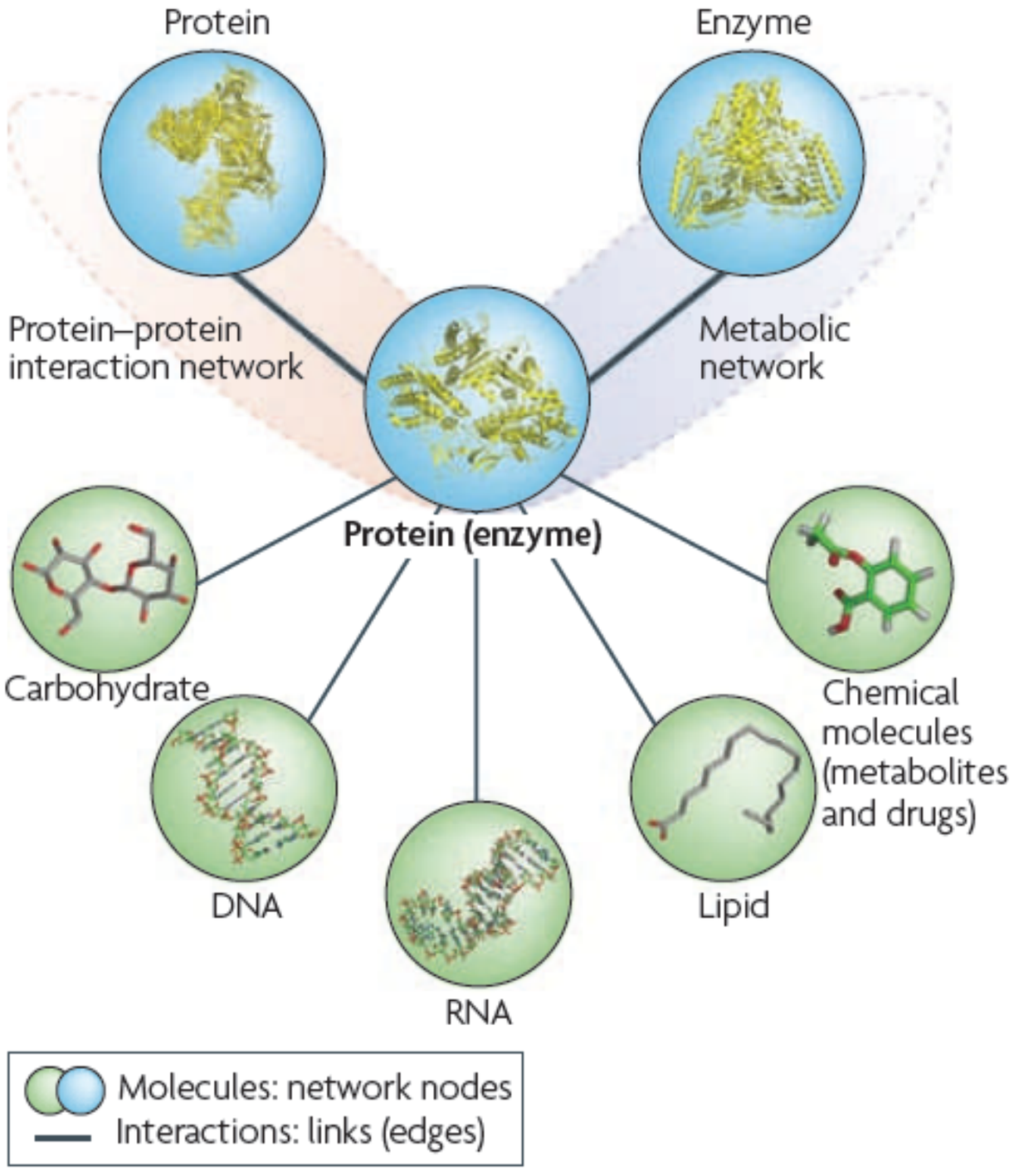
Bioinformatics: Network Analysis

Comparative Network Analysis

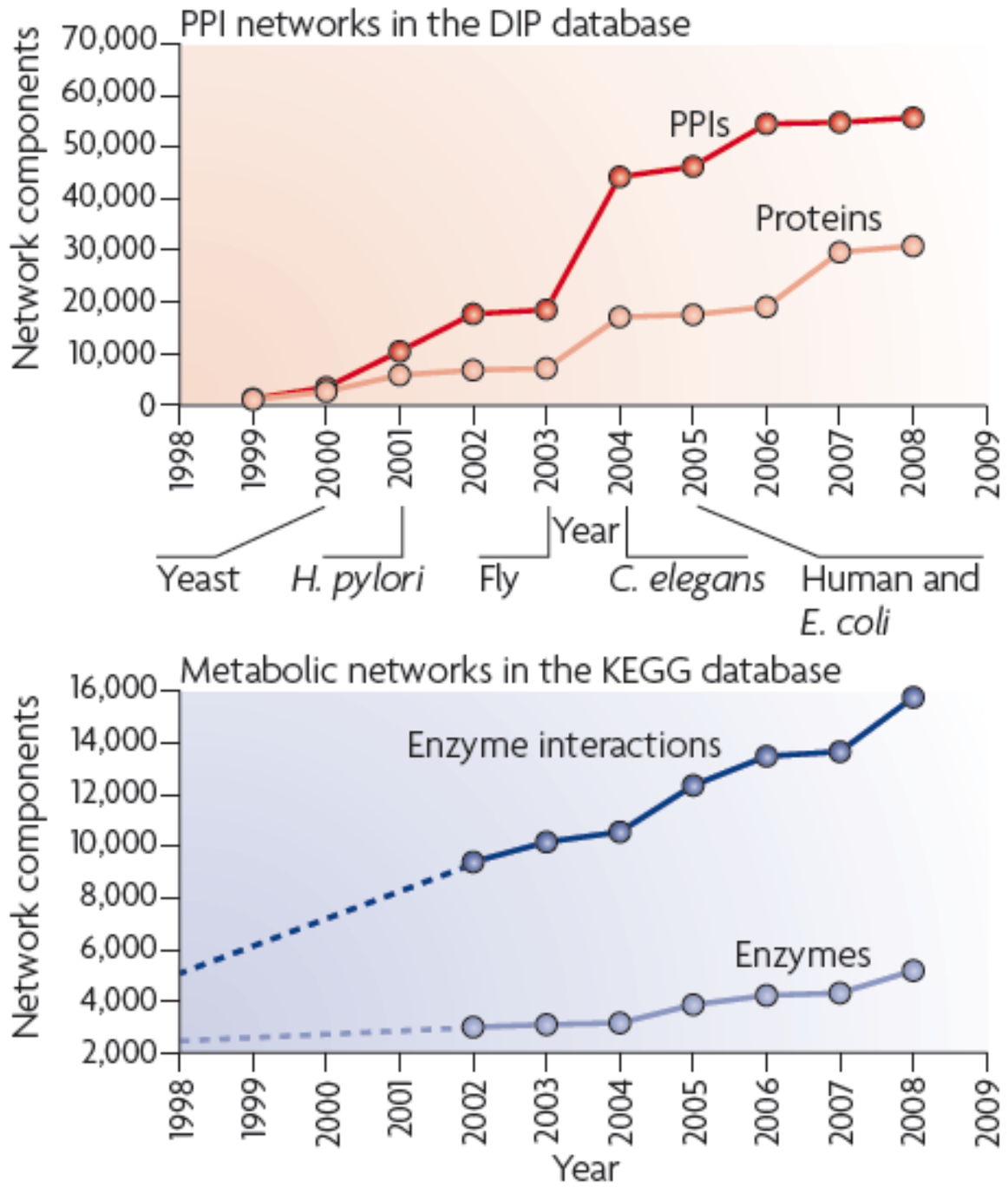
COMP 572 (BIOS 572 / BIOE 564) - Fall 2013

Luay Nakhleh, Rice University

Biomolecular Network Components



Accumulation of Network Components



MINT

Interactions:

- **total**: 103808
- **Saccharomyces**: 41919
- **Mammalia**: 23212
- **Drosophila**: 23205
- **Caenorhabditis**: 4915
- **Viruses**: 1364

Proteins:

- **total**: 28186
- **Mammalia**: 8082
- **Drosophila**: 7165
- **Saccharomyces**: 5223
- **Caenorhabditis**: 2960
- **Viruses**: 267

Pmids:

- **total**: 3385
- **Mammalia**: 2766
- **Viruses**: 304
- **Saccharomyces**: 236
- **Drosophila**: 82
- **Caenorhabditis**: 43

main detection methods: (number of interactions)

- **two hybrid pooling approach**: 43137
- **anti tag coimmunoprecipitation**: 15414
- **tandem affinity purification**: 13123
- **two hybrid**: 13104
- **pull down**: 2891
- **two hybrid fragment pooling approach**: 2706
- **two hybrid array**: 2139
- **coimmunoprecipitation**: 2137
- **anti bait coimmunoprecipitation**: 1501
- **protein array**: 1044
- **gst pull down**: 981
- **ubiquitin reconstruction**: 682
- **experimental knowledge based**: 547
- **experimental interaction detection**: 543
- **x-ray crystallography**: 412
- **colocalization by immunostaining**: 395
- **affinity chromatography technologies**: 375
- **filter binding**: 349
- **interaction prediction**: 294
- **peptide array**: 287
- **colocalization/visualisation technologies**: 255
- **cross-linking studies**: 174
- **protein kinase assay**: 129
- **copurification**: 122
- **enzyme linked immunosorbent assay**: 119
- **surface plasmon resonance**: 99
- **beta lactamase complementation**: 91
- **his pull down**: 81
- **nuclear magnetic resonance**: 71
- **far western blotting**: 58

main organisms: (number of interactions)

- **Saccharomyces cerevisiae**: 41919
- **Drosophila melanogaster**: 23311
- **Homo sapiens**: 19723
- **Caenorhabditis elegans**: 4914
- **Escherichia coli**: 4125
- **Mus musculus**: 3244
- **Plasmodium falciparum (isolate 3D7)**: 2704
- **Escherichia coli O157:H7**: 2668
- **Rattus norvegicus**: 1659
- **Helicobacter pylori**: 1626
- **Escherichia coli O6**: 312
- **Human papillomavirus type 16**: 244
- **Bos taurus**: 226
- **Human herpesvirus 3**: 175
- **Arabidopsis thaliana**: 155
- **Human herpesvirus 8**: 126
- **Gallus gallus**: 126
- **Shigella flexneri**: 122
- **Schizosaccharomyces pombe**: 112
- **Hepatitis C virus**: 101
- **Oryctolagus cuniculus**: 99
- **Human papillomavirus type 18**: 95
- **Xenopus laevis**: 91
- **Human adenovirus 5**: 91
- **Pyrococcus horikoshii**: 89
- **Simian virus 40**: 83
- **Bovine papillomavirus type 1**: 67
- **Epstein-Barr virus (strain B95-8)**: 55
- **Canis familiaris**: 50
- **Human papillomavirus type 11**: 46

(Statistics downloaded March 18, 2008)

Database of Interacting Proteins

Number of proteins	19490
Number of organisms	161
Number of interactions	56186
Number of distinct experiments describing an interaction	64208
Number of data sources (articles)	3257
Number of data sources (other)	34

ORGANISM	PROTEINS	INTERACTIONS	EXPERIMENTS	Details
Drosophila melanogaster (fruit fly)	7066	21004	21106	
Saccharomyces cerevisiae (baker's yeast)	4920	18272	23238	
Escherichia coli	1847	7427	9065	
Caenorhabditis elegans	2643	4037	4082	
Homo sapiens (Human)	1124	1648	2423	
Helicobacter pylori	710	1425	1425	[---]
Mus musculus (house mouse)	264	374	527	
Rattus norvegicus (Norway rat)	102	132	196	
Others (153)	814			

(Statistics downloaded March 18, 2008)

How do we make sense of all this data?

Nothing in Biology Makes Sense Except in the Light of Evolution

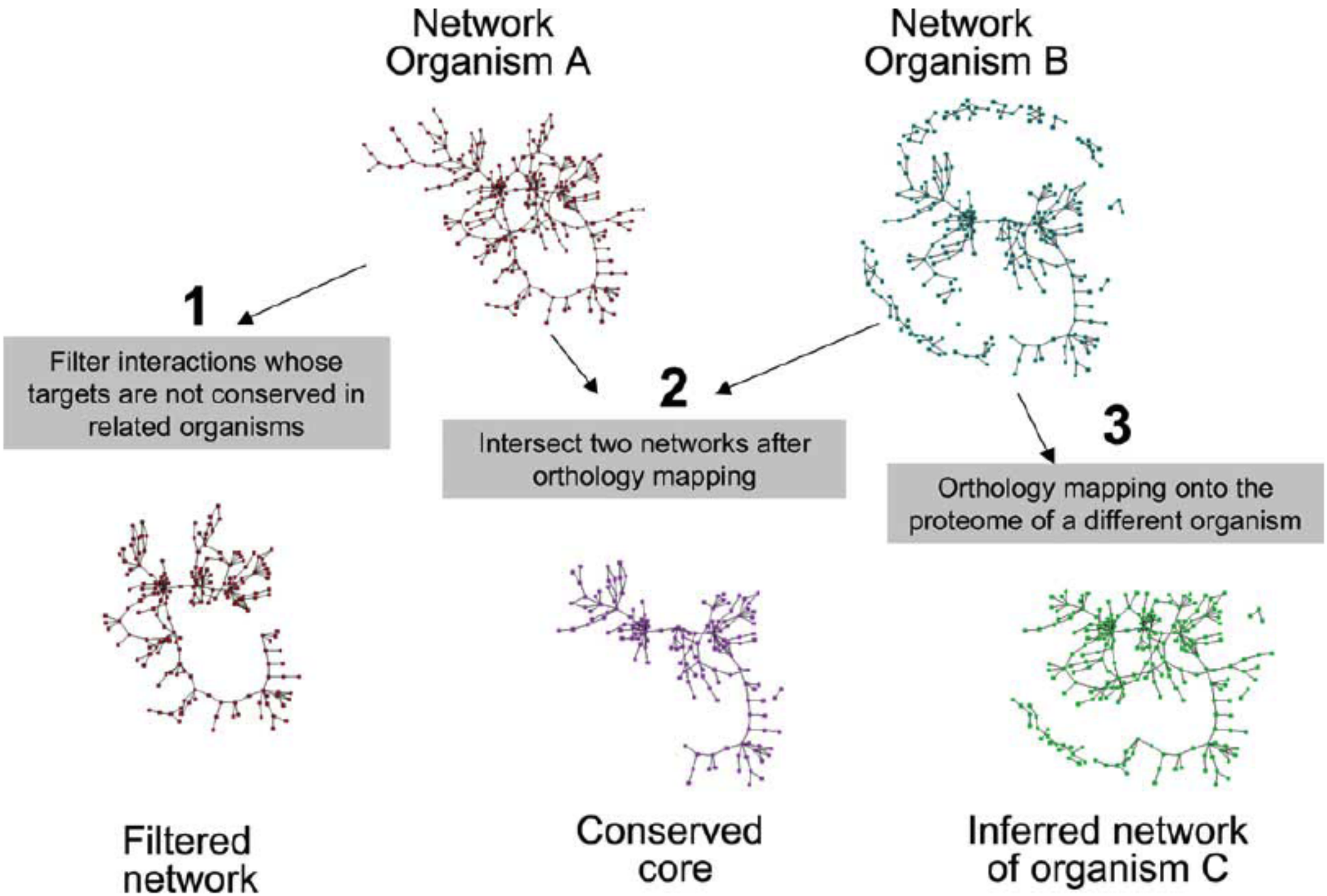
Theodosius Dobzhansky (1900-1975)

- Work over the past 50 years has revealed that molecular mechanisms underlying fundamental biological processes are conserved in evolution and that models worked out from experiments carried out in simple organisms can often be extended to more complex organisms

- This observation forms the basis for using interaction networks derived from experiments in model organisms to obtain information about interactions that may occur between the ortholog proteins in different organisms

- Further the observation allows for identifying "functional" modules based on conservation of network components

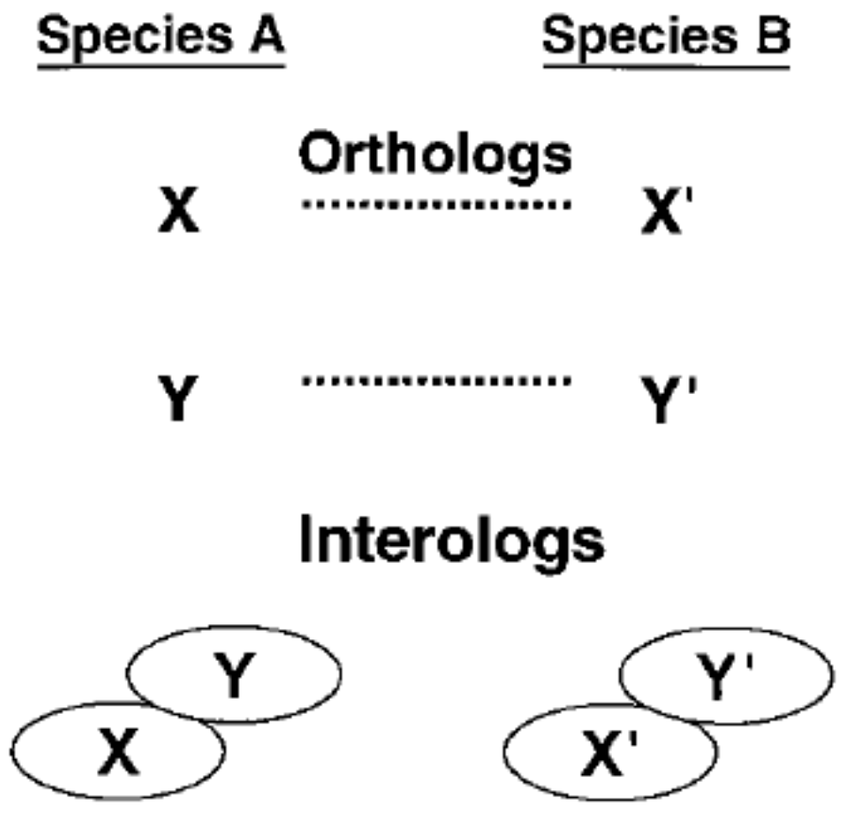
Comparative Interactomics



Protein Interaction Mapping in C. elegans Using Proteins Involved in Vulval Development

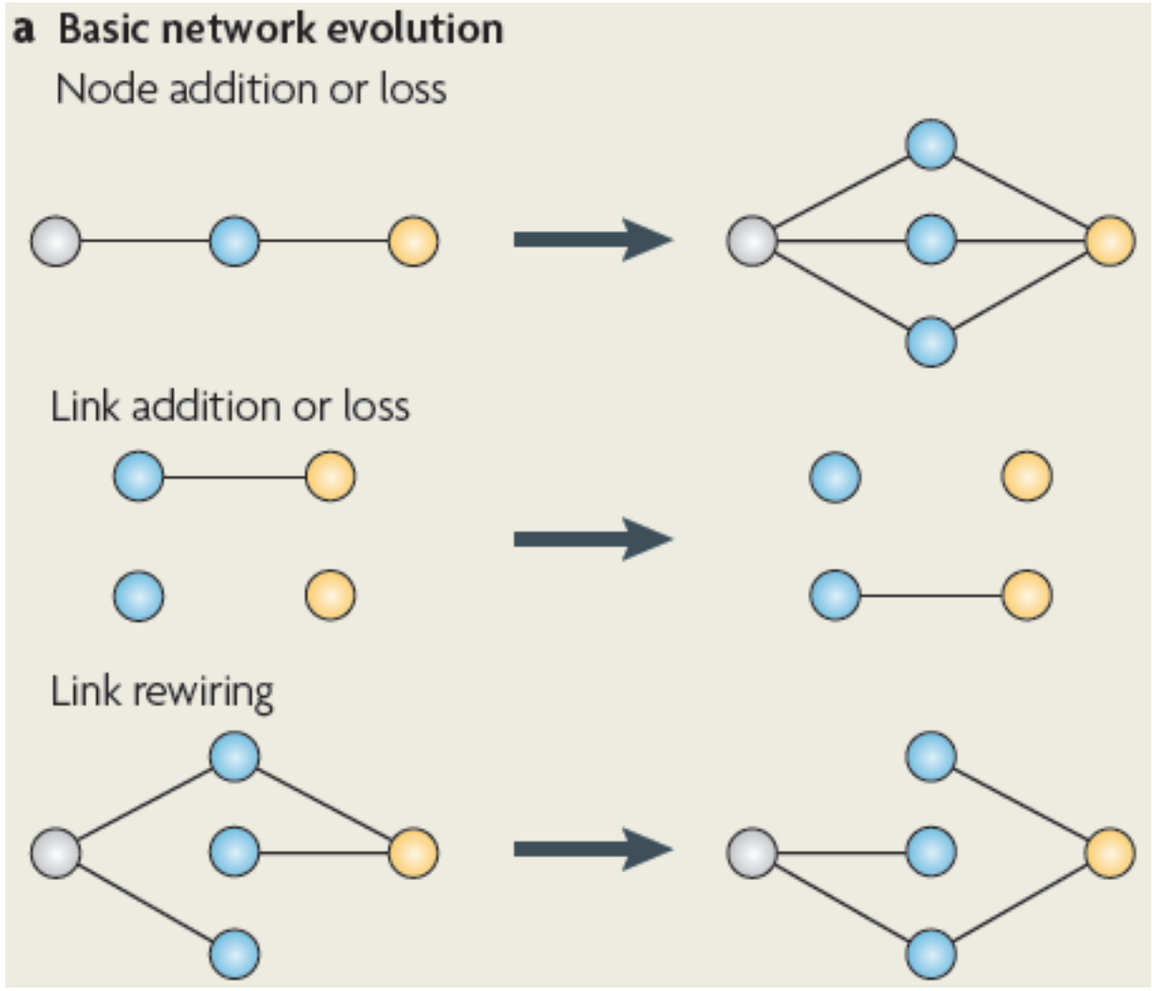
Albertha J. M. Walhout, *et al.*

Science **287**, 116 (2000);



Evolutionary Models for PPI and Metabolic Networks

The evolution of biomolecular networks is coupled to several genetic events. Node addition or loss in PPI and metabolic networks (see the figure, part **a**) usually implies that a gene duplication or loss has occurred and, implicitly, that the addition or loss of links has occurred, because each gene duplicate should keep all of its existing interactions. Horizontal transfer is another means of node addition, in which the impact on links can vary. Link addition or loss usually implies that genetic changes, such as point mutations, domain accretion or loss, alternative splicing, insertions or deletions, have occurred in genes or their regulatory regions. These genetic changes can destroy or create links. Link rewiring is usually a mixture of consequences from link addition or loss, often also involving secondary effects from node addition or loss.

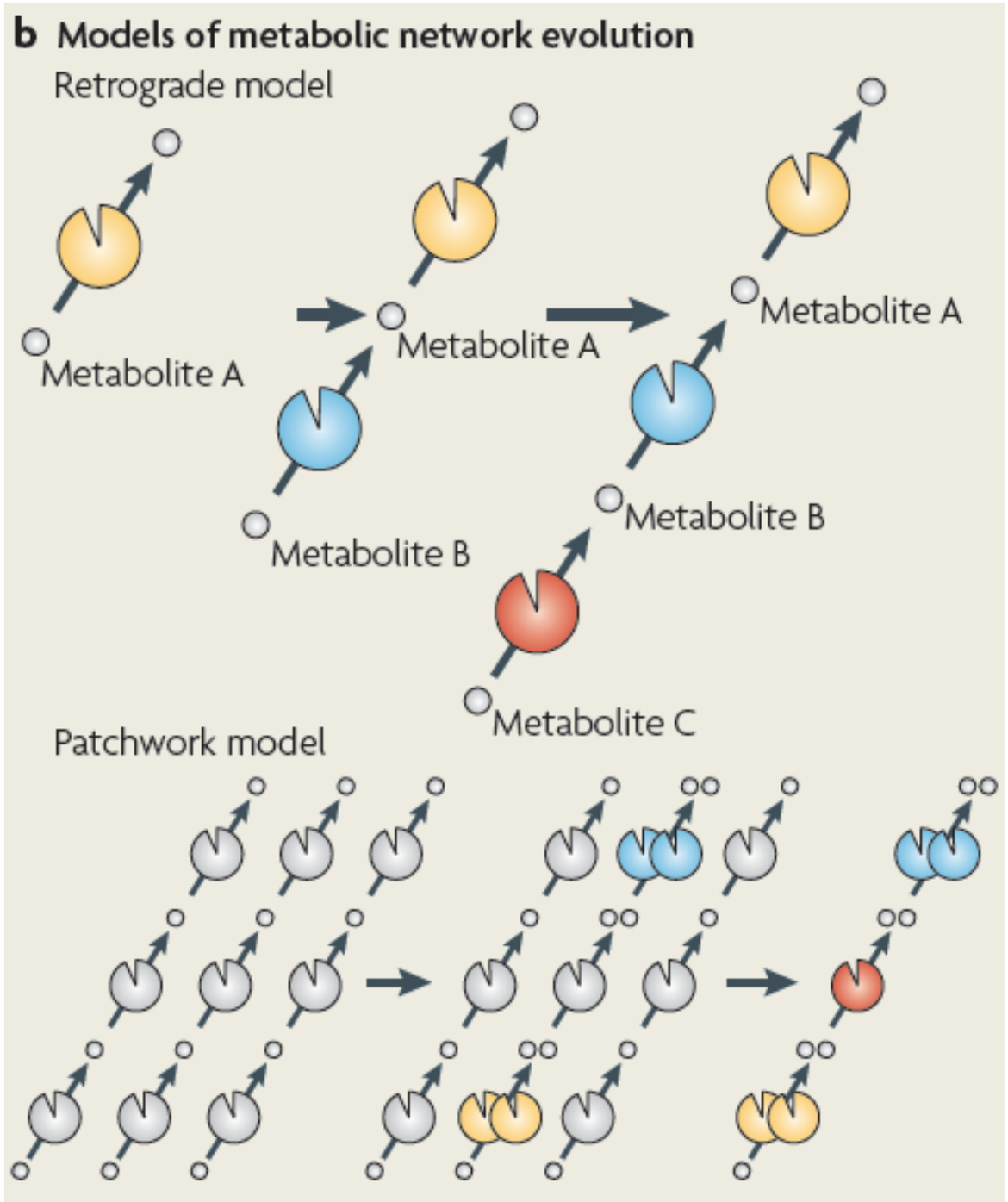


Evolutionary Models for PPI and Metabolic Networks

For the evolution of metabolic networks (see the figure, part **b**), environmental chemical conditions have to be considered. The following two models are the representative models that are specific to metabolism. The enzyme colours in the figure represent the order of recruitment: first yellow, second blue and third red.

Retrograde model. This model assumes that pathways are evolving backwards from a key metabolite. First, an organism that is heterotrophic for key metabolite A uses up all of the environmental supply of A. Second, the recruitment of an enzyme capable of synthesizing A from precursor B brings a selective advantage to the organism. In turn, environmental concentrations of B drop and this is compensated for by the recruitment of enzymes capable of synthesizing B from C.

Patchwork model. This model assumes that enzymes refine their substrate specificity after duplication. Initially, most of the enzymes have broad substrate specificities, which can catalyse multiple reactions. These broad substrate specificities of enzymes enable the generation of many metabolic pathways for the synthesis of the same key metabolites. The duplications of genes in such metabolic pathways bring selective advantage to the pathways because an increased level of the enzyme will generate more of the key metabolites. Finally, enzyme specialization following the gene duplication events will lead to the specialization of the different pathways.



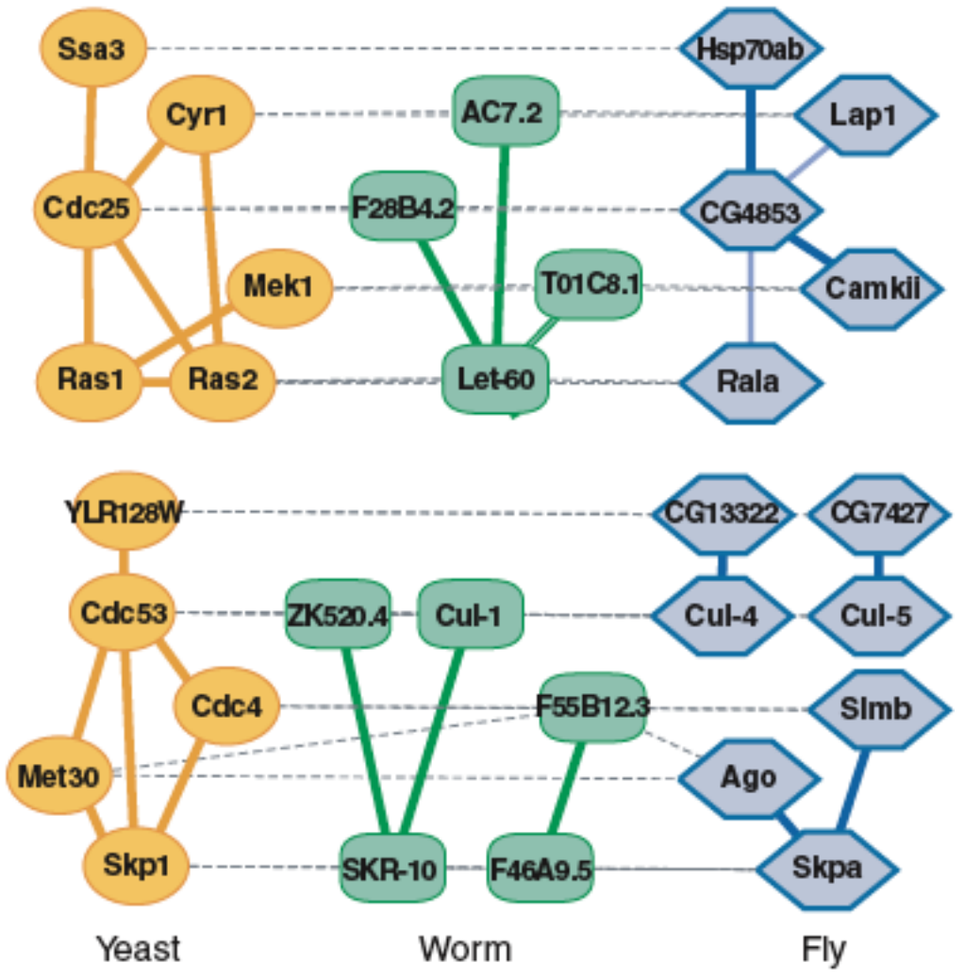
The Network Alignment Problem

- Given a set $\{N_1, N_2, ..., N_k\}$ of PPI networks from k organisms, find subnetworks that are conserved across all k networks
- The problem in general is NP-hard (even for k=2), generalizing subgraph isomorphism
- Several heuristics have been developed

The Network Alignment Problem

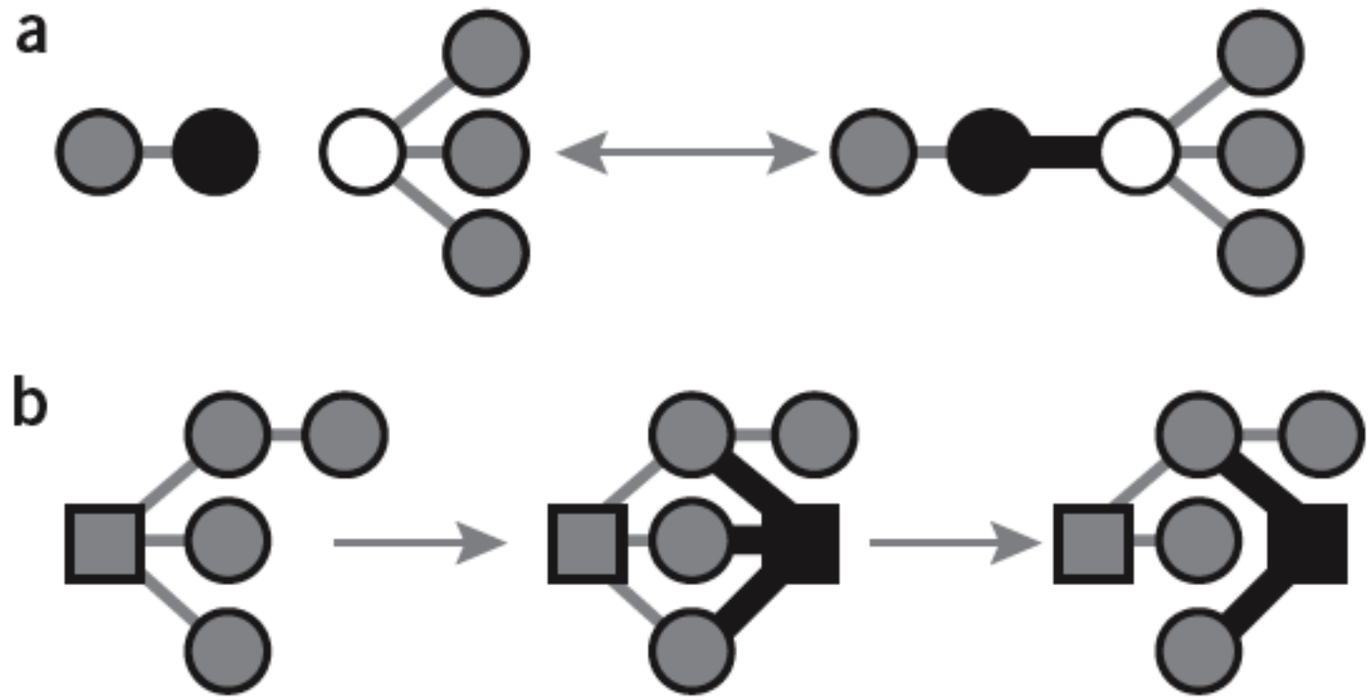
- In general, the output of the network alignment problem is a "conserved subnetwork"
- In particular:
  - a conserved linear path may correspond to a signaling pathway
  - a conserved cluster of interactions may correspond to a protein complex

Multiple alignment of protein interaction networks



Matching proteins are linked by dotted lines, and **yellow**, **green** or **blue** links represent measured protein-protein interactions between yeast, worm or fly proteins, respectively.

Evolutionary Processes Shaping Protein Interaction Networks



Evolutionary processes shaping protein interaction networks. The progression of time is symbolized by arrows. (**a**) **Link attachment and detachment** occur through mutations in a gene encoding an existing protein. These processes affect the connectivity of the protein whose coding sequence undergoes mutation (shown in black) and of one of its binding partners (shown in white). Empirical data shows that attachment occurs preferentially towards partners of high connectivity. (**b**) **Gene duplication** produces a new protein (black square) with initially identical binding partners (gray square). Empirical data suggest that duplications occur at a much lower rate than link attachment/ detachment and that redundant links are lost subsequently (often in an asymmetric fashion), which affects the connectivities of the duplicate pair and of all its binding partners.

Challenges in Comparative Interactomics

- The error rate in genomic data is in the order of 0.1% or less, whereas 30% or more of the interactions reported in high-throughput experiments have been estimated to be artefactual [40,41].
- Genomes are sequenced using a standardized method. By contrast, several different approaches are used to detect interactions. These methods will be biased and are likely to discover different sets of interactions [40,42].
- Many genome sequences are considered complete or almost complete (taking into account assembly errors, possible errors in repetitive sequences, etc.). The same definition of completeness is hardly applicable to interactomes, and only the yeast *S. cerevisiae* has been systematically analyzed for binding partners to each of its ~6000 protein products.
- Genomes are relatively stable over an individual life span and are independent of tissue, although expression activity can vary greatly. Conversely, interactomes vary in space (different tissues, different cell organelles) and time (development, stage in the cell cycle).
- Genes are suitably represented by strings of nucleotides but no such obvious representation exists for interactomes. Although limited by the absence of temporal and spatial information, undirected graphs are currently used, with nodes representing individual proteins and edges representing the interactions between them.
- Finally, although in genomics one compares well defined and 'immutable' chemical entities, namely nucleotides and amino acids, the graph nodes being compared in interactomics represent mutable entities, proteins. This raises the problem of reliably matching functionally and evolutionarily related proteins in the proteomes of unrelated species.

The Rest of This Lecture

- Pairwise network alignment
- Multiple network alignment

Pairwise Network Alignment

One heuristic approach creates a merged representation of the two networks being compared, called a network alignment graph, and then applies a greedy algorithm for

identifying the conserved subnetworks embedded in the merged representation

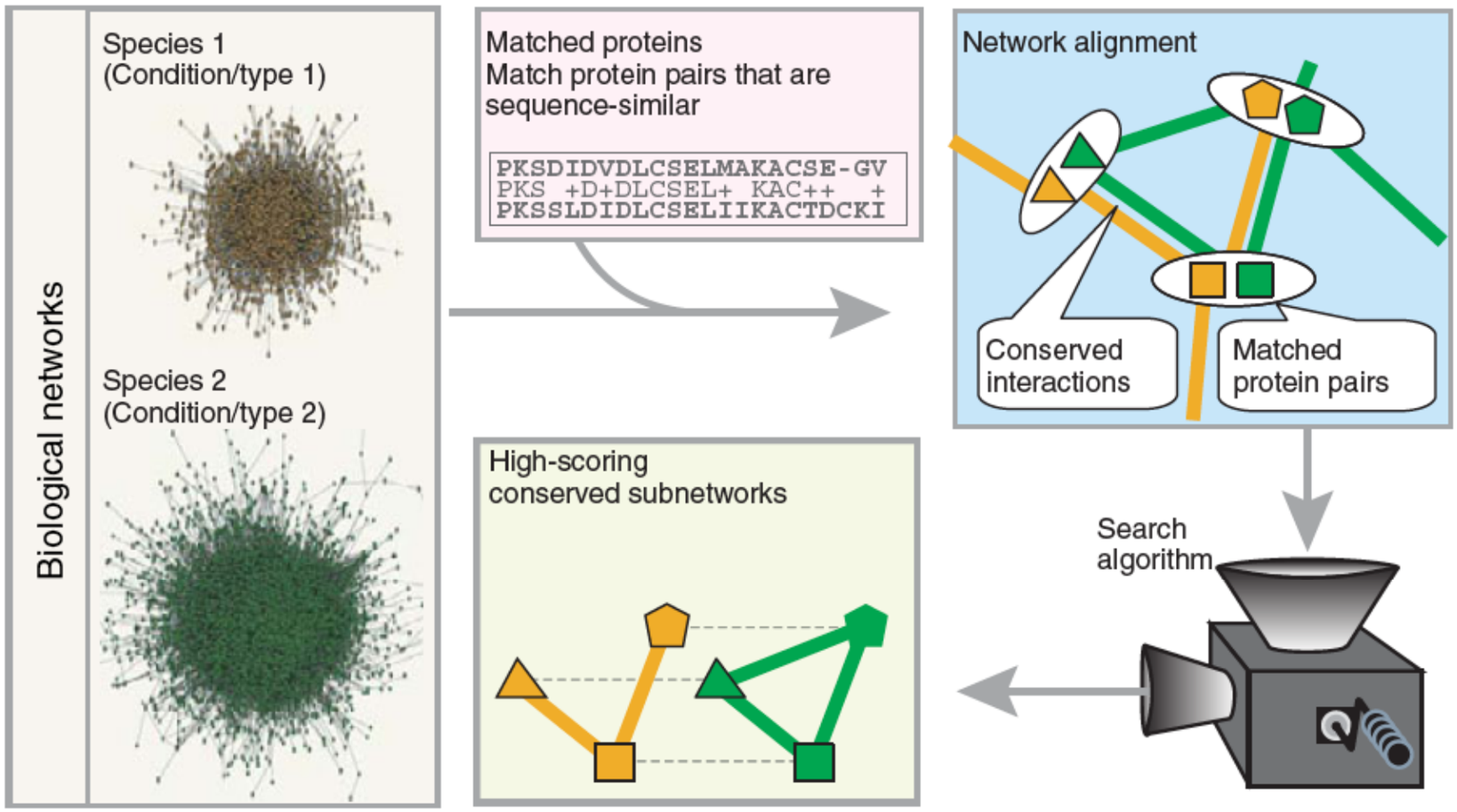


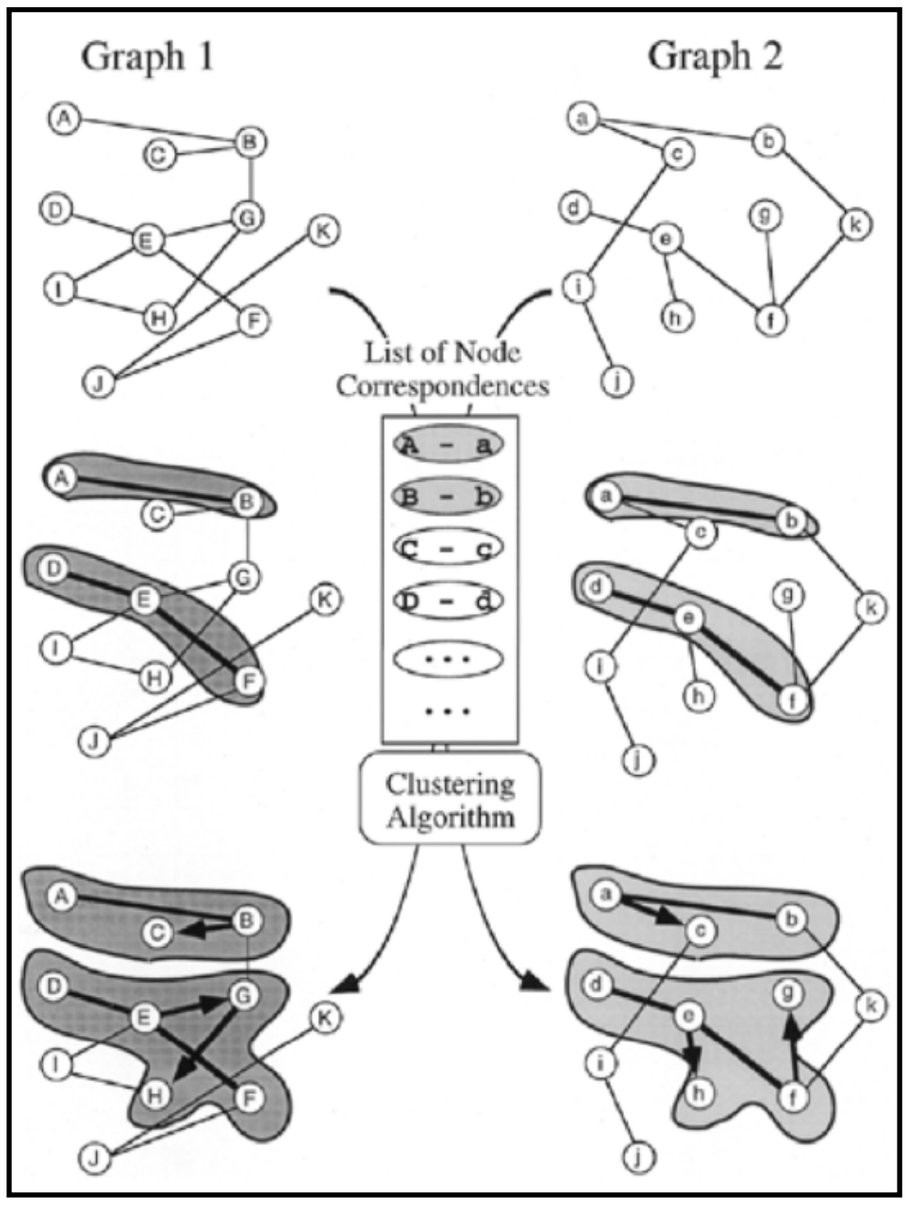
Figure 1 Network alignment. Network alignment combines protein interaction data that are available for each of at least two species with orthology information based on the corresponding protein sequences. A detailed probabilistic model is used to identify protein subnetworks within the aligned network that are conserved across the species. Each node in this aligned network represents a set of sequence-similar proteins (one from each species) and each link represents a conserved interaction. Other than species, the networks being compared can also be sampled across different biological conditions or interaction types.

© 2000 Oxford University Press *Nucleic Acids Research, 2000, Vol. 28, No. 20* ***4021–4028***

A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters

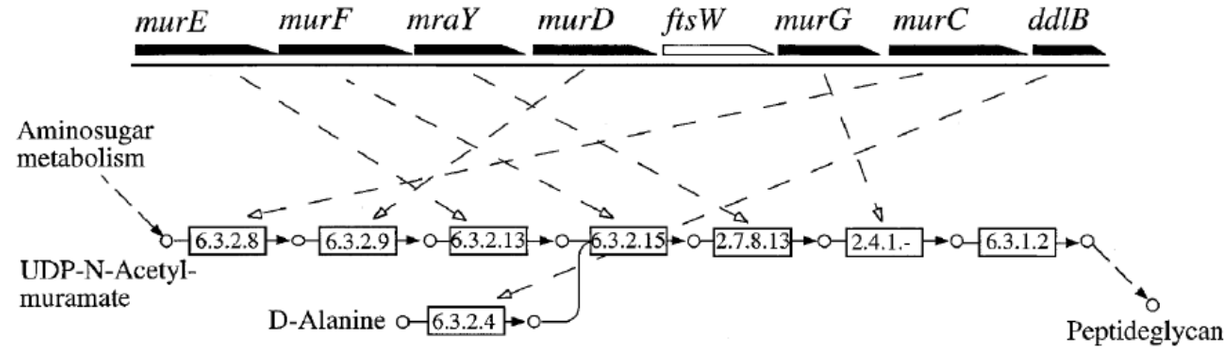
Hiroyuki Ogata, Wataru Fujibuchi, Susumu Goto and Minoru Kanehisa*

Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan



They searched for correspondence between reactions of specific metabolic pathways and the genomic locations of the genes encoding the enzymes catalyzing those reactions

Their network alignment graph combined the genome ordering information (network of genes arranged in a path) with a network of successive enzymes in metabolic pathways



The source code (Perl) and data are available at:
http://kanehisa.kuicr.kyoto-u.ac.jp/Paper/fclust/

PathBLAST

Conserved pathways within bacteria and yeast as revealed by global protein network alignment

Brian P. Kelley*, Roded Sharan†, Richard M. Karp†, Taylor Sittler*, David E. Root*, Brent R. Stockwell*, and Trey Ideker*‡

Kelley et al. applied the concept of network alignment to the study of PPI networks. They translated the problem of finding **conserved pathways** to that of finding high-scoring paths in the alignment graph

The algorithm, PathBLAST, identified five regions that were conserved across the PPI networks of S. cerevisiae and H. pylori

http://www.pathblast.org

Gaps and Mismatches

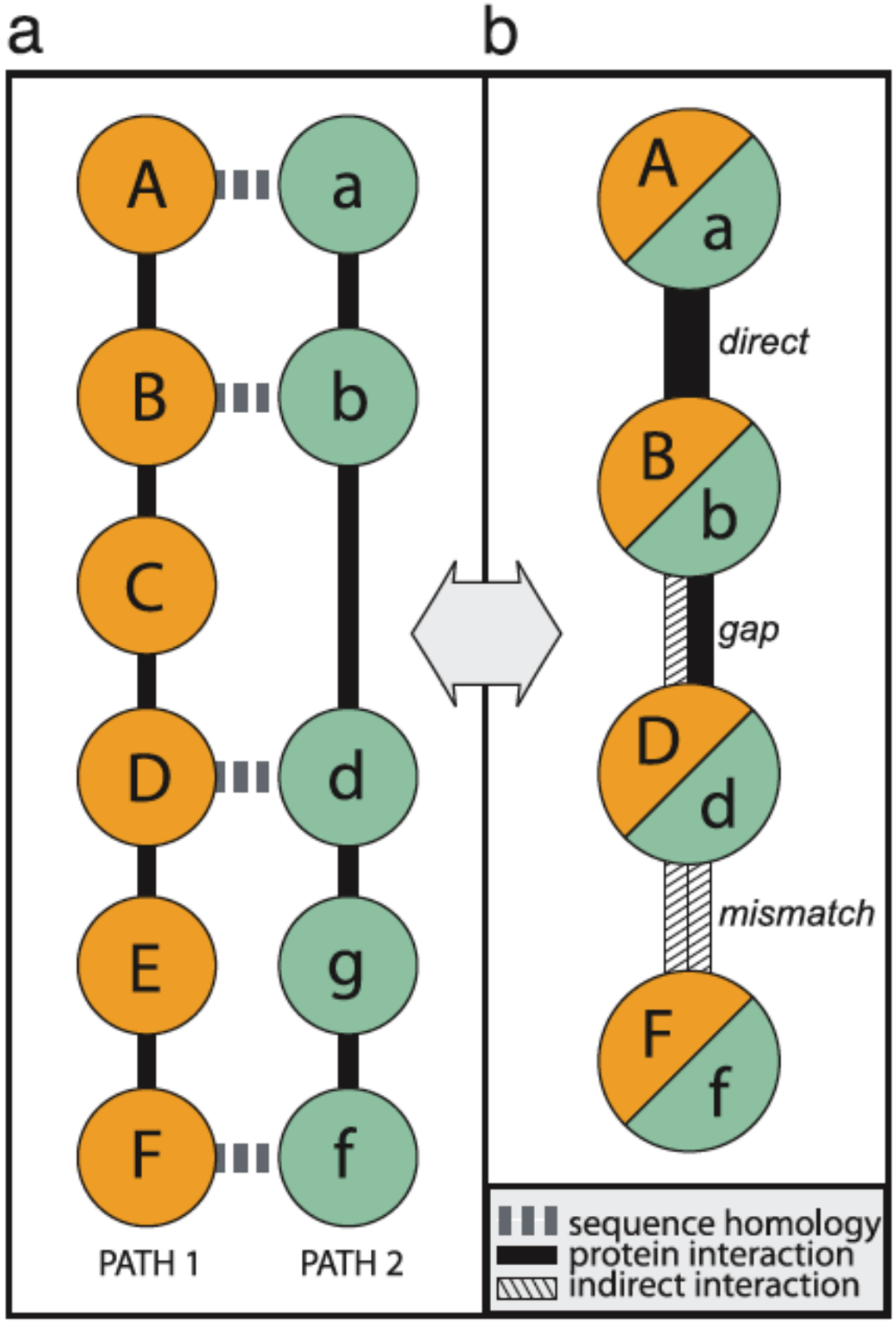


Fig. 1. Example pathway alignment and merged representation. (*a*) Vertical solid lines indicate direct protein–protein interactions within a single pathway, and horizontal dotted lines link proteins with significant sequence similarity (BLAST *E* value $\leq E_{cutoff}$). An interaction in one pathway may skip over a protein in the other (protein *C*), introducing a ''gap.'' Proteins at a particular position that are dissimilar in sequence (*E* value $> E_{cutoff}$, proteins *E* and *g*) introduce a ''mismatch.'' The same protein pair may not occur more than once per pathway, and neither gaps nor mismatches may occur consecutively. (*b*) Pathways are combined as a global alignment graph in which each node represents a homologous protein pair and links represent protein interaction relationships of three types: direct interaction, gap (one interaction is indirect), and mismatch (both interactions are indirect).

Global Alignment and Scoring

- To perform the alignment of two PPI networks, the two networks are combined into a global alignment graph (figure on previous slide), in which each vertex represents a pair of proteins (one from each network) having at least weak sequence similarity (BLAST E value $\leq 10^{-2}$) and each edge represents a conserved interaction, gap, or mismatch

- A path through this graph represents a pathway alignment between the two networks

- A log probability score S(P) is formulated

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

  where *p(v)* is the probability of true homology within the protein pair represented by *v*, given its pairwise protein sequence similarity expressed as BLAST E value, and *q(e)* is the probability that the PPIs represented by e are real

- Protein sequence alignments and associated E values were computed by using BLAST 2.0 with parameters b=0, e=1x10^6, f="C;S", and v=6x10^5. Unalignable proteins were assigned a maximum E value of 5

Optimal Pathway Alignment and Significance

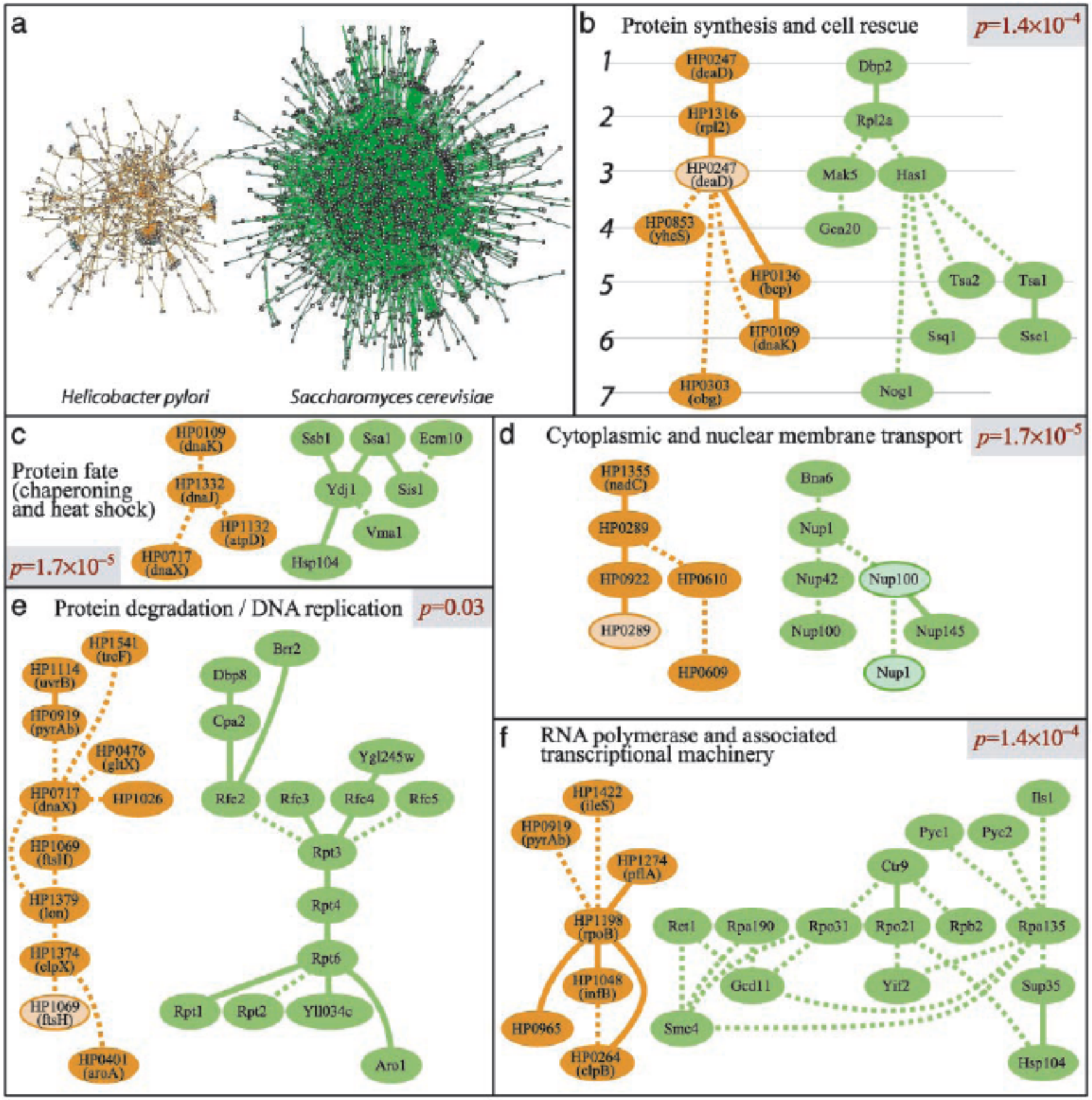
- Once the alignment graph was built, optimal pathway alignment were searched for
- The authors considered simple paths of length 4, and used a dynamic programming algorithm that finds the highest-scoring path of length L in linear time (in acyclic graphs)
- Because the global alignment graph may contain cycles, the authors generated a sufficient number, 5L!, of acyclic subgraphs by random removal of edges from the global alignment graph and then aggregated the results of running dynamic programming on each

Optimal Pathway Alignment and Significance

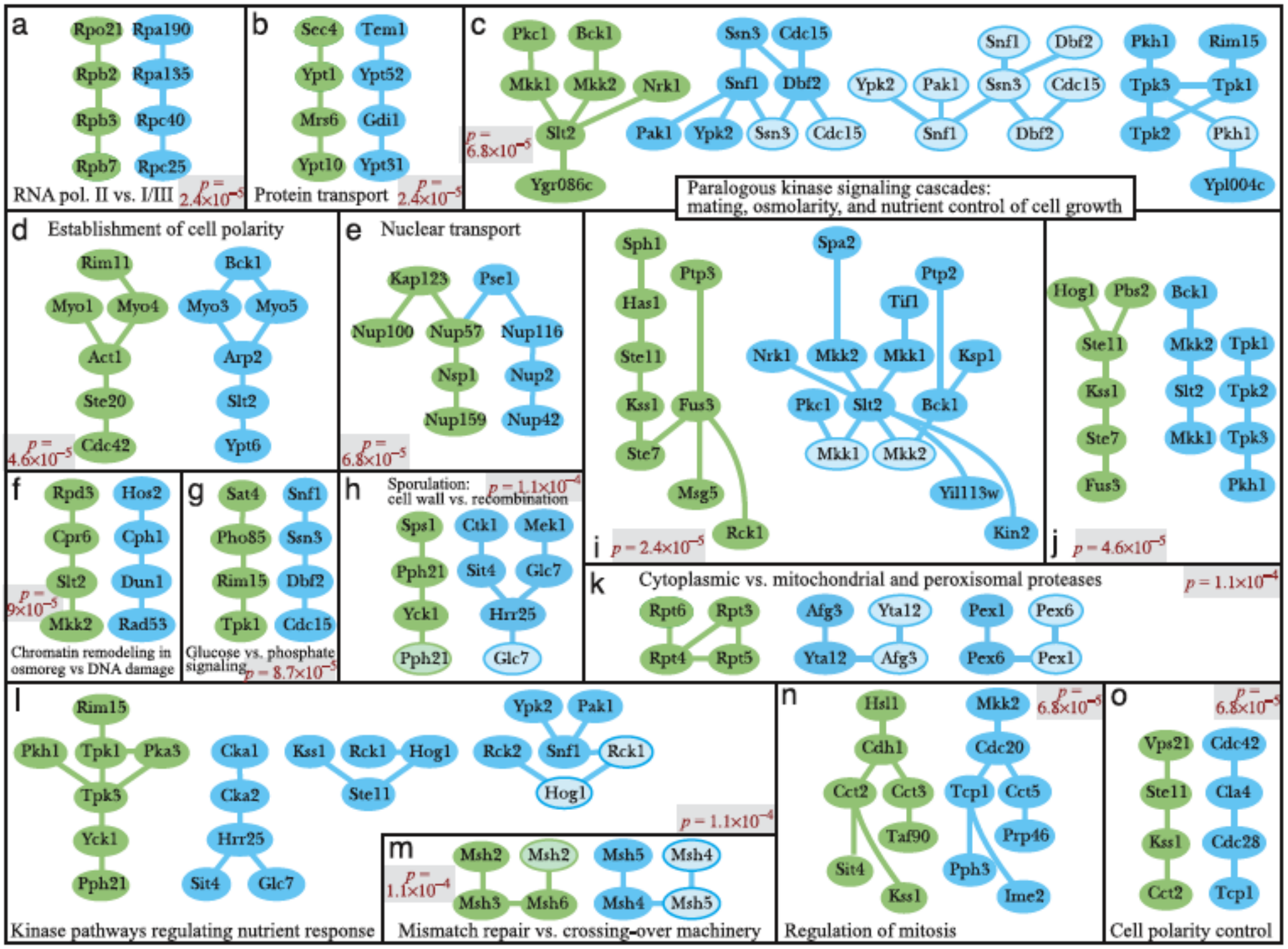
- Because conserved regions of the network could be highly interconnected, it was sometimes possible to identify a large number of distinct paths involving the same small set of proteins
- Rather than enumerate each of these, PathBLAST was used in stages
- For each stage k, the authors recorded the set of 50 highest-scoring pathway alignments (with average score $\langle S_k \rangle$) and then removed their vertices and edges from the alignment graph before the next stage
- The p value of each stage was assessed by comparing <Sk> to the distribution of average scores $\langle S_1 \rangle$ observed over 100 random global alignment graphs and assigned to every conserved network region resulting from that stage

Experimental Results

- Yeast vs. Bacteria: orthologous pathways between the networks of S. cerevisiae and H. pylori
- Yeast vs. Yeast: paralogous pathways within the network of S. cerevisiae

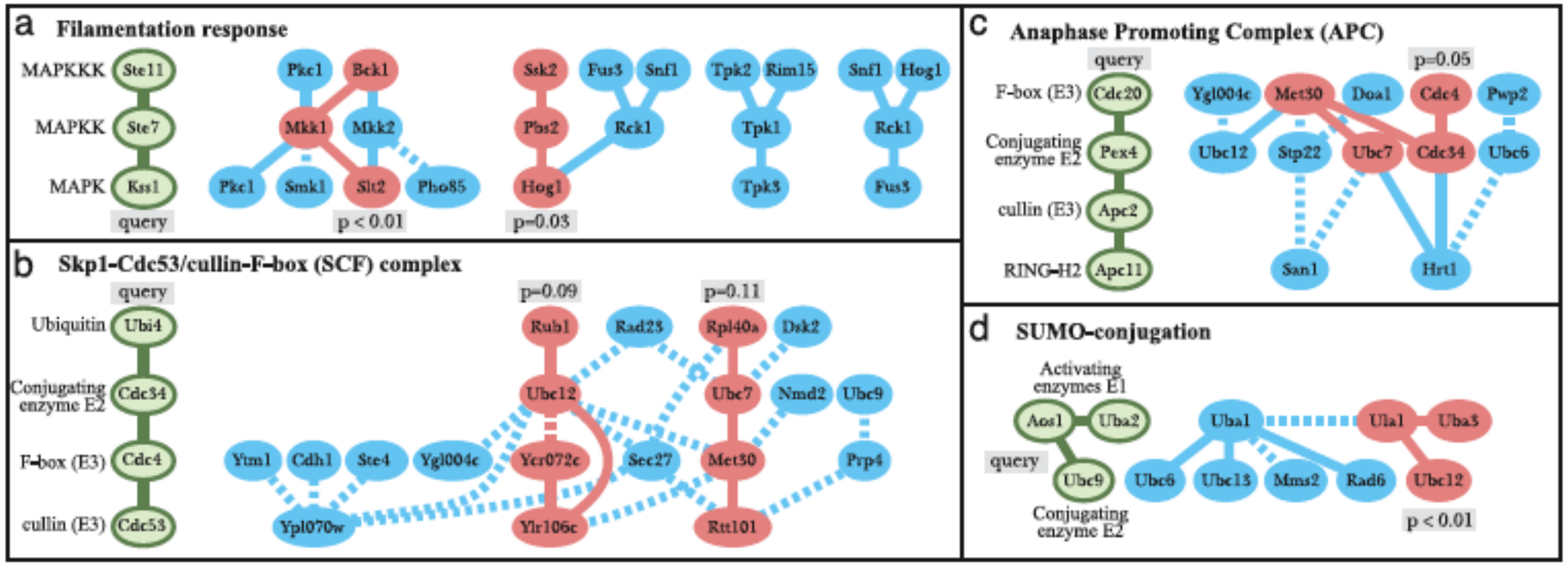


Top-scoring pathway alignments between bacteria and yeast



Paralogous pathways within yeast

(Proteins were not allowed to pair with themselves or their network neighbors)



Querying the yeast network with specific pathways

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 12, Number 6, 2005

Pp. 835–846

Identification of Protein Complexes by Comparative Analysis of Yeast and Bacterial Protein Interaction Data

RODED SHARAN,[1] TREY IDEKER,[2] BRIAN KELLEY,[3] RON SHAMIR,[1] and RICHARD M. KARP[4]

Sharan et al. extended PathBLAST to detect **conserved protein clusters**

The extended method identified eleven complexes that were conserved across the PPI networks of S. cerevisiae and H. pylori

- The method defines a probabilistic model for protein complexes, and search for conserved high probability, high density subgraphs (sub-networks)

A Probabilistic Model for Protein Complexes

- Define two models
  - The protein complex model, M_c: assumes that every two proteins in a complex interact with some high probability β
  - The null model, M_n: assumes that each edge is present with the probability that one would expect if the edges of G were randomly distributed but respected the degrees of the vertices

- A complicating factor in constructing the interaction graph is that we do not know the real protein interactions, but rather have partial, noisy observations of them

- Let T_{uv} denote the event that two proteins u and v interact, and F_{uv} the event that they do not interact

- Denote by O_{uv} the (possibly empty) set of available observations on the proteins u and v, that is, the set of experiments in which u and v were tested for interaction and the outcome of these tests

- Using prior biological information, one can estimate for each protein pair the probability $Pr(O_{uv}|T_{uv})$ of the observations on this pair, given that it interacts, and the probability $Pr(O_{uv}|F_{uv})$ of those observations, given that this pair does not interact

- Further, one can estimate the prior probability $Pr(T_{uv})$ that two random proteins interact

Scoring for Single Species

- Given a subset U of the vertices, we wish to compute the likelihood of U under a protein-complex model and under a null model
- Denote by O_U the collection of all observations on vertex pairs in U. Then

$$Pr(O_U|M_c) = \prod_{(u,v)\in U\times U} Pr(O_{uv}|M_c) \quad (1)$$

[(1) follows from the assumption that all pairwise interactions are independent]

$$= \prod_{(u,v)\in U\times U} (Pr(O_{uv}|T_{uv}, M_c)Pr(T_{uv}|M_c) + Pr(O_{uv}|F_{uv}, M_c)Pr(F_{uv}|M_c)) \quad (2)$$

[(2) is obtained from the law of complete probability]

$$= \prod_{(u,v)\in U\times U} (\beta Pr(O_{uv}|T_{uv}) + (1-\beta)Pr(O_{uv}|F_{uv})) \quad (3)$$

[(3) follows by noting that given the hidden event of whether u and v interact, O_{uv} is independent of any model]

- Next, $Pr(O_U|M_n)$ needs to be computed

- Let $d_1, d_2, ..., d_n$ denote the expected degrees of the vertices in G, rounded to the closest integer

- In order to compute $d_1, ..., d_n$, apply Bayes' rule to derive the expectation of T_{uv} for any pair u,v, given the observations on this vertex pair:

$$Pr(T_{uv}|O_{uv}) = \frac{Pr(O_{uv}|T_{uv})Pr(T_{uv})}{Pr(O_{uv}|T_{uv})Pr(T_{uv}) + Pr(O_{uv}|F_{uv})(1 - Pr(T_{uv}))}$$

- Hence,

$$d_i = [\sum_j Pr(T_{ij}|O_{ij})]$$

where [.] denotes rounding

- The refined null model assumes that G is drawn uniformly at random from the collection of all graphs whose degree sequences is $d_1,...,d_n$

- This induces a probability p_{uv} for every vertex pair (u,v), from which we can calculate the probability of O_U according to the null model

$$Pr(O_U|M_n) = \prod_{(u,v)\in U\times U} (p_{uv}Pr(O_{uv}|T_{uv}) + (1 - p_{uv})Pr(O_{uv}|F_{uv}))$$

- Finally, the log likelihood ratio that we assign to a subset of vertices U is

$$L(U) = \log \frac{Pr(O_U|M_c)}{Pr(O_U|M_n)} \tag{4}$$

$$= \sum_{(u,v)\in U\times U} \log \frac{\beta Pr(O_{uv}|T_{uv}) + (1 - \beta)Pr(O_{uv}|F_{uv})}{p_{uv}Pr(O_{uv}|T_{uv}) + (1 - p_{uv})Pr(O_{uv}|F_{uv})} \tag{5}$$

$$= \sum_{(u,v)\in U\times U} \log \frac{\beta Pr(T_{uv}|O_{uv})(1 - Pr(T_{uv})) + (1 - \beta)(1 - Pr(T_{uv}|O_{uv}))Pr(T_{uv})}{p_{uv}Pr(T_{uv}|O_{uv})(1 - Pr(T_{uv})) + (1 - p_{uv})(1 - Pr(T_{uv}|O_{uv}))Pr(T_{uv})} \tag{6}$$

Scoring for Two Species

- Consider now the case of data on two species 1 and 2, denoted throughout by an appropriate superscript
- Consider two subsets U^1 and V^2 of vertices and some many-to-many mapping $\Theta: U^1 \rightarrow V^2$ between them
- Assuming the interaction graphs of the two species are independent of each other, the log likelihood ratio score for these two sets is simply

$$L(U^1, V^2) = \log \frac{Pr(O_{U^1}|M_c^1)}{Pr(O_{U^1}|M_n^1)} + \log \frac{Pr(O_{V^2}|M_c^2)}{Pr(O_{V^2}|M_n^2)}$$

- However, this score does not take into account the degree of sequence conservation among the pairs of proteins associated by Θ

- In order to include such information, we have to define a conserved complex model and a null model for pairs of proteins from two species

- The conserved complex model assumes that pairs of proteins associated by Θ are orthologous

- The null model assumes that such pairs consist of two independently chosen proteins

- Let E_{uv} denote the BLAST E-value assigned to the similarity between proteins u and v, and let h_{uv}, $\bar{h}_{uv}$ denote the events that u and v are orthologous or non-orthologous, respectively

- The likelihood ratio corresponding to a pair of proteins (u,v) is therefore

$$\frac{Pr(E_{uv}|M_c)}{Pr(E_{uv}|M_n)} = \frac{Pr(E_{uv}|h_{uv})}{Pr(E_{uv}|h_{uv})Pr(h_{uv}) + Pr(E_{uv}|\bar{h}_{uv})Pr(\bar{h}_{uv})} \left(= \frac{Pr(h_{uv}|E_{uv})}{Pr(h)} \right)$$

and the complete score of U^1 and V^2 under the mapping Θ is

$$S_\theta(U^1, V^2) = L(U^1, V^2) + \sum_{i=1}^{k_1} \sum_{v_j^2 \in \theta(u_i^1)} \log \frac{Pr(h_{u_i^1 v_j^2}|E_{u_i^1 v_j^2})}{Pr(h)}$$

(k_1 is the number of vertices in U^1)

prior probability that two proteins are orthologous

Searching for Conserved Complexes

- Using the model just described for comparative interaction data, the problem of identifying conserved protein complexes reduces to the problem of identifying a subset of proteins in each species, and a correspondence between them, such that the score of these subsets exceeds a threshold

The Orthology Graph

- Define a complete edge- and node-weighted orthology graph
- Denote by the superscripts p and y the model parameters corresponding to bacteria and yeast, respectively
- For two proteins y_1 and y_2 define

$$w^{y}_{(y_1,y_2)} = \log \frac{\beta^{y} Pr(O_{y_1y_2}|T_{y_1y_2}) + (1-\beta^{y}) Pr(O_{y_1y_2}|F_{y_1y_2})}{p^{y}_{y_1y_2} Pr(O_{y_1y_2}|T_{y_1y_2}) + (1-p^{y}_{y_1y_2}) Pr(O_{y_1y_2}|F_{y_1y_2})}$$

Similarly, for two bacterial proteins p_1 and p_2 define

$$w^{p}_{(p_1,p_2)} = \log \frac{\beta^{p} Pr(O_{p_1p_2}|T_{p_1p_2}) + (1-\beta^{p}) Pr(O_{p_1p_2}|F_{p_1p_2})}{p^{p}_{p_1p_2} Pr(O_{p_1p_2}|T_{p_1p_2}) + (1-p^{p}_{p_1p_2}) Pr(O_{p_1p_2}|F_{p_1p_2})}$$

- Every pair (y_1,p_1) of yeast and bacterial proteins is assigned a node whose weight reflects the similarity of the proteins:

$$w_{(y_1,p_1)} = \log \frac{Pr(h_{y_1p_1}|E_{y_1p_1})}{Pr(h)}$$

The Orthology Graph

- Every two distinct nodes (y_1,p_1) and (y_2,p_2) are connected by an edge, which is associated with a pair of weights $(w^{y}_{(y_1,y_2)}, w^{p}_{(p_1,p_2)})$
- If $y_1=y_2$ ($p_1=p_2$), set the first (second) weight to 0
- By construction, an induced subgraph of the orthology graph corresponds to two subsets of proteins, one from each species, and many-to-many correspondence between them

The Orthology Graph

- Define the z-score of an induced subgraph with vertex sets U^1 and V^2 and a mapping Θ between them as the log likelihood ratio score $S_\Theta(U^1,V^2)$ for the subgraph, normalized by subtracting its mean and dividing by its standard deviation

- The node and edge weights are assumed to be independent, so the mean and variance of $S_\Theta(U^1,V^2)$ are obtained by summing the sample means and variances of the corresponding nodes and edges

- In order to reduce the complexity of the graph and focus on biologically plausible conserved complexes, certain nodes were filtered from the graph

The Search Strategy

- The problem of searching heavy subgraphs in a graph is NP-hard
- A bottom-up heuristic search is instead performed (in the alignment graph), by starting from high-weight seeds, refining them by exhaustive enumeration, and then expanding them using local search
- An edge in the alignment graph is strong if the sum of its associated weights (the weights within each species graph) is positive

The Search Strategy

1. Compute a seed around each node v, which consists of v and all its neighbors u such that (u,v) is a strong edge

2. If the size of the seed is above a specified threshold, iteratively remove from it the node whose contribution to the subgraph score is minimum, until a desired size is reached

3. Enumerate all subsets of the seed that have size at least 3 and contain v. Each such subset is a refined seed on which a local search heuristic is applied

4. Local search: iteratively add a node whose contribution to the current seed is maximum, or remove a node, whose contribution to the current seed is minimum, as long as this operation increases the overall score of the seed. Throughout the process, the original refined seed is preserved and nodes are not deleted from it

5. For each node in the alignment graph, record up to k heaviest subgraphs that were discovered around that node

The Search Strategy

- The resulting subgraphs may overlap considerably, so the authors used a greedy algorithm to filter subgraphs whose percentage of intersection is above a threshold (60%)
- The algorithm iteratively finds the highest weight subgraph, adds it to the final output list, and removes all other highly intersecting subgraphs

Evaluating the Complexes

- Compute two kinds of p-values
- The first is based on the z-scores that are computed for each subgraph and assumes a normal approximation to the likelihood ratio of a subgraph. The approximation relies on the assumption that the subgraph's nodes ad edges contribute independent terms to the score. The latter probability is Bonferroni corrected for multiple testing.
- The second is based on empirical runs on randomized data. The randomized data are produced by random shuffling of the input interaction graphs of the two species, preserving their degree sequences, as well as random shuffling of the orthology relations, preserving the number of orthologs associated with each protein. For each randomized dataset, the authors used their heuristic search to find the highest-scoring conserved complex of a given size. Then, they estimated the p-value of a suggested complex of the same size, as the fraction of random runs in which the output complex had larger score.

Experimental Setup

- Yeast vs. Bacteria: orthologous complexes between the networks of S. cerevisiae and H. pylori
- The yeast network contained 14,848 pairwise interactions among 4,716 proteins
- The bacterial network contained 1,403 pairwise interactions among 732 proteins
- All interactions were extracted from the DIP database

Experimental Setup

- Protein sequences for both species were obtained from PIR
- Alignments and associated E-values were computed using BLAST 2.0, with parameters b=0; e=1E6; f="C;S"; v=6E5
- Unalignable proteins were assigned a maximum E-value of 5
- Altogether, 1,909 protein pairs had E-value below 0.01, out of which 822 pairs contained proteins with some measured interaction
- Adding 1,242 additional pairs with weak homology and removing nodes with no incident strong edges resulted in a final orthology graph G with 866 nodes and 12,420 edges
- In total, 248 distinct bacterial proteins and 527 yeast proteins participated in G

Experimental Setup

- The authors used a maximum likelihood method to estimate the reliability of observed interactions in yeast
- The reliability of the interactions in H. pylori was estimated at 0.53
- For each species, the probabilities of observing each particular edge in a random graph with the same degree sequence was computed by Monte Carlo simulations
- The authors set β (the probability of observing an interaction in a complex model) to 0.95
- The prior probability of a true interaction was set to 0.001
- The prior probability that a pair of proteins are orthologous was computed as the frequency of protein pairs from both species that are in the same COG cluster, with a value of Pr(h)=0.001611

Experimental Results

- The algorithm identified 11 nonredundant complexes, whose p-values were smaller than 0.05, after correction for multiple testing
- These complexes were also found to be significant when scored against empirical runs on randomized data ($p < 0.05$)

Experimental Results

TABLE 1. CONSERVED PROTEIN COMPLEXES IDENTIFIED BETWEEN YEAST AND BACTERIA[a]

			Yeast enrichment		*Bacterial enrichment*	
ID	*Score*	*Size*	*Purity*	*Complex category*	*Purity*	*Functional category*
1	16.16	12 (12,10)	0.17 (1/6)	Translation (1)	0.56 (5/9)	DNA-metabolism (19)
8	3.31	6 (6,6)	1.00 (4/4)	Respiration (4)	0.33 (2/6)	Energy-metabolism (19)
17	141.31	12 (6,12)	0.90 (9/10)	Proteasome (9)	0.50 (2/4)	Protein-fate (11)
18	37.31	13 (9,13)	0.45 (5/11)	Proteasome (9)	0.25 (2/8)	DNA-metabolism (19)
19	19.09	6 (6,6)	1.00 (6/6)	Translation (10)	0.80 (4/5)	Protein-synthesis (31)
25	40.16	10 (8,10)	0.67 (4/6)	Replication (4)	0.20 (1/5)	DNA-metabolism (19)
28	9.39	9 (9,9)	0.60 (3/5)	Translation (10)	0.50 (4/8)	Protein-synthesis (31)
30	383.52	20 (12,20)	0.55 (6/11)	NUP (6)	0.43 (3/7)	Cell-envelope (7)
31	7.21	6 (6,6)	0	—	1.00 (4/4)	Protein-synthesis (31)
32	3.05	7 (6,7)	0.67 (2/3)	Transcription (3)	0.25 (1/4)	Transcription (4)
33	15.68	13 (12,12)	0.40 (2/5)	RNA-processing (2)	0.33 (3/9)	DNA-metabolism (19)

[a]For each complex the table lists its score ($-\ln p$-value, adjusted for multiple testing); its size (with the numbers of distinct bacterial and yeast proteins in parentheses); purity, as measured using MIPS level 3 categorization of complexes in yeast (with the number of proteins from the most abundant category, and the total number of categorized proteins in the complex, in parentheses); the most abundant category (and its size in parentheses); functional purity, as measured using functional annotation in bacteria; and the most abundant class (with its size among the graph's annotated proteins in parentheses). A zero enrichment for a complex indicates that there is at most one annotated member of the complex. Abbreviations: NUP (nuclear pore complex).

Experimental Results

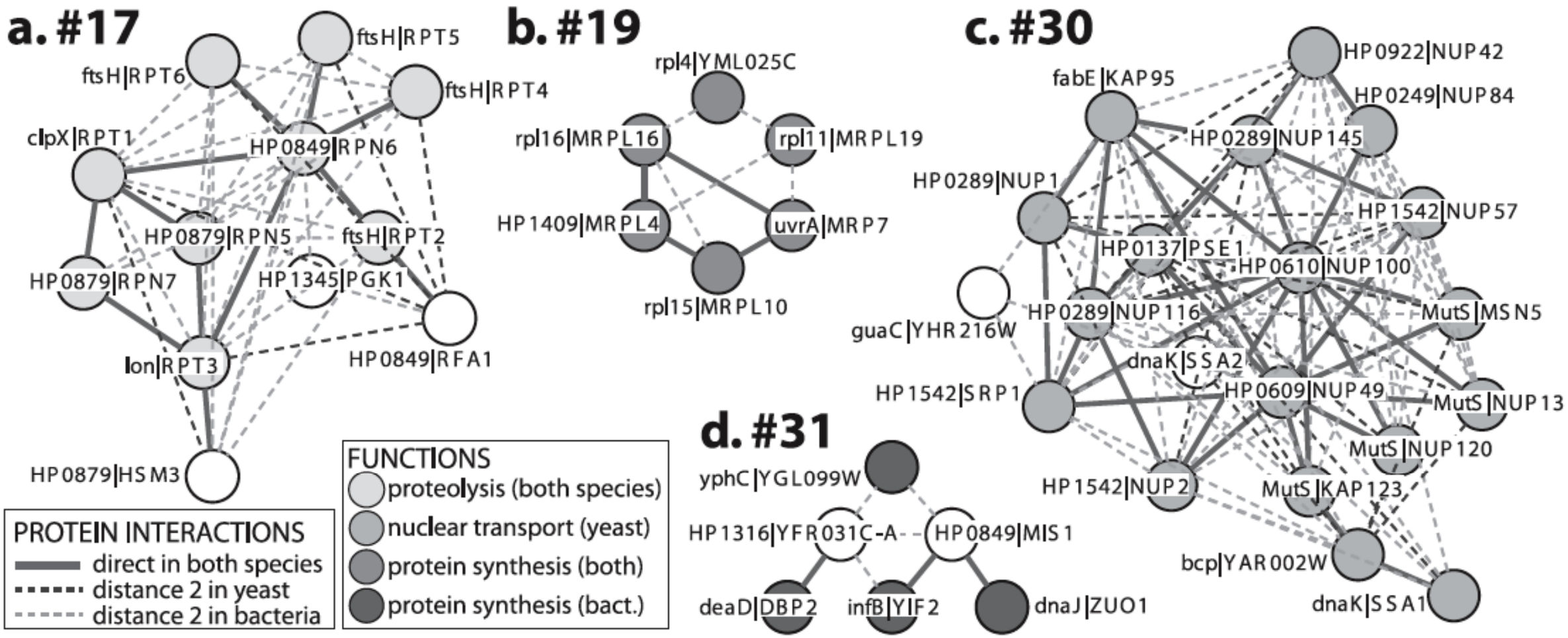


FIG. 2. Conserved protein complexes for proteolysis (**panel a**), protein synthesis (**panels b and d**), and nuclear transport (**panel c**). Conserved complexes are connected subgraphs within the bacteria-yeast orthology graph, whose nodes represent orthologous protein pairs and edges represent conserved protein interactions of three types: direct interactions in both species (solid edges); direct in bacteria but distance 2 in the yeast interaction graph (dark dashed edges); and distance 2 in the bacterial interaction graph but direct in yeast (light dashed edges). In the algorithm, both nodes and edges are assigned weights according to the probabilistic model. The number of each complex indicates the corresponding complex ID listed in Table 1.

MaWISh

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 13, Number 2, 2006

Pp. 182–199

Pairwise Alignment of Protein Interaction Networks

MEHMET KOYUTÜRK,[1] YOHAN KIM,[2] UMUT TOPKARA,[1]
SHANKAR SUBRAMANIAM,[2,3] WOJCIECH SZPANKOWSKI,[1] and ANANTH GRAMA[1]

Koyuturk et al. developed an evolution-based scoring scheme to detect conserved protein clusters, which takes into account interaction insertion/deletion and protein duplication events

The algorithm, MaWISh, identified conserved sub-networks in the PPI networks of human and mouse, as well as conserved sub-networks across S. cerevisiae, C. elegans, and D. melanogaster

http://www.cs.purdue.edu/homes/koyuturk/mawish/

- The authors propose a framework for aligning PPI networks based on the duplication/divergence evolutionary model that has been shown to be promising in explaining the power-law nature of PPI networks

- Like the work of Sharan and colleagues (PathBLAST), the authors here construct an alignment (or, product) graphs by matching pairs of orthologous nodes (proteins)
- Unlike Sharan and colleagues, the authors define matches, mismatches, and duplications, and weight edges in order to reward or penalize these evolutionary events

- The authors reduce the resulting alignment problem to a graph-theoretic optimization problem and propose efficient heuristics to solve it

Outline of the Rest of This Part

- Theoretical models for evolution of PPI networks
- Pairwise local alignment of PPI networks
- Experimental results

Theoretical Models for Evolution of PPI Networks

- Barabasi and Albert (1999) proposed a network growth model based on preferential attachment, which is able to generate networks with degree distribution similar to PPI networks
- According to the BA model, networks expand continuously by addition of new nodes, and these new nodes prefer to attach to well-connected nodes when joining the network

Theoretical Models for Evolution of PPI Networks

- A common model of evolution that explains preferential attachment is the duplication/ divergence model, which is based on gene duplications
- According to this model, when a gene is duplicated in the genome, the node corresponding to the product of this gene is also duplicated together with its interactions

Theoretical Models for Evolution of PPI Networks

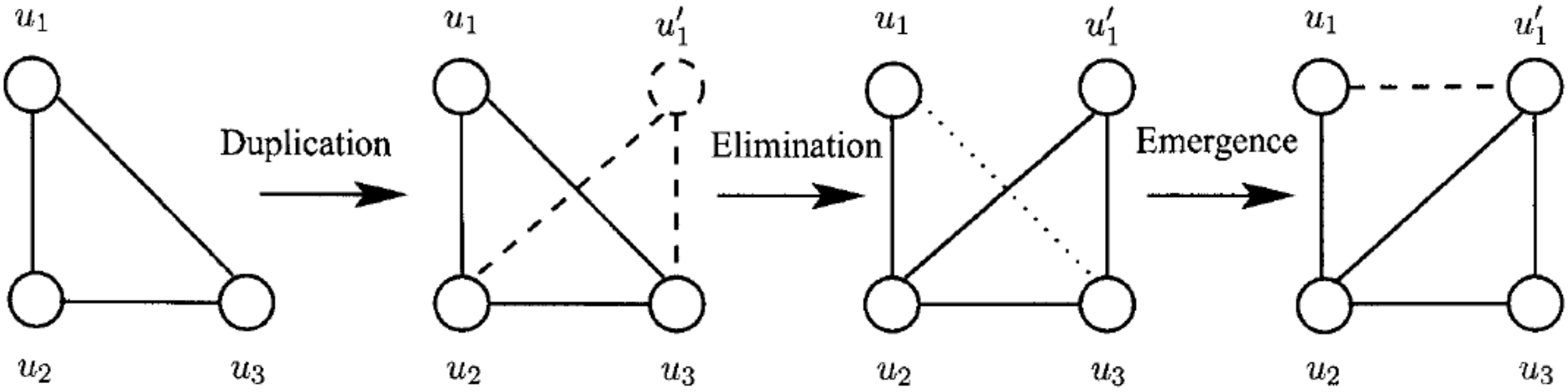


FIG. 1. Duplication/divergence model for evolution of PPI networks. Starting with three interactions between three proteins, protein u_1 is duplicated to add u_1' into the network together with its interactions (dashed circle and lines). Then, u_1 loses its interaction with u_3 (dotted line). Finally, an interaction between u_1 and u_1' is added to the network (dashed line).

Theoretical Models for Evolution of PPI Networks

- A protein loses many aspects of its functions rapidly after being duplicated
- This translates to divergence of duplicated (paralogous) proteins in the interactome through elimination and emergence of interactions

Theoretical Models for Evolution of PPI Networks

- Elimination of an interaction in a PPI network implies the loss of an interaction between two proteins due to structural and/or functional changes
- Similarly, emergence of an interaction in a PPI network implies the introduction of a new interaction between two noninteracting proteins caused by mutations that change protein surfaces

Theoretical Models for Evolution of PPI Networks

- Since the elimination of interactions is related to sequence-level mutations, one can expect a positive correlation between similarity of interaction profiles and sequence similarity for paralogous proteins
- The interaction profiles of duplicated proteins tend to almost totally diverge in about 200 million years, as estimated on the yeast interactome

Theoretical Models for Evolution of PPI Networks

- On the other hand, the correlation between interaction profiles of duplicated proteins is significant for up to 150 million years after duplication, with more than half of the interactions being conserved for proteins that are duplicated less than 50 million yeas back

Theoretical Models for Evolution of PPI Networks

- Consequently, when PPI networks that belong to two separate species are considered, the in-paralogs are likely to have more common interactions than out-paralogs

Pairwise Local Alignment of PPI Networks

- Three items:
    - Define the PPI network alignment problem
    - Formulate the problem as a graph optimization problem
    - Describe an efficient heuristic for solving the problem

The PPI Network Alignment Problem

- Undirected graph G(U,E)
- The homology between a pair of proteins is quantified by a similarity measure S, where S(u,v) measures the degree of confidence in u and v being orthologous, where $0 \leq S(u,v) \leq 1$
- If u and v belong to the same species, then S(u,v) quantifies the likelihood that the two proteins are in-paralogs
- S is expected to be sparse (very few orthologs for each protein)

The PPI Network Alignment Problem

- For PPI networks G(U,E) and H(V,F), a protein subset pair $P = \{\tilde{U}, \tilde{V}\}$ is defined as a pair of protein subsets $\tilde{U} \subseteq U$ and $\tilde{V} \subseteq V$
- Any protein subset pair P induces a local alignment A(G,H,S,P)={M,N,D} of G and H with respect to S, characterized by a set of duplications D, a set of matches M, and a set of mismatches N
- Each duplication is associated with a score that reflects the divergence of function between the two proteins, estimated using their similarity
- A match corresponds to a conserved interaction between two orthologous protein pairs (an interlog), which is rewarded by a match score that reflects confidence in both protein pairs being orthologous

The PPI Network Alignment Problem

- A mismatch is the lack of an interaction in the PPI network of one organism between a pair of proteins whose orthologs interact in the other organism
- Mismatches are penalized to account for the divergence from the common ancestor

The PPI Network Alignment Problem

Definition 1. Local Alignment of PPI Networks. *Given protein interaction networks* $G(U, E)$, $H(V, F)$, *let functions* $\Delta_G(u, u')$ *and* $\Delta_H(v, v')$ *denote the distance between two corresponding proteins in the interaction graphs* G *and* H, *respectively. Given a pairwise similarity function* S *defined over the union of their protein sets* $U \cup V$, *and a distance cutoff* $\bar{\Delta}$, *any protein subset pair* $P = (\tilde{U}, \tilde{V})$ *induces a local alignment* $\mathcal{A}(G, V, S, P) = \{\mathcal{M}, \mathcal{N}, \mathcal{D}\}$, *where*

$$\mathcal{M} = \{u, u' \in \tilde{U}, v, v' \in \tilde{V} : S(u, v) > 0,\ S(u', v') > 0, \\ ((uu' \in E \wedge \Delta_H(v, v') \leq \bar{\Delta}) \vee (vv' \in F \wedge \Delta_G(u, u') \leq \bar{\Delta}))\}, \tag{1}$$

$$\mathcal{N} = \{u, u' \in \tilde{U}, v, v' \in \tilde{V} : S(u, v) > 0,\ S(u', v') > 0, \\ ((uu' \in E \wedge \Delta_H(v, v') > \bar{\Delta}) \vee (vv' \in F \wedge \Delta_G(u, u') > \bar{\Delta}))\}, \tag{2}$$

$$\mathcal{D} = \{u, u' \in \tilde{U} : S(u, u') > 0\} \cup \{v, v' \in \tilde{V} : S(v, v') > 0\}. \tag{3}$$

Each match $M \in \mathcal{M}$, *mismatch* $N \in \mathcal{N}$, *and duplication* $D \in \mathcal{D}$ *are associated with scores* $\mu(M)$, $\nu(N)$, *and* $\delta(D)$, *respectively.*

Scoring Match, Mismatch, and Duplications

- For scoring matches and mismatches, define the similarity between two protein pairs as

$$S(uu', vv') = S(u, v)S(u', v')$$

where $S(uu', vv')$ quantifies the likelihood that the interactions between u and v, and u' and v' are orthologous

- Consequently, a match that corresponds to a conserved pair of orthologous interactions is rewarded as follows:

$$\mu(uu', vv') = \bar{\mu}S(uu', vv')$$

- Here, $\bar{\mu}$ is the match coefficient that is used to tune the relative weight of matches against mismatches and duplications, based on the evolutionary distance between the species that are being compared

Scoring Match, Mismatch, and Duplications

- A mismatch may correspond to the functional divergence of either interacting partner after speciation
- It might also be due to a false positive or negative in one of the networks that is caused by incompleteness of the data or experimental error
- It has been observed that after a duplication event, duplicate proteins that retain similar functions in terms of being part of similar processes are likely to be part of the same subnet
- Moreover, since conservation of proteins in a particular module is correlated with interconnectedness, it is expected that interacting partners that are part of a common functional module will at least be linked by short alternative paths

Scoring Match, Mismatch, and Duplications

- Based on the aforementioned observations, mismatches are penalized for possible divergence in functions as follows:

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

- As for match score, mismatch penalty is also normalized by a coefficient $\bar{\nu}$ that determines the relative weight of mismatches w.r.t. matches and duplications
- With the expectation that recently duplicated proteins, which are more likely to be in-paralogs, show more significant sequence similarity than older paralogs, duplication score is defined as follows:

$$\delta(u, u') = \bar{\delta}(S(u, u') - \bar{d})$$

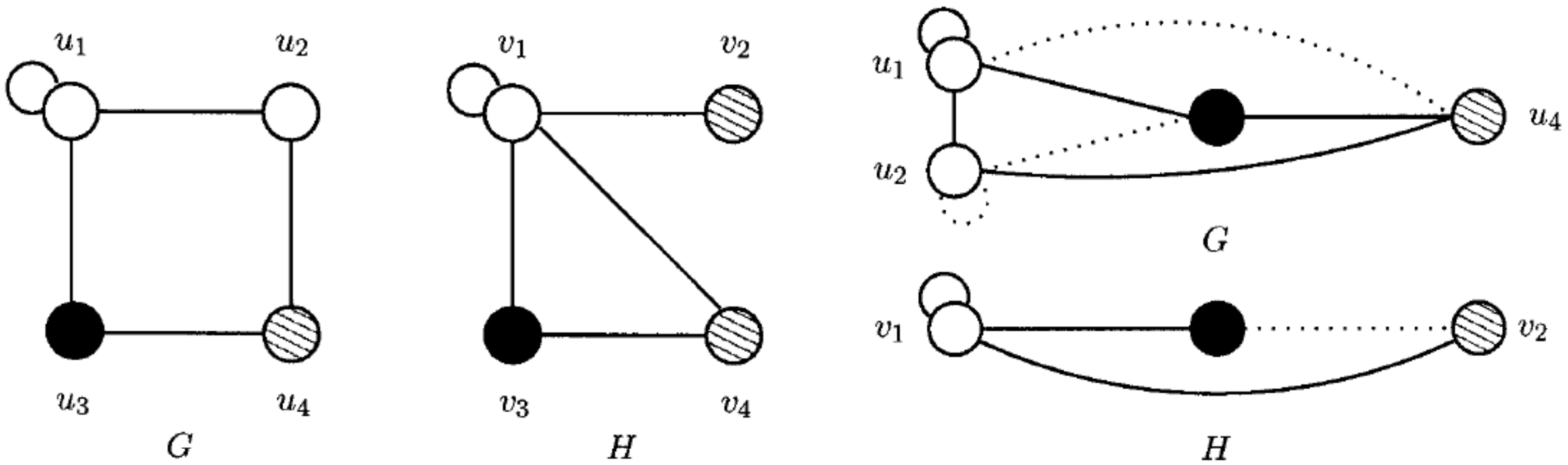
- Here $\bar{d}$ is the cutoff for being considered in-paralogs

Scoring Match, Mismatch, and Duplications

Definition 2. Alignment Score and PPI Network Alignment Problem. *Given PPI networks* G *and* H, *the* score of alignment $\mathcal{A}(G, H, S, P) = \{\mathcal{M}, \mathcal{N}, \mathcal{D}\}$ *is defined as*

$$\sigma(\mathcal{A}) = \sum_{M \in \mathcal{M}} \mu(M) + \sum_{N \in \mathcal{N}} \nu(N) + \sum_{D \in \mathcal{D}} \delta(D). \tag{8}$$

The PPI network alignment problem *is one of finding all maximal protein subset pairs* P *such that* $\sigma(\mathcal{A}(G, H, S, P))$ *is locally maximal, i.e., the alignment score cannot be improved by adding individual proteins to or removing proteins from* P.



(a) (b)

FIG. 2. (**a**) An instance of the pairwise local alignment problem. The proteins that have nonzero similarity scores (i.e., are potentially orthologous), are shaded the same. Note that S does not necessarily induce a disjoint grouping of proteins in practice. (**b**) A local alignment induced by the protein subset pair $\{u_1, u_2, u_3, u_4\}$ and $\{v_1, v_2, v_3\}$. Ortholog and paralog proteins are vertically aligned. Existing interactions are shown by solid lines; missing interactions that have an existing ortholog counterpart are shown by dotted lines. Solid interactions between two aligned proteins in separate species correspond to a match; one solid, one dotted interaction between two aligned proteins in separate species correspond to a mismatch. Proteins in the same species that are on the same vertical line correspond to duplications.

$$\bar{\Delta} \doteq 1$$

Estimating Similarity Scores

- The similarity score S(u,v) quantifies the likelihood that proteins u and v are orthologous
- This likelihood is approximated using the BLAST E-value taking existing ortholog databases as point of reference (similar to the work of Sharan and colleagues)
- Let O be the set of all orthologous protein pairs derived from an orthology database (e.g., COG)
- For proteins u and v with BLAST E-value $E(u, v) < \tilde{E}$, S is estimated as

$$S(u, v) = P(E(u, v) < \tilde{E} | O_{uv}) = \frac{|\{u'v' \in \mathbf{O} : E(u', v') < \tilde{E}\}|}{|\mathbf{O}|}$$

where O_{uv} represents that u and v are orthologous

Alignment Graph and the Maximum-weight Induced Subgraph Problem

Definition 3. Alignment Graph. *For a pair of PPI networks $G(U, E)$, $H(V, F)$, and protein similarity function S, the corresponding weighted* alignment graph $\mathbf{G}(\mathbf{V}, \mathbf{E})$ *is computed as follows:*

$$\mathbf{V} = \{\mathbf{v} = \{u, v\} : u \in U, v \in V \text{ and } S(u, v) > 0\}. \tag{10}$$

In other words, we have a node in the alignment graph for each pair of ortholog proteins. Each edge $\mathbf{vv}' \in \mathbf{E}$, where $\mathbf{v} = \{u, v\}$ and $\mathbf{v}' = \{u', v'\}$, is assigned weight

$$w(\mathbf{vv}') = \mu(uu', vv') + \nu(uu', vv') + \delta(u, u') + \delta(v, v'). \tag{11}$$

Here, $\mu(uu', vv') = 0$ if $(uu', vv') \notin \mathcal{M}$, and similarly for mismatches and duplications.

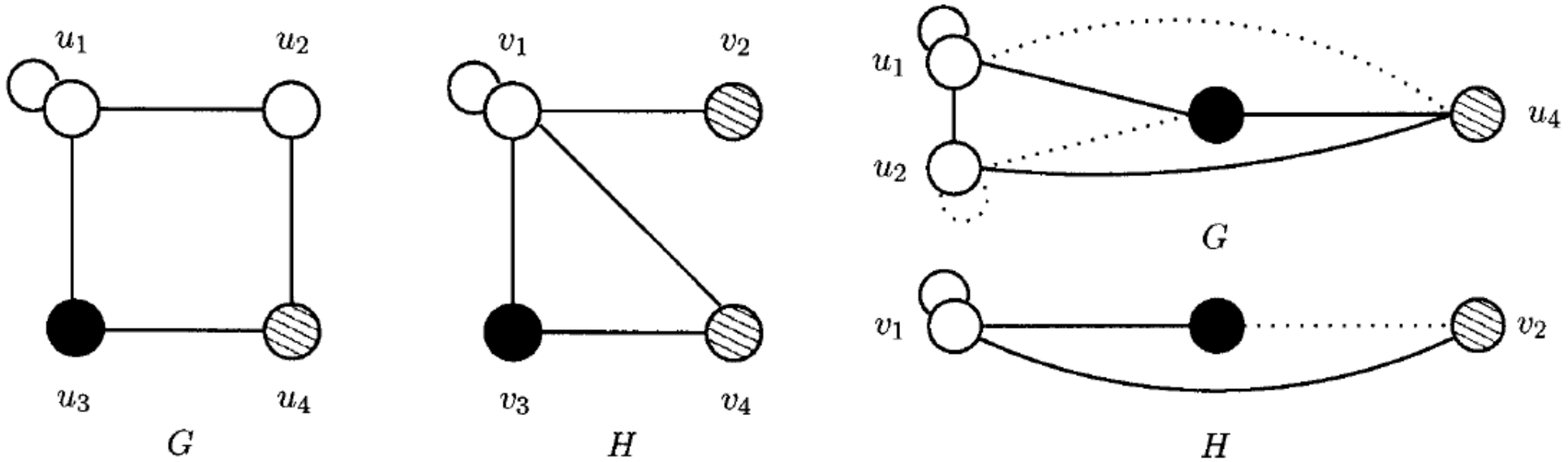


FIG. 2.

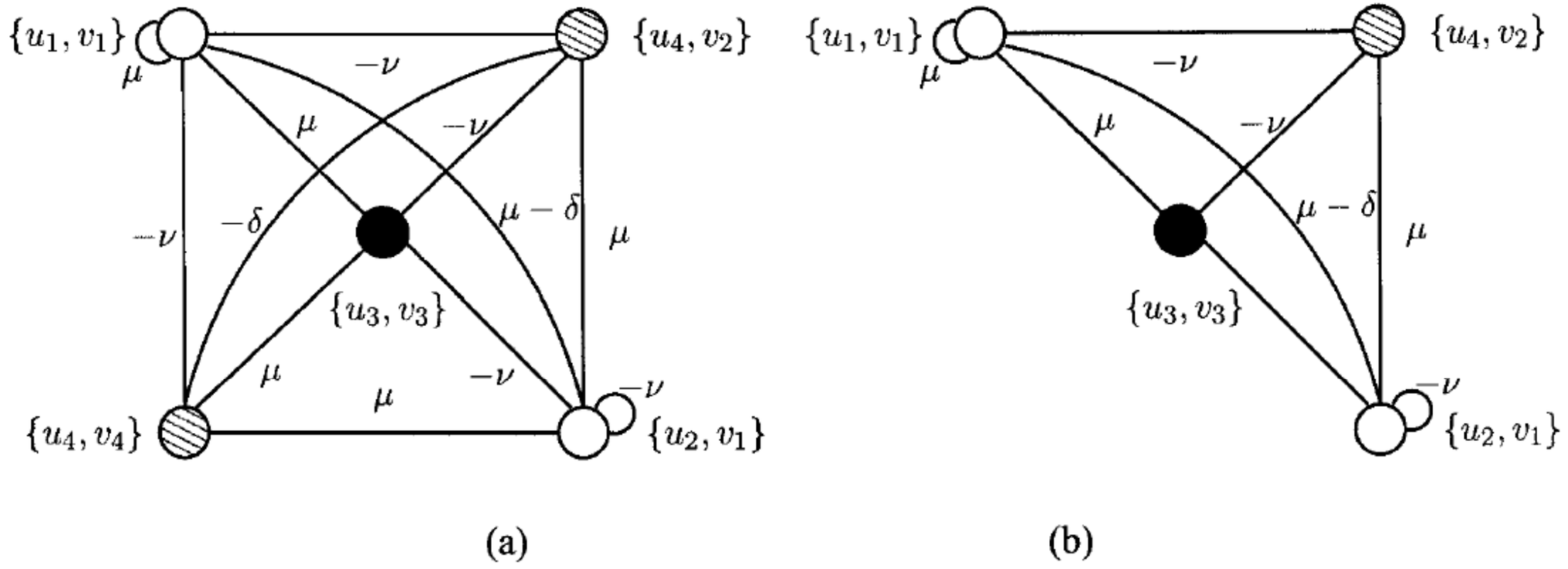


FIG. 3. (**a**) Alignment graph corresponding to the instance of Fig. 2a. Note that match scores, mismatch penalties and duplication scores are functions of incident nodes, which is not explicitly shown in the figure for simplicity. (**b**) Subgraph induced by node set $\tilde{\mathbf{V}} = \{\{u_1, v_1\}, \{u_2, v_1\}, \{u_3, v_3\}, \{u_4, v_2\}\}$, which corresponds to the alignment shown in Fig. 2b.

Alignment Graph and the Maximum-weight Induced Subgraph Problem

- The construction of the alignment graph allows to formulate the alignment problem as a graph optimization problem:

Definition 4. Maximum Weight Induced Subgraph Problem (MAWISH). *Given graph* $\mathbf{G}(\mathbf{V}, \mathbf{E})$ *and a constant* ϵ*, find a subset of nodes,* $\tilde{\mathbf{V}} \in \mathbf{V}$ *such that the sum of the weights of the edges in the subgraph induced by* $\tilde{\mathbf{V}}$ *is at least* ϵ*, i.e.,* $W(\tilde{\mathbf{V}}) = \sum_{\mathbf{v},\mathbf{v}' \in \tilde{\mathbf{V}}} w(\mathbf{v}\mathbf{v}') \geq \epsilon$.

- This problem is equivalent to the decision version of the local alignment problem defined on previous slides. More formally,

Theorem 1. *Given PPI networks* G*,* H*, and a protein similarity function* S*, let* $\mathbf{G}(\mathbf{V}, \mathbf{E}, w)$ *be the corresponding alignment graph. If* $\tilde{\mathbf{V}}$ *is a solution to the maximum weight induced subgraph problem on* $\mathbf{G}(\mathbf{V}, \mathbf{E}, w)$*, then* $P = \{\tilde{U}, \tilde{V}\}$ *induces an alignment* $\mathcal{A}(G, H, S, P)$ *with* $\sigma(\mathcal{A}) = W(\tilde{\mathbf{V}})$*, where* $\tilde{U} = \{u \in U : \exists v \in V \text{ s.t. } \{u, v\} \in \tilde{\mathbf{V}}\}$ *and* $\tilde{V} = \{v \in V : \exists u \in U \text{ s.t. } \{u, v\} \in \tilde{\mathbf{V}}\}$.

Algorithms for Local Alignment of PPI networks

- As in the work of Sharan and colleagues, the authors propose a heuristic that greedily grows a subgraph seeded at heavy nodes

procedure HEAVIESTSUBGRAPH(G)

▷ **Input** $\mathbf{G}(\mathbf{V}, \mathbf{E}, w)$: Alignment graph

▷ **Output** $\tilde{\mathbf{V}}$: Subset of nodes that induces a maximally heavy subgraph in $\mathbf{G}$

1 $\tilde{\mathbf{v}} \leftarrow \text{argmax}_{\mathbf{v} \in \mathbf{V}} |\{\mathbf{v}' \in \mathbf{V} : (\mathbf{v}, \mathbf{v}') \text{ is a match edge}\}|$

2 $\tilde{\mathbf{V}} \leftarrow \{\tilde{\mathbf{v}}\} \cup \{\mathbf{v} \in \mathbf{V} : (\tilde{\mathbf{v}}, \mathbf{v}) \text{ is a match edge}\}$

3 **repeat**

4 $Q \leftarrow \{\mathbf{v} \in \mathbf{V} : key(\mathbf{v}) = -\sum_{\mathbf{v}' \in \tilde{\mathbf{V}}} w(\mathbf{v}, \mathbf{v}')$ **if** $\mathbf{v} \in \tilde{\mathbf{V}}$, $key(\mathbf{v}) = \sum_{\mathbf{v}' \in \tilde{\mathbf{V}}} w(\mathbf{v}, \mathbf{v}')$ **else**$\}$

5 $W_{max} \leftarrow W(\tilde{\mathbf{V}})$

6 **while** $Q \neq \emptyset$

7 $\mathbf{v} \leftarrow$ EXTRACTMAX(Q)

8 **if** $\mathbf{v} \in \tilde{\mathbf{V}}$ **then** $\tilde{\mathbf{V}} \leftarrow \tilde{\mathbf{V}} \setminus \{\mathbf{v}\}$ **else** $\tilde{\mathbf{V}} \leftarrow \tilde{\mathbf{V}} \cup \{\mathbf{v}\}$

9 **if** $W(\tilde{\mathbf{V}}) > W_{max}$ **then** $W_{max} \leftarrow W(\tilde{\mathbf{V}})$, $bestmove \leftarrow \mathbf{v}$

10 **for all** $\mathbf{v}'$ such that $\mathbf{v}\mathbf{v}' \in \mathbf{E}$ update $key(\mathbf{v}')$

11 **endwhile**

12 roll back all moves after $bestmove$

13 **until** $bestmove = \text{NULL}$

14 **return** $\tilde{\mathbf{V}}$

FIG. 4. Fast heuristic for finding a subset of nodes that induces a subgraph of maximal total weight on the alignment graph.

Statistical Significance

- To evaluate the statistical significance of discovered high-scoring alignments, the authors compare them with a reference model generated by a random source
- In the reference model, it is assumed that the interaction networks of the two organisms are independent of each other
- To accurately capture the power-law nature of PPI networks, it is assumed that the interactions are generated randomly from a distribution characterized by a given degree sequence
- If proteins u and u' are interacting with d_u and $d_{u'}$ proteins, respectively, then the probability of observing an interaction between u and u' can be estimated as

$$p_{uu'} = d_u d_{u'} / \sum_{v \in U} d_v$$

Statistical Significance

- In the reference model, the expected value of the score of an alignment induced by $\tilde{\mathbf{V}} \subseteq \mathbf{V}$ is $\mathbf{E}[W(\tilde{\mathbf{V}})] = \sum_{\mathbf{v},\mathbf{v}' \in \tilde{\mathbf{V}}} \mathbf{E}[w(\mathbf{v}\mathbf{v}')]$, where

$$\mathbf{E}[w(\mathbf{v}\mathbf{v}')] = \bar{\mu} S(uu', vv') p_{uu'} p_{vv'} - \bar{\nu} S(uu', vv')(p_{uu'}(1 - p_{vv'}) + (1 - p_{uu'}) p_{vv'}) + \delta(u, u') + \delta(v, v')$$

  is the expected weight of an edge in the alignment graph

- With the simplifying assumption of independence of interactions, we have $\mathbf{Var}[W(\tilde{\mathbf{V}})] = \sum_{\mathbf{v},\mathbf{v}' \in \tilde{\mathbf{V}}} \mathbf{Var}[w(\mathbf{v}\mathbf{v}')]$, which enables computing the z-score to evaluate the statistical significance of each discovered high-scoring alignment

Experimental Results

- Data from BIND and DIP

Organism	*# Proteins*	*# Interactions*
S. cerevisiae	5157	18192
C. elegans	3345	5988
D. melanogaster	8577	28829

- Aligned every pair

Experimental Results

TABLE 2. ALIGNMENT STATISTICS FOR THE THREE PAIRS OF EUKARYOTIC ORGANISMS[a]

Organism pair	*# Nodes*	*# Matched nodes*		*# Matches*		*# Mismatches*	*# Duplications*	
		$\Delta = 1$	$\Delta = 2$	$\Delta = 1$	$\Delta = 2$	$\Delta = 1$	*Org. 1*	*Org. 2*
SC vs CE	2746	312	1230	412	3007	40262	6107	6886
SC vs DM	15884	1730	8622	2061	42781	1054241	6107	32670
CE vs DM	11805	491	3391	455	6626	205593	6886	32670

[a]For each alignment, the number of nodes in alignment graphs (# of orthologous pairs), number of nodes with at least one matched edge, number of matches, number of mismatches, and number of duplications for both organisms are shown. Number of mismatches for $\bar{\Delta} = 2$ can be derived from other statistics.

Experimental Results

TABLE 3. TEN HIGH-SCORING CONSERVED SUBNETS IDENTIFIED BY THE ALIGNMENT OF *S. CEREVISIAE* AND *D. MELANOGASTER* PPI NETWORKS[a]

Rank	*Score*	*z-Score*	*# Proteins*	*# Matches*	*# Mismatches*	*# Duplications*
1	15.97	6.6	18 (16, 5)	28	6	(4, 0)
	Protein amino acid phosphorylation (69%)/JAK-STAT cascade (40%)					
2	13.93	3.7	13 (8, 7)	25	7	(3, 1)
	Endocytosis (50%)/calcium-mediated signaling (50%)					
5	8.22	13.5	9 (5, 3)	19	11	(1, 0)
	Invasive growth (sensu Saccharomyces) (100%)/oxygen and reactive oxygen species metabolism (33%)					
6	8.05	7.6	8 (5, 3)	12	2	(0, 1)
	Ubiquitin-dependent protein catabolism (100%)/mitosis (67%)					
8	6.83	12.4	6 (4, 4)	12	6	(0, 1)
	Protein amino acid phosphorylation (50%, 50%)					
10	6.75	13.7	10 (7, 3)	24	12	(0, 1)
	Ubiquitin-dependent protein catabolism (100%)					
14	5.69	8.7	11 (11, 2)	10	1	(0, 0)
	Regulation of progression through cell cycle (9%, 50%)					
21	4.36	6.2	9 (5, 4)	18	13	(0, 5)
	Cytokinesis (100%, 50%)					
22	4.22	3.9	7 (6, 6)	9	5	(1, 1)
	Protein folding (67%, 17%)					
30	3.76	39.6	6 (3, 5)	5	1	(0, 6)
	DNA replication initiation (100%, 80%)					

[a]The dominant biological process for each organism, in which the majority of proteins in the conserved subnet participate is shown in the second row.

Experimental Results

TABLE 4. FIVE HIGH-SCORING CONSERVED SUBNETS IDENTIFIED BY THE ALIGNMENT OF *S. Cerevisiae* AND *C. Elegans* PPI NETWORKS

Rank	*Score*	*z-Score*	*# Proteins*	*# Matches*	*# Mismatches*	*# Duplications*
1	36.14	7.8	13 (5, 3)	65	24	(0, 3)
	Ubiquitin-dependent protein catabolism (100%)/reproduction (100%)					
2	8.47	6.5	20 (11, 5)	19	4	(1, 1)
	Protein amino acid phosphorylation (82%, 40%)					
3	6.28	10.1	8 (6, 3)	21	12	(0, 0)
	Ubiquitin-dependent protein catabolism (100%, 100%)					
7	3.23	4.9	7 (7, 6)	7	2	(0, 0)
	Glyoxylate cycle (14%, 17%)					
8	3.23	80.1	4 (3, 3)	4	1	(1, 1)
	Mismatch repair (67%, 67%)					

Experimental Results

TABLE 5. FIVE HIGH-SCORING CONSERVED SUBNETS IDENTIFIED BY THE ALIGNMENT OF *C. Elegans* AND *D. Melanogaster* PPI NETWORKS

Rank	*Score*	*z-Score*	*# Proteins*	*# Matches*	*# Mismatches*	*# Duplications*
1	26.75	19.9	17 (4, 9)	52	4	(1, 4)
	Thermosensory behavior (25%)/regulation of transcription from RNA polymerase II promoter (44%)					
2	4.65	31.6	9 (5, 3)	8	0	(2, 1)
	Translational initiation (60%, 67%)					
4	4.37	10.7	11 (3, 6)	10	1	(1, 4)
	Determination of adult life span (33%, 67%)					
5	4.29	16.4	5 (4, 4)	6	0	(1, 1)
	Regulation of transcription, DNA-dependent (50%, 25%)					
6	4.00	12.2	6 (4, 6)	8	2	(0, 2)
	Signal transduction (50%, 17%)					

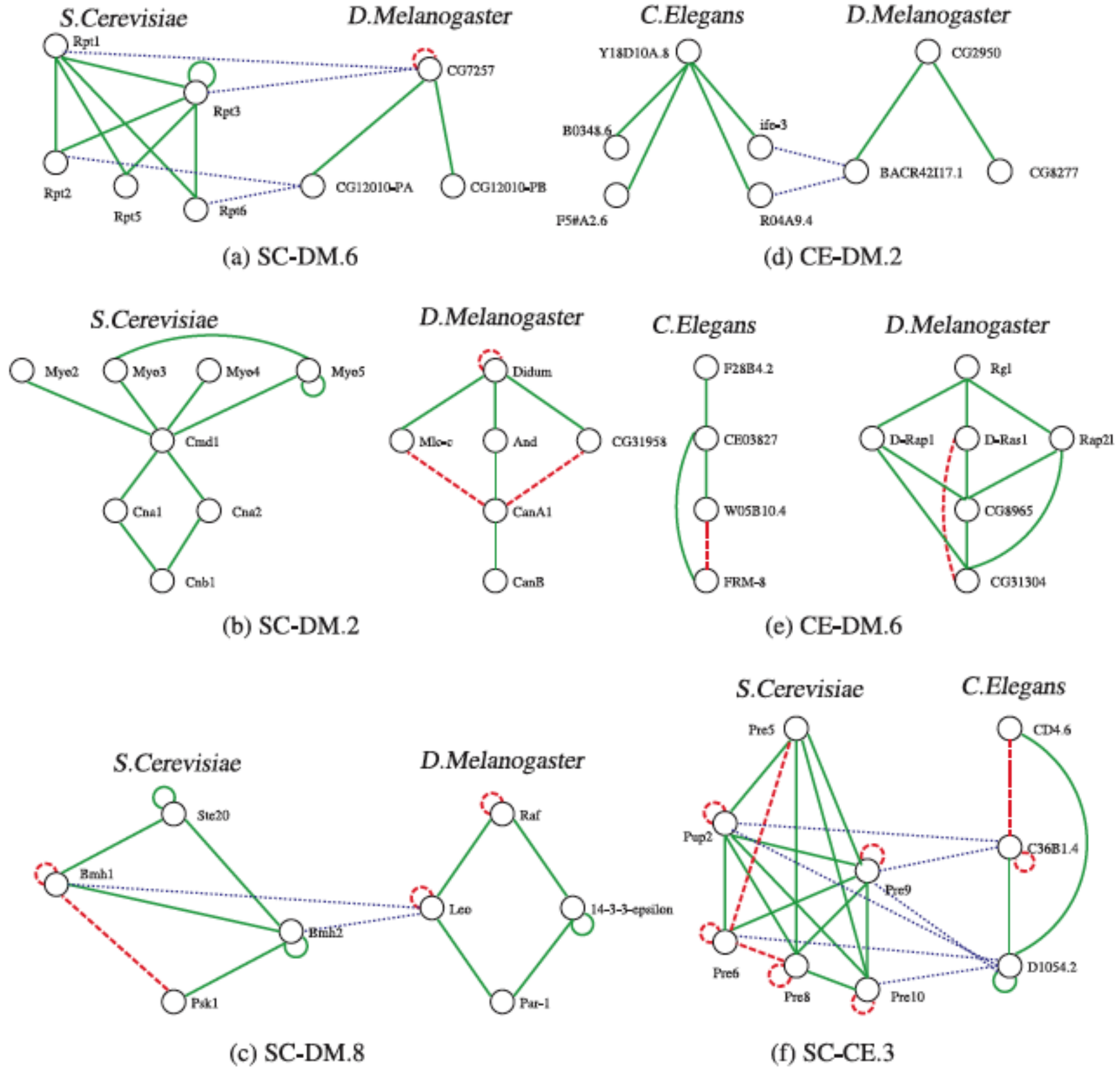


FIG. 5. Sample conserved subnets identified by the alignment algorithm. Orthologous and paralogous proteins are either vertically aligned or connected by blue dotted lines. Existing interactions are shown by green solid lines, missing interactions that have an orthologous counterpart are shown by red dashed lines. The rank of each alignment in the set of alignments discovered for the respective pair of organisms is indicated in its label.

Multiple Network Alignment

Conserved patterns of protein interaction in multiple species

Roded Sharan*†, Silpa Suthram‡, Ryan M. Kelley‡, Tanja Kuhn§, Scott McCuine‡, Peter Uetz§, Taylor Sittler‡, Richard M. Karp*¶, and Trey Ideker‡¶

Græmlin: General and robust alignment of multiple large interaction networks

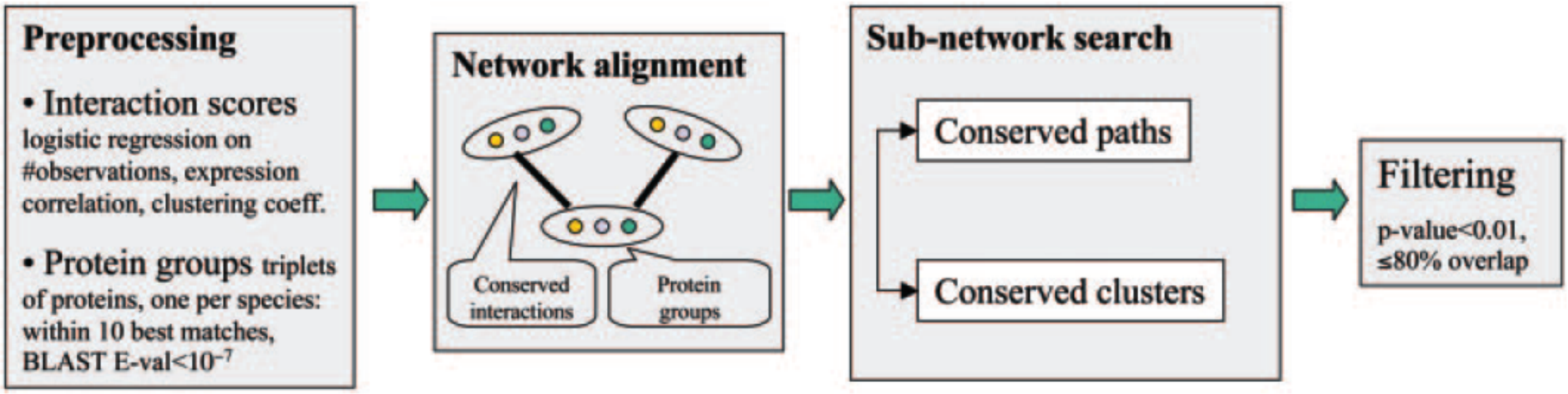
Jason Flannick,[1,4] Antal Novak,[1,4] Balaji S. Srinivasan,[2,3] Harley H. McAdams,[2] and Serafim Batzoglou[1,5]

Conserved patterns of protein interaction in multiple species

Roded Sharan*†, Silpa Suthram‡, Ryan M. Kelley‡, Tanja Kuhn§, Scott McCuine‡, Peter Uetz§, Taylor Sittler‡, Richard M. Karp*¶, and Trey Ideker‡¶

- The authors considered alignments of three PPI networks (C. elegans, D. melanogaster, and S. cerevisiae)

- Their method is almost the same as that for aligning two networks to identify conserved protein complexes, with the only difference that nodes in the alignment graph contain one protein from each of the three species, and an edge between two nodes contains information about interactions among the proteins in the families at both endpoints of the edge



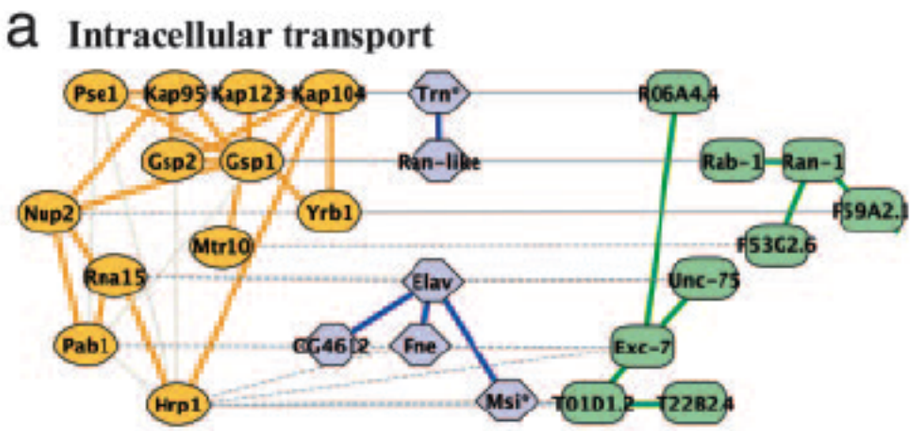
Schematic of the multiple network comparison pipeline. Raw data are preprocessed to estimate the reliability of the available protein interactions and identify groups of sequence-similar proteins. A protein group contains one protein from each species and requires that each protein has a significant sequence match to at least one other protein in the group (BLAST *E* value $< 10^{-7}$; considering the 10 best matches only). Next, protein networks are combined to produce a network alignment that connects protein similarity groups whenever the two proteins within each species directly interact or are connected by a common network neighbor. Conserved paths and clusters identified within the network alignment are compared to those computed from randomized data, and those at a significance level of $P < 0.01$ are retained. A final filtering step removes paths and clusters with >80% overlap.

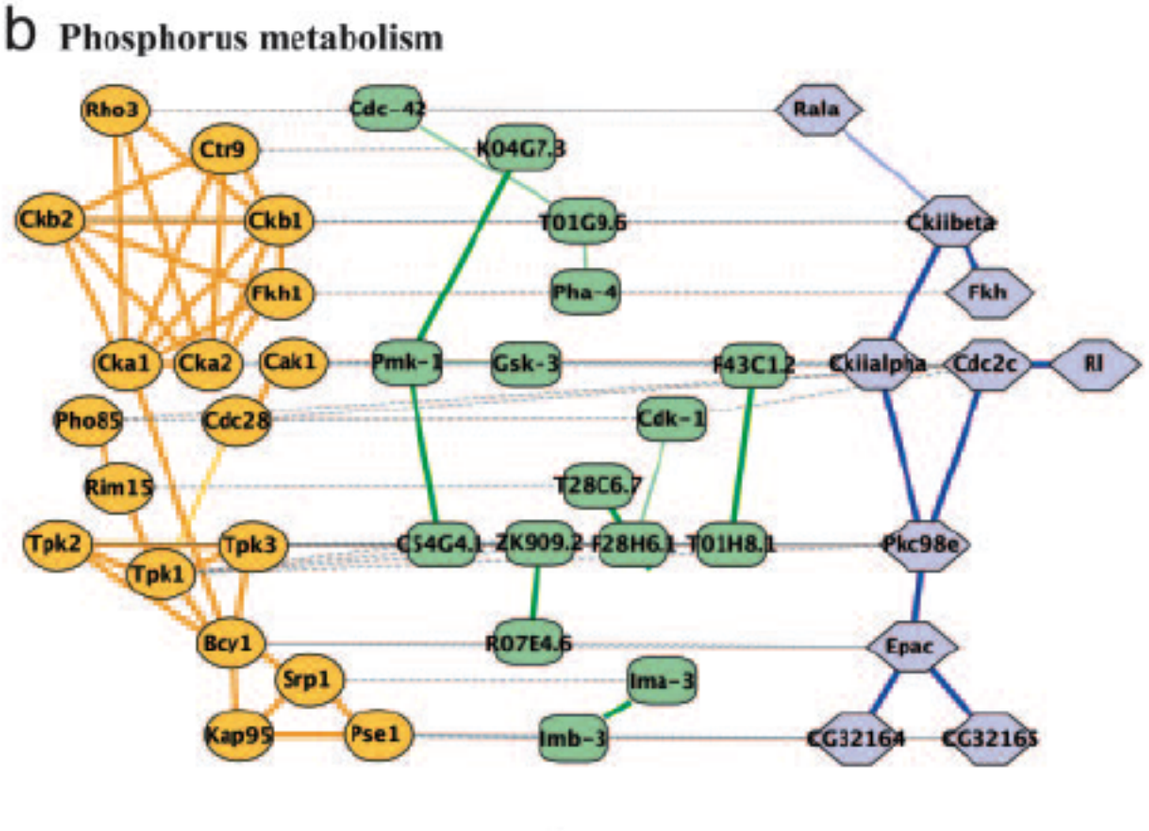
Experimental Setup

- Data was downloaded from DIP
- Yeast: 14,319 interactions among 4,389 proteins
- Worm: 3,926 interactions among 2,718 proteins
- Fly: 20,720 interactions among 7,038 proteins
- Protein sequences obtained from the Saccharomyces Genome Database, WormBase, and FlyBase were combined with the protein interaction data to generate a network alignment of 9,011 protein similarity groups and 49,688 conserved interactions for the three networks

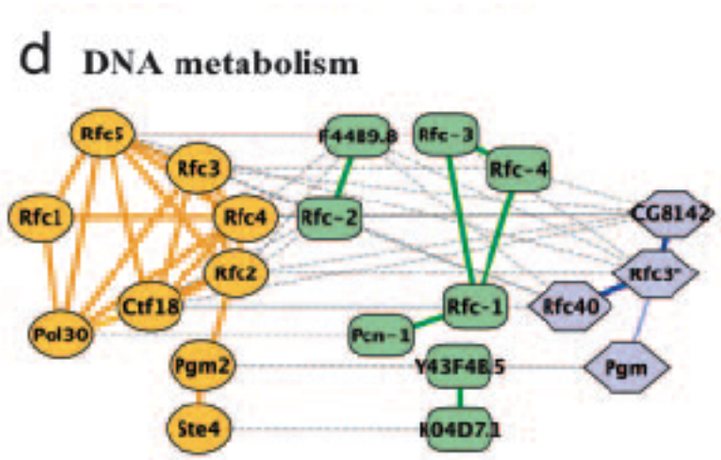
Experimental Results

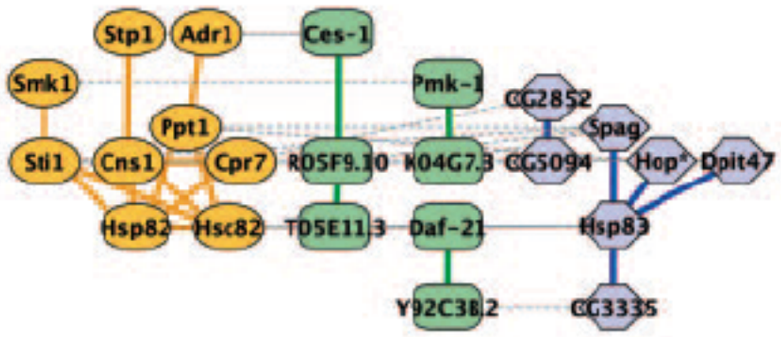
- A search over the network alignment identified 183 protein clusters and 240 paths conserved at a significance level of P<0.01
- These covered a total of 649 proteins among yeast, worm, and fly

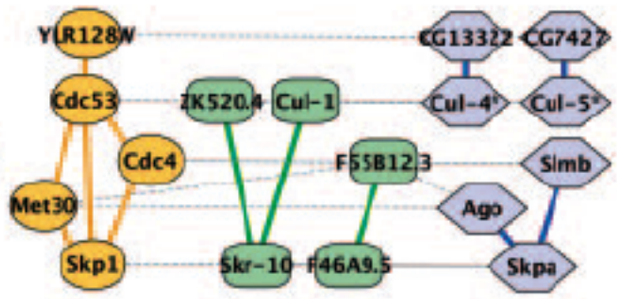


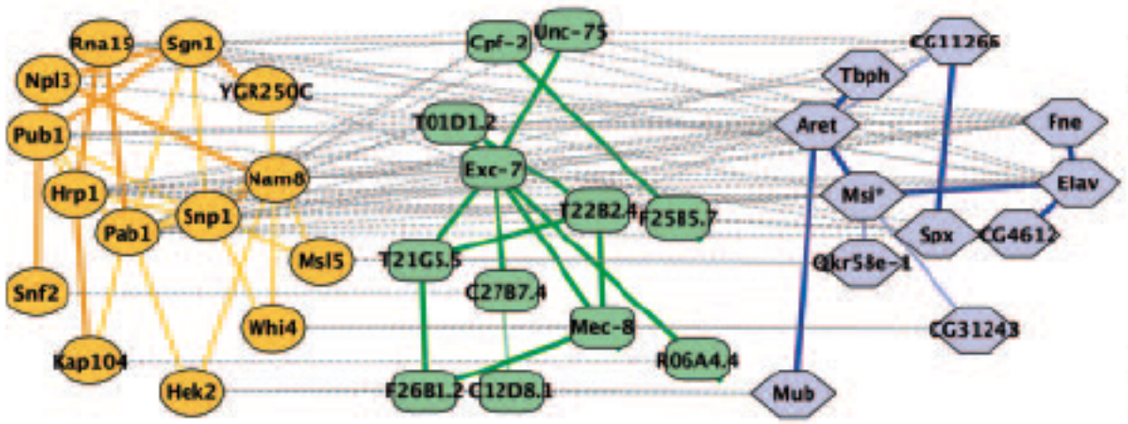


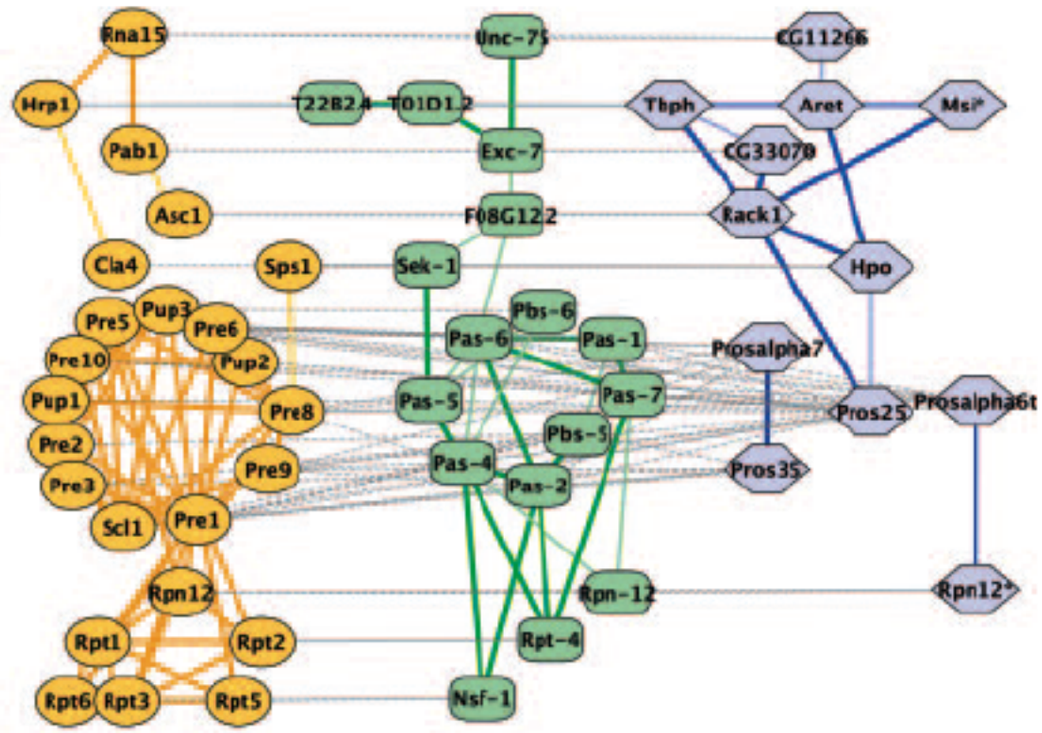
yeast worm fly





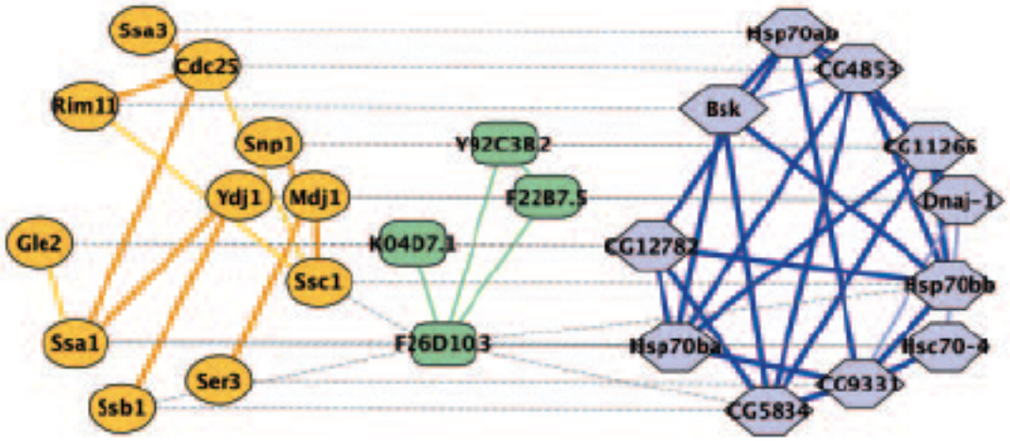




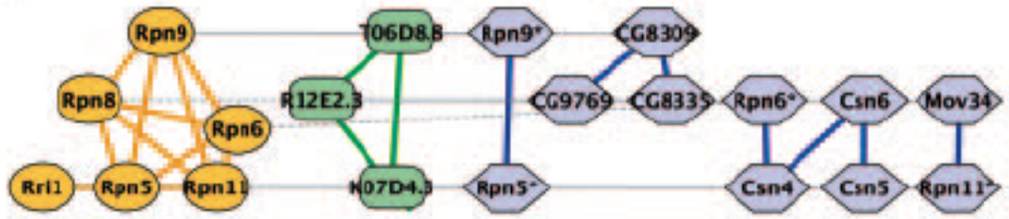


yeast worm fly

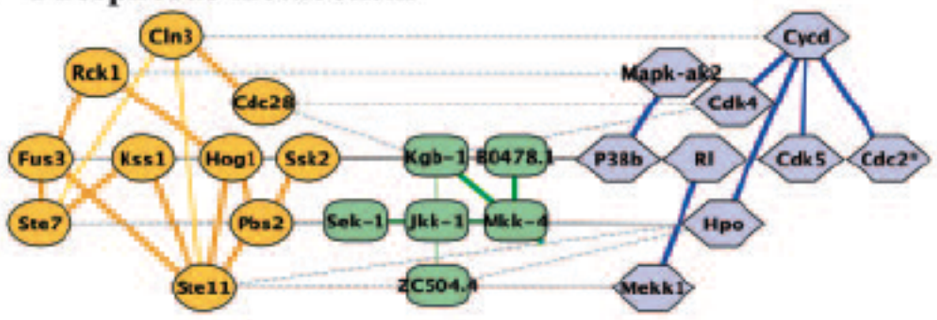
i **Protein folding**



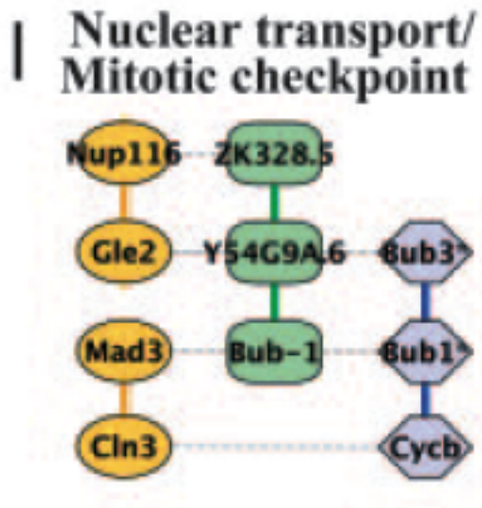
j **Protein metabolism**

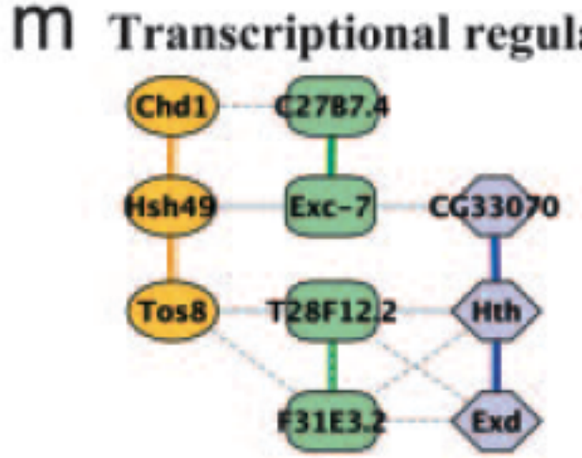


k **Phosphorus metabolism**



yeast worm fly





yeast worm fly

- In addition to the three-way comparison, the authors performed all possible pairwise alignments: yeast/worm, yeast/fly, and worm/fly

- The process identified 220 significant conserved clusters for yeast/worm, 835 for yeast/fly, and 132 for worm/fly

Græmlin: General and robust alignment of multiple large interaction networks

Jason Flannick,[1,4] Antal Novak,[1,4] Balaji S. Srinivasan,[2,3] Harley H. McAdams,[2] and Serafim Batzoglou[1,5]

- Work described so far is limited to two (or three) PPI networks

- Graemlin is capable of multiple alignment of an arbitrary number of networks, searches for conserved functional modules, and provides a probabilistic formulation of the topology-matching problem

- Available from: http://graemlin.stanford.edu

Graemlin's Features

- Multiple alignment
- Local and global
- Network-to-network alignment (an exhaustive list of conserved modules) and query-to-network alignment (matches to a particular module within a database of interaction networks)

The Alignment Problem

- Each interaction network is represented as a weighted graph $G_i=(V_i,E_i)$, where nodes correspond to proteins and each weighted edge specifies the probability that two proteins interact
- A network alignment is a set of subgraphs chosen from the interaction networks of different species, together with a mapping between aligned proteins
- The mapping is required to be transitive (if protein A is aligned to proteins B and C, then protein B must also be aligned to protein C)
- It follows that the groups of aligned proteins are disjoint, and are referred to as equivalence classes

The Alignment Problem

- It is also required that all aligned proteins be homologous, hence all proteins in the same equivalence class are in general members of the same protein family
- In other words, an alignment is a collection of protein families whose interactions are conserved across a given set of species
- Because the members of a protein family descend from a common ancestor, this allows to reconstruct the evolutionary events leading from each ancestral protein to its extant descendants

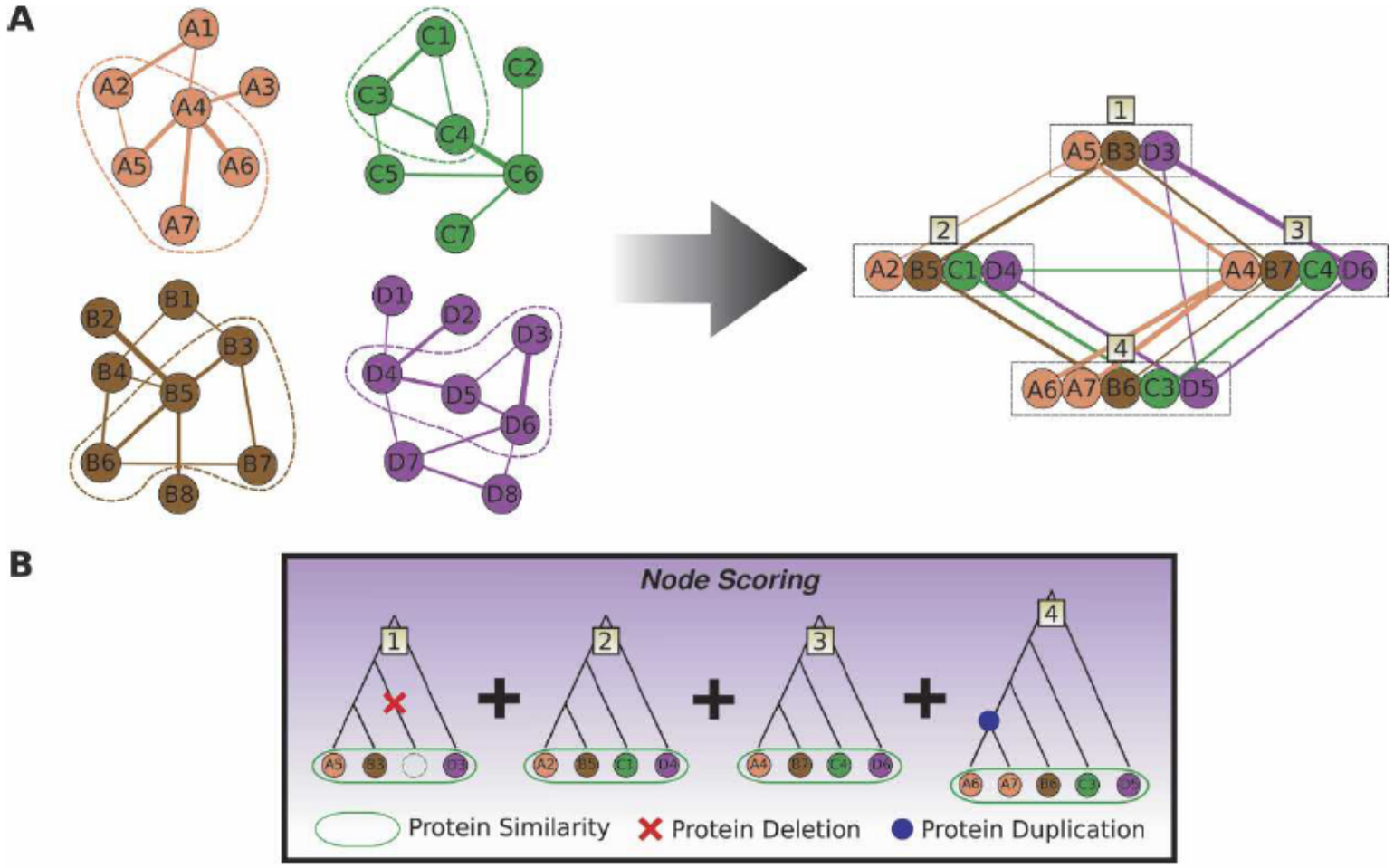
The Alignment Problem

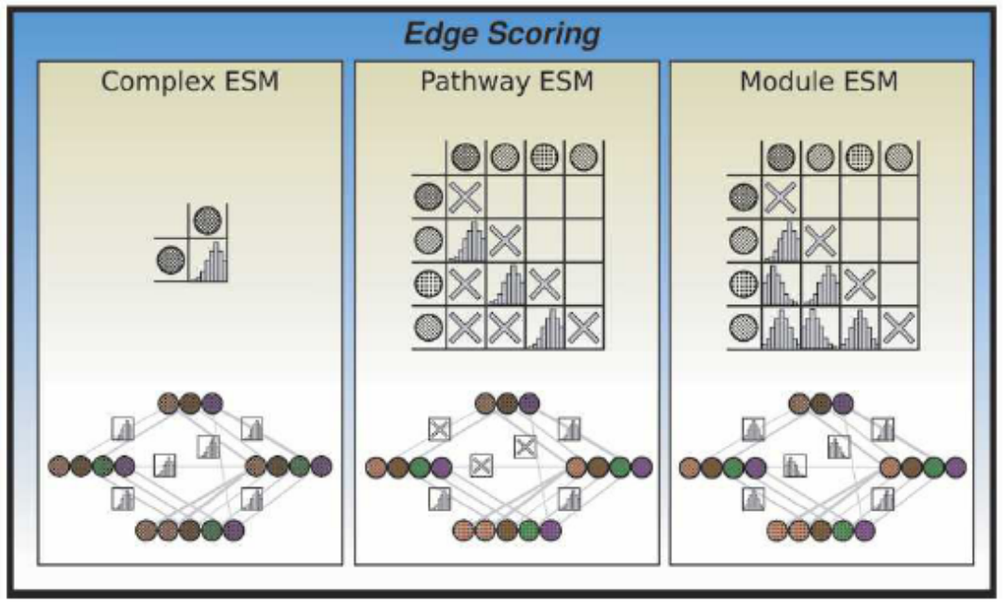
- Two elements are needed:
  - A scoring framework that captures the knowledge about module evolution
  - An algorithm to rapidly identify high-scoring alignments

Scoring an Alignment

- Define two models that assign probabilities to the evolutionary events leading from the hypothesized ancestral module to modules in the extant species
  - The alignment model, M, posits that the module is subject to evolutionary constraint
  - The random model, R, assumes that the proteins are under no constraints
- The score of an alignment is the log-ratio of the two probabilities

An Overview of the Scoring Scheme

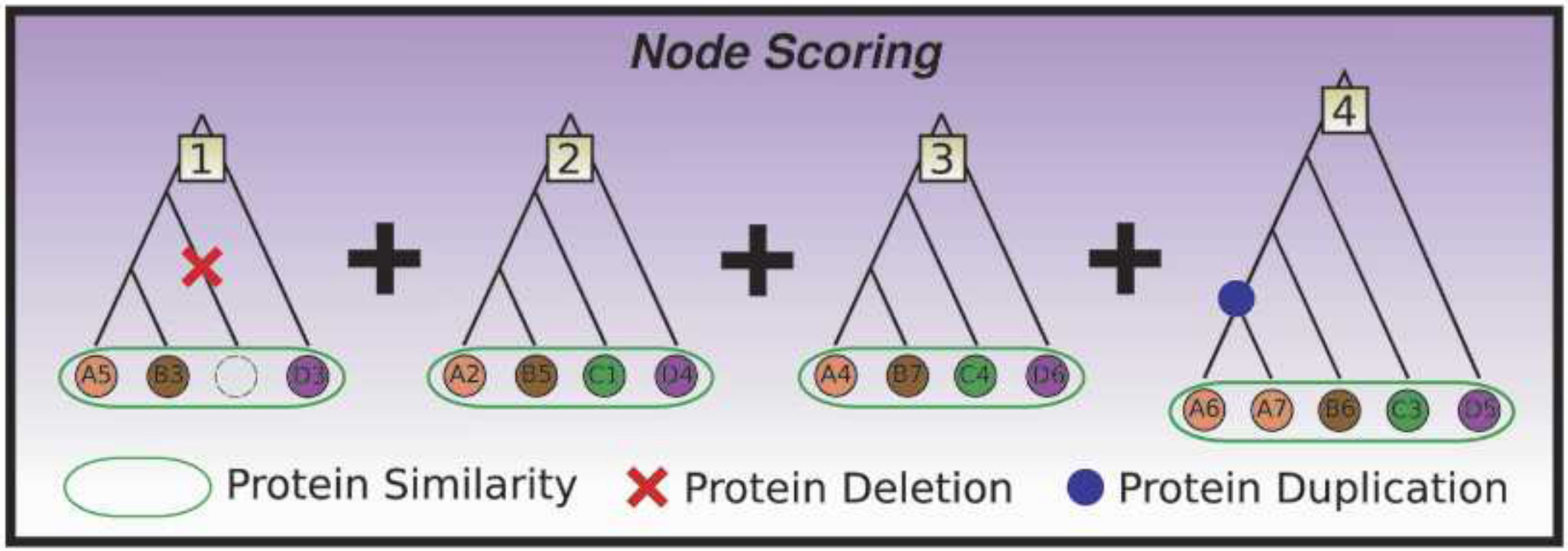




Node Scoring

- To score an equivalence class, Graemlin uses a scheme that reconstructs the most parsimonious ancestral history of an equivalence class, based on five types of evolutionary events: protein sequence mutations, proteins insertions and deletions, protein duplications, and protein divergences
- The models M and R give each of these events a different probability
- Graemlin uses weighted sum-of-pairs scoring to determine the probabilities for sequence mutations

Node Scoring



Edge Scoring

- Each edge e is assigned a score S_e=log(Pr_M(e)/Pr_R(e))
- The random model R assigns each edge a probability parametrized by its weight and degrees of its endpoints (captures the notion that two nodes of high degree are more likely to interact by chance than two nodes of low degree)
- The alignment model M is more involved

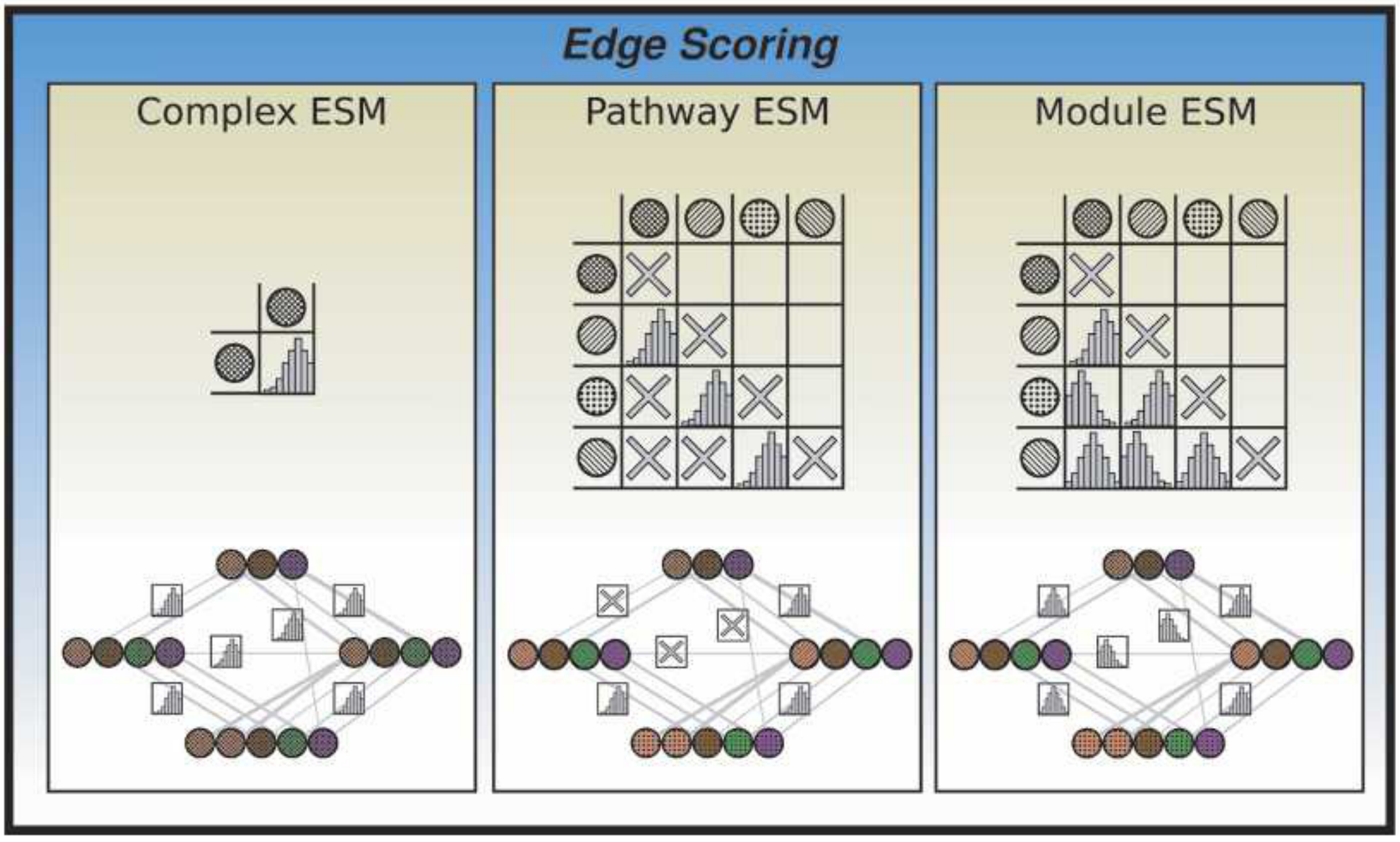
Edge Scoring

- The alignment model M uses an Edge Scoring Matrix, or ESM, to encapsulate the desired module structure into a symmetric matrix
- An ESM has a set of labels by which its rows and columns are indexed, and each cell in the matrix contains a probability distribution over edge weights
- To score edges in an alignment, Graemlin first assigns to each equivalence class one of the labels from the ESM. Then, it scores each edge e using the cell in the matrix indexed by the labels of the two equivalence classes to which its endpoints belong: the function in the cell maps the weight of the edge to a probability $Pr_M(e)$, which is used to compute the score S_e

Edge Scoring

- To search for conserved protein complexes, Graemlin uses a Complex ESM, which consists of a single label with an alignment distribution assigning high probabilities to high edge weights
- A Pathway ESM has one label for each protein in the pathway and rewards high edge weights between adjacent proteins; between all other proteins, the alignment and random distributions are the same, so that Graemlin neither rewards nor penalizes edges connected nonadjacent proteins
- A Module ESM is used for query searching: it has a label for each node in the query and generates the alignment distribution based on the edges that are present or absent in the query

Edge Scoring



Alignment Algorithm

- Graemlin uses slightly different methodologies for pairwise and multiple alignments

Pairwise Alignment Algorithm

- To search for high-scoring alignments between a pair of networks, Graemlin first generates a set of seeds (d-clusters), which it uses to restrict the size of the search space
- The seeds consist of *d* proteins that are close together in a network
- For each network, Greamlin constructs one d-cluster for each node by finding the d-1 nearest neighbors of that node, where the length of an edge is the negative logarithm of its weight

Pairwise Alignment Algorithm

- Graemlin compares two d-clusters D_1 and D_2 by mapping a subset of nodes in D1 to a subset of nodes in D2 and reporting a score equal to the sum of all pairwise scores induced by the mapping; the score of two d-clusters is the highest-scoring such mapping
- Graemlin identifies pairs of d-clusters, one from each network, that score higher than a threshold *T* and uses these as seeds

Pairwise Alignment Algorithm

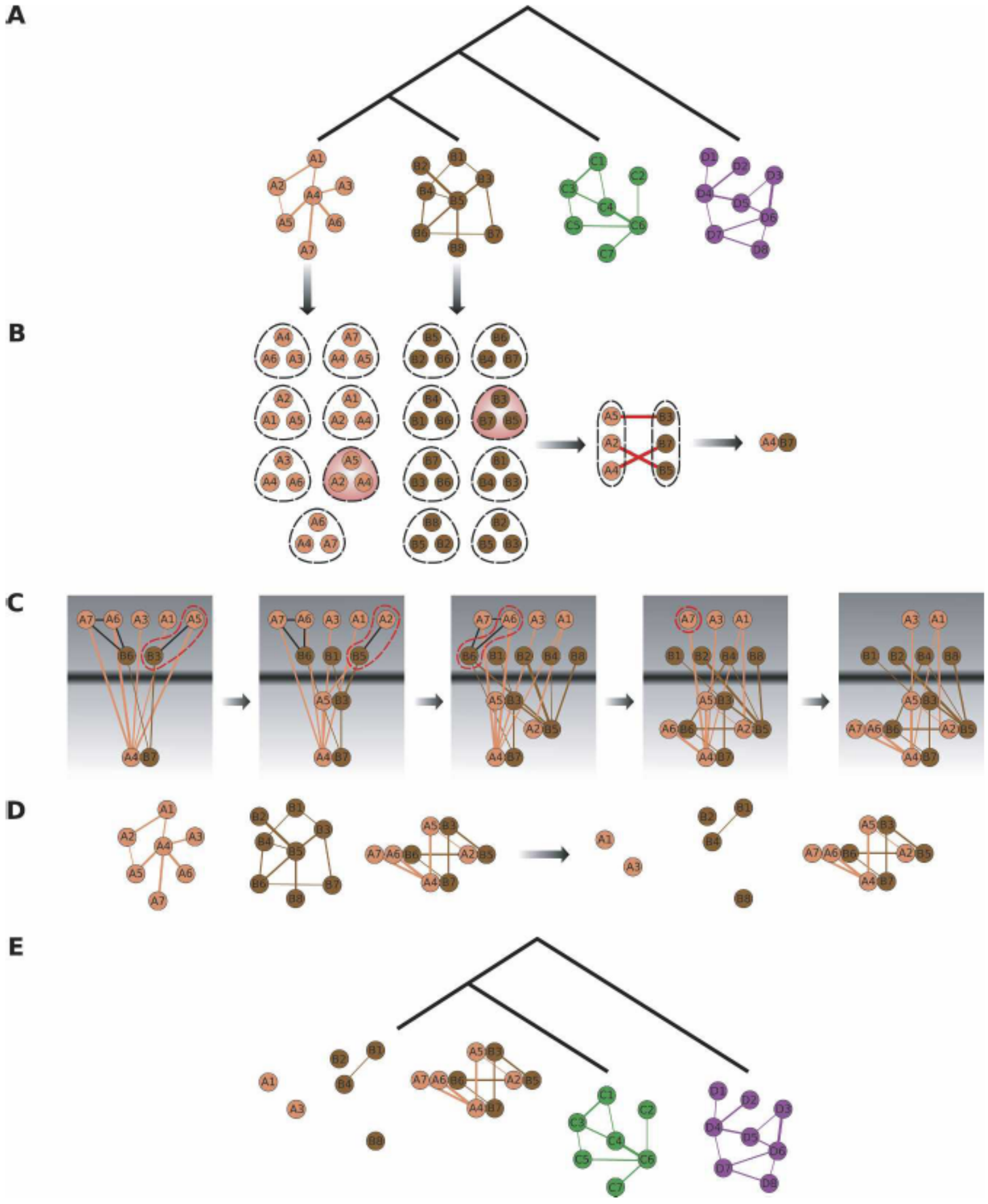
- Benefits of using d-clusters:
  - Graemlin can compare them rapidly, since the comparison neglects edge scores
  - The parameters *d* and *T* allow for a speed-sensitivity trade-off
  - High-scoring alignments are likely to contain high-scoring d-clusters, since a high node score of an alignment is usually a prerequisite to a high overall score

Pairwise Alignment Algorithm

- Given two networks, Graemlin enumerates the set of seeds between them and tries to transform each, in turn, into a high-scoring alignment
- The seed extension phase is greedy and occurs in successive rounds
- At each step, all proteins adjacent to some node in the alignment constitute the "frontier," which contains candidates to be added to the alignment

Pairwise Alignment Algorithm

- Graemlin selects from the frontier the pair of proteins that, when added to the alignment, yields the maximal increase in score
- The extension phase stops when no pair of proteins on the frontier can increase the score of the alignment
- Graemlin uses several heuristics to control for the exponential increase in the size of the frontier as it adds more nodes to the alignment


A
B
C
D
E

Multiple Alignment Algorithm

- Graemlin performs multiple alignment using an analog of the progressive alignment technique commonly used in sequence alignment
- Using a phylogenetic tree, it successively aligns the closest pair of networks, constructing several new networks from the resulting alignments
- Graemlin places each new network at the parent of the pair of networks that it just aligned
- The constructed networks contain nodes that are no longer proteins but equivalence classes
- Graemlin continues this process until the only remaining networks are at the root of the phylogenetic tree

Multiple Alignment Algorithm

- To enable comparisons of unaligned parts of a network to more distant species as it traverses the phylogenetic tree, rather than construct a network only from the high-scoring alignments, Graemlin also maintains two additional networks composed of the unaligned nodes from the two original networks
- The end result is that after completion of the entire multiple alignment, Graemlin produces multiple alignments of all possible subsets of species
- Graemlin avoids exponential running time in practice because after each pairwise alignment, the networks it constructs have small overlaps (the total number of nodes in all networks therefore does not increase significantly)

Experimental Setup

- Graemlin was tested on a set of 10 microbial protein interaction networks constructed via the SRINI algorithm
- They also used PPI networks from S. cerevisiae, C. elegans, and D. melanogaster, to compare the performance of the method to other methods that had used these three species

Experimental Setup

Species	Color	# Nodes	Edge Threshold	# Edges	# Edges per Node	# Alignable KEGGs
Campylobacter jejuni NCTC 11168		1629	0.25	22116	13.58	42
			0.5	6171	3.79	29
Caulobacter crescentus		3737	0.25	40568	10.86	70
			0.5	6018	1.61	55
Escherichia coli K12		4242	0.25	216426	51.02	72
			0.5	35132	8.28	70
Helicobacter pylori 26695		1576	0.25	12960	8.22	32
			0.5	3723	2.36	26
Mycoplasma tuberculosis H37Rv		3991	0.25	129183	32.37	75
			0.5	17380	4.35	61
Salmonella typhimurium LT2		4527	0.25	94609	20.90	61
			0.5	18149	4.01	55
Streptococcus pneumoniae TIGR4		2094	0.25	25732	12.29	29
			0.5	4607	2.20	23
Streptomyces coelicolor		8154	0.25	230467	28.26	76
			0.5	60852	7.46	54
Synechocystis PCC 6803		3166	0.25	69439	21.93	47
			0.5	13963	4.41	32
Vibrio cholerae		3835	0.25	36087	9.41	61
			0.5	7886	2.06	45
Saccharomyces cerevisiae	N/A	4766	N/A	15200	3.19	22
Caenorhabditis elegans	N/A	2629	N/A	3950	1.50	0
Drosophila melanogaster	N/A	7067	N/A	21822	3.09	4

Experimental Setup

Table 2. KEGG pathway conservation statistics

Species set	Threshold	No. of alignable KEGGs
E. coli, C. crescentus	0.25	55
	0.5	44
E. coli, M. tuberculosis	0.25	60
	0.5	54
E. coli, V. cholerae	0.25	54
	0.5	39
E. coli, S. coelicolor	0.25	57
	0.5	43
E. coli, C. crescentus, V. cholerae	0.25	47
	0.5	27
C. jejuni, E. coli, H. pylori	0.25	28
	0.5	15

This table shows the number of alignable KEGG pathways that are present for various subsets of species. An alignable KEGG pathway is present for a given subset of species if the pathway is alignable in each of the species in the subset.

Experimental Setup

- The sensitivity (TP/(TP+FN)) of a method was assessed by counting the number of KEGG pathways that it aligned between two species (a "hit" occurs if the method aligns at least three proteins in the pathway to their counterparts in the other species)
- The "coverage" of a pathway is the fraction of proteins correctly aligned within that pathway

Experimental Setup

- To measure the specificity (TN/(FP+TN)) of a method, the authors computed the number of "enriched" alignments
- To calculate enrichment, the authors first assign to each protein all of its annotations from level eight or deeper in the GO hierarchy
- Given an alignment, the authors then discarded unannotated proteins and calculated its enrichment using the GO TermFinder
- They considered an alignment to be enriched if the P-value of its enrichment was < 0.01

Experimental Setup

- An alternative measure of specificity counts the fraction of nodes that have KEGG orthologs but were aligned to any nodes other than their KEGG orthologs

Experimental Results

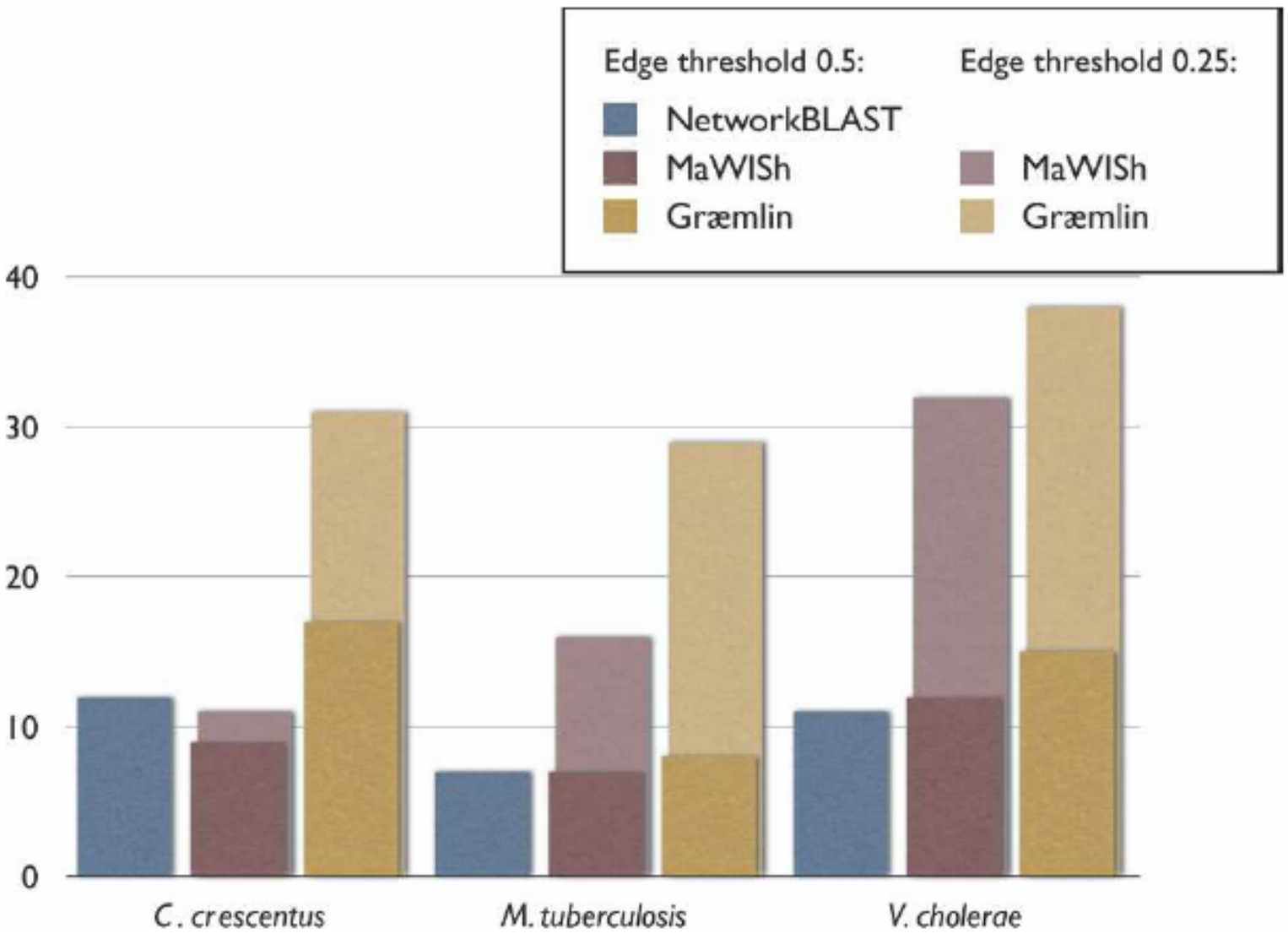


Figure 4. Sensitivity comparison of methods. For three pairwise alignments of *E. coli*, shown are the number of KEGGs hit by each aligner. For Græmlin and MaWISh, this graph includes results on networks with edge thresholds of both hold and 0.5. For NetworkBLAST, however, we only include results on networks thresholded at 0.5, as it did not scale to denser inputs.

Experimental Results

Table 3. Results on pairwise alignment of complete networks thresholded at 0.5

		KEGGs hit	KEGG coverage	Alignments enriched	Running time (sec)
E. coli vs. *C. crescentus*					
MaWISh		9 (20%)	32%	72%	3
NetworkBLAST	Pathway	6 (14%)	28%	61%	9624
	Complex	12 (27%)	49%	72%	
Græmlin	Pathway	15 (34%)	47%	68%	21
	Complex	17 (39%)	45%	67%	11
E. coli vs. *M. tuberculosis*					
MaWISh		7 (13%)	20%	85%	3
NetworkBLAST	Pathway	7 (13%)	24%	88%	301
	Complex	7 (13%)	32%	88%	
Græmlin	Pathway	8 (15%)	36%	89%	11
	Complex	8 (15%)	39%	89%	8
E. coli vs. *V. cholerae*					
MaWISh		12 (31%)	35%	64%	3
NetworkBLAST	Pathway	10 (26%)	35%	58%	8797
	Complex	11 (28%)	41%	64%	
Græmlin	Pathway	19 (49%)	48%	75%	13
	Complex	15 (38%)	55%	74%	12
E. coli vs. *S. coelicolor*					
MaWISh		N/A	N/A	N/A	N/A
NetworkBLAST	Pathway	6 (14%)	23%	46%	122,168
	Complex	10 (23%)	67%	95%	
Græmlin	Pathway	8 (19%)	58%	88%	734
	Complex	9 (21%)	59%	85%	829

For each pair of species, we performed complete network-to-network alignment using MaWISh and Græmlin. For each tested method, shown, from left, is the total number of KEGG pathways hit by an alignment, the fraction of KEGG pathways hit by an alignment, the average coverage of a KEGG pathway, the percentage of enriched alignments, and the total running time. We calculated the average coverage of KEGGs with respect to only those KEGGs that an aligner hit, and measured running time in CPU-seconds.

Experimental Results

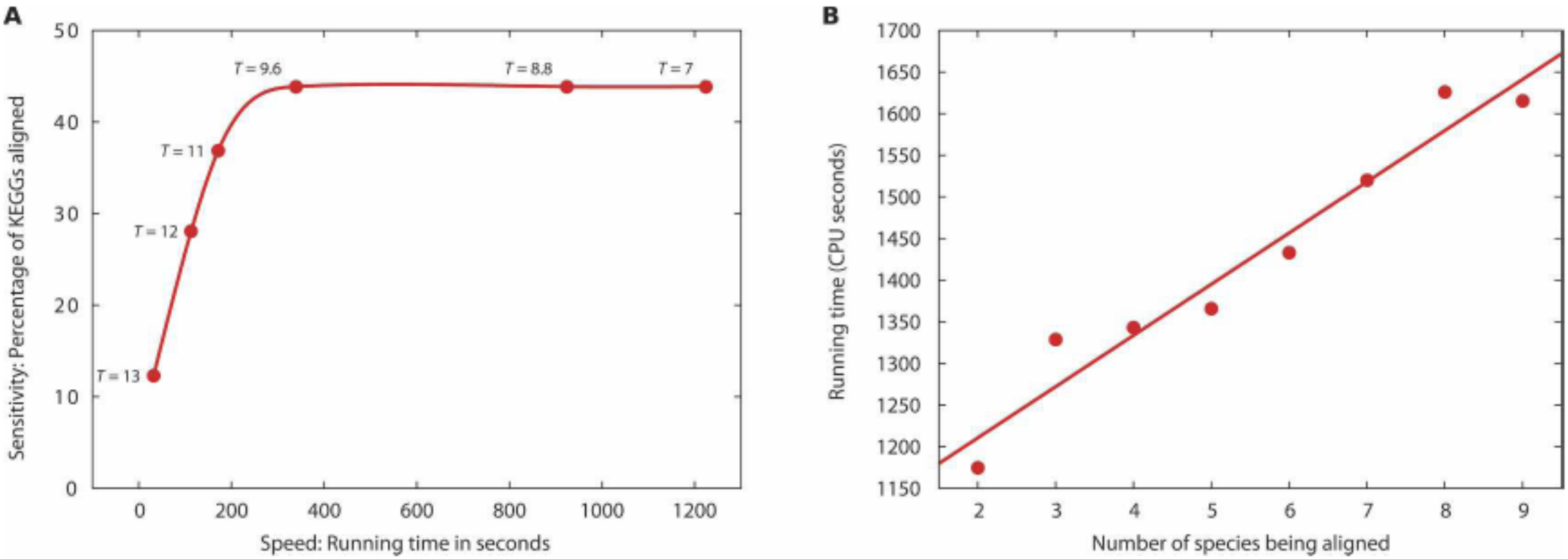


Figure 5. Running-time performance of Græmlin. (*A*) The speed sensitivity trade-off. Each point represents a run of Græmlin with $d = 4$ and different values of *T*. For each set of parameters, the *x*-axis plots the running time, and the *y*-axis plots the fraction of alignable KEGGs hit. (*B*) Progressive multiple alignment. Beginning with *E. coli*, we added species of increasing evolutionary distance to the multiple alignment. The pairwise running time is comparatively high because the two species aligned, *E. coli* and *S. typhimurium*, are the two most similar species and have many high-scoring alignments. In this manner, adding particularly close species to the alignment can lead to higher-than-average increases in running time, but over all species the average scaling will remain roughly linear.

Experimental Results

Table 4. Results on multiple alignment of complete networks

		KEGGs hit	KEGG coverage	Alignments enriched	Running time (sec)
0.25 threshold					
E. coli vs. *C. crescentus* vs. *V. cholerae*					
Græmlin	Pathway	27 (57%)	68%	72%	329
	Complex	29 (62%)	71%	79%	251
E. coli vs. *C. jejuni* vs. *H. pylori*					
Græmlin	Pathway	16 (57%)	57%	87%	44
	Complex	17 (61%)	63%	89%	43
0.5 threshold					
E. coli vs. *C. crescentus* vs. *V. cholerae*					
NetworkBLAST	Pathway Complex	N/A	N/A	N/A	$>10^6$
Græmlin	Pathway	7 (26%)	67%	72%	63
	Complex	9 (33%)	62%	75%	38
E. coli vs. *C. jejuni* vs. *H. pylori*					
NetworkBLAST	Pathway	5 (33%)	41%	94%	32,394
	Complex	4 (27%)	38%	96%	
Græmlin	Pathway	3 (20%)	74%	82%	12
	Complex	3 (20%)	72%	79%	12

We performed three-way multiple network alignment using NetworkBLAST and Græmlin; the columns in this table are analogous to those in Table 3.

Experimental Results

Table 5. Results on alignment of a query network to a database thresholded at 0.5

		KEGGs hit	KEGG coverage	Running time (sec)
E. coli vs. *C. crescentus*				
MaWISh		15 (34%)	31%	37
NetworkBLAST	Pathway	8 (18%)	32%	3453
	Complex	10 (23%)	49%	
Græmlin	Pathway	20 (45%)	45%	17
	Complex	20 (45%)	47%	3
	Module	20 (45%)	48%	23
C. crescentus vs. *E. coli*				
MaWISh		9 (20%)	32%	130
NetworkBLAST	Pathway	10 (23%)	37%	4788
	Complex	10 (23%)	41%	
Græmlin	Pathway	15 (34%)	39%	6
	Complex	15 (34%)	42%	5
	Module	15 (34%)	42%	33
E. coli vs. *M. tuberculosis*				
MaWISh		10 (19%)	19%	93
NetworkBLAST	Pathway	12 (22%)	23%	3947
	Complex	12 (22%)	29%	
Græmlin	Pathway	17 (31%)	31%	3
	Complex	17 (31%)	35%	3
	Module	17 (31%)	35%	22
M. tuberculosis vs. *E. coli*				
MaWISh		6 (11%)	12%	138
NetworkBLAST	Pathway	10 (19%)	19%	5047
	Complex	7 (13%)	22%	
Græmlin	Pathway	13 (24%)	25%	5
	Complex	14 (26%)	26%	5
	Module	14 (26%)	27%	28

For each pair of species, using MaWISh, NetworkBLAST, and Græmlin, we successively aligned each KEGG pathway in the query species to the complete network of the database species. For each tested method, shown, from left, is the total number of KEGG pathways with a database hit, the fraction of KEGG pathways with a database hit, the average coverage of a KEGG pathway, and the total running time. As NetworkBLAST does not have an option to search separately for pathways and complexes, the table lists the combined running time of both searches.

Experimental Results

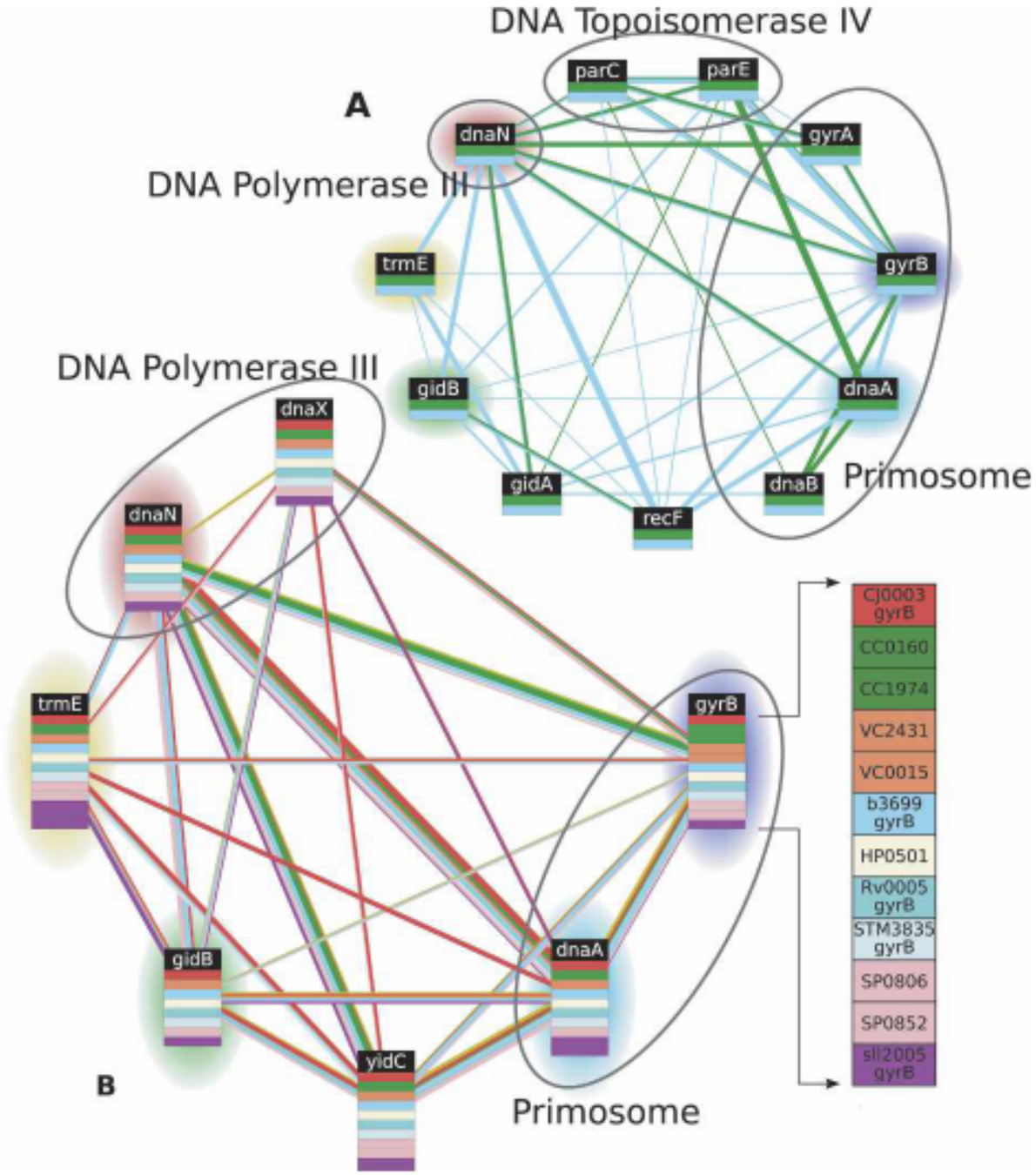


Figure 6. Two alignments of proteins involved in DNA replication. (*A*) A pairwise alignment between *E. coli* and *C. crescentus* includes several proteins involved in cell division as well as a conserved thiophene and furan oxidation protein. (*B*) A multiple alignment extends the pairwise alignment to include *S. typhimurium, V. cholerae, C. jejuni, H. pylori, M. tuberculosis, S. pneumoniae,* and *Synechocystis*. In this and subsequent figures, each colored box represents a protein and each vertical array of boxes represents an equivalence class; Græmlin hypothesizes that proteins in the same equivalence class performed the same function in the most recent common ancestor of the aligned species. To avoid clutter, individual proteins are not labeled, and, instead, each equivalence class is labeled with the consensus gene name of the proteins in it; as an example of the set of proteins aligned in an equivalence class, the detailed *inset* shows the specific proteins aligned to gyrB. Each protein is colored according to species, using the color code in Table 1; edges are also colored using the same scheme, and the width of each edge is proportional to its weight. In this figure, equivalence classes in the multiple alignment are highlighted the same color as the pairwise equivalence classes that they subsume.

Experimental Results

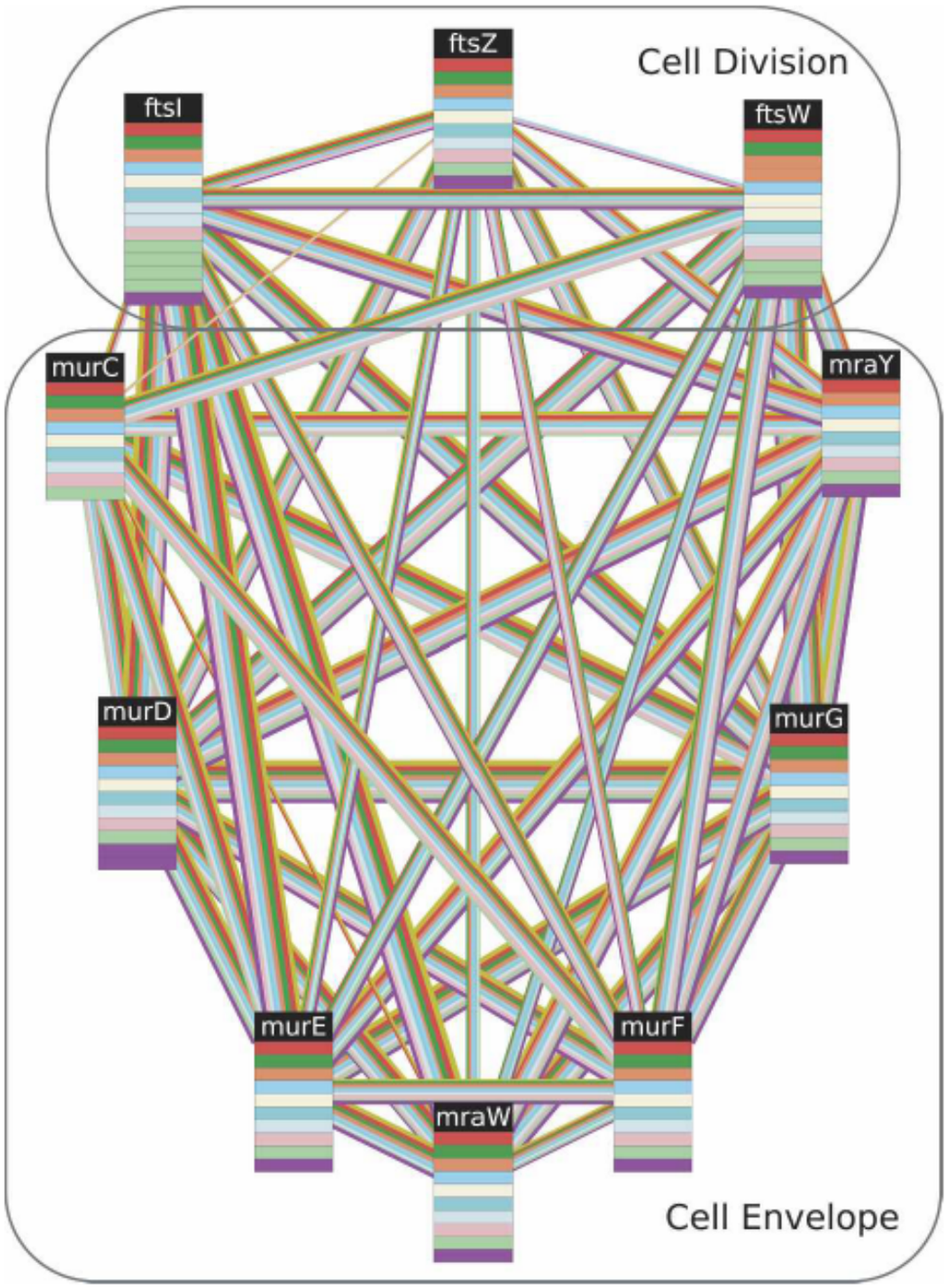


Figure 7. An alignment including proteins involved in cell division. This alignment implicates several proteins in bacterial cell division; it includes all species listed in Table 1.

Experimental Results

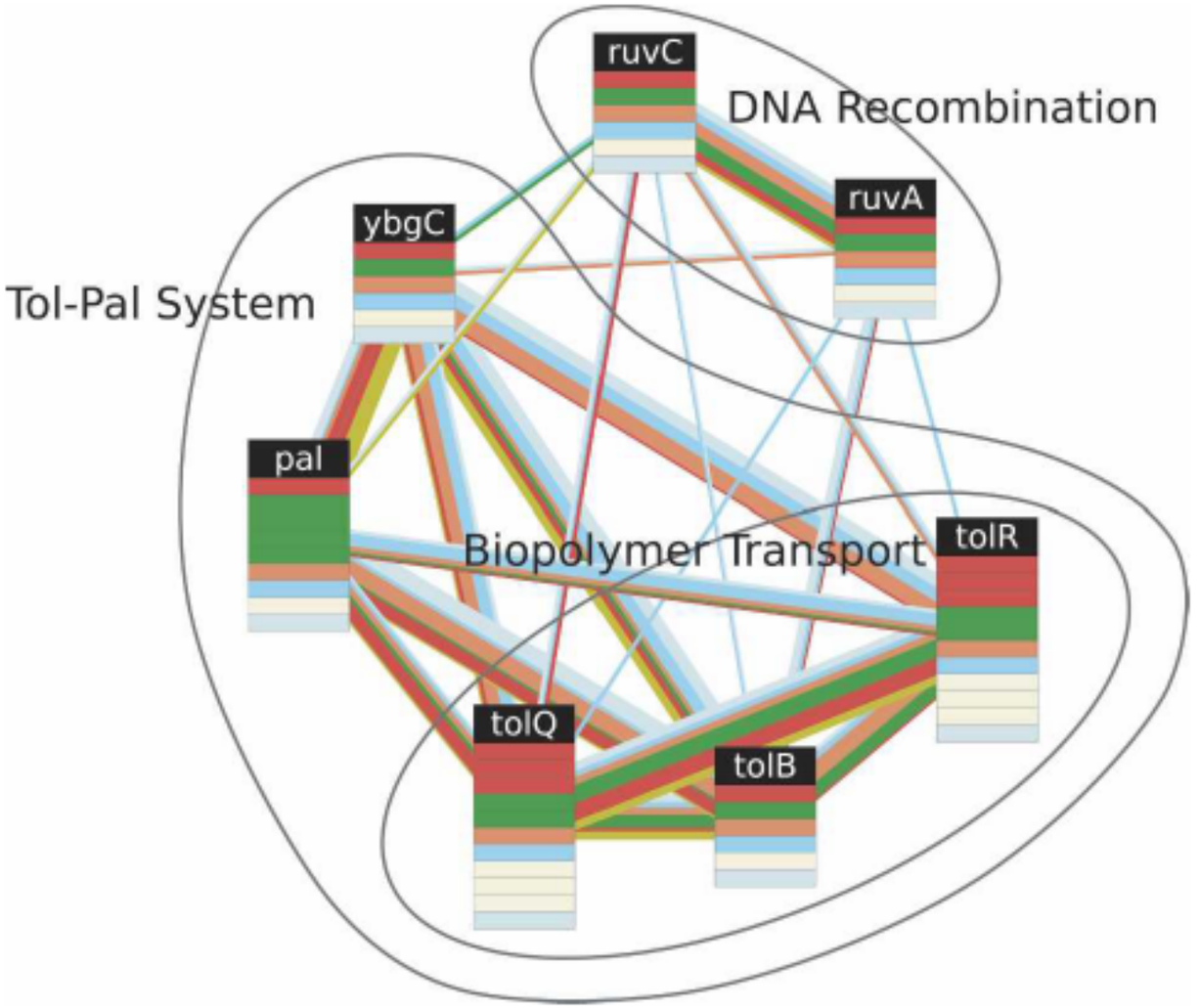


Figure 8. An alignment of a hypothetical functional module. In this alignment, proteins involved in biopolymer transport interact with proteins involved in DNA recombination. The sum total of these interactions in six species suggests that the proteins may be a part of a conserved functional module responsible for transformation.