Lecture 7
Protein-Protein Interaction

Instructor: Teresa Przytycka

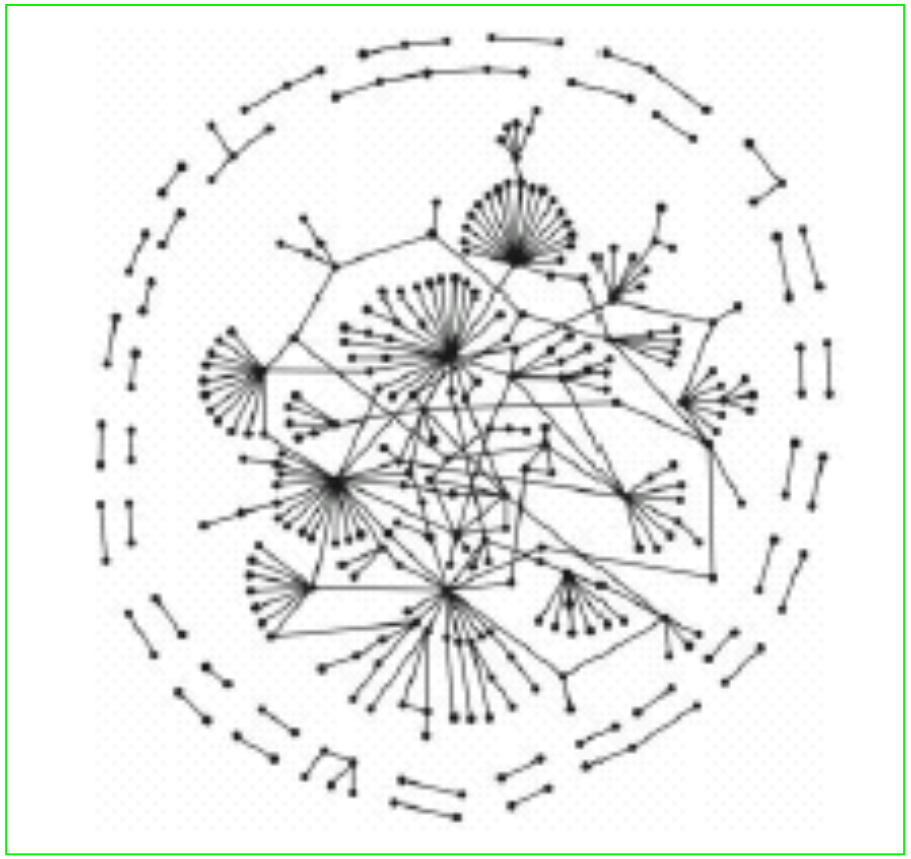
Protein-protein interaction network

- Molecular processes are sequences of events mediated by proteins that act in a cooperative manner. This cooperation requires that proteins to interact and form protein complexes.
- Protein – protein interaction network:
  - Nodes – proteins
  - Edges - interactions

How do we know that a pair of proteins interact?

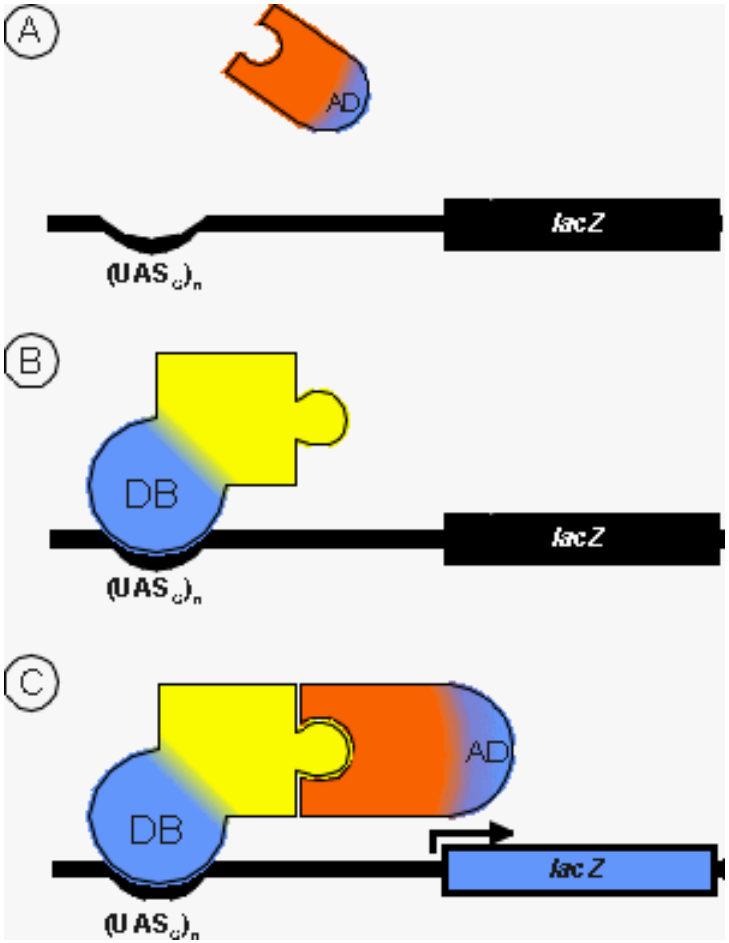
- The two proteins have been crystallized together.
- High throughput interaction screening methods:
  - Yeast two hybrid experiments (Y2H)
  - Protein complex purification (PCP)
- Problem with high throughput method:
  - significant amount of false positives and false negatives

Protein-protein interaction network in yeast (nuclear proteins)



From
Maslov & Sneppen
Science 2002

Y2H



Principle of the Two-hybrid system. (A), (B) Two chimeras, one containing the DNA-binding domain (DB: blue circle) and one that contains an activation domain (AD: half blue circle), are co-transfected into an appropriate host strain. (C) If the fusion partners (yellow and red) interact, the DB and AD are brought into proximity and can activate transcription of reporter genes (here *LacZ*).

From *Yeast Two-Hybrid: State of the Art* *Wim Van Criekinge*[1*] *and Rudi Beyaert*[2]; *http://www.biologicalprocedures.com/bpo/arts/1/16/m16f1lg.htm*

CPC

- Take a set of proteins “baits”
- Expose each “bait” protein so to a set of “pray” proteins that potentially can form complexes with it.
- Allow the complexes to form
- Identify proteins in each complex
- Only complexes containing the “bait” protein are analyzed.

Computational Challenges

- Propose Computational Methods for detecting PPI and domain interactions
- Analyze such PPI networks
  - What properties of these networks tell us about interactions – any surprising properties?
  - Put some confidence measures on such interaction
  - Comparative analysis interaction networks

Computational Methods for predicting PPI

- Phylogenetic Profiles
- Rosetta Stone
- Gene Neighbors
- Co-evolution
- Gene clusters

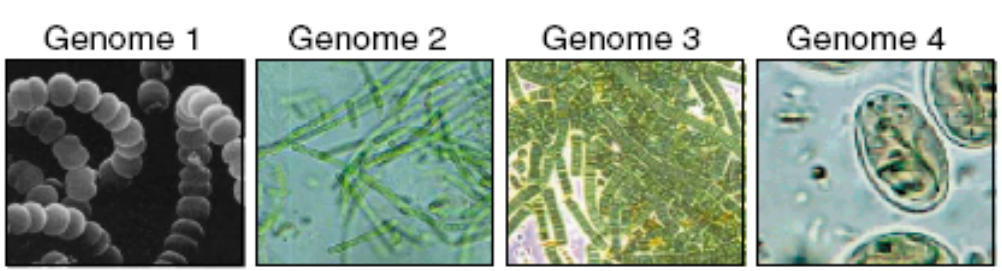
Predicting domain-domain interaction from protein-protein interaction

- Association method
- Maximum Expectation
- E-value (Eisenberg)

Phylogenetic Profile

Figure form Bowers at al., Genome Biology 2004

(c)



	Genome 1	Genome 2	Genome 3	Genome 4
Protein A	1	1	1	0
Protein B	1	1	1	0
Protein C	1	0	1	1
Protein D	1	1	0	1

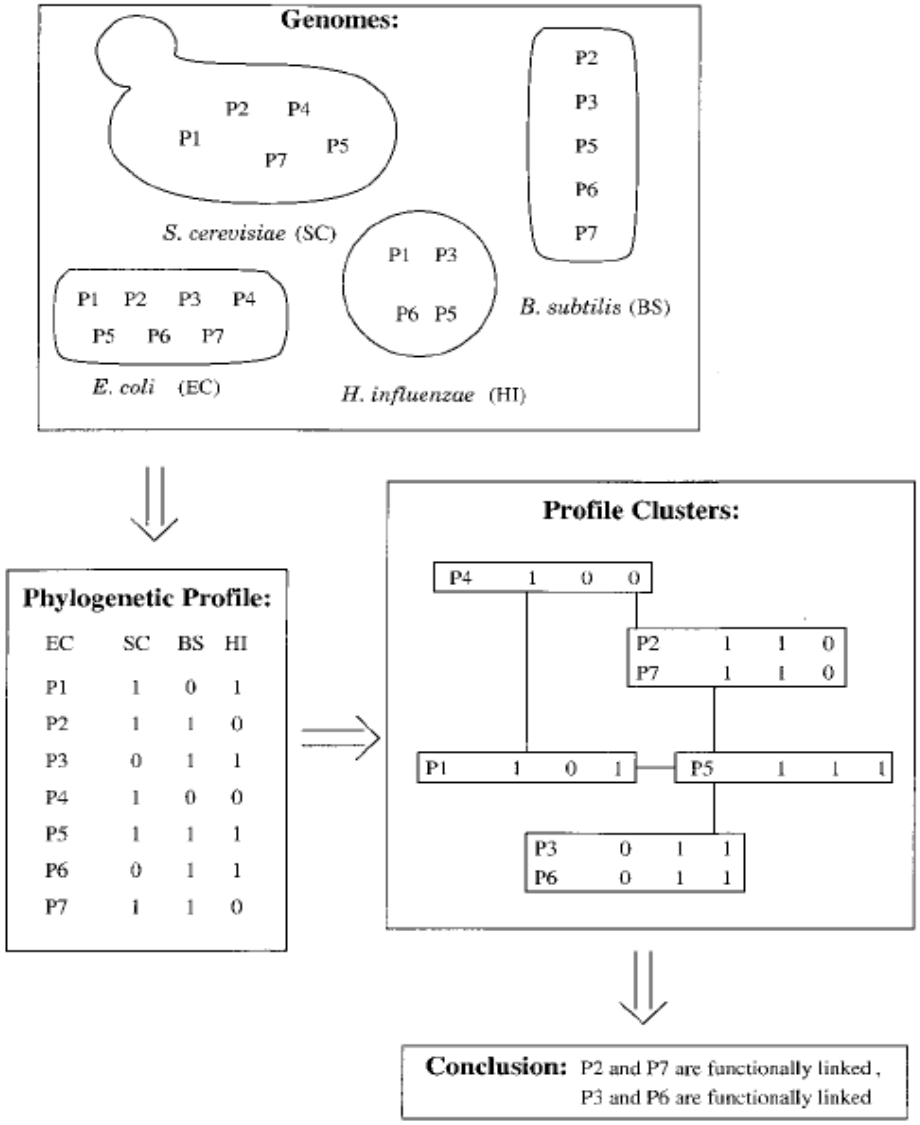
The same profiles

- Idea: Pairs of non-homologous proteins that are always both present or both absent in a genome suggest their functional dependence → possible interaction
- Profile of a protein: A vector of 0/1 where each position corresponds to one genome: 1- protein present 0-protein absent

Finding profile clusters

Pioneered by:
Pellegirni, Marcotte, Thompson, Eisenberg, Yeates PNAS 1999
(see also earlier paper by Huynen, Bork PNAS 1998)

Demonstrated that Proteins with same or similar evolutionary profiles are strongly functionally linked



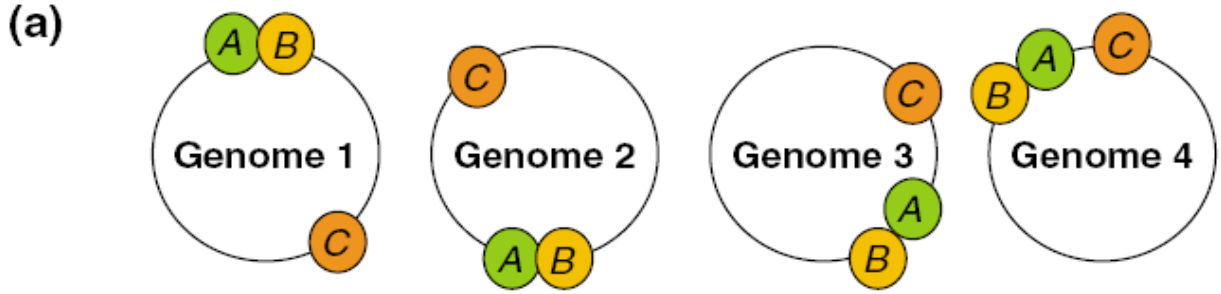
Gene Cluster Method

Within bacteria, proteins of closely related function are often transcribed from a single functional unit known as an operon. Operons contain two or more closely spaced genes located on the same DNA strand. These genes are often in proximity to a transcriptional promoter that regulates operon expression.

Advantage: Each operon is informative (multiple genome comparison is not necessary)

Gene Neighbors

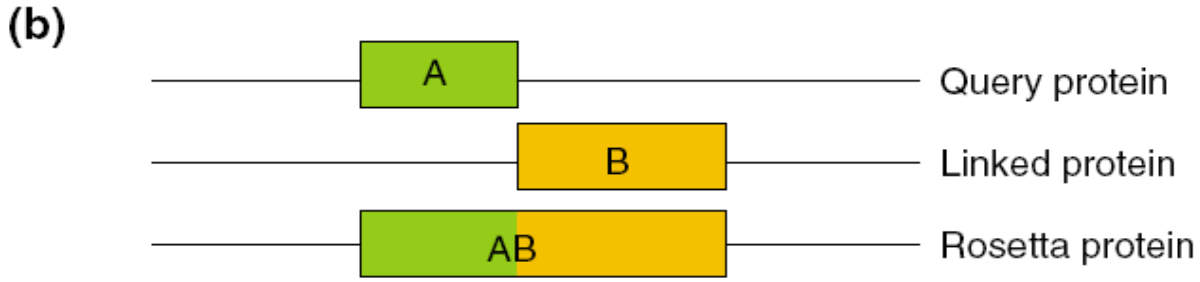
Figure form Bowers et al, Genome Biology 2004



- Gene A is a neighbor of B in several genomes - potential functional link

Rosetta Stone

Figure form Bowers, Genome Biology 2004



- A, B – two domains that
- Rosetta protein – protein containing both domains in some organism – indication that in anther organism these two domains (which now are in different proteins) may interact

SCIENCE VOL 285 30 JULY 1999

Detecting Protein Function and Protein-Protein Interactions from Genome Sequences

Edward M. Marcotte, Matteo Pellegrini, Ho-Leung Ng, Danny W. Rice, Todd O. Yeates, David Eisenberg*

A computational method is proposed for inferring protein interactions from genome sequences on the basis of the observation that some pairs of interacting proteins have homologs in another organism fused into a single protein chain. Searching sequences from many genomes revealed 6809 such putative protein-protein interactions in *Escherichia coli* and 45,502 in yeast. Many members of these pairs were confirmed as functionally related; computational filtering further enriches for interactions. Some proteins have links to several other proteins; these coupled links appear to represent functional interactions such as complexes or pathways. Experimentally confirmed interacting pairs are documented in a Database of Interacting Proteins.

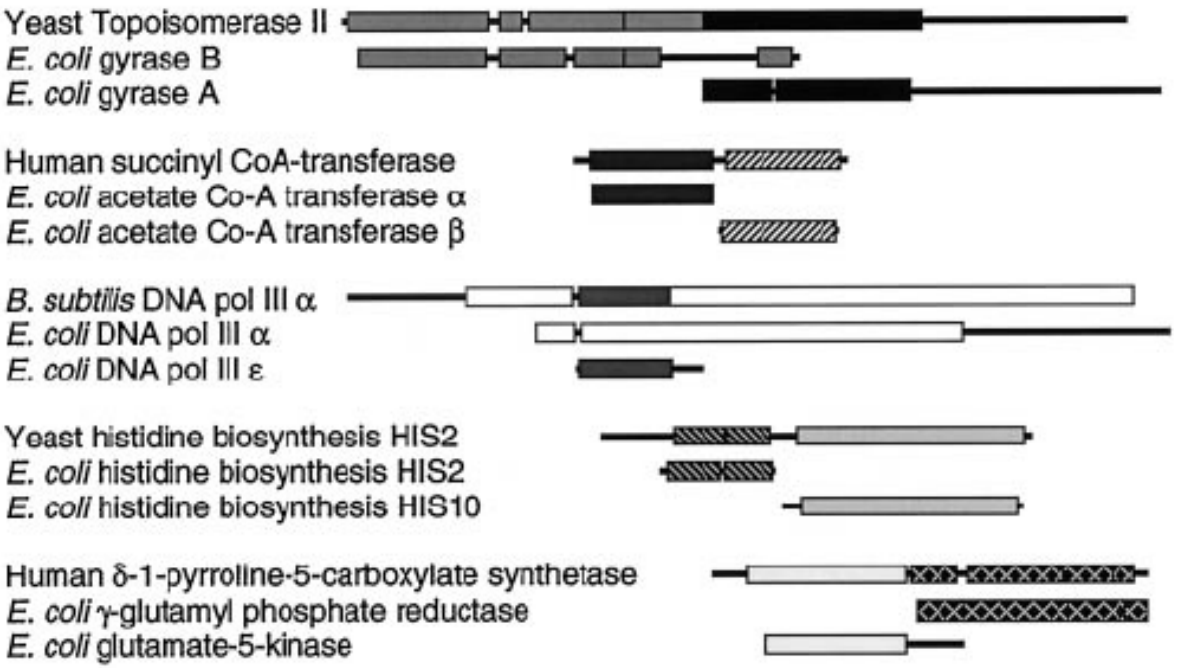


Fig. 1. Five examples of pairs of *E. coli* proteins predicted to interact by the domain fusion analysis. Each protein is shown schematically with boxes representing domains [as defined in the ProDom domain database (*17*)]. For each example, a triplet of proteins is pictured: The second and third proteins are predicted to interact because their homologs are fused in the first protein (called the Rosetta Stone protein in the text). The first three predictions are known to interact from experiments (*18*). The final two examples show pairs of proteins from the same pathway (two nonsequential enzymes from the histidine biosynthesis pathway and the first two steps of the proline biosynthesis pathway) that are not known to interact directly.

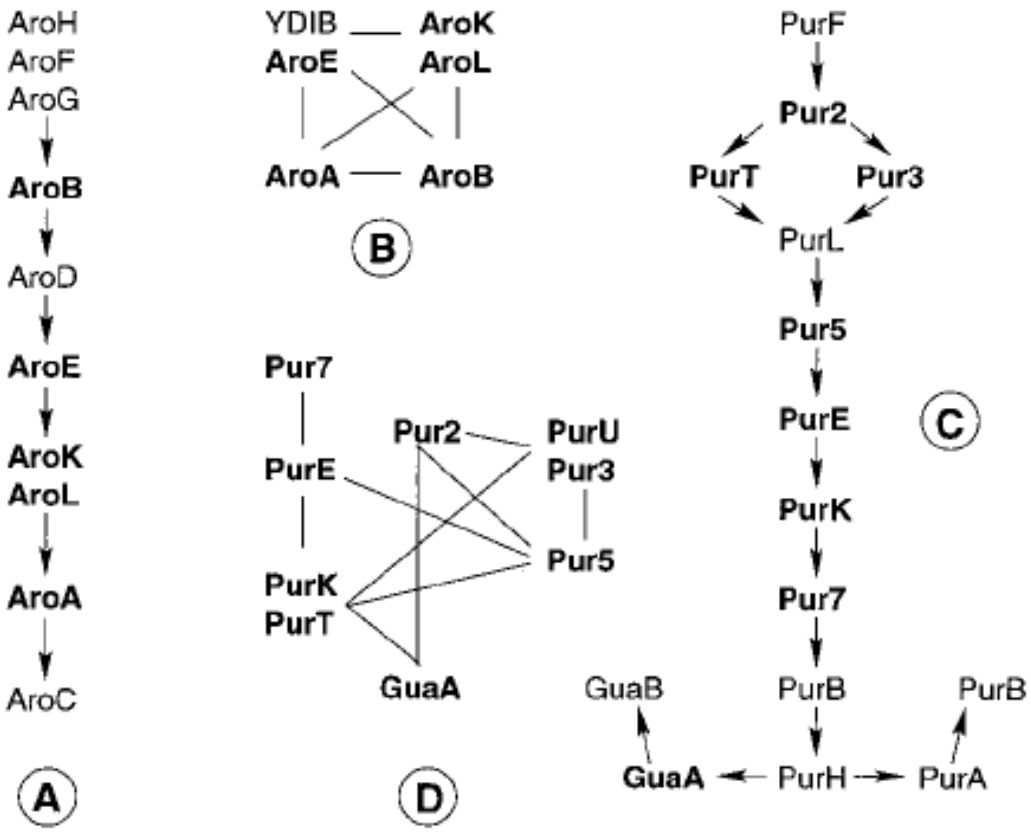


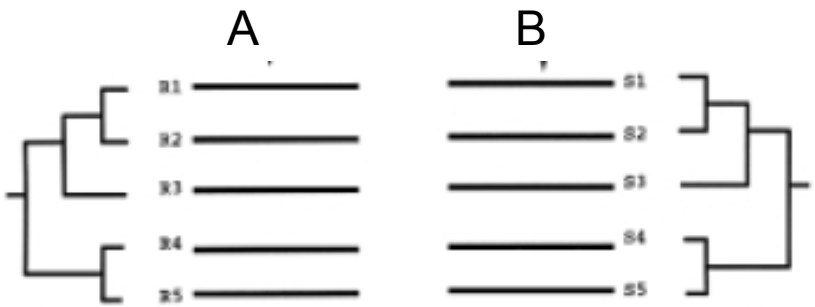
Fig. 2. Reconstruction of two metabolic pathways in *E. coli*, with only interactions predicted by the domain fusion method. Pathways A and C are the known pathways for biosynthesis of shikimate and purine, respectively; they are ordered by the traditional method of successive action of the enzymes on the known metabolites. Pathways B and D are constructed from the proteins in pathways A and C with connections predicted by the domain fusion method. In both cases, more than half of the proteins in the biochemical pathway are predicted by the domain fusion method to interact with other proteins of the pathway. It is possible that these groupings represent multiprotein complexes. Enzymes stacked together (for example, AroK and AroL) are homologs.

Co-evolution method

- Idea: Assume that protein A and B interact.
- If A and B are both present in several organisms and perform the same role in these organisms they interact in all these organisms
- Evolution of A and B should be correlated

Mirror Tree method

- Given two proteins A and B find a set proteins orthologous to A and orthologous to B so that both families contain the proteins for same species
- Construct gene trees of set of A-orthologs and B-orthologs
- Compare the trees



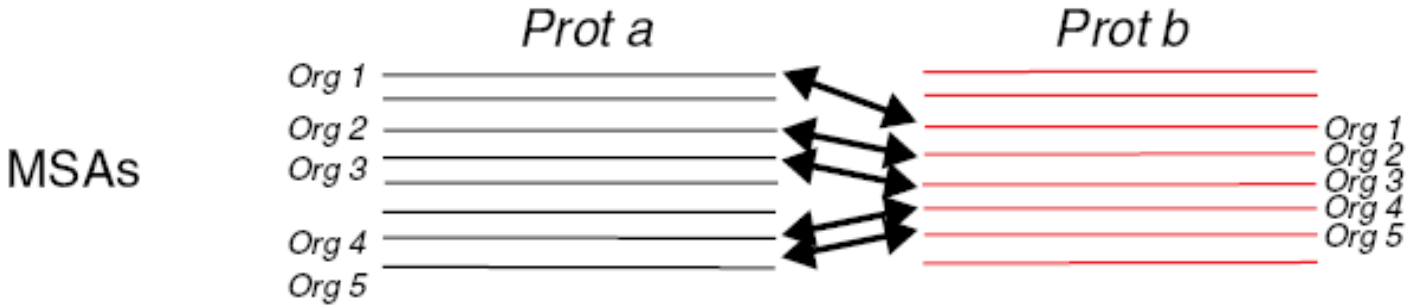
Finding orthologs

Protein Engineering vol.14 no.9 pp.609–614, 2001

Similarity of phylogenetic trees as indicator of protein–protein interaction

Florencio Pazos and Alfonso Valencia[1]

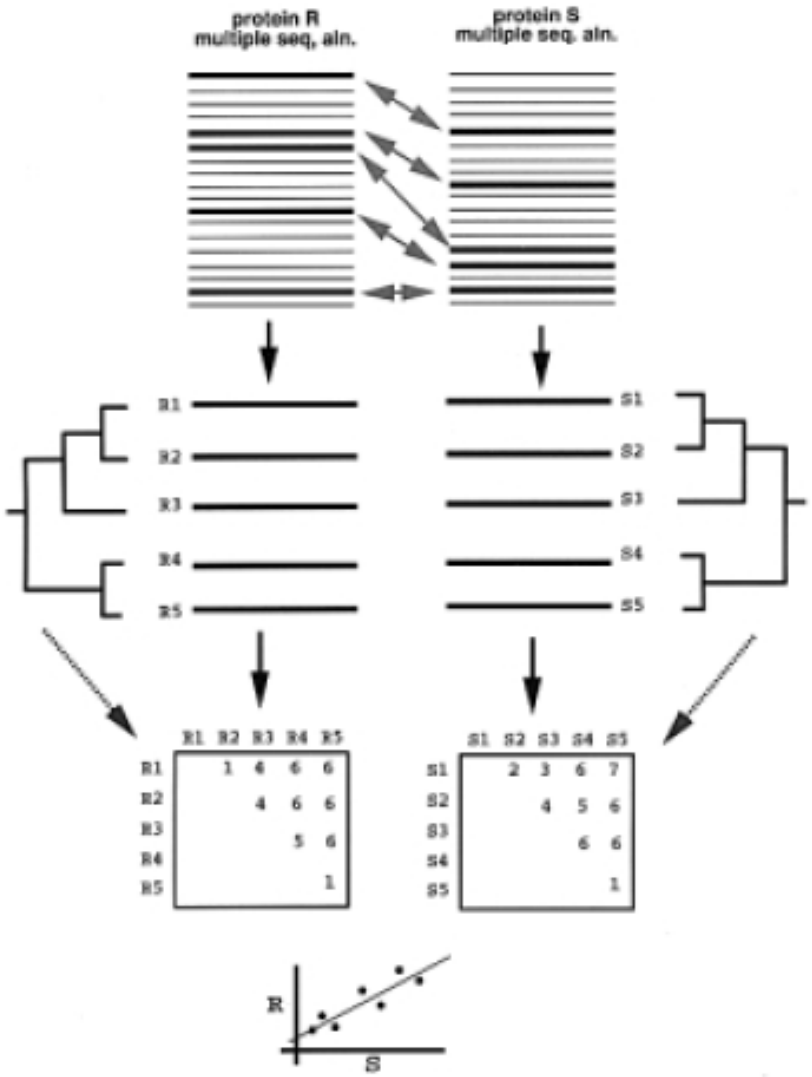
general divergence between the corresponding species under



Assumption (not always correct): The best blast hit from an organism is an ortholog.

Remark: There are other more sophisticated methods of finding orthologs.

Comparing the trees



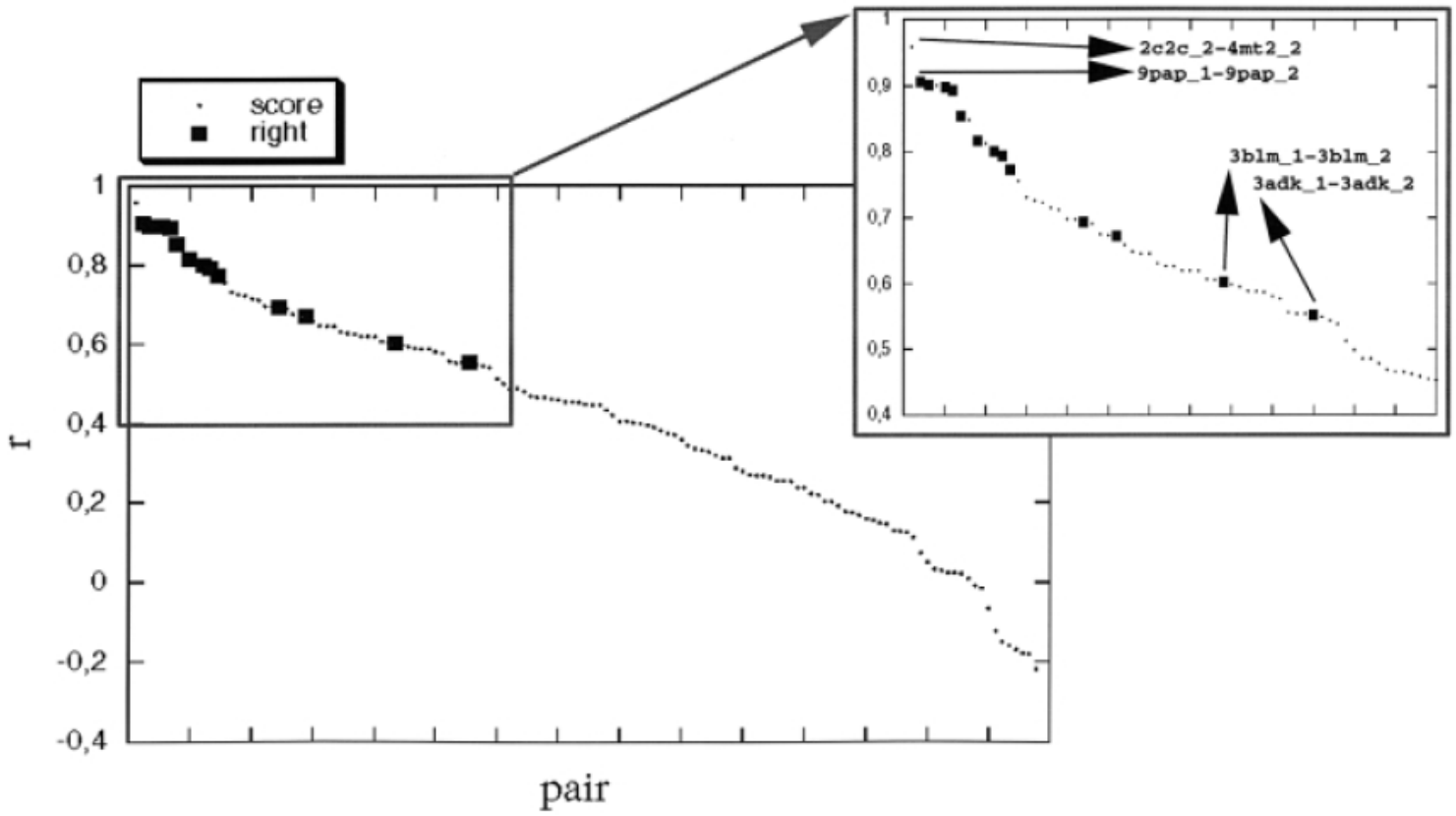
Rather than computing similarity of two trees compute correlation coefficient of the two distance matrices

$$r = \frac{\sum_{i=1}^{n} (R_i - \bar{R})\,(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n} (R_i - \bar{R})^2}\ \sqrt{\sum_{i=1}^{n} (S_i - \bar{S})^2}}$$

Results

Distance matrices for families of two interacting proteins are have high correlation coefficient

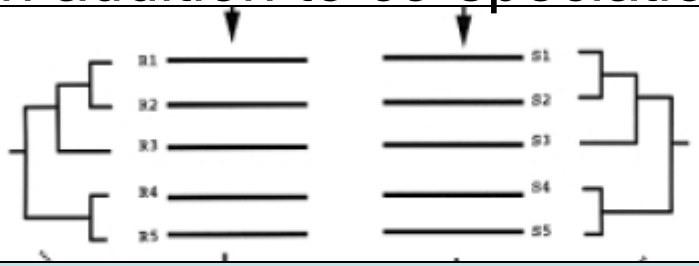
F.Pazos and A.Valencia



Sources of errors and improved method

Any pair of such “mirror trees” will be correlated because of common speciation history

We would like to consider only co-evolution that occurs in addition to co-speciation



Idea: Subtract co-speciation from co-evolution signal and what is left should be co-evolution due to common evolutionary pressure for preserving the functionality of the interacting partners

Subtracting Common Speciation

doi:10.1016/j.jmb.2005.07.005 J. Mol. Biol. (2005) 352, 1002–1015

JMB

Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

Assessing Protein Co-evolution in the Context of the Tree of Life Assists in the Prediction of the Interactome

Florencio Pazos[1]*†, Juan A. G. Ranea[2], David Juan[3] and Michael J. E. Sternberg[1]

BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 17 2005, pages 3482–3489
doi:10.1093/bioinformatics/bti564

Sequence analysis

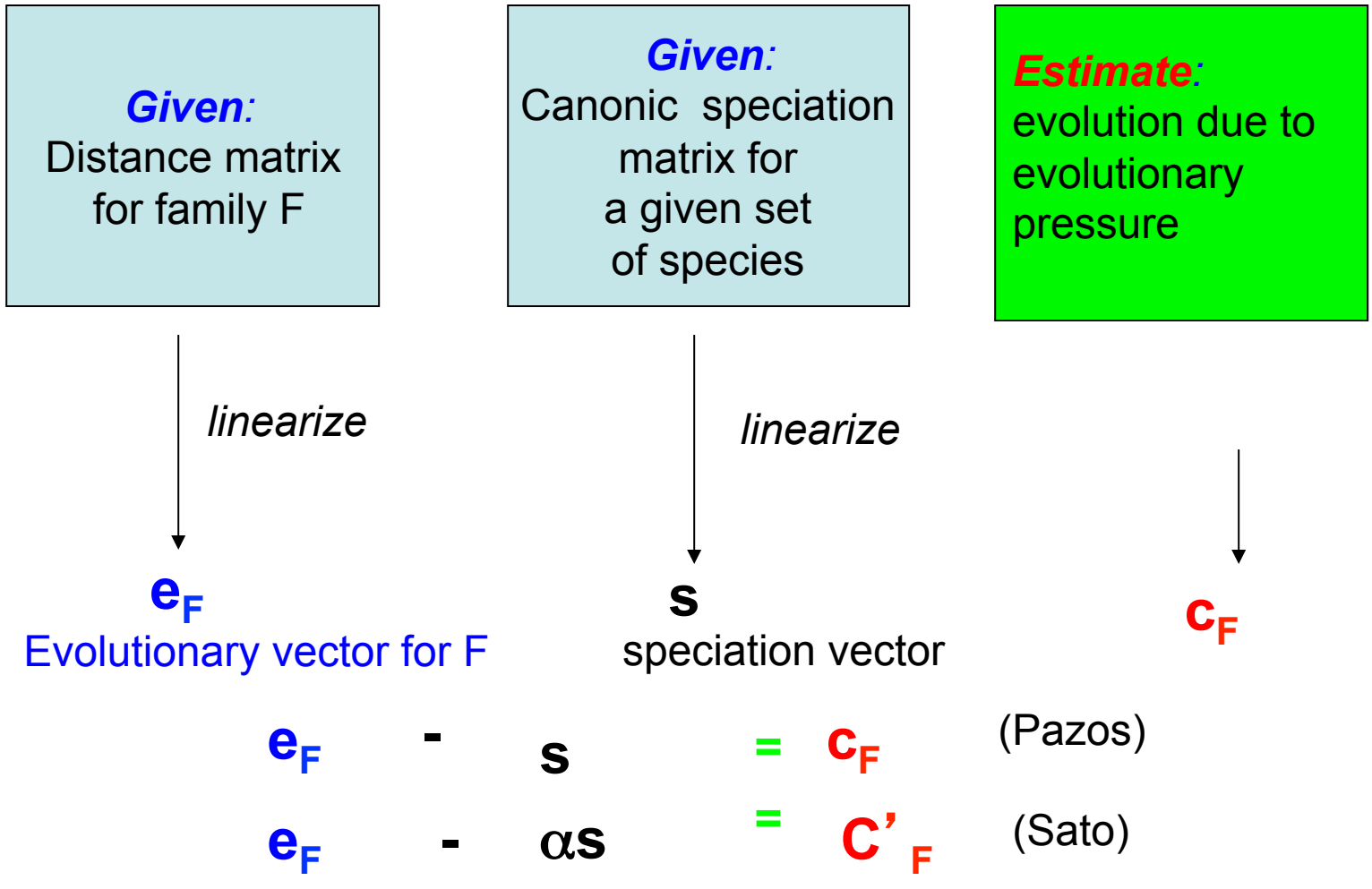
The inference of protein–protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships

Tetsuya Sato[1,*], Yoshihiro Yamanishi[2], Minoru Kanehisa[1] and Hiroyuki Toh[3]

[1]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan, [2]Centre de Géostatistique, Ecole des Mines de Paris, 35 rue Saint-Honoré, 77305 Fontainebleau cedex, France and [3]Division of Bioinformatics, Medical Institute of Bioregulation, Kyushu University, Fukuoka, Fukuoka 812-8582, Japan

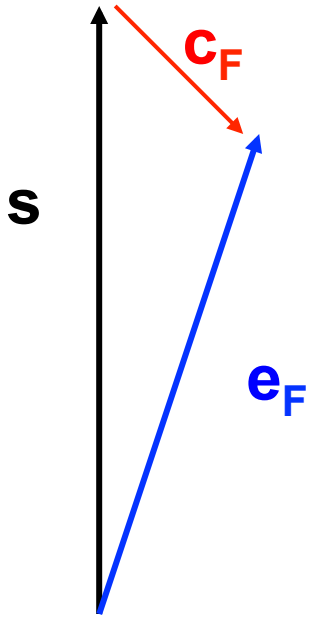
Received on April 23, 2005; revised on June 23, 2005; accepted on June 28, 2005

Evolutionary vectors



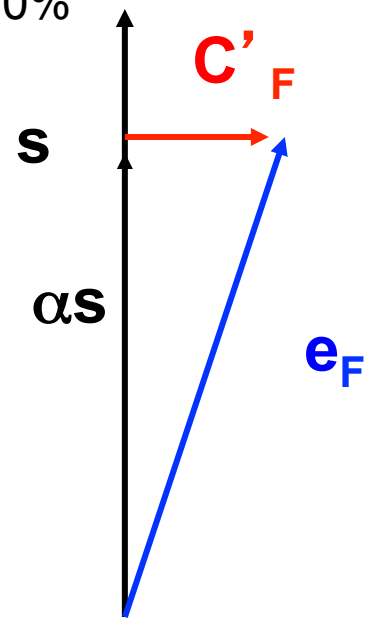
Difference between the two methods

Significant improvement 15-20%



$e_F - s = c_F$

(Pazos)
Will be able to discover
If two proteins evolve with different speed



αs = projection of e_F on s

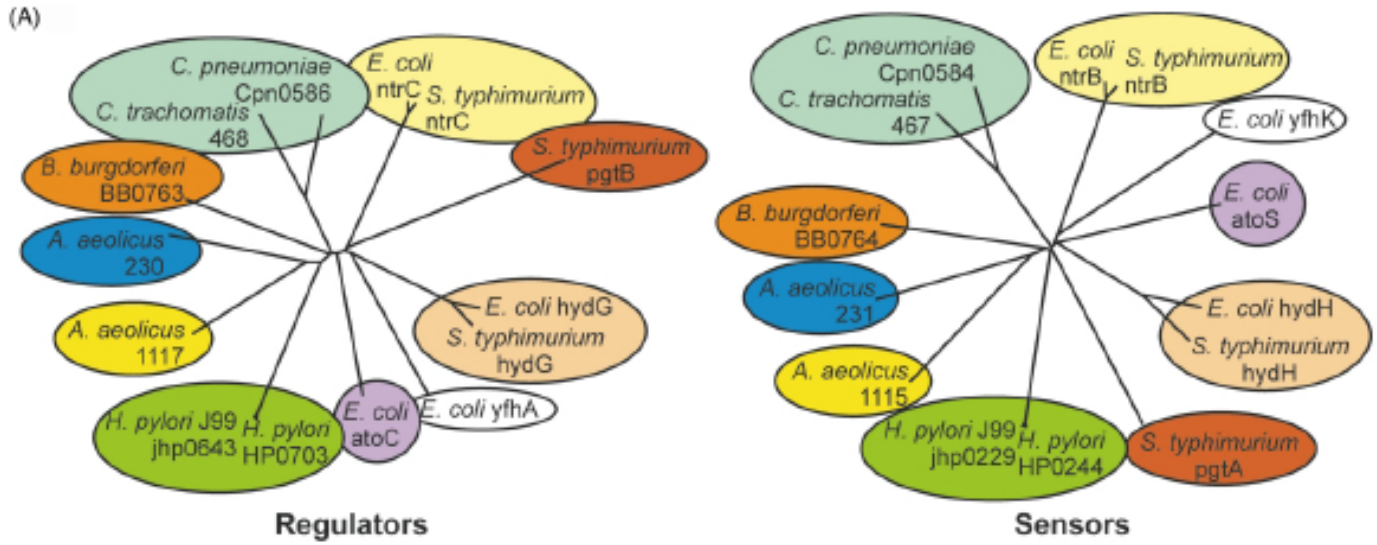
$e_F - \alpha s = C'_F$

(Sato)
Separate completely
speciation

Using Co-evolution technique for predicting specific interaction

- Previously we asked the question whether protein A and B interact, by reducing it to a a question whether family of orthologs containing A interact with family of orthologs containing B
- Now we have two families of proteins A and B, we know that each protein in A has an interacting partner in B and we try to figure out which protein from A interact with which protein form B.

Example: Which goes with which?



doi:10.1016/S0022-2836(03)00114-1 J. Mol. Biol. (2003) 327, 273–284

JMB

Available online at www.sciencedirect.com

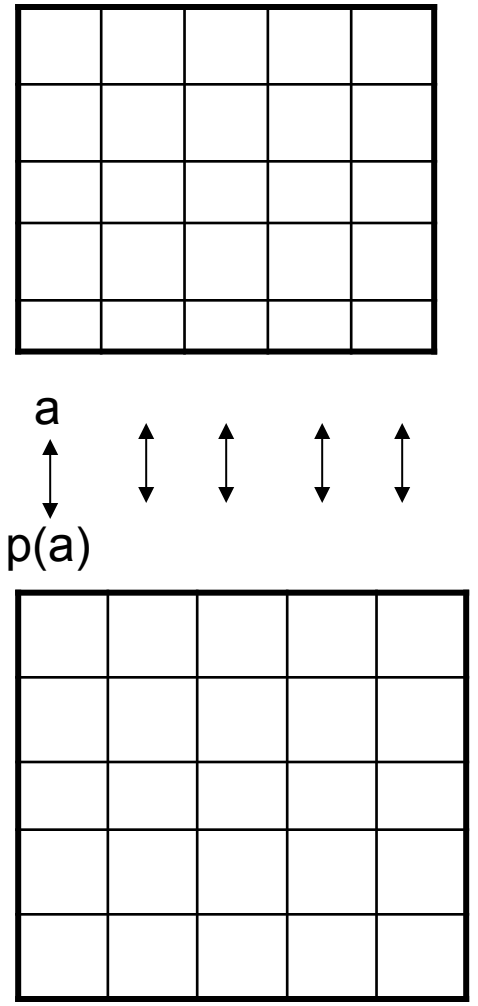
SCIENCE DIRECT®

Exploiting the Co-evolution of Interacting Proteins to Discover Interaction Specificity

Arun K. Ramani[1] and Edward M. Marcotte[1,2*]

Idea

- Find mapping between the leaves of the two tree so that if both distance matrices (the matrices used to compute the trees) are ordered so that the corresponding leaves have the same index
- And the correlation coefficient is maximized
- The mapping is found by Monte Carlo Metropolis algorithm



Column swapping method

doi:10.1016/S0022-2836(03)00114-1 J. Mol. Biol. (2003) 327, 273–284

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Exploiting the Co-evolution of Interacting Proteins to Discover Interaction Specificity

BIOINFORMATICS

Vol. 19 no. 16 2003, pages 2039–2045
DOI: 10.1093/bioinformatics/btg278

Inferring protein interactions from phylogenetic distance matrices

Jason Gertz[1], Georgiy Elfond[2], Anna Shustrova[2], Matt Weisinger[3], Matteo Pellegrini[4,], Shawn Cokus[5] and Bruce Rothschild[5]*

BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–9

Predicting Protein-Protein Interaction by Searching Evolutionary Tree AutoMORPHism Space

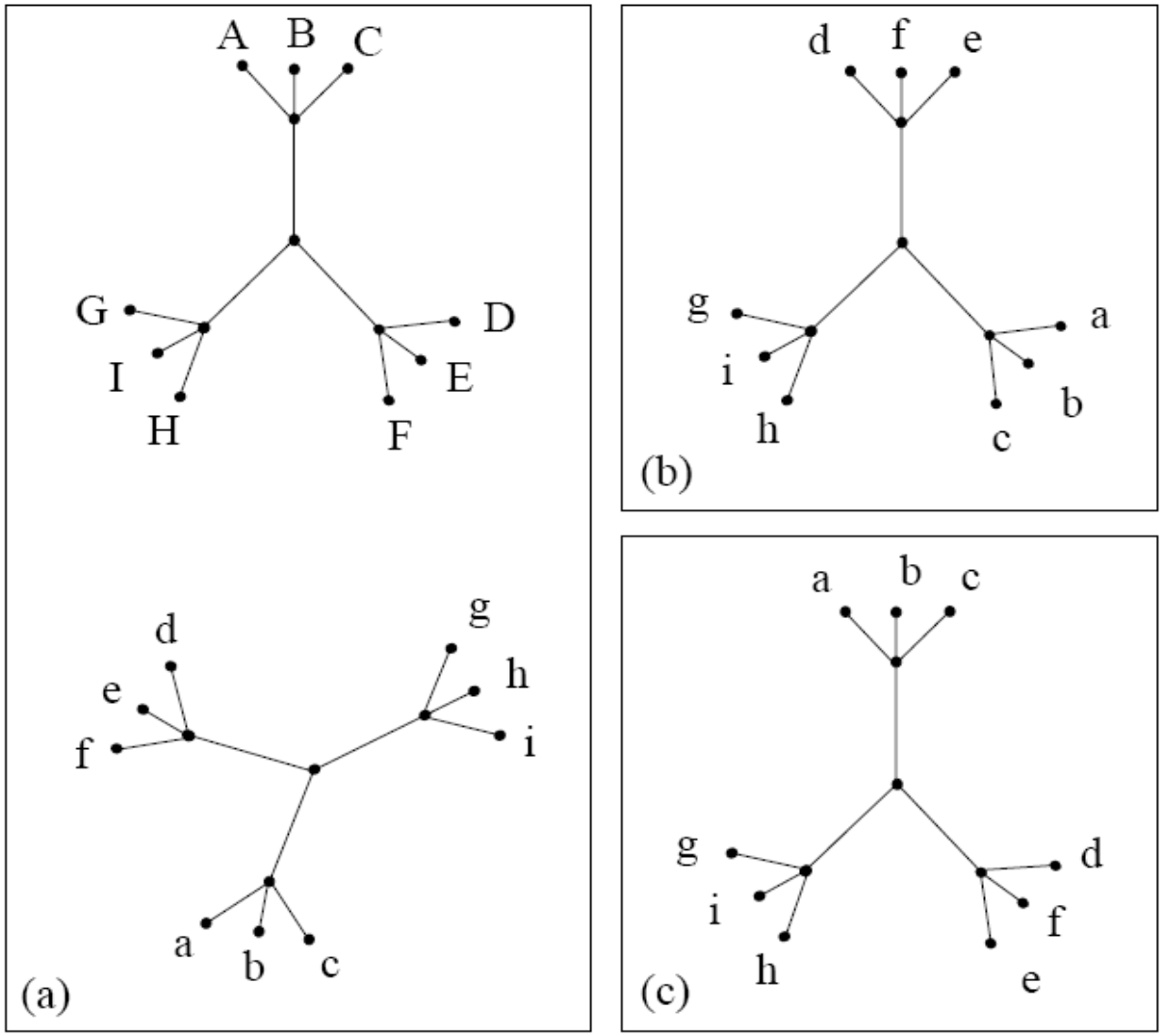
Raja Jothi, Maricel G. Kann, Teresa M. Przytycka

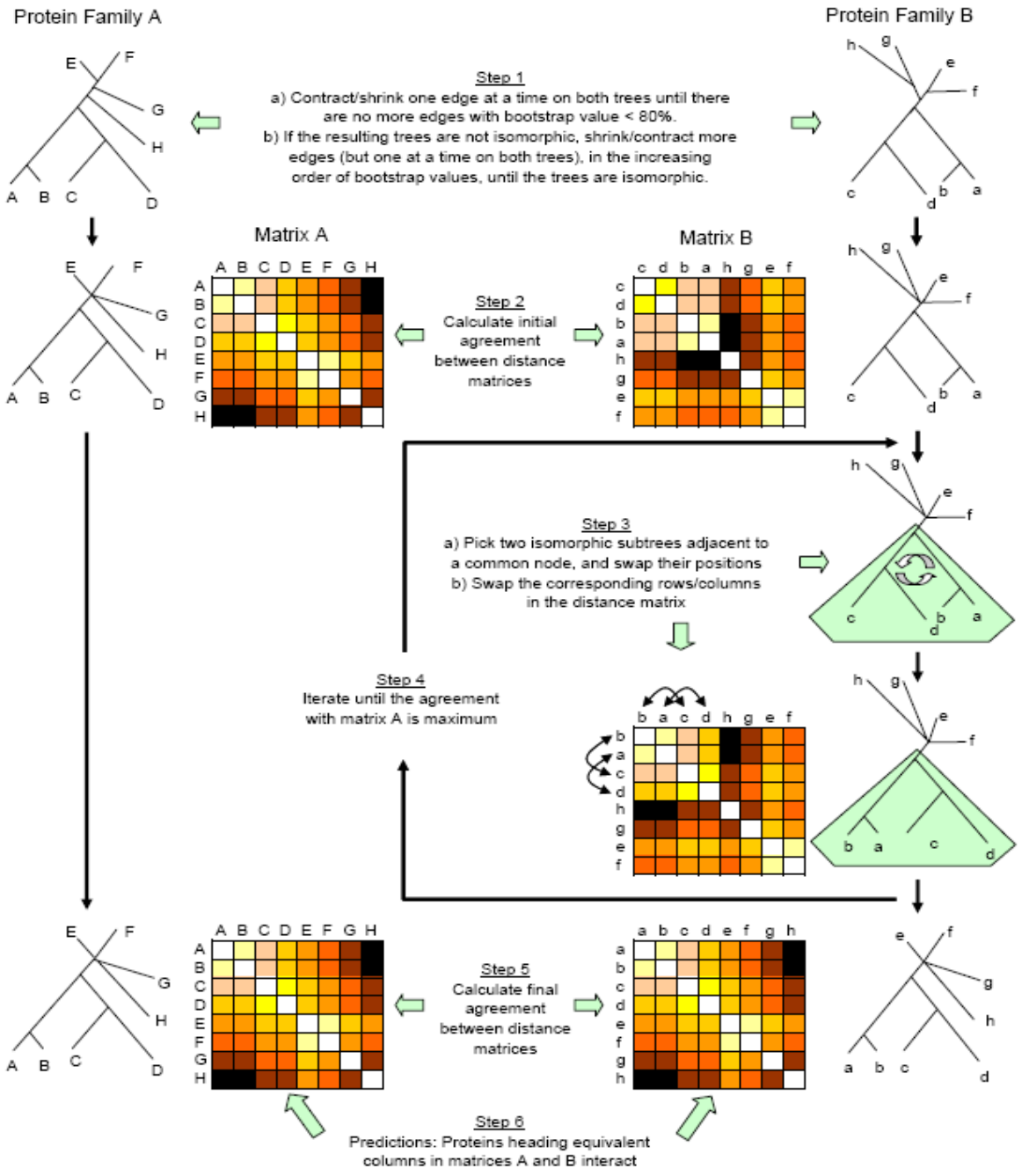
(ISMB 2005)

Metropolis Column Swapping Algorithm

- Move set – select randomly a pair of column (and corresponding rows)
- Acceptance /rejection test: test if swapping the columns increase correlation coefficient.
- Do the swap using Metropolis Criterion.

Column Flipping can get you to local optimum




Protein Family A
Protein Family B
Step 1
a) Contract/shrink one edge at a time on both trees until there are no more edges with bootstrap value < 80%.
b) If the resulting trees are not isomorphic, shrink/contract more edges (but one at a time on both trees), in the increasing order of bootstrap values, until the trees are isomorphic.
Matrix A
Matrix B
Step 2
Calculate initial agreement between distance matrices
Step 3
a) Pick two isomorphic subtrees adjacent to a common node, and swap their positions
b) Swap the corresponding rows/columns in the distance matrix
Step 4
Iterate until the agreement with matrix A is maximum
Step 5
Calculate final agreement between distance matrices
Step 6
Predictions: Proteins heading equivalent columns in matrices A and B interact

ing protein families	proteins[a]	MATRIX[†]	MORPH	MATRIX[†]	MORPH	MATRIX[†]	MORPH	IC[c]	Correct pairing[d]	MORPH[e]	Factor (%
cine/receptor–mouse/human/rat	31	48.4	**51.6**	12.9	**22.6**	112.7	44.9	67.8	0.868233	0.870028	57.1
cine/receptor–human	13	NA	NA	23.1	23.1	32.5	26.3	6.2	0.540785	0.812693	90.0
pe chemokine/receptor–mouse/human/rat	18	55.5	55.5	33.3	33.3	52.5	23.9	28.6	0.843352	0.856464	53.3
pe chemokine/receptor–mouse/human	6	100	100	33.3	33.3	9.5	5.6	3.9	0.884096	0.887253	0.00
pe regulator/sensors–*E. coli*	14	NA	NA	21.4	21.4	36.3	36.3	0.0	0.387540	0.681353	100.0
pe regulator/sensors–*B. subtilis*	13	NA	NA	7.7	7.7	32.5	32.5	0.0	0.382881	0.715536	100.0
pe regulator/sensors–5 bacteria	16	43.8	**56.3**	31.3	**56.3**	44.3	24.8	19.5	0.889139	0.915159	69.2
pe regulator/sensors–*E. coli/B. subtilis*	27	NA	NA	18.5	18.5	93.1	93.1	0.0	0.418321	0.655748	100.0
e regulator/sensors–8 bacteria	22	**36.4**	9.1	**36.4**	9.1	69.9	44.9	25.0	0.825362	0.872074	78.9
e regulator/sensors–8 bacteria	14	85.7	85.7	57.1	**71.4**	36.3	18.1	18.2	0.902306	0.907389	63.6
e regulator/sensors–*E. coli/B. subtilis*	5	100	100	100.0	100.0	6.9	3.0	3.9	0.780736	0.780736	0.00
e regulator/sensors–*E. coli/B. subtilis*	4	50	100	50.0	**100.0**	4.6	4.6	0.0	0.987242	0.987242	100.0
mponent sensor/regulators–*E. coli*	27	NA	NA	7.4	**14.8**	93.1	93.1	0.0	0.541882	0.706465	100.0
e-, and "other"-type regulator/sensors–8 bacteria	20	NA	NA	5.0	**10.0**	61.1	61.1	0.0	0.112060	0.521943	100.0
heY–11 bacteria	13	69.2	**100**	69.2	**100.0**	32.5	21.5	11.0	0.837894	0.838406	80.0
nsporter membrane protein 1/2–*E. coli*	19	NA	NA	26.3	26.3	56.8	36.7	20.1	0.589718	0.683888	87.5
nsporter memb./binding prot.–*E. coli*	17	NA	NA	0.0	**17.6**	48.3	41.3	7.0	0.625733	0.679630	92.8
nsporter membrane protein 1/2–*H. influenzae*	14	NA	NA	0.0	**28.6**	36.3	29.8	6.5	0.430913	0.768785	90.9
nsporter memb./binding prot.–*H. influenzae*	13	NA	NA	7.7*	**38.5**	35.5	32.5	3.0	0.548655	0.691065	100.0
,ParC/E–α-proteobacteria	20	100	100	50*	50.0	61.1	11.0	50.1	0.992959	0.993684	0.00
,ParC/E–Gram positive bacteria	28	100	100	17.9*	17.9	97.9	32.0	65.9	0.944774	0.947315	52.0
nteraction partners from multiple organisms											
heB–bacteria	8	NA	NA	100.0	100.0	15.3	7.6	7.7	0.962251	0.962285	60.0
CoA carboxylase α/β Gram positive bacteria	9	NA	NA	33.3	**55.5**	18.5	4.6	13.9	0.872684	0.884890	33.3
CoA carboxylase α/β proteo bacteria	16	NA	NA	75.0	75.0	44.3	37.3	7.0	0.975810	0.978088	92.3
te CoA synthetase α/β proteo bacteria	22	NA	NA	81.8	81.8	69.9	55.5	14.4	0.897055	0.897446	89.4
te CoA synthetase α/β archaea	13	NA	NA	30.8	**38.5**	32.5	13.5	19.0	0.917747	0.942711	60.0
yrB–α-proteobacteria	20	NA	NA	70*	**80.0**	61.1	36.5	24.6	0.972145	0.972901	70.5
yrB–Gram positive bacteria	18	NA	NA	50.0	**55.5**	52.5	11.6	40.9	0.981282	0.981444	40.0
yrB–archaea	10	NA	NA	20.0	20.0	21.8	16.3	5.5	0.808534	0.919128	85.7
e dehydrogenase α/β–bacteria	17	NA	NA	52.9	**82.4**	48.3	27.4	20.9	0.953532	0.961960	78.5
rE–bacteria	26	NA	NA	**61.5**	23.1	88.4	42.5	45.9	0.972655	0.696493	73.9
rE–α-proteobacteria	12	NA	NA	66.6	**83.3**	28.8	7.6	21.2	0.992246	0.992580	11.1
rE–Gram positive bacteria	14	NA	NA	57.1	57.1	36.3	8.0	28.3	0.968444	0.972985	27.2
olymerase III E2/E3–bacteria	20	NA	NA	45.0	**65.0**	61.1	19.3	41.8	0.939153	0.951563	52.9

r of proteins in a family of interacting proteins (number of columns in the corresponding similarity matrix)
rch space)
tion content
tion coefficient for correct pairing of interaction partners
tion coefficient of the maximal agreement of similarity matrices found by MORPH
tage of internal edges in the phylogenetic tree that were shrunk to reach isomorphism
interaction prediction algorithm proposed by Ramani and Marcotte (2003).* Results in Ramani and Marcotte (2003) could not be reproduced using their MATRIX web-server

MORPH's performance at predicting protein interaction partners between two protein families that are known to interact

Predicting domain-domain interactions from PPI network

- Most proteins contain more than one domain
- Protein-protein interaction is mediated by domain-domain interaction for one or more domain pairs
- High throughput experiments can discover interaction on protein-protein level. Can we deduct from it domain-domain interactions?

Protein-Protein Interaction from network alignment

- Given: Interaction networks from three organisms (yeast, fruit fly, worm)
- Idea: Construct alignment graph:
  - Nodes – triples of sequentially similar proteins (each from one organism)
  - Edges – conserved protein interactions

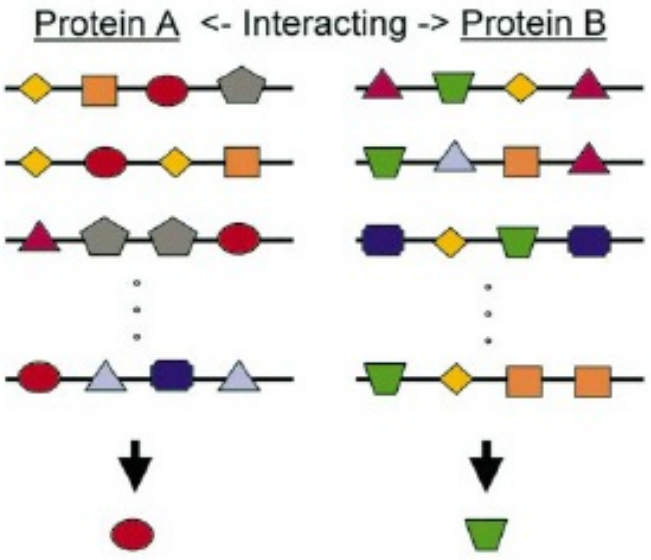
Association method

doi:10.1006/jmbi.2001.4920 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2001) 311, 681–692

JMB

Correlated Sequence-signatures as Markers of Protein-Protein Interaction

Einat Sprinzak and Hanah Margalit*



Idea: probability p(Int(A,B)) that domains A and B interact is approximated by:

interacting protein pairs where one contains A and the other B
---
possible protein pairs where one contains A and the other B

Probability that two proteins X,Y interact is
1-probability_they_don't_itneract =
$1 - \Pi_{\text{all domain pairs A, B where A in X and B in Y}}(1\text{-Int}(A,B))$

Expectation Maximization

Letter

Inferring Domain–Domain Interactions From Protein–Protein Interactions

Minghua Deng, Shipra Mehta, Fengzhu Sun,[1,2] Ting Chen[1,3]

Program in Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles, California 90089, USA

Idea:

Assume each domain pair has some interaction probability.

Use Expectation Maximization to estimate the probabilities that maximize the likelihood of the observed protein-protein interaction network.

Setting

Let $D_1, \ldots, D_M$ denote the M domains, and $P_1, \ldots, P_N$ denote the N proteins. Let P_{ij} denote the protein pair of P_i and P_j, and D_{ij} denote the domain pair of D_i and D_j. Let P_{ij} be the set of domain pairs formed by proteins P_i and P_j. For example, assume that protein P_1 contains domains $\{D_1, D_2, D_3\}$ and protein P_2 contains domains $\{D_1, D_4\}$. Then $P_{12} = \{D_{11}, D_{12}, D_{13}, D_{14}, D_{24}, D_{34}\}$.

Setting:

Observation ij is 1 (interaction)

$$\Pr(P_{ij} = 1) = 1.0 - \prod_{D_{mn} \in P_{ij}} (1 - \lambda_{mn}), \quad (1)$$

in which $\lambda_{mn} = \Pr(D_{mn} = 1)$ denotes the probability that domain D_m interacts with domain D_n.

Probability of false positive and false negative

$$fp = \Pr(O_{ij} = 1 \mid P_{ij} = 0),$$
$$fn = \Pr(O_{ij} = 0 \mid P_{ij} = 1).$$

Thus, the probability for the observed protein–protein interaction is

$$\begin{aligned} &\Pr(O_{ij} = 1) \quad (2)\\ &= \Pr(O_{ij} = 1, P_{ij} = 1) + \Pr(O_{ij} = 1, P_{ij} = 0)\\ &= \Pr(O_{ij} = 1 \mid P_{ij} = 1)\Pr(P_{ij} = 1)\\ &\quad + \Pr(O_{ij} = 1 \mid P_{ij} = 0)(1 - \Pr(P_{ij} = 1))\\ &= \Pr(P_{ij} = 1)(1 - fn) + (1 - \Pr(P_{ij} = 1))fp. \end{aligned}$$

The likelihood function, i.e., the probability of the observed whole proteome interaction data is

$$L = \prod (\Pr(O_{ij} = 1))^{O_{ij}} (1 - \Pr(O_{ij} = 1))^{1-O_{ij}} \qquad (3)$$

where

$$O_{ij} = \begin{cases} 1 & \text{if the interaction of } P_i \text{ and } P_j \text{ is observed,} \\ 0 & \text{otherwise} \end{cases}$$

The likelihood L is a function of $\theta = (\lambda_{mn}, fp, fn)$. In the following, we fix fp and fn.

Theta is then estimated using **Expectation Maximization approach:**

- Start with some imitation assumption about lambda
- Compute Expectation of the data given observation

$$E(D_{mn}^{(ij)} \mid O_{kl} = o_{kl}, \forall\, k, l, \theta^{(t-1)})$$

- Estimate lambdas

$$\lambda_{mn}^{(t)} = \frac{1}{N_{mn}} \sum_{i \in A_m, j \in A_n} E(D_{mn}^{(ij)} | O_{kl} = o_{kl}, \forall\, k, l, \theta^{(t-1)})$$

- Iterate last two steps

DPEA-Domain-Pair exclusion

Method

Open Access

Inferring protein domain interactions from databases of interacting proteins

Robert Riley*, Christopher Lee†, Chiara Sabatti* and David Eisenberg*†‡

Addresses: *Department of Human Genetics, David Geffen School of Medicine at UCLA, University of California Los Angeles, Los Angeles, CA 90095, USA. †Institute for Genomics and Proteomics, University of California Los Angeles, Los Angeles, CA 90095, USA. ‡Howard Hughes Medical Institute, University of California Los Angeles, Los Angeles, CA 90095-1570, USA.

Correspondence: David Eisenberg. E-mail: david@mbi.ucla.edu

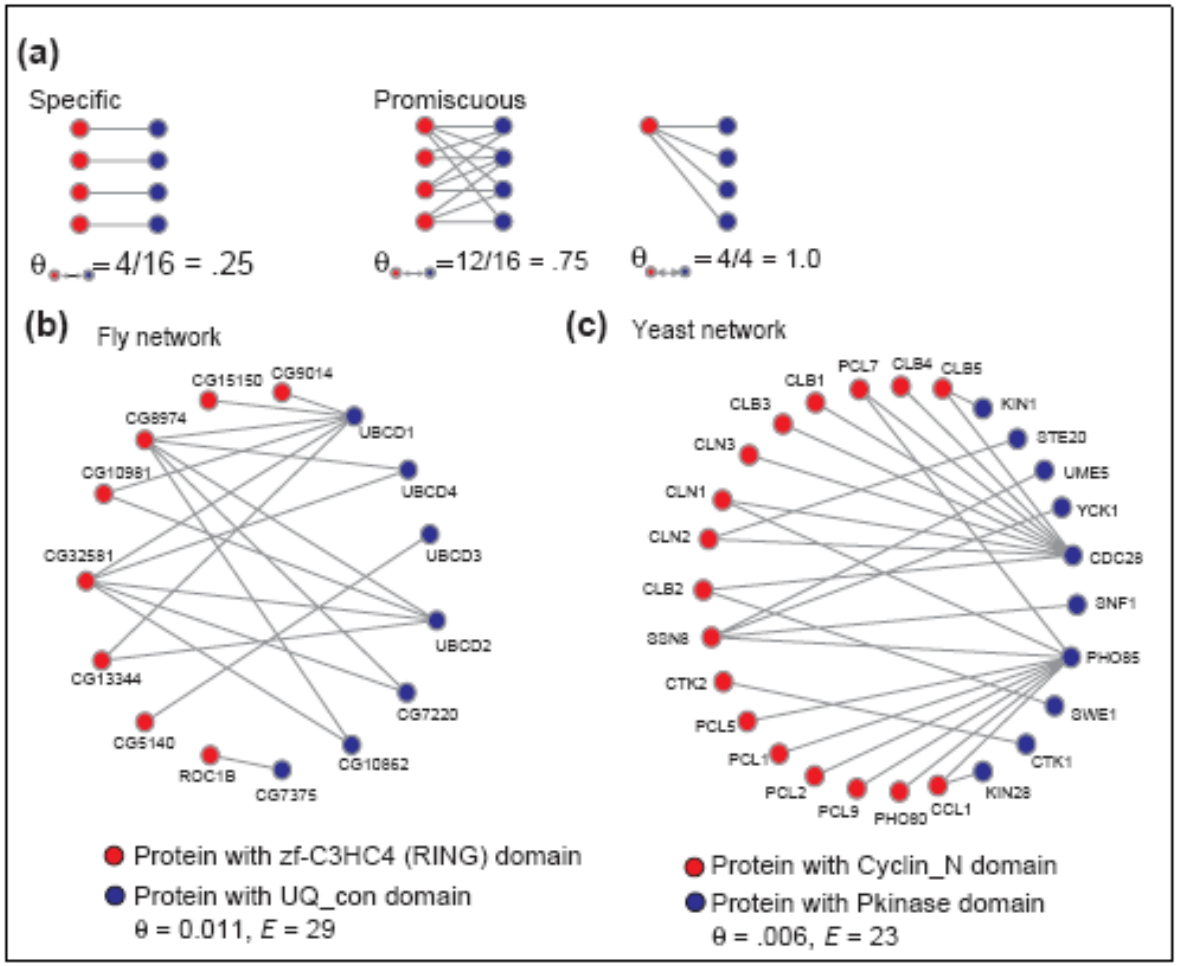
Published: 19 September 2005

Received: 15 April 2005

Problems with association and EM methods:

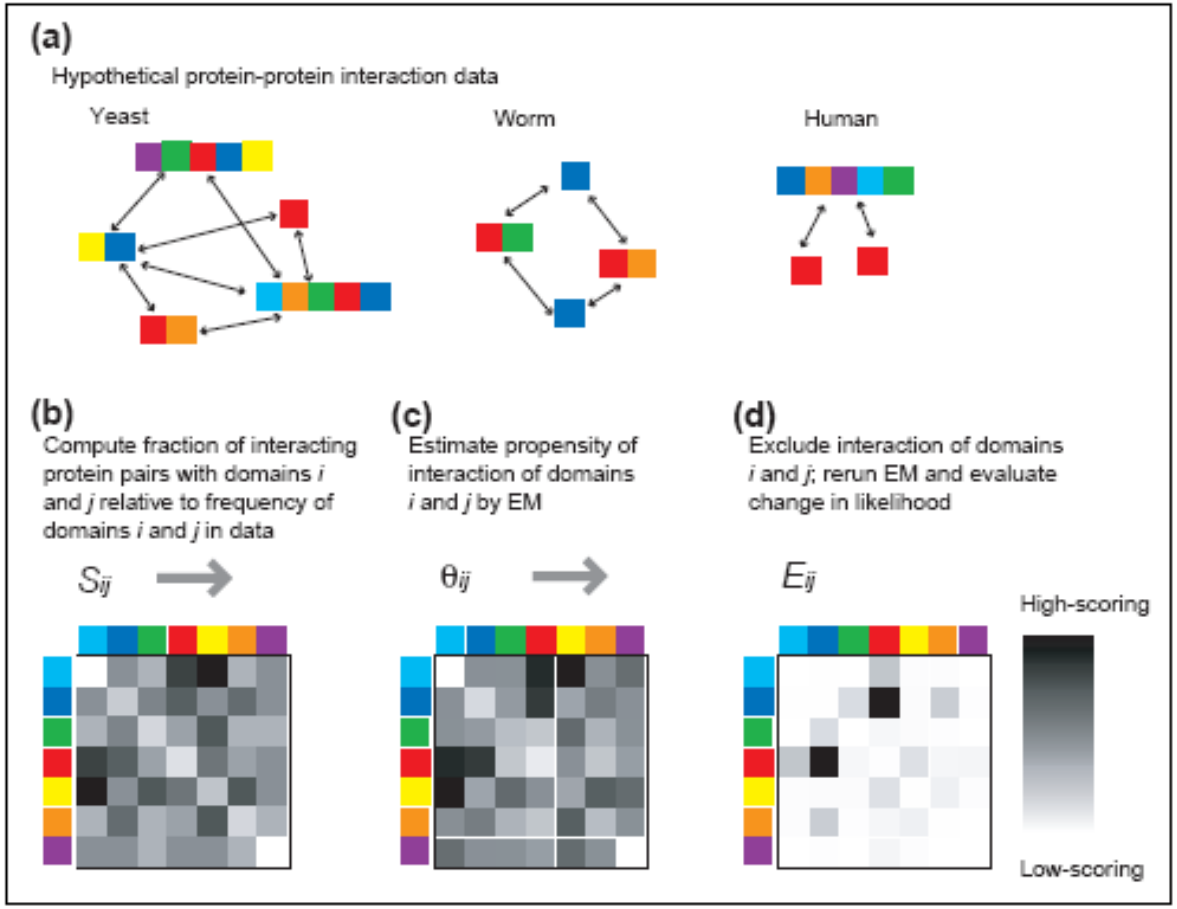
Many domain-domain interactions are highly specific; that is the same domain pair may interact in one context but not in anther.

Difficult examples for Association and EM but not for DPEA



Idea behind DPEA method

For every potential domain-domain interaction run the expectation maximization approach for under assumption that the given domain-domain interaction can occur and under the assumption that it cannot. If the expectation drops significantly, it means that given domain-domain interaction was necessary in explaining the network.,


(a)
Hypothetical protein-protein interaction data
Yeast
Worm
Human
(b)
Compute fraction of interacting protein pairs with domains i and j relative to frequency of domains i and j in data
(c)
Estimate propensity of interaction of domains i and j by EM
(d)
Exclude interaction of domains i and j; rerun EM and evaluate change in likelihood
S_{ij}
θ_{ij}
E_{ij}
High-scoring
Low-scoring

Comparing Protein Interaction Networks

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‡¶

*Computer Science Division, University of California, and International Computer Science Institute, 1947 Center Street, Berkeley, CA 94704; ‡Department of Bioengineering, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093; and §Institute of Genetics, Research Center Karlsruhe, Postfach 3640, D-76021 Karlsruhe, Germany

Contributed by Richard M. Karp, December 22, 2004

To elucidate cellular machinery on a global scale, we performed a multiple comparison of the recently available protein–protein interaction networks of *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Saccharomyces cerevisiae*. This comparison integrated protein interaction and sequence information to reveal 71 network regions that were conserved across all three species and many exclusive to the metazoans. We used this conservation, and found statistically significant support for 4,645 previously undescribed protein functions and 2,609 previously undescribed protein interactions. We tested 60 interaction predictions for yeast by two-hybrid analysis, confirming approximately half of these. Significantly, many of the predicted functions and interactions would not have been identified from sequence similarity alone, demonstrating that network comparisons provide essential biological information beyond what is gleaned from the genome.

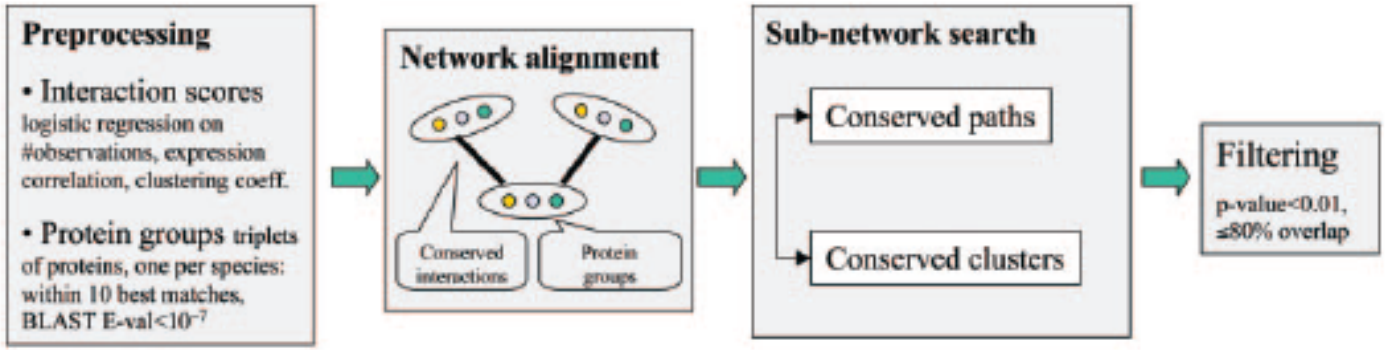
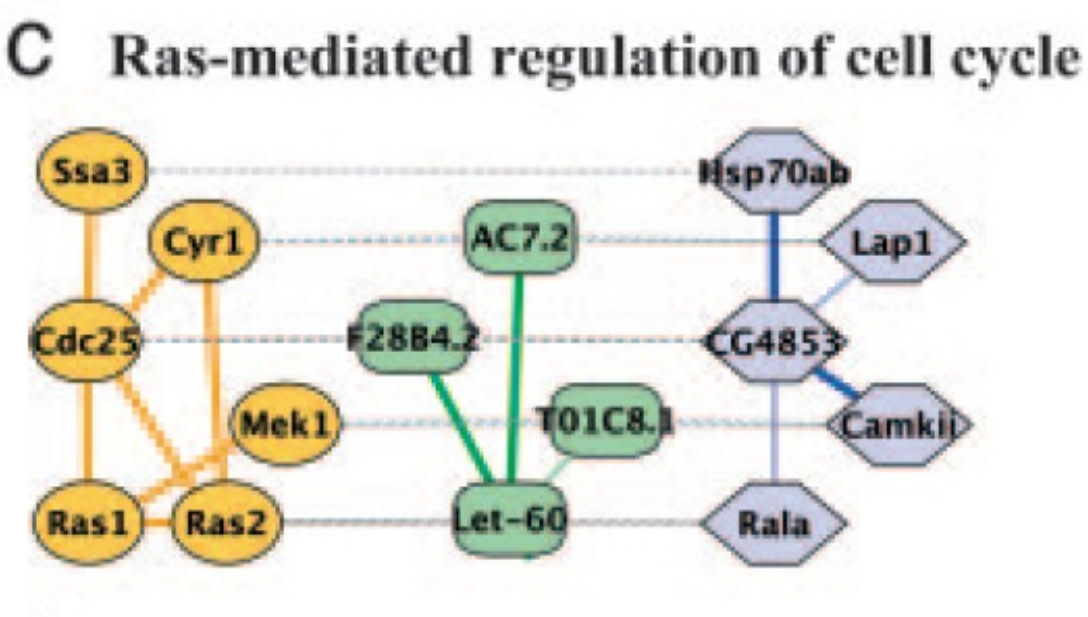


Fig. 1. Schematic of the multiple network comparison pipeline. Raw data are preprocessed to estimate the reliability of the available protein interactions and identify groups of sequence-similar proteins. A protein group contains one protein from each species and requires that each protein has a significant 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.

Example of an alignment



Prediction of interaction:

based on sequence similarity

occurrence within the same conserved cluster