计算系统生物学

王 勇

中国科学院数学与系统科学研究院

Aligning Molecular Biological Networks across various species

Yong Wang

Academy of Mathematics & Systems Science

Questions?

- Molecular networks are of current interest. Previous analyses have focused on topologic structures of individual networks.
- Different biological networks by their molecular types, species organisms, or tissues, under varying conditions.
- We should take a comparative approach toward interpreting these networks.

Sequence alignment--→Network Alignment

- Sequence alignment seeks to identify conserved DNA or protein sequence
  - Intuition: conservation implies functionality

–	**EFTPPVQAAYQKVVAGV**	(human)
–	**DFNPNVQAAFQKVVAGV**	(pig)
–	**EFTPPVQAAYQKVVAGV**	(rabbit)

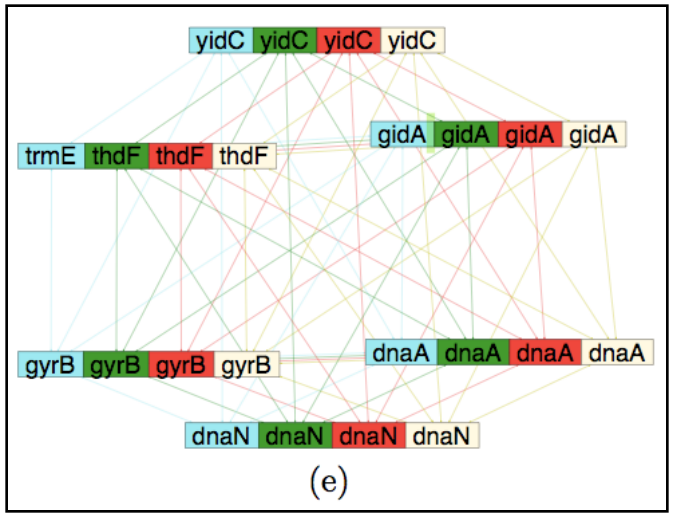
Network comparison Locally

Table 1 Modes of network comparison

Mode	Common application	Main goals	Some current limitations
Alignment	At least two networks of the same type across species	Identification of functional (conserved) protein modules; study of network evolution; interaction prediction	Limited to few (five or fewer) species; nonevolution-based scores
Integration	At least two networks of different types for the same species	Identification of modules (supported by several networks); study of interrelations between data types; interaction prediction	No agreed-upon way to combine scores over different networks
Querying	Subnetwork module versus a network	Identification of duplicated/conserved instances of the module; knowledge transfer	Query is limited to a tree topology; nonevolution-based scores

Sharan, R., and Ideker, T. Modeling cellular machinery through biological network comparison. *Nature Biotechnology.* Review. **24(4)**:427-33. Apr (2006).

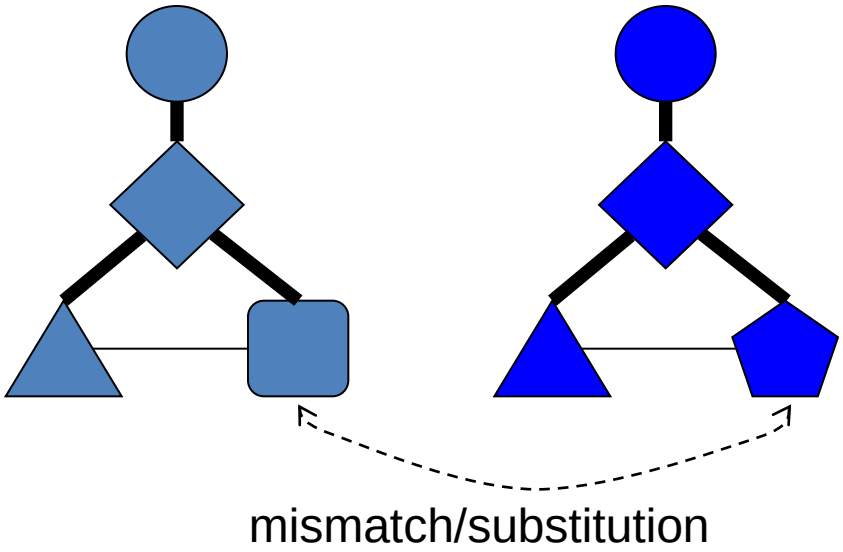
Motivation



- By similar intuition, subnetworks conserved across species are likely functional modules
- Conserved linear paths may correspond to signaling pathways, and conserved clusters of interactions may be indicative of protein complexes.
- When the two networks being compared represent linear chains of interactions, the network alignment problem admits efficient algorithmic solutions.

Network Alignment

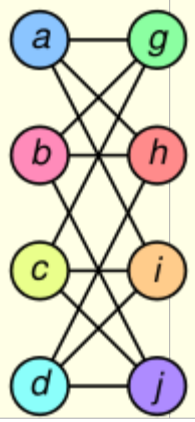
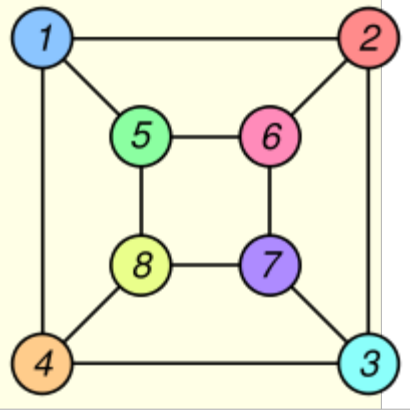
- "Conserved" means two subgraphs contain proteins serving **similar** functions, having **similar** interaction profiles
  - Key word is similar, not identical



SubGraph isomorphism

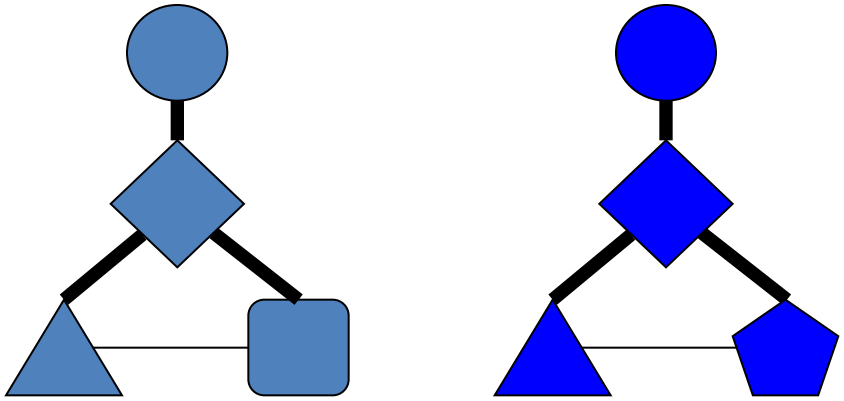
In graph theory, a graph isomorphism is a bijection (a one-to-one and onto mapping) between the vertices of two graphs *G* and *H*， *f:V(G)→V(H)*, with the property that any two vertices *u* and *v* from *G* are adjacent if and only if *f(u)* and *f(v)* are adjacent in *H*.

- The subgraph isomorphism problem, is known to be NP-complete.

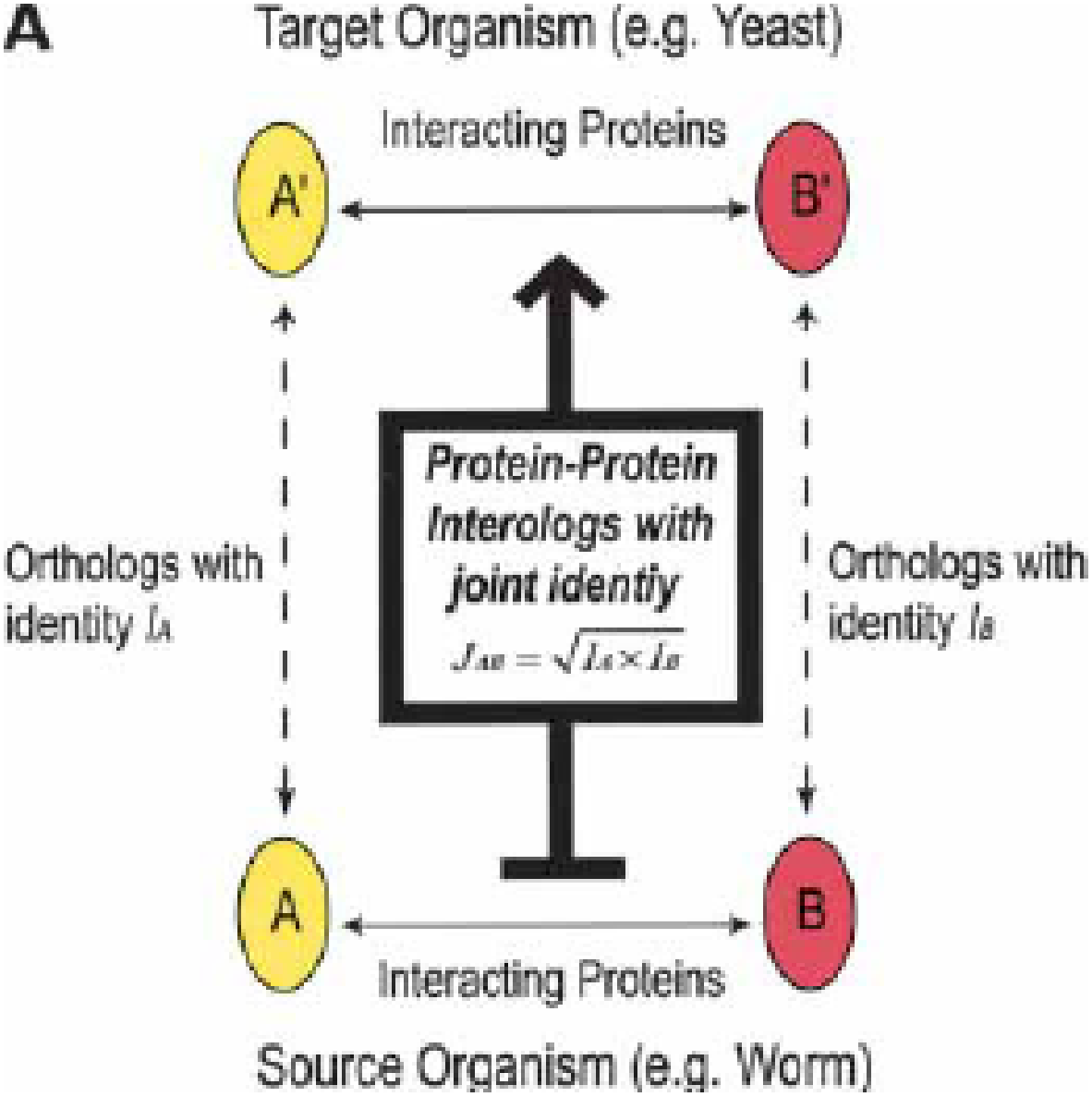
Graph G	Graph H	An isomorphism between G and H
		*f*(*a*) = 1 f(b) = 6 f(c) = 8 f(d) = 3 f(g) = 5 f(h) = 2 f(i) = 4 f(j) = 7

The simplest case: interologs

- **Inter**actions conserved in ortho**logs**
  - Orthology is a fuzzy notion
  - Sequence similarity not necessary for conservation of function



Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. H Yu, NM Luscombe, HX Lu, X Zhu, Y Xia, JD Han, N Bertin, S Chung, M Vidal, M Gerstein (2004) *Genome Res* 14: 1107-18.


A
Target Organism (e.g. Yeast)
Interacting Proteins
A'
B'
Protein-Protein
Interologs with
joint identiy
$J_{AB} = \sqrt{I_A \times I_B}$
Orthologs with
identity I_A
Orthologs with
identity I_B
A
B
Interacting Proteins
Source Organism (e.g. Worm)

Network Alignment framework

- In general, the problem is computationally hard (generalizing subgraph isomorphism under certain formulations), but heuristic approaches have been devised for it.
- A merged representation of the two networks is created, called a network alignment graph. In a network alignment graph, the nodes represent sets of molecules, one from each network, and the links represent conserved molecular interactions across the different networks.
- A greedy algorithm is applied for identifying the conserved subnetworks embedded in the merged representation.

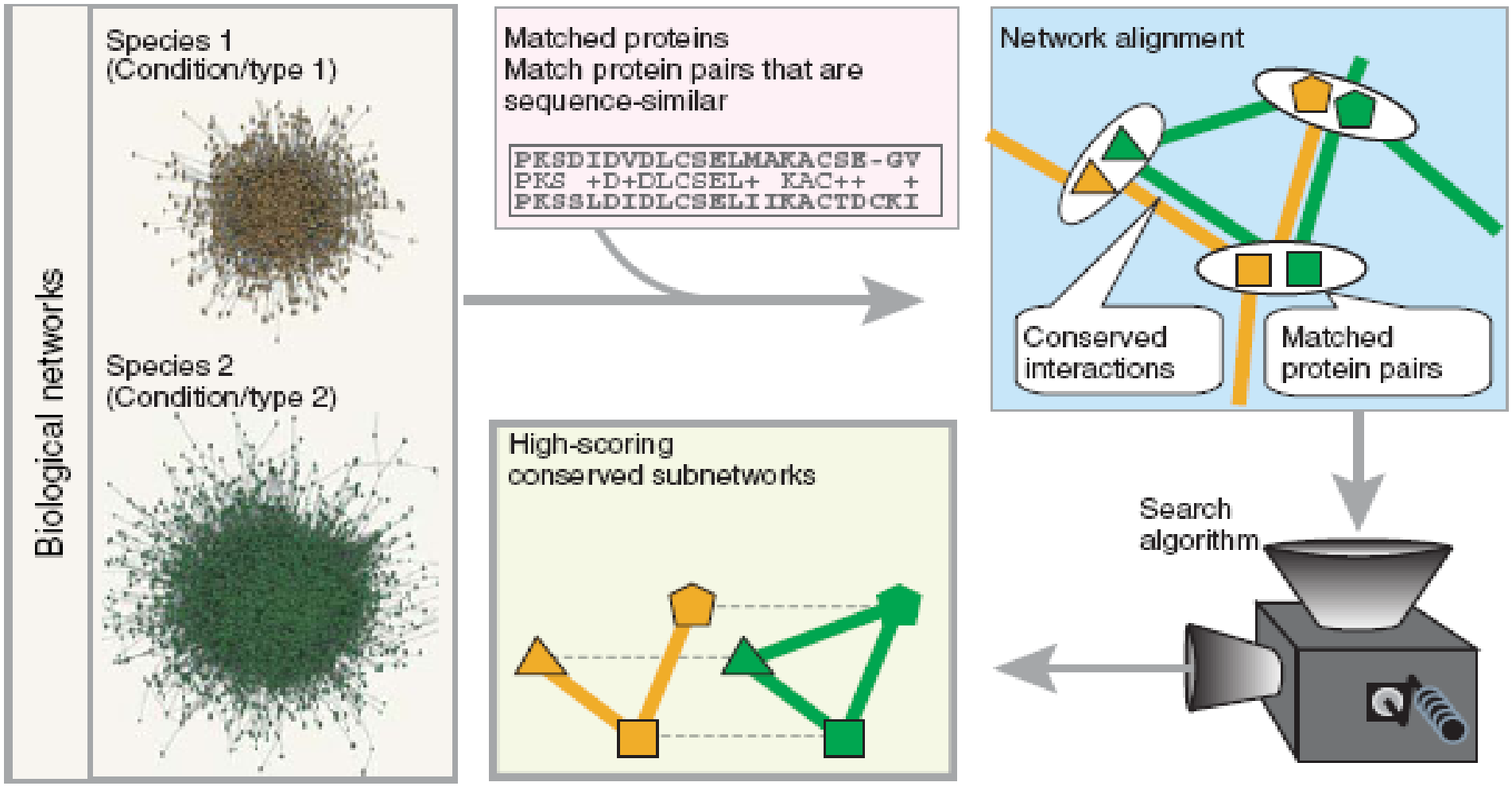
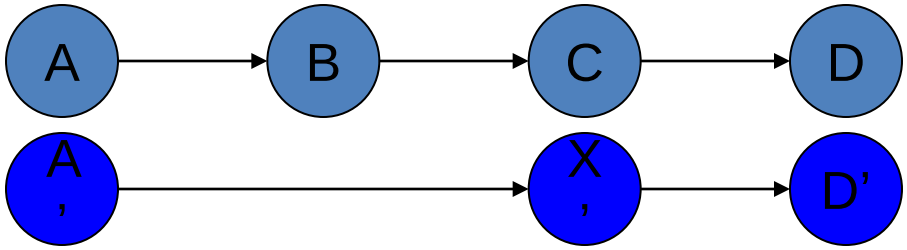


Figure 1 Network alignment. Network alignment combines protein interaction data that are available for each of at least two species with orthology information based on the corresponding protein sequences. A detailed probabilistic model is used to identify protein subnetworks within the aligned network 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.

Earlier approaches: PathBLAST

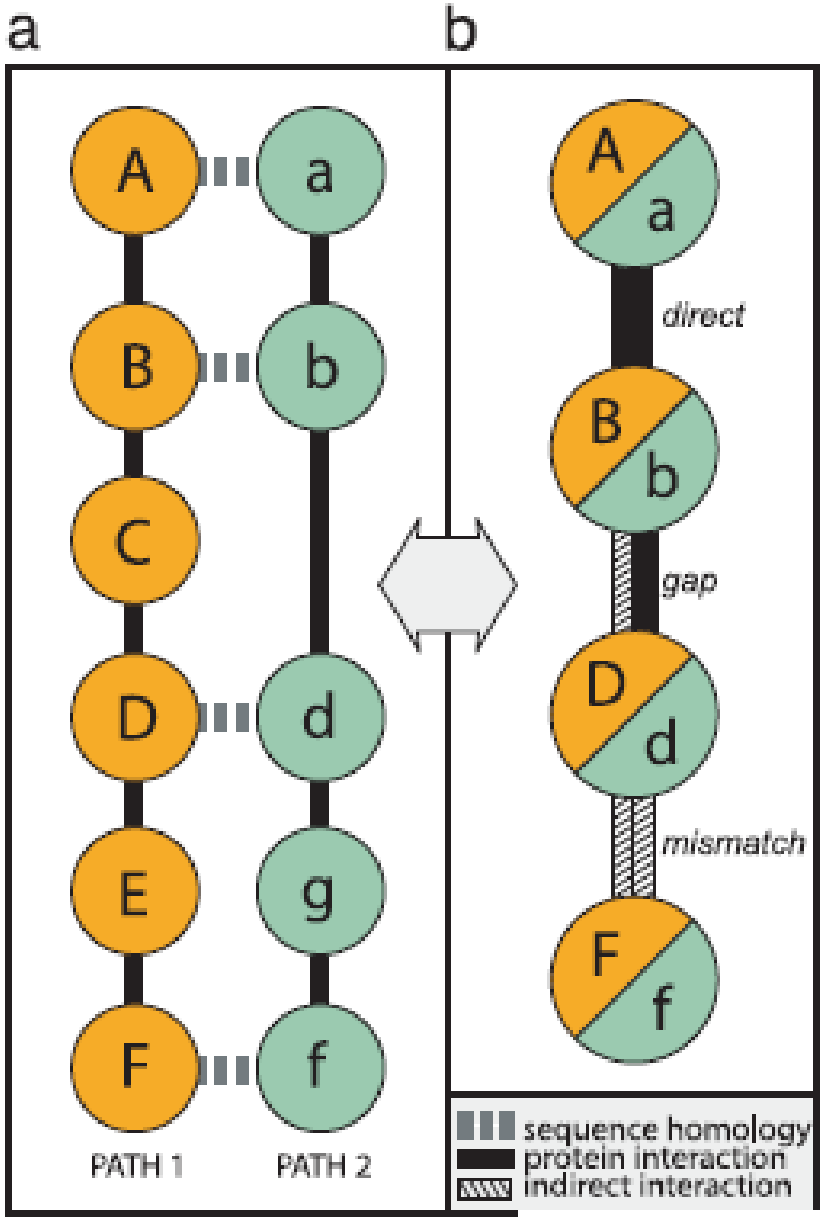
- Goal: identify conserved *pathways* (chains)
- Idea: can be done efficiently by dynamic programming if networks are DAGs



Score: match + gap + mismatch + match

Kelley, B. P., Sharan, R., Karp, R., Sittler, E. T., Root, D. E., Stockwell, B. R., and Ideker, T. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc Natl Acad Sci U S A 100, 11394-9 (2003).

Kelley, B. P., Yuan, B., Lewitter, F., Sharan, R. Stockwell, B. R., Ideker, T. PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Research* **1;32:** W83-8 (2004).

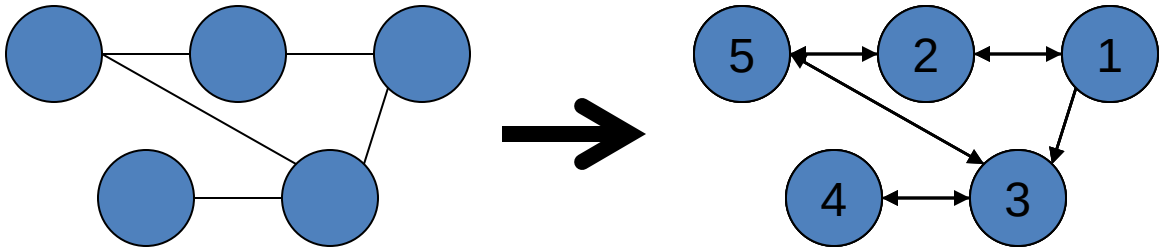


Comment: One of the drawbacks of the alignment graph is that it includes a node for every pair (or triplet) of similar proteins (one from each input network). The commonly used similarity functions (e.g. BLAST E-value threshold) generally impose a many-to-many correspondence between proteins, which causes the size of the alignment graph to grow exponentially with the number of aligned networks.

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

Earlier approaches: PathBLAST

- Problem: Networks are neither acyclic nor directed
- Solution: eliminate cycles by imposing random ordering on nodes, perform DP; repeat many times

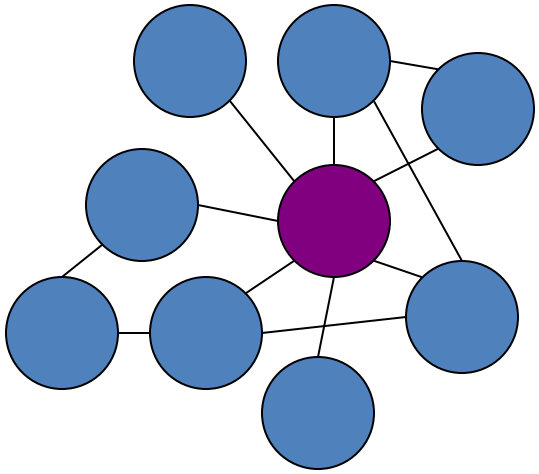


- In expectation, finds conserved paths of length *L* within networks of size *n* in $O(L!n)$ time
- Drawbacks
  - Computationally expensive
  - Restricts search to specific topology

Kelley et al (2003)

Earlier approaches: MaWISh

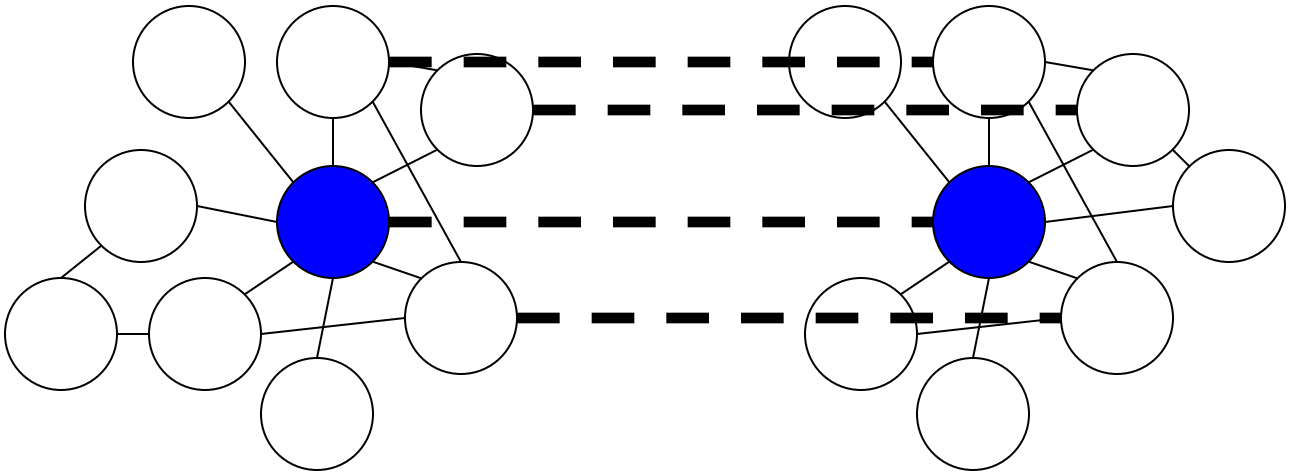
- Goal: identify conserved *multi-protein complexes* (clique-like structures)
- Idea: such structures will likely contain at least one *hub* (high-degree node)



Koyuturk, M., Grama, A. & Szpankowski, W. in Proceedings of the Ninth Annual International Conference on Research in Computational Molecular Biology (RECOMB) 48–65 (2005).

Earlier approaches: MaWISh

- Algorithm: start by aligning a pair of homologous hubs, extend greedily



Efficient running time, but also only solves a specific case

Koyuturk et al (2004)

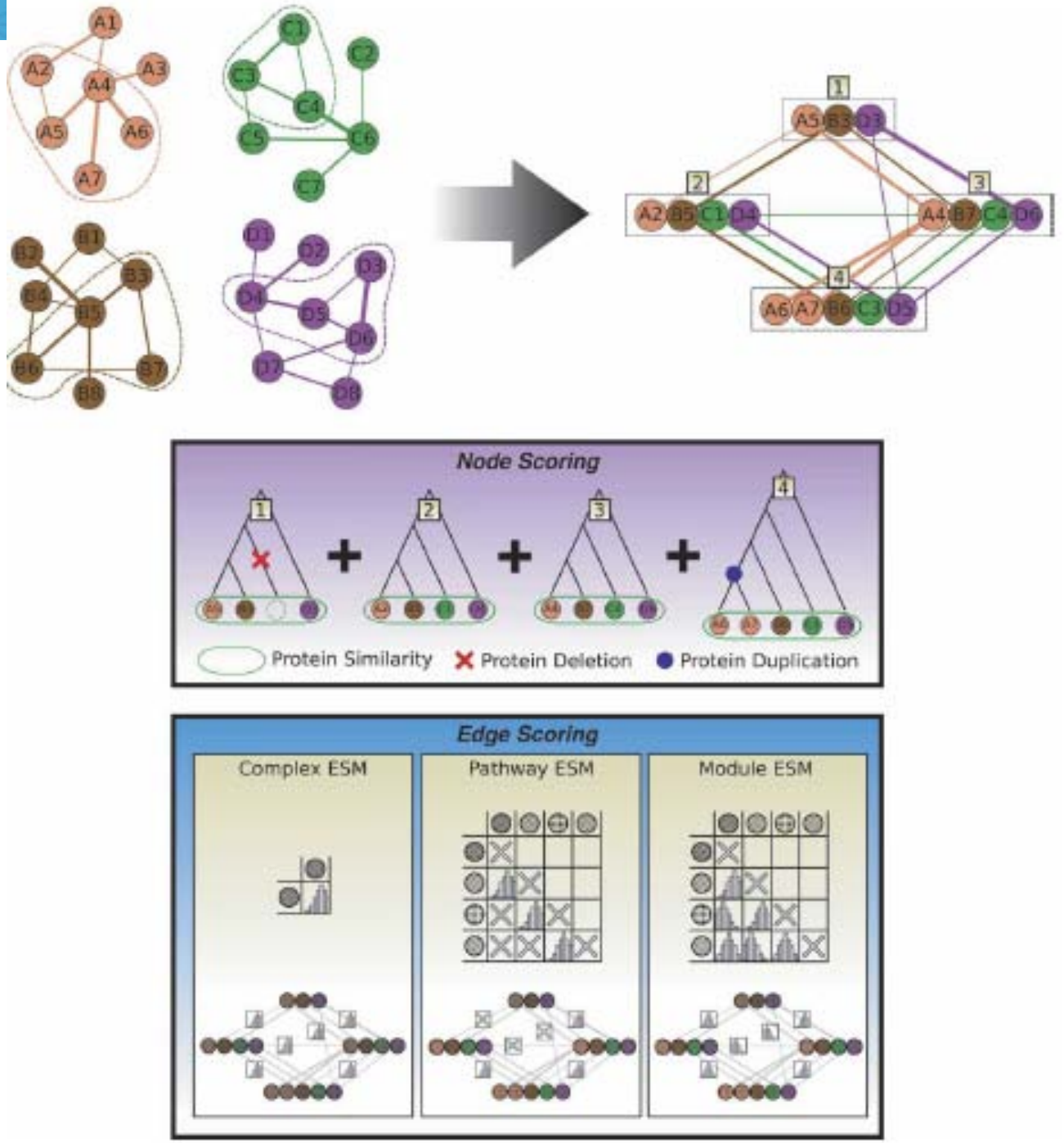
$$\sum_{\alpha \in M} m(\alpha) - \sum_{\beta \in N} n(\beta) - \sum_{\chi \in D} d(\chi)$$

- Koyuturk *et al.* suggested an evolution-based scoring scheme for the alignment of protein interaction networks of two species.
- Define *M* to be the set of interologs (matches) among the two subnetworks being compared (that is, two pairs of interacting proteins, one in each subnetwork, with orthology relations between them).
- Define *N* to be the set of mismatched interactions (that is, two pairs of proteins with orthology relations between them, such that only one pair interacts).
- Define *D* to be the union of the sets of duplicated protein pairs within each subnetwork.

Earlier approaches: Graemlin

- a novel network alignment framework that is fast, scalable, and capable of searching large sets of dense networks for conserved functional modules.
- Græmlin's probabilistic formulation of the topology-matching problem eliminates earlier restrictions on the possible architecture of conserved modules.
- Most important, Græmlin is the first program capable of multiple alignment of an arbitrary number of networks.

Flannick, Jason, Novak, Antal, Srinivasan, Balaji S., McAdams, Harley H., Batzoglou, Serafim, **Graemlin: General and robust alignment of multiple large interaction networks,** Genome Res. 2006.


A1
A2
A3
A4
A5
A6
A7
C1
C2
C3
C4
C5
C6
C7
B1
B2
B3
B4
B5
B6
B7
B8
D1
D2
D3
D4
D5
D6
D7
D8
1
2
3
4
A5 B3 D3
A2 B5 C1 D4
A4 B7 C4 D6
A6 A7 B6 C3 D5
Node Scoring
Protein Similarity
Protein Deletion
Protein Duplication
Edge Scoring
Complex ESM
Pathway ESM
Module ESM

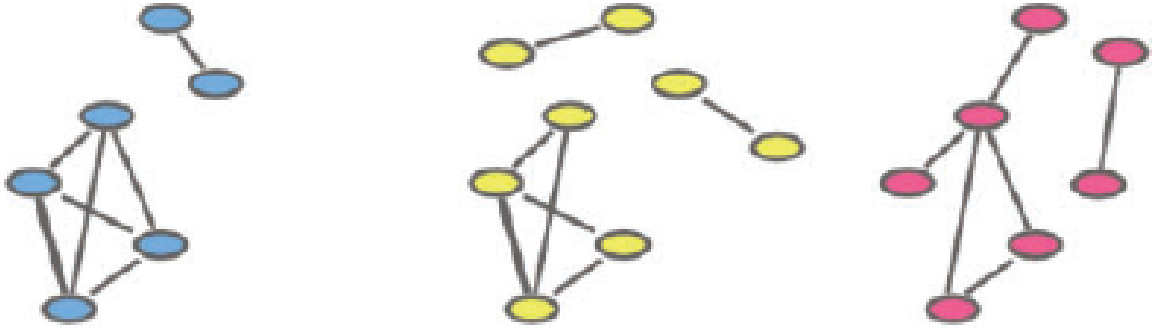
- The efficient performance of Græmlin is due to the use of several strategies common in sequence alignment.
- First, its variant of "progressive alignment" allows it to scale linearly with the number of networks compared.
- Second, Græmlin searches for pairwise alignments between networks using a modification of the "seed extension" method popularized by BLAST.
- Finally, it allows an explicit speed-sensitivity trade-off through the control of a parameter analogous to the BLAST word size.

Earlier approaches: CAPPI

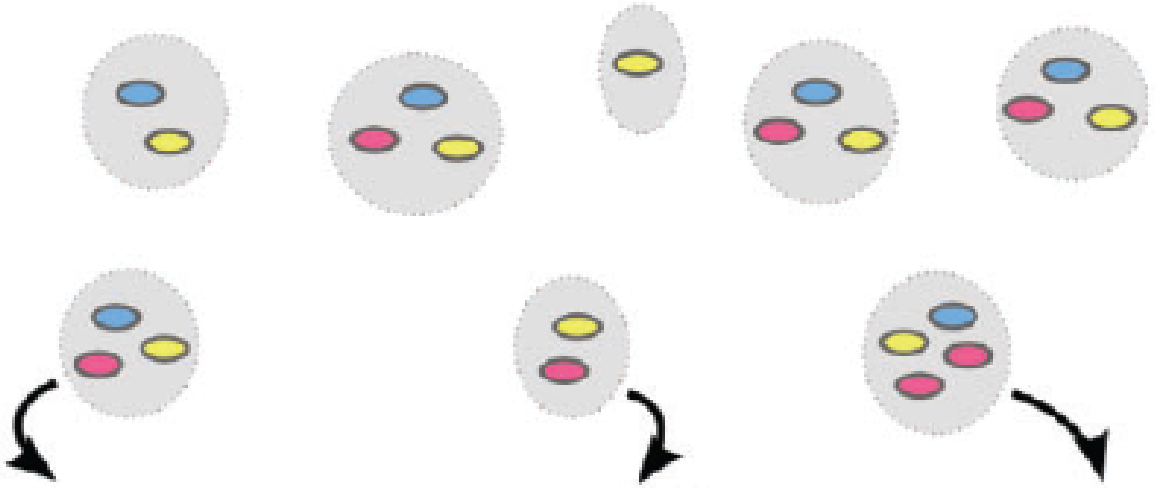
- They develop a new framework for protein network alignment, based on reconstruction of an ancestral PPI network. The reconstruction algorithm is built upon a proposed model of protein network evolution, which takes into account phylogenetic history of the proteins and the evolution of their interactions.

Janusz Dutkowski , and Jerzy Tiuryn **Identification of functional modules from conserved ancestral protein–protein interactions**
Bioinformatics 23: i149-i158, 2007.

1) Input PPI networks

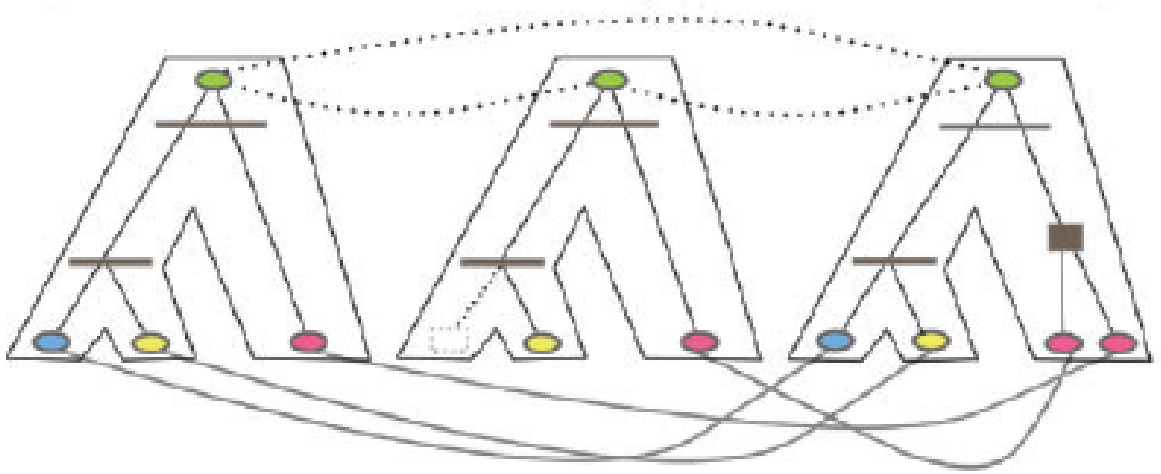


2) Cluster proteins with MCL using BLAST *E*-values as pairwise distances



3) Build reconciled gene trees

4) Compute the probability of each ancestral interaction given protein history, observed interaction data and model of network evolution



Our motivation

(1) A general framework to deal with all kind of networks. Directed and undirected, weighted or unweighted.

(2) The combined network alignment graph should be optimized and one protein should correspond to only one protein.

Our method——MNAligner

Given two networks $G_1=(V_1, E_1)$, $G_2=(V_2,E_2)$,

$$V_1 = \{v_1^1, v_2^1, ..., v_m^1\},$$

$$V_2 = \{v_1^2, v_2^2, ..., v_n^2\},$$

The adjacent matrix are

$$A = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1m} \\ a_{21} & a_{22} & ... & a_{2m} \\ ... & ... & ... & ... \\ a_{m1} & a_{m2} & ... & a_{mm} \end{pmatrix} \qquad B = \begin{pmatrix} b_{11} & b_{12} & ... & b_{1n} \\ b_{21} & b_{22} & ... & b_{2n} \\ ... & ... & ... & ... \\ b_{n1} & b_{n2} & ... & b_{nn} \end{pmatrix}$$

$$a_{ij} = \begin{cases} 1, & if\ (v_i^1, v_j^1) \in E_1 \\ 0, & otherwise \end{cases} \qquad b_{ij} = \begin{cases} 1, & if\ (v_i^2, v_j^2) \in E_2 \\ 0, & otherwise \end{cases}$$

Zhenping Li, Shihua Zhang, Yong Wang, Xiang-Sun Zhang and Luonan Chen. Alignment of molecular networks by integer quadratic programming. Bioinformatics, 2007.

Node similarity

$$S = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ s_{m1} & s_{m2} & \cdots & s_{mn} \end{pmatrix}$$

where S_{ij} is the node v_i^1 in the first network and v_j^2 in the second netowrk

（1）sequence similarity, such as BLAST

（2）protein evolution similarity, such as ortholog information

（3）functional similarity, such as the similarity between enzymes can determined by their EC number difference

Defining variables as

$$x_{ij} = \begin{cases} 1 & \text{if } v_i^1 \in V_1 \text{ matches } v_j^2 \in V_2 \\ 0 & \text{otherwise} \end{cases}$$

Then the network alignment problem is formulated as an Integer quadratic programming problem

$$\max_X \quad f(G_1, G_2) = \lambda \sum_{i=1}^{m}\sum_{j=1}^{n} s_{ij}x_{ij} + (1-\lambda)\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{l=1}^{n} a_{ik}b_{jl}x_{ij}x_{kl}$$

$$s.t. \begin{cases} \sum_{j=1}^{n} x_{ij} \le 1 & i = 1, 2, \cdots m \\ \sum_{i=1}^{m} x_{ij} \le 1 & j = 1, 2, \cdots n \\ x_{ij} = 0, 1 & i = 1, 2, \cdots m; j = 1, 2, \cdots, n \end{cases}$$

Comment of model

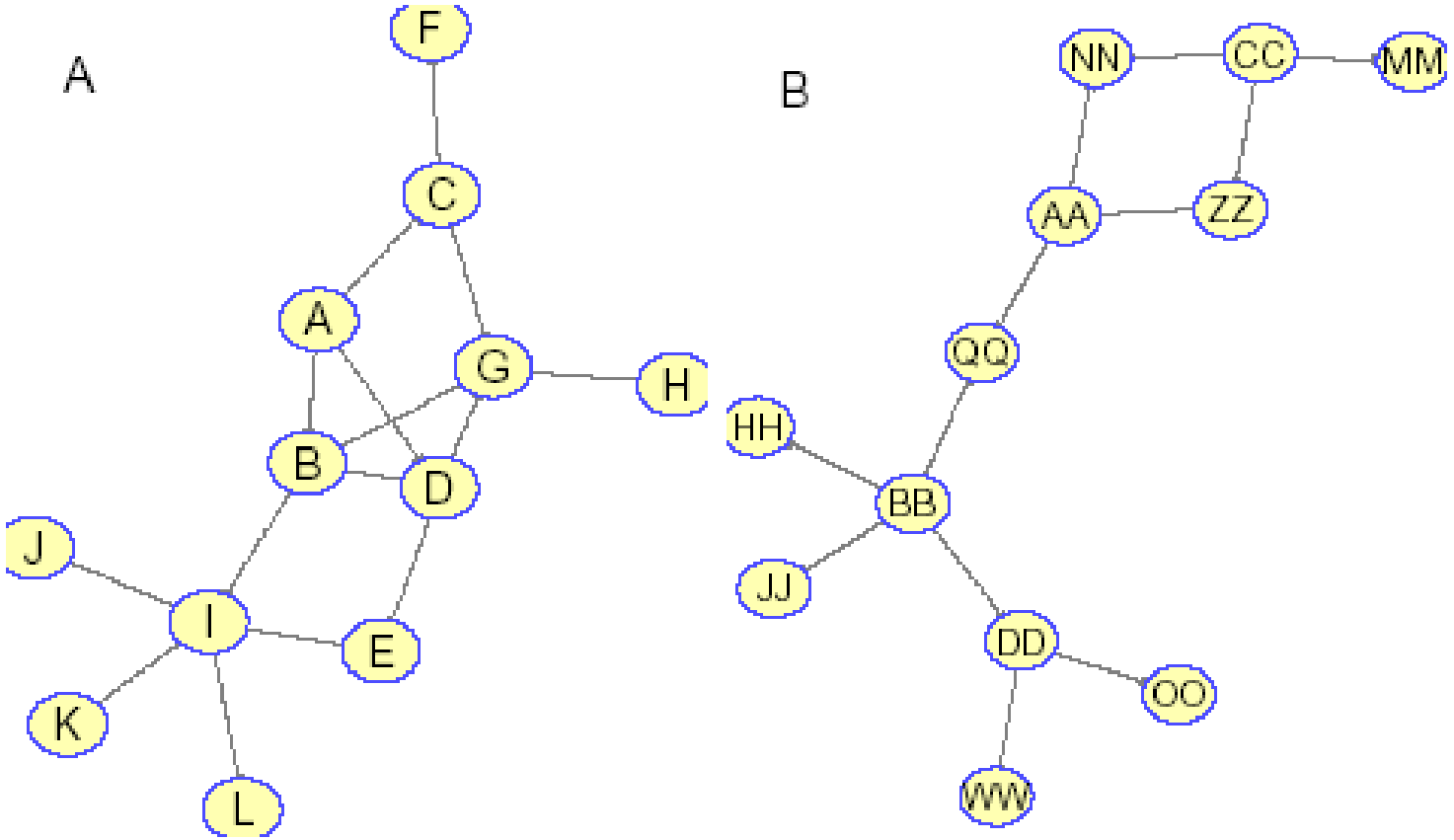
Object function：The first term is total node similarity and the second term is the edge similarity.

The parameter λ is to balance the importance of node similarity and edge similarity

Constraints：One node in one network can correspond to at most one node in the other network

Some results

An example from website of PathBLAST (http://www.cytoscape.org/plugins1.php)



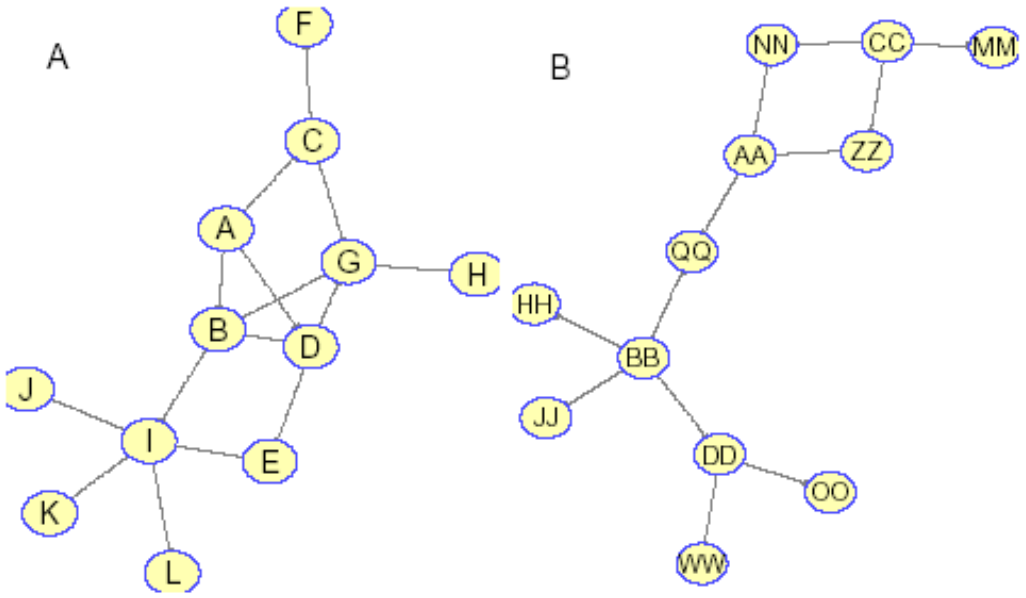
Adjacent matrix

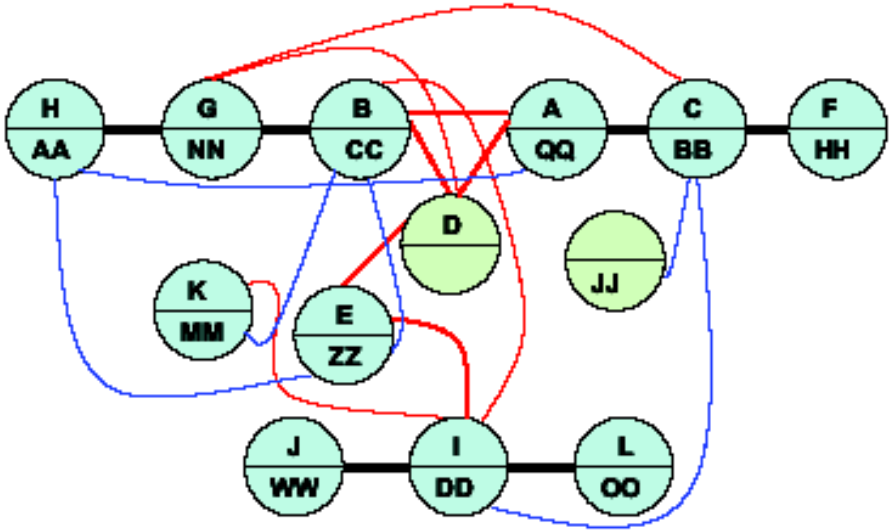
$$A = \begin{pmatrix}
0 & 0.10 & 0.70 & 0.01 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.10 & 0 & 0 & 0.30 & 0 & 0 & 0.01 & 0 & 0.02 & 0 & 0 & 0 \\
0.70 & 0 & 0 & 0 & 0 & 0.20 & 0.01 & 0 & 0 & 0 & 0 & 0 \\
0.01 & 0.30 & 0 & 0 & 0.20 & 0.01 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.20 & 0 & 0 & 0 & 0 & 0.01 & 0 & 0 & 0 \\
0 & 0 & 0.20 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.01 & 0.01 & 0.01 & 0 & 0 & 0 & 0.70 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.70 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.02 & 0 & 0 & 0.01 & 0 & 0 & 0 & 0 & 0.30 & 0.01 & 0.60 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.30 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.01 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.60 & 0 & 0 & 0
\end{pmatrix}$$

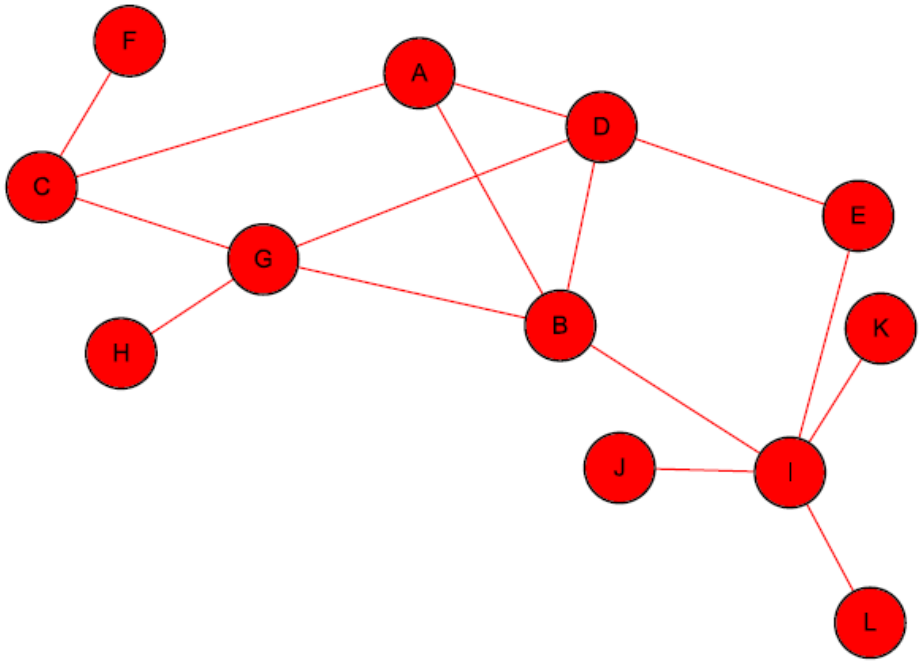
$$B = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0.01 & 0.20 & 0.10 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.01 & 0.70 & 0 & 0 & 0 & 0.70 & 0.01 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.02 & 0.20 & 0.10 & 0 & 0 & 0 & 0 \\
0 & 0.01 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.10 & 0.01 \\
0 & 0.70 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.02 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.01 & 0 & 0.20 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.20 & 0 & 0.10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.10 & 0.70 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.01 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.01 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}$$

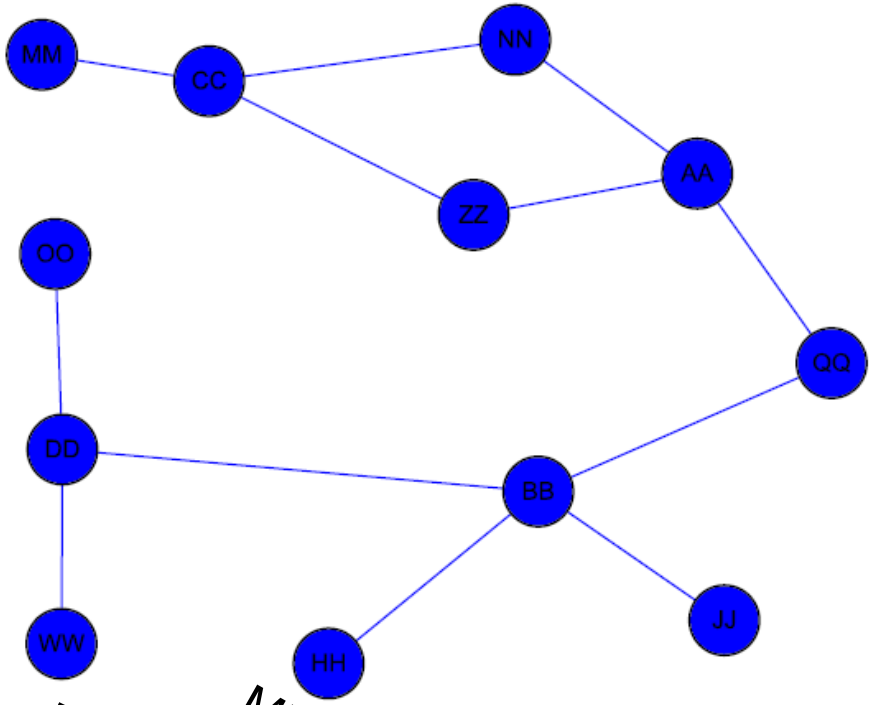
Node similarity matrix

$$S = \begin{pmatrix}
0.1 & 0.1 & 0.1 & 0.8 & 0.5 & 0.1 & 0.1 & 0.8 & 0.8 & 0.1 & 0.1 & 0.1 \\
0.1 & 0.1 & 0.8 & 0.1 & 0.1 & 0.1 & 0.8 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\
0.1 & 0.8 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\
0.1 & 0.8 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\
0.1 & 0.8 & 0.1 & 0.1 & 0.1 & 0.1 & 0.8 & 0.1 & 0.1 & 0.1 & 0.8 & 0.1 \\
0.1 & 0.1 & 0.1 & 0.1 & 0.8 & 0.1 & 0.1 & 0.8 & 0.1 & 0.1 & 0.1 & 0.1 \\
0.8 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.8 & 0.1 \\
0.8 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\
0.1 & 0.1 & 0.1 & 0.8 & 0.1 & 0.8 & 0.8 & 0.1 & 0.1 & 0.1 & 0.1 & 0.8 \\
0.1 & 0.1 & 0.1 & 0.8 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.8 \\
0.8 & 0.1 & 0.1 & 0.1 & 0.1 & 0.8 & 0.1 & 0.1 & 0.8 & 0.1 & 0.1 & 0.8 \\
0.1 & 0.1 & 0.8 & 0.8 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.8 & 0.8
\end{pmatrix}$$


A
F
C
A
G
H
B
D
J
I
E
K
L
B
NN
CC
MM
AA
ZZ
QQ
HH
BB
JJ
DD
OO
WW


H
AA
G
NN
B
CC
A
QQ
C
BB
F
HH
D
JJ
K
MM
E
ZZ
J
WW
I
DD
L
OO



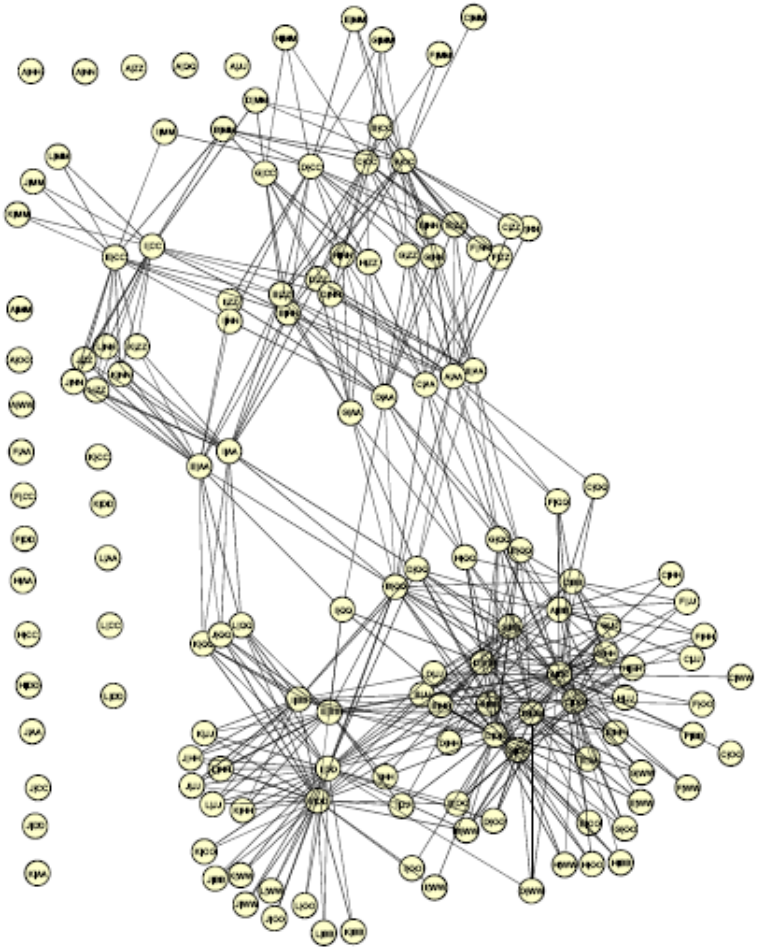


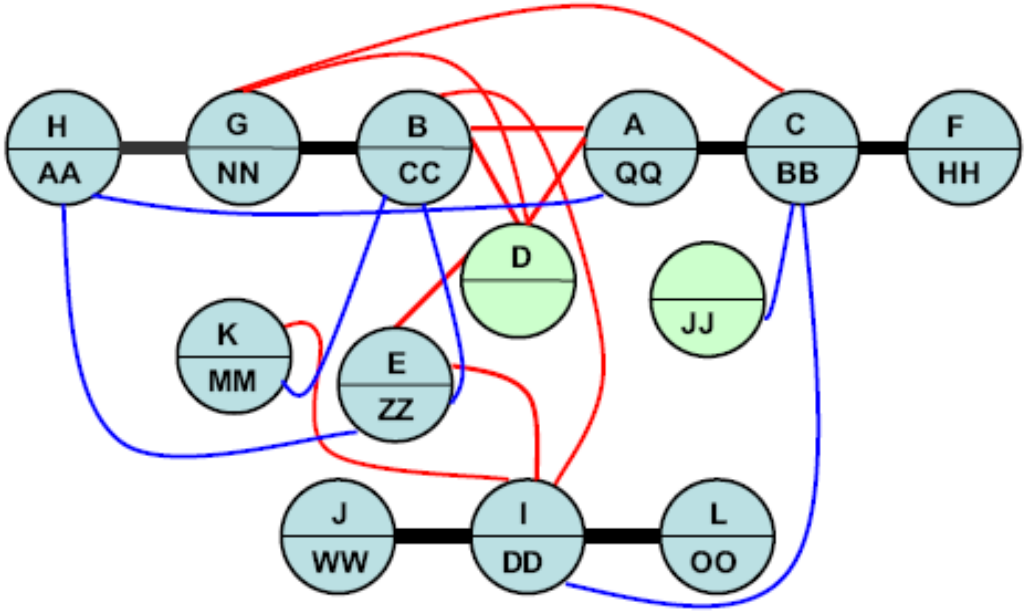
PathBLAST

MNAligner

Global Alignment Graph

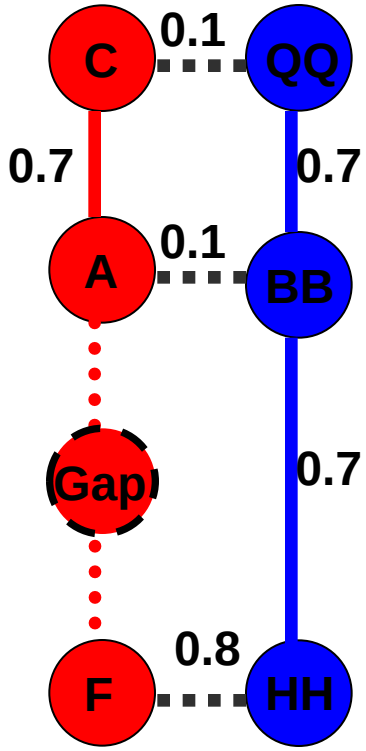
Global Alignment Graph





Conserved Pathways

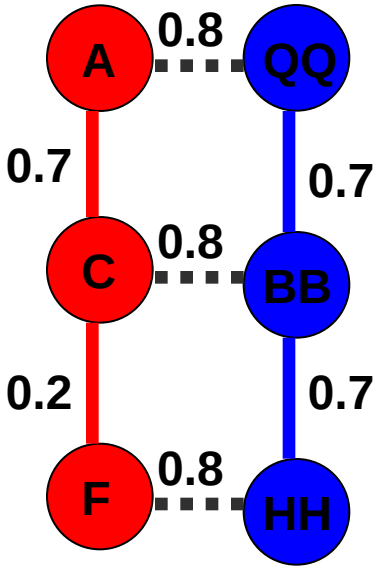
Conserved Pathways



The first

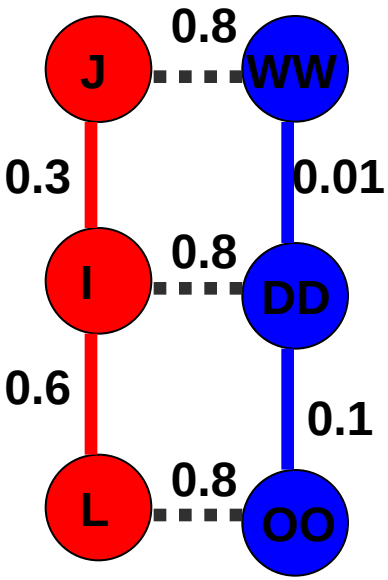
SP	-3.41548
SM	0.815

Conserved Pathways



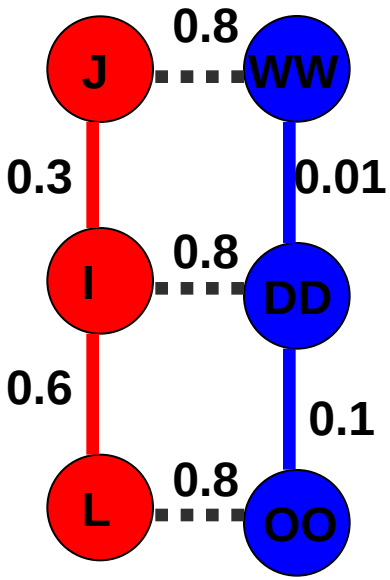
The first

SP	-1.4544
SM	1.5015



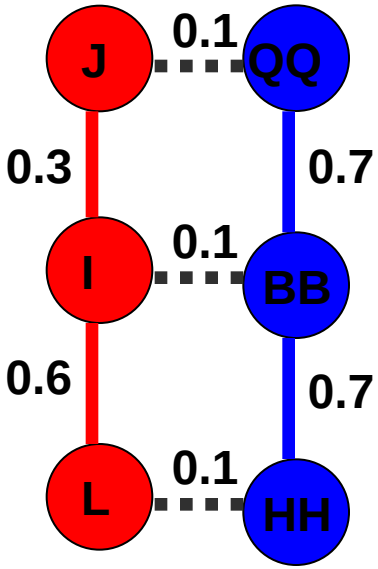
The second

SP	-4.03545
SM	1.2315



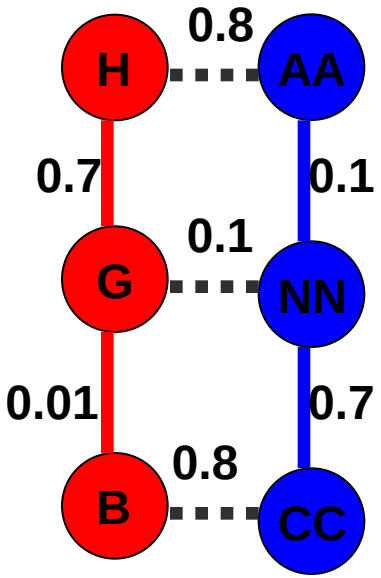
The second

SP	-4.03545
SM	1.2315



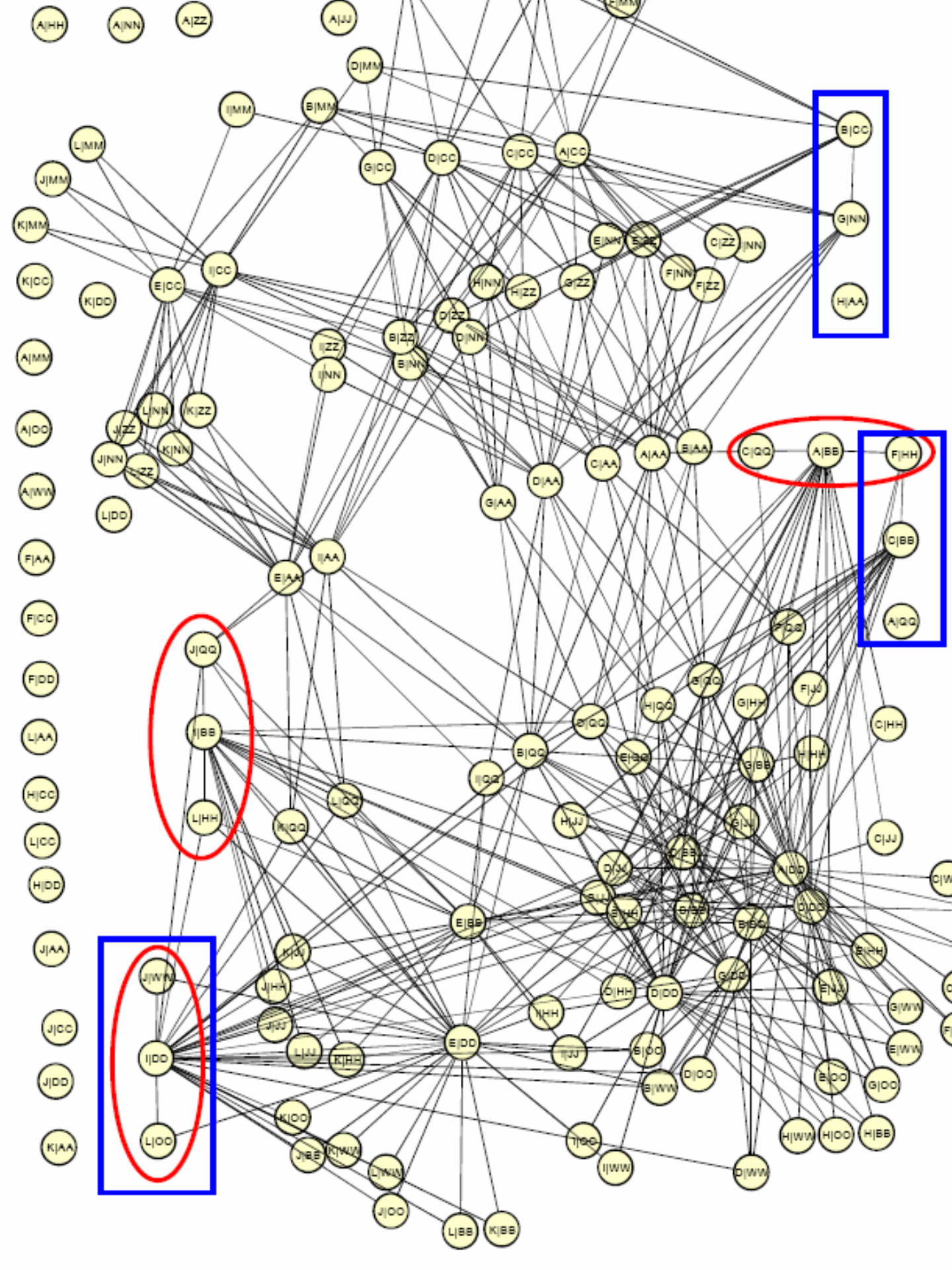
The third

SP	-4.05453
SM	0.465



The third

SP	-5.04769
SM	0.9205


A|HH
A|NN
A|ZZ
A|JJ
D|MM
I|MM
B|MM
L|MM
J|MM
K|MM
G|CC
D|CC
C|CC
A|CC
B|CC
G|NN
H|AA
E|NN
B|ZZ
C|ZZ
I|NN
F|NN
F|ZZ
H|NN
H|ZZ
G|ZZ
I|CC
E|CC
K|CC
K|DD
D|ZZ
B|ZZ
D|NN
I|ZZ
B|NN
I|NN
A|MM
L|NN
K|ZZ
A|CC
J|ZZ
K|NN
J|NN
L|ZZ
A|WW
L|DD
C|QQ
A|BB
F|HH
C|AA
A|AA
B|AA
D|AA
G|AA
C|BB
A|QQ
I|AA
E|AA
F|AA
F|CC
J|QQ
F|DD
I|BB
L|AA
H|CC
L|HH
L|CC
H|DD
J|AA
J|WW
J|CC
I|DD
J|DD
L|OO
K|AA
E|BB
B|QQ
H|JJ
C|JJ
D|JJ
A|DD
E|HH
B|CC
G|DD
D|HH
D|DD
G|WW
E|WW
B|OO
G|OO
H|WW
H|OO
H|BB
D|WW
I|WW
I|OO
L|BB
K|BB

We can align two directed networks

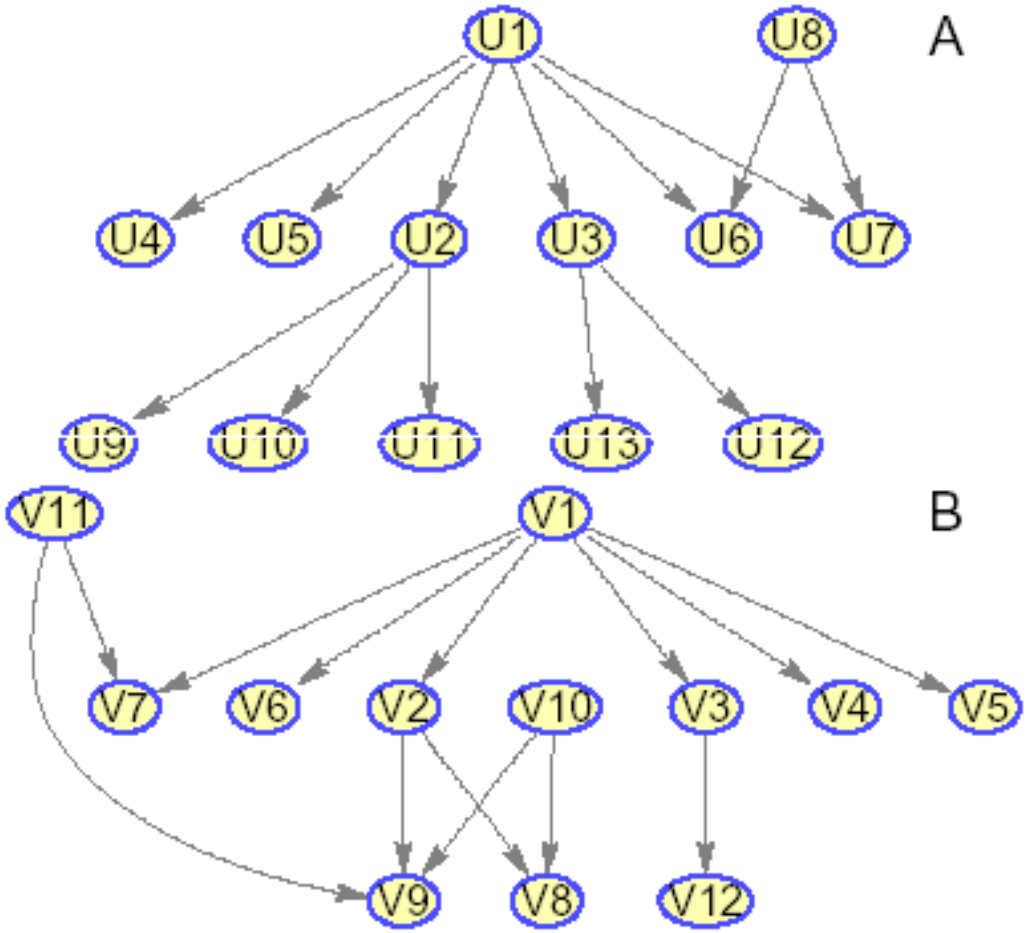
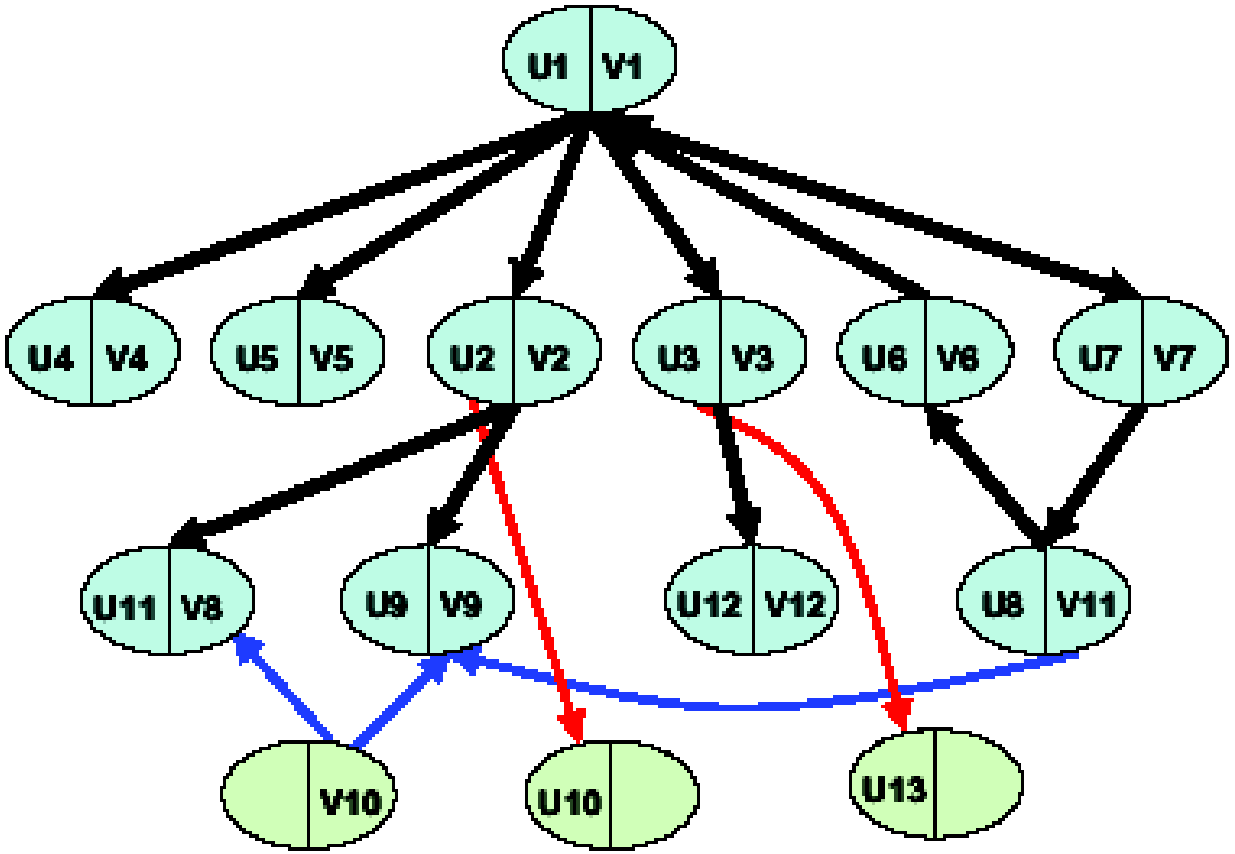
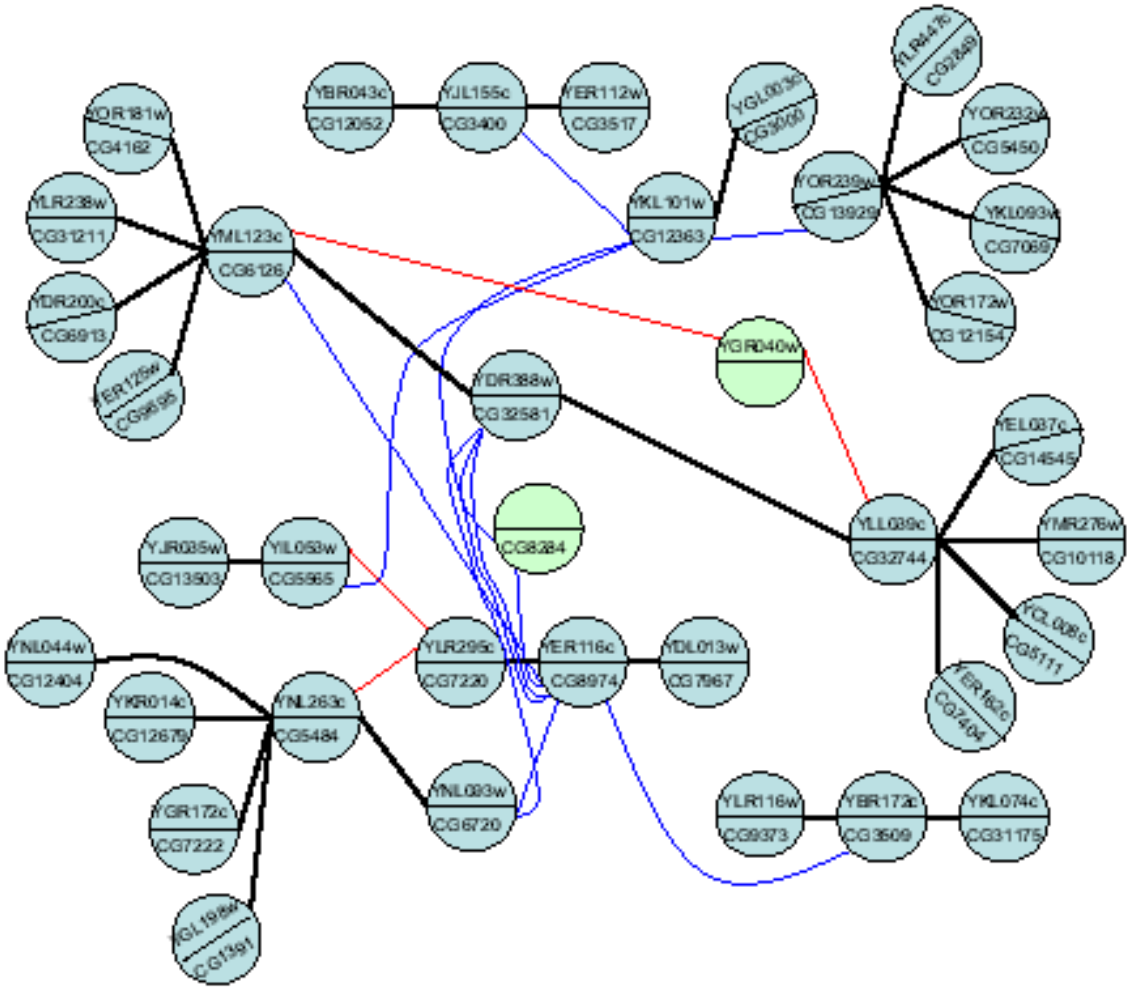


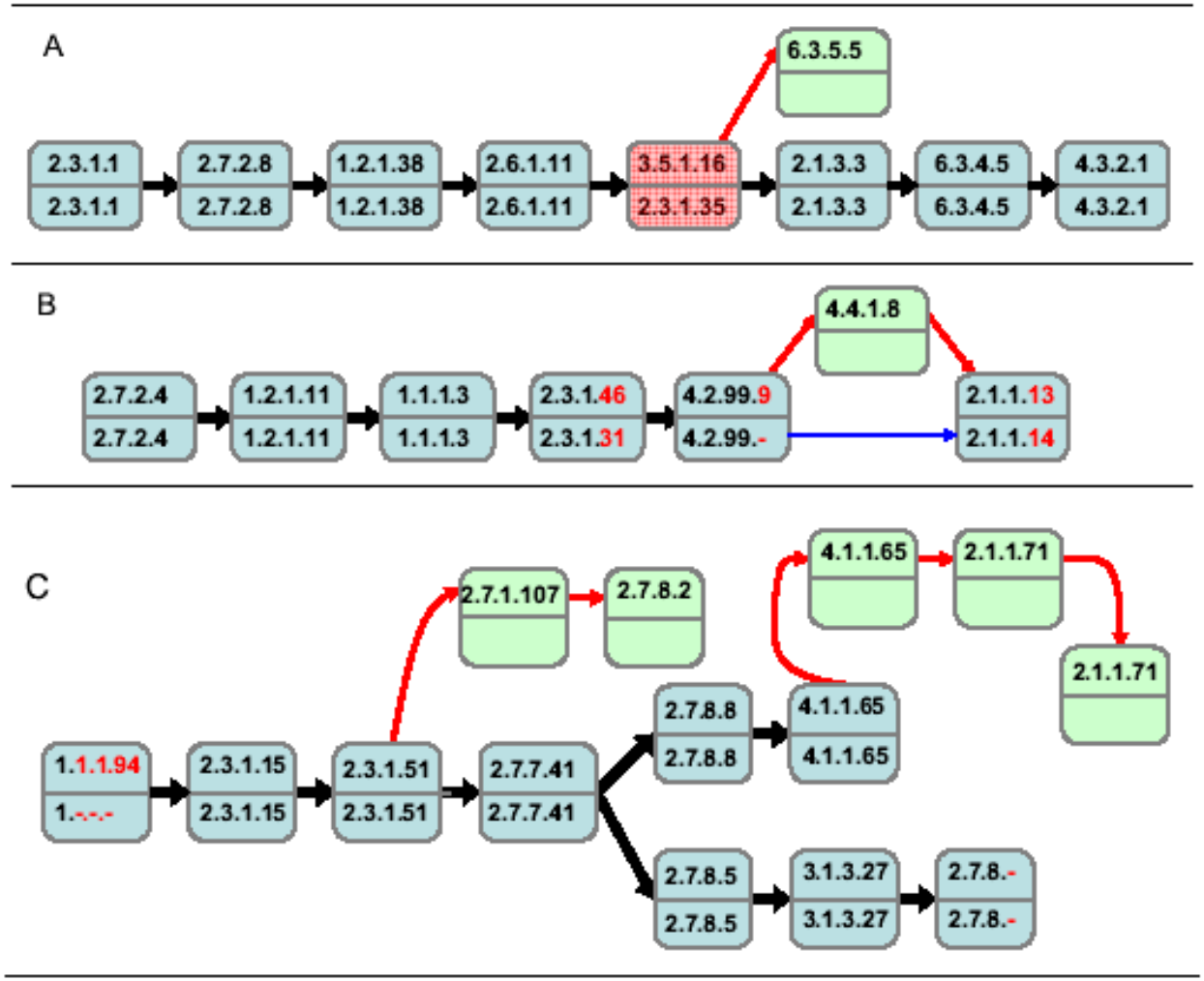
Fig. 2. The simulated example of two directed networks


U1 V1
U4 V4
U5 V5
U2 V2
U3 V3
U6 V6
U7 V7
U11 V8
U9 V9
U12 V12
U8 V11
V10
U10
U13

Real data example on PPI network



Metabolic pathway alignment



Network comparison Globally

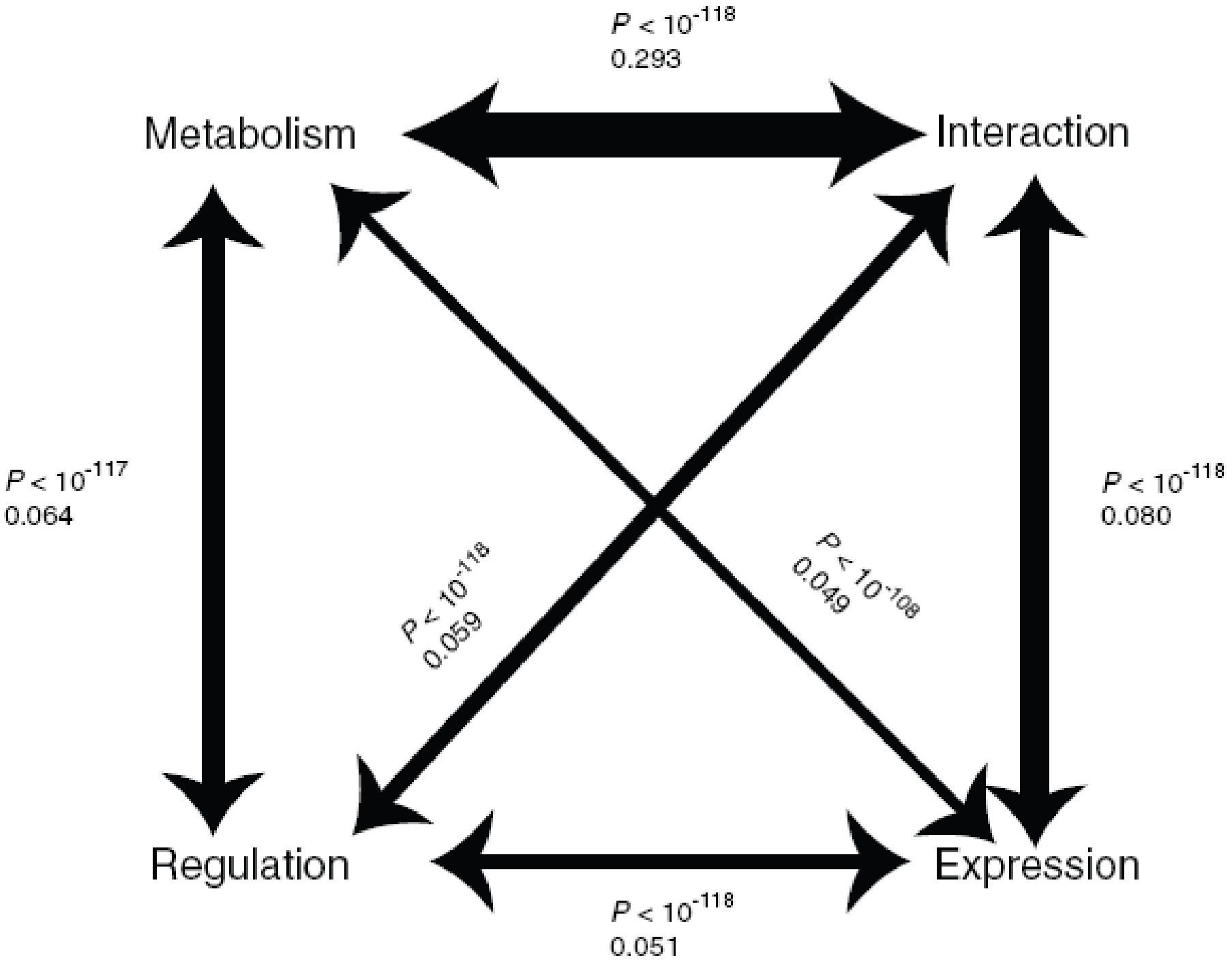
1. Directly to find the isomorphism is NP-complete, thus this measure can not be used to practically test similarity of two networks.
2. The feasible way is to extract features or global properties from the network, then compute the similarity between the vectors or distributions.

- It is very common to use some of the topological features of networks as a basis of checking their similarity.
- For example, the degree distribution, the k-hop reachability, the graphlet frequency, the betweenness distribution and the closeness distribution.

A global comparison of four basic molecular networks: regulatory, co-expression, interaction, and metabolic. In terms of overall topologic correlation

Network name		Network Type	Number of proteins (N)	Number of links	Power-law distribution $N = \alpha K^{-\gamma}$		Average degree (K)	Clustering coefficient (C)	Characteristic path length (L)	Diameter (D)
					α	γ				
Expression		undirected	5,205	70,201	2,542	1.358	26.97	0.3585	5.518	19
Interaction			4,743	23,294	2,601	1.588	9.822	0.2321	4.358	11
Metabolism		directed	852	5,933	486.6	1.341	13.93	0.434	4.659	20
Regulation	Regulator		248	7,231	16.01	0.5835	29.14	0.1087	3.766	9
	Target		3,271		-	-	2.209			

Yu H, Xia Y, Trifonov V, Gerstein M. **Design principles of molecular networks revealed by global comparisons and composite motifs**. *Genome Biology* 7: R55 (2006).

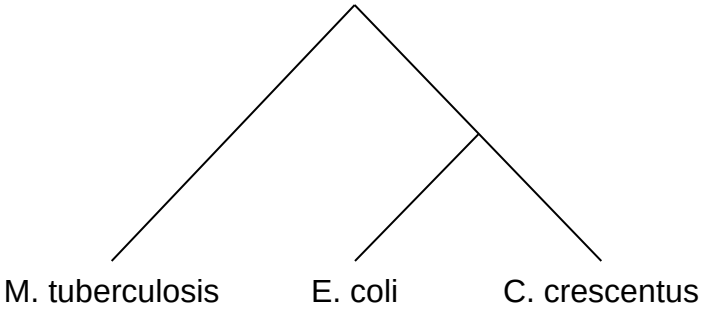

$P < 10^{-118}$
0.293
Metabolism
Interaction
$P < 10^{-117}$
0.064
$P < 10^{-118}$
0.080
$P < 10^{-118}$
0.059
$P < 10^{-108}$
0.049
Regulation
Expression
$P < 10^{-118}$
0.051

Construct phylogenetic tree

- Basically use the sequence or structure similarity to get the distance matrix.
- Can we use the network data of different species (PPI, co-expression)?
- Relate network with evolution
- Network evolution? (Understanding how network evolves is a fundamental issue) sequence mutation+ duplication

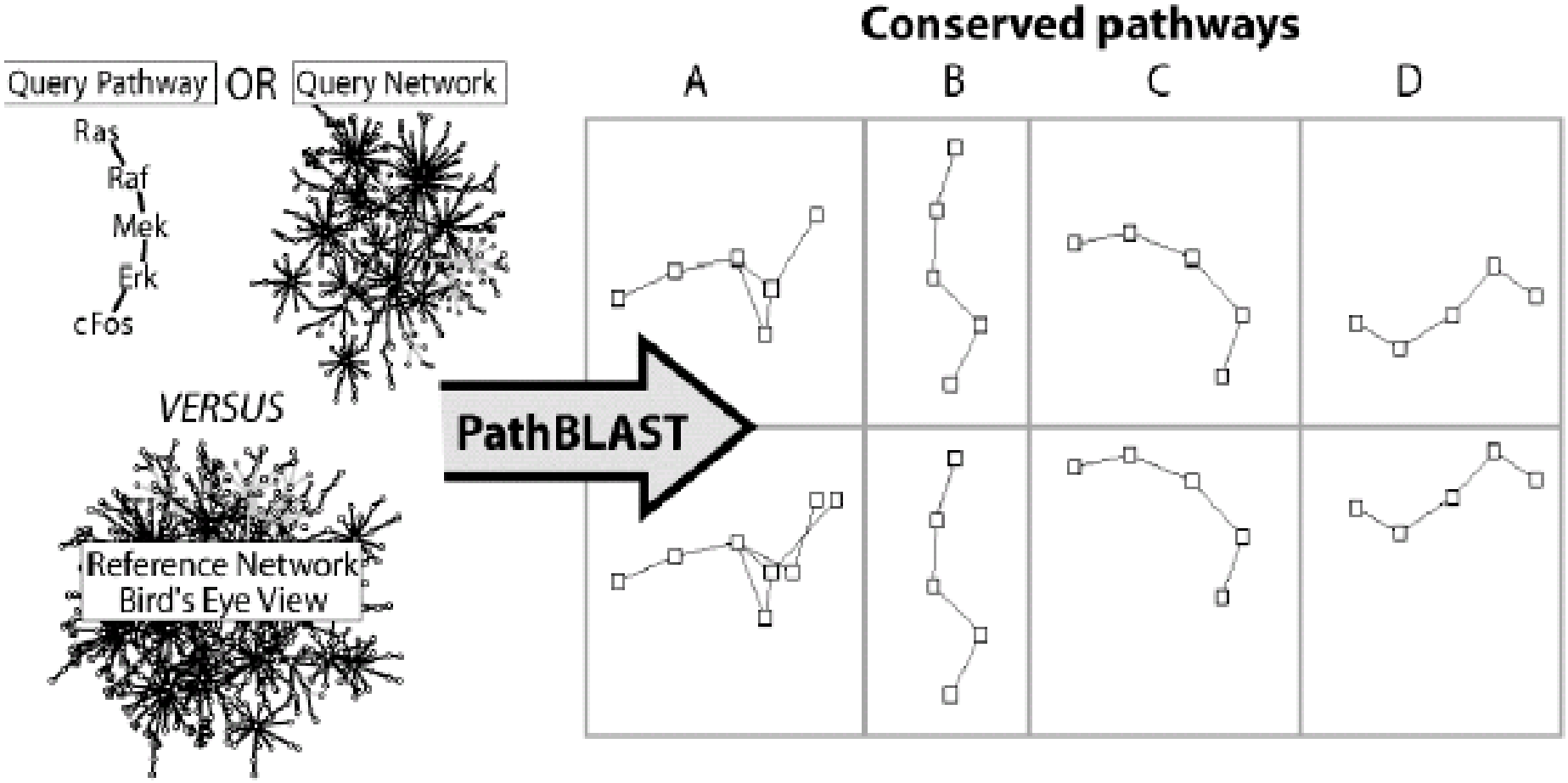
Multiple Alignment?

- Progressive alignment technique
  - Used by most multiple sequence aligners



- Simple modification of implementation to align *alignments* rather than *networks*
  - Node scoring already uses weighted SOP
  - Edge scoring remains unchanged

Network query



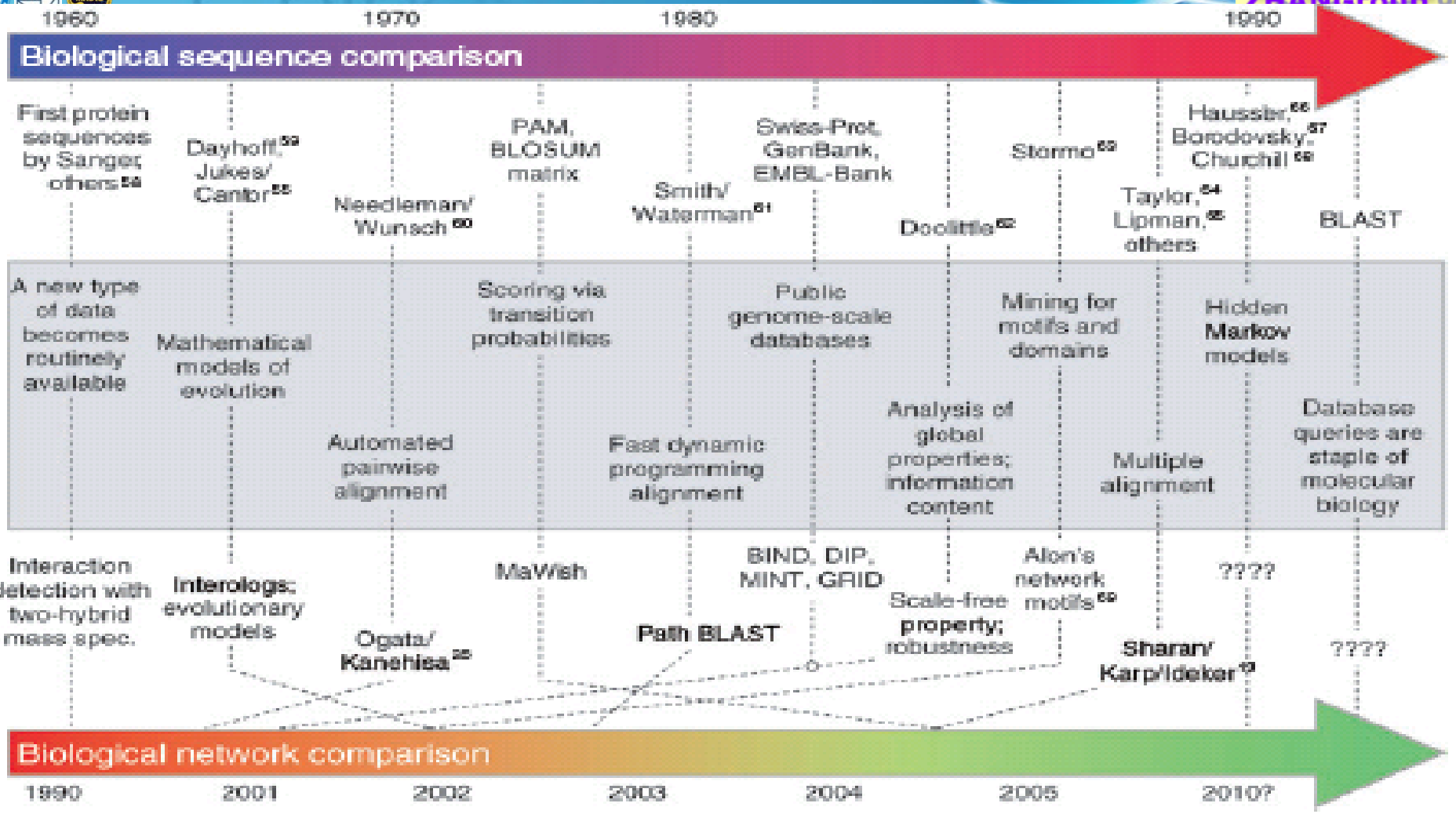


Figure 4 Parallels between sequence and network comparison on a timeline. The recent and possibly future developments in methods for network comparison are shown in the context of the analogous developments as they occurred in the field of sequence comparison. General milestones for both fields are shown in the middle (gray box), with the specific instances for sequence versus network comparison appearing directly above or below, respectively.

Linearity of sequences as opposed to the nonlinearity of networks

Take-home messages

- Network alignment: NP hard problem
- Heuristic methods
- Global Vs local; alignment Vs comparison