生物信息学与系统生物学

张世华

中国科学院数学与系统科学研究院

Aligning Biological Molecular Networks across various species

Shihua Zhang

Questions?

- Molecular networks are of current interest. Previous analyses have focused on topologic structures of **individual network**.

- **Biological networks are different** (molecular types, species organisms, or tissues, under varying conditions).

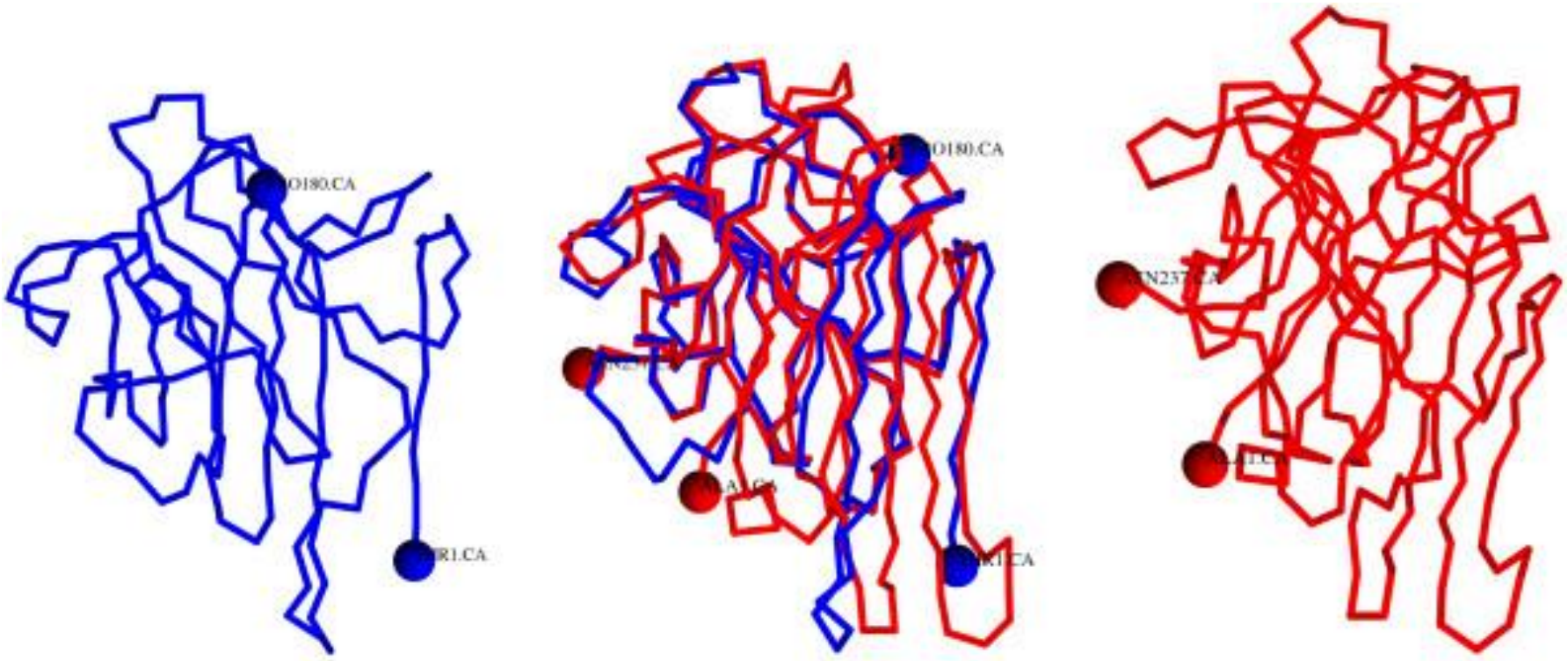
- We should take **a comparative approach** toward interpreting these networks.

Sequence alignment——

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

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

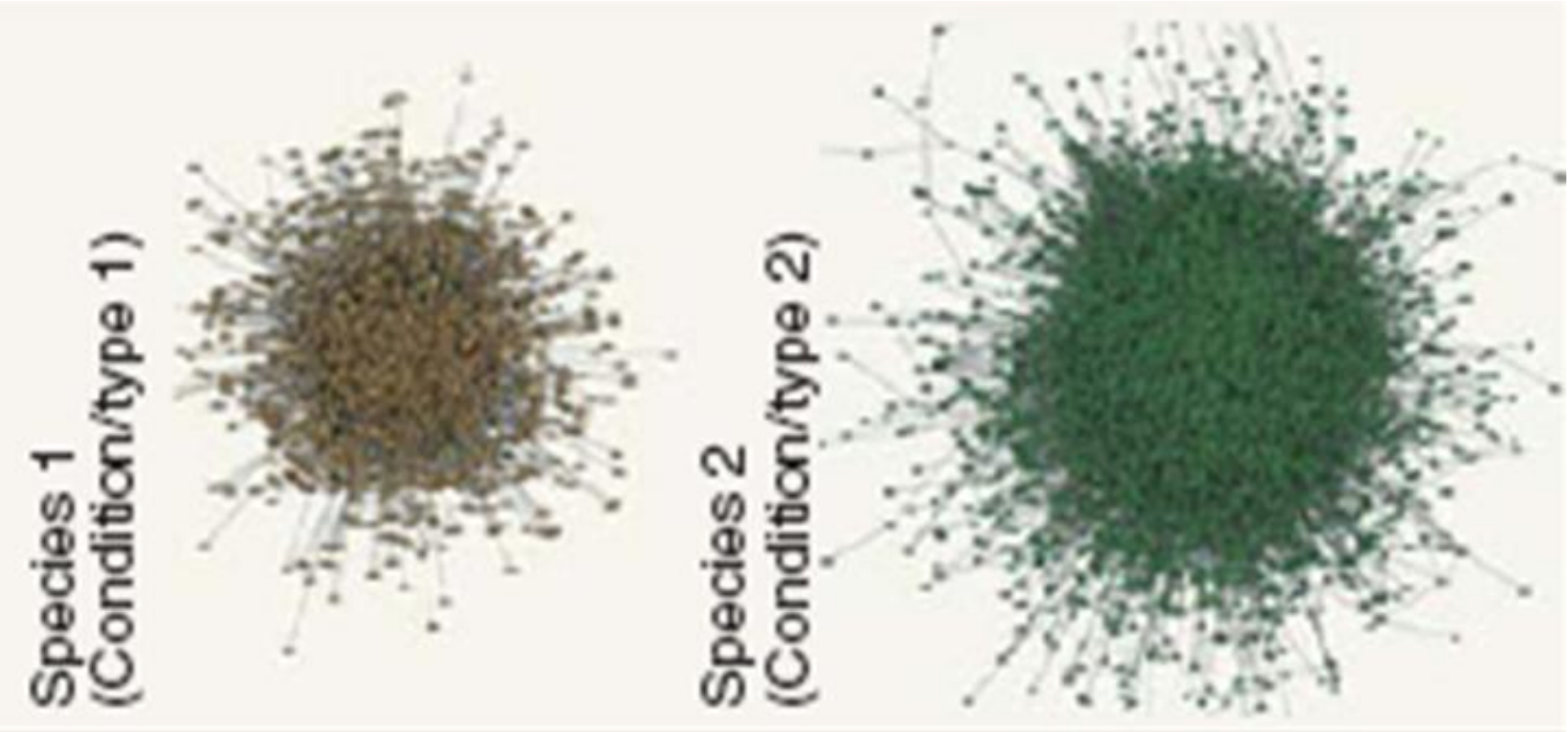
Structural alignment——



(a) Backbone of 1rin. (b) Alignment result. (c) Backbone of 2cna.

L. Chen, L.Y. Wu, Y. Wang, S. Zhang and X.S. Zhang.
BMC Structural Biology, 2006, 6:18.

Network alignment——???



Linearity of sequences as opposed to the nonlinearity of networks

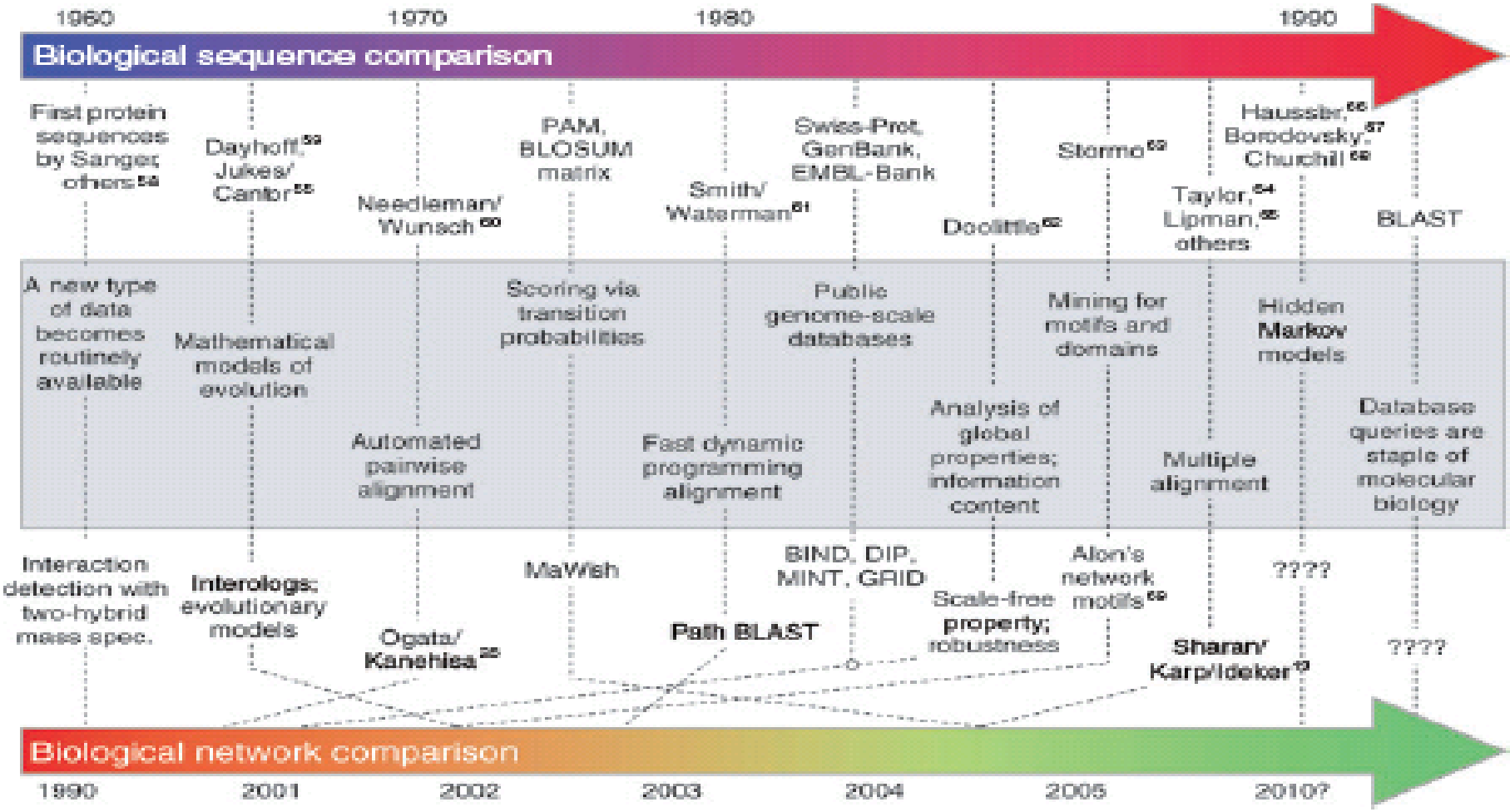
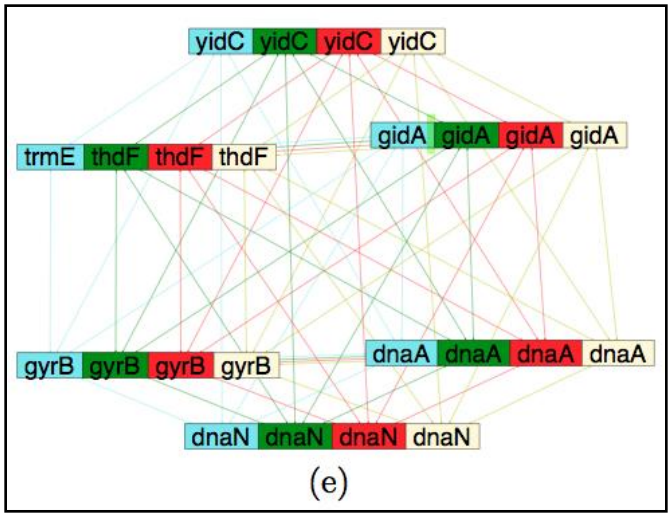


Figure 4 Parallels between sequence and network comparison on a timeline. The recent and possibly future developments in methods for network comparison are shown in the context of the analogous developments as they occurred in the field of sequence comparison. General milestones for both fields are shown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

Nature Biotechnology. 24(4):427-33 (2006).

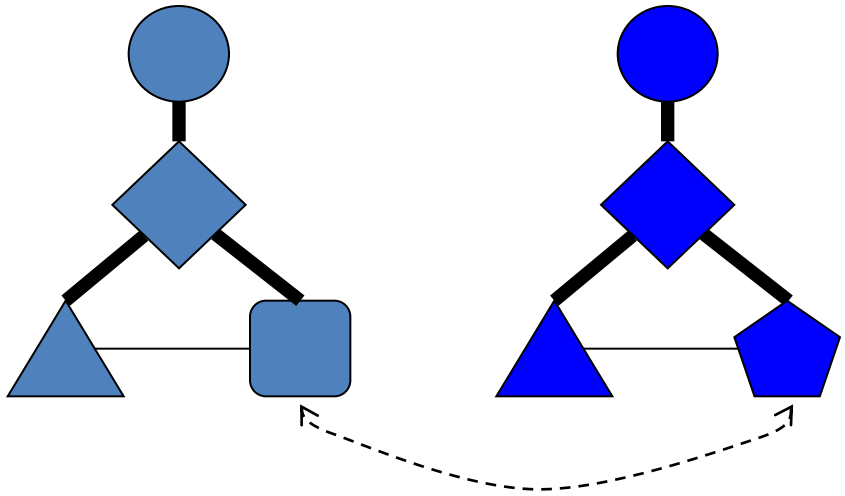
Motivation



- 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

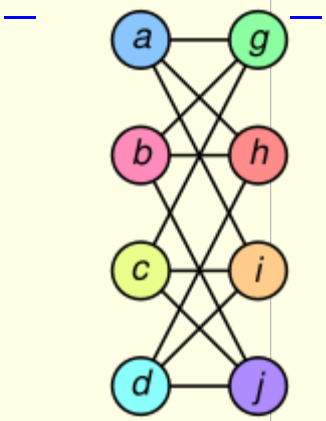
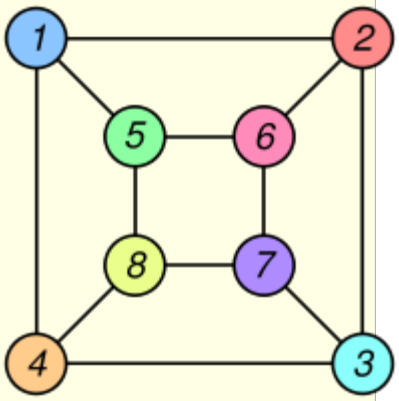
- "Conserved" means two subgraphs contain proteins serving **similar** functions, having **similar** interaction profiles
  - Key word is similar, not identical



SubGraph isomorphism

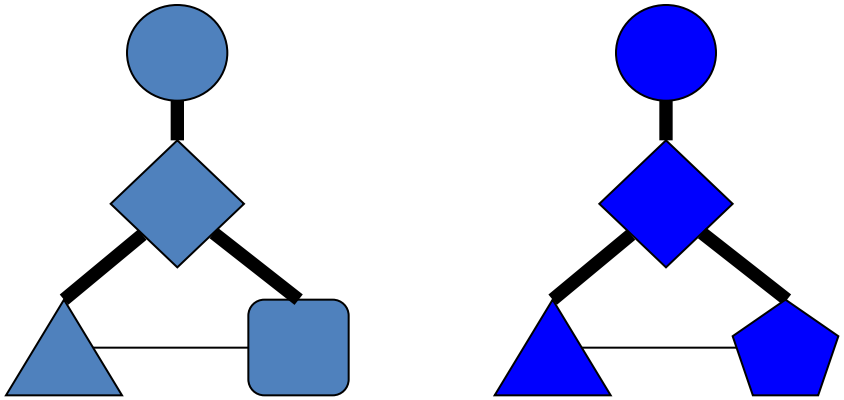
In graph theory, a graph isomorphism is a bijection (a one-to-one and onto mapping) between the vertices of two graphs *G* and *H*， *f:V(G)→V(H)*, with the property that any two vertices *u* and *v* from *G* are adjacent if and only if *f(u)* and *f(v)* are adjacent in *H*.

•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

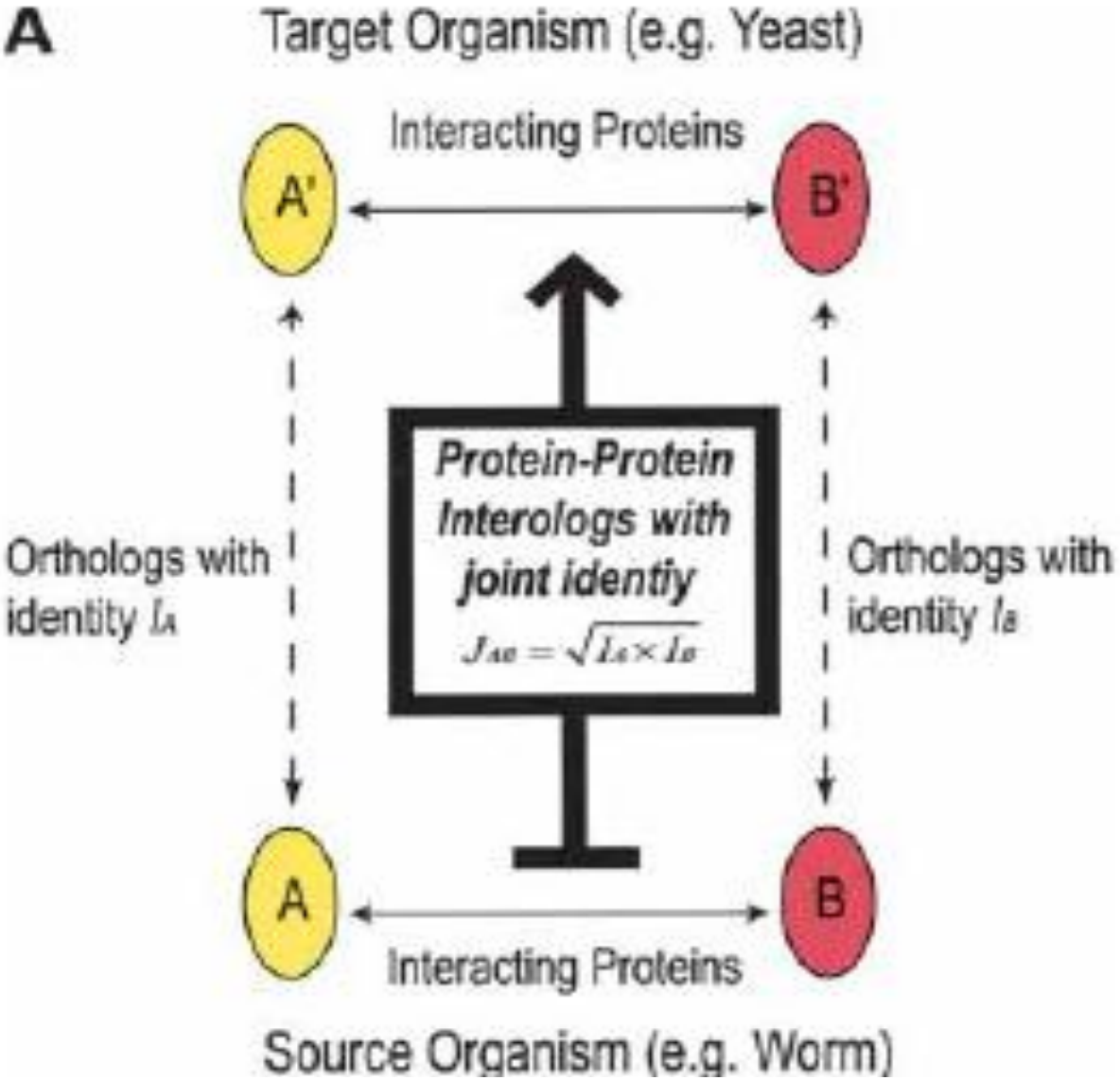
The simplest case: interologs

- **Inter**actions conserved in ortho**logs**
  - Orthology is a fuzzy notion
  - Sequence similarity is 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.

interolog


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 (PNAS, 2003).
- A greedy algorithm is applied for identifying the conserved subnetworks embedded in the merged representation.

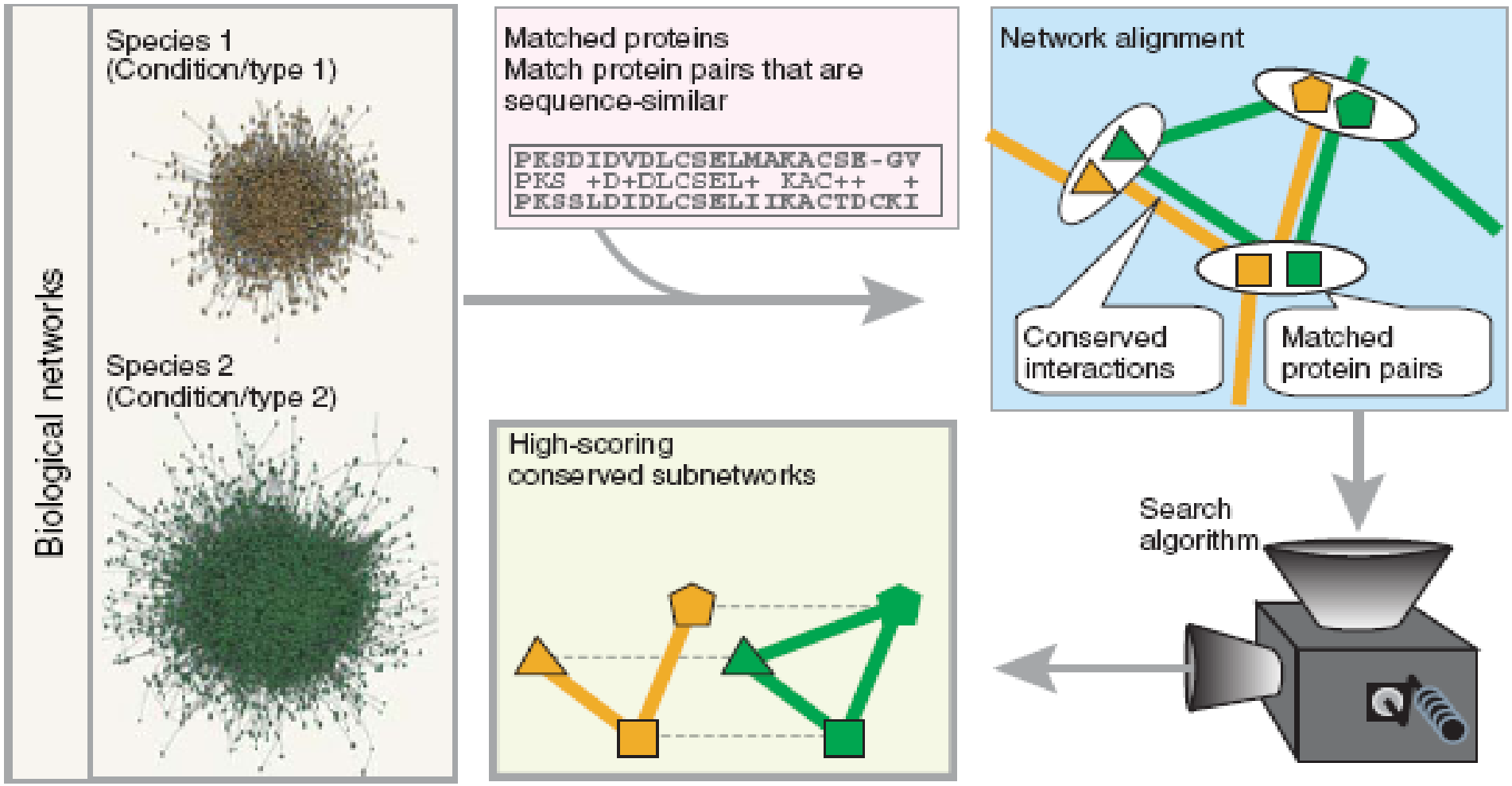
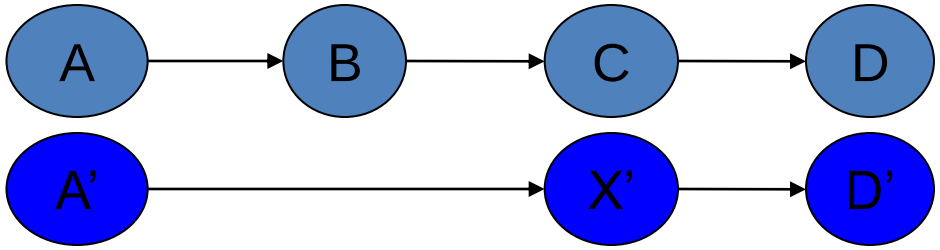


Figure 1 Network alignment. Network alignment combines protein interaction data that are available for each of at least two species with orthology information based on the corresponding protein sequences. A detailed probabilistic model is used to identify protein subnetworks within the aligned network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

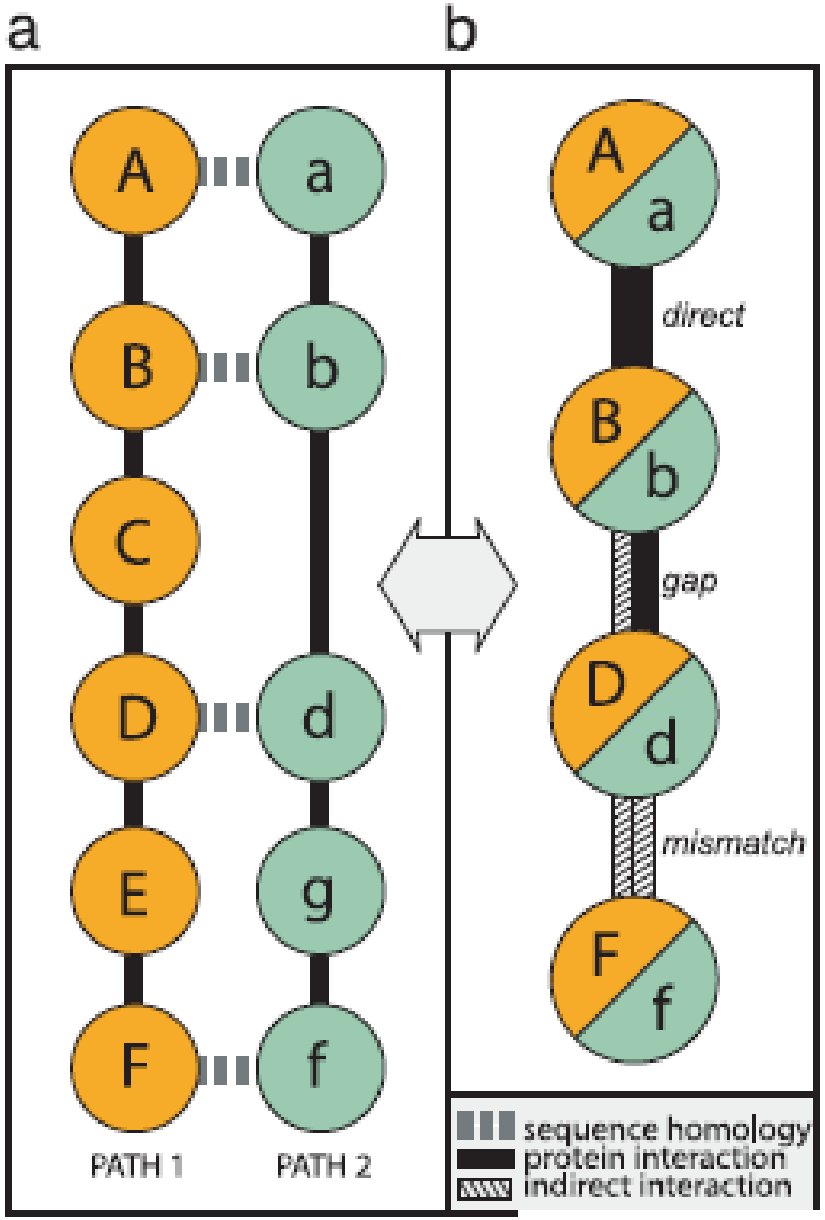
- Goal: identify conserved *pathways* (chains)
- Idea: can be done efficiently by dynamic programming if networks are DAGs



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).

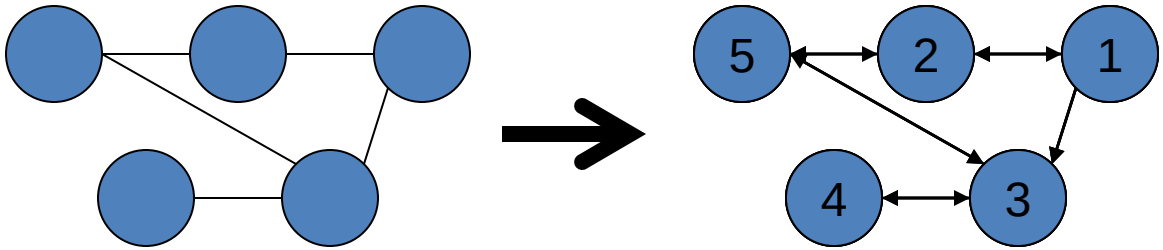


Comment: One of the **drawbacks** of the alignment graph is that it includes a node for every pair (or triplet) of similar proteins (one from each input network). The commonly used similarity functions (e.g. BLAST E-value threshold) generally impose a **many-to-many correspondence** between proteins,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

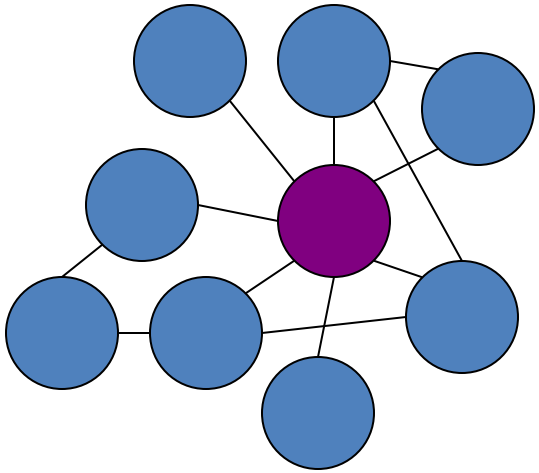


- In expectation, finds conserved paths of length *L* within networks of size *n* in $O(L!n)$ time
- Drawbacks
  - Computationally expensive
  - Restricts search to specific topology

Kelley et al (2003)

Earlier approaches: MaWISh

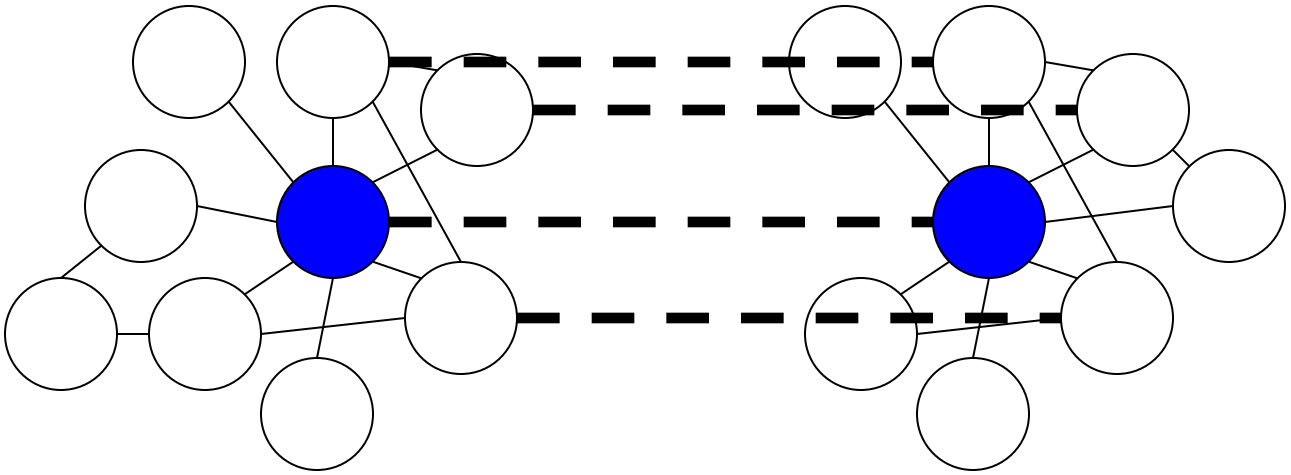
- Goal: identify conserved *multi-protein complexes* (clique-like structures)
- Idea: such structures will likely contain at least one *hub* (high-degree node)



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 importantly, 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.

- 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.

Our motivation

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

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

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

Z. Li*, S. Zhang*, Y. Wang, X.S Zhang and L. Chen. 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

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

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

Comments on the model

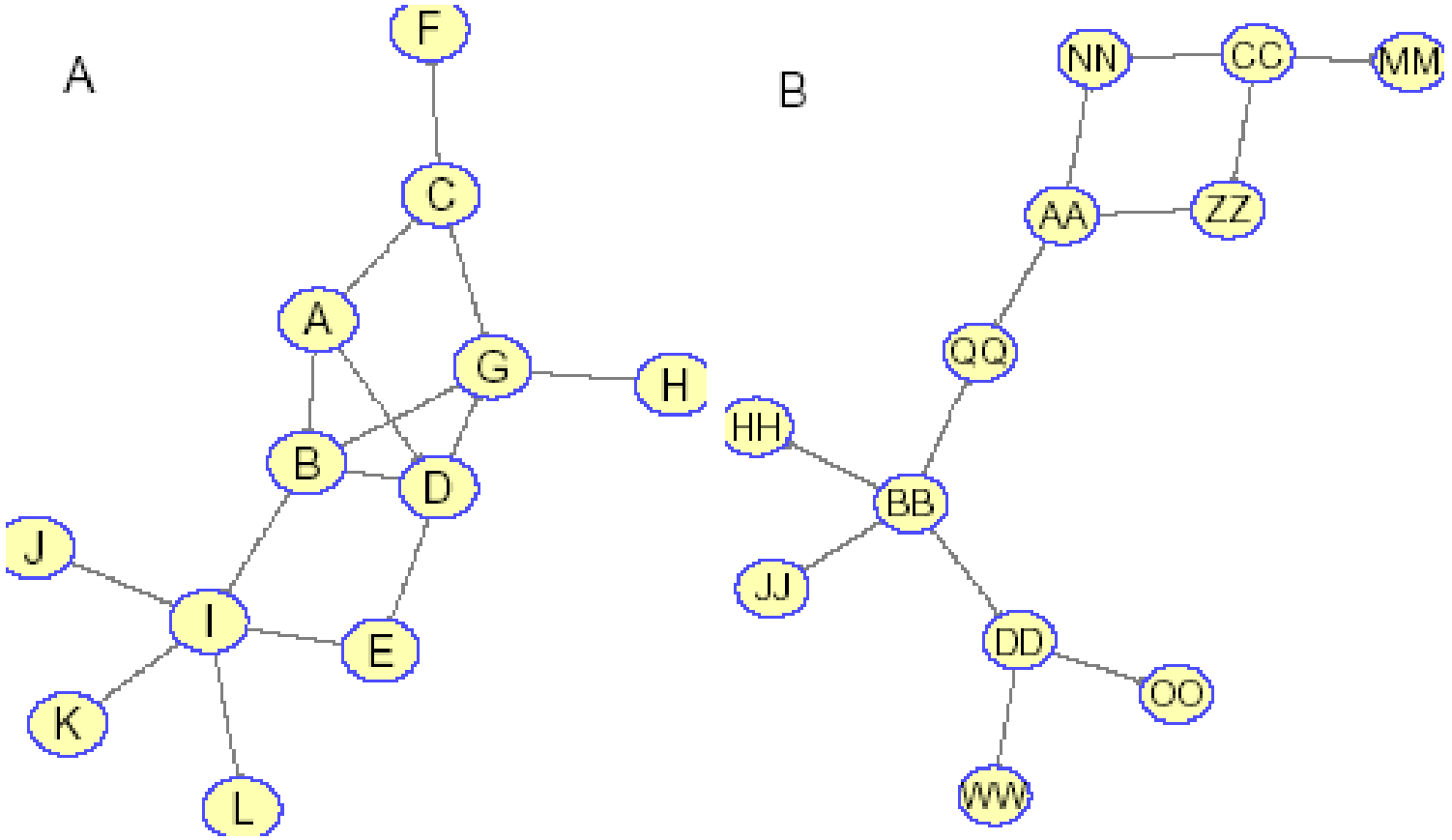
◆**Object function**：**The first term is total node similarity and the second term is the edge similarity.**

◆**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)



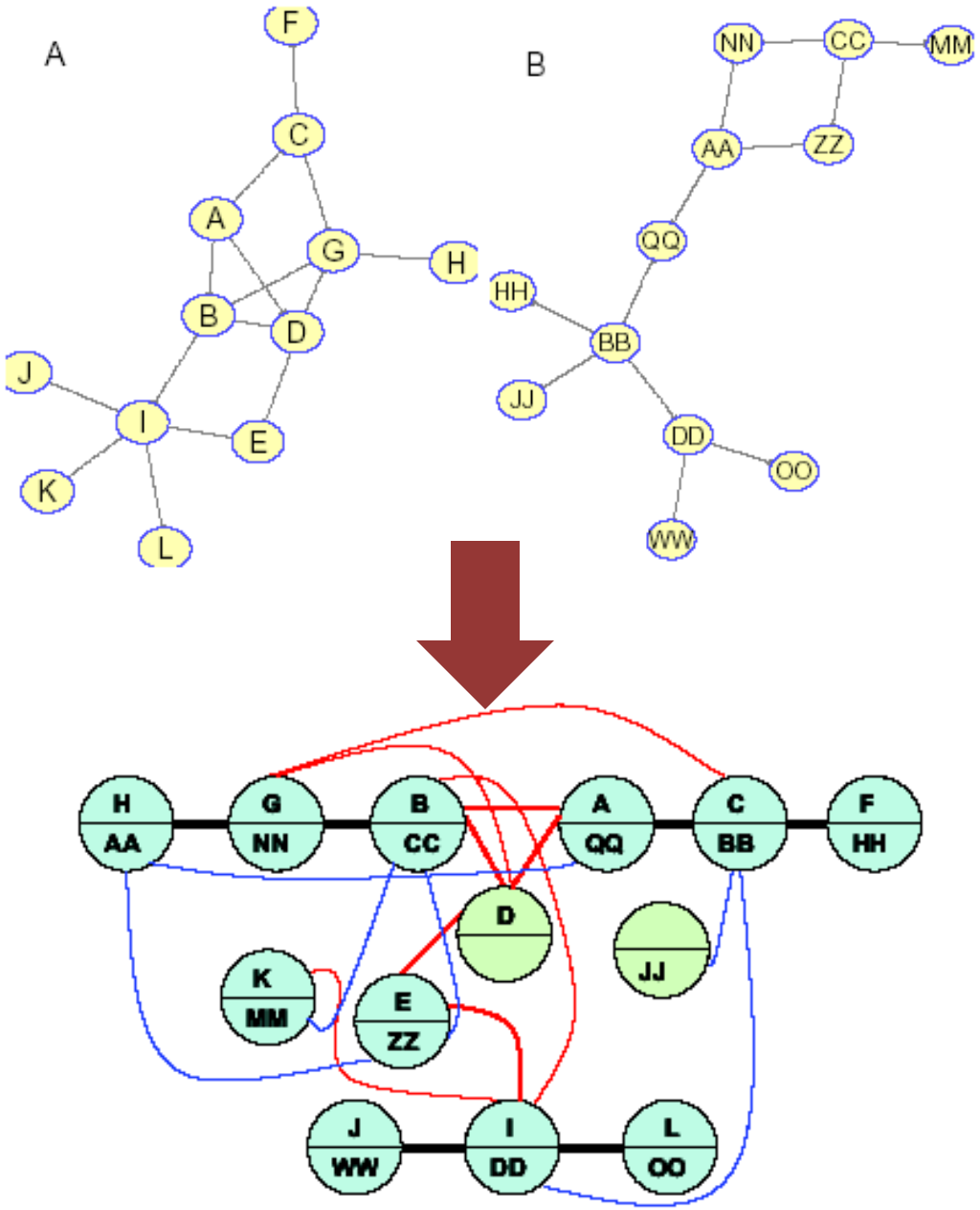
Adjacent matrix

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

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

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

We can align two directed networks

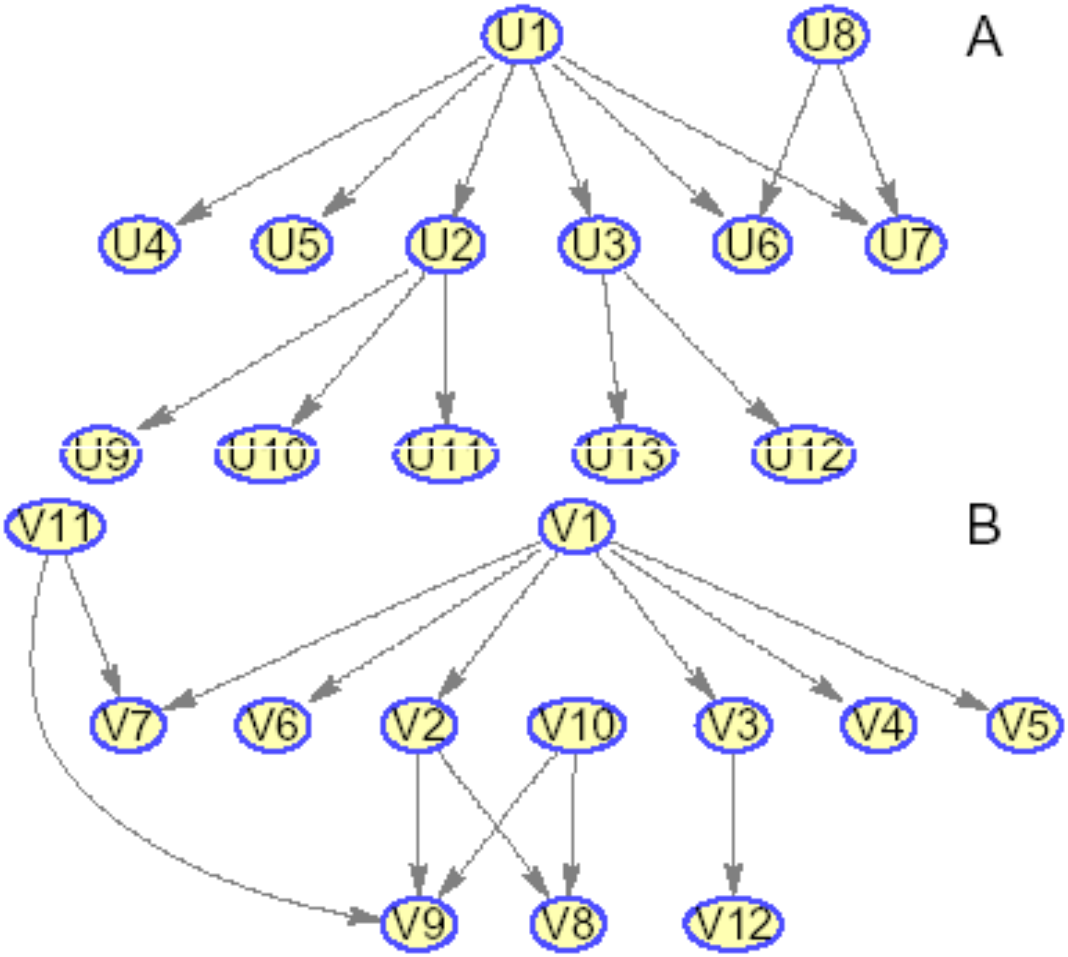
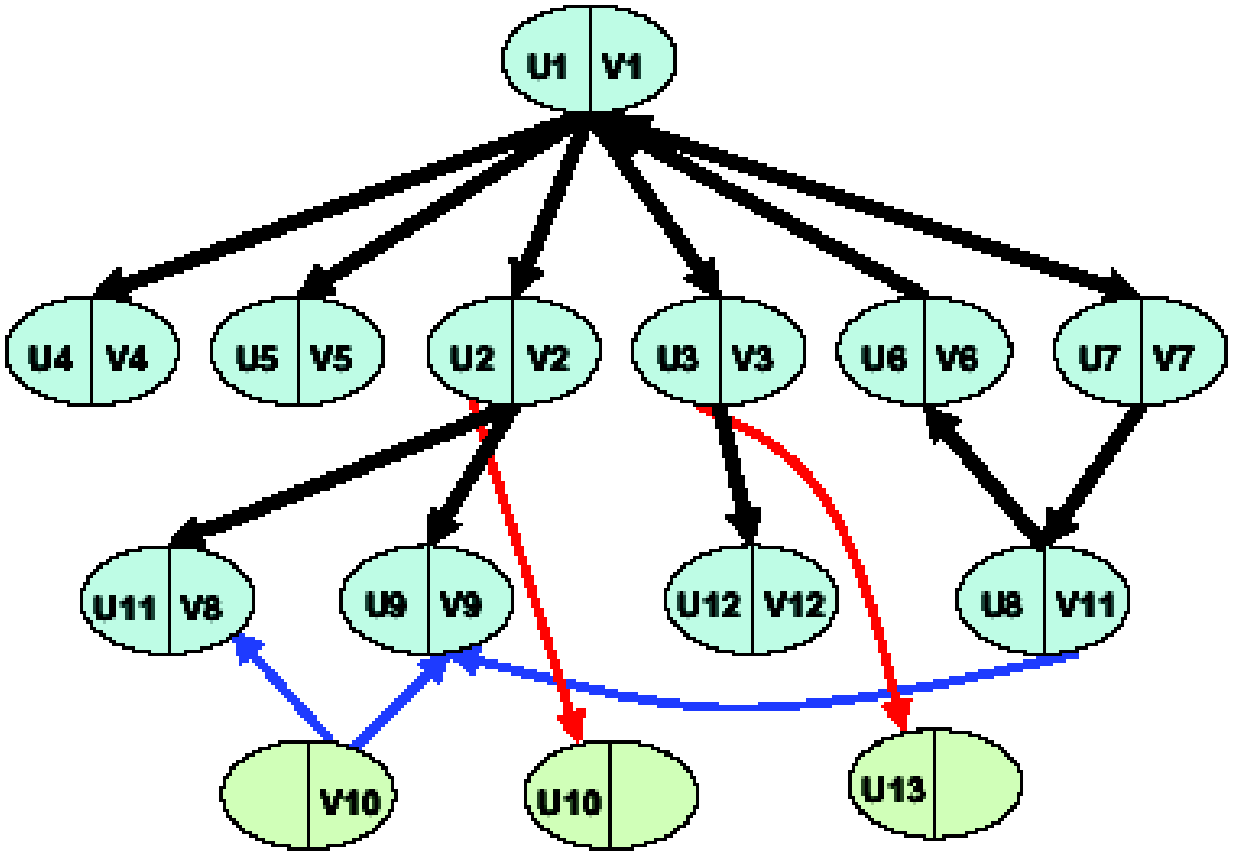
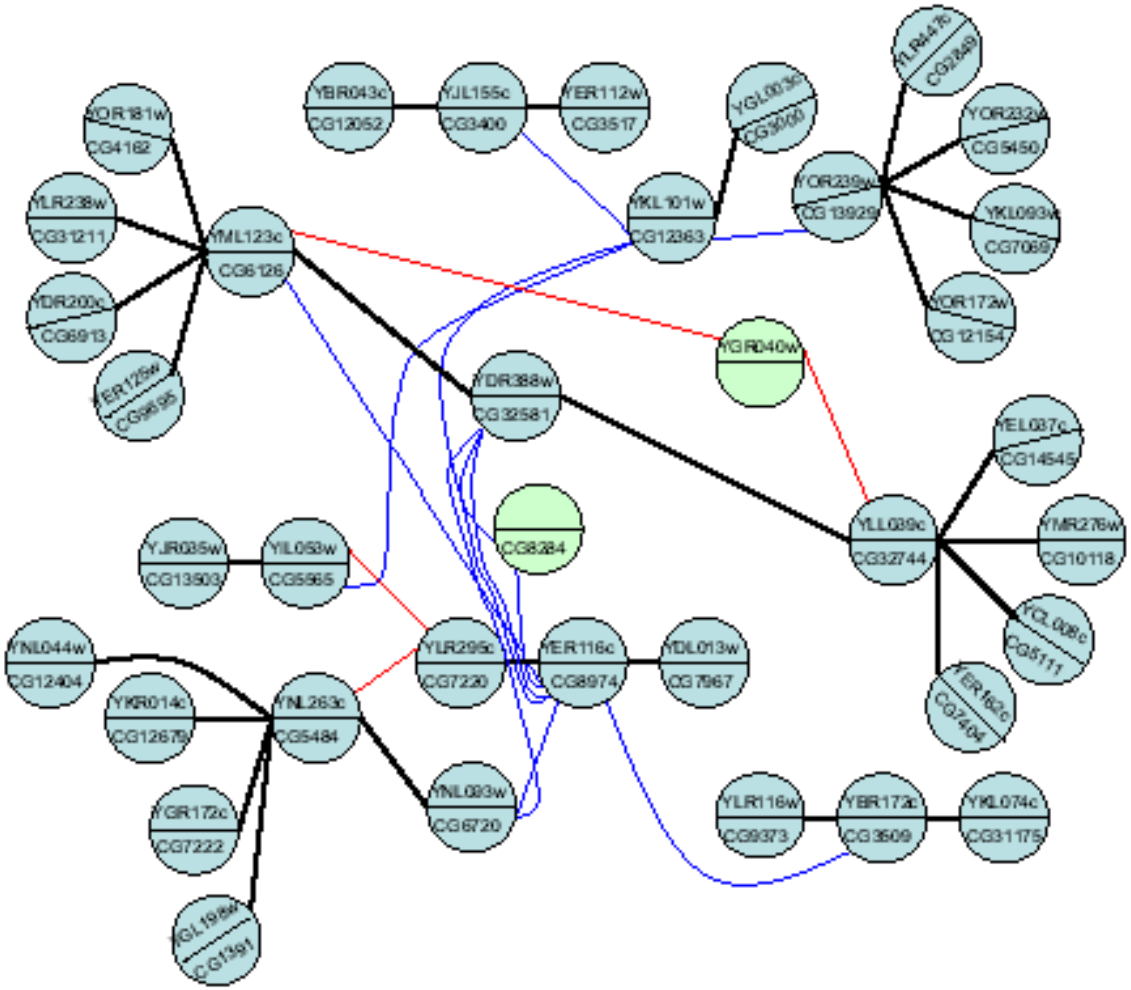


Fig. 2. The simulated example of two directed networks

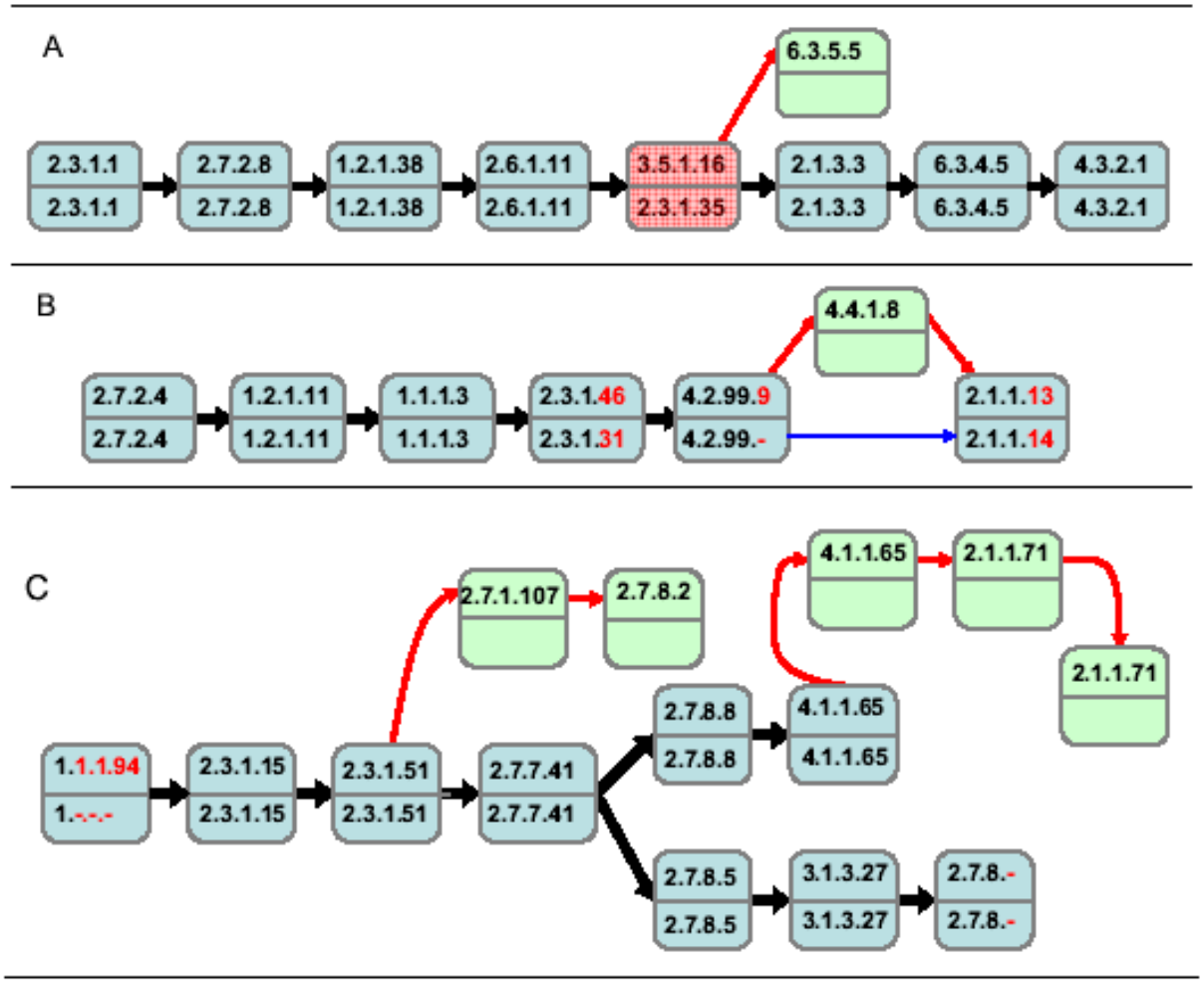
We can align two directed networks



Example on PPI networks



Metabolic pathway alignment



Network comparison globally

- Directly to find the isomorphism is NP-complete, thus this measure can not be used to practically test similarity of two networks.
- 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			

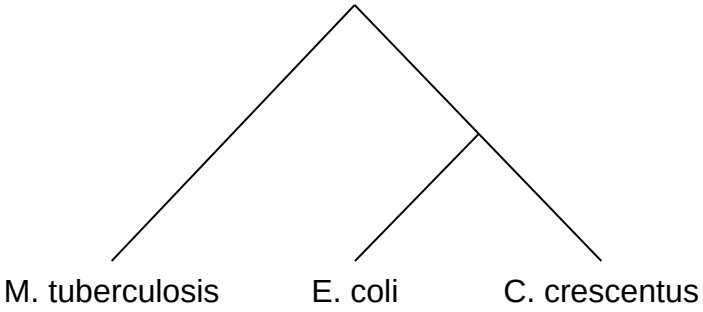
Yu H, Xia Y, Trifonov V, Gerstein M. **Design principles of molecular networks revealed by global comparisons and composite motifs**. *Genome Biology* 7: R55 (2006).

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

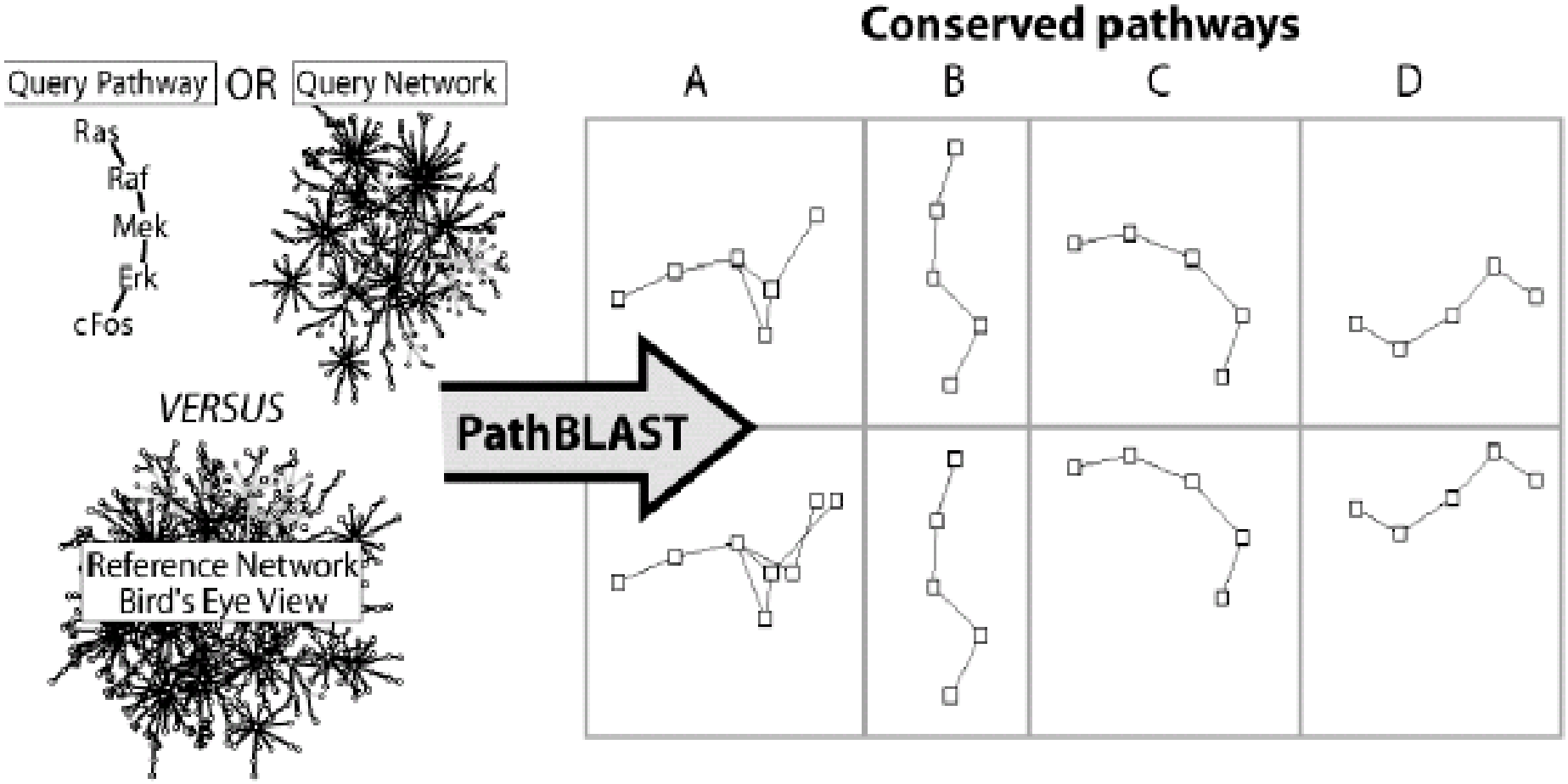
Multiple Alignment?

- 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



Take-home messages

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

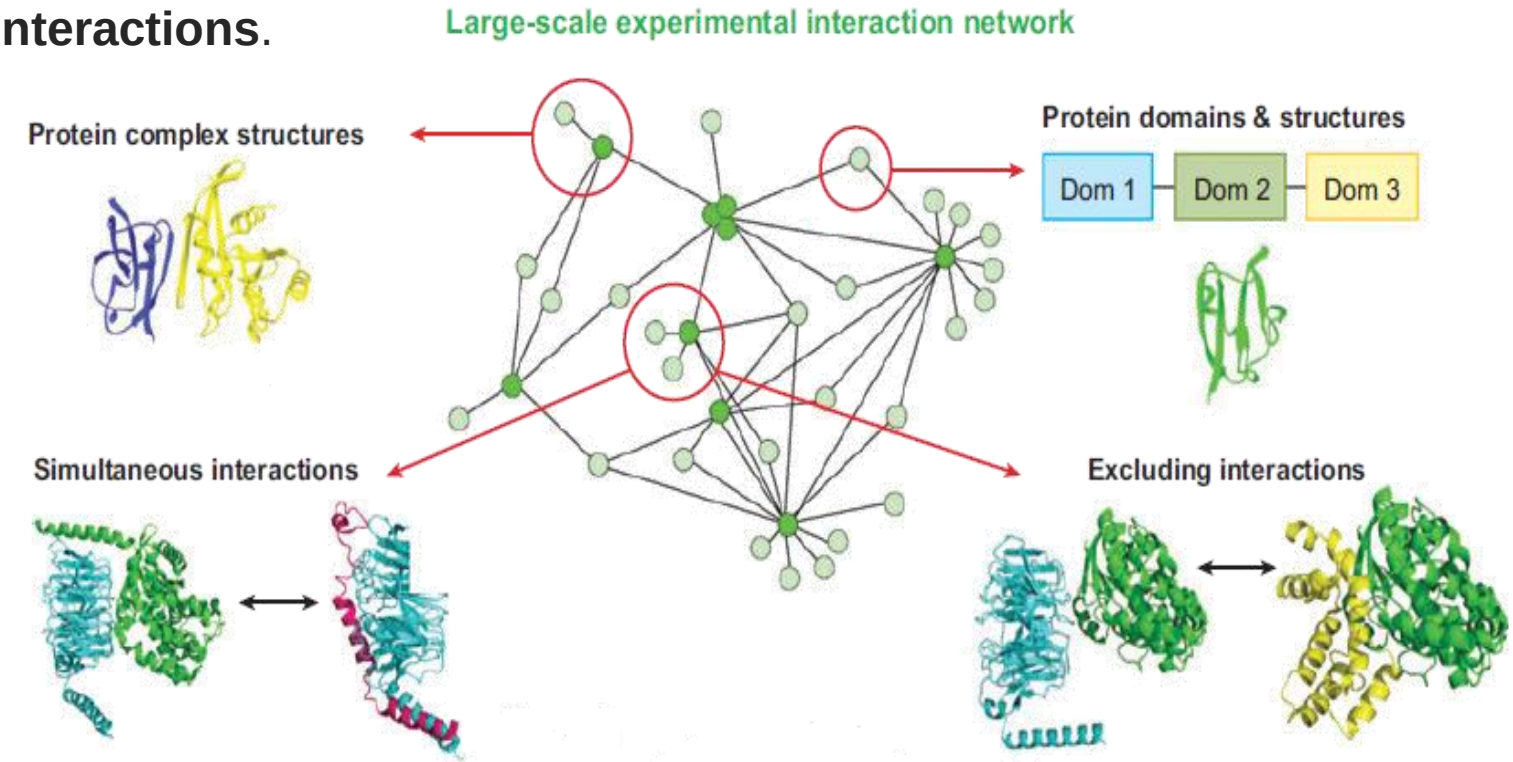
Simultaneous fitting of assembly components into cryo-EM density maps

Overview

- Background
- Simultaneous fitting problem
- Our method
  - Vector quantization
  - Integer Quadratic Programming (IQP)
  - Scoring of candidate structures
  - Weighted ICP refinement
- Results
- Summary

Background

The function of biological macromolecules is often driven by their interactions.

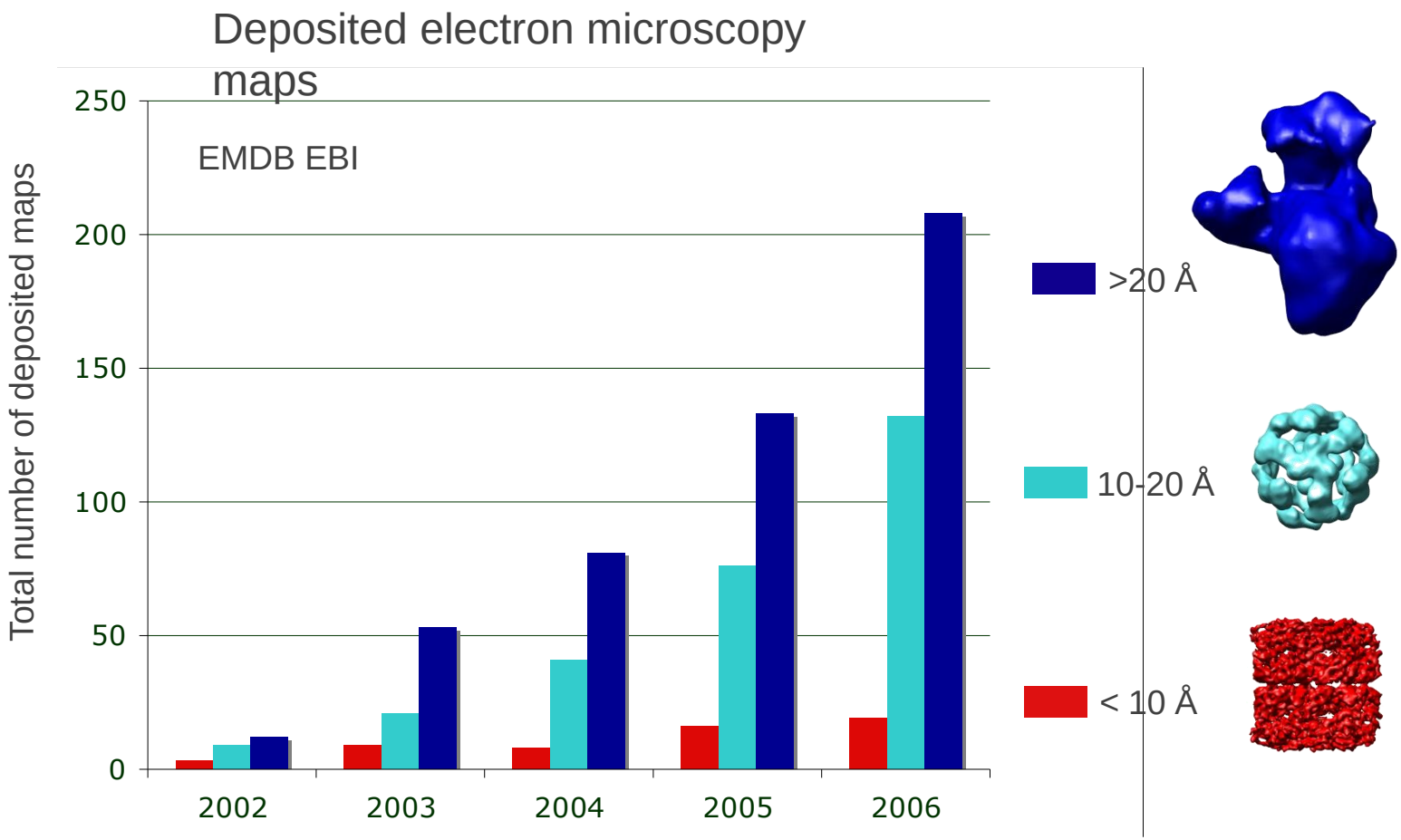


-- There are thousands of protein assemblies/complexes with unknown structure.

-- Structure determination of assemblies is difficult because of the limitations in experimental technologies.

Cryo-electron microscopy (Cryo-EM) is a promising tool to generate low-resolution (>4 Å) density maps of large protein assemblies.

Background: cryo-EM maps



Sequential fitting problem

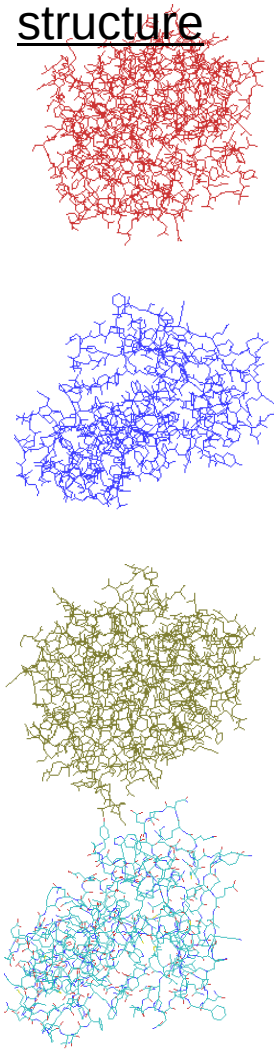
Component structure

a

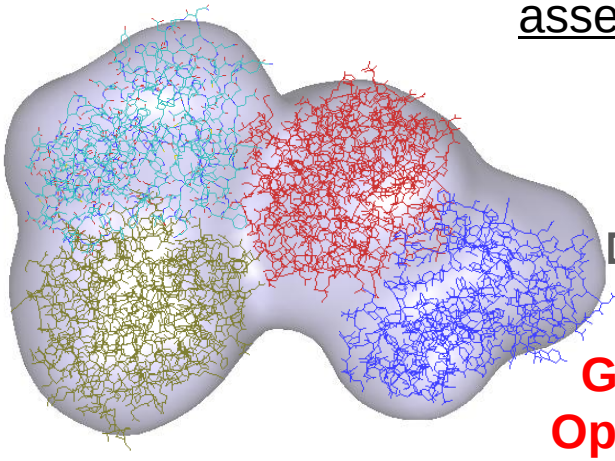
b

c

d

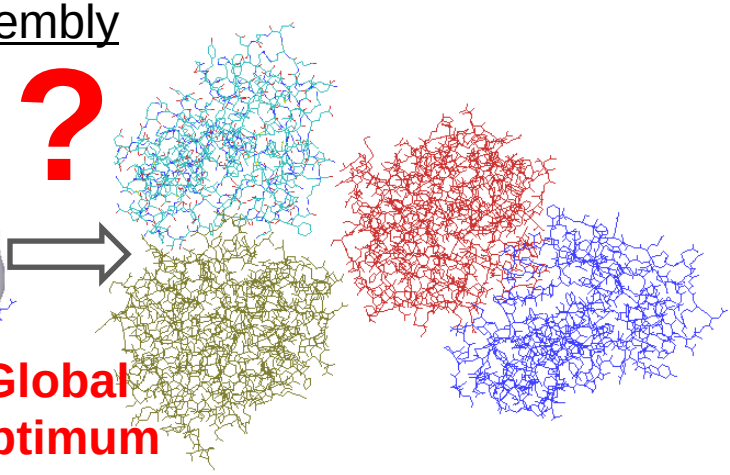


3D density map of assembly



Pesudo-atomic structure of assembly

?

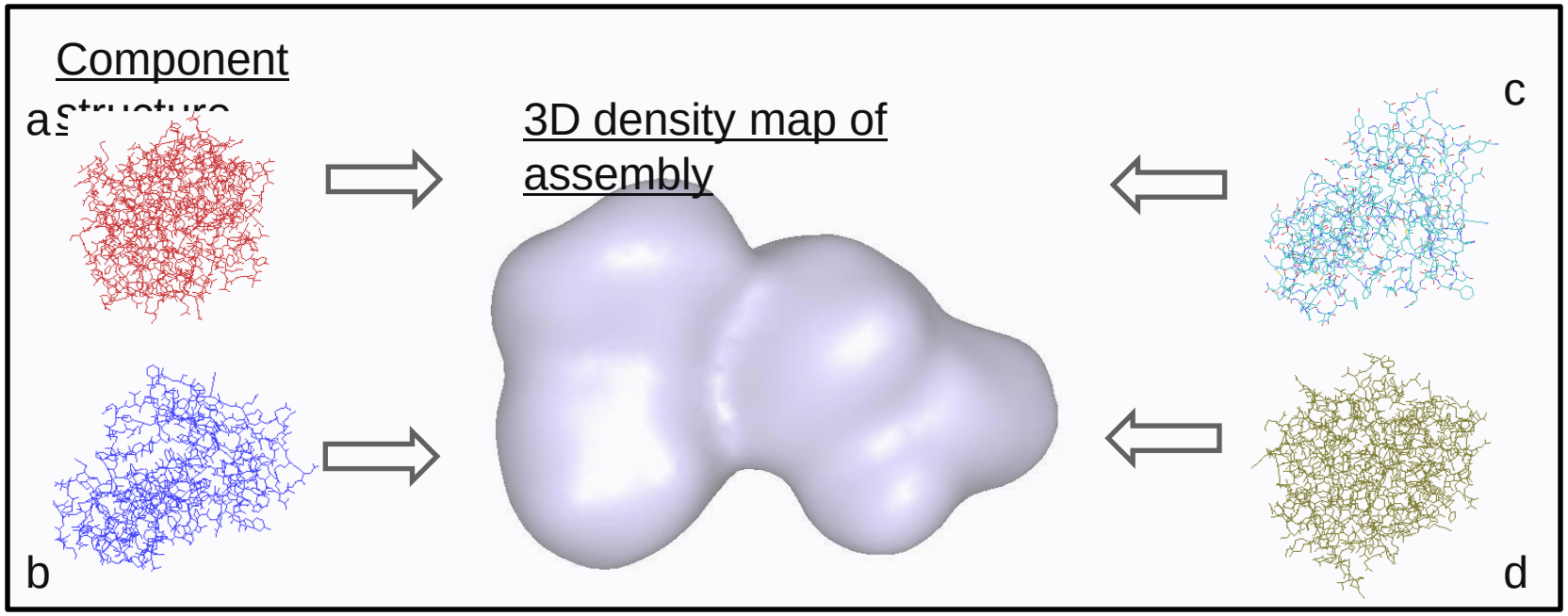


Global Optimum

Scoring function: cross-correlation score between the EM map of the assembly and the atomic structure at similar resolution.

Search method: exhaustive search over all possible orientations.

Simultaneous fitting

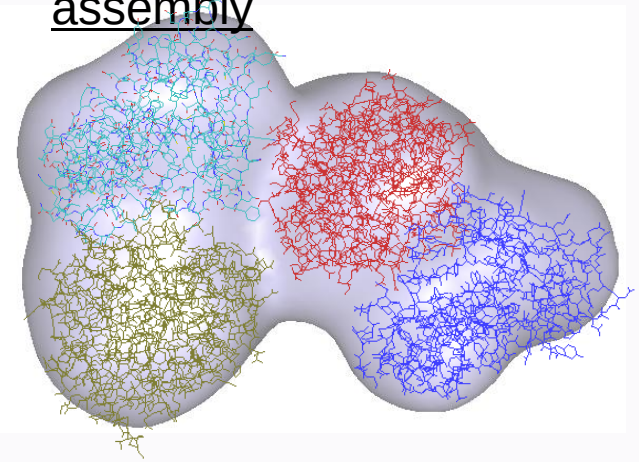


Simultaneous fitting

a Component structure

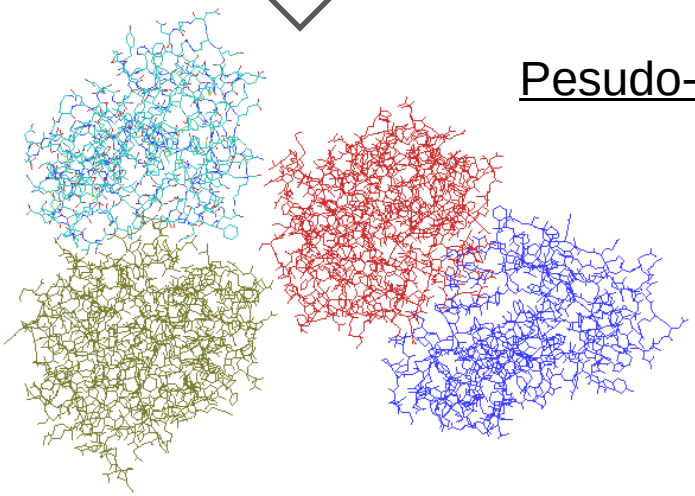
3D density map of assembly

c



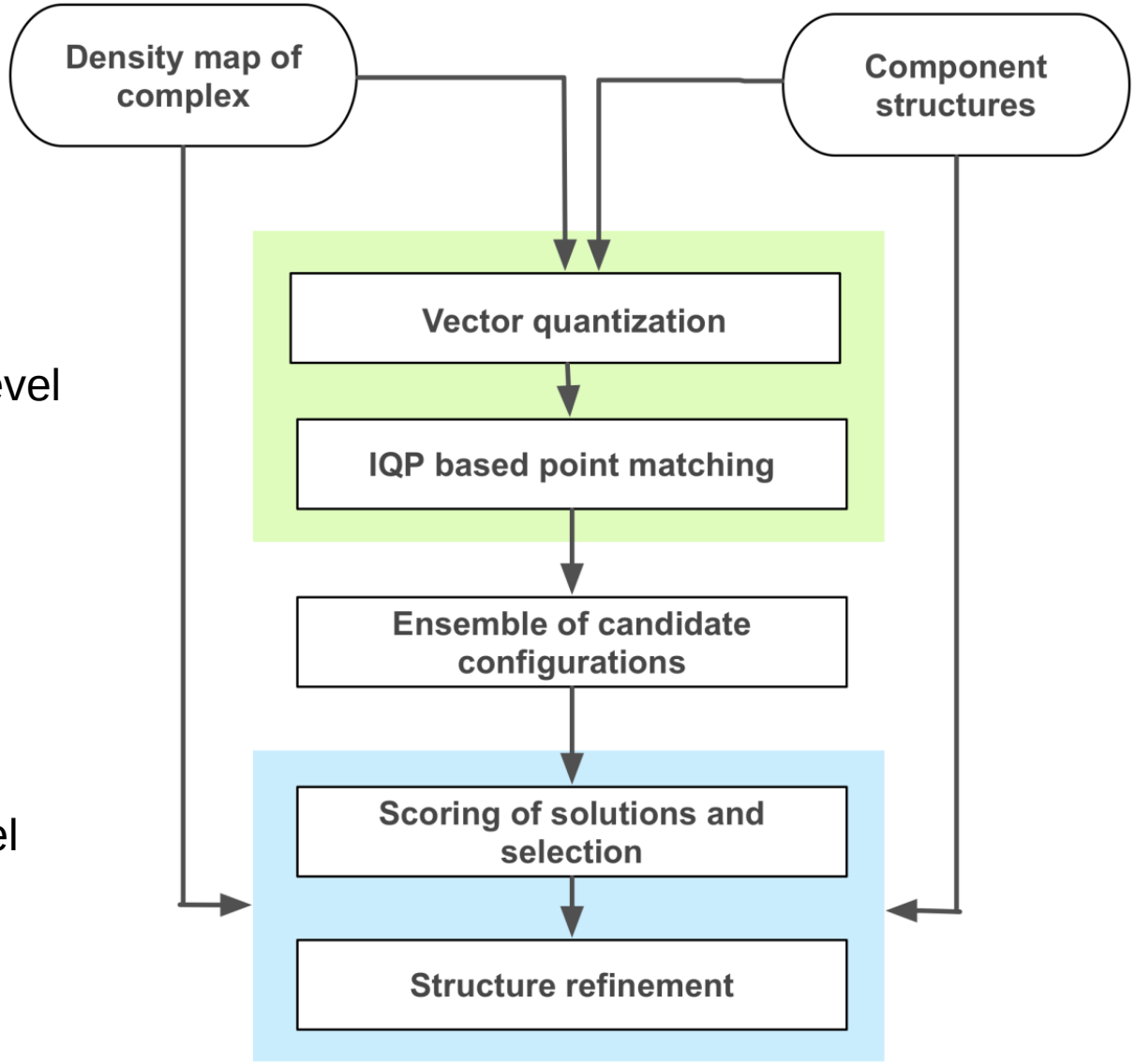
b

d



Pesudo-atomic structure of assembly

Simultaneous fitting method



Reduced representation using feature points

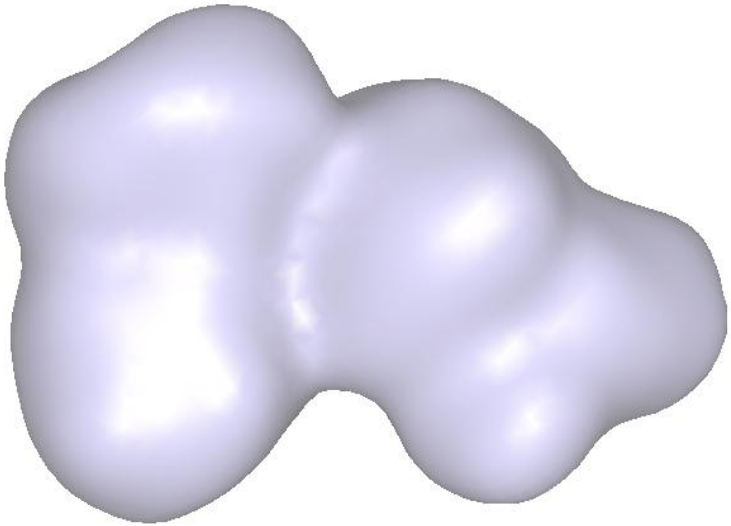
Vector Quantization

IQP Point Matching

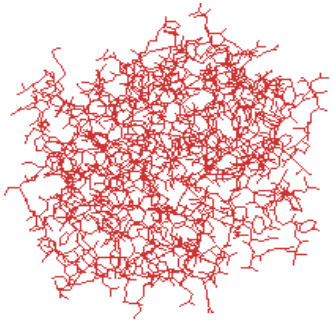
Independent Scoring System

Weighted ICP Refinement

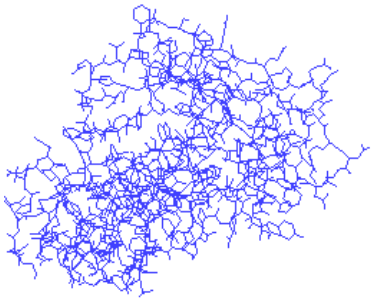
Assembly map



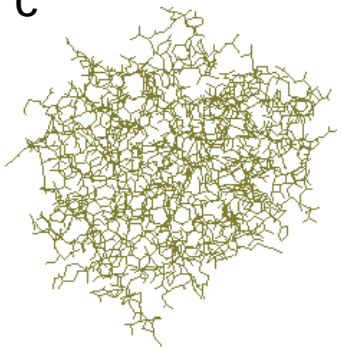
a



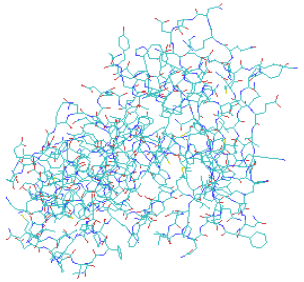
b



c



d

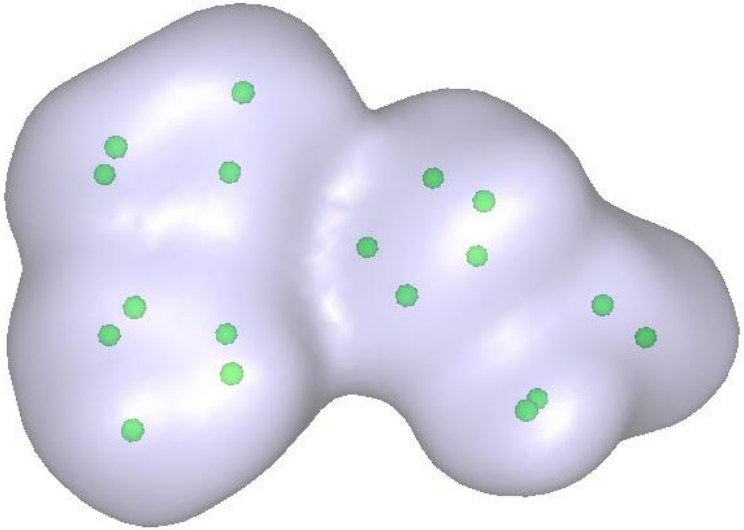


Component structure

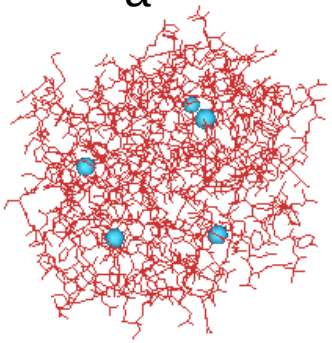
Reduced representation using feature points

- Vector Quantization
- IQP Point Matching
- Independent Scoring System
- Weighted ICP Refinement

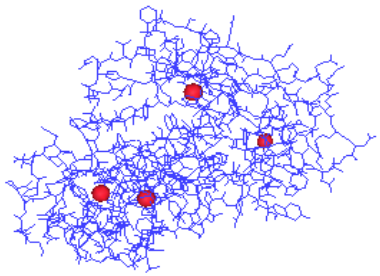
Assembly map



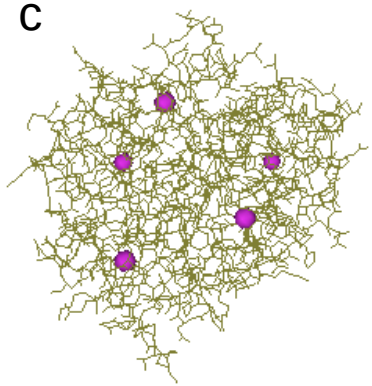
a



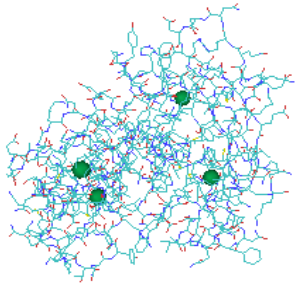
b



c



d

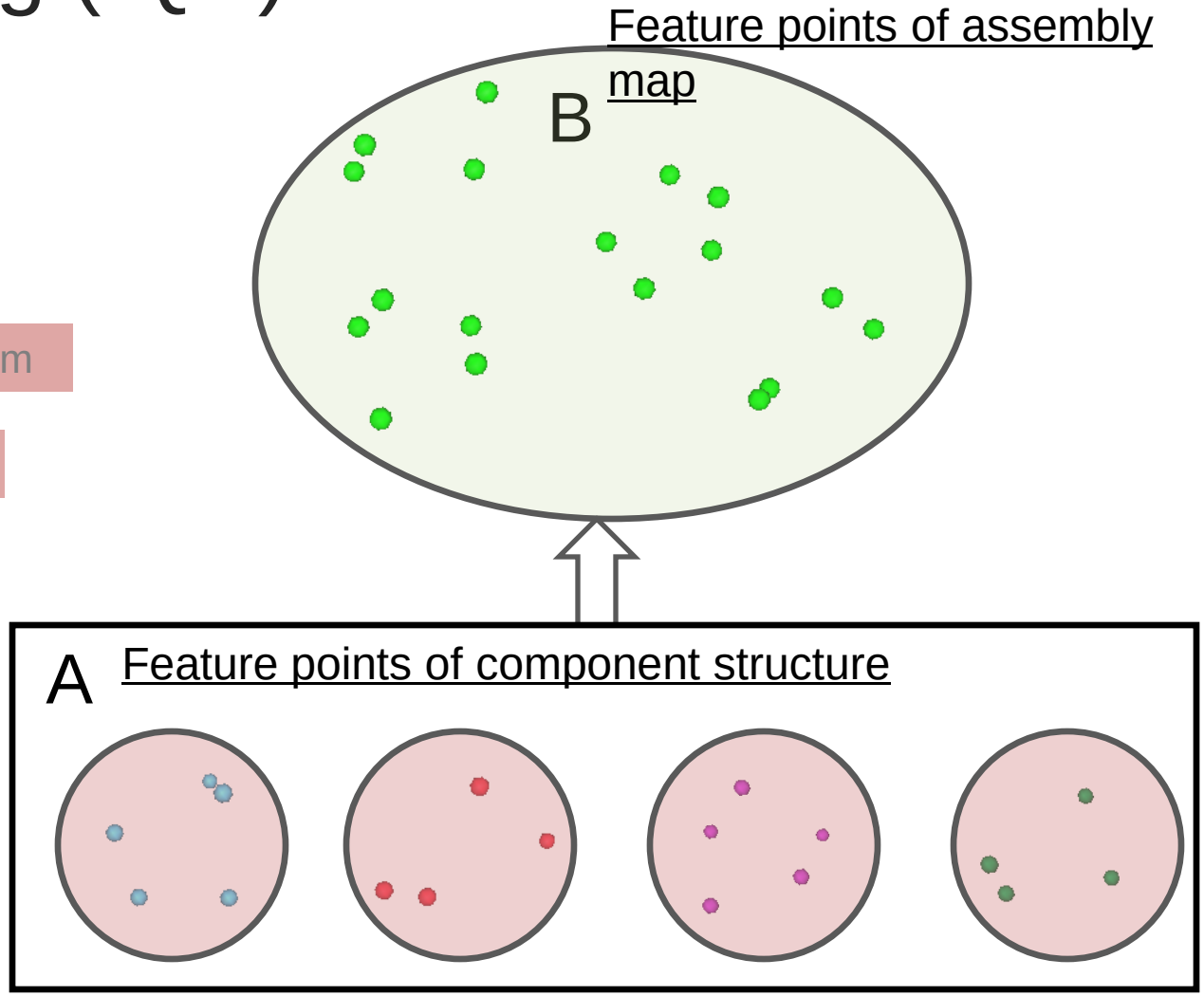


Component structure

Wriggers, W. et al. (1998)

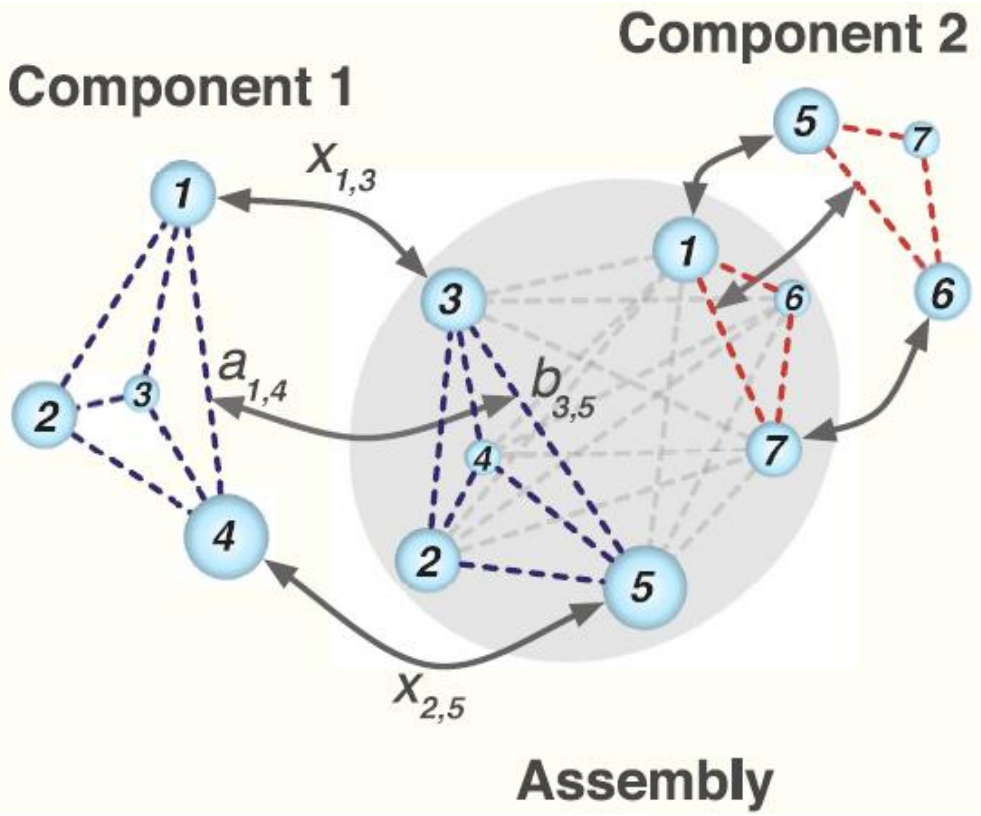
Point matching by Integer Quadratic Programming (IQP)

- Vector Quantization
- IQP Point Matching
- Independent Scoring System
- Weighted ICP Refinement



Integer Quadratic Programming (IQP) method for point matching

Point matching by Integer Quadratic Programming (IQP)



Point matching based on: 1) **Geometric distance**
2) **Density information**

Integer Quadratic Programming (IQP)

The matching problem is to maximize a similarity score $F(U_1,U_2,V_1,V_2)$ between point sets V_1 and V_2 with density values U_1 and U_2 among all feasible combinations X. A solution can be found using integer quadratic programming (IQP).

The objective function is subject to three constraints. First, each point in V_1 can match at most one point in V_2. Second, each point in V_2 can match at most one point in V_1. Third, the variable x_{ij} is binary.

$$\max_X \; F(U_1,U_2,V_1,V_2)=\sum_{i=1}\sum_{j=1} S(\mathbf{u}_i^1,\mathbf{u}_j^2)x_{ij} + \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{l=1}^{n} G(a_{ik},b_{jl})x_{ij}x_{kl}$$

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

where $S(a,b)=G(a,b)=e^{-\frac{2\times|a-b|}{a+b}}$

Scoring of candidate structures

Vector Quantization

IQP Point Matching

Independent Scoring System

Weighted ICP Refinement

An *ensemble* of candidate structures is generated by running VQ and IQP multiple times

The best structure is selected using a normalized cross-correlation function

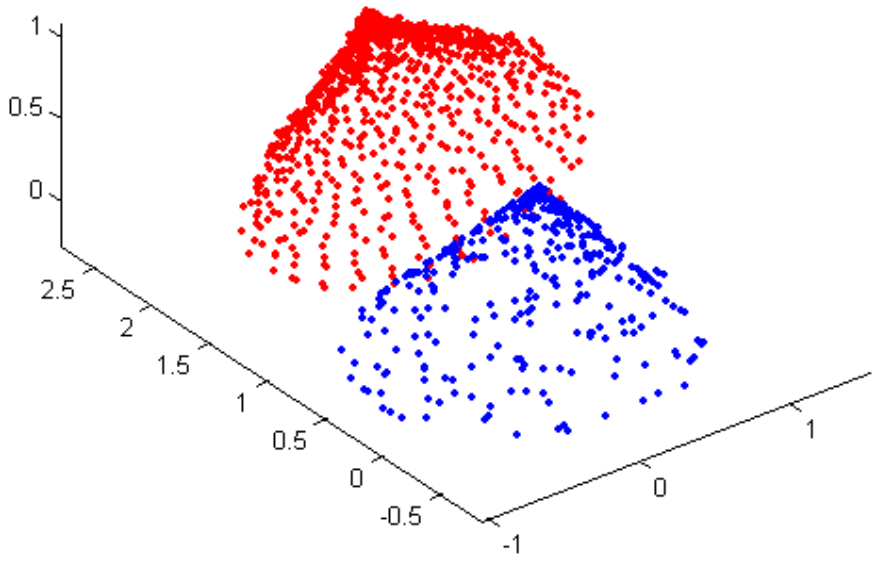
$$CCF = \frac{1}{N}\sum_{i=1}^{N}\frac{(\rho^t(i) - <\rho^t>)(\rho^p(i) - <\rho^p>)}{\sigma^t\sigma^p}$$

Scoring is based on density maps using a normalized cross-correlation function (CCF)

Chacon and Wriggers (2002), Topf, et al. (2005)

Structure refinement using density maps

- Vector Quantization
- IQP Point Matching
- Independent Scoring System
- Weighted ICP Refinement



Weighted Iterative Closest Point (wICP) registration method for refinement

Weighted ICP refinement

- A refinement of the selected structures using the initial density map is needed to improve the accuracy.
- Each density grid is represented as a point with an associated density value.
- Then we expand the **Iterative Closest Point (ICP) registration** method (Besl and McKay, 1992) to incorporate density map information in the optimization process by introducing a ***weighted*** error metric *wRMSD*.
  - Alternatively iterative method
  - Gradually decrease the error metric.

Zhang et al., ISMB2010

Benchmark set

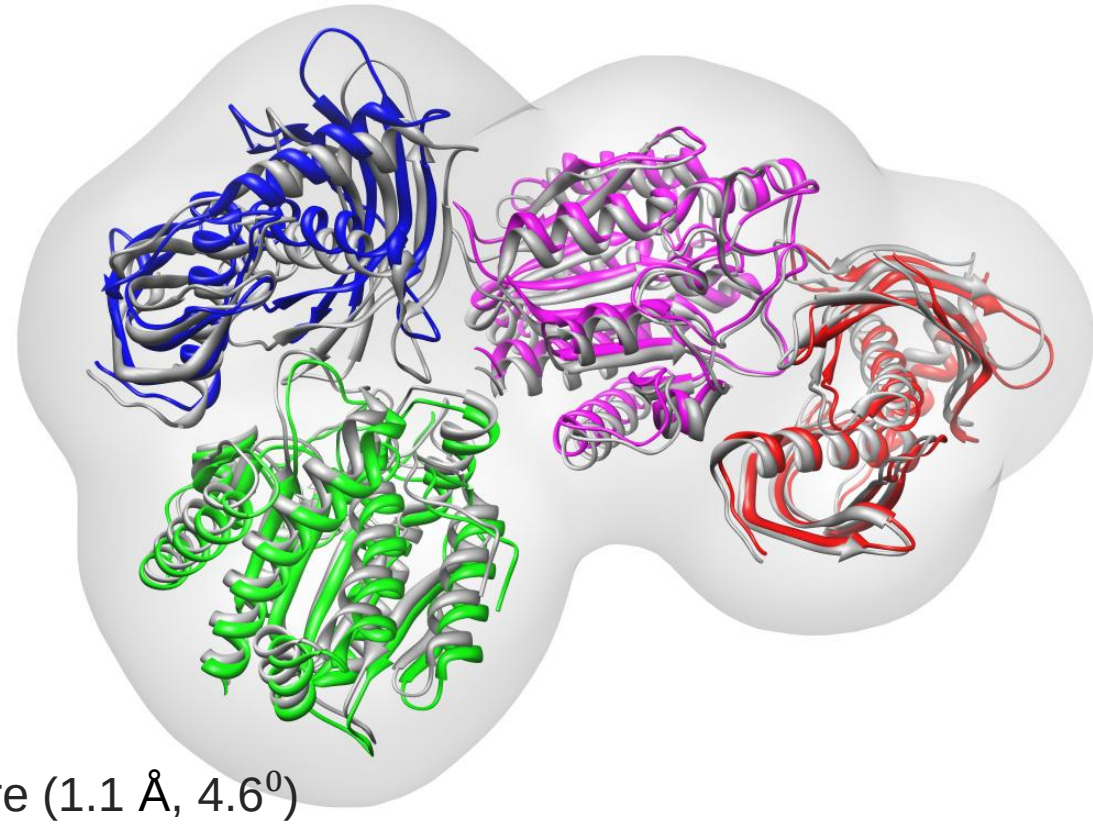
- 11 protein assemblies that are diverse in total size, global shape, number of components (2-7 components), and symmetry.

- Density maps were simulated at 20 Å resolution (PDB2VOL program of the Situs 2.0 package)

Wriggers W. , Biophysical Reviews (2010), 2:21-27.

Results: an example

2BO9: Human carboxypeptidase A4 in complex with human latexin



Assessment:
RMSD (Cα): 1.7Å
RMSD*: 1.1Å
Component placement score (1.1 Å, 4.6^0)
Time (0.75s/1 IQP run)

(Grey ribbon diagram) native assembly, (colour ribbon diagrams) fitted components

Results

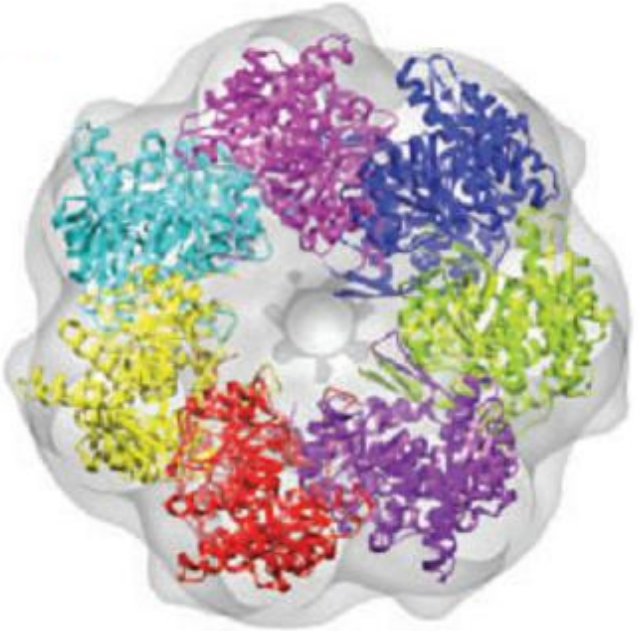
Table 1. Summary

Assembly	Comp.	Lowest RMSD structure				Best CCF ranking structure		
		CCF (Lapl.-CCF)	CPS (Å, °)	RMSD	RMSD*	CPS (Å, °)	RMSD	RMSD*
1DOR	2	2 (1)	(1.1, 6.8)	2.1	1.1	(0.6, 9.5)	2.5	1.2
1AFW	2	2 (1)	(2.3, 14.4)	4.8	0.9	(2.5, 15.0)	4.9	0.9
1PC8	2	6 (10)	(1.1, 3.1)	1.3	0.5	(0.8, 6.4)	1.6	0.5
1TX4	2	8 (6)	(1.2, 2.8)	2.6	0.4	(0.7, 2.9)	3.0	0.4
1NIC	3	1 (1)	(5.6, 5.1)	5.9	1.1	(5.6, 5.1)	5.9	1.1
1CS4	3	8 (7)	(2.4, 24.0)	6.5	1.8	(2.3, 55.5)	12.8	11.7
2DQJ	3	34(11)	(2.0, 21.1)	4.5	1.7	(1.4, 62.1)	9.5	7.8
1F1X	4	2 (18)	(2.4, 14.6)	4.6	0.9	(2.3, 168.4)	28.2	26.1
2BO9	4	1 (1)	(1.1, 4.6)	1.7	1.1	(1.1, 4.6)	1.7	1.1
2REC	6	1 (1)	(1.3, 4.2)	1.7	1.0	(1.3, 4.2)	1.7	1.0
1J2P	7	1 (3)	(1.6, 16.2)	4.4	1.5	(1.6, 16.2)	4.4	1.5

Apo- Gro-EL experimental density map at 23.5 Å resolution

Experimental map

Ranson et al. (2001)



Fitted atomic model,

RMS error of 8.6 Å with respect to the ‘native’ structure

Summary

- We have developed a fast method for ***simultaneous*** fitting of multiple components into cryo-EM density maps of assemblies.
- Our approach relies on a fast mathematical programming method and an efficient refinement procedure.
- Our approach matches two point sets *not only* based on their geometrical equivalence, but also based on the similarity of the density in the immediate point neighborhood.
- In principle our approach allows the integration of other information, *e.g*. the knowledge about specific binding interfaces of a protein interaction.