Global alignment of protein-protein interaction networks by graph matching methods.

Mikhail Zaslavkiy [1] Francis Bach[2]
Jean-Philippe Vert[1]

[1]Mines ParisTech / Institut Curie / INSERM

[2]INRIA / Ecole normale superieure de Paris

Kyushu University, Department of Informatics, July 14, 2009.

Outline

Outline

Functional orthologs

Species 1	Species 2
f_1: MKQALAAADDDDAQ...	y_1: MDDDDALGLLLLA...
f_2:MGDXLLMMAALLLL...	y_2: MHHAAKLLDDAS...
...	...

Definition

Functional orthologs are pairs of proteins directly inherited from a common ancestor and which play functionally equivalent roles.

Our goal

Automatic identification of functional orthologs (useful for annotation transfer)

Identification of functional orthologs by best-best hit

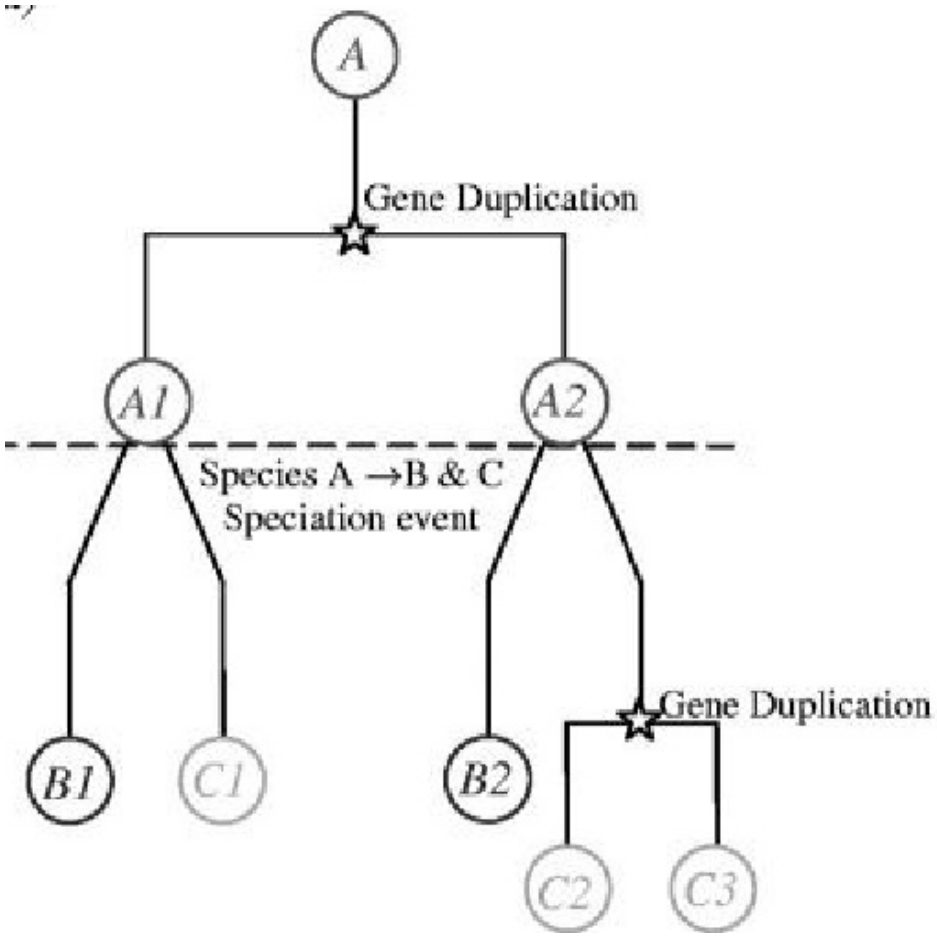
Species 1	Species 2
f_1: MKQDLARIEQFLDALF...	y_1: MSRLPVLLLLQLLVRGA...
f_2: MSKLKIAVSDSCPDCF...	y_2: MELAALCRAGLLLALDA...
...	...

$$\mathbf{C} = \begin{array}{c|cc} & \mathbf{y_1} & \mathbf{y_2} \\ \hline \mathbf{f_1} & 10 & 50 \\ \mathbf{f_2} & 27 & 10 \end{array}$$

C_{ij}-BLAST similarity scores

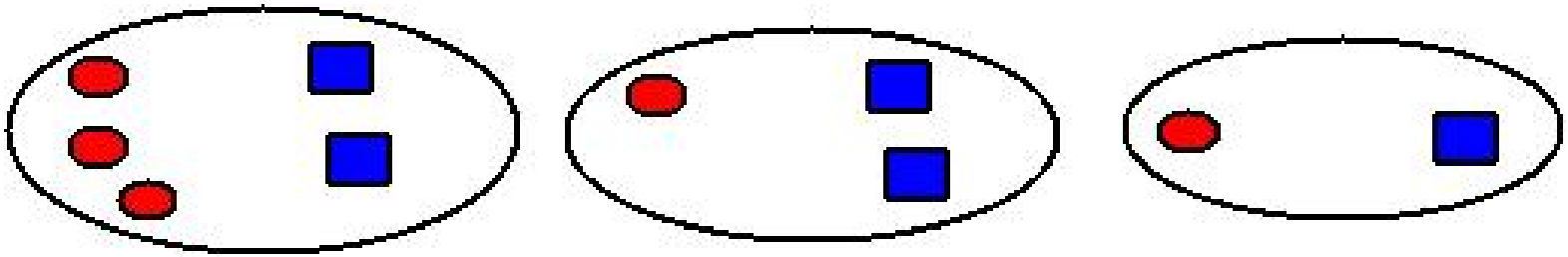
Optimal assignment : $f_1 \rightarrow y_2,\ f_2 \rightarrow y_1$

Limitations of sequence comparison-based methods



- **y** may be the best hit for **f**, but **f** may not be the best hit for **y**...
- (y_1, f) **and** (y_2, f) may produce very similar blast scores...

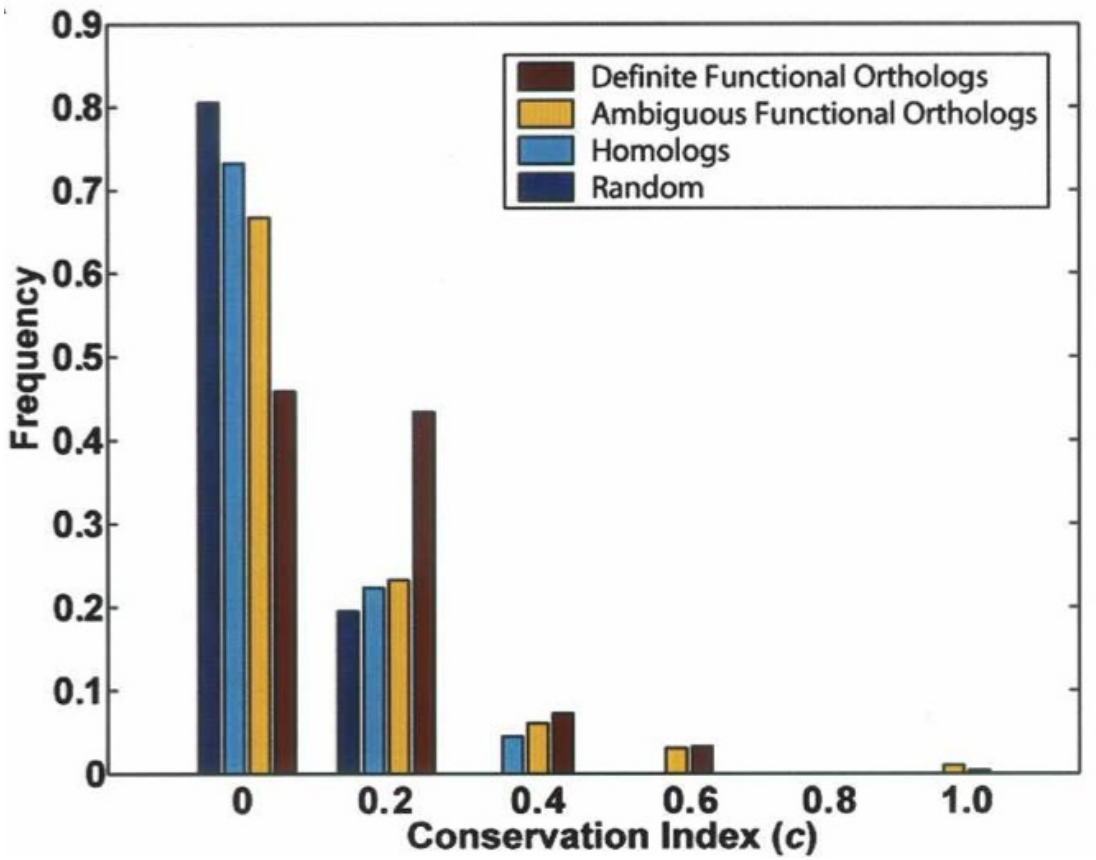
Clusters of orthologs



- Many programs produce clusters of orthologous genes from sequence comparison only (COG, KEGG, Inparanoid, ...)
- Several genes of each species may be in the same cluster
- How to find functional orthologs within the clusters?

Ideas to solve ambiguous functional orthologs

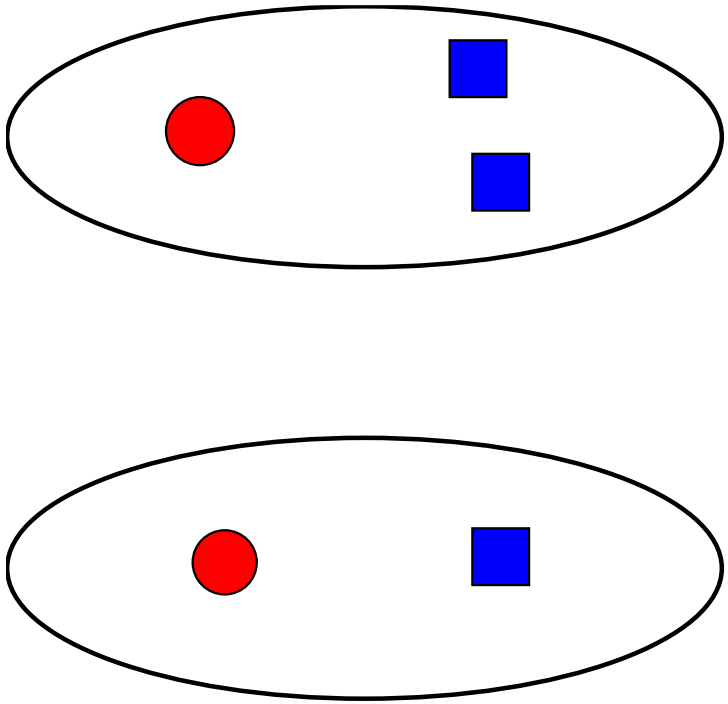
- Increase the similarity of similarity scores / phylogenetic approaches
- Comparison of expression profiles across species
- Functional orthologs tend to have more conserved protein-protein interactions (PPI) across species



(Bandyopadhyay et al., 2006)

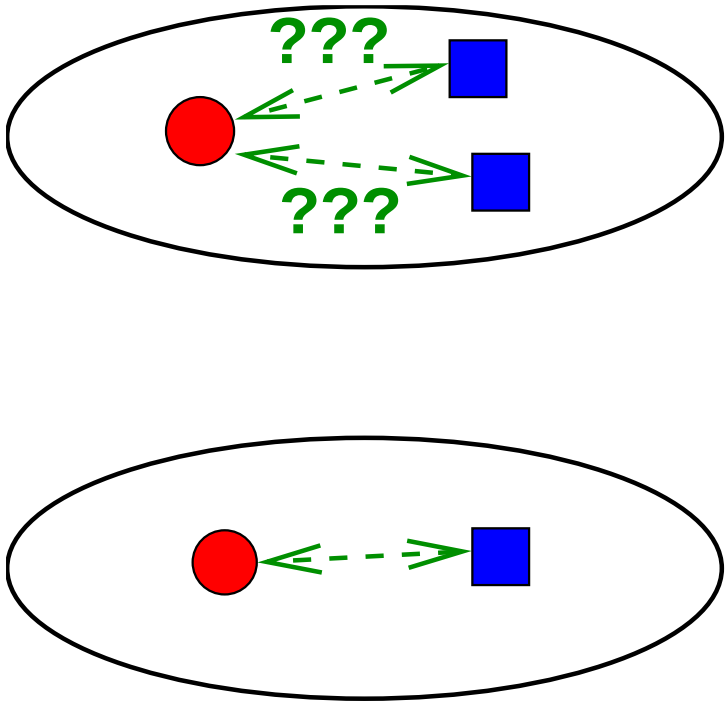
Disambiguation by PPI conservation

Idea: If we know that y^* and f^* are functional orthologs, and there exist interactions $f^* - f$ and $y^* - y_2$. Then the assignment $y_2 - f$ is more likely because it conserves one interaction.



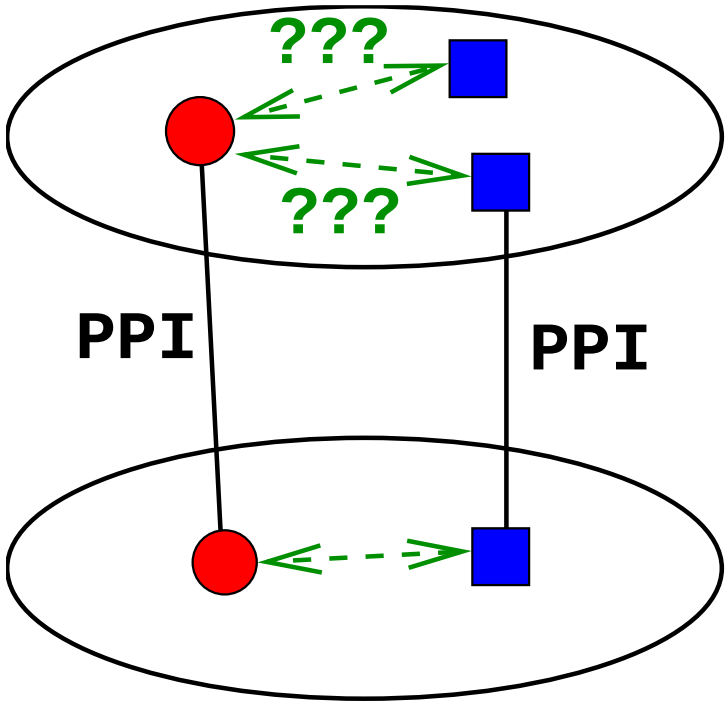
Disambiguation by PPI conservation

Idea: If we know that y^* and f^* are functional orthologs, and there exist interactions $f^* - f$ and $y^* - y_2$. Then the assignment $y_2 - f$ is more likely because it conserves one interaction.



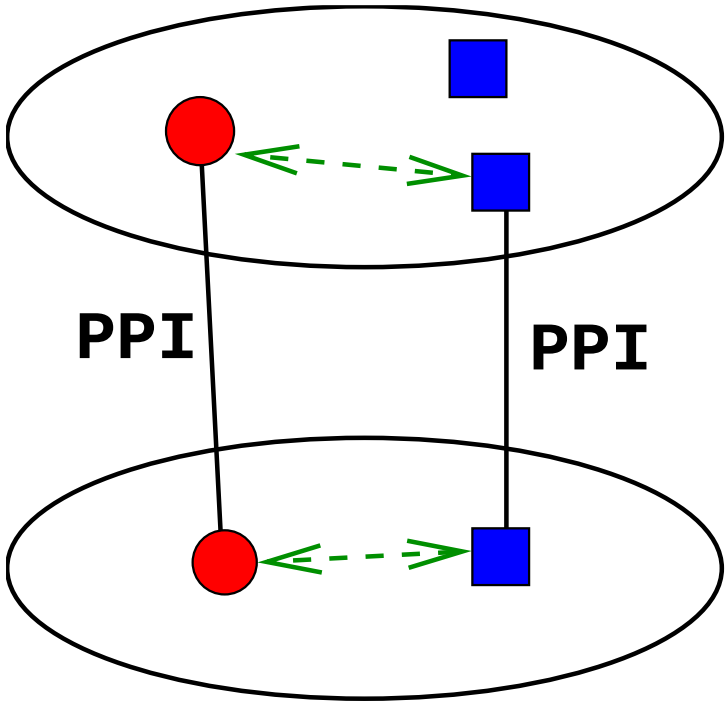
Disambiguation by PPI conservation

Idea: If we know that y^* and f^* are functional orthologs, and there exist interactions $f^* - f$ and $y^* - y_2$. Then the assignment $y_2 - f$ is more likely because it conserves one interaction.

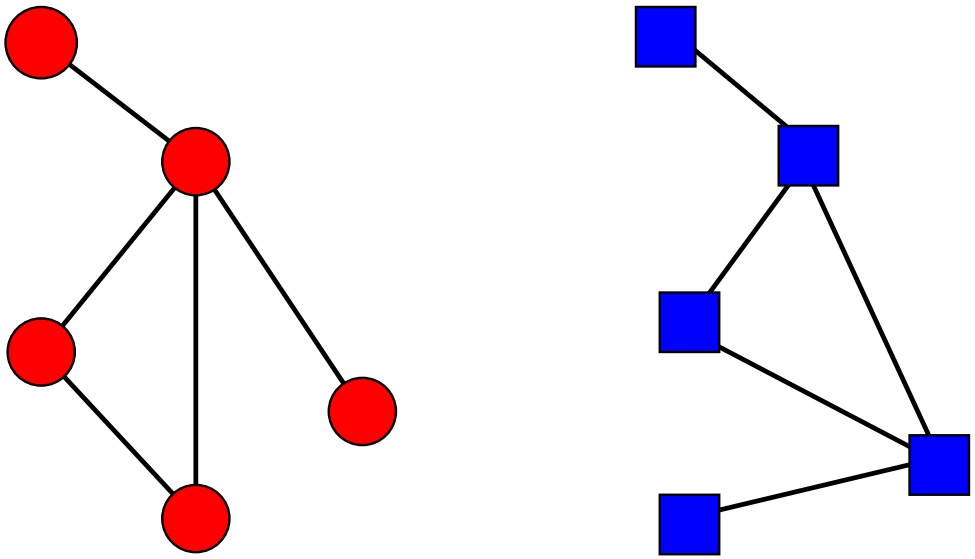


Disambiguation by PPI conservation

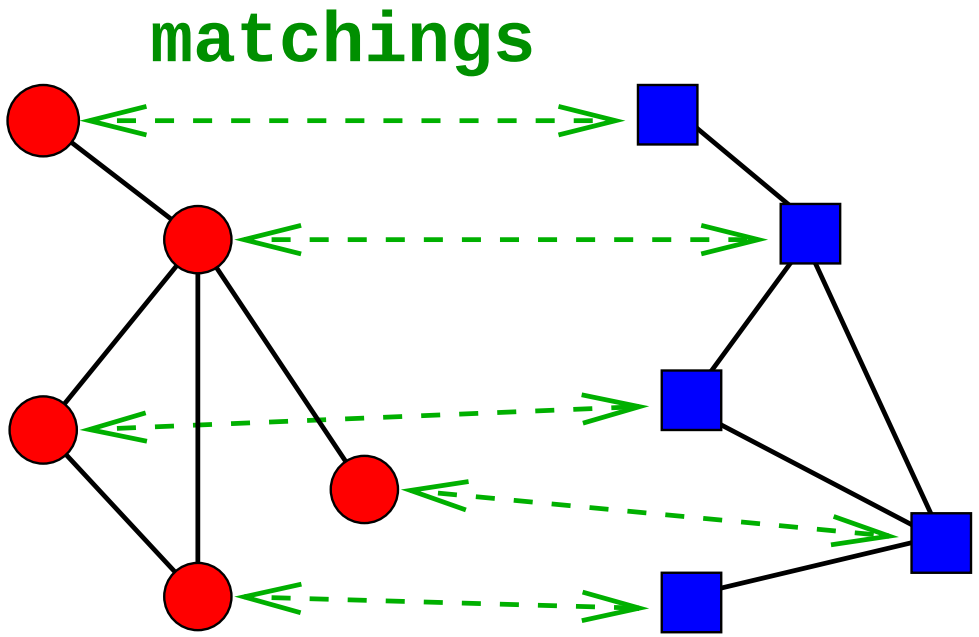
Idea: If we know that y^* and f^* are functional orthologs, and there exist interactions $f^* - f$ and $y^* - y_2$. Then the assignment $y_2 - f$ is more likely because it conserves one interaction.



Extension to PPI networks

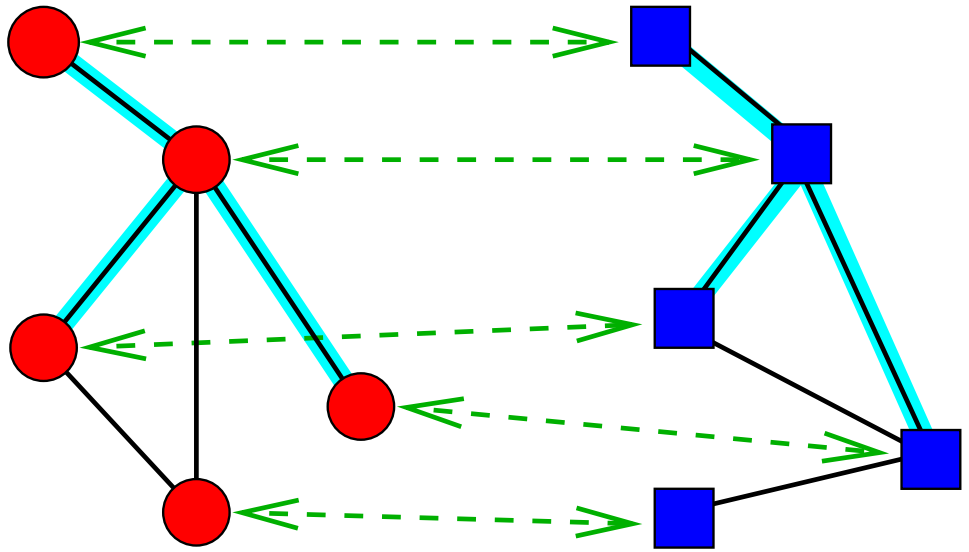


Extension to PPI networks

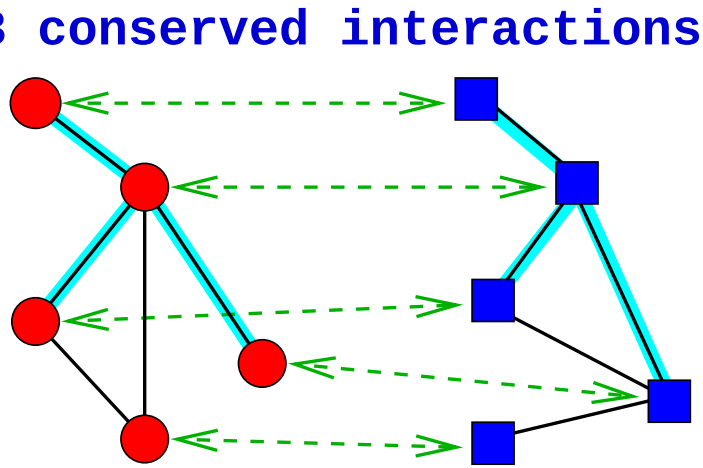


Extension to PPI networks

3 conserved interactions



Global Network Alignment (GNA)



Given two PPI networks and the all-vs-all sequence similarity matrix, find a global matching that maximizes the number of conserved interactions subject to:

- Constraint GNA: matchings only occur within clusters of orthologs.
- Balanced GNA: the mean sequence similarity between matched pairs is as large as possible.

Complexity of the problems (bad news)

- Both problems are NP-hard for general graphs and similarity matrix.
- Therefore we must use algorithms that approximately optimize the criteria, e.g:
  - MRF method (Bandyopadhyay et al., *MSB 2006*) for constrained GNA
  - IsoRank (Singh et al., *PNAS 2008*) for balanced GNA
- *We investigate other algorithms for these problems, borrowing ideas from state-of-the-art graph matching algorithms.*

Complexity of the problems (bad news)

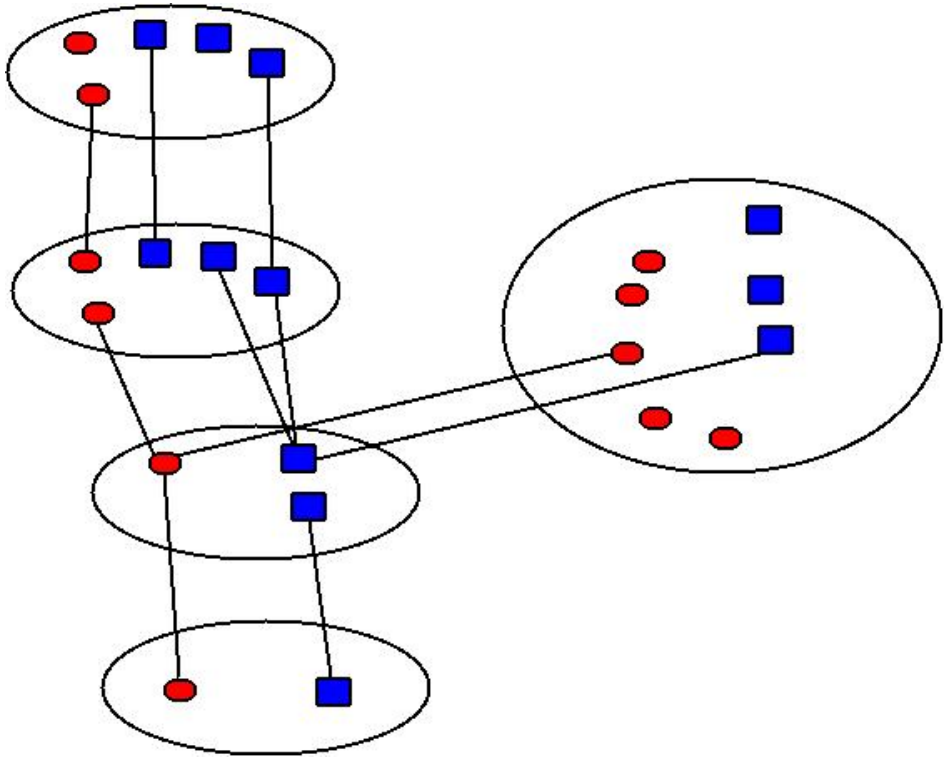
- Both problems are NP-hard for general graphs and similarity matrix.
- Therefore we must use algorithms that approximately optimize the criteria, e.g:
  - MRF method (Bandyopadhyay et al., *MSB 2006*) for constrained GNA
  - IsoRank (Singh et al., *PNAS 2008*) for balanced GNA
- *We investigate other algorithms for these problems, borrowing ideas from state-of-the-art graph matching algorithms.*

Complexity of the problems (bad news)

- Both problems are NP-hard for general graphs and similarity matrix.
- Therefore we must use algorithms that approximately optimize the criteria, e.g:
  - MRF method (Bandyopadhyay et al., *MSB 2006*) for constrained GNA
  - IsoRank (Singh et al., *PNAS 2008*) for balanced GNA
- *We investigate other algorithms for these problems, borrowing ideas from state-of-the-art graph matching algorithms.*

Outline

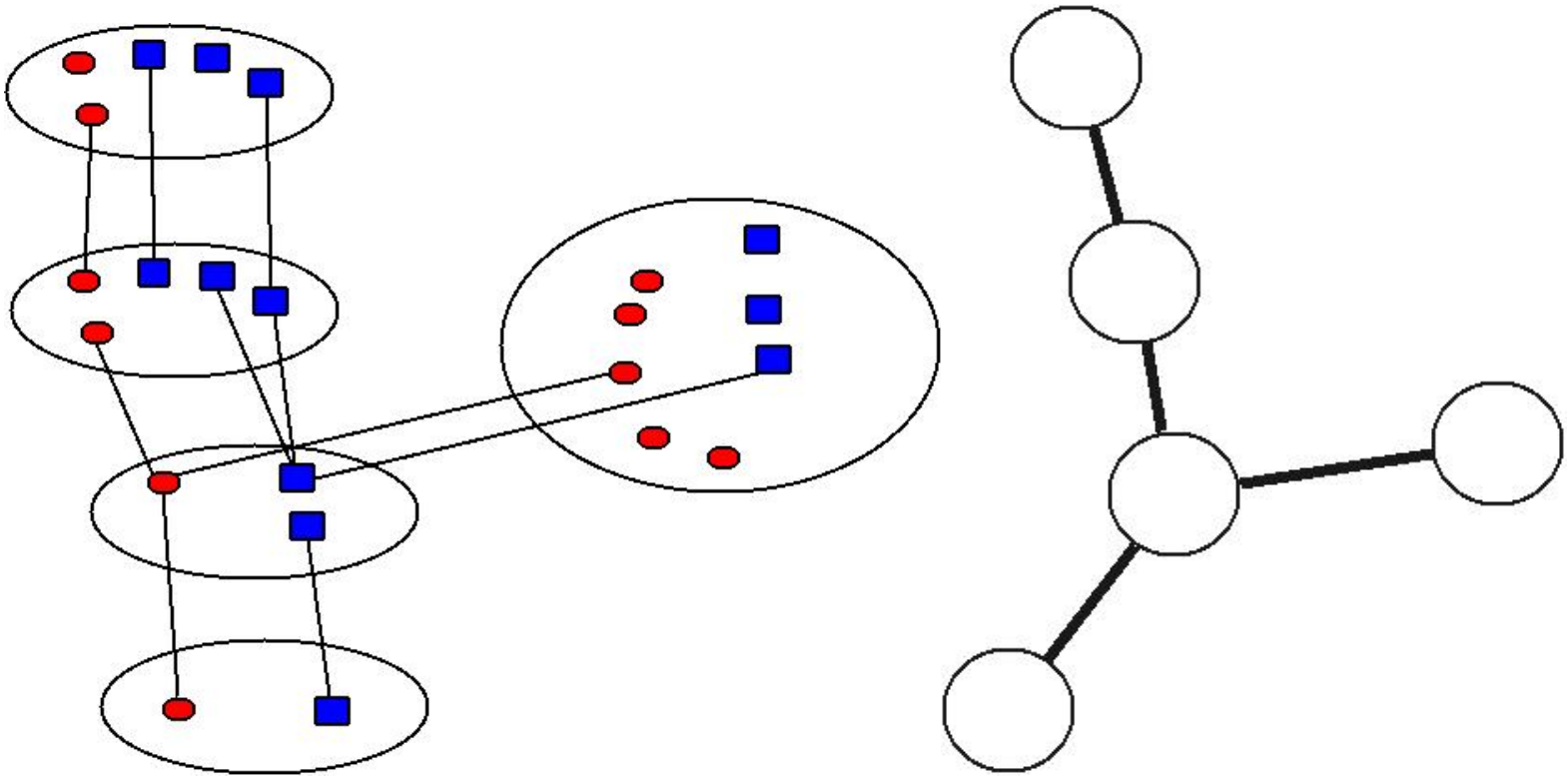
Constrained GNA



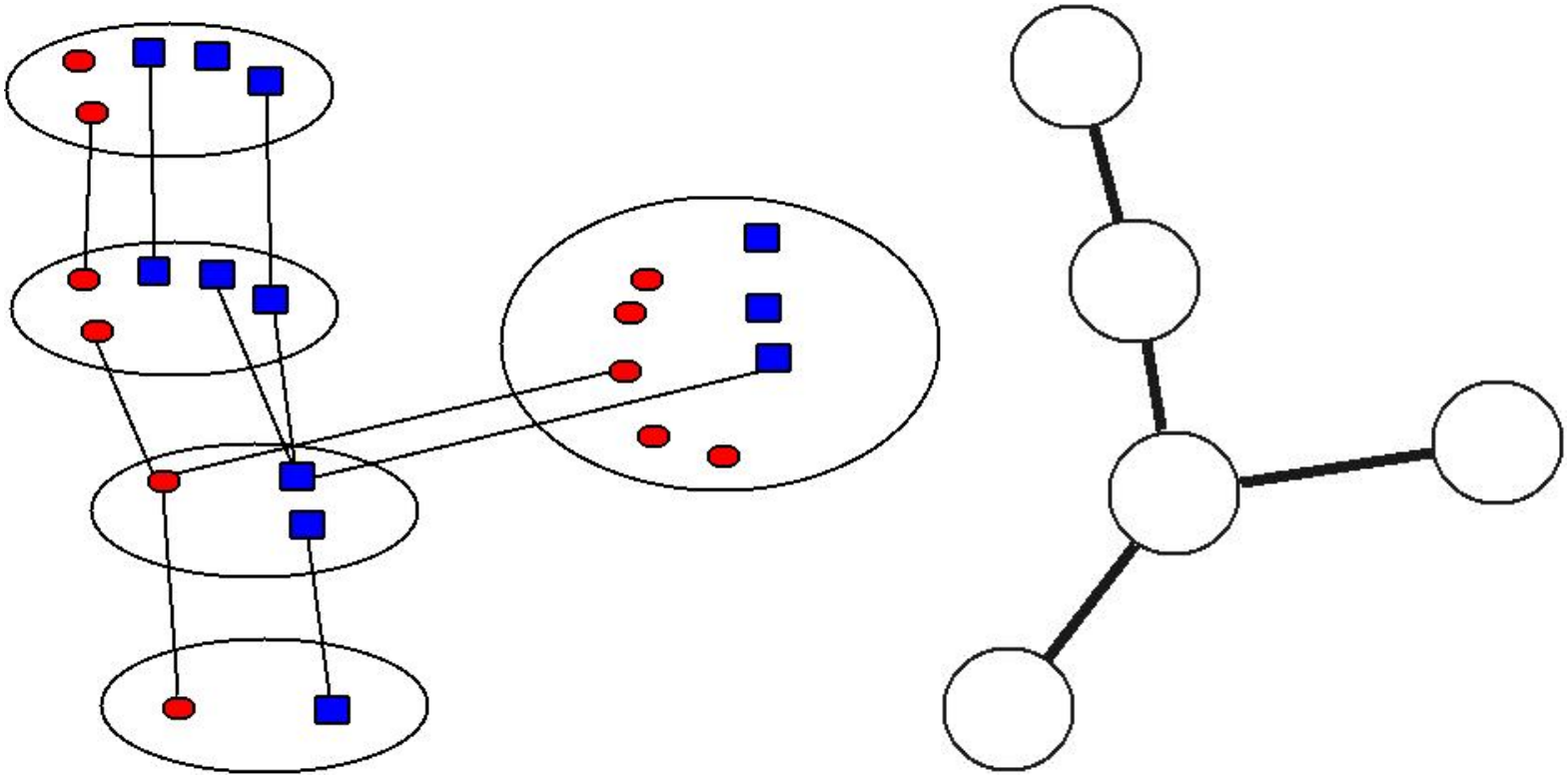
Problem

Find matchings within the clusters that maximise the number of conserved interactions

Graph of clusters induced by PPI



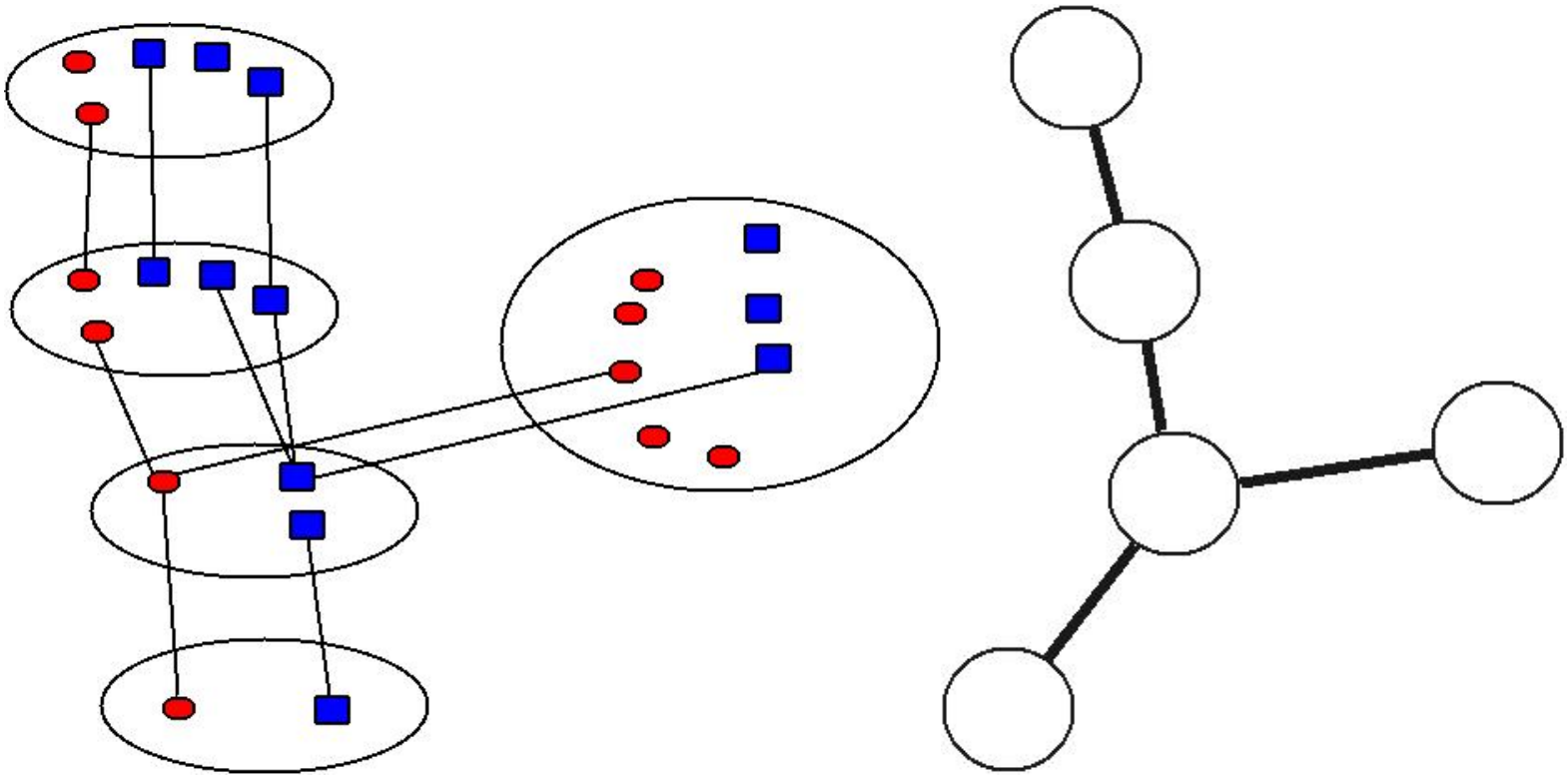
Global optimum



Proposition

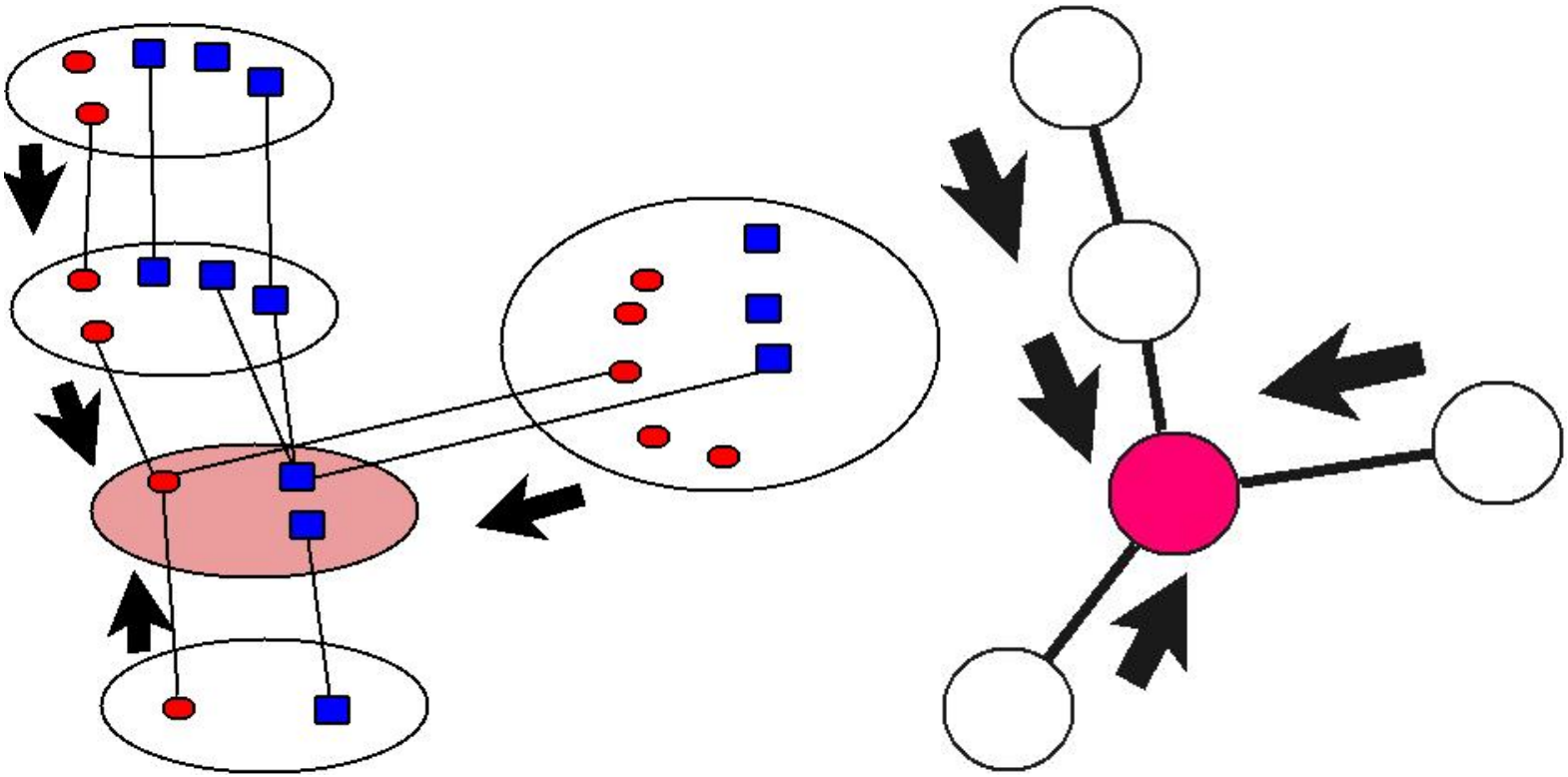
If the graph of clusters generated by the PPI has no cycle, then we can find the optimal matching efficiently with a message passing algorithm.

Global optimum by message passing



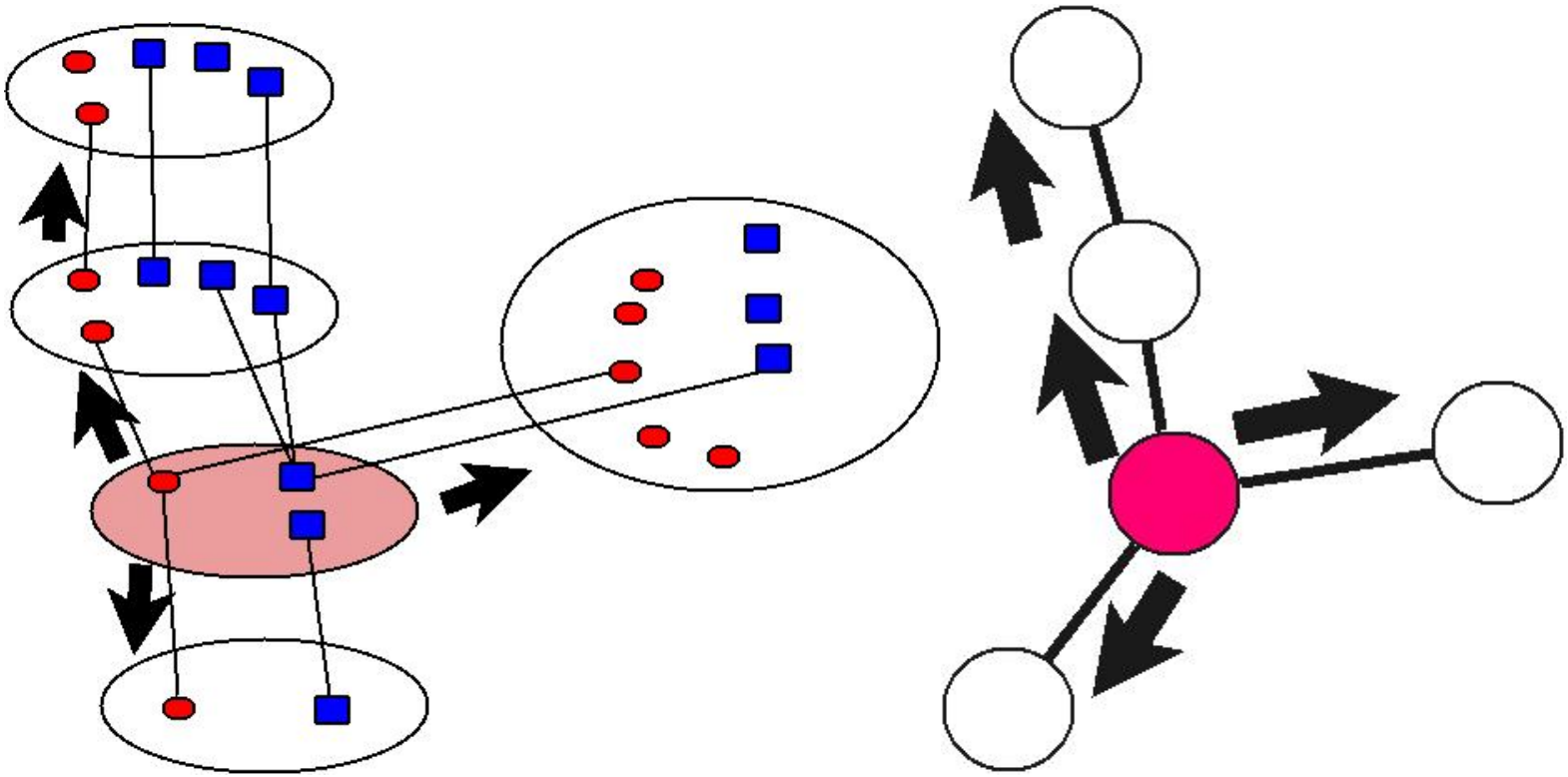
(Similar to Viterbi's algorithm for HMM)

Global optimum by message passing



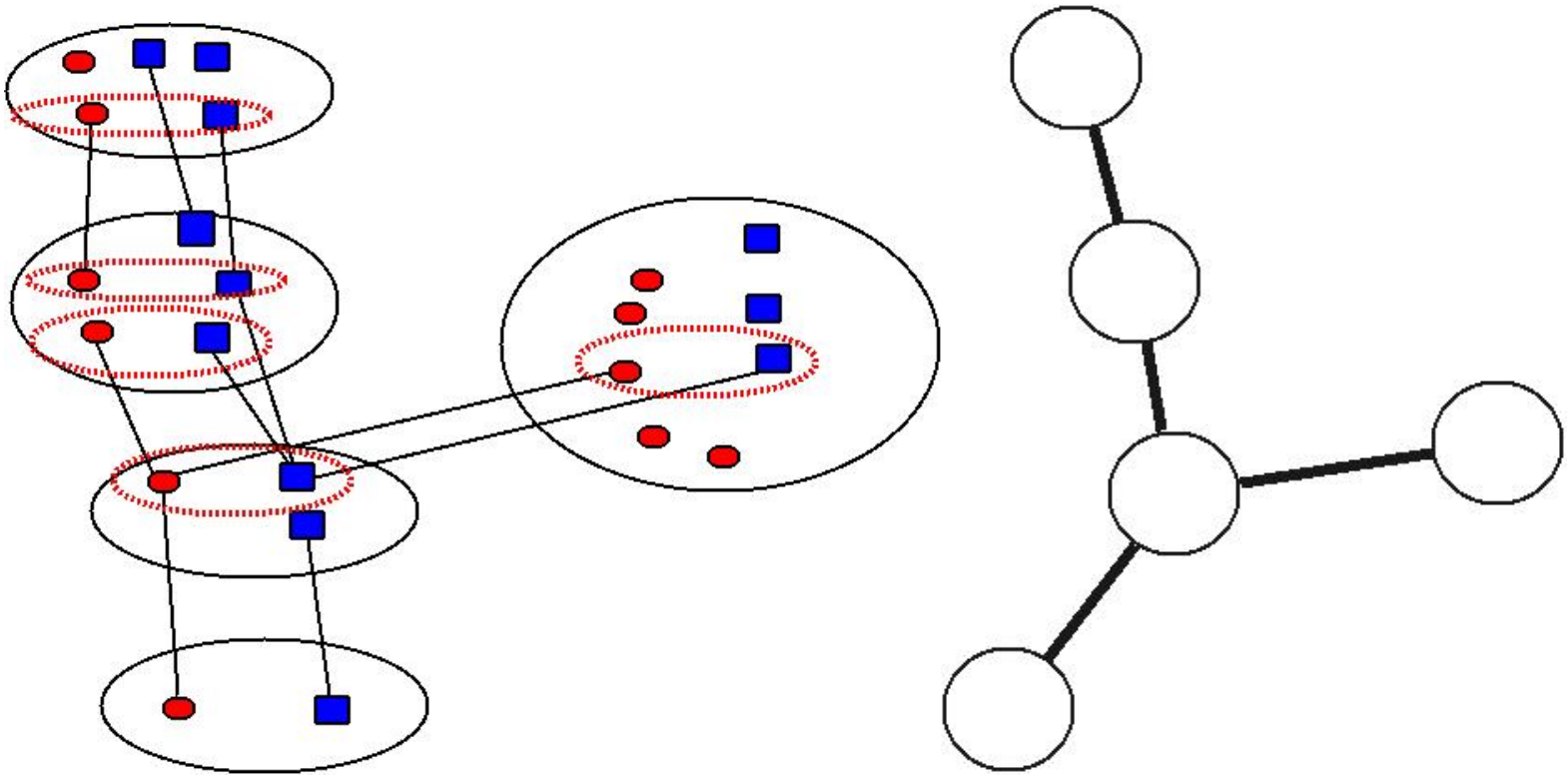
(Similar to Viterbi's algorithm for HMM)

Global optimum by message passing



(Similar to Viterbi's algorithm for HMM)

Global optimum by message passing



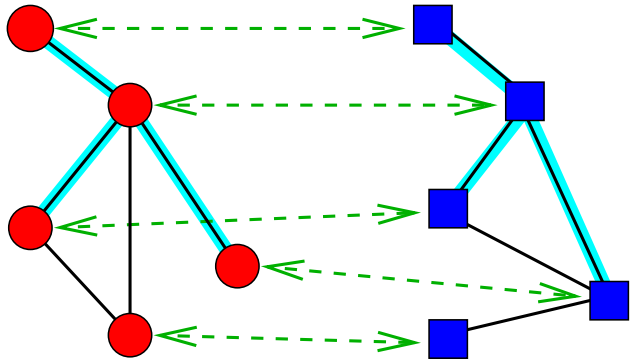
(Similar to Viterbi's algorithm for HMM)

What if the graph of clusters has cycle?

- The message passing method can not be used...
- Instead we reformulate the constrained GNA problem as a balanced GNA by setting similarity between proteins in different clusters to $-\infty$, and use algorithms for balanced GNA.

Outline

Balanced GNA

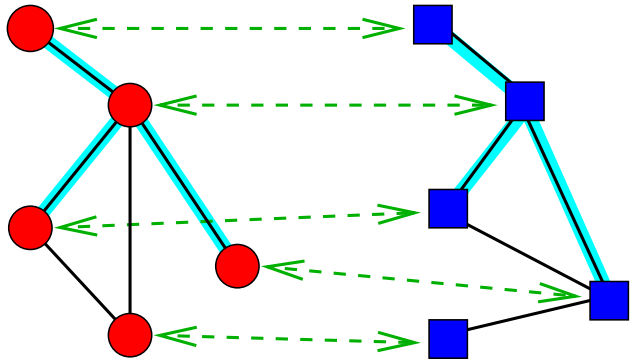


- Given two graphs and a matrix of all-vs-all similarities, find a matching $P \in \mathcal{P}$ that jointly maximizes:
  - the number of conserved interaction $CI(P)$,
  - the mean similarity of matched pairs $S(P)$.
- The trade-off can be found by maximizing over $\mathcal{P}$:

$$\min_{P \in \mathcal{P}} F(P) = (1-\alpha) CI(P) + \alpha S(P) \,,$$

where $\alpha \in [0,1]$ determines the balance between both objectives.

Balanced GNA



- Given two graphs and a matrix of all-vs-all similarities, find a matching $P \in \mathcal{P}$ that jointly maximizes:
  - the number of conserved interaction $CI(P)$,
  - the mean similarity of matched pairs $S(P)$.
- The trade-off can be found by maximizing over $\mathcal{P}$:

$$\min_{P \in \mathcal{P}} F(P) = (1-\alpha) CI(P) + \alpha S(P)\,,$$

where $\alpha \in [0,1]$ determines the balance between both objectives.

Existing methods for balanced GNA

$$\min_{P \in \mathcal{P}} F(P) = (1-\alpha) CI(P) + \alpha S(P)\,,$$

- When $\alpha = 1$ this is an optimal assignment problem efficiently solved by the Hungarian algorithm (Kuhn, 1955).
- When $\alpha < 1$ this is a general graph matching problem, usually computationally intractable. Existing algorithms include:
  - Exact solution by incomplete enumeration (only for small graphs)
  - Spectral methods (Umeyama, 1986; Singh et al., 2008)
  - Relaxations of the problem into a continuous optimization problem (Almohamad and Duffuaa, 1993; Gold and Rangarajan, 1996).

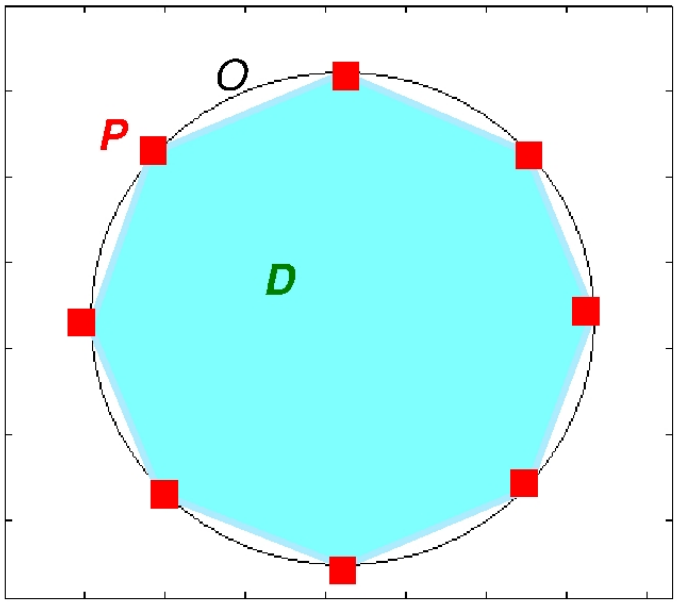
Existing methods for balanced GNA

$$\min_{P \in \mathcal{P}} F(P) = (1-\alpha) CI(P) + \alpha S(P) \,,$$

- When $\alpha = 1$ this is an optimal assignment problem efficiently solved by the Hungarian algorithm (Kuhn, 1955).
- When $\alpha < 1$ this is a general graph matching problem, usually computationally intractable. Existing algorithms include:
  - Exact solution by incomplete enumeration (only for small graphs)
  - Spectral methods (Umeyama, 1986; Singh et al., 2008)
  - Relaxations of the problem into a continuous optimization problem (Almohamad and Duffuaa, 1993; Gold and Rangarajan, 1996).

Relaxation algorithms

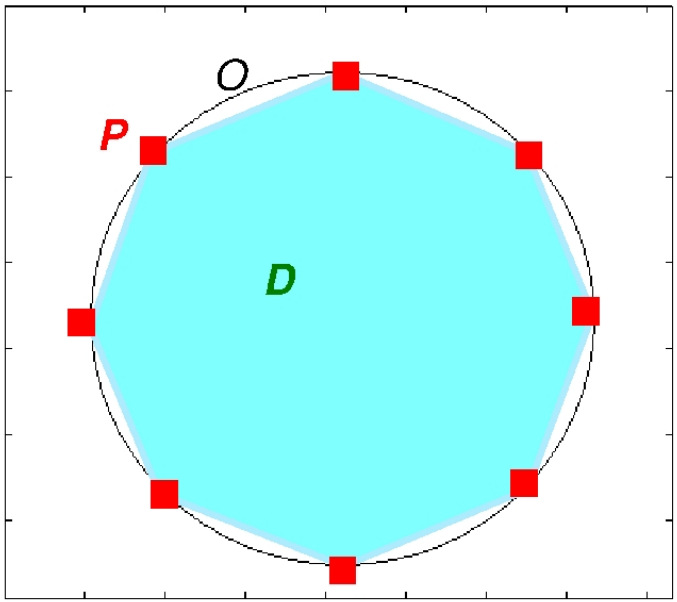
$$\min_{P \in \mathcal{P}} F(P)$$



- Embed the discrete set $\mathcal{P}$ into a continuous space $\mathcal{D}$
- Extend the function $F(P)$ to $\mathcal{D}$
- Minimize $F(P)$ over $\mathcal{D}$
- Map back the solution to $\mathcal{P}$

Relaxation algorithms

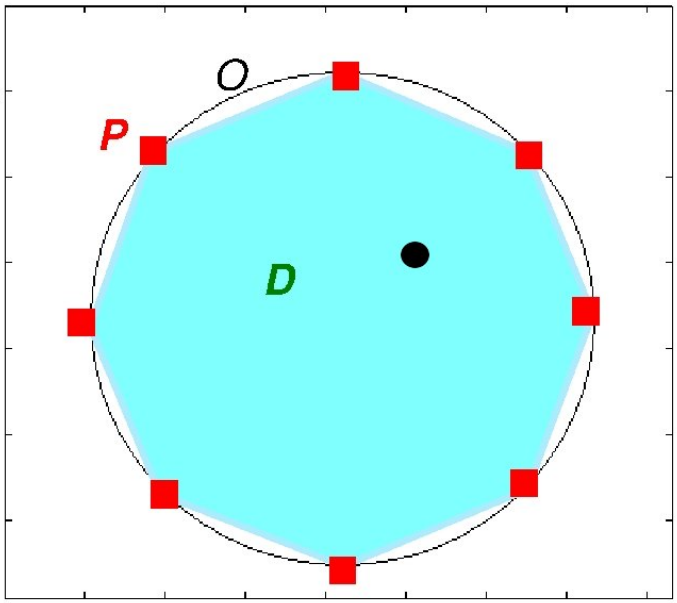
$$\min_{P \in \mathcal{P}} F(P)$$



- Embed the discrete set $\mathcal{P}$ into a continuous space $\mathcal{D}$
- Extend the function $F(P)$ to $\mathcal{D}$
- Minimize $F(P)$ over $\mathcal{D}$
- Map back the solution to $\mathcal{P}$

Relaxation algorithms

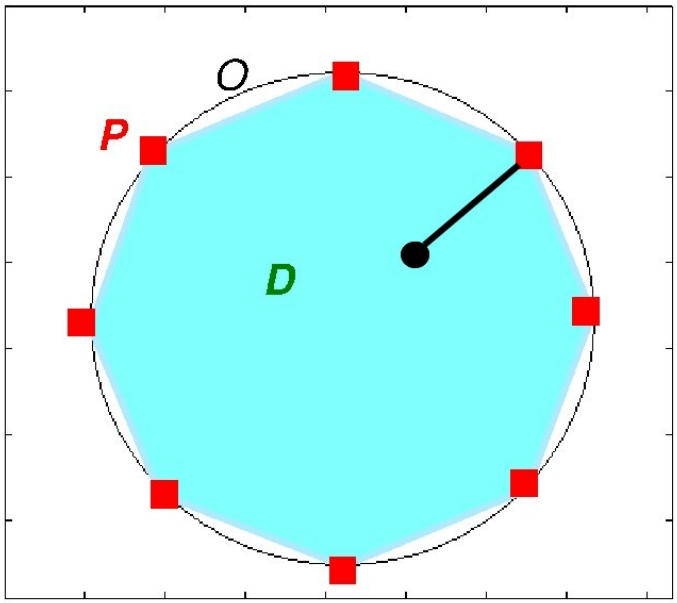
$$\min_{P \in \mathcal{P}} F(P)$$



- Embed the discrete set $\mathcal{P}$ into a continuous space $\mathcal{D}$
- Extend the function $F(P)$ to $\mathcal{D}$
- Minimize $F(P)$ over $\mathcal{D}$
- Map back the solution to $\mathcal{P}$

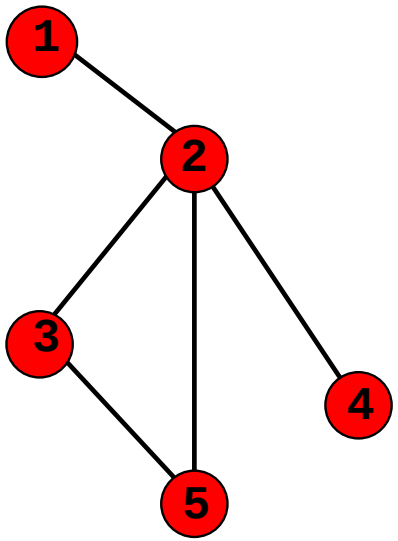
Relaxation algorithms

$$\min_{P \in \mathcal{P}} F(P)$$



- Embed the discrete set $\mathcal{P}$ into a continuous space $\mathcal{D}$
- Extend the function $F(P)$ to $\mathcal{D}$
- Minimize $F(P)$ over $\mathcal{D}$
- Map back the solution to $\mathcal{P}$

Mathematical formulation

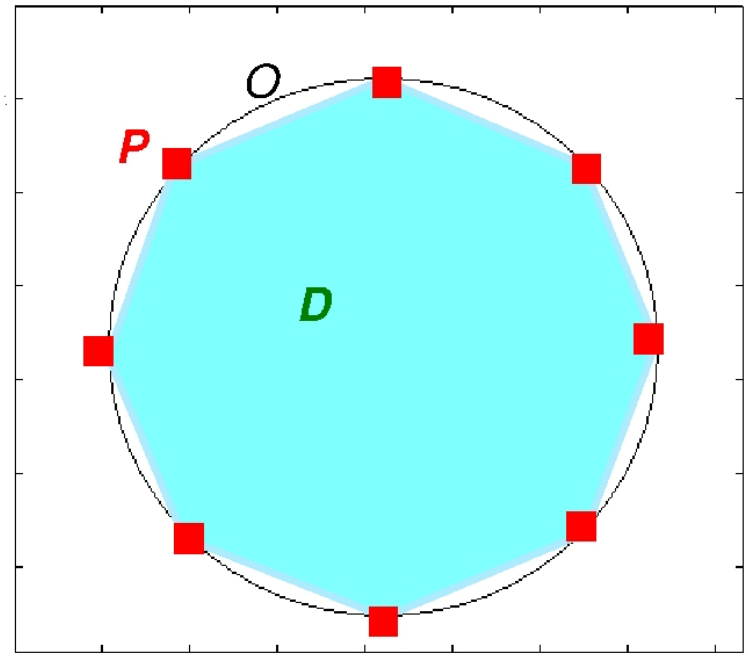


$$A_G = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} \quad P = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

- $\mathcal{P}$ = permutation matrices ($P_{ij} = 1$ if i is matched to j)
- $\mathcal{D}$ = doubly stochastic matrices ($P \geq 0$, $P1_N = 1_N$, $1_N^\top P = 1_N$)
- Classical relaxation:

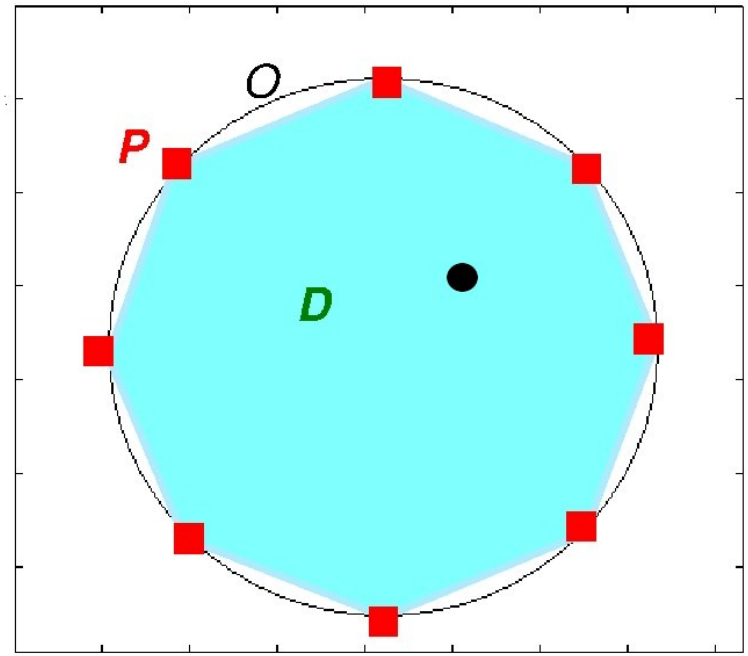
$$CI(P) = ||A_G - A_{P(H)}|| = ||A_G - PA_H P^T||$$

Quadratic convex relaxation (QCV)



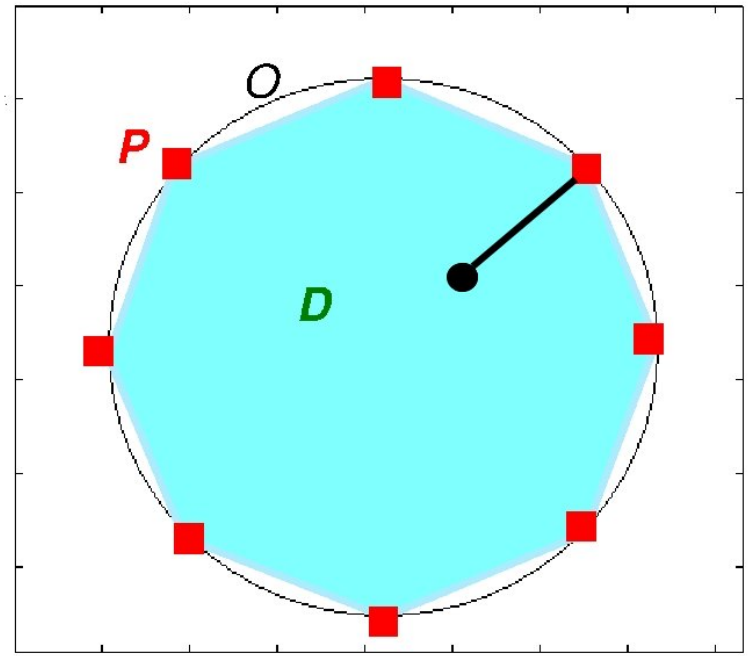
- Minimize $F_0(P) = ||A_G P - P A_H||_F^2 = \text{vec}(P)^T Q \text{vec}(P)$ over $\mathcal{D}$ (convex QP)
- Project the solution D^* to $\mathcal{P}$ (Hungarian algorithm)
- Not very good if D^* is far from $\mathcal{P}$...

Quadratic convex relaxation (QCV)



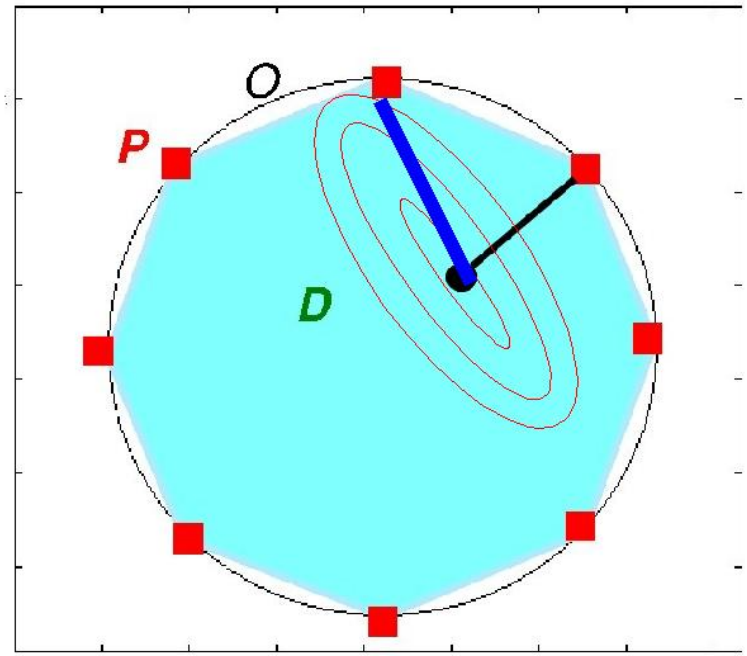
- Minimize $F_0(P) = ||A_G P - P A_H||_F^2 = \text{vec}(P)^T Q \text{vec}(P)$ over $\mathcal{D}$ (convex QP)
- Project the solution D^* to $\mathcal{P}$ (Hungarian algorithm)
- Not very good if D^* is far from $\mathcal{P}$...

Quadratic convex relaxation (QCV)



- Minimize $F_0(P) = ||A_G P - P A_H||_F^2 = \text{vec}(P)^T Q \text{vec}(P)$ over $\mathcal{D}$ (convex QP)
- Project the solution D^* to $\mathcal{P}$ (Hungarian algorithm)
- Not very good if D^* is far from $\mathcal{P}$...

Quadratic convex relaxation (QCV)



- Minimize $F_0(P) = ||A_G P - P A_H||_F^2 = \text{vec}(P)^T Q \text{vec}(P)$ over $\mathcal{D}$ (convex QP)
- Project the solution D^* to $\mathcal{P}$ (Hungarian algorithm)
- Not very good if D^* is far from $\mathcal{P}$...

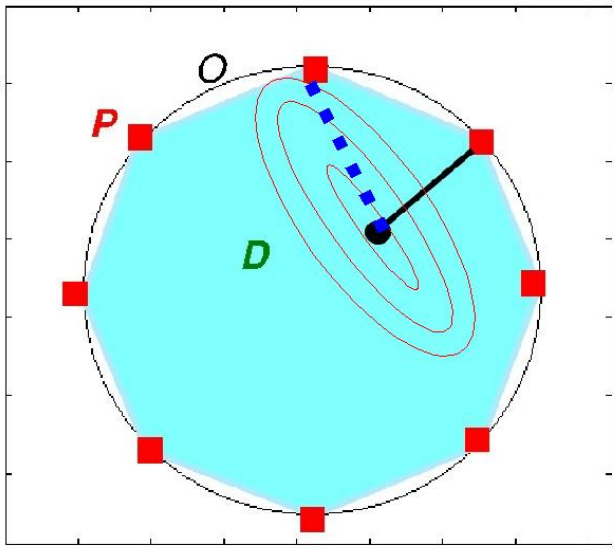
A new concave relaxation

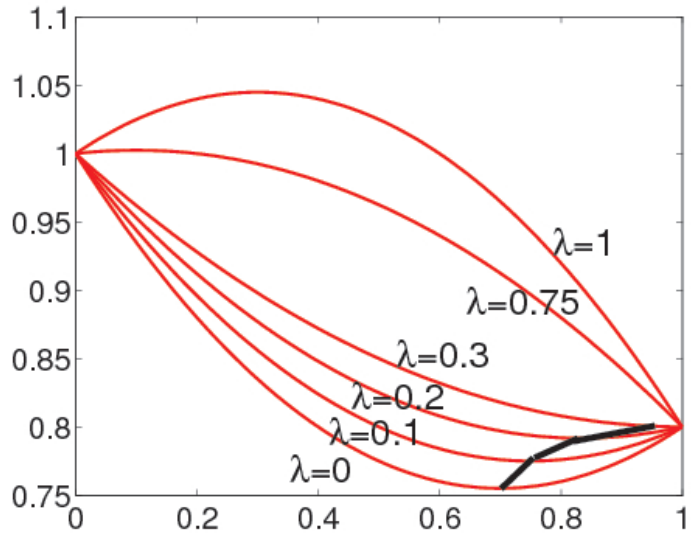
- On $\mathcal{P}$ we also have:

$$CI(P) = F_1(P) = -tr(\Delta P) - \text{vec}(P)^T (L_G \otimes L_H)\text{vec}(P)$$

- This is a concave function, therefore its global minimum over $\mathcal{D}$ is on $\mathcal{P}$ (extreme points)
- Idea: starting from a "good solution" on $\mathcal{D}$, we can project to $\mathcal{P}$ by gradient ascent (GA) to maximize $-F_1(P)$

The PATH algorithm





$$F_0(P) = ||A_G P - P A_H||_F^2 = \text{vec}(P)^T Q \text{vec}(P)$$
$$F_1(P) = -tr(\Delta P) - \text{vec}(P)^T (L_G \otimes L_H) \text{vec}(P)$$
$$F_\lambda(P) = (1-\lambda) F_0(P) + \lambda F_1(P)$$

(Zaslavskyi et al., IEEE PAMI, 2009.)

Outline

Random graphs:N=8

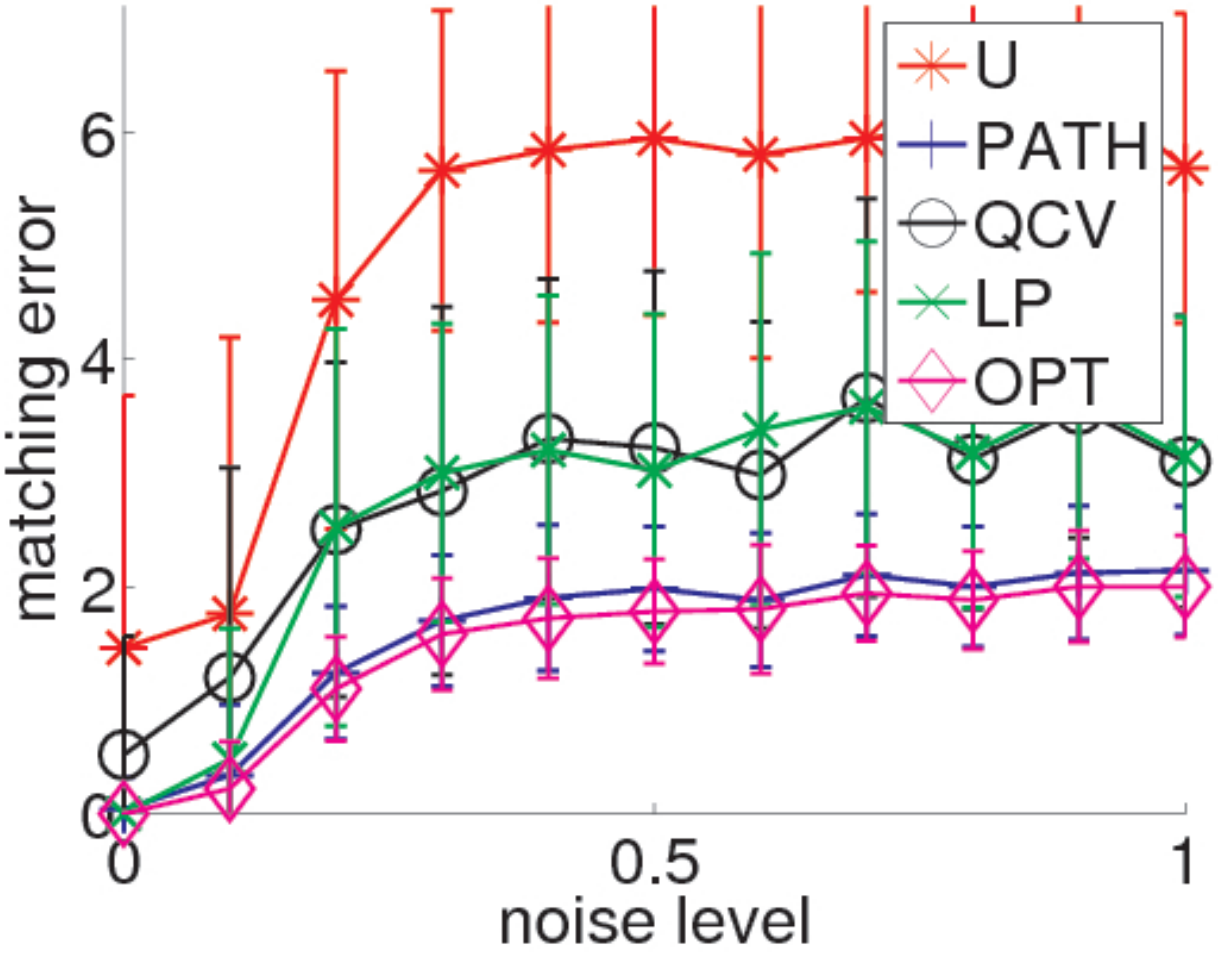


Figure: Precision as a noise function, U — Umeyama algorithm results, LP — linear programming algorithm, QCV — convex function approach (F_0), PATH — path minimization algorithm, OPT — an exhaustive search (the global minimum).

Random graphs:N=20

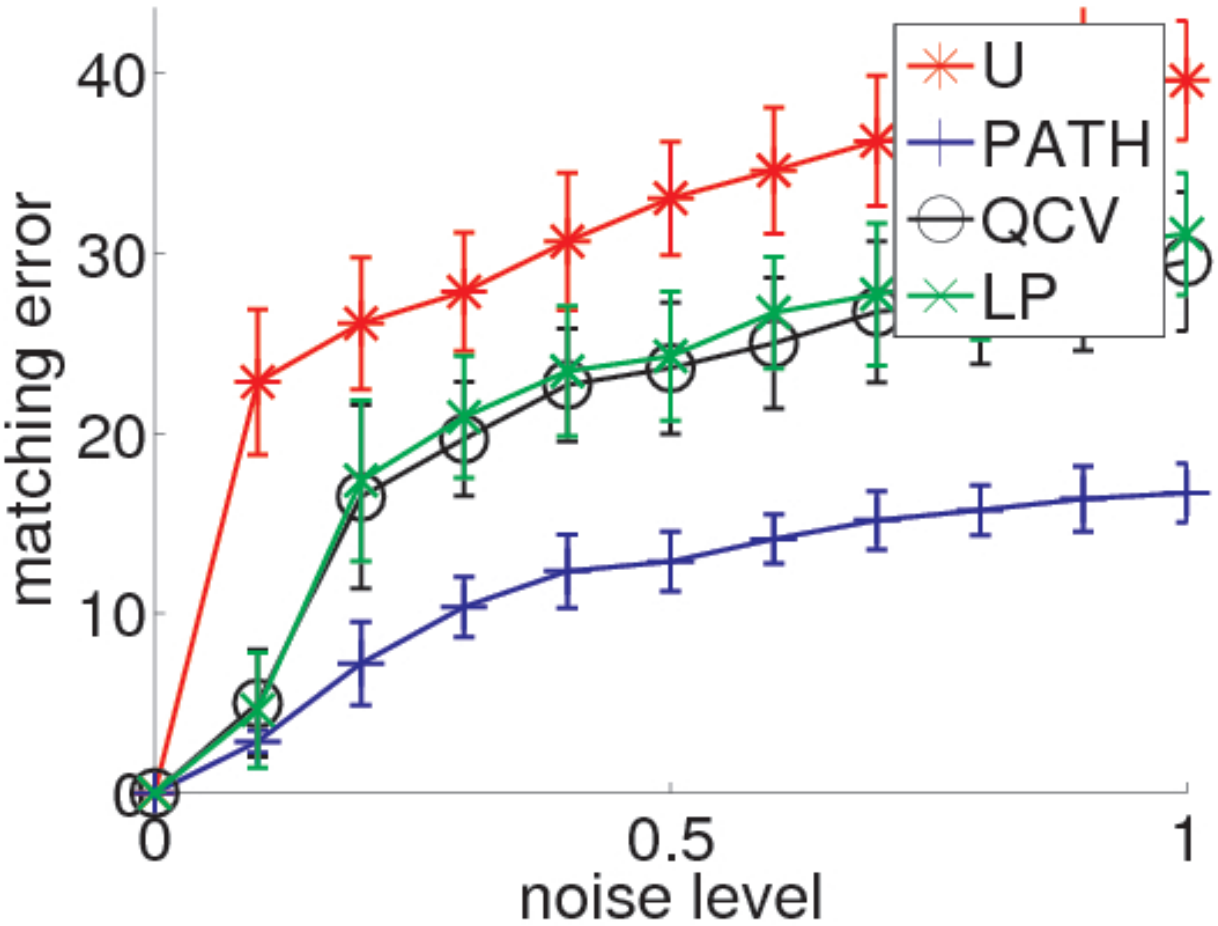


Figure: Precision as a noise function, U — Umeyama algorithm results, LP — linear programming algorithm, QCV — convex function approach (F_0), PATH — path minimization algorithm.

Random graphs:N=100

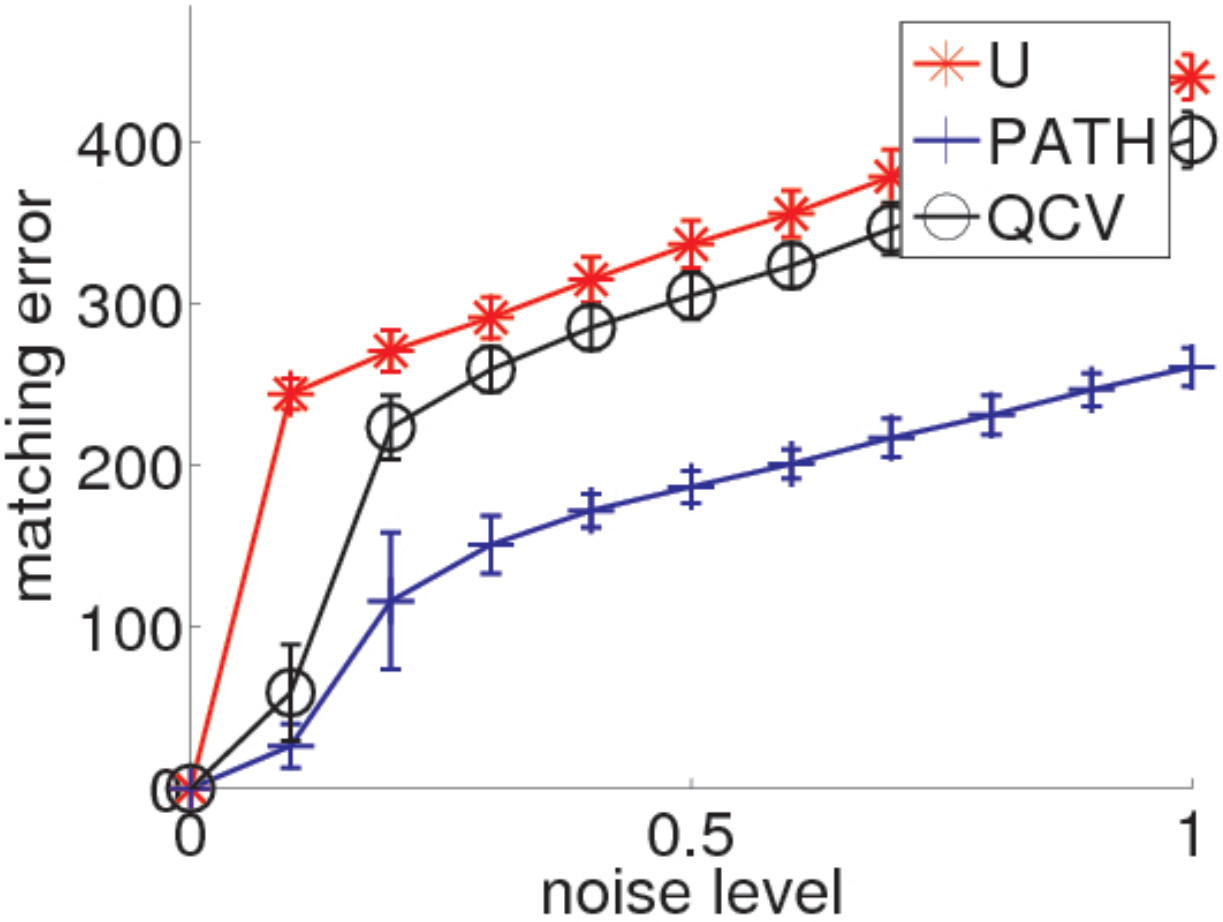


Figure: Precision as a noise function, U — Umeyama algorithm results, QCV — convex function approach (F_0), PATH — path minimization algorithm.

Algorithm complexity

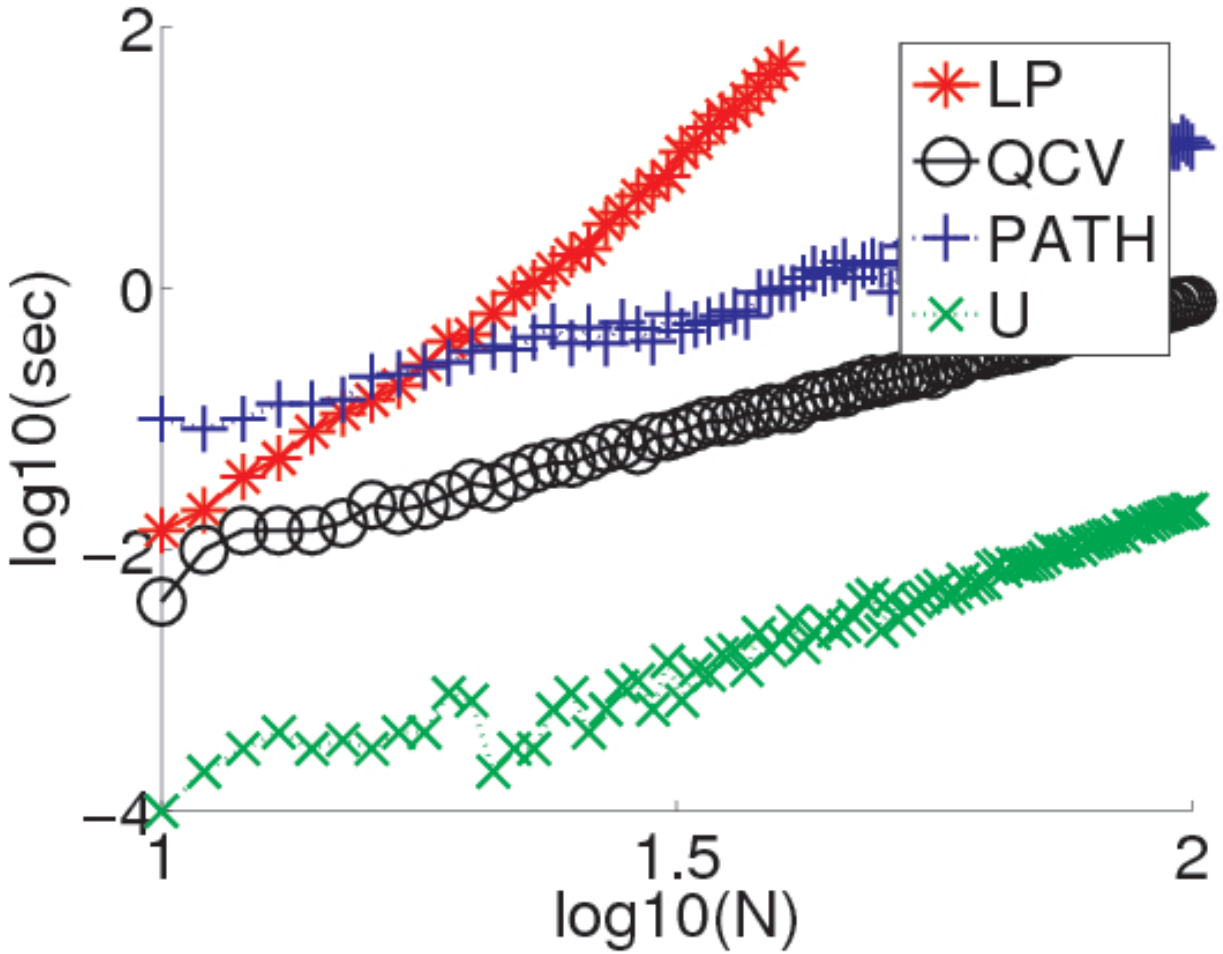
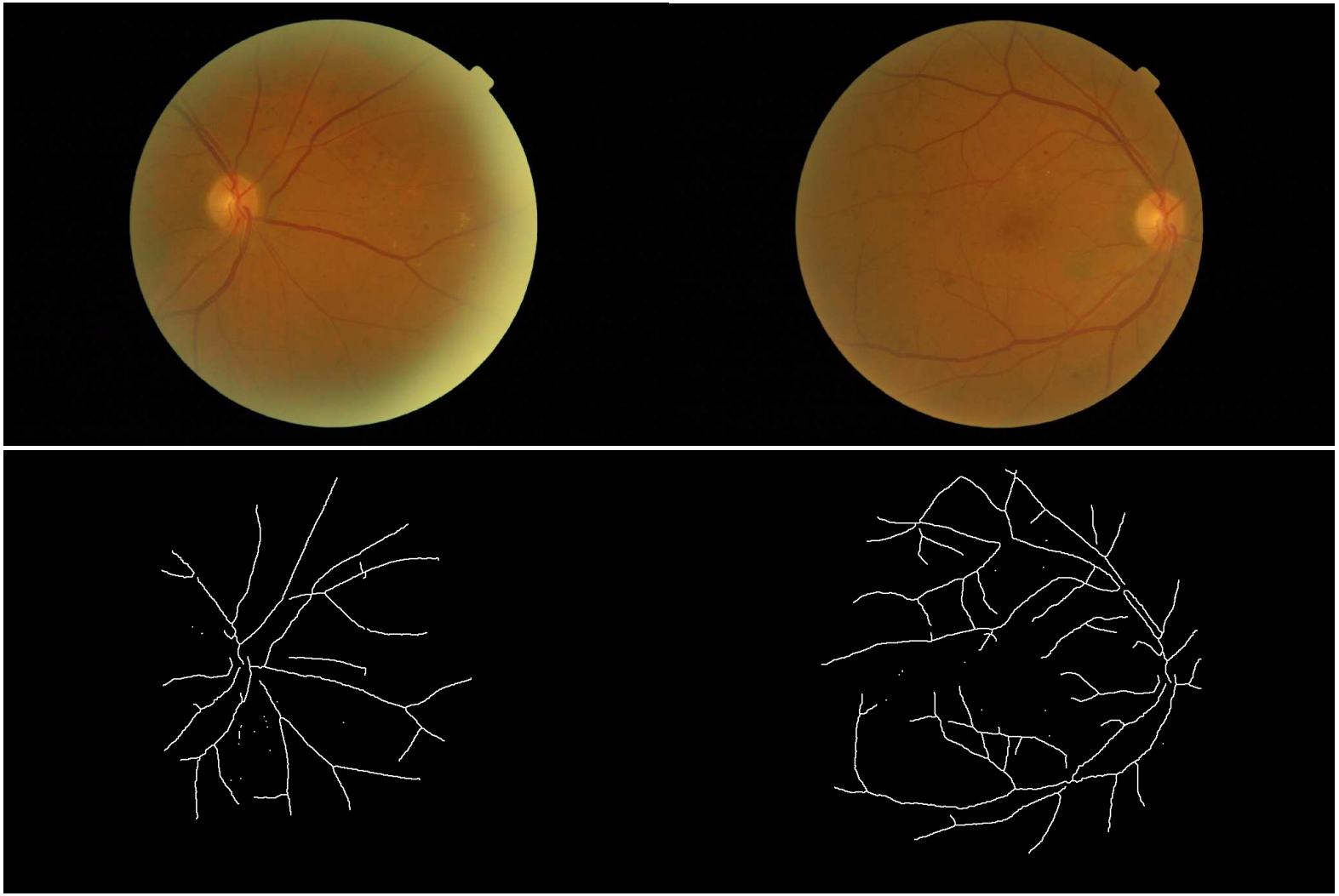


Figure: Timing of U, LP, QCV and PATH algorithms as a function of graph size. Noise level is 0.3. Slope: $tan_{LP} = 6.67, tan_U = tan_{QCV} = tan_{PATH} = 3.3$

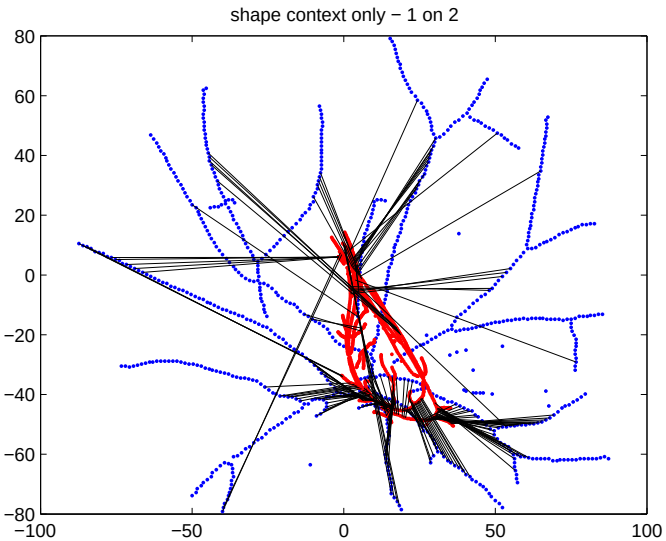
Experiment results for QAPLIB benchmark

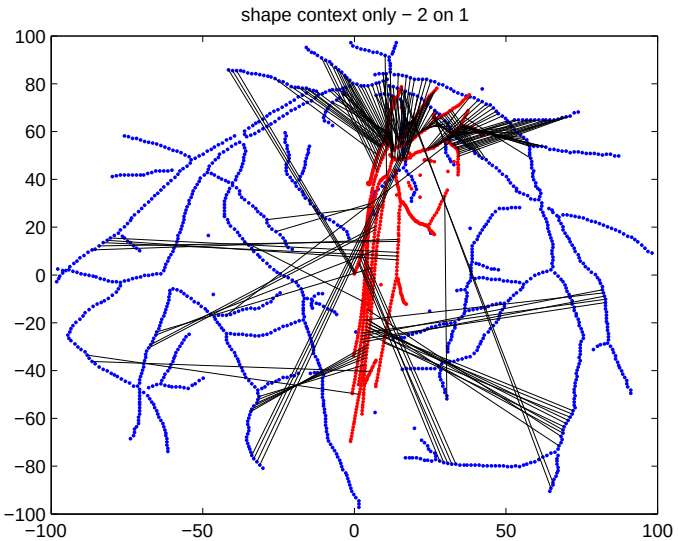
QAP	MIN	PATH	QPB	GRAD	U
chr12c	11156	**18048**	20306	19014	40370
chr15a	9896	**19086**	26132	30370	60986
chr15c	9504	**16206**	29862	23686	76318
chr20b	2298	**5560**	6674	6290	10022
chr22b	6194	**8500**	9942	9658	13118
esc16b	292	300	**296**	298	306
rou12	235528	**256320**	278834	273438	295752
rou15	354210	391270	**381016**	457908	480352
rou20	725522	**778284**	804676	840120	905246
tai10a	135028	**152534**	165364	168096	189852
tai15a	388214	**419224**	455778	451164	483596
tai17a	491812	**530978**	550852	589814	620964
tai20a	703482	**753712**	799790	871480	915144
tai30a	1818146	**1903872**	1996442	2077958	2213846
tai35a	2422002	**2555110**	2720986	2803456	2925390

Eye vessels image processing

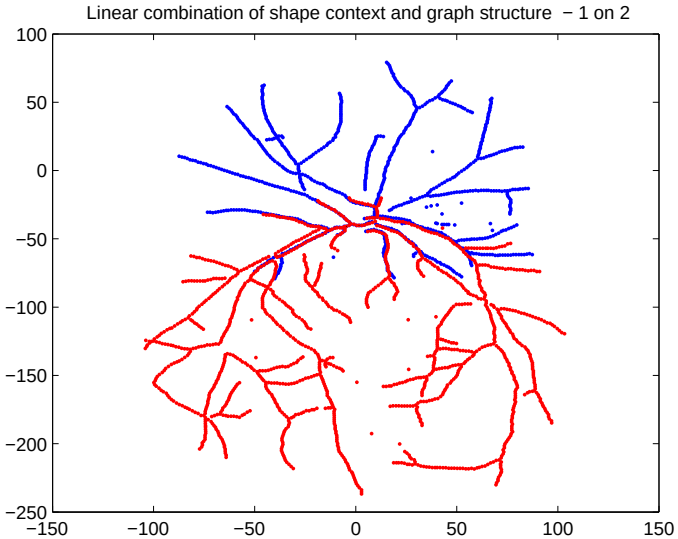


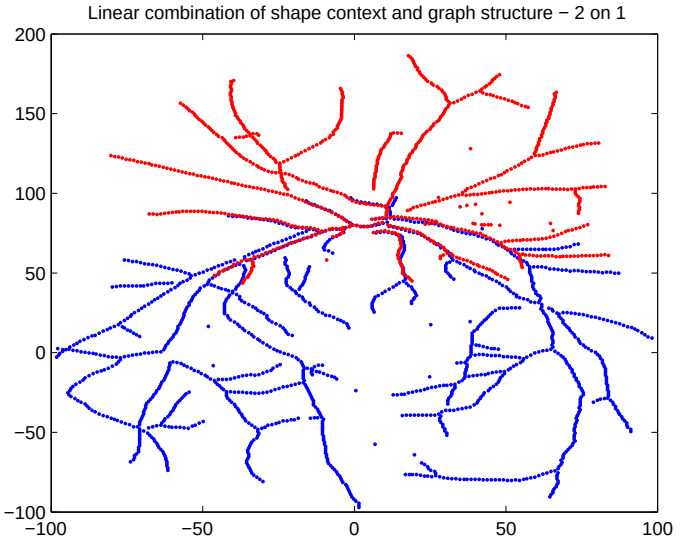
Eye vessels image processing: Shape context





Combination of shape context and structural information





Recognition of chinese characters

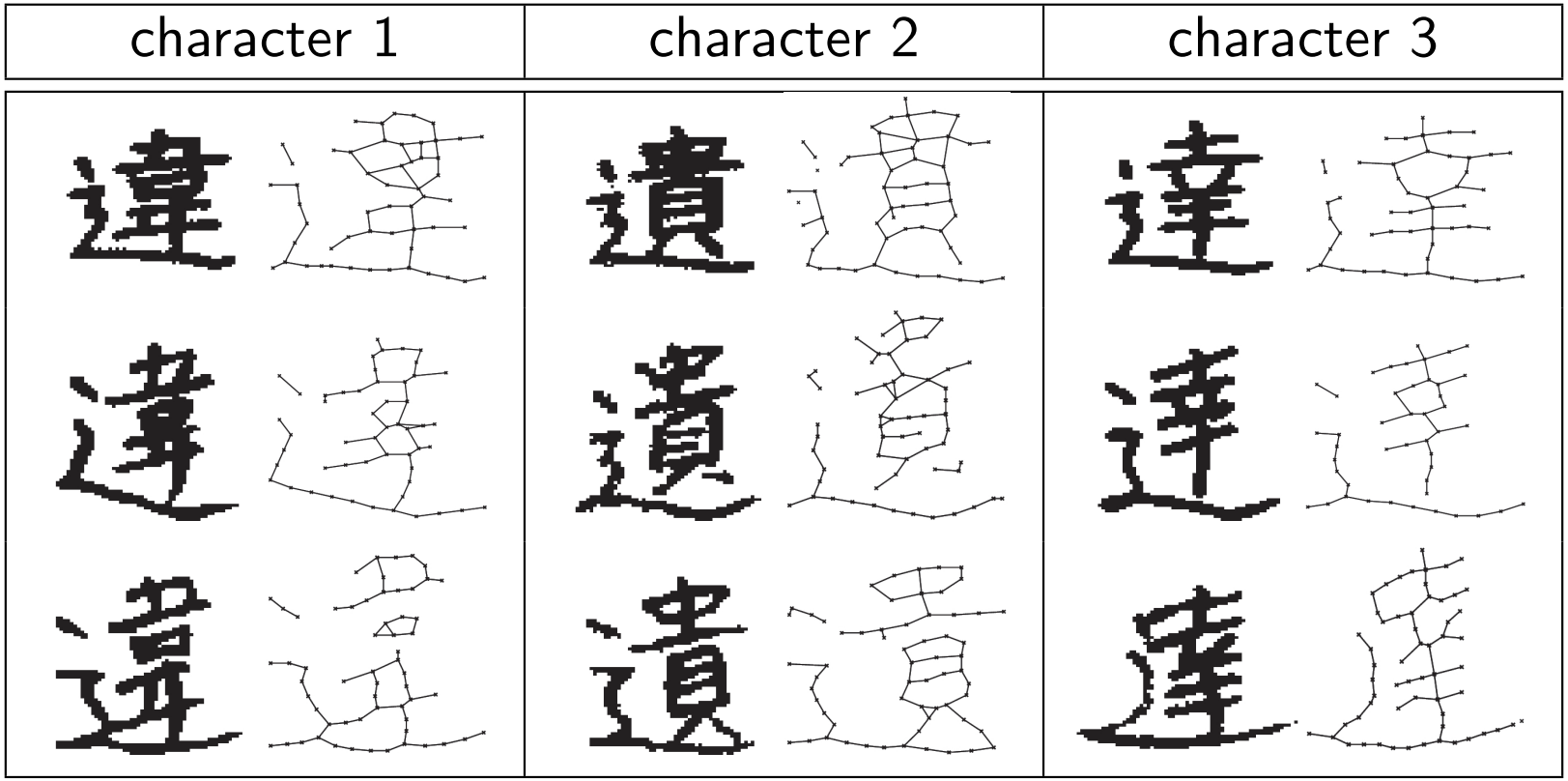


Figure: Chinese characters from the ETL9B dataset.

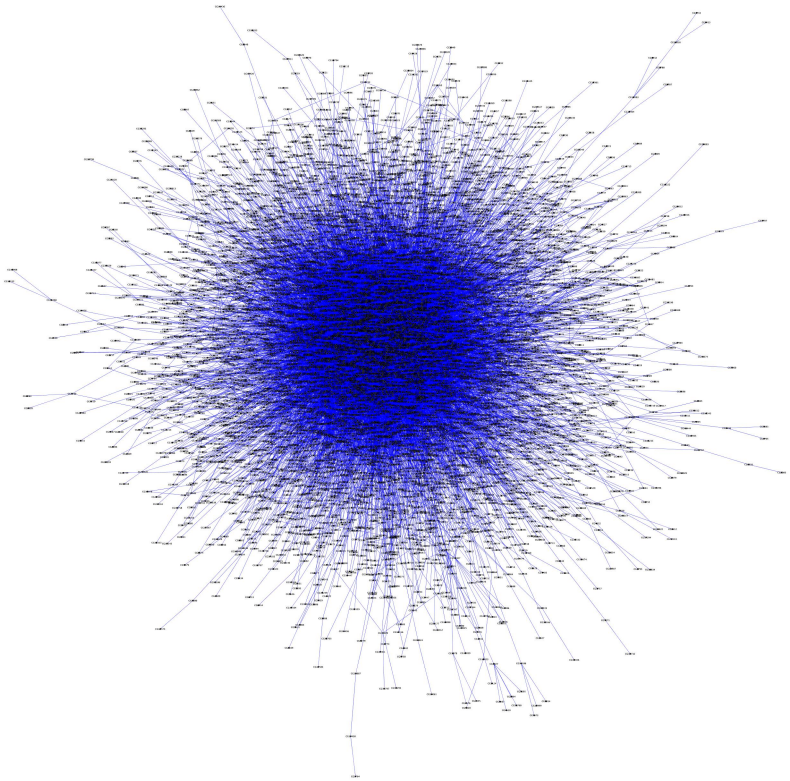
Recognition of chinese characters

Table: Classification of chinese characters. (CV, STD)—mean and standard deviation of test error over cross-validation runs (five folds, 50 repetitions)

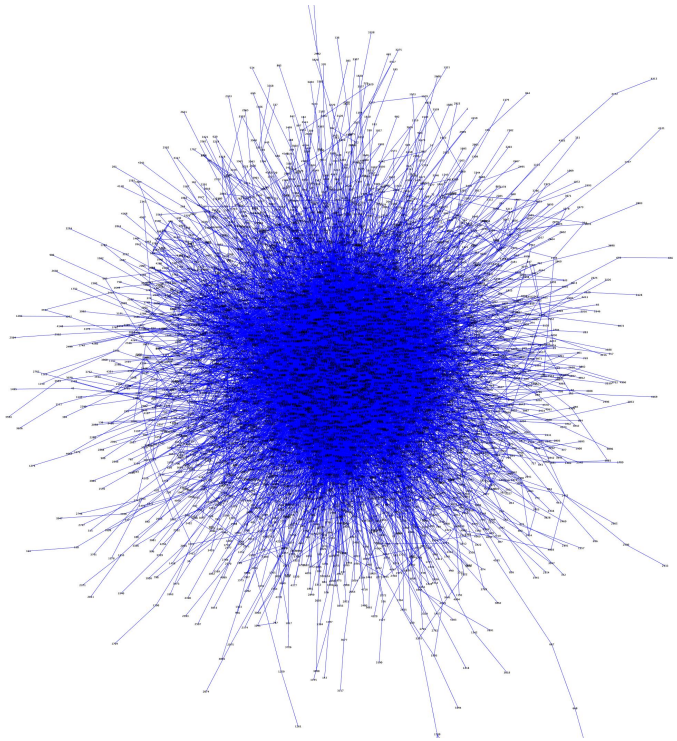
Method	CV	STD
Linear SVM	0.377	$\pm$ 0.090
SVM with gaussian kernel	0.359	$\pm$ 0.076
KNN (PATH) (α=1): shape context	0.399	$\pm$ 0.081
KNN (PATH) (α=0.4)	**0.248**	$\pm$ **0.075**
KNN (PATH) (α=0): pure graph matching	0.607	$\pm$ 0.072
KNN (U) (α=0.9): α best choice	0.382	$\pm$ 0.077
KNN (QCV) (α=0.3): α best choice	0.295	$\pm$ 0.061

Alignment of PPI networks: Fly vs. Yeast

- PPI networks and all-vs-all BLAST / Inparanoid clusters for D. melanogaster (fly) vs. S. cerevisiae (yeast)
- Data provided by Bandyopadhyay et al. (MSB 2006)



Fly (7k nodes, 20k edges)

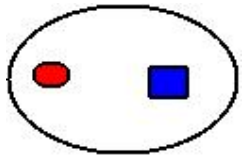


Yeast (4k nodes,15k edges)

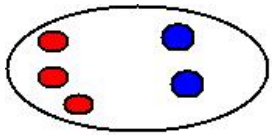
Experiments: Constrained Alignment

There are Inparanoid 2244 clusters:

1552 clusters with only two proteins



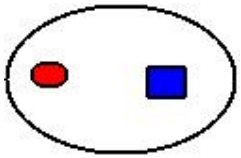
692 ambiguous clusters



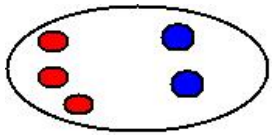
Experiments: Constrained Alignment

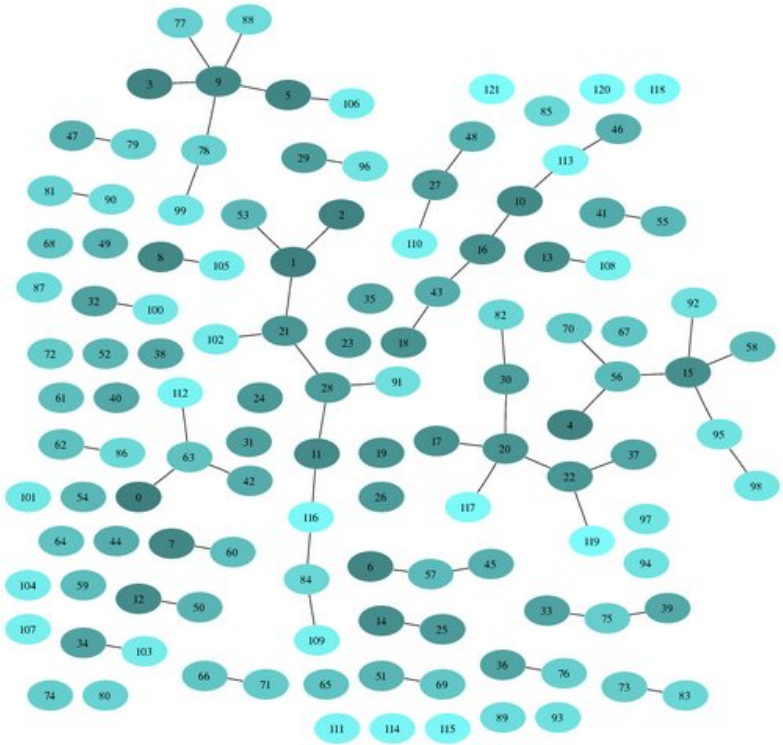
There are 2244 clusters:

1552 clusters with only two proteins



692 ambiguous clusters





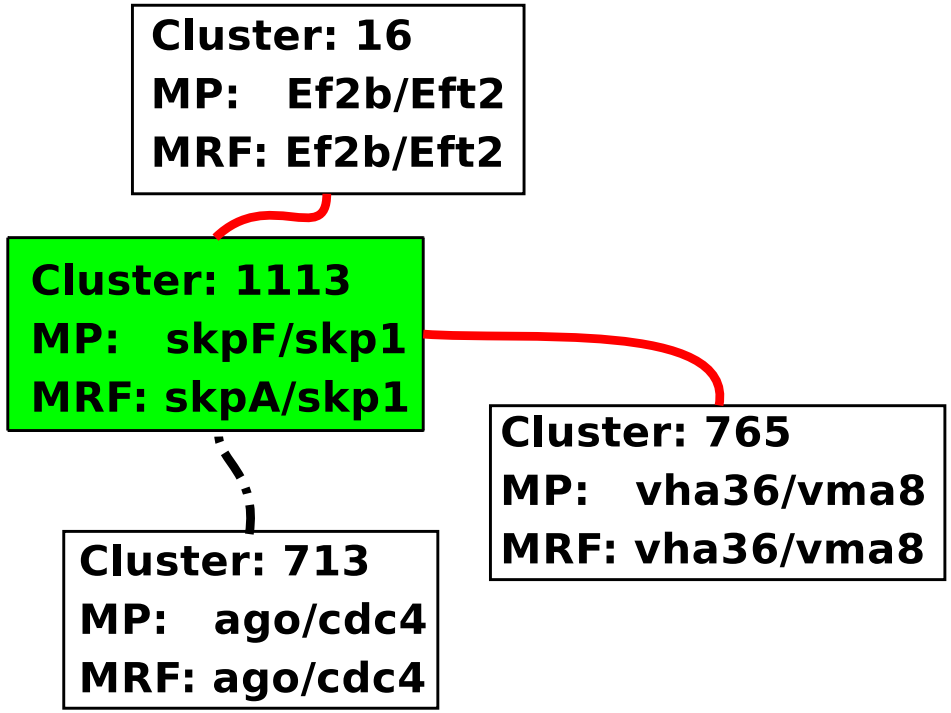
There is no cycles in the graph of clusters!

Experiments: Constrained Alignment

- InParanoid clusters: 2244 clusters (1552 clusters with only two proteins + 692 ambiguous clusters)
- Message Passing Algorithm (MP) provides the optimal solution
- MRF (Bandyopadhyay et al., 2006), IsoRank (Singh et al., 2008), PATH and GA methods may be used as well
- Measure the number of conserved interactions
- Validation: count the number of Homologene pairs (gold standard for functional orthologs?)

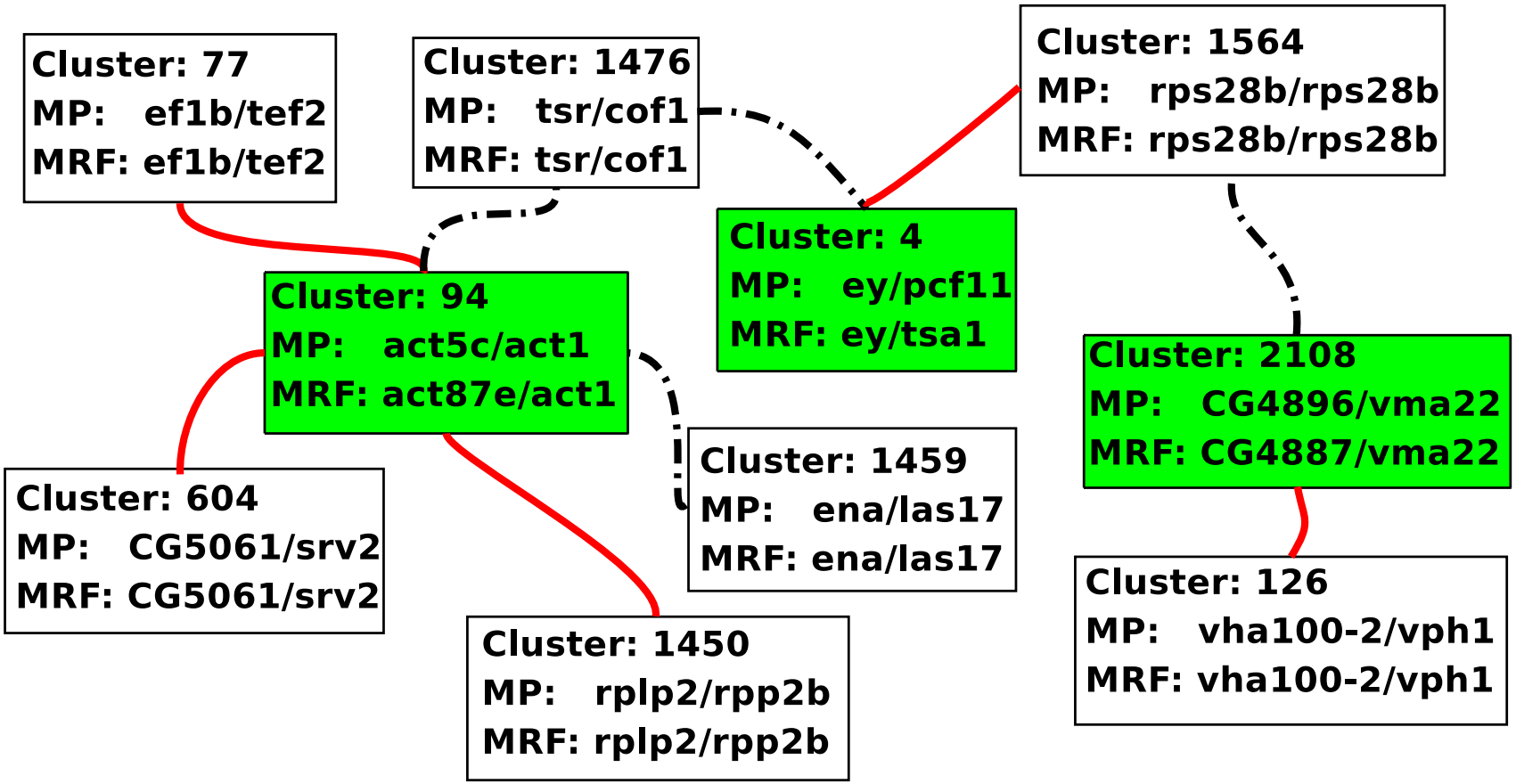
Algorithm	MP	GA	PATH	MRF	IsoRank
#cons. interactions	**238**	**238**	**238**	233	228
#HomoloG pairs	41	41	41	36	39
Timing(sec)	1-2	1-2	80	10	1-2

Differences MP vs MRF: example



Solid red: interaction conserved by MP; Dotted black: interactions conserved by MRF.

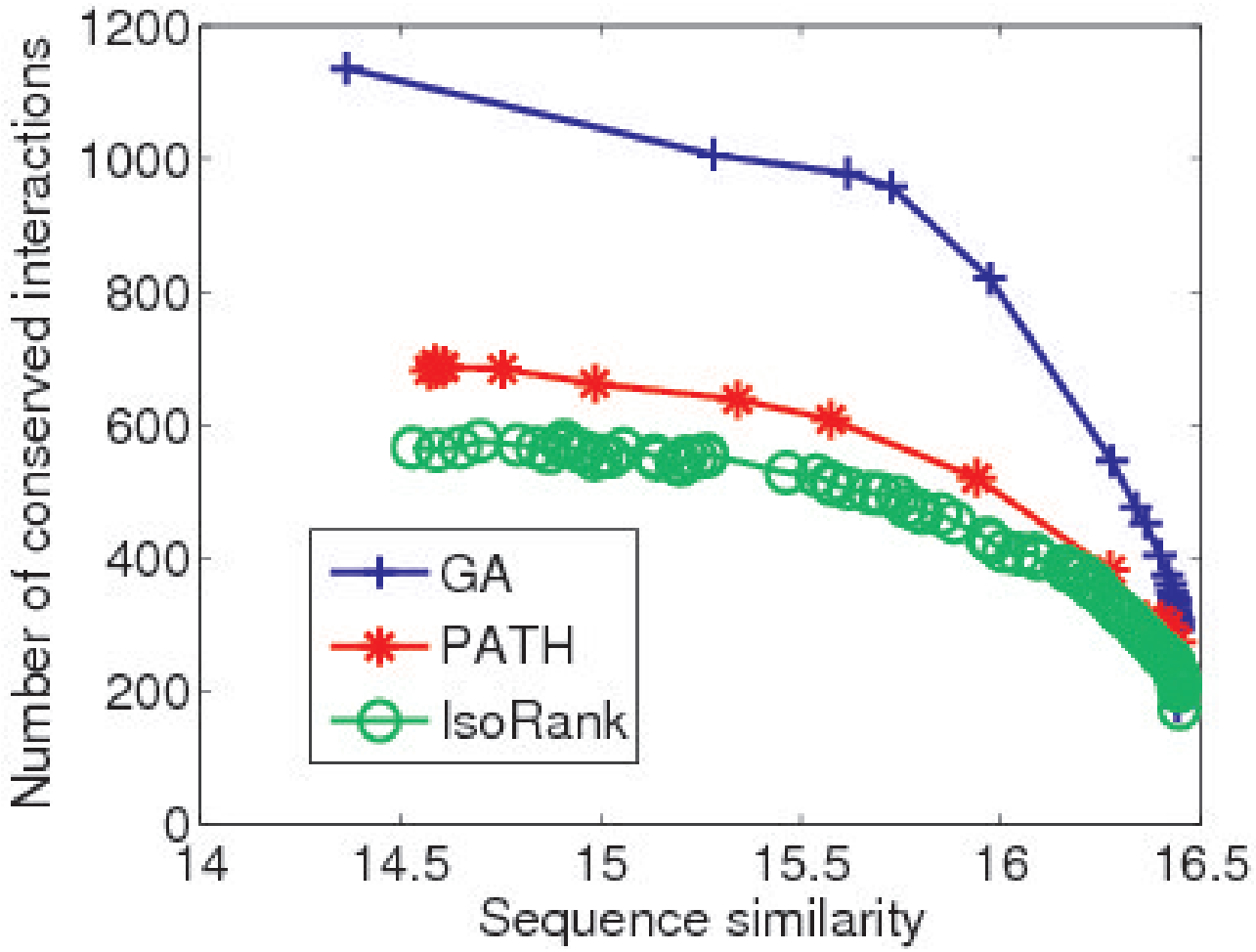
Differences MP vs MRF: example



Solid red: interaction conserved by MP; Dotted black: interactions conserved by MRF.

Experiments: Balanced Alignment

Maximize: $(1-\lambda)J + \lambda S$



Number of conserved interaction J versus sequence similarity S.

Outline

Conclusion

What we did

- Formulation of biological network alignment as a graph matching problem
- Message passing algorithm: exact solution for the constrained alignment problem
- Graph matching algorithms: good performance in the case of balanced alignment.

Future work

- Interactions of a higher order (see paper)
- Synchronized alignment of several networks
- Many-to-Many graph matching

Acknowledgements

Misha Zaslavskiy

Francis Bach

This presentation is supported by a JSPS Invitation Fellowship Program for Research in Japan, hosted by Tatsuya Akutsu (Kyoto University)