GRAPHLETALIGN AND MCDA: COMPUTATIONAL EXPLORATION OF PROTEIN FUNCTIONS BASED ON BIOMOLECULAR INTERACTIONS

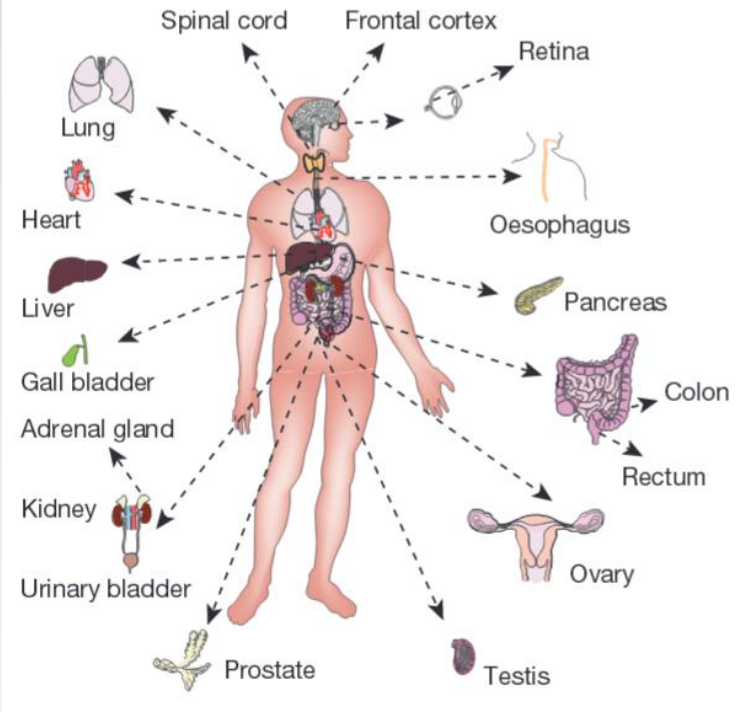
Mu-Fen Hsieh, Sing-Hoi Sze

Department of Computer Science and Engineering

Texas A&M University

Introduction: Computational Protein Function Analysis

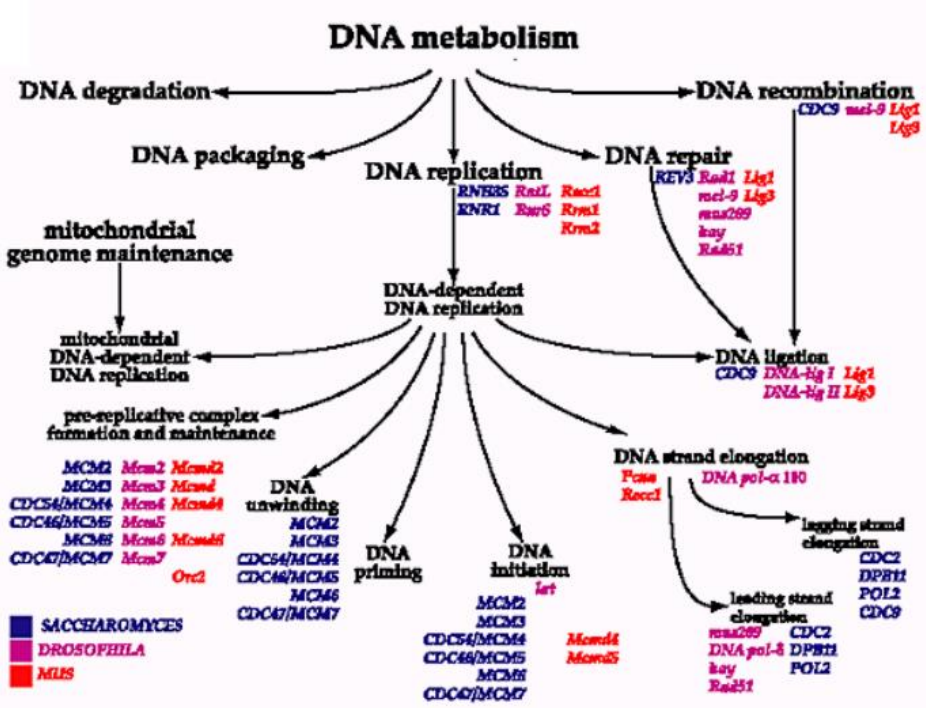
- Proteins: Basic working units of cell machinery. Participate in basic physiology, cell regulation, flow of genetic information, cell cycle, transportation, response to environments, cell structure, and more
- Protein function analysis: Sequence the protein, infer the protein structure, and explain the protein properties and functions
- Computational (automated) protein function annotation: compare protein sequences, domains, structures or phylogeny to transfer annotations



Kim *et al*., "A draft map of the human proteome", *Nature*, '14

Introduction: Protein Function Classification Schemes

- Gene Ontology (GO): The “Biological Process” ontology has 26752 terms at 15 levels in May 2014
- MIPS FunCat: 1307 categories at 5 levels



Ashburner *et al.*, 2000

ID	Description	#proteins
01	metabolism	1390
02	energy	327
10	cell cycle and DNA processing	965
11	transcription	995
12	protein synthesis	460
14	protein fate	1106
16	protein with binding function	1008
18	regulation of metabolism	240
20	cellular transport	974
30	cellular communication	230
32	cell rescue, defense and virulence	509
34	interaction with the environment	443
38	transposable elements	29
40	cell fate	264
41	development	66
42	biogenesis of cellular components	822
43	cell type differentiation	430

Introduction: Protein Function Classification Schemes (cont')

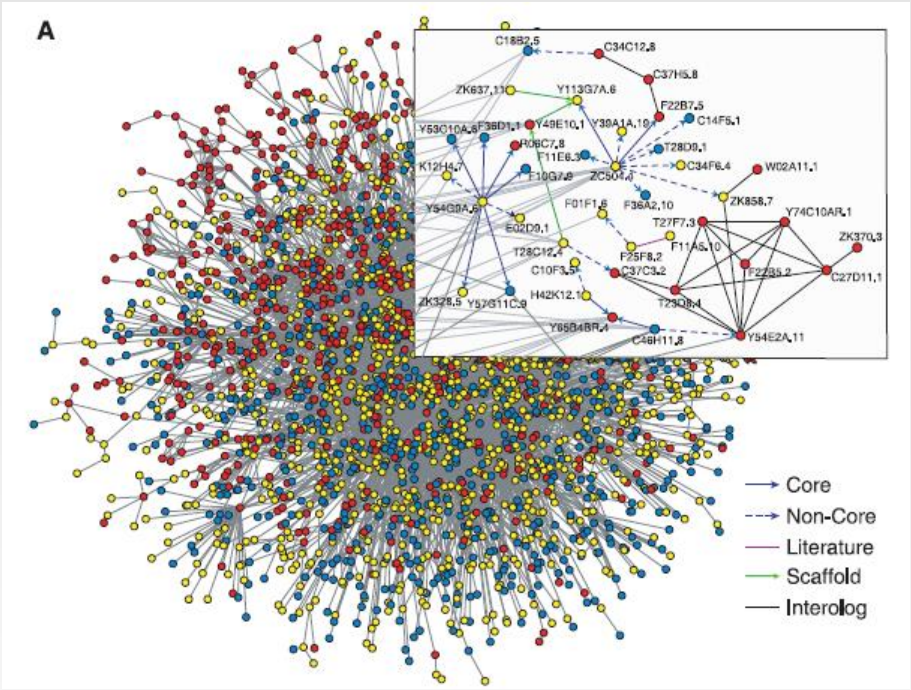
- A protein can have more than one function
- Representative GO terms shared by a group of proteins can be determined by sample probabilities from a hypergeometric distribution

Suppose GO term *g* annotates *l* out of *n* genes, the probability of seeing >*k* out of *s* genes annotated by g:

$$\Pr(x \geq k|g) = 1 - \sum_{i=0}^{k-1} \frac{\binom{l}{i}\binom{n-l}{s-i}}{\binom{n}{s}}$$

Introduction: Biomolecular Interactions

- Protein-Protein Interaction (PPI): physical contact of two proteins; Performs protein functions
- Genetic Interaction: observation of phenotype associated with processing of two genes; Represents how genes regulate cellular processes together



C. elegans network (Li *et al*., 2004)

- Network/graph, $G(V(G), E(G))$
  V: protein/node/vertex
  E: interaction/link/edge
  Neighbors, $N : V(G) \rightarrow \{V(G)\}$
- Database: BIND, MINT, **DIP**, **SNDB**, **BioGRID**, **IntAct**, the human protein specific database (HPRD) and I2D

Introduction: Biomolecular Interactions and Protein Function Analysis

- Hypotheses about interactions and functions
  - Local density enrichment: Physically interacting proteins are more likely to share a protein function
  - A more frequent function in a gene/protein's neighborhood can probably be the function for this gene/protein
  - Similar genetic profiles (i.e. similar sets of genetic interacting partners) are associated with similar cellular functions. Two proteins are likely to have the same function if they share interaction partners.

Outlines: Biomolecular Interactions and Protein Function Analysis

- **Local network alignment**: Identification of protein functional modules in a network or across multiple networks
  - Graphlet alignment of protein-protein interaction networks (GraphletAlign)
- **Prediction of protein function**
  - Regularized multi-label canonical discriminant analysis (MCDA) based on genetic and protein-protein interactions

Unknown protein	Predicted function
u_1	g_1, g_2, g_3
u_2	g_1
u_3	g_2
u_4	g_1, g_3
u_5	g_3

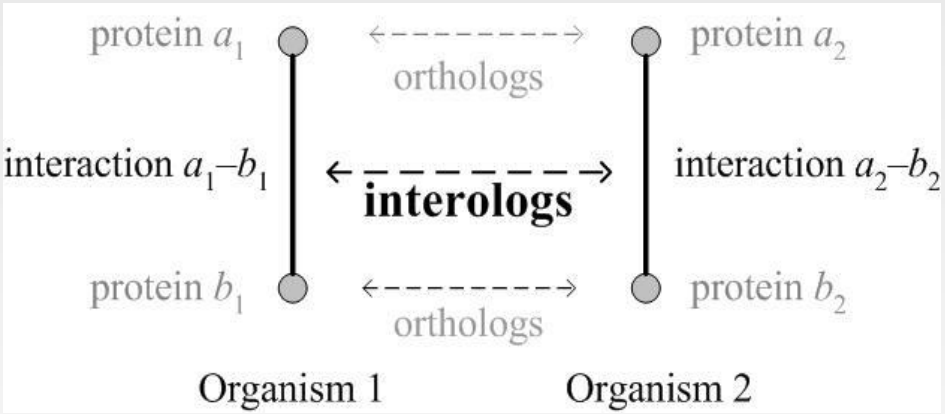
Functional module	Predicted function
H_{11}	g_1, g_2
H_{12}	g_1
H_{13}	g_2
H_{21}	g_1, g_3
H_{22}	
H_{23}	g_3

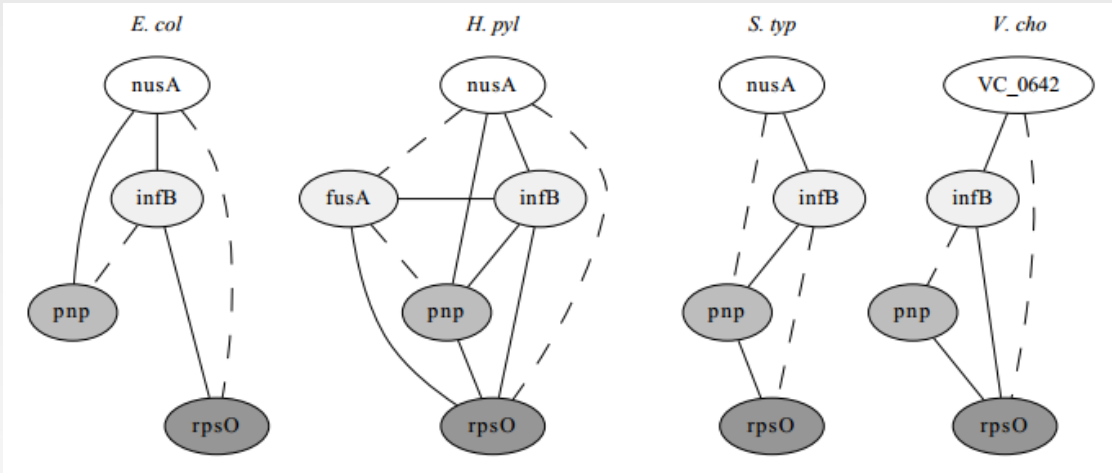
Local Network Alignment

GraphletAlign: Graphlet alignment of protein-protein interaction networks

Local Alignment of Protein-Protein Interaction Networks

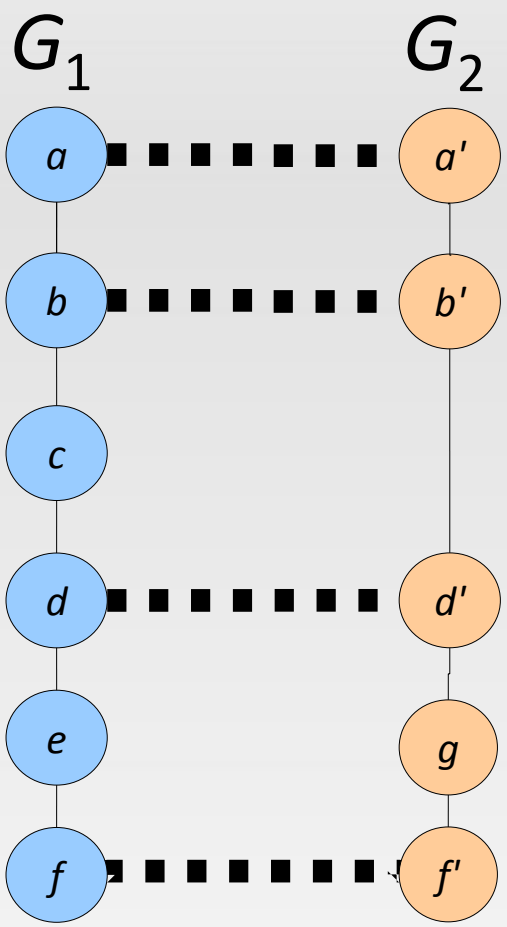
- A computational method to compare PPI networks across different species or within a single species
- Input: One or multiple PPI networks
- Output: Alignment of conserved subnetworks
- Applications: Identification of conserved functional modules and protein function prediction





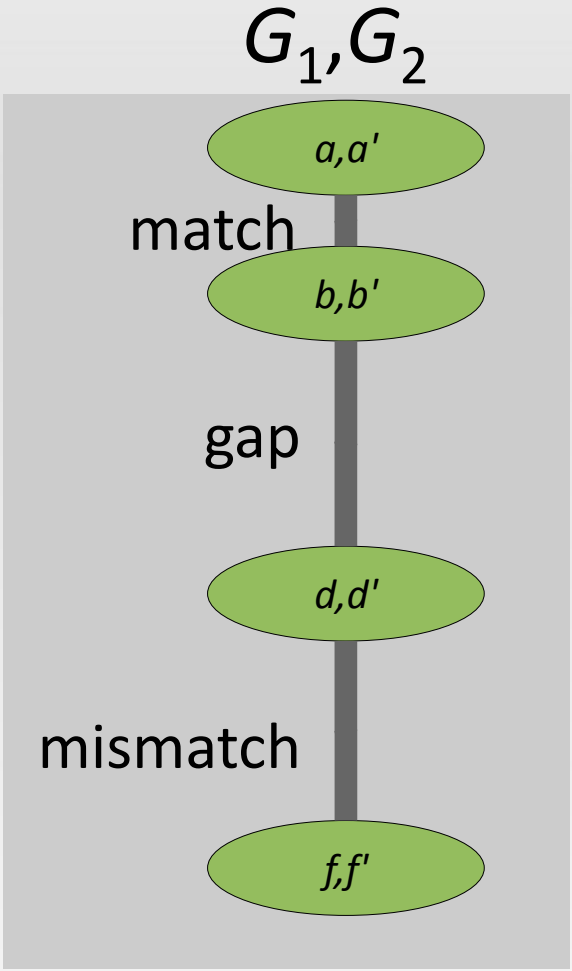
Local Alignment of Protein-Protein Interaction Networks (cont')

Two PPI Networks



PathBLAST (Kelley *et al.*, 2003)

Alignment Graph



A Pathway Alignment

Challenges: Local Network Alignment

- Numerous possibilities of vertex matches and the edge matches; Not possible to store an alignment graph of multiple species
- Select good network alignments
- A protein may appear in many network alignments

Related Work: Local Network Alignment

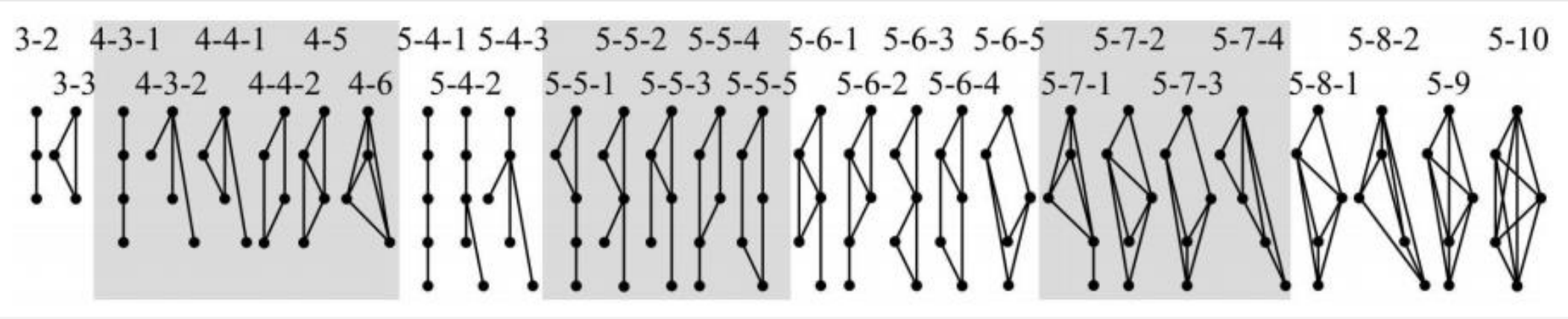
- PathBLAST (Kelley *et al*., 2003): Identifies maximum weighted paths in directed acyclic pairwise alignment graph
- NetworkBLAST (Sharan *et al*., 2005) and MaWISh (Koyutürk *et al*., 2006): Identifies maximum weighted induced graph in the pairwise alignment graph by seed-and-extension
- Græmlin (Flannick *et al*., 2006): Pairwise seed-and-extension alignment consists of equivalent homolog classes; Progressive multiple network alignment following phylogeny
- MNAligner (Li *et al*., 2007): Solves integer quadratic programming (IQP) problem to find max-scoring matching nodes and edges between two small dense networks

Related Work: Local Network Alignment

- CAPPI (Dutkowski *et al*., 2009): Finds multiple network alignment corresponding to a common ancestral network inferred using tree-structured Bayesian network
- NetworkBLAST-M (Kalaev *et al*., 2009): Finds high-scoring d-subnet connected by k-spines homology association in a ”layered alignment graph”
- DOMAIN (Guo and Hartemink, 2009): Finds high-scoring subgraph in a pairwise APE graph (alignable pairs of edge) considering domain-domain interactions

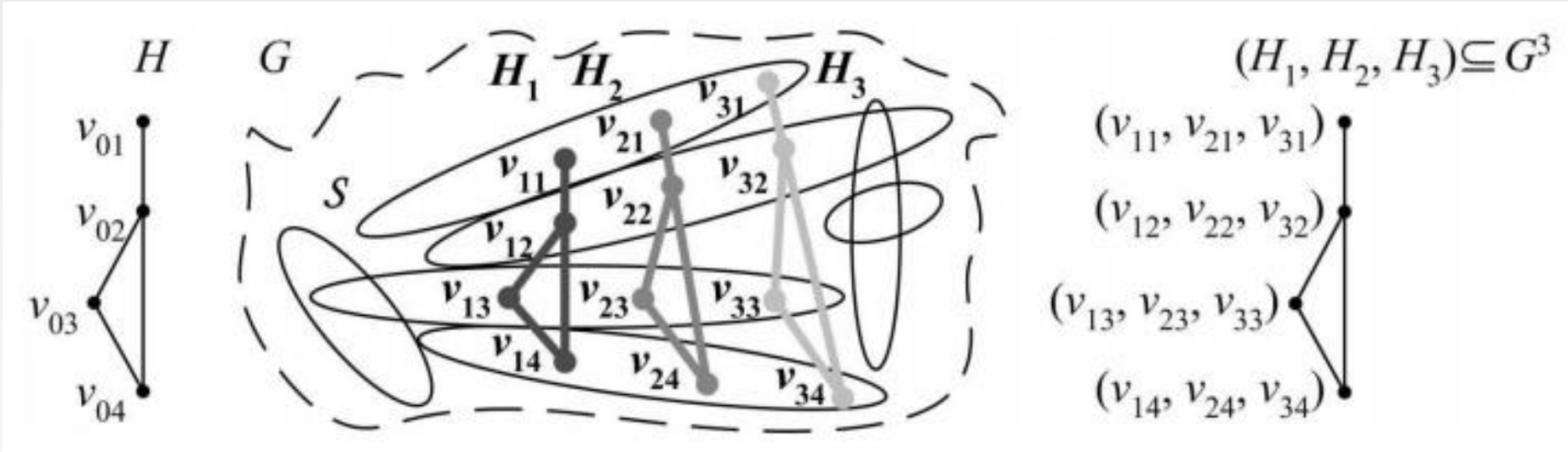
GraphetAlign, Graphlet Alignment of PPI Networks: Motivations

- Incorporate more proteins in the network alignment and cover more protein functions
- Align subnetworks of a single species network
- Based on the concept of graphlet (Prulj *et al.*, 2004)



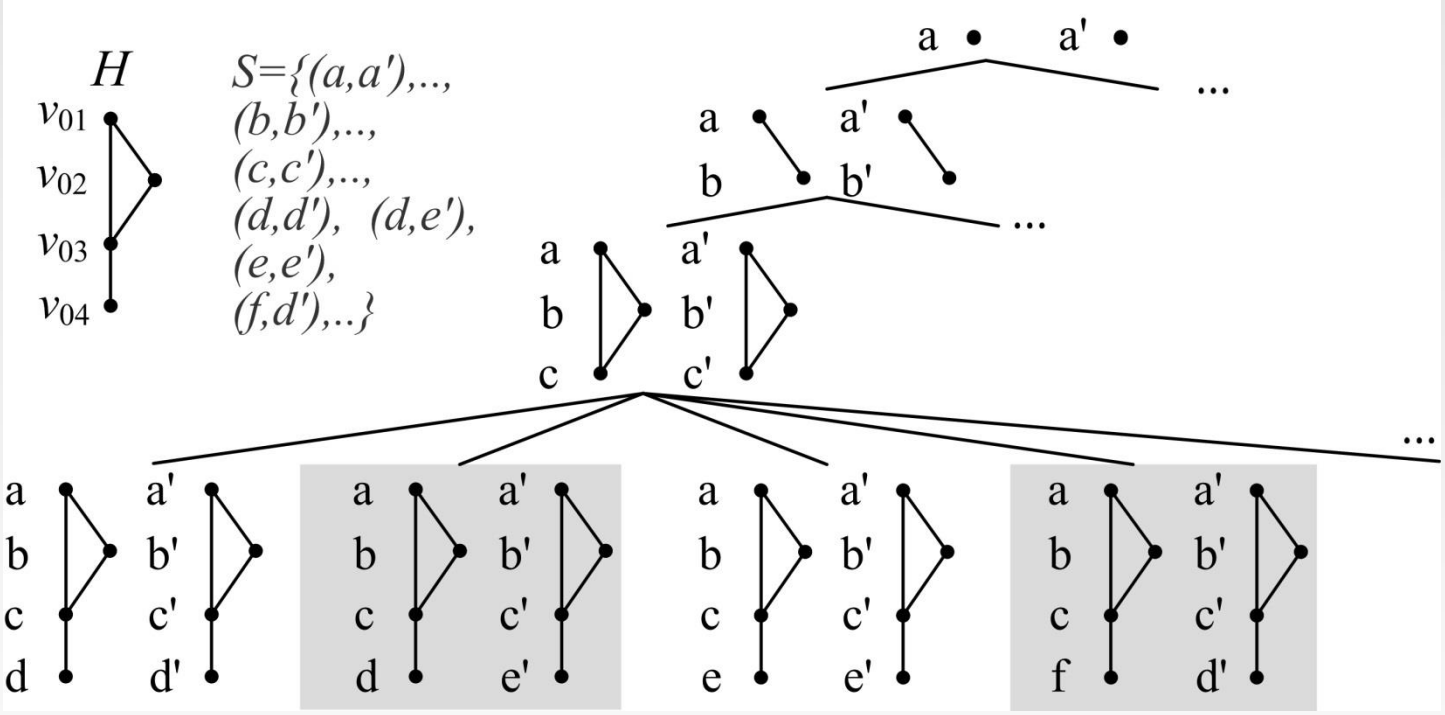
Problem Formulation: Graphlet Alignment and Clustering

- Enumerates $H_1,..,H_m$ graphlet alignments in a joined networks $G(V,E)$ of m networks, given a target graphlet $H(V_0, E_0)$, and S, the size-m sets of homologs
  - Allows distance-2 gaps between two proteins in at most m-1 species
  - Considers only the induced subgraphs on the alignment graph
- Performs clustering on overlapping graphlet alignments



Enumerating Graphlet Alignments: Branch-and-Bound Algorithm

- Bounding conditions: No adjacent nodes that
  - Have disjoint vertices
  - Satisfy the induced topology of the next node in given *H*
  - Have a label satisfy symmetry breaking conditions
- $O\left(|V|^m \delta^{m(|V_0|-1)}\right)$, δ: the largest vertex degree

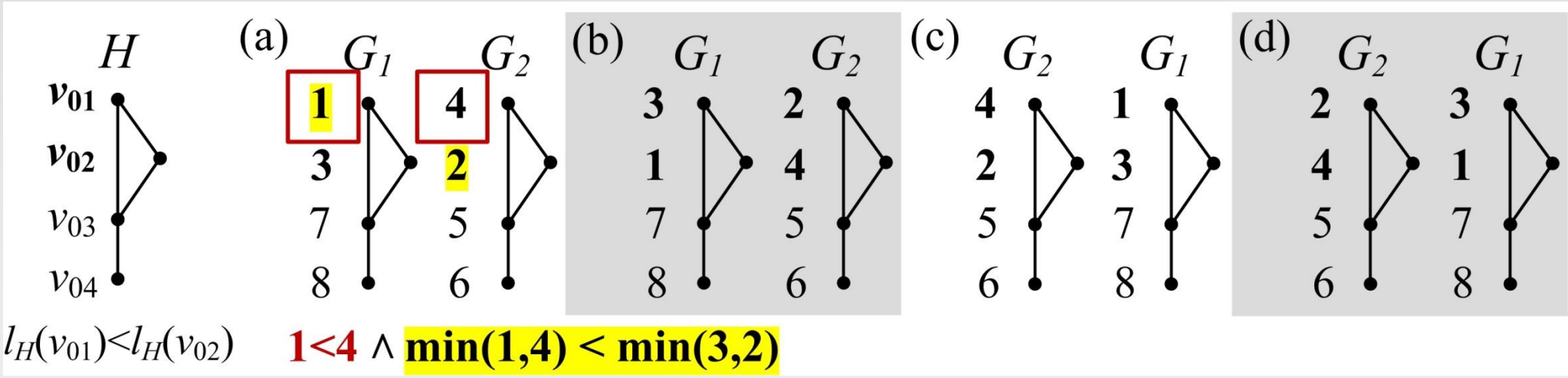


Enumerating Graphlet Alignments: Symmetry Breaking Conditions

Proteins in G_1 and G_2 are randomly and distinctly numbered

Given ordered sets of homologous proteins

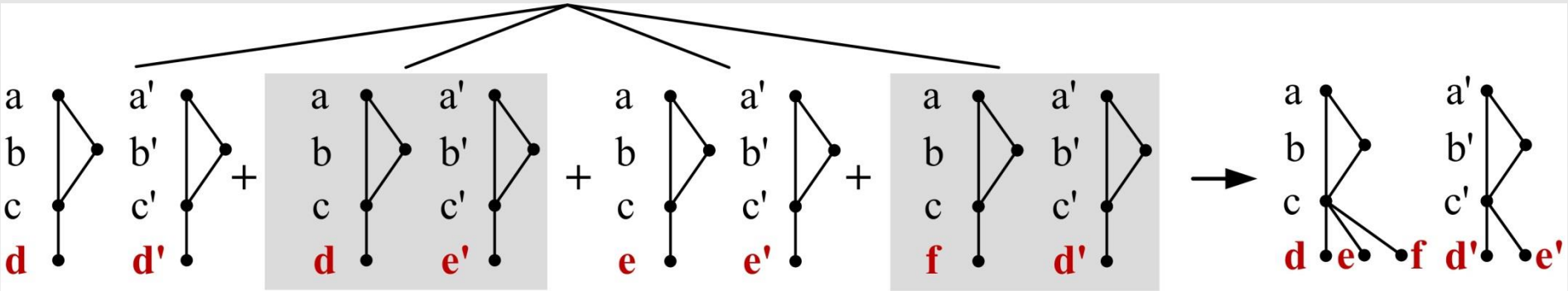
S={(1,4),(4,1),(2,3),(3,2),(5,7),(7,5),(6,8),(8,6)}



- Two types of symmetry:
  - Automorphism: (a) vs (b), (c) vs (d)
  - Instance symmetry: (a) vs (c), (b) vs (d)

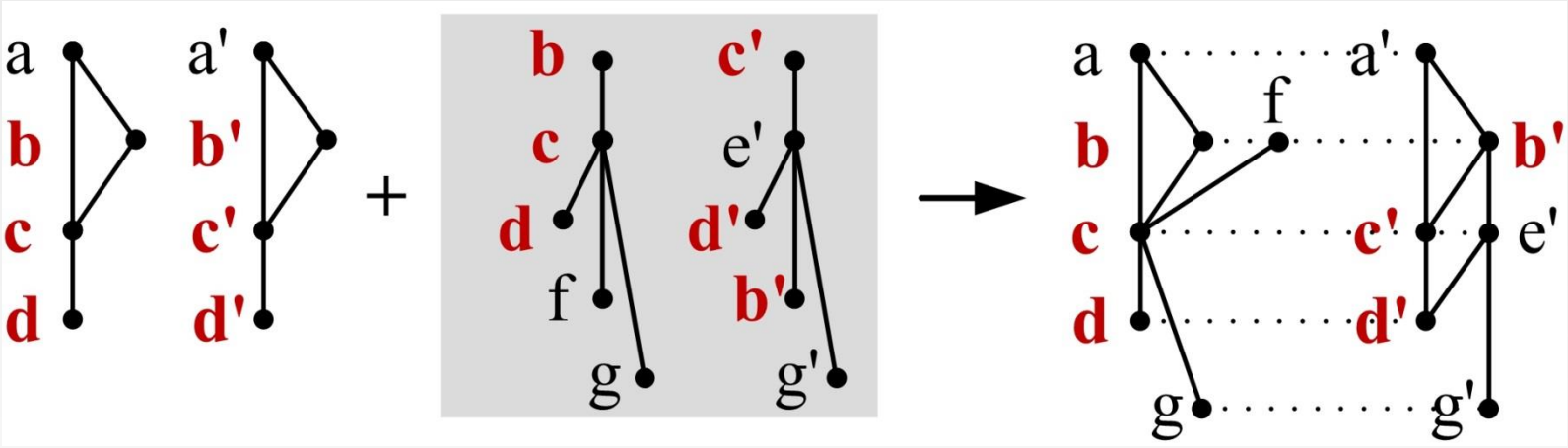
Post-processing: Clustering of Graphlet Alignments

- Stage 1: Preclustering at the last level of the search tree



Post-processing: Clustering of Graphlet Alignments (cont')

- Stage 2: Greedily combining large high-scoring alignments that contain a certain protein
  - Alignments sorted by decreasing total number of proteins, conserved edges, and minus log product of BLAST e-value of every aligned protein pairs
  - Iteratiely combine alignments if the overlap is > 50% of average protein size on every instance
    - Max 100 proteins per instance



GraphletAlign: Experiment Data and Output Statistics

- 2, 3 and 4-species PPI networks come from IntAct, DIP, and Stanford Network (SNDB), respectively
- BLAST *E*-value < -10^7
- Targets to generate at least 2×10^7 graphlet alignments

Table 1: Input data for local network alignment, and and parameters and output statistics for the Graphlet Alignment.

	Input		Graphlet Alignment Parameters			Graphlet Alignment Output				
Species networks	Pro -teins	Inter -actions	Grphlt copies aligned	Max grphlt size	Blastp top BH†	Sets of grphlts	Sets of mod -ules	Avg. mod. size	Max. mod. size	Aligning time (m)
Yeast	7446	22734	2	3	10	3.55e7	427	59.02	117	662.24
Human	6294	13455	2	3	5	2.97e7	435	44.96	177	883.16
Fly + yeast	12374	40111	2	5	15	4.79e7	923	35.12	100	2286.81
Human + mouse	7455	14545	2	4	276	2.2e7	485	21.45	109	1246.76
Fly + worm + yeast	14254	42361	3	4	40	3.45e7	506	17.92	100	1011.89
Four microbes*	10339	64890	4	3	35	7.59e7	74	9.29	41	1635.11

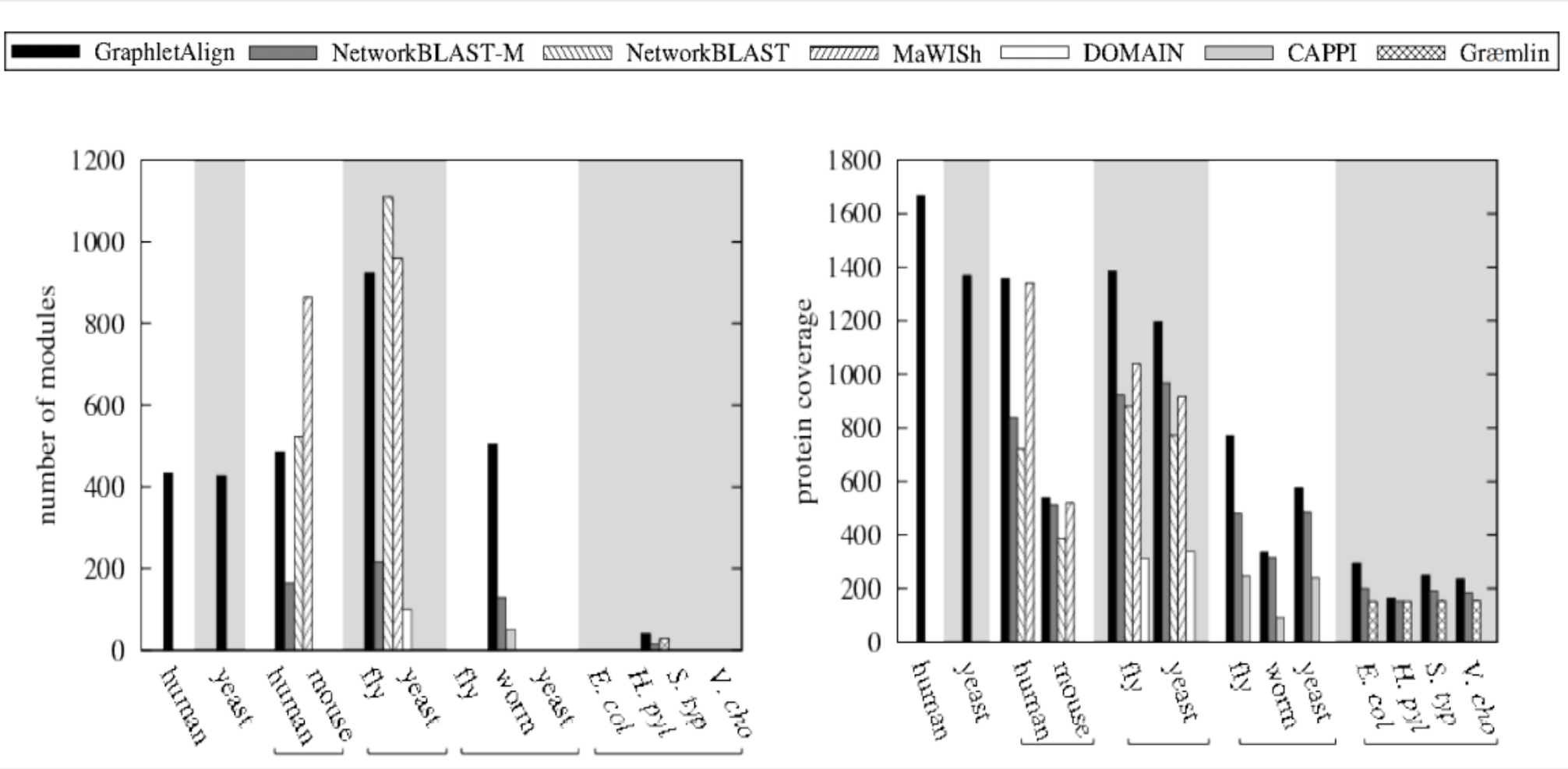
* The four microbes aligned are E.col + H. pylori + S. typhimurium + V. cholerae.

† BH denotes "birectional hits".

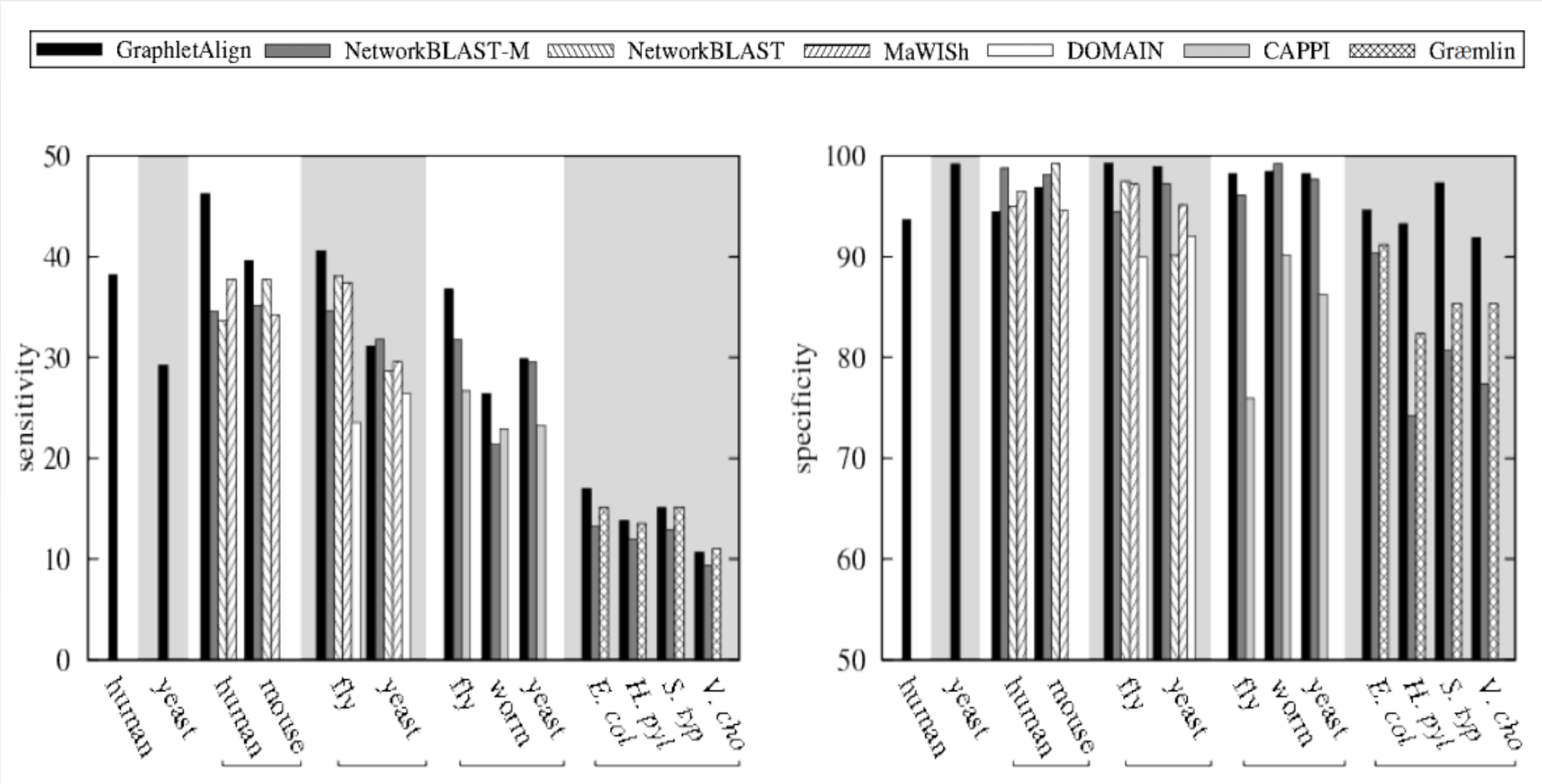
GraphletAlign: Experiment Measures

- Accuracy and coverage of functional modules
  - No ”benchmarks” for protein functional modules
  - Hypothesis review: ”Physically interacting proteins are more likely to share a protein function”
  - Used Gene Ontology (GO) Term Finder to find ”Biological Process” terms significantly presented in a functional module with corrected p-value < 0.05
- Measures for each functional module of each species
  - **Coverage**: Total number of proteins
  - **Specificity**: % of functional modules significantly associated with a GO term
  - **Sensitivity**: % of the 318 ”biological process” level-3 GO terms associated with a functional module

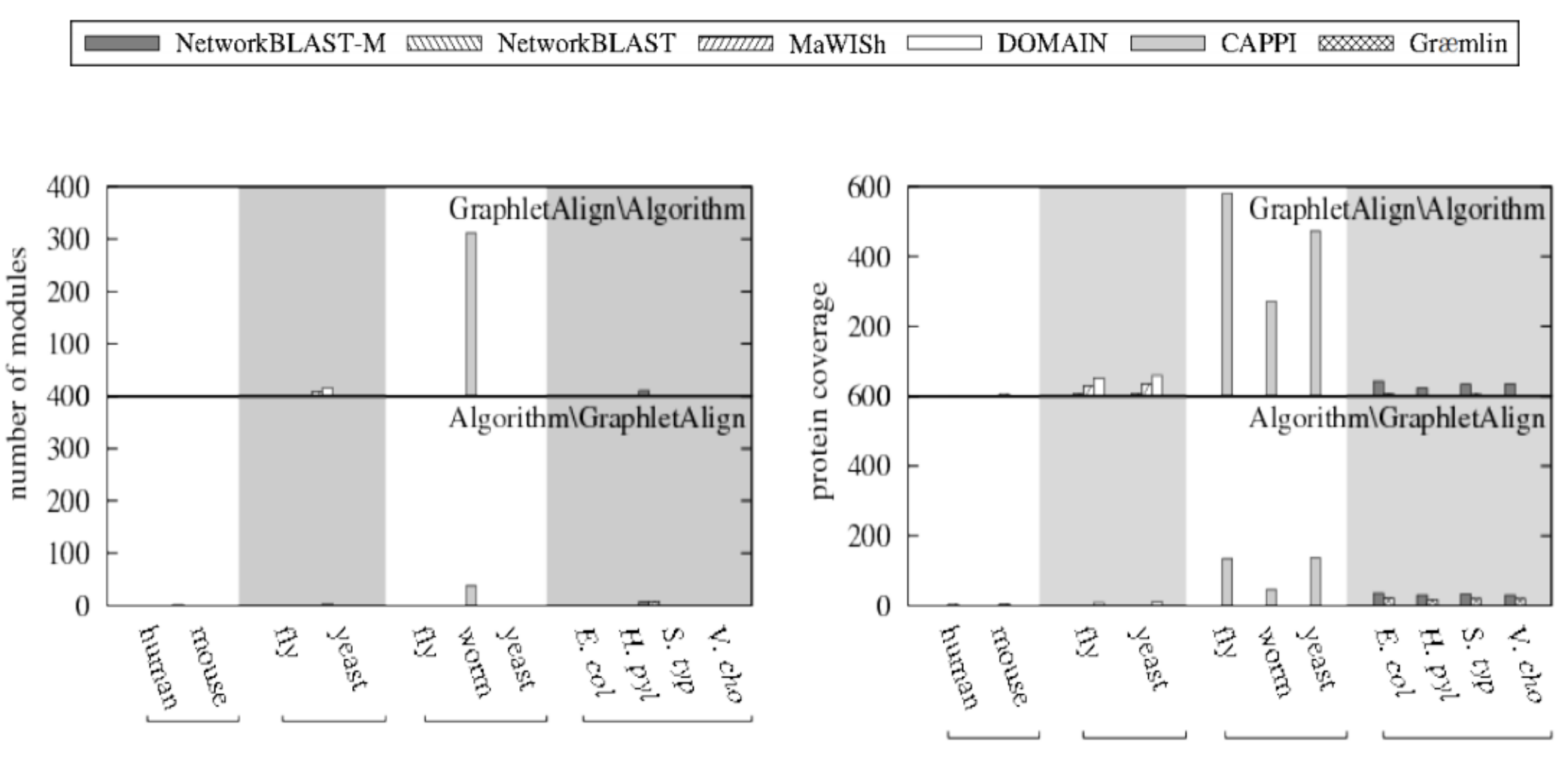
GraphletAlign: Number of Modules and Protein Coverage



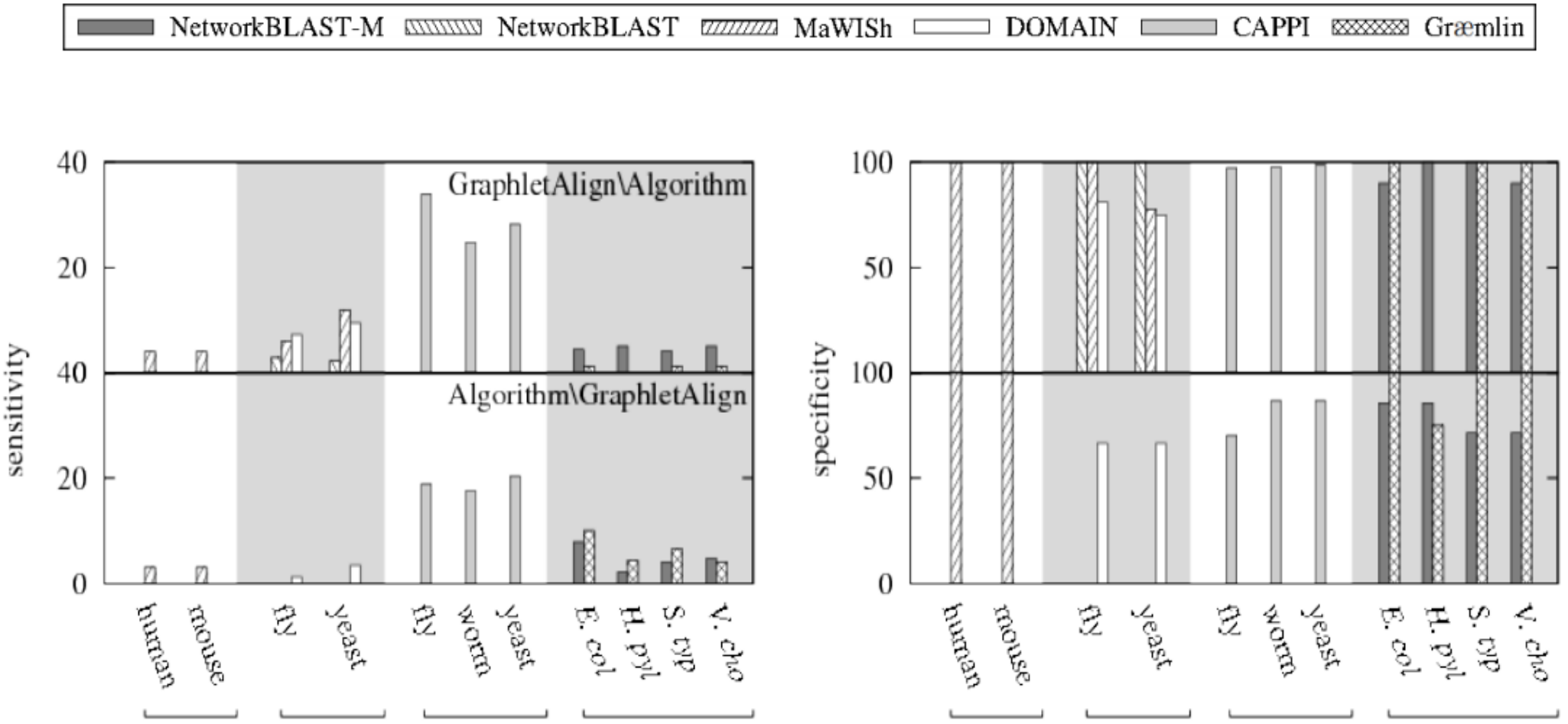
GraphletAlign: Sensitivity and Specificity



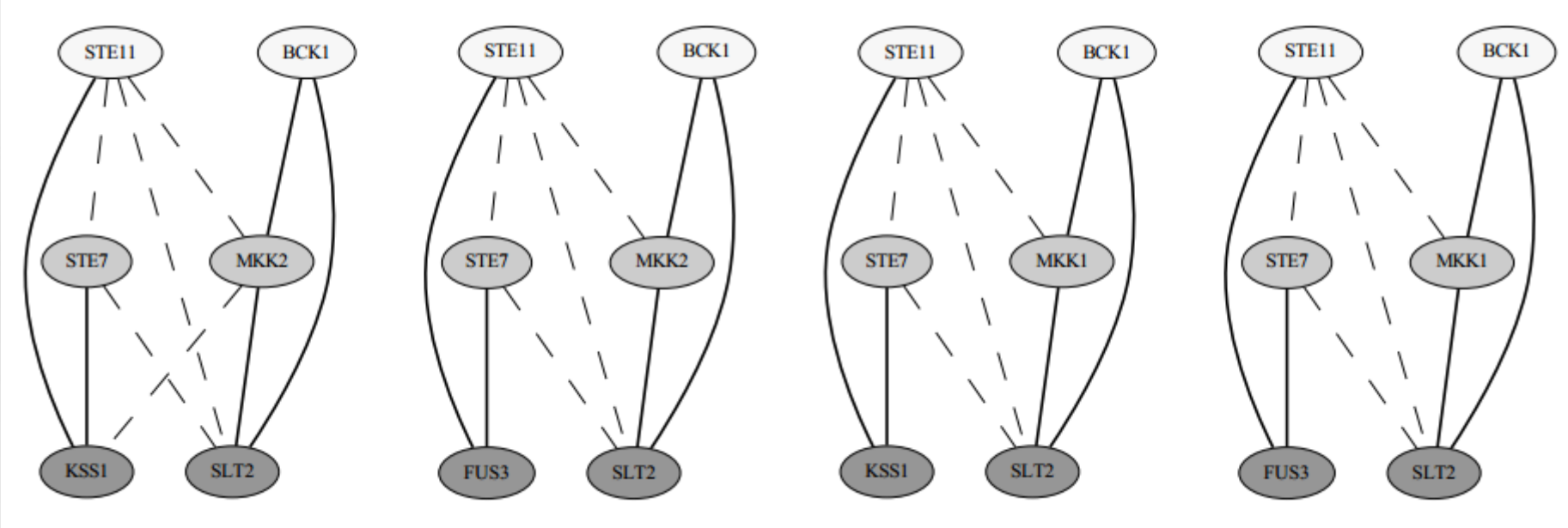
Unique Functional Modules: Number of Modules and Protein Coverage

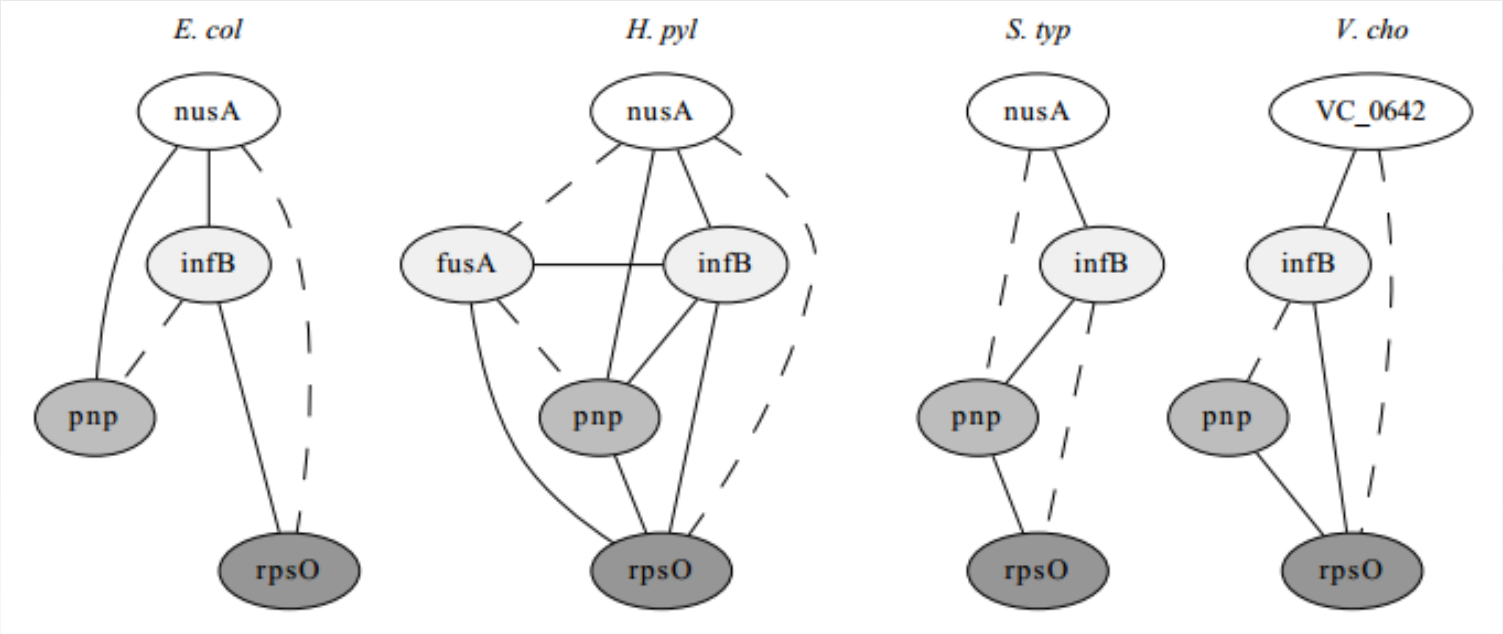


Unique Functional Modules: Sensitivity and Specificity

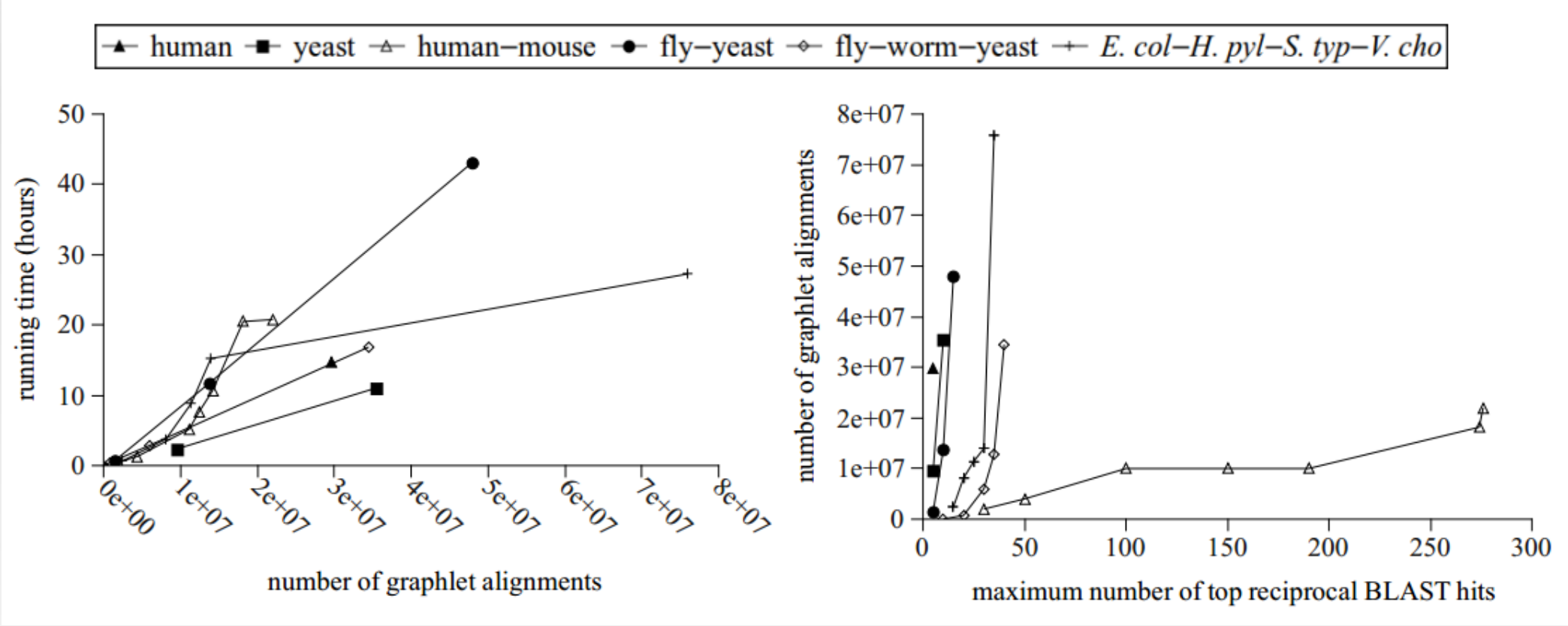


Examples of a Graphlet Alignment and Conserved Functional Modules

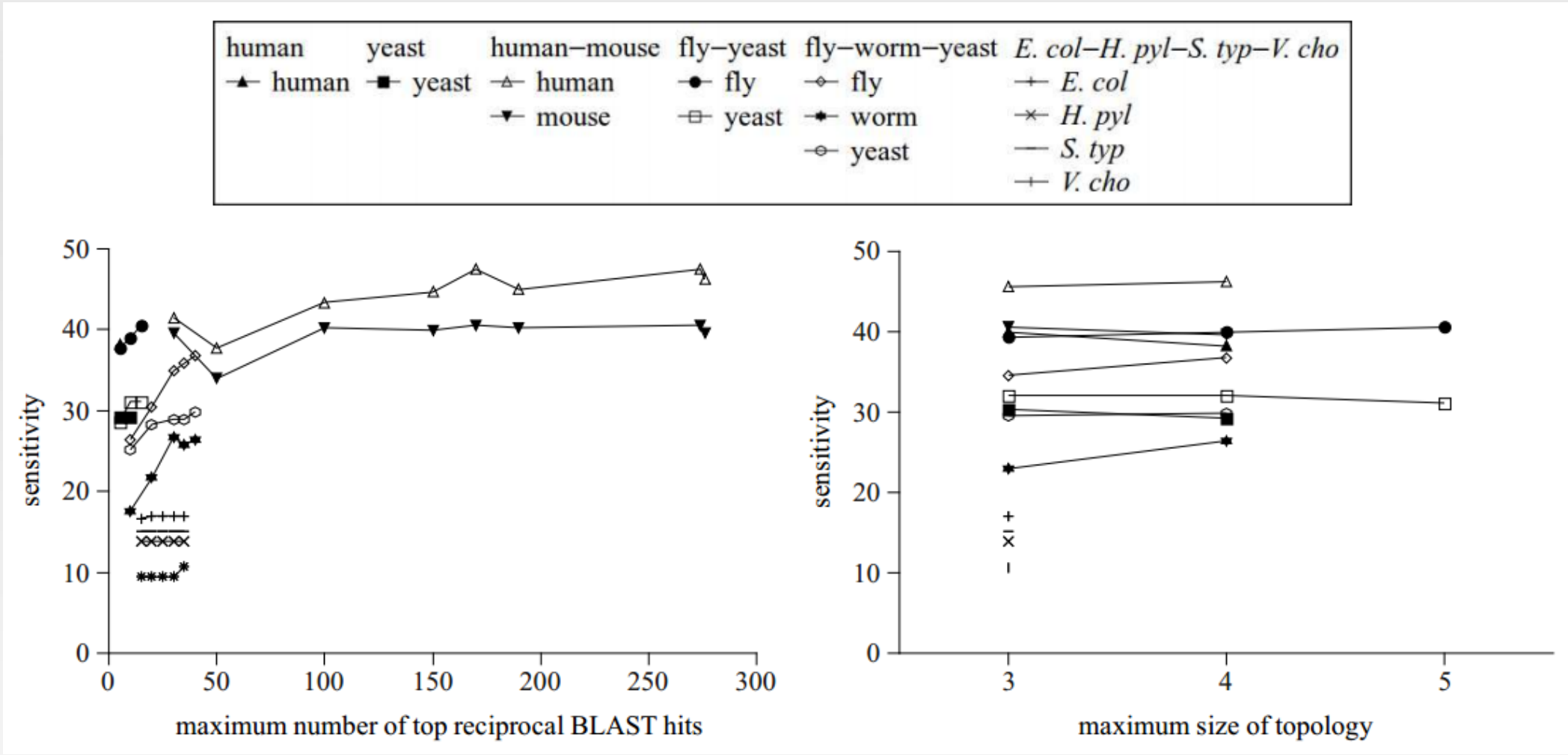




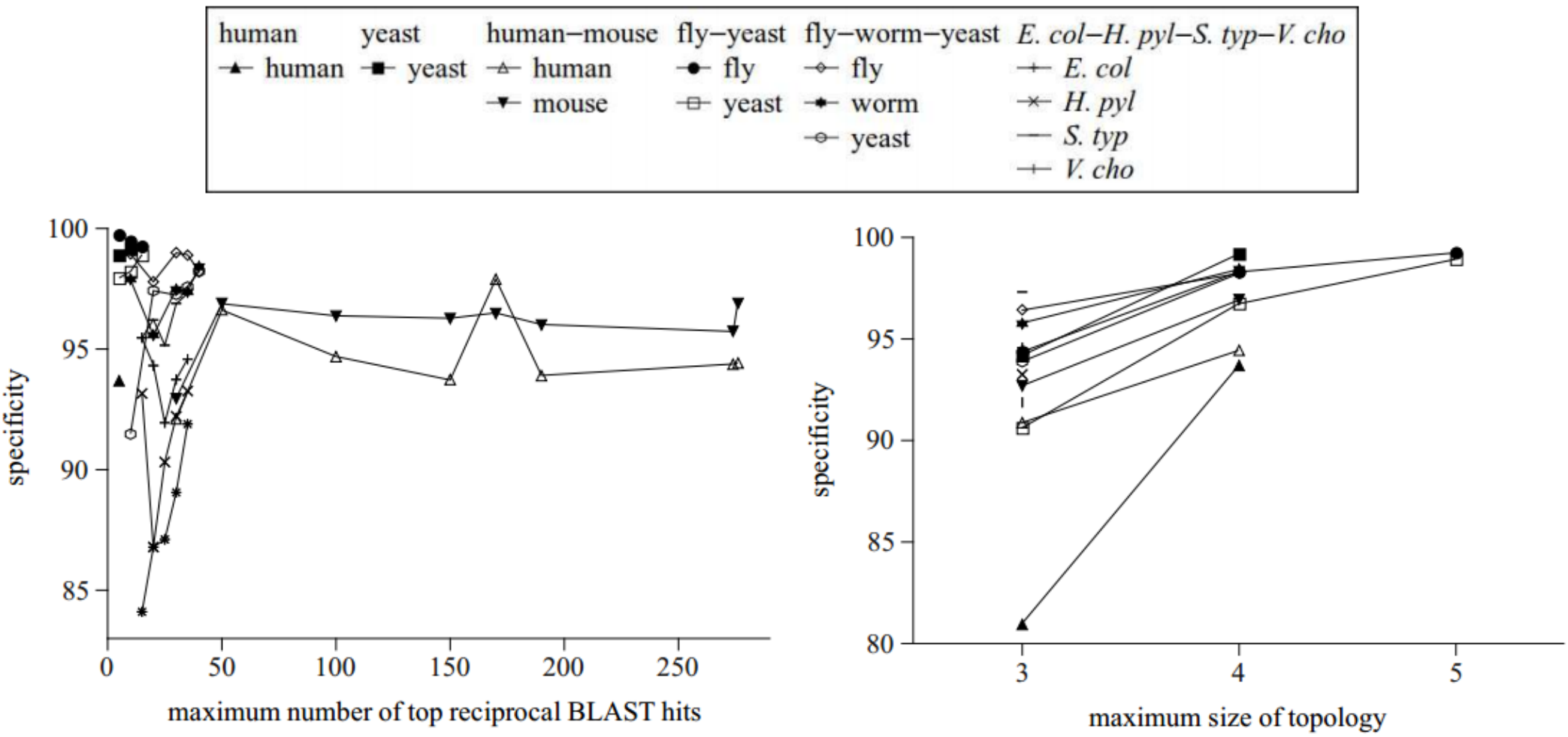
GraphletAlign: Running Time and Number of Alignments



GraphletAlign: Trends of Sensitivity



GraphletAlign: Trends of Specificity



GraphletAlign: Summary and Future Work

- GraphletAlign, graphlet alignment of protein-protein interaction networks and clustering is presented
- GraphletAlign identified functional modules that have higher protein coverage, higher or comparable sensitivity and specificity to other local network alignment software
- GraphletAlign can be used to align subnetworks of a single network
- Future work: Produce non-overlapping functional modules, and score based on distribution of conserved graphlets under real model and random model

Protein Function Prediction

Regularized multi-label canonical discriminant analysis (MCDA) based on genetic and protein-protein interactions

Protein Function Prediction: Formulation

- Input

  K: known protein, *U*: unknown protein

  F: collection of biological data for *K* and *U*

  G: labels of functional groups

  L_K: protein-label association for *K*

$$\hat{L_U}(p) = \arg\max_{g \in G}\{\hat{\Pr}(g|p)\}, p \in U$$

- Output

  L_U: protein-label association for *U*

$$\hat{L_U}(p) = \{g|\hat{\Pr}(g|p) > t\}, p \in U$$

Multiclass classification

Protein	Functional group
p_1	g_1
p_1	g_2
p_1	g_3
p_2	g_1
p_3	g_2
p_4	g_1
p_4	g_3
p_5	g_3

Multiclass binary classification

Function	Member	Non-member
g_1	p_1, p_2, p_4	p_3, p_5
g_2	p_1, p_3	p_2, p_4, p_5
g_3	p_1, p_4, p_5	p_2, p_3

Multi-label classification

	g_1	g_2	g_3
p_1	1	1	1
p_2	1	0	0
p_3	0	1	0
p_4	1	0	1
p_5	0	0	1

Protein Function Prediction: Challenges

- How to integrate protein sequence data and interaction data for protein function prediction
- How to associate interactions of protein pairs with individual protein functions
  - Hypotheses of interactions and protein funtions
- A protein can have more than one function, and the functions can be correlated and have unbalanced distribution

Protein Function Prediction: Related Work

- Neighborhood counting, guilt-by-association, majority voting method (Schwikowski *et al.*, 2000): Predicts for a known protein the top 3 frequent functions among its interacting partners
- Global neighborhood optimization and minimum multiway cut approach (Vazquez *et al.*, 2003): Predicts a function *g* for an unknown protein *p* that maximizes its sharing of *g* with neighboring proteins
- Multiple-kernel and SVM approach (Lanckriet *et al.*, 2004): Applies semi-definite programming (SDP) to combine pairwise similarity matrices (kernel matrices) computed from different types of data, and uses the support vector machine (SVM) classification algorithm on the kernel matrix and binary label matrix.

Protein Function Prediction: Related Work

- Multiple-kernel and min-max approach (Tsuda and Noble, 2005): Formulates a convex optimization problem for combining multiple kernels and solved the equivalent min-max problem using Lagrange multipliers.
- Maximization of data-knowledge consistency (MDKC) (Wang *et al*., 2015): Solves non-negativity matrix factorization problem that minimizes the difference between data similarity and functional label similarity. Calculates functional group correlation using cosine product over design matrix of labels.

Protein Function Prediction: Goals

- Input: Protein sequences, genetic and protein-protein interactions of baker's yeast; MIPS Funcat categories, Yeast Gene Ontology Slim terms for yeast proteins
- Output: Prediction of Funcat, Yeast Gene Slim terms for yeast proteins
- Propose a 3-stage framework that enables application of traditional methods of multi-label classification
- Propose a multi-label prediction method, MCDA, that considers function correlation
  - Apply inner product over design matrix of functional labels

Design matrix

	g_1	g_2	g_3
p_1	1	1	1
p_2	1	0	0
p_3	0	1	0
p_4	1	0	1
p_5	0	0	1

Protein Function Prediction
Stage 1: Distance Matrix Δ

- Protein sequence distance: Smith and Waterman alignment

$$\mathrm{PAM}(s_i, s_j) = 10 \times (10 \hat{} Pr(\frac{s_i \text{ and } s_j \text{ have common ancestor}}{s_i \text{ and } s_j \text{ are random alignment}}) - 1)$$

$$\begin{aligned} \mathrm{SW}(p_i, p_j) &= 1 - Pr(\frac{s_i \text{ and } s_j \text{ have common ancestor}}{s_i \text{ and } s_j \text{ are random alignment}}) \\ &= 1 - \log_{10}(\mathrm{PAM}(s_i, s_j)/10 + 1) \end{aligned}$$

- Distance between genetic or protein-protein interactions: Sørensen-Dice (SD) dissimilarity or Jaccard (Jc) distance

$$\mathrm{SD}(p_i, p_j) = 1 - \frac{|N(p_i) \cap N(p_j)|}{2|N(p_i) \cup N(p_j)|} \qquad \mathrm{Jc}(p_i, p_j) = 1 - \frac{|N(p_i) \cap N(p_j)|}{|N(p_i) \cup N(p_j)|}$$

  - Hypothesis: Similar sets of genetic interacting partners are associated with similar cellular functions. Two proteins are likely to have the same function if they share interaction partners.

Stage 2: Multi-dimensional scaling (Mardia, 1978) to *X*

- Given a double-centered distance matrix Δ , decomposed into eigenvectors *V* and a matrix of eigenvalues at its diagonal Λ,

$$X_l = V_l \Lambda_l^{1/2} R$$

$$\text{minimize } \mathrm{tr}[(\Delta^* - X_l X_l^T)^2] \text{ subject to } RR^T = I$$

the MDS variable X_l, axes 1..*l*, are chosen by minimizing square error

Stage 3: Multi-label Canonical Discriminant Analysis (MCDA) to *Z*

- Background: classic canonical discriminant analysis (CDA) for multiclass prediction
  - Find canonical coefficients *C* to transform MDS variable *X* to canonical variable *Z*

$$Z = XC^T$$

$$C\Sigma_B C^T \quad \text{subject to} \quad C\Sigma_W C^T = I$$

$$(C\Sigma_W^{\frac{1}{2}})\Sigma_W^{-\frac{1}{2}}\Sigma_B\Sigma_W^{-\frac{1}{2}}(C\Sigma_W^{\frac{1}{2}})^T = F\Sigma_W^{-\frac{1}{2}}\Sigma_B\Sigma_W^{-\frac{1}{2}}F^T$$

$$\text{subject to} \quad FF^T = I$$

$$C = \Sigma_W^{-\frac{1}{2}}F$$

  - Decomposes $SS_E^{-\frac{1}{2}}SS_HSS_E^{-\frac{1}{2}}$ into eigenvectors *F* and eigenvalues λ

$\Sigma_B = SS_H/(m-1)$, $\Sigma_W = SS_E/(n-1)$

m: # functional classes, *n*: # proteins

Canonical correlation at *l*-axis: $\rho_l = \sqrt{\lambda_l/(1+\lambda_l)}$, $l = 1..q$.

Stage 3: Multi-label Canonical Discriminant Analysis (MCDA) to *Z*

- Background: classic canonical discriminant analysis (CDA) for multiclass prediction
  - Assume multivariate general linear model (GLM)

$$X = Y^T\beta + \varepsilon, \quad H_0 : H\beta = 0,$$

$$\hat{\beta} = (Y^TY)^{-1}Y^TX \text{ (by linear regression)}$$

$$\hat{SS}_H = (H\hat{\beta})^T\left(H(Y^TY)^{-1}(H)^T\right)^{-1}(H\hat{\beta})$$

$$\hat{SS}_E = \hat{\epsilon}^T\hat{\epsilon}$$

  - Predict g_j for canonical variable *z*

$$\hat{Pr}(g_j|z) = \left(q_j \exp\left(-\frac{1}{2}\text{Maha}^2(z, \hat{\mu}_{C,j})\right)\right) \Big/ \sum_{k=1}^{m}\left(q_k \exp\left(-\frac{1}{2}\text{Maha}^2(z, \hat{\mu}_{C,k})\right)\right)$$

Y

Protein	Functional group
p_1	g_1
p_1	g_2
p_1	g_3
p_2	g_1
p_3	g_2
p_4	g_1
p_4	g_3
p_5	g_3

Stage 3: Multi-label Canonical Discriminant Analysis (MCDA) to *Z*

- Assume multivariate general linear model (GLM)

$X = D^T\beta + \varepsilon,$ $H_0 : H^{(j)}\beta = 0$ for each group g_j

$\hat{\beta} = (D^T D)^{-1} D^T X$ (by linear regression)

D

	g_1	g_2	g_3
p_1	1	1	1
p_2	1	0	0
p_3	0	1	0
p_4	1	0	1
p_5	0	0	1

√ $\hat{SS}_H^{(j)} = (H^{(j)}\hat{\beta})^T \left(H^{(j)}(D^T D)^{-1}(H^{(j)})^T \right)^{-1} (H^{(j)}\hat{\beta})$

√ $\hat{SS}_H = \sum_{j=1}^{m} \hat{SS}_H^{(j)}$

$\hat{SS}_E = \hat{\epsilon}^T \hat{\epsilon}$

- Problem: $(D^T D)$ or $\widehat{SS_E}$ can be singular
  - Pseudo-inverse
  - Regularization

Stage 3: "Regularized" Multi-label Canonical Discriminant Analysis to *Z*

- Regularization form: $A_r = \alpha A + (1 - \alpha)I$
- Assume multivariate general linear model (GLM)

$$X = D^T\beta + \varepsilon, \quad H_0 : H^{(j)}\beta = 0 \text{ for each group } g_j$$

$$\hat{\beta}^{\text{ridge}} = (D^T D + \lambda^{\text{ridge}} I)^{-1} D^T X,\ 10^{-2} \le \lambda^{\text{ridge}} \le 10^{10}$$

$$(\hat{SS}_H^{\text{ridge}})^{(j)} = (H^{(j)}\hat{\beta}^{\text{ridge}})^T \left(H^{(j)}(D_r^T D_r)^{-1}(H^{(j)})^T\right)^{-1} (H^{(j)}\hat{\beta}^{\text{ridge}})$$

$$\hat{SS}_H = \sum_{j=1}^{m} \hat{SS}_H^{(j)}$$

$$(\hat{SS}_E)_r = \hat{d}^{\frac{1}{2}} \hat{R}_r \hat{d}^{\frac{1}{2}}$$

$$\hat{R}_r = \lambda\hat{R} + (1 - \lambda)I,\ 0 \le \lambda \le 1$$

$$\hat{R} = \hat{d}^{-\frac{1}{2}} \hat{SS}_E \hat{d}^{-\frac{1}{2}}, \text{ where } \hat{d} = \text{diag}(\hat{SS}_E)$$

- Decompose $(\hat{SS}_E)_r^{-\frac{1}{2}} \hat{SS}_H^{\text{ridge}} (\hat{SS}_E)_r^{-\frac{1}{2}}$

D

	g_1	g_2	g_3
p_1	1	1	1
p_2	1	0	0
p_3	0	1	0
p_4	1	0	1
p_5	0	0	1

Multi-label Protein Function Prediction: Experiment Data

- Baker's yeast: sequences from SGD and interactions from BioGRID
- Prediction of 17 MIPS FunCat level-1 categories
  - 4292 proteins
  - 47492 genetic interactions
  - 53393 protein-protein interactions
  - 98142 genetic and protein-protein mixed interactions
  - 500 MDS variables
- Prediction of 91 Yeast Slim Gene Ontology terms
  - 3698 proteins
  - 43893 genetic interactions
  - 32446 protein-protein interactions
  - 73512 genetic and protein-protein mixed interactions
  - 100 MDS variables

Multi-label Protein Function Prediction: Evaluation

- For each function label *g*:

		True label: g	True label: not g
Prediction label	g	True positive (TP)	False positive (FP)
	not g	False negative (FN)	True negative (TN)

$$\text{precision} = \frac{\text{TP}}{\text{TP+FP}}$$

$$\text{recall} = \frac{\text{TP}}{\text{P}}$$

$$\text{false positive rate (FPR)} = \frac{\text{FP}}{\text{N}}$$

$$\text{true positive rate (TPR)} = \frac{\text{TP}}{\text{P}}$$

$$\text{F1} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

$$\max \text{F1} = \max_{0 \le t \le 1} \{\text{F1}_t\} = \max_{0 \le t \le 1} \left\{ \frac{2 \times \text{precision}_t \times \text{recall}_t}{\text{precision}_t + \text{recall}_t} \right\}$$

- Area under curve (AUC) of ROC: (FPR*t*, TPR*t*) points, $0 \le t \le 1$
- *k*-fold cross-validation, *k*=5

$$Per = \frac{1}{k} \sum_{i=1}^{k} \left(\frac{1}{|G|} \sum_{j=1}^{|G|} FuncPrior_j Per_{ij} \right)$$

Multi-label Protein Function Prediction: Effect of MDS Variable Numbers

#MDS variables	(a) Best F1 score	(b) AUC-ROC
10	35.36±0.51	69.20±0.42
20	37.58±0.52	70.97±0.42
50	42.47±0.75	73.40±0.52
100	45.58±0.77	74.36±0.42
200	46.93±0.62	75.56±0.54
500	**48.85±0.55**	**77.02±0.38**
1000	47.73±0.60	75.57±0.46
2000	42.94±0.40	72.51±0.42

Regularized MDCA: Comparison of Distance Measures

	MIPS FunCat		SGD Yeast Slim	
Distance/similarity	(a) F1 score	(b) AUC-ROC	(c) F1 score	(d) AUC-ROC
SW	33.11±0.33	62.47±0.30	16.21±0.14	61.08±0.35
KL (by MDKC)	**42.17**[1]			
SD-GI	**38.83±0.86**	69.94±0.56	30.85±0.52	76.47±0.44
Jc-GI	38.73±0.90	**70.25±0.61**		
SD-GI-1hop	35.22±0.32	66.31±0.33		
SD-PPI	41.70±0.39	72.52±0.34	36.61±0.57	77.62±0.39
Jc-PPI	**42.38±0.53**	**72.96±0.28**		
SD-PPI-1hop	37.55±0.28	69.84±0.64		
SD-GI/PPI	45.27±0.74	75.07±0.62	39.48±0.48	81.78±0.45
Jc-GI/PPI	**45.75±0.64**	**75.22±0.63**		
TM-GI/PPI (by MDKC)	40.03[1,2]			
SW + SD-GI[3]	**43.40±0.58**	**72.85±0.59**	32.31±0.49	78.22±0.64
SW + Jc-GI[3]	42.42±0.46	72.23±0.47		
SW + SD-PPI[3]	**43.75±0.56**	**73.68±0.12**	36.98±0.56	78.52±0.34
SW + Jc-PPI[3]	42.91±0.46	73.26±0.12		
SW + SD-GI/PPI[3]	47.04±0.34	75.85±0.42	40.35±0.43	**82.33±0.54**
SW + Jc-GI/PPI[3]	45.90±0.44	74.97±0.22		
SW + SD-GI + SD-PPI[4]	**49.50±0.42**	**77.00±0.51**	**40.92±0.37**	81.28±0.49
SW + Jc-GI + Jc-PPI[4]	46.93±0.39	76.18±0.26		
KL + SD-GI/PPI[4] (by MDKC)	≤45[5]		40.11[1,6]	

Comparison of Regularized and Non-regularized MCDA

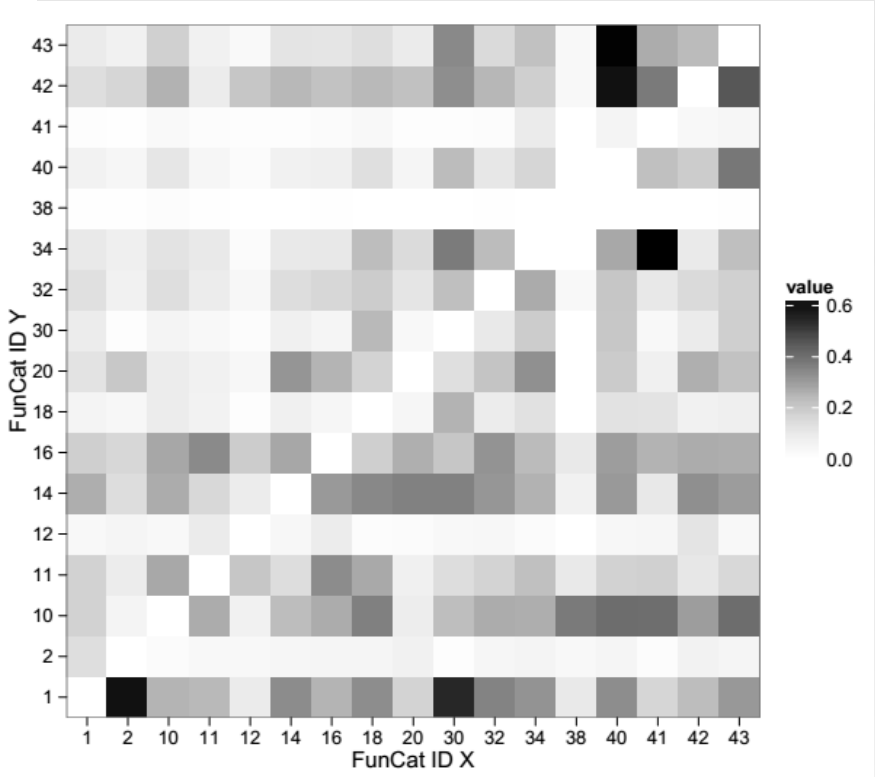
RegH	RegE	Prior	Corr	(a) Best F1 score	(b) AUC-ROC
				47.18	75.04
×				47.43	75.43
	×			49.00	76.48
×	×			**49.50**	**77.00**
×	×	×		45.87	73.76
×	×		×	37.96	67.54

RegH: Ridge regression and regularization of sum of squares for hypothesis SS_H
RegE: Regularization of sum of squares for error SS_E
Prior: Prior of function distribution
Corr: Cosine function-function correlation (Wang *et al.*, 2015)

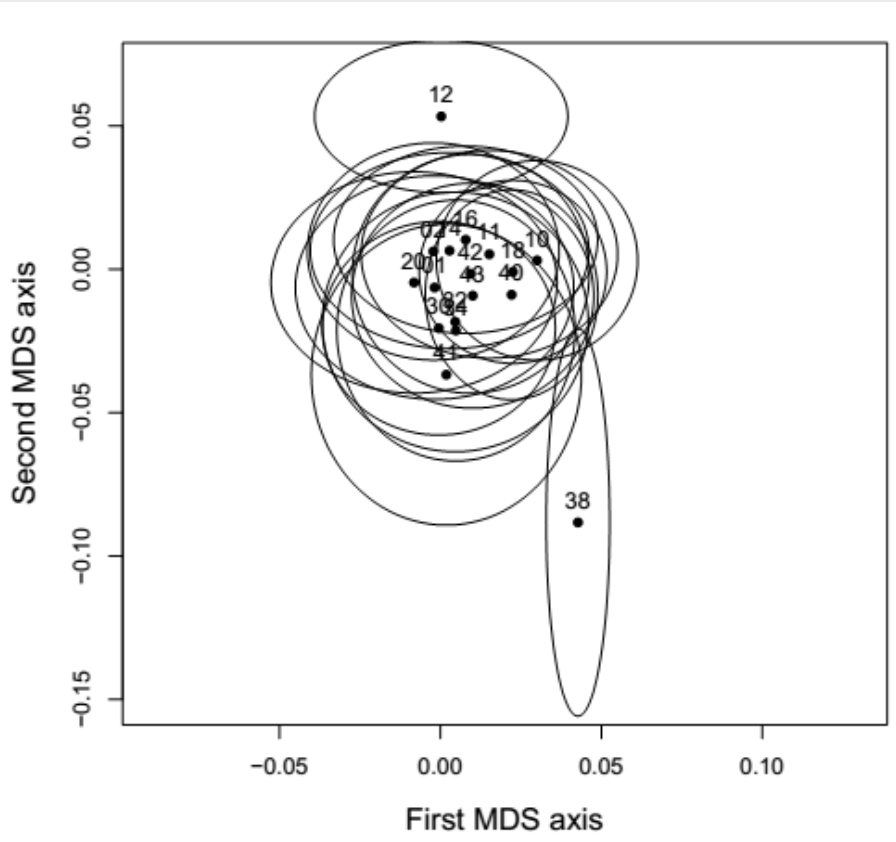
Regularized MCDA: Performance by Category

ID	Description	#proteins	(a) MCDA	(b) MDKC[1]
01	metabolism	1390	**60.14**	≤40
02	energy	327	44.53	≤45
10	cell cycle and DNA processing	965	**62.85**	≤∼60
11	transcription	995	**68.37**	≤60
12	protein synthesis	460	71.52	≤70
14	protein fate	1106	**53.33**	≤40
16	protein with binding function	1008	**44.15**	≤30
18	regulation of metabolism	240	**32.41**	≤20
20	cellular transport	974	**63.47**	≤∼60
30	cellular communication	230	51.17	≤∼50
32	cell rescue, defense and virulence	509	**38.37**	≤∼30
34	interaction with the environment	443	33.45	≤∼**45**
38	transposable elements	29	**65.59**	≤∼45
40	cell fate	264	44.08	≤∼45
41	development	66	16.14	≤**30**
42	biogenesis of cellular components	822	45.79	≤∼45
43	cell type differentiation	430	46.16	≤∼45
All		10258	**49.50**	≤45

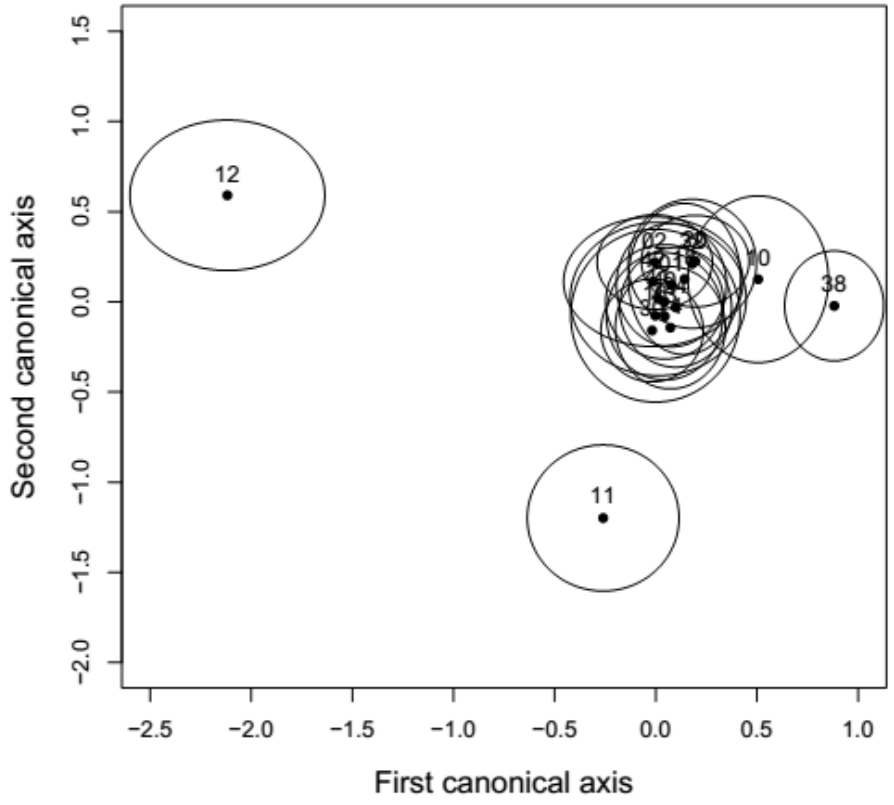
1: Performances estimated from Figure 3b in Wang *et al.*, (2015)



Regularized MCDA: Visualization of Canonical Variables

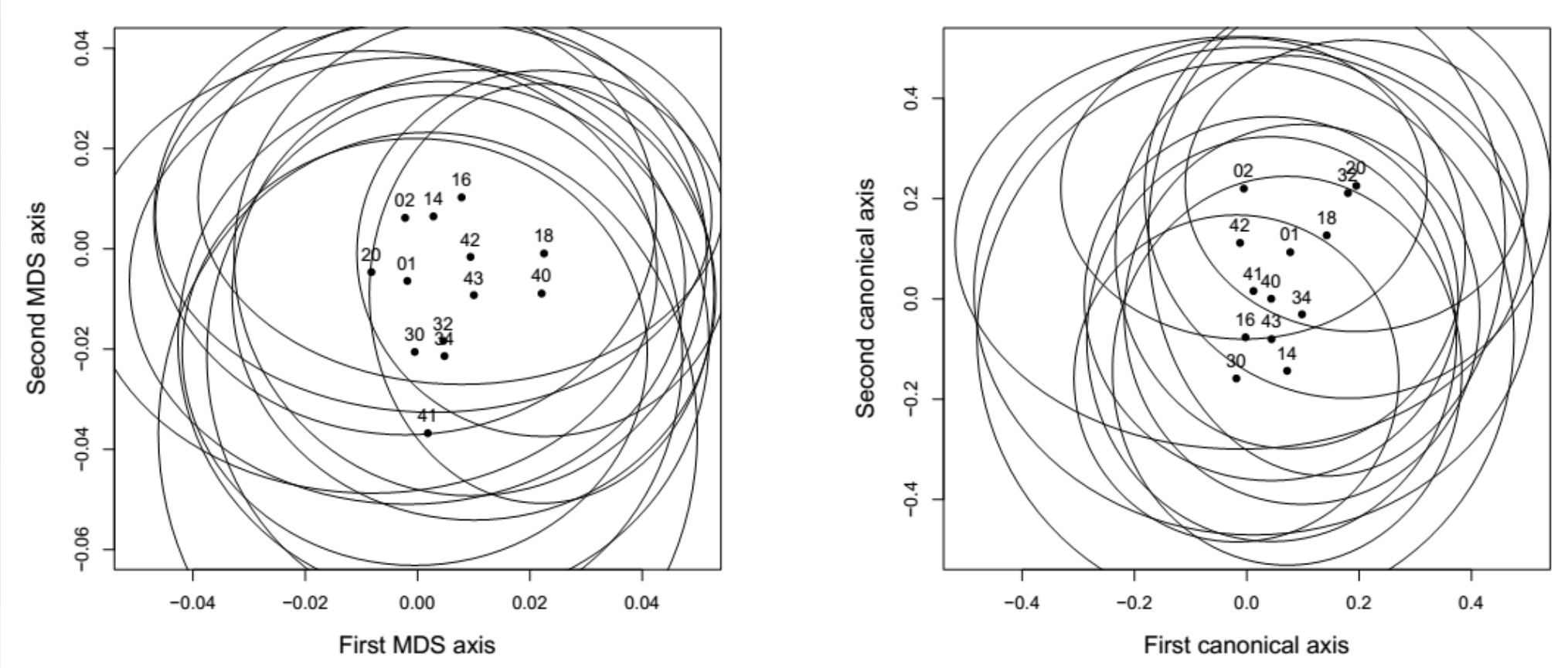


(a) FunCat groups at MDS space



(b) FunCat groups at canonical space

Regularized MCDA: Visualization of Canonical Variables



Regularized MCDA: Summary and Future Work

- Generalize the canonical discriminant analysis from multiclass prediction to multi-label prediction method that considers function correlation
- Regularization of covariance matrices of MCDA improves performance
- Smith-Waterman alignment distance + Sørensen-Dice dissimilarity + MDS + regularized MCDA: outperforms MDKC and other four algorithms
- Enables application of any traditional multi-label classification algorithm on MDS variables
- Visualize biological features of proteins and functional groups on two-dimensional plots
- Future work:
  - Problem: Sensitive to large protein overlap among functional groups
  - Extend to multiple species

Conclusions

- With the advance of high-throughput experiments, more and more biomolecular interaction data will be available. It will continue to be an important subject: the computational functional analysis of proteome based on biomolecular interactions
  - Protein network comparison, and identification of protein functional modules
    - GraphletAlign
  - Protein function prediction
    - MCDA and a multi-label function prediction framework