Sparse Learning based Linear Coherent Bi-clustering

Presenter: Yi Shi

Sept. 10, 2012

Greenland
Iceland
Finland
Sweden
Norway
Russia
United Kingdom
Poland
Germa
France
Ukraine
Kazakhstan
Mongolia
Spain
Italy
Turkey
China
South Korea
Afghanistan
Iraq
Iran
Pakistan
Algeria
Libya
Egypt
Saudi Arabia
India
Thailand
Mali
Niger
Chad
Sudan
Nigeria
Ethiopia
Kenya
DR Congo
Tanzania
Indonesia
Angola
Namibia
Botswana
Madagascar
South Africa
Canada
United States
Mexico
Venezuela
Colombia
Brazil
Peru
Bolivia
Chile
Argentina
North Pacific Ocean
North Atlantic Ocean
South Pacific Ocean
South Atlantic Ocean
Indian Ocean
A
B
Map
Traffic
2000 mi
2000 km
Imagery ©2012 NASA, TerraMetrics, Map data ©2012 MapLink, Tele Atlas

Outline

- Introduction
- Method
- Results & Discussion

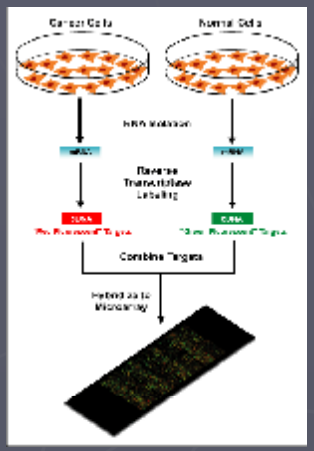
Introduction

Introduction - Gene Expression

Towards understanding of life in a system level...

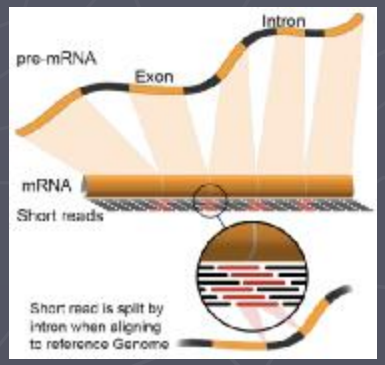
Introduction - Gene Expression

- DNA Microarray
  - Advantage: Can monitor the expression of thousands of genes simultaneously.
  - Drawback: Noise; genes are pre-assumed.



(Picture from wikipedia.org)

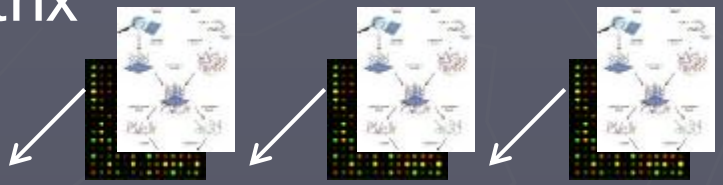
- RNA-Seq
  - Advantage: Even more high throuput (can generate gigabytes of data per run); accurate; can discover new genes.
  - Drawback: Still expensive; need a lot of softwares to assemble the primary short reads.



(Picture from wikipedia.org)

Introduction - Gene Expression

- Gene expression matrix



samples

genes

	Sample1	Sample2	Sample3	...
Gene1	1.681793	1.356604	0.784584	...
Gene2	0.907519	1.109569	1.028114	...
Gene3	1.042466	0.870551	0.946058	...
Gene4	0.806642	0.933033	1.197479	...
Gene5	0.747425	0.939523	0.823591	...
Gene6	0.926588	0.737135	0.952638	...
Gene7	1.337928	1.101905	0.882703	...
Gene8	0.913831	1.301342	1.443929	...
Gene9	0.864537	1.042466	1.117287	...
Gene10	0.784584	0.952638	0.876606	...
...	...	...	...	...

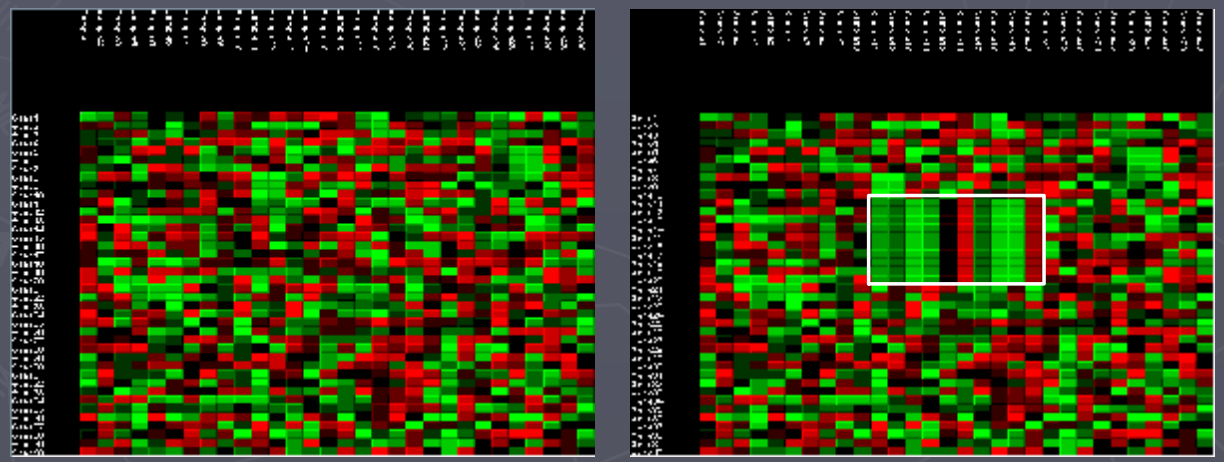
A gene expression matrix

Introduction - Gene Expression

- Gene expression applications
  - Gene expression profiling techniques provide an innovative tool for approaching the system-level understanding of biological systems.
  - It has become indispensable in numerous studies covering a broad range of biological and medical disciplines such as:
    - Hierarchical gene clustering
    - Gene selection (biomarker identification)
    - Sample class prediction
    - Genetic and regulatory network prediction
    - …

Introduction - Bi-clustering

- Why need bi-clustering algorithms?
  - § Grouping genes with similar expression behavior is an important step for downstream analysis.
  - § Genes only exhibit similar behaviors over a subset of conditions.
  - § Genes may participate in more than one functional process and therefore belong to multiple groups.



An example of bi-cluster

Introduction - Bi-clustering

- Six different bi-cluster models

x	y	z	w
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	1.0	1.0

(a) Constant value

x	y	z	w
1.2	1.2	1.2	1.2
0.8	0.8	0.8	0.8
1.5	1.5	1.5	1.5
0.6	0.6	0.6	0.6

(b) Constant row

x	y	z	w
1.2	0.8	1.5	0.6
1.2	0.8	1.5	0.6
1.2	0.8	1.5	0.6
1.2	0.8	1.5	0.6

(c) Constant column

x	y	z	w
1.2	0.8	1.5	0.6
1.0	0.6	1.3	0.4
2.0	1.6	2.3	1.4
0.7	0.3	1.2	0.3

(d) Additive coherent

x	y	z	w
2.0	4.0	8.0	1.0
1.0	2.0	4.0	0.5
4.0	8.0	16.0	2.0
1.0	2.0	4.0	0.5

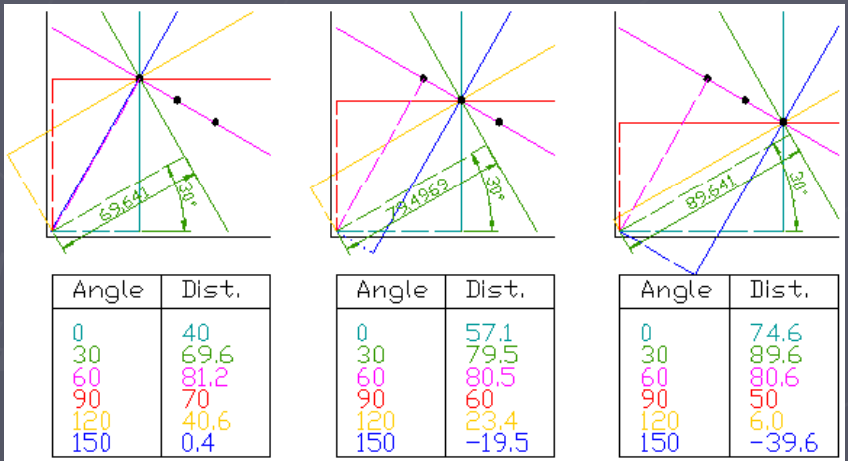
(e) Multiplicative coherent

x	y	z	w
2.0	4.0	3.0	5.0
1.5	2.5	2.0	3.0
2.3	4.3	3.3	5.3
4.5	8.5	6.5	10.5

(f) Linear coherent

Introduction - Bi-clustering

- Line detection based approach

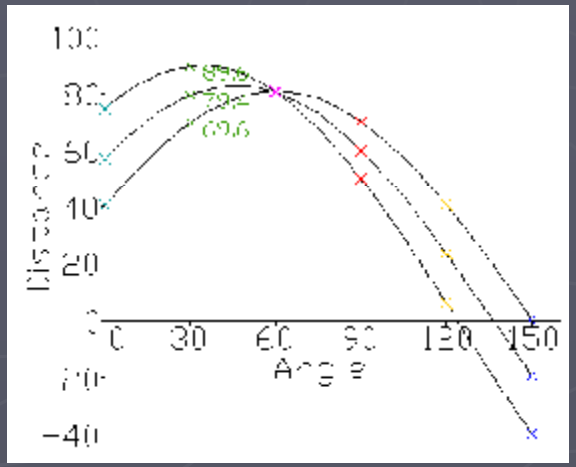


Angle	Dist.
0	40
30	69.6
60	81.2
90	70
120	40.6
150	0.4

Angle	Dist.
0	57.1
30	79.5
60	80.5
90	60
120	23.4
150	-19.5

Angle	Dist.
0	74.6
30	89.6
60	80.6
90	50
120	6.0
150	-39.6

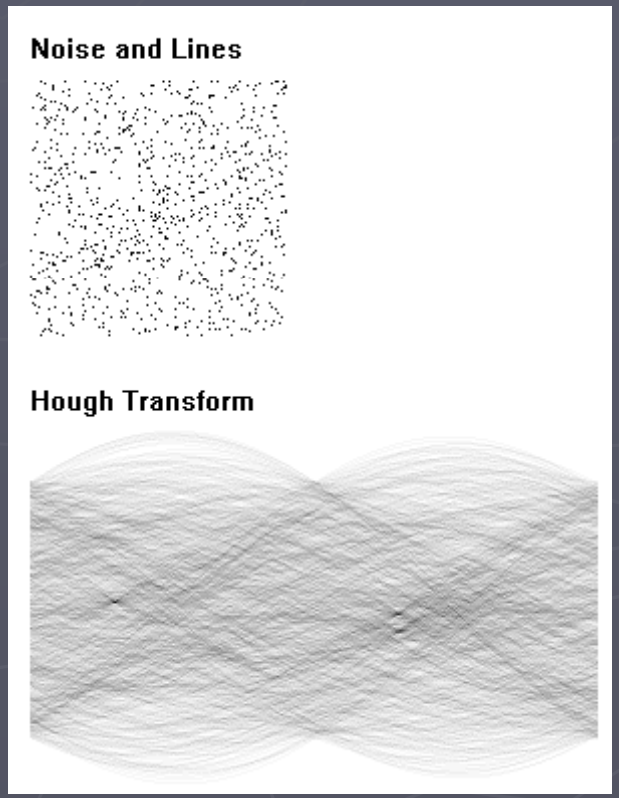
Under X-Y coordinate



Under r-theta coordinate

Introduction - Bi-clustering

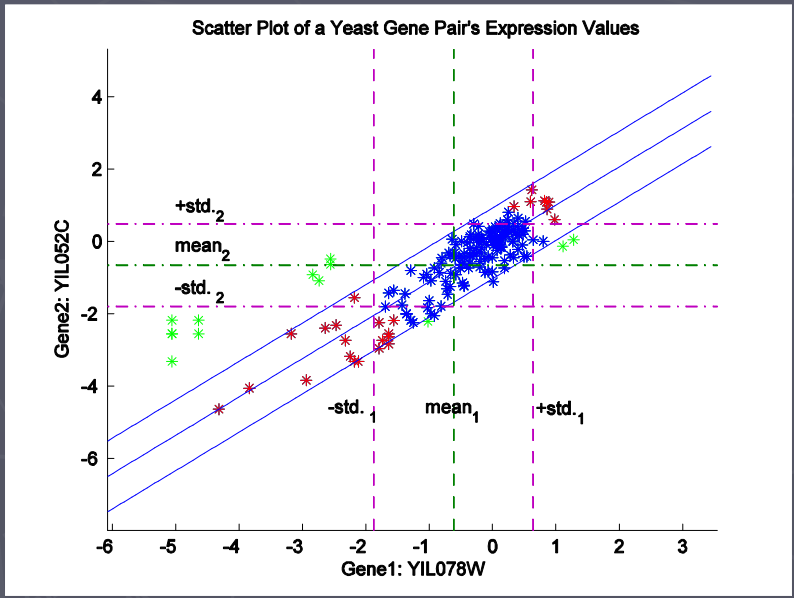
- Illustration of hough transform



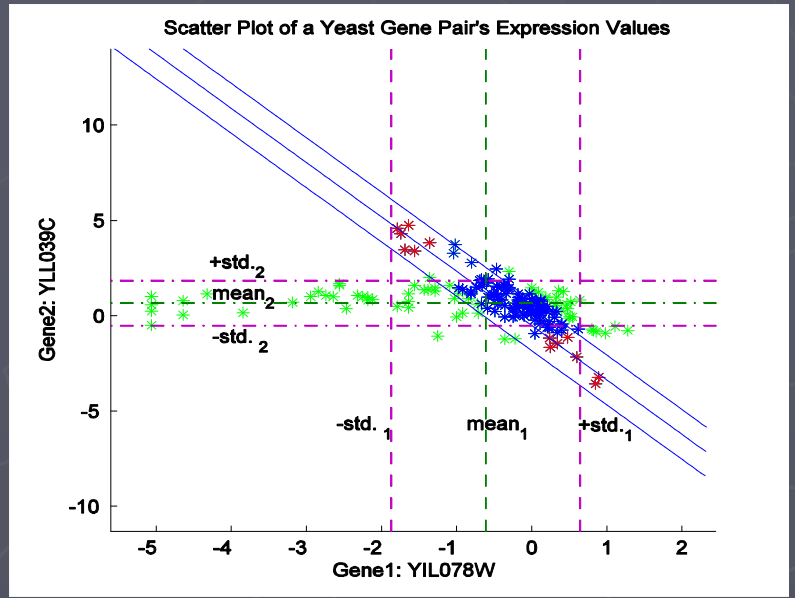
Example of Hough Transform in practical application

Introduction - Bi-clustering

- Beam detection based approach
  - Noise on data.
  - Place of the line/beam due to co-regulation.



Positive linear correlation

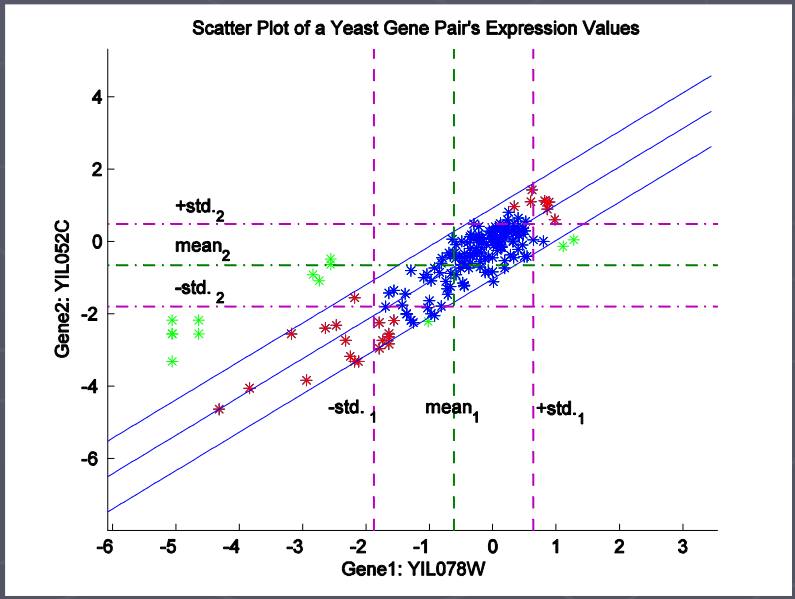


Negative linear correlation

Introduction - Bi-clustering

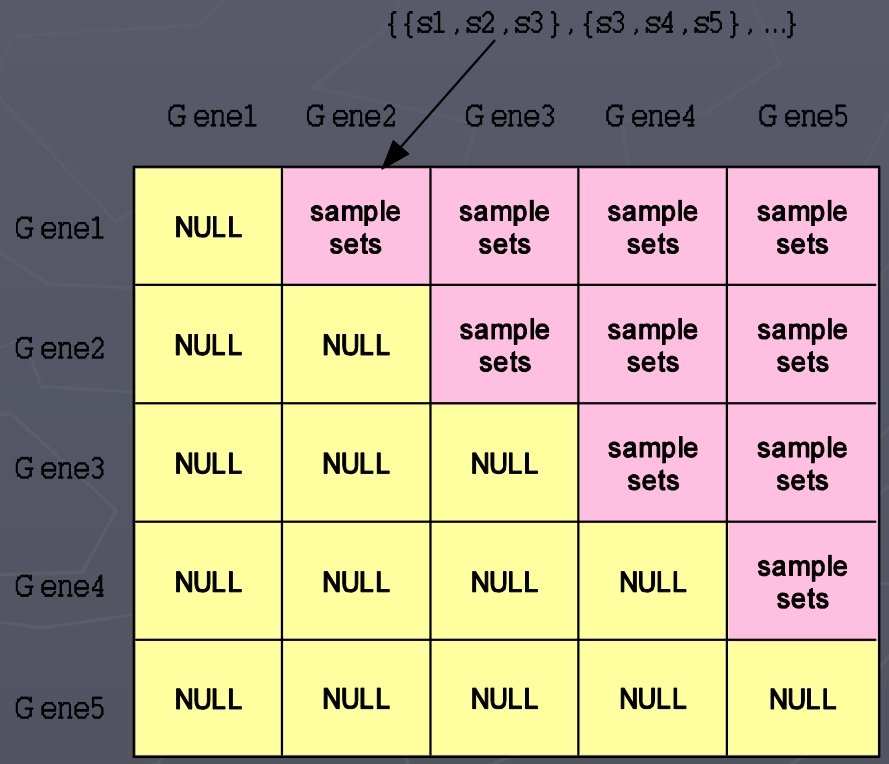
- Step 1. Beam detection
  - On a gene pair's 2D space, search the beam that optimize the following function. It awards "far" points (stronger witnesses of linear coherence) and penalize "close" points.

$$\begin{aligned} &\text{Objective:} && \max_{\theta,\beta,\gamma} \; W_{S_{\theta,\beta,\gamma}} \cdot D_{S_{\theta,\beta,\gamma}} \\ &\text{subject to} && \left| corr\left(A(p, S_{\theta,\beta,\gamma}), A(q, S_{\theta,\beta,\gamma}) \right) \right| \geq T_{cc} \end{aligned}$$



Introduction - Bi-clustering

- Step 2. Constructing bi-clusters based on sample set clustering



Introduction - Sparse Learning

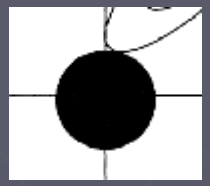
- Sparse learning methods
  - Have recently attracted a lot of interest in areas of statistics, machine learning and signal processing.
  - Many bioinformatics problems can be represented as sparse learning problems.
  - A standard sparse learning model for regression:

$$\min_{\mathbf{w}} \sum_{i} L(y_i, \mathbf{w}^T \mathbf{x_i}) + \lambda \|\mathbf{w}\|_1$$

Introduction - Sparse Learning

- Ridge regression

$$arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$

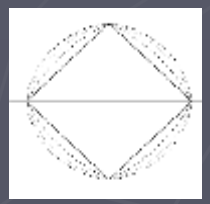


- LASSO

$$arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

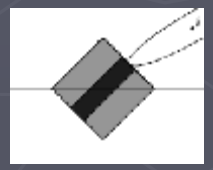
- Elastic net

$$arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda_1\|\beta\|_2^2 + \lambda_2\|\beta\|_1$$



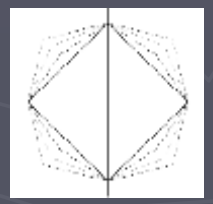
- Fused LASSO

$$arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\sum_{j=2}^{d}\|\beta_j - \beta_{j-1}\|_1$$



- OSCAR

$$arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 + \lambda_2\sum_{i<j}\max\{|\beta_i|, |\beta_j|\}$$



Method

Method

- Sparse learning bi-clustering

$$\min_{\mathbf{s_i}, a_{ij}, b_{ij}} \sum_k s_{ik}(m_{ik} - a_{ij}m_{jk} - b_{ij})^2 + \beta_1 \|\mathbf{1} - \mathbf{s_i}\|_1 \quad (1)$$
$$s.t. \quad s_{ik} \in \{0, 1\}$$

For a gene pair (m_i, m_j), there is a linear regression of the data points on their 2D space, represented by a slop a_{ij} and a intercept b_{ij}. On a selected set of features (with s_{ik} = 0 means not selected and 1 means selected), they will exhibit linear coherency. Model (1) will do the feature selection by minimizing the sum of squared residuals. To avoid trivial solution, i.e., selecting zero feature, we add the L_1 norm regularizer.

Method

- Sparse learning bi-clustering

  Now, we not only select features, but also select data points (genes), so (1) becomes (2):

$$\min_{W,S,A,B} \sum_{i,j} w_{ij} \sum_{k} s_{ik}(m_{ik} - a_{ij}m_{jk} - b_{ij})^2 + \beta_1 \|\mathbf{1} \cdot \mathbf{1}^T - S\|_{1,1} + \beta_2 \|\mathbf{1} \cdot \mathbf{1}^T - W\|_{1,1} \quad (2)$$

$$s.t. \quad w_{ij} \in \{0, 1\}, s_{ik} \in \{0, 1\}$$

  After relaxing **w** and **s**, and adding vector **d** that favor far end features, (2) becomes (3):

$$\min_{W,S,A,B} \sum_{i,j} w_{ij} \sum_{k} s_{ik} \frac{1}{d_{ijk}} (m_{ik} - a_{ij}m_{jk} - b_{ij})^2 + \beta_1 \|\mathbf{1} \cdot \mathbf{1}^T - S\|_{1,1} + \beta_2 \|\mathbf{1} \cdot \mathbf{1}^T - W\|_{1,1} \quad (3)$$

$$s.t. \quad w_{ij} \in [0, 1], s_{ik} \in [0, 1]$$

Method

- Sparse learning bi-clustering

  Re-writing (3) in matrix operation form, we get:

$$\min_{W,S,A,B} \sum_i \left\| \triangle(\mathbf{w_{i:}})^{1/2}(\mathbf{1} \otimes \mathbf{m_{i:}} - \triangle(\mathbf{a_{i:}})M - \triangle(\mathbf{b_{j:}})\mathbf{1} \otimes \mathbf{1}^T) \div D_i^* \triangle(\mathbf{s_{i:}})^{1/2} \right\|_F^2$$
$$+\beta_1 \|\mathbf{1} \cdot \mathbf{1}^T - S\|_{1,1} + \beta_2 \|\mathbf{1} \cdot \mathbf{1}^T - W\|_{1,1}$$
$$s.t. \quad w_{ij} \in [0,1], s_{ik} \in [0,1] \tag{4}$$

Method

- SLLB in detail

Algorithm 1 The SLLB Algorithm

Input: M, β_1, β_2, ϵ.
Output: a set of linear coherent bi-clusters.

$A = \mathbf{1} \cdot 1^T$
$b_{ij} = \bar{m}_{i:} - a_{ij}\bar{m}_{j:}$ for $i, j \in [1, 2, ...n]$.
$s_{ik} = 1$ if $d'_{ijk} <= e^{d''_{ijk}} - 1$, $s_{ik} = 0$ otherwise, for $i \in [1, 2, ...n]$, $k \in [1, 2, ...p]$.
while $\Delta L > \epsilon$ **do**
  for each i, j **do**
    if $\sum_k \frac{s_{ik}}{d_{ijk}}(m_{ik} - a_{ij}m_{jk} - b_{ij})^2 < \beta_2$ **then**
      $w_{ij} = 1$
    else
      $w_{ij} = 0$
    end if
  end for
  for each i, j **do**
    if $\sum_j \frac{w_{ij}}{d_{ijk}}(m_{ik} - a_{ij}m_{jk} - b_{ij})^2 < \beta_1$ **then**
      $s_{ik} = 1$
    else
      $s_{ik} = 0$
    end if
  end for
  for each i, j **do**
    $(a_{ij}, b_{ij})^T = (X_{ij}^T \triangle(\mathbf{s_{i:}} \bullet \mathbf{d_{ij:}})X_{ij})^{-1} X_{ij}^T \triangle(\mathbf{s_{i:}} \bullet \mathbf{d_{ij:}})\mathbf{y_{ij}}$
  end for
  Calculate loss change ΔL.
end while
Construct bi-clusters from W, S and remove redundant bi-clusters.

Results & Discussion

- Results evaluations:
  - On synthetic datasets
    - The Prelic's match score function: Let M1, M2 be two sets of bi-clusters. The gene match score of M1 with respect to M2 is given by the function.

$$S^*_G(M1, M2) = \frac{1}{M_1} \sum_{(G_1, C_1) \in M_1} max_{(G_2, C_2) \in M_2} \frac{|G_1 \cap G_2|}{|G_1 \cup G2|}$$

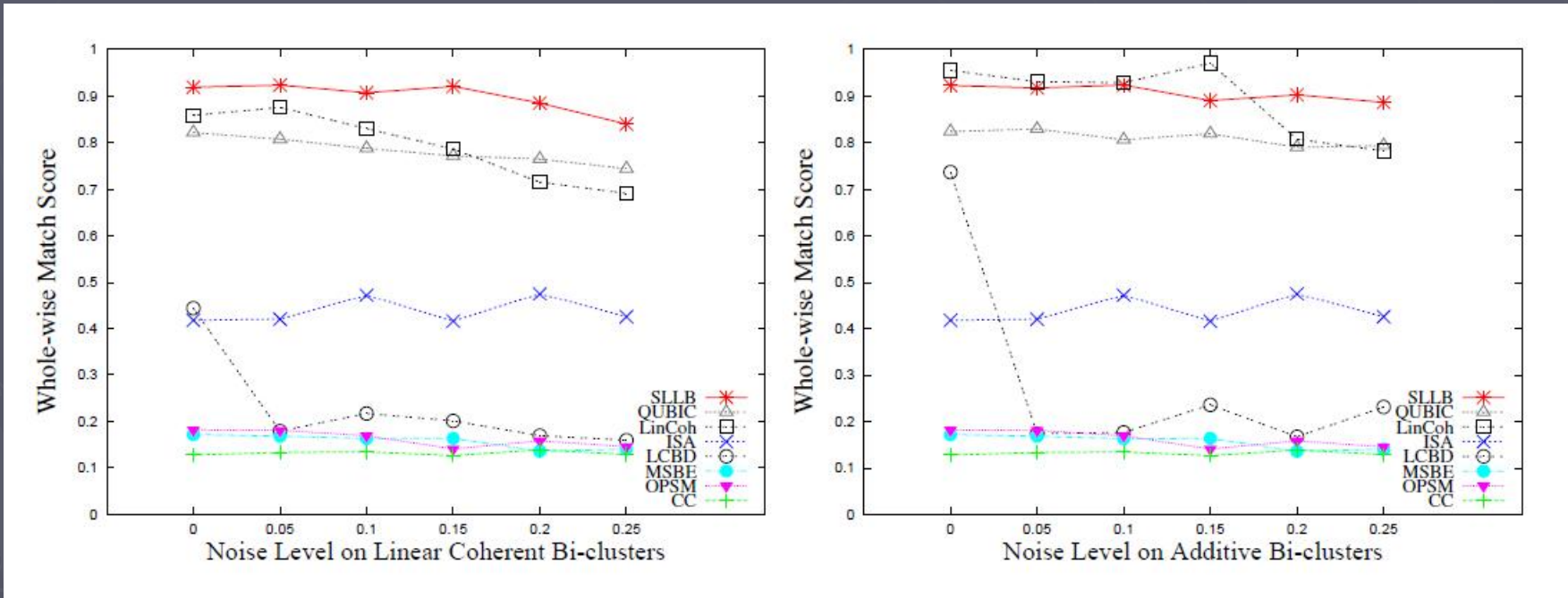
    - Overall Match Score

$$\text{score}(\mathcal{C}, \mathcal{C}^*) = \sqrt{\text{score}_G(\mathcal{C}, \mathcal{C}^*) \times \text{score}_S(\mathcal{C}, \mathcal{C}^*)}.$$

    - Gene discovery rate
  - On real datasets
    - Functional enrichment score.

Results & Discussion

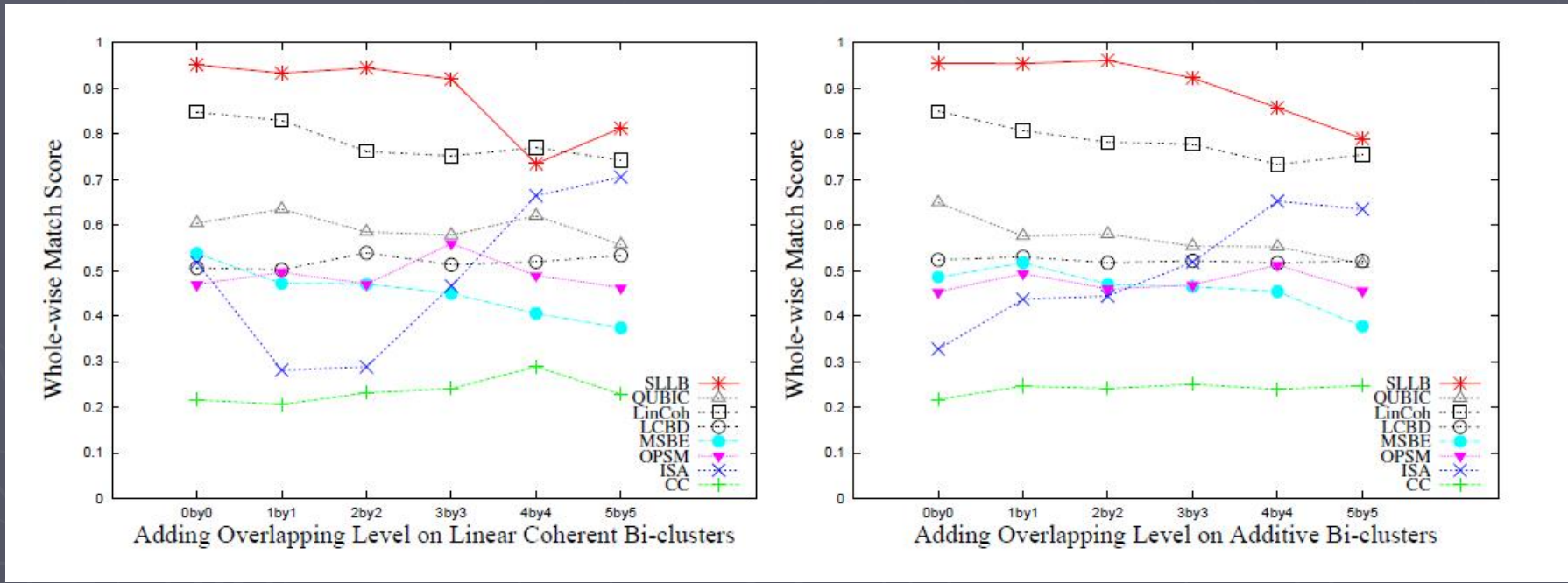
- Results on synthetic dataset



The whole match scores of the eight algorithms on recovering linear coherent bi-clusters and additive bi-clusters at six different noise levels.

Results & Discussion

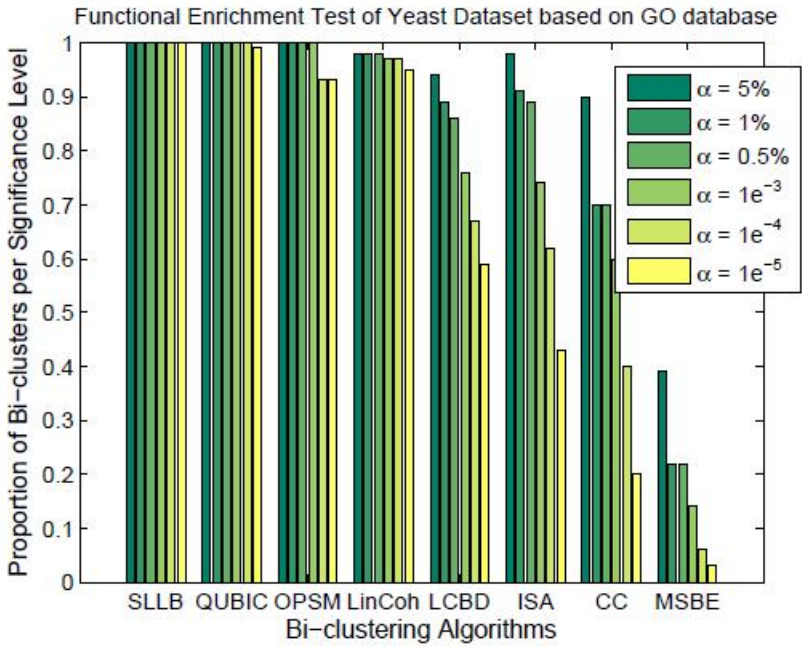
- Results on synthetic dataset

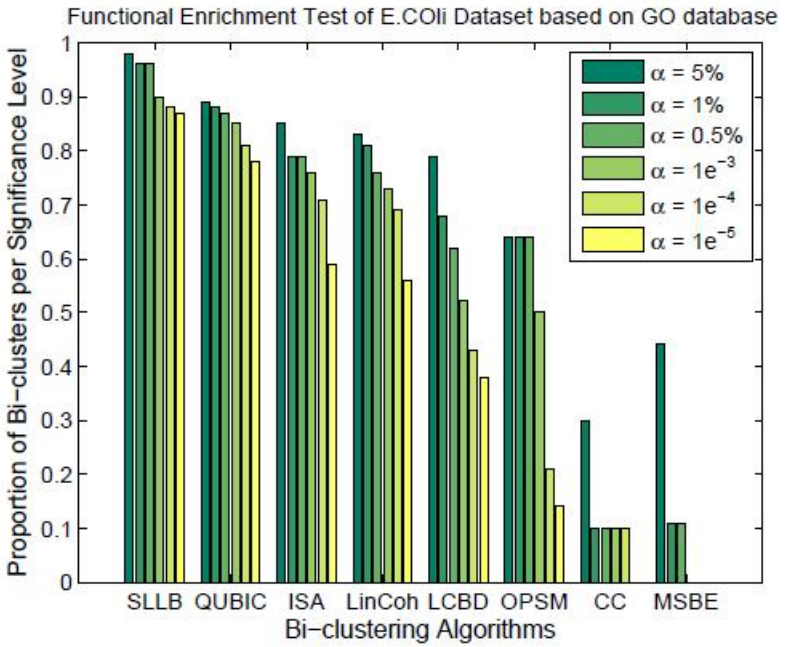


The whole match scores of the eight algorithms for recovering the overlapping linear coherent and additive bi-clusters, under the adding overlap model.

Results & Discussion

- Results on real dataset

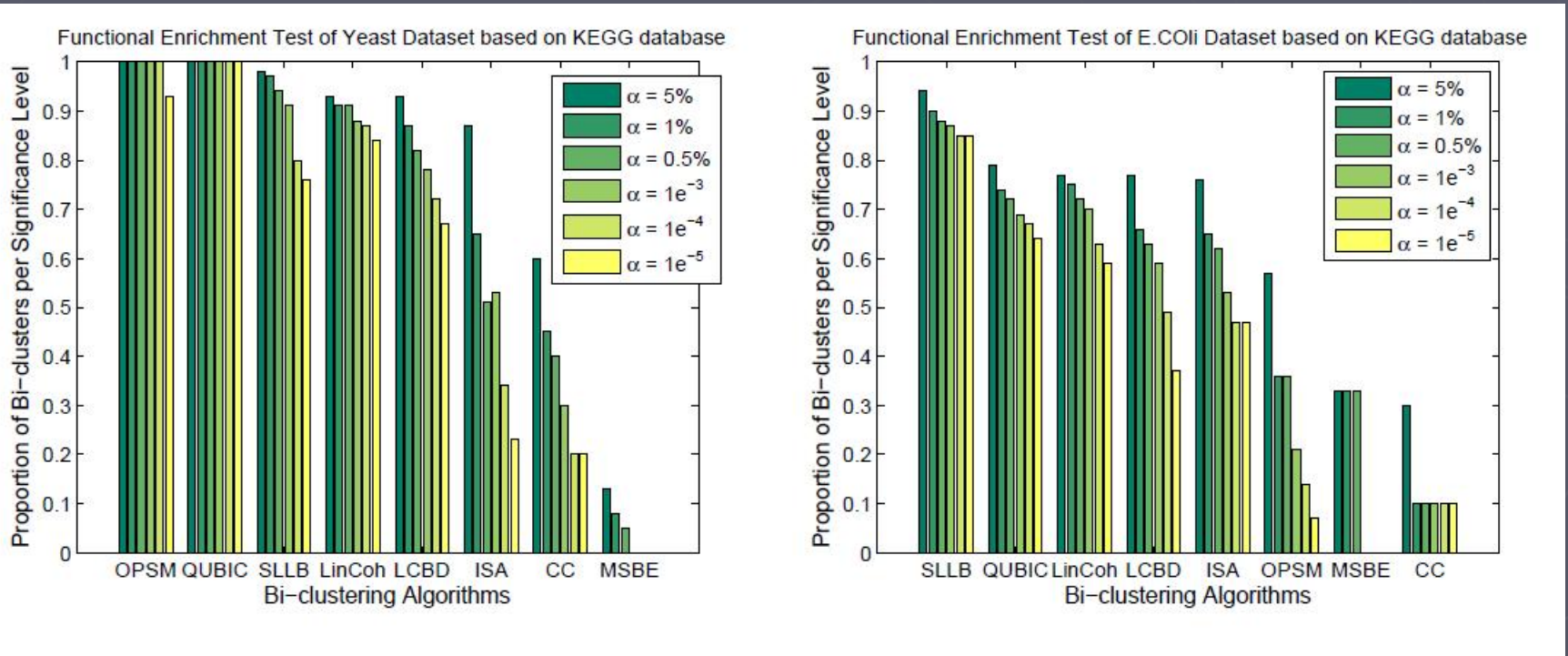




Portions of discovered bi-clusters by the eight algorithms on the two real datasets that are significantly enriched in the GO biological process, using six different *P-value cutoffs.*

Results & Discussion

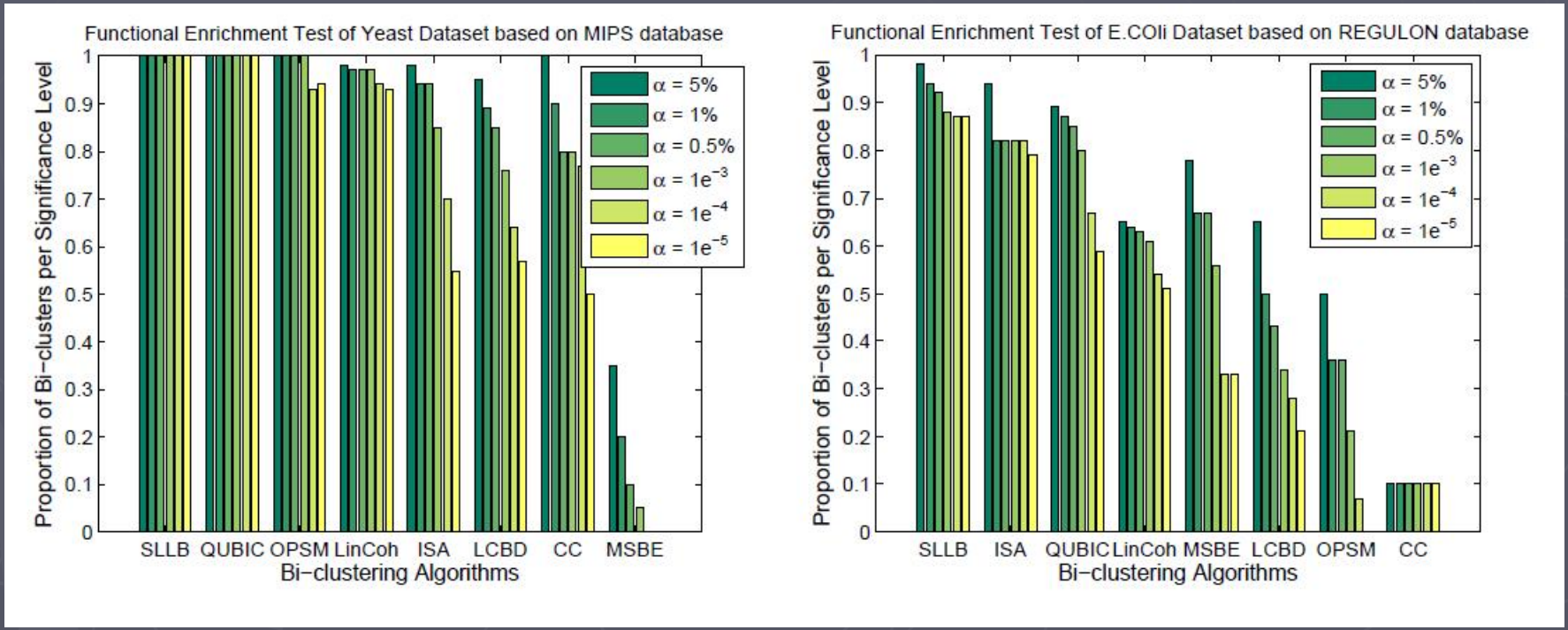
- Results on real dataset



Portions of discovered bi-clusters by the eight algorithms on the two real datasets that are significantly enriched in the KEGG pathway, using six different *P-value cutoffs.*

Results & Discussion

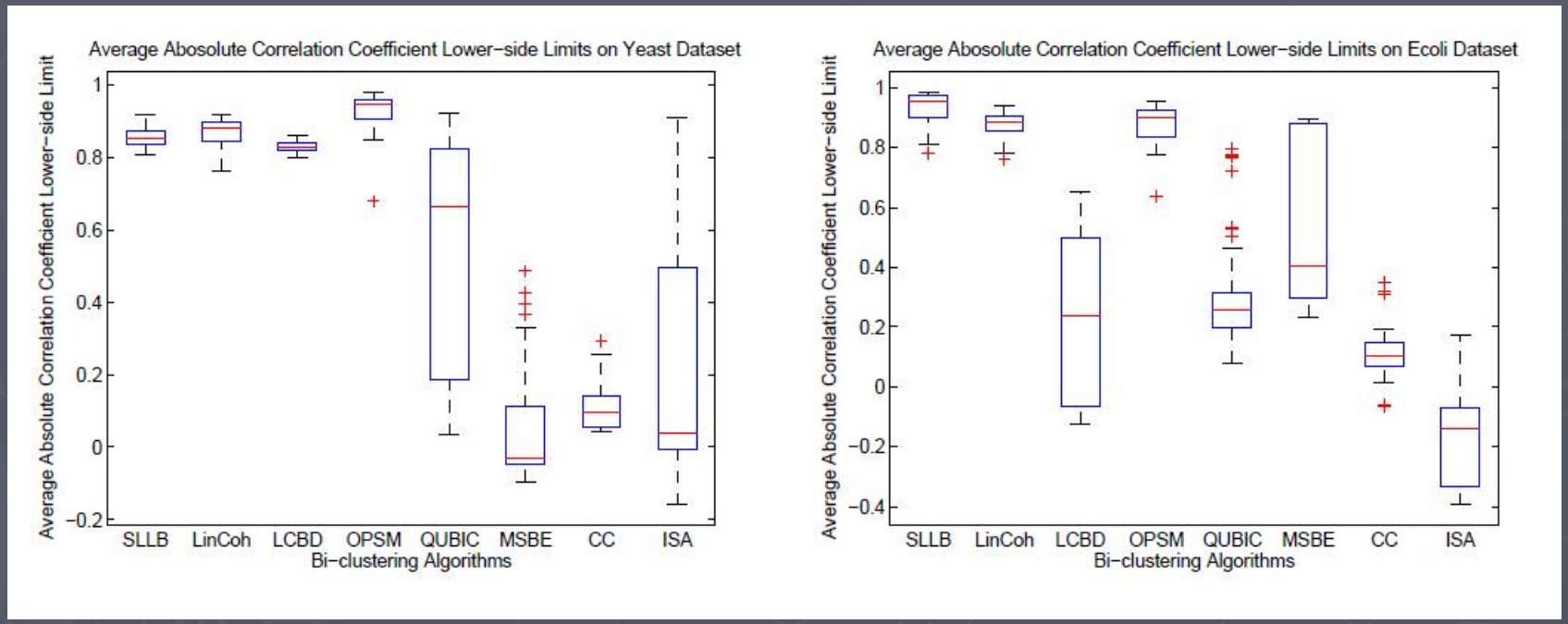
- Results on real dataset



Portions of discovered bi-clusters by the eight algorithms on the two real datasets that are significantly enriched in the MIPS pathway experimentally verified REGULONS, respectively, using six different *P-value cutoffs.*

Results & Discussion

- Results on real dataset



Box plots of the average absolute correlation coefficients obtained by the eight bi-clustering algorithms on Yeast and E.coli datasets, respectively

Results & Discussion

- Statistics

	#Bi-clusters	$\mu_{\|gene\|}$	$\sigma_{\|gene\|}$	$\mu_{\|sample\|}$	$\sigma_{\|sample\|}$	#Terms enriched (GO, KEGG, MIPS/regulons)
Yeast:						
SLLB	101	83.40	36.40	85.56	25.98	3, 7, 5
QUBIC	100	119.08	59.69	23.21	6.21	3, 2, 3
LinCoh	100	61.84	38.43	133.09	18.09	5, 7, 5
LCBD	132	46.46	17.53	13.35	4.58	10, 6, 11
ISA	47	67	34.54	8.4	1.78	15, 13, 18
OPSM	14	423.29	728.95	9.07	5.14	1, 1, 2
MSBE	40	19.25	8.32	18.68	8.22	8, 4, 6
CC	10	297.7	304.18	60.8	23.46	6, 4, 8
E.Coli:						
SLLB	52	43.79	16.23	106.58	51.70	8, 11, 13
QUBIC	100	73.91	33.45	33.51	14.51	14, 8, 15
LinCoh	100	9.63	7.66	141.43	34.04	24, 24, 22
LCBD	155	485.05	366.63	15.37	22.95	23, 22, 33
ISA	34	124.21	42.18	13.88	6.11	11, 10, 13
OPSM	14	419.29	744.35	8.93	4.8	8, 4, 5
MSBE	9	82.67	18.1	80.22	19.18	1, 3, 4
CC	10	309.9	950.15	31.4	81.74	2, 2, 2

Statistics of different algorithms' bi-clustering results

Results & Discussion

- Future works
  - Better parameter training approach.
  - Investigate other types of bi-clusters.
  - Applying the method in other bi-clustering problems, e.g. document bi-clustering and image bi-clustering.

Acknowledgement

Prof. Dale Schuurmans

Prof. Guohui Lin

Dr. Xinhua Zhang

Dr. Xiaoping Liao

ALBERTA INGENUITY CENTRE FOR MACHINE LEARNING

Thank You!

Questions?