Comparative Analysis of Biclustering Algorithms

Doruk Bozdag[1], Ashwin S Kumar[1], Umit V Catalyurek[1,2]

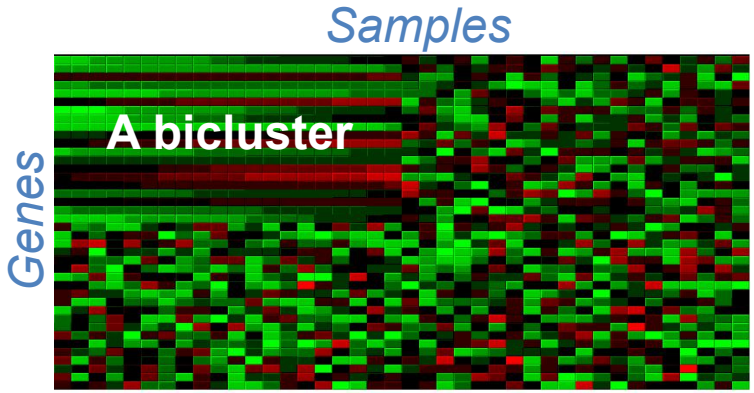
[1] Department of Biomedical Informatics

[2] Department of Electrical and Computer Engineering

The Ohio State University

Introduction

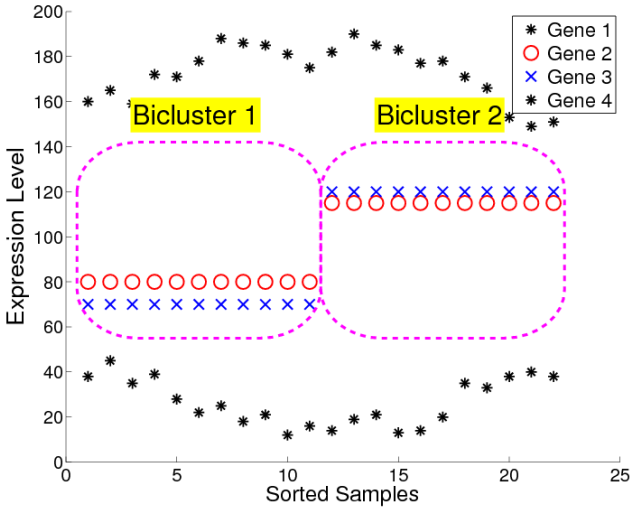
- **Objective:** Analysis of biclustering algorithms that use microarray data sets for identifying functionally related genes.
- Several approaches to identify genes that have related expression levels
  - Related expression => Related biological functions
- **Clustering**: gene behavior across all samples
  - Drawback: Functionally related genes may not exhibit similar pattern in all samples
- **Biclustering**: gene behavior across a subset of samples
  - Introduced by Cheng and Church (2000)



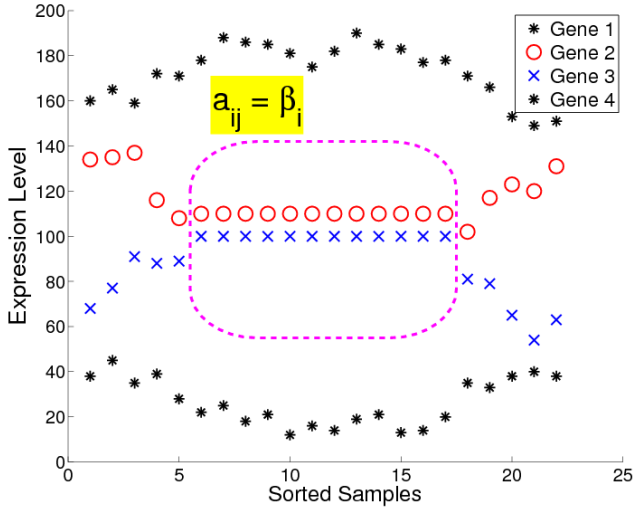
Basis of Comparison

- Comparing biclustering algorithms is very challenging
  - Numerous algorithms with different objectives and search strategies.
- Identified three aspects of algorithms and corresponding methods to evaluate these aspects independently.
  - **Bicluster patterns sought**
    - Patterns that optimize the objective function of an algorithm
    - Theoretical analysis
  - **Search technique**
    - Success of algorithms in finding the patterns that they target
    - Experimental analysis on synthetic data sets
  - **Biological relevance**
    - Biological significance of identified biclusters
    - Experimental analysis on real data sets

- Bicluster patterns can be classified into two:
  - Global: Defined on multiple biclusters. Membership of a row/column depends on external elements and other clusters
  - **Local**: Defined on single clusters. No information required about elements outside the bicluster.

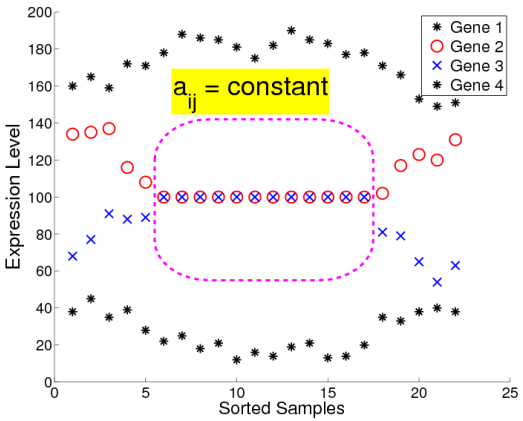


Global pattern example

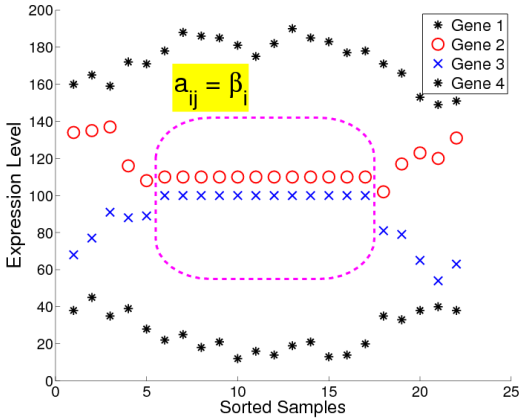


Local pattern example

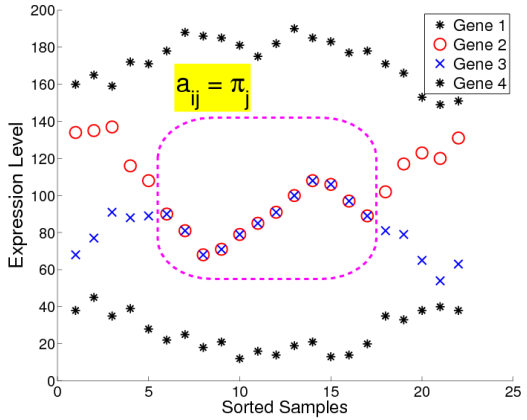
Well-known Local Patterns



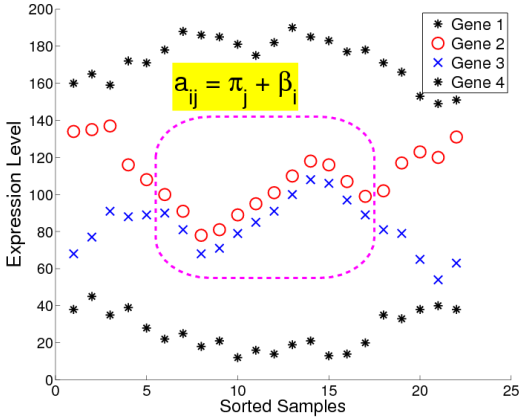
Constant bicluster



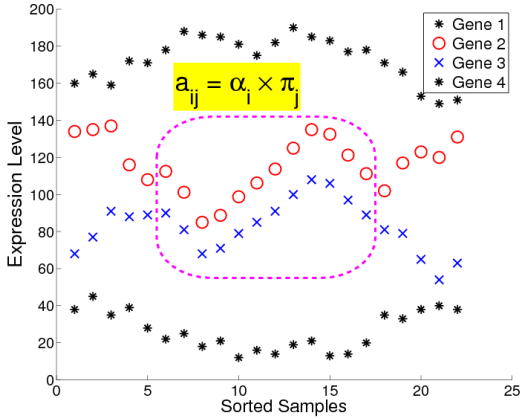
Constant rows



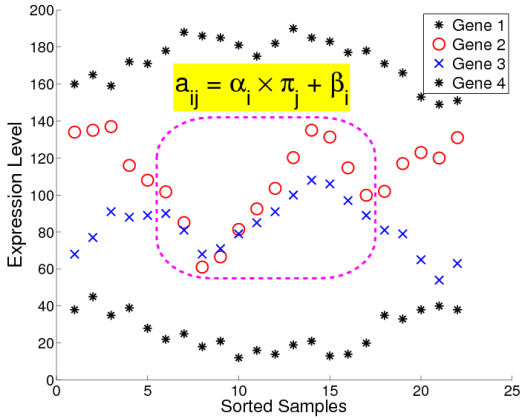
Constant columns



Shifting



Scaling



Shift-scale

Analyzing patterns sought

1. Assume that the bicluster has a shift-scale pattern (the most general local pattern)
2. Plug in $a_{ij} = \alpha_i \times \pi_j + \beta_i$ the objective function
3. Find constraints on α_i , π_j and β_i to optimize obj. function.
4. Lookup for the constraints below to find patterns.

α_i	π_j	β_i	Simplified $\hat{a}_{ij}$ $(i \in I, j \in J)$	Bicluster pattern
0 any constant constant	any 0 constant constant	constant constant 0 constant	a constant	Constant bicluster
0 any varying any	any 0 constant constant	varying varying any varying	β_i	Constant rows
constant constant	varying varying	0 constant	π_j	Constant columns
constant	varying	varying	$\pi_j + \beta_i$	Shifting
varying	varying	0	$\alpha_i \times \pi_j$	Scaling
varying	varying	varying	$\alpha_i \times \pi_j + \beta_i$	Shift-scale

Cheng and Church (CC)

- **Objective:** Minimize $$MSR = \frac{1}{|I||J|} \sum_{j \in J, i \in I} \epsilon_{ij}^2$$

  where $$\epsilon_{ij} = a_{ij} - a_{iJ} - a_{Ij} + a_{IJ}$$

- **Criteria for perfect biclusters:** $$(\alpha_i - \overline{\alpha})(\pi_j - \overline{\pi}) = 0$$

α_i	π_j	β_i	Bicluster pattern
0	any	any	constant bicluster, constant rows
constant	any	any	constant bicluster, constant columns, shifting
any	0	any	constant bicluster, constant rows
any	constant	any	constant bicluster, constant rows

  - Not optimized for detecting scaling and shift-scale patterns

- **Objective:** Maximize relevance indices $R_{Ij} = 1 - \frac{\sigma_{Ij}^2}{\sigma_{\cdot j}^2}$

where $\sigma_{Ij} = \frac{1}{|I|} \sum_{i \in I} (a_{ij} - a_{Ij})^2$

- **Criteria for perfect biclusters:** $(\alpha_i - \overline{\alpha}) \times \pi_j - (\beta_i - \overline{\beta}) = 0$

α_i	π_j	β_i	Bicluster pattern
0	any	0	none
0	any	constant	constant bicluster
constant	any	0	constant bicluster, constant columns
constant	any	constant	constant bicluster, constant columns
any	0	0	none
any	0	constant	constant bicluster

  - Not optimized for detecting constant rows, shifting, scaling and shift-scale patterns

Correlated Pattern Biclusters (CPB)

- **Objective:** PCC between every pair of rows in the bicluster should be greater than a threshold, with respect to the columns in the bicluster.

$$PCC = \frac{\sum_{j \in J}(a_{ij} - a_{iJ})(a_{\ell j} - a_{\ell J})}{\sqrt{\sum_{j \in J}(a_{ij} - a_{iJ})^2 \sum_{j \in J}(a_{\ell j} - a_{\ell J})^2}}$$

- **Criteria for perfect biclusters:** $\alpha_i^2 \alpha_\ell^2 \left(\sum_{j \in J} (\pi_j - \overline{\pi})^2 \right)^2$ is non-zero
  - PCC = 1 if the denominator is non-zero

α_i	π_j	β_i	Bicluster pattern
constant	varying	any	constant columns, shifting
varying	varying	any	scaling, shift-scale

  - Cannot capture constant biclusters and constant rows
  - Biclusters with shift-scale patterns have perfect correlation between any pair of rows with respect to columns in the bicluster

Order Preserving Submatrix (OPSM)

- **Objective:** Find a set of columns s.t. the order of the columns is the same in all rows: $a_{ij} < a_{ik}$ **iff** $a_{\ell j} < a_{\ell k}$

- **Criteria for perfect biclusters:** $\alpha_i(\pi_j - \pi_k) < 0$ **iff** $\alpha_\ell(\pi_j - \pi_k) < 0$
  - Potentially identifies the same type of biclusters as CPB.
  - Distribution of PCC values between rows in an OPSM is the same as distribution of PCC values between pairs of random vectors that have the same column ordering.
    - *Smallest* PCC between 20 pairs of random vectors with the same ordering in 20 (60) columns is 0.83 (0.96).
    - **Same column ordering => high PCC values**

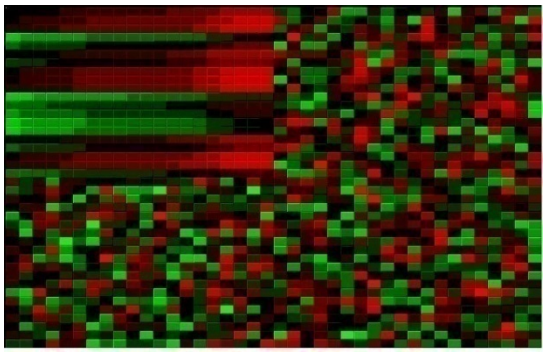
Experiments: Algorithms Considered

- Algorithms that seek local patterns
  - **CC** – threshold MSR = 0.01, 100 runs
  - **HARP** – no implementation available
  - **CPB** – threshold PCC = 0.9, 100 runs
  - **OPSM** – number of partial models = 100

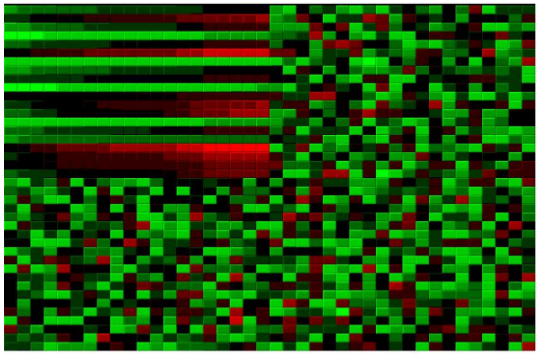
- Algorithms that seek global patterns
  - **SAMBA** – biclusters with large variance
  - **MSSRCC** – biclusters with small combined MSR, 100 runs

Synthetic Dataset Generation

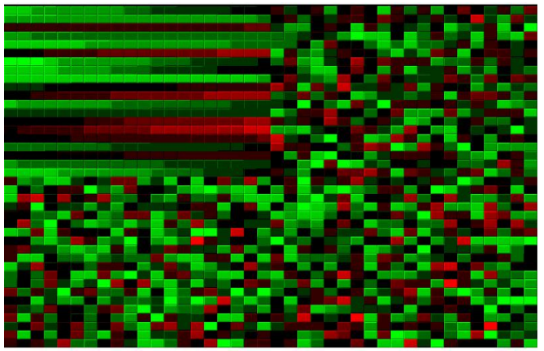
1. Generate a 1000x120 matrix filled with random values [0 1].
2. Generate an NxN bicluster (where N is 20, 40 or 60) with perfect:
   - Shift pattern, or
   - Shift-scale pattern, or
   - Order preserving pattern.
3. Implant the bicluster into the matrix and shuffle rows & columns



Shift



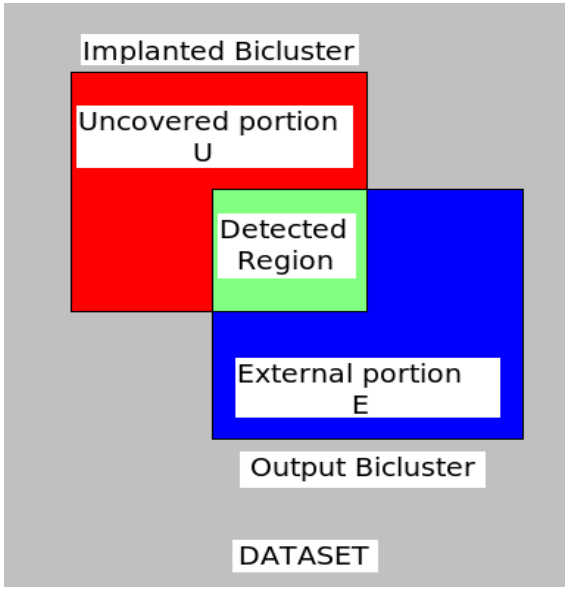
Shift-scale



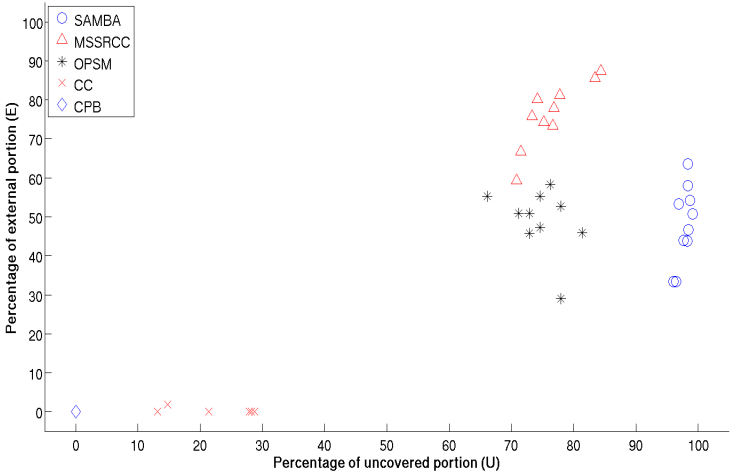
Order-preserving

Evaluating the search strategies

- Compare each bicluster returned by an algorithm against the implanted bicluster.
- The smaller are the uncovered portion (U) and extarnal portion (E), the better.

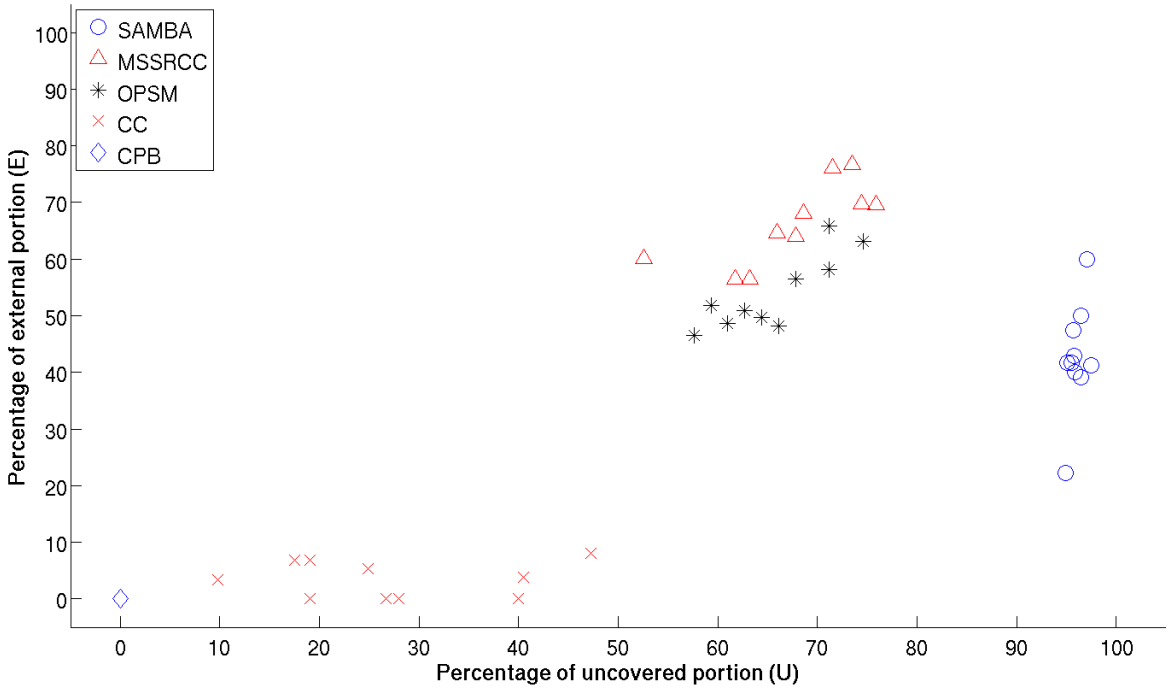


- In the result charts, U and E are given on the x-axis and y-axis, respectively.
  - Each point represents the best bicluster found in a dataset
  - Total of 10 points (datasets) per algorithm



Effect of Bicluster Pattern

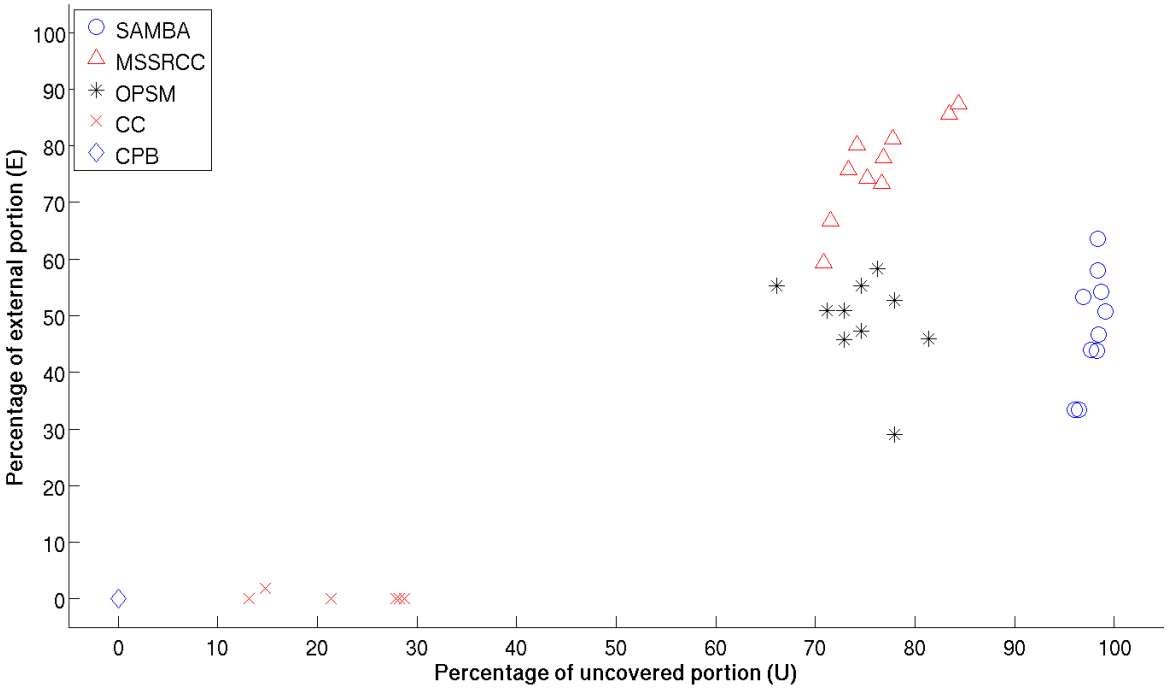
- Implanted a 60x60 bicluster with **shift pattern**



- CPB and CC are the best to detect shift patterns.
- U > 0 for CC, but E < 10%

Effect of Bicluster Pattern

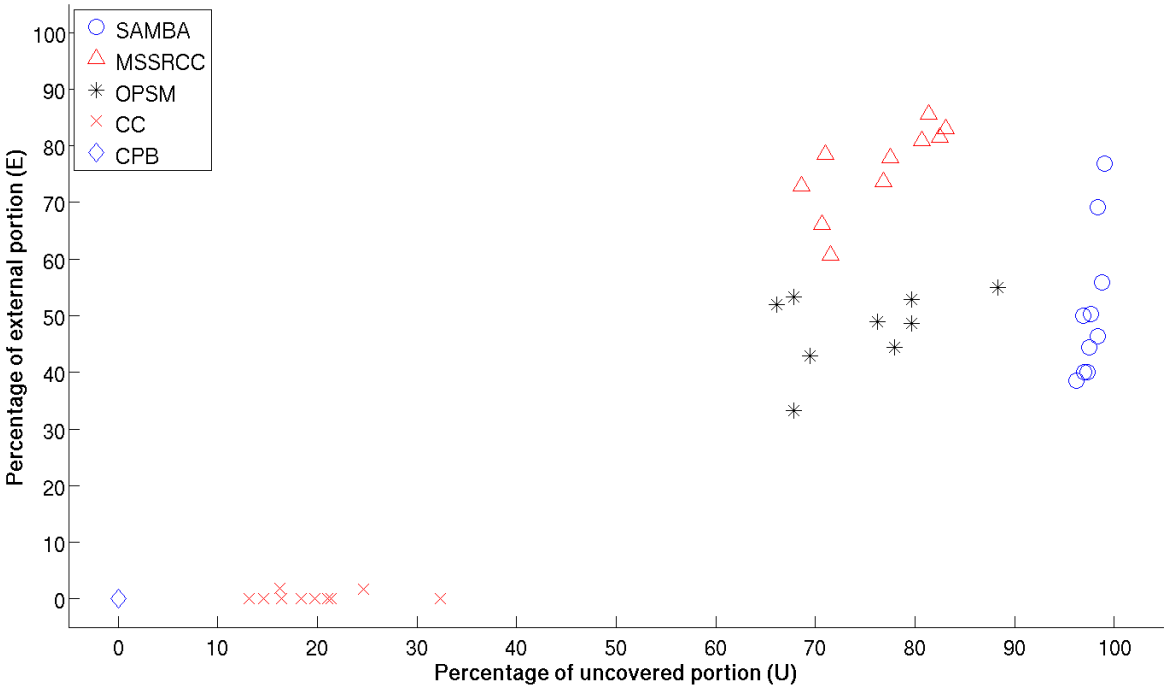
- Implanted a 60x60 bicluster with **shift-scale pattern**



- CPB and CC are again the best to detect shift-scale patterns.
- Other algorithms perform slightly worse compared to shift pattern

Effect of Bicluster Pattern

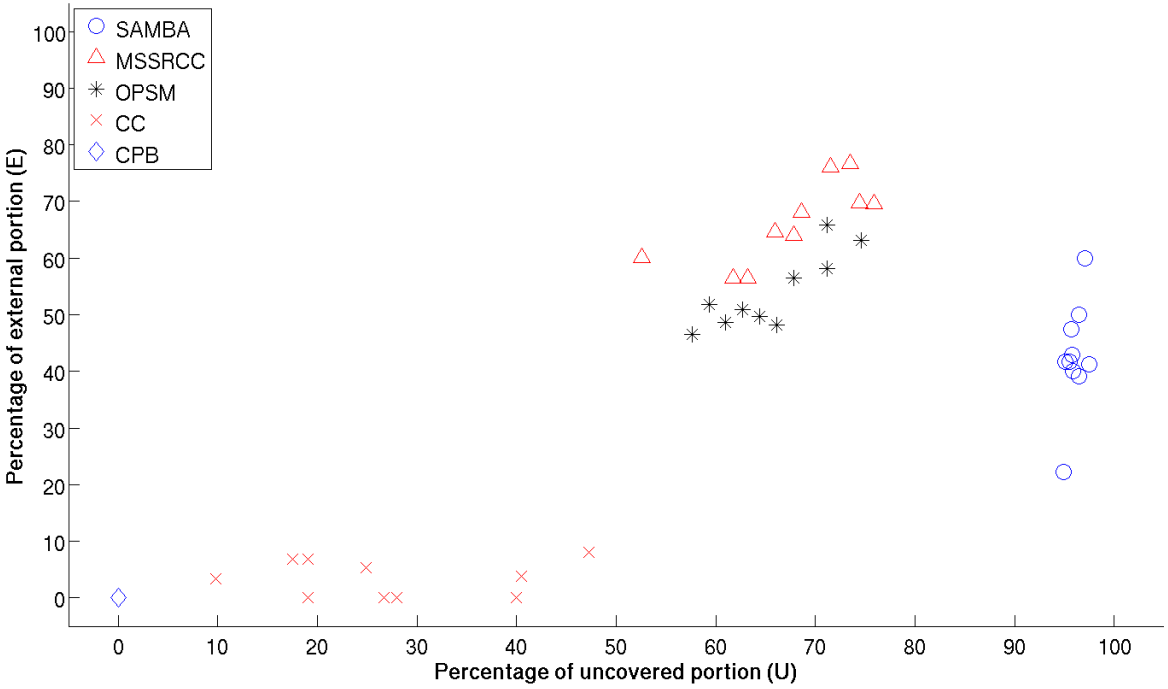
- Implanted a 60x60 bicluster with **order-preserving pattern**



- Results are similar to shift-scale, due to high PCC between rows
- The best cluster found by OPSM is slightly better than shift-scale

Effect of Bicluster Size

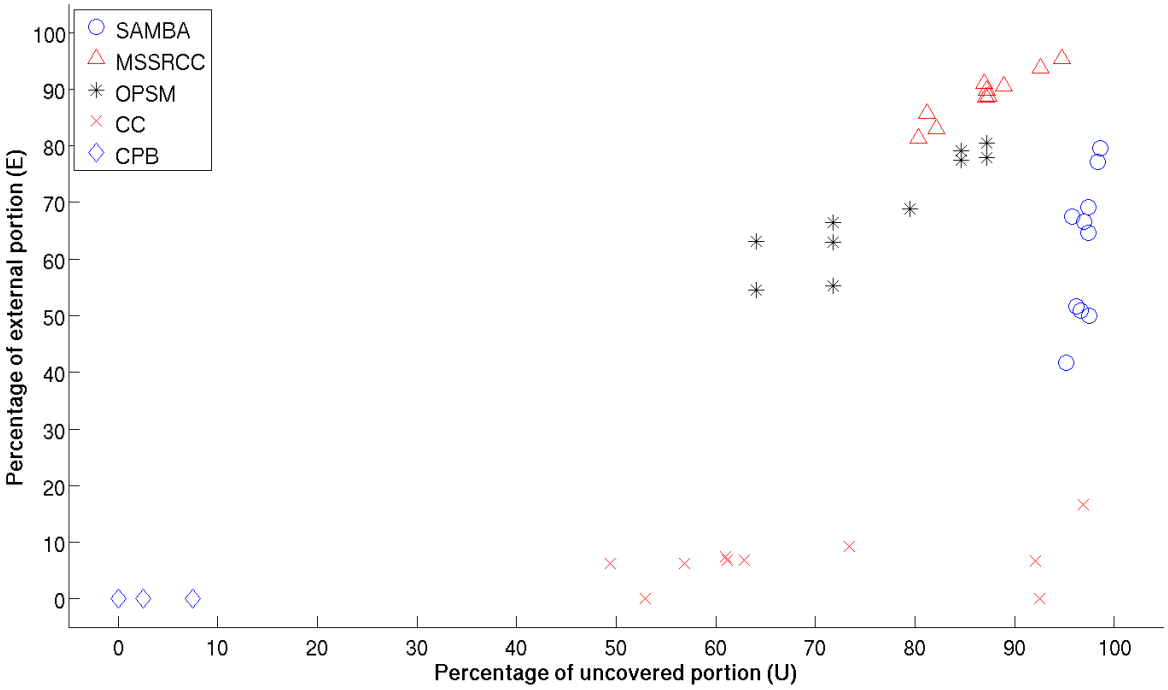
- Implanted a **60x60** bicluster with shift pattern



- The same results shown before.

Effect of Bicluster Size

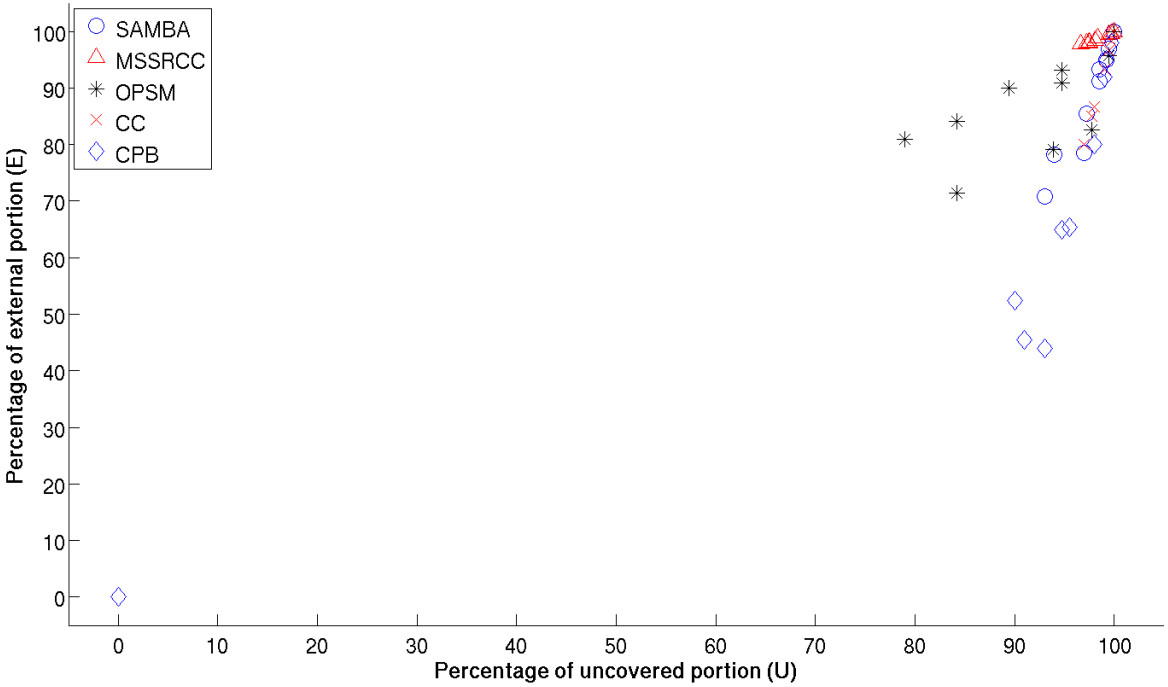
- Implanted a **40x40** bicluster with shift pattern



- It gets harder to detect a smaller bicluster
- CPB still perfectly identified a 40x40 bicluster in 8 datasets

Effect of Bicluster Size

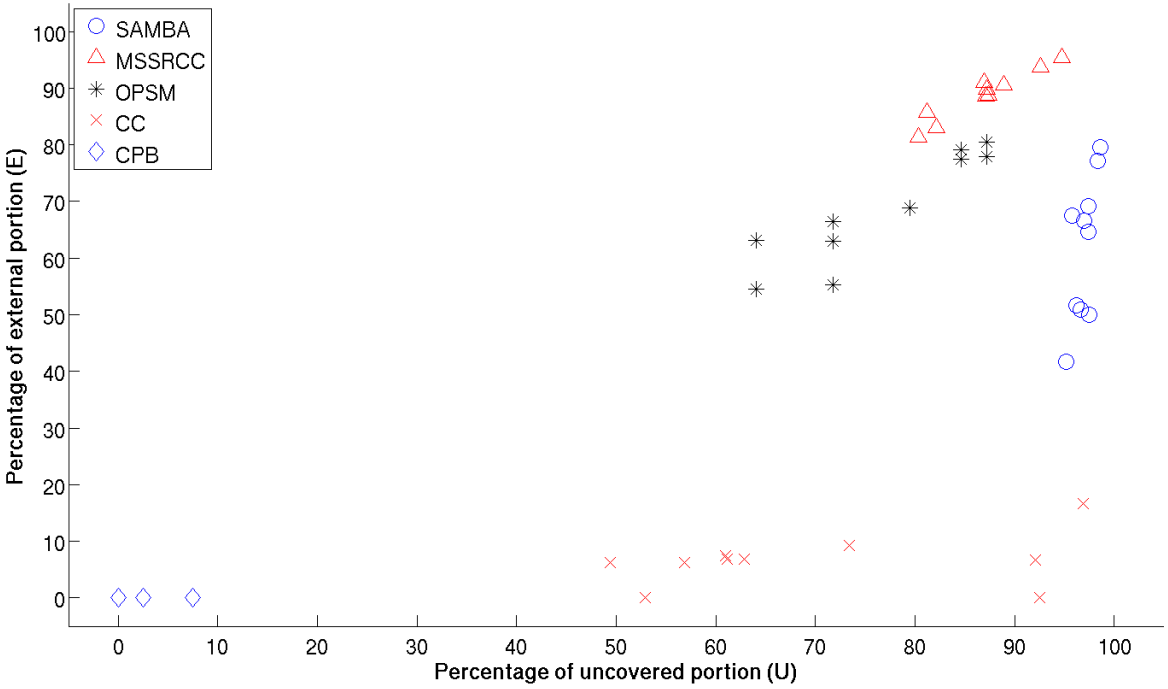
- Implanted a **20x20** bicluster with shift pattern



- It gets harder to detect a smaller bicluster
- CPB perfectly identified a 20x20 bicluster in only 1 dataset

Effect of Noise

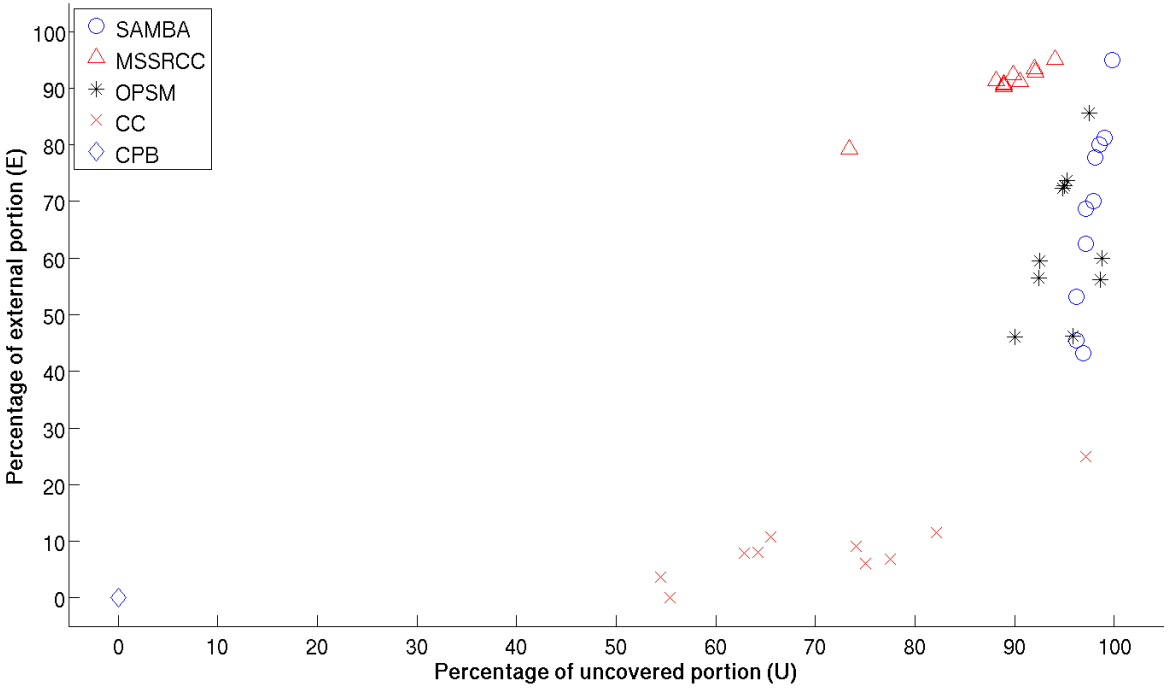
- Implanted a 40x40 bicluster with shift pattern **without noise**



- The same results shown before.
- Next: each value is randomly incremented to simulate noise.

Effect of Noise

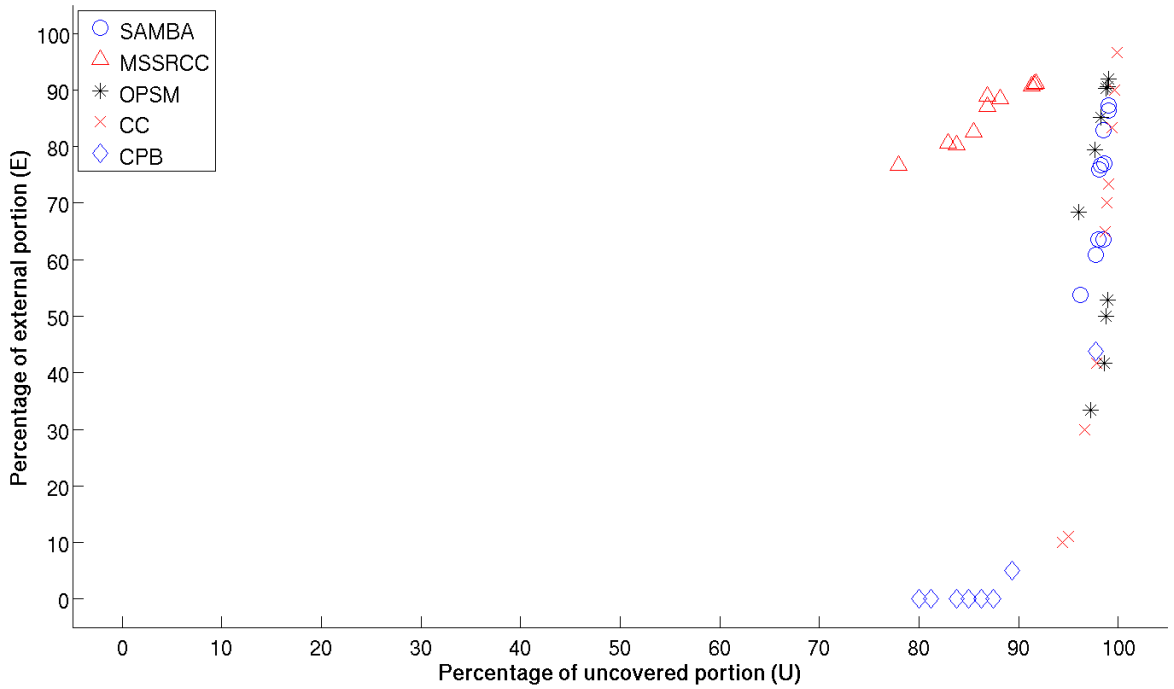
- Implanted a 40x40 bicluster with shift pattern **with 5% noise**



- Performance drops in general
- CPB is least affected, OPSM is most affected

Effect of Noise

- Implanted a 40x40 bicluster with shift pattern **with 20% noise**



- Performance drops dramatically
- CPB is still successful at returning minimal external portion

Effect of Overlap

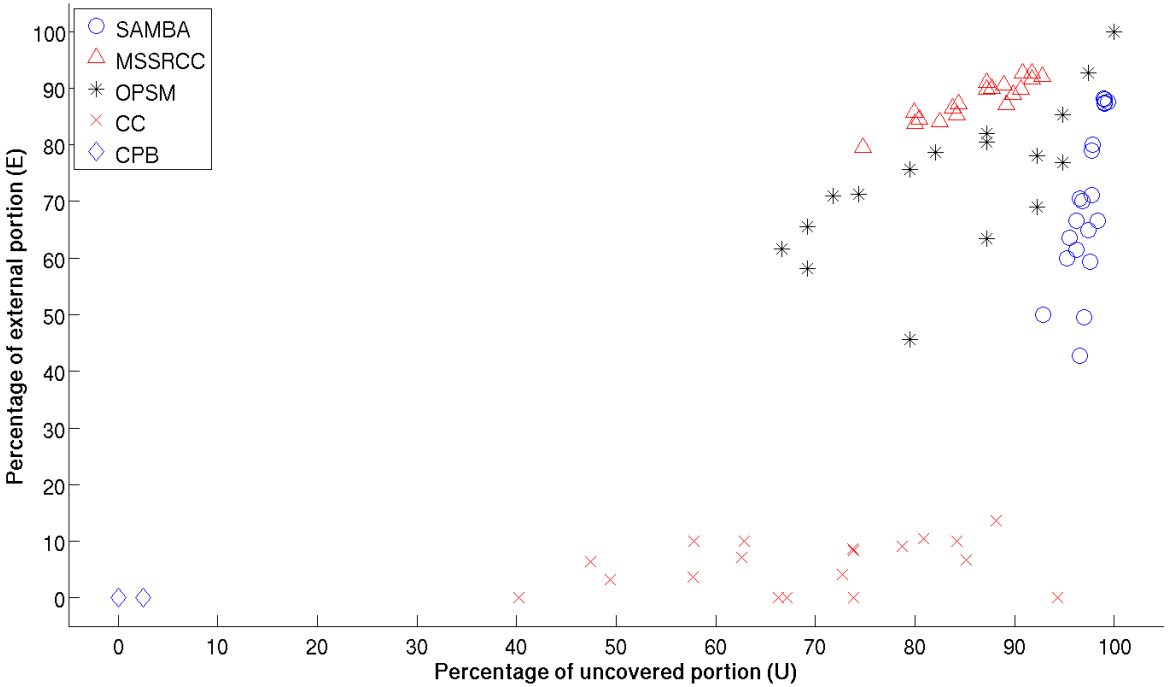
- Implanted 2 overlapping biclusters in each dataset
  - Total of 20 over 10 datasets

1	2	3	4		
5	6	7	8		
9	10	11	12	13	14
13	14	15	16	17	18
		19	20	21	22
		23	24	25	26

- 50% overlap => 50% of rows and 50% of columns overlap

Effect of Overlap

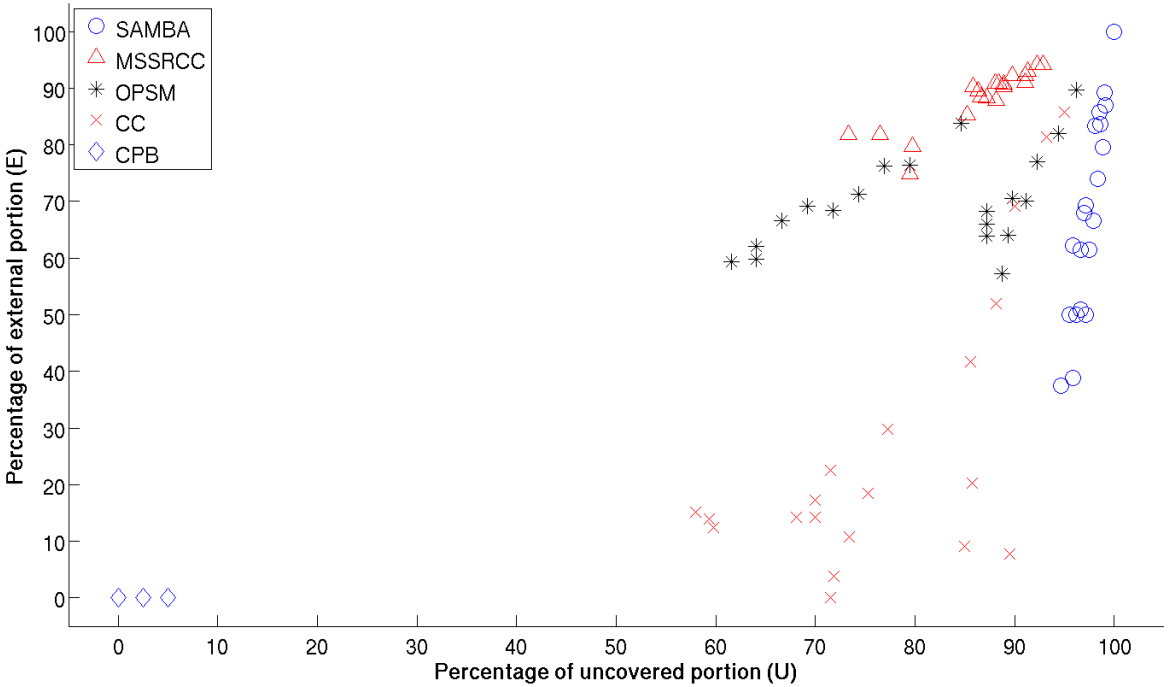
- Implanted 40x40 biclusters with shift pattern **without overlap**



- Similar results as before

Effect of Overlap

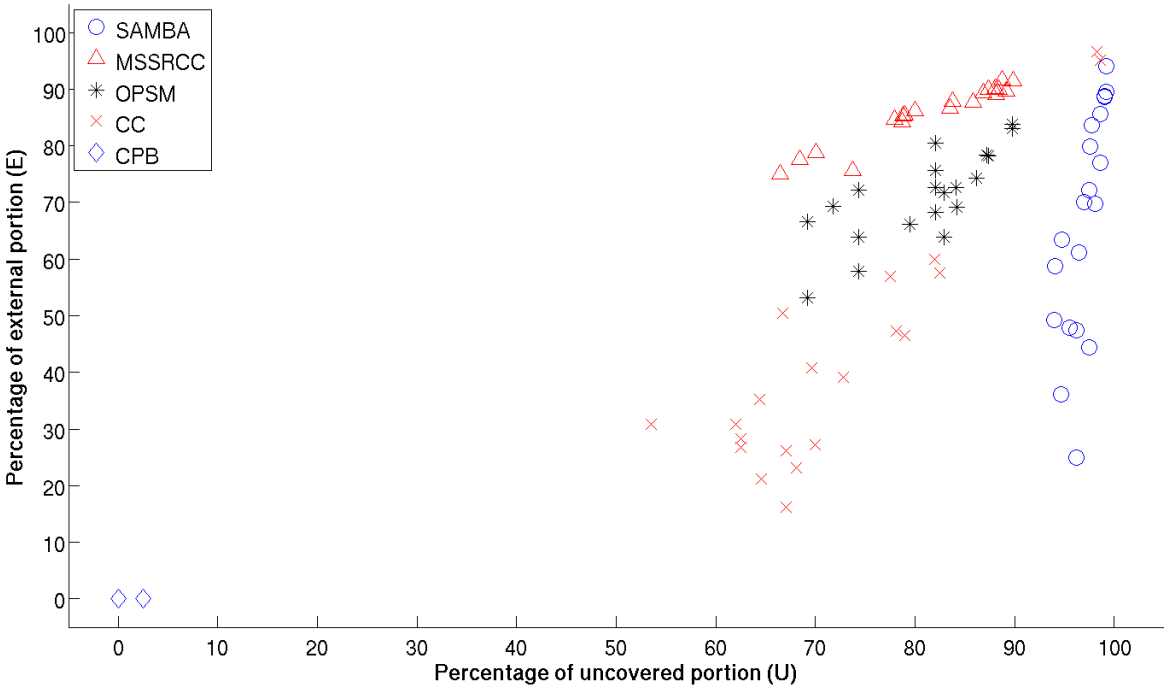
- Implanted 40x40 biclusters with shift pattern **with 25% overlap**



- Performance of CPB and OPSM was not affected significantly
- Performance of CC drops due to random masking

Effect of Overlap

- Implanted 40x40 biclusters with shift pattern **with 50% overlap**



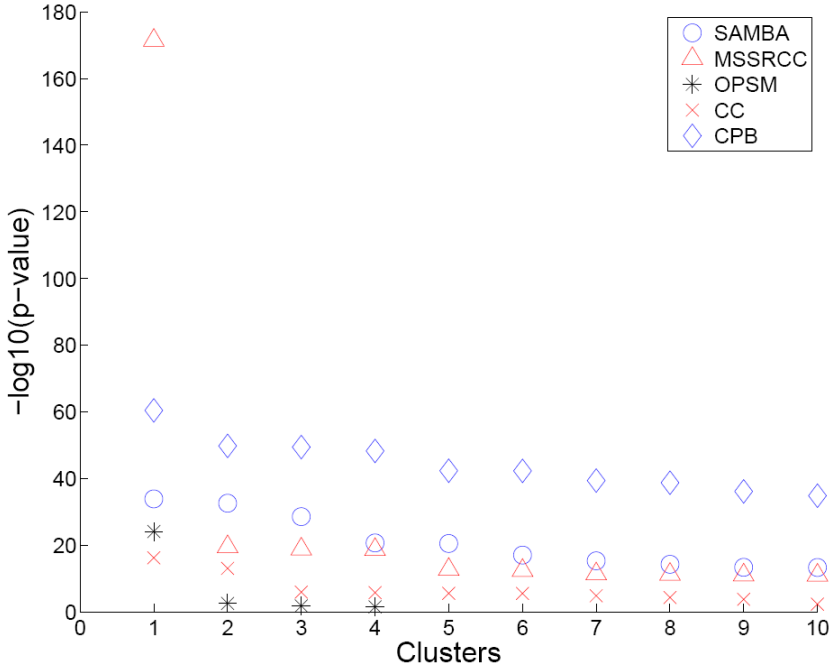
- Performance of CPB and OPSM was not affected significantly
- Performance of CC drops due to random masking

Experiments on Real Datasets

- Datasets from the Gene Expression Omnibus (GEO) database
  - Yeast (GDS1611) – 9275 genes, 96 conditions
  - Mouse (GDS1406) – 12422 gens, 87 conditions
  - Drosophila (GDS1739) – 13966 genes, 54 conditions
- Evaluation based on Gene Ontology (GO) term enrichment.
- Top 10 clusters with the most enriched GO terms are reported
  - For each cluster, -10log(p-value) of the most enriched term is reported

Experiments on Real Datasets

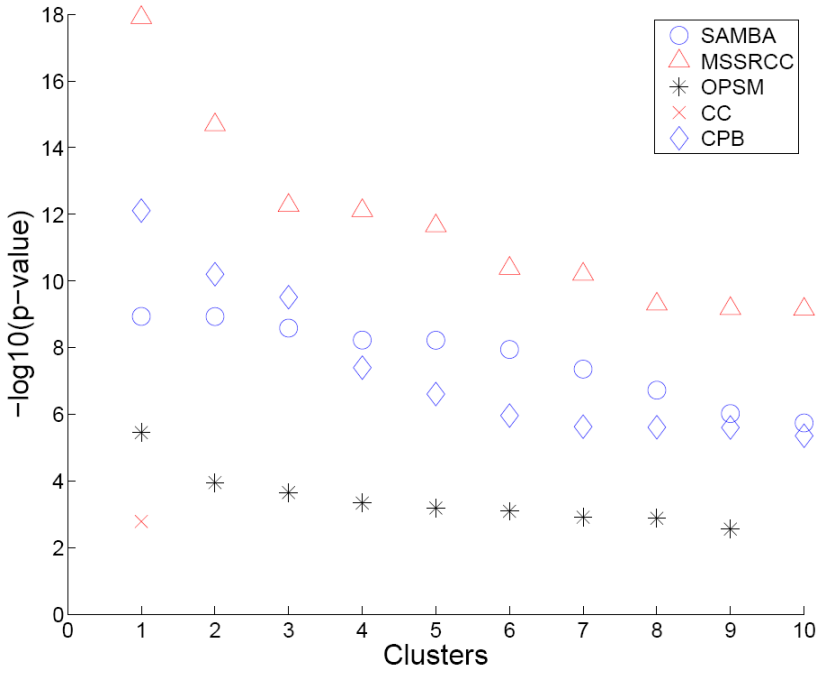
- Yeast dataset



- CPB was better in general, but MSSRCC found the best cluster
- SAMBA clusters and one of the OPSM clusters were also good.

Experiments on Real Datasets

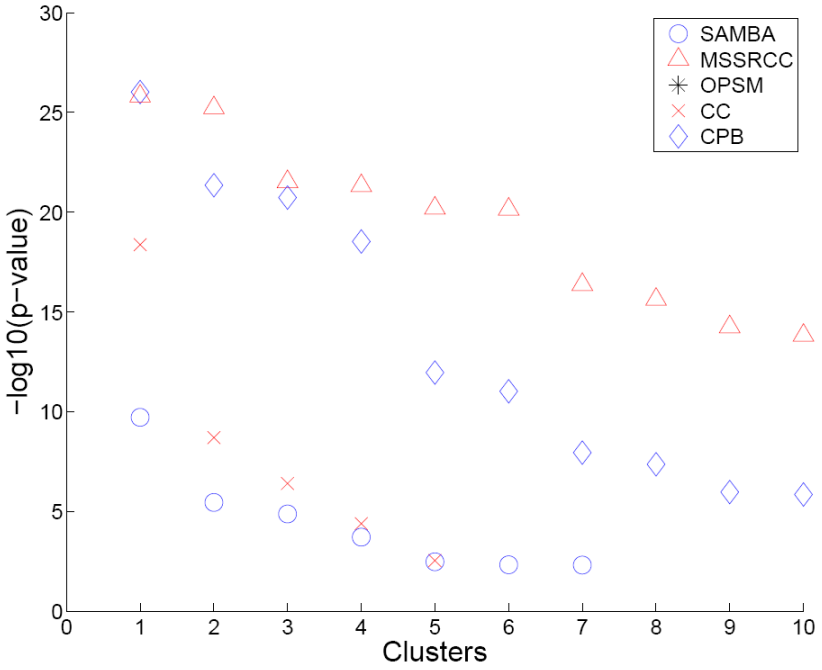
- Mouse dataset



- Most enriched clusters by MSSRCC, followed by CPB and SAMBA.
- Algorithms that seek global patterns are also strong

Experiments on Real Datasets

- Drosophila dataset



- MSSRCC and CPB again performed the best
- One of the CC clusters was also good

Conclusions

- Compared biclustering algorithms on the basis of bicluster patterns and power of search technique
  - Focused on local patterns
- CPB performs significantly better, good candidate to detect shifting and scaling patterns
  - Robust against noise, overlaps and varying in bicluster sizes
- Clusters found by CPB and MSSRCC on real datasets were more significantly enriched
  - Patterns sought by CPB and MSSRCC may have higher biological relevance

Thanks

- For more information visit
  - umit@bmi.osu.edu
  - http://bmi.osu.edu/~umit or http://bmi.osu.edu/hpc

- Research at the HPC Lab is funded by