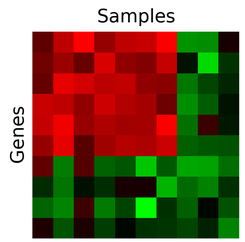
A comparison of recent biclustering algorithms

Mehmet Deveci, Kemal Eren, Ümit V. Çatalyürek

Introduction

Biclustering: an unsupervised data analysis method for simultaneous clustering of both rows and columns in a data matrix. Often applied to microarray data.



Datasets

Synthetic datasets with six bicluster models were generated: constant, constant upregulated, shift, scale, shift-scale, and plaid. Results on ten microarray experiments were also obtained.

Experiments

- Ability to recover each type of bicluster.
- Varying numbers of biclusters.
- Sensitivity to random noise.
- Overlapping biclusters.
- Overfitting synthetic data.
- GEO enrichment on microarray data.

Scoring

Biclusters b_1 and b_2 with Jaccard coefficient:

$$s(b_1, b_2) = \frac{|b_1 \cap b_2|}{|b_1 \cup b_2|}$$

Sets of biclusters M_1 and M_2 compared with:

$$S^*(M_1, M_2) = \frac{1}{|M_1|} \sum_{b_1 \in M_1} \max_{b_2 \in M_2} s(b_1, b_2)$$

Let E be set of expected, F set of found: recovery = $S^*(E, F)$; relevance = $S^*(F, E)$.

Results

Figure 1: Bicluster type experiment. Each data point represents the mean of 100 datasets.

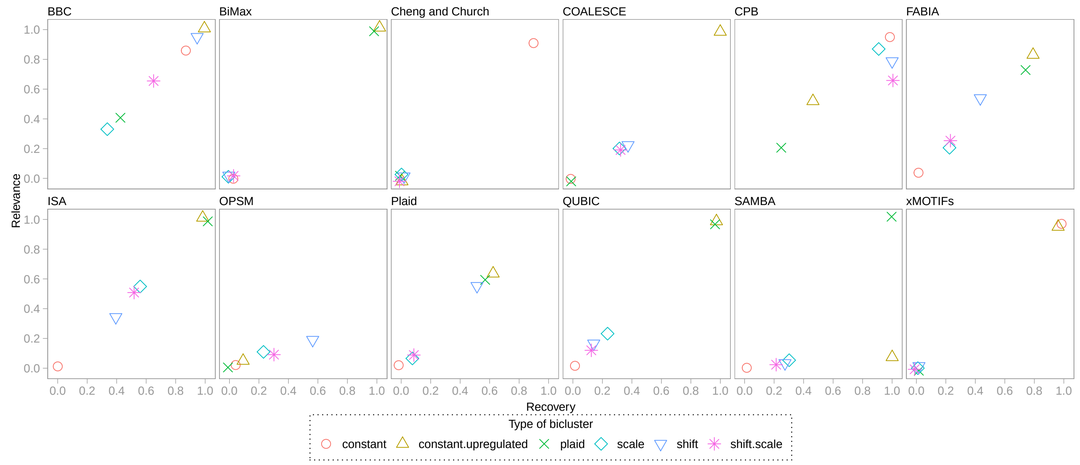


Figure 2: Noise experiment: recovery scores for constant and constant upregulated biclusters.

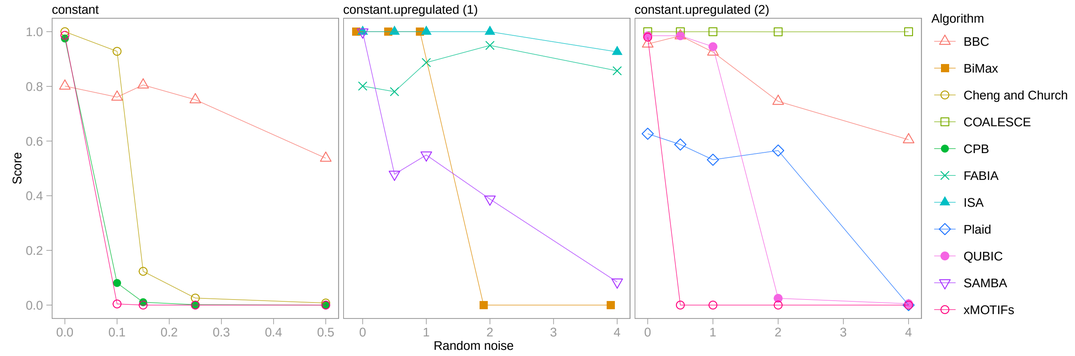


Table 1: Summary of bicluster types. **o** = supported & found. **x** = supported & not found. **!** = not supported & found. **bold**: best. Distance threshold: 0.75

Algorithm	const	const-up	plaid	shift	scale	shift-scale
BBC	o	**o**	x	o		!
BiMax		**o**	!			
Cheng and Church	**o**	x				
COALESCE		**o**				
CPB	**o**	x		o	o	o
FABIA	x	**o**	!		x	
ISA		**o**	!		!	
OPSM	x	x		**x**	x	x
Plaid	x	**o**	o	o		
QUBIC	x	**o**	!			
SAMBA		**o**	!			
xMOTIFs	**o**	o				

Table 2: Microarray results. Bayesian binomial test for % of significant biclusters, $\alpha = 0.05$. *n*: number of biclusters found; *x*: number of enriched biclusters; *lower* and *upper*: 95% posterior interval; *mean*: expectation of the posterior.

algorithm	n	x	lower	mean	upper
BBC	400	44	0.08	0.11	0.14
BiMax	14	8	0.32	0.57	0.80
Cheng and Church	68	5	0.03	0.08	0.15
COALESCE	5522	319	0.00	0.06	0.25
CPB	186	14	0.04	0.08	0.12
FABIA	202	30	0.10	0.15	0.20
ISA	24	6	0.11	0.26	0.44
OPSM	59	11	0.10	0.19	0.30
Plaid	55	12	0.13	0.22	0.34
QUBIC	48	3	0.02	0.07	0.16
SAMBA	199	38	0.14	0.19	0.25
xMOTIFs	420	60	0.11	0.14	0.18

Overall grades

Algorithm	Avg.Dist	Exp.Avg.Dist	Number	Noise	Overlap	Overfit	Real	Mean	Relative	Grade
BBC	0.70	0.80	0.93	0.90	0.80	0.00	0.08	0.60	0.90	A
BiMax	0.34	1.00	1.00	1.00	1.00	0.00	0.32	0.67	1.00	A
Cheng and Church	0.15	0.46	1.00	0.96	0.87	0.04	0.03	0.50	0.75	C
COALESCE	0.31	1.00	1.00	1.00	0.77	0.00	0.00	0.58	0.88	B
CPB	0.72	0.82	0.98	0.83	0.97	0.26	0.04	0.66	0.99	A
FABIA	0.42	0.35	0.75	0.79	0.75	0.00	0.10	0.45	0.68	D
ISA	0.57	1.00	1.00	1.00	0.84	0.00	0.11	0.65	0.97	A
OPSM	0.15	0.18	0.00	0.28	0.57	0.16	0.10	0.21	0.31	F
Plaid	0.32	0.44	0.78	0.56	0.61	0.00	0.13	0.40	0.61	D
QUBIC	0.41	0.49	0.98	0.99	0.91	0.71	0.02	0.64	0.97	A
SAMBA	0.38	0.71	1.00	0.74	1.00	0.00	0.14	0.57	0.85	B
xMOTIFs	0.33	0.98	0.94	0.49	0.94	0.00	0.11	0.54	0.81	B

Table 3: Grades for each experiment.

Conclusions

Disagreement between results on synthetic and real data. GEO enrichment probably not accurate for biclustering because:

- GEO enrichment only scores bicluster rows.
- Unknown biclusters receive low scores.

Best: BBC, CPB, ISA, and QUBIC. Only CPB capable of finding shift-scale biclusters.

Funding: This work was supported in parts by the NIH/NCI Grant R01CA141090; by the DOE SciDAC Institute Grant DE-FC02-06ER2775; and by the NSF Grants CNS-0643969, OCI-0904809 and OCI-0904802.

HPC Lab, Biomedical Informatics, Ohio State University, Columbus, OH {mdeveci, ekemal, umit}@bmi.osu.edu http://bmi.osu.edu/hpc