Comparison of sparse biclustering algorithms for gene expression datasets

Katherine Nicholls [1 2] and Chris Wallace [1 2]

[1]Cambridge Institute for Therapeutic Immunology and Infectious Disease, University of Cambridge, Cambridge, CB2 0AW, UK [2]MRC Biostatistics Unit, Cambridge Biomedical Campus, Forvie Site, Robinson Way, Cambridge, CB2 0SR, UK

Why biclustering?

Biclusters: groups of **genes that covary** in a **subset of the samples**.

- Adjusts for confounders
- **Gene signatures** only in **subset of patients**
- **Pathways specific to cell type** or condition

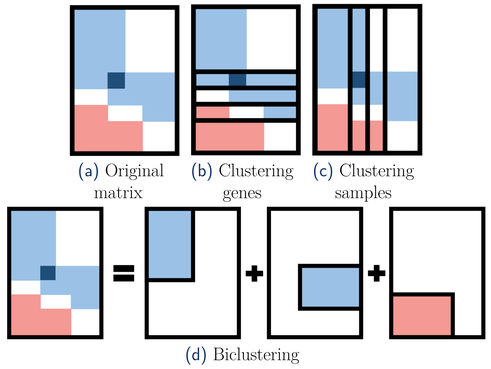

Figure 1: The same matrix is used for each type of clustering, with rows as genes and columns as samples. Only biclustering captures the true structure.

Algorithm classes

Table 1: Examples of the four classes of biclustering algorithm included.

Class	Advantages
Adaptive	Mixture of sparse and dense
SSLB, BicMix	biclusters, learn K automatically
NMF	Fast, interpretable
nsNMF, SNMF	
Popular	Benchmark - in previous
FABIA, Plaid	comparison studies
Tensor	Share information across cell
MultiCluster, SDA	types

Novel study features

- **Algorithm classes** not previously compared
- **Range of complexity** of simulated datasets
- **Direct evaluation** of biclustering on **real datasets**

Results

Novel thresholding step reveals biclusters

- Raw output had only trivial biclusters containing every gene and every sample
- After thresholding, diverse biclusters revealed
- Unnecessary for *Adaptive* algorithms

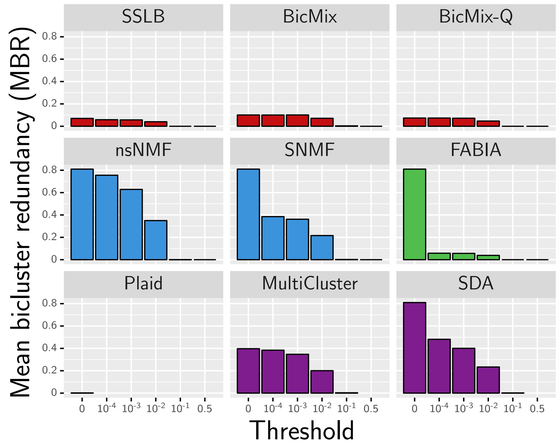

Figure 2: Novel metric MBR (Mean Bicluster Redundancy) measures similarity between biclusters within a run. Lower values preferred. This is shown for different threshold values. Raw output (threshold 0) of FABIA, *NMF* and *Tensor* algorithms contained many copies of same (trivial) biclusters, but after more severe thresholding this improves. We chose threshold 10^{-2} for analysis.

Adaptive algorithms most accurate on simulated datasets

- Using robust metric Clustering Error [1]

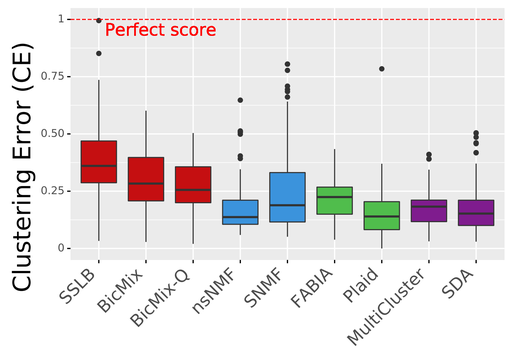

Figure 3: Clustering error (CE) across all simulated datasets. Larger values indicate higher accuracy. K_{init} is $K_{\text{true}} + 10$ for *Adaptive* algorithms and K_{true} otherwise. Thresholding has been applied. Runs that failed are discarded.

SSLB, Plaid and *NMF* algorithms best recovery of biclusters in knockout-mouse dataset

- Results vary by normalisation method

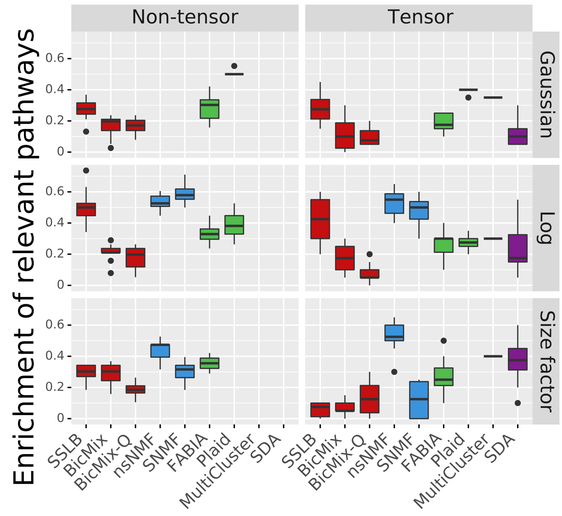

Figure 4: Bicluster recovery on IMPC knockout-mouse dataset [2]. Measured by mean proportion of knocked-out genes for which the bicluster best matching the samples where the gene was knocked out is enriched for at least one pathway containing the knocked-out gene. The measure is shown for tensor and non-tensor forms of the dataset and three different normalisation methods. *NMF* algorithms can't use Gaussian datasets, *Tensor* algorithms can't use non-tensor datasets.

nsNMF fastest, *Adaptive* algorithms slower

Table 2: Time in seconds for each algorithm to run on the largest simulated dataset and the main real dataset. Plaid failed to find any biclusters in the largest simulated dataset. Times under 5 minutes are underlined.

	Time to run (s)	
Algorithm	**Simulated**	**Real**
SSLB	6801	3904
BicMix	11250	354
BicMix-Q	29587	837
nsNMF	*263*	*6*
SNMF	*146*	29107
FABIA	1459	749
Plaid	*	*90*
MultiCluster	696	*40*
SDA	4746	3330

NMF algorithms and Plaid most robust

- Overall fairly low similarity between pairs of runs which differ only by seed

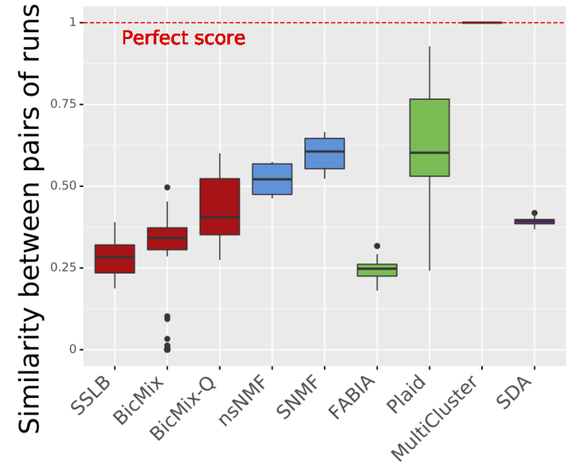

Figure 5: Each algorithm was run on the IMPC dataset with 10 different seeds. Plot shows similarity between pairs of such runs for each algorithm, as measured by Clustering Error. It thus gives a measure of robustness of the biclusters recovered by the algorithms.

Conclusion

- Novel post-processing thresholding invaluable
- *Adaptive* algorithms best for dataset with unknown K and without processing
- *NMF* algorithms have potential - fast and robust

Preprint

For full details, see the preprint on bioRxiv:
`https://doi.org/10.1101/2020.12.15.422852`

[1] Danilo Horta and Ricardo J.G.B. Campello. "Similarity Measures for Comparing Biclusterings". en. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 11.5 (Sept. 2014), pp. 942–954. ISSN: 1545-5963. DOI: 10.1109/TCBB.2014.2325016.

[2] Gautier Koscielny et al. "The International Mouse Phenotyping Consortium Web Portal, a Unified Point of Access for Knockout Mice and Related Phenotyping Data". In: *Nucleic Acids Research* 42.Database issue (Jan. 2014), pp. D802–D809. ISSN: 0305-1048. DOI: 10.1093/nar/gkt977.

UNIVERSITY OF CAMBRIDGE

MRC Medical Research Council

wellcome