Biclustering of very large datasets with GPU technology using CUDA

J. Arnedo-Fdez, I. Zwir, R. Romero-Zaliz

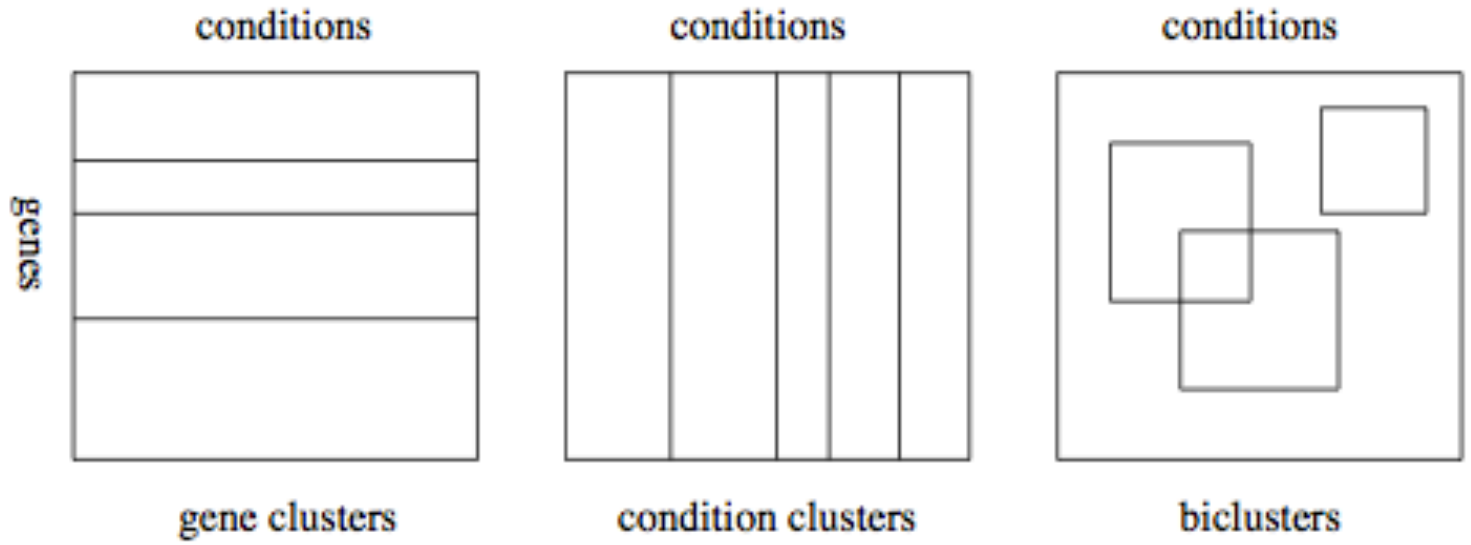
Dpto. de Ciencias de la Computación e I.A.

Universidad de Granada - España

MMLab

Biclustering

- Simultaneous clustering of rows and columns of a matrix.



Biclustering

- Cheng and Church original heuristics suffer of ***random interference***.
  - *…produced the masking of null values and discovered biclusters with random numbers.*

- Yang *et. al.* developed a probabilistic move-based algorithm called FLOC.
  - *…by locating biclusters simultaneously rather than sequentially.*

FLOC

- Starts from a set of seeds (initial biclusters).
- At each iteration, each row and column is moved among biclusters to produce a better biclustering in terms of lower mean squared residues.
- The best biclustering obtained during each iteration will serve as the initial biclustering for the next iteration.
- The algorithm ends when the current iteration fails to improve the overall biclustering quality.

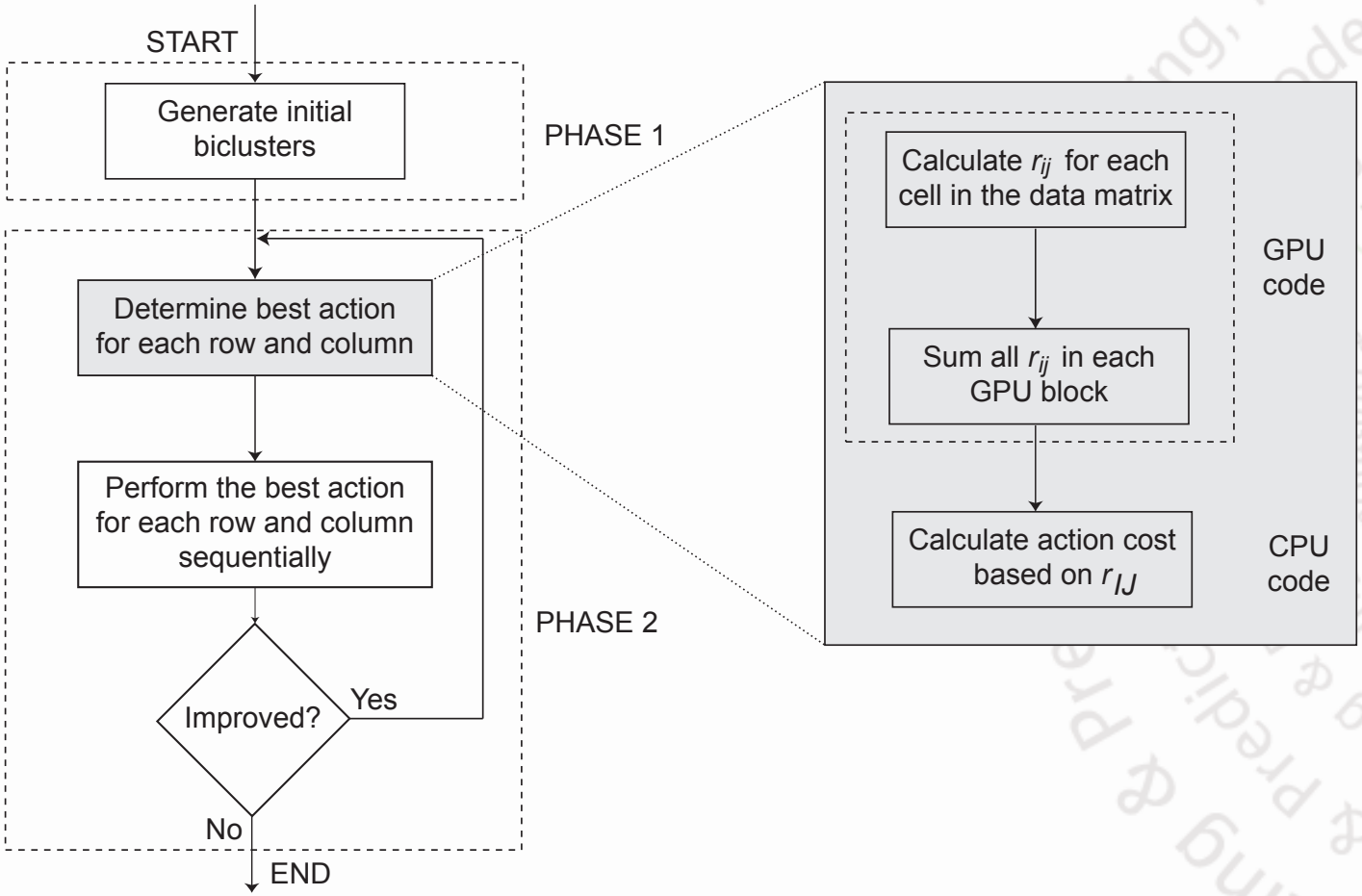
FLOC

- **Profiling...**

Each sample counts as 0.01 seconds.

% time	cumulative seconds	self seconds	calls	self s/call	total s/call	name
97.44	19.57	19.57	220010	0.00	0.00	residu
1.49	19.87	0.30	100	0.00	0.18	bestgain
0.65	20.01	0.13	440040	0.00	0.00	count_row_col
0.40	20.09	0.08	1020	0.00	0.00	sum
0.10	20.11	0.02	100	0.00	0.02	action
0.00	20.11	0.00	36817	0.00	0.00	echange
0.00	20.11	0.00	100	0.00	0.00	order
0.00	20.11	0.00	100	0.00	0.00	tri
0.00	20.11	0.00	1	0.00	20.11	floc

GP-GPU implementation



Experiments

CPU FLOC

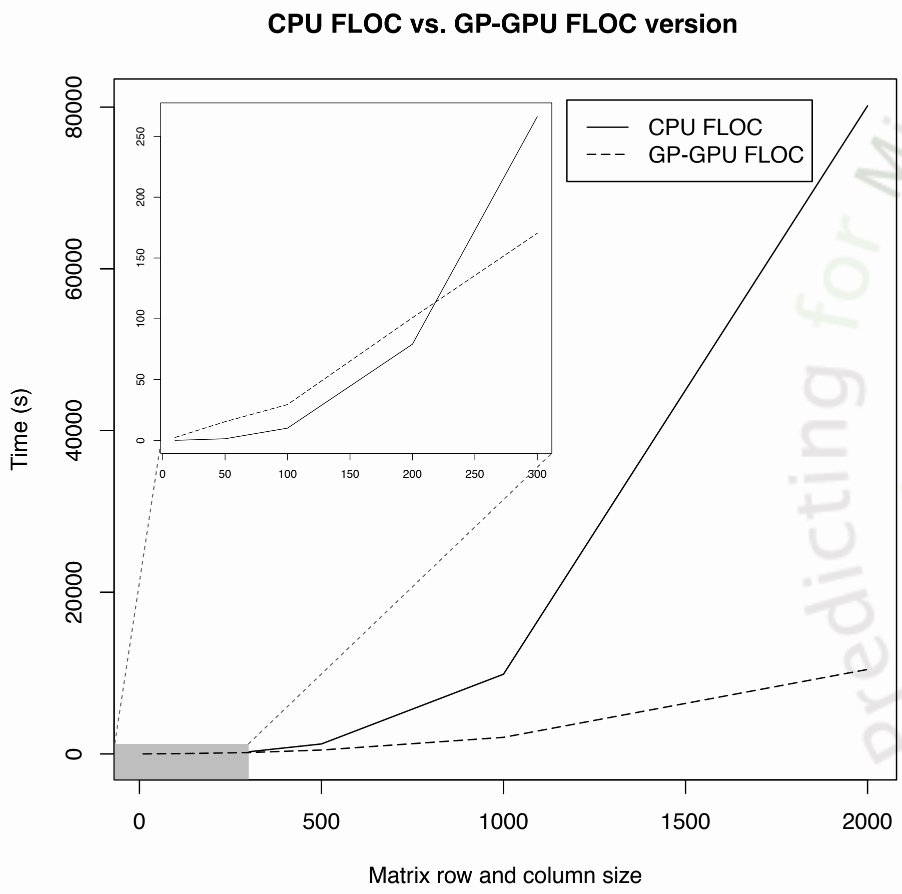
Matrix size	Time consumption (s)	
	Mean	*Standard Deviation*
10 × 10	0.01	1.79e-03
50 × 50	1.32	2.43e-02
100 × 100	10.06	4.47e-03
200 × 200	79.10	2.53e-01
300 × 300	266.34	6.79e-01
500 × 500	1233.60	1.41e+00
1000 × 1000	9859.20	8.47e+00
2000 × 2000	80152.90	1.56e+02

GP-GPU FLOC

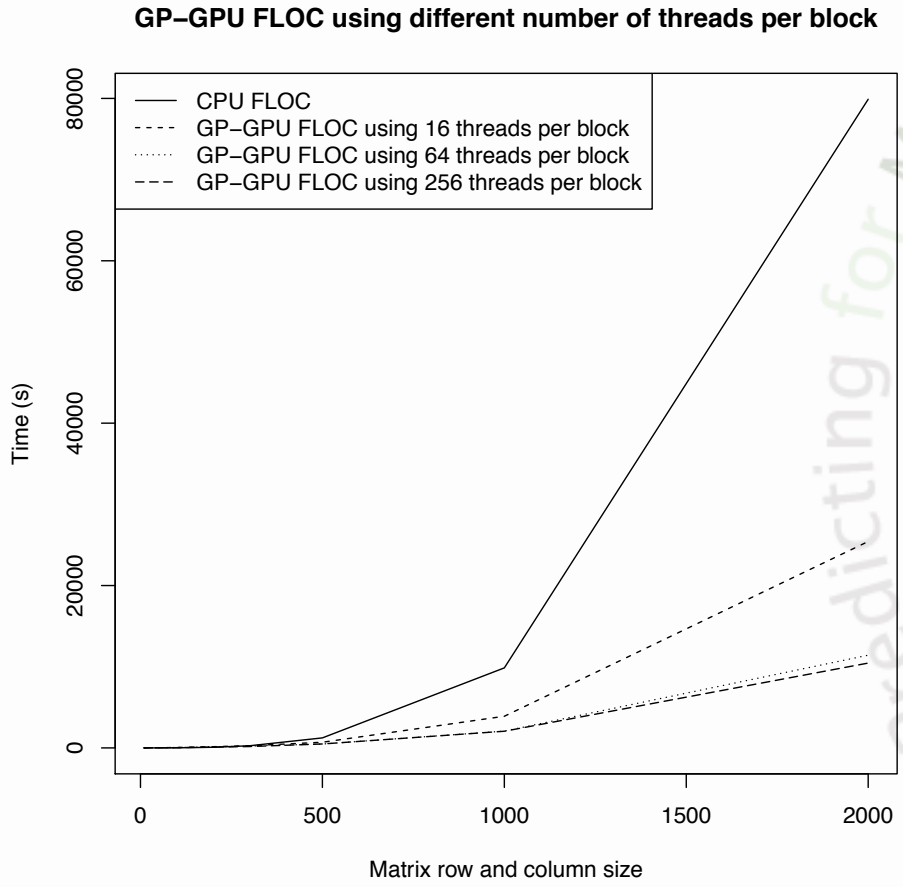
Matrix size	Time consumption (s)	
	Mean	*Standard Deviation*
10 × 10	2.40	6.98e-02
50 × 50	15.36	1.23e-01
100 × 100	29.47	2.18e-01
200 × 200	100.88	2.54e-01
300 × 300	170.40	4.34e-01
500 × 500	483.94	8.48e-01
1000 × 1000	2050.00	3.15e+00
2000 × 2000	10449.70	1.47e+01

** All experiments were run in an Intel i7 980 machine with 16 GB of RAM and Gainward GeForce GTX 480 video cards with 1.5 GB of RAM each.*

Experiments



Experiments



Experiments

- Larger matrices: 5000 x 5000
  - GPU implementation spent two days.
  - CPU version spent more than a week (and still running…).

Experiments

- Real world data: Gasch (2000).
- Dataset contains 6152 genes and 173 diverse environmental transition conditions such as temperature shock, amino acid starvation, and nitrogen source depletion.
  - GPU implementation spent 2 hours.
  - CPU version spent 9 hours.

Discussion

- Preliminary results show that the use of GPU acceleration can substantially improve the performance of biclustering methods.
  - …help bioinformatic software to cope with the large amount of data that Next Generation Sequencing technology is providing.

Discussion

- Memory transfers from the host CPU to the GPU devices over the PCI- Express bus is the main issue when programming GPU applications.
  - …bottleneck of the system, especially if large amounts of data need to be transferred over the bus.

Discussion

- Limit in the number of threads per block and blocks per grid that has to be considered.
- Not all algorithms are suitable for GPU acceleration.
- Future work will test all possible GPU parameter's configuration including the use of more than one GPU, and to compare them with other parallel architectures.

Thanks!

This work is supported by University of Granada
GREIB.PT.2011.20 - GREIB.AL.2011.06

rocio@decsai.ugr.es

MMLab

BarraCUDA

BarraCUDA

```
OpenSSH_5.2p1, OpenSSL 0.9.8r 8 Feb 2011
Last login: Tue Jul 17 14:01:27 2012 from 190.244.13.106


888888b.         d8888 8888888b.  8888888b.         d8888  .d8888b.  888     888 8888888b.         d8888
888  "88b       d88888 888   Y88b 888   Y88b       d88888 d88P  Y88b 888     888 888  "Y88b       d88888
888  .88P      d88P888 888    888 888    888      d88P888 888    888 888     888 888    888      d88P888
8888888K.     d88P 888 888   d88P 888   d88P     d88P 888 888        888     888 888    888     d88P 888
888  "Y88b   d88P  888 8888888P"  8888888P"     d88P  888 888        888     888 888    888    d88P  888
888    888  d88P   888 888 T88b   888 T88b     d88P   888 888    888 888     888 888    888   d88P   888
888   d88P d8888888888 888  T88b  888  T88b   d8888888888 Y88b  d88P Y88b. .d88P 888  .d88P  d8888888888
8888888P" d88P     888 888   T88b 888   T88b d88P     888  "Y8888P"   "Y88888P"  8888888P"  d88P     888

                     o
                                              ___..                       ,
                                      __..--''__ (                      .';
                  o      __.-------.-'            `--..__             .'  ;
                    _.--'   0)         .--._             ``--...____.'   .'
                   (vvvvv _.      )).   .__.-''                          <
                    ^^^^^___.....______.....-  -..___       _____...--'-.'.
                                      `-.___.'     ```````             `.;

[rocio@barracuda ~]$
```