Parallel large scale inference of protein domain families

Clément Rezvoy [1,5,6] Frédéric Vivien [3,6,5] Daniel Kahn [2,6,4]

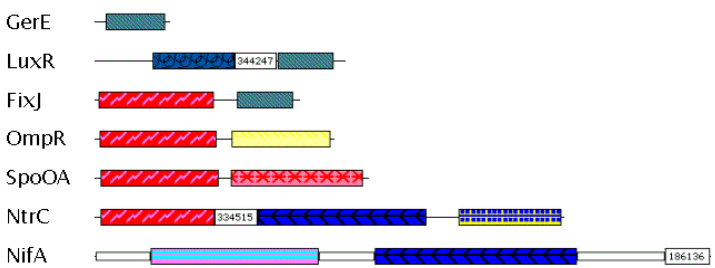
[1] ENS Lyon [2] INRA [3] INRIA [4] Lab. de Biométrie et Biologie Evolutive

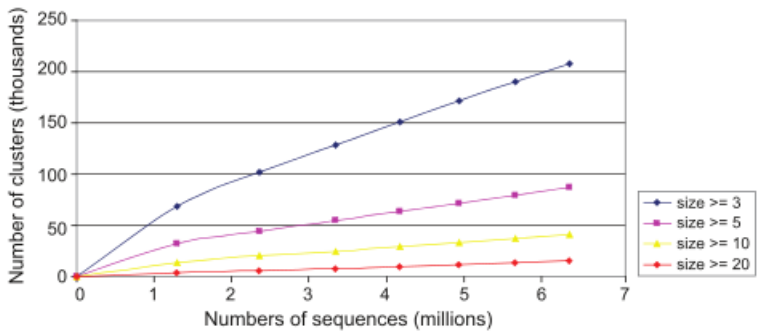
[5] Lab. de l'Informatique du Parallélisme [6] Université de Lyon

Wednesday September 8, 2010

Protein modularity

- Domains: conserved units of protein structure
- Proteins often composed of combinatorial arangement of domains





Yooseph *et al.* 2007

MPI_Mkdom2

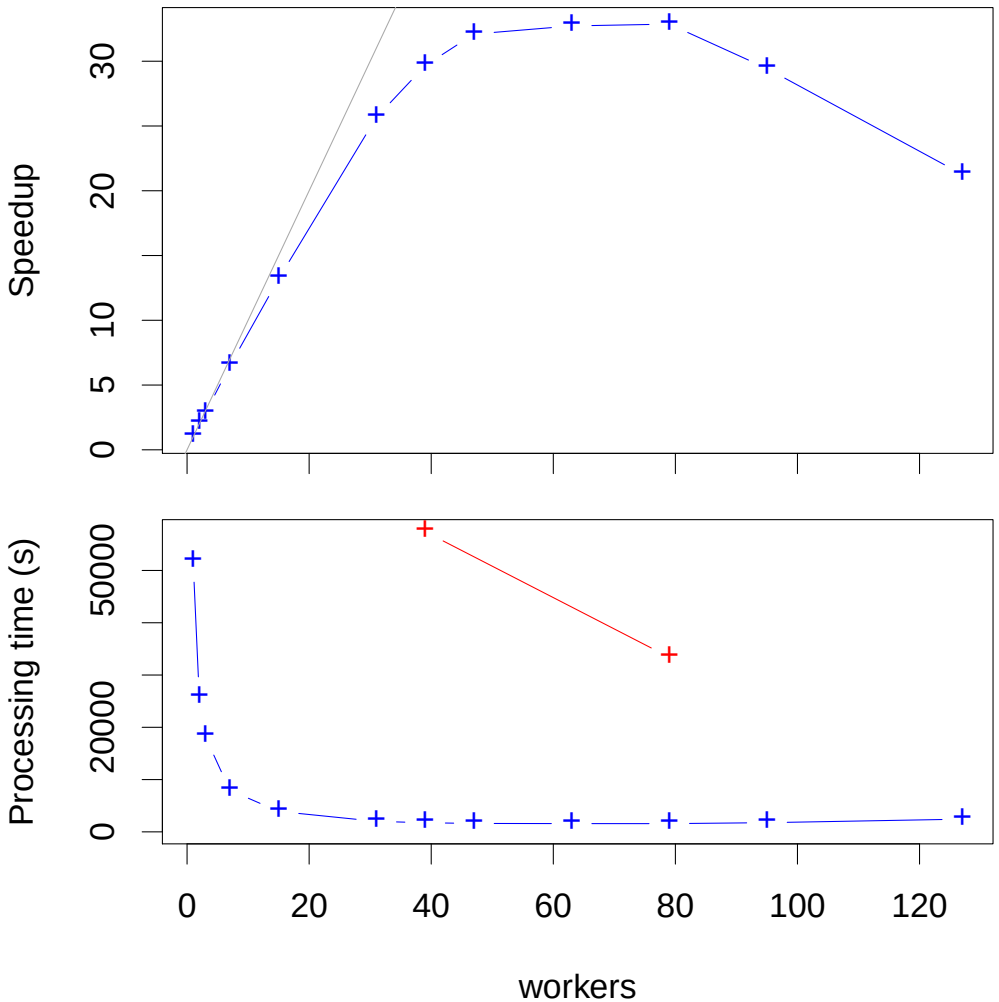
Mkdom2

- Greedy algorithm
- At each step: create a family around the shortest sequence.

MPI_Mkdom2

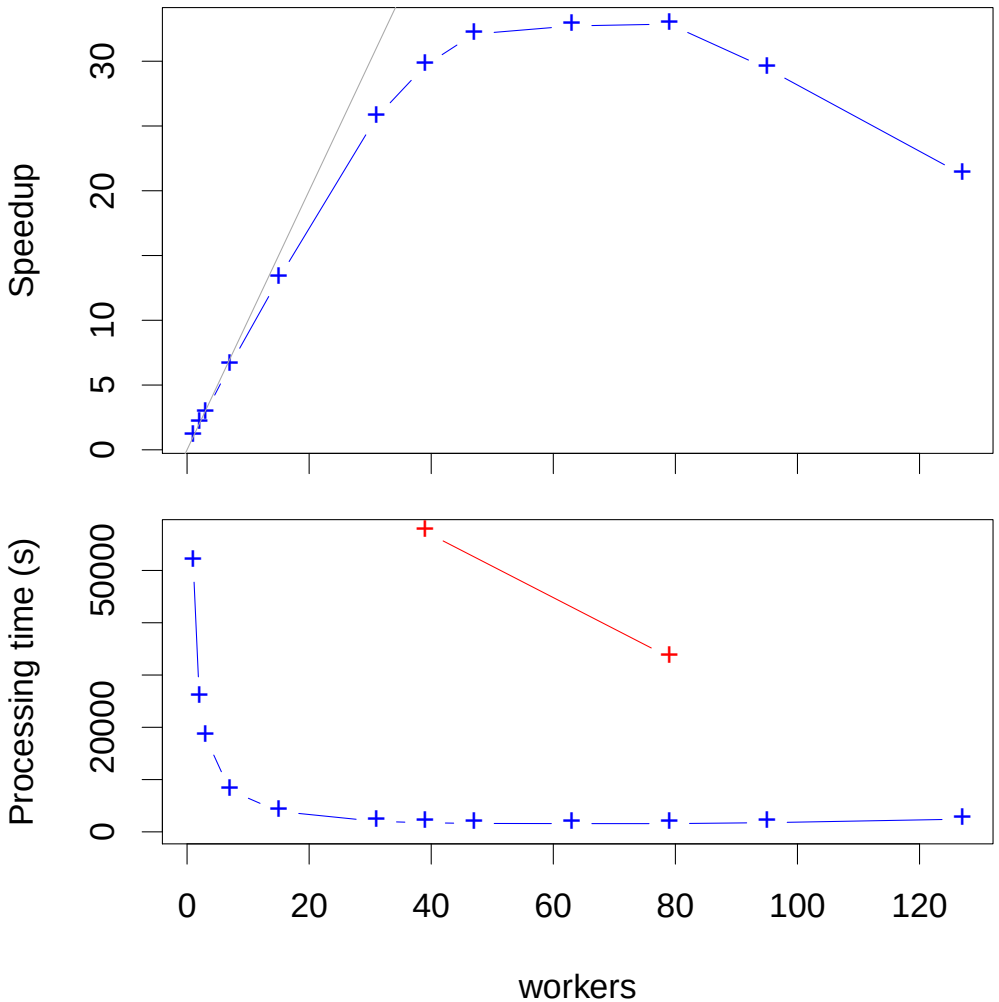
- Targeted at clusters, grids.
- Process multiple queries at once, check for overlapping results afterwards
- Try to avoid running interdependent queries on the basis of a global comparison.
- Handle large variations in query processing times.

Speedup



69,621 sequence dataset

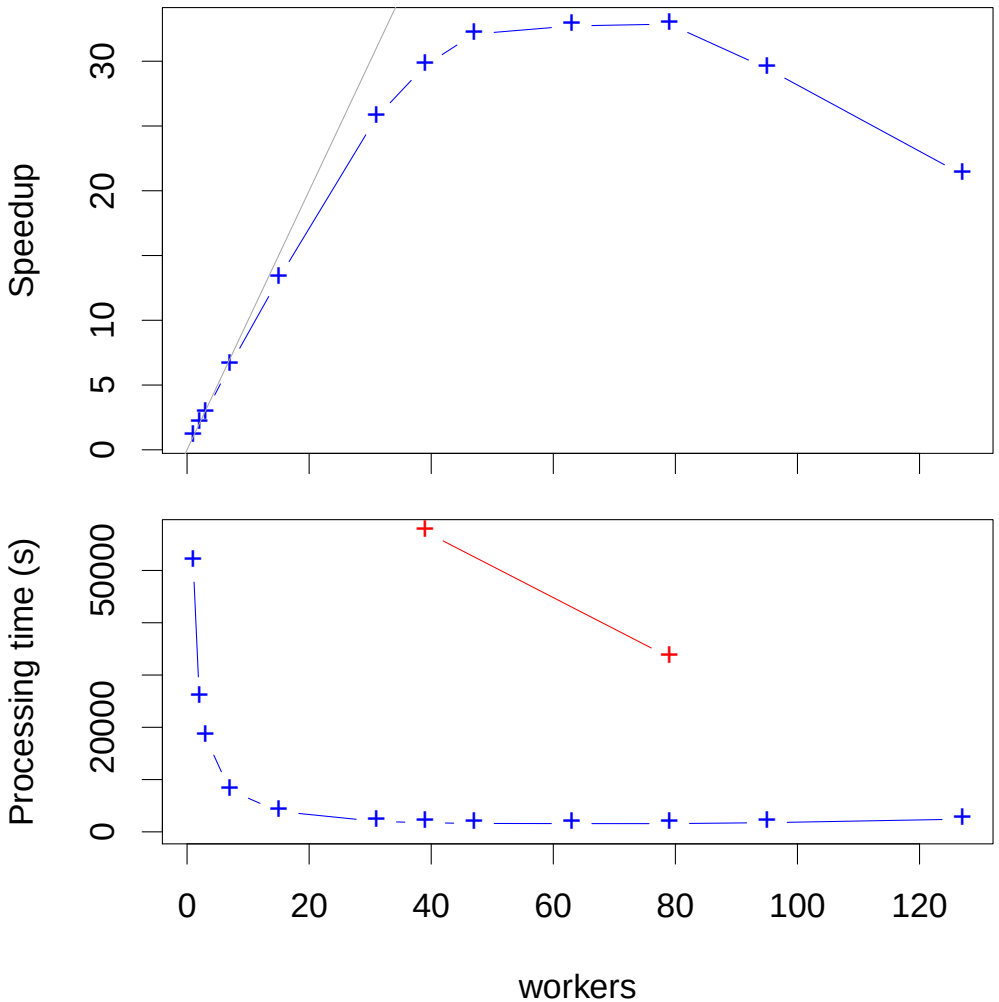
Speedup



69,621 sequence dataset

556,964 sequence dataset

Speedup



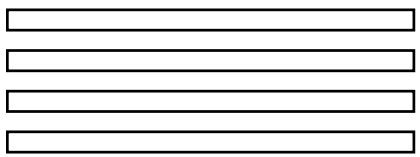
69,621 sequence dataset

556,964 sequence dataset

The larger the database, the larger the speedup

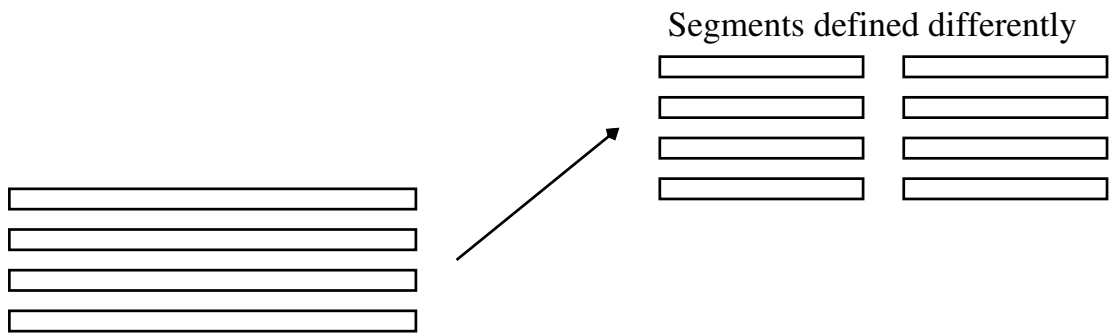
Comparing clusterings

- Need to compare both domain extents and domain clustering.



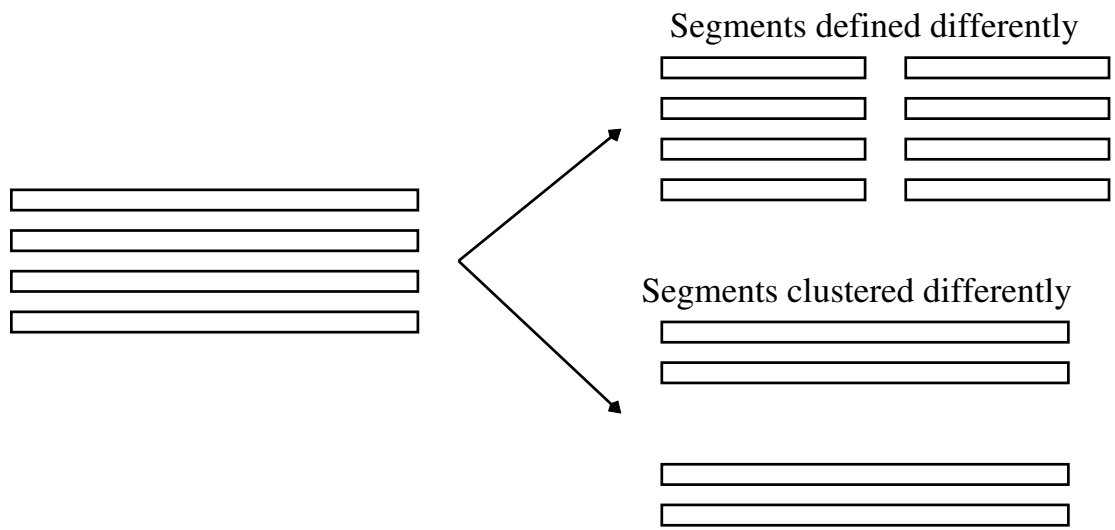
Comparing clusterings

- Need to compare both domain extents and domain clustering.



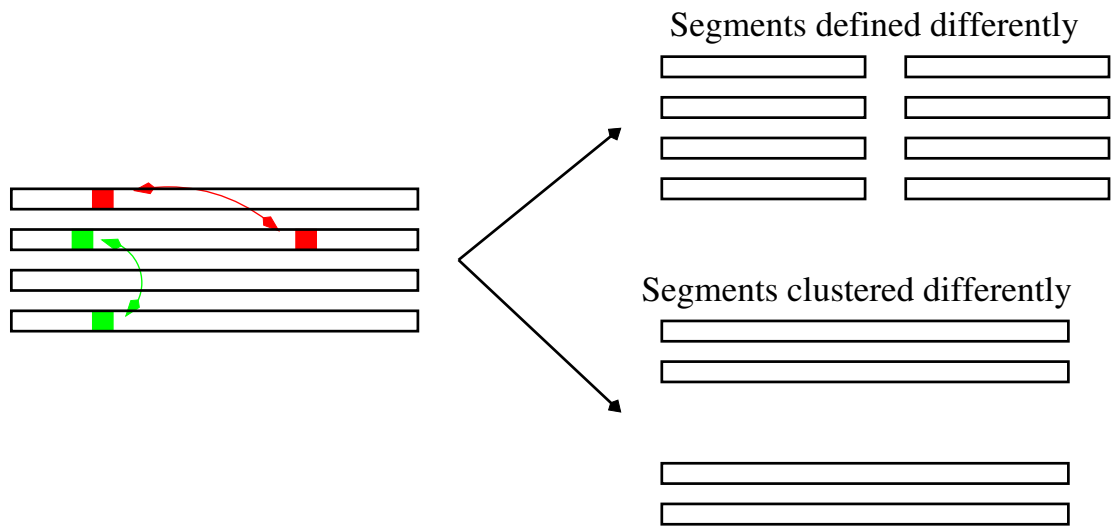
Comparing clusterings

- Need to compare both domain extents and domain clustering.



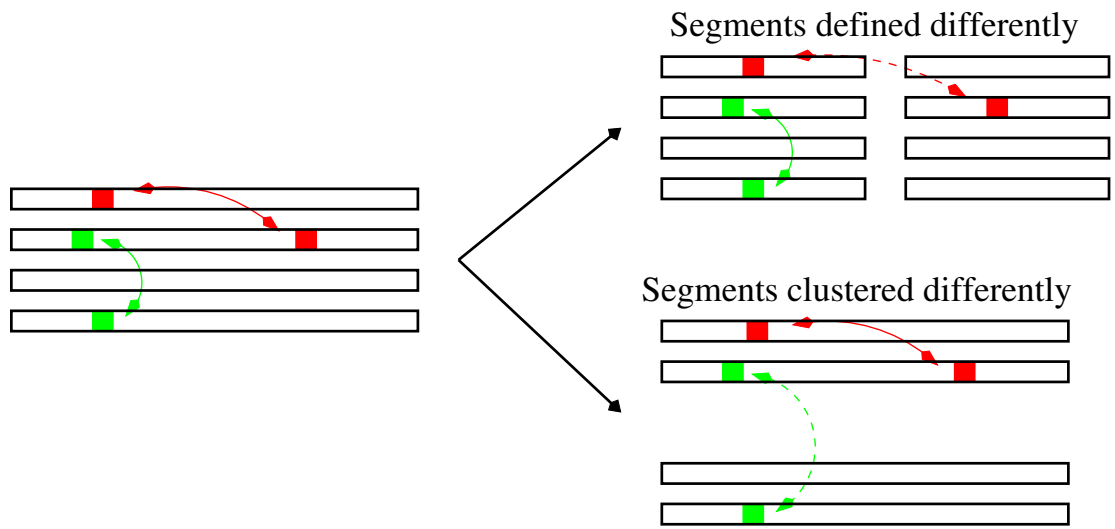
Comparing clusterings

- Need to compare both domain extents and domain clustering.
- Wallace index: counting preserved pairs.

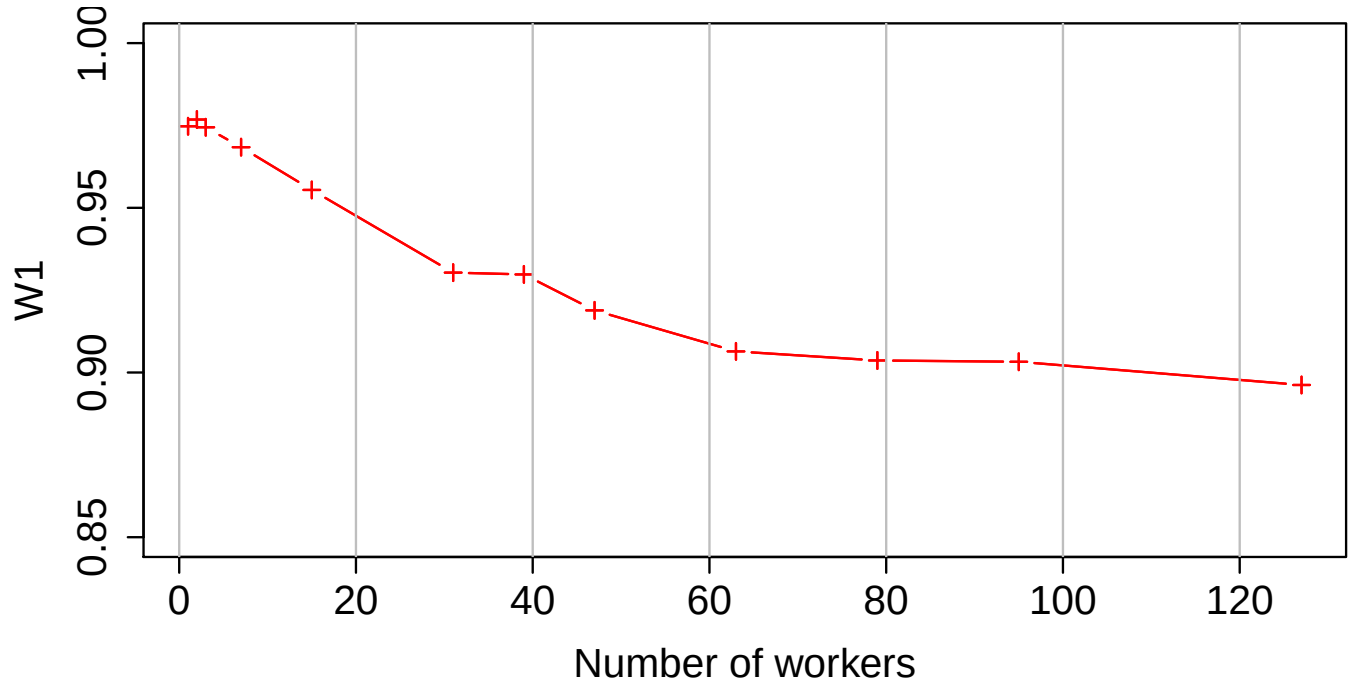


Comparing clusterings

- Need to compare both domain extents and domain clustering.
- Wallace index: counting preserved pairs.



Result validation: comparing parallel and sequential results



- Clustering reasonably preserved, even with a large number of workers.
- Compute ProDom 2010 (set of 6,118,869 sequences, in progress)