Whole genome alignment in High Performance Computing environments

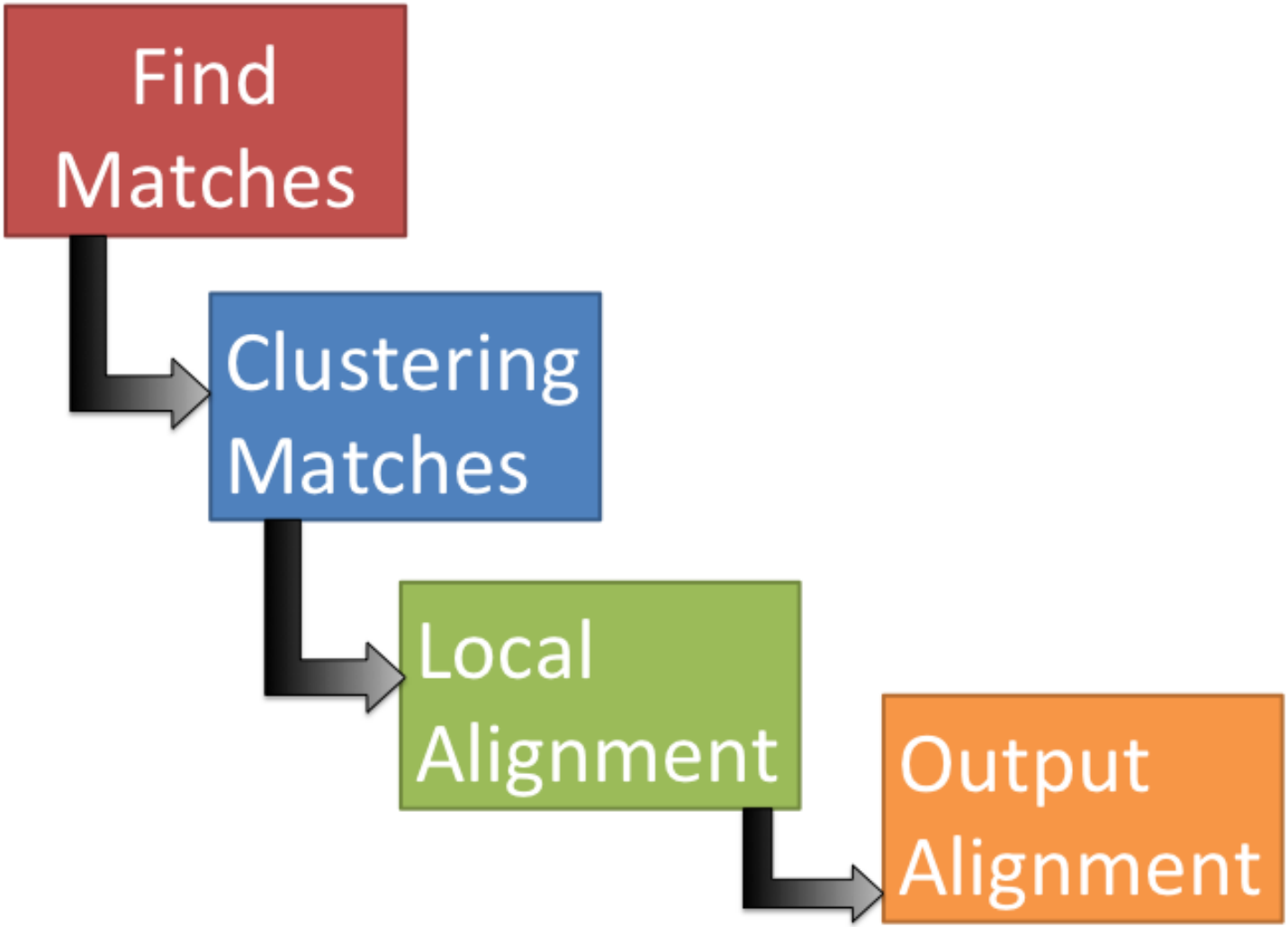
Julio César García Vizcaíno Directores: Antonio Espinosa, Juan Carlos Moure

Computer Architecture & Operating Systems Department
Universitat Autónoma de Barcelona

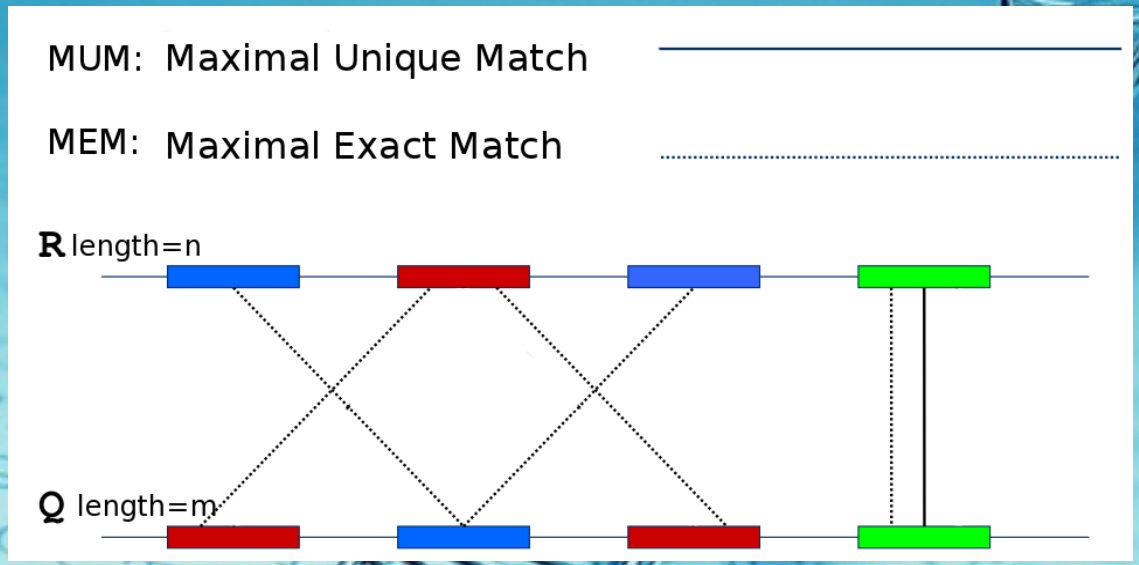
21 de enero de 2013

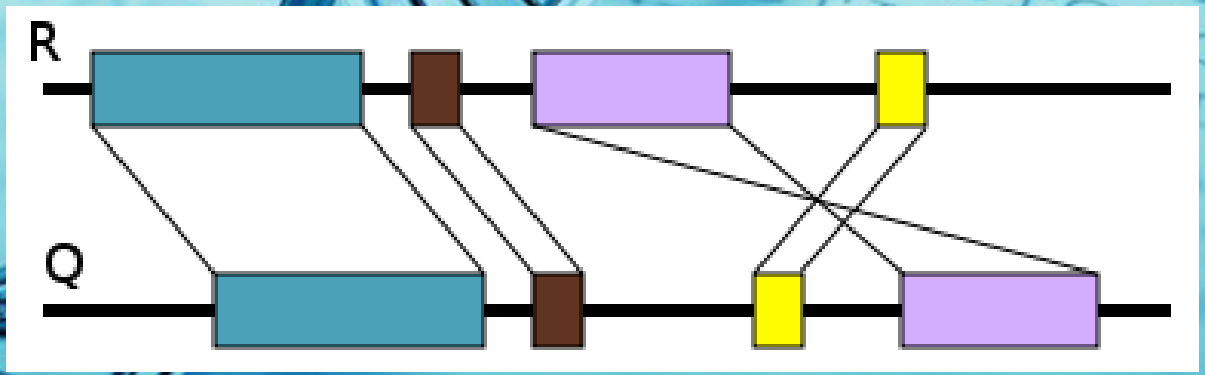
Contents

Whole Genome Alignment in MUMmer

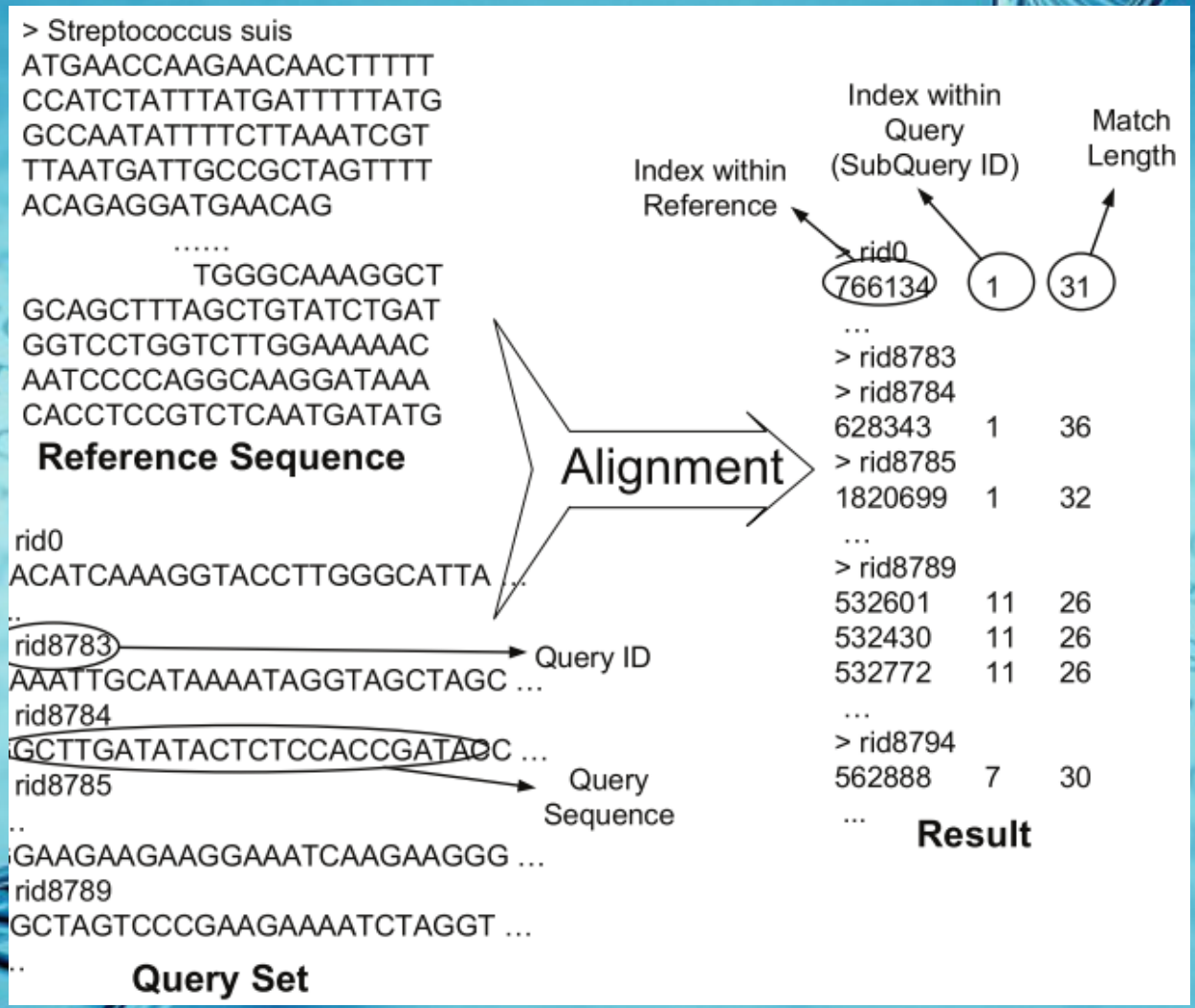


Search of Maximal Unique Matches





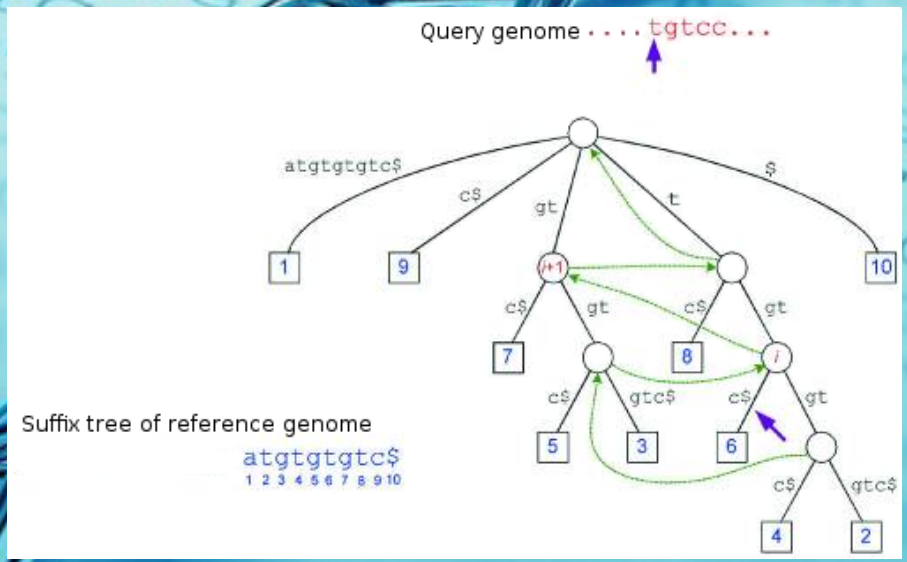
Genome alignment: search of Maximal Unique/Exact Matches



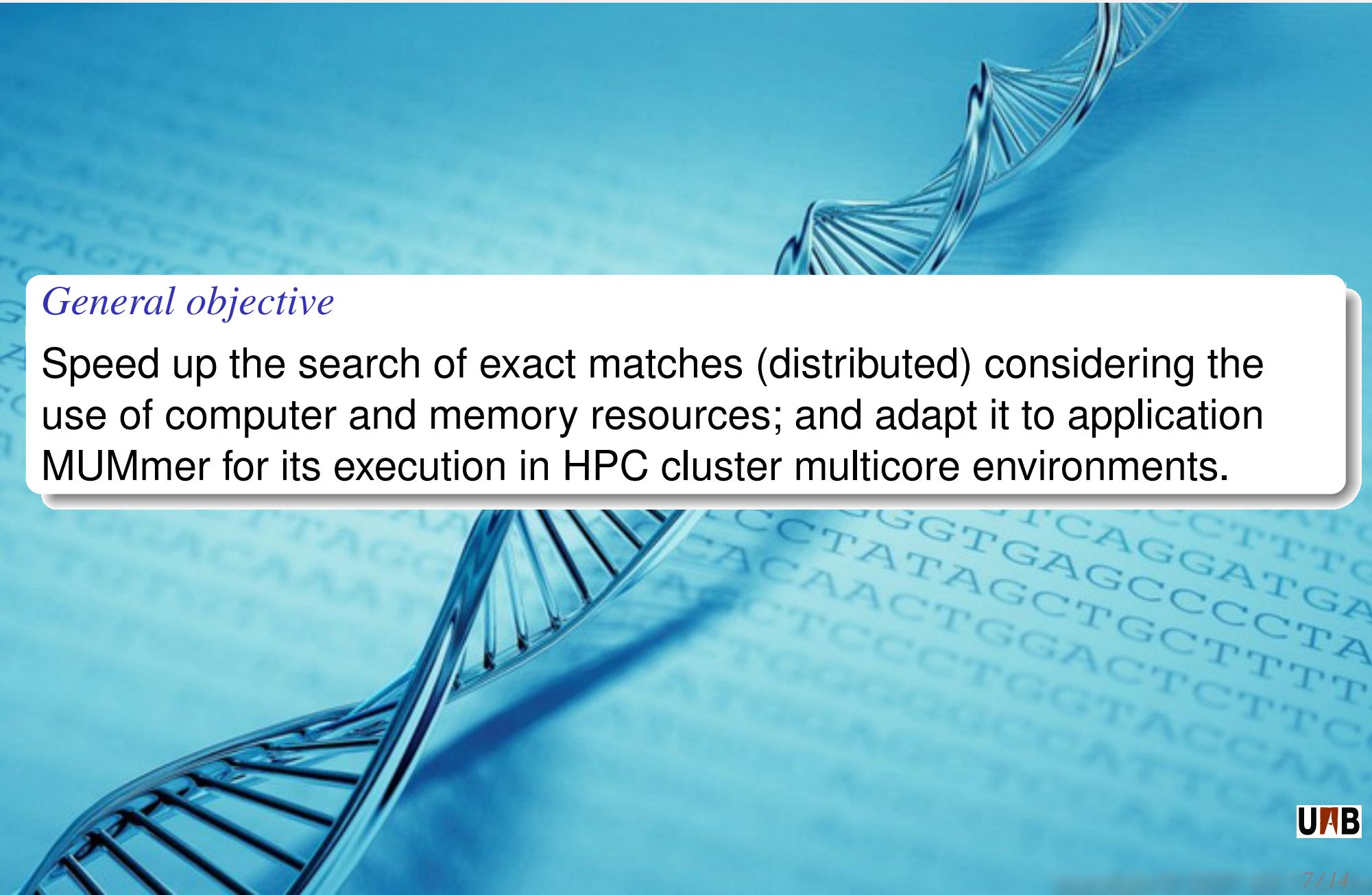
Traversal of suffix tree

Search of MUMs

1: **for each** Suffix in Q **do**
2: Compare suffix by traversing the suffix tree of R.
3: **if** $R[i-1] \neq Q[j-1]$ && node is a leaf **then**
4: The path of this traversal is a MUM-candidate.
5: **end if**
6: **end for**



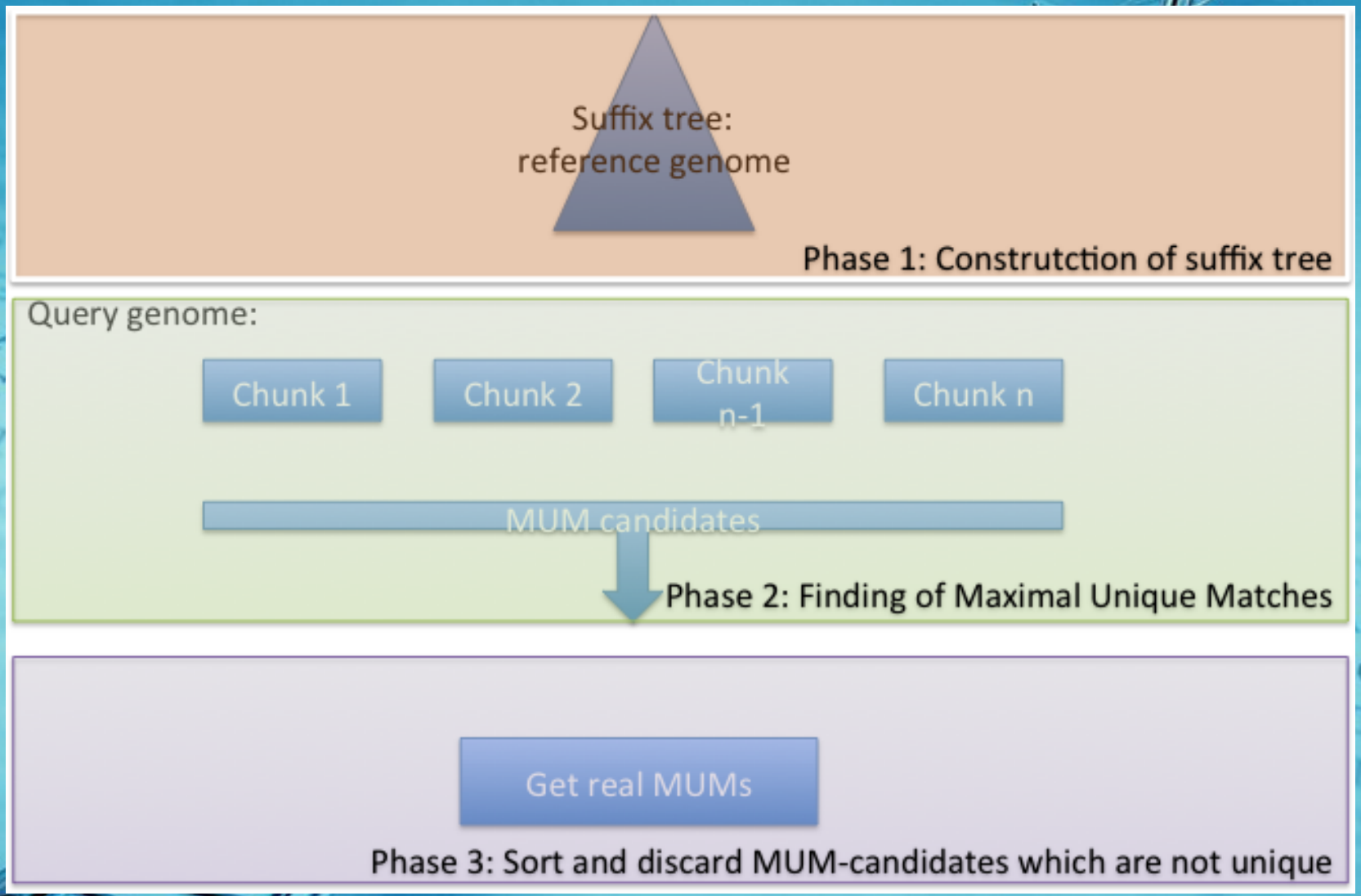
General objective



General objective

Speed up the search of exact matches (distributed) considering the use of computer and memory resources; and adapt it to application MUMmer for its execution in HPC cluster multicore environments.

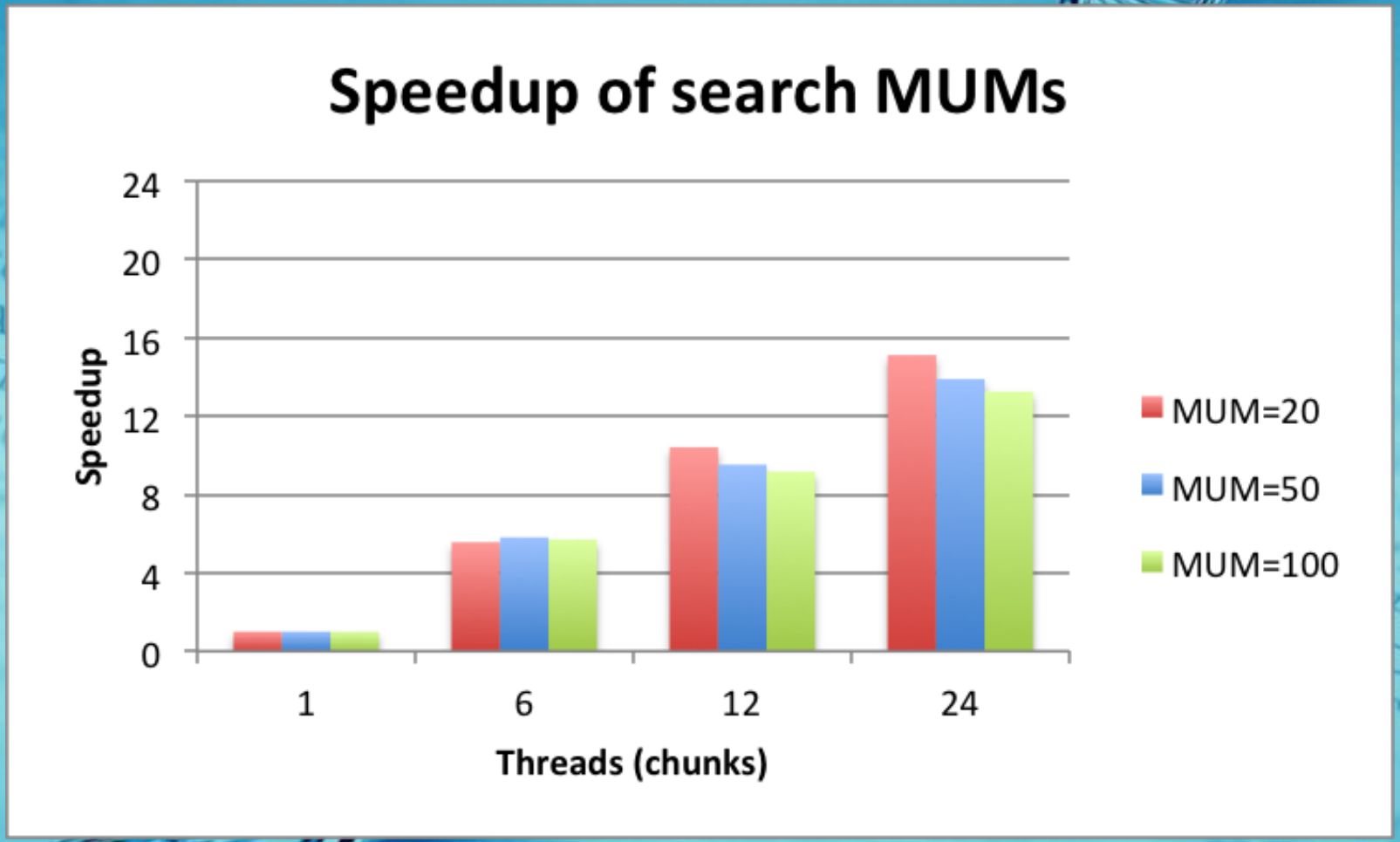
Search of MUMs in Multicore architectures



Parallel search of maximal unique matches

- Hardware:
  - 2 Processor Intel(R) Xeon(R) E5645 @ 2.4GHz of 6 cores each one, 32KB L1 cache, 256KB L2 and 12MB L3 shared cache per socket.
  - RAM: 96 GB
- Software:
  - Linux Kernel 2.6.32-220.el6.x86_64 #1 SMP
  - gcc 4.7.0 with OpenMP support
  - PAPI 5.0.1
- Genomes:
  - Reference: Human chromosome 21 single fasta file
  - Query: Mouse chromosome 16 single fasta file

Parallel search of maximal unique matches

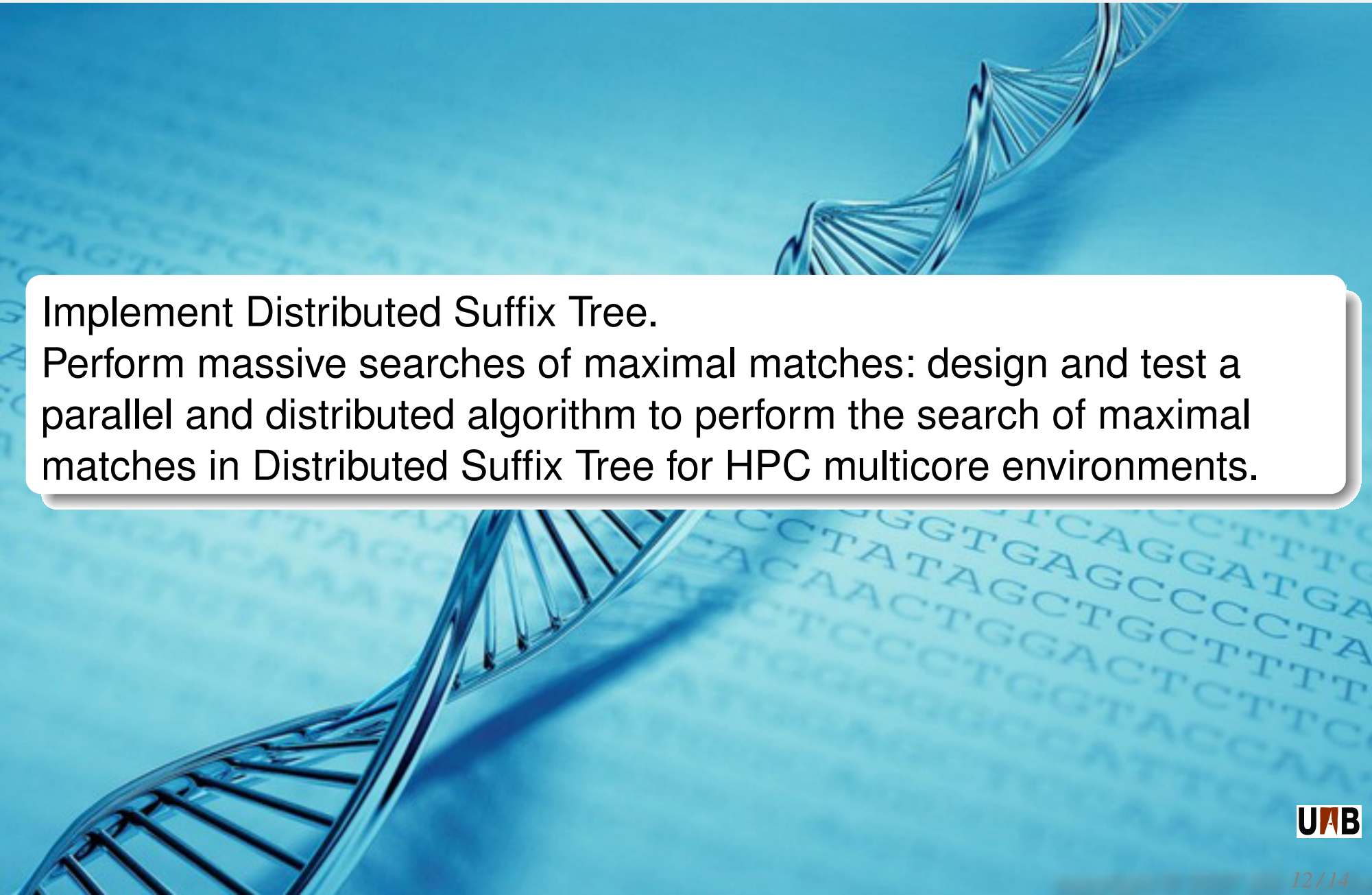


Conclusions

Current work

- Evaluation of performance to search MUMs of a query and reference genome in multi-core architectures with OpenMP.
- Results shows that the heaviest section of searching MUMs in a suffix tree is improved with the use of a multi-core architecture.
- Bottleneck is in suffix tree: traverse a suffix tree in multi-core architectures.

Future work



Implement Distributed Suffix Tree.
Perform massive searches of maximal matches: design and test a parallel and distributed algorithm to perform the search of maximal matches in Distributed Suffix Tree for HPC multicore environments.


Thanks!

Whole genome alignment in High Performance Computing environments

Julio César García Vizcaíno Directores: Antonio Espinosa, Juan Carlos Moure

Computer Architecture & Operating Systems Department
Universitat Autónoma de Barcelona

21 de enero de 2013