dries.decap@intec.ugent.be

Flexible motif discovery using feature selection trees: a high performance computing approach

Dries Decap, Bart Dhoedt, Jan Fostier and Yvan Saeys

Exhaustive motif discovery

Look for the motif that best distinguishes between a positive and negative group of sequences

Positive

AAGACCCGAGTAAACCCTGACCAAGTAGA
GGTGAGATAAACCCTAGACCAGTTGACCA
GTGAGATAAACCCTATACTCGTAGGGACG
TTGAGAGTTACCGAAACCCTACCCAGTTA
...

Negative

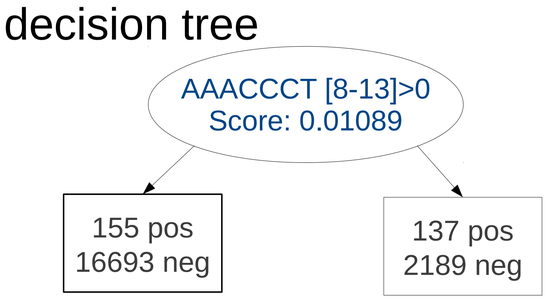
AAGAGCCCAGTAGAGATAGACCAAGTAGA
GGTGAGATAGACCGTAGACCAGTTGACCA
GTGAGATATACCCGGATACCGTAGGGACG
TGAGAGTTACCAGATATGAGACCAGTCTA
...

- Exhaustively loop over all motifs
- Calculate a score for each motif selecting the best range and threshold
- Score measures how much the collections are divided

motif	range	threshold	score
AAACCCTA	8-13	> 0	0.01047
AAACCCT	8-13	> 0	0.01089
AACCCTA	9-14	> 0	0.01076
...	...	...	...
AC	3-25	> 1	0.01044
...	...	...	...

The locations of motifs are saved in a generalized suffix tree to be easy accessible

- Select the motif (combination) that best distinguishes between the positive and negative sequences
- Apply this recursively to obtain a decision tree



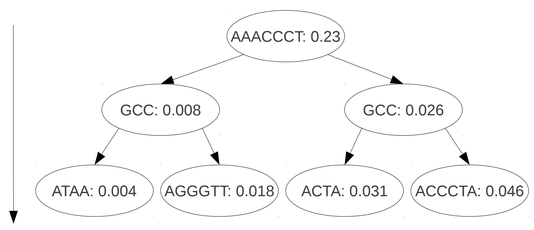
Sequences that **do not** contain motif "AAACCCT"

Sequences that contain motif "AAACCCT"

Adjustable Parameters

- Maximum motif length

- Depth of decision tree



- Use of IUPAC alphabet

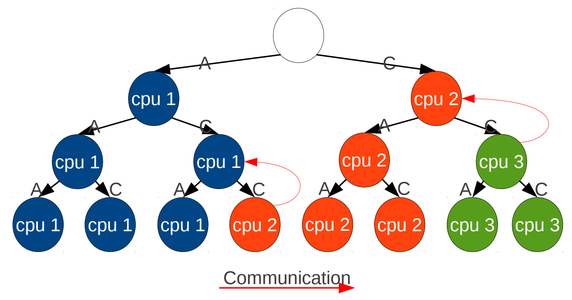
W	A or T
S	C or G
M	A or C
K	G or T
R	A or G
Y	C or T
N	A, C, G or T

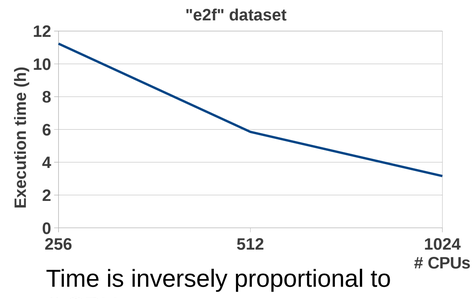
- Use of positional information

AAGACCCGAGTAAACCCTGACCAAGTAGA
GGTGAGATAAACCCTAGACCAGTTGACCA
GTGAGATAAACCCTATACTCGTAGGGACG
TTGAGAGTTACCGAAACCCTACCCAGTTA

High Performance Computing

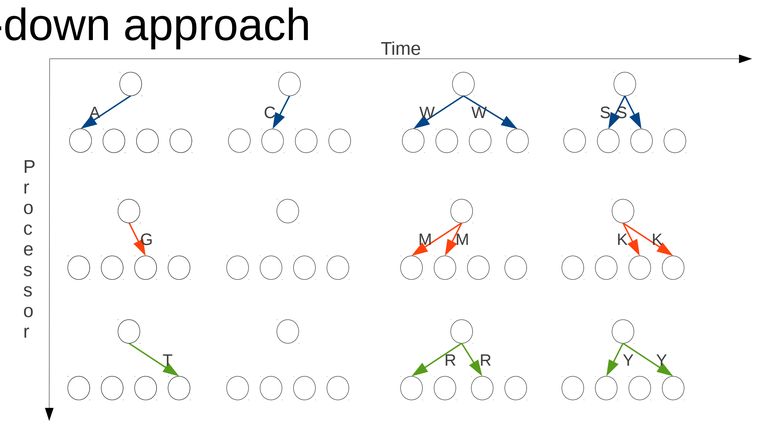
- For parallel execution MPI is used
- Suffix tree is used to partition motifs among nodes
- Bottom-up approach





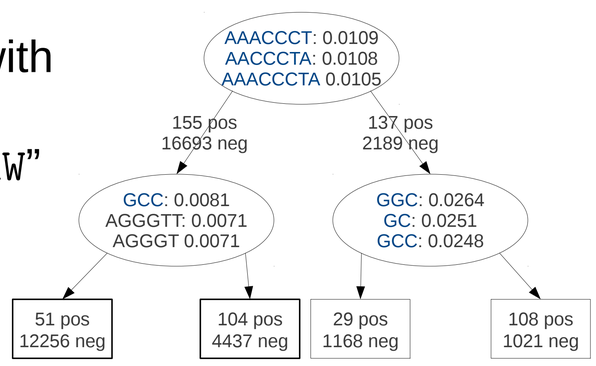
Time is inversely proportional to # CPUs

- OpenMP used for IUPAC characters
- Top-down approach



Results

- Benchmark "ribo" dataset with known motifs:
  "AAACCCTA" and "GGCCCAW"
- Top 3 motifs in motif tree
- Both motifs are found



- Benchmark dataset with known splice site:
  near position 200
- Important splice site features are found