UNIVERSITEIT GENT

FACULTY OF ENGINEERING AND ARCHITECTURE

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

Dries Decap, Bart Dhoedt, Jan Fostier and Yvan Saeys
dries.decap@intec.ugent.be

Exhaustive motif discovery

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

Positive

AAGACCCGAGTAAACCCTGACCAAGTAGA
GGTGAGATAAACCCTAGACCAGTTGACCA
GTGAGATAAACCCTATACTCGTAGGGACG
TTGAGAGTTACCGAAACCCTACCCAGTTA
….

Negative

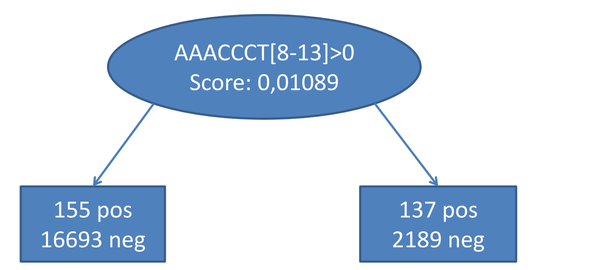
AAGAGCCCAGTAGAGATAGACCAAGTAGA
GGTGAGATAGACCGTAGACCAGTTGACCA
GTGAGATATACCCGGATACCGTAGGGACG
TGAGAGTTACCAGATATGAGACCAGTCTA
….

- Exhaustively loop over all motifs
- Calculate a score for each motif selecting the best range and threshold

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 based on the score
- Apply this recursively to obtain a decision tree (typical depth of 3)



Sequences that **do not** contain motif “AAACCCT”

Sequences that contain motif “AAACCCT”

OpenMP multithreading

- One thread does suffix tree traversal in parallel openMP region:

```
#pragma omp parallel
  #pragma omp single
  {
    //recursive traversal in suffix tree
  }
```

- Each branch assigned to different OpenMP threads with tasks:

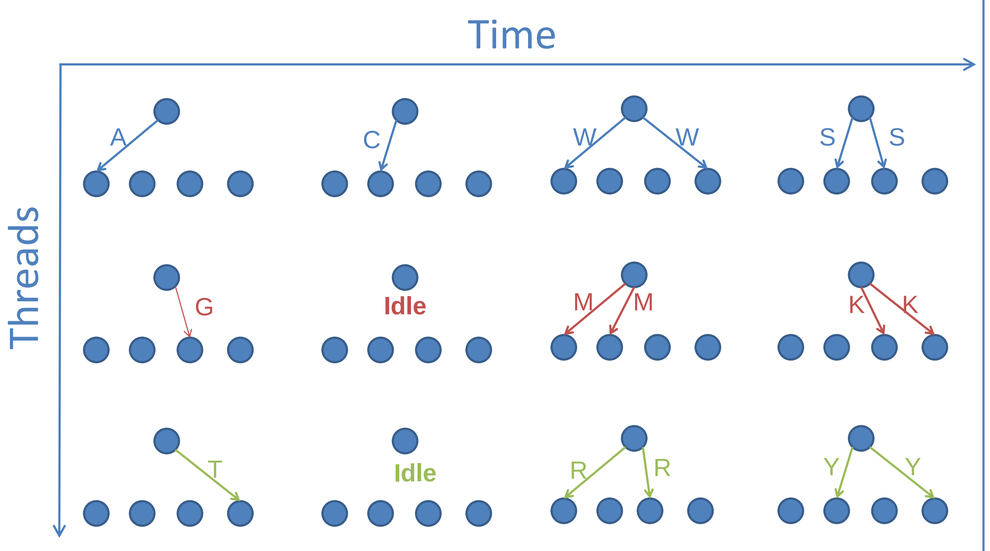
```
#pragma omp task
{
  //calculate score for motif
}
```

IUPAC support:

- Characters that represent two or more nucleotides
- Multiple branches for one thread
- Top-down approach
- Range omitted

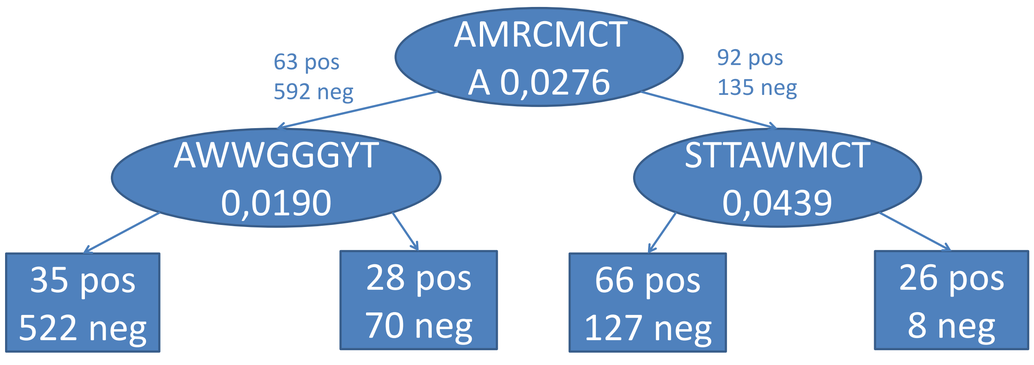
W	A or T	S	C or G
M	A or C	K	G or T
R	A or G	Y	C or T

Parallelization of suffix tree traversal:



Results

- Benchmark “ribo” dataset with known motifs:
  - “AAAACCCTA”
  - “GGCCCAW”
- Limit IUPAC characters to degeneracy 2
- Maximum motif length 8



- 1,3MB sequences takes 40 min (1 thread)
- Idle threads cause suboptimal speedup
- Speedup on personal computer: