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
AAGACCCGAGTAACACCTGACCAAGTAGA
GGTGAGATAAACCTAAGAAGGGTACCA
GGTGATATAACCCGATAGAAGGGTAGA
TTGAGATTACGAAACCTATCCAGTCTA

Negative
AAGACCCGAGTAAGAAGTAGA
GGTGAGATAAGAAGGGTACCA
GGTGATATAACCCGATAGAAGGGTAGA
TTGAGATTACGAAACCTATCCAGTCTA

- 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

<table>
<thead>
<tr>
<th>motif</th>
<th>range</th>
<th>threshold</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAACCCTA</td>
<td>8-13</td>
<td>&gt; 0</td>
<td>0.01247</td>
</tr>
<tr>
<td>AAACCCT</td>
<td>8-13</td>
<td>&gt; 0</td>
<td>0.01289</td>
</tr>
<tr>
<td>AAACCCTA</td>
<td>9-14</td>
<td>&gt; 0</td>
<td>0.01076</td>
</tr>
<tr>
<td>AAACCCTA</td>
<td></td>
<td>&gt; 1</td>
<td>0.02204</td>
</tr>
</tbody>
</table>

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

Adjustable Parameters

- Maximum motif length
- Depth of decision tree
- Use of IUPAC alphabet
- Use of positional information

AAAGACCCGAGTAACACCTGACCAAGTAGA
GGTGAGATAAACCTAAGAAGGGTACCA
GGTGATATAACCCGATAGAAGGGTAGA
TTGAGATTACGAAACCTATCCAGTCTA

High Performance Computing

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

- 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

Score measures how much the positive and negative sequences differ. 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” are marked.

AAACCCT

Sequences that contain motif “AAACCCT” are marked.

Score : 0.13

Time is inversely proportional to # CPUs

Benchmark “ribo” dataset

Score: 0.01089

“TTGAGAGTTACCG
GTGAGAT
GGTGAGAT
AAGACCCGAGT”

Score: 0.34

Update: 0.5354

Score: 0.53

Score: 0.01076

Score: 0.0105

Score: 0.01047