Efficient Extraction of Structured Motifs Using Box-links

Alexandra M. Carvalho

INESC-ID, Lisbon, Portugal

joint work with

Arlindo L. Oliveira

Ana T. Freitas

Marie-France Sagot

Background and related work

- Data structures:
  - Suffix tree

    [Ukkonen, *Algorithmica*, 1995]

    [McCreight, *Journal of the ACM*, 1976]

    [Weiner, *14th IEEE Symposium on Switching and Automata Theory*, 1973]
  - Factor tree

    [J. Allali and M.-F. Sagot, *Submitted for publication*, 2003]
- Algorithms:
  - Single motif extraction

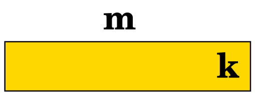
    [M.-F. Sagot, *3rd Latin American Symposium on Theoretical Informatics*, 1998]
  - Structured motif extraction
    - SMILE1 and SMILE2

      [L. Marsan and M.-F. Sagot, *Journal of Computational Biology*, 2000]

Structured motif

Definition. ***single motif***

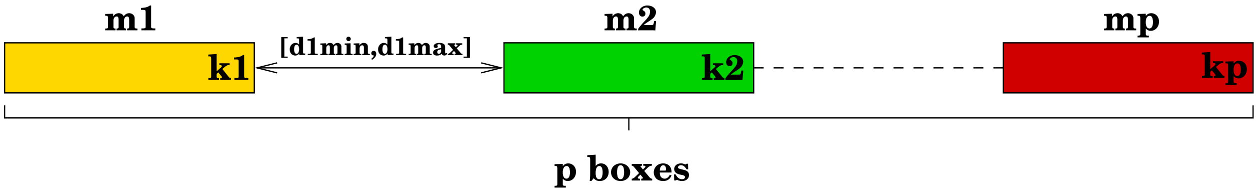
A single motif is a non-empty string over the DNA alphabet: A, C, G and T.



Definition. ***structured motif***

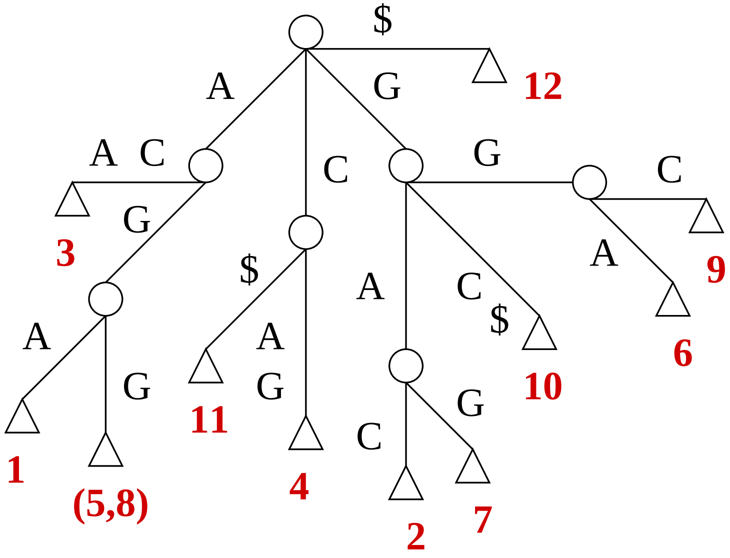
A structured motif is a pair (m, d) where:

- $m = (m_i)_{1 \le i \le p}$, denoting p single motifs (**boxes**)
- $d = (d_{\min_i}, d_{\max_i})_{1 \le i \le p-1}$, denoting $p - 1$ intervals of distance



Factor Tree

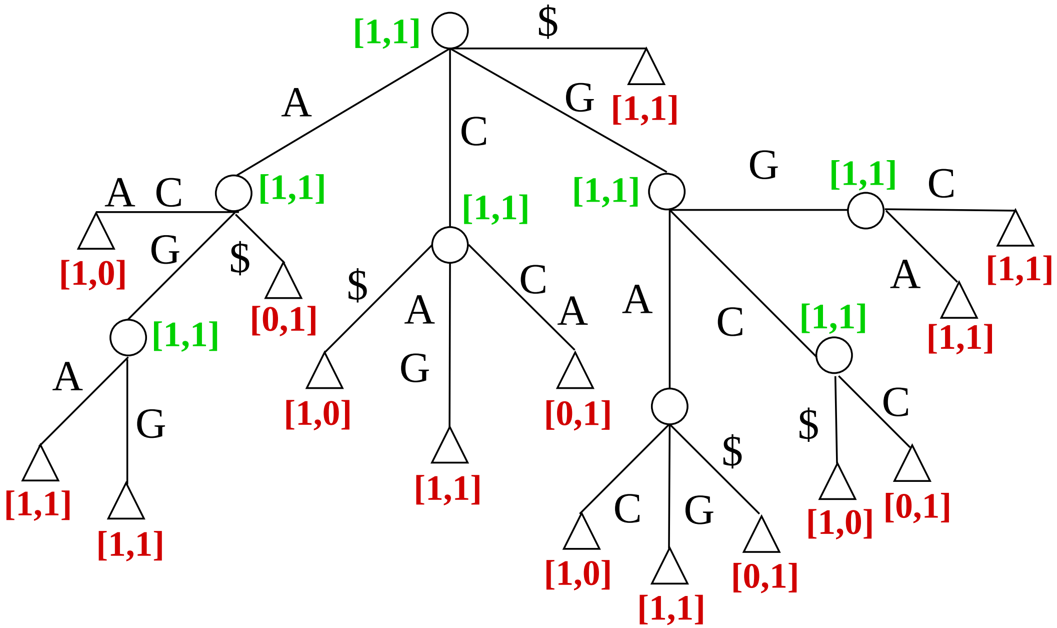
Factor tree for the string AGACAGGAGGC$



3-factor tree

Factor Tree

Generalized factor tree for the strings AGACAGGAGGC$ and AGAGGCCAGGA$



generalized 3-factor tree with $Colors$

Box-link data structure

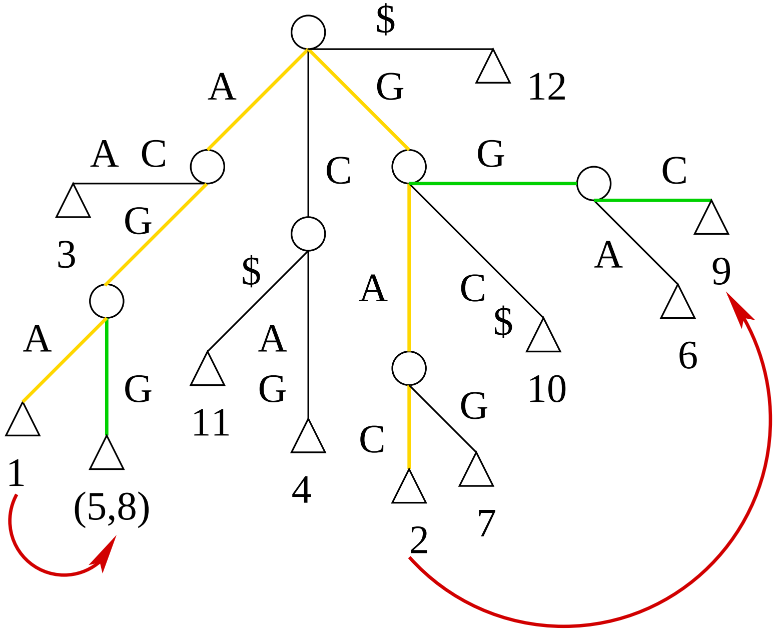
Factor tree for the string AGACAGGAGGC$

A G A C A G G A G G C

A G A   A G G

G A C   G G C

box-links for 2 boxes of size $k = 3$ distanced by $d = 4$



3-factor tree with box-links

Box-link data structure

Generalized factor tree for the strings AGACAGGAGGC$ and AGAGGCCAGGA$

A G A C A G G A G G C

A G A A G G

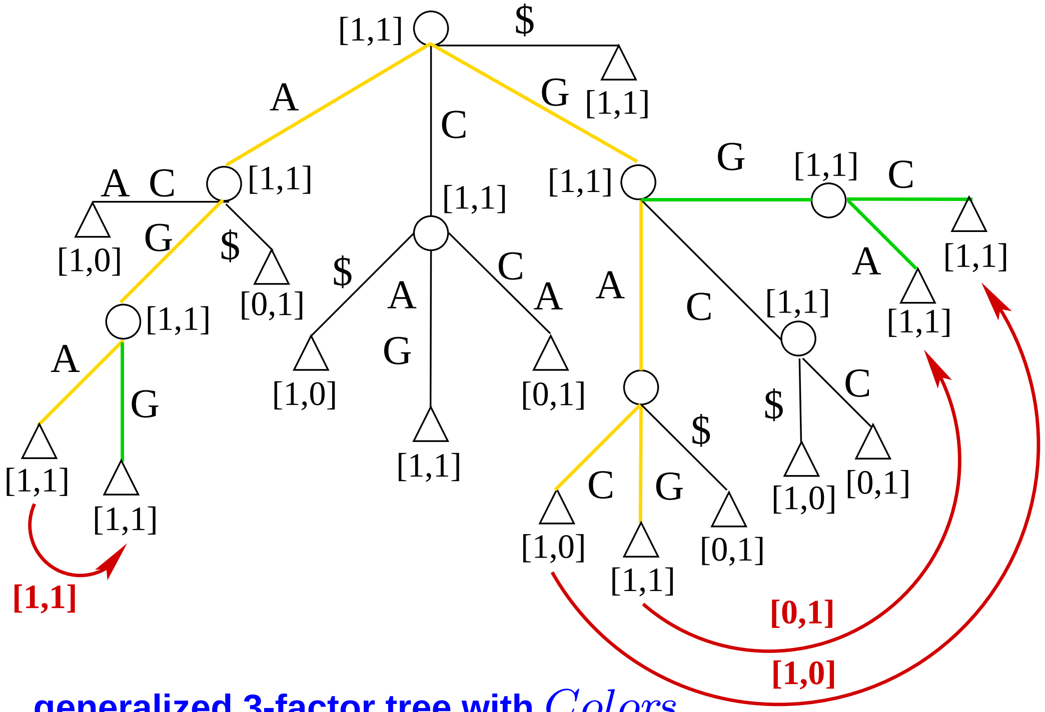
G A C G G C

A G A G G C C A G G A

A G A A G G

G A G G G A

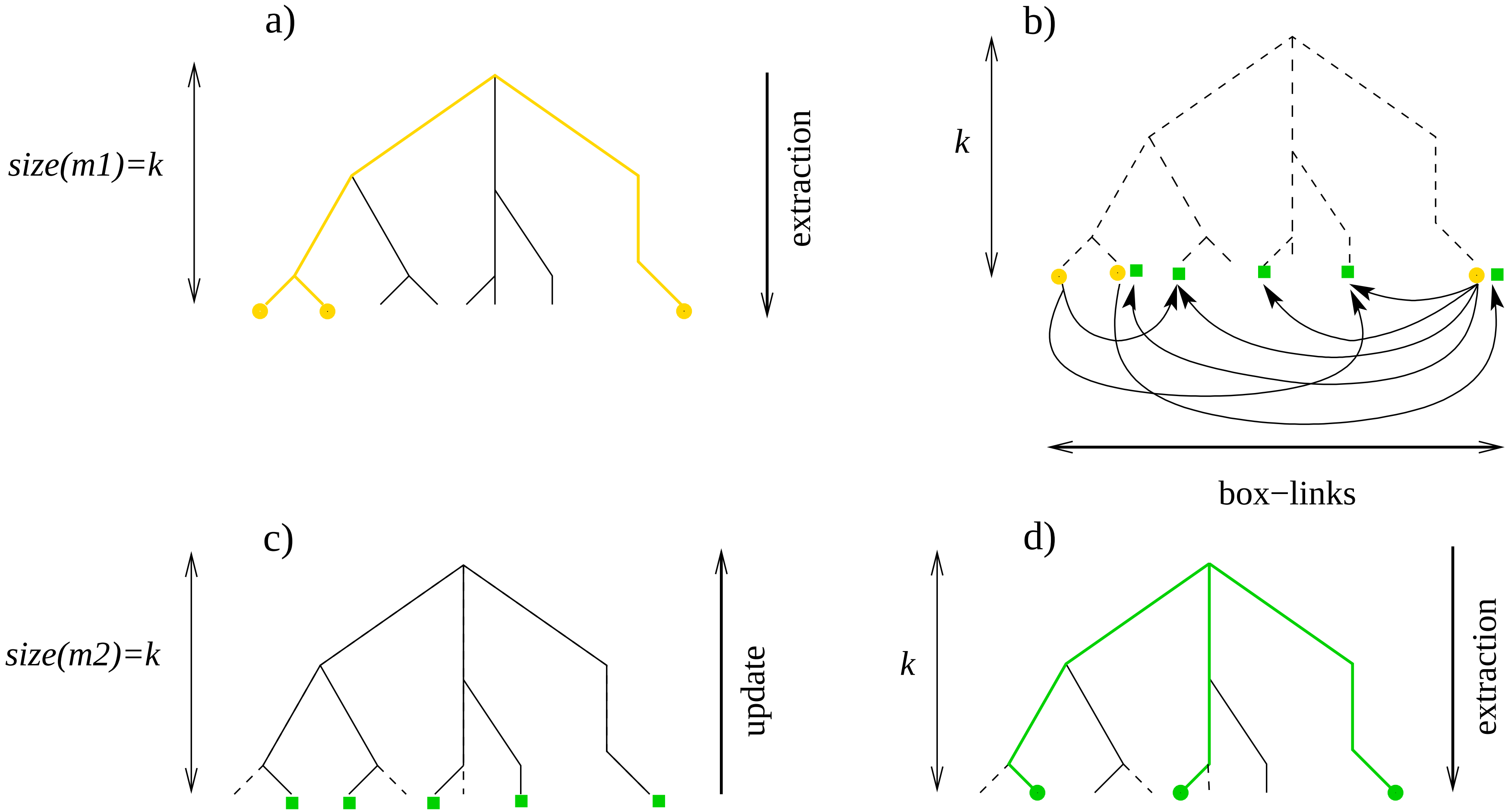
box-links for 2 boxes of size
$k = 3$ distanced by $d = 4$



generalized 3-factor tree with $Colors$

and box-links

Extraction of Structured Models: **RISO**



Experimental results

Extraction of the $CGGn_{11}CCG$ and $CGGAn_9TCCG$ motifs

68 genes that are known to be regulated by zinc cluster factors

# Errors		CPU Times (in seconds)		
Box 1	Box 2	SMILE1	SMILE2	RISO
1	1	44.72	7.4	**0.12**
2	2	1612.68	60.71	**12.12**

Extraction of the $TTGACAn_{17}TATAAT$ motif

1148 sequences from the *E. coli* genome

# Errors		CPU Times (in seconds)		
Box 1	Box 2	SMILE1	SMILE2	RISO
1	2	1429.81	1983.41	**942.42**