Extraction of structured motifs

Alexandra Carvalho

INESC-ID

Plan of the talk

- Single motifs extraction

  [M.-F. Sagot, *Latin*, 1998]

- Structured motifs extraction
  - SMILE

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

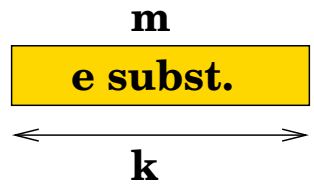
    [A. Carvalho, A. Freitas, A. Oliveira and M.-F. Sagot, *submitted*, 2004]

- Ongoing and future work

Structured model

Definition. ***single model***

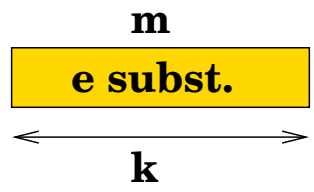
A single model is a string over the DNA alphabet: A, C, G and T.



Structured model

Definition. ***single model***

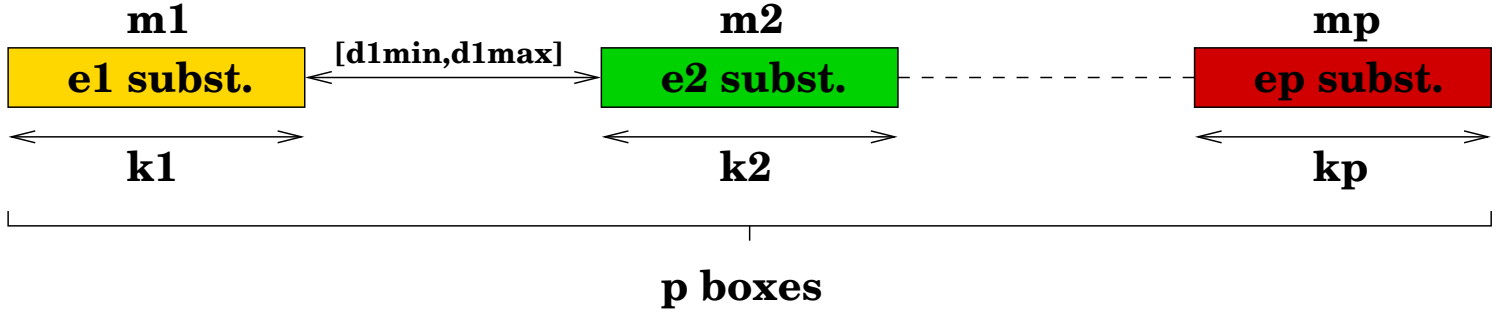
A single model is a string over the DNA alphabet: A, C, G and T.



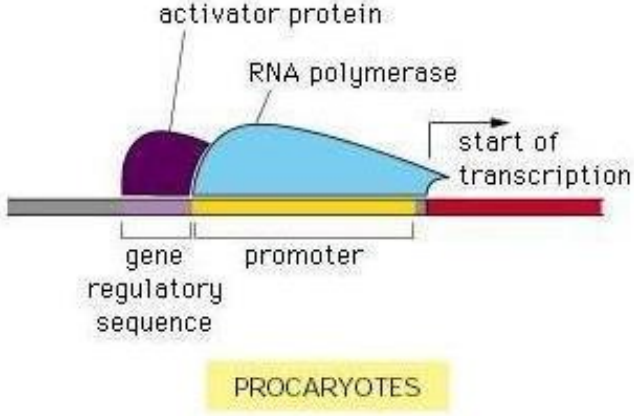
Definition. ***structured model***

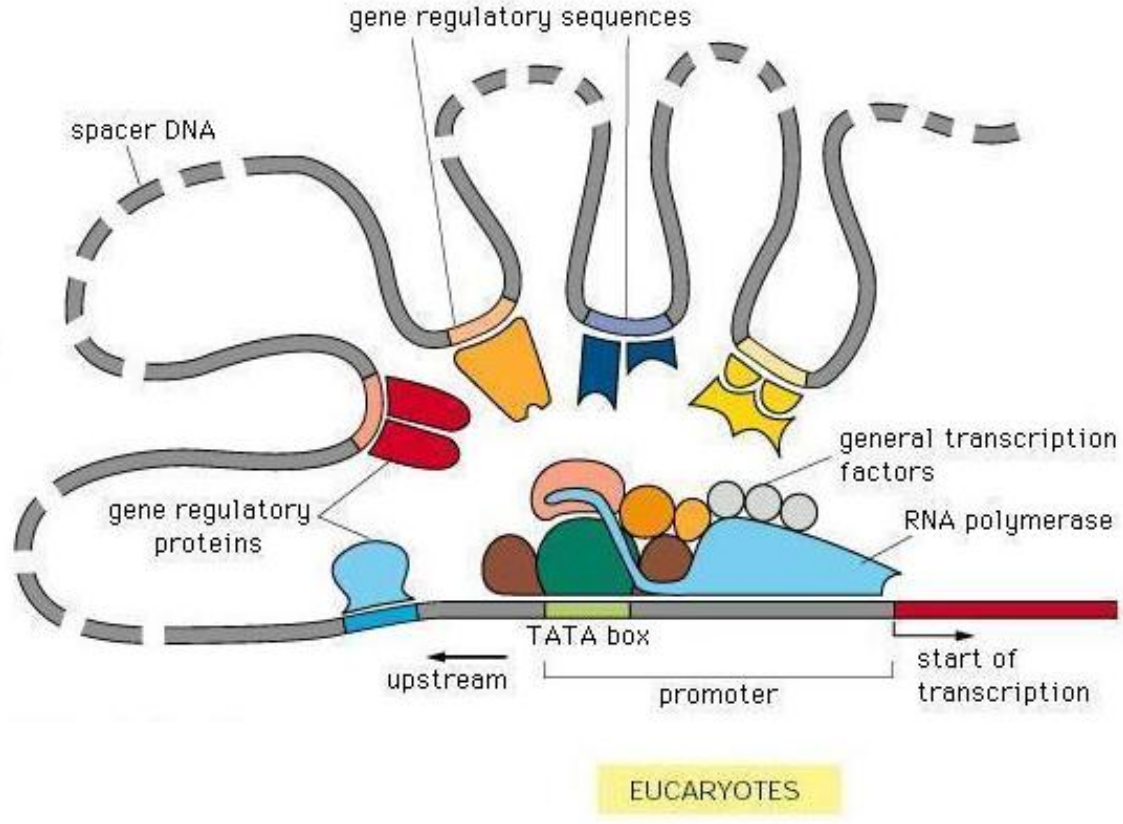
A structured model is a pair (m, d) where:

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



Promoter and Regulatory Sequences





Structured motif

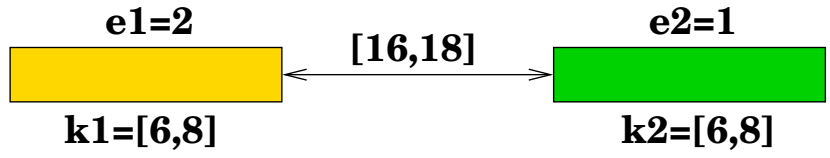
Definition. ***valid model, quorum***

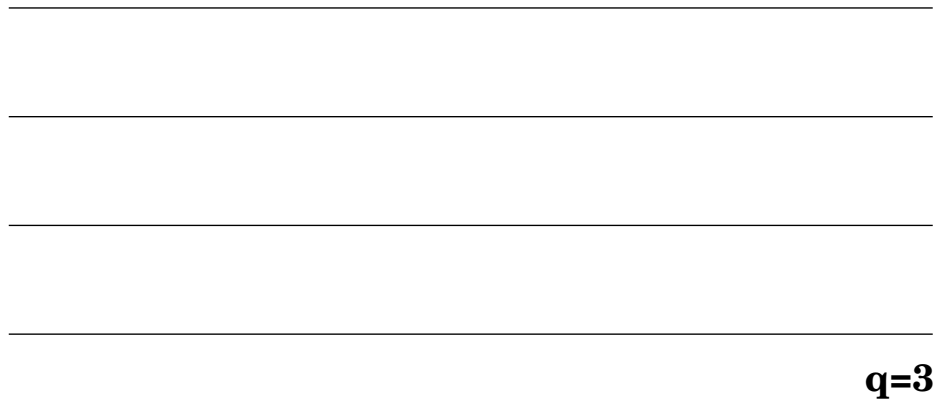
A model is valid if occurs in at least q input sequences, where q is called the quorum.

Structured motif

Definition. ***valid model, quorum***

A model is valid if occurs in at least q input sequences, where q is called the quorum.

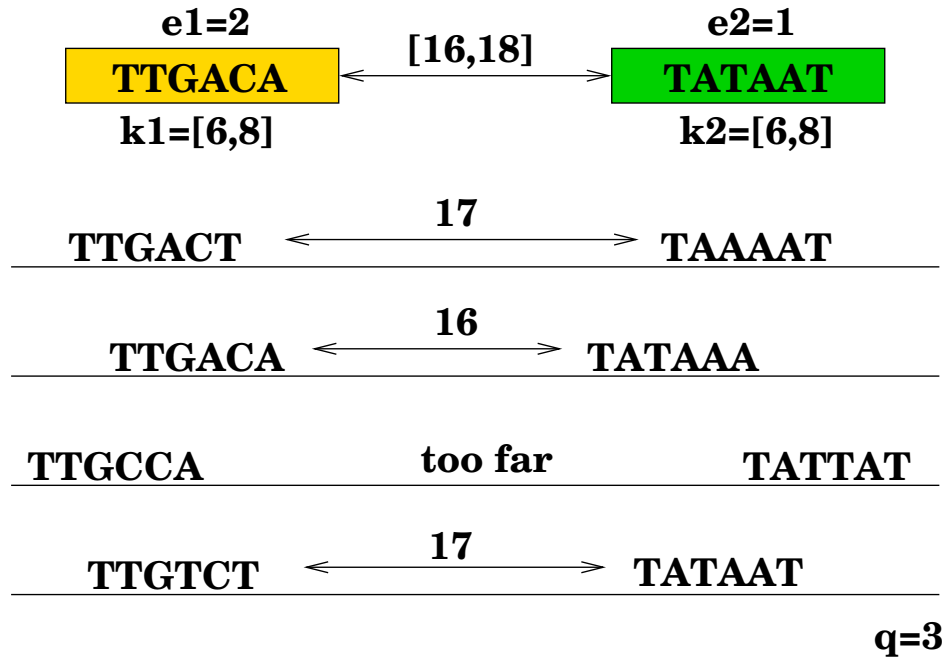




Structured motif

Definition. ***valid model, quorum***

A model is valid if occurs in at least q input sequences, where q is called the quorum.



Input sequences

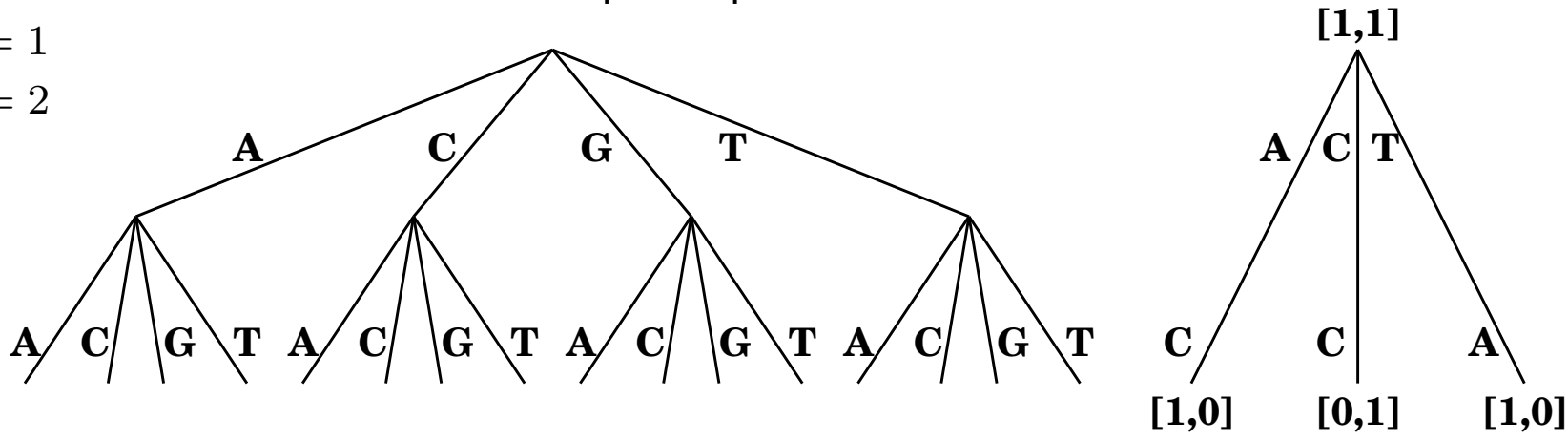
```
>strand + guaB inositol-monophosphate dehydrogenas
CTTTCCGTTATCTAAATATTTCAACTCTTTCCCGCTTCCTTGACATGCTCTTGGCTAGTTGATAATCT
ACATATAATATTTTGCCGAAAA
>strand - yaaC yaaC
TTTTCGGCAAAATATTATATGTAGATTATCAACTAGCCAAGAGCATGTCAAGGAAGCGGGAAAGAGTT
GAAATATTTAGATAACGGAAAG
>strand + yaaJ similar to hypothetical proteins
CCGTTTCAGTTATAGTTAATATGTAGCCTTTTTAGGCAATGAAAAAACTTTGAAA
>strand - yaaI similar to isochorismatase
TTTCAAAGTTTTTTCATTGCCTAAAAAGGCTACATATTAACTATAACTGAAACGG
>strand + metS methionyl-tRNA synthetase
ATTTTATAAATATTTAATAAAGCTATTATCCTACTAAAAATCCTTTTAAATCAAGACTTTCGAACCAA
AGTTTTTTATTTCATTTGATTATATACGACAAAATTCGACACGAACAGACTTTTTTTATTTTCATTAA
AGATTTTTAATTTTAATTATTCTTTTTCAGGGCGTATGTATATATTCTTGATCTTAAAGGCTAAGATG
GTATCATAGATAAAGGATAAATATAAATAATATTCATATATGATTTGCACTTATCGCCGCTCTCGTCC
TTTGGGCGGGAGCTTTTTGACATTCTGA
```

Extraction of Single Models

M.-F. Sagot, *Latin*, 1998

$k = 2$
$e = 1$
$q = 2$

Input sequences: TAC and CCC

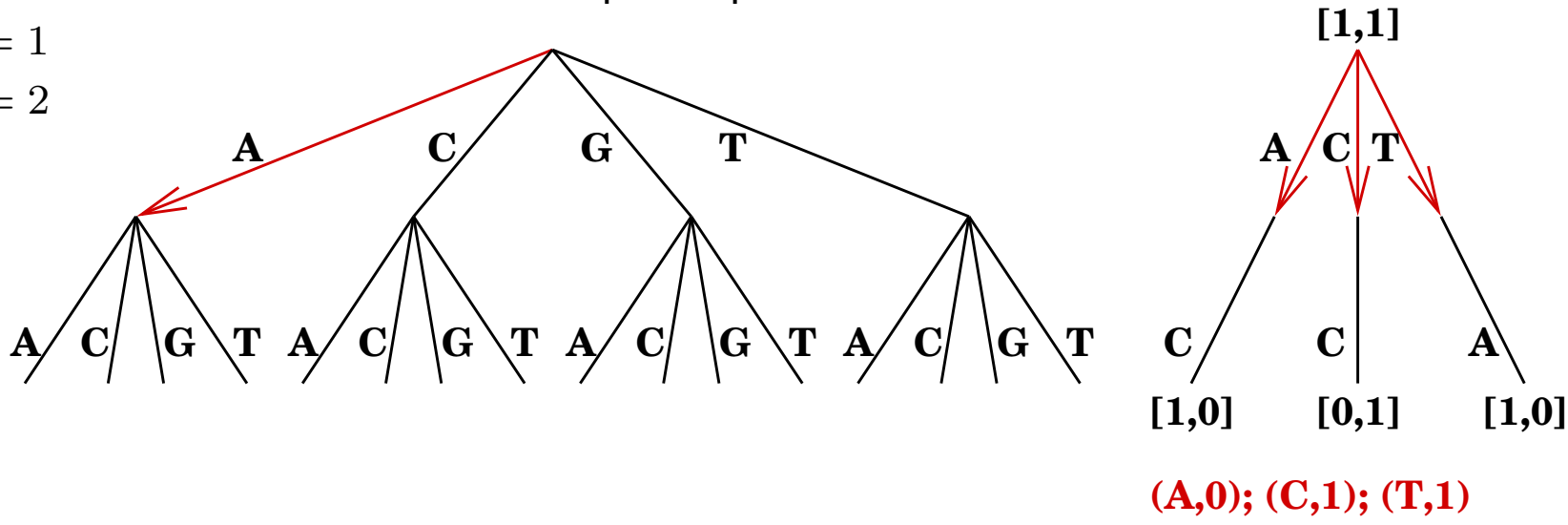


Extraction of Single Models

M.-F. Sagot, *Latin*, 1998

$k = 2$
$e = 1$
$q = 2$

Input sequences: TAC and CCC

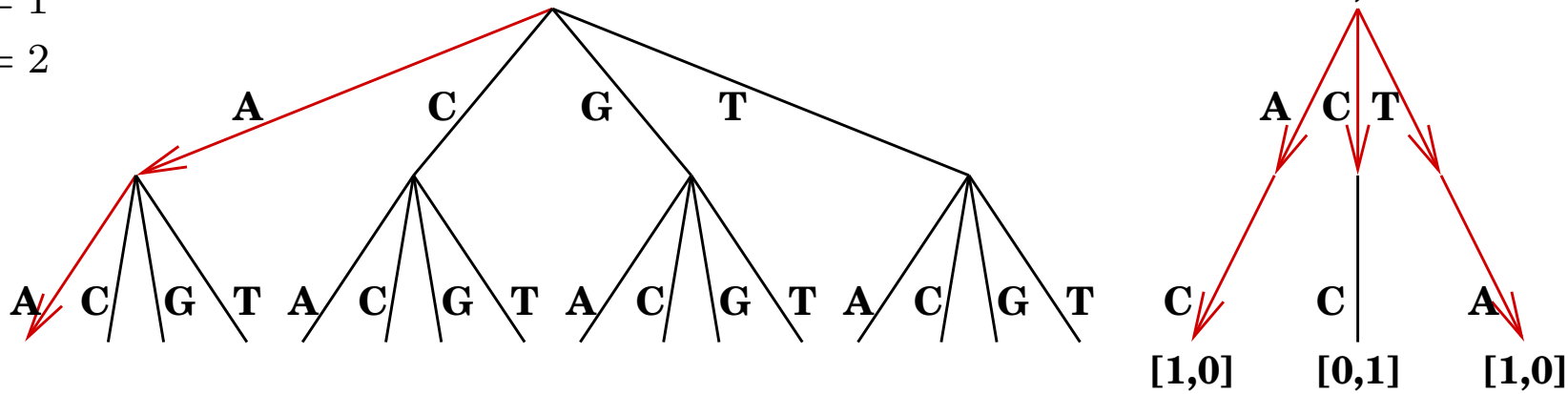


Extraction of Single Models

M.-F. Sagot, *Latin*, 1998

$k = 2$
$e = 1$
$q = 2$

Input sequences: TAC and CCC



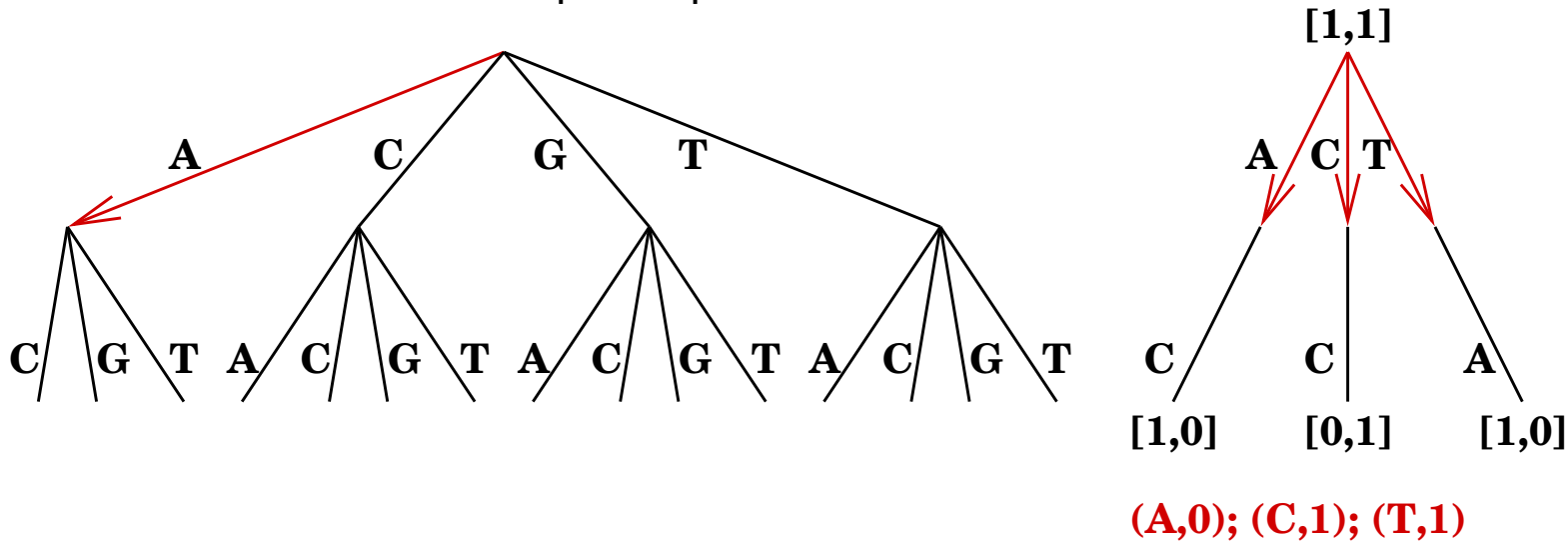
(A,0); (C,1); (T,1)
(AC,1); (TA,1)

Extraction of Single Models

M.-F. Sagot, *Latin*, 1998

$k = 2$
$e = 1$
$q = 2$

Input sequences: TAC and CCC

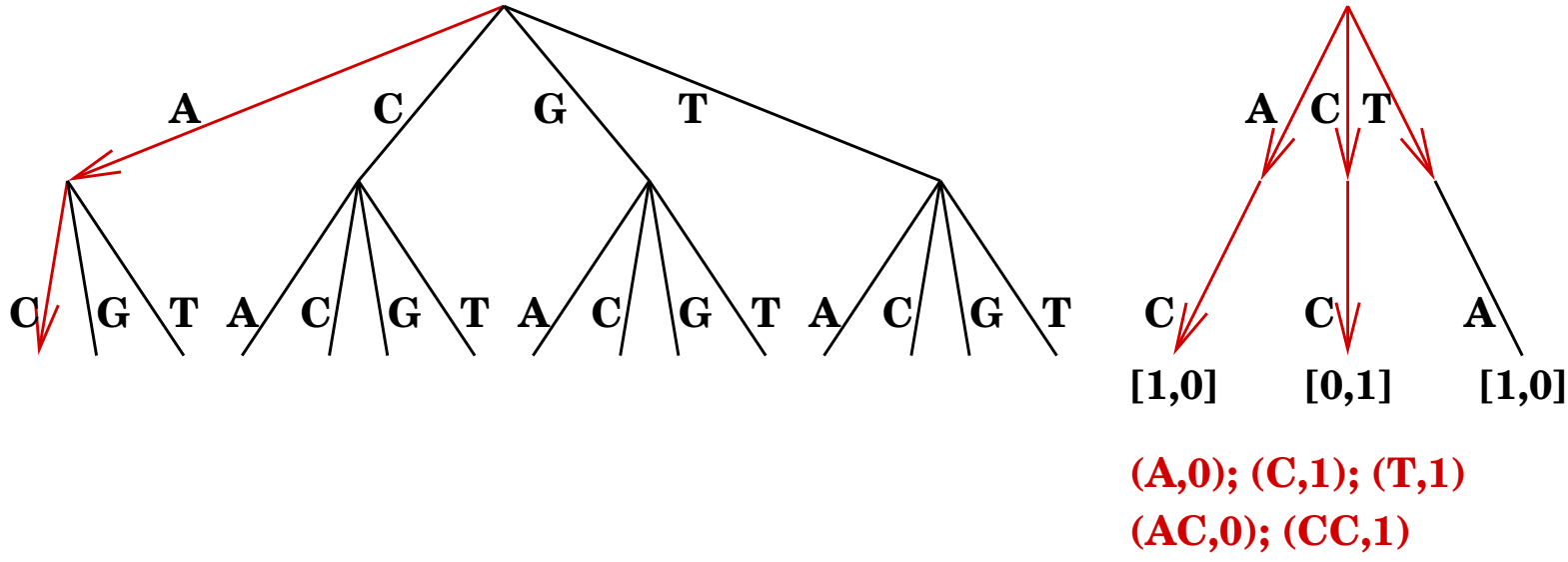


Extraction of Single Models

M.-F. Sagot, *Latin*, 1998

$k = 2$
$e = 1$
$q = 2$

Input sequences: TAC and CCC

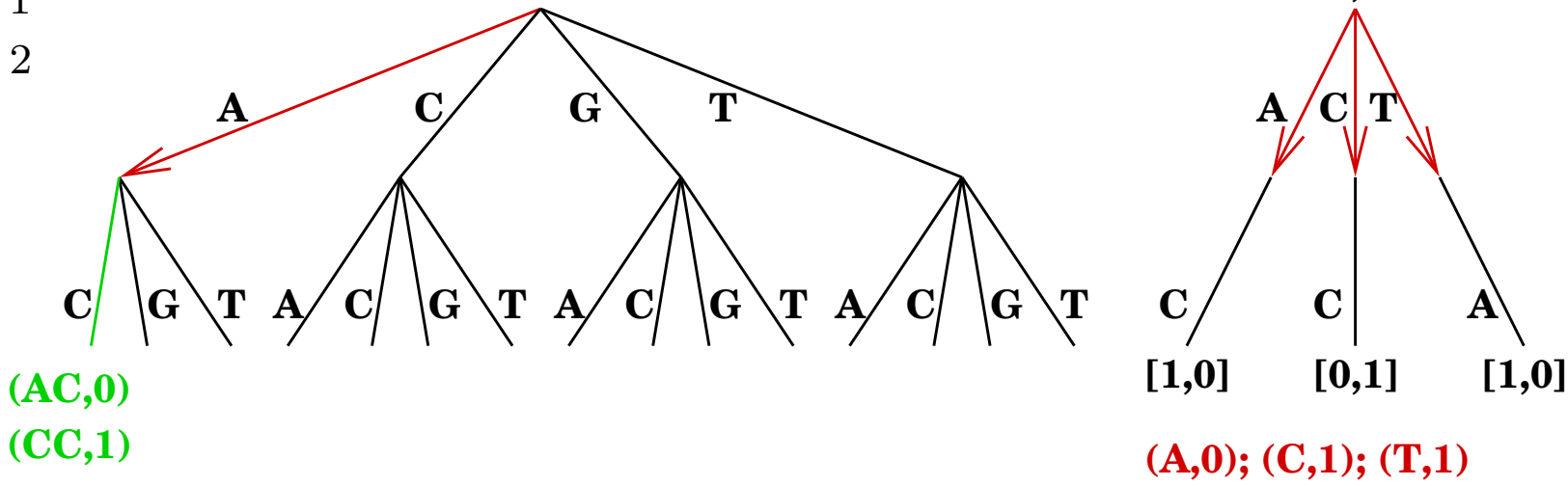


Extraction of Single Models

M.-F. Sagot, *Latin*, 1998

$k = 2$
$e = 1$
$q = 2$

Input sequences: TAC and CCC

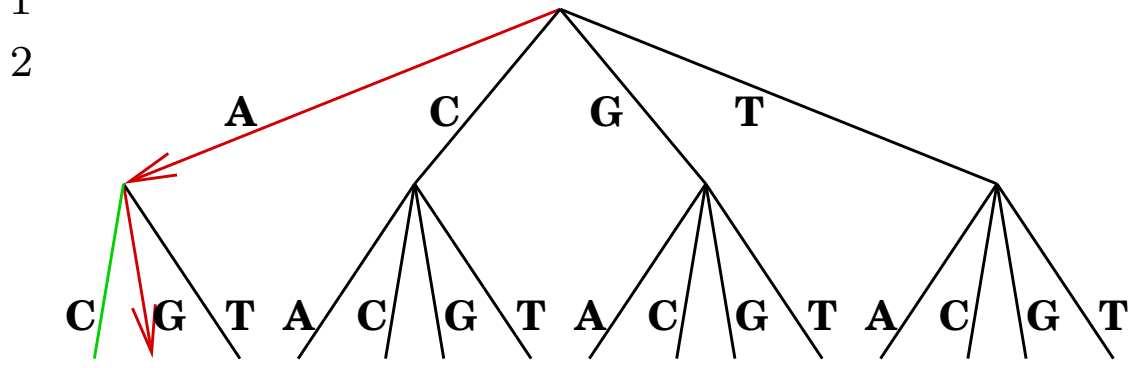


Extraction of Single Models

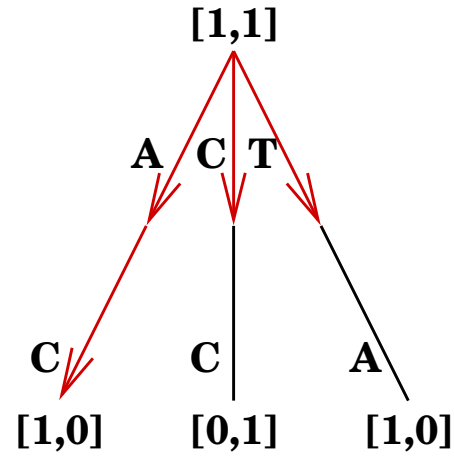
M.-F. Sagot, *Latin*, 1998

$k = 2$
$e = 1$
$q = 2$

Input sequences: TAC and CCC



(AC,0)
(CC,1)



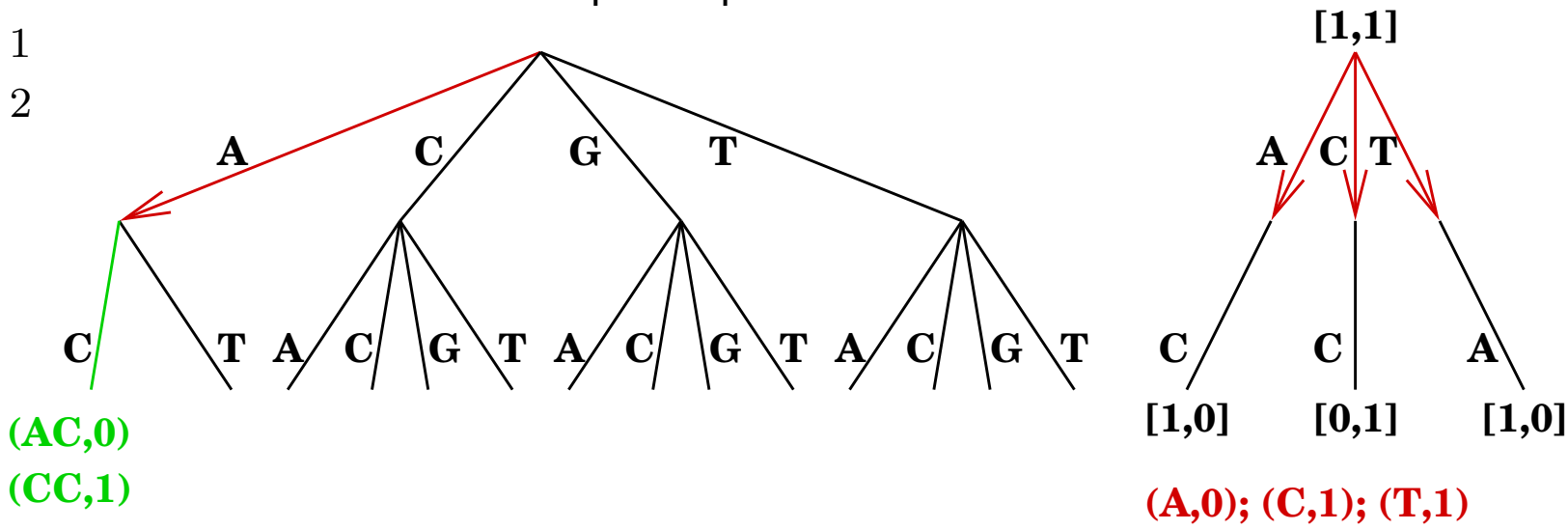
(A,0); (C,1); (T,1)
(AC,1)

Extraction of Single Models

M.-F. Sagot, *Latin*, 1998

$k = 2$
$e = 1$
$q = 2$

Input sequences: TAC and CCC

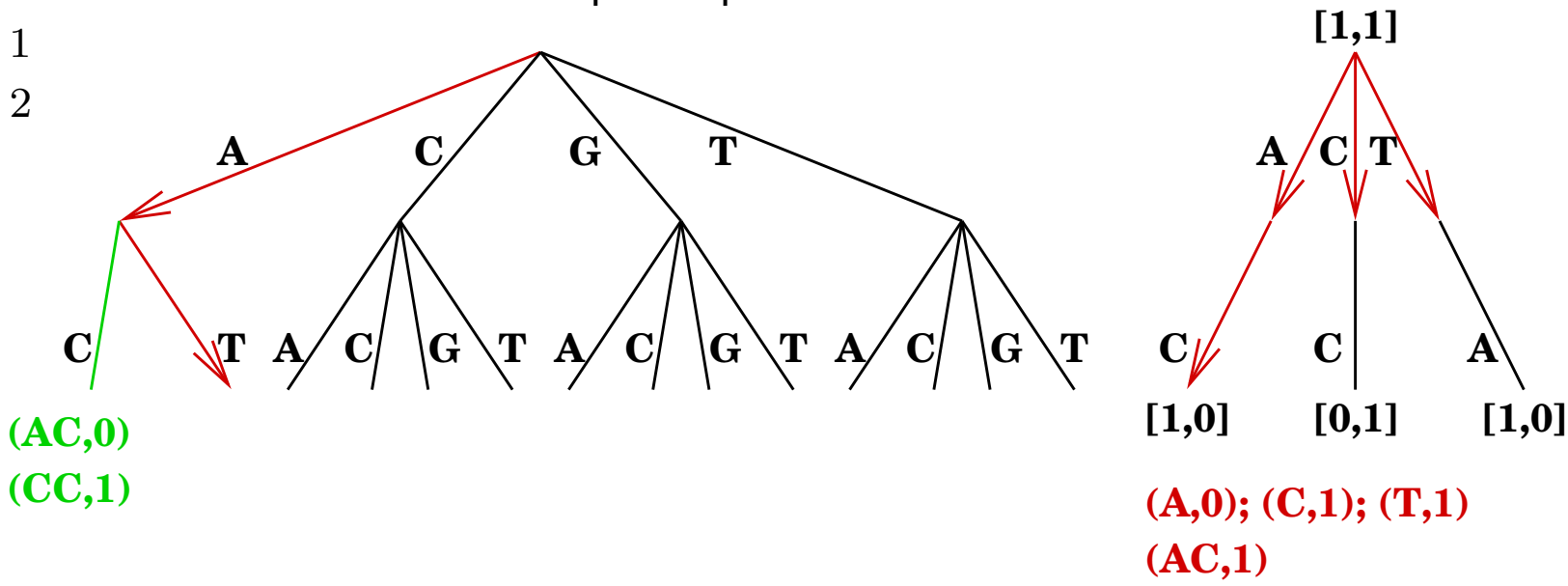


Extraction of Single Models

M.-F. Sagot, *Latin*, 1998

$k = 2$
$e = 1$
$q = 2$

Input sequences: TAC and CCC

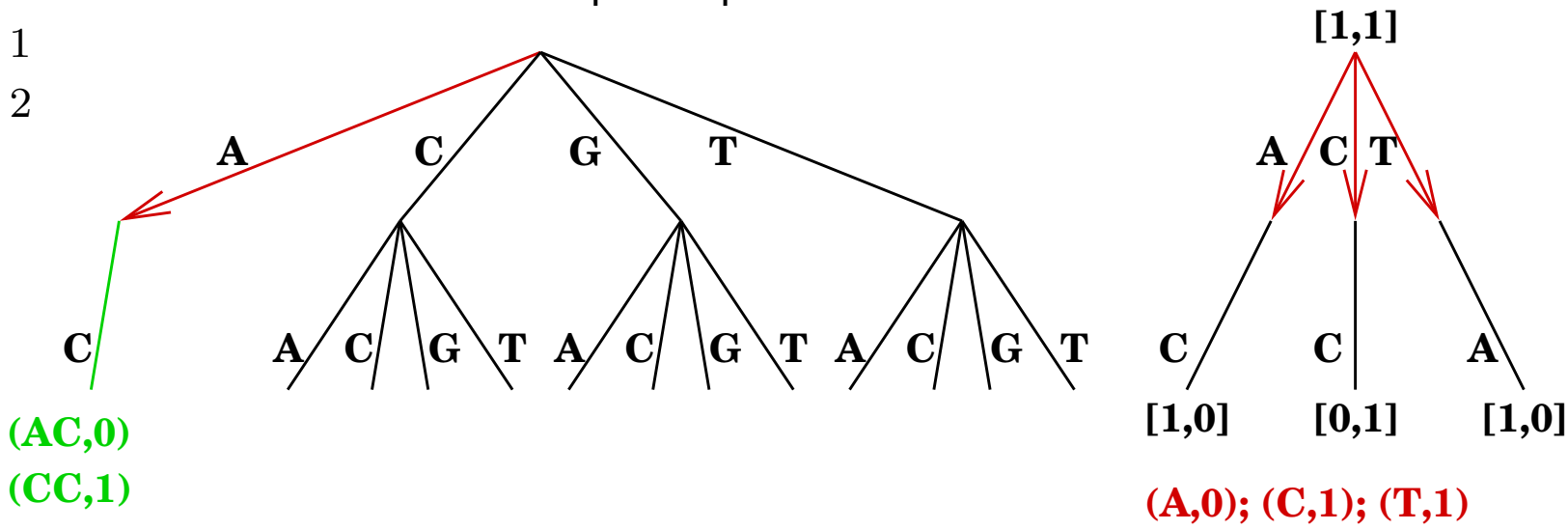


Extraction of Single Models

M.-F. Sagot, *Latin*, 1998

$k = 2$
$e = 1$
$q = 2$

Input sequences: TAC and CCC

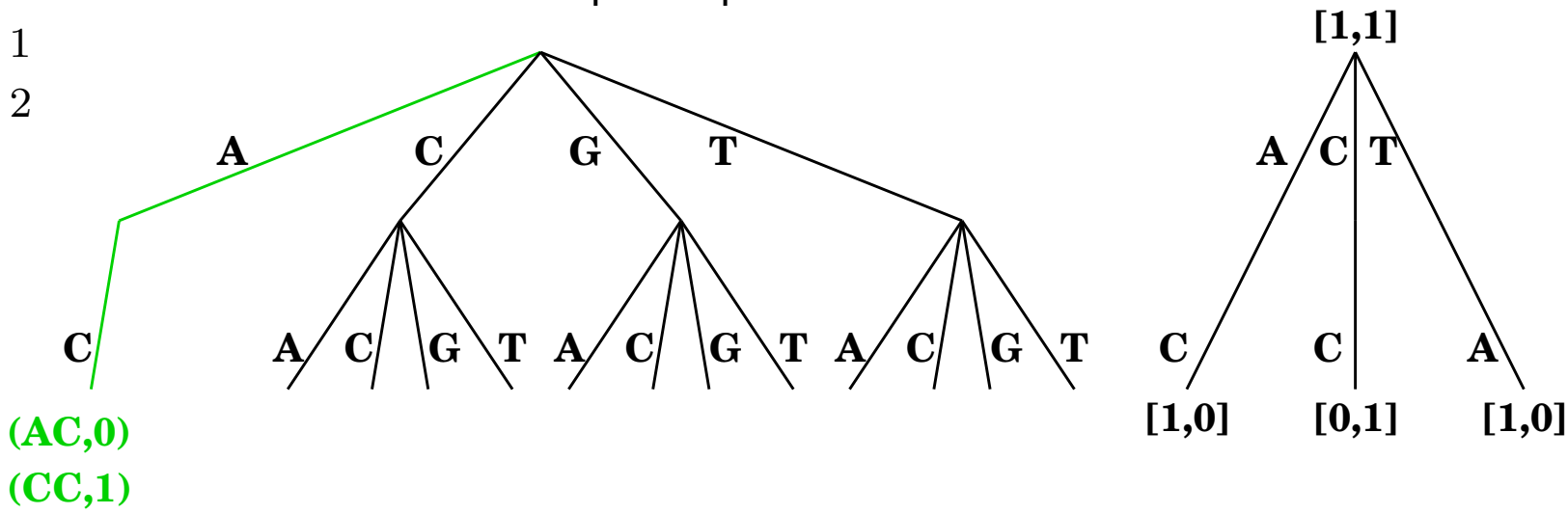


Extraction of Single Models

M.-F. Sagot, *Latin*, 1998

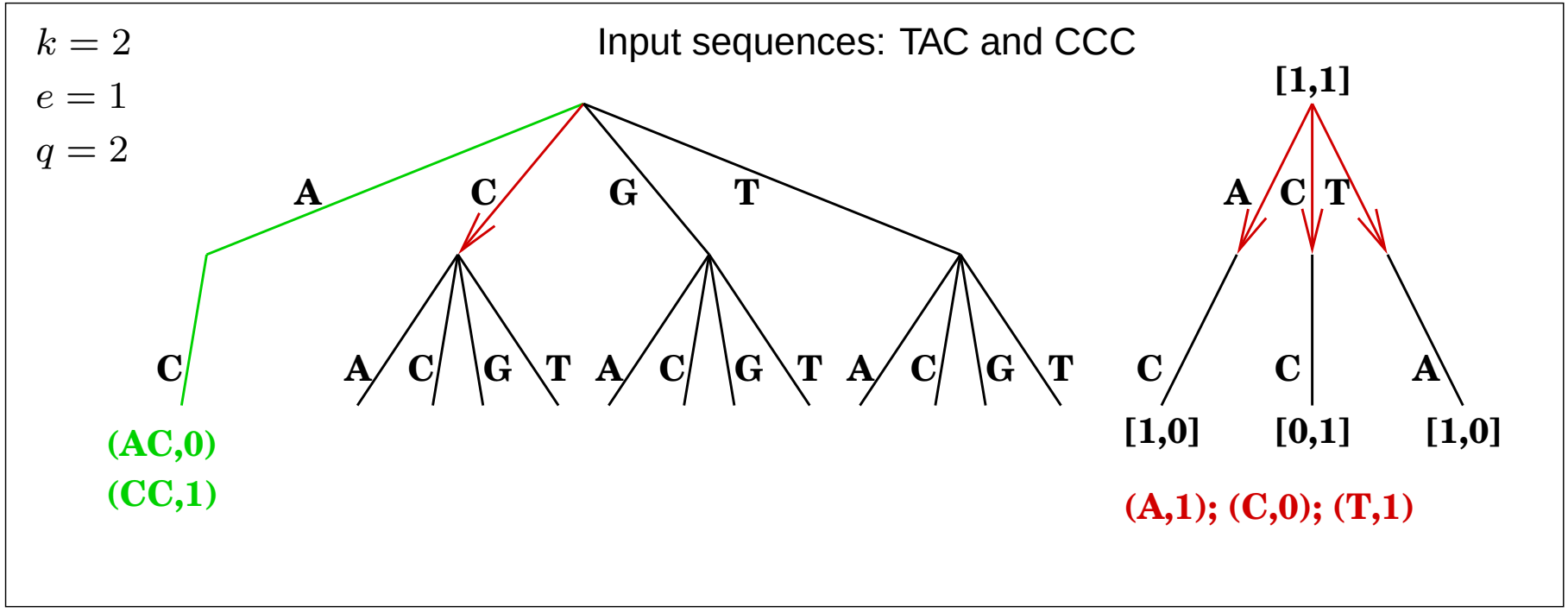
$k = 2$
$e = 1$
$q = 2$

Input sequences: TAC and CCC



Extraction of Single Models

M.-F. Sagot, *Latin*, 1998

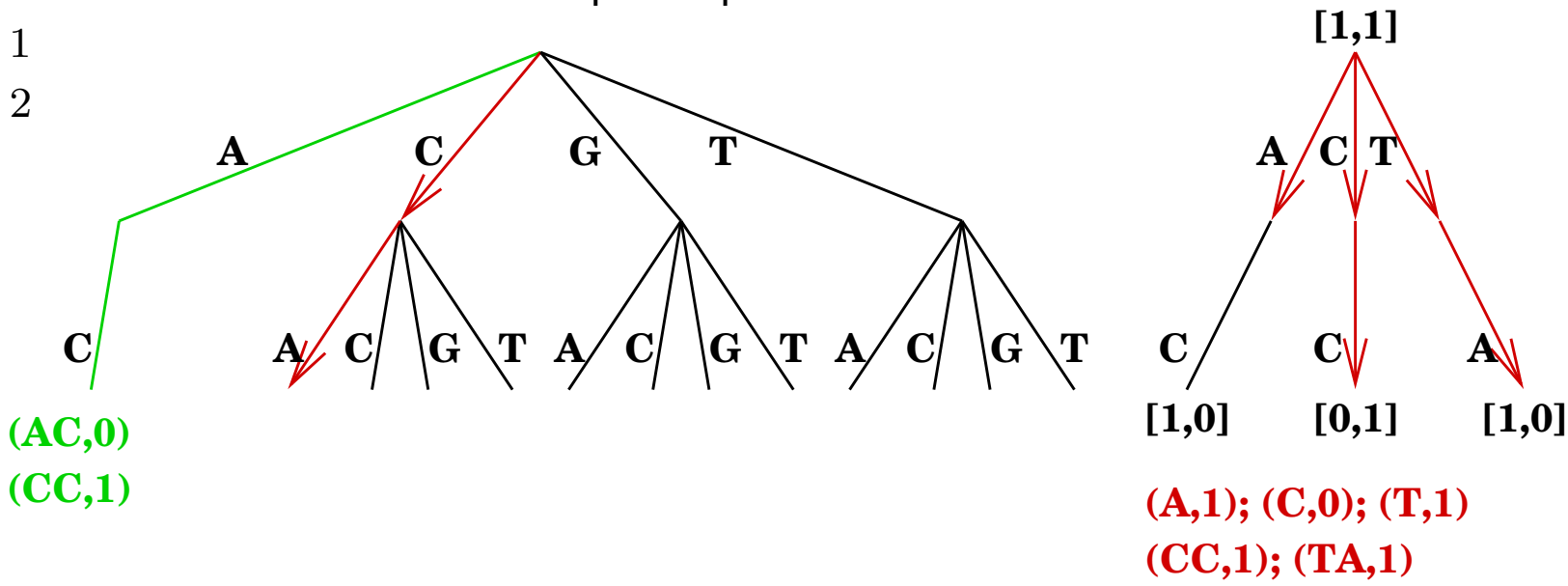


Extraction of Single Models

M.-F. Sagot, *Latin*, 1998

$k = 2$
$e = 1$
$q = 2$

Input sequences: TAC and CCC

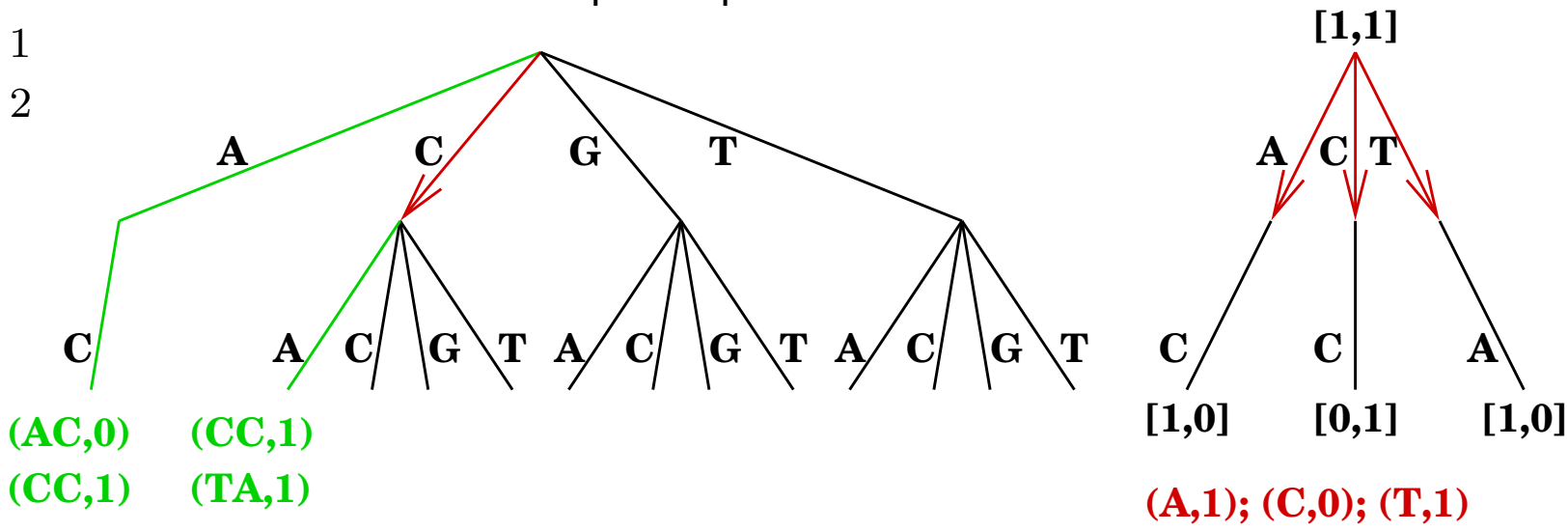


Extraction of Single Models

M.-F. Sagot, *Latin*, 1998

$k = 2$
$e = 1$
$q = 2$

Input sequences: TAC and CCC

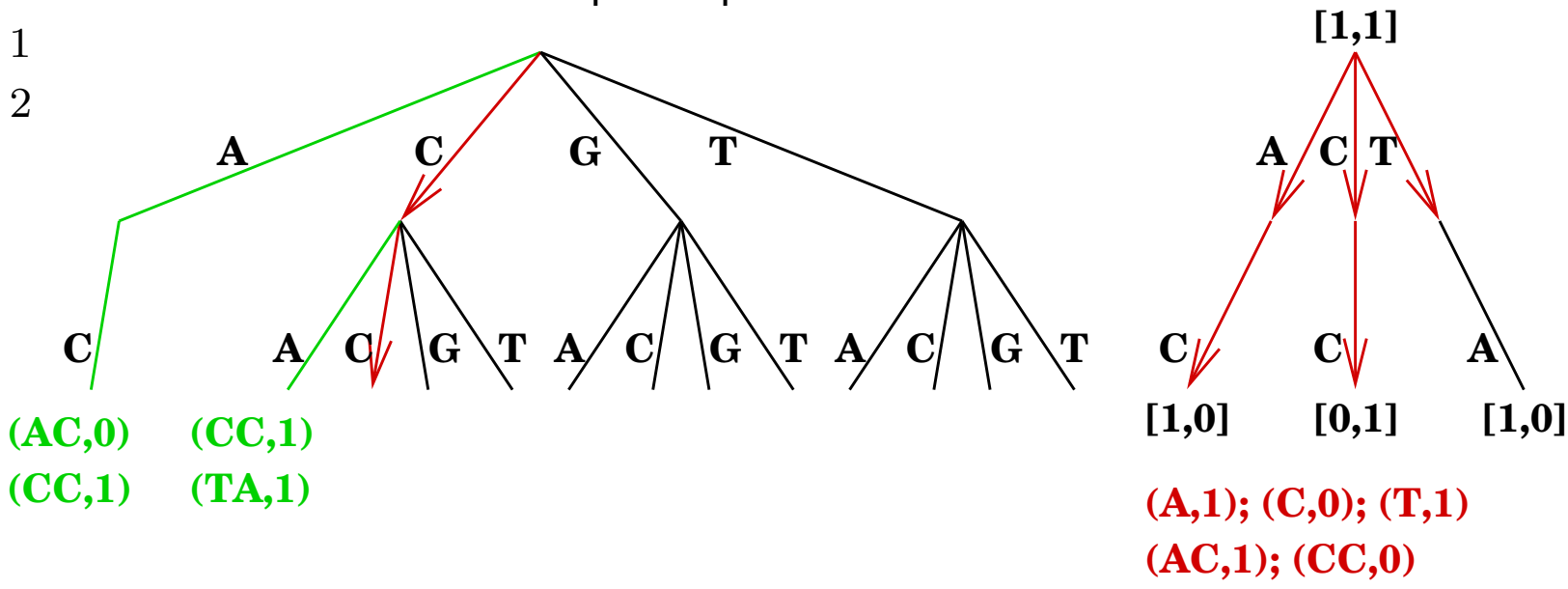


Extraction of Single Models

M.-F. Sagot, *Latin*, 1998

$k = 2$
$e = 1$
$q = 2$

Input sequences: TAC and CCC

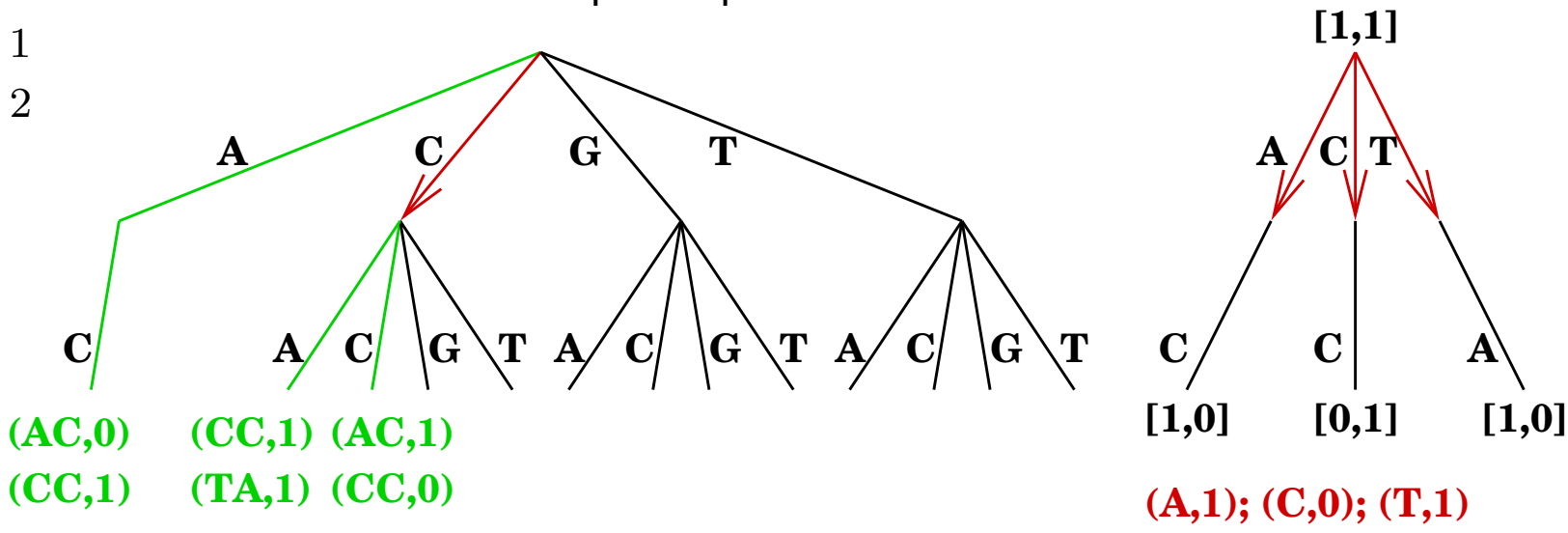


Extraction of Single Models

M.-F. Sagot, *Latin*, 1998

$k = 2$
$e = 1$
$q = 2$

Input sequences: TAC and CCC

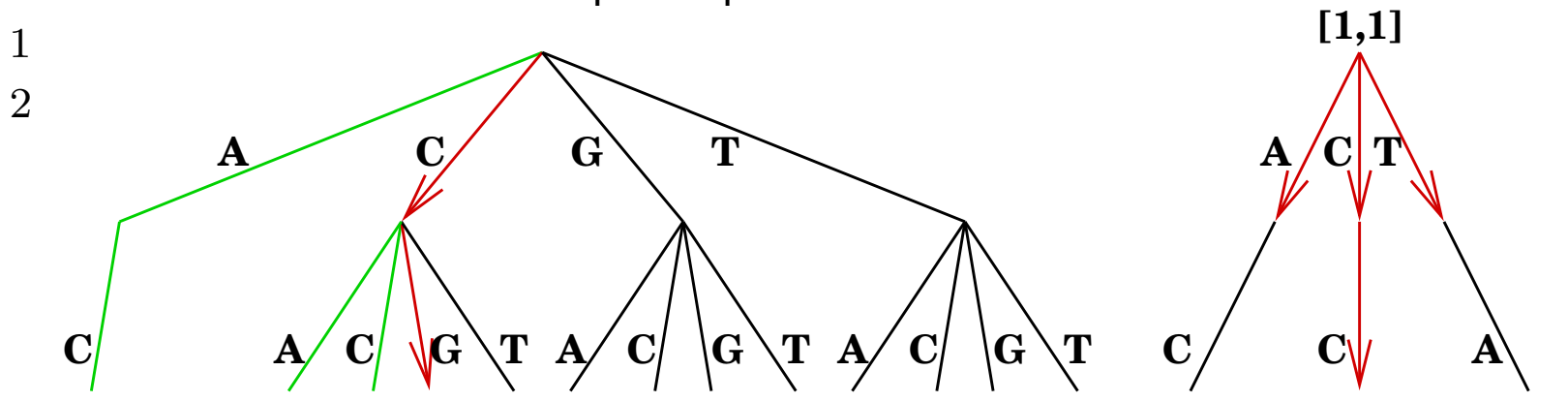


Extraction of Single Models

M.-F. Sagot, *Latin*, 1998

$k = 2$
$e = 1$
$q = 2$

Input sequences: TAC and CCC

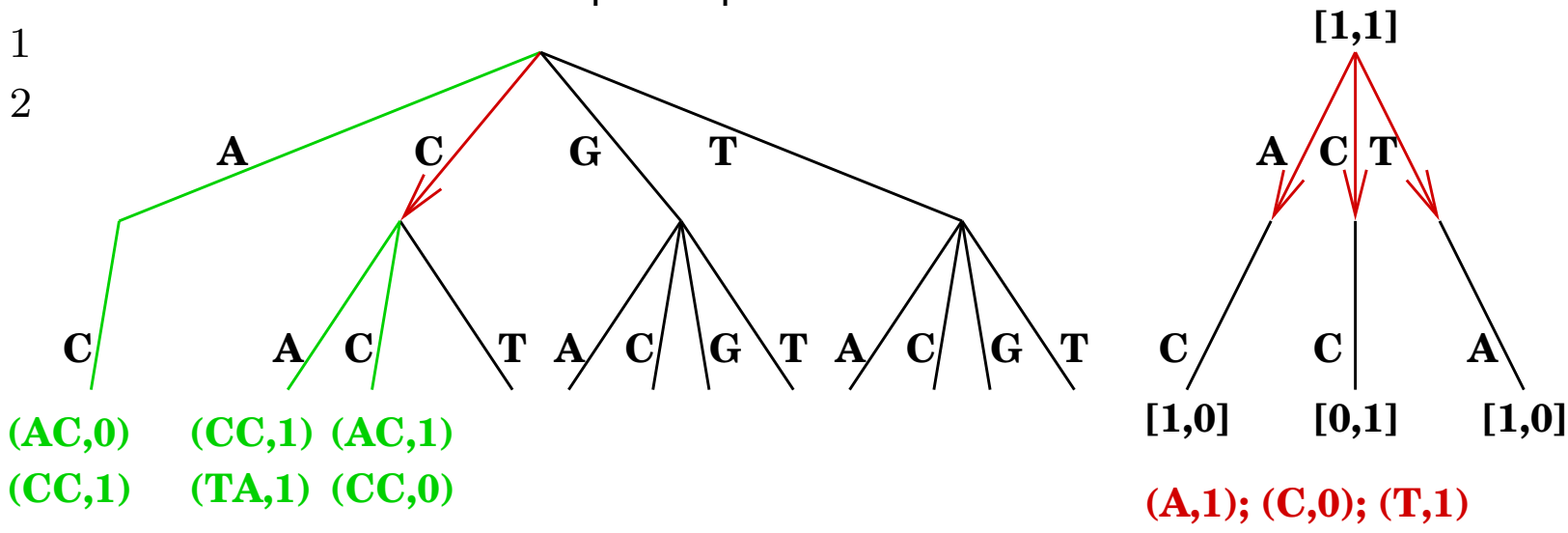


Extraction of Single Models

M.-F. Sagot, *Latin*, 1998

$k = 2$
$e = 1$
$q = 2$

Input sequences: TAC and CCC

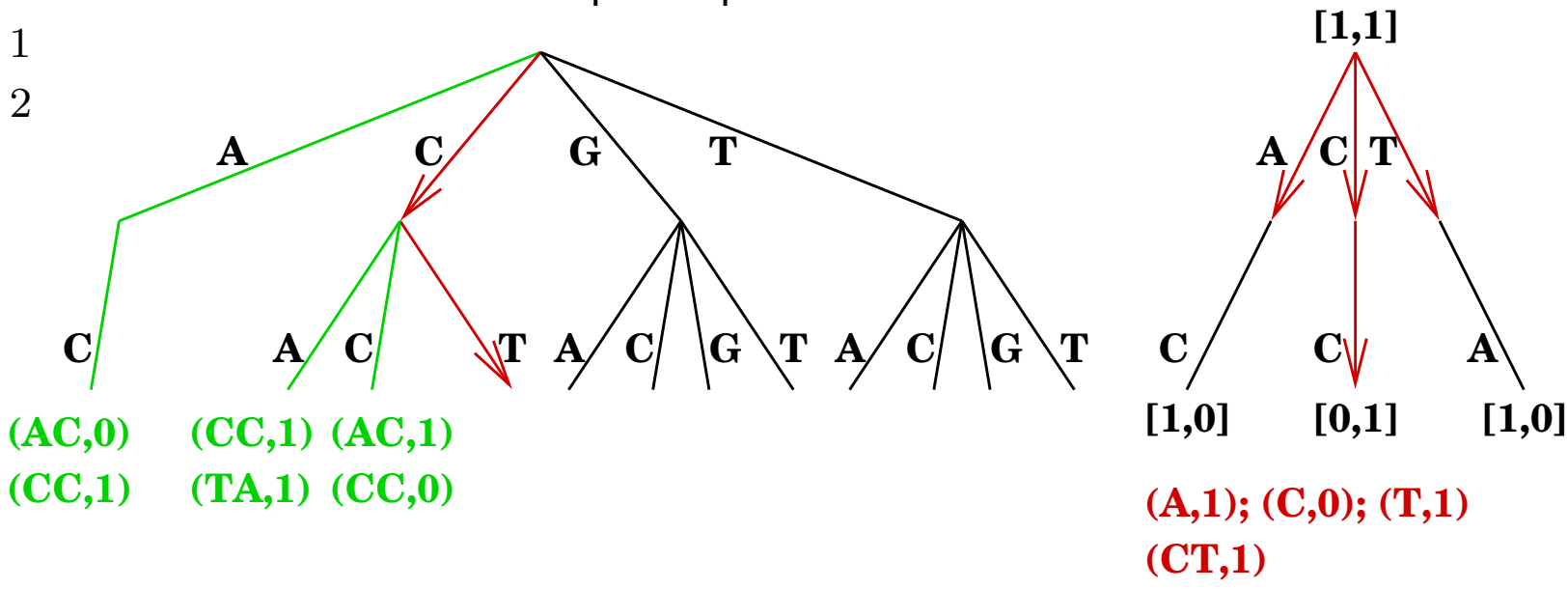


Extraction of Single Models

M.-F. Sagot, *Latin*, 1998

$k = 2$
$e = 1$
$q = 2$

Input sequences: TAC and CCC

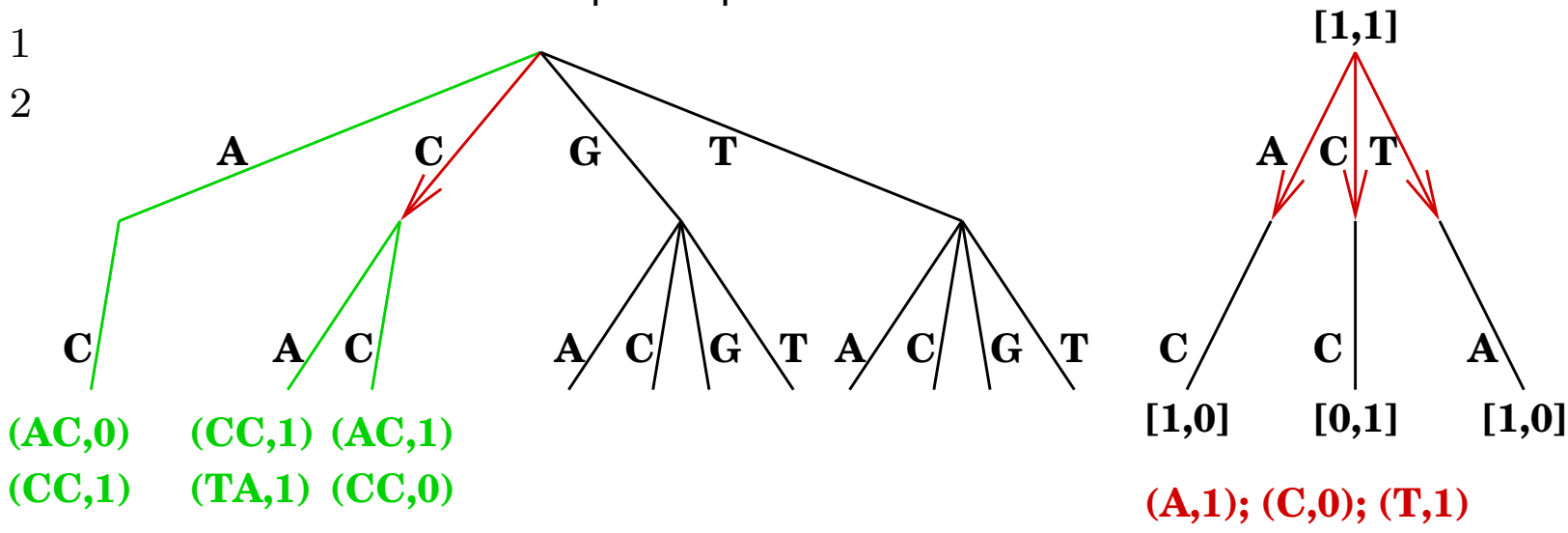


Extraction of Single Models

M.-F. Sagot, *Latin*, 1998

$k = 2$
$e = 1$
$q = 2$

Input sequences: TAC and CCC

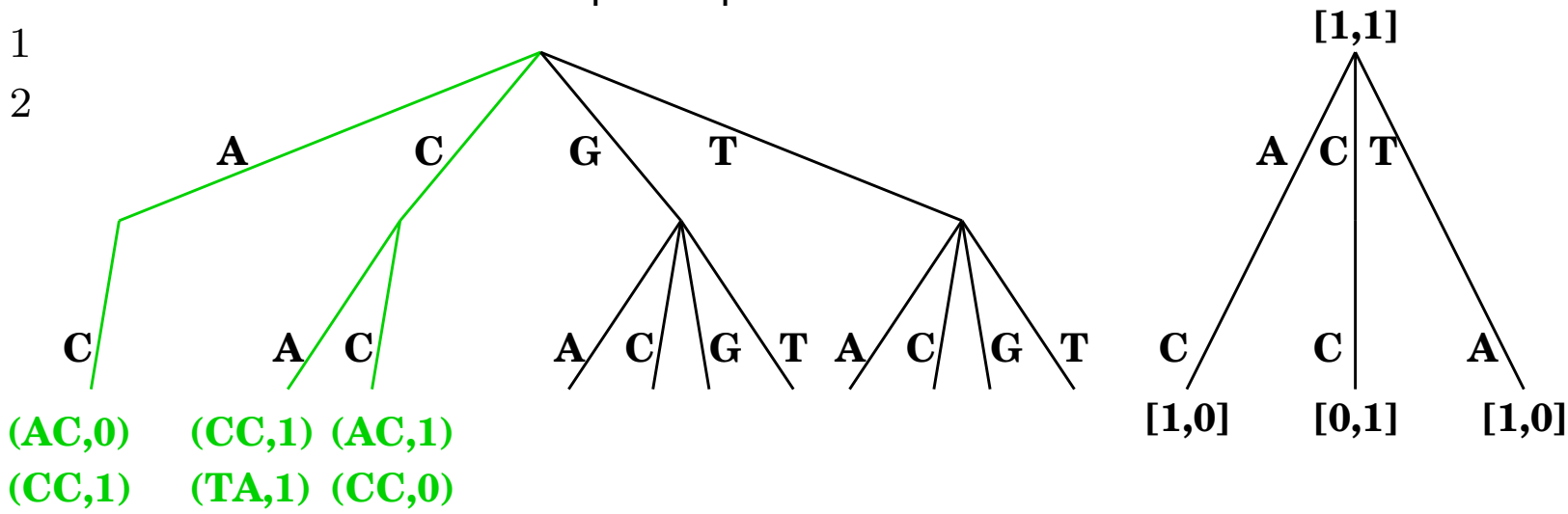


Extraction of Single Models

M.-F. Sagot, *Latin*, 1998

$k = 2$
$e = 1$
$q = 2$

Input sequences: TAC and CCC

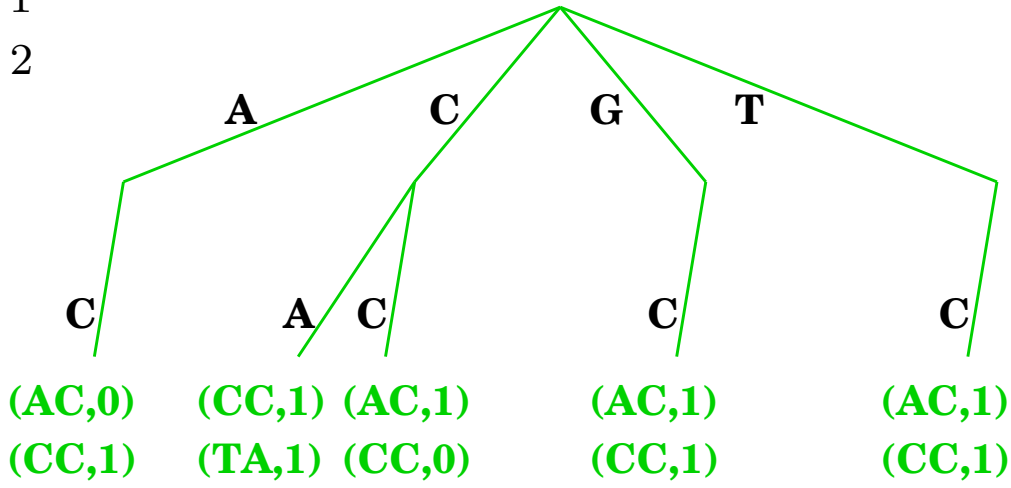


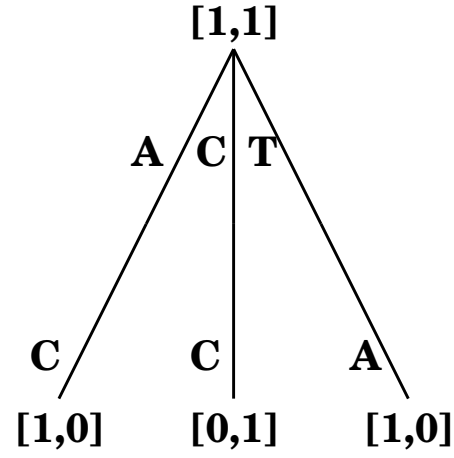
Extraction of Single Models

M.-F. Sagot, *Latin*, 1998

$k = 2$
$e = 1$
$q = 2$

Input sequences: TAC and CCC





Extraction of Structured Models: SMILE

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

ExtractModels (**Model** m, **Block** i)

1. for each node-occurrence v of m
2. if (i > 1)
3. put in $PotencialStarts$ the children of v at levels $(i-1)k + (i-1)d_{min_{i-1}}$ to $(i-1)k + (i-1)d_{max_{i-1}}$
4. else
5. put v in $PotencialStarts$
6. for each model m_i obtained by doing a recursive depth-first traversal from the root of the virtual model tree $\mathcal{M}$ while simultaneously traversing $\mathcal{T}$ from the node-occurrences in $PotencialStarts$
7. if (i < p)
8. **ExtractModels**(m = $m_1 \ldots m_i$, i + 1)
9. else
10. **KeepModel**($\langle (m_1, \ldots, m_p), ((d_{min_1}, d_{max_1}), \ldots, (d_{min_p}, d_{max_p})) \rangle$)

Extraction of Structured Models: SMILE

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

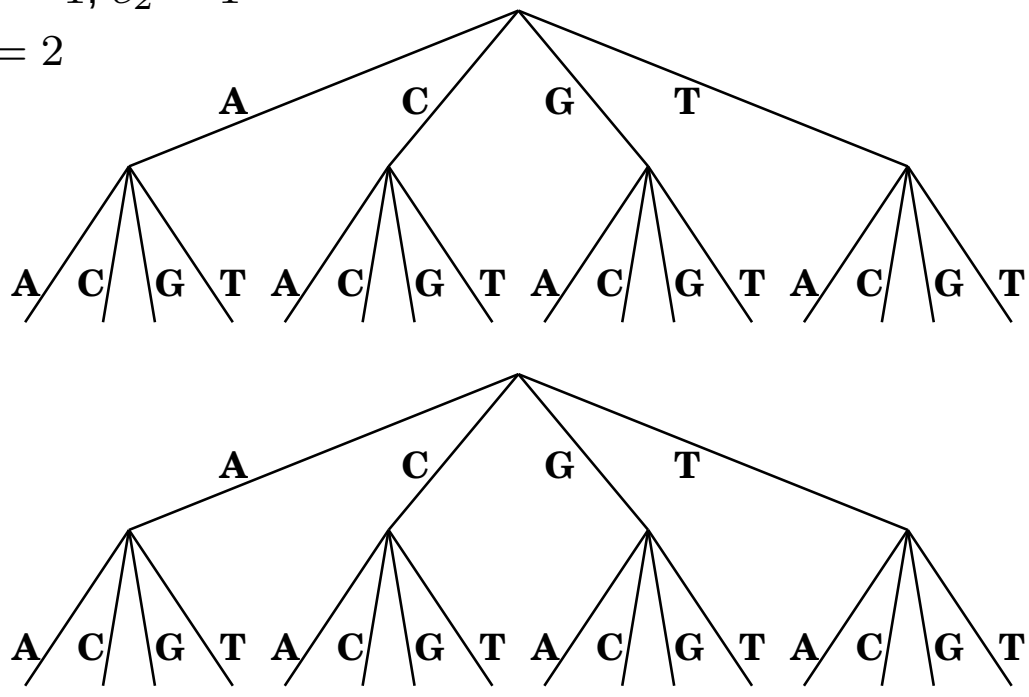
$p = 2$

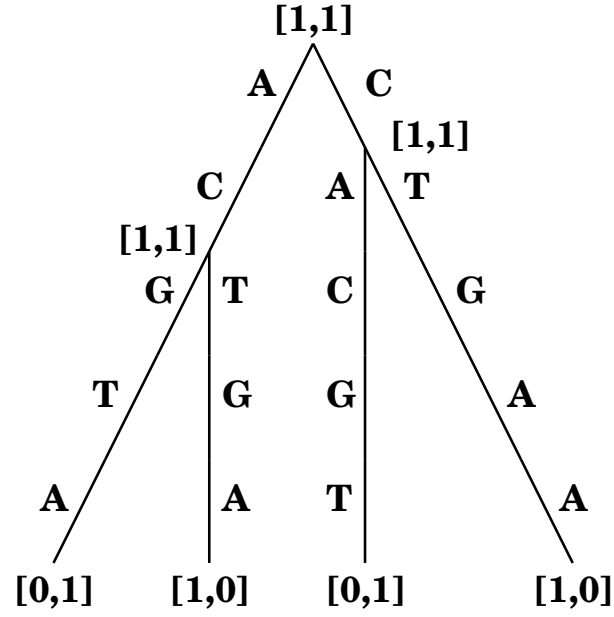
$k_1 = 2$, $d = 1$, $k_2 = 2$

$e_1 = 1$, $e_2 = 1$

$q = 2$

Input sequences: ACTGAA and CACGTA





Extraction of Structured Models: SMILE

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

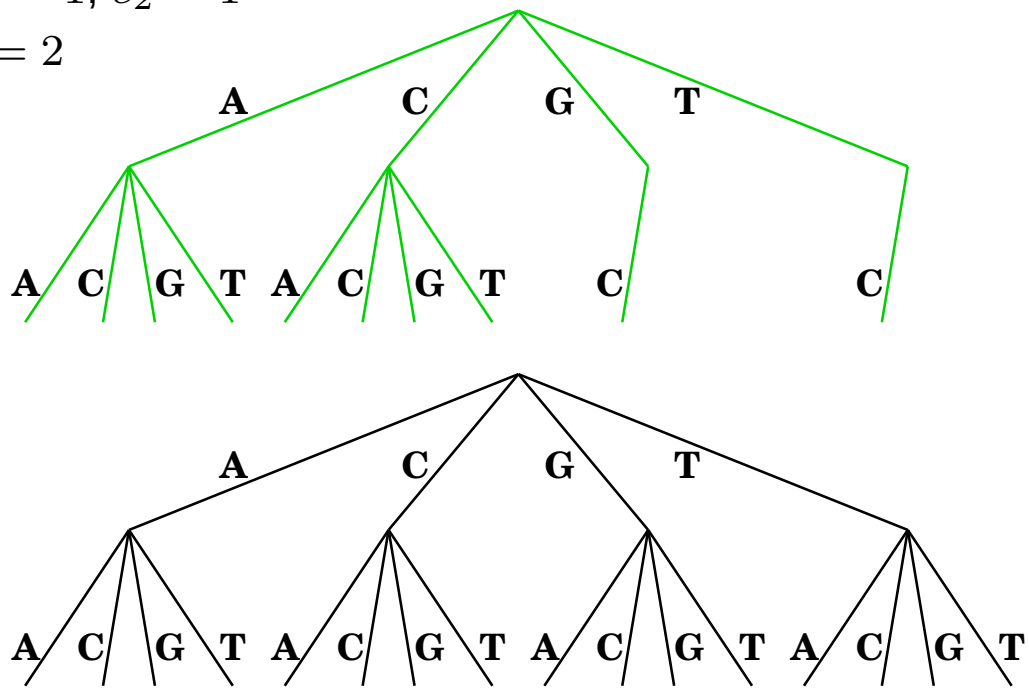
$p = 2$

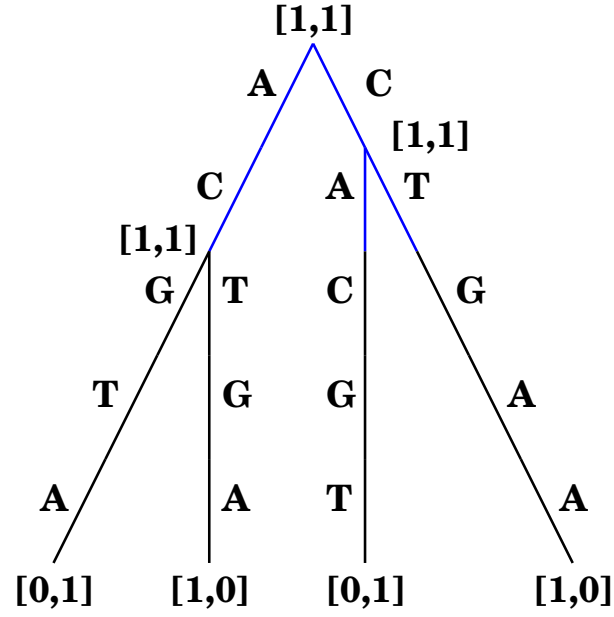
$k_1 = 2, d = 1, k_2 = 2$

$e_1 = 1, e_2 = 1$

$q = 2$

Input sequences: ACTGAA and CACGTA





Extraction of Structured Models: SMILE

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

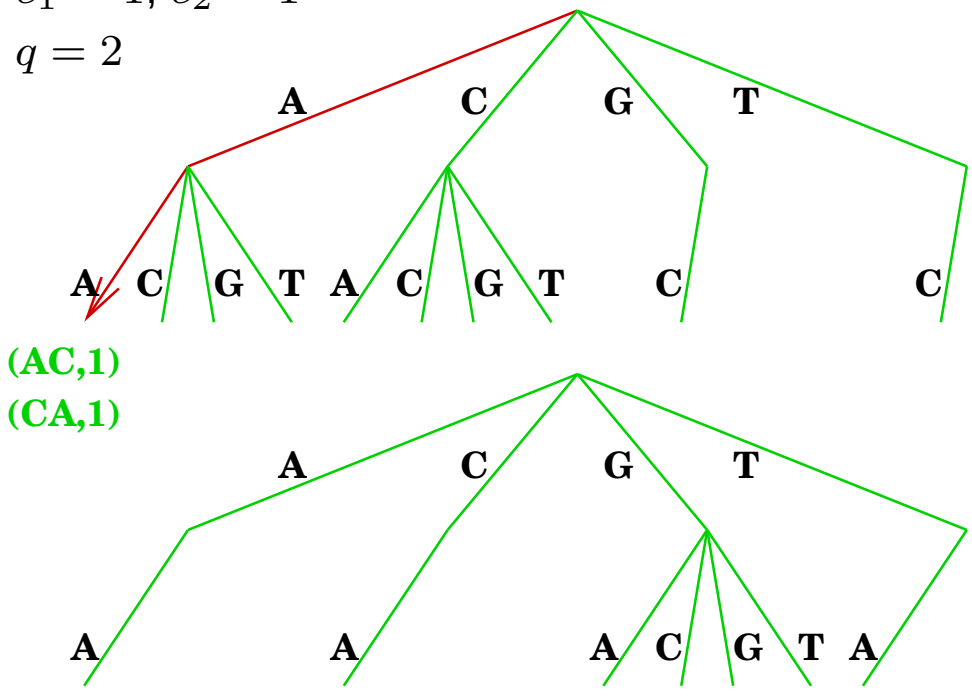
$p = 2$

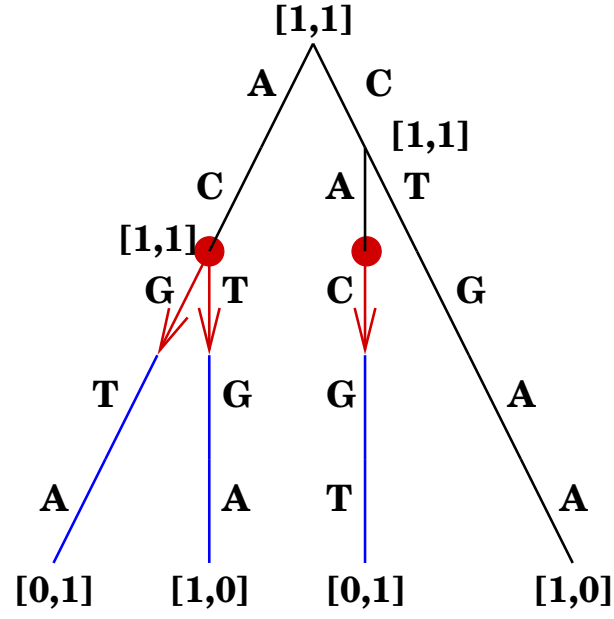
$k_1 = 2, d = 1, k_2 = 2$

$e_1 = 1, e_2 = 1$

$q = 2$

Input sequences: ACTGAA and CACGTA





Extraction of Structured Models: SMILE

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

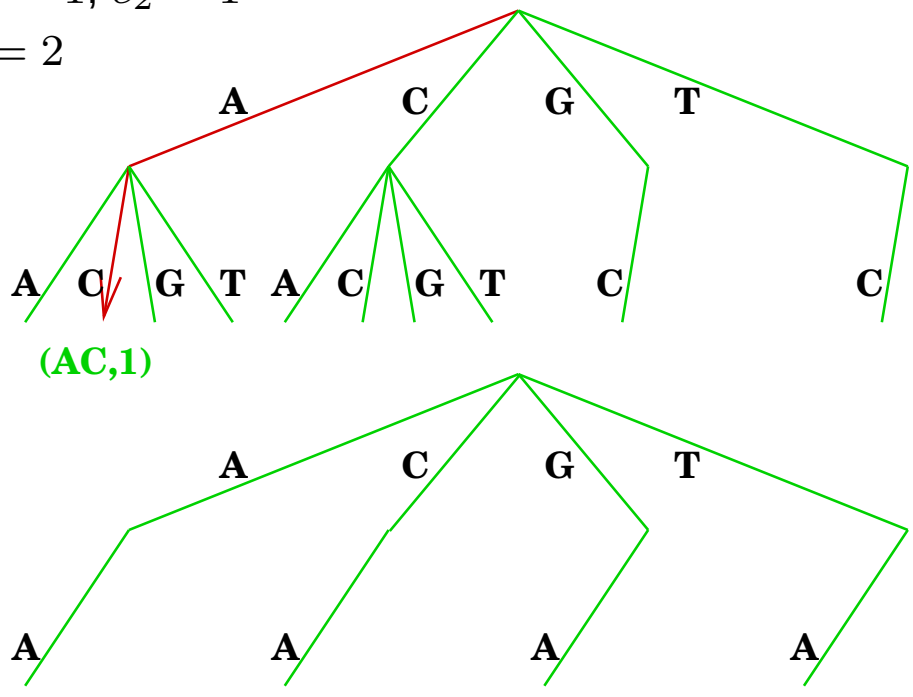
$p = 2$

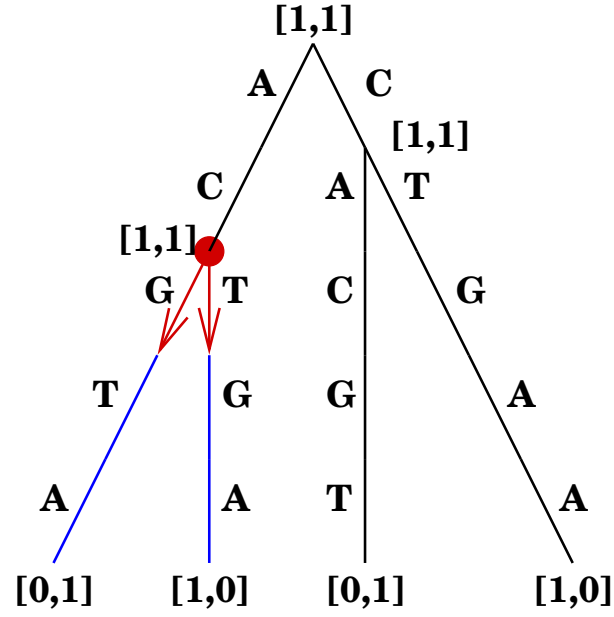
$k_1 = 2, d = 1, k_2 = 2$

$e_1 = 1, e_2 = 1$

$q = 2$

Input sequences: ACTGAA and CACGTA





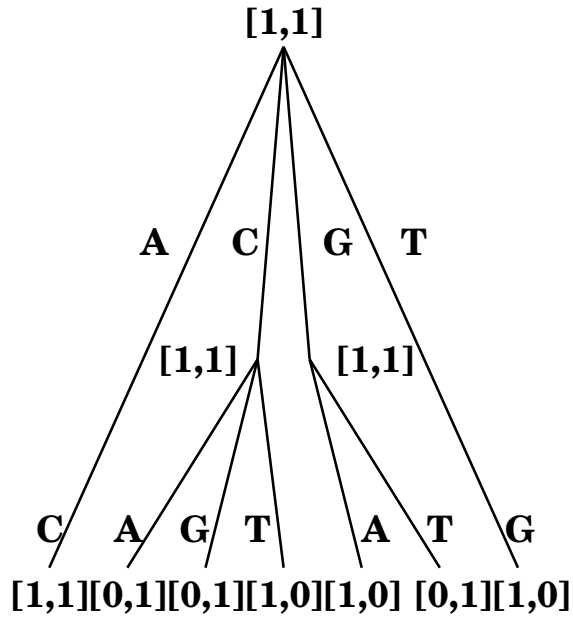
Box-links

A. Carvalho, A. Freitas, A. Oliveira and M.-F. Sagot, *submitted*, 2004

$p = 2$

$k_1 = 2, d = 1, k_2 = 2$

Input sequences: ACTGA and CACGT



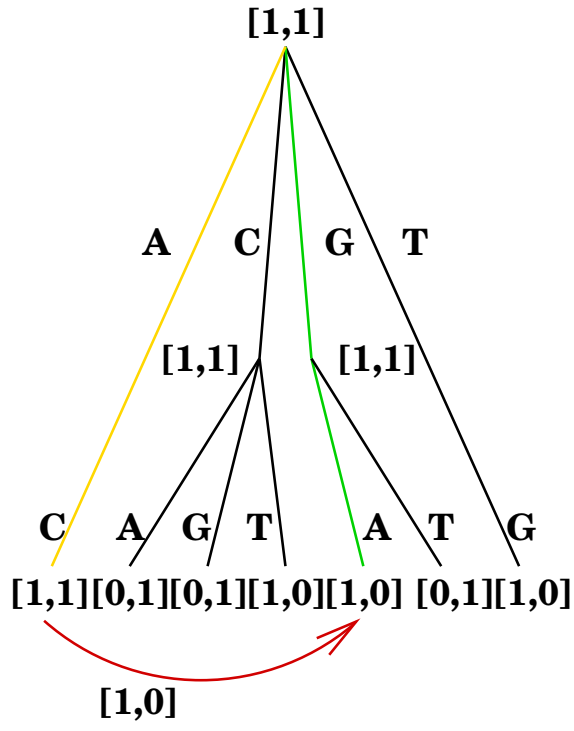
Box-links

A. Carvalho, A. Freitas, A. Oliveira and M.-F. Sagot, *submitted*, 2004

$p = 2$

$k_1 = 2, d = 1, k_2 = 2$

Input sequences: ACTGA and CACGT



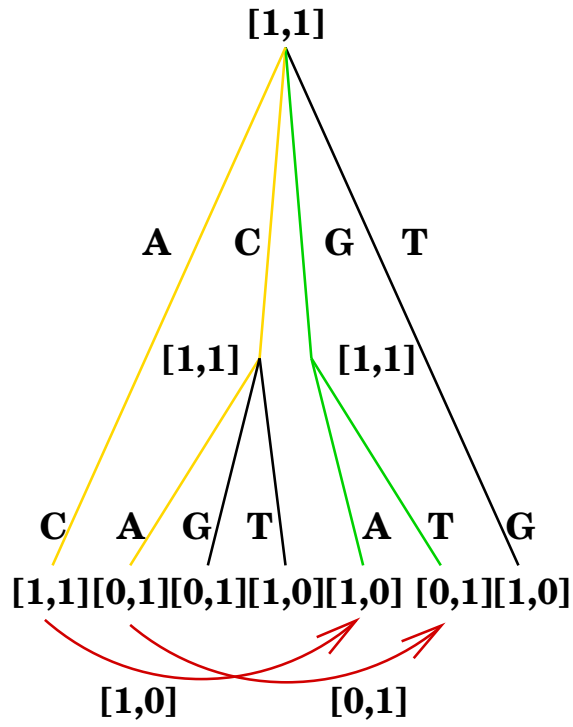
Box-links

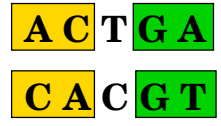
A. Carvalho, A. Freitas, A. Oliveira and M.-F. Sagot, *submitted*, 2004

$p = 2$

$k_1 = 2, d = 1, k_2 = 2$

Input sequences: ACTGA and CACGT





Extraction of Structured Models: RISO

A. Carvalho, A. Freitas, A. Oliveira and M.-F. Sagot, *submitted*, 2004

ExtractModels (**Model** m, **Block** i)

1. for each node-occurrence v of m
2. follow box-links from v and update the tree $\mathcal{T}_i$
3. for each model m_i obtained by doing a recursive depth-first traversal from the root of the virtual model tree $\mathcal{M}$ while simultaneously traversing $\mathcal{T}$ from the root
4. if (i < p)
5. **ExtractModels**(m = $m_1 \ldots m_i$, i + 1)
6. else
7. **KeepModel**($\langle (m_1, \ldots, m_p), ((d_{min_1}, d_{max_1}), \ldots, (d_{min_p}, d_{max_p})) \rangle$)
8. restore the tree $\mathcal{T}_i$

Extraction of Structured Models: RISO

A. Carvalho, A. Freitas, A. Oliveira and M.-F. Sagot, *submitted*, 2004

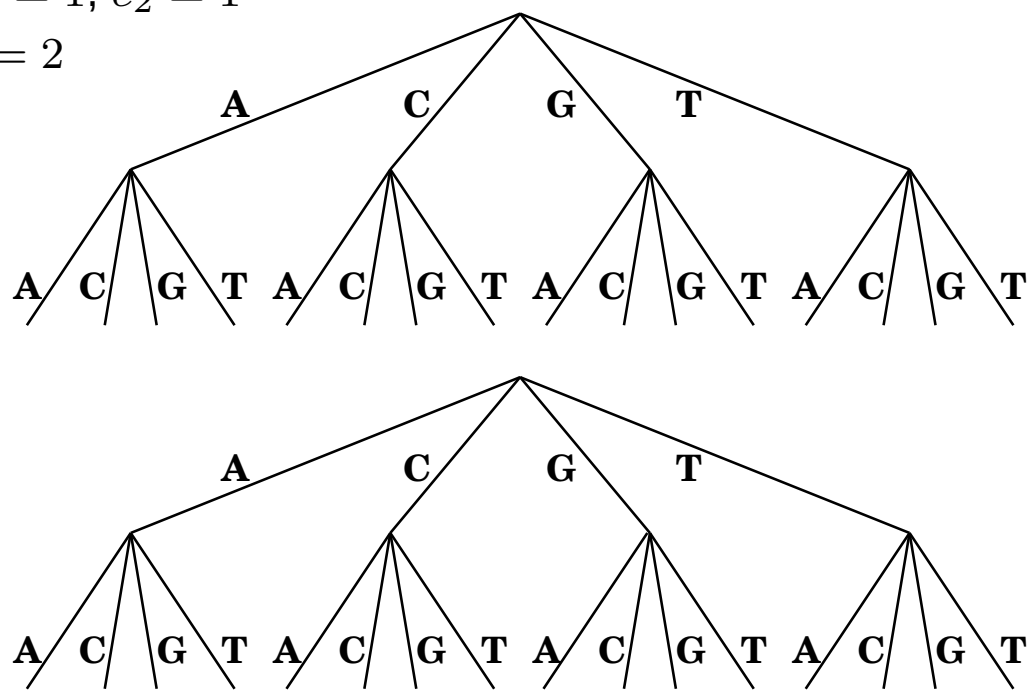
$p = 2$

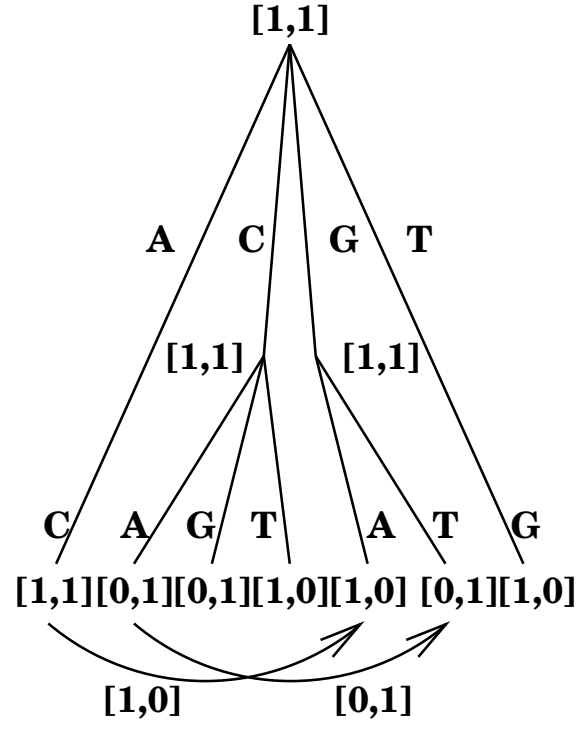
$k_1 = 2, d = 1, k_2 = 2$

$e_1 = 1, e_2 = 1$

$q = 2$

Input sequences: ACTGA and CACGT





Extraction of Structured Models: RISO

A. Carvalho, A. Freitas, A. Oliveira and M.-F. Sagot, *submitted*, 2004

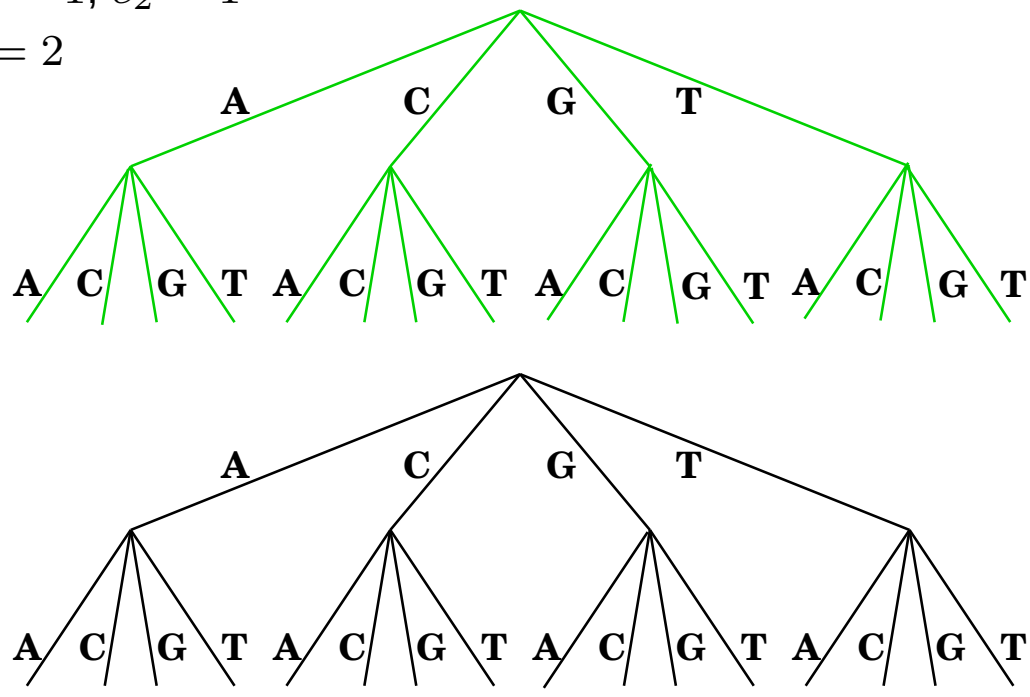
$p = 2$

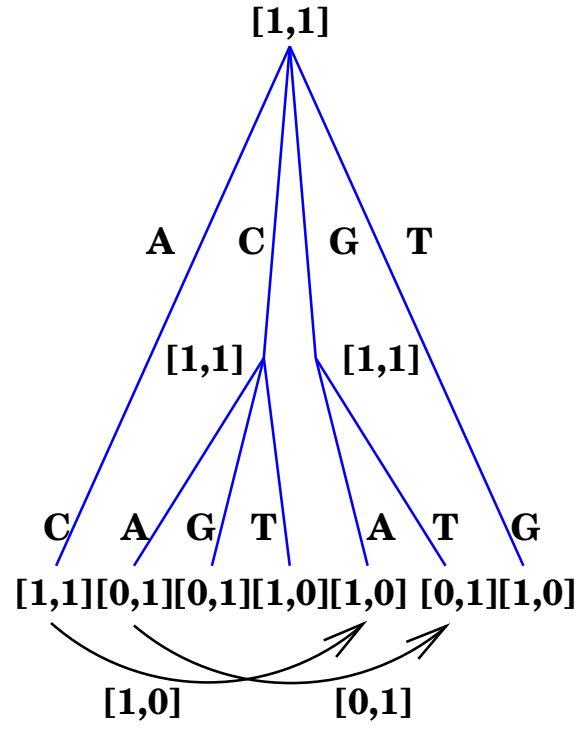
$k_1 = 2, d = 1, k_2 = 2$

$e_1 = 1, e_2 = 1$

$q = 2$

Input sequences: ACTGA and CACGT





Extraction of Structured Models: RISO

A. Carvalho, A. Freitas, A. Oliveira and M.-F. Sagot, *submitted*, 2004

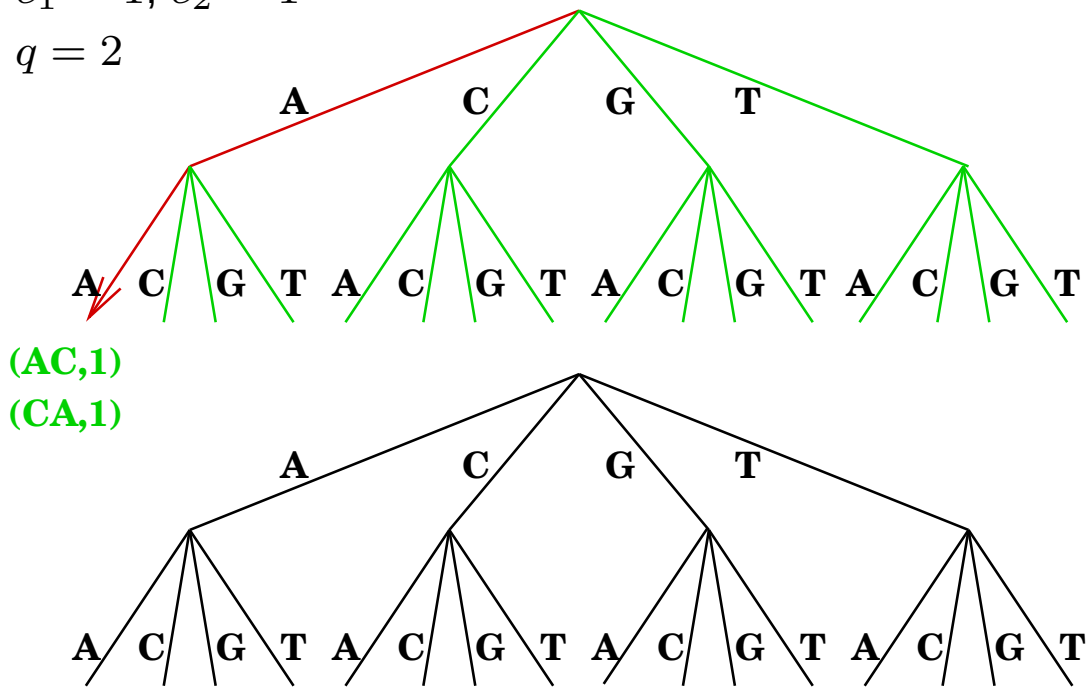
$p = 2$

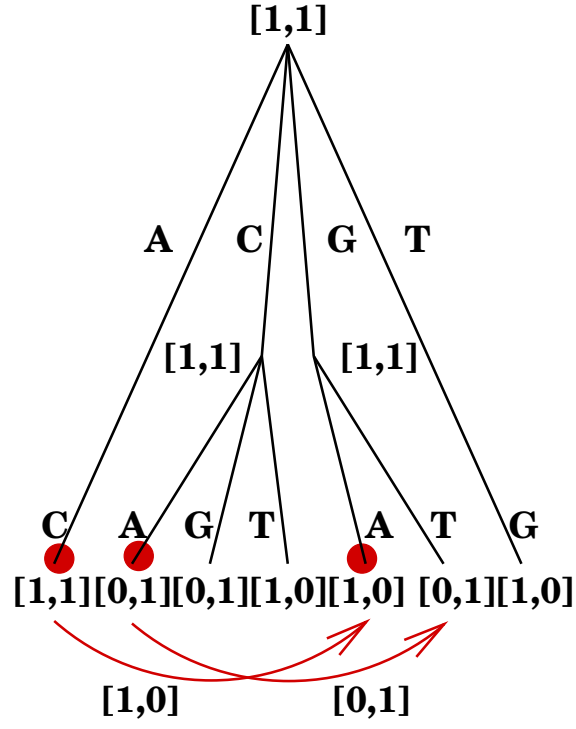
$k_1 = 2, d = 1, k_2 = 2$

$e_1 = 1, e_2 = 1$

$q = 2$

Input sequences: ACTGA and CACGT





Extraction of Structured Models: RISO

A. Carvalho, A. Freitas, A. Oliveira and M.-F. Sagot, *submitted*, 2004

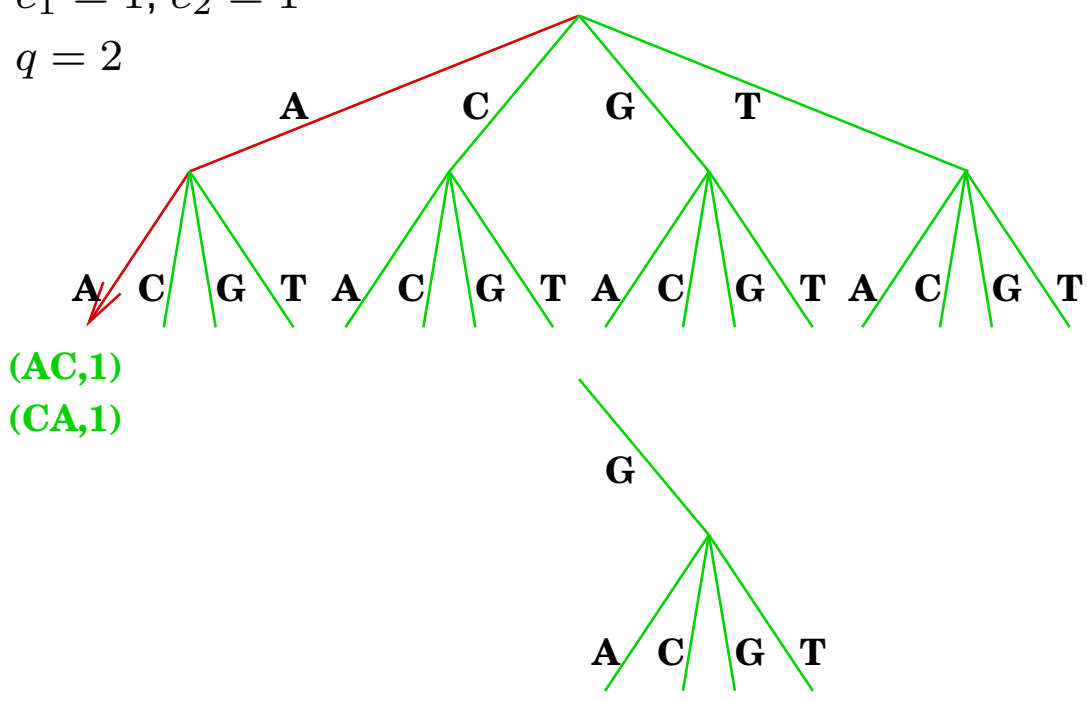
$p = 2$

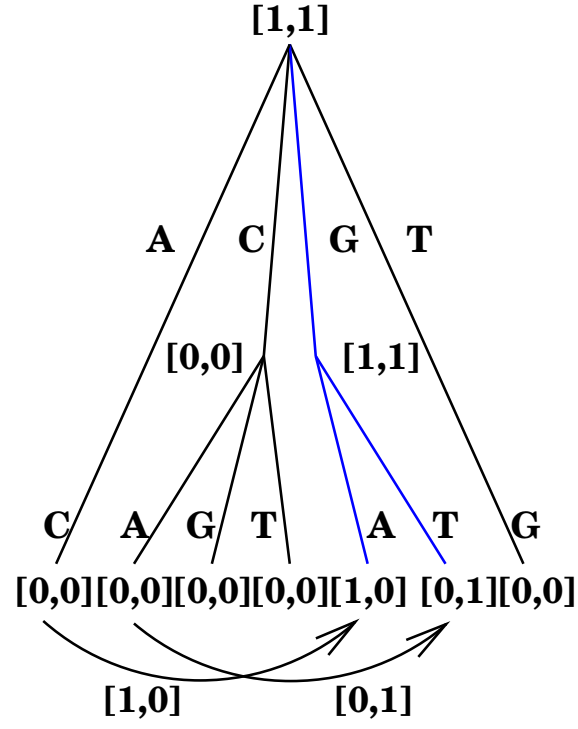
$k_1 = 2, d = 1, k_2 = 2$

$e_1 = 1, e_2 = 1$

$q = 2$

Input sequences: ACTGA and CACGT





Extraction of Structured Models: RISO

A. Carvalho, A. Freitas, A. Oliveira and M.-F. Sagot, *submitted*, 2004

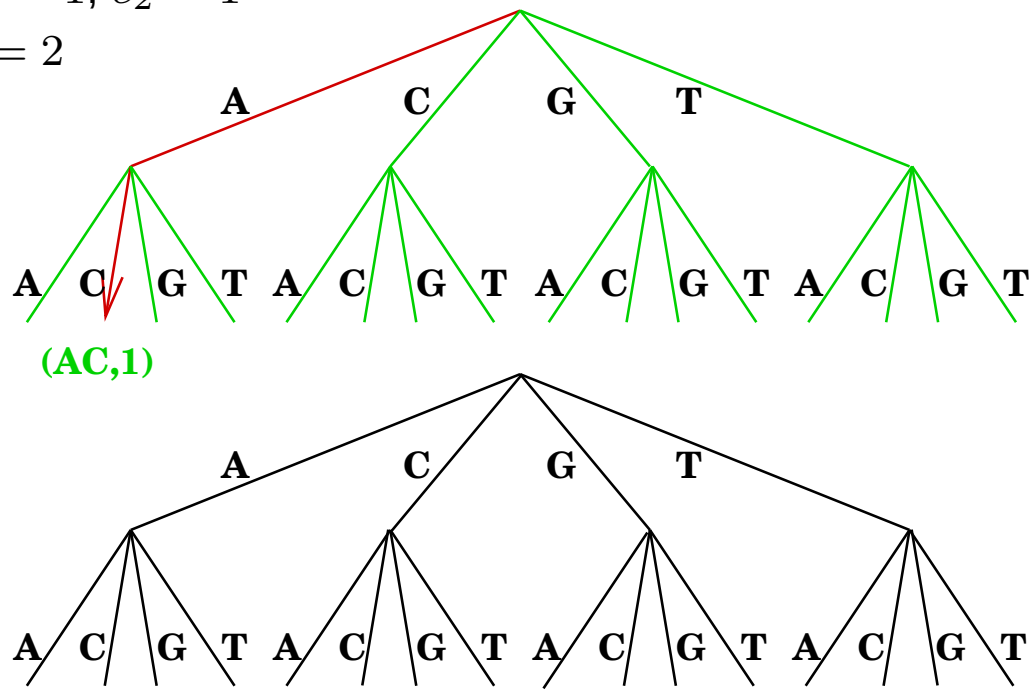
$p = 2$

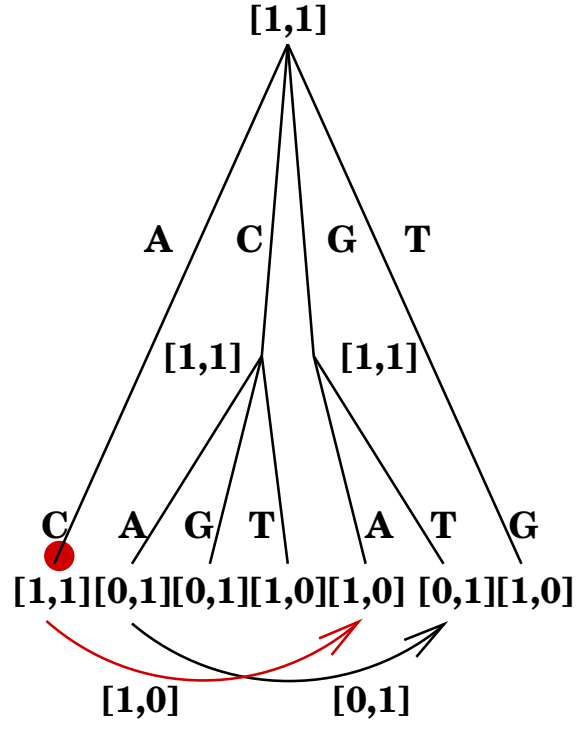
$k_1 = 2, d = 1, k_2 = 2$

$e_1 = 1, e_2 = 1$

$q = 2$

Input sequences: ACTGA and CACGT





Extraction of Structured Models: RISO

A. Carvalho, A. Freitas, A. Oliveira and M.-F. Sagot, *submitted*, 2004

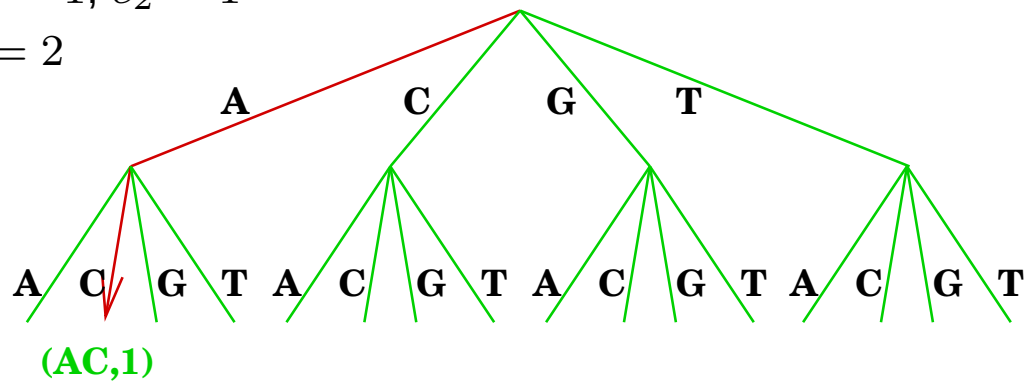
$p = 2$

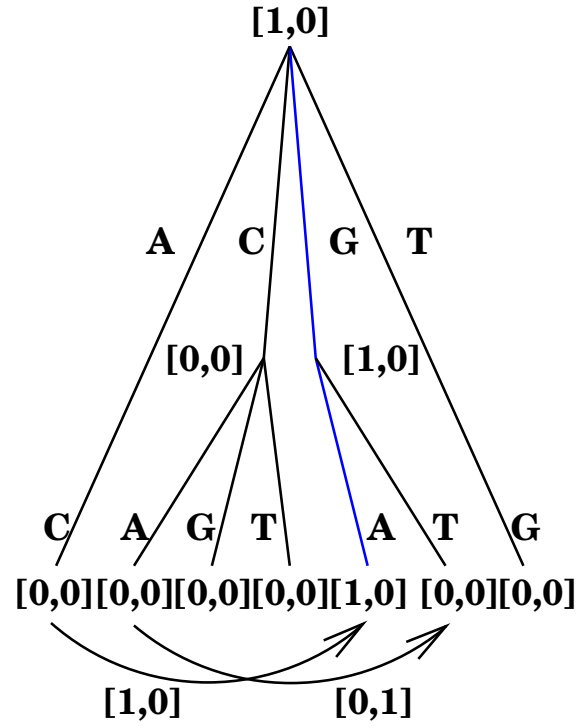
$k_1 = 2, d = 1, k_2 = 2$

$e_1 = 1, e_2 = 1$

$q = 2$

Input sequences: ACTGA and CACGT





Comparing the algorithms

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		# models
Box 1	Box 2	SMILE	RISO	
1	1	44.72	0.12	4096
2	2	1612.68	12.12	65536

Extraction of the $TTGACAn_{17}TATAAT$ motif

1148 sequences from the *E. coli* genome

# Errors		CPU Times		# models
Box 1	Box 2	SMILE	RISO	
1	2	1429.81	942.42	11147160

Ongoing and future work

- Proposal of new and more flexible biological models for promoter regions and development of efficient algorithms to extract them
- Integration of these algorithms with a database of transcription factors and respective promoter consensus motifs for the several organism, in order to:
  - provide semi-automatic methods for processing experimental results
  - allow users to analyze complex interactions between gene networks and proteins