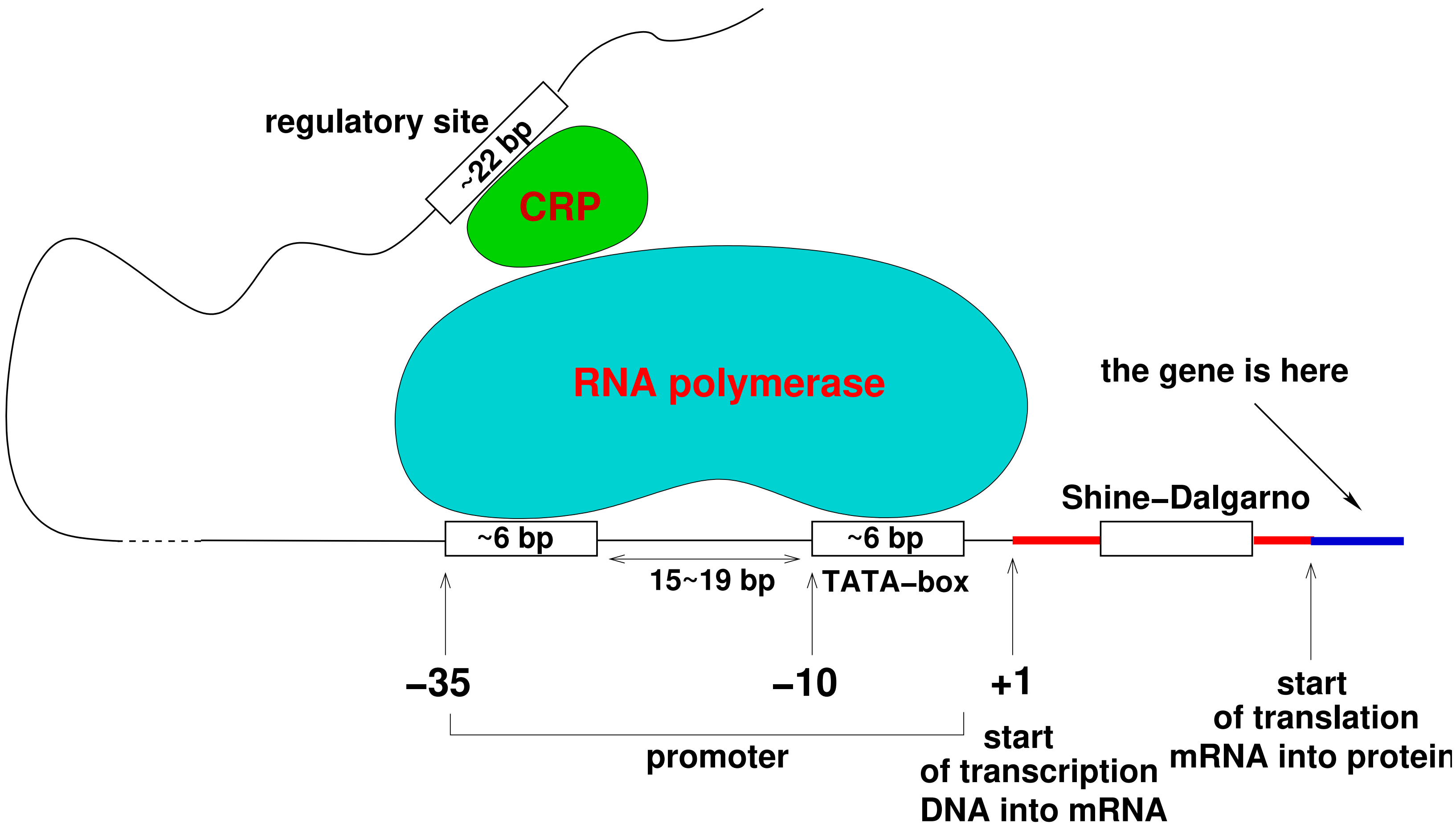
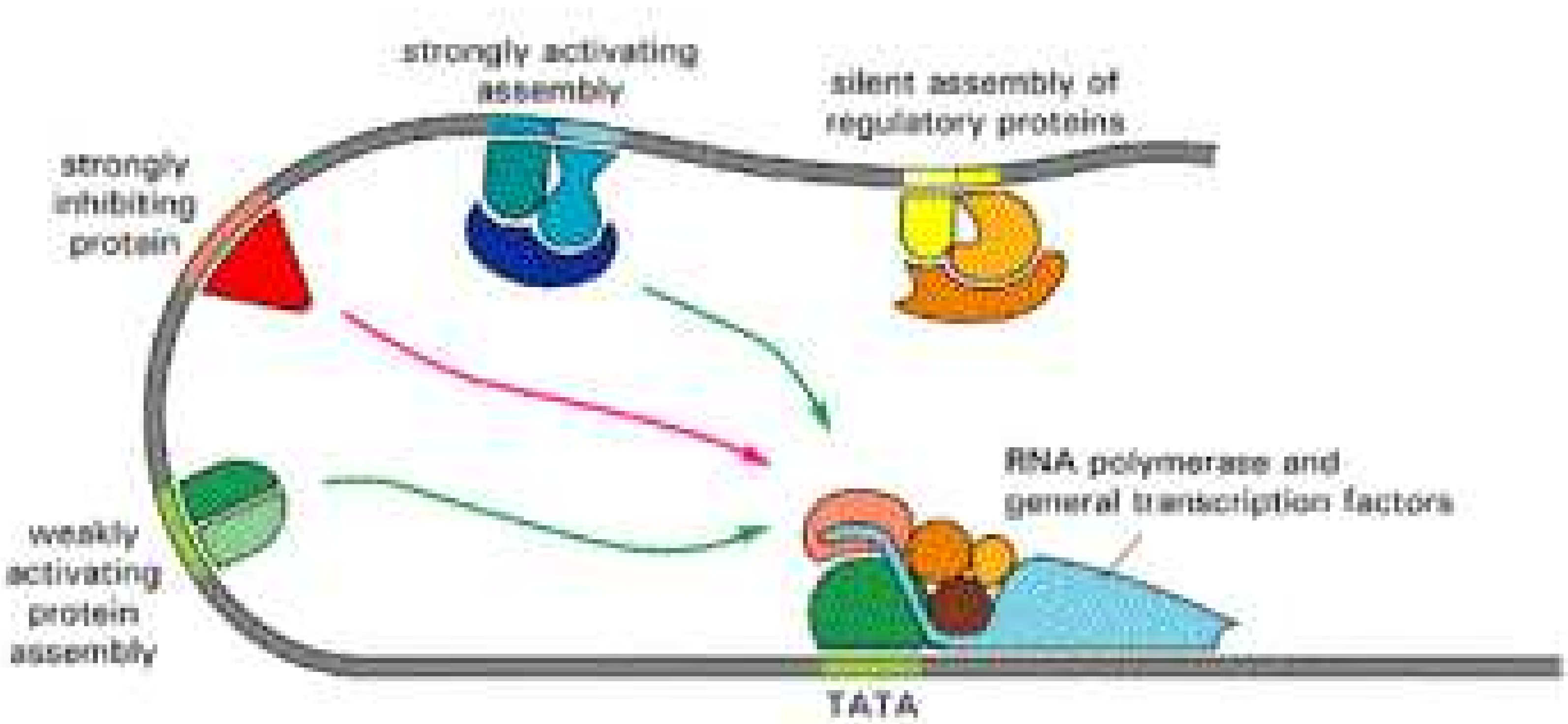
On bases of motifs

Maxime Crochemore (IGM, University of Marne-la-Vallée)
Roberto Grossi (University of Pisa, Italy)
Nadia Pisanti (University of Pisa, Italy)
Marie-France Sagot (BAOBAB Team, INRIA & Université Claude Bernard, Lyon)

MOTIVATION

Biology: motifs represent important signals





Life is the first combinatorist

Motifs are repeats, short or long, “dispersed” or in tandem as here

...gaaagctcttctgctccagtaccaactccatccagctcgacgact
gaaagctcttctgctccagt aa ccagctcgaccact
gaaagctcttctgctccagtaccaactccatccagctcgacgact
gaaagctcttctgctccagctccaacccatcaagctccaccact
gaaagctcttctgctccagt aa ccagctcgaccact
gaaagctcttctgctccagtaccaactccatccagctcgacgact
gaaagctcttctactccagt aa ccagctccaccact
gaaagctcttctgctccagtgccaactccatcaagctcgacgact
gaaagctcttctgctccagtaccaactccatcaagctcgacgact
gaaagctcttctgctccagctccaacccatcaagctcgacgact
gaaagctcttctgctccagt aa ccagctcgaccact
gaaagctcttctgctccagtaccaactccatccagctcgacgact
gaaagctcttctgctccagtaccaacccatccagctcgacgact
gaaagctcttctgctccagtaccaactccatccagctcgacgact...

THE INFERENCE PROBLEM

Motif search $\times$ Motif inference

Search

a known motif **a text**	$\Longrightarrow$	**all positions in the text where the motif is "found"**

Inference

a set of properties **a text**	$\Longrightarrow$	**(all) motifs satisfying the properties**

Text: A set of biological sequences (strings) or a single sequence

Set of properties

main property: motifs of interest = "conserved" elements

mutations (substitutions and in some cases insertions and deletions) may happen that do not destroy function

and may even enable to modulate it

Set of properties

main property: motifs of interest = “conserved” elements

mutations (substitutions and in some cases insertions and deletions) may happen that do not destroy function

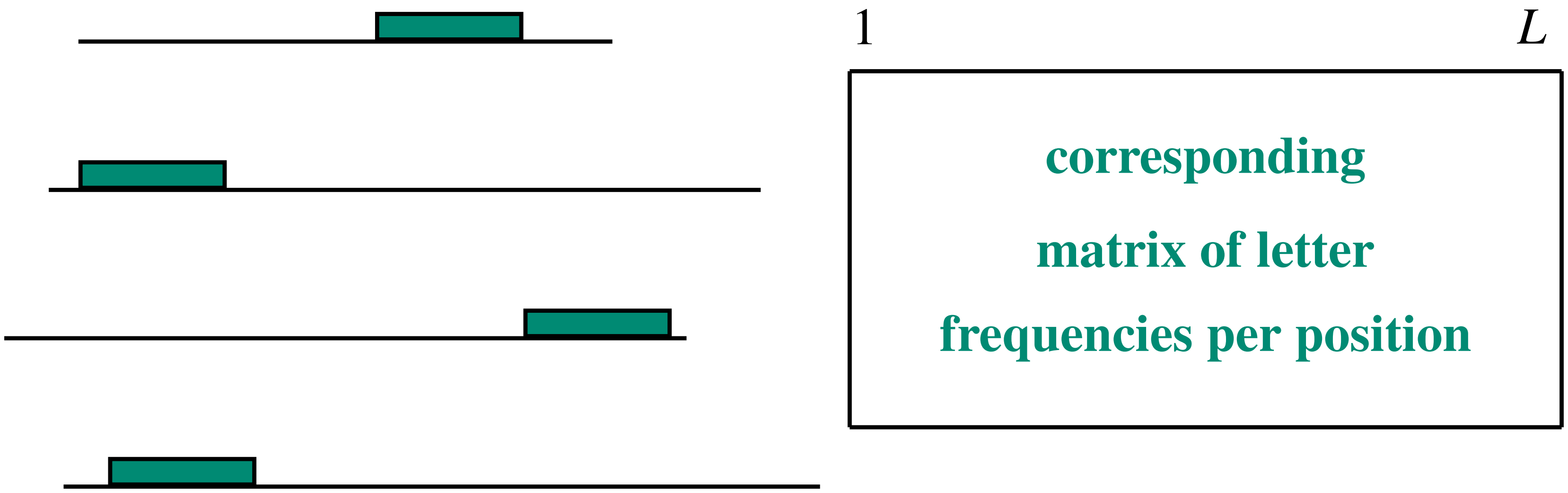
and may even enable to modulate it

(“la vie comme un bordel contrôlé”)

“Conserved” element(s): “Most surprising set(s) of words”

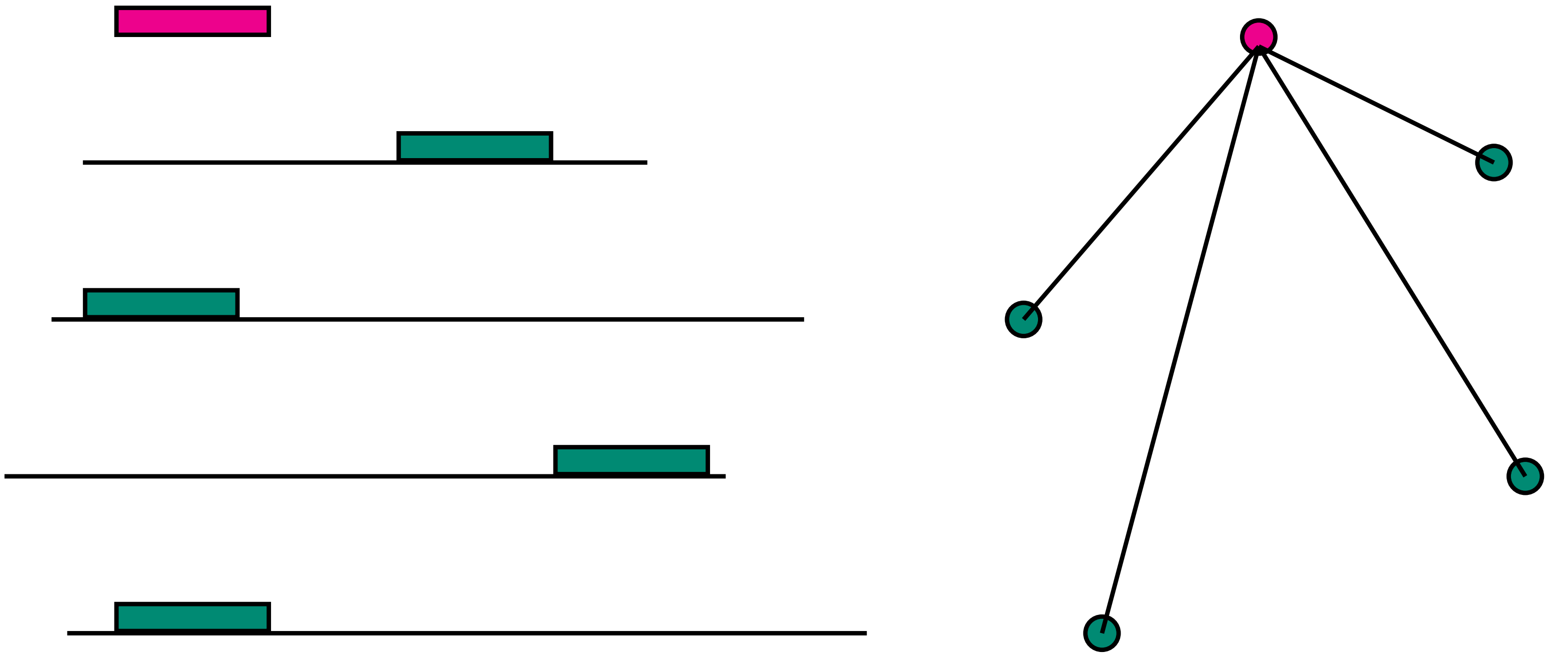
$$\sum_{i=1}^{L} \sum_{\alpha \in \Sigma} f_{i\alpha} \log_2 \frac{f_{i\alpha}}{f_\alpha}$$

relative entropy



“Conserved” elements: “Approximate repeat”

Pattern + “distance” (for instance, Hamming distance)



Pattern

simple pattern	limited regular expression
TATAAT	**TA[AT]●[AT]T**
TTGACA	
RFMCP	**[ILMV][ASG]●●C[ILMV]H**

where ● is the don’t care symbol

MOTIFS WITH WILD CARDS

A more formal definition of the problem

input:

data: one string s over alphabet Σ

output:

all motifs over $\Sigma \cup \Sigma(\Sigma \cup \{\bullet\})^{*}\Sigma$ having at least q occurrences in s

Motifs

specificity relation:

$\bullet \preceq \sigma$

relation $\preceq$ **extends to words of equal length**

occurrence:

the pattern x **occurs in** s **at position** d **if** $x \preceq s[d\,..\,d+|x|-1]$

location list:

$\mathcal{L}(x) = \{$**positions of** x **in** $s\}$

Example

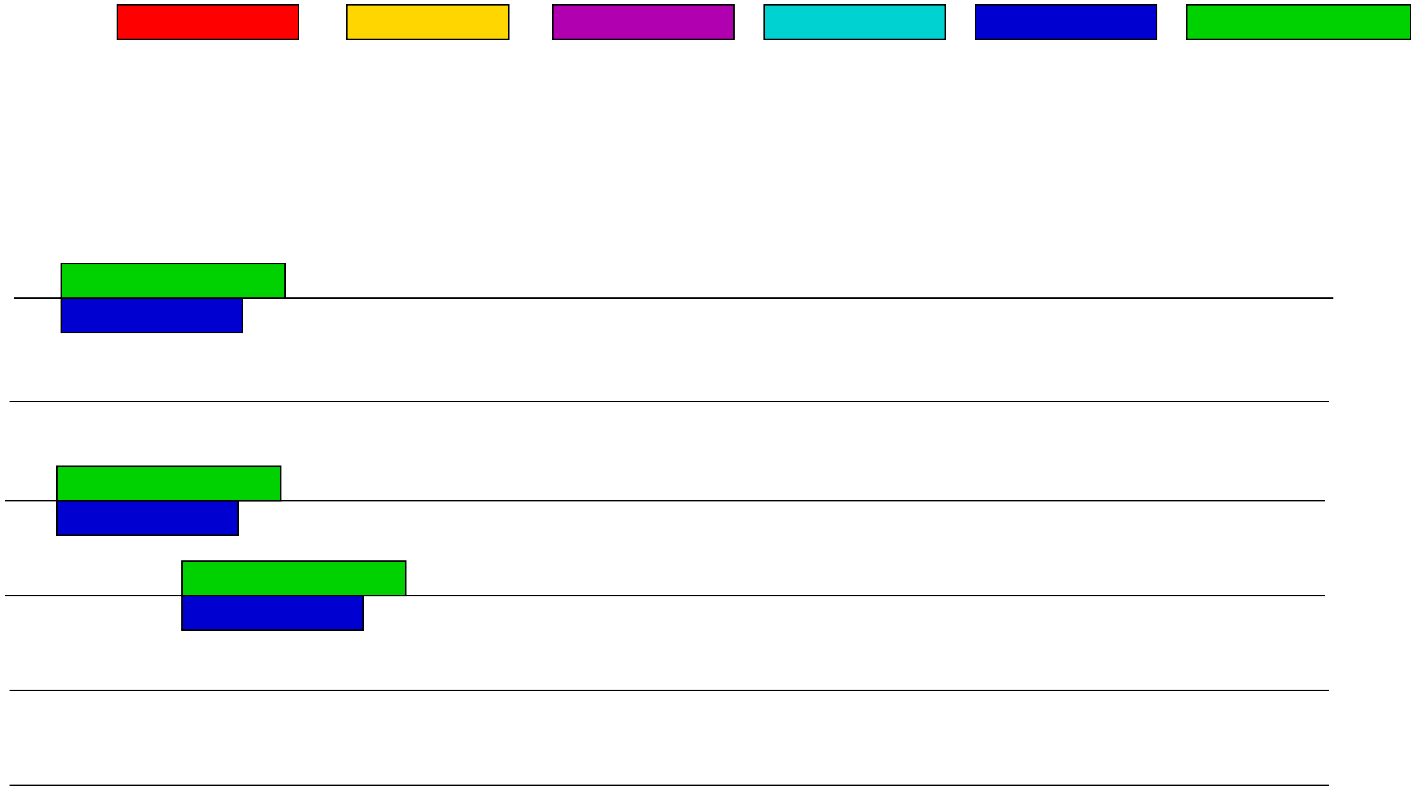
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	positions
s	F	A	B	C	X	F	A	D	C	Y	Z	E	A	D	C	E	A	D	C	
x_1								D	C					D	C			D	C	$\{7, 13, 17\}$
x_2							A	D	C				A	D	C		A	D	C	$\{6, 12, 16\}$
x_3		A	•	C			A	•	C				A	•	C		A	•	C	$\{1, 6, 12, 16\}$
x_4	F	A	•	C		F	A	•	C											$\{0, 5\}$
x_5												E	A	•	C	E	A	•	C	$\{11, 15\}$

Initial observations

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	positions
s	F	A	B	C	X	F	A	D	C	Y	Z	E	A	D	C	E	A	D	C	
x_1								D	C					D	C			D	C	$\{7, 13, 17\}$
x_2							A	D	C				A	D	C		A	D	C	$\{6, 12, 16\}$
x_3		A	•	C			A	•	C				A	•	C		A	•	C	$\{1, 6, 12, 16\}$
x_4	F	A	•	C		F	A	•	C											$\{0, 5\}$
x_5												E	A	•	C	E	A	•	C	$\{11, 15\}$

- x_1 **is not maximal**
- x_2, x_3, x_4, x_5 **are maximal**

“ Maximal” motifs



Counting motifs with wild cards

A few facts

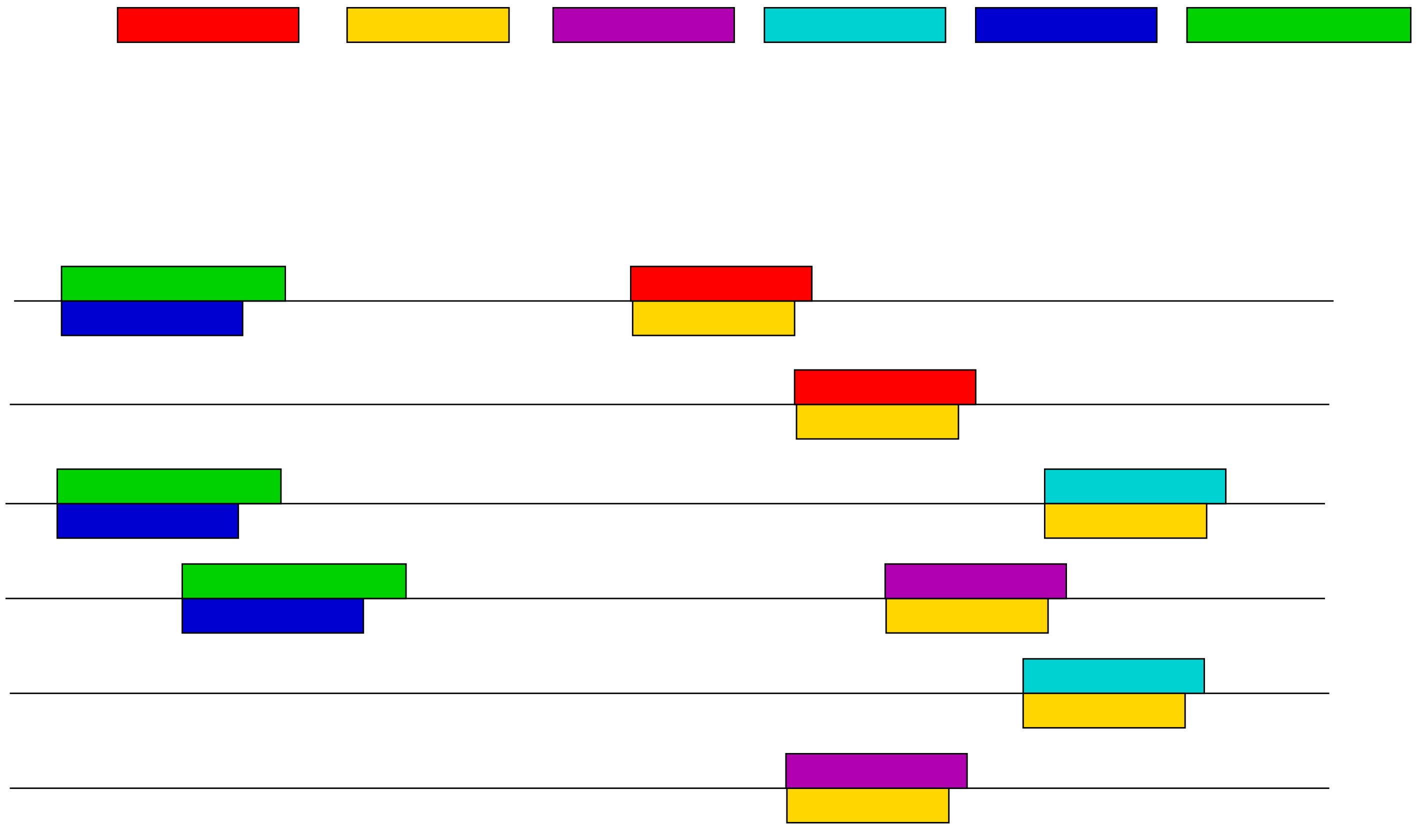
- **there can be an exponential number of motifs over Σ' for $q = 2$**
- **there can be an exponential number of maximal such motifs**
  example: $\mathbf{A}^p\mathbf{TA}^p$

Exponential number of maximal motifs

		1	2	3	4	5	6	7	8	9	10	11	12	13
s	=	A	A	A	A	A	A	T	A	A	A	A	A	A

A●AAAA	1,6,8	A● ●AAA	1,5,6,8	A●●●AA	1,4,5,6,8
AA●AAA	1,5,8	A●A●AA	1,4,6,8	A●●A●A	1,3,5,6,8
AAA●AA	1,4,8	A●AA●A	1,3,6,8	A●A●●A	1,3,4,6,8
AAAA●A	1,3,8	AA●●AA	1,4,5,8	AA●●●A	1,3,4,5,8
		AA●A●A	1,3,5,8		
		AAA● ●A	1,3,4,8		

Introducing the notion of "irredundant motifs" (Parida *et al.*)



More formally

A maximal motif x **is redundant if**

$$\mathcal{L}(x) = \mathcal{L}(y_1) \cup \mathcal{L}(y_2) \cup \ldots \cup \mathcal{L}(y_k)$$

for $y_1, y_2, \ldots, y_k$ **maximal motifs** $\neq x$

x **is covered by** $y_1, y_2, \ldots, y_k$

A maximal motif x **is irredundant if it is not redundant**

Counting motifs with wild cards

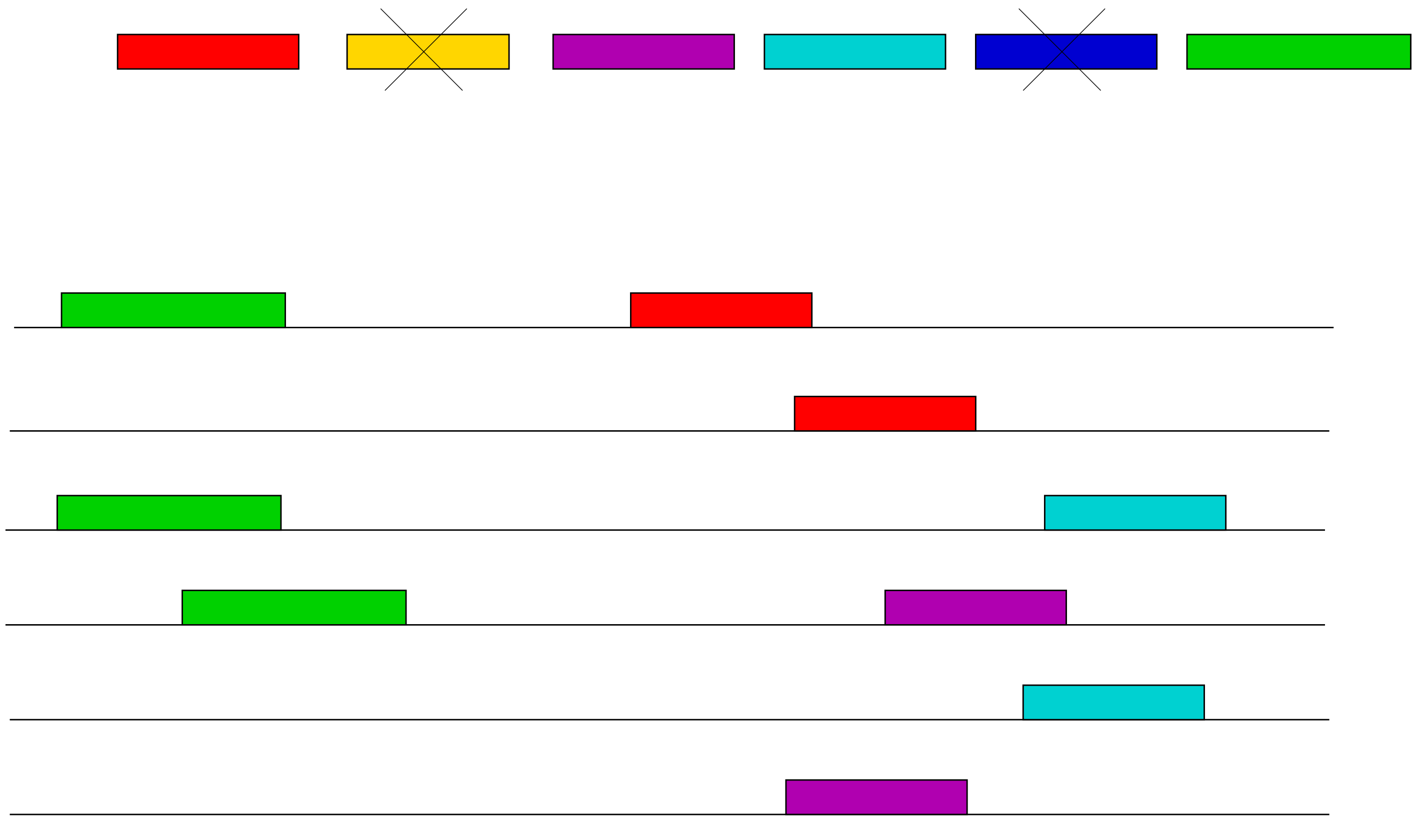
A few facts

- **there can be an exponential number of motifs over Σ' for $q = 2$**
- **there can be an exponential number of maximal such motifs**
  example: $\mathbf{A}^p\mathbf{TA}^p$

What was believed initially

- **there are at most $3n$ irredundant motifs over Σ' for any q, in particular 2**

Irredundant motifs form a "basis"



“Basis” is to be understood in the usual algebraic way

$s =$ **AAATAAA**

Non maximal motifs: A●●A, A●●AA, AA●●A

Non irredundant motifs: A●A, A● ● ●A

Basis of irredundant motifs: AA, AAA, A●A●A, AA●●AA

$$
\begin{array}{ccccccc}
 & A & A & \bullet & \bullet & A & A \\
\oplus & A & \bullet & A & \bullet & A & \bullet \\
= & A & \bullet & \bullet & \bullet & A & \bullet
\end{array}
$$

Basis of irredundant motifs

What was shown later

- there are $\Omega(n^2)$ irredundant motifs over Σ' for $q = 2$

  Ex.: **AAAAAATAAAAAA**

A•AAAA	1,6,8	A• •AAA	1,5,6,8
AA•AAA	1,5,8	A•A•AA	1,4,6,8
AAA•AA	1,4,8	A•AA•A	1,3,6,8
AAAA•A	1,3,8	AA••AA	1,4,5,8
		AA•A•A	1,3,5,8
		AAA• •A	1,3,4,8

eliminated!

Basis of irredundant motifs

What was shown later

- **there are $\Omega(n^2)$ irredundant motifs over Σ' for $q = 2$**

Ex.: $a_4 a_2 u a_2 a_4 vw a_4 a_2 a_3 a_1 xy a_1 a_3 z a_3 a_1$AAAAAATAAAAAA

$a_2 \bullet^8$A•AAAA	-8,-3,-1	A• •AAA	1,5,6,8
$a_4 \bullet^{17}$AA•AAA	-17,-15,-10	A•A•AA	1,4,6,8
$a_1 \bullet^7$AAA•AA	-7,-4,0	A•AA•A	1,3,6,8
$a_3 \bullet^{16}$AAAA•A	-16,-14,-9	AA••AA	1,4,5,8
		AA•A•A	1,3,5,8
		AAA• •A	1,3,4,8

not eliminated anymore!

Basis of irredundant motifs

What was shown later

- there are $\Omega(n^2)$ irredundant motifs over Σ' for $q = 2$

  Ex.: AAAAAATAAAAAA

A•AAAA	1,6,8	A• •AAA	1,5,6,8
AA•AAA	1,5,8	A•A•AA	1,4,6,8
AAA•AA	1,4,8	A•AA•A	1,3,6,8
AAAA•A	1,3,8	AA••AA	1,4,5,8
		AA•A•A	1,3,5,8
		AAA• •A	1,3,4,8

eliminated!

		•							•	A	A	A	A	A	A	T	A	A	A	A	A	A

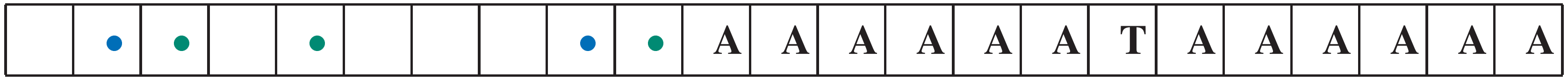
A•AAAA	**1,6,8**	**A• •AAA**	**1,5,6,8**
AA•AAA	**1,5,8**	**A•A•AA**	**1,4,6,8**
AAA•AA	**1,4,8**	**A•AA•A**	**1,3,6,8**
AAAA•A	**1,3,8**	**AA••AA**	**1,4,5,8**
		AA•A•A	**1,3,5,8**
		AAA• •A	**1,3,4,8**

		•		•					•	A	A	A	A	A	A	T	A	A	A	A	A	A

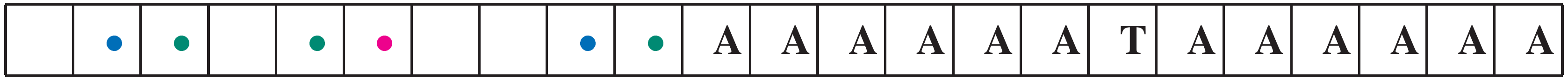
A•AAAA	**1,6,8**	**A• •AAA**	**1,5,6,8**
AA•AAA	**1,5,8**	**A•A•AA**	**1,4,6,8**
AAA•AA	**1,4,8**	**A•AA•A**	**1,3,6,8**
AAAA•A	**1,3,8**	**AA••AA**	**1,4,5,8**
		AA•A•A	**1,3,5,8**
		AAA• •A	**1,3,4,8**

		•		•					•	A	A	A	A	A	A	T	A	A	A	A	A	A

A•AAAA	**1,6,8**	**A• •AAA**	**1,5,6,8**
AA•AAA	**1,5,8**	**A•A•AA**	**1,4,6,8**
AAA•AA	**1,4,8**	**A•AA•A**	**1,3,6,8**
AAAA•A	**1,3,8**	**AA••AA**	**1,4,5,8**
		AA•A•A	**1,3,5,8**
		AAA• •A	**1,3,4,8**



A●AAAA	1,6,8	A● ●AAA	1,5,6,8
AA●AAA	1,5,8	A●A●AA	1,4,6,8
AAA●AA	1,4,8	A●AA●A	1,3,6,8
AAAA●A	1,3,8	AA●●AA	1,4,5,8
		AA●A●A	1,3,5,8
		AAA● ●A	1,3,4,8



A●AAAA	1,6,8	A● ●AAA	1,5,6,8
AA●AAA	1,5,8	A●A●AA	1,4,6,8
AAA●AA	1,4,8	A●AA●A	1,3,6,8
AAAA●A	1,3,8	AA●●AA	1,4,5,8
		AA●A●A	1,3,5,8
		AAA● ●A	1,3,4,8

	•	•		•	•			•	•	A	A	A	A	A	A	T	A	A	A	A	A	A

A•AAAA	**1,6,8**	**A• •AAA**	**1,5,6,8**
AA•AAA	**1,5,8**	**A•A•AA**	**1,4,6,8**
AAA•AA	**1,4,8**	**A•AA•A**	**1,3,6,8**
AAAA•A	**1,3,8**	**AA••AA**	**1,4,5,8**
		AA•A•A	**1,3,5,8**
		AAA• •A	**1,3,4,8**

⋆		•							•	A	A	A	A	A	A	T	A	A	A	A	A	A

A•AAAA	**1,6,8**	**A• •AAA**	**1,5,6,8**
AA•AAA	**1,5,8**	**A•A•AA**	**1,4,6,8**
AAA•AA	**1,4,8**	**A•AA•A**	**1,3,6,8**
AAAA•A	**1,3,8**	**AA••AA**	**1,4,5,8**
		AA•A•A	**1,3,5,8**
		AAA• •A	**1,3,4,8**

★		•			•				•	**A**	**A**	**A**	**A**	**A**	**A**	**T**	**A**	**A**	**A**	**A**	**A**	**A**

A•AAAA	**1,6,8**	**A• •AAA**	**1,5,6,8**
AA•AAA	**1,5,8**	**A•A•AA**	**1,4,6,8**
AAA•AA	**1,4,8**	**A•AA•A**	**1,3,6,8**
AAAA•A	**1,3,8**	**AA••AA**	**1,4,5,8**
		AA•A•A	**1,3,5,8**
		AAA• •A	**1,3,4,8**

⋆		•			•				•	**A**	**A**	**A**	**A**	**A**	**A**	**T**	**A**	**A**	**A**	**A**	**A**	**A**

A•AAAA	**1,6,8**	**A• •AAA**	**1,5,6,8**
AA•AAA	**1,5,8**	**A•A•AA**	**1,4,6,8**
AAA•AA	**1,4,8**	**A•AA•A**	**1,3,6,8**
AAAA•A	**1,3,8**	**AA••AA**	**1,4,5,8**
		AA•A•A	**1,3,5,8**
		AAA• •A	**1,3,4,8**

⋆	•	•			•			•	•	A	A	A	A	A	A	T	A	A	A	A	A	A

A•AAAA	1,6,8	A• •AAA	1,5,6,8
AA•AAA	1,5,8	A•A•AA	1,4,6,8
AAA•AA	1,4,8	A•AA•A	1,3,6,8
AAAA•A	1,3,8	AA••AA	1,4,5,8
		AA•A•A	1,3,5,8
		AAA• •A	1,3,4,8

★	•	•			•	•		•	•	A	A	A	A	A	A	T	A	A	A	A	A	A

A•AAAA	**1,6,8**	**A• •AAA**	**1,5,6,8**
AA•AAA	**1,5,8**	**A•A•AA**	**1,4,6,8**
AAA•AA	**1,4,8**	**A•AA•A**	**1,3,6,8**
AAAA•A	**1,3,8**	**AA••AA**	**1,4,5,8**
		AA•A•A	**1,3,5,8**
		AAA• •A	**1,3,4,8**

⋆	•	•			•	•		•	•	A	A	A	A	A	A	T	A	A	A	A	A	A

A•AAAA	1,6,8	A• •AAA	1,5,6,8
AA•AAA	1,5,8	A•A•AA	1,4,6,8
AAA•AA	1,4,8	A•AA•A	1,3,6,8
AAAA•A	1,3,8	AA••AA	1,4,5,8
		AA•A•A	1,3,5,8
		AAA• •A	1,3,4,8

Number of maximal irredundant motifs at least quadratic

What was shown later

- there are $\Omega(n^2)$ irredundant motifs over Σ' for $q = 2$

  Ex.: $a_4a_2ua_2a_4vwa_4a_2a_3a_1xya_1a_3za_3a_1$AAAAAATAAAAAA

$a_2\bullet^8$A•AAAA	-8,-3,-1	A• •AAA	1,5,6,8
$a_4\bullet^{17}$AA•AAA	-17,-15,-10	A•A•AA	1,4,6,8
$a_1\bullet^7$AAA•AA	-7,-4,0	A•AA•A	1,3,6,8
$a_3\bullet^{16}$AAAA•A	-16,-14,-9	AA••AA	1,4,5,8
		AA•A•A	1,3,5,8
		AAA• •A	1,3,4,8

not eliminated anymore!

Tiling motifs

A maximal motif x is tiled if

$$\mathcal{L}(x) = (\mathcal{L}(y_1) + d_1) \cup (\mathcal{L}(y_2) + d_2) \cup \ldots \cup (\mathcal{L}(y_k) + d_k)$$

for $y_1, y_2, \ldots, y_k$ maximal motifs $\neq x$, and integers $d_1, d_2, \ldots, d_k$

A maximal motif x is tiling if it is not tiled

(Pisanti *et al.*, 2002, related definition by Pelfrene *et al.*, 2002)

What happens to the previous example?

- $a_4 a_2 u a_2 a_4 v w a_4 a_2 a_3 a_1 x y a_1 a_3 z a_3 a_1$ AAAAAATAAAAAA

$a_2 \bullet^{8} A \bullet AAAA$	-8,-3,-1	A• •AAA	1,5,6,8
$a_4 \bullet^{17} AA \bullet AAA$	-17,-15,-10	A•A•AA	1,4,6,8
$a_1 \bullet^{7} AAA \bullet AA$	-7,-4,0	A•AA•A	1,3,6,8
$a_3 \bullet^{16} AAAA \bullet A$	-16,-14,-9	AA••AA	1,4,5,8
		AA•A•A	1,3,5,8
		AAA• •A	1,3,4,8

eliminated again!!

Some properties of tiling motifs for $q = 2$

There are at most $n-1$ tiling motifs

They generate all motifs, *i.e.* they form a basis

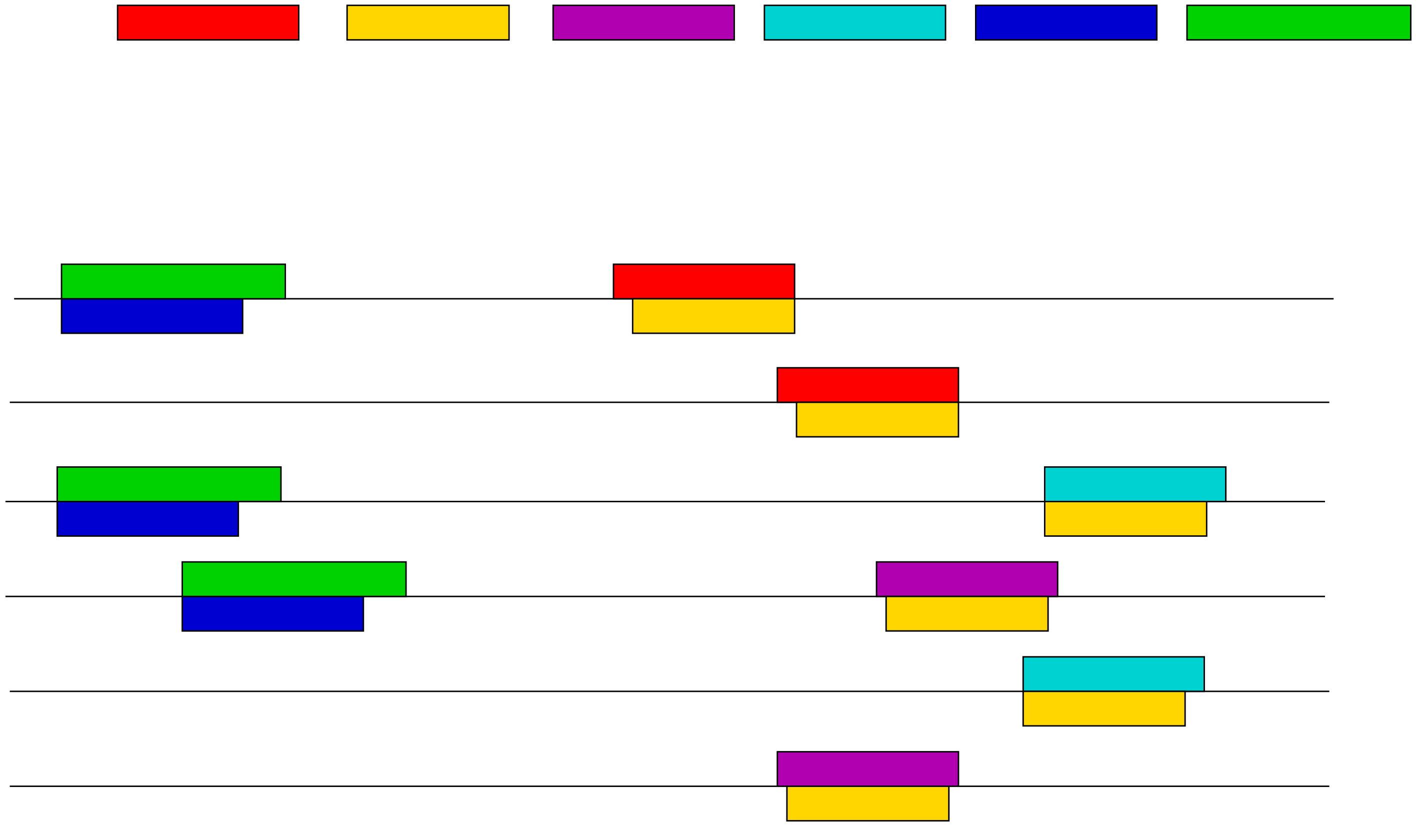
The set of tiling motifs is symmetric

Tiling motifs are not redundant

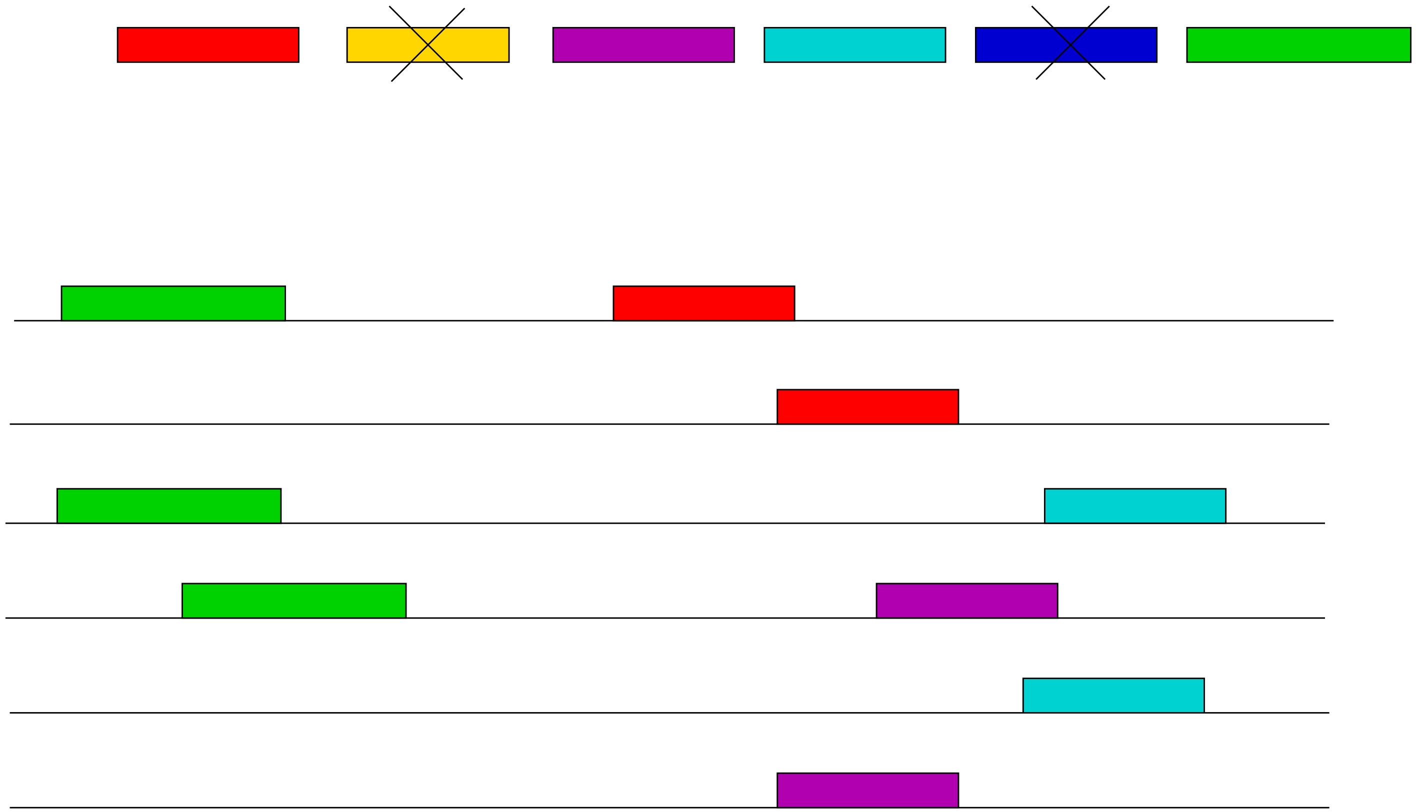
Tiling motifs may be computed in $O(n^2 \log n \log |\Sigma|)$ time

Tiling motifs may be represented in linear space

Tiling motifs



Basis of tiling motifs



(Pisanti *et al.*, MFCS and submitted)

Example (continued)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	positions
s	F	A	B	C	X	F	A	D	C	Y	Z	E	A	D	C	E	A	D	C	
x_1								D	C					D	C			D	C	{7, 13, 17}
x_2							A	D	C				A	D	C		A	D	C	{6, 12, 16}
x_3		A	•	C			A	•	C				A	•	C		A	•	C	{1, 6, 12, 16}
x_4	F	A	•	C		F	A	•	C											{0, 5}
x_5												E	A	•	C	E	A	•	C	{11, 15}

- x_1 **is not maximal**
- x_2, x_3, x_4, x_5 **are maximal**
- x_3 **is tiled by** x_4 **and** x_5: $\mathcal{L}(x_3) = (\mathcal{L}(x_4)+1) \cup (\mathcal{L}(x_5)+1)$
- **Set of tiling motifs**: $\{x_2, x_4, x_5\} = \{ADC, FA\bullet C, EA\bullet C\}$

Merges

- **Merge operator**

$$\sigma \oplus \tau = \begin{cases} \sigma & \texttt{if } \sigma = \tau \\ \bullet & \texttt{otherwise} \end{cases}$$

extends to words

- **Merges of s are trimmed merges between s and its non-empty suffixes**

Example

	A	A	A	A	A	T	A	A	A	A	A	
1		A	A	A	A	•	•	A	A	A	A	{0, 1}
2			A	A	A	•	A	•	A	A	A	{0, 2}
3				A	A	•	A	A	•	A	A	{0, 3}
4					A	•	A	A	A	•	A	{0, 4}
5						•	A	A	A	A	•	{1, 6}
6							A	A	A	A	A	{0, 6}
7								A	A	A	A	{0, 7}
8									A	A	A	{0, 8}
9										A	A	{0, 9}
10											A	{0, 10}

Tiling motifs

1	A A A A • • A A A A	{0, 1}
2	A A A • A • A A A	{0, 2}
3	A A • A A • A A	{0, 3}
4	A • A A A • A	{0, 4}
6	A A A A A	{0, 6}

Properties of merges

- **Merges are maximal motifs**
- **Tiling motifs are merges**
  (therefore there are no more than $n - 1$ tiling motifs)
- **Tiled merges are tiled by other merges**

Computing tiling motifs

Strategy

1. Compute the merges
2. Compute the positions of the merges
3. Discard the merges tiled by other merges

Running time

1. $O(n^2)$
2. $O(n^2 \log n \log |\Sigma|)$ using string matching with don't cares (Fischer, Paterson, 1974)
3. $O(n^2)$ using binary vectors

Number of tiling motifs for any q

- there are $\Omega\left(\frac{1}{2^q}\binom{n-1}{q-1}\right)$ tiling (and irredundant) motifs over Σ' for any q
  Example: $t_k = \mathbf{A}^k\mathbf{T}\mathbf{A}^k$
- the number of tiling motifs is upper bounded by $\binom{n-1}{q-1}$
- the total number of occurrences of all tiling motifs is upper bounded by $q\binom{n-1}{q-1}$

(Tiling basis and counting results by Pisanti *et al.*; similar basis independently found in one case by Pelfrene *et al.*; algorithm by Pisanti *et al.*, related algorithms by Apostolico *et al.*, Pelfrene *et al.*)

OPEN PROBLEMS