An overview of past, current and future work on

MOTIF INFERENCE

In this talk

I'll consider as information

essentially the sequence only

that is, I'll address the problem of sequence motif inference in a string (dispersed repeats) or set of strings (common motifs)

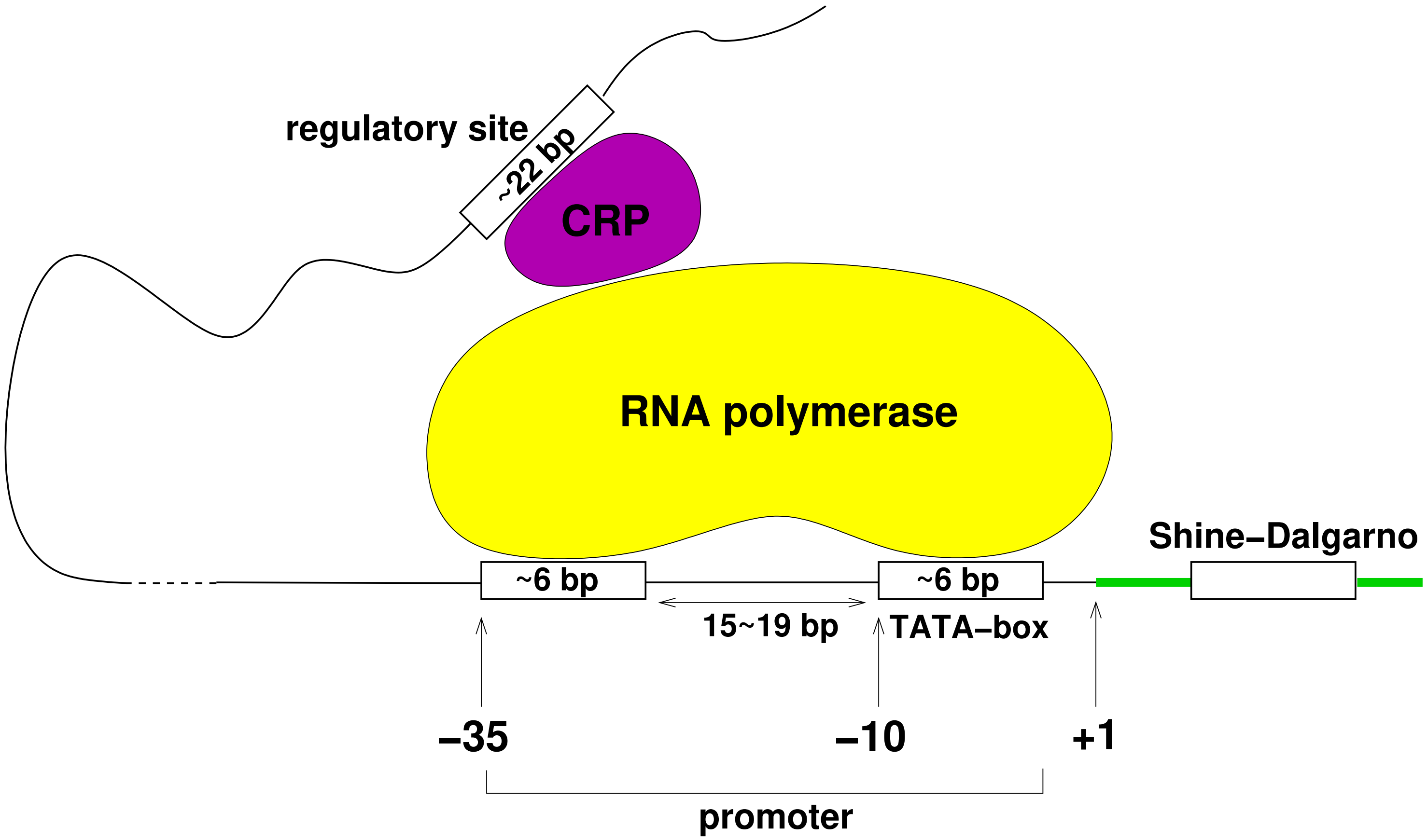
One possible (not unique !) motivation

given a genome (or various genomes)

what are ALL the elements involved in the regulation of the expression of its genes ?

in other words, can we “model” recognition and capture the combinatorics underlying regulation ?

A very simplified (and partial) view on the regulation of gene expression (in the “ simpler” case of a prokaryote)



Motif search $\times$ Motif inference

Search

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

Motif search × 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**

Set of properties

main property: motifs of interest = “conserved” elements

“*I’m afraid*, **said Alice timidly,** *some of the words have got altered.*”

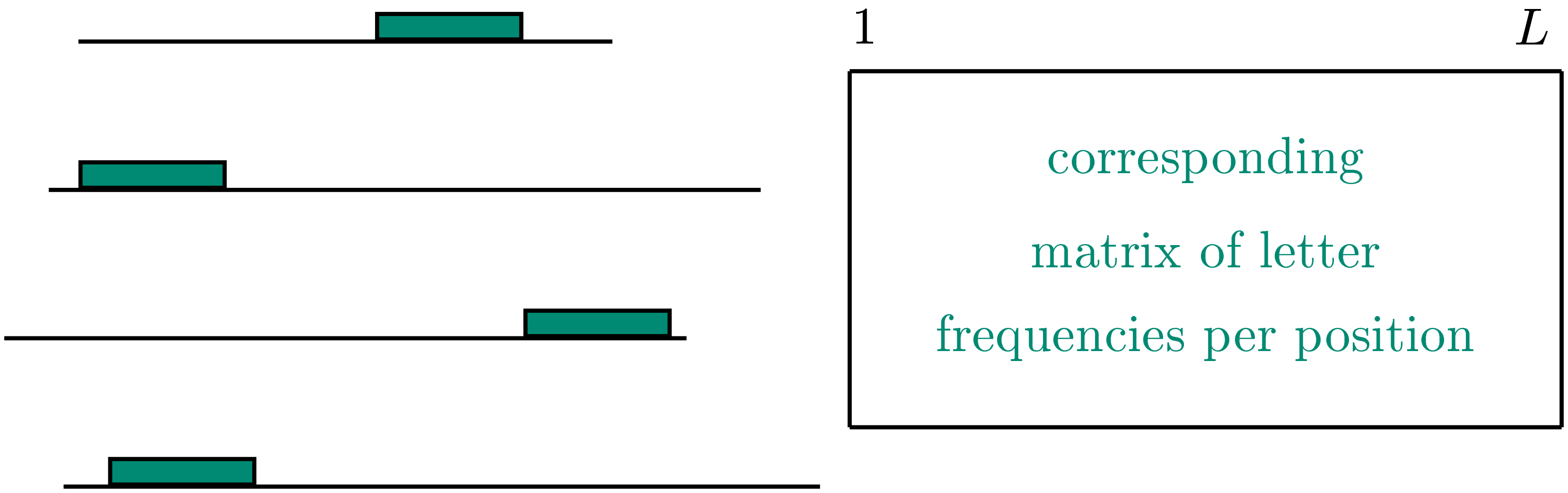
mutations (substitutions, insertions, deletions) may happen that do not destroy function and may even enable to modulate it

in some cases, one would have also to consider man-generated “errors”

“Conserved” elements: “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

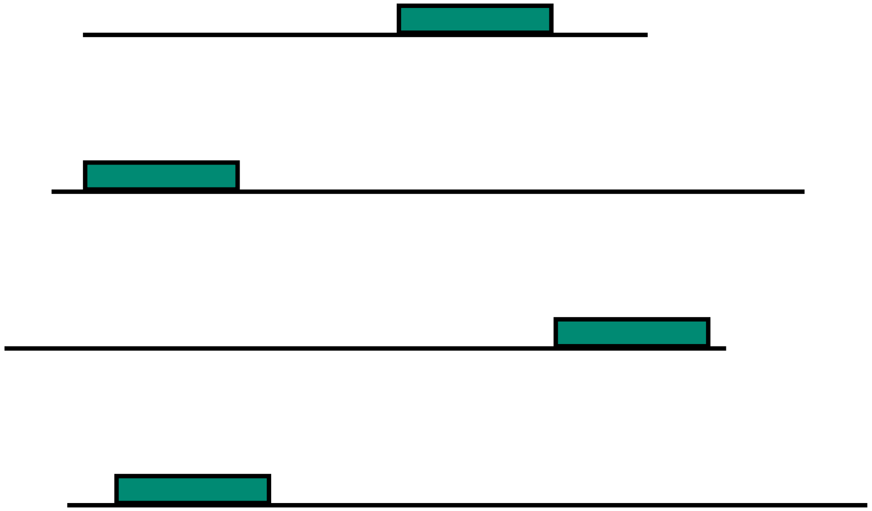


Lawrence (EM 1990, Gibbs 1993); Stormo & Hertz (greedy 1989)
Bailey (MEME 1995); Buhler and Tompa (Projection 2000)
Thijs (MotifSampler 2001); Keich & Pevzner (MultiProfiler 2002)

“Conserved” elements: “Most surprising” motifs

In general, correspond to motifs with lowest probability of occurring by chance given a background

In general also, exact motifs are considered



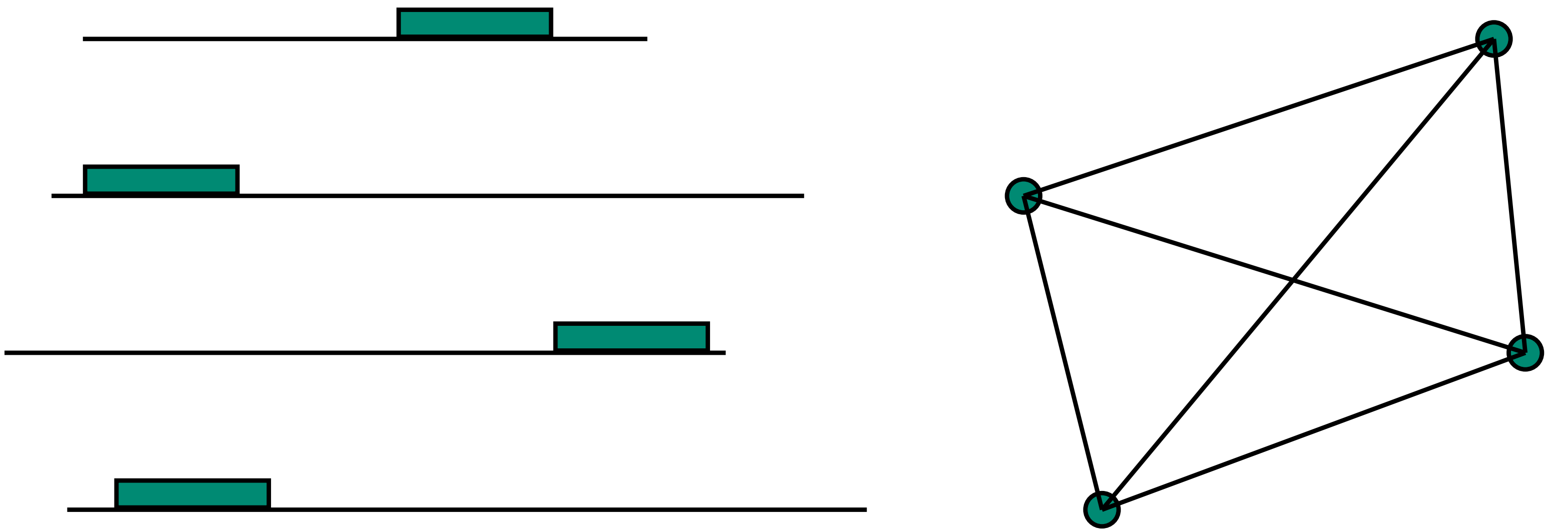
Motifs most often obtained by enumeration in one case, using suffix trees

Schbath (RMES 1996); Apostolico (Verbumculus 1999)
Robin (1999, 2001, 2002); Barash (hyper geometric approach 2001)
Blanchette and Sinha (2001); Denise and Régnier (2001)

“Conserved” elements: Cliques in a multipartite graph

Nodes of the graph: k-mers in the sequences

Edges: between any two nodes if Hamming distance
between k-mers in distinct sequences
is no more than $2e$

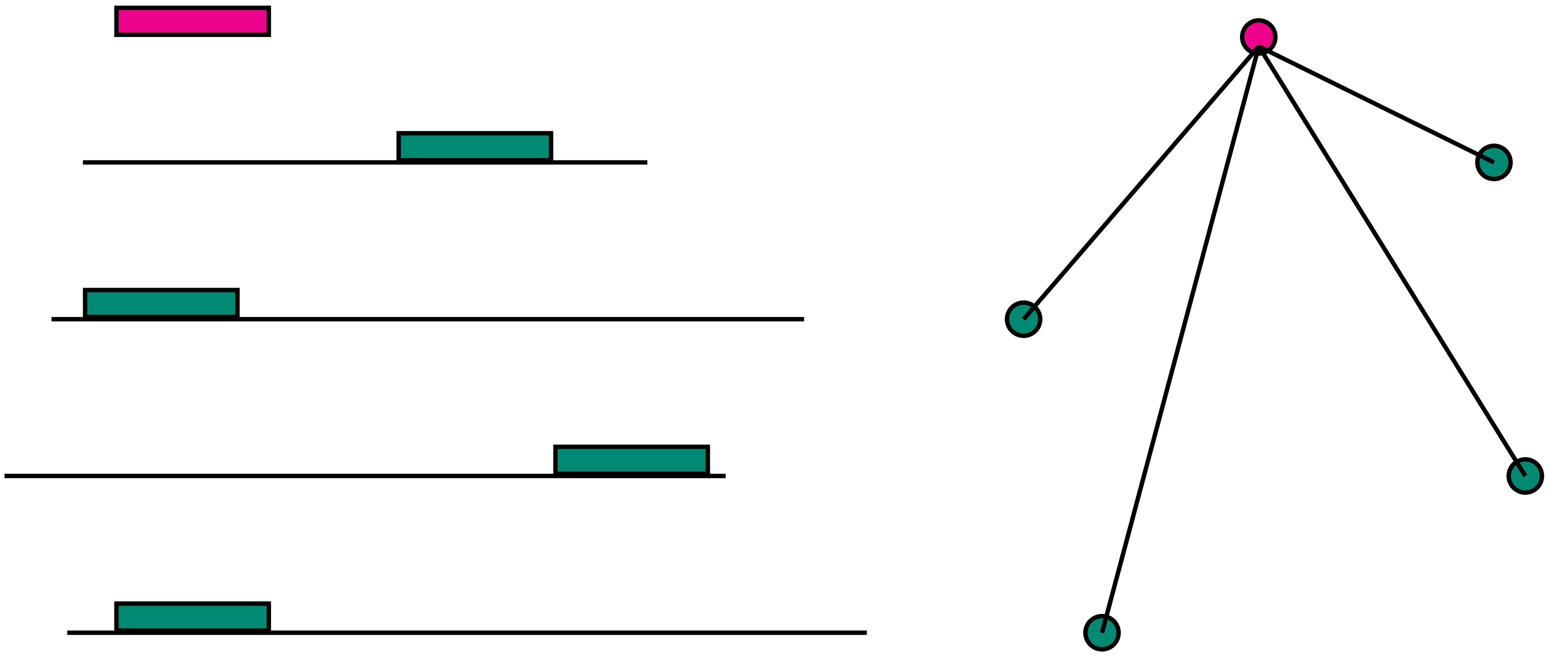


Soldano *et al.* (KMRC 1995) with $e = 0$ but special relation on Σ
Sagot *et al.* (GoK 1994) for common protein structural motifs
Pevzner with Sze (Winnower 2000) or Eskin (MITRA 2002)

“Conserved” elements: “Models”

Model = **Motif + constraints (main one: max. error rate e)**

= “Closest Substring” (in some cases)



Sagot (Moivre/Poivre 1994/5, Klast 1995); Jonassen (Pratt 1995)
Sagot (Combi 1996, Suffix trees 1998); Vilo (Suffix trees 1998)
Marsan (SMILE 1999); Pavesi *et al.* (Weeder 2001)

Motif

"simple pattern"	**"network expression" (G. Myers)**
TATAAT	**TA[AT]N[AT]T**
TTGNCA	
RFXCP	**[ILMV][ASG]XXC[ILMV]H**

where N or X is the don't care symbol

$\Sigma_{DNA/RNA}$ = **{A, C, G, T(U for RNA)} (nucleotides)**
$\Sigma_{protein}$ = **{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y} (amino acids)**

Further properties for the last two or for all definitions

Quorum:	**minimum number of sequences containing the string**
Length of motif:	• **within an interval (Sagot; Marsan)** • **estimated (Bailey, Lawrence)** • **fixed value (Pevzner, Pavesi etc)**

The choice between definitions is not "algorithmically innocent"

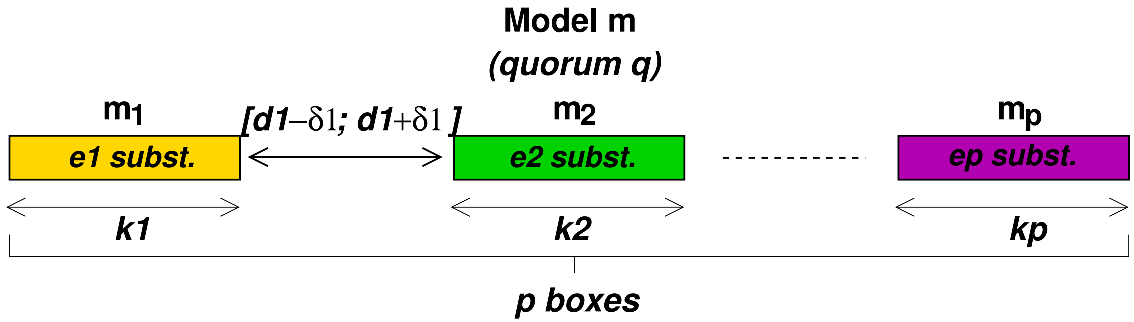
"Most surprising" set of words:	**NP-hard**
Cliques in a multipartite graph:	**NP-complete** **but reasonable ($O(n^{2.46})$)** **in practice**
"Models":	**Exponential but** **in max. error rate** **not in number sequences** **linear in tot. seq. length**

Open question: "models" fixed-parameter tractable if parameterized by max. error rate ?

Towards modelling the combinatorics of regulation

"Structured models"

an ordered collection of p boxes, p maximum numbers of differences (e), p intervals of lengths (k), $p-1$ intervals of distances (between successive boxes in the collection) ($d \pm \delta$) (this is called *spacers*)

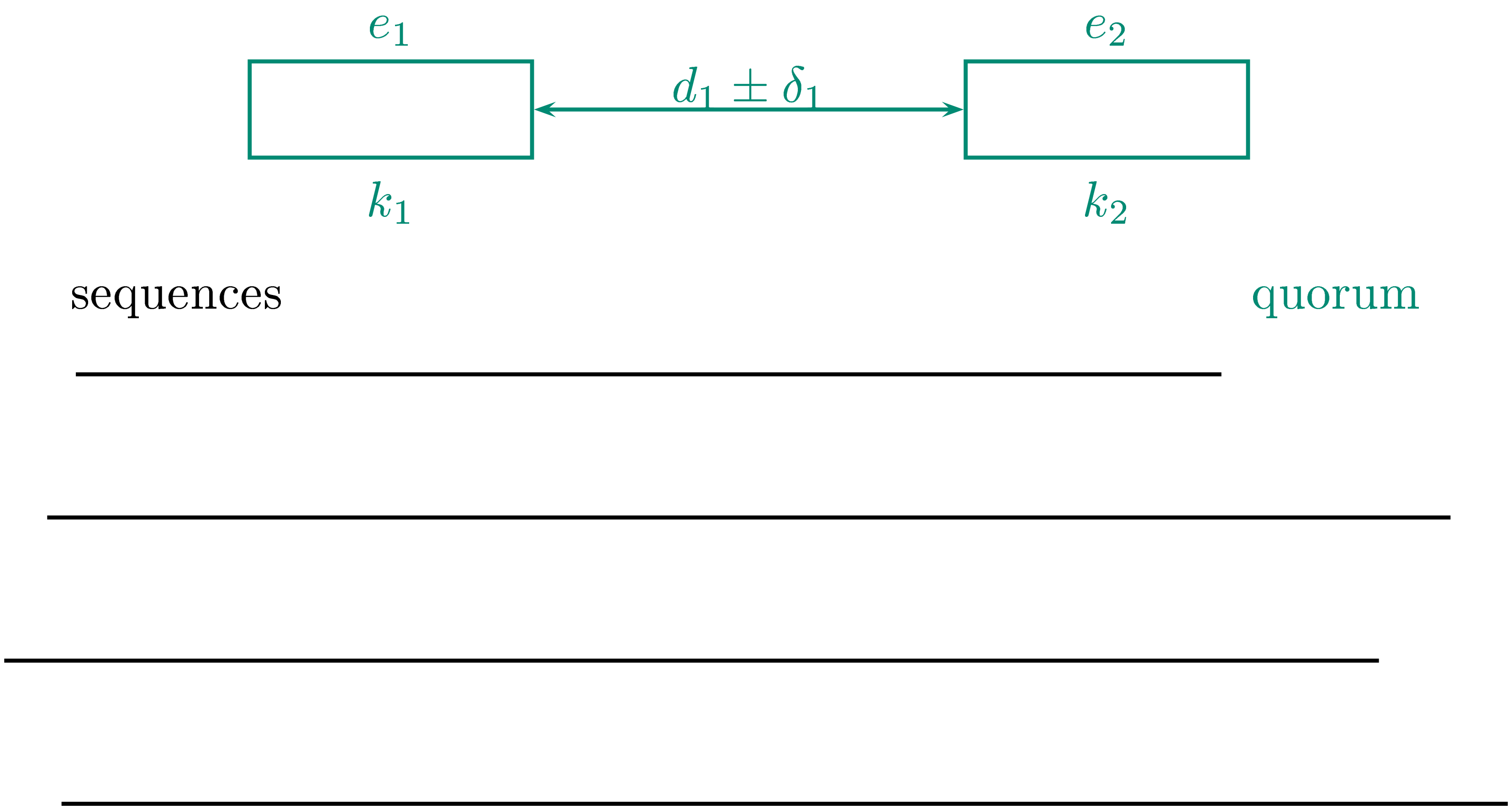


Smile (LNCS, 1380:111-127, 1998; *JMB*, 297:335-353, 1999; *Res. Microb.*, 150:779-799, 2000; Recomb'00; *JCB*, 7:345-362, 2000)

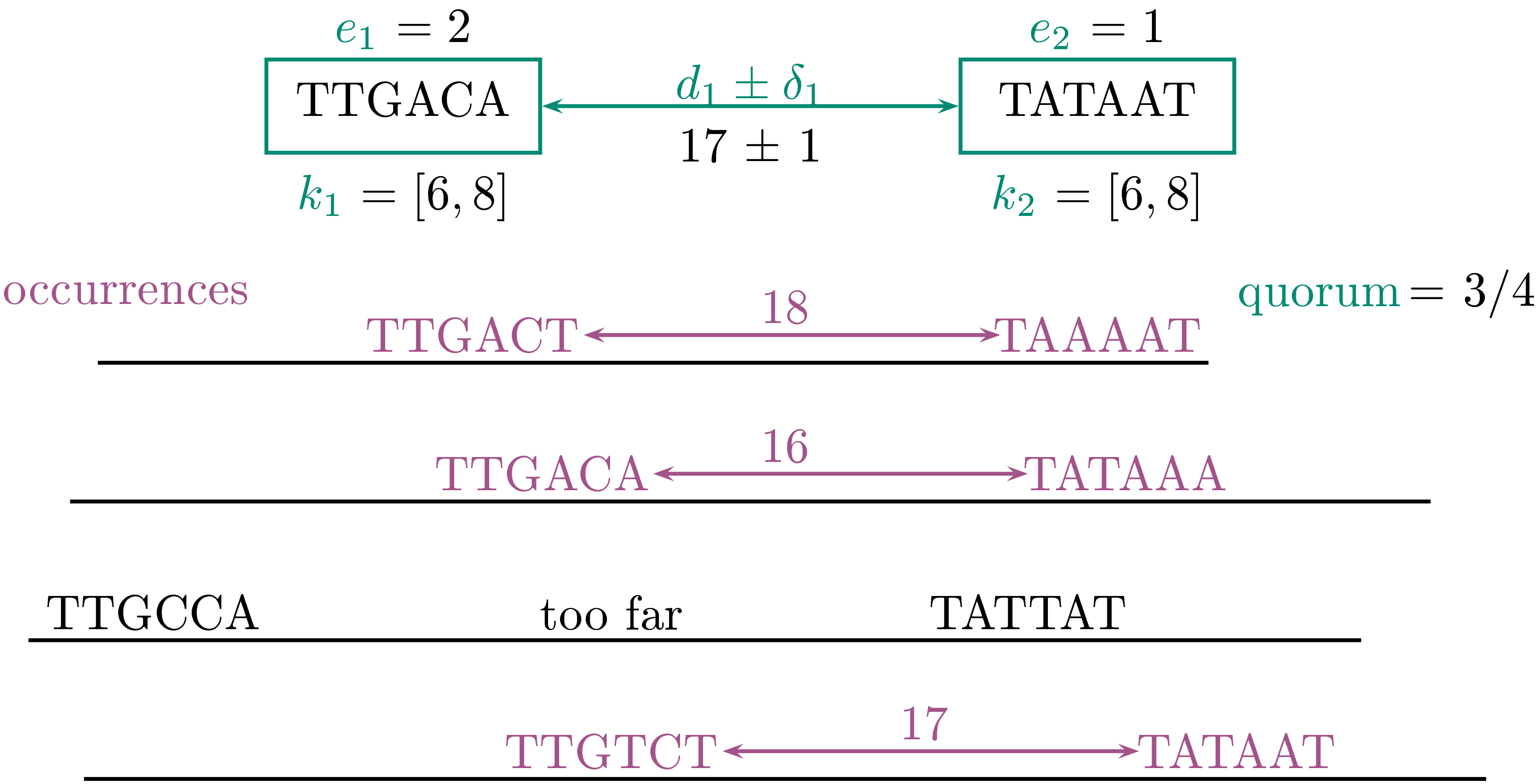
Extended Smile (L. Marsan *et al.*, paper in preparation)

Laurent Marsan

An example with two boxes



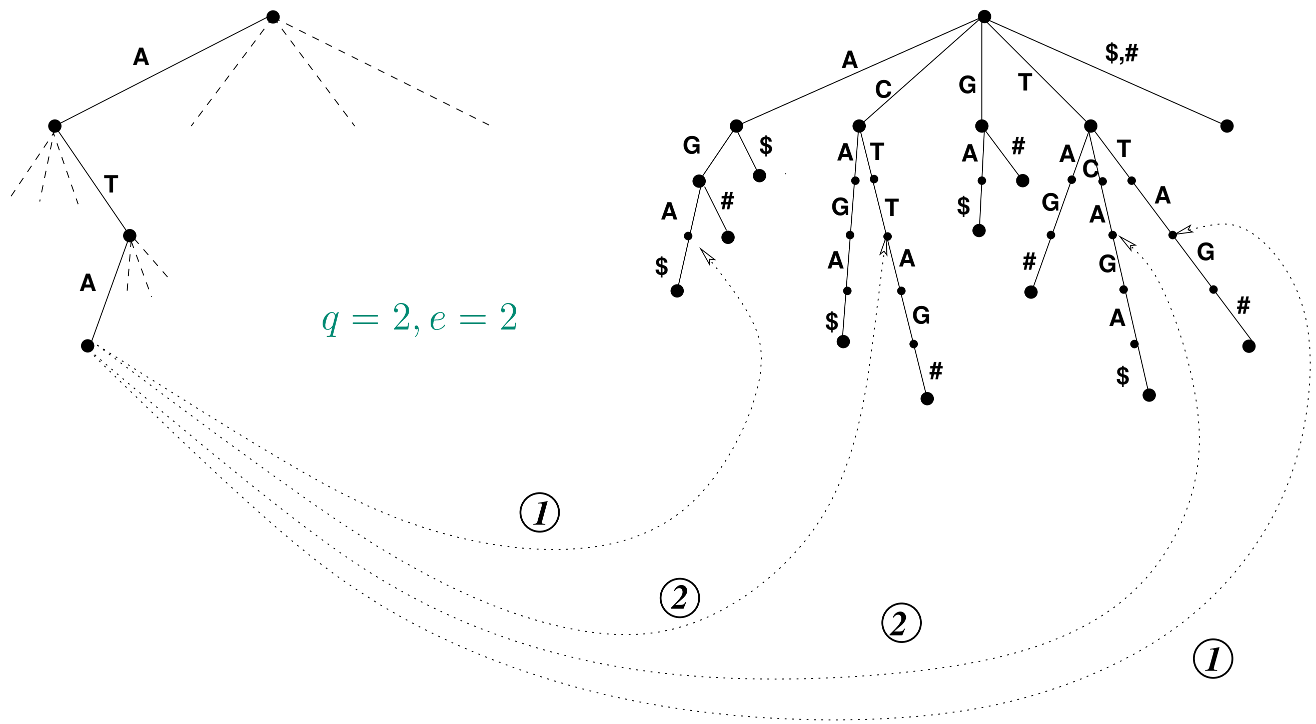
An example with two boxes



A whiff of the algorithm: A story of traversing two trees with a dynamic programming algorithm

Virtual trie of models

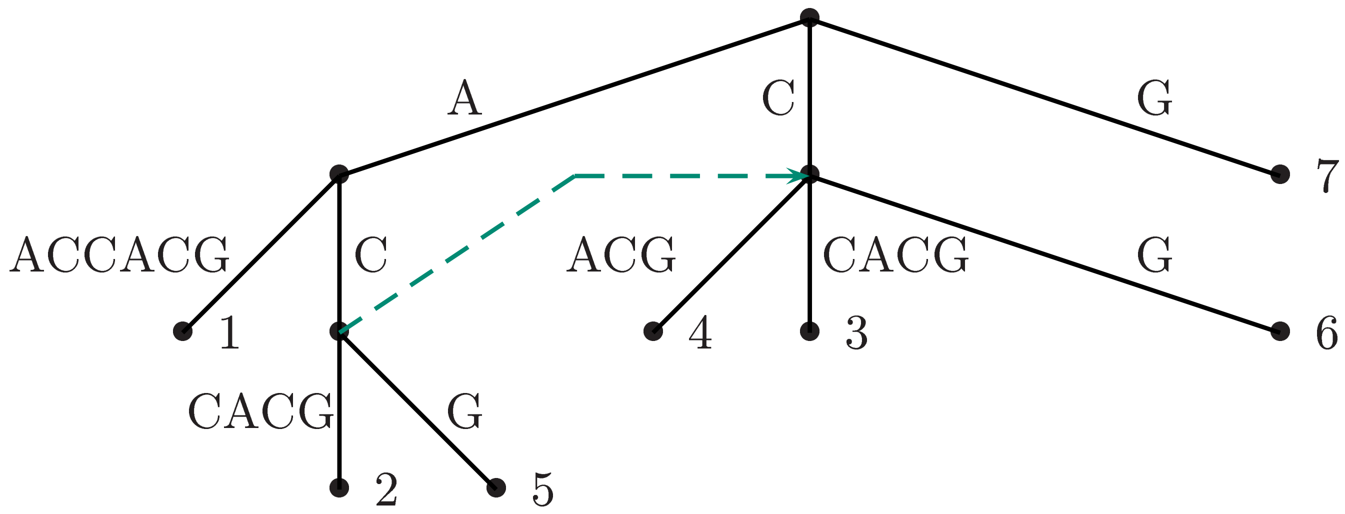
(Generalized) Suffix tree of strings TCAGA\$ and CTTAG#



$O(nNk^e)$ for constant-sized alphabet

Suffix tree: A reminder

Suffix tree for string AACCACG: a tree representing all suffixes of the string

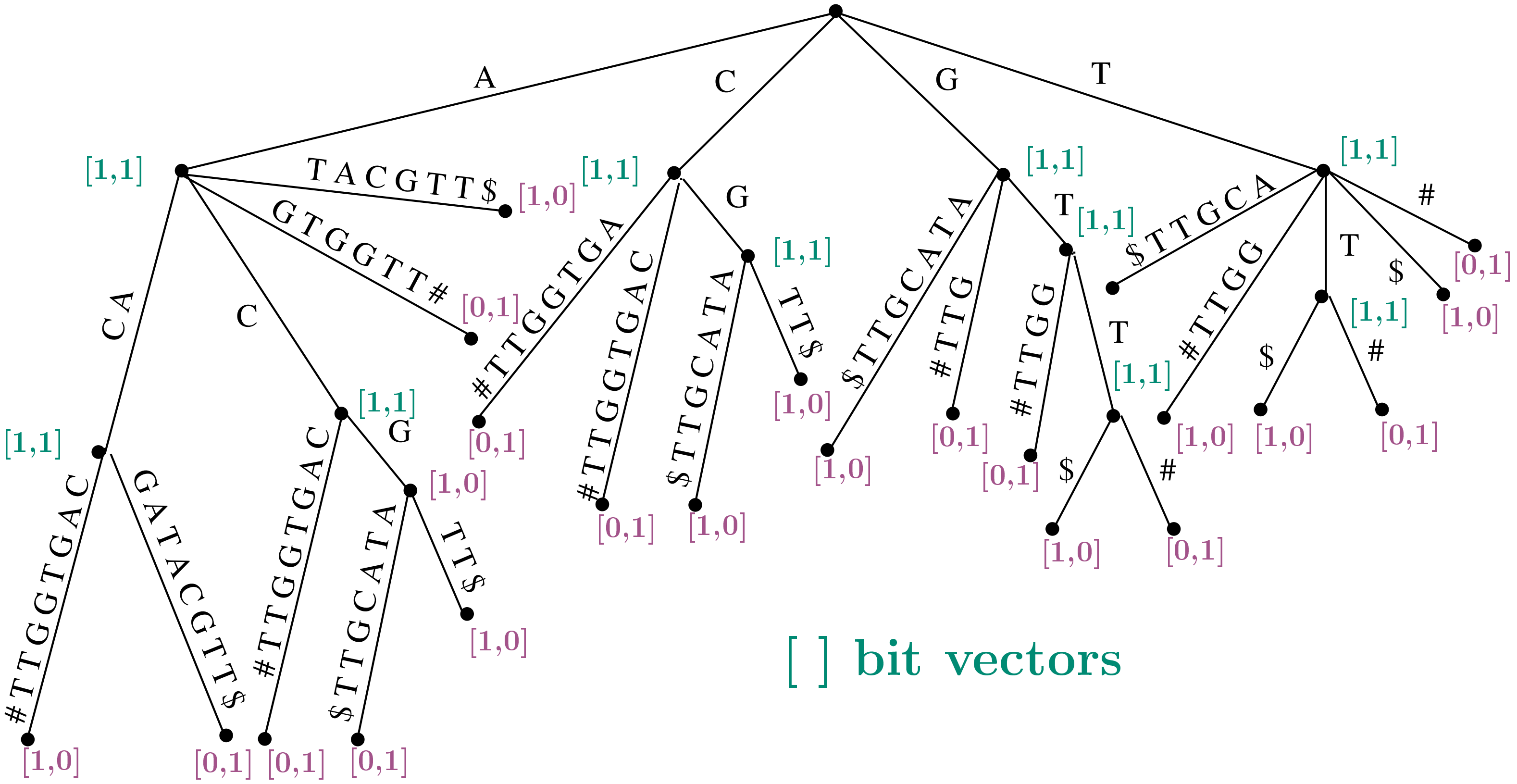


key feature: concatenation of labels of arcs from root to leaf i spells suffix of input string starting at position i

$\rightarrow$ **(suffix link) links node spelling αm to node spelling m**

Generalized suffix tree

Strings AACGATACCTT$ and AACCAGTGGTT#



[ ] bit vectors

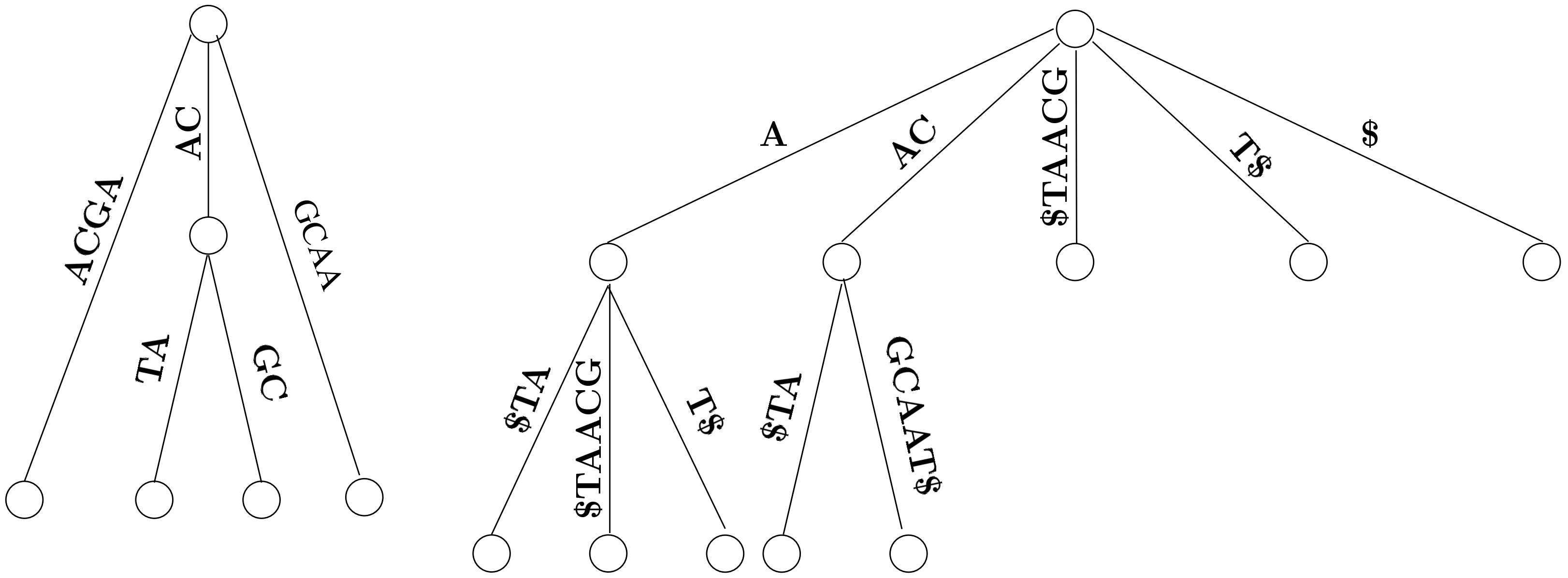
Key ideas

Use of an index (a generalized suffix tree) for representing the strings

> **in fact, what is used is a new data structure: a generalized k-depth factor tree**
>
> > **(Allali and Sagot, submitted)**

Factor tree (example for all factors of length *exactly* 4 of the string CAGCAAT)

Construction in the same time and space complexity: $O(n)$



factor tree

suffix tree for the same string

Can be "added" to any suffix tree (or CDAWG ?) construction, including Kurtz *et al.*'s space economical one

Julien Allali

Key ideas

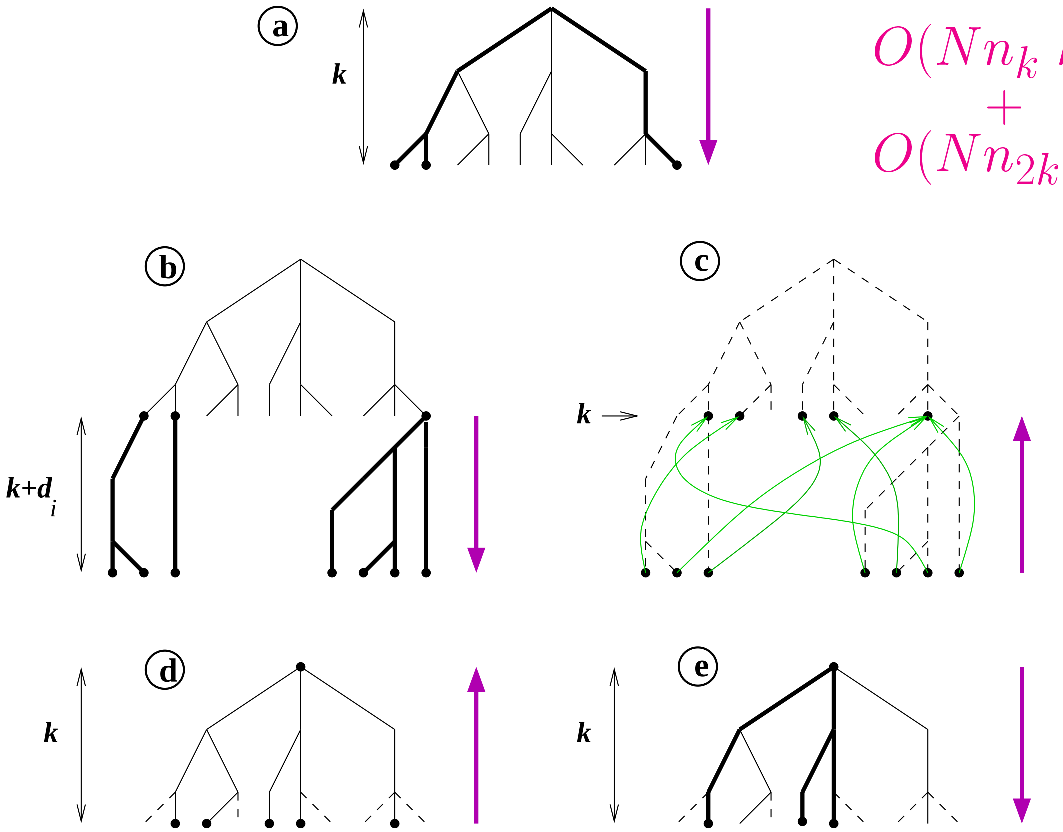
Use of an index (a generalized suffix tree) for representing the strings

in fact, what is used is a new data structure: a generalized k-depth factor tree

(Allali and Sagot, submitted)

For models of two or more boxes, use of a generalized dynamic factor tree

Generalized dynamic factor tree



$$O(Nn_k\, k^{pe}) + O(Nn_{2k+d_{max}}\, k^{(p-1)e})$$

Key ideas (continuation)

Models having one prefix which does not meet the constraints will not meet it either

descent down the trees may end at first such prefix

Key ideas (continuation)

Models having one prefix which does not meet the constraints will not meet it either

descent down the trees may end at first such prefix

Models m having a suffix which we know (from previous calculations) will not enable m to reach the minimum length constraint need not be extended further

descent down the trees may end as soon as such situation is met

(Marsan, Pisanti and Sagot, paper in preparation)

Statistical evaluation of the motifs found by SMILE

Mostly (so far) by simulation

Against random sequences

probability calculated based on the frequency of occurrence, with the same max. error rate, of the motifs in random sequences with same k-mer composition for a given value of k

Against "counter-examples"

probability calculated based on the frequency of occurrence, with the same max. error rate, of the motifs in sequences not supposed to contain them

Statistical evaluation of the motifs found

More recently

A priori **probability occurrence, with the same max. error rate, of motif(s) in random sequences under a Markov Model of order 0 or 1 (2 boxes maximum for now)**

(S. Robin, J.J. Daudin, H. Richard, M.-F.Sagot, S. Schbath, JCB, in press)

Still a LOT of work to be done

Biological applications: the first (really serious) one

Helicobacter pylori **with Anne Vanet**

Helicobacter pylori

Main result

Detected motif:
TTAAGC x(3-4) TTTTAA x(11-13) TATAAT

A. Vanet, L. Marsan, A. Labigne, M.-F. Sagot, *JMB*, 297:335-353, 2000

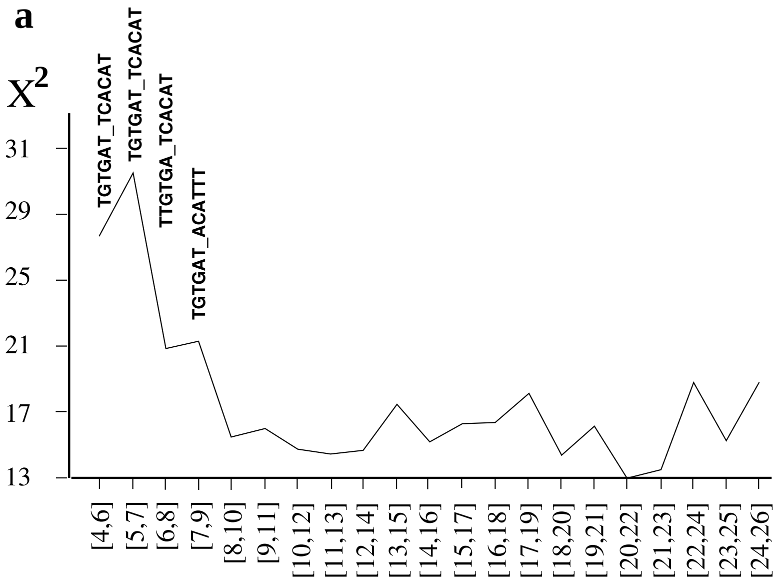
SMILE initially tested in "known" situations with a comparison to other approaches

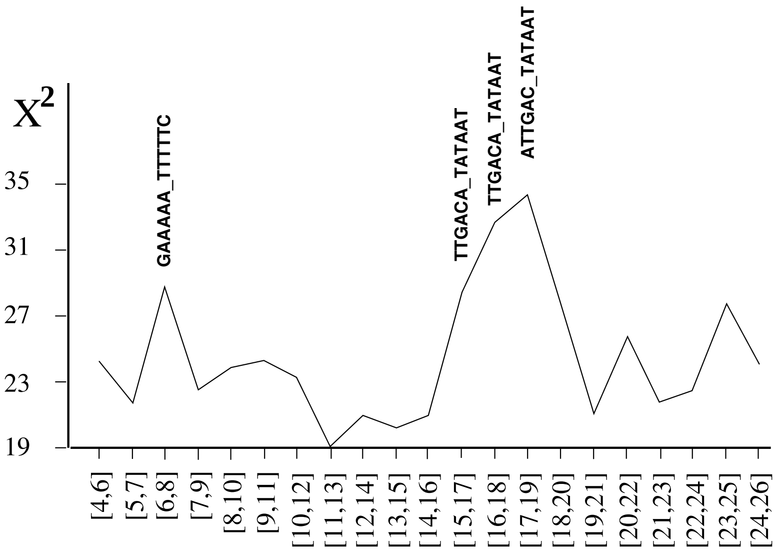
Escherichia coli **and** *Bacillus subtilis*

both "clean" reduced datasets and big "genomic" datasets

A. Vanet, L. Marsan, M.-F. Sagot, *Res. in Microb.*, **150:779-799, 1999**

Basic result





Some further, currently in progress, biological applications

Vertebrates

splicing enhancer sequences

Mycobacterium tuberculosis **and** *leprae*

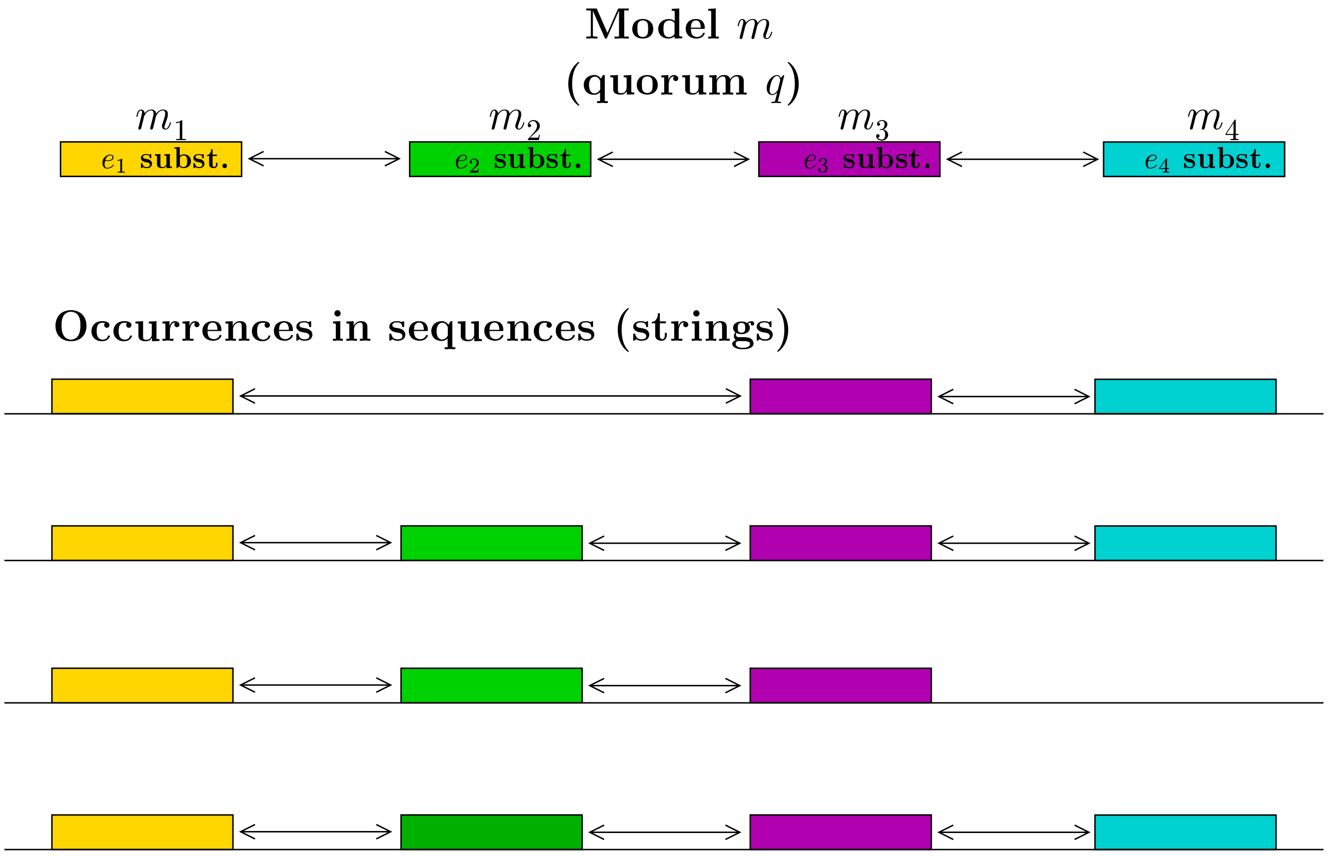
promoter and regulatory sequences in two very closely related genomes, one of which seems to be "shrinking" (around 1500 genes have already "disappeared" – but 165 are "new")

Synechocystis **PCC 6803**

promoter and regulatory sequences in a bacterial genome that seems not very "bacterial-like"

etc

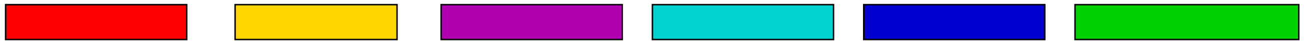
Part of the work in progress: "meta"-conservation



Marsan *et al.*, in preparation

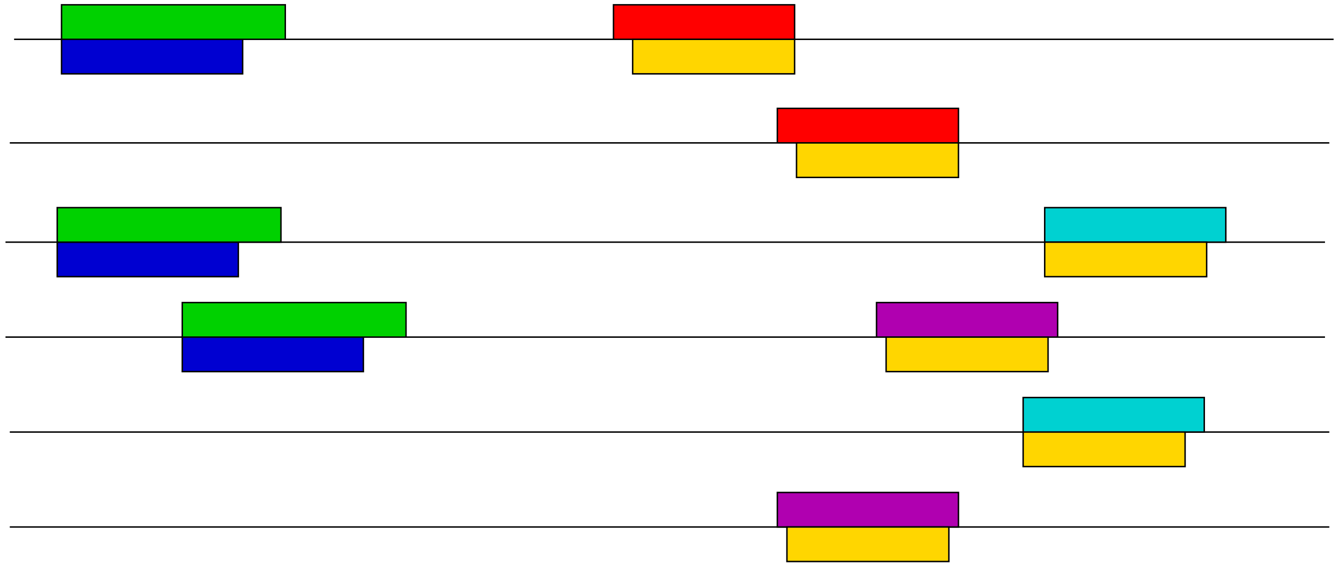
Part of the work in progress: basis of motifs

Models (one box) found with no differences and written over alphabet DNA/protein plus wild card



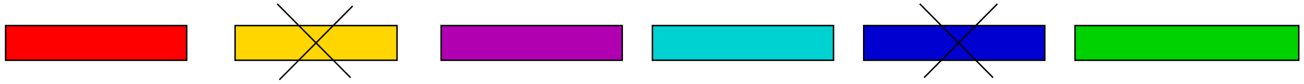
Occurrences in sequences (strings)

keep only
"tiling" motifs



Part of the work in progress: basis of motifs

Models (one box) found with no differences and written over alphabet DNA/protein plus wild card



Occurrences in sequences (strings)

keep only “tiling” motifs

Pisanti *et al.*, A new basis for repeated motifs in biological sequences, submitted

Nadia Pisanti

Some surprising findings in *E. coli*: the road ahead is long

CRP binding site

Data sets

1. non coding sequences upstream of genes regulated by CRP according to RegulonDB (Collado-Vides *et al.*)

2. non coding sequences (on both strands) in the genome (*i.e.*, upstream of all genes of *E. coli*)

Preliminary results

TGTGA-x(6)-TCACA "very well" identified (by all) in the genome: 156 occ. with at most 1 difference

same consensus with at most 1 difference present in only one third of the sequences of RegulonDB

How to interpret ? Which consequences ?

A summarized view of the quality of the results obtained by us, and by other methods

The nice points

we seem to be still ahead at least in some cases and we can treat sometimes much bigger datasets

The provocative points

everybody is reasonably good at detecting what the human eye would (almost) be capable of detecting without any help

everybody is as bad as the others at detecting subtle things, certainly at getting near the original goal

A few more perspectives

“Rare” motifs

Long repeats

Motif + Gene Inference

RNA structural motifs (palindromes, pseudo-knots and triple helices) past and future

What about the “challenge” problem ?