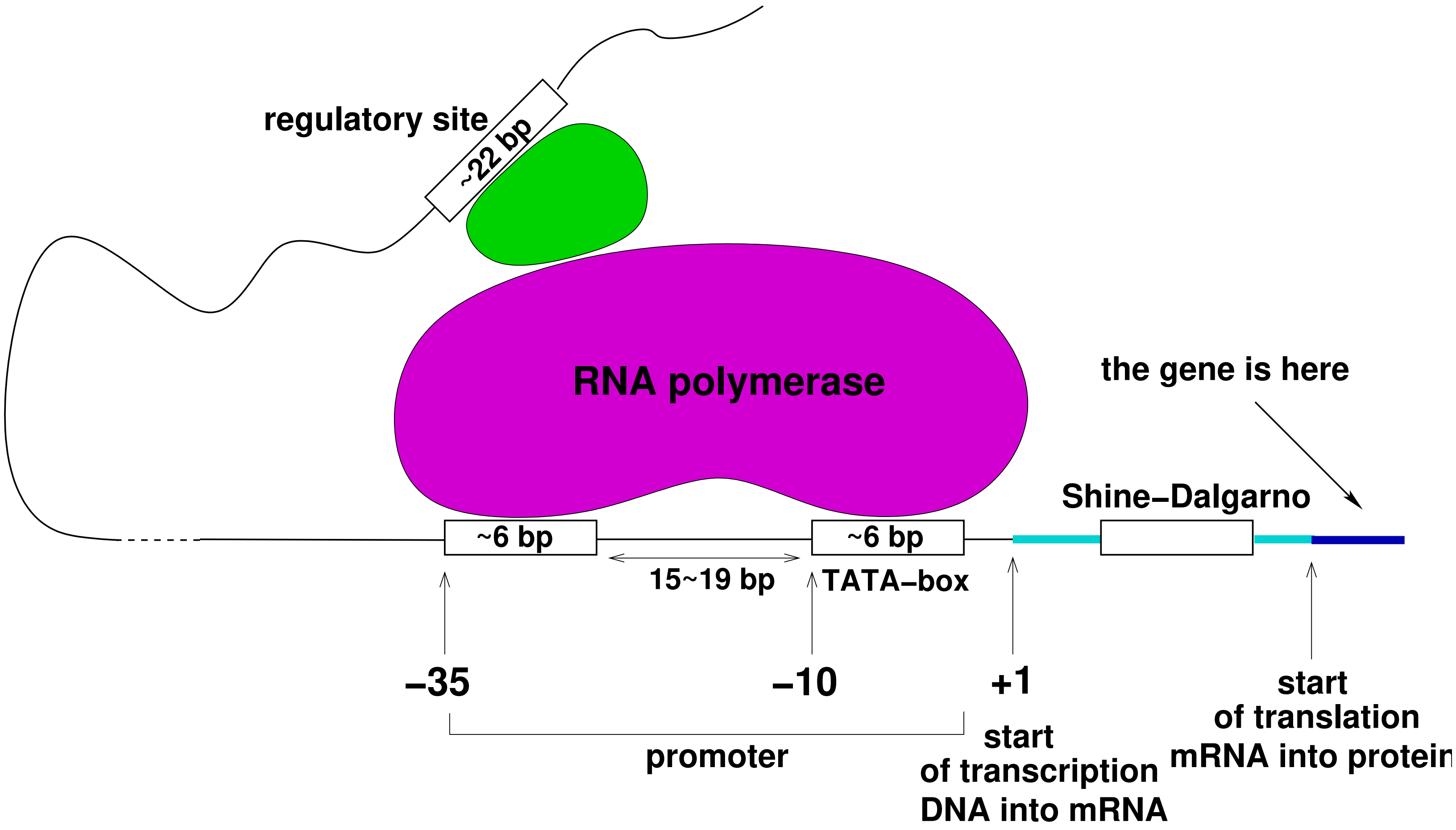
Biological motif inference

Marie-France Sagot

Inria Rhône-Alpes – Projet HELIX
Laboratoire de Biométrie et Biologie Évolutive
Université Claude Bernard, Lyon I

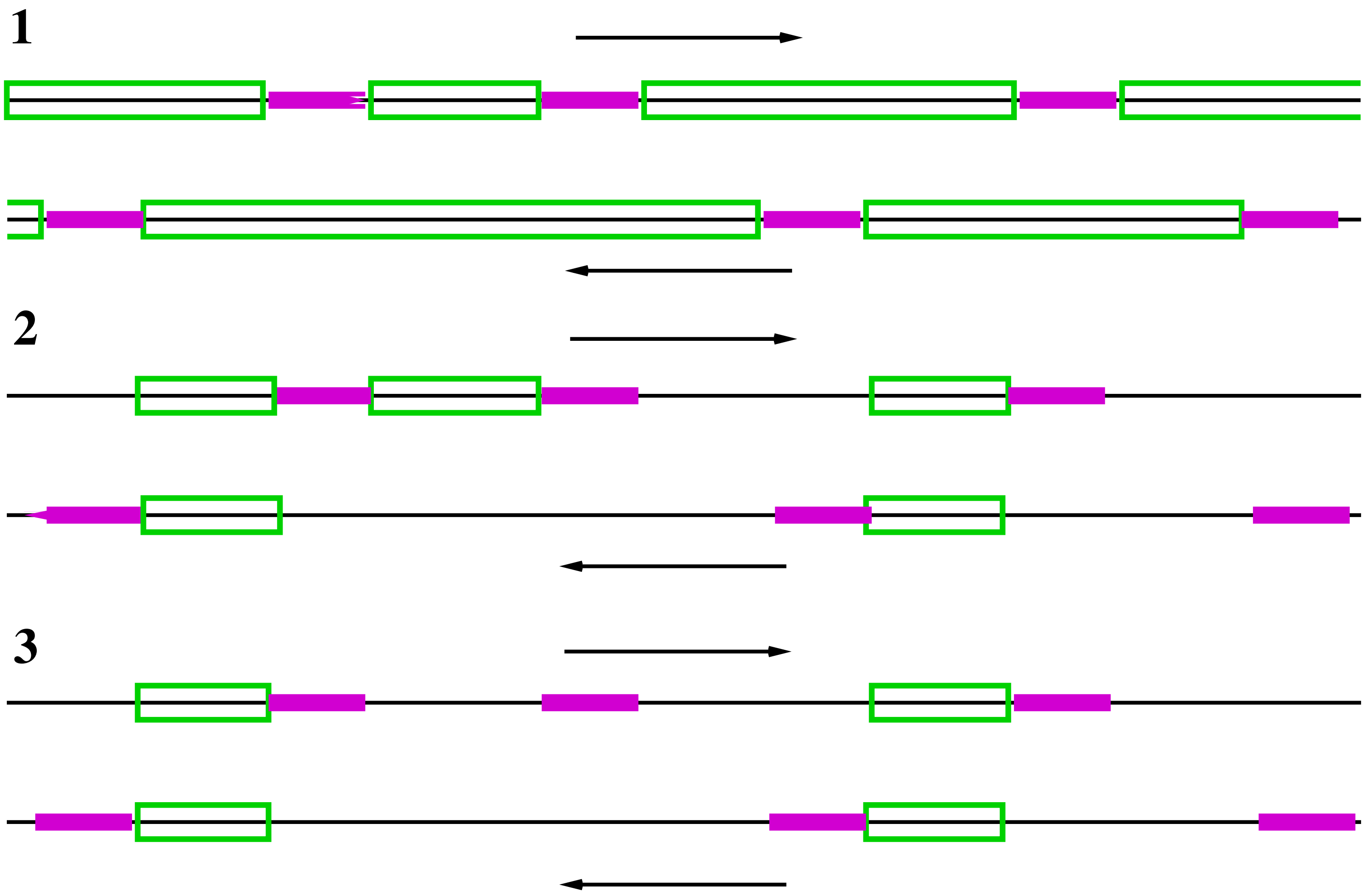
Initial biological motivation

Promoter sequences in bacteria, in particular *Helicobacter pylori*



“Typical” input

A set of non coding sequences upstream of genes in a bacterium



Motif search x 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: See previous slide

Properties

main property: motifs of interest = "conserved" elements

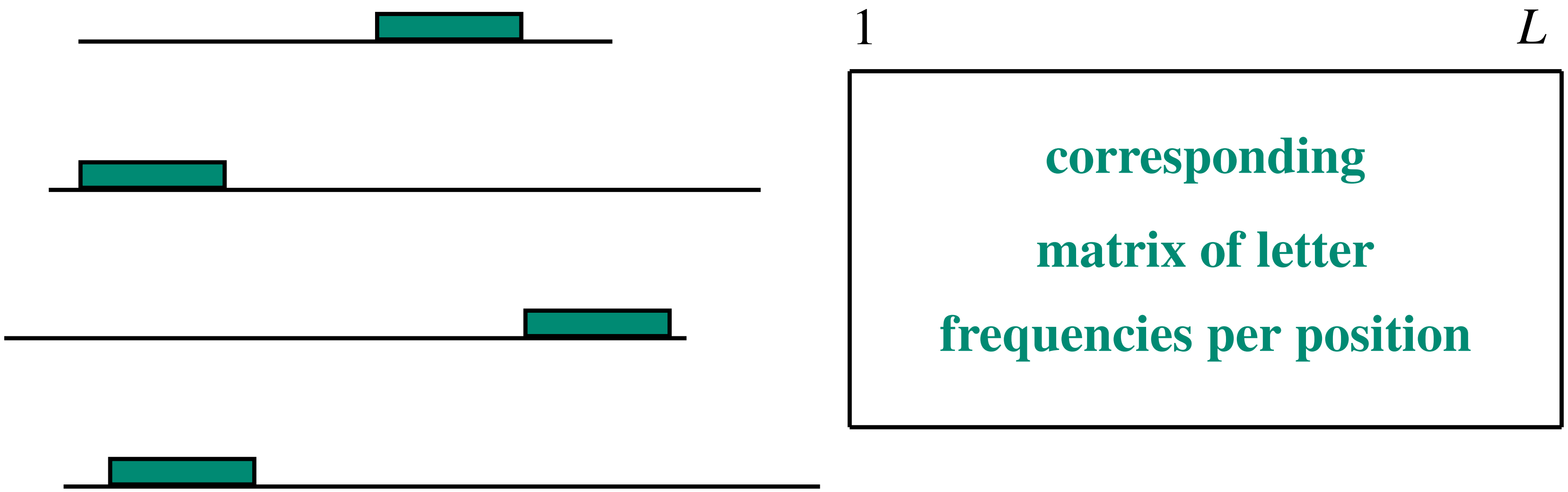
mutations (substitutions and in some cases insertions and deletions) may happen that do not destroy function and may even enable to modulate it

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

The question is: how to model "conservation"?

“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



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

Formal definition of the: “Most Surprising Set(s) of Words Problem”

INPUT:

data: a set of N sequences

parameters: a length k

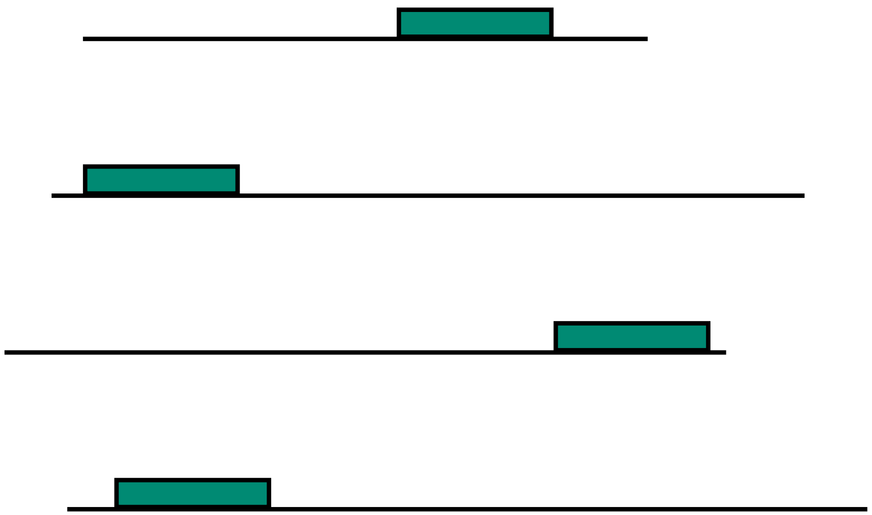
OUTPUT:

the set(s) of N words of length k, each belonging to a distinct sequence, that has maximum relative entropy

“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)

Formal definition of the: “Most Surprising Motifs Problem”

INPUT:

data: a set of N sequences

hypothesis: a certain statistical model

parameters: a length k

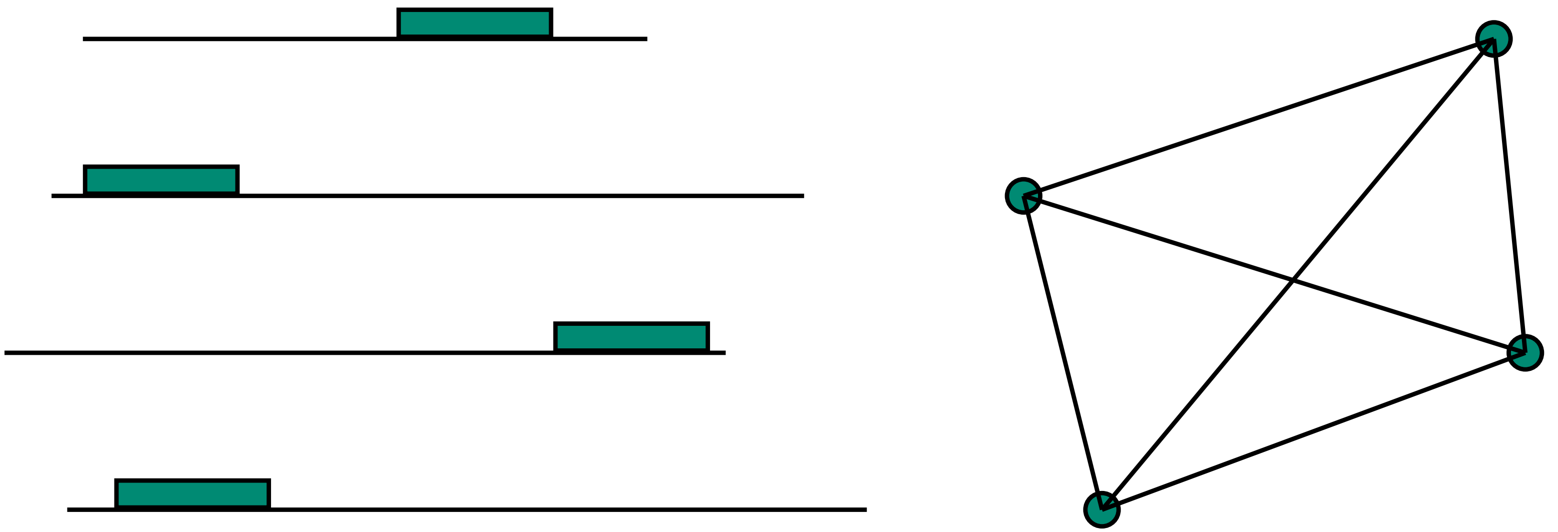
OUTPUT:

all words of length k, sorted according to their probability

“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)

Formal definition of the: “Motifs as Cliques Problem”

INPUT:

data: a set of N sequences

parameters: a length k, a maximum allowed difference $2e$

MODELLING:

- N**-partite graph** $G(V,E)$ **with** $V = \{V_1, \ldots, V_N\}$
- **Nodes** $v \in V_i$ **represent all** k**-mers in sequence** i
- $(v,w) \in E$ **with** $v \in V_i, w \in V_j$ **if corresponding** k**-mers at Hamming distance at most** $2e$

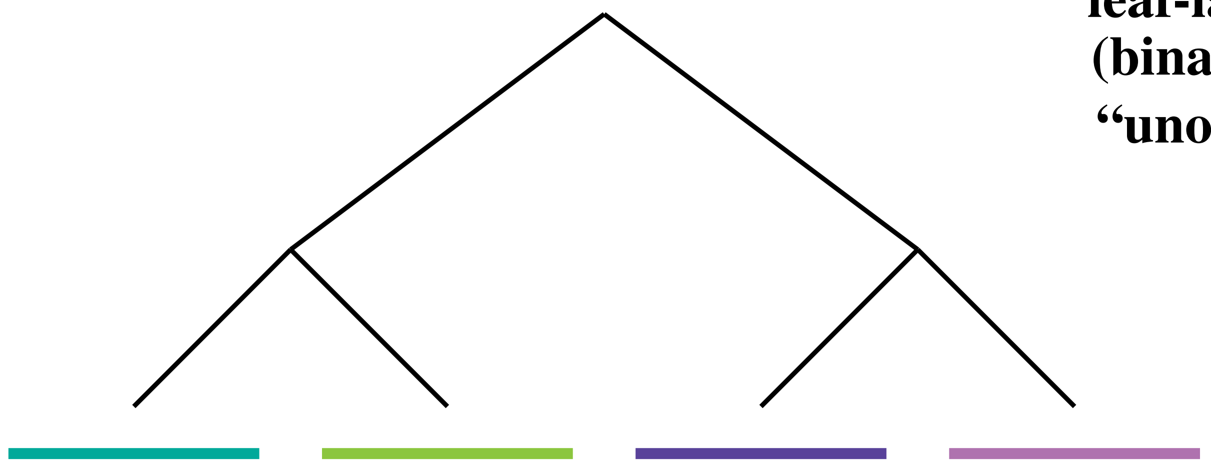
OUTPUT:

all N**-cliques in** G

Trees representing the evolutionary history of a set of organisms

Data (here): for each organism, a set of characters over a discrete set of states

(rooted)
leaf-labeled
(binary)
"unordered"



Phylogenetic tree

Find most parsimonious tree – "Small Phylogeny Problem"

* ancestors

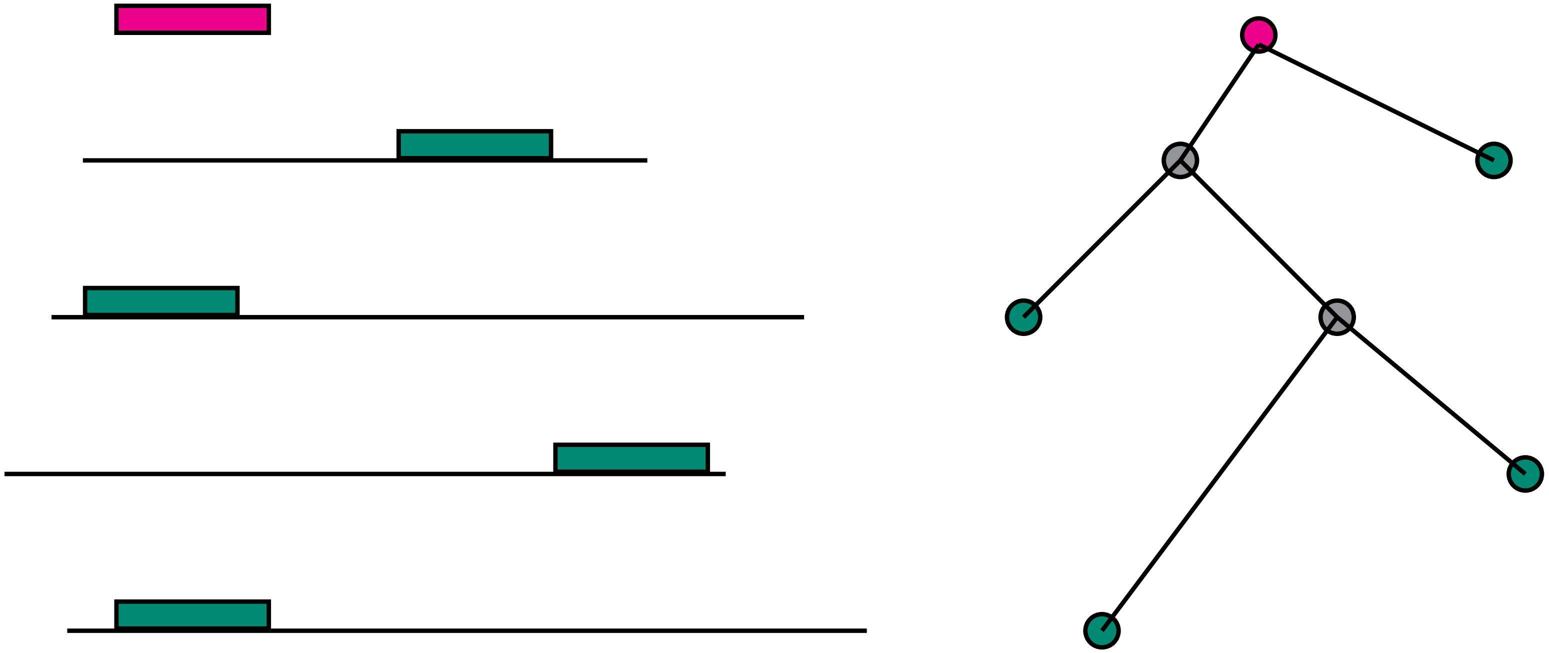
(rooted)
leaf-labeled
(binary)
"unordered"

"Conserved" elements: Ancestors in most parsimonious tree

Leafs are one *k*-mer for each sequence

Motifs correspond to ancestors of most parsimonious trees for which sum of mutations along all edges is at most *E*

Requires orthologous sequences and a tree



Blanchette and Blanchette *et al.* (Phylogenetic footprinting 2000-2003)

Formal definition of the: “Motifs as Ancestors of Most Parsimonious Trees Problem”

INPUT:

data: a set of N sequences and a phylogenetic tree

parameters: a length k, a max. (global!) allowed difference E

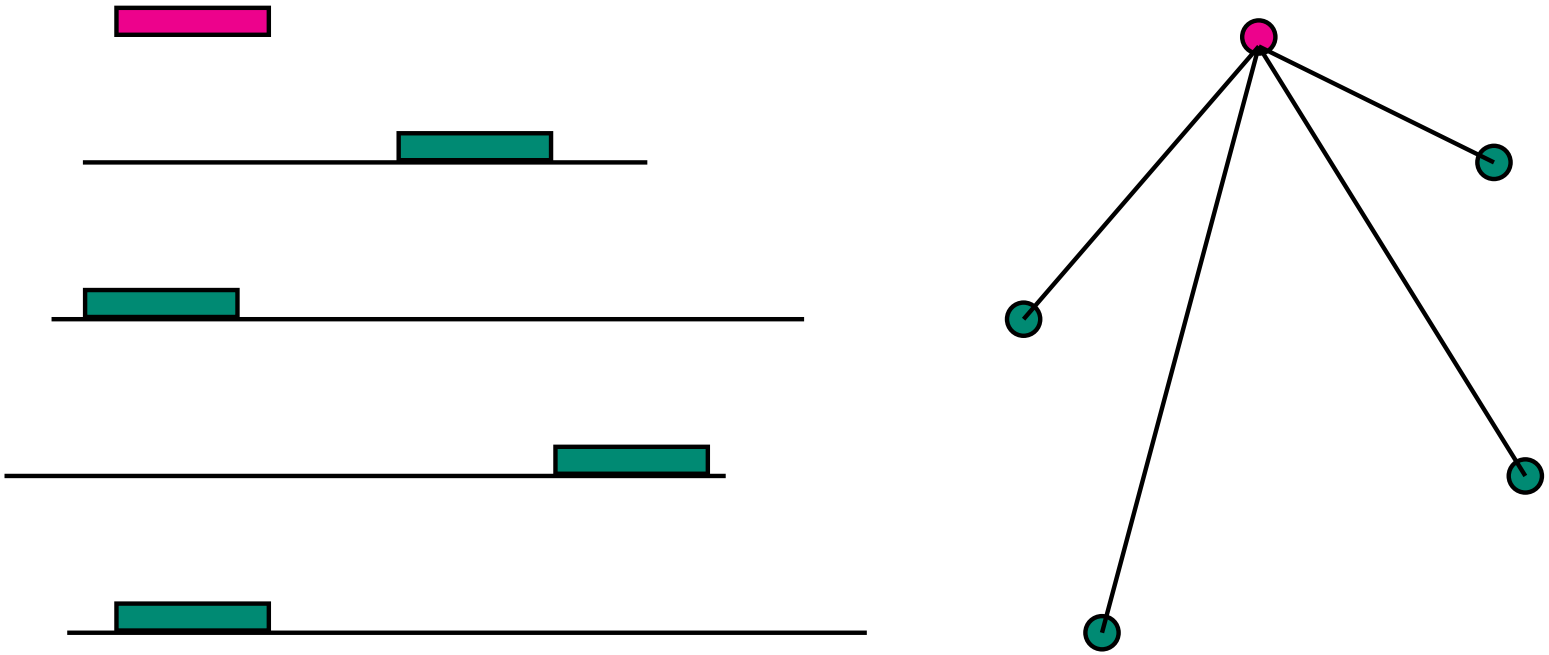
OUTPUT:

ancestors of all most parsimonious trees whose total length is at most e

“Conserved” elements: “Models”

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

= “Closest Substring” (in some cases)



Sagot (Moivre/Poivre 1995, 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

$$\Sigma_{DNA/RNA} = \{\text{A, C, G, T(U for RNA)}\}\ \text{(nucleotides)}$$

$$\Sigma_{protein} = \{\text{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}\}$$

(amino acids)

Formal definition of the: “Models Problem”

INPUT:

data: a set of N sequences

parameters: a length k, a max. (“local”!) allowed difference e, an alphabet Σ for the motifs

OUTPUT:

all models, that is, all motifs over Σ that have at least one “occurrence” in each sequence s (*i.e* k-mer in s at Hamming distance at most e from motif)

(alternatively) ancestors of all star trees with edges of length at most e

Possible modifications of some properties

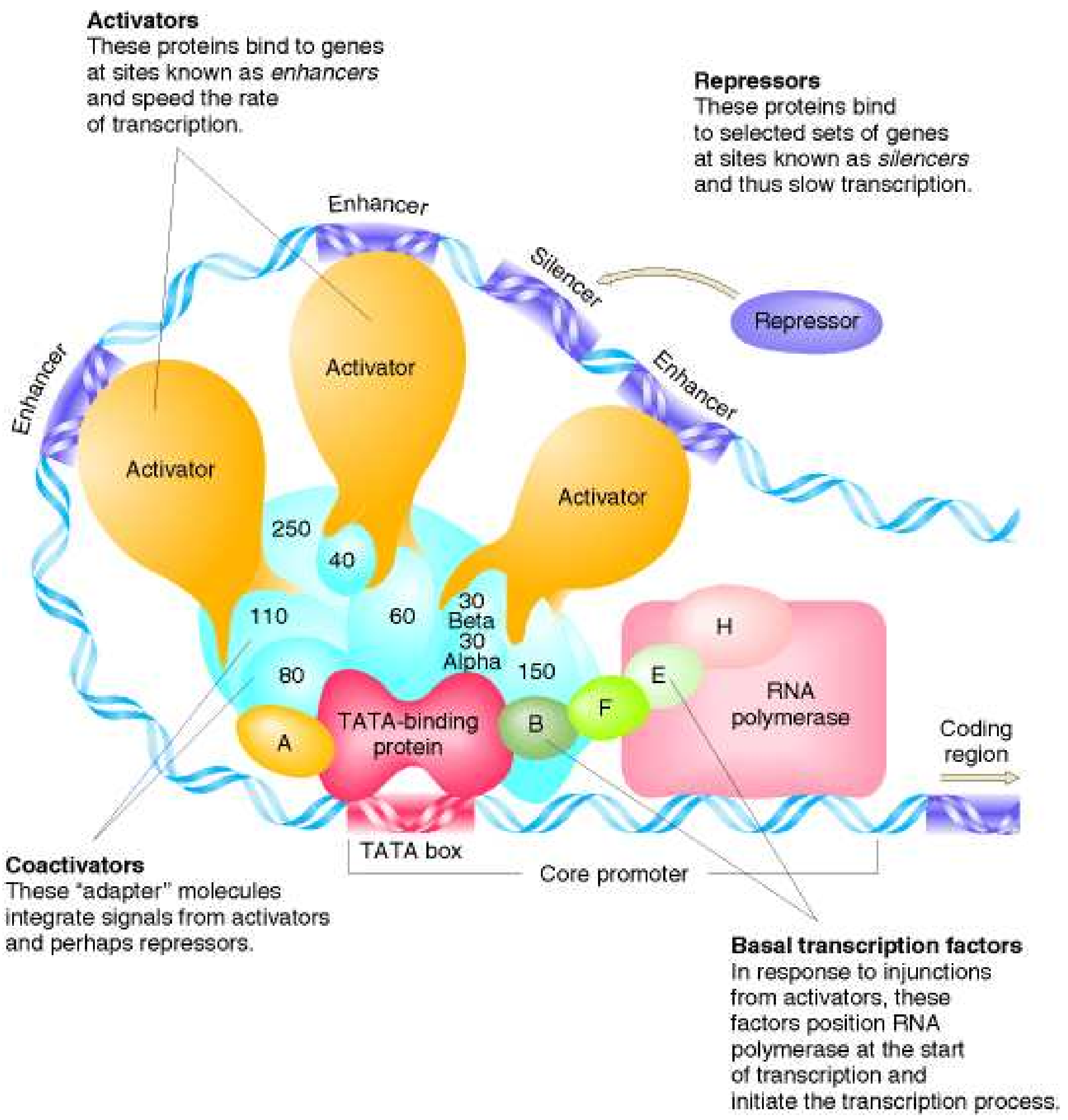
Quorum: that is, minimum number of sequences "containing the motif": may be less than N

Length of motif:

- within an interval (Sagot, Marsan)
- estimated (Bailey, Lawrence)
- fixed value (Pevzner, Pavesi, Blanchette)

The choice between definitions is not "algorithmically innocent"

"Most surprising" set of words:	**NP-hard**
"Cliques":	**NP-complete but reasonable ($O(n^{2.46})$) in practice**
"Ancestors":	**Exponential in length motif: $O(N \min(n(3k)^{\frac{E}{2}}, k(4^k+n)))$, but reasonable for small ("global") max. allowed differences**
"Models":	**Exponential in ("local") max. allowed differences: $O(N^2 n k^e)$**

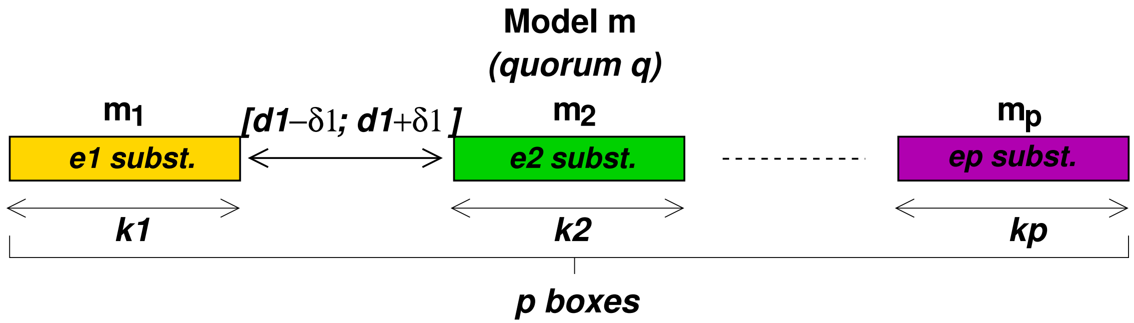


Our choice motivated by trying to model combinatorics of regulation

Towards modelling the combinatorics of regulation

“Structured models”

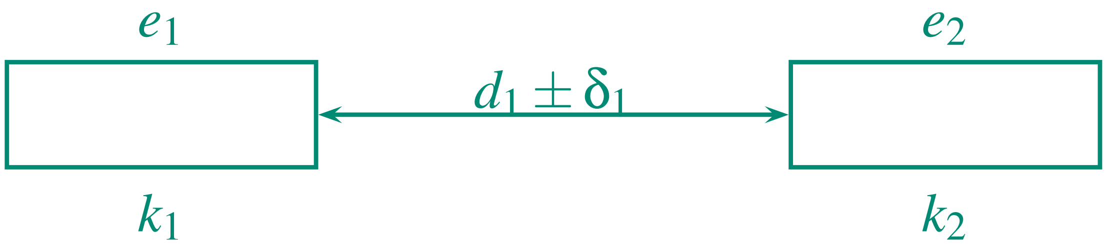
an ordered collection of p boxes, p maximum numbers of differences (e), p intervals of lengths (k), $p-1$ intervals of distances (between successive boxes in the collection) ($d \pm \delta$) (this is called *spacers*)



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, N. Pisanti *et al.*, paper in preparation)

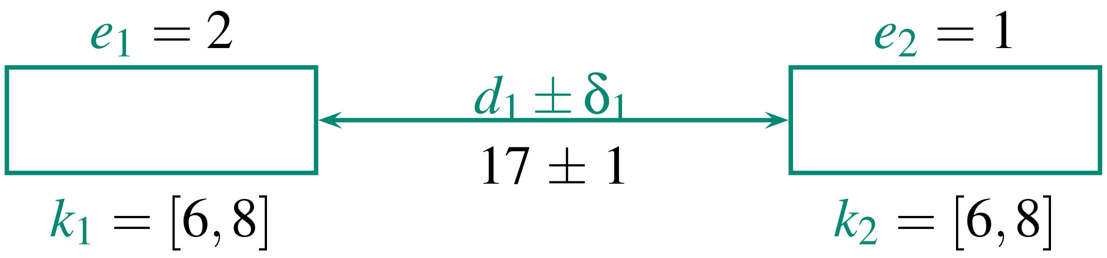
An example with two boxes



sequences

quorum

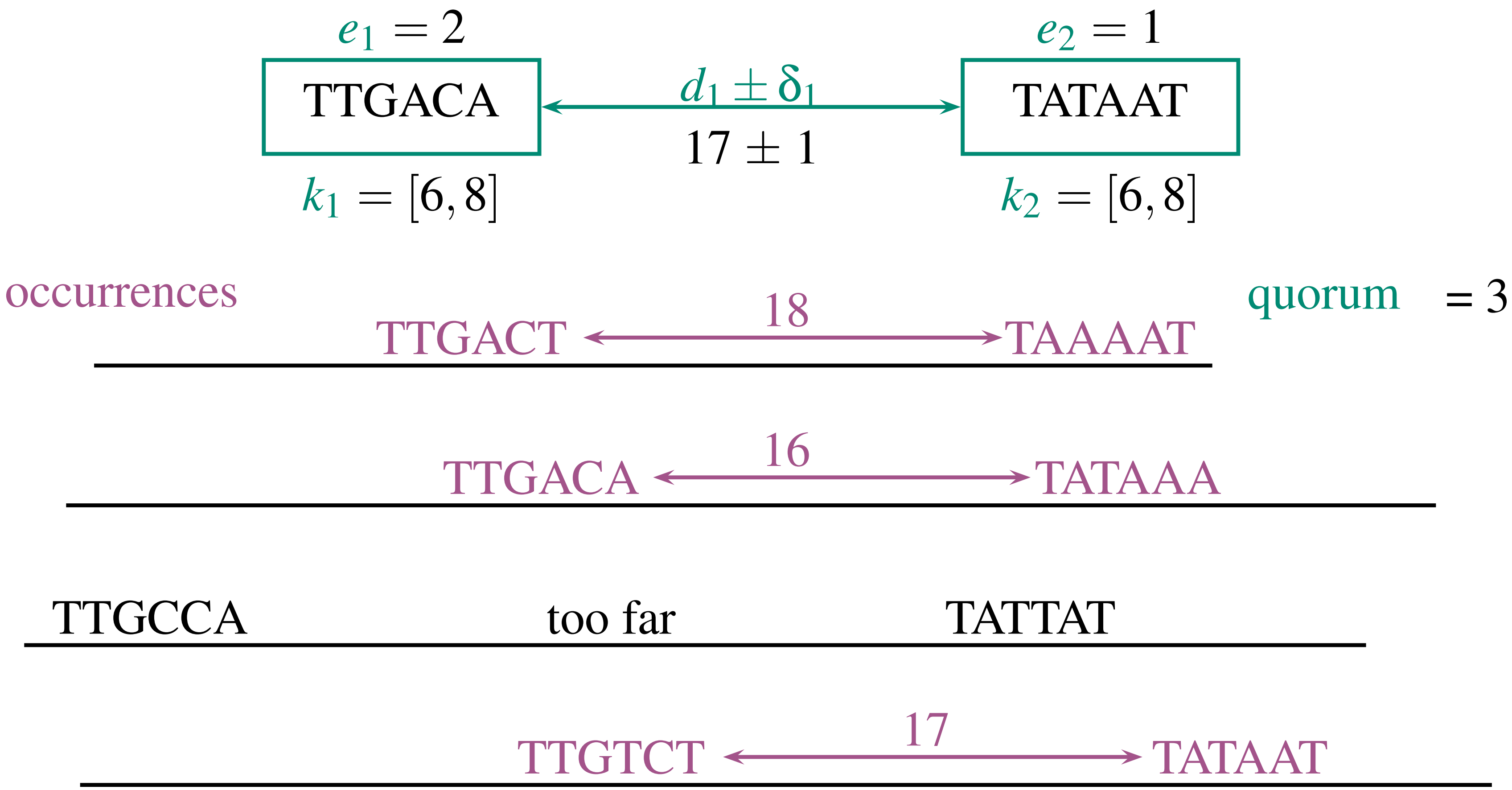
An example with two boxes



sequences

quorum $= 3$

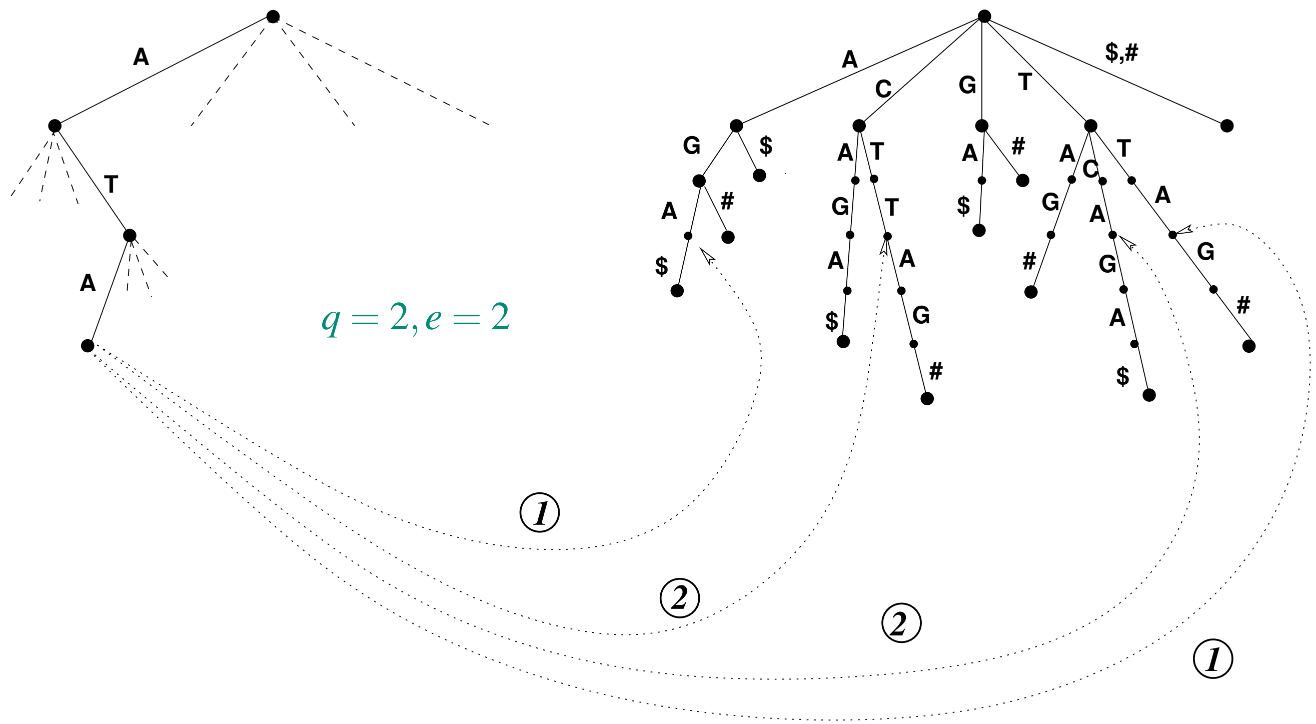
An example with two boxes



An idea of the SMILE algorithm for models detection: A story of traversing two trees with a dynamic programming algorithm

Trie of models (search tree)

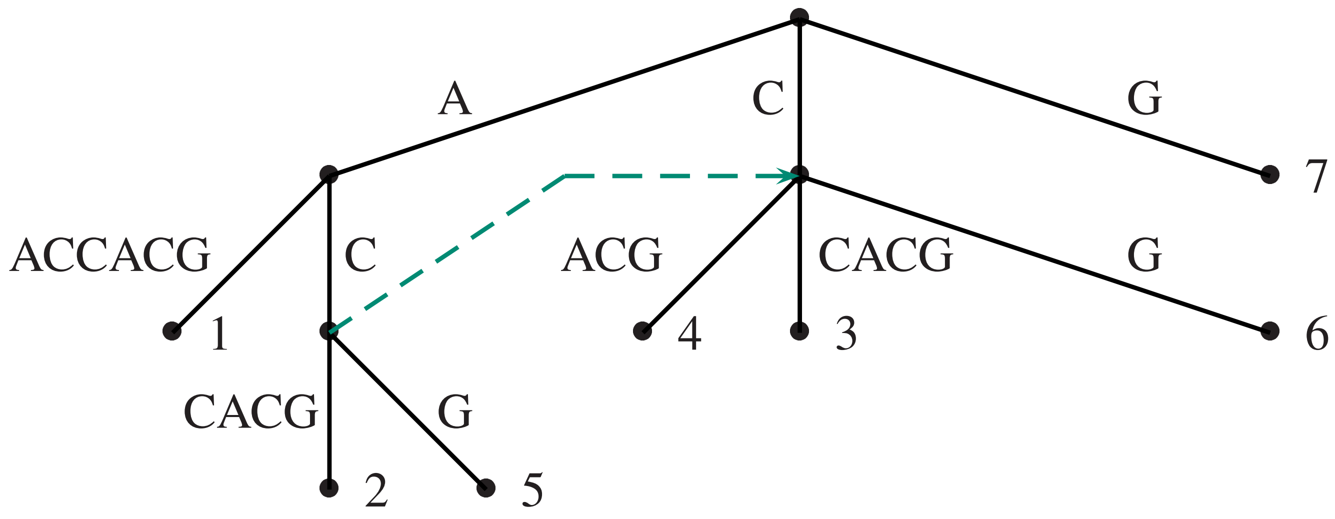
(Generalized) Suffix tree of strings TCAGA$ and CTTAG#



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

Suffix tree

Suffix tree for AACCACG

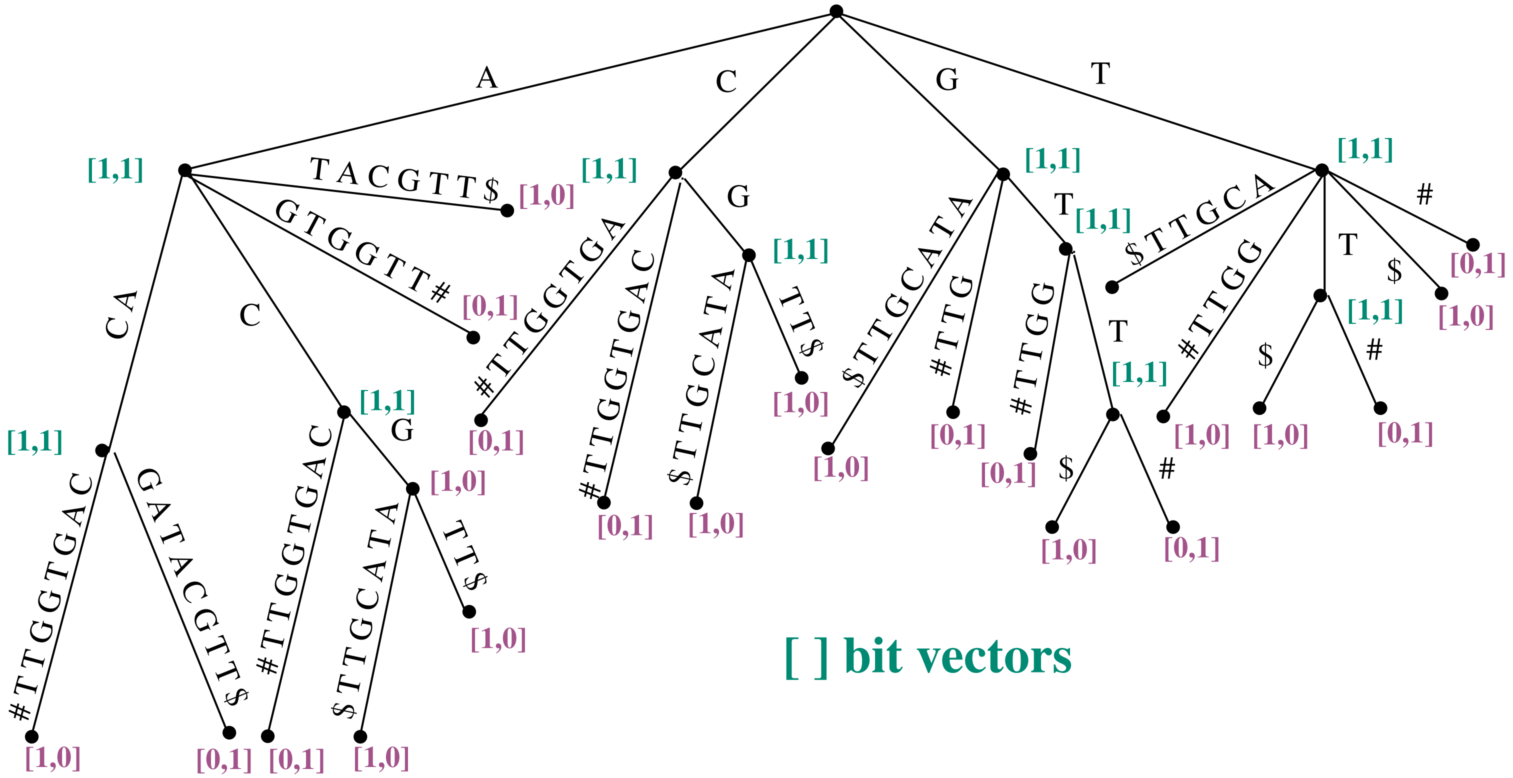


key feature: concatenation of labels of arcs from root to leaf i spells suffix of input string starting at position i

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

Generalized suffix tree

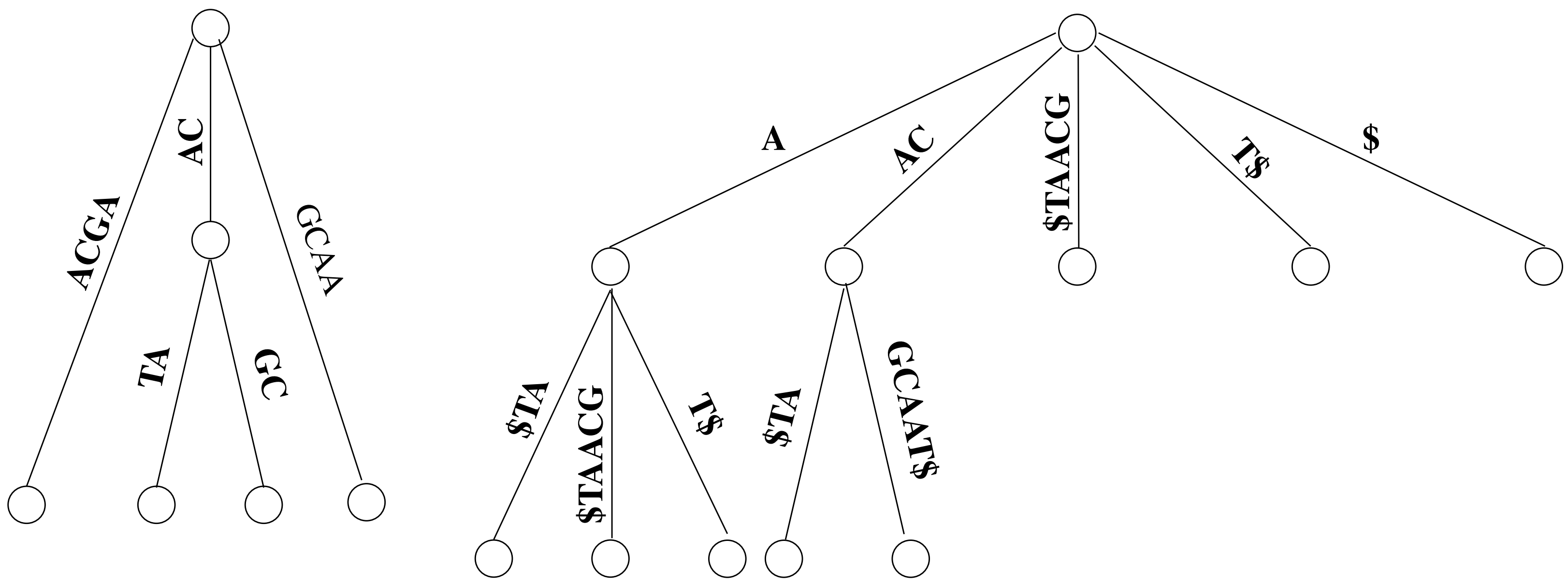
Strings AACGATACCTT$ and AACCAGTGGTT#



[ ] bit vectors

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

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

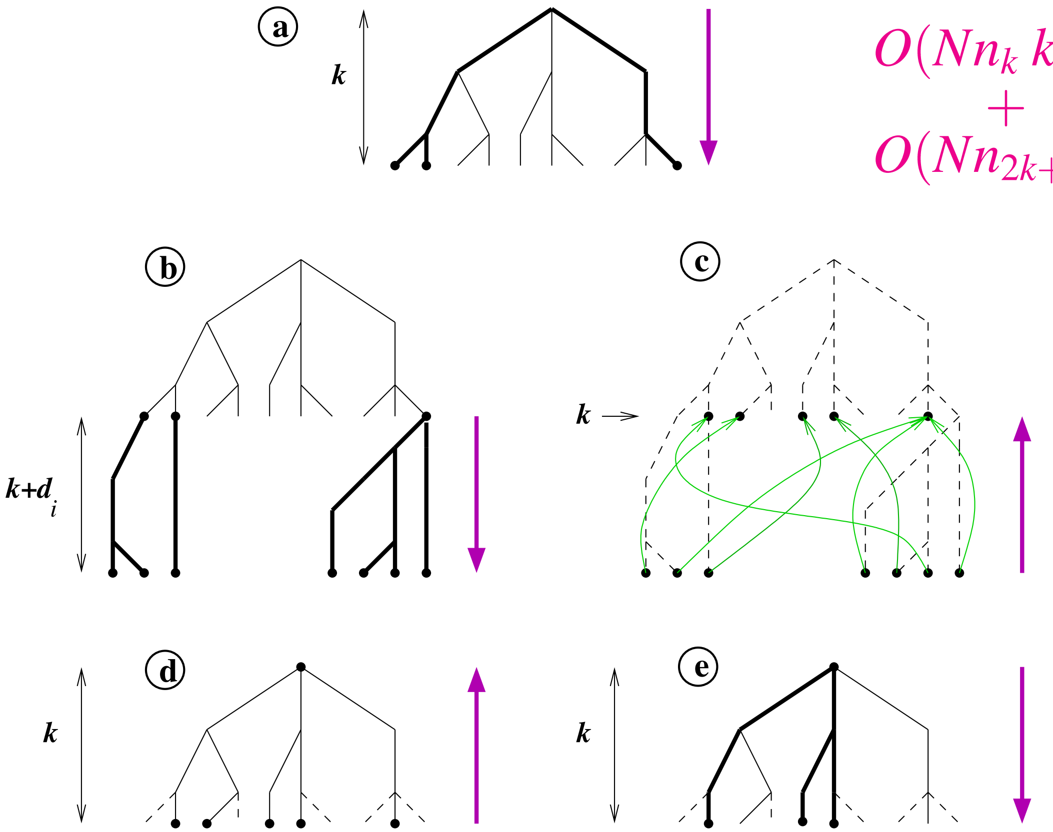


factor tree — **suffix tree for the same string**

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

(Allali and Sagot, submitted)

Generalized dynamic factor tree



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

In fact, all of this is in full evolution

Among the things we have been working on

more efficient algorithms

better indexes

correlated boxes (for instance, palindromic)

boxes skipped or order changed

obtaining hints on how to set the parameters

etc.

Getting back to what exists: Statistical evaluation of 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, 9:761-773, 2003)

Still a LOT of work to be done

A possible alternative to statistical evaluation: Basis of motifs

Main ideas

Find a subset of all motifs that is of "small" (polynomial) size and computable in polynomial time

and is a generator of all motifs (basis)

(Parida *et al.*, 2000)

Darling, shrinking the set of motifs is a hard problem

Parida's basis is at least quadratic in size even for quorum 2

Later bases, although linear for $q = 2$, are exponential in $q - 1$ for quorum q; this is the case also for Parida's basis

(proofs in Pisanti *et al.*, submitted 2002 and MFCS; related work by Apostolico *et al.*, 2003 and Pelfrene *et al.*, 2003)

Back to the beginning of the story: *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

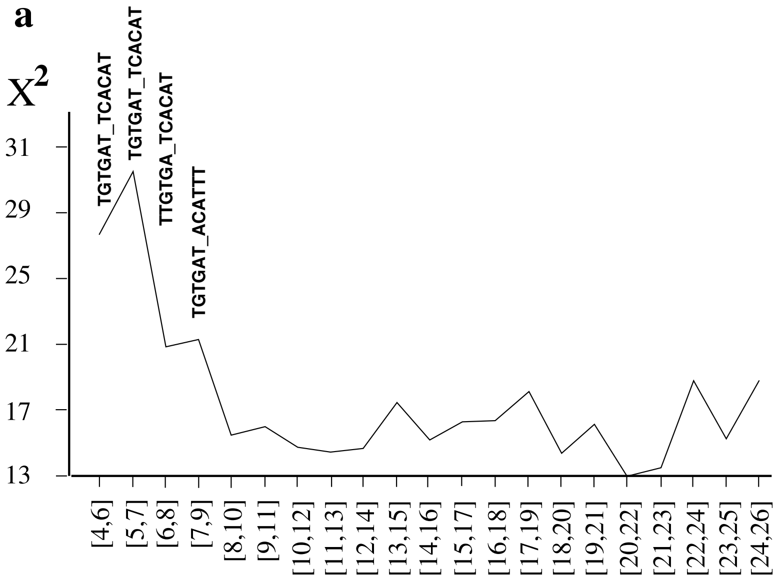
Escherichia coli **and** *Bacillus subtilis*

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

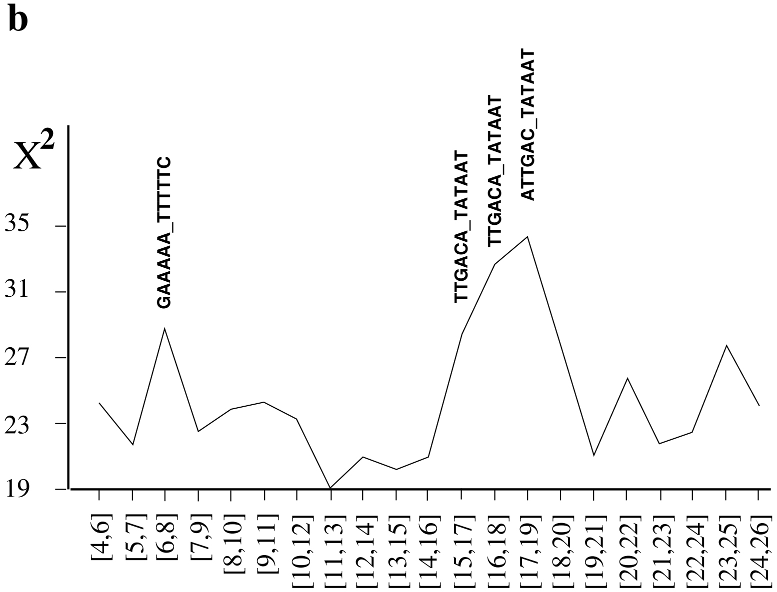
(A. Vanet, L. Marsan, M.-F. Sagot, *Res. in Microb.***, 150:779-799, 1999)**

An example of the results with 2 boxes

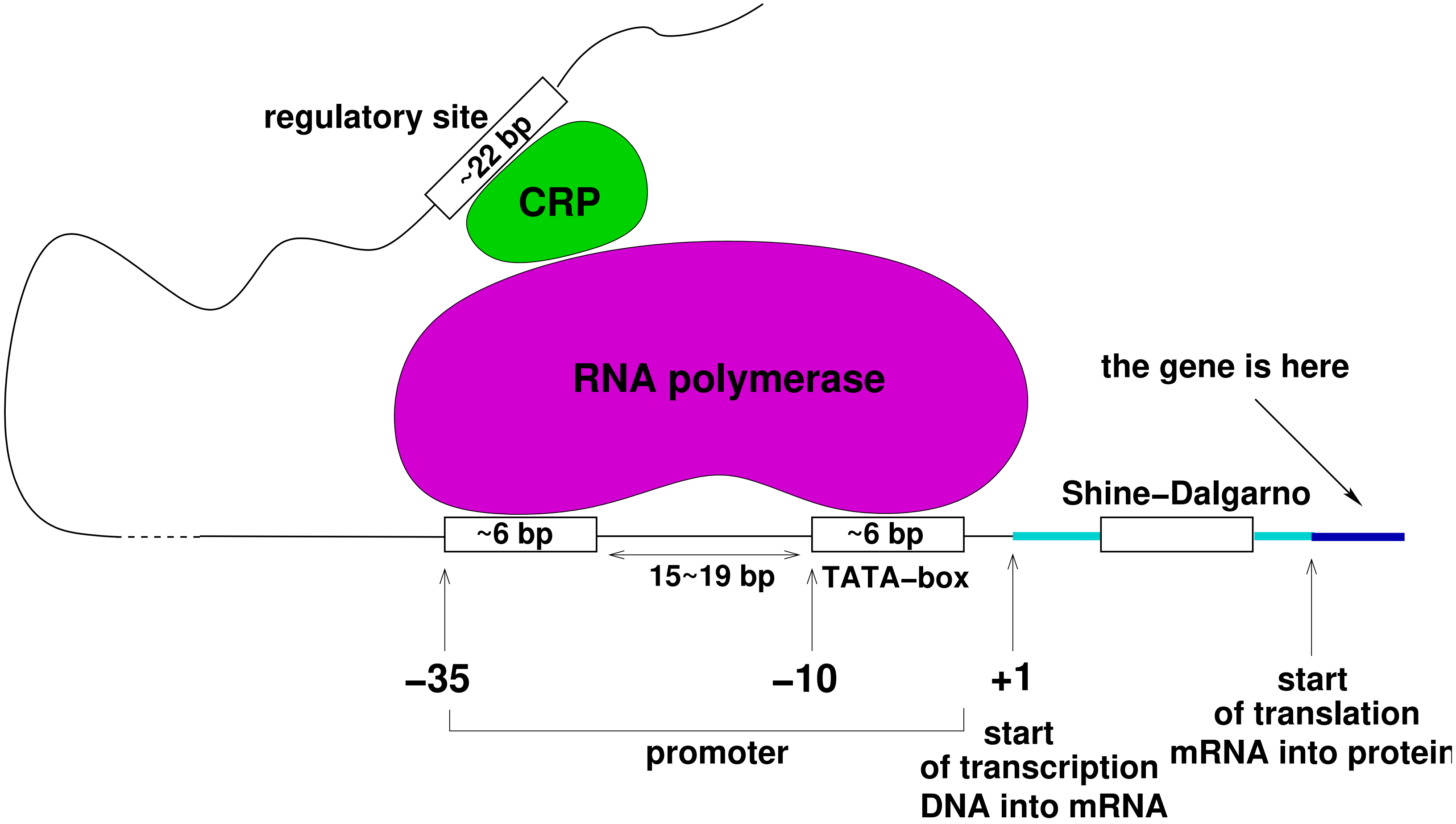
Escherichia coli



Bacillus subtilis



Some surprising findings in *E. coli*: CRP binding site



The road ahead is long

Data sets

1. non coding sequences (81) upstream of genes regulated by CRP according to RegulonDB (Collado-Vides *et al.*)

2. non coding sequences (on both strands, 2652) in the genome (*i.e.*, upstream of all genes of *E. coli*)

The road ahead is long (contd.)

Two boxes, each of length 5, separated by a strict distance of 6

RegulonDB			Non coding sequences		
# Subs	# Occ	# Occ / # Seq	# Subs	# Occ	# Occ / # Seq
0	6	7.41%	0	23	0.87%
1	30	37.03%	1	142	5.35%
2	65	80.25%	2	444	16.74%
3	81	100.00%	3	1399	52.75%

The problem is the same for all other approaches

How then to interpret ? Which consequences ?

Current work on 3 to 4 boxes motifs (CRP + promoter sequences) on RegulonDB set

Other subjects we have been working on

Biological structures

Comparing and inferring RNA (small or large) structures

Enumerating protein folds to understand protein evolution

Inferring protein sequence/structure motifs

Genome structure

Inferring eukaryotic genes

Identifying long repeats (also devising new filtering techniques)

Inferring repeats and duplication trees simultaneously

Studying the relation of repeats and regulation

Other subjects we have been working on (contd.)

Genome dynamics

Detecting recombination points

Revisiting definitions of “segments conserved by rearrangements”

Calculating rearrangement distances between evolutionary trees and genomes

Biological networks

Defining and inferring “network motifs”

Studying the evolution of networks