Motifs and Regulation

Marie-France Sagot

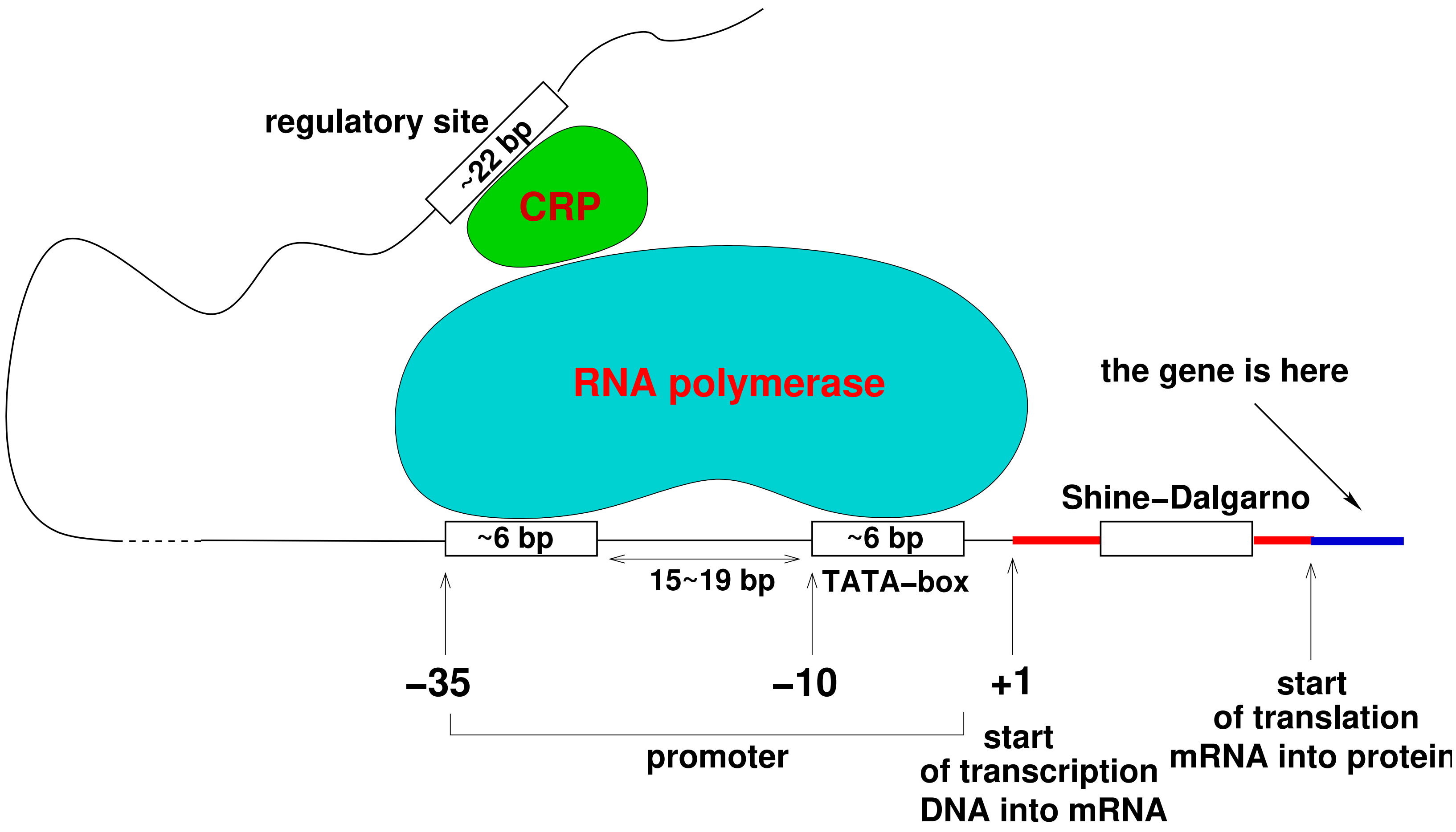
Inria Rhône-Alpes – Projet HELIX
Laboratoire de Biométrie et Biologie Évolutive
Université Claude Bernard, Lyon I

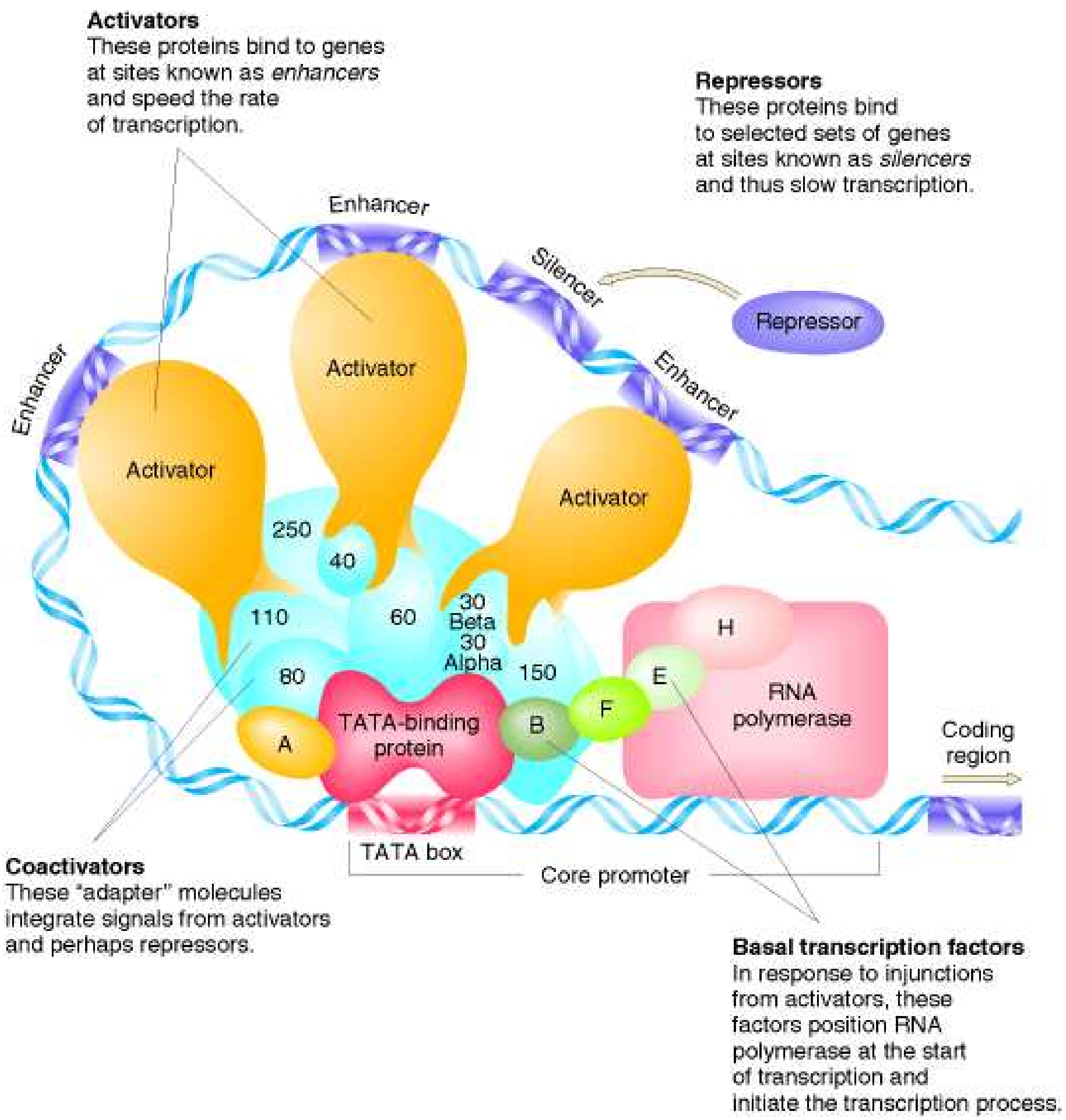
Participants

Lyon	
Raquel Tavares	
Paris (with Maxime Crochemore)	
Julien Allali **Benoît Olivieri**	**Laurent Marsan** **Pierre Peterlongo**
Canada	
Sébastien Provencher	
Italy	**Portugal**
Nadia Pisanti	**Susana Vinga**
Brazil	
Said Sadique Adi	**Paulo Gustavo Fonseca**

and many collaborators...

Biological motivation (Promoters/regulatory elements in prokaryotes)





Biological motivation (Promoters/regulatory elements in eukaryotes)

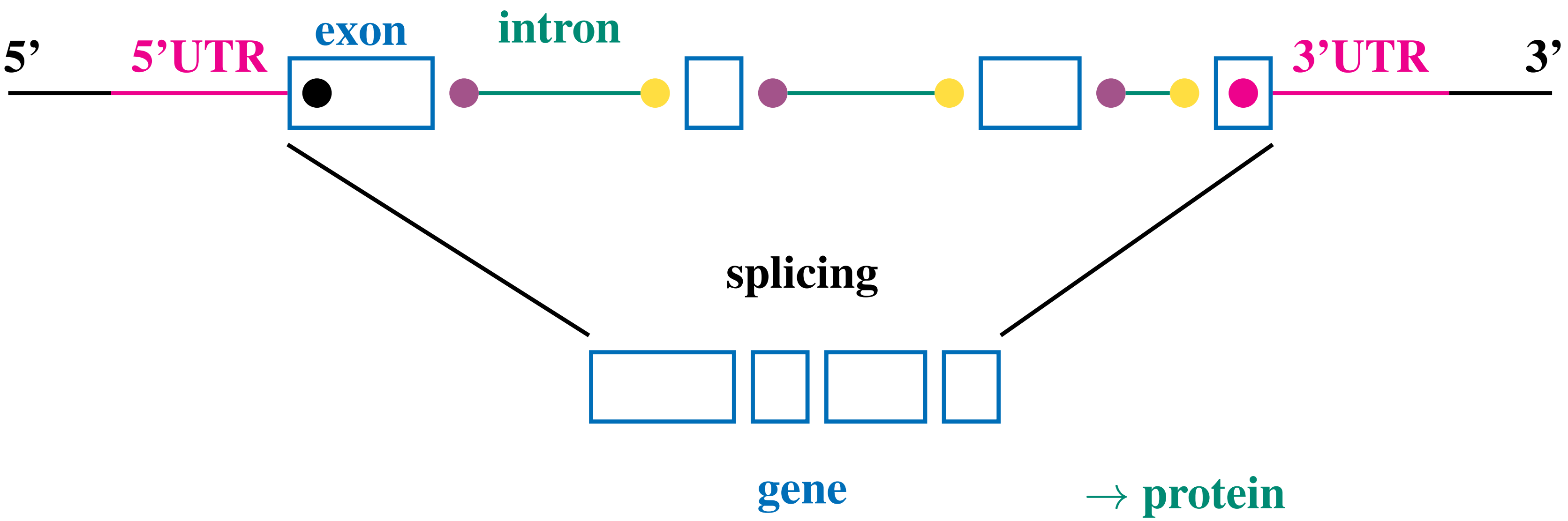
Biological motivation (Repeats: dispersed or in tandem as here)

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

Biological motivation (Splicing sites and enhancer sequences in eukaryotes)

start codon GT (donor site) AG (acceptor site) stop codon

3 nucleotides (codon) → 1 amino acid



5'UTR and 3'UTR: transcribed into RNA but not translated

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

Properties

main property: motifs of interest = “conserved” elements

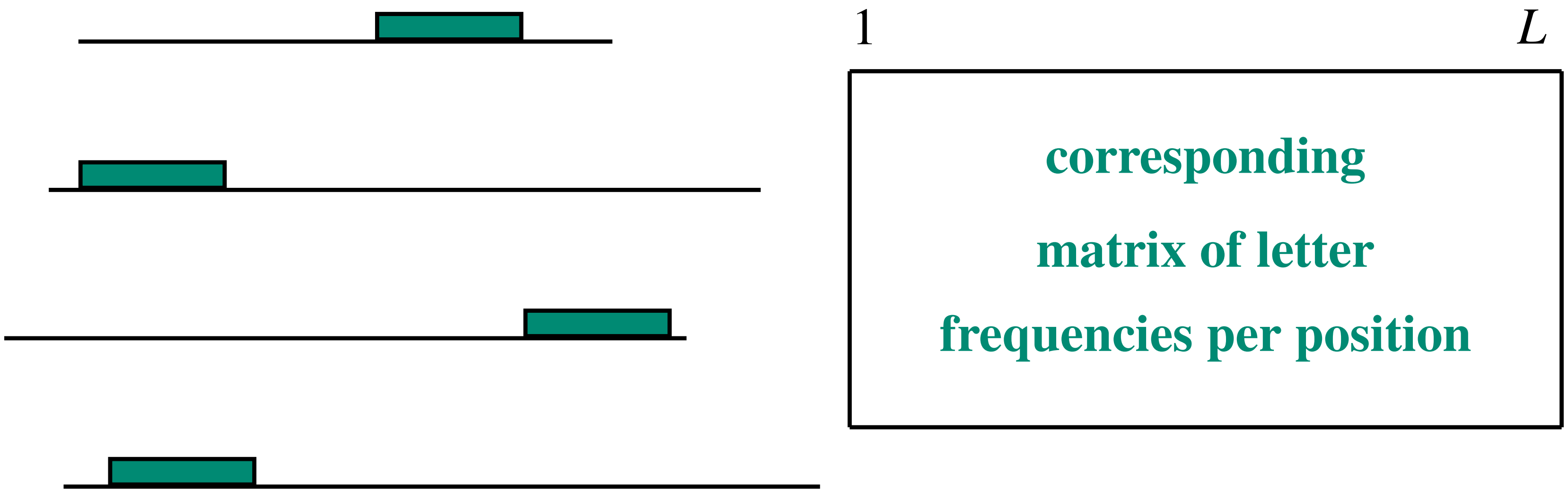
mutations (substitutions and in some cases insertions and deletions) may happen that do not destroy function and may even enable to modulate it

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

“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, a "quorum" of N

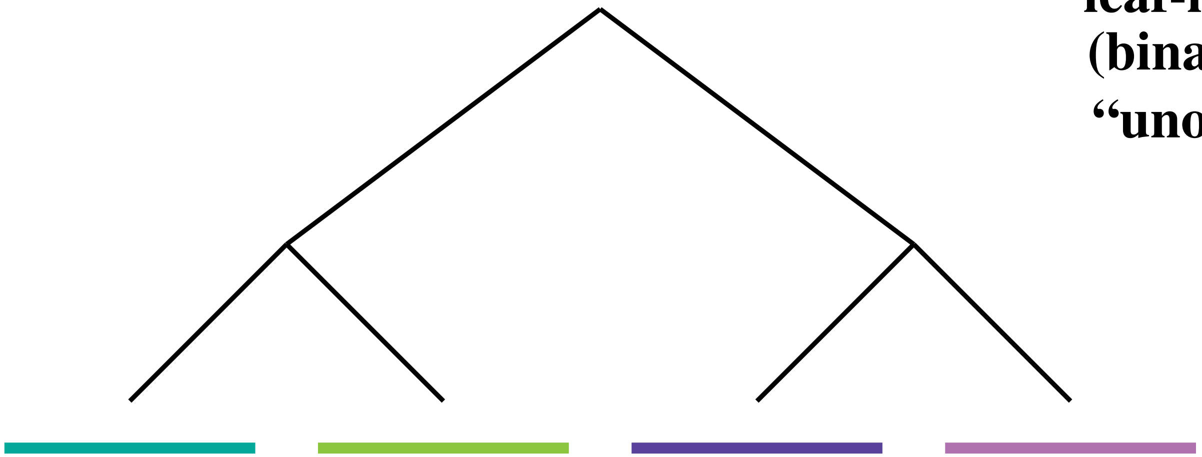
OUTPUT:

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

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”

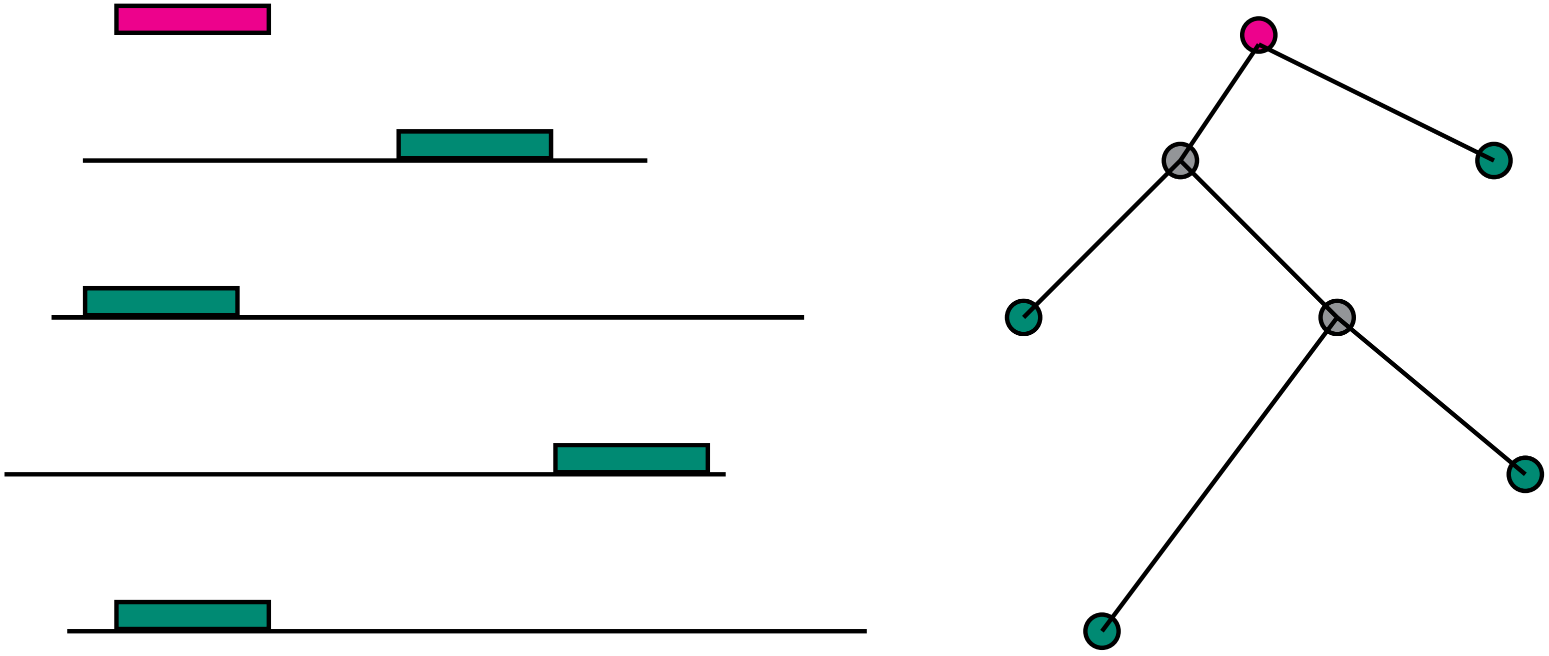
*** ancestors**

(rooted)
leaf-labeled
(binary)
“unordered”

“Conserved” elements: Ancestors in most parsimonious tree

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 “quorum” of N, a maximum (global!) allowed difference e

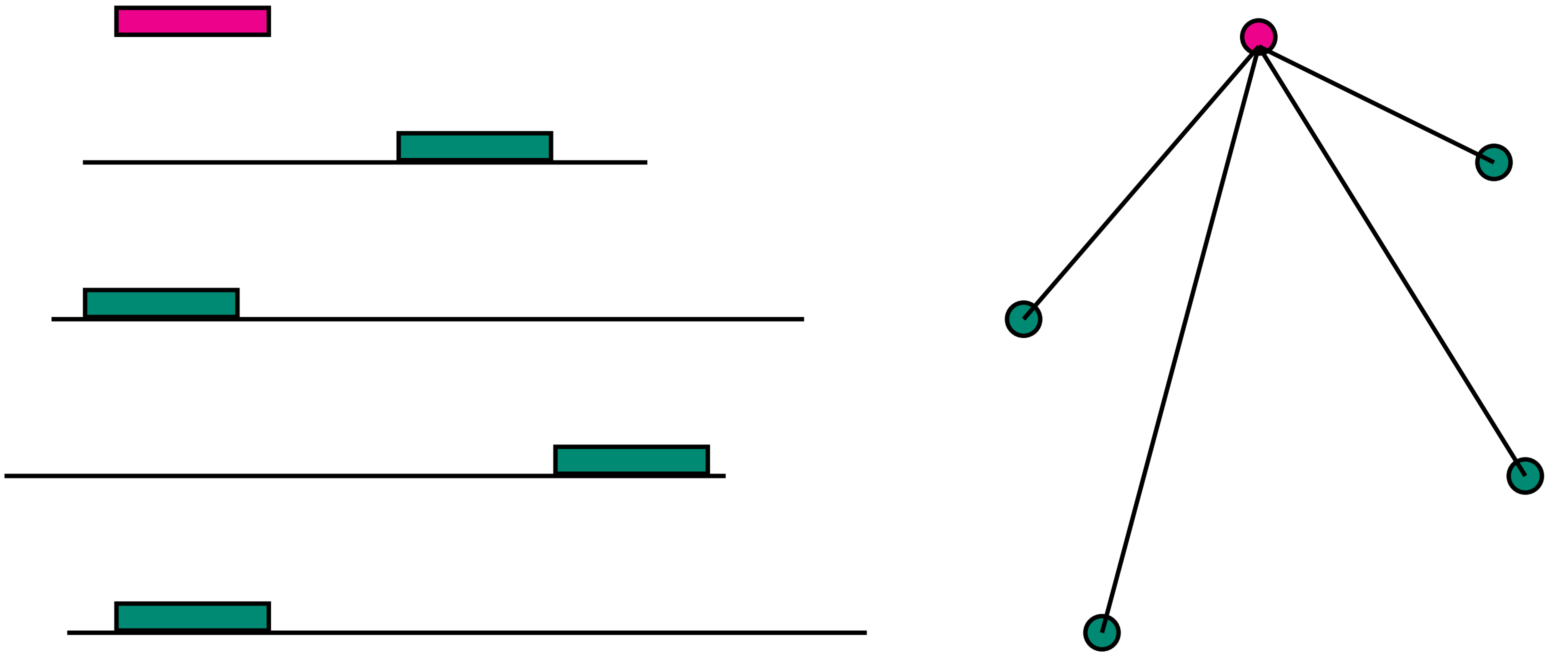
OUTPUT:

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 “quorum” of N, a maximum (“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 modification 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: $O(N^{(2)} n k^e)$**

One more word about "cliques"

Main ideas:

A few definitions

- **vertex** v **"neighbour" of an** l**-clique** $\{u_1, \ldots, u_l\}$ **if** $\{v, u_1, \ldots, u_l\}$ **is a clique**
- **"extendable** l**-clique" with** $l < N$**: an** l**-clique that has at least one neighbour in every part of multipartite graph** G
- **"spurious edge": edge that does not belong to any extendable** l**-clique for a given** l

Procedure

- **given** l**, eliminate spurious edges**
- **find cliques in resulting (hopefully sparse) graph**

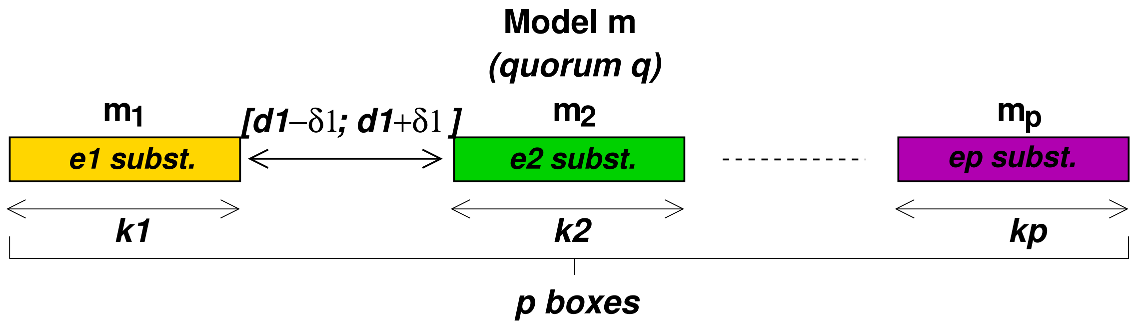
More recent developments

Mixing ideas from “cliques” and “models” definitions of conservation

Towards modelling the combinatorics of regulation

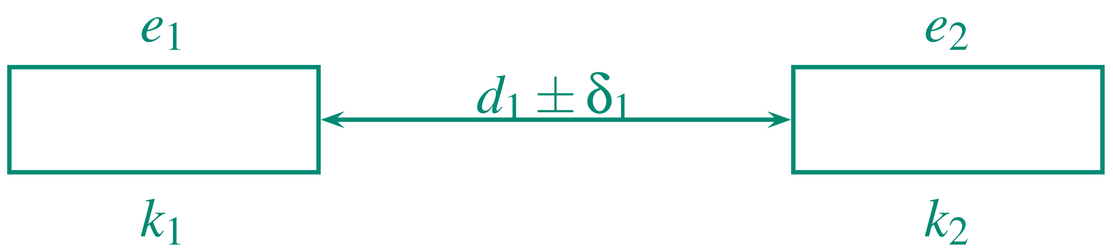
"Structured models"

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 (S. Provencher, R. Tavares *et al.*, paper in preparation)

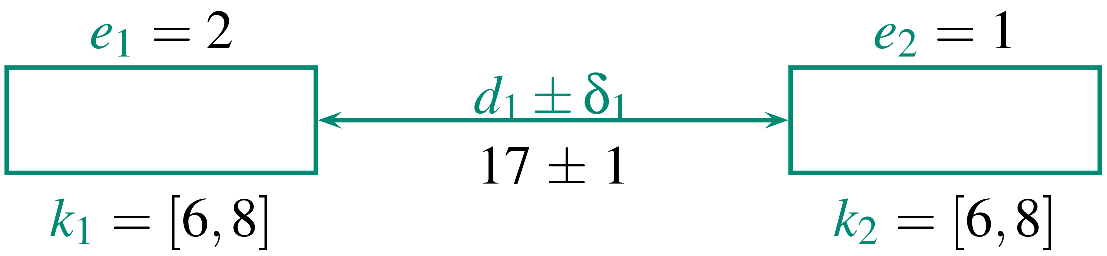
An example with two boxes



sequences

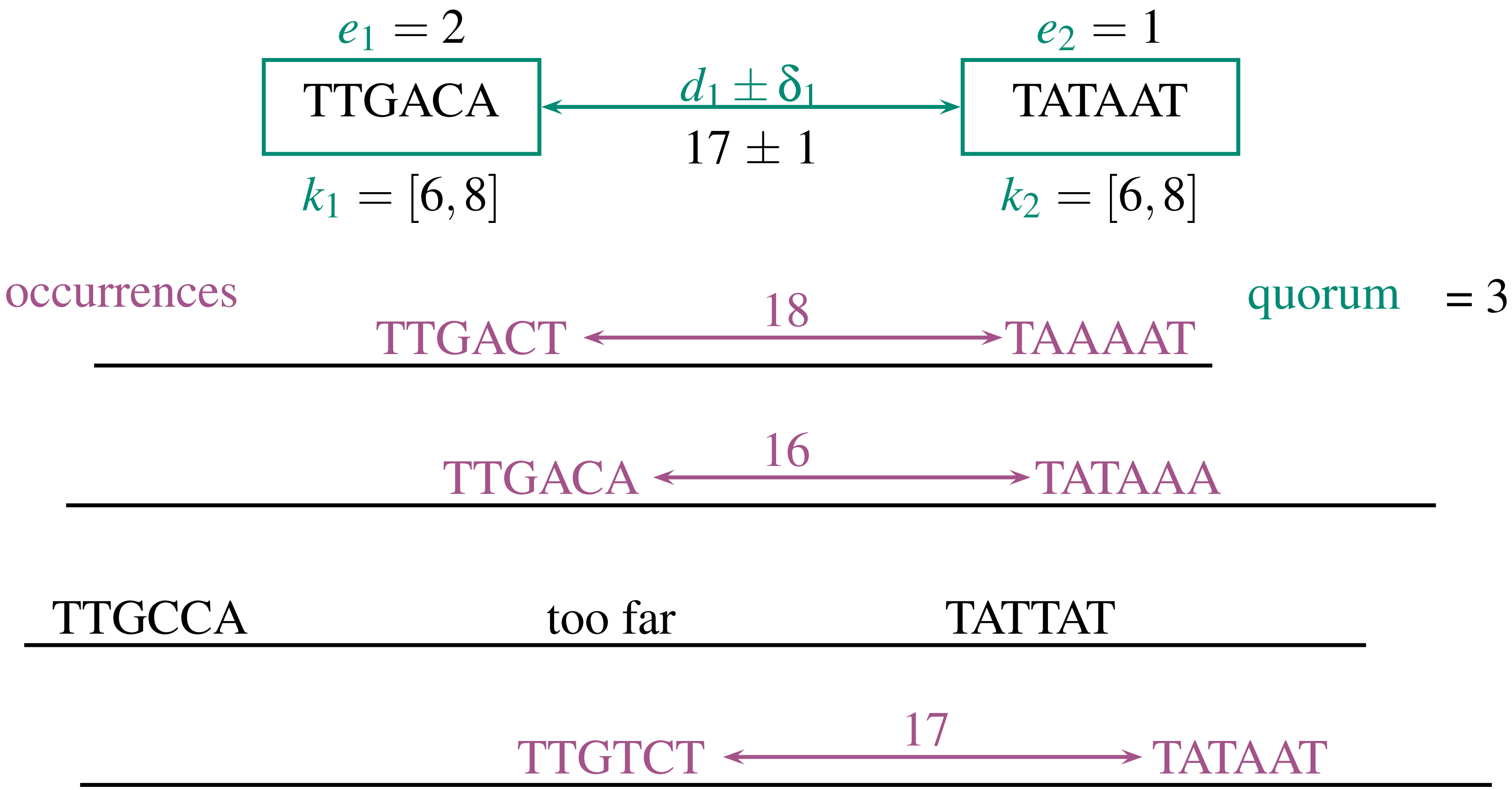
quorum

An example with two boxes

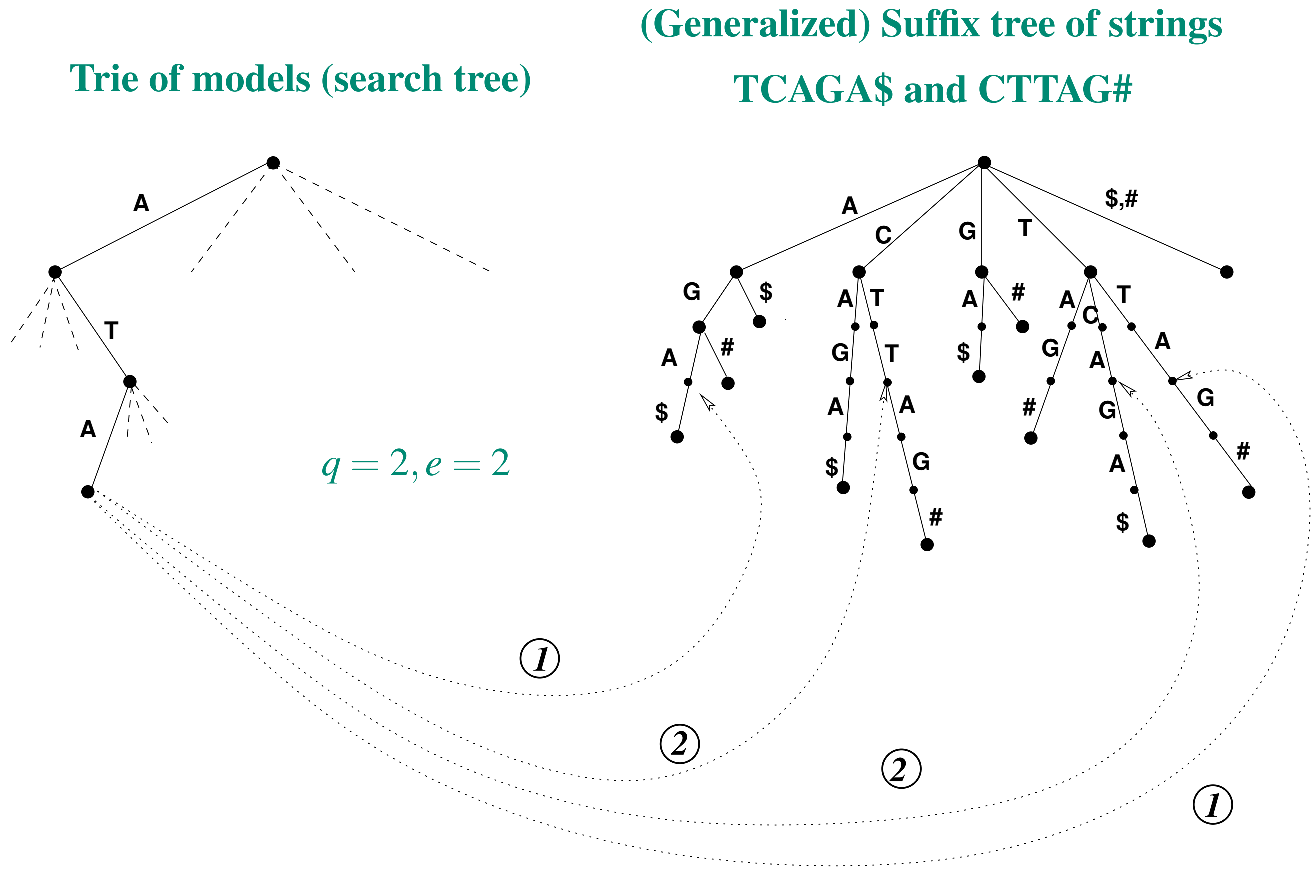


quorum $= 3$

An example with two boxes

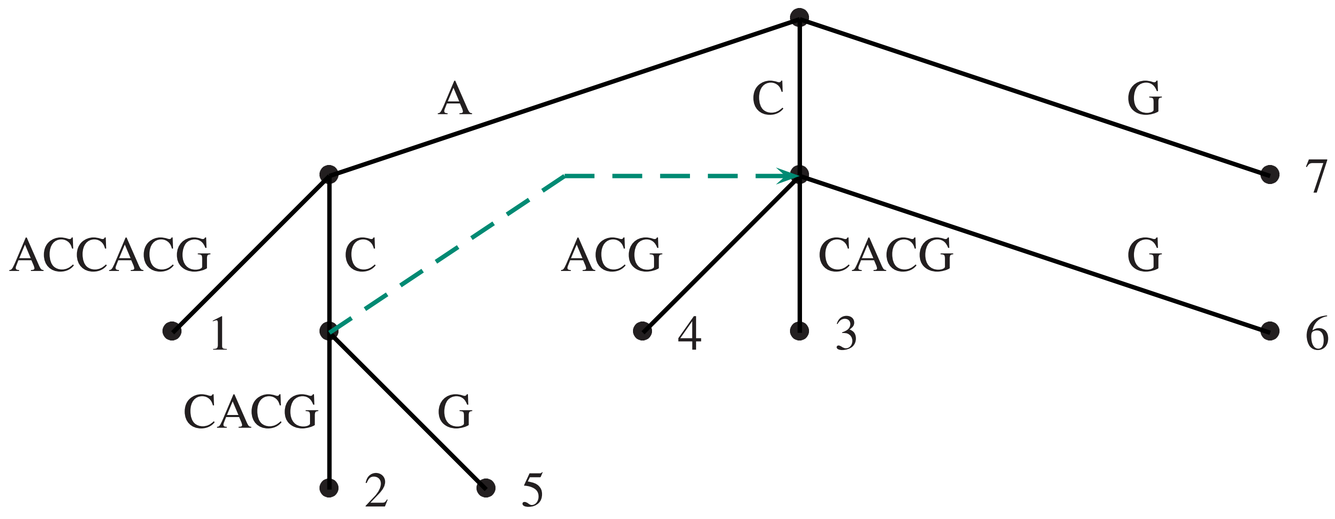


An idea of the SMILE algorithm for models detection: A story of traversing two trees with a dynamic programming algorithm



Suffix tree

Suffix tree for AACCACG

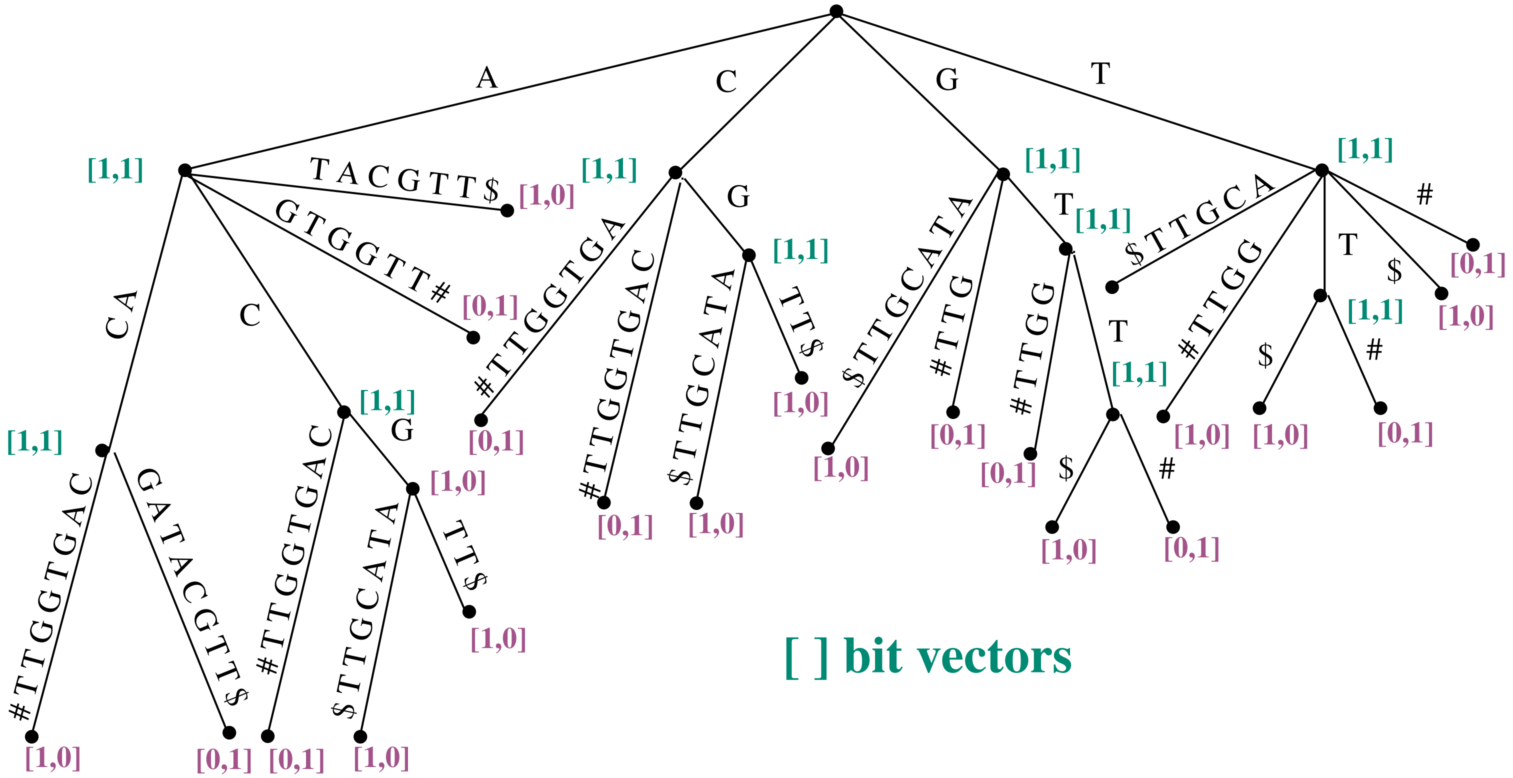


key feature: concatenation of labels of arcs from root to leaf *i* spells suffix of input string starting at position *i*

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

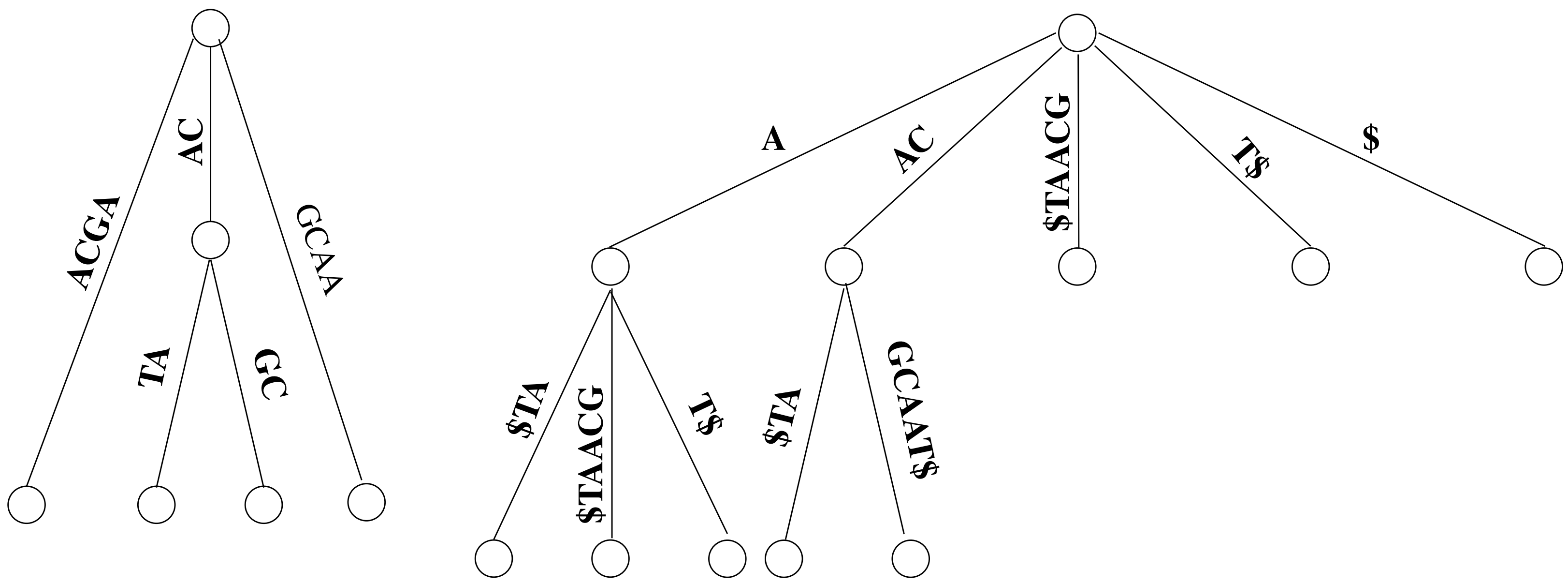
Generalized suffix tree

Strings AACGATACCTT$ and AACCAGTGGTT#



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

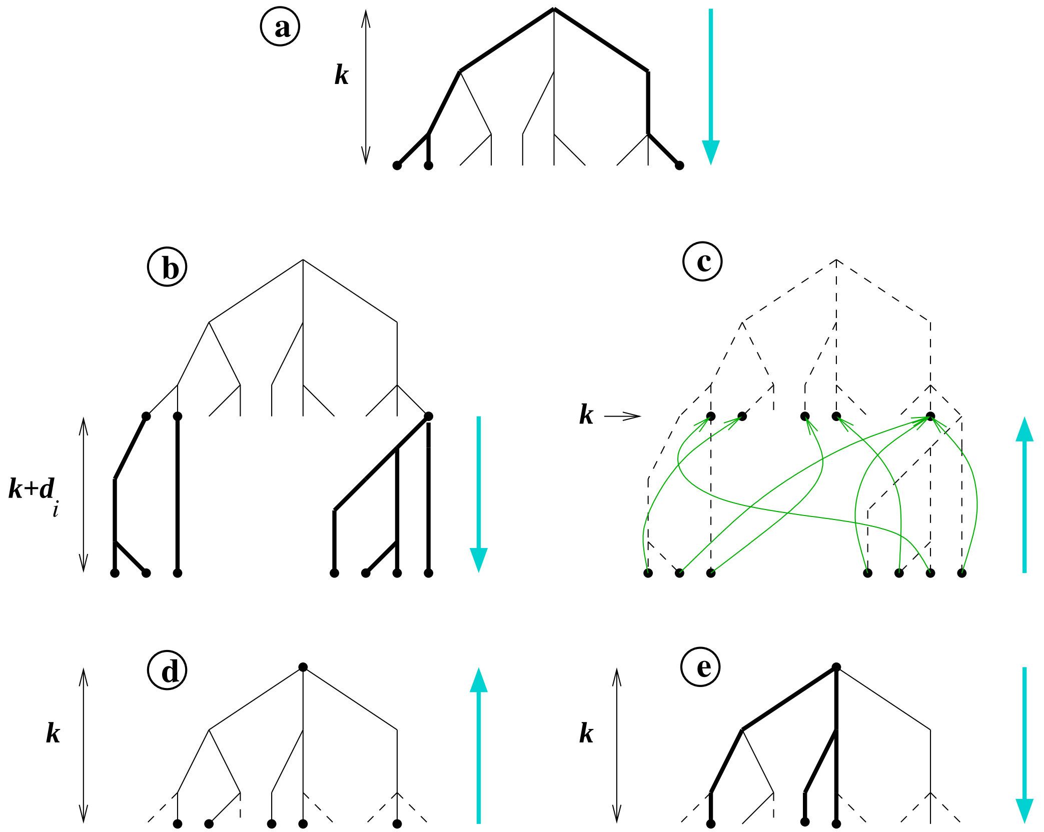
(Allali and Sagot, submitted)

A few results (Canterbury corpus / Calgary corpus)

file	length	$\lvert\Sigma\rvert$	suffix tree		k=20		k=10	
			time	space	time	gain	time	gain
bible (t)	4047393	64	13.56	9.64	14.17	9.94	12.50	40.60
pic (b)	513217	160	1.20	8.95	0.85	44.01	0.74	47.30
C14	87164908	5	319.98	12.43	342.30	7.35	243.73	67.34
C22	34553832	5	116.28	12.29	124.82	9.83	92.30	66.20

file	length	pattern	suffix tree		factor tree		
		length	mean	max	k	mean	max
bible(t)	4047393	10	14.90	368.34	17	12.28	183.40
bible(t)	4047393	15	8.29	134.81	17	8.79	17.54
pic(b)	513217	60	8423.35	23333.33	204	4782.99	12999.99
U00096	4639221	10	3.44	17.55	12	3.21	4.80
U00096	4639221	12	3.22	12.68	12	3.13	4.93

Generalized dynamic factor tree



Bounds on the search

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

Bounds on the search (continuation)

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

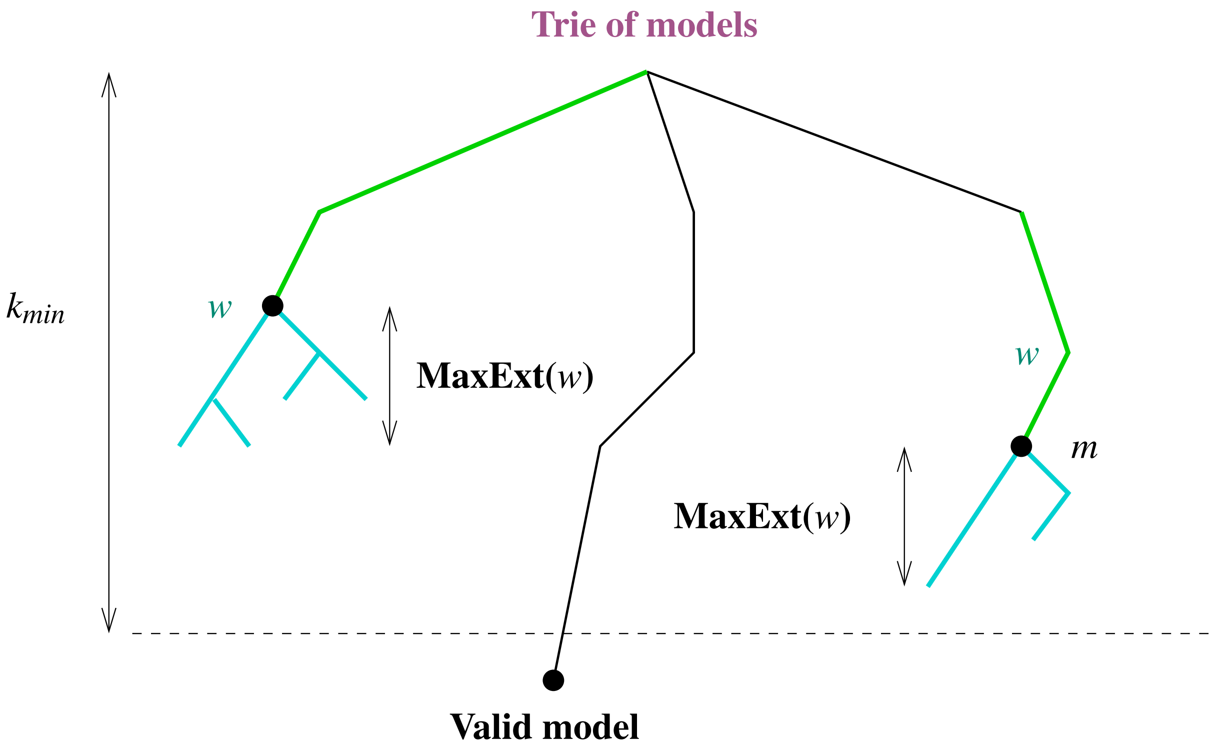
descent down the trees may end at first such prefix

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

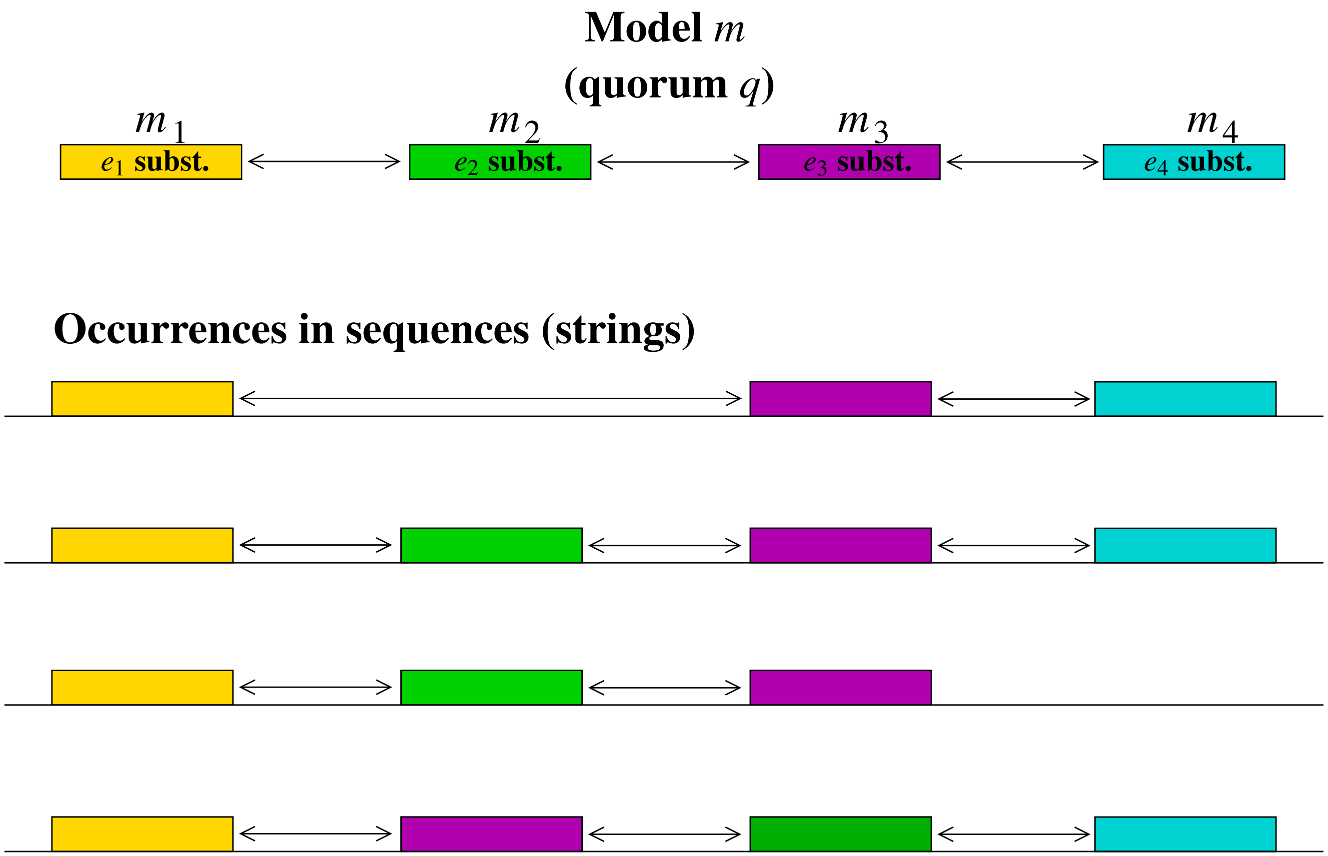
descent down the trees may end as soon as such situation is met

(Marsan, Pisanti and Sagot, paper in preparation)

Maximum "extendability"



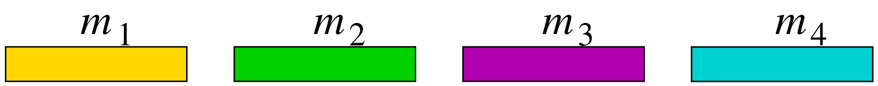
Skipping boxes or changing order



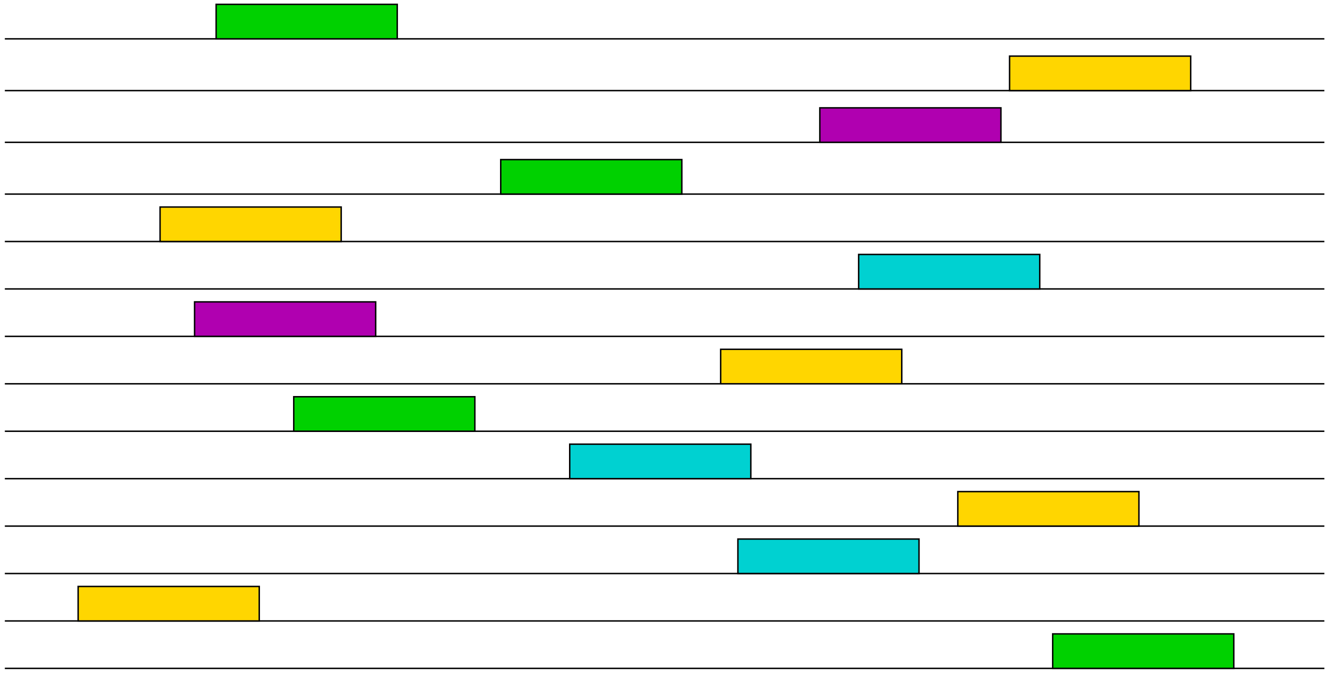
Skipping boxes: (Marsan *et al.*, in preparation)

Covering motifs (1)

Minumum set of models (one box) satisfying the properties



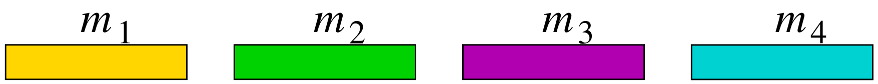
that “covers exactly” a set of sequences (strings)



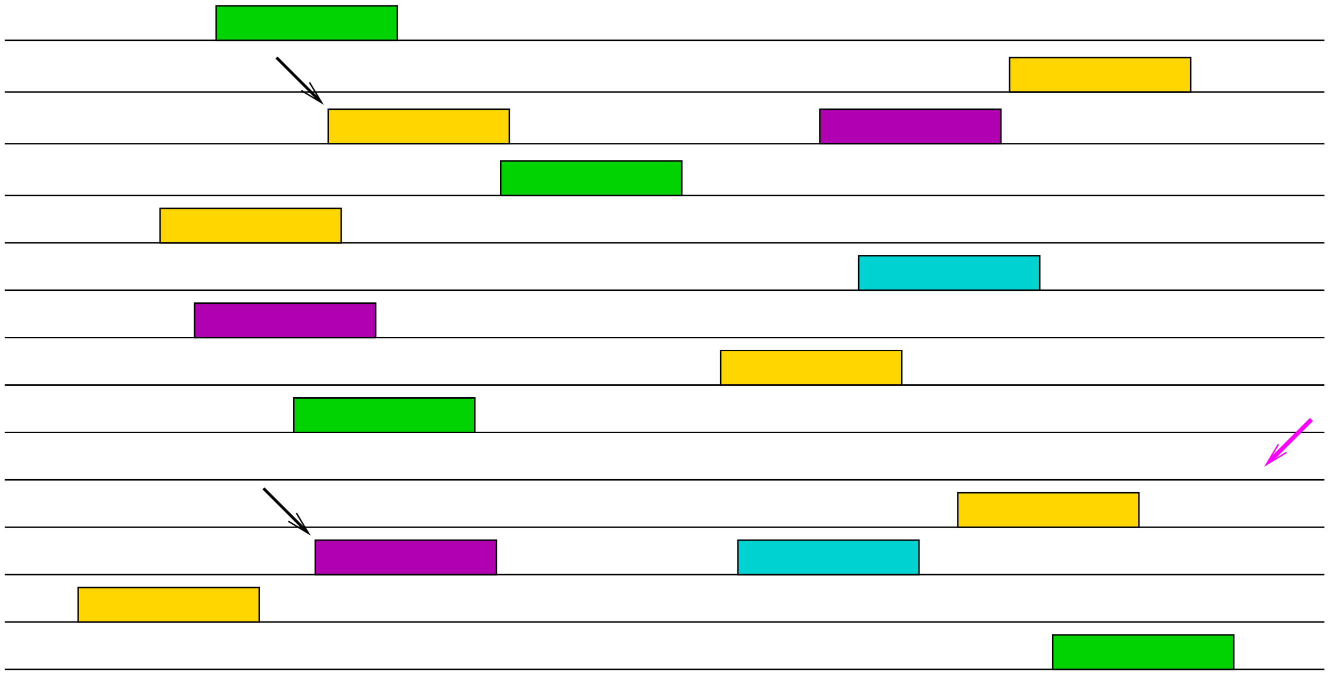
(Vialette, Sagot, work in progress)

Covering motifs (2)

Minumum set of models (one box) satisfying the properties



that "covers approximately" a set of sequences (strings)



(Vialette, Sagot, work in progress)

Basis of motifs

Formal definition of the initial problem

INPUT:

data: one sequence s over alphabet Σ

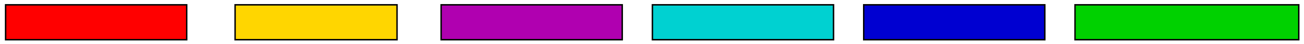
parameters: a quorum q, no differences allowed, an alphabet for the motifs $\Sigma' = \Sigma \cup \{\bullet\}$ where $\bullet$ is the don't care (wild card) symbol

OUTPUT:

all motifs over Σ' having at least q occurrences in the sequence s

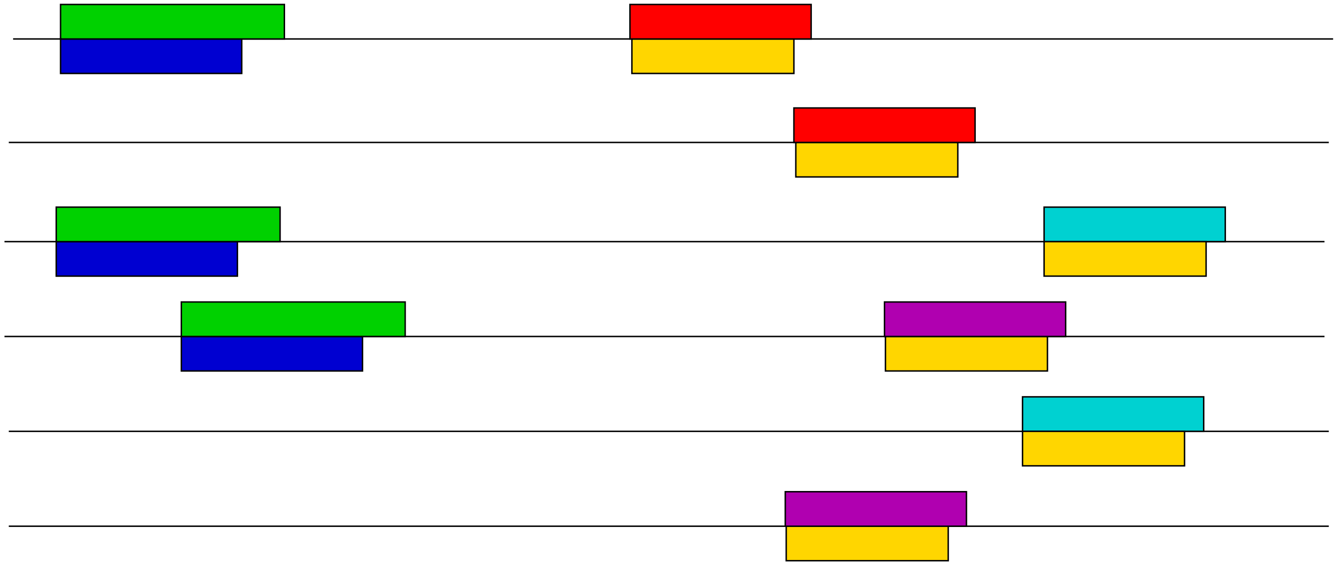
Basis of “irredundant” motifs

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



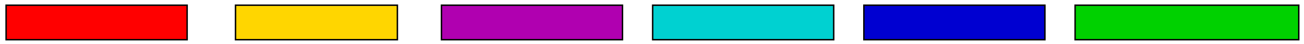
Occurrences in sequences (strings)

keep only “irredundant” motifs



Basis of “irredundant” motifs

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



Occurrences in sequences (strings)

keep only
“irredundant” motifs

(Parida *et al.*, 2000)

An example

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

Basis of irredundant motifs

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: A^{p}TA^p

What was believed initially

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

Basis of irredundant motifs

What was proven 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 proven later

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

  Ex.: a_1 ●●a_1 ●●●a_1 AAAAAATAAAAAA

A●AAAA	1,6,8	A● ●AAA	1,5,6,8
AA●AAA	1,5,8	A●A●AA	1,4,6,8
a_1 $\bullet^7$AAA●AA	-7,-4,0	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

Basis of irredundant motifs

What was proven later

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

  Ex.: $a_3 a_1 \bullet\bullet a_1 a_3 \bullet a_3 a_1$AAAAAATAAAAAA

$a_3 \bullet^8$A•AAAA	-8,-3,-1	A• •AAA	1,5,6,8
AA•AAA	1,5,8	A•A•AA	1,4,6,8
$a_1 \bullet^7$AAA•AA	-7,-4,0	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

Basis of irredundant motifs

What was proven later

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

  Ex.: $a_2 \bullet a_2 \bullet\bullet\bullet\bullet a_2 a_3 a_1 \bullet\bullet a_1 a_3 \bullet a_3 a_1$AAAAAATAAAAAA

$a_2 \bullet^8$A•AAAA	-8,-3,-1	A• •AAA	1,5,6,8
AA•AAA	1,5,8	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

Basis of irredundant motifs

What was proven 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 proven later

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

  Example: $s_k = u_k v_k t_k$ **with** $t_k = \mathrm{A}^k\mathrm{T}\mathrm{A}^k$ **and, if** k **is even:**

$$
\begin{aligned}
ev_k &= a_2 a_4 \cdots a_{k-2}, \\
od_k &= a_1 a_3 \cdots a_{k-3}, \\
u_k &= ev_k\, u\, \widetilde{ev_k}\, v w\, ev_k, \\
v_k &= od_k\, xy\, od_k\, z\, od_k.
\end{aligned}
$$

  or, if k **is odd:**

$$
\begin{aligned}
ev_k &= a_2 a_4 \cdots a_{k-3}, \\
od_k &= a_1 a_3 \cdots a_{k-2}, \\
u_k &= ev_k\, uv\, \widetilde{ev_k}\, wx\, ev_k, \\
v_k &= od_k\, y\, od_k\, z\, od_k.
\end{aligned}
$$

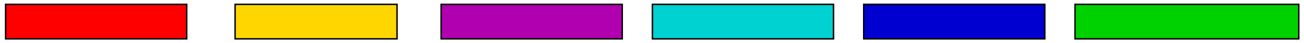
- **there are $\Omega\left(\frac{1}{2^q}\binom{n-1}{q-1}\right)$ irredundant motifs over Σ' for any q**

  Example: $t_k = \mathrm{A}^k\mathrm{T}\mathrm{A}^k$

(Pisanti, Crochemore, Grossi, Sagot, MFCS'03 and submitted)

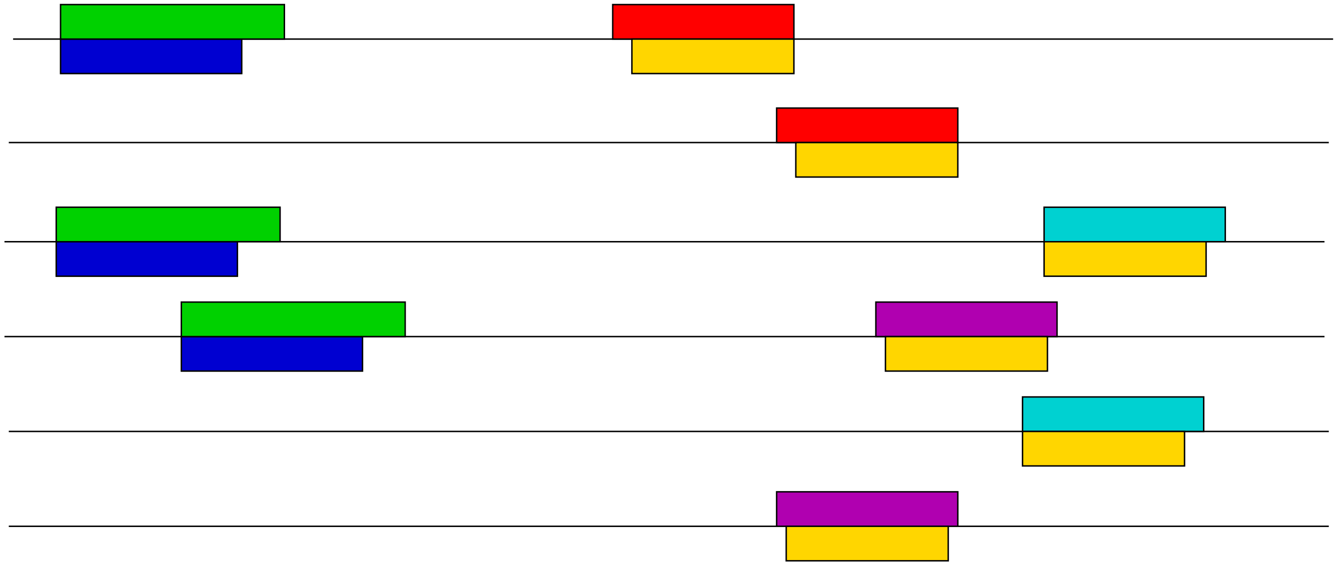
Basis of “tiling” motifs

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



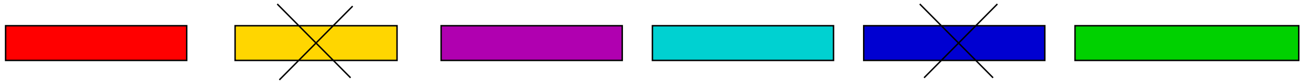
Occurrences in sequences (strings)

keep only
“tiling” motifs



Basis of “tiling” 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, Crochemore, Grossi, Sagot, MFCS’03 and submitted)

An example

$s = \mathbf{AAATAAA}$

Non maximal motifs: A••A, A••AA, AA••A

Non tiling motifs: AA, A•A, A• • •A

Basis of tiling motifs: AAA, A•A•A, AA••AA

Basis of tiling motifs

What is proven

- there are $\Omega\left(\frac{1}{2^q}\binom{n-1}{q-1}\right)$ tiling 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 $q\binom{n-1}{q-1}$
- in particular, for $q = 2$:
  - there are at most $n-1$ tiling motifs
  - the total number of occurrences over all tiling motifs is at most $2n-2$
  - there exists an algorithm in $O(n^2 \log n \log |\Sigma|)$ to find the basis

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, 9:761-773, 2003)

Still a LOT of work to be done

In particular concerning the "BACKGROUND": are we sure we know it ??

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

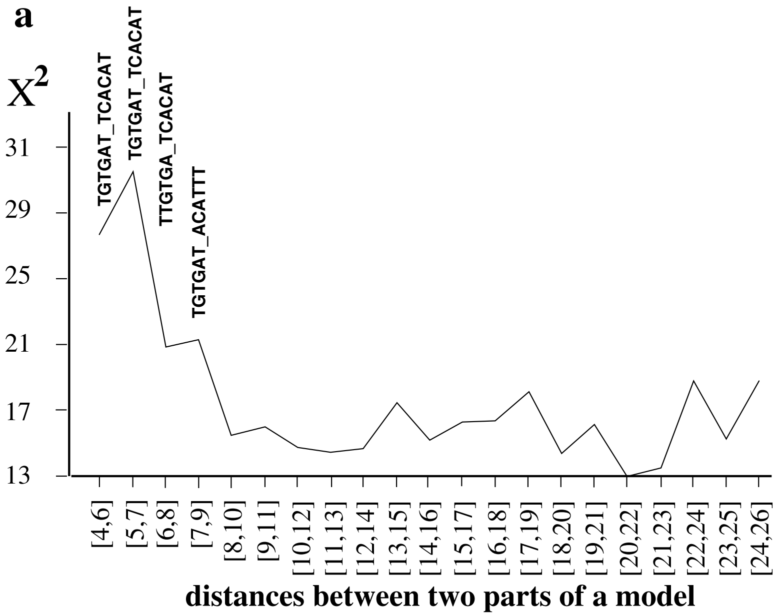
Escherichia coli **and** *Bacillus subtilis*

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

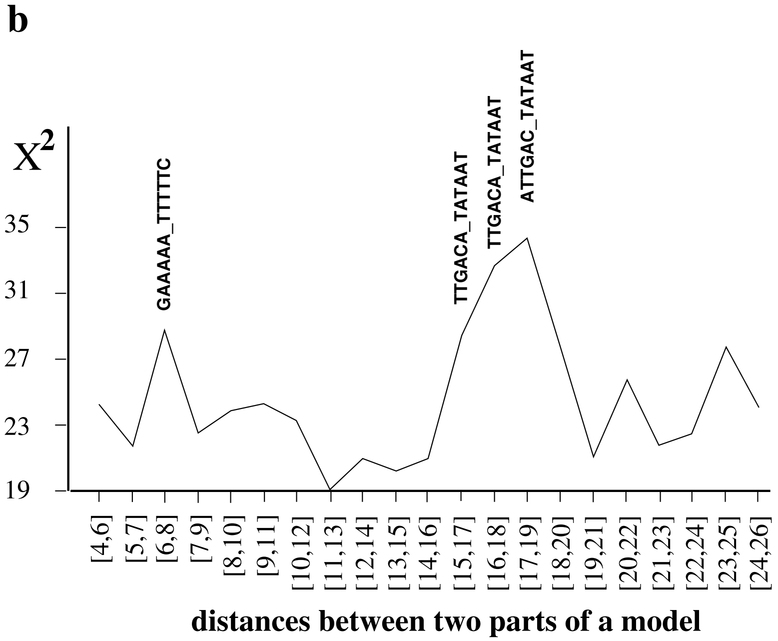
(A. Vanet, L. Marsan, M.-F. Sagot, *Res. in Microb.***, 150:779-799, 1999)**

Basic result

Escherichia coli



Bacillus subtilis



(S. Provencher, R. Tavares *et al.*, paper in preparation)

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 ?

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**

(Collaboration with Hidde de Jong and Hans Geiselmann)

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

etc

What we are aiming at

Biological questions

obtain a catalog, as complete as possible, of all the elements involved in the expression and regulation of genes in any type of organism

study the evolution of the regulation system and its relation with the structure of a genome

A related problem: repeats and regulation

Question: is there a relation between the distribution of repeats (tandem, palindromic or dispersed) along a genome and the gene expression level ?

requires a mapping as accurate and exhaustive as possible of ALL repeats in one (various) genome(s)

and therefore elaborating EXACT algorithms, possibly different ones for different types of repeats

What exists already:

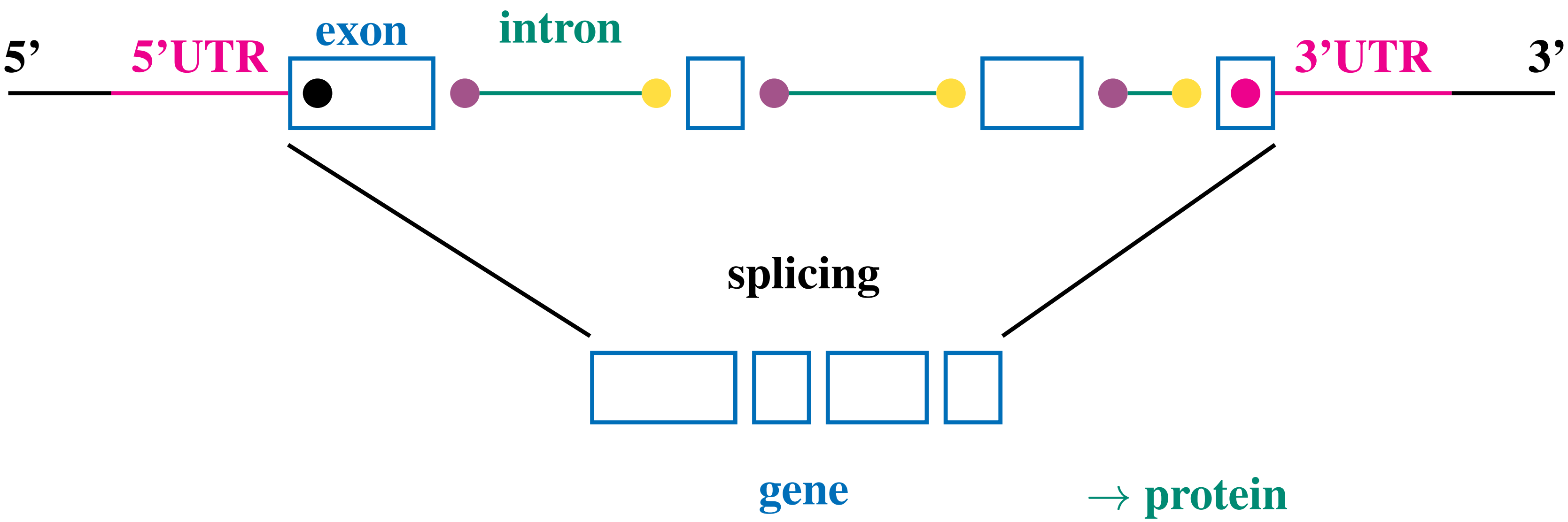
satellites (Sagot and Myers, 1998); Reputer (Kurtz *et al.*, 1999)

Another somewhat related problem: motifs and genes

Reminder: A very elementary view of an eukaryotic gene

start codon **GT (donor site)** **AG (acceptor site)** **stop codon**

3 nucleotides (codon) → 1 amino acid



5’UTR and 3’UTR: transcribed into RNA but not translated

Gene finding initially in a very particular context: that of "orphan genes"

An orphan gene is a gene for which no homology (in the sense here of having "strong enough" similarity) has been detected with the sequences stored in the databases

Main idea around the problem

An orphan may have undetected "parents", that is other orphans like itself which are its HOMOLOGS (having common ancestor)

ORTHOLOGS more similar ?

PARALOGS having diverged more ?

in both cases, having possibly different gene structures (*i.e.* a different number of exons)

Additional hypothesis
(important but is it always justified ?)

Exons are "better conserved" or, more accurately, "differently conserved" than introns or 5'UTRs or 3'UTRs or intergenic sequences

Flavour of method

Find gene structure by comparing orphans which are undetected homologs using as information only the bare essentials

(method by "pure homology")

Using a dynamic programming approach with a few twists

Sequences are composed of coding and non coding regions

There are two potential types of "errors"

Nature's (gaps, substitutions)

Man's (sequencing reading errors = "frameshifts")

UTOPIA (P. Blayo *et al.*, TCS 290:1407-1431, 2003)
(C. Mathé *et al.* NAR, 30:4103-4117, 2002)
(R. Tavares, S. Provencher *et al.*, in preparation)

Objective

Find best assembly of exons

which satisfies “bare essentials” gene model

where “best” means highest scoring assemblage of exons

Why a combinatorial, “bare essentials” type of approach ?

It does not substitute for other, statistical in particular, approaches

It cannot (perhaps) even compete with them (it was not meant to)

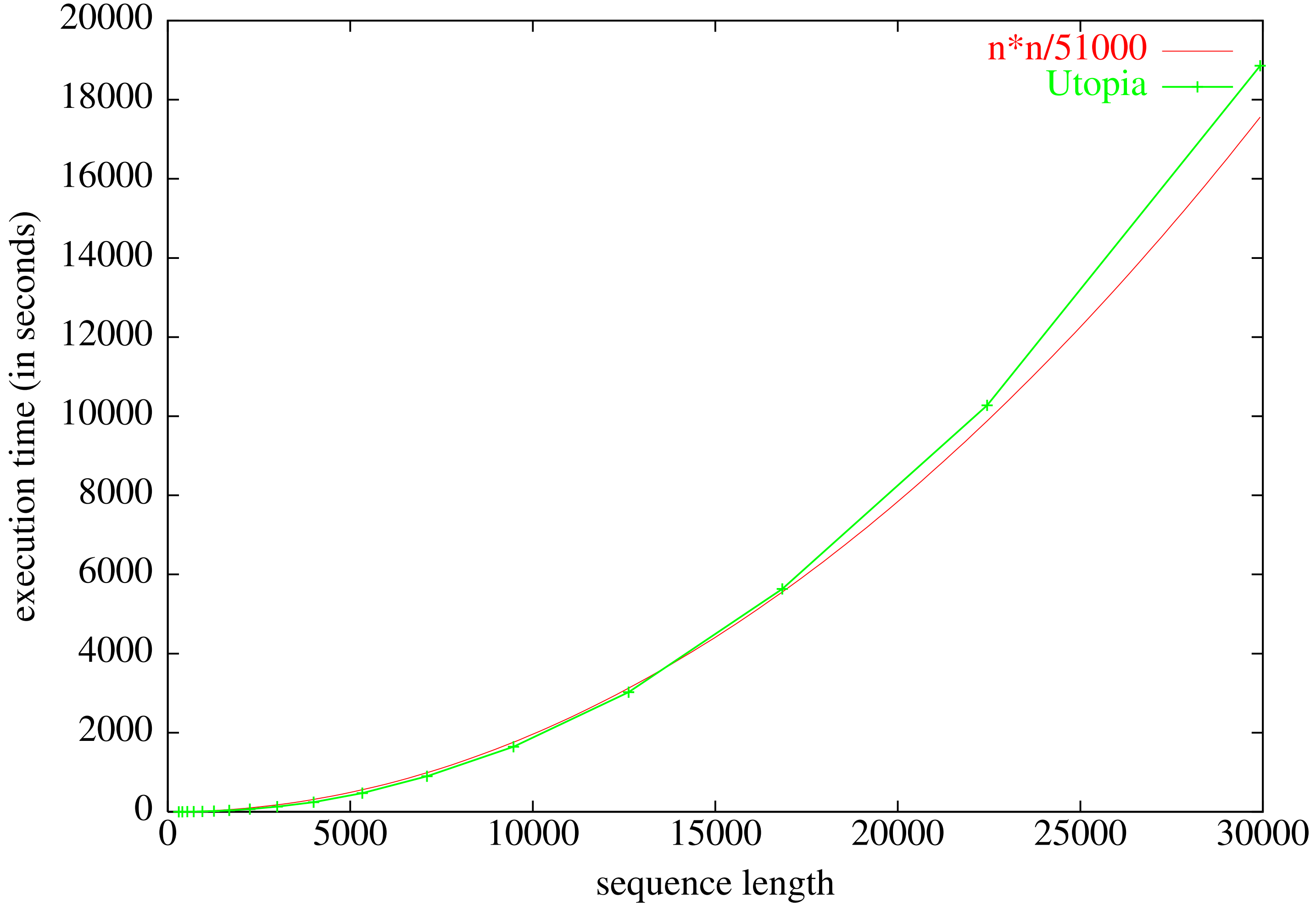
BUT

It allows, indeed obliges to think over again our notions of “conservation” and, in particular, of the non-conservation of non-coding regions

It is GENERIC

in particular, it is independent of what can be learned from specific characteristics of known examples of genes

Cannot even compete with them



Preliminary applications (1)

7 genes from a multigene family in *Arabidopsis thaliana*

of unknown function

going by the name of MYST

of different gene structures and lengths

13 to 15 exons

from 1851 bp. to 2034 bp.

(average) identity 56%

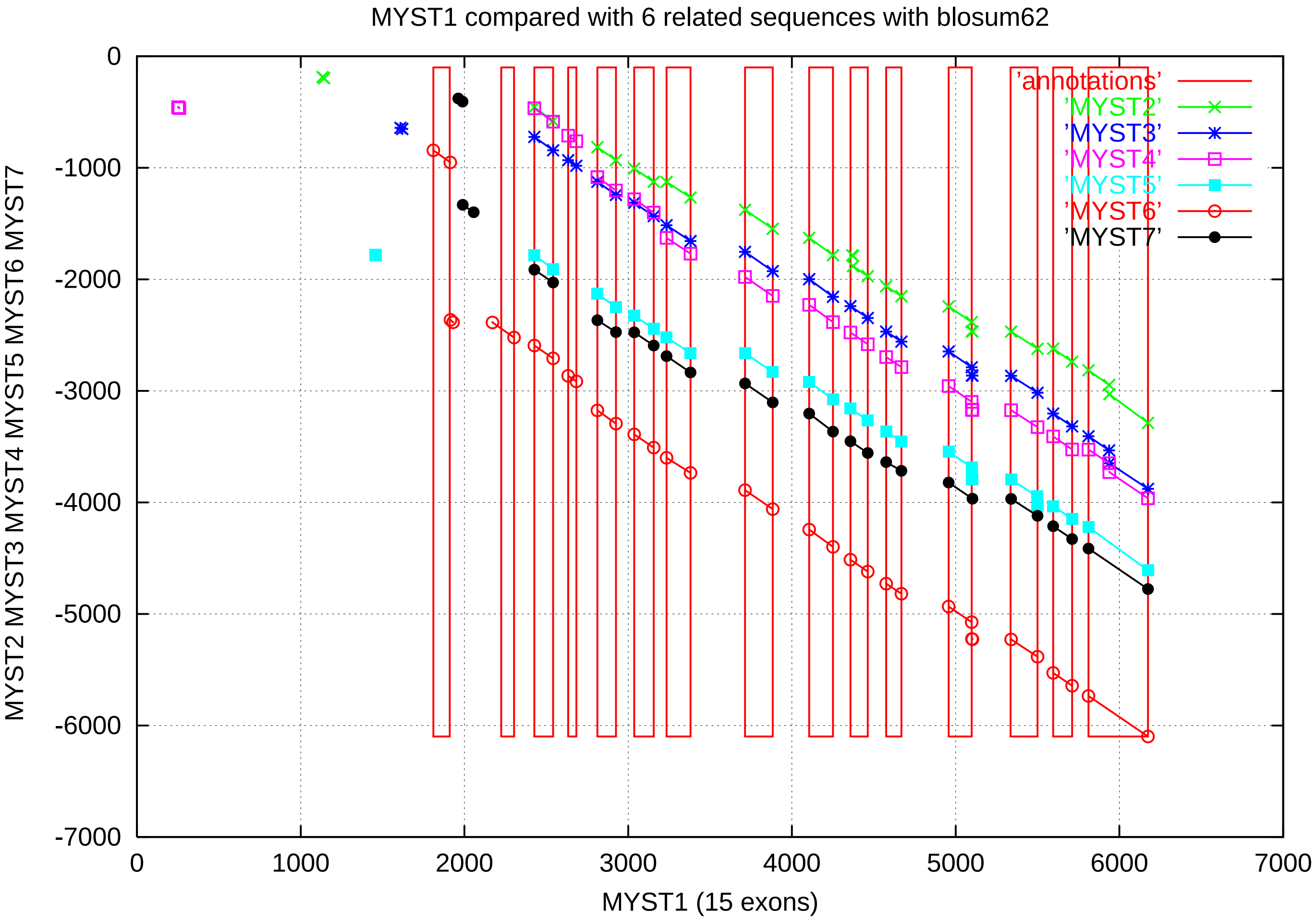
Sensitivity and specificity

Sensitivity

$$sensitivity = \frac{number\ of\ correctly\ predicted\ items}{number\ of\ actual\ items} = \frac{TP}{TP + FN}$$

Specificity

$$specificity = \frac{number\ of\ correctly\ predicted\ items}{number\ of\ predicted\ items} = \frac{TP}{TP + FP}$$



Specificity: 98%, 87%, 79%; Sensitivity: 93%, 81%, 73%

Preliminary applications (2)

5 genes from a multigene family in *Arabidopsis thaliana*

of unknown function

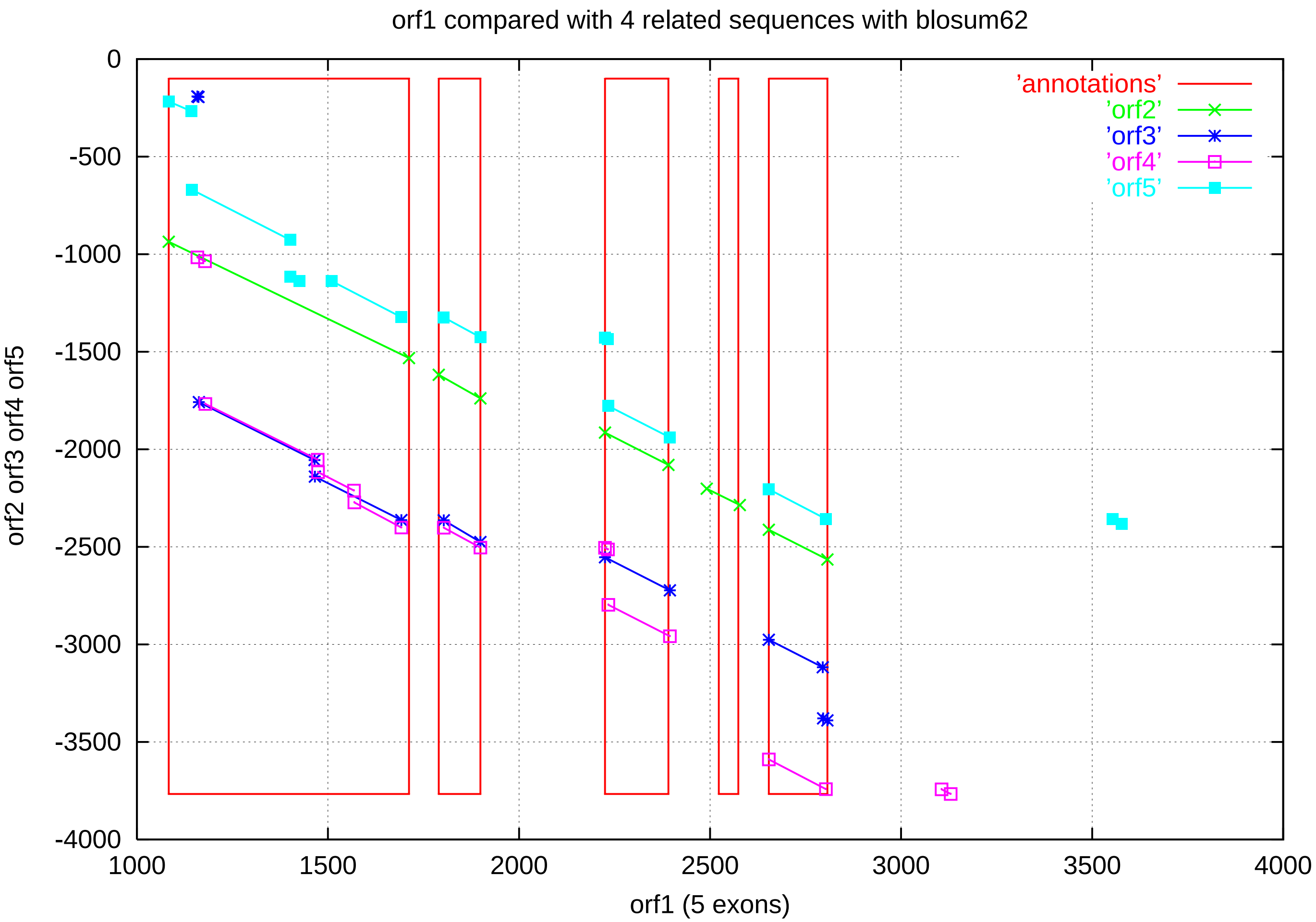
going by the name of ORF

of different gene structures and lengths

3 to 4 exons

from 1122 bp. to 1389 bp.

(average) identity 45%



Specificity: 93%, 45%, 29%; Sensitivity: 82%, 51%, 32%

Preliminary applications (3)

28 genes from a multigene family in *Arabidopsis thaliana*

function: terpene synthase

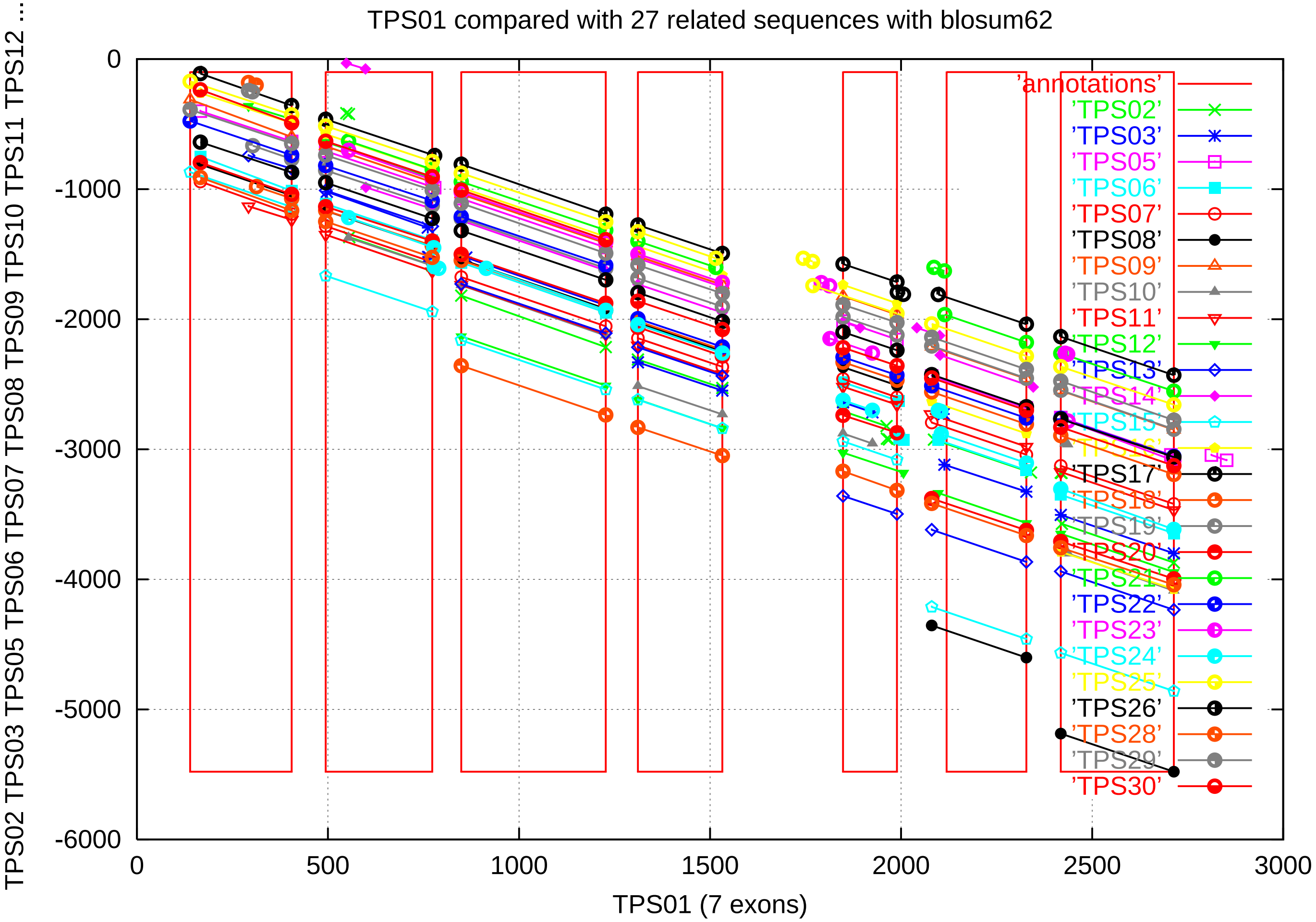
going by the name of TSP

of different gene structures and lengths

5 to 6 exons

from 1641 bp. to 1851 bp.

(average) identity 38%



Specificity: 98%, 83%, 70%; Sensitivity: 92%, 79%, 67%

Preliminary applications (4): Annotation error detection

8 genes from a multigene family in *Arabidopsis thaliana*

function: cinnamyl alcohol deshydrogenase

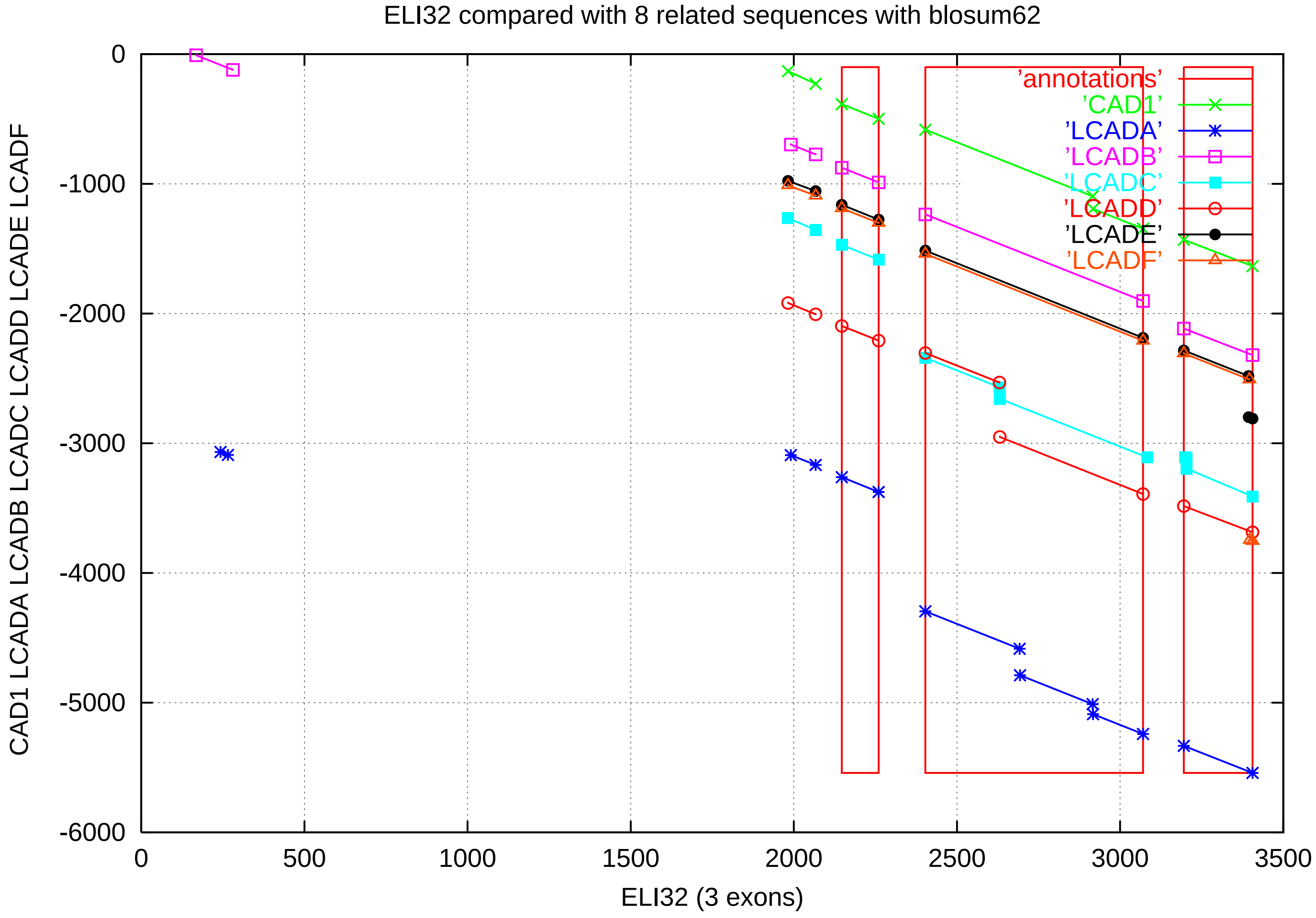
going by the name of CAD

of different gene structures and lengths

3 to 6 exons

from 1071 bp. to 1128 bp.

(average) identity 59%



Specificity: 93%, 87%, 81%; Sensitivity: 99%, 93%, 87%

What we are aiming at as concerns gene finding

Biological aims

develop a gene finding method that is as generic (*i.e* as organism independent) as possible

introduce a multiple (-gene and -organism) approach to the problem (includes DNA multiple alignment): for that, we probably need to put together motif inference and gene detection