Un lac, deux villes et trois rivières

sous-titre: Algorithmique combinatoire et biologie moléculaire

(et le reste du texte, mais pas de l'exposé, sera en anglais)

I am not a biologist (sorry)

I am not (really) a computer scientist (sorry)

I am an algorithmicist

that is:

someone who likes solving problems

by trying to understand their deeper structure

and thus the structure of the objects studied

The lac* which joined three rivers in two different towns

Lyon	Paris (Nov. 2001)
Nadia Pisanti	Julien Allali
Sébastien Provencher	Philippe Blayo
Marie-France Sagot	Frédéric Chataigner
Marina Zelwer	Sophie Créno
Brazil	Laurent Marsan
Estela M. Rodrigues	Raquel Tavares
Collaboration (Paris)	
Anne Vanet	Anne Cariou
Katja Schuerer	

computer scientist — biologist

*Laboratoire d'Algorithmique Combinatoire, Inst. Pasteur

with time, of course, people become

Lyon	**Paris (Nov. 2001)**
Nadia Pisanti	Julien Allali
Sébastien Provencher	Philippe Blayo
Marie-France Sagot	Frédéric Chataigner
Marina Zelwer	Sophie Créno
Brazil	Laurent Marsan
Estela M. Rodrigues	Raquel Tavares
Collaboration (Paris)	
Anne Vanet	Anne Cariou
Katja Schuerer	

MUTANTS

The way we (try to) work:

at the beginning

there is always (we hope)

a biological question

what are the areas we wish to cover ?

- **ALL AREAS at a greater or lesser depth of detail**

various (personal) reasons for that:

- **a need to see the "whole" in order to "structure" the details**
- **everything is related**
  - **biological questions**
  - **algorithmical/mathematical models and methods**

(still at a more personal level)

why the wish to join the INRIA ?

because algorithmics and mathematics are beautiful

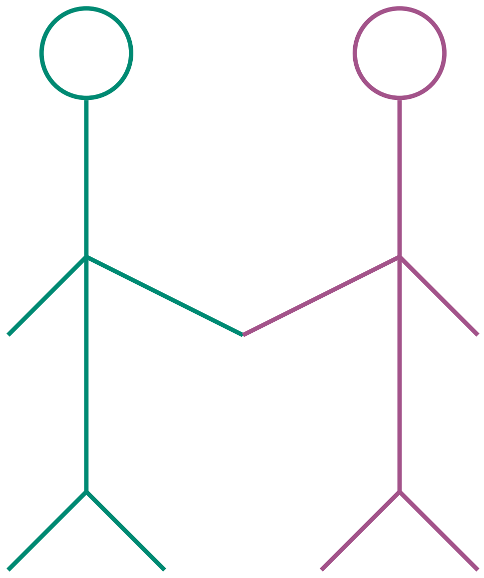
and the Équipe de Biométrie Moléculaire, Évolution et Structure des Génomes du Laboratoire de Biométrie et Biologie Évolutive ?

because biology is beautiful too !

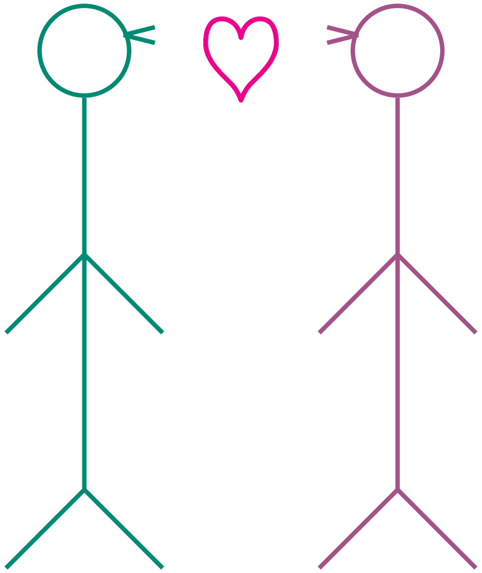
but one can easily grow too confortable (and stale)

and so one needs to look for some extreme experiences

Sometimes algorithmics and biology advance hand in hand

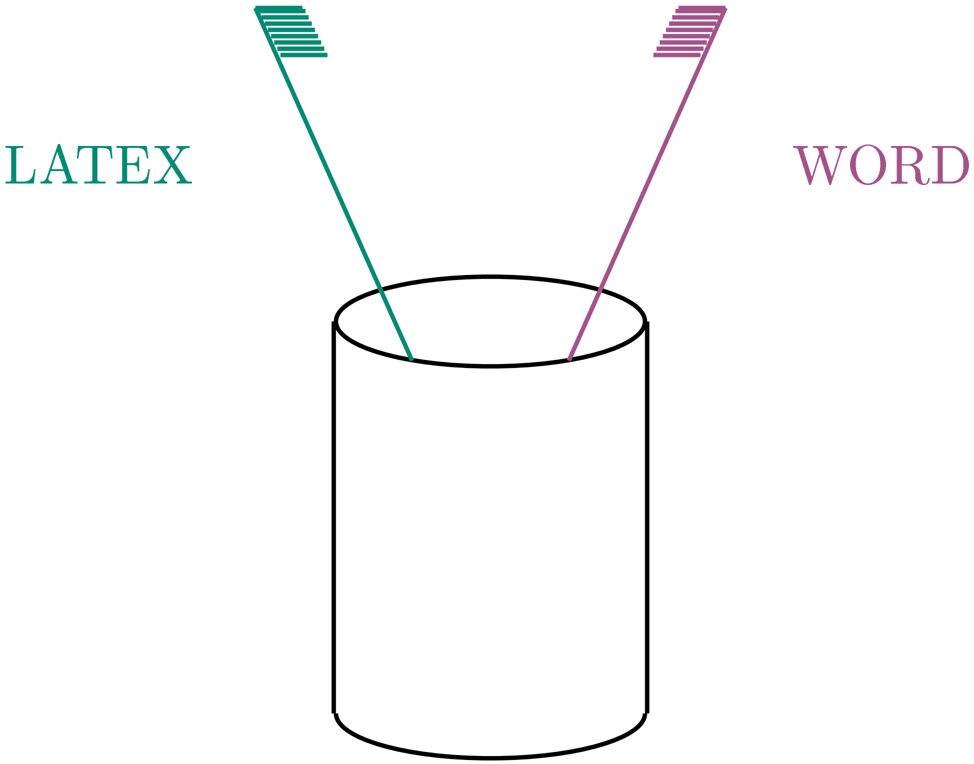


Sometimes in parallel



Each, of course, keeping its own personality

(or, later in life...)



what will this talk be about ?

a non technical survey

of the work we've been doing and of the perspectives

which tries to show our way of doing algorithmics

and why we wished to work with you

If any of these subjects interests you

there are more details in the papers

you can also come talk to us (for instance, over lunch or a cup of coffee/chocolate ?)

finally, there are (rudimentary) implementations of some algorithms available for use

First (VERY broad) area of research:

GENE REGULATION

AND

EXPRESSION

more precisely, initially, we'd like to be able to answer the question:

given a genome (or various genomes)

what are ALL the elements

involved in the expression and regulation of its genes ?

Can we “model” recognition/interaction ?

belongs to the (very broad) class of

INFERENCE algorithms

A journey of a thousand miles

starts with a single step

buddhist saying

Let us start

using as information

the sequence only

starting point: **a (very preliminary) MODEL**

elements sought = CONSERVED sequence elements

first question:

conservation at the level of the "letters" representing the monomers (nucleotides, amino acids) composing the sequences ?

conservation at the level of the physico-chemical properties of the monomers composing the sequences ?

"letter" conservation	physico-chemical conservation
TATAAT	run of pyrimidines
TTGNCA	
RFXCP	run of hydrophilic aa

or a mixture of both

TA[AT]N[AT]T

[ILMV][ASG]XXC[ILMV]H[FYW]P

in this talk: **(but we can treat all cases)**

“letter” conservation	**physico-chemical conservation**
TATAAT	**run of pyrimidines**
TTGNCA	
RFXCP	**run of hydrophilic aa**

or a mixture of both

TA[AT]N[AT]T

[ILMV][ASG]XXC[ILMV]H[FYW]P

Another (big) question

“statistical” conservation measure ?

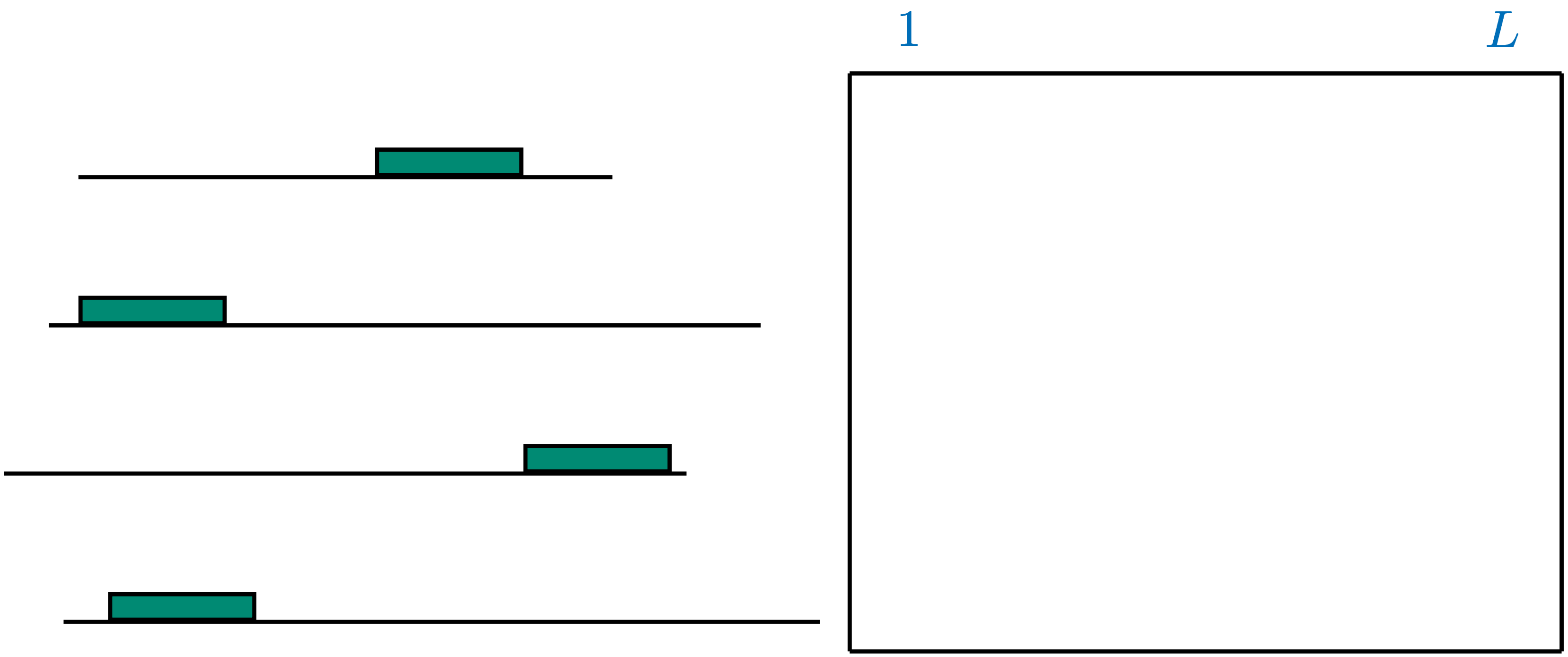
“deterministic” conservation measure ?

"Statistical" conservation measure

	G	T	T	T	T	T	C	T
	C	T	G	C	A	T	C	T
	G	T	G	T	A	A	C	C
	G	G	G	T	A	T	G	T
	T	T	G	T	C	T	C	T
	G	C	T	T	A	T	C	T
	A	T	G	T	C	T	C	T
	G	A	G	T	A	T	C	A
	G	T	G	T	A	G	G	T
	G	T	G	A	A	T	C	A
A	1	1	0	1	7	1	0	2
C	1	1	0	1	2	0	8	1
G	7	1	8	0	0	1	2	0
T	1	7	2	8	1	8	0	7

Sought: "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)



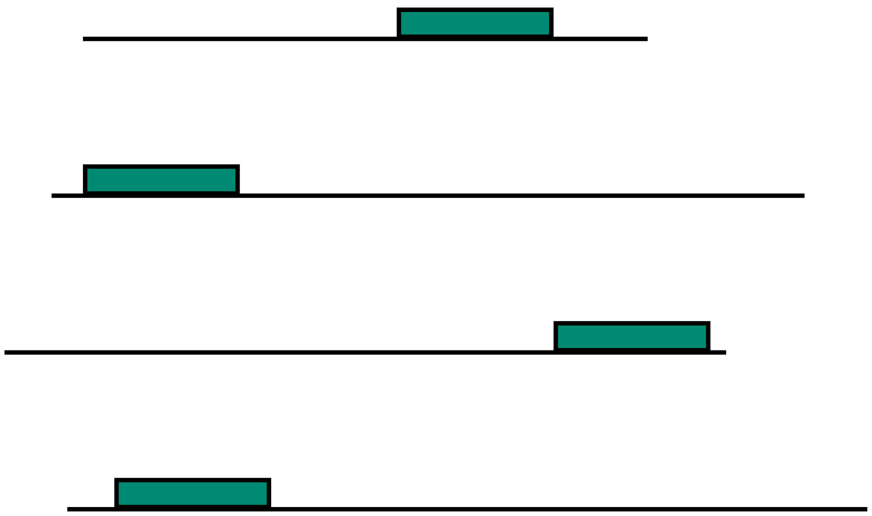
weighted average of the log-likelihood ratio

(the weights are the frequencies)

“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)



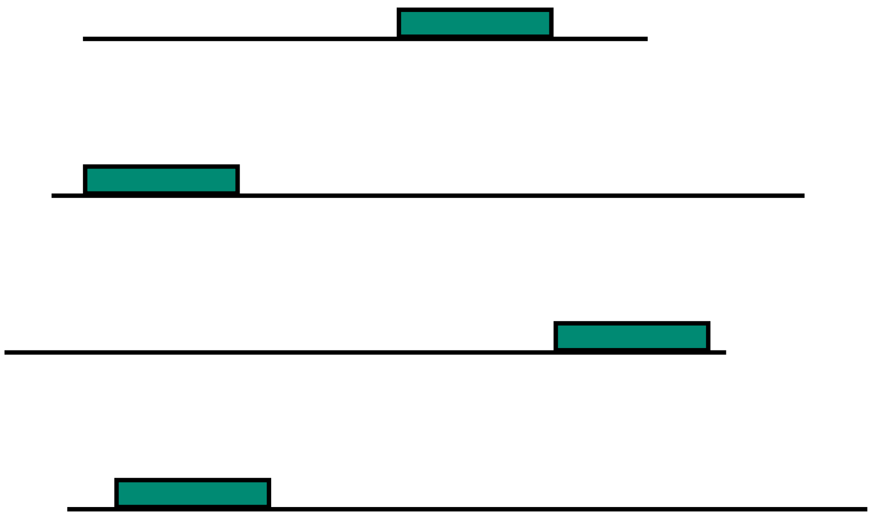
1				L
C	T	A	T	T
T	C	C	A	A
G	A	G	G	C
A	G	T	C	G

$$f_\alpha = \frac{1}{4}$$

$$0 = 0 \quad 0 \quad 0 \quad 0 \quad 0$$

“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)



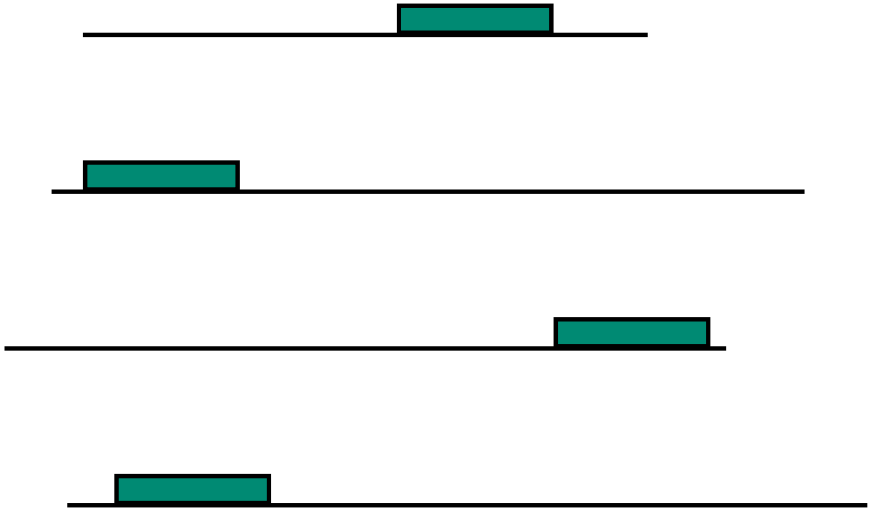
1				L
A	T	C	G	C
A	T	C	G	C
A	T	C	G	C
A	T	C	G	C

$$f_\alpha = \frac{1}{4}$$

$10 =$ 2 2 2 2 2

“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)



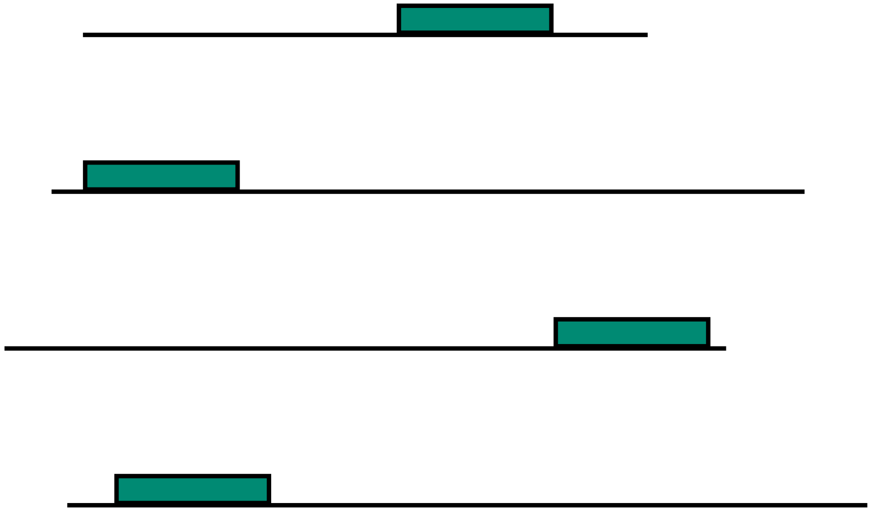
1				L
A	A	A	A	A
A	A	A	A	A
A	A	A	A	A
A	A	A	A	A

$$f_A = \frac{1}{16}$$

$20 =$ 4 4 4 4 4

“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)



$$\begin{array}{ccccc} 1 & & & & L \\ \hline A & A & A & A & A \\ A & A & A & A & A \\ A & A & A & A & A \\ A & A & A & A & A \end{array}$$

$$f_A = \frac{3}{4}$$

$$2 \; = 0.4 \quad 0.4 \quad 0.4 \quad 0.4 \quad 0.4$$

“Deterministic” conservation measure

“Model”

	G	T	G	T	A	T	C	T
2	G	T	T	T	T	T	C	T
2	C	T	G	C	A	T	C	T
2	G	T	G	T	A	A	C	C
2	G	G	G	T	A	T	G	T
2	T	T	G	T	C	T	C	T
2	G	C	T	T	A	T	C	T
2	A	T	G	T	C	T	C	T
2	G	A	G	T	A	T	C	A
2	G	T	G	T	A	G	G	T
2	G	T	G	A	A	T	C	A

Sought: ALL models where a model is:

A motif

a word written over the same alphabet as the text, or over a degenerate (physico-chemical) alphabet

A number of properties (constraints)

for each occurrence, maximum number of differences allowed in relation to the motif

(substitutions only, or substitutions and indels)

minimum number of occurrences the motif must have (quorum)

The question is algorithmically not innocent

“statistical” conservation $\rightarrow$ **“heuristical” algorithms**

“deterministic” conservation $\rightarrow$ **“exact” algorithms**

How does the algorithm based on a “deterministic” conservation measure work ?

It does not matter since the algorithm is exact !

(for the algorithmicist: it is a tale of two trees, one a suffix – in fact, factor – tree, the other a trie)

However, this has in general to be followed by

a STATISTICAL EVALUATION

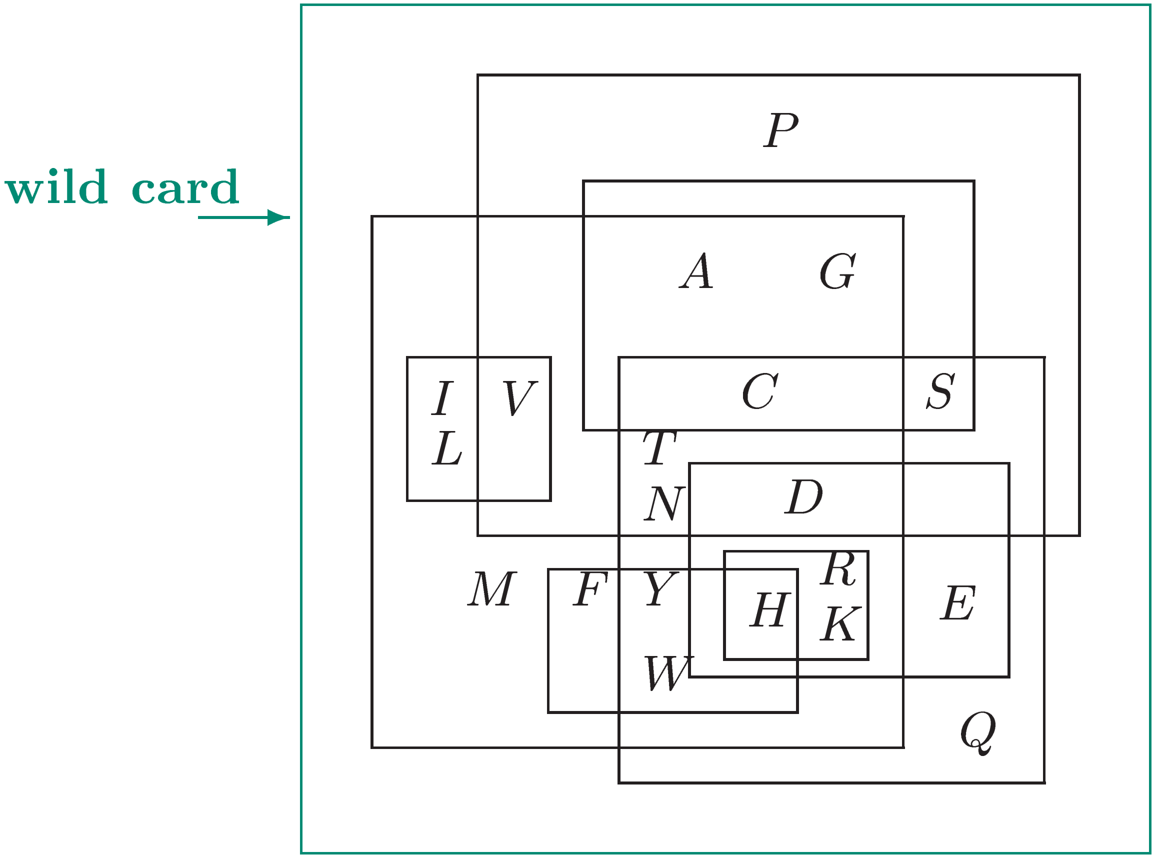
of the models found to classify them according to

how SURPRISING they are given the rest of the sequences

(there is still a LOT of work to be done in this area)

One can make the models more complex: motif inference with differences and a non transitive relation

Alphabet of models corresponds to groups of amino acids



One can make the models more complex: motif inference with differences and a non transitive relation

Example

models written over a physico-chemical alphabet

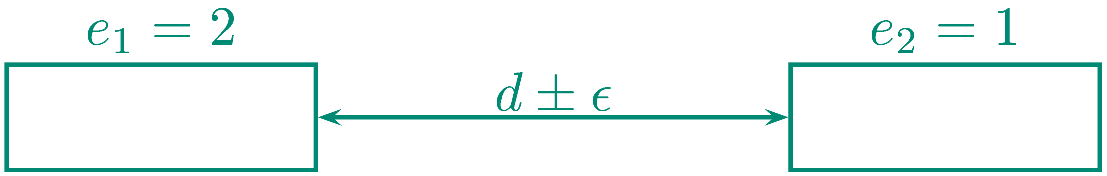
[A S T] [I L M V] X X [F Y W] [H K R] X [P G] C

occurrences

0 difference	1 substitution	1 deletion
AIAGWHAPC	ATTAYHSPC	SVMLFLPC

One can make the models more complex: structured models

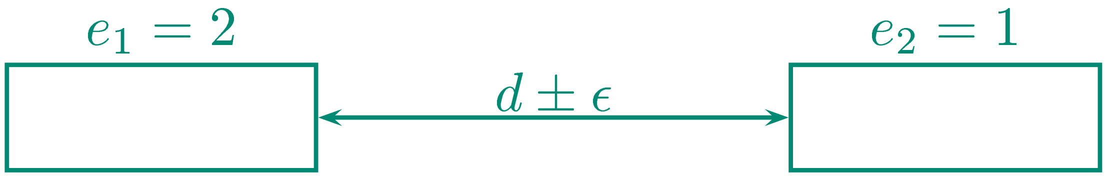
an ordered collection of p boxes, p maximum rates of differences (e), $p-1$ intervals of distances (between successive boxes in the collection) ($d \pm \epsilon$)



quorum

One can make the models more complex: structured models

an ordered collection of p boxes, p maximum rates of differences (e), $p-1$ intervals of distances (between successive boxes in the collection) ($d \pm \epsilon$)

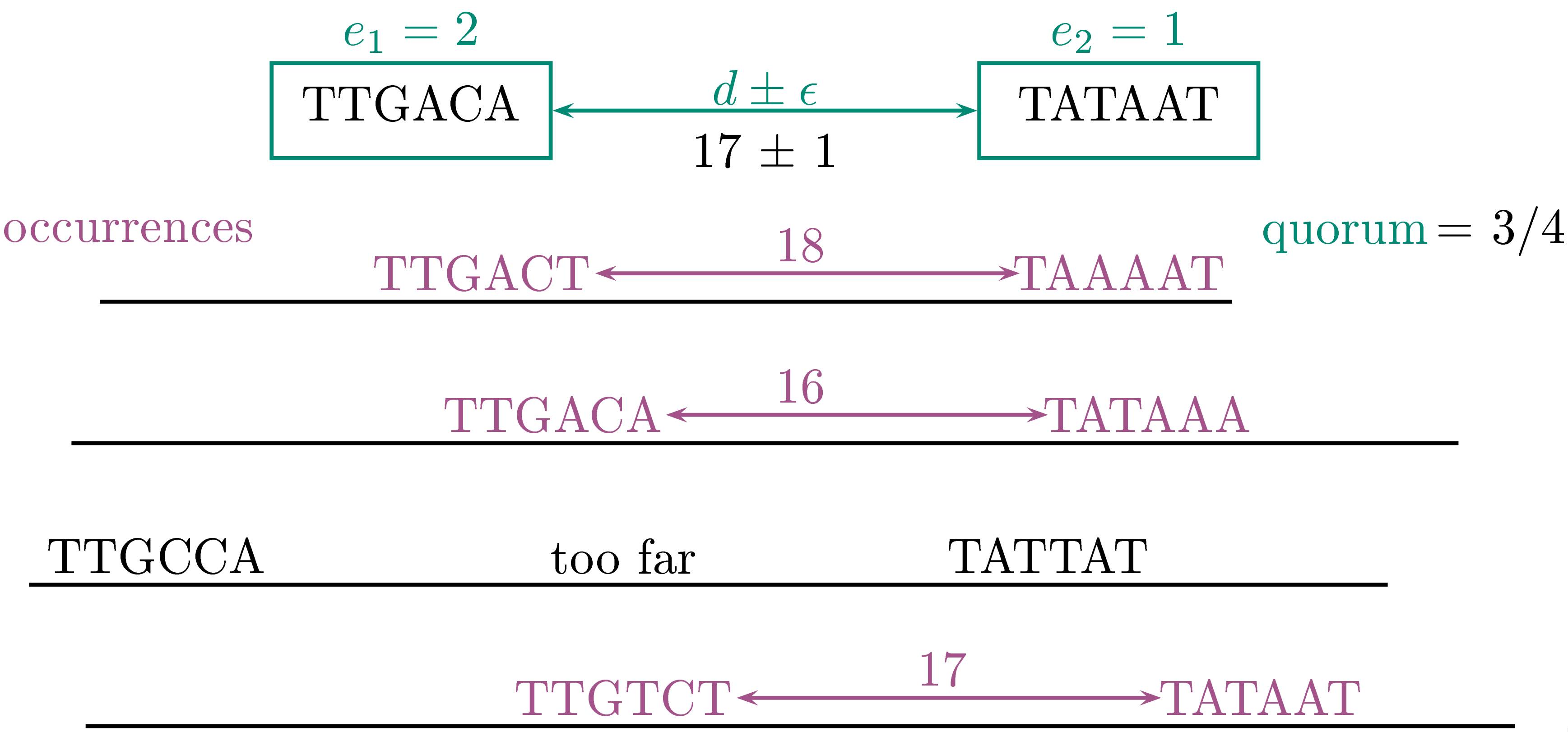


quorum

Smile (L. Marsan *et al.*, JCB, 7:345-362, 2000)

One can make the models more complex: structured models

an ordered collection of p boxes, p maximum rates of differences (e), $p-1$ intervals of distances (between successive boxes in the collection) ($d \pm \epsilon$)



Applications

A comparison of the two approaches

and initial application to:

Escherichia coli, Bacillus subtilis, Helicobacter pylori

both "clean" reduced datasets and "genomic" datasets

may be found in:

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

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

and we are preparing another paper with applications

Let us just mention in passing an example which may illustrate a possible difference (among others) between the two main types of approaches described

“Genomic” set of *Bacillus subtilis*

1148 sequences, 226928 nucleotides

“Genomic” set – Smile (2 boxes)

“Genomic” set – MEME

```
MOTIF  1  width =  12  sites = 308.7

         bits 2.2
              2.0
              1.7
              1.5
Information   1.3
content       1.1      **
(5.6 bits)    0.9      **
              0.7      **
              0.4     ***** *
              0.2    ********
              0.0 ------------

Multilevel        AAAAAAAGGAGG
consensus         TGG    ACGAA
sequence          CT     T   T
```

“Genomic” set – MEME

```
MOTIF  2  width =  22  sites = 54.4

         bits 2.2
              2.0
              1.7
              1.5
Information   1.3
content       1.1
(8.3 bits)    0.9
              0.7  **    *             *
              0.4 ***  * **       *    *
              0.2 ****** *****  * **  **
              0.0 ----------------------

Multilevel        GGCAGCAGCCCGTGCAGAGCGA
consensus          C T  C AAA GAATACCGAG
sequence             G  T     A   CTAAC
```

“Genomic” set – MEME

```
MOTIF  3  width =  43  sites = 173.2

          bits 2.2
               2.0
               1.7
               1.5
 Information   1.3
 content       1.1
 (11.3 bits)   0.9
               0.7  *                          *
               0.4 **  *           **          *  *
               0.2 *****     *   *****  ******** ** * * *   *
               0.0 -------------------------------------------

 Multilevel        TTTTTTCATAATTTTTTTTTTTTTCTTTTTTTATTTAATATTT
 consensus         CCCCCAACACCAACCACACACC CTCA  ACCTCAACTATAGA
 sequence               CTT TT CA         G     C    G T C   C
```

Further, currently in progress, applications

Vertebrates

- **on a dataset used by Ingrid Lafontaine for testing a completely structural approach to promoter and regulatory sequences detection (Lafontaine, Lavery)**
- **to identify splicing enhancer sequences (work with Claude Thermes and Yves d'Aubenton Carafa)**

Synechocystis **PCC 6803**

- **a bacterial genome that seems not very "bacterial-like"**

Mycobacterium tuberculosis & leprae **(A. Cariou, K. Schuerer)**

- **two very closely related genomes, one of which seems to be "shrinking" (around 1500 genes have already "disappeared" – but 165 are "new")**

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

Algorithmical and statistical issues

there are MANY, some exceedingly beautiful (and sometimes powerful) ones

curious ? $\rightarrow$ come talk to us !!

I shall mention just one

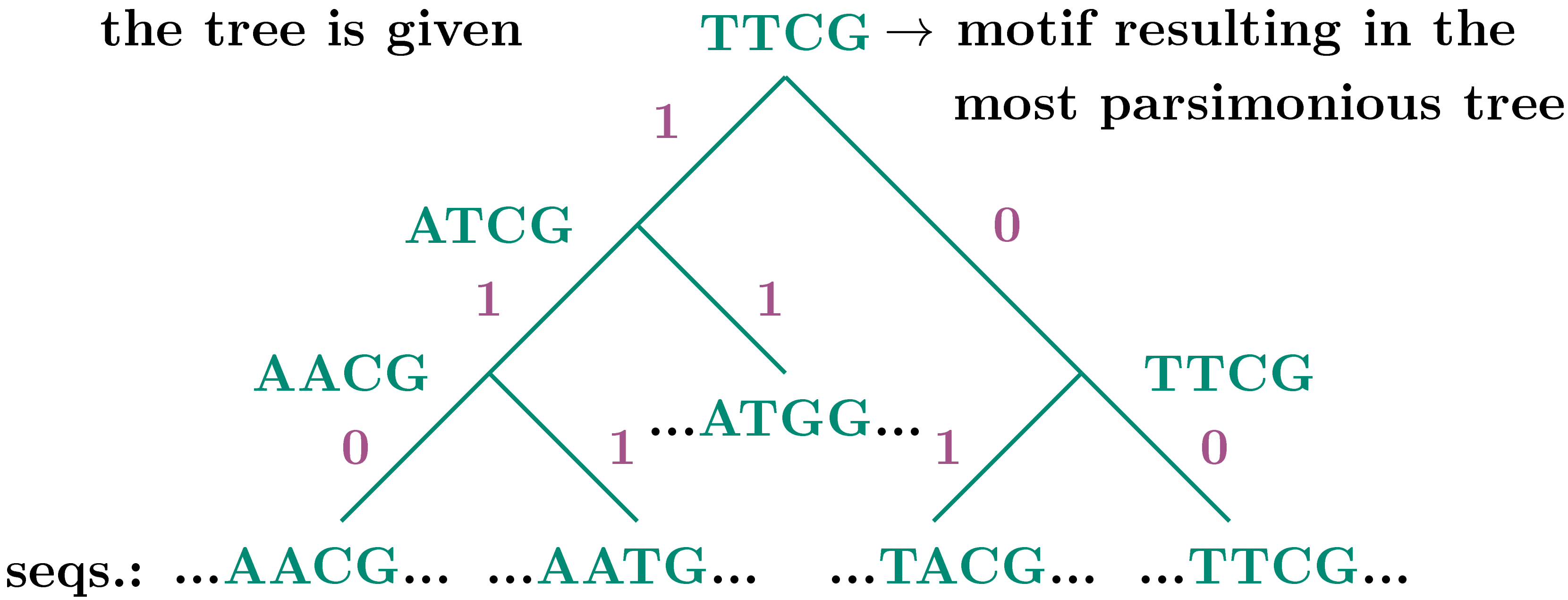
because it concerns

something I shall talk a little about later on

Phylogenetic footprinting

Main idea

A set of phylogenetically related sequences



total: 5

Second area of research:

GENOME STRUCTURE

what is meant by "genome structure" ?

a division of the genome into well-identified (big) units

genes

repeated elements (dispersed or in "organized fashion", in particular transposons and satellites)

(of course, there may be other units)

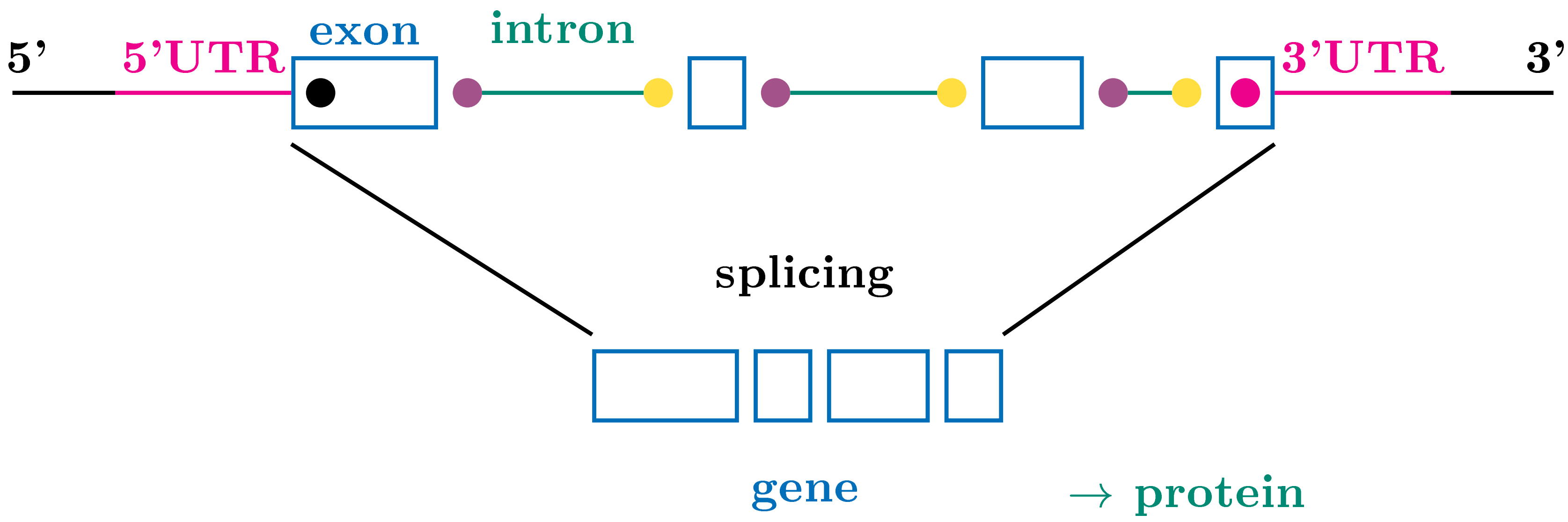
And first, gene finding in a very particular context: that of "orphan genes"

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

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

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")**

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.*, accepted TCS, 2001)

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 ?

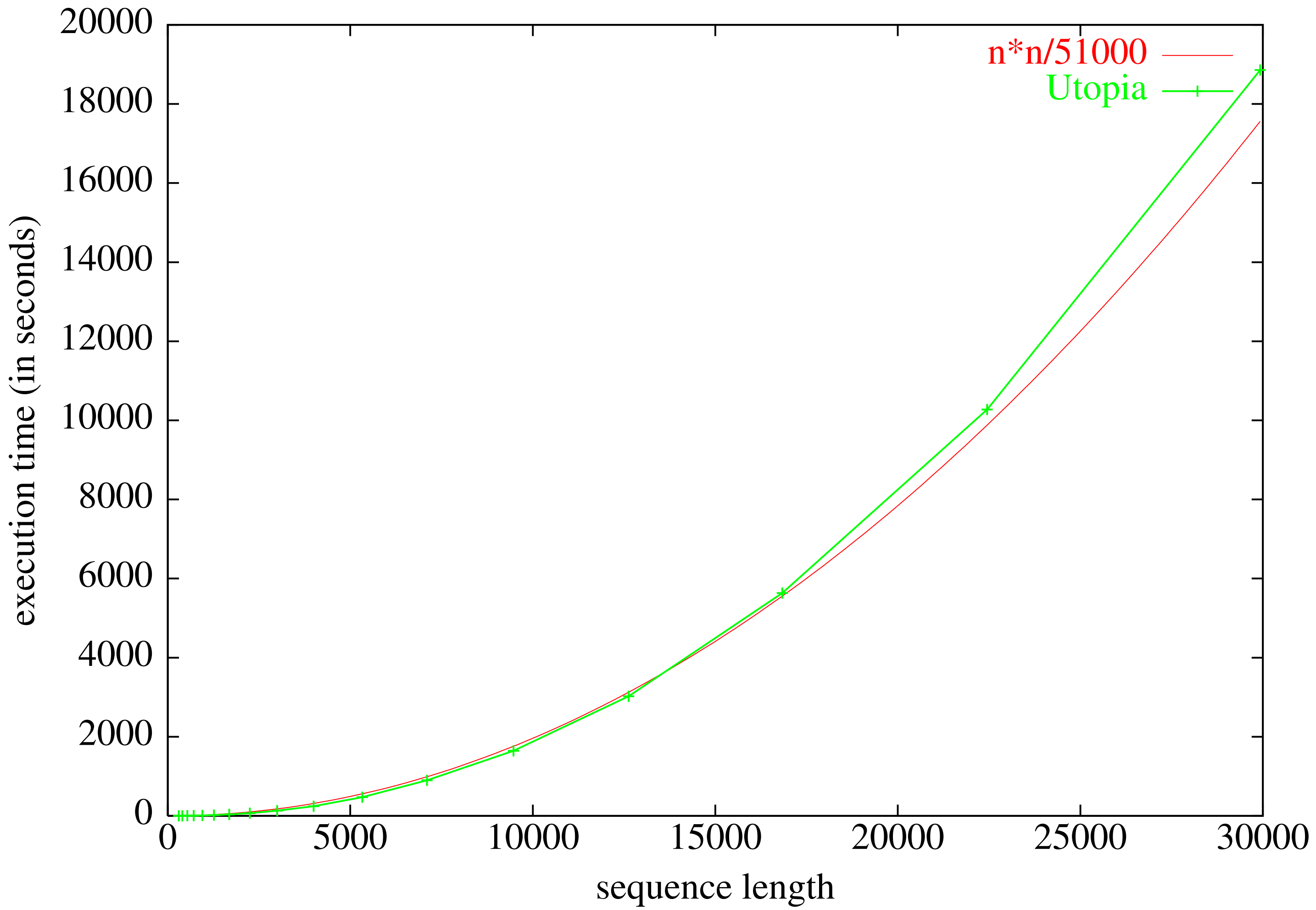
It does not substitute for other, statistical in particular, approaches

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

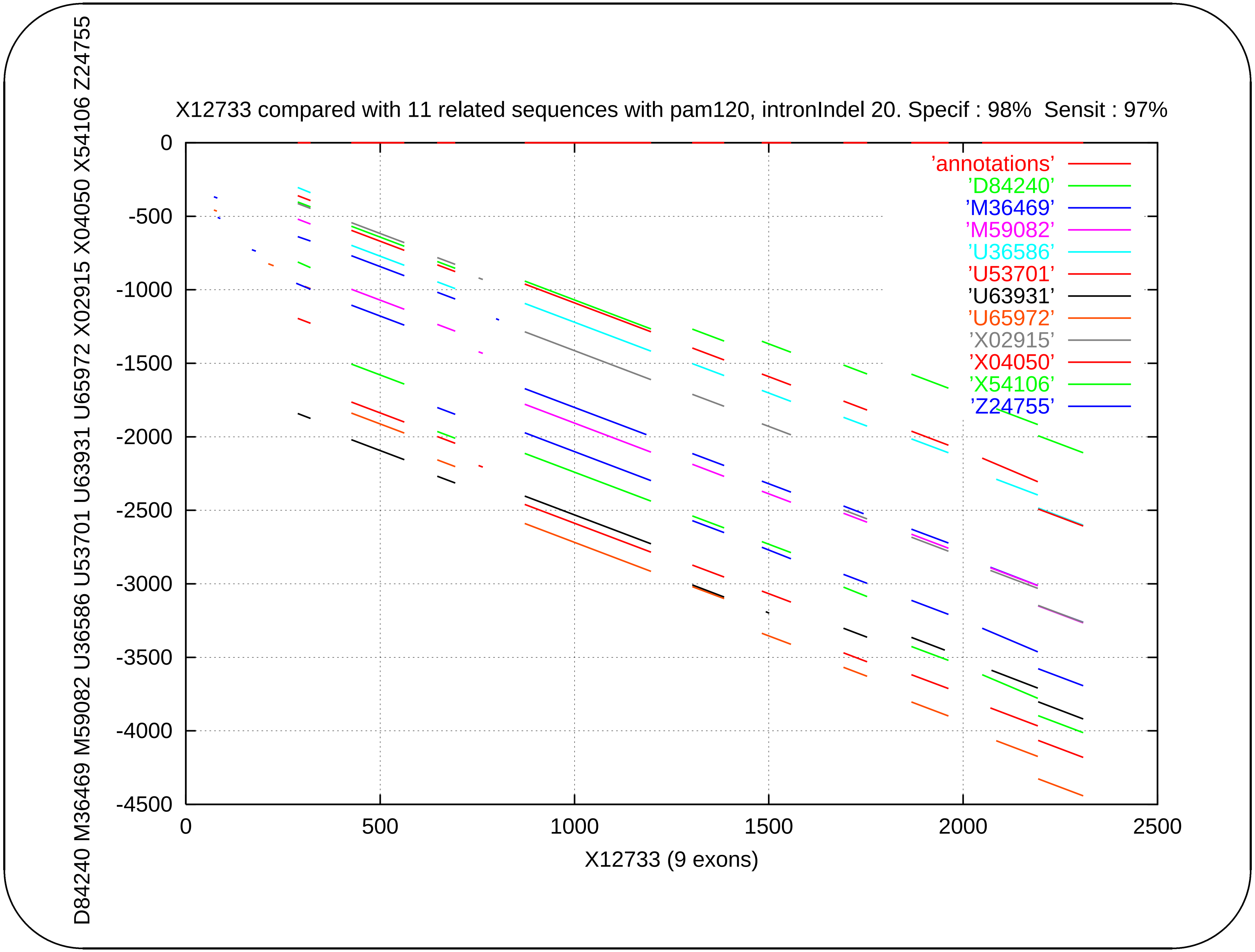
Cannot even compete with them:



Preliminary applications (1)

12 ADH protein genes of plants (among them *Arabidopsis thaliana*)

- **dico and monocotyledones**
- **one paralog and 11 orthologs**
- **of different gene structures and lengths**
  - **7 to 10 exons**
  - **from 982 bp. to 1143 bp.**



X12733 compared with 11 related sequences with pam120, intronIndel 20. Specif : 98% Sensit : 97%
'annotations'
'D84240'
'M36469'
'M59082'
'U36586'
'U53701'
'U63931'
'U65972'
'X02915'
'X04050'
'X54106'
'Z24755'
0
-500
-1000
-1500
-2000
-2500
-3000
-3500
-4000
-4500
0
500
1000
1500
2000
2500
X12733 (9 exons)
D84240 M36469 M59082 U36586 U53701 U63931 U65972 X02915 X04050 X54106 Z24755

Sensitivity and specificity

Sensitivity

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

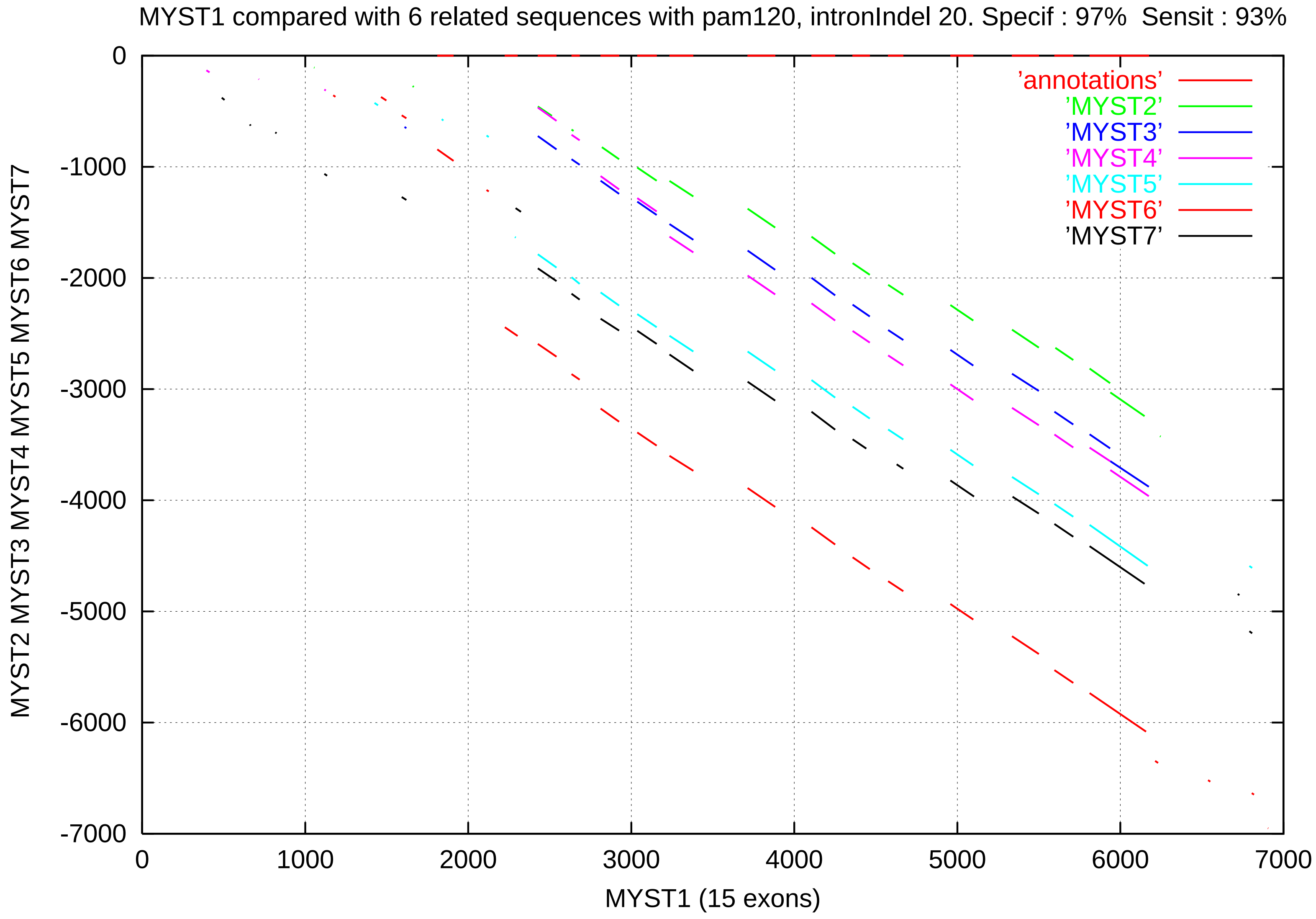
Specificity

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

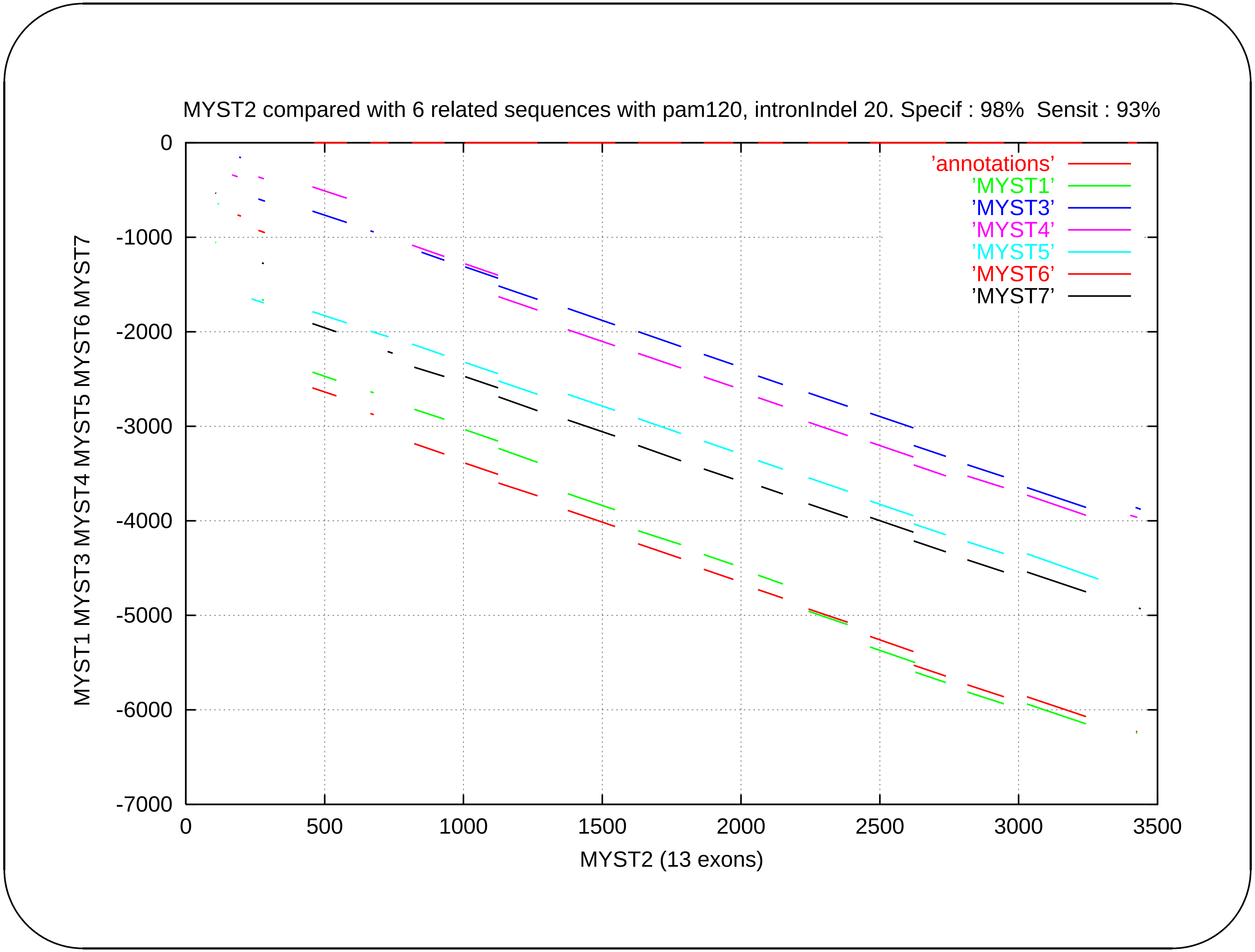
Preliminary applications (2)

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 1848 bp. to 2040 bp.**



MYST1 compared with 6 related sequences with pam120, intronIndel 20. Specif : 97% Sensit : 93%
0
-1000
-2000
-3000
-4000
-5000
-6000
-7000
0
1000
2000
3000
4000
5000
6000
7000
MYST1 (15 exons)
MYST2 MYST3 MYST4 MYST5 MYST6 MYST7
'annotations'
'MYST2'
'MYST3'
'MYST4'
'MYST5'
'MYST6'
'MYST7'


MYST2 compared with 6 related sequences with pam120, intronIndel 20. Specif : 98% Sensit : 93%
'annotations'
'MYST1'
'MYST3'
'MYST4'
'MYST5'
'MYST6'
'MYST7'
MYST1 MYST3 MYST4 MYST5 MYST6 MYST7
MYST2 (13 exons)
0
-1000
-2000
-3000
-4000
-5000
-6000
-7000
0
500
1000
1500
2000
2500
3000
3500

What we are aiming at

(Radical) biological aim

- **carefully think over again all that has been done around this subject so far**

(Sensible) biological aims

- **introduce a multiple (-gene and -organism) approach to the problem (includes DNA multiple alignment)**
- **for that, we probably need to:**
  - **put together motif inference (perhaps using a phylogenetic footprinting method ? – there are difficulties in doing that !) and gene detection**

Algorithmical issues

- **there are MANY, some exceedingly beautiful ones**
- **curious ? → come talk to us !!**

What preceded gives an idea of our way of doing algorithmics

(elaborating algorithms that are exact and as free from *a priori* as possible)

and of addressing biological questions with algorithms

What follows will go faster over what we have been doing

just suggesting the problems we have been treating and our perspectives

for more details concerning these, or what preceded, you may invite any of the group for another seminar

or come talk to us (over lunch or a cup of coffee/chocolate ?)

Still within "GENOME STRUCTURE"

Second type of unit:

Repeats, in particular, tandem arrays (satellites)

The problem informally put

```
gaaagctcttctgctccagtaccaactccatccagctctactact
gaaagctcttctgctccagt   aa      ccagctctacCact
gaaagctcttctgctccagtaccaactccatccagctctactact
gaaagctcttctgctccagCTccaacCccatcAagctcCacCact
gaaagctcttctgctccagt   aa      ccagctctacCact
gaaagctcttctgctccagtaccaactccatccagctctactact
gaaagctcttctactccagt   aa      ccagctccacCact
gaaagctcttctgctccagtGccaactccatcAagctctactact
gaaagctcttctgctccagtaccaactccatcAagctctactact
gaaagctcttctgctccagCTccaacCccatcAagctctactact
gaaagctcttctgctccagt   aa      ccagctctacCact
gaaagctcttctgctccagtaccaactccatccagctctactact
gaaagctcttctgctccagtaccaacCccatccagctctactact
gaaagctcttctgctccagtaccaactccatccagctctactact
```

The problem formally put

model for a satellite: a word m not necessarily present in the sequence s which approximately "matches" the "wagons of a train"

wagon of a model m: a word in s whose edit distance (minimum number of substitutions, insertions and deletions to obtain one word from another) from m is no more than e

maximum number of errors permitted:

$e = 1$

sequence

model

wagon

ATCCATCC **ACCCATCC**	**ACTCATCC** **ACCCATCC**	**ACCCATCC** **ACCCATCC**
ACCCATCC **ACCCATCC**	**ATCCATCC** **ACCCATCC**	**A CC T** **ACCCATCC**
ATCCATCC **ACCCATCC**	**ATCCATCC** **ACCCATCC**	**ACCCATCC** **ACCCATCC**
ACCCATCC **ACCCATCC**	**TTCCATTT** **ACCCATCC**	**ACCCATCC** **ACCCATCC**

The problem formally put

train associated with a model m: an ordered collection of wagons, $u_1, u_2, \ldots, u_p$ such that:

Property 1 there are enough wagons

maximum number of errors permitted:
$e = \mathbf{1}$, $min\_repeat = \mathbf{10}$

sequence

model

wagon

1 **ATCCATCC** **ACCCATCC**	**2** **ACTCATCC** **ACCCATCC**	**3** **ACCCATCC** **ACCCATCC**
4 **ACCCATCC** **ACCCATCC**	**5** **ATCCATCC** **ACCCATCC**	**A CC T** **ACCCATCC**
6 **ATCCATCC** **ACCCATCC**	**7** **ATCCATCC** **ACCCATCC**	**8** **ACCCATCC** **ACCCATCC**
9 **ACCCATCC** **ACCCATCC**	**TTCCATTT** **ACCCATCC**	**10** **ACCCATCC** **ACCCATCC**

The problem formally put

train associated with a model m: an ordered collection of wagons, $u_1, u_2, \ldots, u_p$ such that:

Property 1 **there are enough wagons**

Property 2 **the distance between the starts of two successive wagons is inside an interval** $[min\_range, max\_range]$

maximum number of errors permitted:

$e = \mathbf{1}$, $min\_repeat = \mathbf{10}$

sequence $min\_range = \mathbf{4}$, $max\_range = \mathbf{10}$

model

ATCCATCC	**ACTCATCC**	**ACCCATCC**
ACCCATCC	**ACCCATCC**	**ACCCATCC**
8	**8**	**8**
ACCCATCC	**ATCCATCC**	**A CC T**
ACCCATCC	**ACCCATCC**	**ACCCATCC**
8		
ATCCATCC	**ATCCATCC**	**ACCCATCC**
ACCCATCC	**ACCCATCC**	**ACCCATCC**
8	**8**	**8**
ACCCATCC	**TTCCATTT**	**ACCCATCC**
ACCCATCC	**ACCCATCC**	**ACCCATCC**

The problem formally put

train associated with a model m: an ordered collection of wagons, $u_1, u_2, \ldots, u_p$ such that:

Property 1 **there are enough wagons**

Property 2 **the distance between the starts of two successive wagons is inside an interval $[min\_range, max\_range]$, or a multiple $max\_jump > 1$ of that interval**

maximum number of errors permitted:

$e = 1$, $min\_repeat = 10$, $max\_jump = 2$

$min\_range = 4$, $max\_range = 10$

sequence

model

jump ("bad wagons" between two "good" ones)

ATCCATCC **ACTCATCC** **ACCCATCC**

ACCCATCC **ACCCATCC** **ACCCATCC**

8 8 8

ACCCATCC **ATCCATCC** **A CC T**

ACCCATCC **ACCCATCC** **ACCCATCC**

8 $2 \cdot 4 \leq 12 \leq 2 \cdot 10$

ATCCATCC **ATCCATCC** **ACCCATCC**

ACCCATCC **ACCCATCC** **ACCCATCC**

8 8 8

ACCCATCC **TTCCATTT** **ACCCATCC**

ACCCATCC **ACCCATCC** **ACCCATCC**

$2 \cdot 4 \leq 16 \leq 2 \cdot 10$

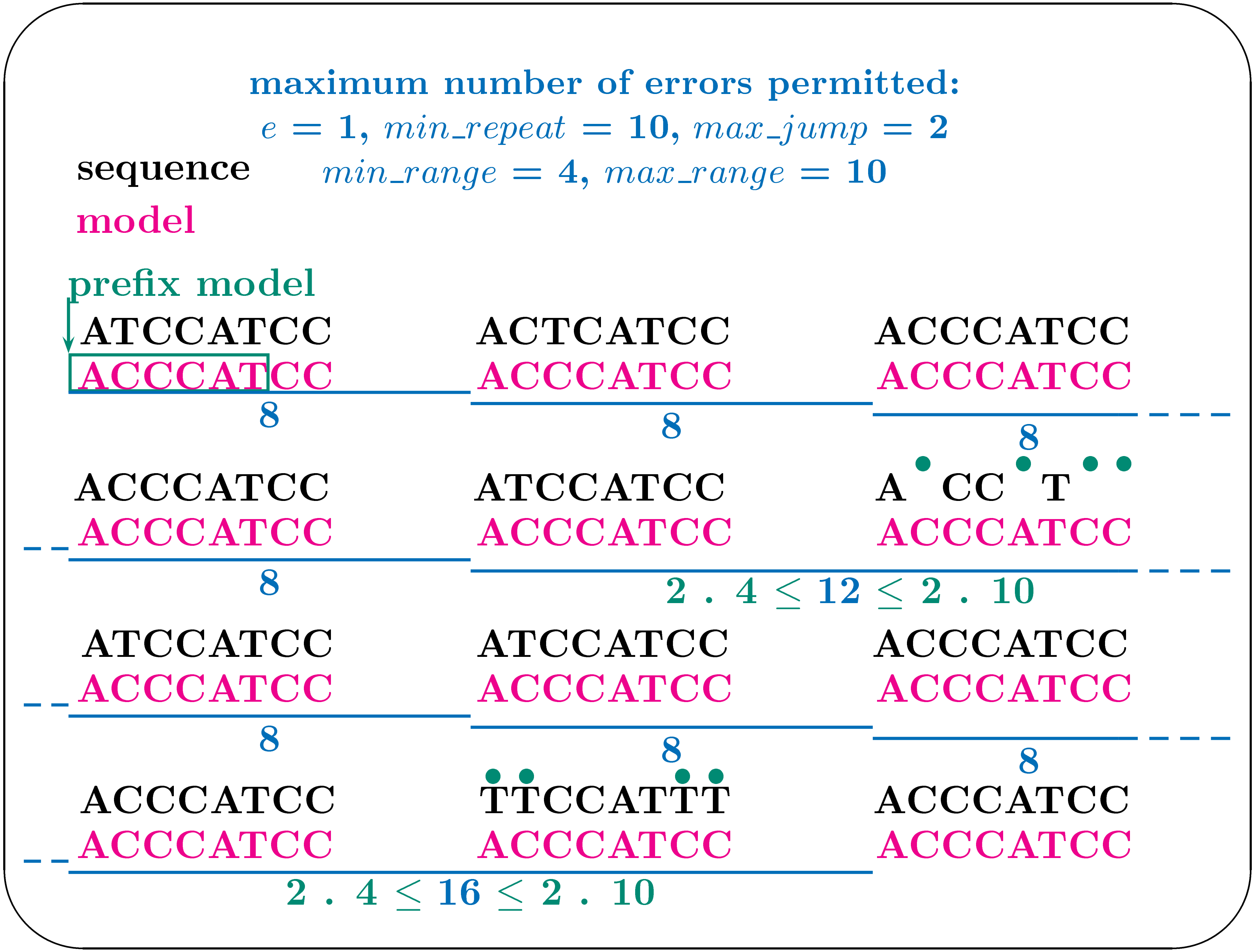
The problem formally put

train associated with a model m: an ordered collection of wagons, $u_1, u_2, \ldots, u_p$ such that:

- **Property 1** there are enough wagons
- **Property 2** the distance between the starts of two successive wagons is inside an interval $[min\_range, max\_range]$, or a multiple $max\_jump > 1$ of that interval

prop. 1 + prop. 2 = satellite prefix model:

- **this is what is checked during all the first part of the algorithm**


maximum number of errors permitted:
e = 1, min_repeat = 10, max_jump = 2
min_range = 4, max_range = 10
sequence
model
prefix model
ATCCATCC
ACCCATCC
8
ACTCATCC
ACCCATCC
8
ACCCATCC
ACCCATCC
8
ACCCATCC
ACCCATCC
8
ATCCATCC
ACCCATCC
A CC T
ACCCATCC
2 . 4 ≤ 12 ≤ 2 . 10
ATCCATCC
ACCCATCC
8
ATCCATCC
ACCCATCC
8
ACCCATCC
ACCCATCC
8
ACCCATCC
ACCCATCC
TTCCATTT
ACCCATCC
ACCCATCC
ACCCATCC
2 . 4 ≤ 16 ≤ 2 . 10

The problem formally put

train associated with a model m: an ordered collection of wagons, $u_1, u_2, \ldots, u_p$ such that:

Property 3 the distance between the end of a wagon and the beginning of the wagon coming just after it is either 0 (no bad wagons are allowed)

maximum number of errors permitted:
$e = \mathbf{1}$, $min\_repeat = \mathbf{10}$

sequence

model

ATCCATCC **ACTCATCC** **ACCCATCC**
ACCCATCC **ACCCATCC** **ACCCATCC**
0 **0** **0**

ACCCATCC **ATCCATCC** **A CC T**
ACCCATCC **ACCCATCC** **ACCCATCC**
0

ATCCATCC **ATCCATCC** **ACCCATCC**
ACCCATCC **ACCCATCC** **ACCCATCC**
0 **0** **0**

ACCCATCC **TTCCATTT** **ACCCATCC**
ACCCATCC **ACCCATCC** **ACCCATCC**

The problem formally put

train associated with a model m: an ordered collection of wagons, $u_1, u_2, \ldots, u_p$ such that:

> **Property 3 the distance between the end of a wagon and the beginning of the wagon coming just after it is either 0 (no bad wagons are allowed) or is within the intervals possible given the number of bad wagons allowed between two good ones**

maximum number of errors permitted:

$e = 1$, $min\_repeat = 10$, $max\_jump = 2$

$min\_range = 4$, $max\_range = 10$

sequence

model

jump (“bad wagons” between two “good” ones)

ATCCATCC	ACTCATCC	ACCCATCC
ACCCATCC	ACCCATCC	ACCCATCC
0	0	0
ACCCATCC	ATCCATCC	A CC T
ACCCATCC	ACCCATCC	ACCCATCC
0		$1 . 4 \leq 4 \leq 1 . 10$
ATCCATCC	ATCCATCC	ACCCATCC
ACCCATCC	ACCCATCC	ACCCATCC
0	0	0
ACCCATCC	TTCCATTT	ACCCATCC
ACCCATCC	ACCCATCC	ACCCATCC
	$1 . 4 \leq 8 \leq 1 . 10$	

The problem formally put

train associated with a model m: an ordered collection of wagons, $u_1, u_2, \ldots, u_p$ such that:

> **Property 3 the distance between the end of a wagon and the beginning of the wagon coming just after it is either 0 (no bad wagons are allowed) or is within the intervals possible given the number of bad wagons allowed between two good ones**

prop. 1 + prop. 2 + prop. 3 = satellite consensus model:

> **prop. 3 is checked at the end of the algorithm only**

maximum number of errors permitted:

$e = \mathbf{1}$, $min\_repeat = \mathbf{10}$, $max\_jump = \mathbf{2}$

sequence $min\_range = \mathbf{4}$, $max\_range = \mathbf{10}$

model

consensus model

ATCCATCC **ACTCATCC** **ACCCATCC**

$\boxed{\textbf{ACCCATCC}}$ **ACCCATCC** **ACCCATCC**

0 **0** **0**

ACCCATCC **ATCCATCC** **A CC T**

ACCCATCC **ACCCATCC** **ACCCATCC**

0 $\mathbf{1 . 4 \leq 4 \leq 1 . 10}$

ATCCATCC **ATCCATCC** **ACCCATCC**

ACCCATCC **ACCCATCC** **ACCCATCC**

0 **0** **0**

ACCCATCC **TTCCATTT** **ACCCATCC**

ACCCATCC **ACCCATCC** **ACCCATCC**

$\mathbf{1 . 4 \leq 8 \leq 1 . 10}$

What the algorithm does

Given the parameters:

min_repeat, max_jump, min_range, max_range **and** e

Finds ALL valid consensus models

Currently works for short to medium-sized repeats (up to 50 bp long if 10 to 15% error rate is tolerated)

What the algorithm does

Given the parameters:

min_repeat, max_jump, min_range, max_range **and** e

Finds ALL valid consensus models

Currently works for short to medium-sized repeats (up to 50 bp long if 10 to 15% error rate is tolerated)

"satellites" (Sagot and Myers, JCB, 5:539-554, 1998)

What we are aiming at

Biological questions (part of the work of R. Tavares)

- 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 for longer repeats
- is junk DNA really junk ?

Algorithmical issues

- there are MANY, some exceedingly beautiful ones
- curious ? → come talk to us !!

Just starting and still within "GENOME STRUCTURE" (coding parts)

Detecting and studying the combinatorics of protein domains

In order to infer functional links

(part of the work of R. Tavares)

Third area of research:

MOLECULES IN SPACE

what is covered by this (very broad) title ?

longer range interactions which may or may not be related, directly or indirectly, to a structure

the structure itself of a macromolecule

using known or predicted information concerning it

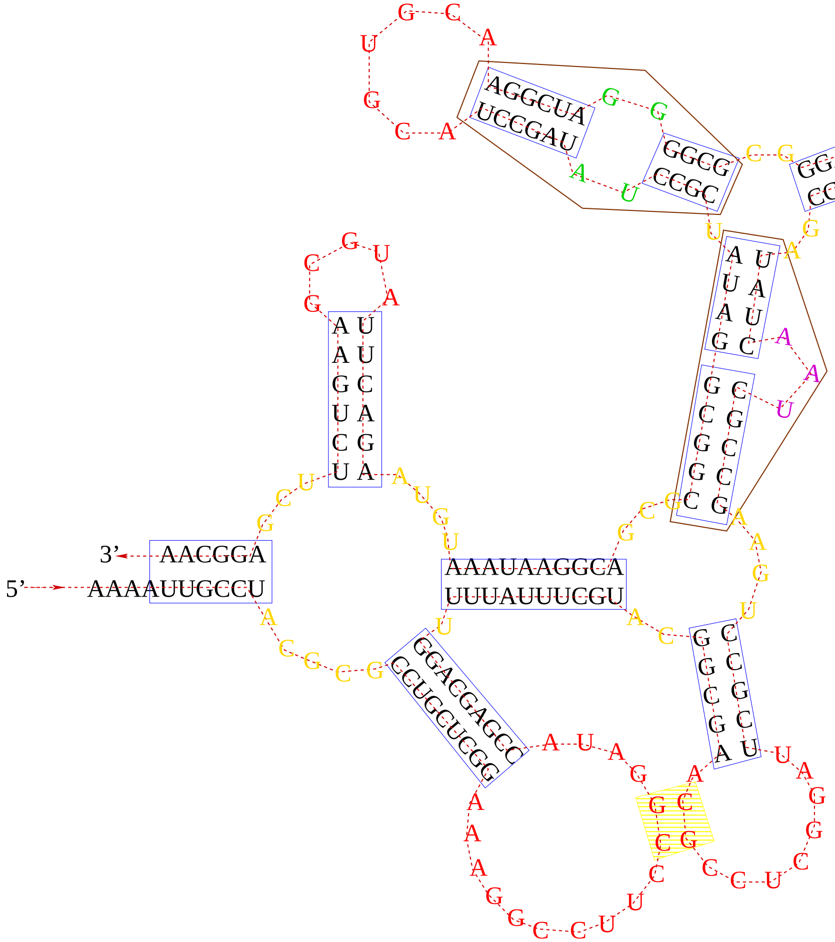
inferring it

The problems we have been adressing for now

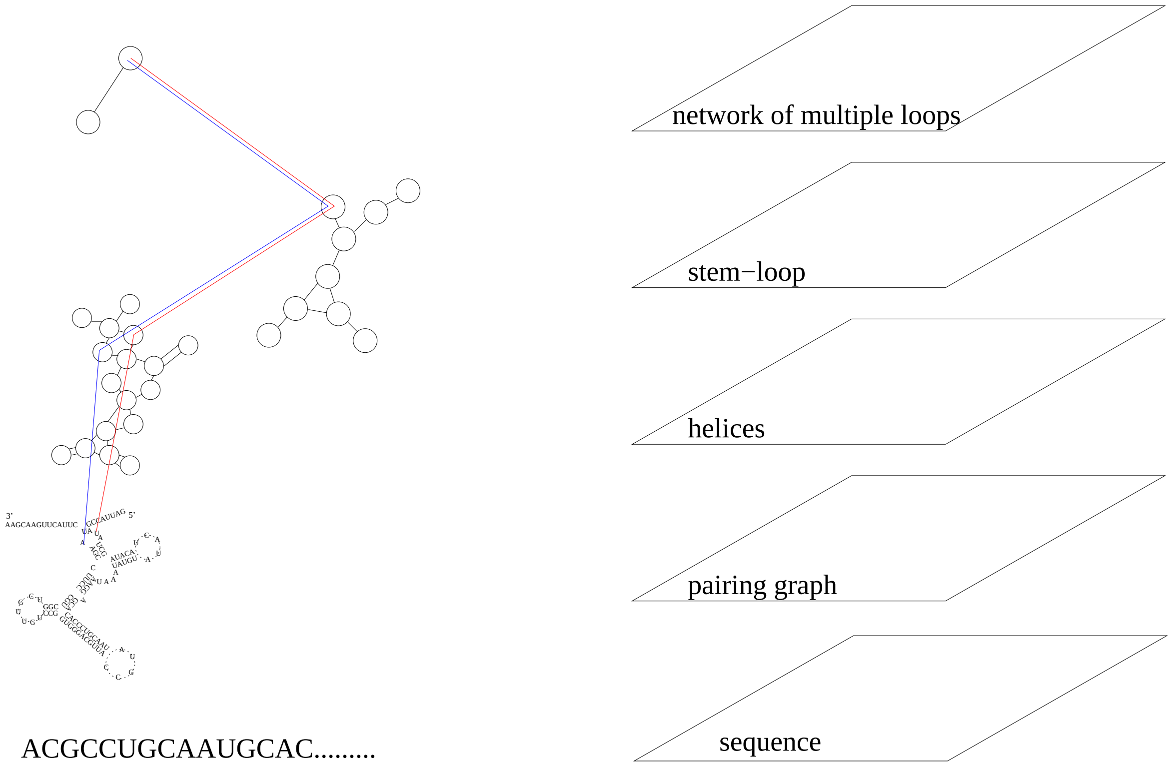
obtaining a represention for comparing RNA secondary (tertiary) structures

exploring tree and graph (local) distances

A simplified view of an RNA secondary structure



The representation we have obtained: MIGAL (“MultIple GrAph Layer”)



An implementation exists (J. Allali, mémoire de DEA)

What we are aiming at

Biological questions

given two or more structures, calculate a local distance between them and find a model which represents what they have in common

given a (possibly vague) description of a structure, find all sequences which may fold according to it

given a set of sequences, find the element(s) of structure they may have in common

Algorithmical issues

there is a HUGE number, some exceedingly beautiful ones

in fact, two complete areas of research to develop

curious ? → come talk to us !!

But just to give a, we hope seducing, hint

Introduce notions of “allowed” differences that may represent two very distinct concepts

biological variability

errors (in general man-made) in the data

into tree and graph theory and algorithms

Another problem within "MOLECULES IN SPACE"

introducing structural considerations into the protein motif inference problem

Another problem within "MOLECULES IN SPACE"

introducing structural considerations into the protein motif inference problem

beyond the mere use of a physico-chemical alphabet

for which an algorithm is, however, available

CoSamp (for Combinatorial Sampling – Créno *et al.*, not published yet)

What we are aiming at

Biological issues

explore the physico-chemical alphabet in a combinatorial manner

be able to take 2D or 3D information (known or predicted) into account

Algorithmical issues

there are MANY, some exceedingly beautiful ones

curious ? $\rightarrow$ come talk to us !!

Fourth area of research:

GENOME DYNAMICS

Listen ! a frog

Jumping in the stillness

Of an ancient pond !

Basho (Spring days)

Our concern will be, not with frogs jumping about and into ponds, but with

genomic fragments "jumping" (sometimes making copies of themselves, or disappearing) intra or inter-strains (species)

how to identify with a fair degree of confidence and accuracy the rearrangements and genetic transfers that took place ?

how to use rearrangement events to calculate a distance or infer a phylogeny between organisms ?

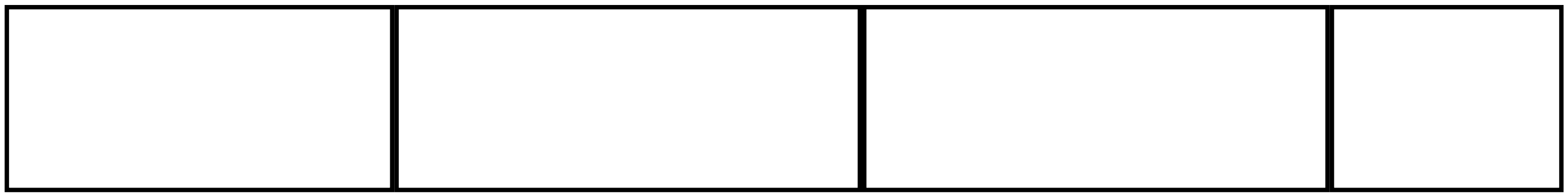
Genetic transfer identification and location

Data

a set of phylogenetically "correctly" aligned sequences from closely related organisms

The (ideal ?) result

alignment

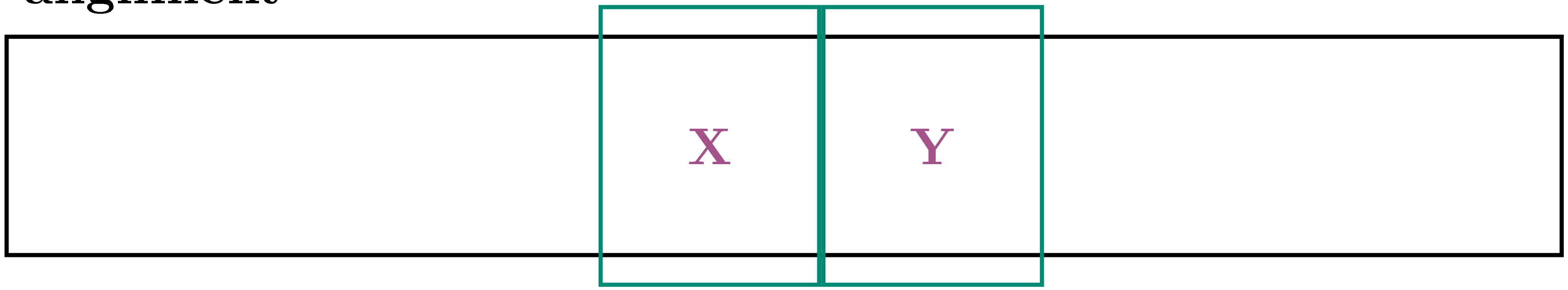


tree 1 tree 2 tree 3 tree 4

"optimal" partition: sum of lengths of trees + penalty for genetic transfers is **MINIMUM**

Main (crucial) questions

alignment



How to calculate the distance between two trees in this specific case ?

- **length (number of mutations) difference**
- **genetic transfer (inter-species or inter-strains rearrangement) distance**

How to determine whether the obtained distance is statistically significant ?

Genetic transfer identification and location: a first pragmatic approach

Slide a double (small) window along the alignment

species \ sites	1	2	3	4	5	6	7	8	9	10	11	12	...	n
1	A	T	G	G	A	G	C	T	A	C	T	C	...	T
2	A	T	A	G	A	G	A	T	A	T	A	C	...	T
3	A	C	A	T	C	C	A	T	A	T	A	C	...	T
4	C	C	T	T	C	C	A	G	C	C	T	G	...	G
5	C	C	T	T	A	C	C	G	C	C	T	G	...	G

matrix X (sites 2–6) matrix Y (sites 7–11)

Genetic transfer identification and location: a first pragmatic approach

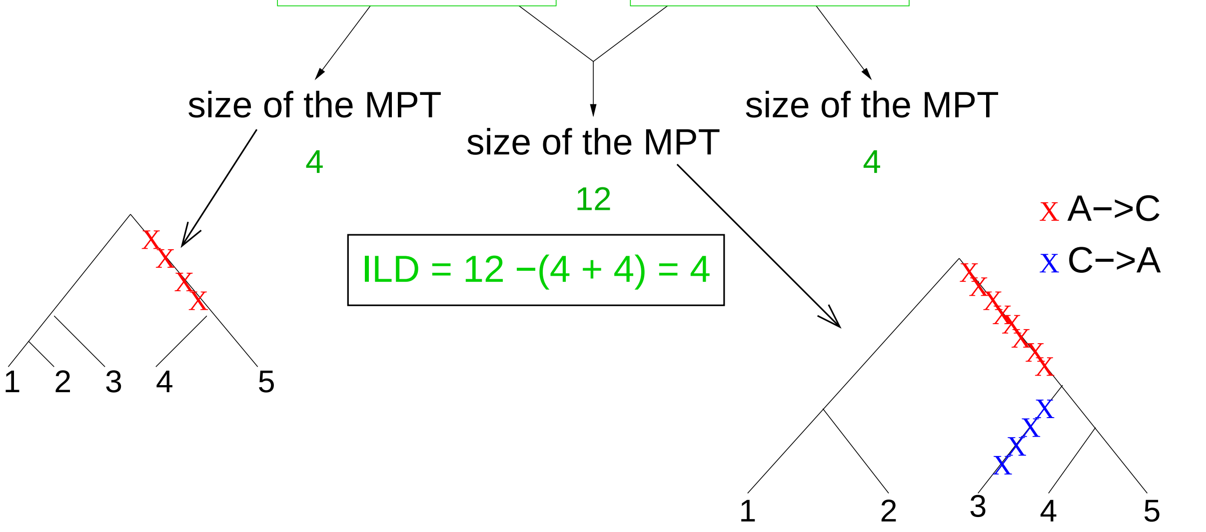
Use an Incongruence Length Difference (ILD) test

matrix 1

species 1	A	A	A	A
species 2	A	A	A	A
species 3	A	A	A	A
species 4	C	C	C	C
species 5	C	C	C	C

matrix 2

species 1	A	A	A	A
species 2	A	A	A	A
species 3	C	C	C	C
species 4	C	C	C	C
species 5	C	C	C	C



DRUID (M. Zelwer *et al.* – accessible at Pasteur web site)

Genetic transfer distance between two phylogenetic trees

One possible distance

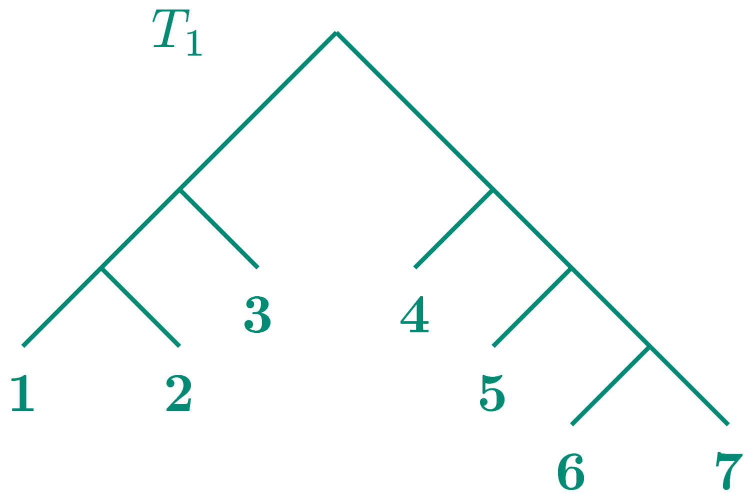
Maximum Agreement Forest between two trees

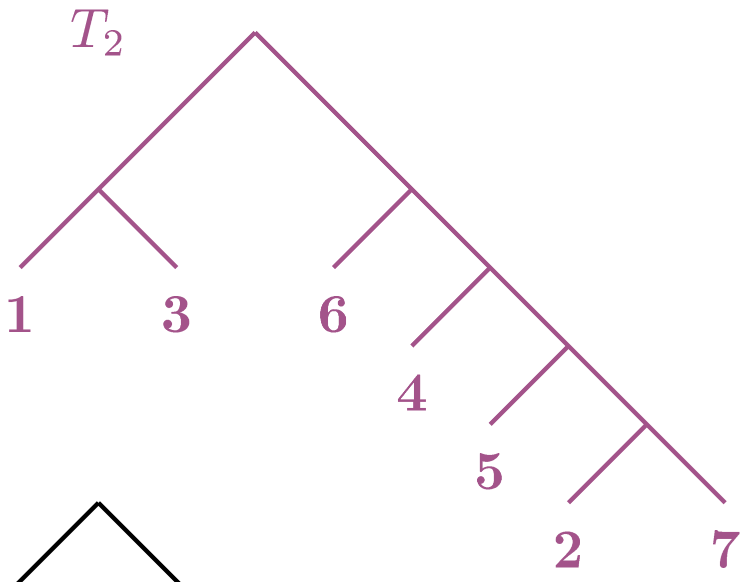
(E. M. Rodrigues *et al*, Approx'2001 – theoretical work)

(currently undergoing implementation)

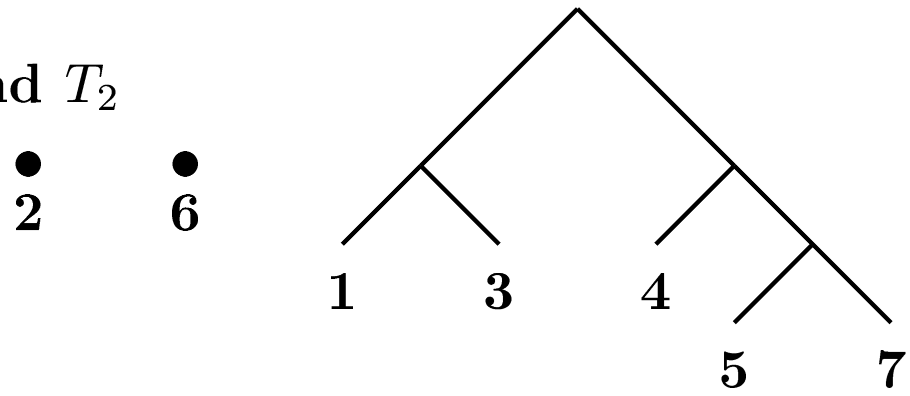
Maximum Agreement Forest (MAF) between two trees

rooted, leaf-labeled, binary, "unordered" trees



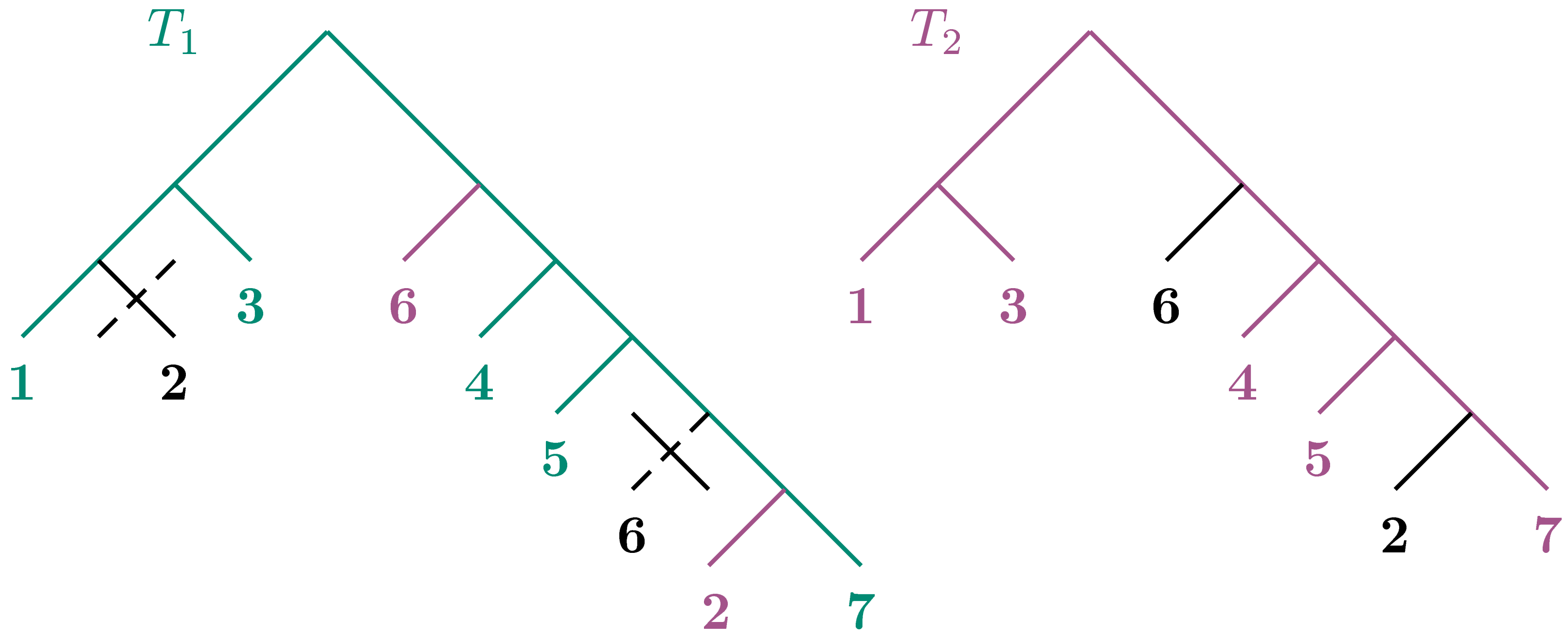


MAF of T_1 and T_2



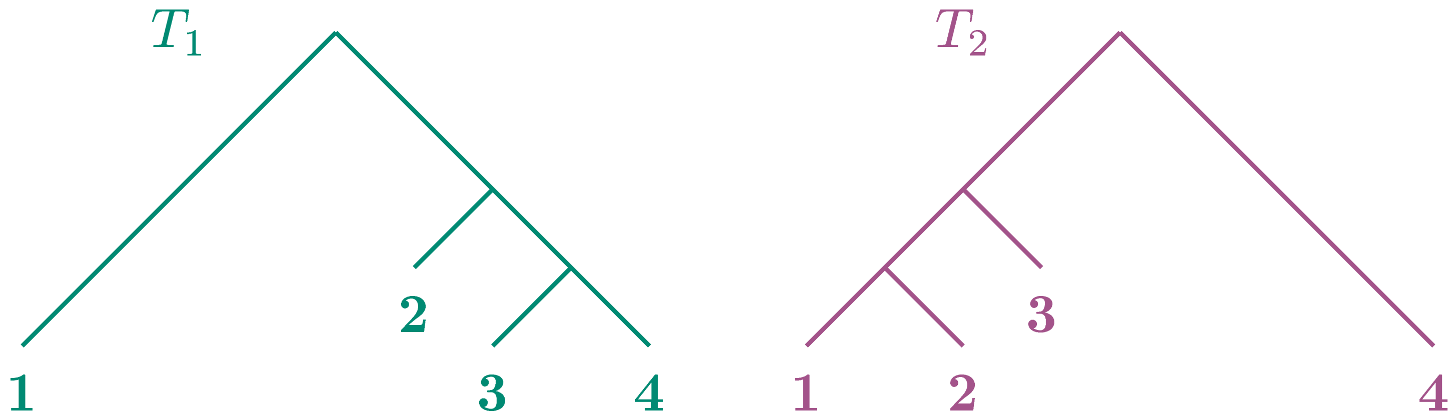
Size of Maximum Agreement Forest minus 1:

lower bound for the "Subtree Prune and Regraft" (SBR) distance between trees

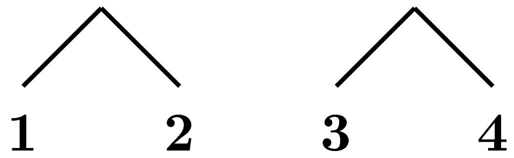


Size of Maximum Agreement Forest minus 1:

but NOT identical to the "Subtree Prune and Regraft" (SBR) distance between trees

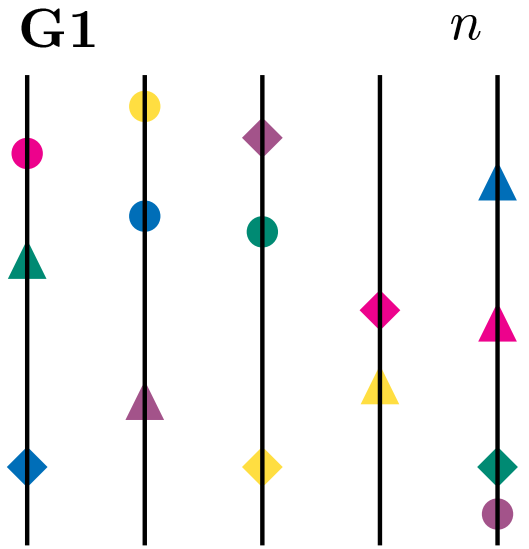


MAF of T_1 and T_2



Rearrangement distance between two genomes

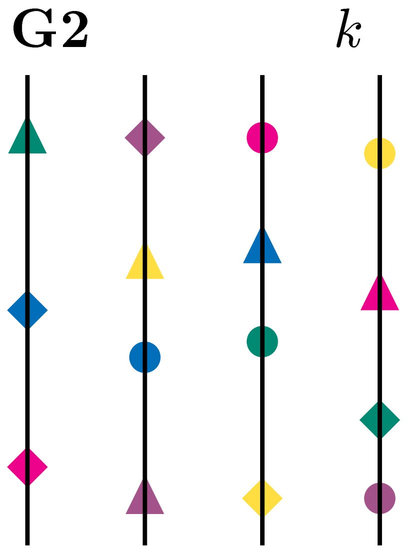
(a very special case of syntenic distance)



genome

=

set of
bags of genes



distance = minimum number of 3 types of operations
to transform one genome into another

fusion

fission

translocation

(N. Pisanti *et al.*, submitted to *Algorithmica*)

What we are aiming at

Biological issues

- **improve and extend the detection of rearrangements and their location**
- **explore the various tree distances and their relations**
- **address the problem of a phylogenetic reconstruction in the presence of rearrangements (F. Chataigner, M. Zelwer)**
- **introduce more realistic distances between genomes**
- **model and identify groups of synteny in eukaryotes**

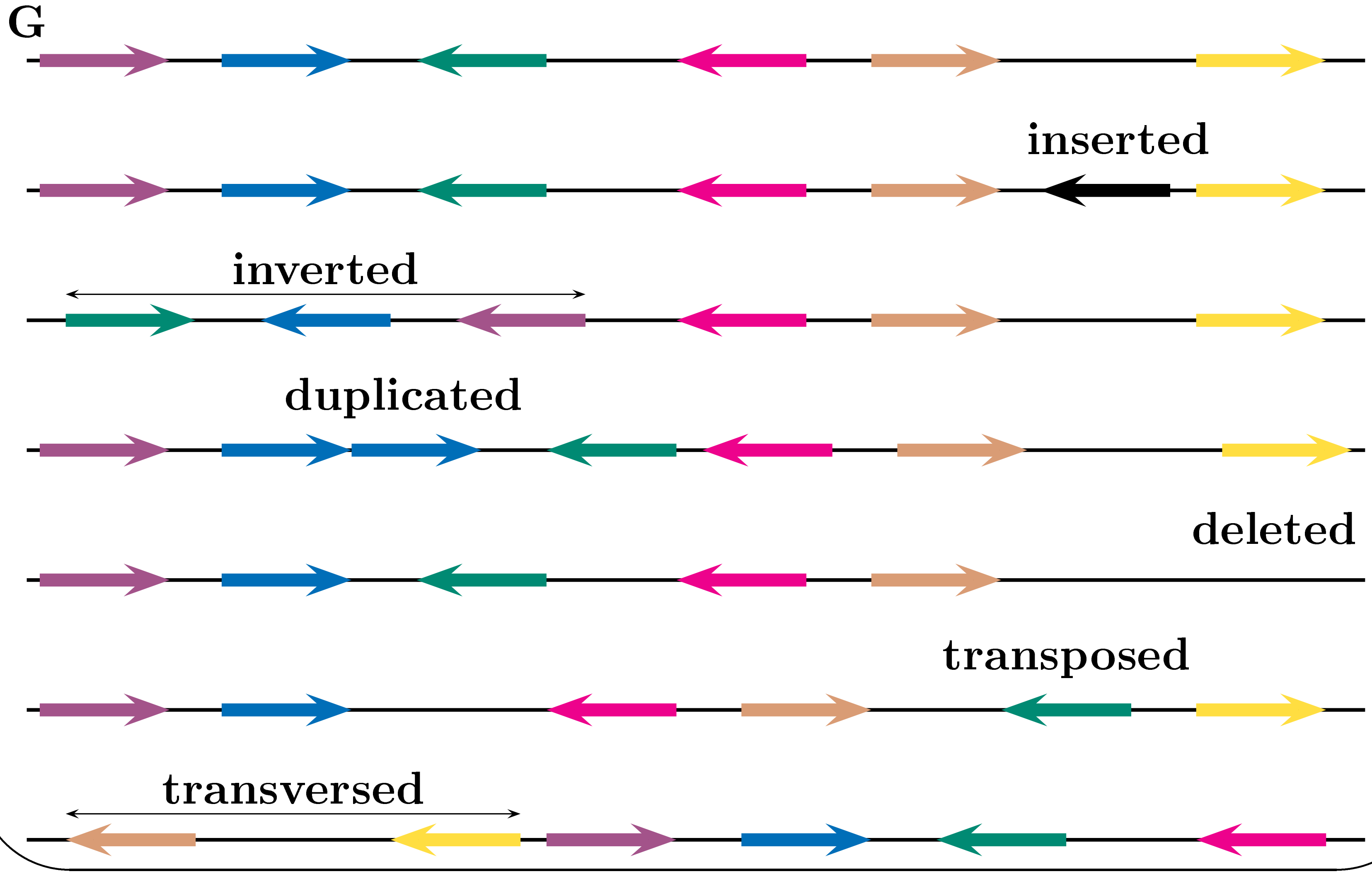
Algorithmical issues

- **there is a HUGE number, some exceedingly beautiful ones**

curious ? $\rightarrow$ come talk to us !!

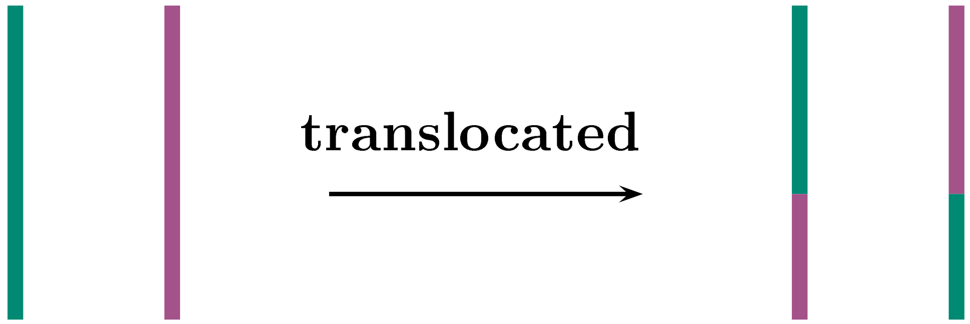
More realistic rearrangement distances between genomes

Segments (*e.g.* genes) are frequently



And, in the case of multichromosomal genomes,

Segments (*e.g.* genes, often various) are frequently



and chromosomes are frequently

Fifth area of research:

CELL MODELLING

A journey of a thousand miles

starts with a single step

and continues with a second

slight modification of a buddhist saying

The main idea so far

A multi-layered approach

(inspired on MIGAL for RNA structures)

S. Provencher

First layer **Environments and interaction between environments**

extra-cellular environment and various cellular regions of interest

Second layer **States of a cell**

catalog of genes/proteins with their expression levels and spatial positions

Third layer **The cell**

receptors and targets

signalling pathways and interactions

Fourth layer **States of a signalling pathway**

information about which proteins are activated/inhibited and what are the environment conditions inside the cell

Fifth layer **Genetic regulation networks and interactions**

Sixth layer **State of a genetic regulation network**

Seventh layer **Genome structure**

What we are aiming at

Biological issues

obtaining a conceptual framework for

- **structuring information**
- **developing simulations that will guide experimental research**
- **inferring properties**

Algorithmical issues

we suspect there are many, possibly related to graphs and exploration thereof

curious ? → come talk to us !!

Recalling the main participants

Lyon	Paris (Nov. 2001)
Nadia Pisanti	Julien Allali
Sébastien Provencher	Philippe Blayo
Marie-France Sagot	Frédéric Chataigner
Marina Zelwer	Sophie Créno
Brazil	Laurent Marsan
Estela M. Rodrigues	Raquel Tavares
Collaboration (Paris)	
Anne Vanet	Anne Cariou
Katja Schuerer	

Mentioning the main collaborators past and present

S. Aubourg	Y. D'Aubenton	J. Berstel
S. Cole	M. Crochemore	P. Darlu
J.-J. Daudin	H. Geiselmann	J. Houmard
C. Iliopoulos	H. de Jong	A. Marcel
C. Mathé	A. Morgat	L. Mouchard
G. Myers	J. Pothier	M. Raffinot
M. Régnier	S. Robin	P. Rouzé
S. Schbath	T. Schiex	H. Soldano
C. Thermes	A. Viari	Y. Wakabayashi

N. Tandeau de Marsac

and soon, we hope, some of you !

trois petits tours et puis s'en vont – c'est la fin !