A few combinatorial problems and algorithms for molecular biology

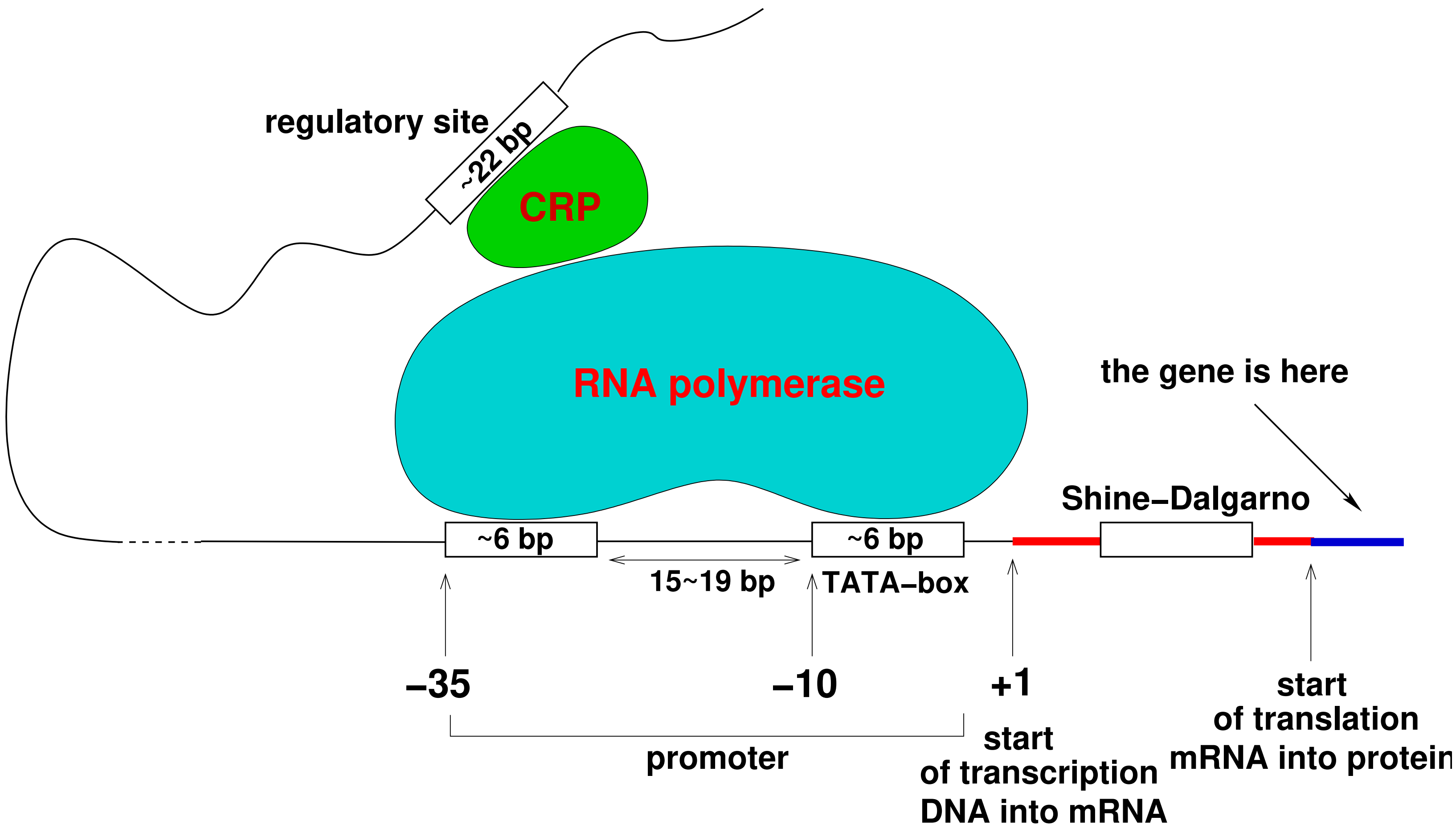
Marie-France Sagot

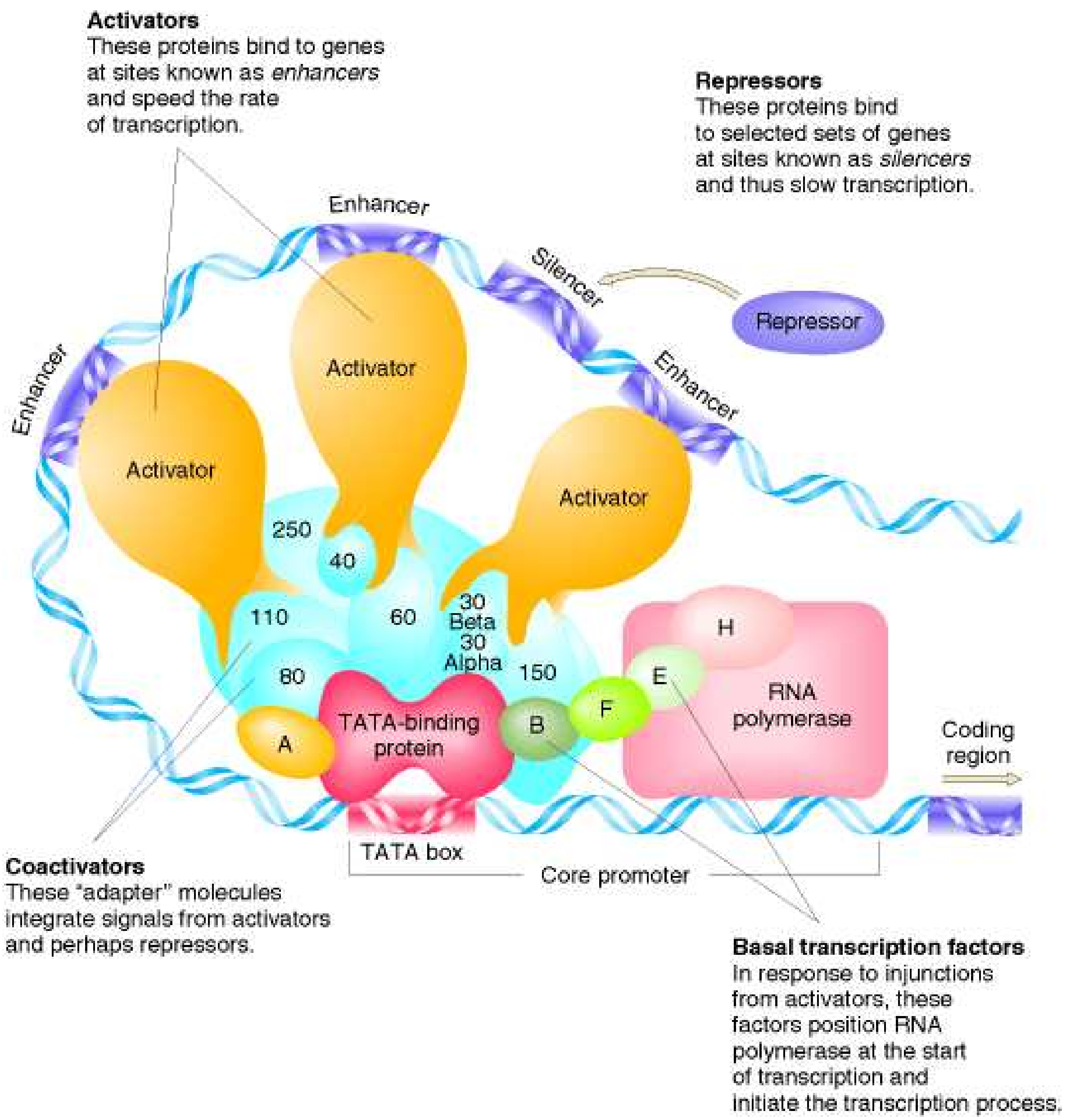
Inria Rhône-Alpes – Projet HELIX
Laboratoire de Biométrie et Biologie Évolutive
Université Claude Bernard, Lyon I

First problem:

MOTIF INFERENCE

Biological motivation (1. Prokaryotes)





Biological motivation (2. Eukaryotes)

Biological motivation (3. Repeats: dispersed or in tandem as here)

gaaagctcttctgctccagtaccaactccatccagctcgacgact

gaaagctcttctgctccagt aa ccagctcgaccact

gaaagctcttctgctccagtaccaactccatccagctcgacgact

gaaagctcttctgctccagctccaaccccatcaagctccaccact

gaaagctcttctgctccagt aa ccagctcgaccact

gaaagctcttctgctccagtaccaactccatccagctcgacgact

gaaagctcttctactccagt aa ccagctccaccact

gaaagctcttctgctccagtgccaactccatcaagctcgacgact

gaaagctcttctgctccagtaccaactccatcaagctcgacgact

gaaagctcttctgctccagctccaaccccatcaagctcgacgact

gaaagctcttctgctccagt aa ccagctcgaccact

gaaagctcttctgctccagtaccaactccatccagctcgacgact

gaaagctcttctgctccagtaccaacccccatccagctcgacgact

gaaagctcttctgctccagtaccaactccatccagctcgacgact

Motif search $\times$ Motif inference

Search

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

Inference

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

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

Set of properties

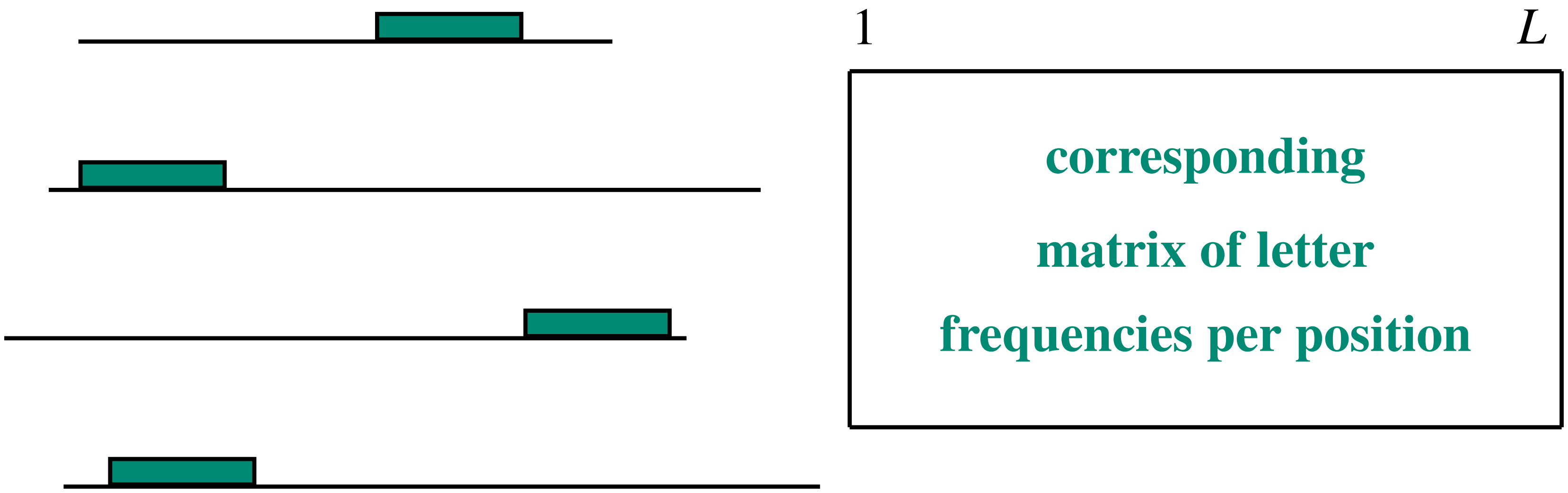
main property: motifs of interest = "conserved" elements

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

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*

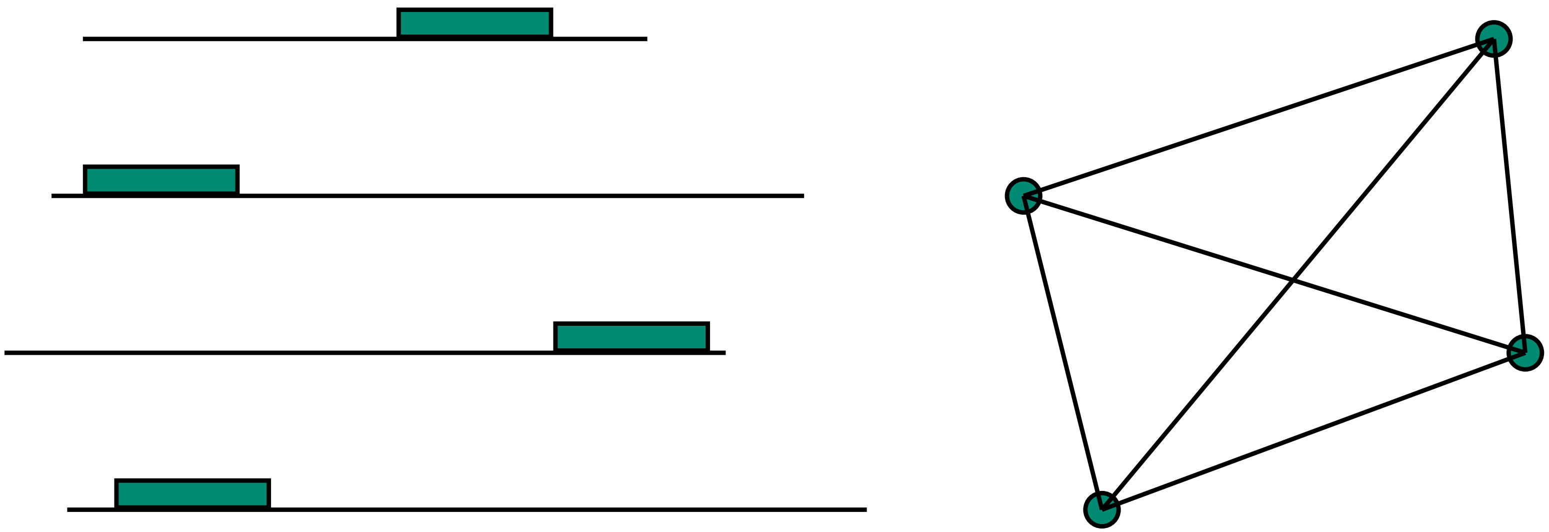
OUTPUT:

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

“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 "quorum" of N, 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 $2e$**

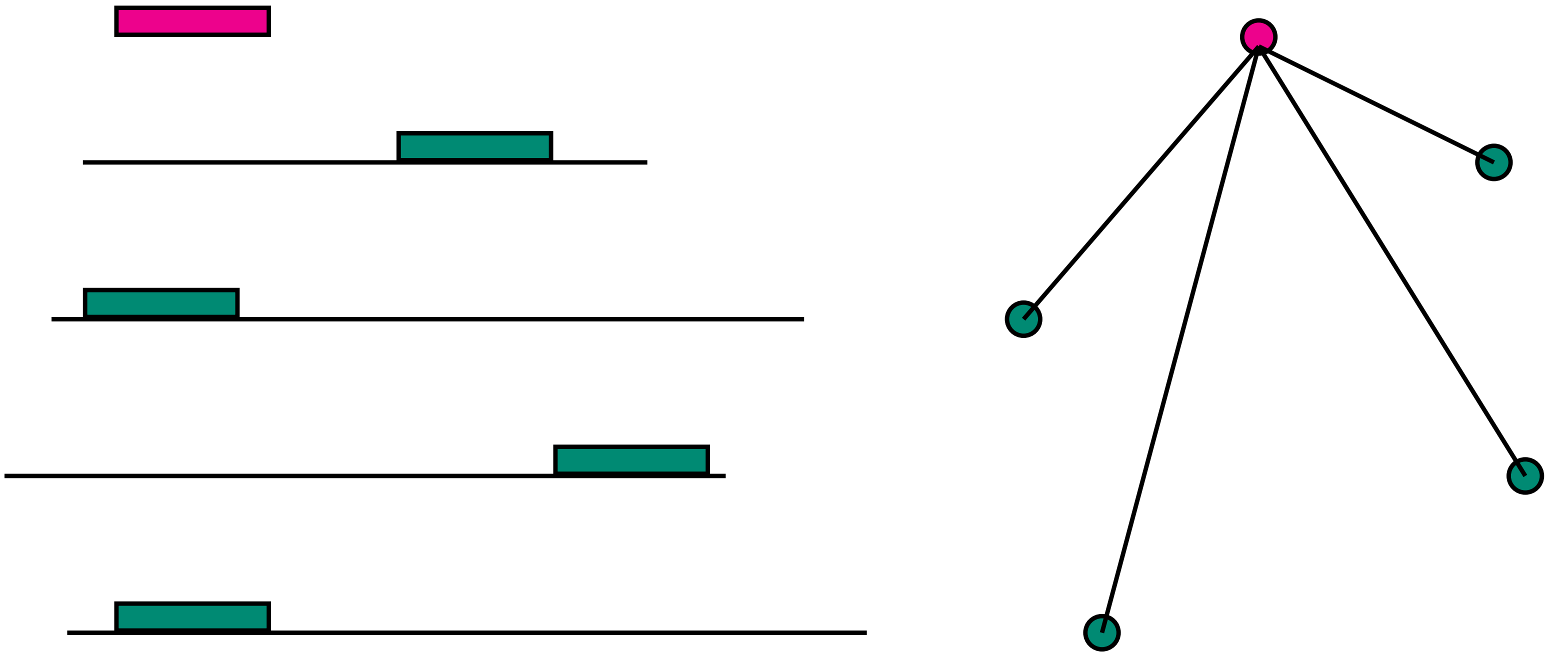
OUTPUT:

all N-cliques in G

“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

$\Sigma_{DNA/RNA}$ = {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)

Formal definition of the: “Models Problem”

INPUT:

data: a set of N sequences

parameters: a length k, a “quorum” of N, a maximum 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)

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**
"Models":	**Exponential in ("local) max. allowed differences: $O(N^{(2)}nk^{e})$**

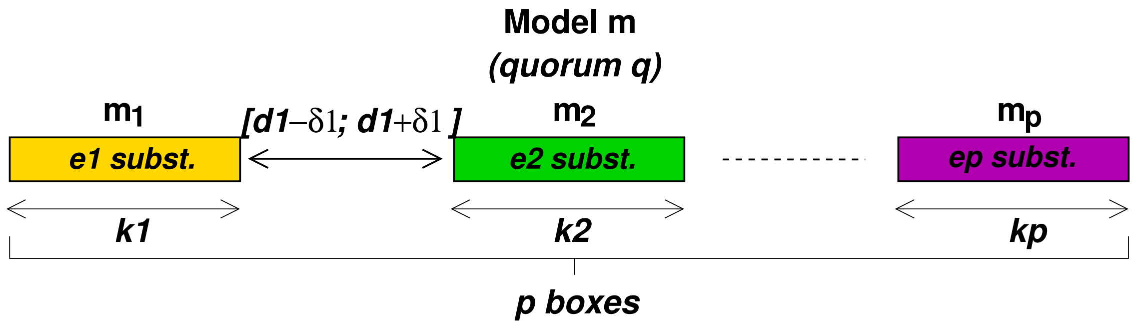
More recent developments

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

Towards modelling the combinatorics of regulation

“Structured models”

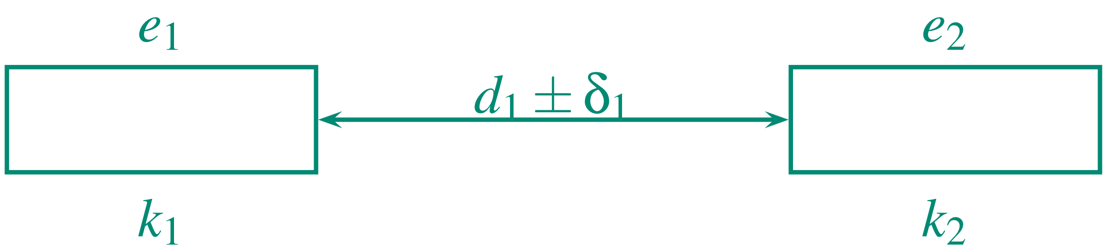
an ordered collection of p boxes, p maximum numbers of differences (e), p intervals of lengths (k), $p-1$ intervals of distances (between successive boxes in the collection) ($d \pm \delta$) (this is called *spacers*)



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)

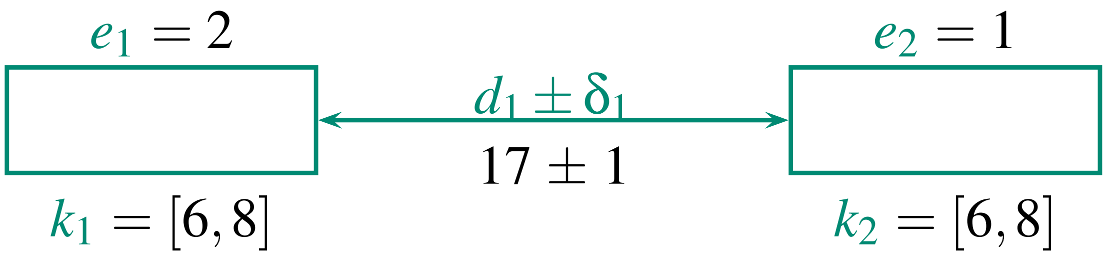
An example with two boxes



sequences

quorum

An example with two boxes



quorum $= 3$

An example with two boxes

$e_1 = 2$ $\qquad\qquad\qquad\qquad$ $e_2 = 1$

$\boxed{\text{TTGACA}} \xleftrightarrow[17 \pm 1]{d_1 \pm \delta_1} \boxed{\text{TATAAT}}$

$k_1 = [6,8]$ $\qquad\qquad\qquad\qquad$ $k_2 = [6,8]$

occurrences $\qquad\qquad\qquad\qquad$ quorum $= 3$

TTGACT $\xleftrightarrow{18}$ TAAAAT

TTGACA $\xleftrightarrow{16}$ TATAAA

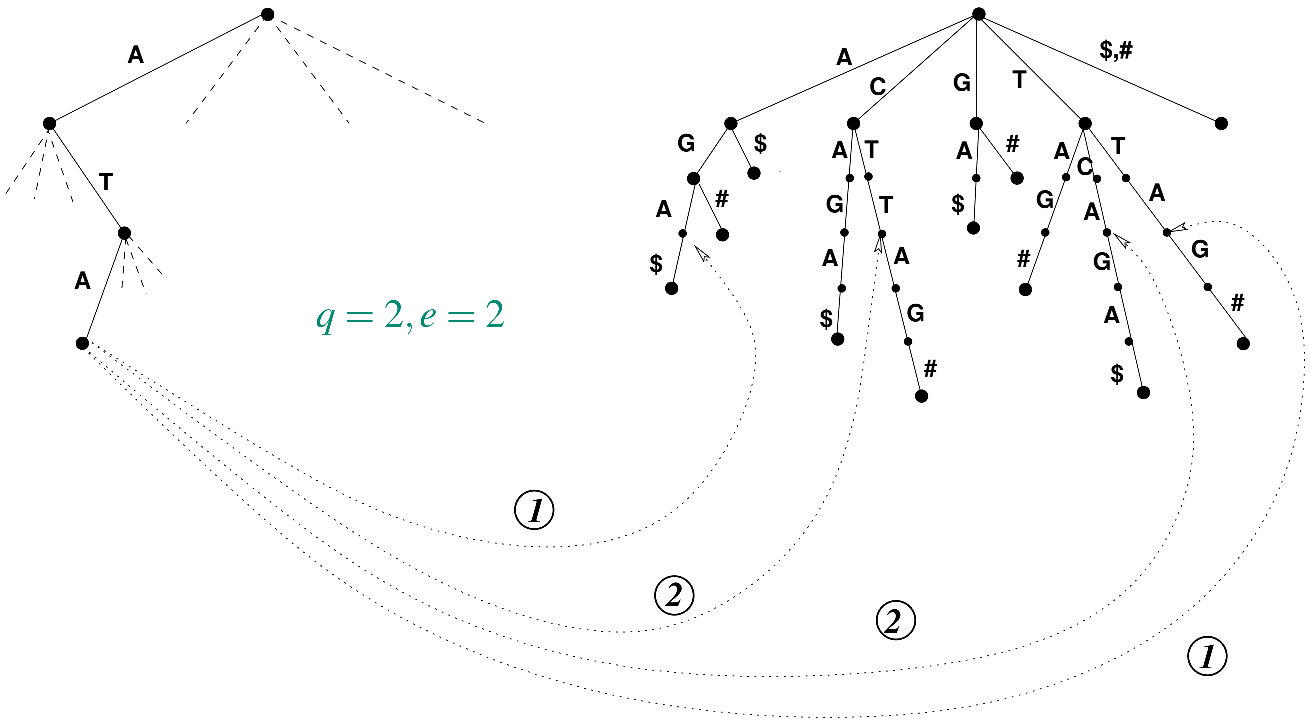
TTGCCA too far TATTAT

TTGTCT $\xleftrightarrow{17}$ TATAAT

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

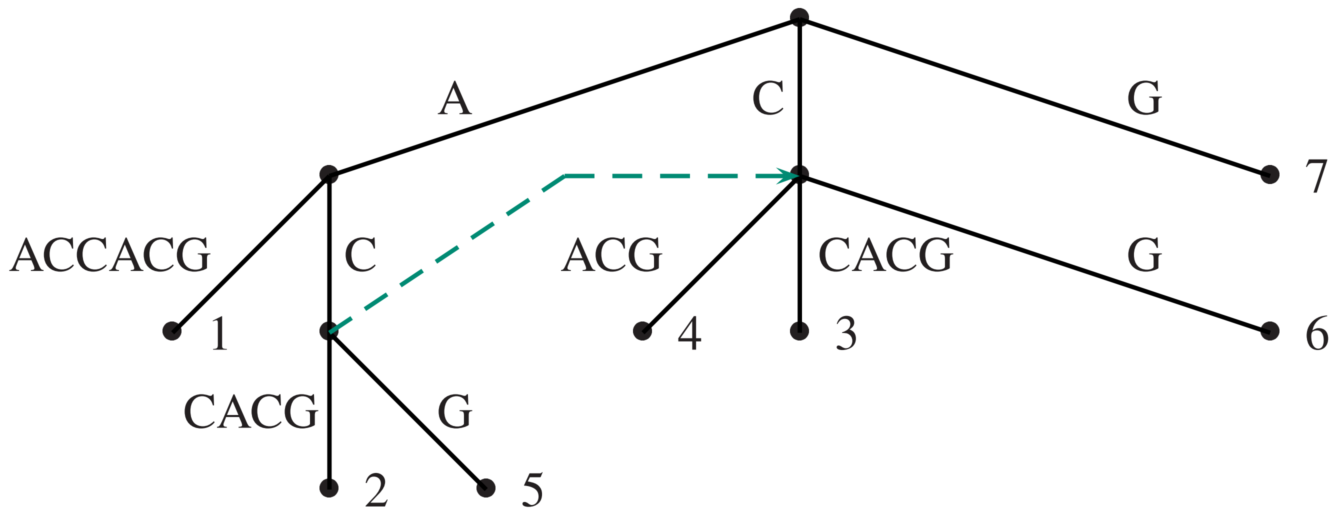
Trie of models (search tree)

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



Suffix tree

Suffix tree for AACCACG

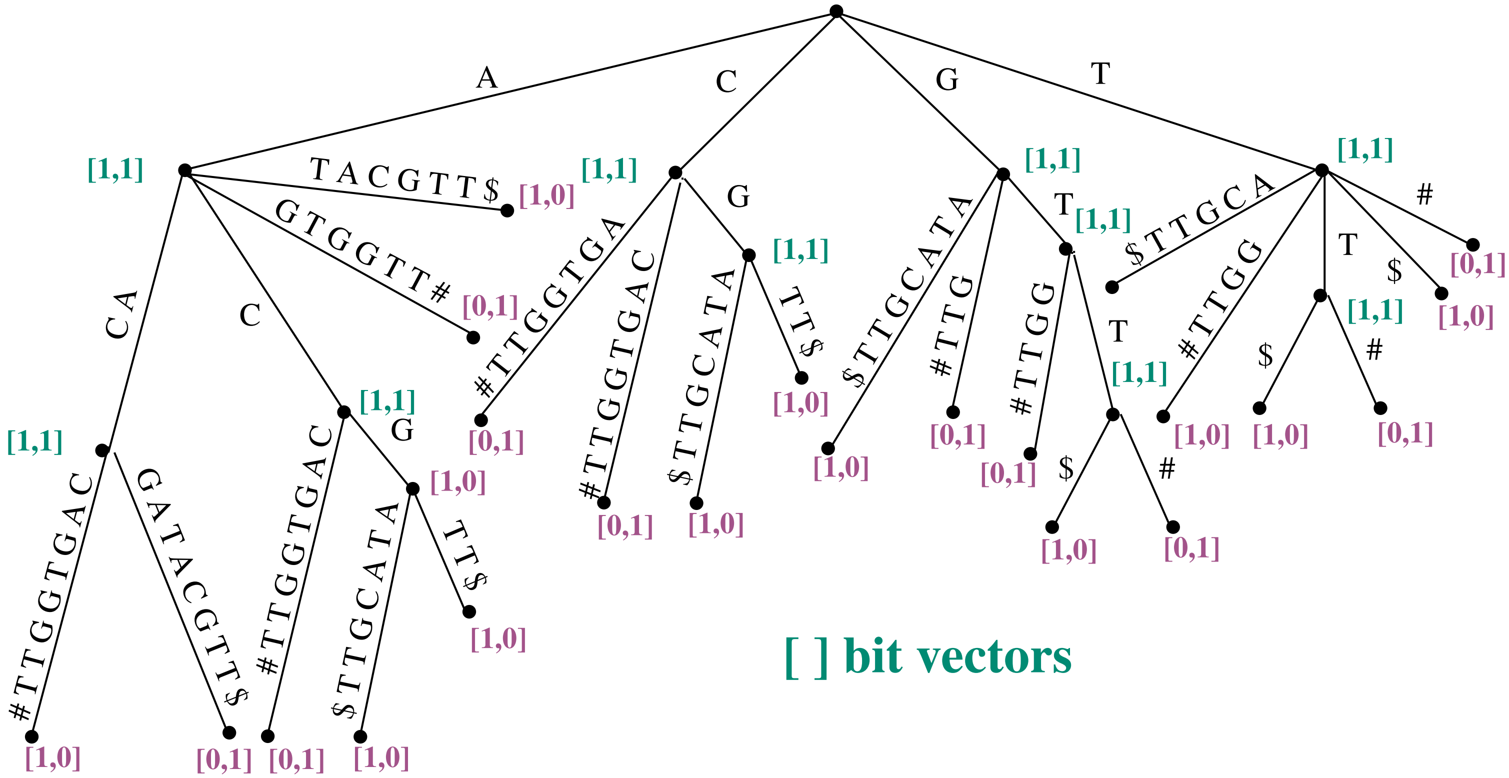


key feature: concatenation of labels of arcs from root to leaf *i* spells suffix of input string starting at position *i*

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

Generalized suffix tree

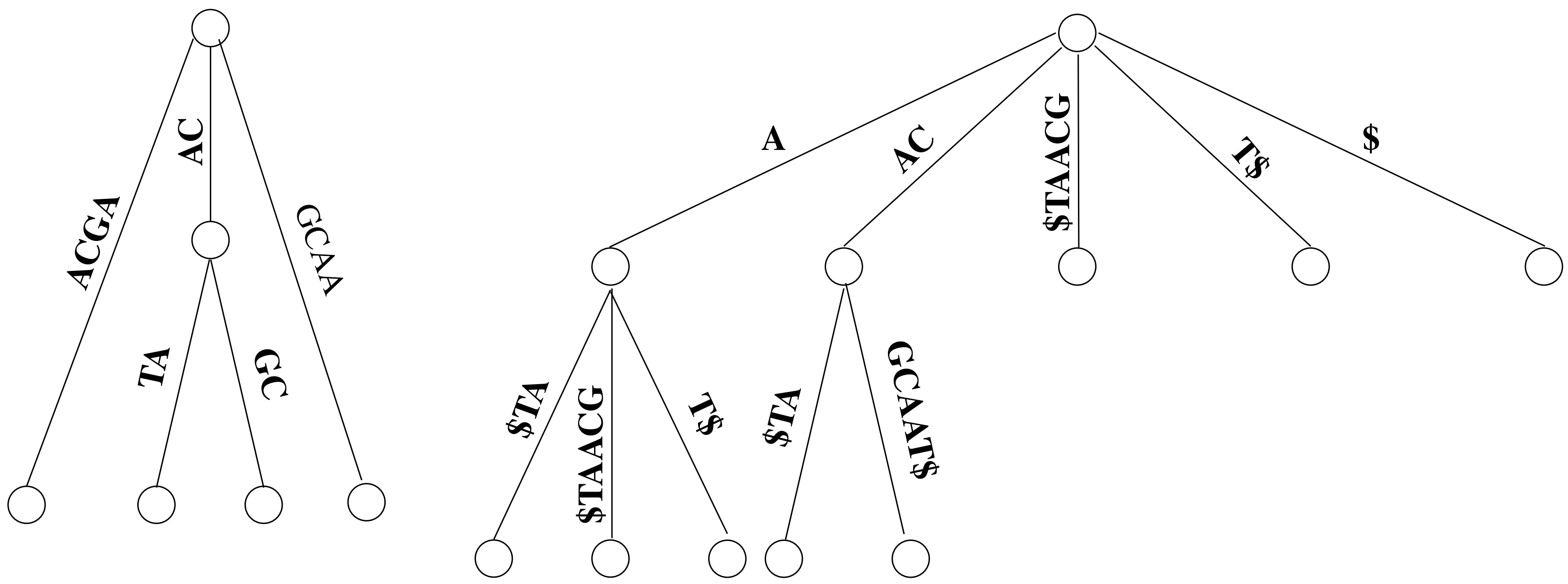
Strings AACGATACCTT$ and AACCAGTGGTT#



[ ] bit vectors

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

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



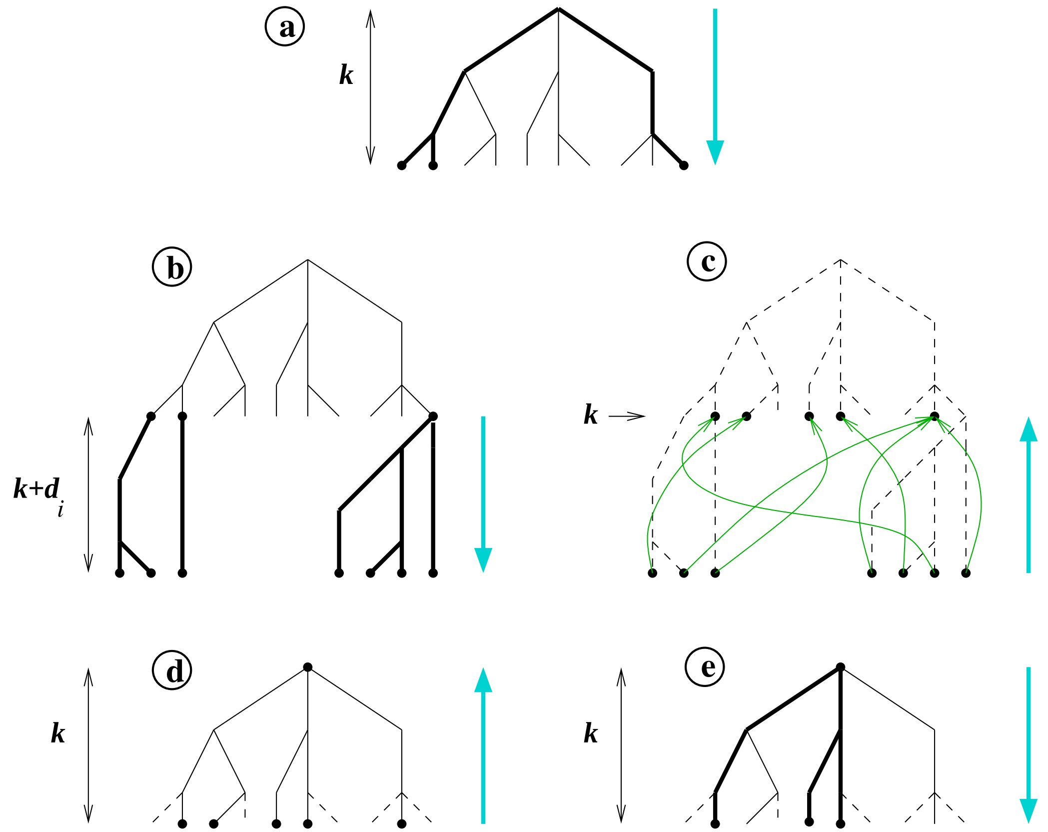
factor tree

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



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

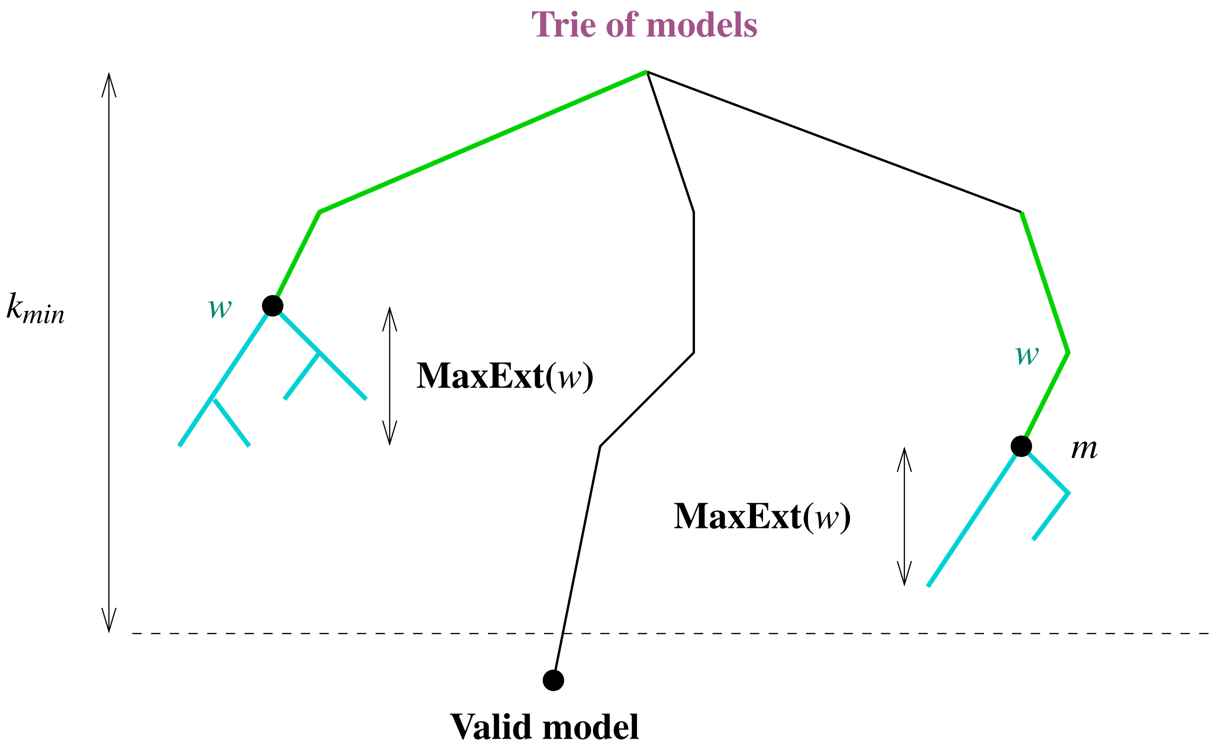
descent down the trees may end at first such prefix

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

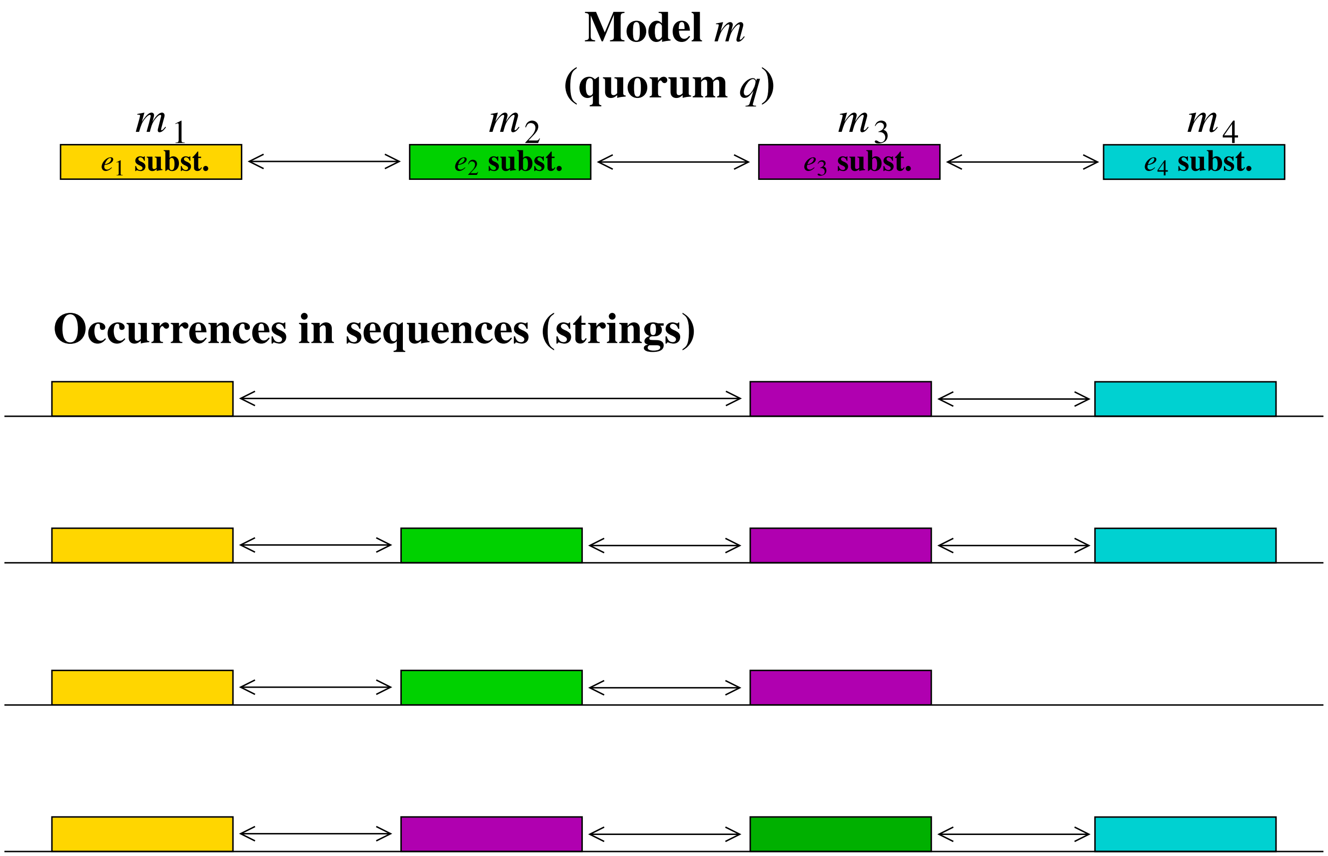
descent down the trees may end as soon as such situation is met

(Marsan, Pisanti and Sagot, paper in preparation)

Maximum “extendability”



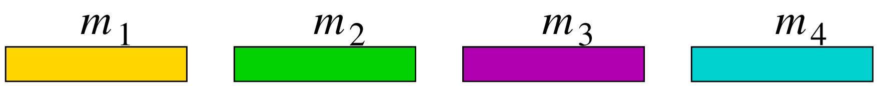
Variants (1. Skipping boxes or changing order)



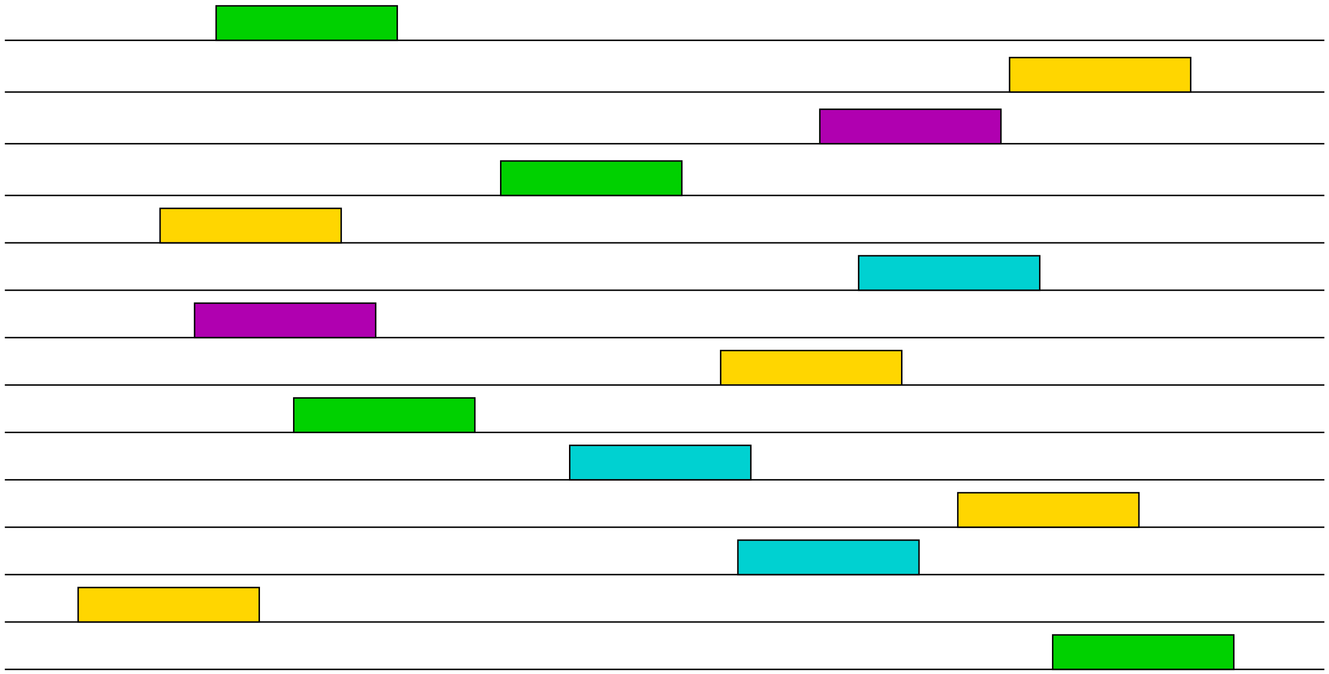
Skipping boxes: (Marsan *et al.*, in preparation)

Variants (2. Covering motifs)

Minumum set of models (one box) satisfying the properties



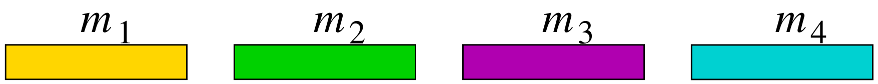
that "covers exactly" a set of sequences (strings)



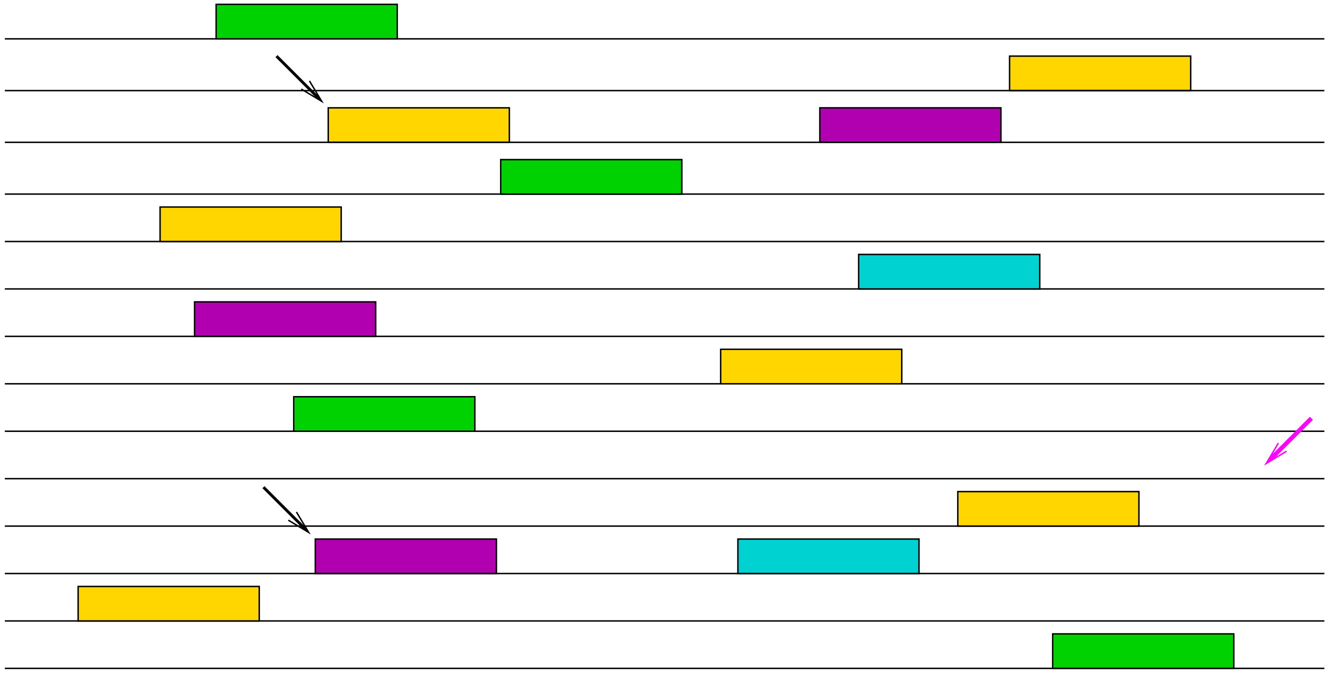
(Vialette, Sagot, work in progress)

Variants (2. Covering motifs)

Minumum set of models (one box) satisfying the properties



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



(Vialette, Sagot, work in progress)

Variant (3. Basis of motifs)

Formal definition of the initial problem

INPUT:

data: one sequence s over alphabet Σ

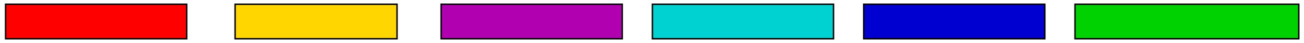
parameters: a quorum q, no differences allowed, an alphabet for the motifs $\Sigma' = \Sigma \cup \{\bullet\}$ where $\bullet$ is the don't care (wild card) symbol

OUTPUT:

all models, that is, all motifs over Σ' that have at least q occurrences in the sequence s

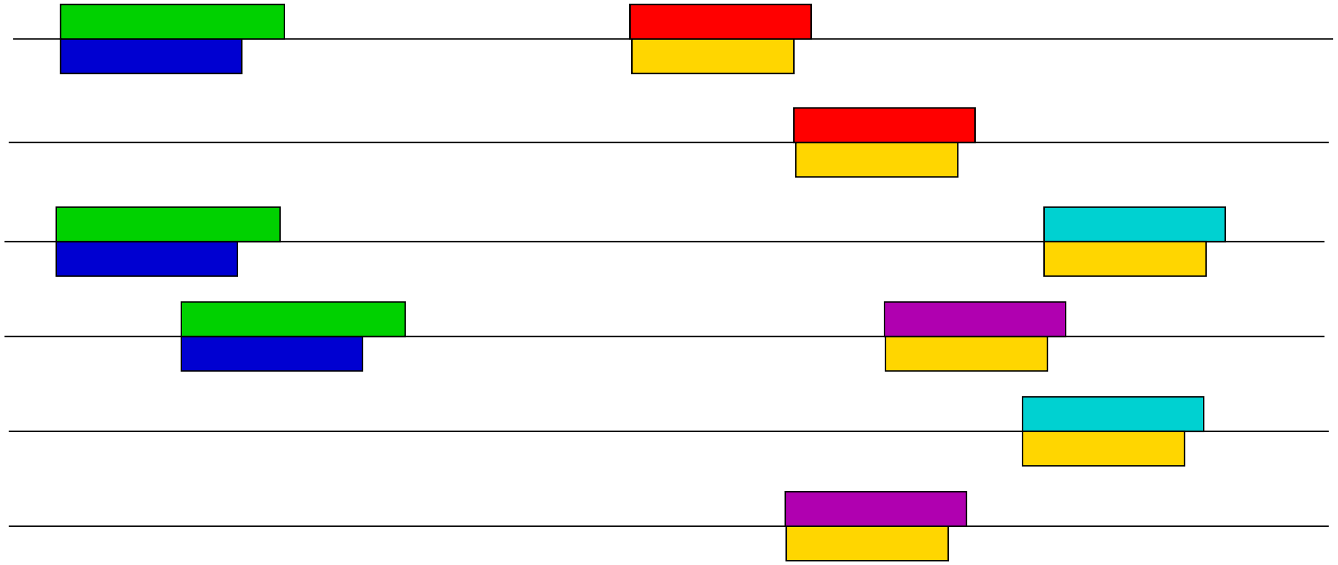
Basis of “irredundant” motifs

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



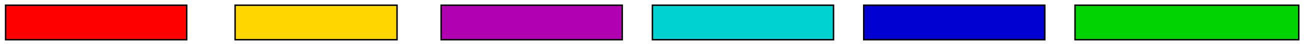
Occurrences in sequences (strings)

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

$s = \mathbf{AAATAAA}$

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: $\mathbf{A}^p\mathbf{TA}^p$

What was believed initially

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

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_1AAAAAATAAAAAA

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	-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•AAAA	1,6,8	A• •AAA	1,5,6,8
AA•AAA	1,5,8	A•A•AA	1,4,6,8
AAA•AA	-7,-4,0	A•AA•A	1,3,6,8
AAAA•A	-8,-3,-1	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•AAAA	1,6,8	A• •AAA	1,5,6,8
AA•AAA	-16,-14,-9	A•A•AA	1,4,6,8
AAA•AA	-7,-4,0	A•AA•A	1,3,6,8
AAAA•A	-8,-3,-1	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.: $a_4a_2ua_2a_4vwa_4a_2a_3a_1xya_1a_3za_3a_1$AAAAAATAAAAAA

A•AAAA	-17,-15,-10	A• •AAA	1,5,6,8
AA•AAA	-16,-14,-9	A•A•AA	1,4,6,8
AAA•AA	-7,-4,0	A•AA•A	1,3,6,8
AAAA•A	-8,-3,-1	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 $q = 2$**

  Ex.: $a_4a_2ua_2a_4vwa_4a_2a_3a_1xya_1a_3za_3a_1$AAAAAATAAAAAA

A•AAAA	**-17,-15,-10**	**A• •AAA**	**1,5,6,8**
AA•AAA	**-16,-14,-9**	**A•A•AA**	**1,4,6,8**
AAA•AA	**-7,-4,0**	**A•AA•A**	**1,3,6,8**
AAAA•A	**-8,-3,-1**	**AA••AA**	**1,4,5,8**
		AA•A•A	**1,3,5,8**
		AAA• •A	**1,3,4,8**

not eliminated anymore !

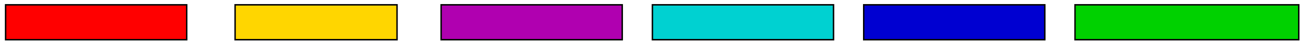
- **there are $\Omega\left(\frac{1}{2^q}\binom{n-1}{q-1}\right)$ irredundant motifs over Σ' for any q**

  Ex.: $t_k = \mathbf{A}^k\mathbf{T}\mathbf{A}^k$

(Pisanti *et al.*, Bases of motifs for generating repeated patterns with don't cares, submitted)

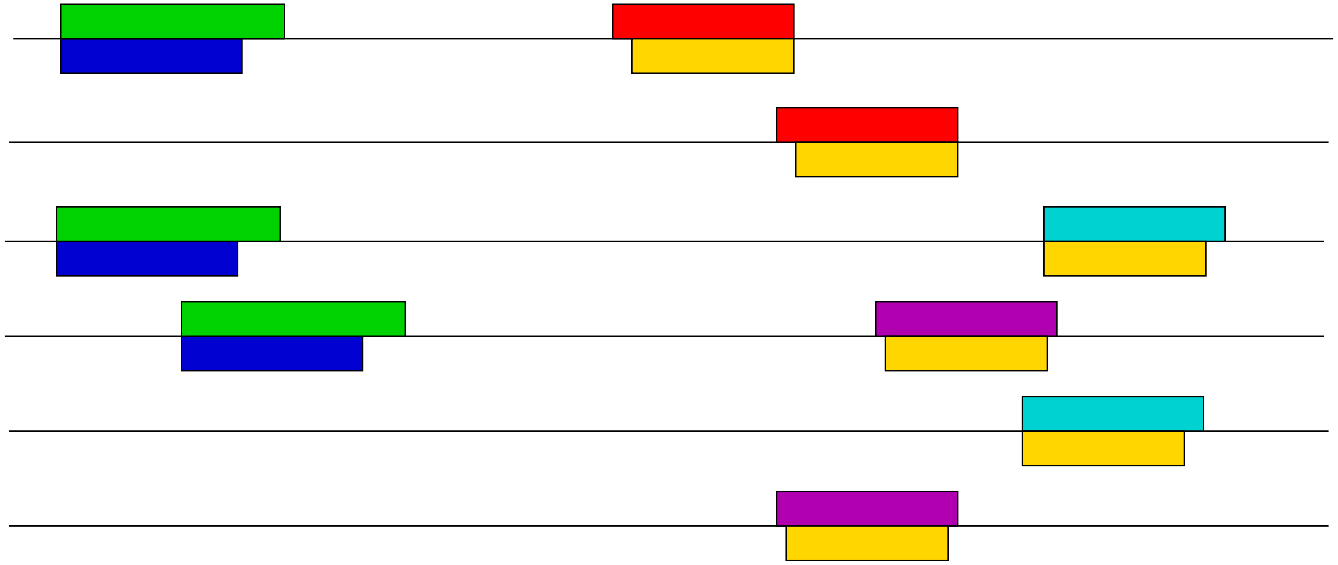
Basis of “tiling” motifs

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



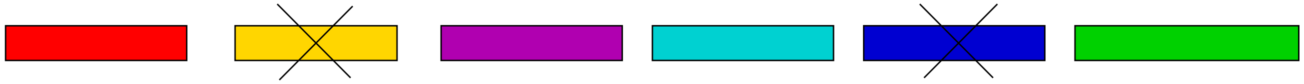
Occurrences in sequences (strings)

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 *et al.*, Bases of motifs for generating repeated patterns with don’t cares, submitted)

An example

s = **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**

Second problem:

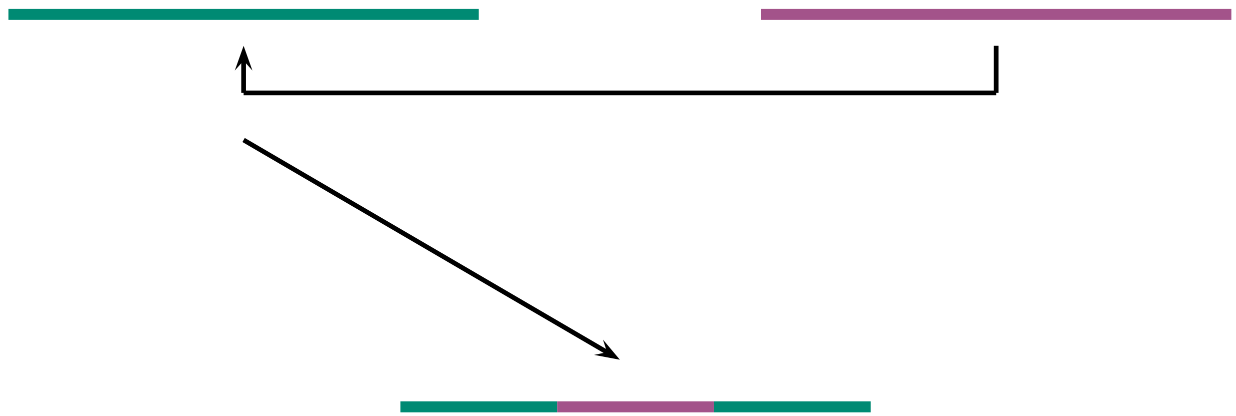
DISTANCE BETWEEN (PHYLOGENETIC) TREES

(linked to what is called "genome dynamics")

A rapid word on genetic transfer

Horizontal transfer

species or strain X_1 species or strain X_2



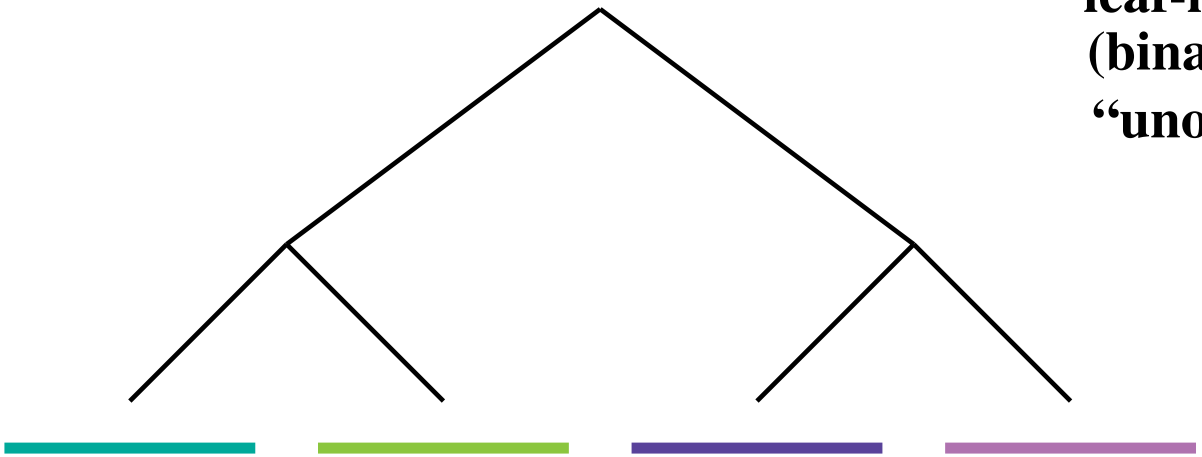
A rapid word on genetic transfer (continuation)

Conversion

species X

Trees representing the evolutionary history of a set of organisms

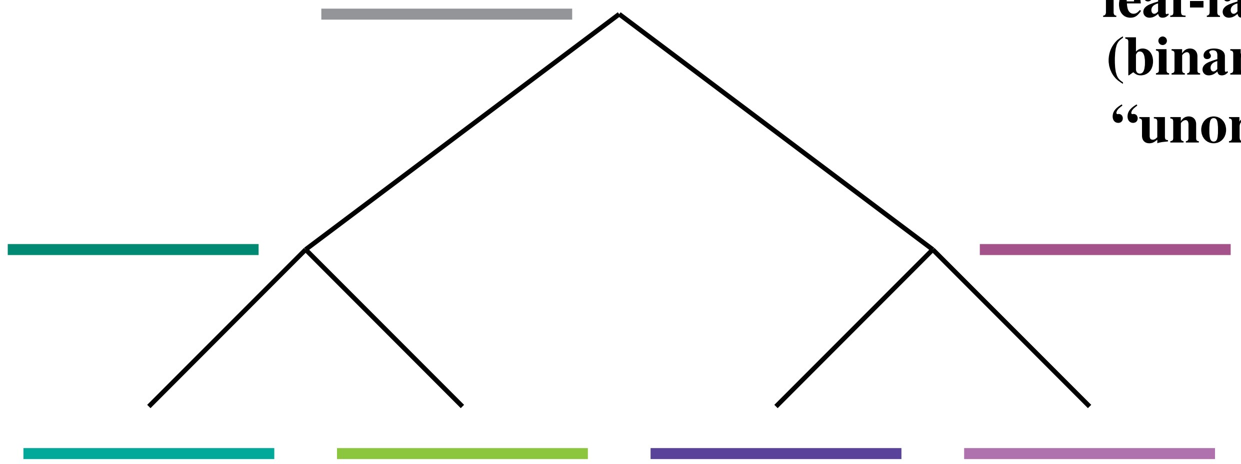
Data (here): for each organism, a set of characters over a discrete set of states



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

Phylogenetic tree

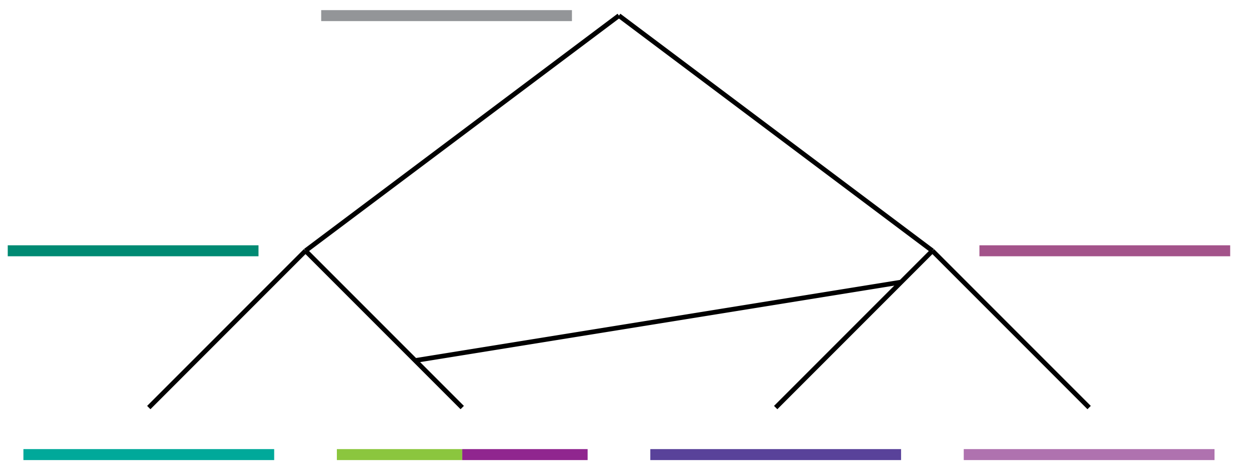
Find most parsimonious tree – “Small Phylogeny Problem”



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

Effect of transfer on the history of the organisms

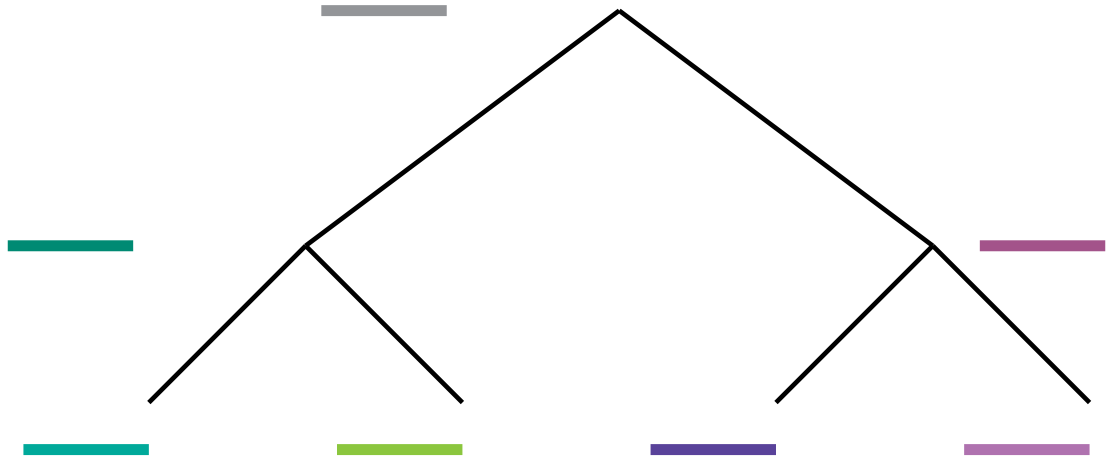
Not a single tree anymore



Effect of transfer on the history of the organisms

Not a single tree anymore but two ! (or more)

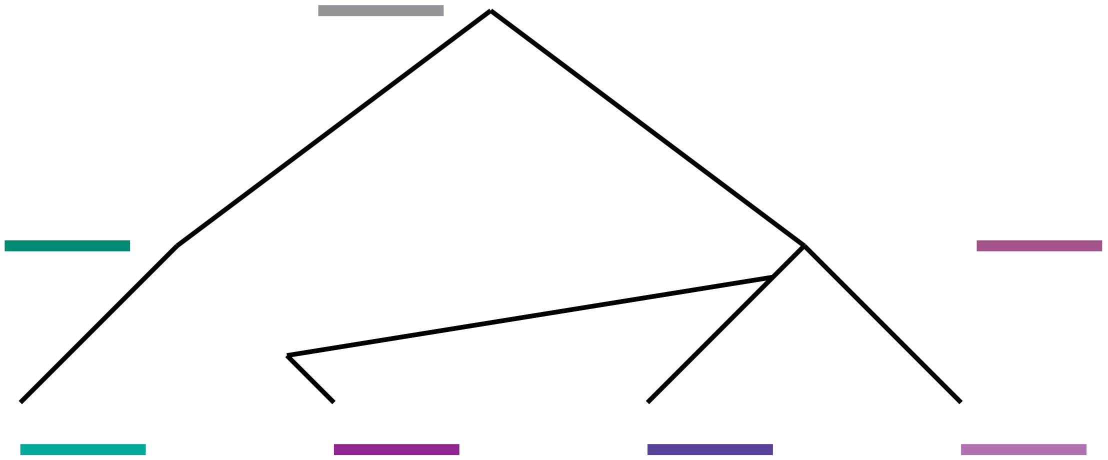
the first:



Effect of transfer on the history of the organisms

Not a single tree anymore but two ! (or more)

the second:



Genetic transfer detection and localization

Data

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

Genetic transfer detection and localization as an optimisation problem

The (ideal ?) result

alignment

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

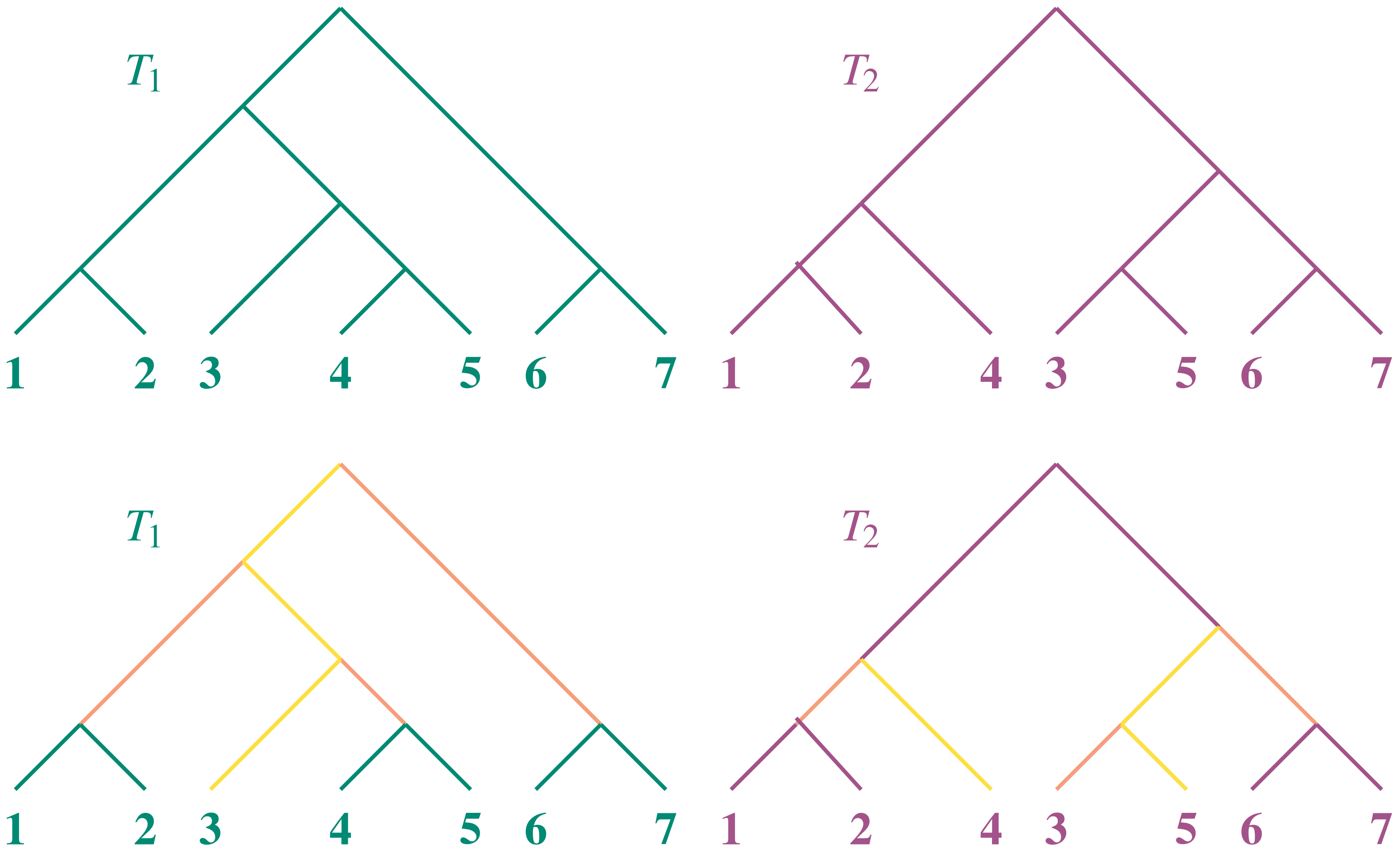
penalty should depend on genetic transfer distance between two trees

Genetic transfer distance between two phylogenetic trees

Maximum Agreement Forest of two trees

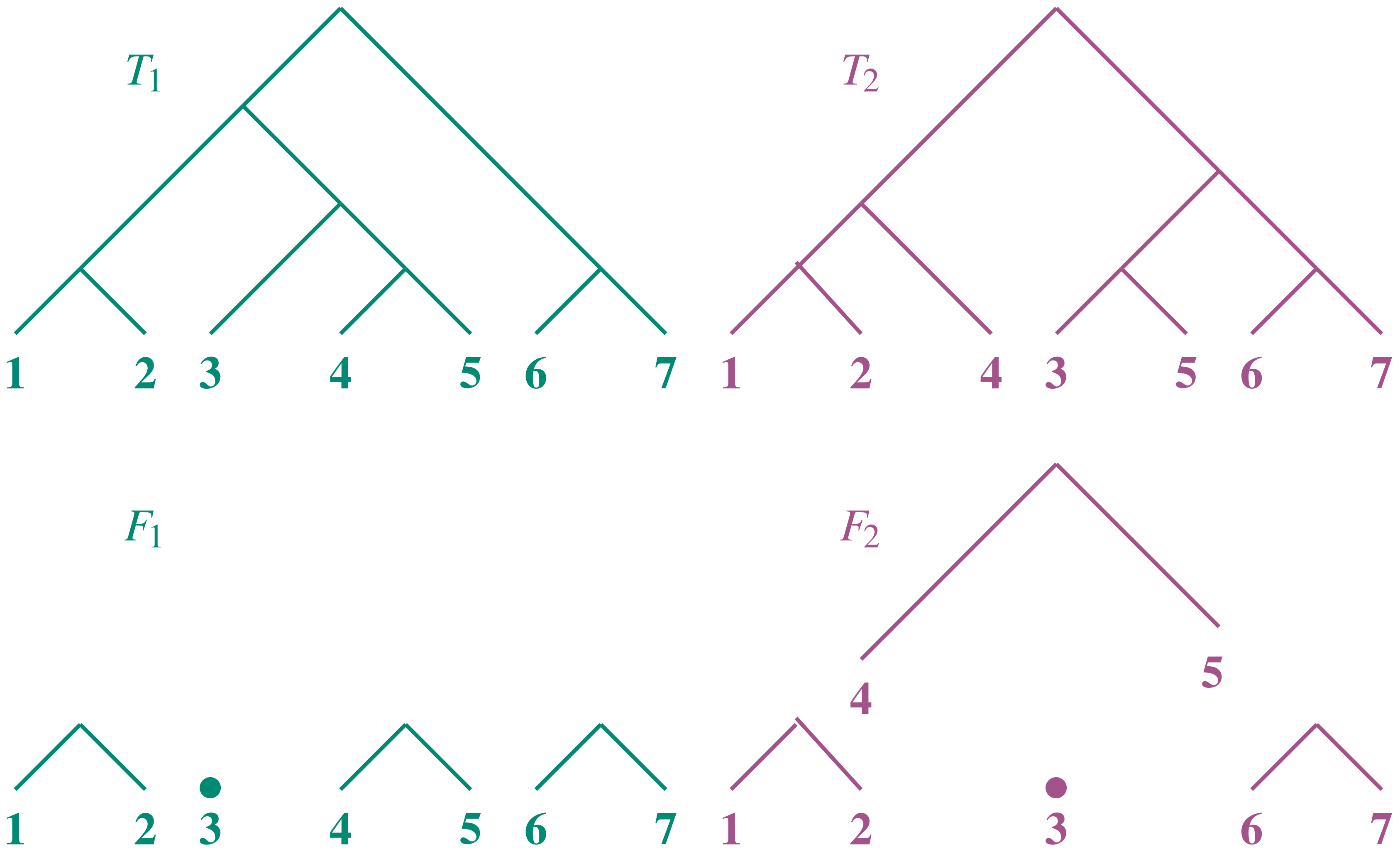
MAFALDA (E. M. Rodrigues *et al*, LNCS, 2129:159-169)
(E. M. Rodrigues *et al*, The Maximum Agreement Forest problem: approximation algorithms and computational experiments, submitted)

Agreement Forest (AF)



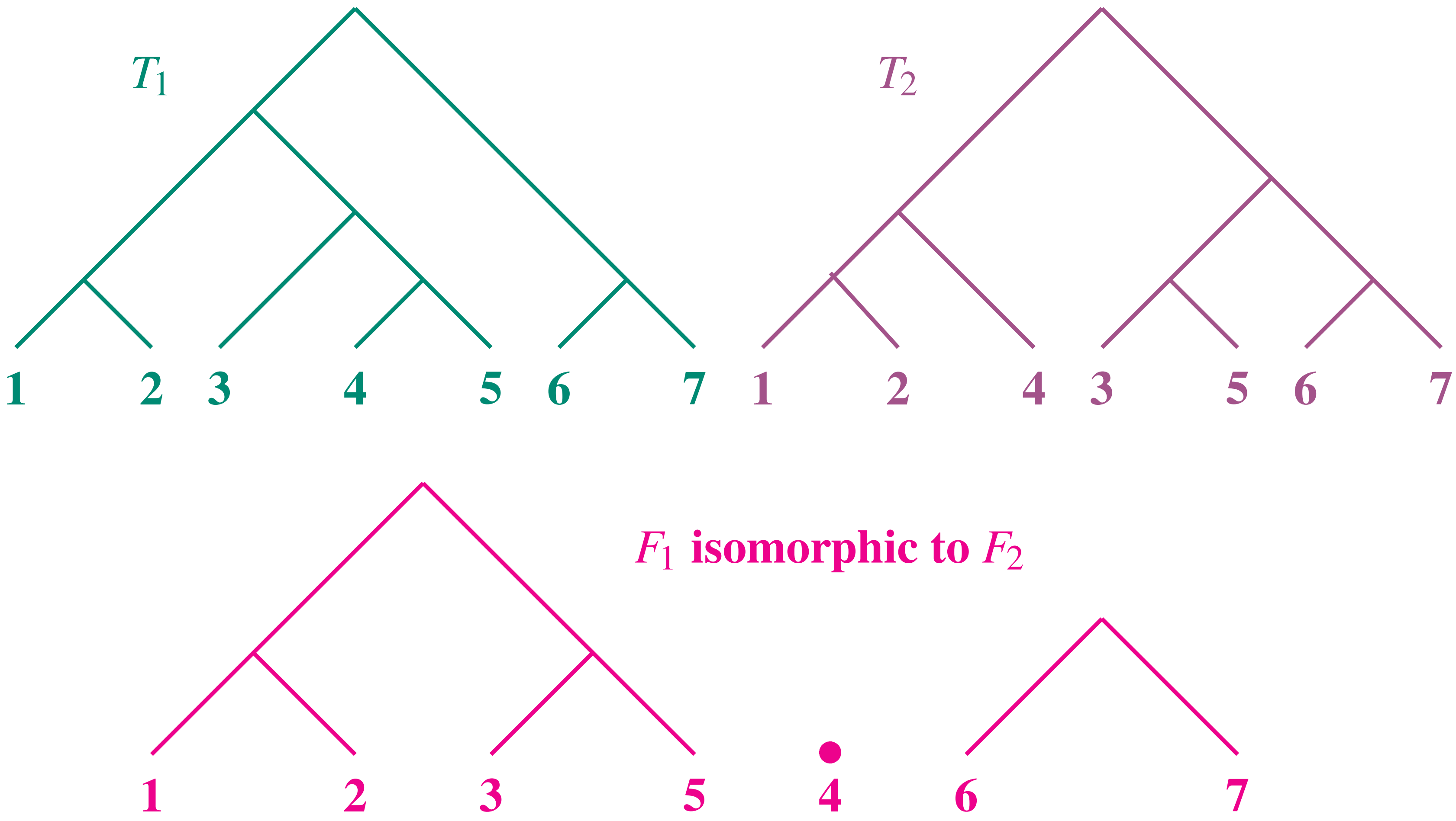
Elimination = Removal + Contraction

Agreement Forest (AF)



size (*i.e.* number of components) = 4

Maximum Agreement Forest (MAF)



size of MAF = 3

Maximum Agreement Forest problem for binary trees

Shown to be NP-Hard by Hein *et al.*, 1996

(Suggested to be MAX-SNP-Hard)

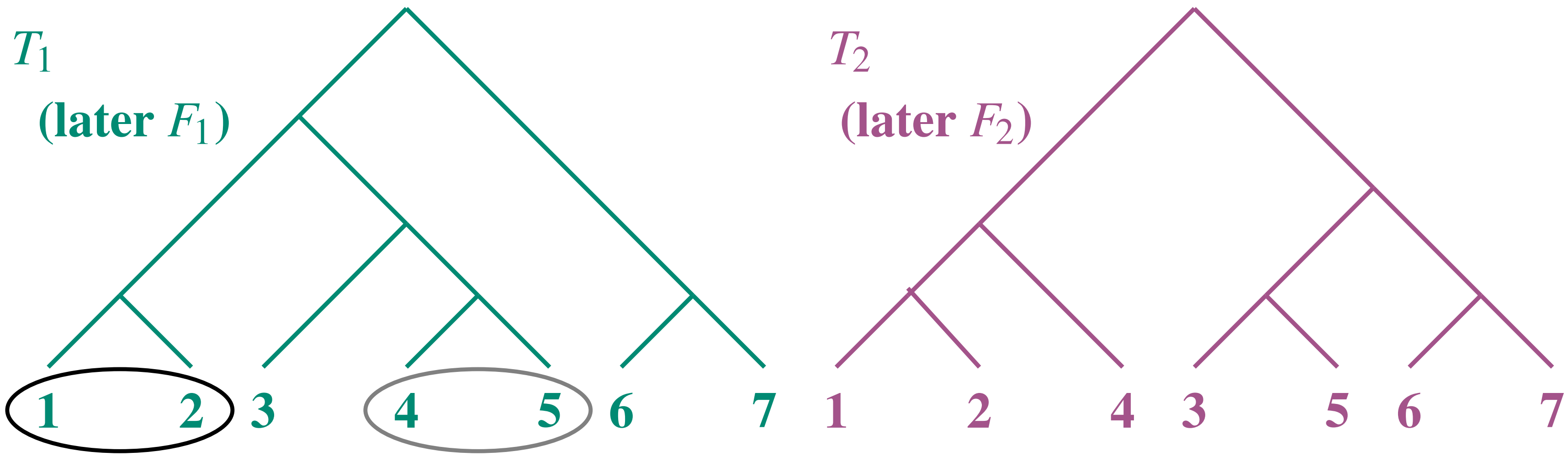
Who also provided a 3-approximation algorithm for the problem

Which, in fact, we showed is a 4-approximation

We introduced three algorithms all of which are 3-approximations

and extended the best to deal with trees of degree d (the ratio is $d + 1$)

Our (so far) best algorithm for the MAF problem



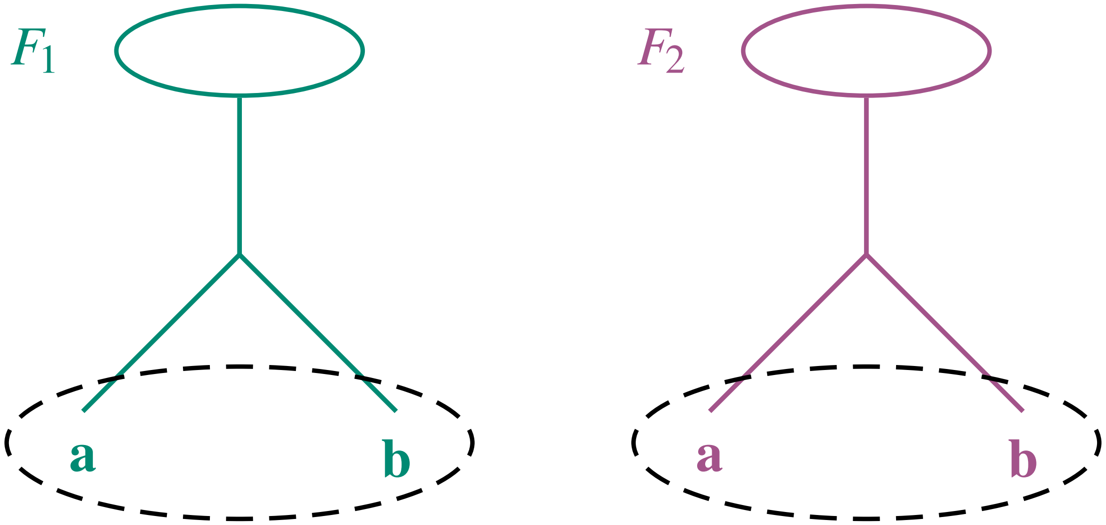
Search for pairs of sibling leaves in T_1 (F_1)

For each pair of sibling leaves in T_1 (F_1)
Consider the positions of the corresponding leaves in T_2 (F_2)

Various cases must be considered

The easiest cases

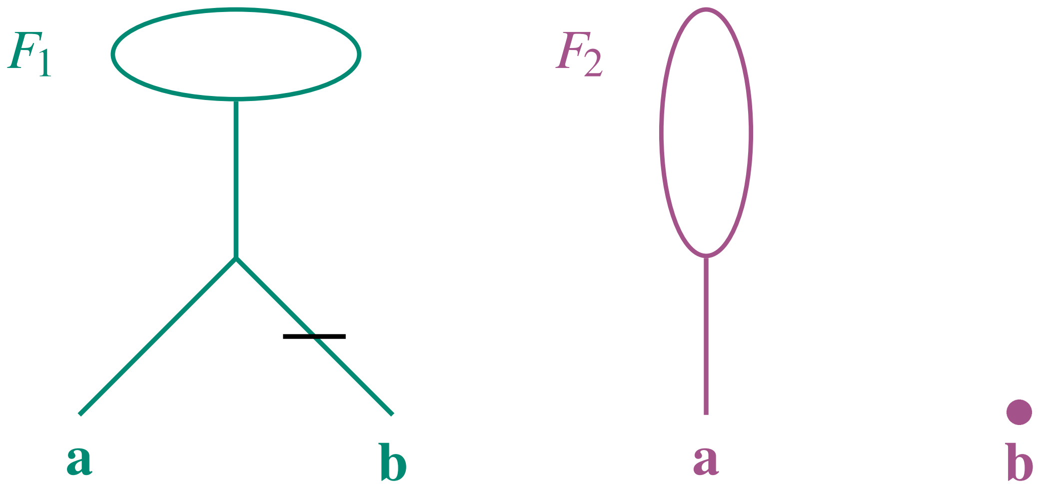
Case 1:



shrink

The easiest cases

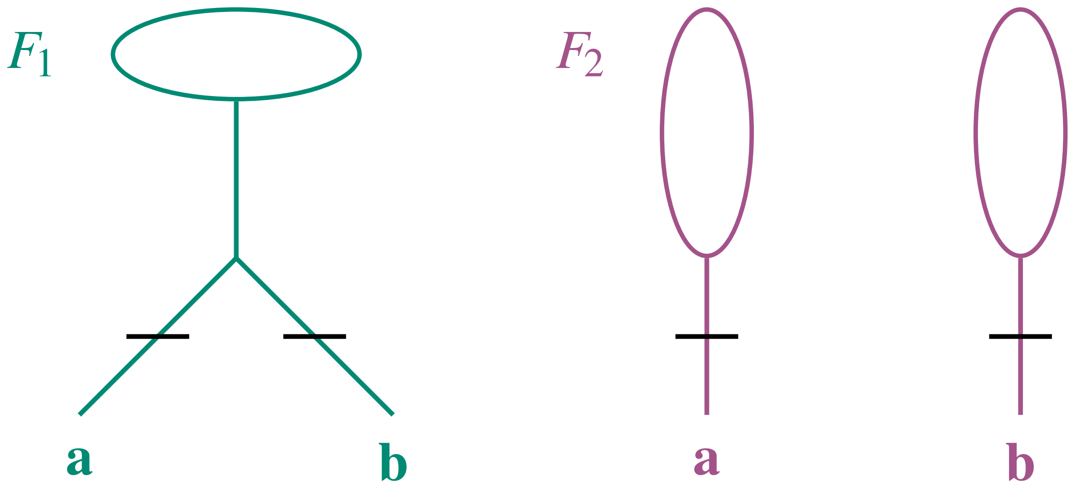
Case 2:



cut an edge in F_1 and contract

The easiest cases

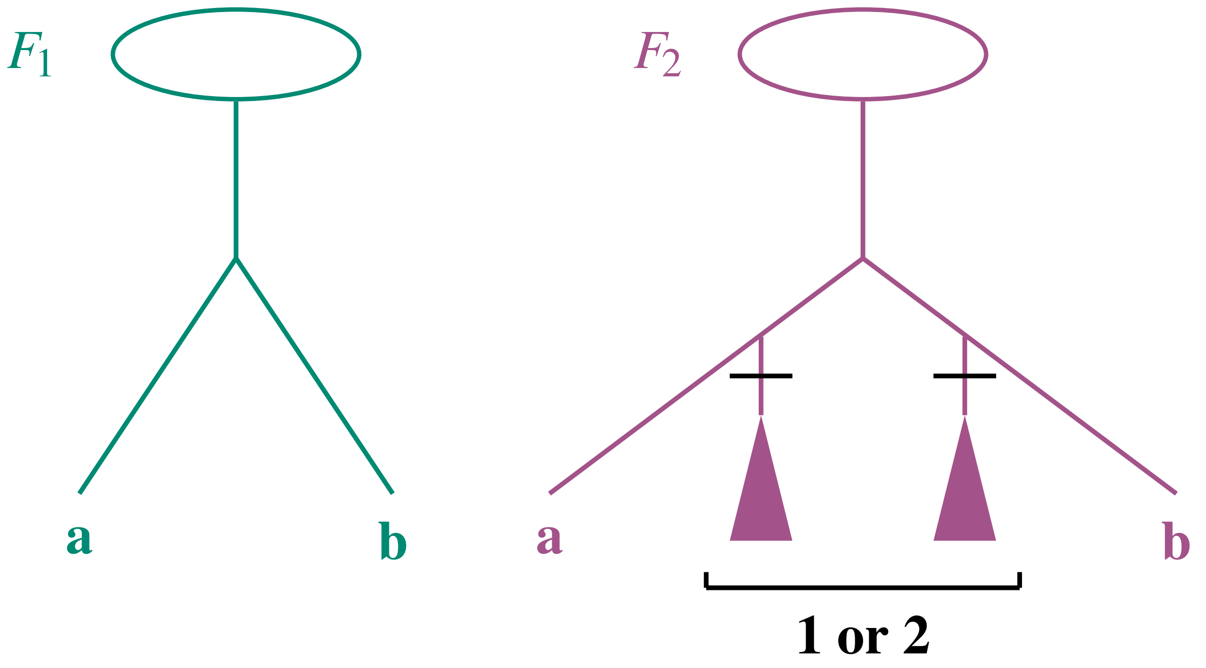
Case 3:



cut two edges each in F_1 and F_2 and contract in F_1

The two more delicate cases

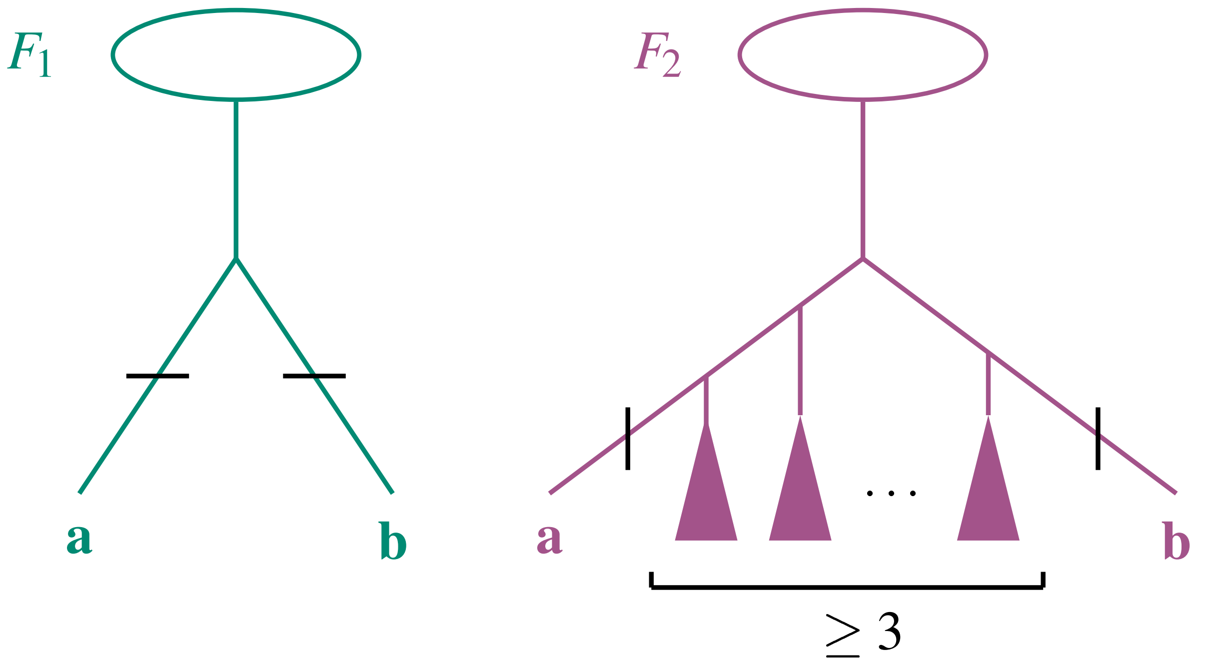
Case 4:



cut the subtrees hanging from the path from a to b in F_2

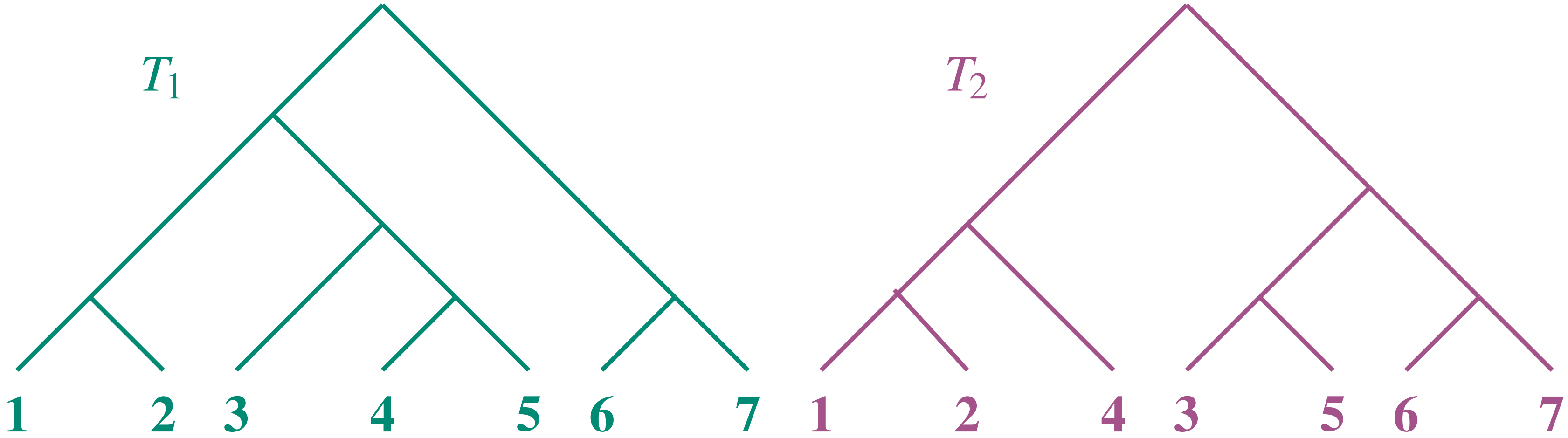
The two more delicate cases

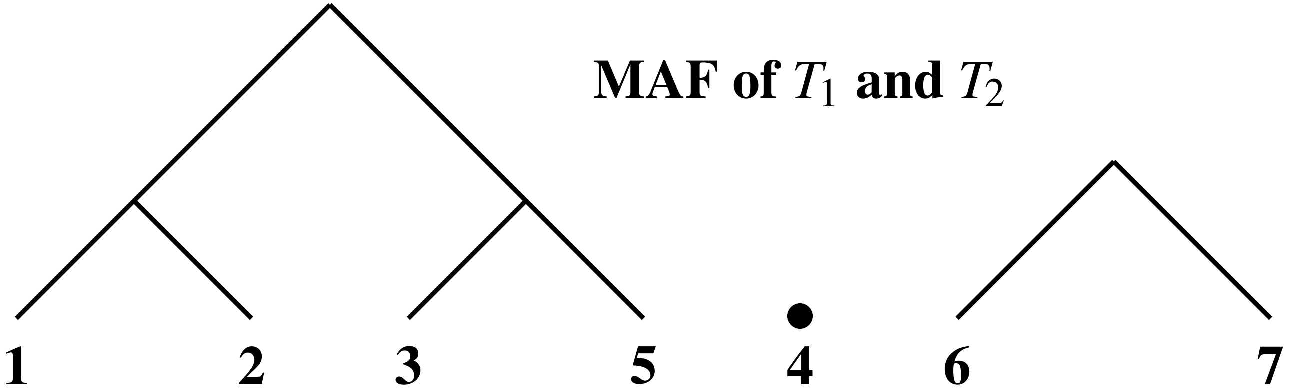
Case 5:



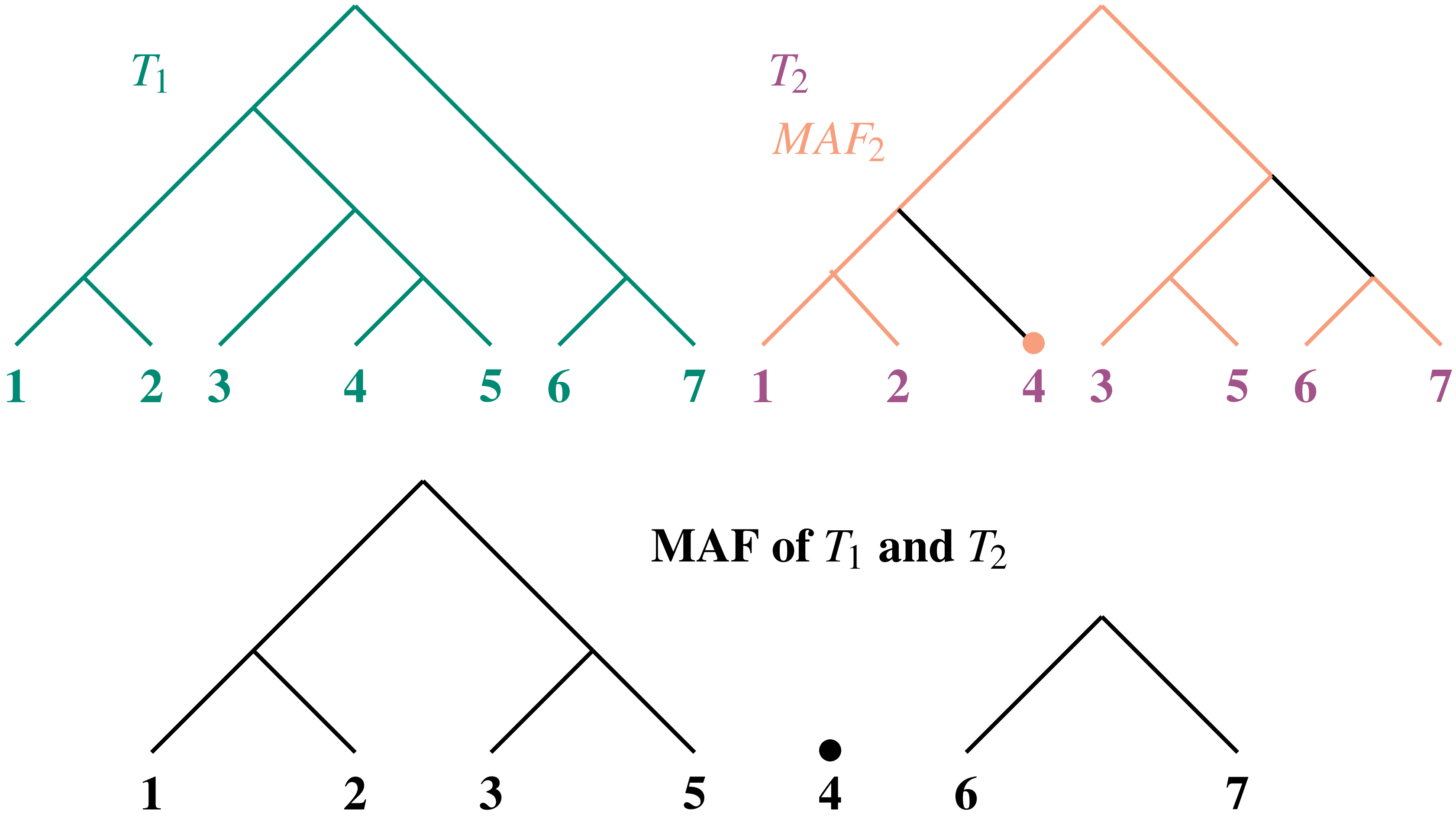
cut two edges each in F_1 and F_2
(those leading to the leaves labelled a and b)
and contract in F_1 and, possibly, F_2

Outline of the ratio analysis



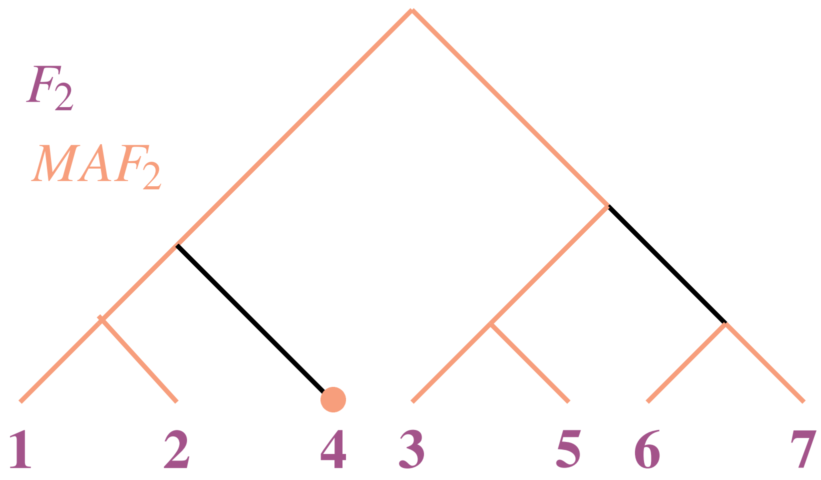


Outline of the ratio analysis



edges NOT belonging to MAF_2 are called links

Outline of the ratio analysis



algorithm does "good" and "bad" eliminations

good elimination: elimination of a link

number of good eliminations is MAF-1

idea: prove at least one third of all eliminations are good

again various cases to analyse

Where is the main difficulty

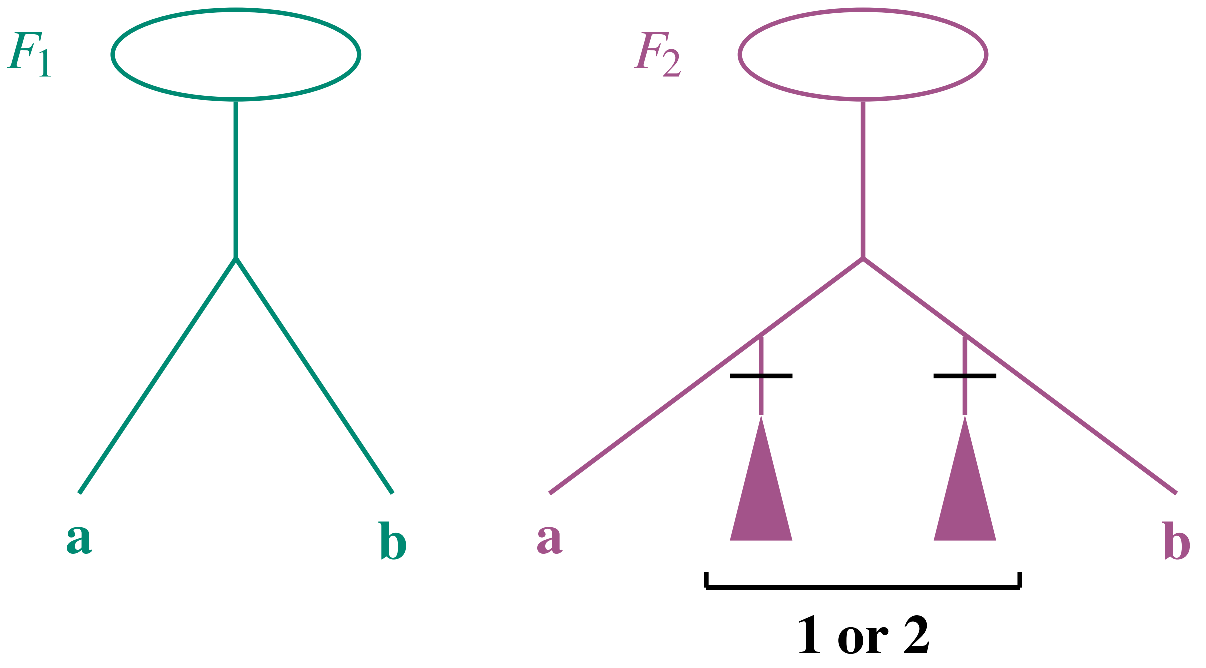
Bad eliminations must be compensated (possibly later only) with the good elimination of an edge

“belonging to the same case”

as led to the elimination because we do not know how many times each case will appear

The two more delicate cases

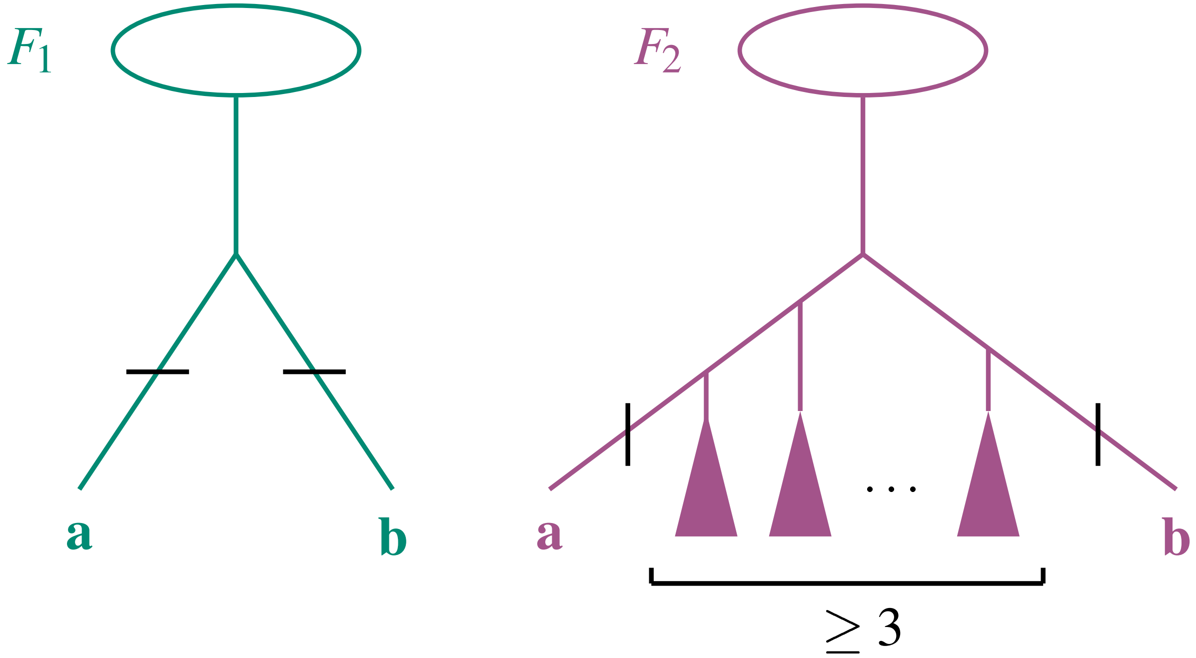
Case 4:



possibly both are bad eliminations but...

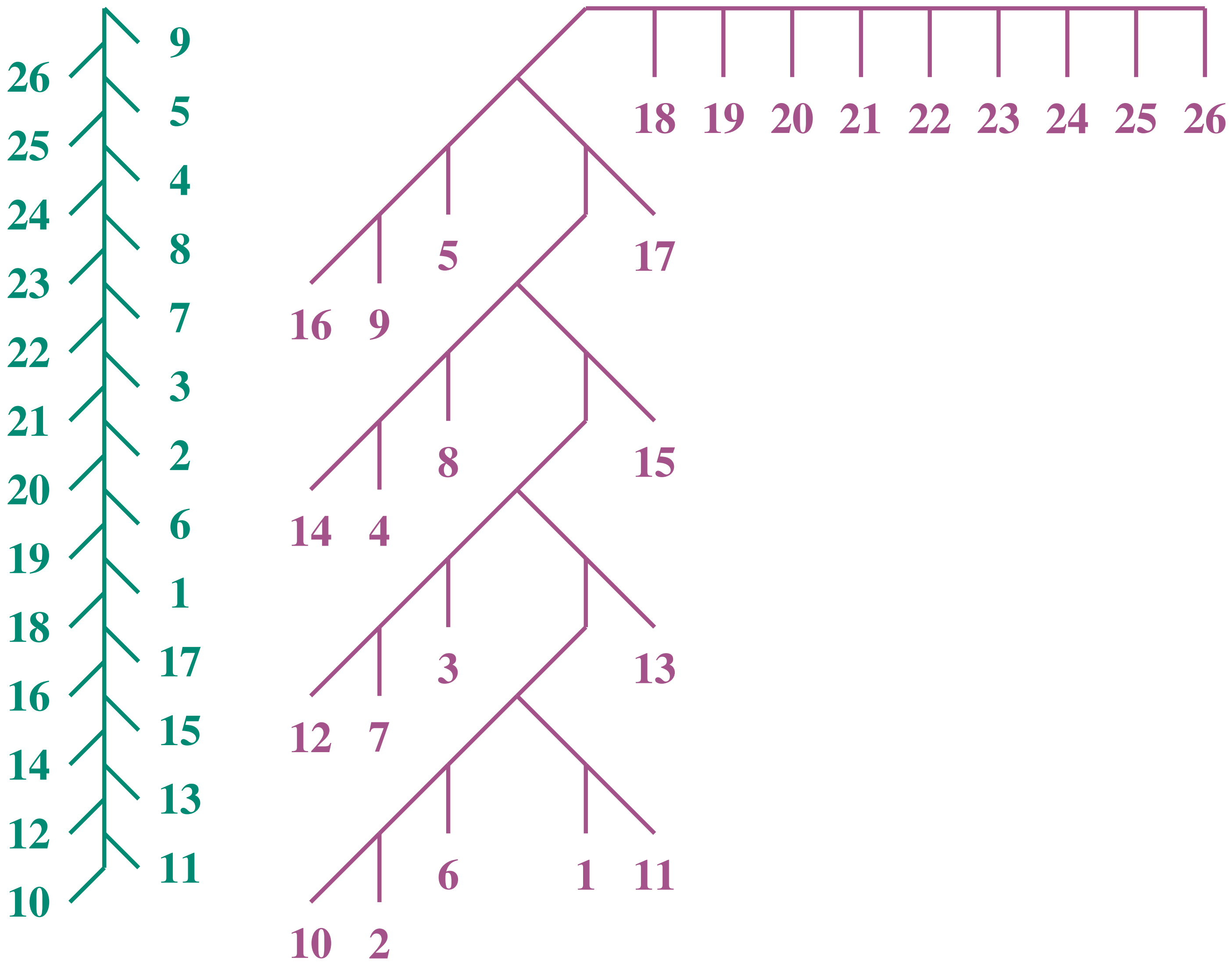
The two more delicate cases

Case 5:

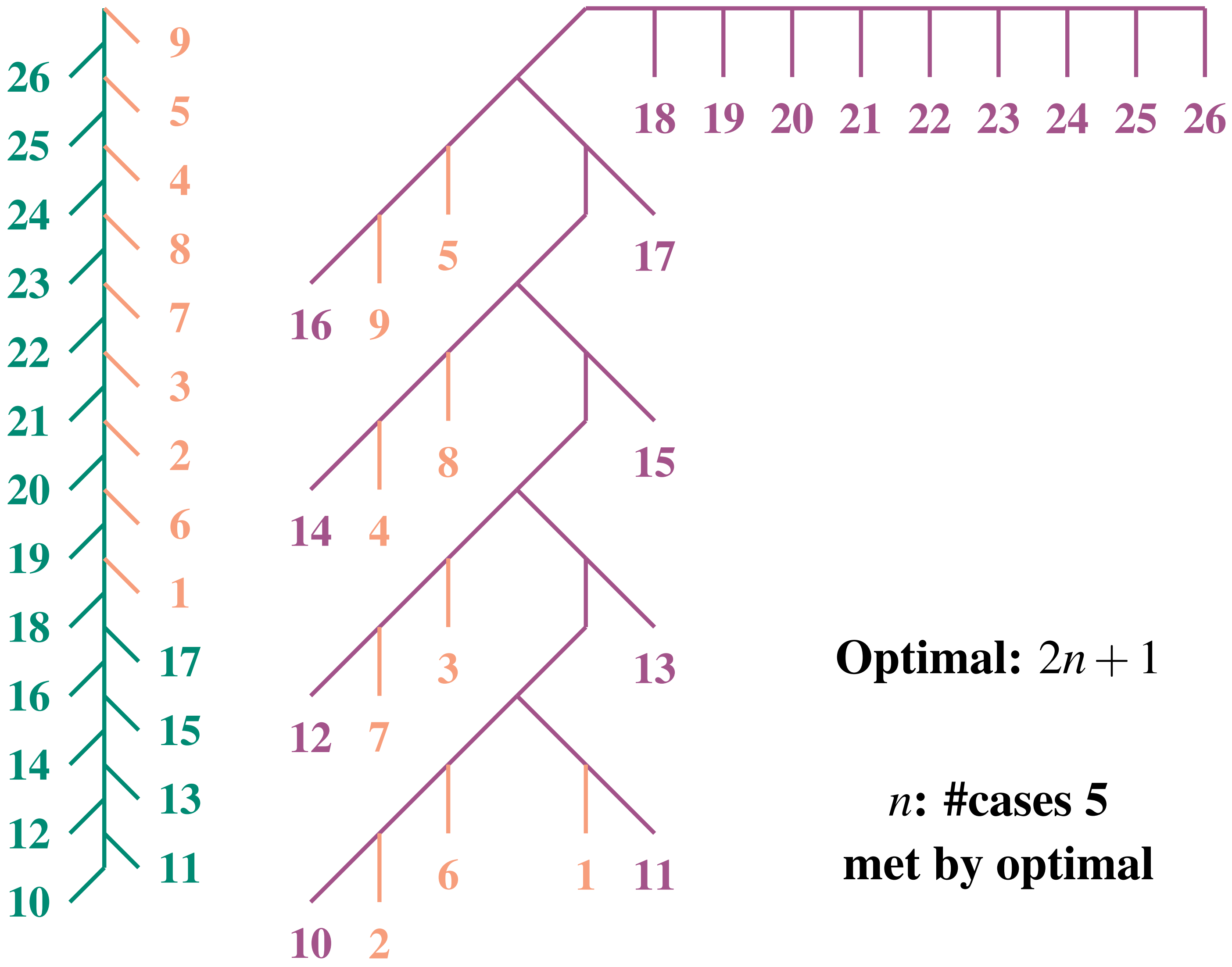


possibly both are bad eliminations but...

This is a tight ratio



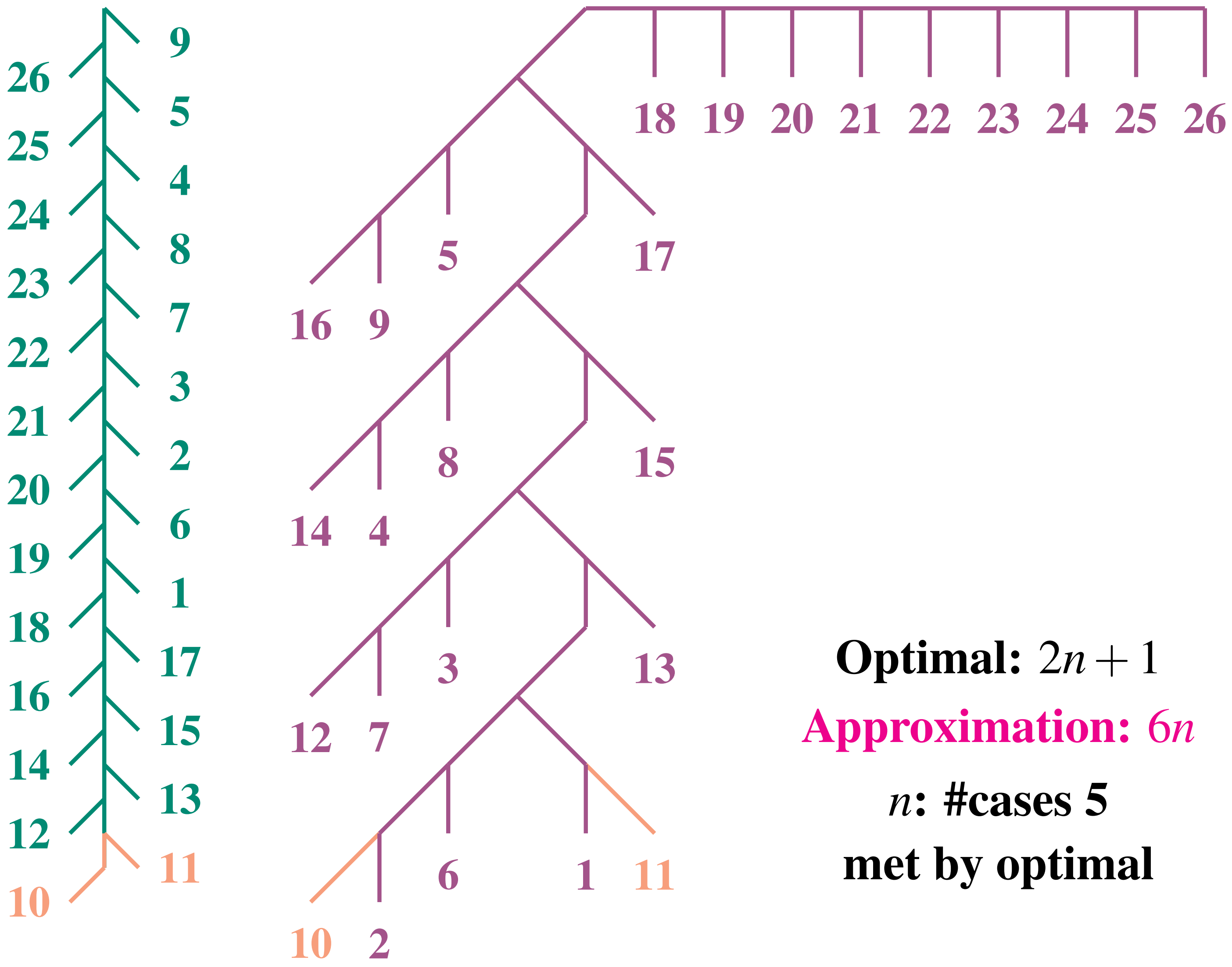
This is a tight ratio



Optimal: $2n+1$

n: **#cases 5**
met by optimal

This is a tight ratio



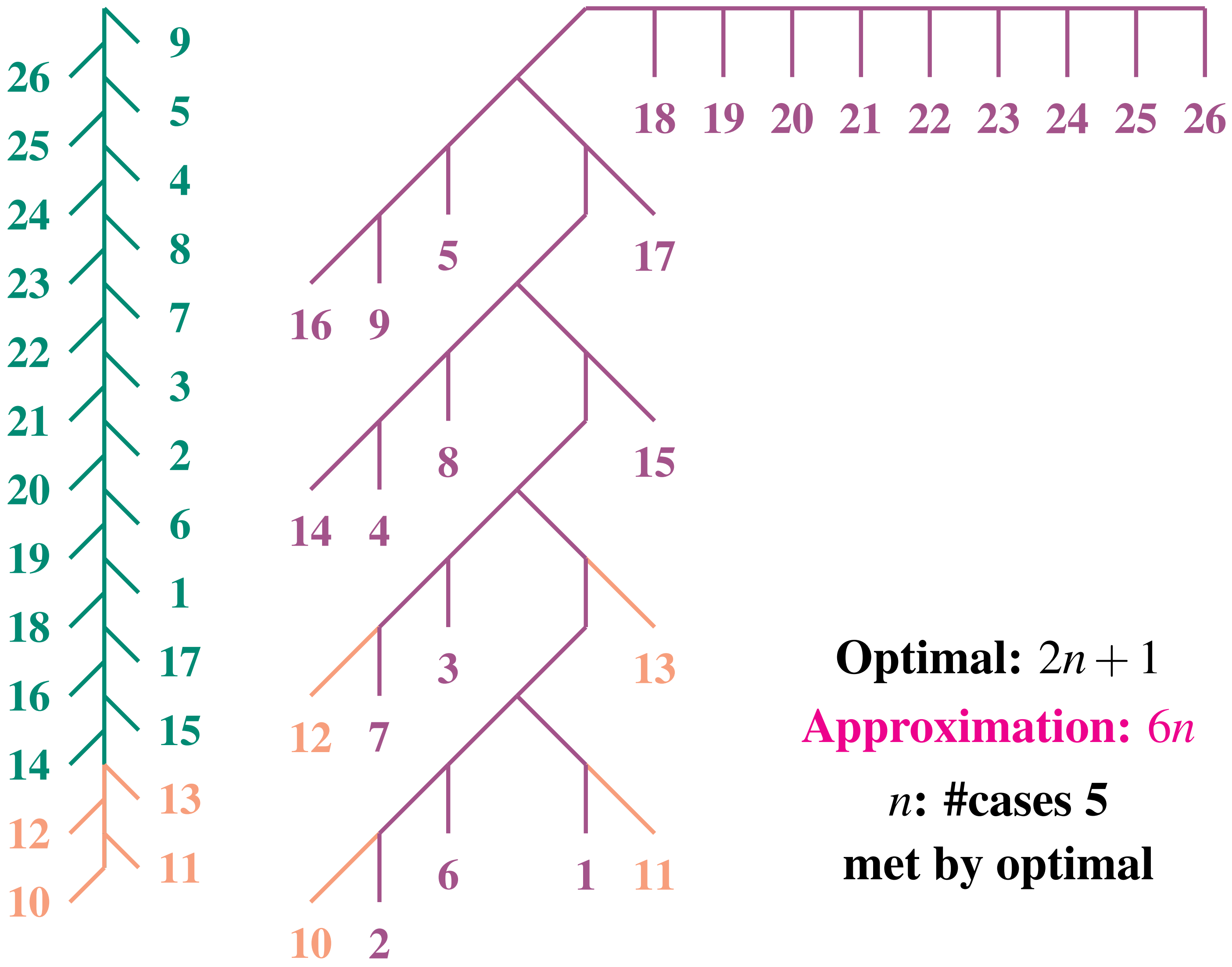
Optimal: $2n+1$

Approximation: $6n$

n**: #cases 5**

met by optimal

This is a tight ratio



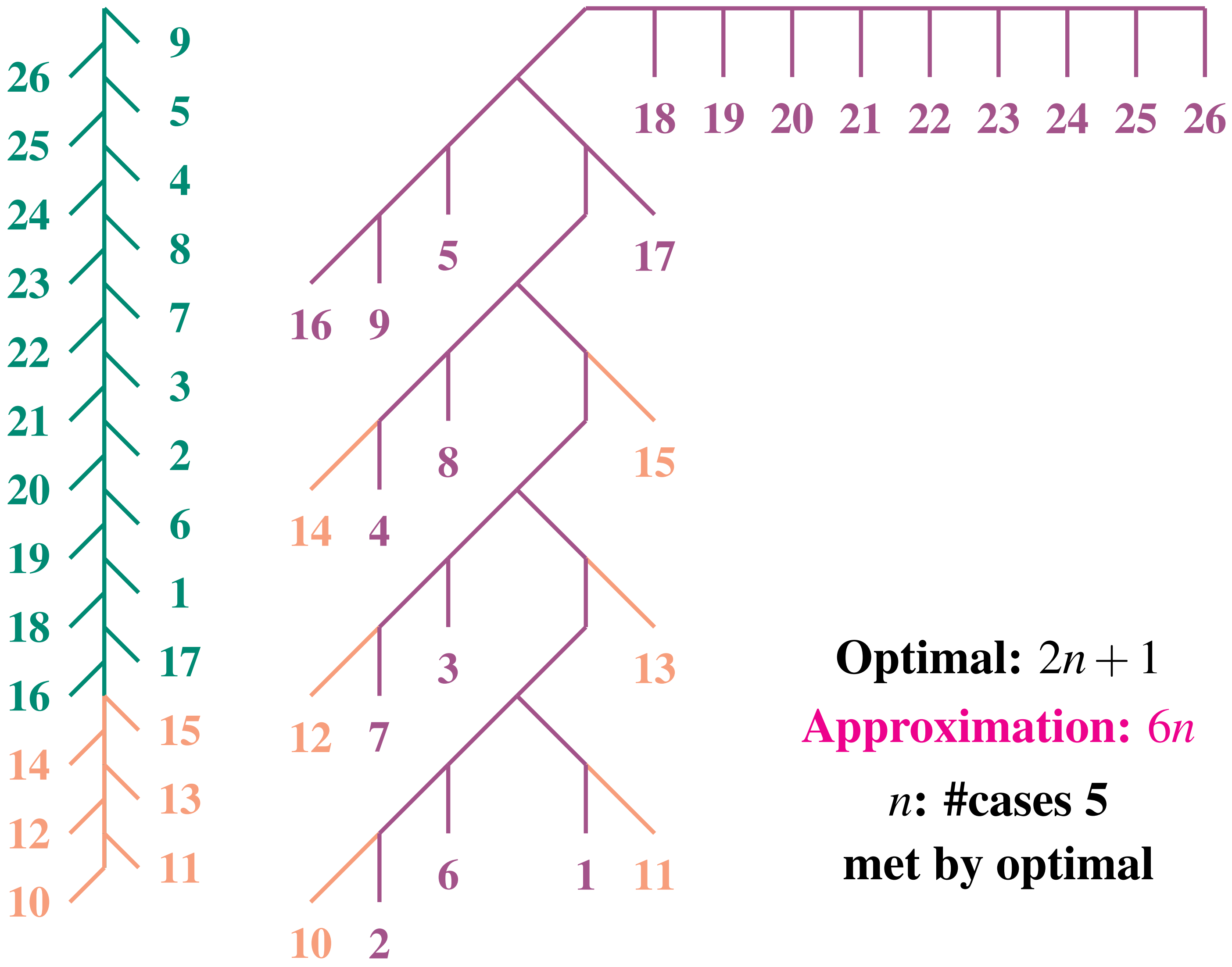
Optimal: $2n + 1$

Approximation: $6n$

n**: #cases 5**

met by optimal

This is a tight ratio

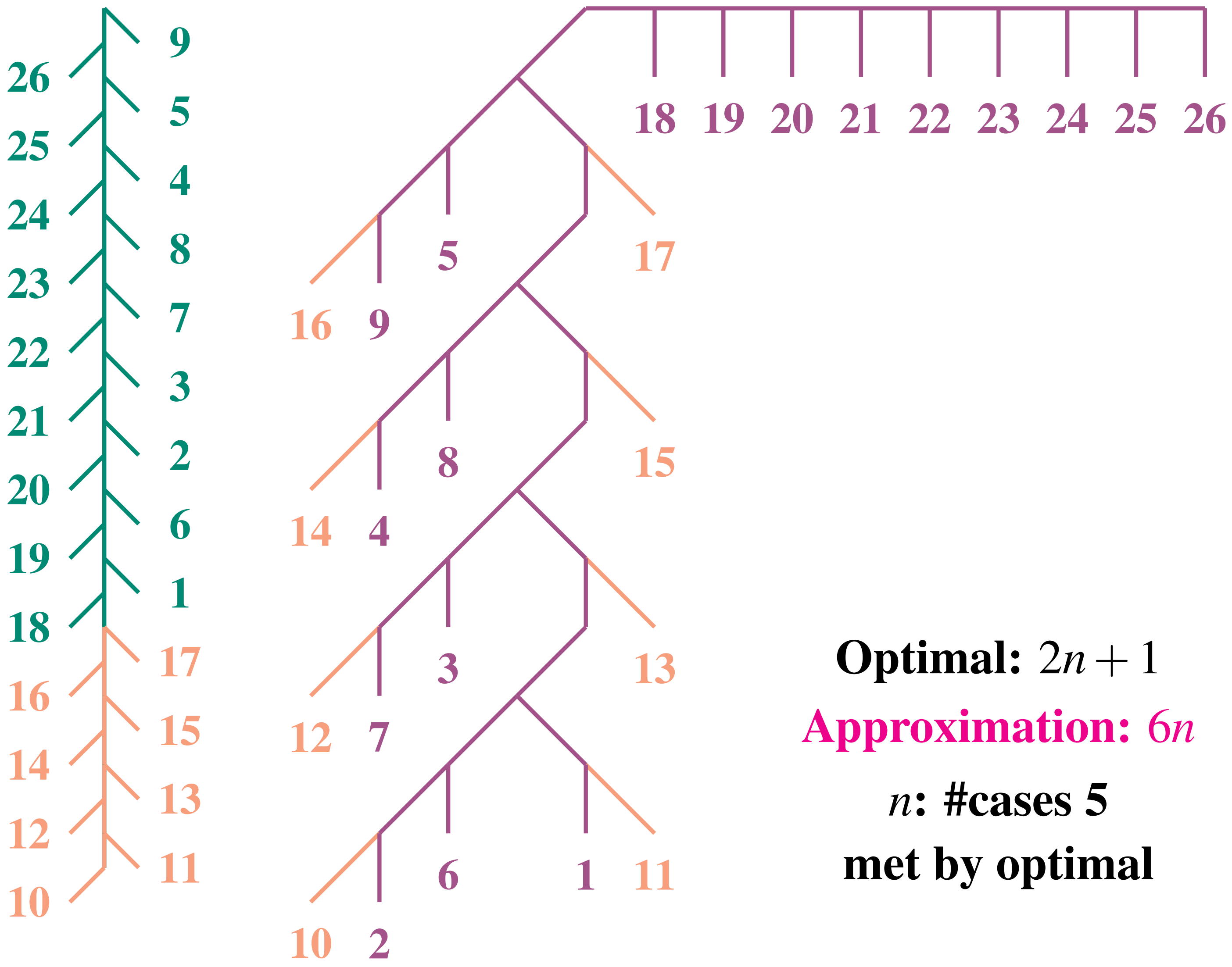


Optimal: $2n+1$

Approximation: $6n$

n**: #cases 5 met by optimal**

This is a tight ratio

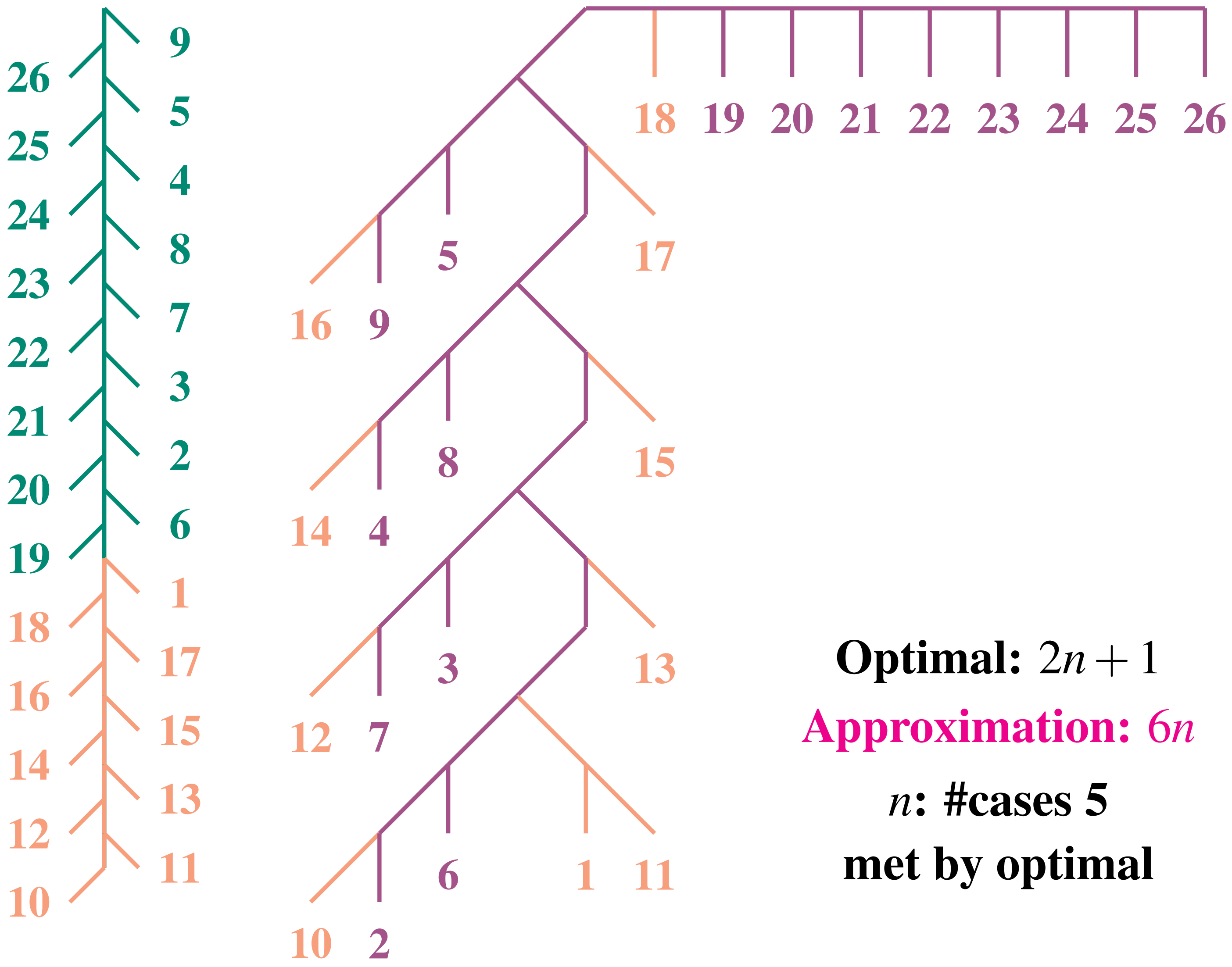


Optimal: $2n+1$

Approximation: $6n$

n**: #cases 5 met by optimal**

This is a tight ratio

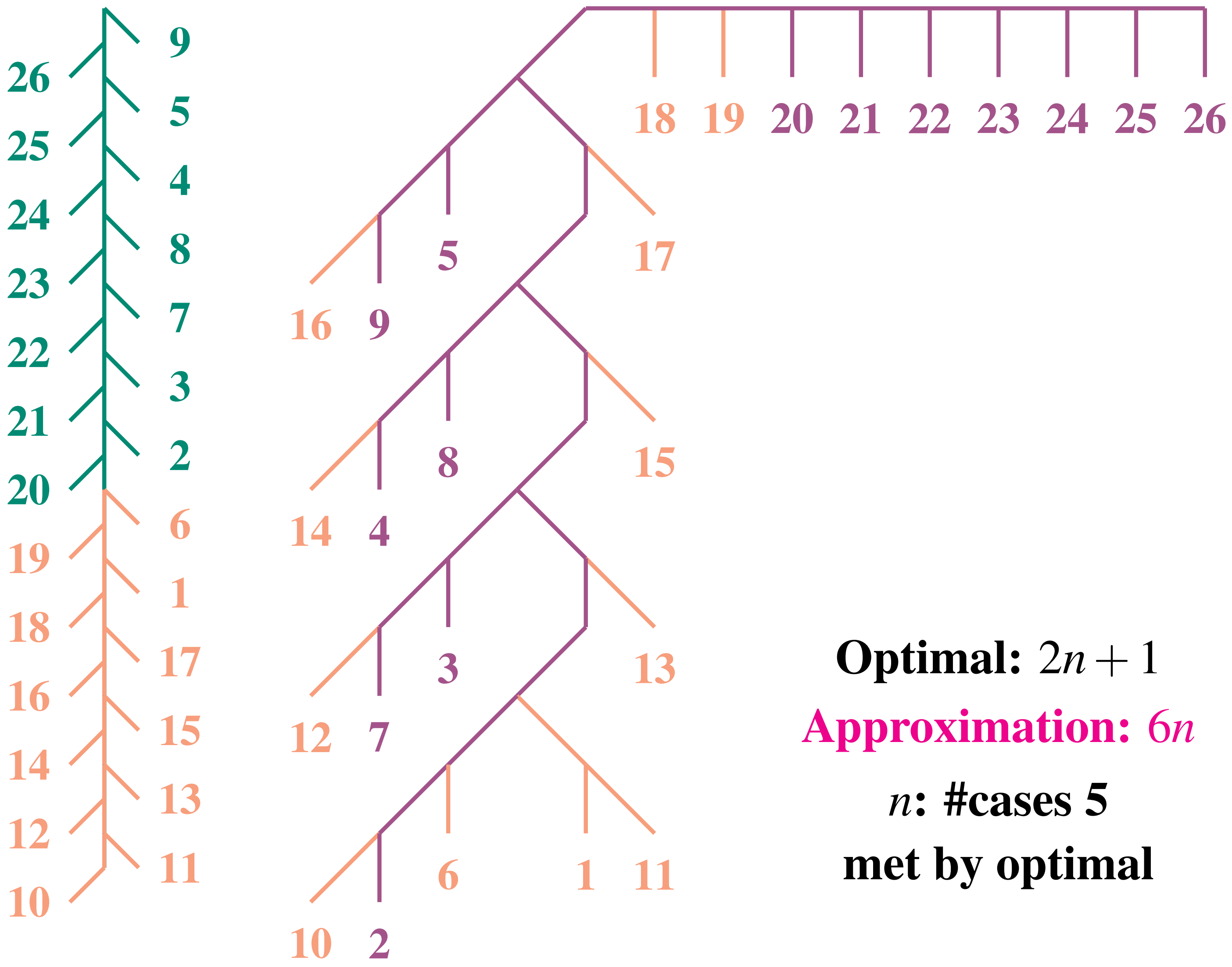


Optimal: $2n+1$

Approximation: $6n$

n**: #cases 5 met by optimal**

This is a tight ratio

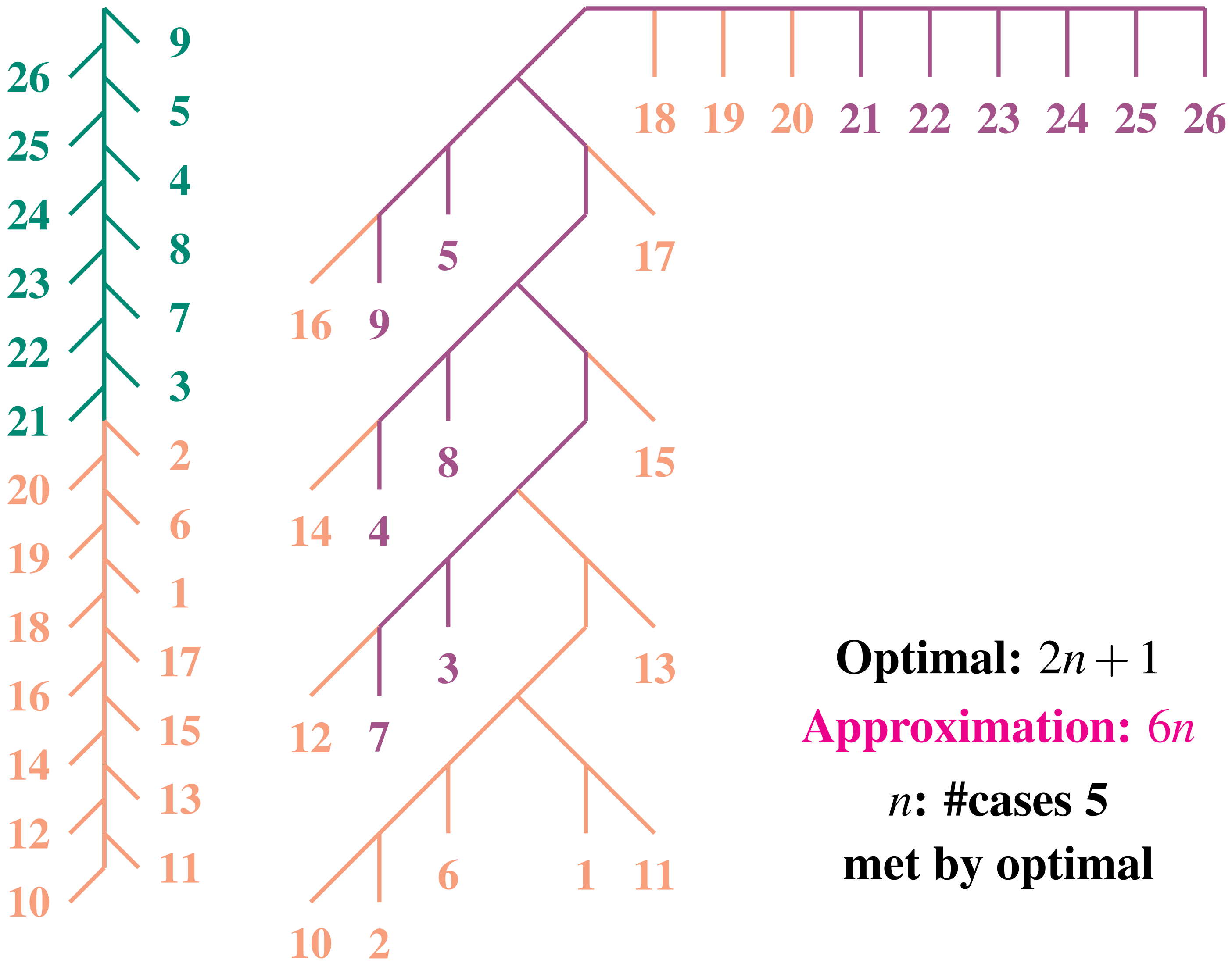


Optimal: $2n+1$

Approximation: $6n$

n: **#cases 5 met by optimal**

This is a tight ratio

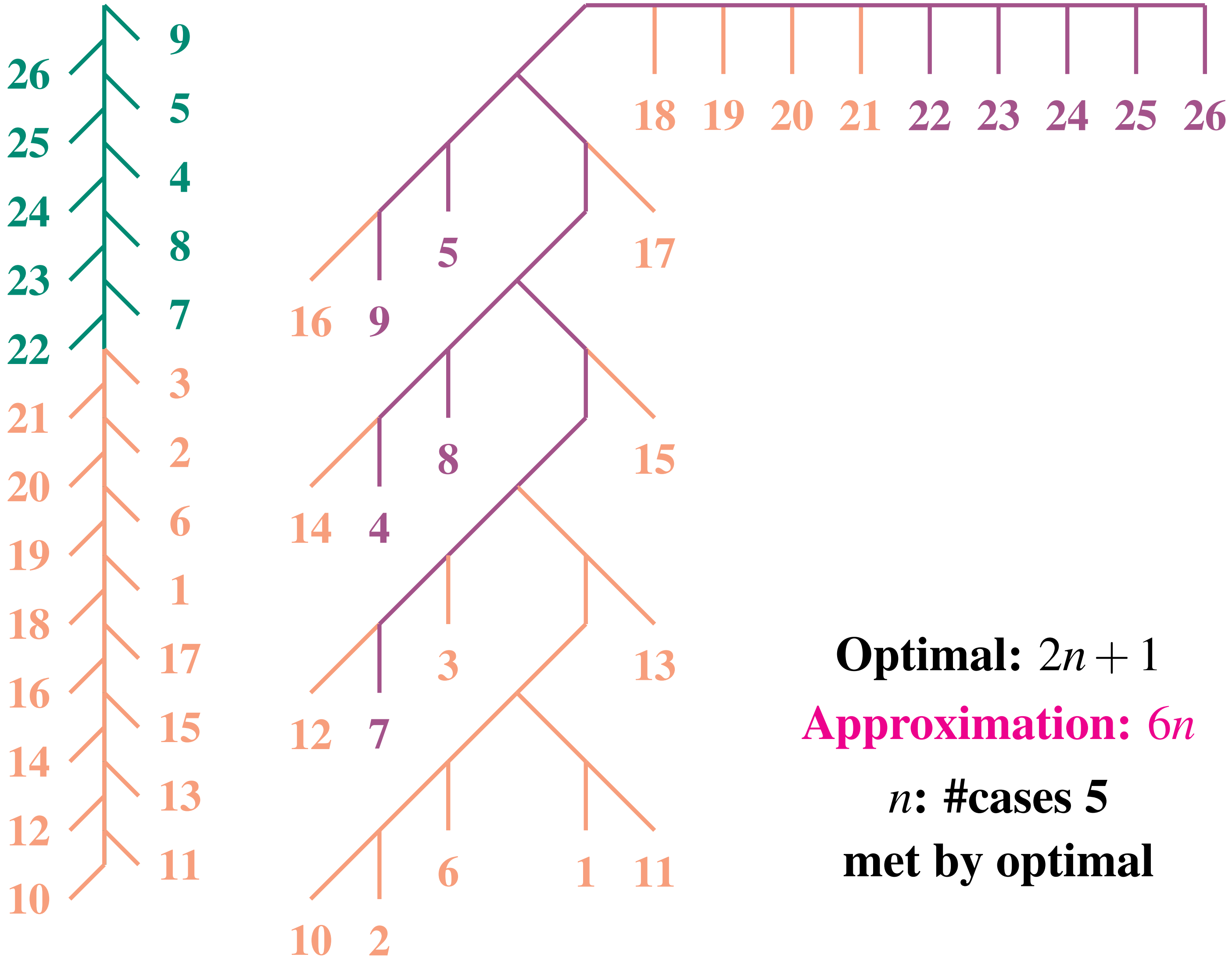


Optimal: $2n+1$

Approximation: $6n$

n**: #cases 5 met by optimal**

This is a tight ratio



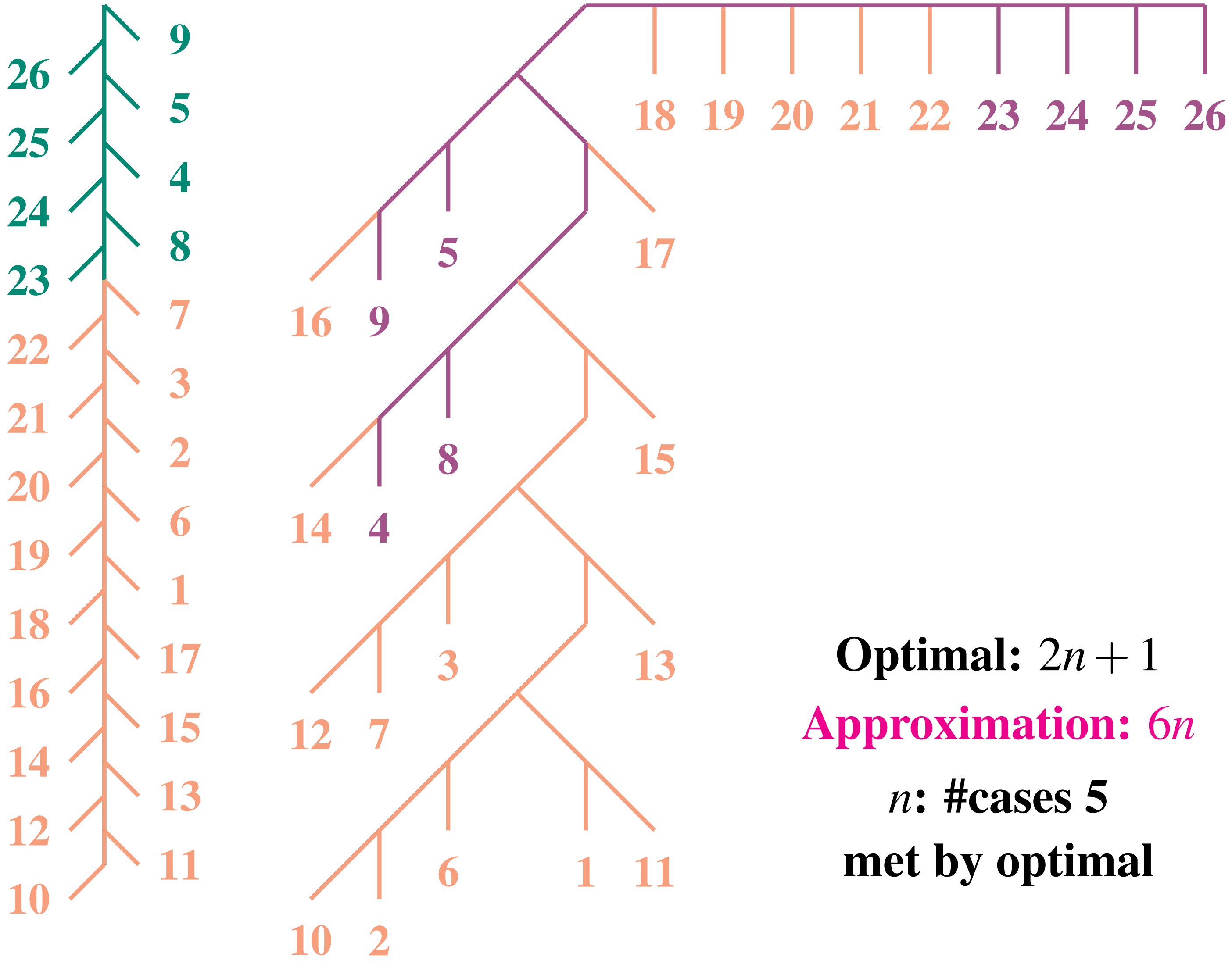
Optimal: $2n+1$

Approximation: $6n$

n: **#cases 5**

met by optimal

This is a tight ratio

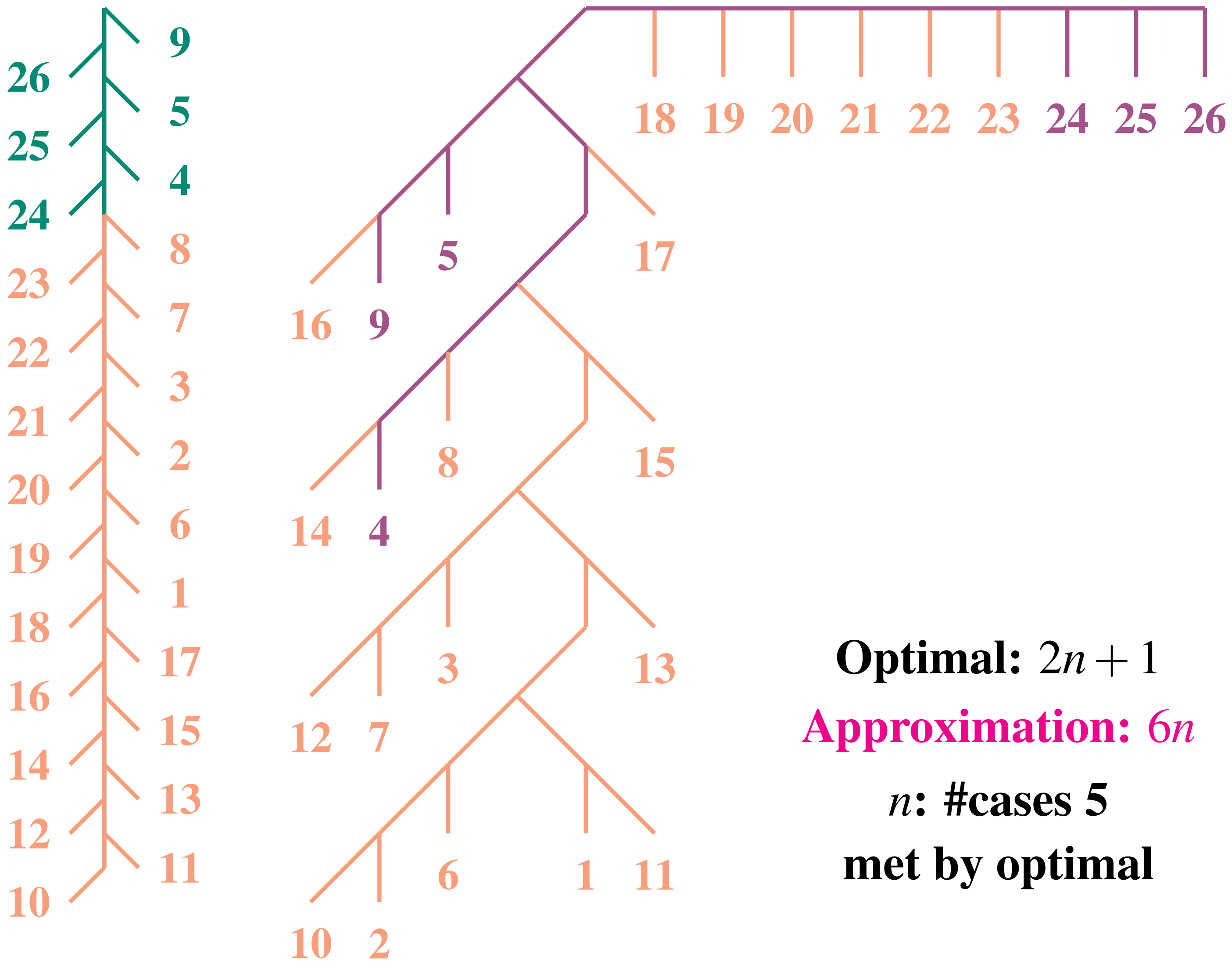


Optimal: $2n+1$

Approximation: $6n$

n**: #cases 5**
met by optimal

This is a tight ratio

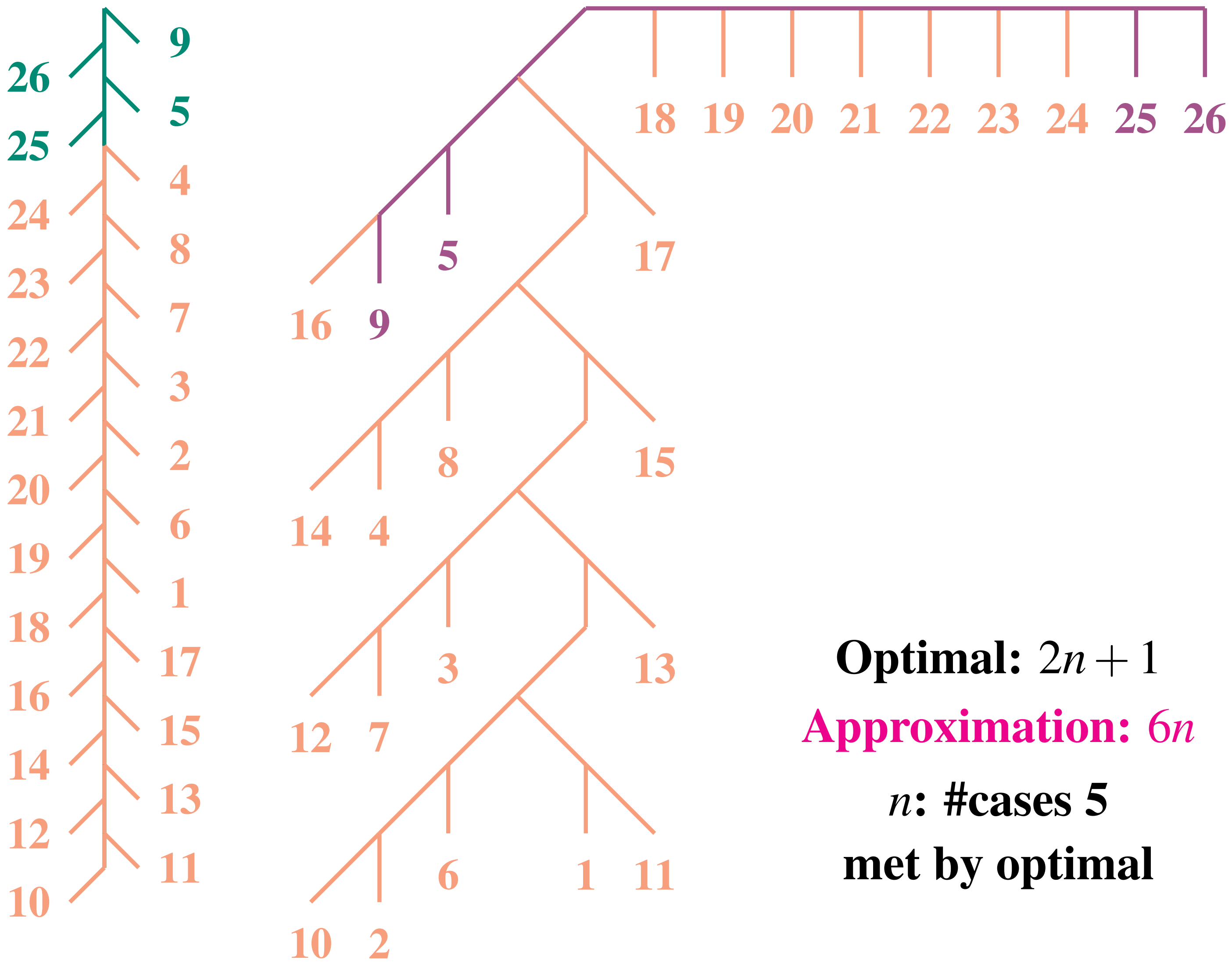


Optimal: $2n+1$

Approximation: $6n$

n**: #cases 5 met by optimal**

This is a tight ratio

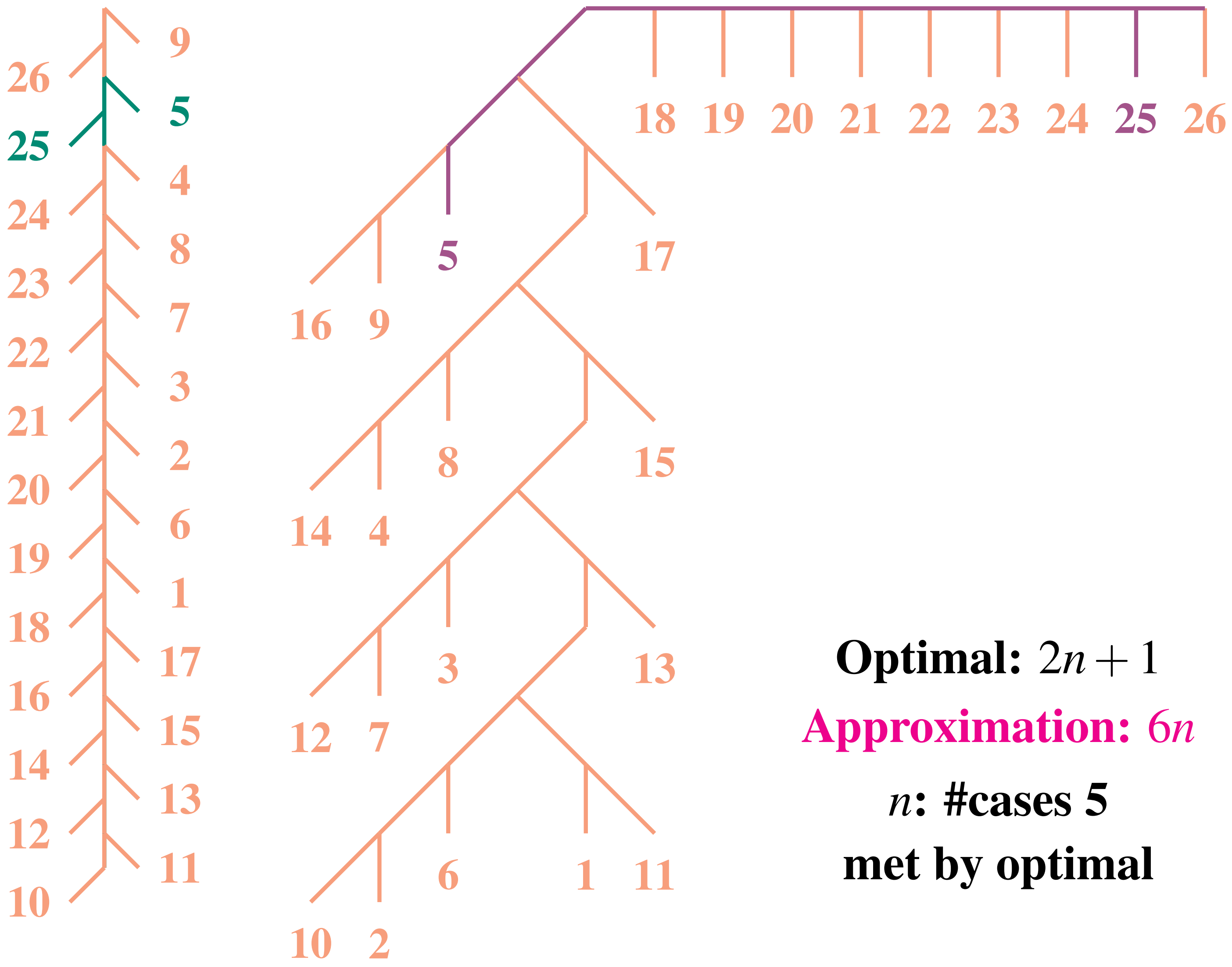


Optimal: $2n+1$

Approximation: $6n$

n: **#cases 5 met by optimal**

This is a tight ratio



Optimal: $2n+1$

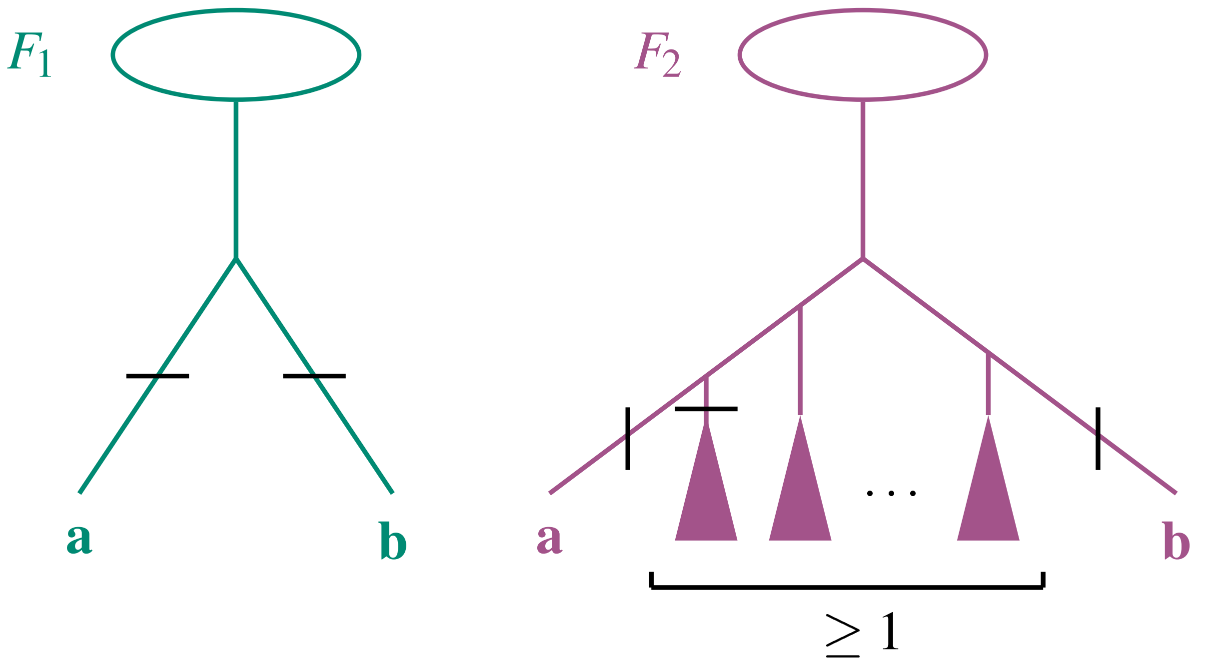
Approximation: $6n$

n**: #cases 5**

met by optimal

A simpler algorithm with the same ratio

Cases 4 and 5:



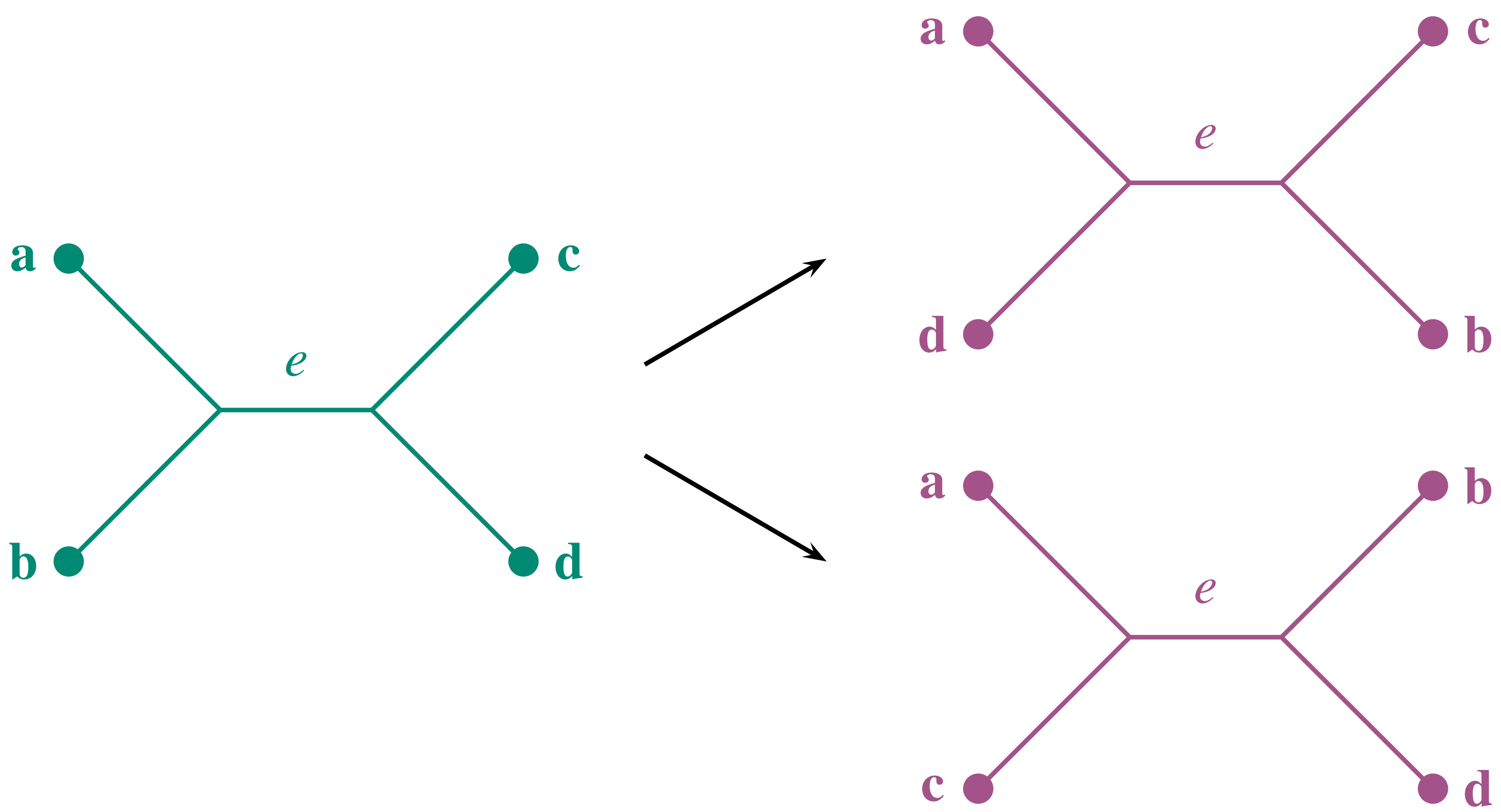
cut the two edges in F_1 and F_2
leading to the leaves labelled a and b
plus one subtree (any one) hanging from the path from a to b

Results on simulated data

File	Alg.2	Alg.1	Alg.3	lb	(Alg.2)/lb
Optimum $\leq$ 100 (99 subtree tranfers)					
001-100	139	150	265	89	1.57
002-100	150	153	270	91	1.65
003-100	152	156	277	93	1.65
004-100	137	141	255	86	1.60
005-100	148	151	267	90	1.65
006-100	142	148	250	84	1.70
007-100	142	148	265	89	1.60
008-100	144	147	266	90	1.60
009-100	146	146	264	89	1.65
010-100	147	153	259	87	1.69
Optimum $\leq$ 400 (399 subtree tranfers)					
001-400	515	556	744	249	2.07
002-400	526	562	731	245	2.15
003-400	513	553	731	245	2.10
004-400	492	546	723	242	2.04
005-400	514	562	733	245	2.10
006-400	526	569	708	237	2.22
007-400	508	558	758	254	2.00
008-400	501	551	711	238	2.11
009-400	529	570	747	250	2.12
010-400	539	582	751	251	2.15

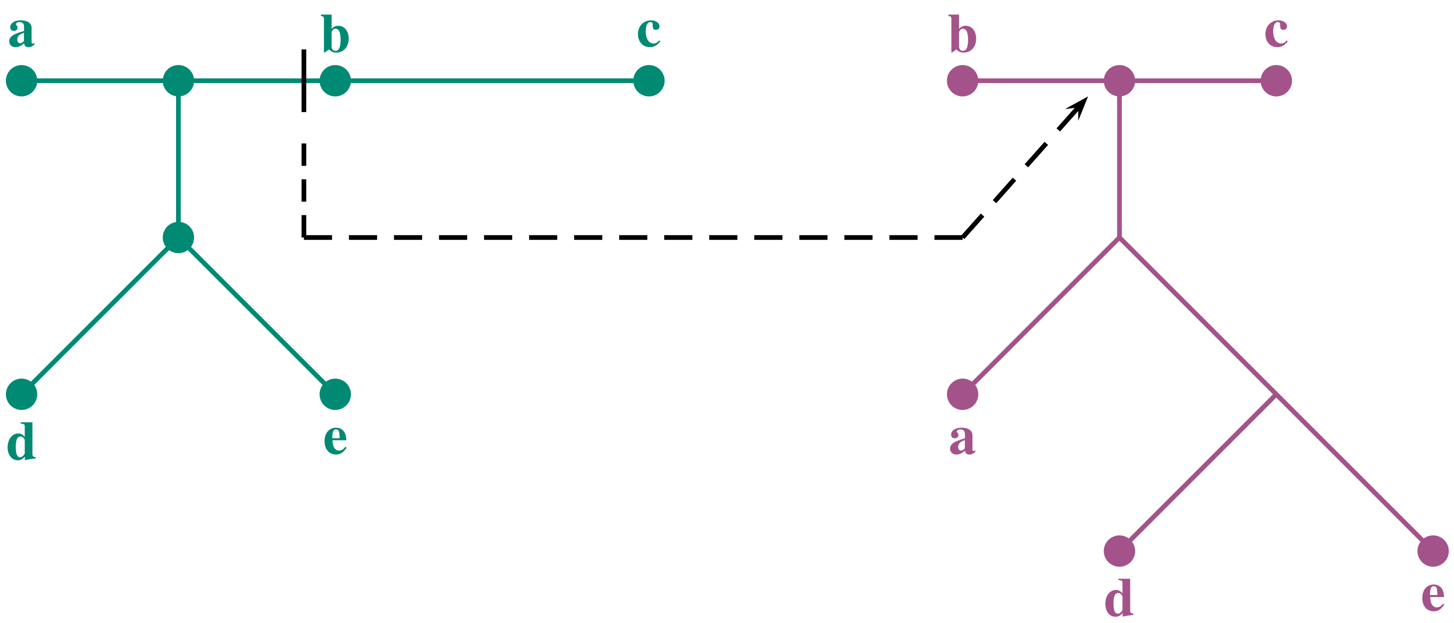
There are other genetic transfer distances between two trees

Nearest Neighbour Interchange (NNI)



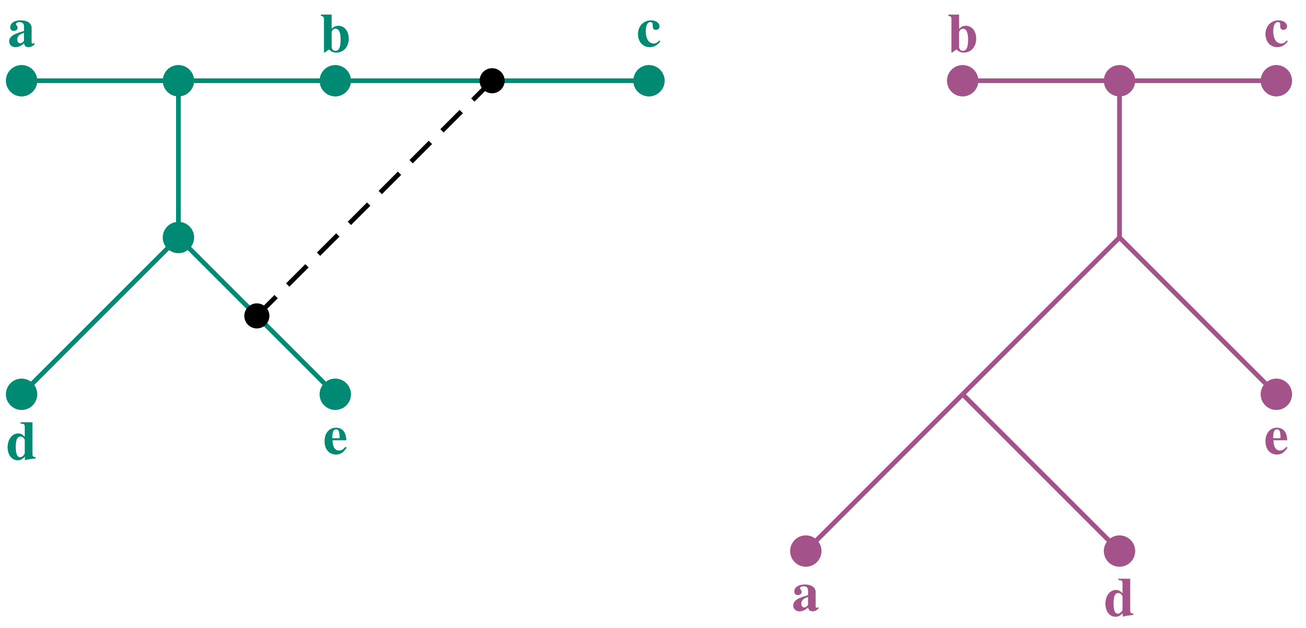
There are other genetic transfer distances (continuation)

Subtree Prune and Regraft (SPR)



There are other genetic transfer distances (continuation)

Tree Bisection and Reconnection (TBR)



a SPR operation is always a TBR but not conversely

Relation between the various distances

For any two binary trees T_1 and T_2 labelled at the leaves

$$D_{TBR}(T_1,T_2) \leq D_{SPR}(T_1,T_2) \leq D_{NNI}(T_1,T_2)$$

$$D_{SPR}(T_1,T_2) \leq 2 \times D_{TBR}(T_1,T_2)$$

$$MAF(T_1,T_2) - 1 = D_{TBR}(T_1,T_2)$$

(B. Allen and M. Steel, 2000)

What we have proven already

$$D_{SPR}(T_1,T_2) = \quad MAF(T_1,T_2)$$

or $MAF(T_1,T_2) - 1$

Our conjecture

this depends on whether the root is preserved or not

What are some of our perspectives concerning this ?

Prove the conjecture and improve the bounds

Obtain a better than 3-approximation for the MAF

Examine the "fixed parameter tractability" of each distance

> **Ex: When the maximum distance d between the two trees is (small), can we devise an algorithm whose complexity is $O(n^{O(1)}f(d))$ where n is the number of leaves and $f(d)$ can be any function (even exponential, but the exponential function involves only d) ?**

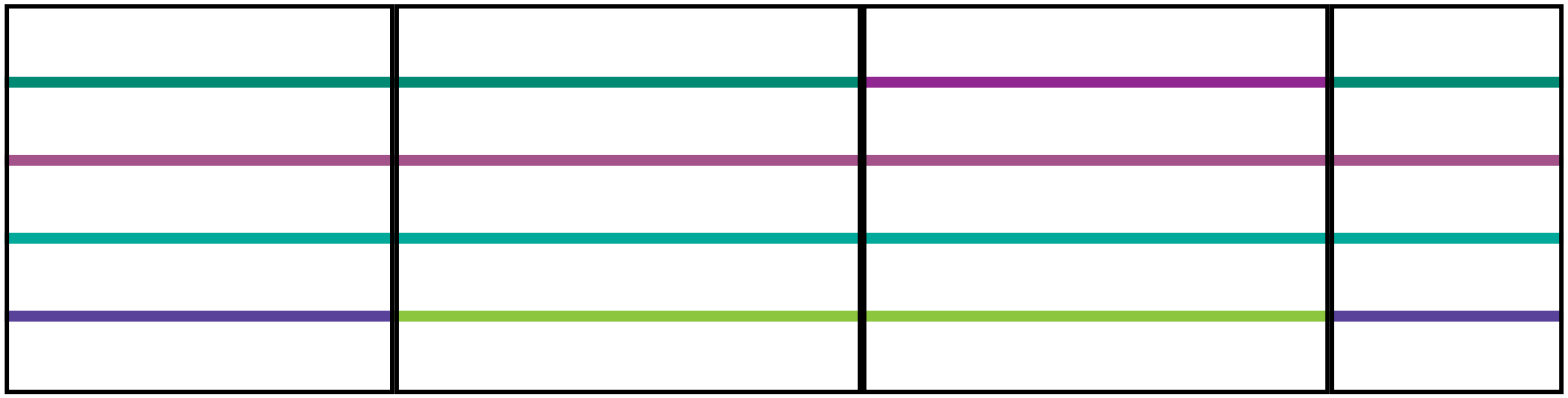
Examine further what is (perhaps ?) the most interesting genetic transfer distance: weighted SPR

The most difficult ? $\rightarrow$ go back to where we came from

Genetic transfer detection and localization as an optimisation problem

The (ideal ?) result

alignment



tree 1 **tree 2** **tree 3** **tree 4**

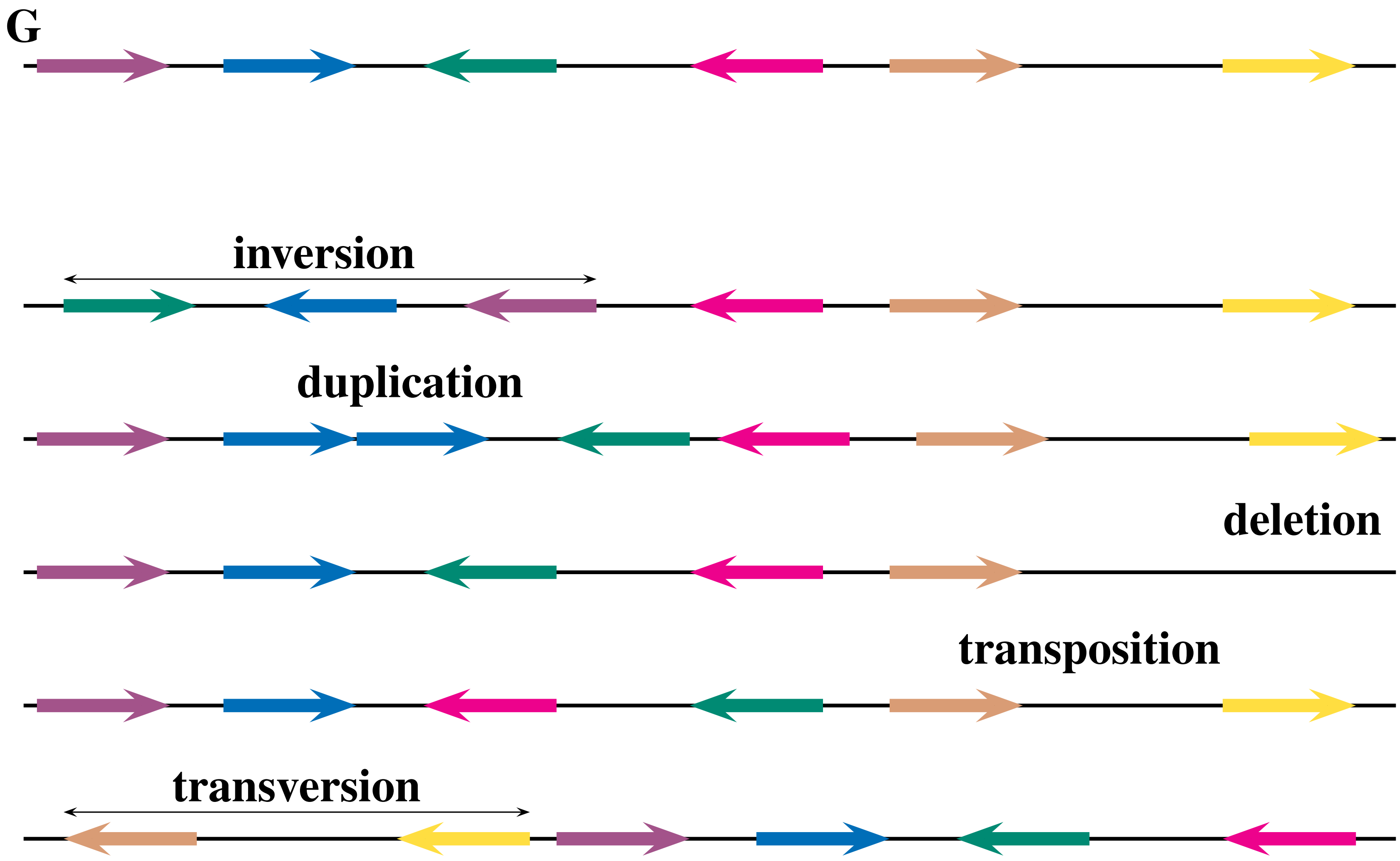
"optimal" partition: sum of lengths of trees + penalty for genetic transfers is MINIMUM

Third problem:

REARRANGEMENT DISTANCE BETWEEN GENOMES

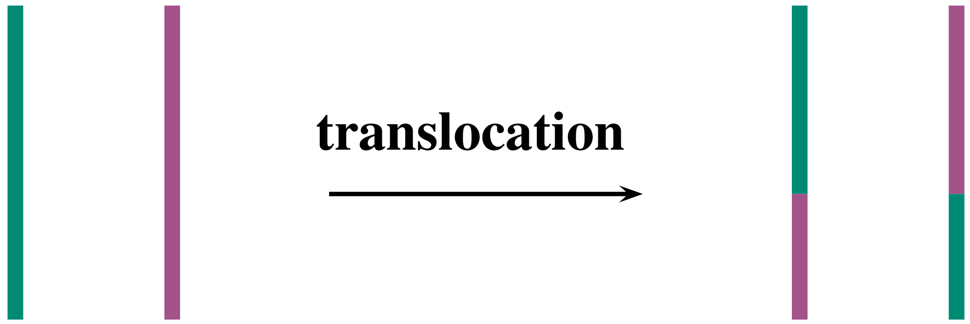
A rapid word on rearrangements

here, the “considered unit” is a gene but it could be something else



And, in the case of multichromosomal genomes

here, “considered unit” = gene



The most usual (but biologically simplified) formulations (1)

Permutation

INPUT:

data: a permutation of $[1, n]$

allowed operation: inversion

OUTPUT:

the minimum number of operations needed to obtain the identity permutation

Known complexity results:

- **NP-hard**
- **best approximation algorithm: ratio $\frac{3}{2}$**
- **can it be parameterized ?**

(Kececioglu and Sankoff; Hannenhali and Pevzner)

The most usual (but biologically simplified) formulations (2)

“Signed” permutation

INPUT:

data: a permutation of $[1, n]$ with some of the elements signed plus and others signed minus

allowed operation: inversion

OUTPUT:

the minimum number of operations needed to obtain the identity permutation with all elements signed plus

Known complexity results:

- **polynomial**
- **best algorithm is quadratic but complicated**
- **simplest algorithm is cubic; can it be improved ?**

(Kececioglu *et al.*; Hannenhali *et al.*; Kaplan *et al.*; Moret *et al.*)

The most usual (but biologically simplified) formulations (3)

Permutation

INPUT:

data: a permutation of $[1, n]$

allowed operations: inversion and transposition between adjacent elements

OUTPUT:

the minimum number of operations needed to obtain the identity permutation with all elements signed plus

Known complexity results:

- **unknown (same for transpositions alone), diameter unknown**
- **best approximation algorithm for transpositions: ratio** $\frac{3}{2}$
- **sorting by "block-interchange" is NP-hard**

(Bafna and Pevzner; Christie and Irving)

The most usual (but biologically simplified) formulations (4)

Partition

INPUT:

data: a partition of $[1,n]$ into k sets

allowed operations: fission, fusion and transposition

OUTPUT:

the minimum number of operations needed to obtain a partition into singletons

Known complexity results:

- **NP-hard**
- **best approximation algorithm: ratio 2**

(B DasGupta *et al.* 1997; D. Liben-Nowell *et al.* 1999 and 2000; N. Pisanti *et al.*, *Algorithmica* 34:157-180, 2000-2002)

The most usual (but biologically simplified) formulations (5)

Partition

INPUT:

data: two partitions of $[1, n]$ into k and l sets, the elements in each set are ordered and signed (plus or minus)

allowed operations: fusion, fission, inversion and tail-transposition

OUTPUT:

the minimum number of operations needed to obtain the same partitions

Known complexity results:

- polynomial

(Hannenhali and Pevzner; Tesler and Pevzner; Ozery-Flato and Shamir)

Some of the (many) missing things

Theoretical and algorithmical results on the above formulations

Formulations that may give different weights to different operations and that include also duplications, deletions and insertions (*i.e.* we do not have permutations and partitions *stricto sensu*)

Satisfactorily addressing the problem of "segments (approximately) conserved by rearrangements" (formulation(s), algorithms, theoretical results – although some exist already)

But all this could be the subject of (at least) another full talk, without even mentioning many other subjects closely related to rearrangements

<table>
<tr><th colspan="2">Lyon</th></tr>
<tr><td>Eric Tannier</td><td>Raquel Tavares
Marina Zelwer</td></tr>
<tr><th colspan="2">Paris (with Maxime Crochemore)</th></tr>
<tr><td>Julien Allali
Philippe Blayo
Benoît Olivieri</td><td>Gérald Balayssac
Laurent Marsan
Pierre Peterlongo</td></tr>
<tr><th colspan="2">Canada</th></tr>
<tr><td colspan="2">Sébastien Provencher</td></tr>
<tr><th>Italy</th><th>Portugal</th></tr>
<tr><td>Nadia Pisanti</td><td>Susana Vinga</td></tr>
<tr><th colspan="2">Brazil</th></tr>
<tr><td>Said Sadique Adi
Jeane B. B. de Melo</td><td>Paulo Gustavo Fonseca
Estela M. Rodrigues</td></tr>
</table>

And many collaborators...