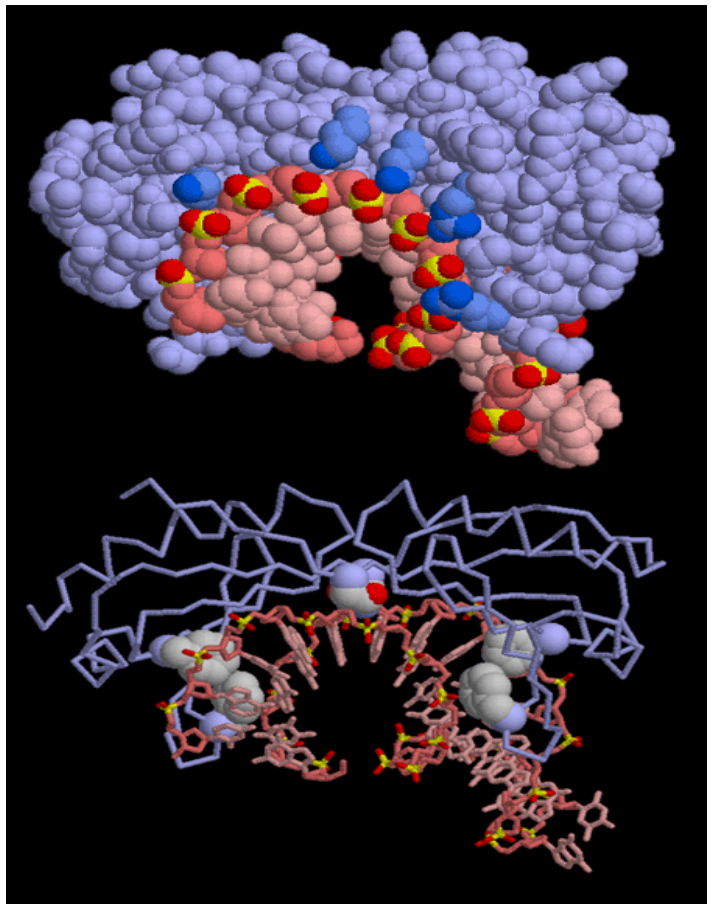
Introduction à la découverte de motifs en biologie moléculaire



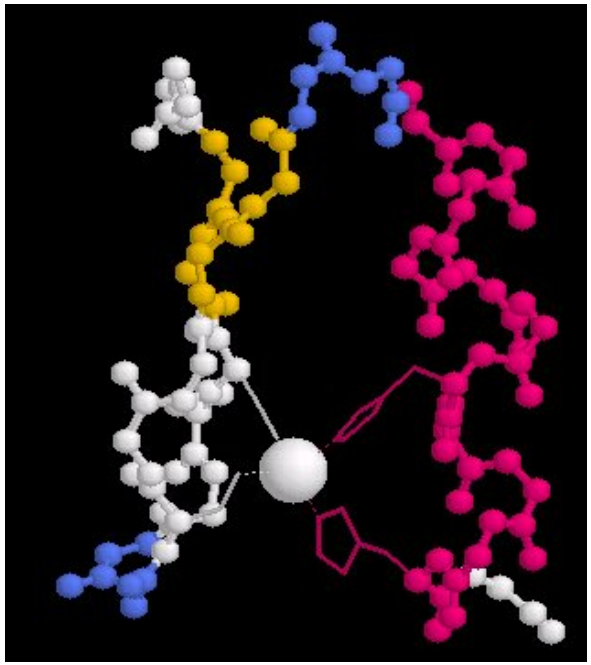
François Coste, Nov. 2005, École chercheurs BioInfo

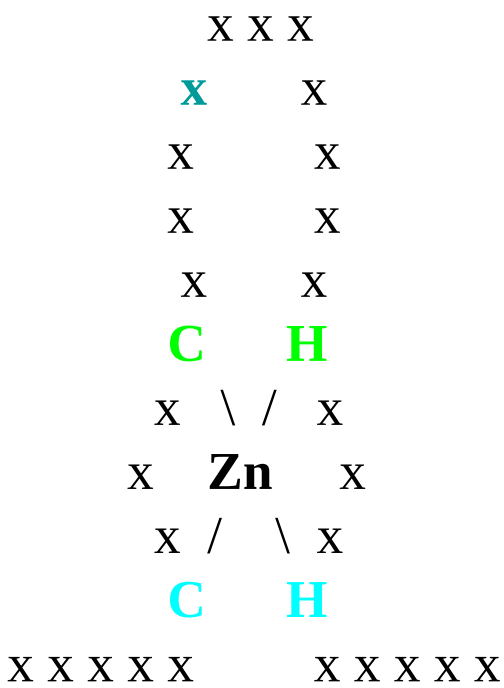
Caractérisation d'une famille de protéines

Exple : Protéines « en doigt de zinc »

Z	...YLGPLNCKSCWQKFDSFSKCHDHYLCRHCLNLLL...
ZFH2	...ILMCFICKLSFGNVKSFSLHANTEHRLNL...
ZNF236	...HKCEICLLSFPKESQFQRHMRDHE...

C-x(2,4)-**C**-x(3)-[LIVMFYWC]-x(8)-**H**-x(3,5)-**H**





Motif C2H2 pour la famille Zinc finger

C-x(2,4)-**C**-x(3)-[LIVMFYWC]-x(8)-**H**-x(3,5)-**H**

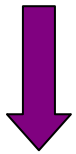
- 416 séquences Zinc finger
- motif C2H2 contenu dans :
  - 372 Zinc finger
  - 34 protéines non Zinc finger
  - 6 protéines candidates Zinc finger

Pattern discovery overview

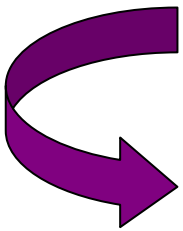
- Pattern Discovery consists to build a model (motif) of a family from a set of sequences of this family;
- Such a model may be either characteristic (one looks for a definition of the set of sequences) or discriminant (one looks for a difference between two sets of sequences);
- In all cases, the motif has a predictive value and may be checked against new sequences with a pattern matching algorithm.
- Pattern discovery may be applied either to nucleic or amino-acids sequences, generally with different algorithms.

Découverte de motifs

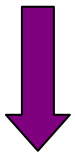
Famille : Set de séquences ' reliées '
(fonction/structure identique)



Contraintes de ' conservation ' sur
les zones impliquées



Identification d'un pattern associé
à une fonction biologique



- Annotation des génomes
- Caractérisation de familles fonctionnelles
- Recherche de nouvelles séquences

Usefulness of Patterns in Genomics

DNA / RNA : regulation, genes, diseases…

- Transcription Factors;
- Site of fixation of sigma factors;
- Regulation patterns specific of a tissue or a development stage;
- Transcription Terminators;
- Frameshift;
- Repeats (tandem, inverted…).

Usefulness of Patterns in Proteomics

Proteins : Function, activity, localization, alignment…

- Patterns of intra-molecular links;
- Patterns of interaction protein/ ion, DNA, Protein;
- Pattern specific of tissues or addressing inside a cell;
- Signatures of functions
- Signatures of structures.

Specific signatures of families : comparison is not sufficient…

- Two situations where motifs are necessary
  1. Annotation of a new sequence;
  2. Search for candidates of a family of interest.

- In both cases, people generally use various versions of Blast. It is not the most efficient way to look at motifs since some positions are far more important than others and corresponding positions may even change from one sequence to another one.
  Multiple alignements, for instance with ClustalW is a better solution when possible but suffers basically from the same drawbacks.

Pattern discovery methods : a bench of algorithms…

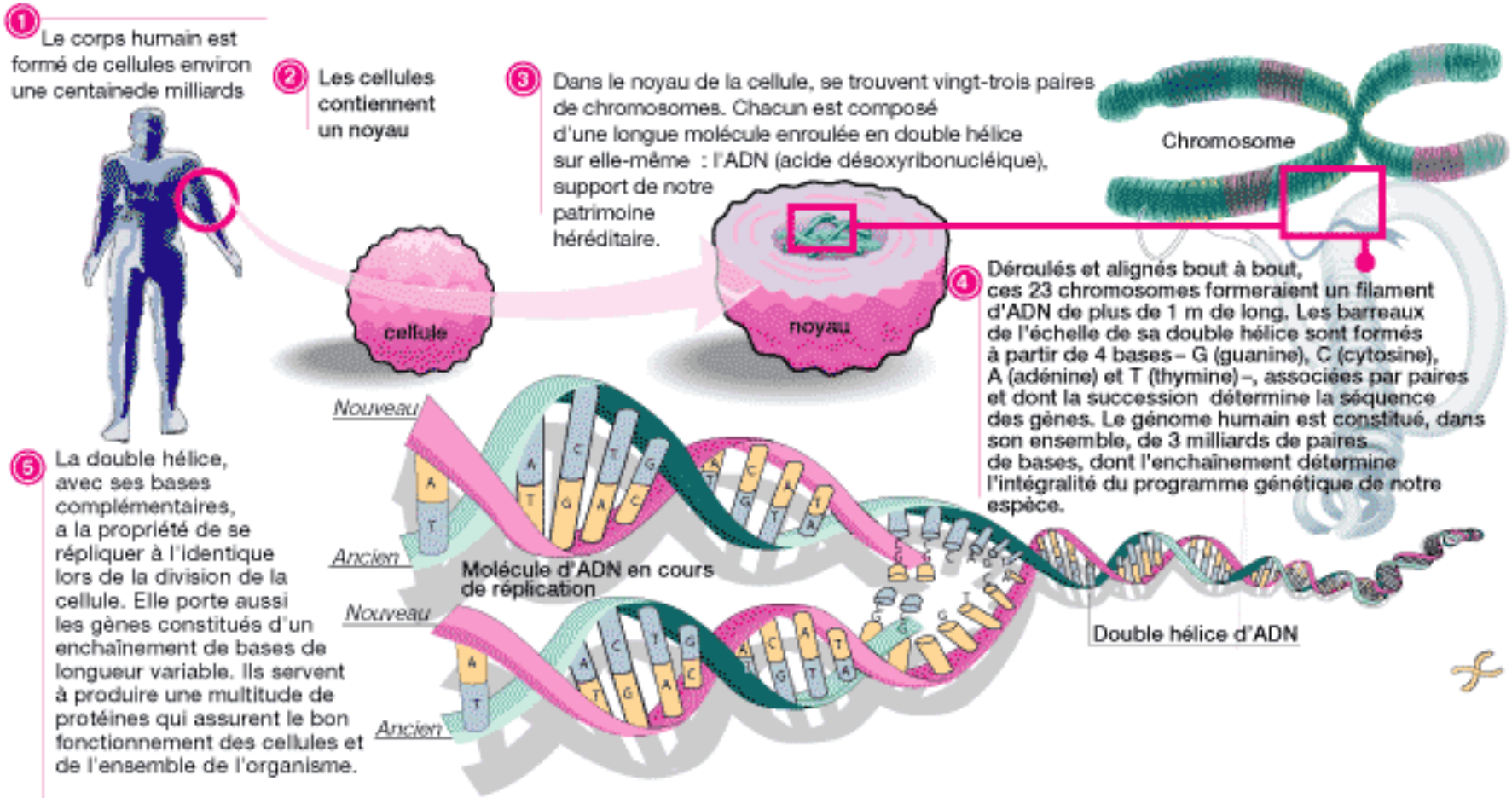
MEME	EM	1994
MACAW	Gibbs	1994
CoResearch	Enum/EM	1996
R'MES	Markov?	1997
AlignACE	Gibbs	1998
TEIRESIAS	Cliques	1998
Yebis	Markov	1998
CONSENSUS	Enum	1999
Winnower	Cliques	2000
SP-STAR	Cliques	2000
Ann-Spec	ANN	2000
SMILE	Suffixtrees	2000
SMILE (dyads)	Suffixtrees	2000
Verbumculus	Suffixtrees	2000
MobyDick	Dictionary	2000
Dyad and Oligo-Analysis	Enum	2000
YMF	Enum	2000
Kimono	Gibbs/Express	2000
BioProspector	Gibbs	2001
Co-Bind		2001
ITB	Enum	2001
(Barash et al)	EM	2001
Mitra	Cliques	2002
MultiProfiler	Cliques	2002
Spexs	Suffixtrees	2002
MotifSampler	Gibbs	2002
SeSiMCMC	Gibbs	2002
AHAB	Dictionary	2002
Projection	Projection	2002
Footprinter	Enum/Phylo	2002
Improbizer	EM	2002
PhyloCon	?/Phylo	2002
MDScan	Enum	2002
FindModels	SuffixTrees	2002
PROCSE	Clustering/Phyl	2002
Mitra-PSSM	Cliques	2003
IRSA	Cliques	2003
Gibbs Recursive Sample	Gibbs	2003
cWinnower	Cliques	2003
YMF3	Enum	2003
REDUCE	Enum/Express	2003
LOGOS	Dictionary	2003
SDDA	Dictionary	2003
MotifRegressor	MDScan/Expres	2003
BMC	Gibbs	2003
MERMAID	Enum	2003
MOPAC	Enum	2003
(Mwangi et al)	Enum	2003
Stars	Comparison	2003

De l'ADN aux protéines

La régulation des gènes,
vue par un informaticien…

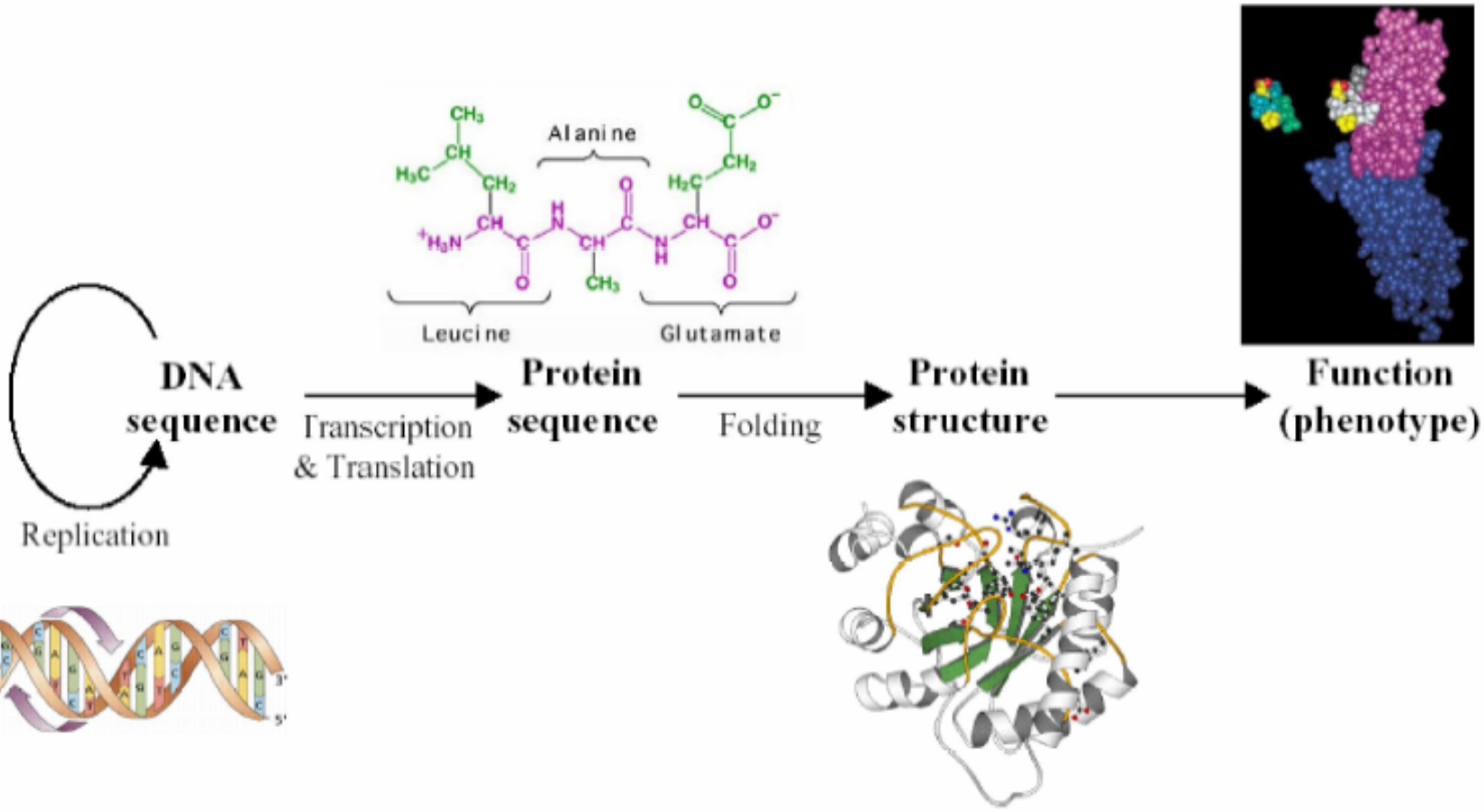
ADN

Le « livre de la vie » : de la cellule au chromosome



Source : Le Monde ; Infographie Le Monde avec G.N/A.F.P.

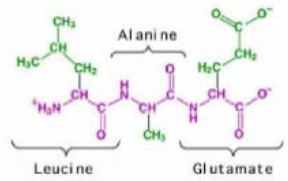
De l'ADN à la fonction

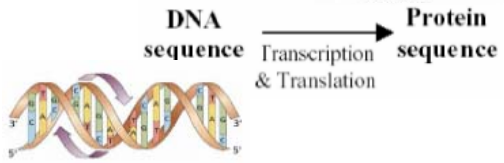


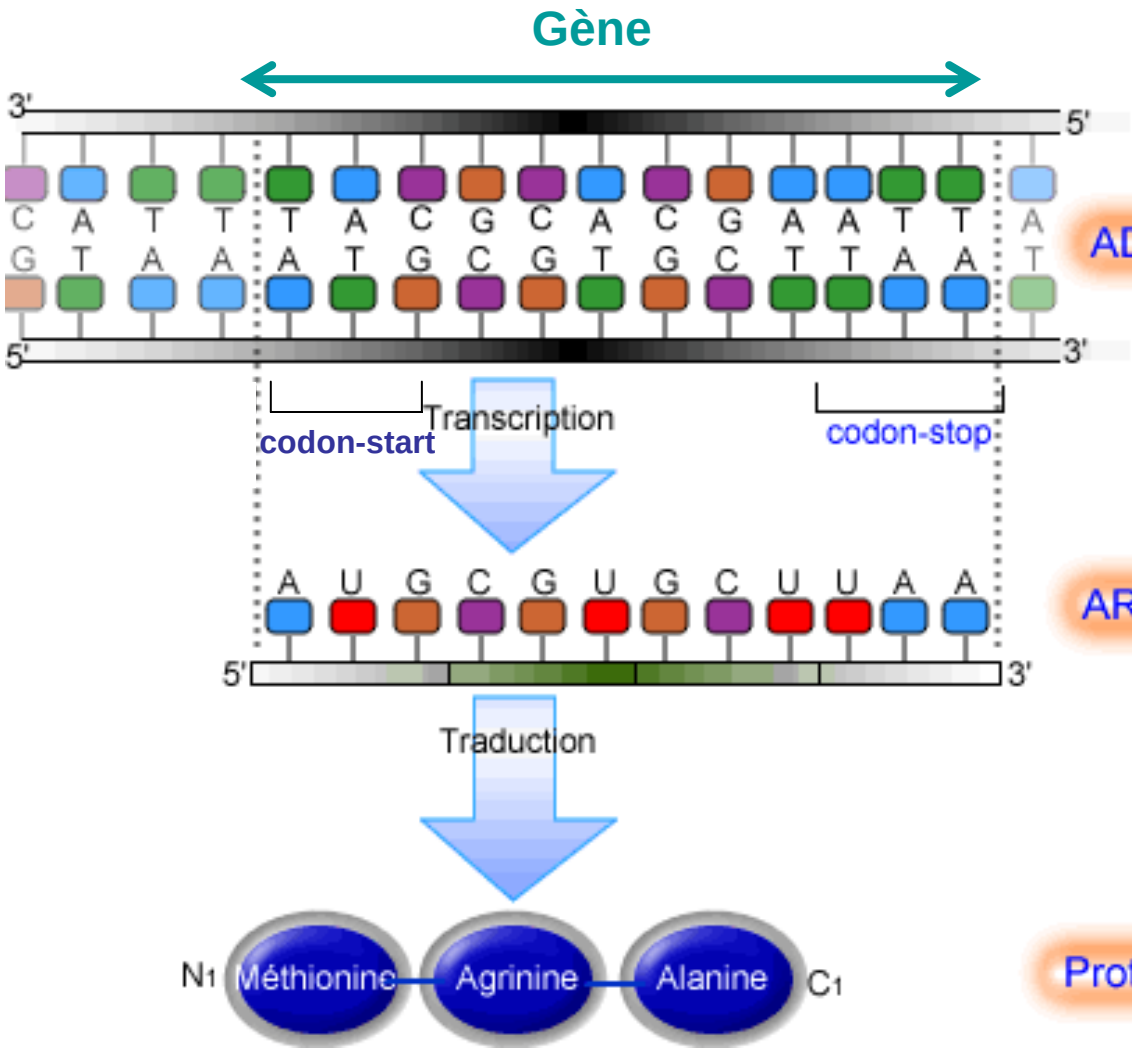
Source: JC Latombe et al…

Code source —— Compilation ——→ Exécutable

De l'ADN à la protéine



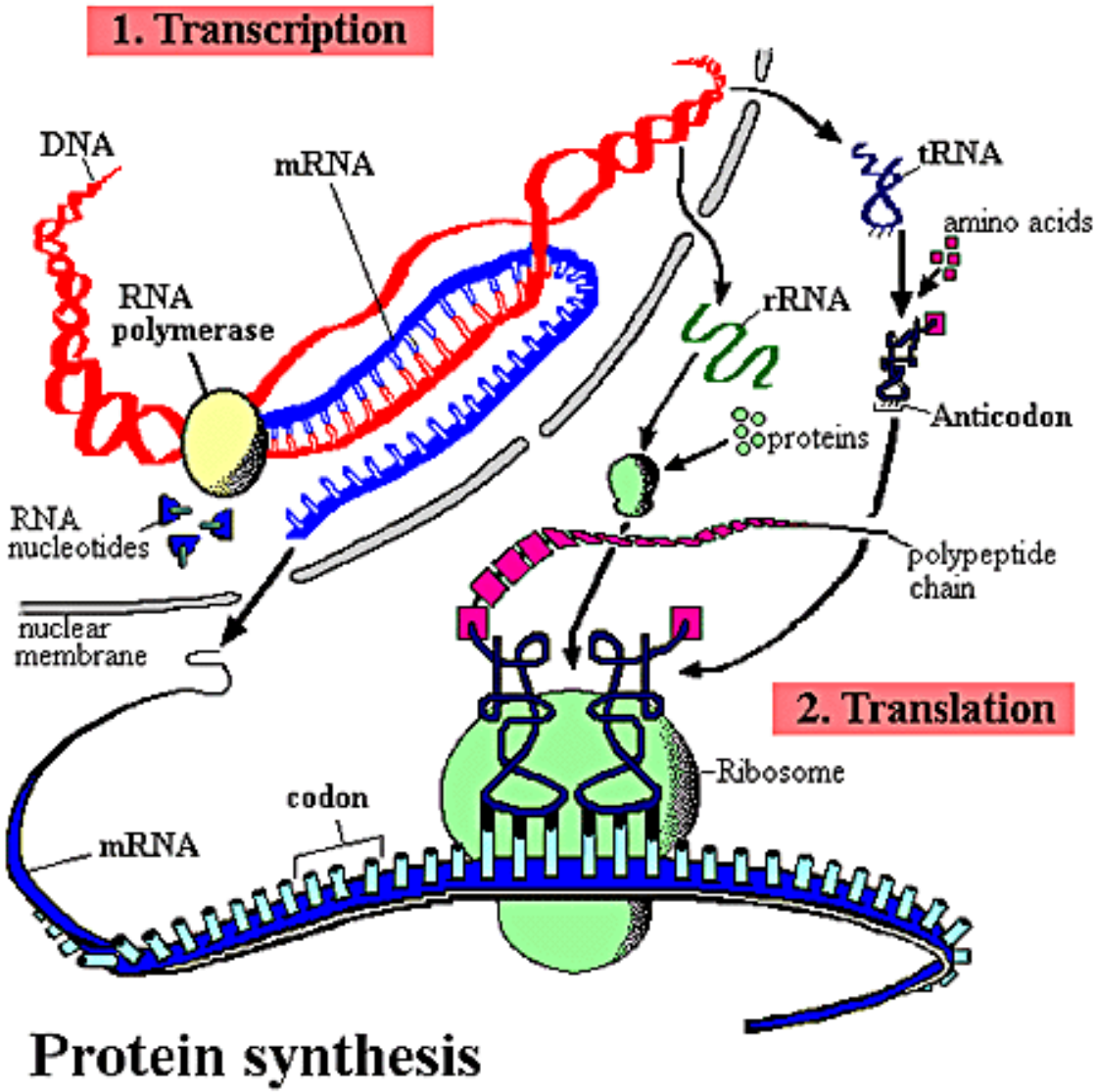




D'après Guillaume Bokiau

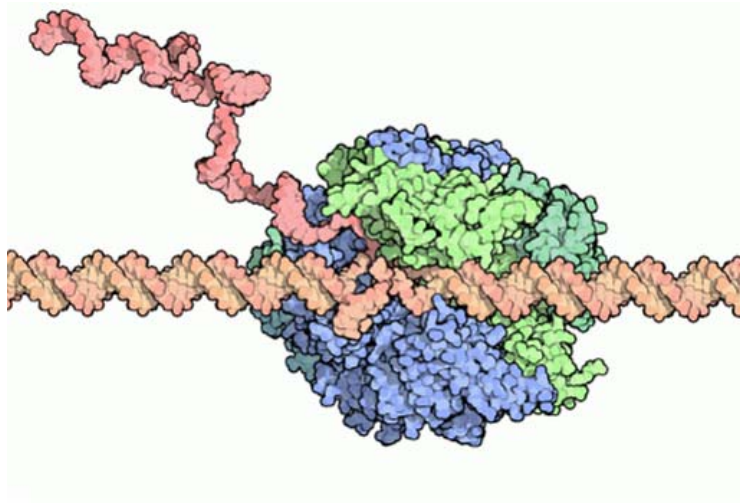
Synthèse des protéines

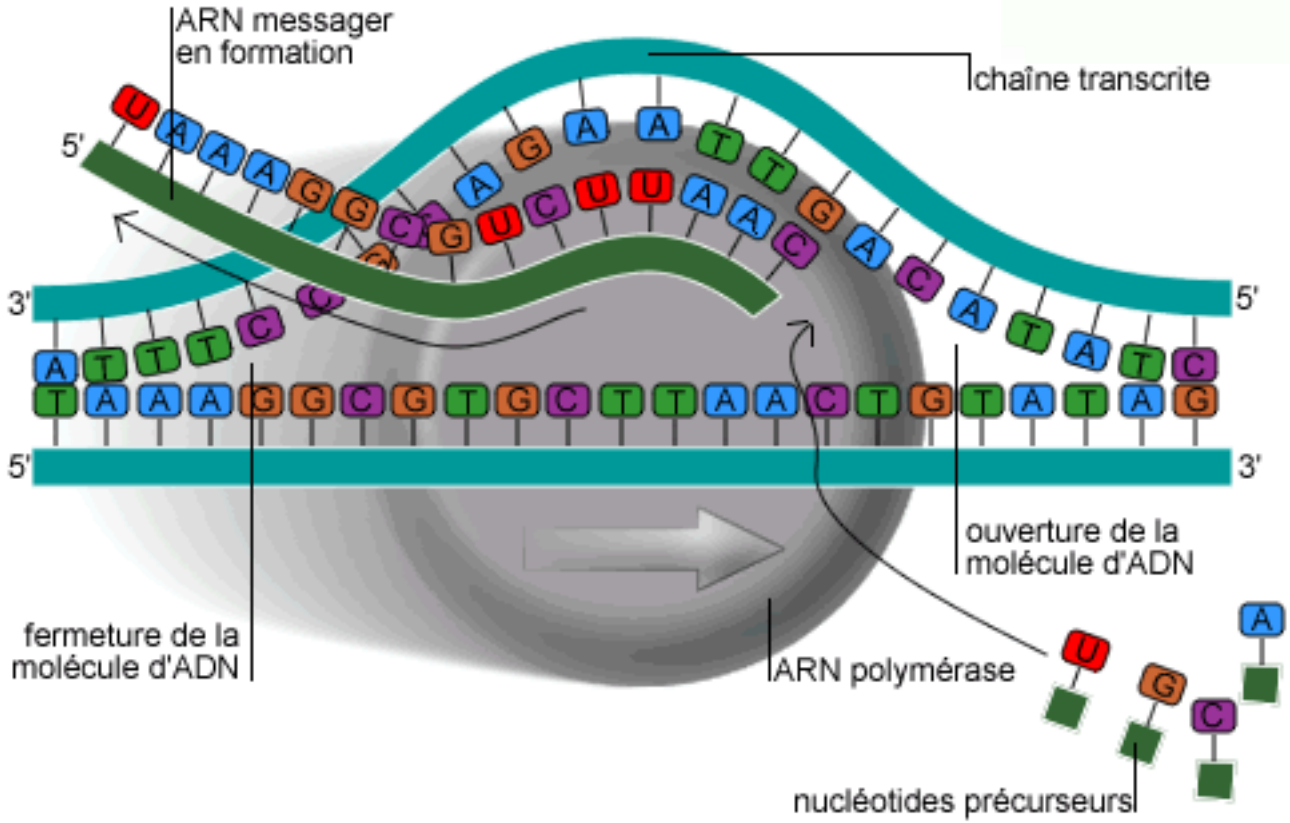
Alphabet des acides aminés : 20 lettres



ARN Polymérase

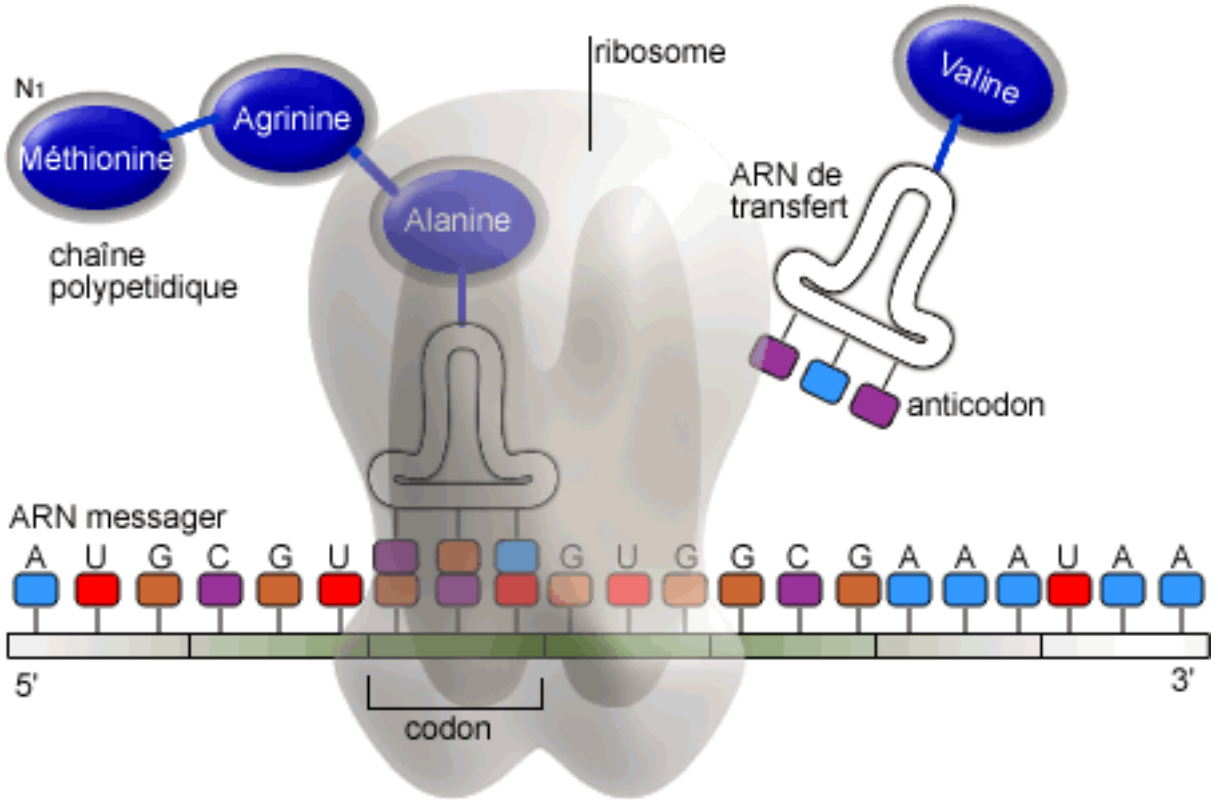
Transcription de l'ADN en ARN





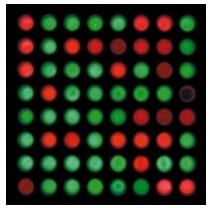
Ribosome

Traduction de l'ARN en protéine



Source : Guillaume Bokiau

Comment est régulée l'expression des gènes ?



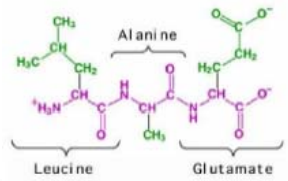
ARN-Polymerase : ADN →ARN (→Protein)

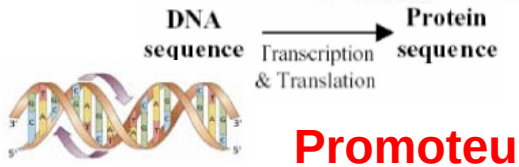
ADN : ~30000 gènes

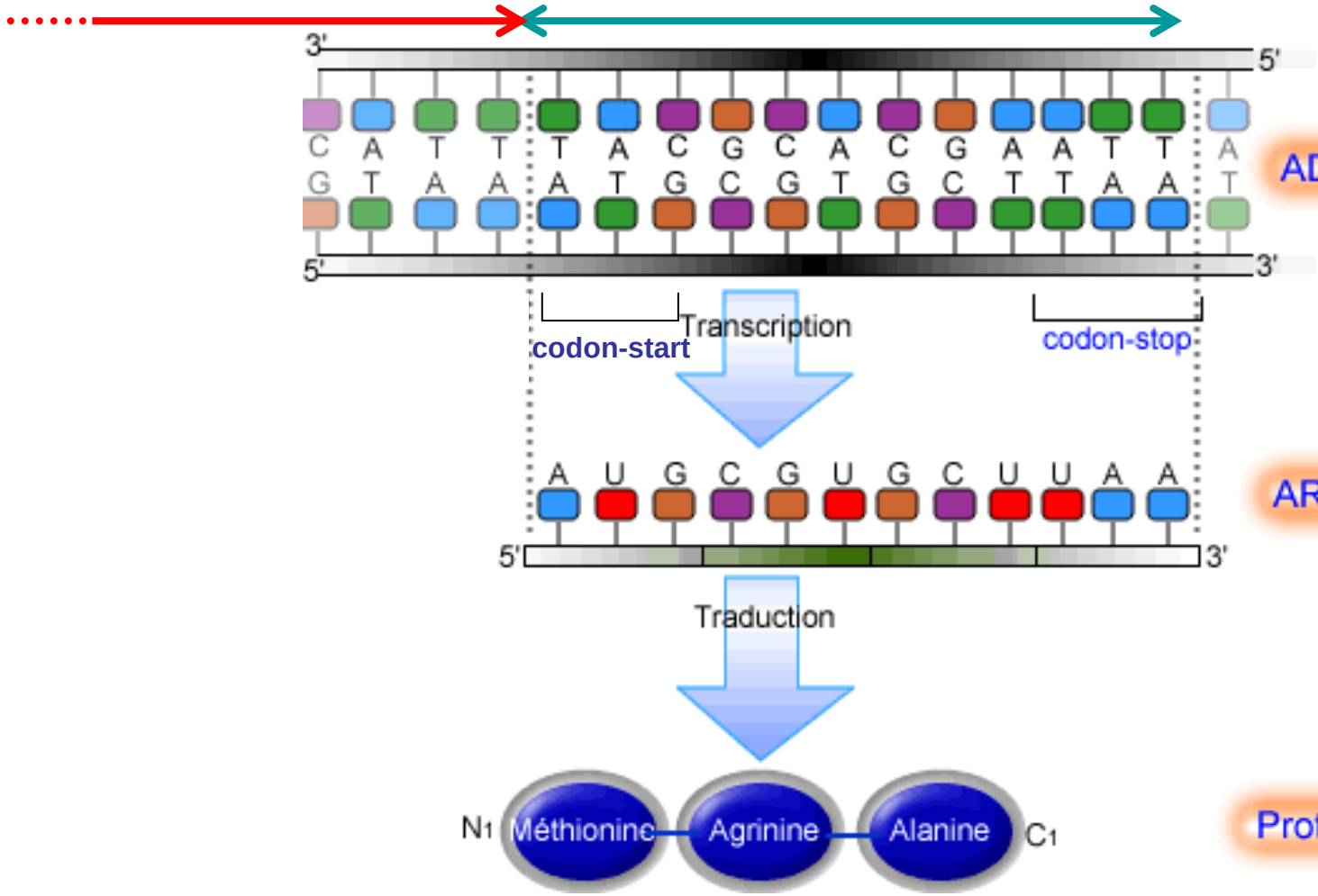
- Comment repérer les gènes ?
- Quand activer un gène ?
- Combien de copies ?
- Où ? (dans quelle cellule ?)

Transcription au bon moment et au bon endroit….

De l'ADN à la protéine

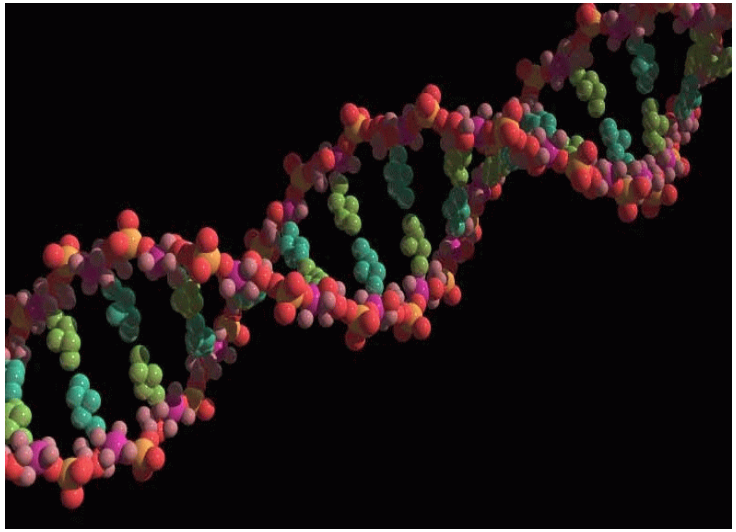


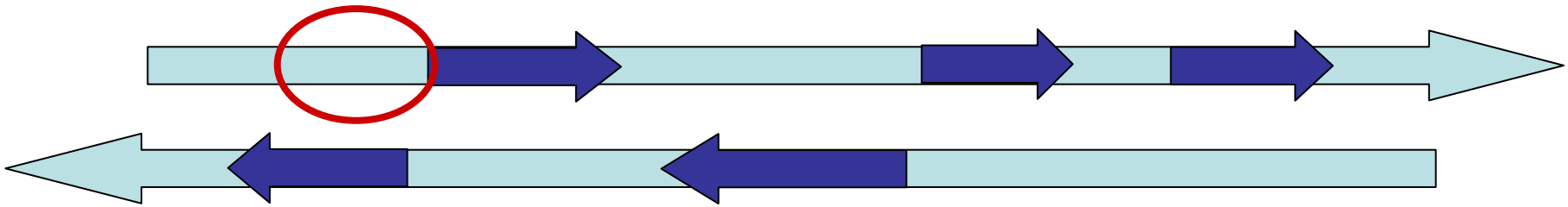




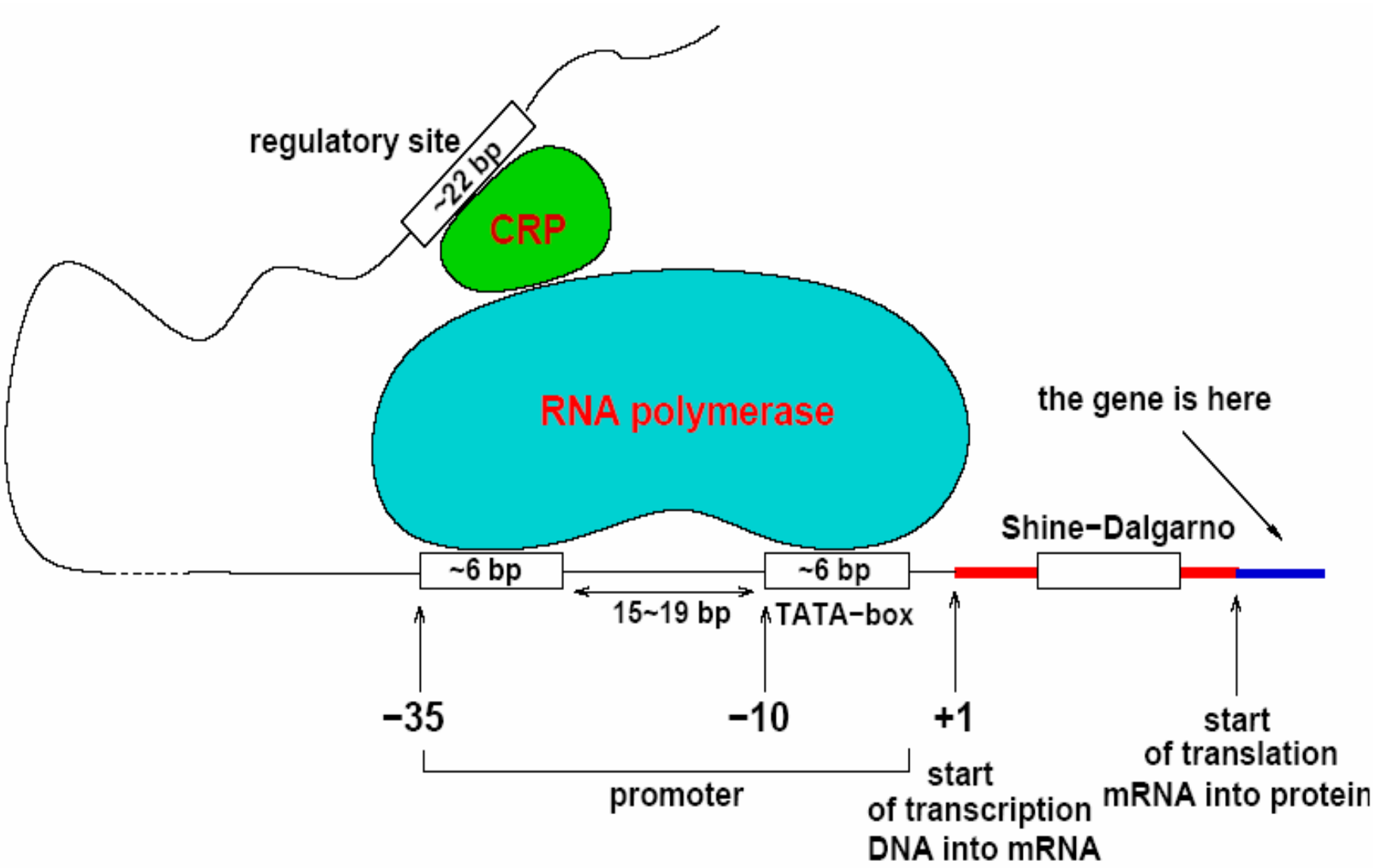
D'après Guillaume Bokiau

Gènes



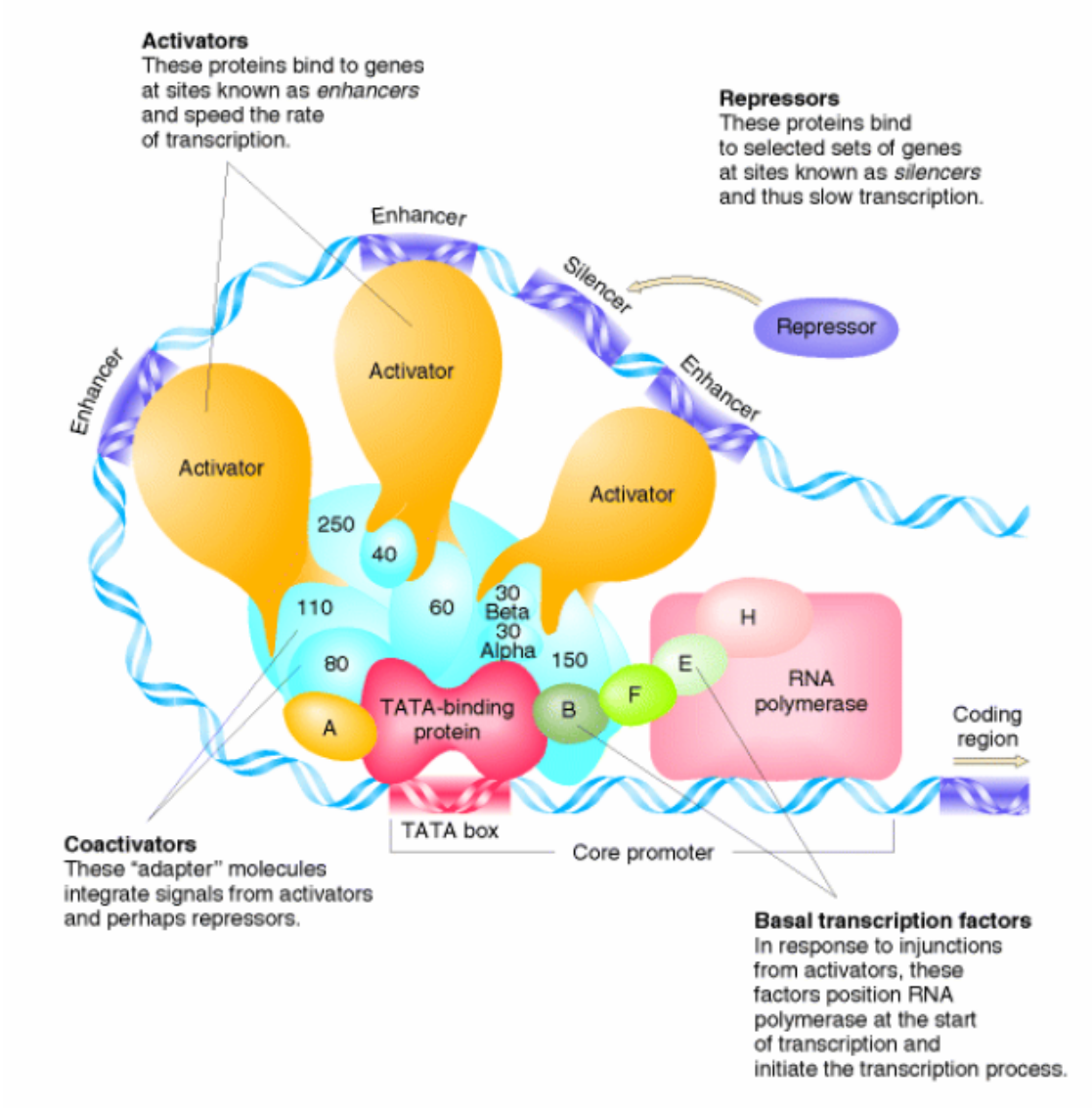


Initiation de la transcription (procaryotes)



Source Marie-France Sagot

Initiation de la transcription (eucaryotes)

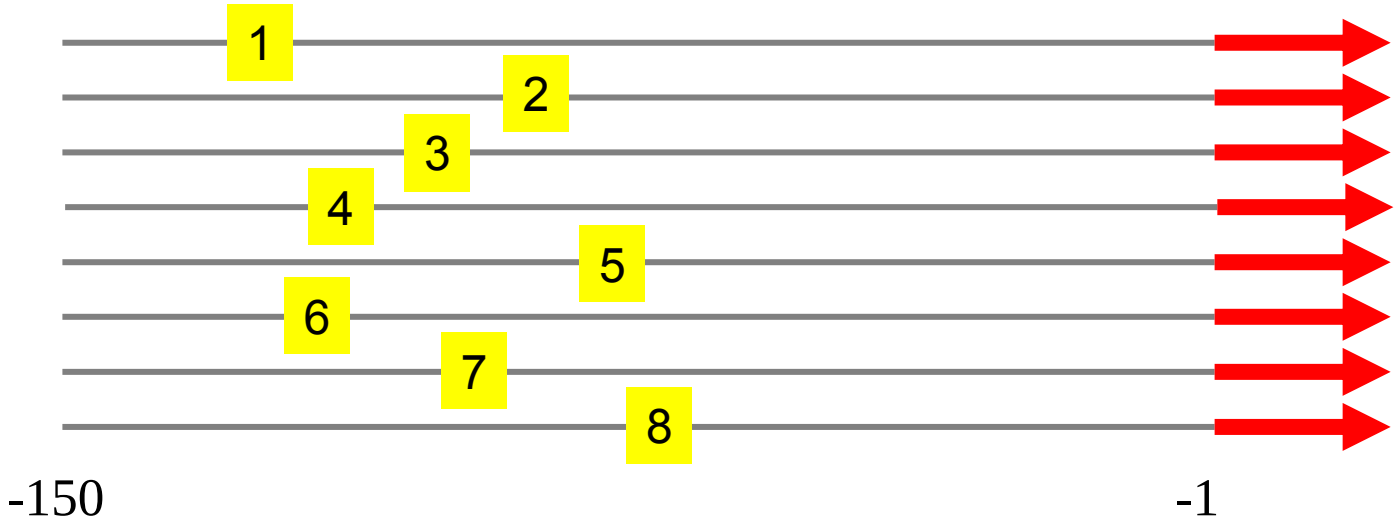


Source : Marie-France Sagot

TATA-Box (Pribnow-Box) : TATAAT

- Incluse dans :
  - 20% des promoteurs de la levure
  - 30% " " de l'humain
- D'autres types de sites de fixations :
  - GC-Box, CAAT-Box,…
- Tolérance à certaines mutations :
  - TATTAT fonctionnel
    (avec un affaiblissement potentiel du signal)
  - TAATAAT non fonctionnel
  - TATAAT présent sans mutation dans 14 des 291 TATA-Box connues…

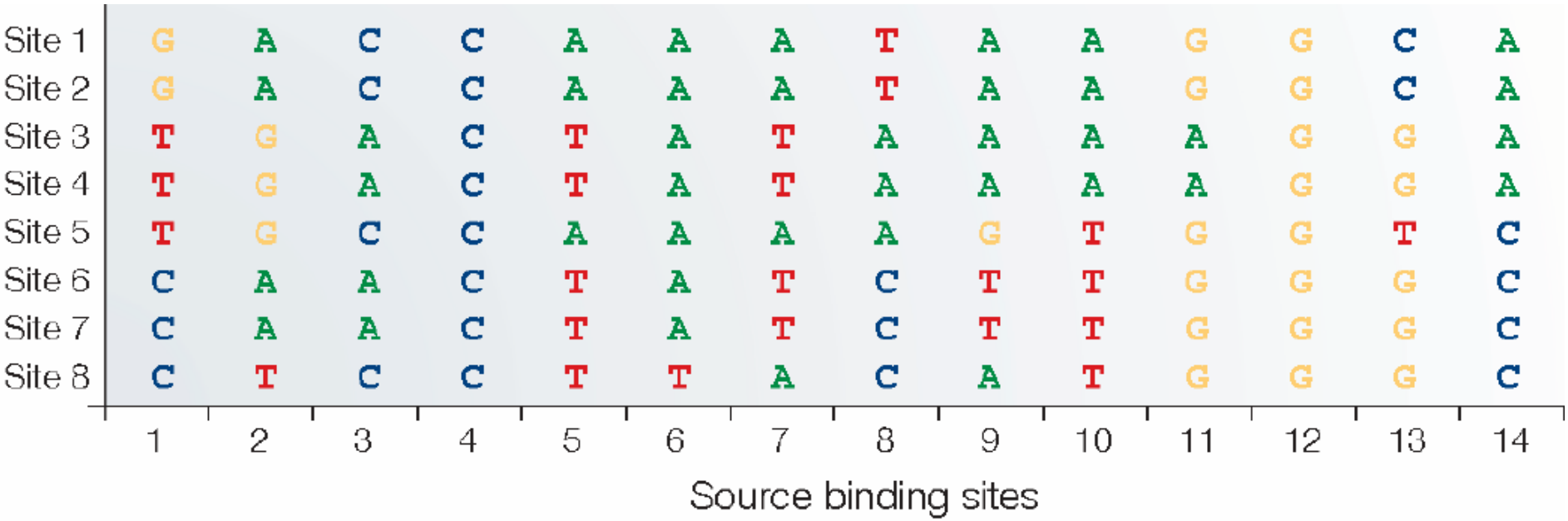
Exemples de sites de fixation



Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C
	1	2	3	4	5	6	7	8	9	10	11	12	13	14

Source binding sites

Motif consensus



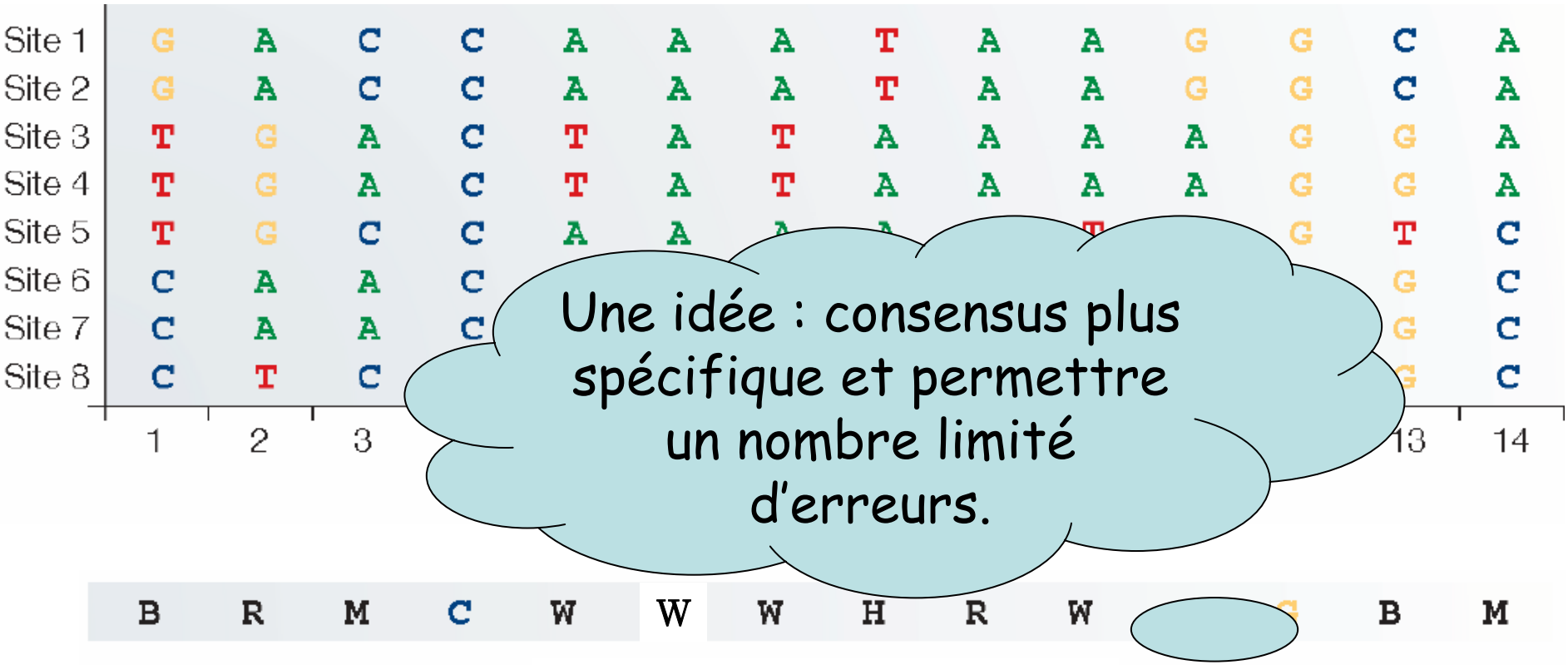
Motif (séquence) consensus

Utilisation du code IUPAC

[TCG] [ATG] [AC] C [AT] [AT] [AT] [ATC] [ATG] [AT] G G [TCG] [AC]

D'après Maximilian Häußler

GTACATTTGAAGTA vs TAACTATAATGGGA ?

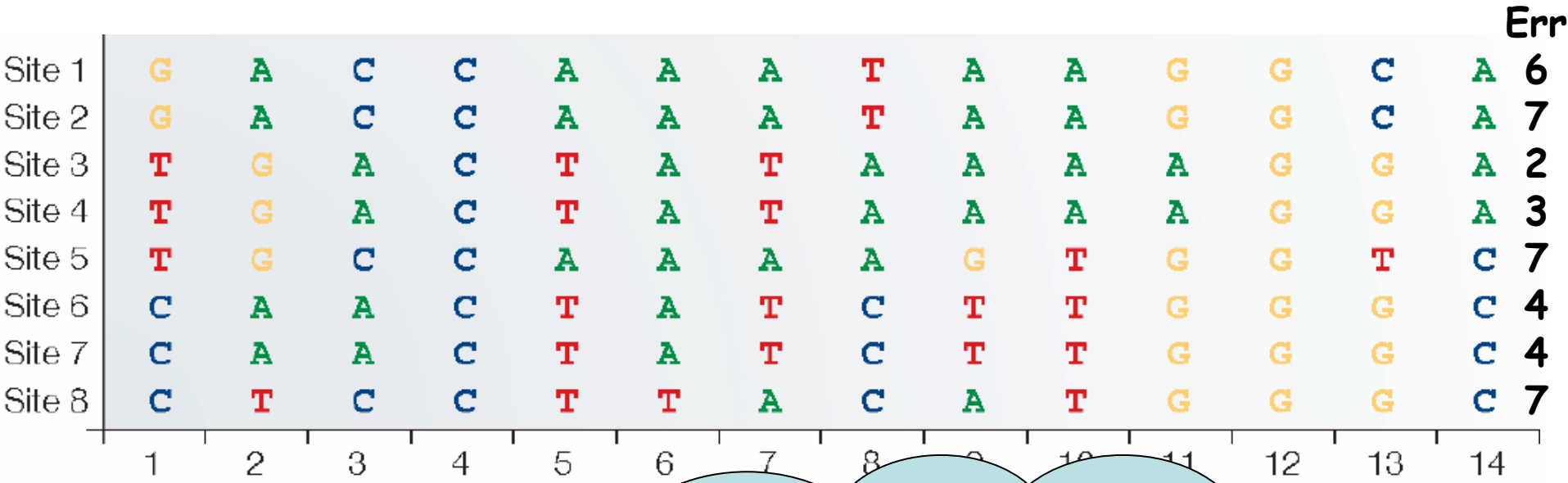


Motif (séquence) consensus
Utilisation du code IUPAC

[TCG] [ATG] [AC] C [AT] A [AT] [ATC] [ATG] [AT] G G [TCG] [AC]

D'après Maximilian Häußler

Séquence consensus



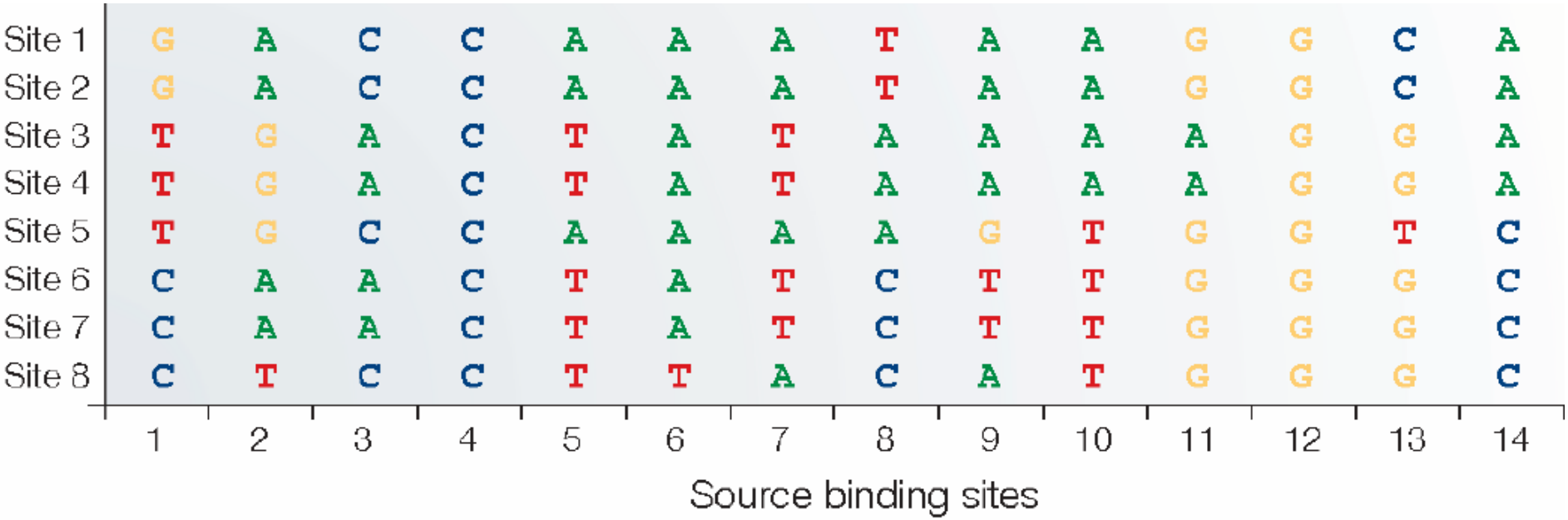
T A A G A

Mais:

Des positions supportent mieux les mutations que d'autres…

Il peut y avoir des préférences de mutations…

Position frequency matrix



Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

Source Maximilian Häußler

Probabilité d'une sous-séquence ?

Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

Exercices :

- Utiliser la PFM ci-dessus pour calculer :

  P(GTACATTTGAAGTA) = ?

  P(TAACTATAATGGGA) = ?

  P(AAACTATAATGGGA)= ?
- Comment savoir si une probabilité est significative ?
- Comment se comportent ces probabilités si la composition en nucléotides est biaisée ?

Position weights

- Probabilité du nucléotide b en la position i :

$$p(b,i) = \frac{f_{b,i} + s}{N + 4s}$$

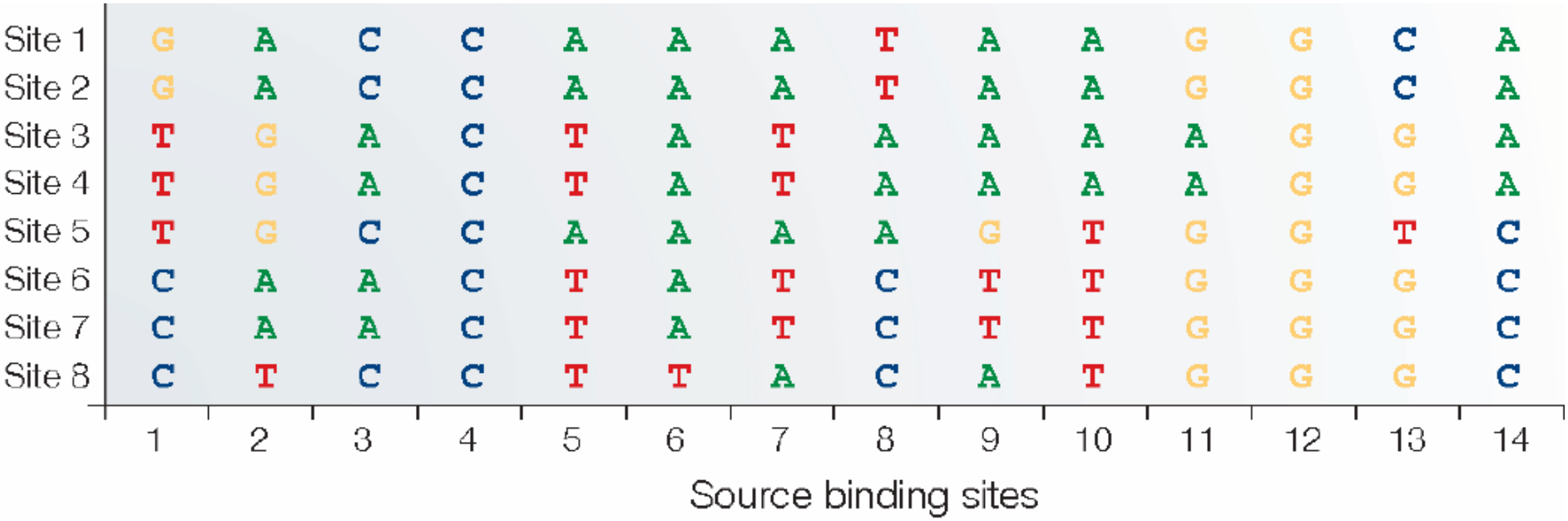
pseudo compte

- Poids du nucléotide b en la position i (*Log odds*):

$$W_{b,i} = \log \frac{p(b,i)}{p(b)}$$

background probability

Position weight matrix (PWM)



Position weight matrix (PWM)

A	−1.93	0.79	0.79	−1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	−1.93	−1.93	0.79
C	0.45	−1.93	0.79	1.68	−1.93	−1.93	−1.93	0.45	−1.93	−1.93	−1.93	−1.93	0.00	0.79
G	0.00	0.45	−1.93	−1.93	−1.93	−1.93	−1.93	−1.93	−0.66	−1.93	1.30	1.68	1.07	−1.93
T	0.15	−0.66	−1.93	−1.93	1.07	−0.66	0.79	0.00	0.00	0.79	−1.93	−1.93	−0.66	−1.93

$p(A)=p(T)=p(G)=p(C)=\frac{1}{4}$

Source Maximilian Häußler

Score d'un site

Position weight matrix (PWM)

A	–1.93	0.79	0.79	–1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	–1.93	–1.93	0.79
C	0.45	–1.93	0.79	1.68	–1.93	–1.93	–1.93	0.45	–1.93	–1.93	–1.93	–1.93	0.00	0.79
G	0.00	0.45	–1.93	–1.93	–1.93	–1.93	–1.93	–1.93	–0.66	–1.93	1.30	1.68	1.07	–1.93
T	0.15	–0.66	–1.93	–1.93	1.07	–0.66	0.79	0.00	0.00	0.79	–1.93	–1.93	–0.66	–1.93

Site scoring

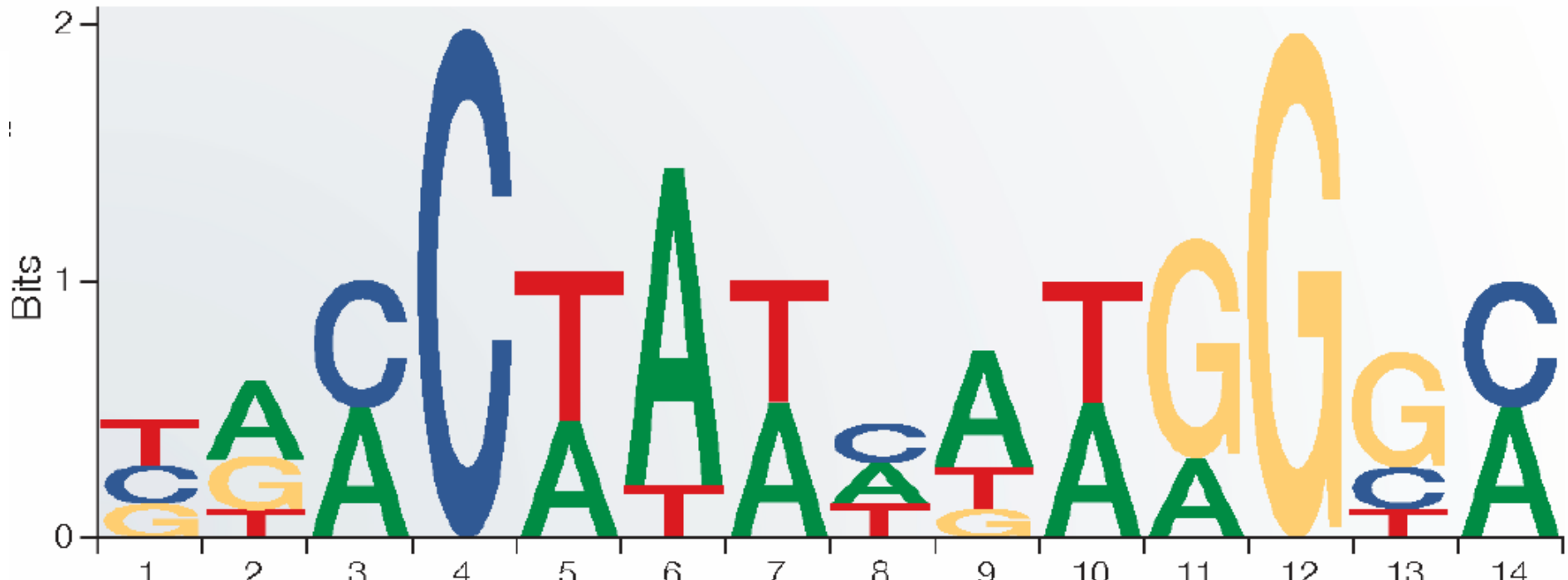
0.45	–0.66	0.79	1.68	0.45	–0.66	0.79	0.45	–0.66	0.79	0.00	1.68	–0.66	0.79
C	T	A	C	A	T	A	A	G	T	A	G	T	C

$\Sigma = 5.23$, 78% of maximum

Source Maximilian Häußler

Sequence Logo [Schneider]

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C



Entropie relative (information content)

- D'une position :

$$IC_{pos} = \sum_b f_b log_2 \frac{f_b}{p_b}$$

entre 0 et 2 bits (ADN, $\frac{1}{4}$)

- D'une matrice :

$$IC_{matrix} = \sum_{pos=1}^{len} IC_{pos}$$

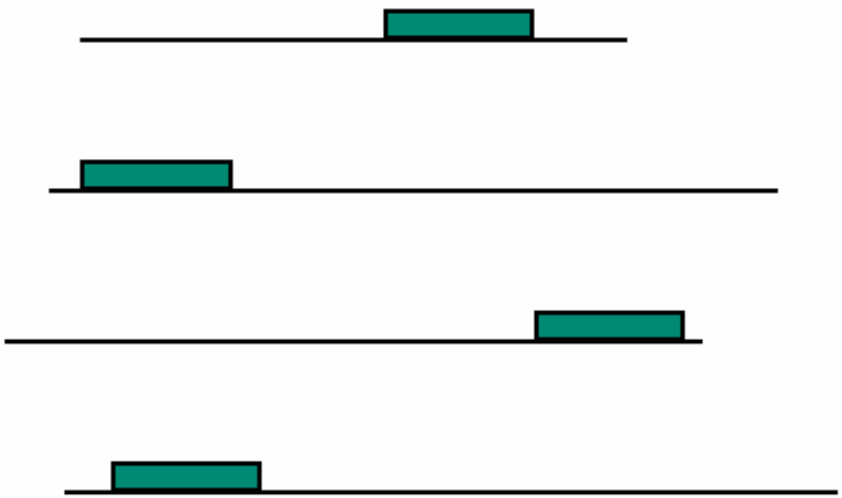
max = len $\times$ 2 (ADN , $\frac{1}{4}$)

Mesure de la conservation du motif

Entropie relative (information content)

$$\sum_{i=1}^{L} \sum_{\alpha \in \Sigma} f_{i\alpha} \log_2 \frac{f_{i\alpha}}{f_\alpha}$$

relative entropy



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

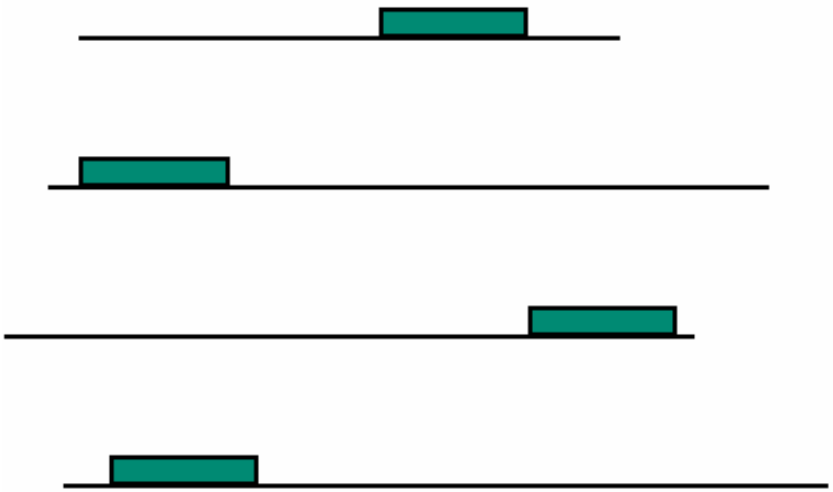
$f_\alpha = \frac{1}{4}$

0 = 0 0 0 0 0

Entropie relative (information content)

$$\sum_{i=1}^{L} \sum_{\alpha \in \Sigma} f_{i\alpha} \log_2 \frac{f_{i\alpha}}{f_\alpha}$$

relative entropy



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

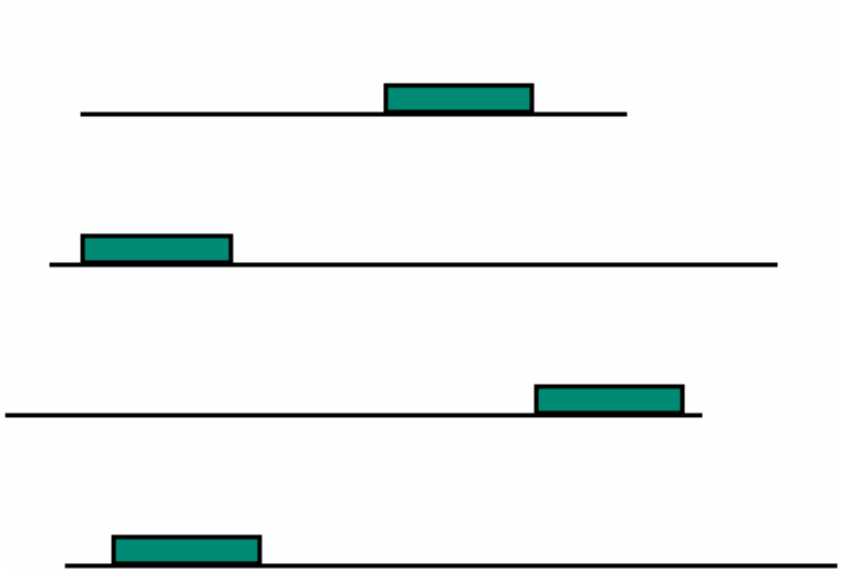
$f_\alpha = \frac{1}{4}$

$10 =$ 2 2 2 2 2

Entropie relative (information content)

$$\sum_{i=1}^{L} \sum_{\alpha \in \Sigma} f_{i\alpha} \log_2 \frac{f_{i\alpha}}{f_\alpha}$$

relative entropy



A	A	A	A	A
A	A	A	A	A
A	A	A	A	A
A	A	A	A	A

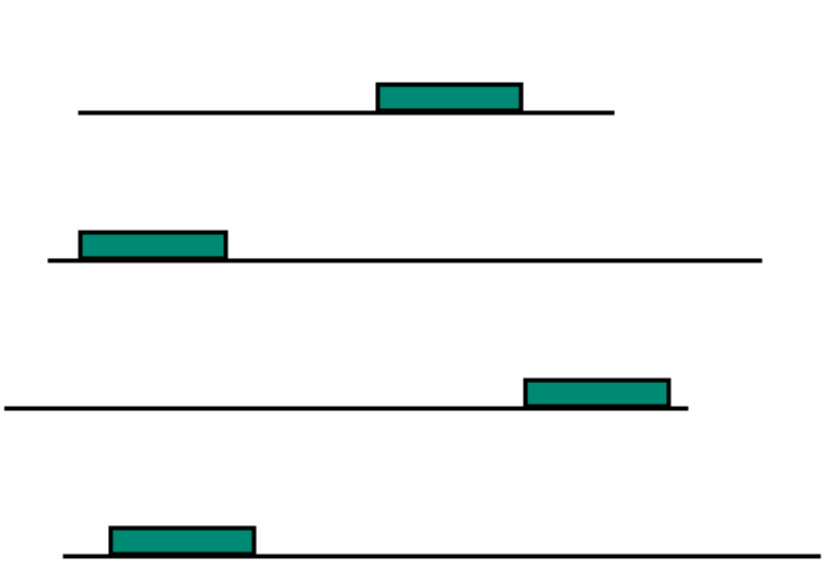
$$f_A = \frac{1}{16}$$

$$20 = 4 \quad 4 \quad 4 \quad 4 \quad 4$$

Entropie relative (information content)

$$\sum_{i=1}^{L} \sum_{\alpha \in \Sigma} f_{i\alpha} \log_2 \frac{f_{i\alpha}}{f_\alpha}$$

relative entropy



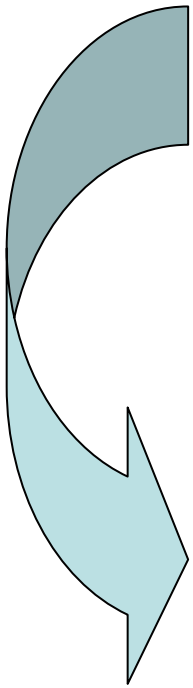
A	A	A	A	A
A	A	A	A	A
A	A	A	A	A
A	A	A	A	A

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

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

Insertions

```
A C A - - - A T G
T C A A C T A T C
A C A C - - A G C
A G A - - - A T C
A C C G - - A T C
```

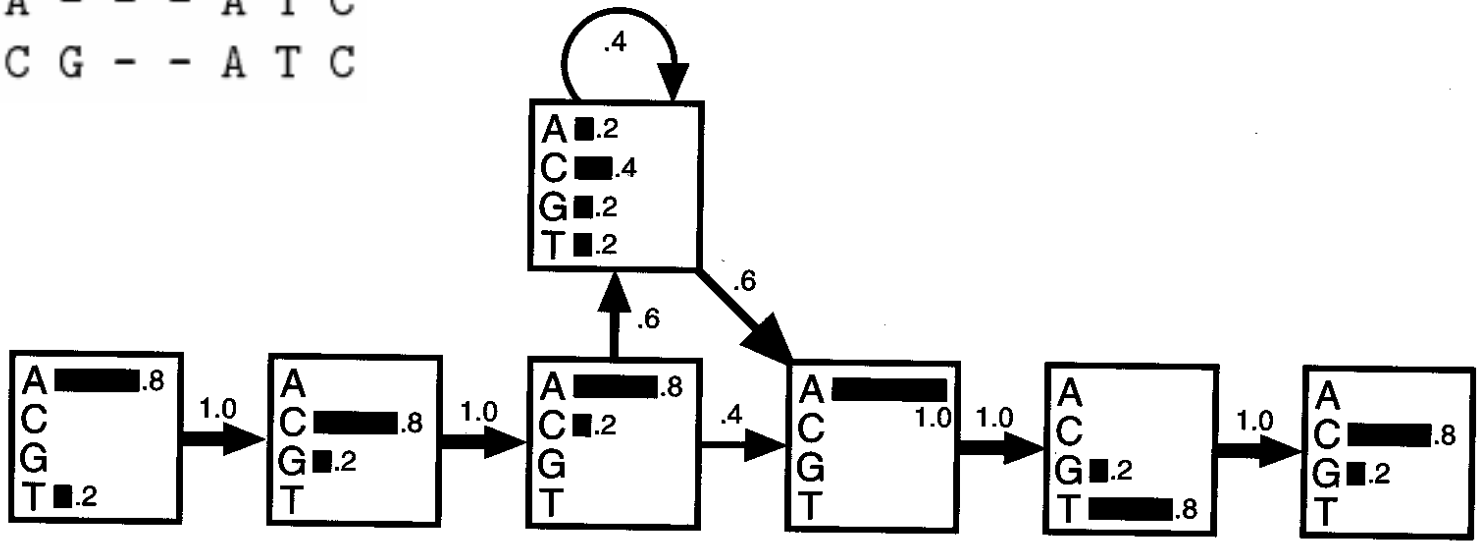


[AT] [CG] [AC] [ACGT]* A [TG] [GC]

Gap

« Généralisation » des PWMs

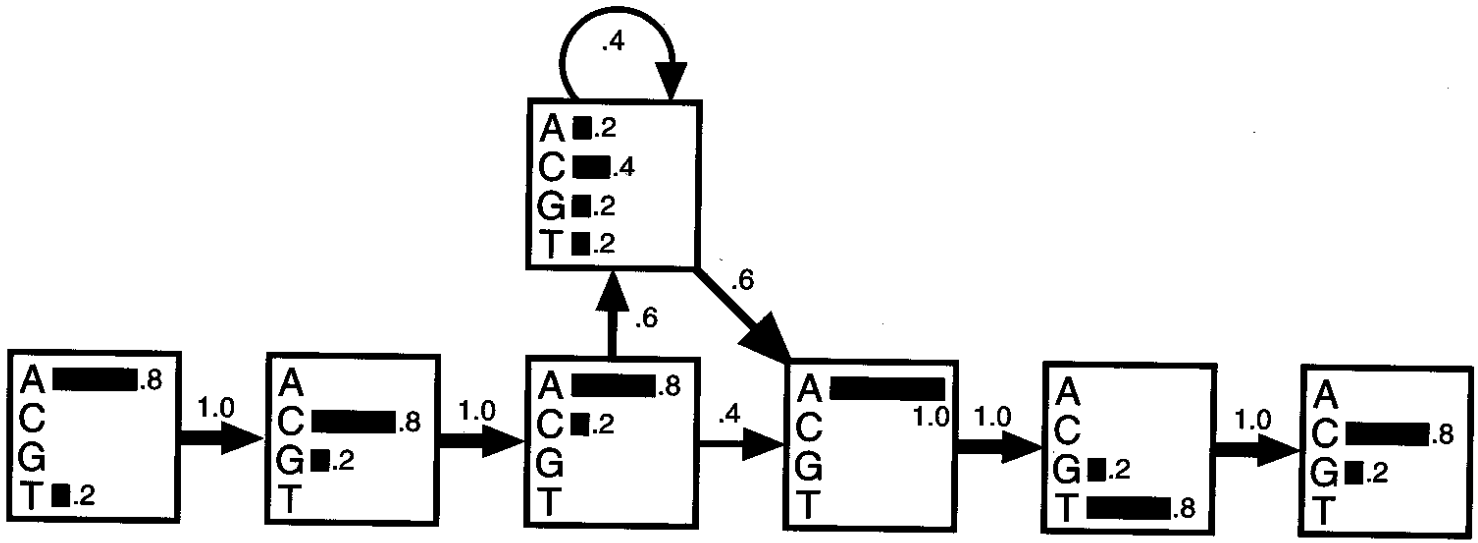
```
A C A - - - A T G
T C A A C T A T C
A C A C - - A G C
A G A - - - A T C
A C C G - - A T C
```



A HMM model for a DNA motif alignments, The transitions are shown with arrows whose thickness indicate their probability. In each state, the histogram shows the probabilities of the four bases.

Probabilité de séquences

To score a sequence using probability:

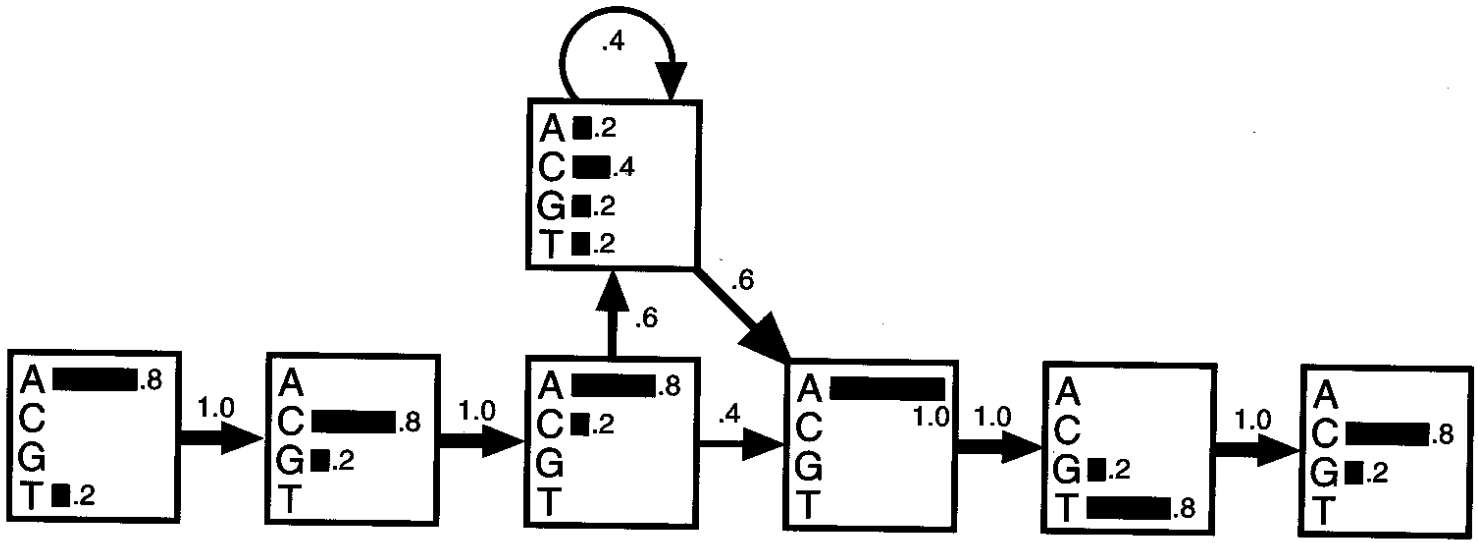


Consensus sequence: ACAC - - ATC

P (ACACATC) = 0.8x1 x 0.8x1 x 0.8x0.6 x 0.4x0.6 x 1x1 x 0.8x1 x 0.8 = 4.7 x 10^{-2}

Probabilité de séquences

To score a sequence using probability:



Highly implausible sequence: TGCT - - AGG

$$P(TGCTAGG) = 0.0023 \times 10^{-2}$$

Probabilité de séquences

To score the sequence using log-odds:
log-odds for sequence S = log [P(S)/(0.25) ^{L}] = log P(S) - L Log 0.25

	Sequence	Probability x 100	Log odds
Consensus	A C A C - - A T C	4.7	6.7
Original sequences	A C A - - - A T C	3.3	4.9
	T C A A C T A T C	0.0075	3.0
	A C A C - - A G C	1.2	5.3
	A G A - - - A T C	3.3	4.9
	A C C G - - A T C	0.59	4.6
Exceptional	T G C T - - A G G	0.0023	- 0.97

Probabilities and log-odds scores for the 5 sequences in the alignment and for the consensus sequence and the exceptional sequence.

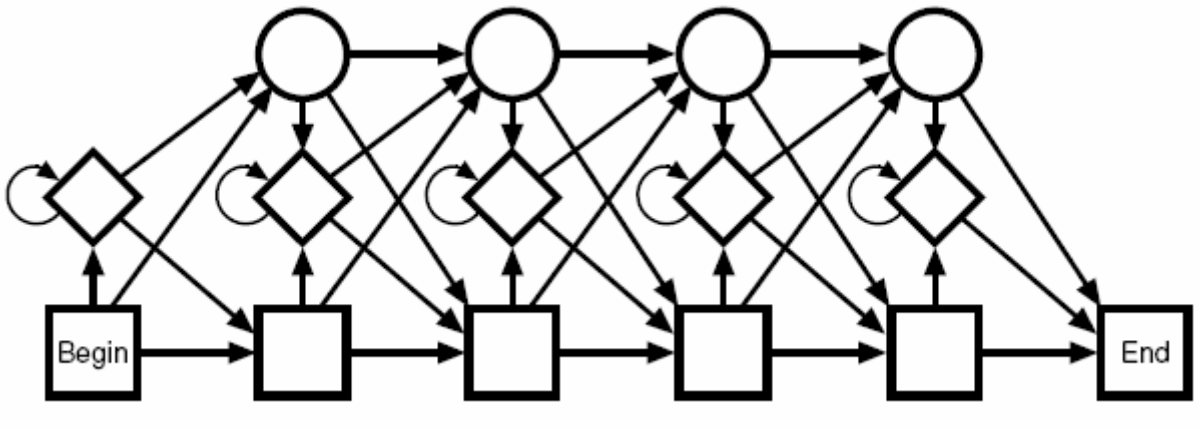
Profile HMM

- Insertions-délétions :

Delete states

Insert states

Match states



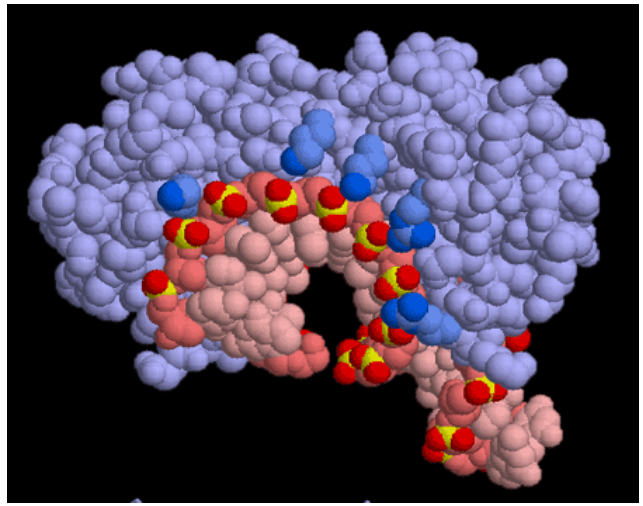
P20226 Human TATA-box Binding Protein

Séquence (format Fasta):

```
>UniProt/Swiss-Prot|P20226|TBP_HUMAN TATA-box binding protein
MDQNNSLPPYAQGLASPQGAMTPGIPIFSPMMPYGTGLTPQPIQNTNSLSILEEQQRQQQ
QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQAVAAAAVQQSTSQQATQGTSGQAPQ
LFHSQTLTTAPLPGTTPLYPSPMTPMTPITPATPASESSGIVPQLQNIVSTVNLGCKLDL
KTIALRARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEEQSRLAARKYARV
VQKLGFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHQQFSSYEPELFPGLIYRMIKPRI
VLLIFVSGKVVLTGAKVRAEIYEAFENIYPILKGFRKTT
```

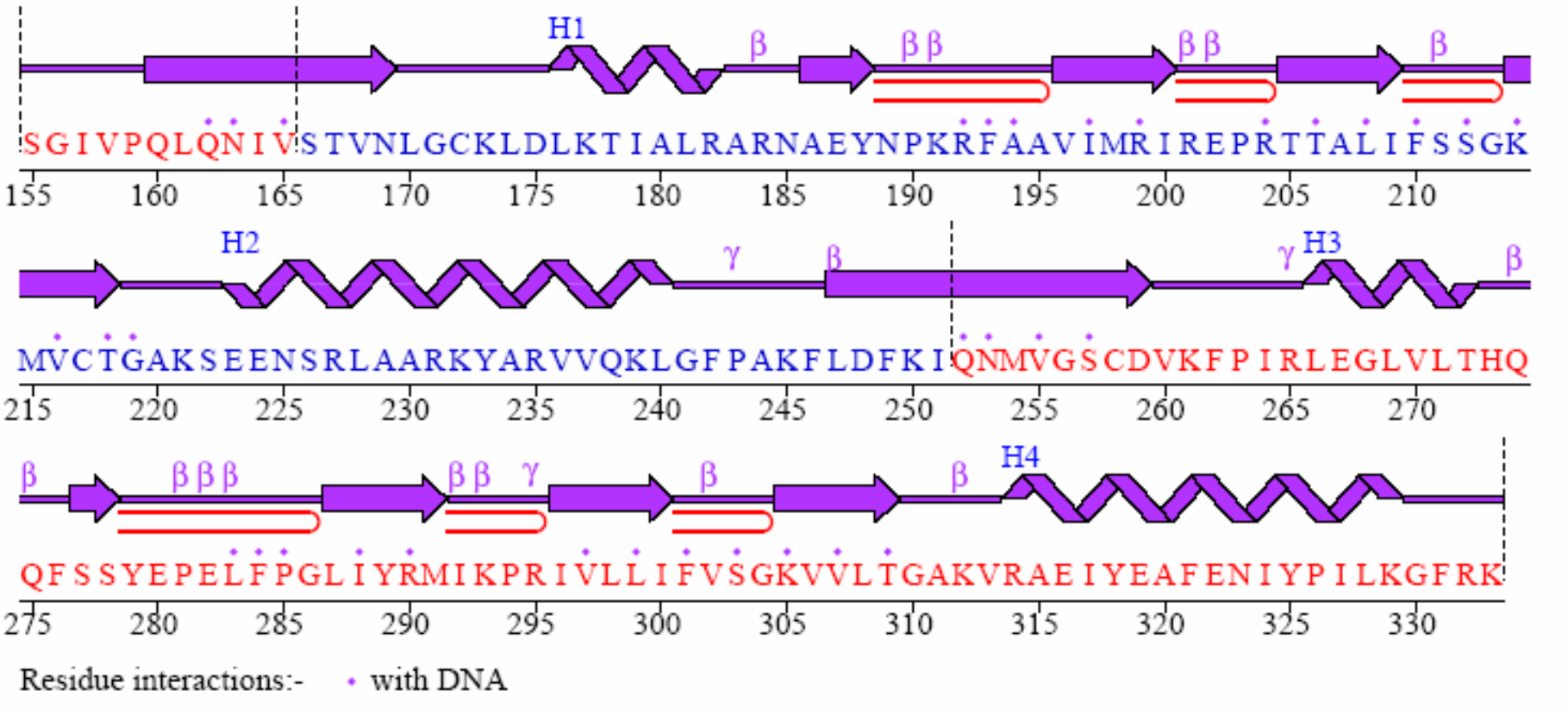
Extrait du fichier PDB :

```
REMARK 350   BIOMT2   1   0.00000   1.00000   0.00000        0.00000
REMARK 350   BIOMT3   1   0.00000   0.00000   1.00000        0.00000
REMARK 350
CRYST1   45.800   78.000   97.400  90.00  90.00  90.00 P 1           1
SCALE1      0.021834  0.000000  0.000000        0.00000
SCALE2      0.000000  0.012821  0.000000        0.00000
SCALE3      0.000000  0.000000  0.010267        0.00000
ATOM      1  N   SER A 155      79.567  95.989 -35.807  1.00 31.29
ATOM      2  CA  SER A 155      78.596  95.092 -36.391  1.00 28.46
ATOM      3  C   SER A 155      79.183  93.711 -36.578  1.00 28.17
ATOM      4  O   SER A 155      78.463  92.713 -36.552  1.00 37.13
ATOM      5  CB  SER A 155      78.144  95.636 -37.734  1.00 24.93
ATOM      6  OG  SER A 155      79.256  95.759 -38.586  1.00 31.23
ATOM      7  N   GLY A 156      80.498  93.655 -36.761  1.00 29.06
ATOM      8  CA  GLY A 156      81.158  92.381 -36.994  1.00 27.86
ATOM      9  C   GLY A 156      81.094  92.099 -38.486  1.00 30.80
ATOM     10  O   GLY A 156      81.434  91.013 -38.934  1.00 30.19
ATOM     11  N   ILE A 157      80.677  93.104 -39.254  1.00 33.09
ATOM     12  CA  ILE A 157      80.554  92.985 -40.701  1.00 33.55
```

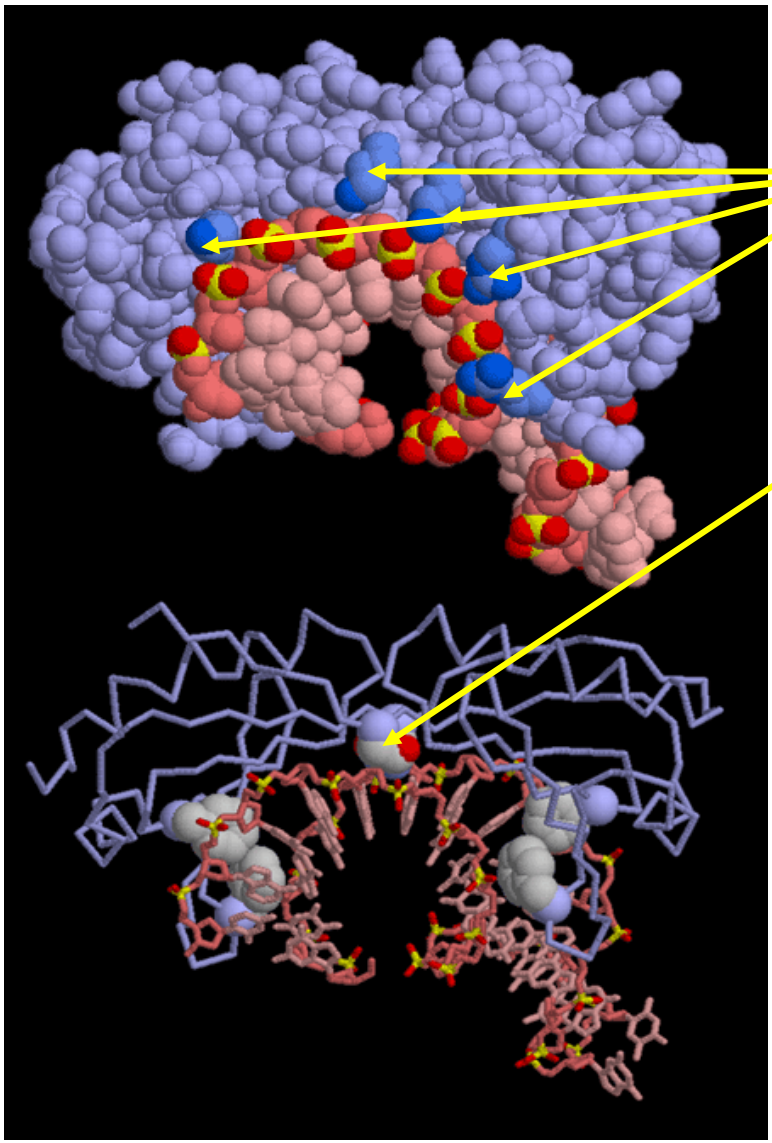


P20226 Human TATA-box Binding Protein

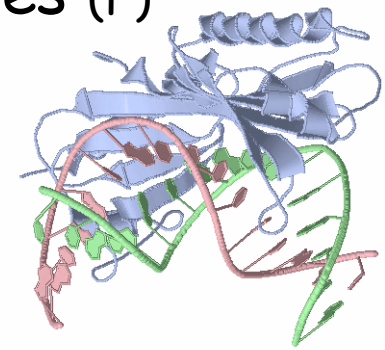
Vue synthétique :



Reconnaissance de la TATA-Box par la TATA Binding Protein



- Interactions Lysines (K) et Arginines (R) avec les Groupes phosphates de l'ADN
- Pont hydrogène entre Asparagine (N) et l'ADN
- 2×2 Phenylalalines (F) se glissent dans le sillon de l'ADN :

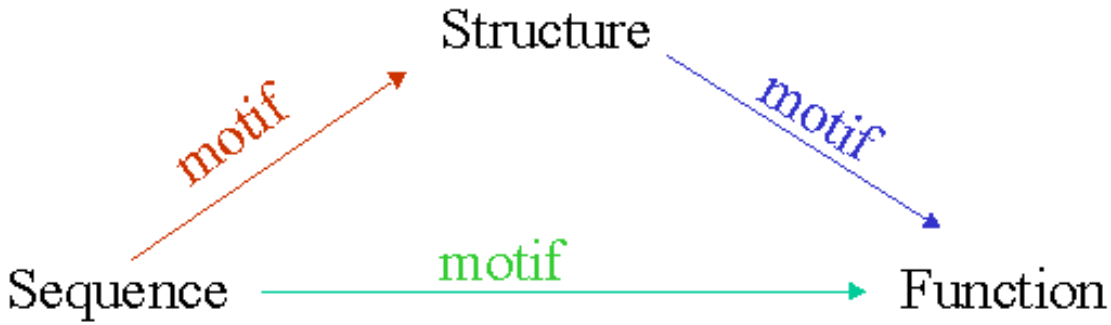


  - interaction avec les bases
  - Courbure de l'ADN (flexibilité De TATA)

Other TATA-box binding proteins

There are different TATA-box binding proteins that have been identified, including TBP1, TBP2, TBP3 and TBPL (TATA-box binding protein like). All of these proteins are related in terms of sequence and structure. The TBP is composed of an N-terminal that varies in both length and sequence, and a highly conserved C-terminal region that binds to the TATA box. The C-terminal region contains two 77-amino acid repeats that produce a saddle-shaped structure that straddles the DNA. In addition, the C-terminal core interacts with a variety of transcription factors as well as regulatory proteins. The N-terminal region appears to modulate DNA binding of the TBP molecule, in addition to other more specific functions.

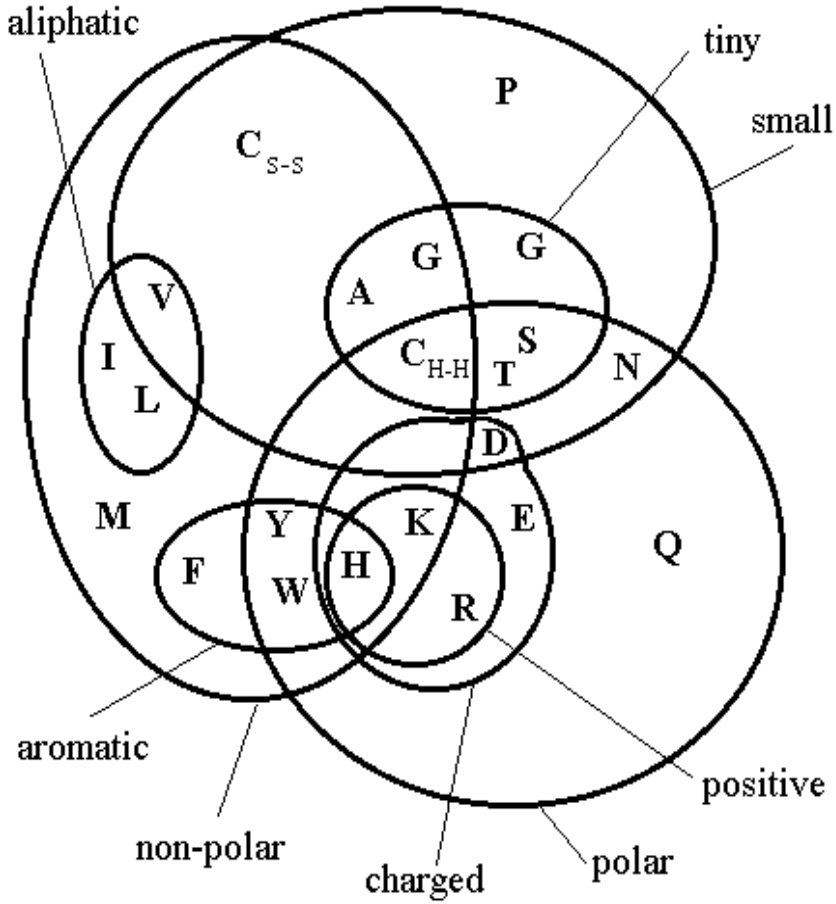
Motifs in Protein Analysis



Sequence -> structure motifs
Sequence -> function motifs
} Sequence motifs

Structure -> function motifs
Structure motifs

Acides aminés

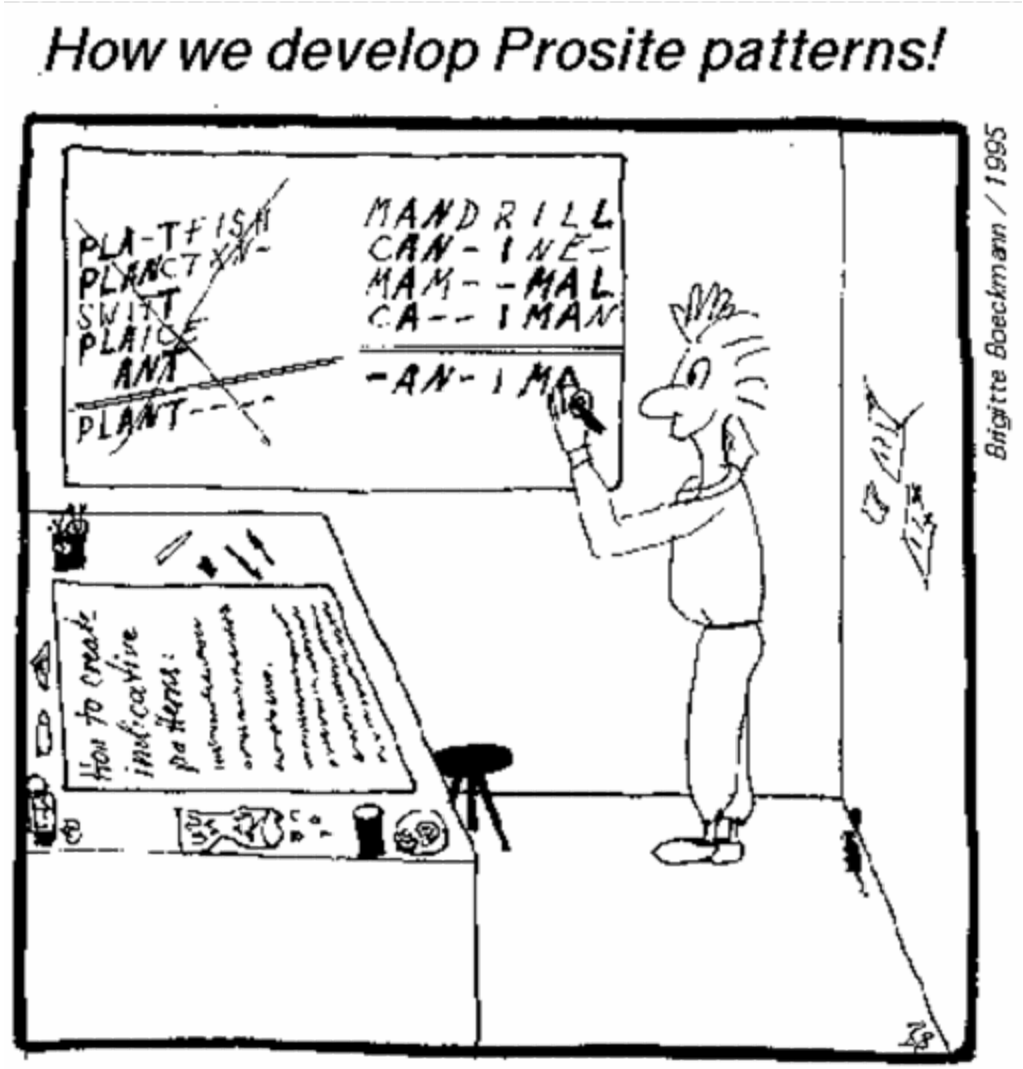


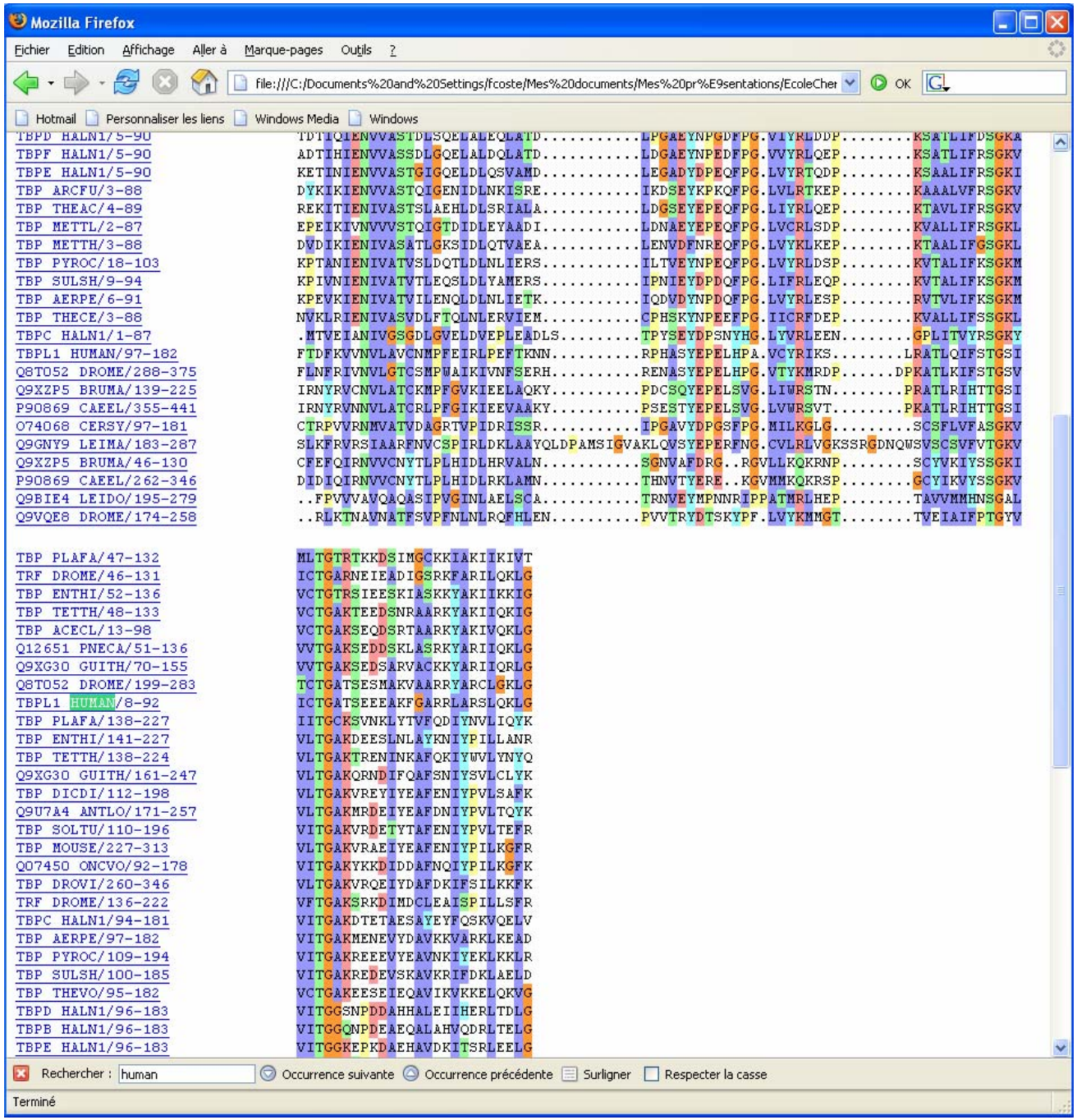
Matrice de substitution (Blosum)

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

L'outil de découverte de motifs le plus utilisé sur les Protéines

- ClustalW !




Mozilla Firefox
Fichier Edition Affichage Aller à Marque-pages Outils ?
file:///C:/Documents%20and%20Settings/fcoste/Mes%20documents/Mes%20pr%E9sentations/EcoleCher
OK
Hotmail Personnaliser les liens Windows Media Windows
TBPD_HALN1/5-90 TDTIQIENVVASTDLSQELALEQLATD.....LPGAEYNPGDFPG.VIYRLDDP.....KSATLIFDSGKA
TBPF_HALN1/5-90 ADTIHIENVVASSDLGQELALDQLATD.....LDGAEYNPEDFPG.VVYRLQEP.....KSATLIFRSGKV
TBPE_HALN1/5-90 KETINIENVVASTGIGQELDLQSVAMD.....LEGADYDPEQFPG.LVYRTQDP.....KSAALIFRSGKI
TBP_ARCFU/3-88 DYKIKIENVVASTQIGENIDLNKISRE.....IKDSEYKPKQFPG.LVLRTKEP.....KAAALVFRSGKV
TBP_THEAC/4-89 REKITIENIVASTSLAEHLDLSRIALA.....LDGSEYEPEQFPG.LIYRLQEP.....KTAVLIFRSGKV
TBP_METTL/2-87 EPEIKIVNVVVSTQIGTDIDLEYAADI.....LDNAEYEPEQFPG.LVCRLSDP.....KVALLIFRSGKL
TBP_METTH/3-88 DVDIKIENIVASATLGKSIDLQTVAEA.....LENVDFNREQFPG.LVYKLKEP.....KTAALIFGSGKL
TBP_PYROC/18-103 KPTANIENIVATVSLDQTLDLNLIERS.....ILTVEYNPEQFPG.LVYRLDSP.....KVTALIFKSGKM
TBP_SULSH/9-94 KPIVNIENIVATVTLEQSLDLYAMERS.....IPNIEYDPDQFPG.LIFRLEQP.....KVTALIFKSGKM
TBP_AERPE/6-91 KPEVKIENIVATVILENQLDLNLIETK.....IQDVDYNPDQFPG.LVYRLESP.....RVTVLIFKSGKM
TBP_THECE/3-88 NVKLRIENIVASVDLFTQLNLERVIEM.....CPHSKYNPEEFPG.IICRFDEP.....KVALLIFSSGKL
TBPC_HALN1/1-87 .MTVEIANIVGSGDLGVELDVEPLEADLS.....TPYSEYDPSNYHG.LYVRLEEN.....GPLITVYRSGKY
TBPL1_HUMAN/97-182 FTDFKVVNVLAVCNMPFEIRLPEFTKNN.....RPHASYEPELHPA.VCYRIKS.....LRATLQIFSTGSI
Q8T052_DROME/288-375 FLNFRIVNVLGTCSMPWAIKIVNFSERH.....RENASYEPELHPG.VTYKMRDP.....DPKATLKIFSTGSV
Q9XZP5_BRUMA/139-225 IRNYRVCNVLATCKMPFGVKIEELAQKY.....PDCSQYEPELSVG.LIWRSTN.....PRATLRIHTTGSI
P90869_CAEEL/355-441 IRNYRVNNVLATCRLPFGIKIEEVAAKY.....PSESTYEPELSVG.LVWRSVT.....PKATLRIHTTGSI
O74068_CERSY/97-181 CTRPVVRNMVATVDAGRTVPIDRISSR.....IPGAVYDPGSFPG.MILKGLG.....SCSFLVFASGKV
Q9GNY9_LEIMA/183-287 SLKFRVRSIAARFNVCSPIRLDKLAAYQLDPAMSIGVAKLQVSYEPERFNG.CVLRLVGKSSRGDNQWSVSCSVFVTGKV
Q9XZP5_BRUMA/46-130 CFEFQIRNVVCNYTLPLHIDLHRVALN.....SGNVAFDRG..RGVLLKQKRNP.....SCYVKIYSSGKI
P90869_CAEEL/262-346 DIDIQIRNVVCNYTLPLHIDLRKLAMN.....THNVTYERE..KGVMMKQKRSP.....GCYIKVYSSGKV
Q9BIE4_LEIDO/195-279 ..FPVVVAVQAQASIPVGINLAELSCA.....TRNVEYMPNNRIPPATMRLHEP.....TAVVMMHNSGAL
Q9VQE8_DROME/174-258 ..RLKTNAVNATFSVPFNLNLRQFHLEN.....PVVTRYDTSKYPF.LVYKMMGT.....TVEIAIFPTGYV
TBP_PLAFA/47-132 MLTGTRTKKDSIMGCKKIAKIIKIVT
TRF_DROME/46-131 ICTGARNEIEADIGSRKFARILQKLG
TBP_ENTHI/52-136 VCTGTRSIEESKIASKKYAKIIKKIG
TBP_TETTH/48-133 VCTGAKTEEDSNRAARKYAKIIQKIG
TBP_ACECL/13-98 VCTGAKSEQDSRTAARKYAKIVQKLG
Q12651_PNECA/51-136 VVTGAKSEDDSKLASRKYARIIQKLG
Q9XG30_GUITH/70-155 VVTGAKSEDSARVACKKYARIIQRLG
Q8T052_DROME/199-283 TCTGATSESMAKVAARRYARCLGKLG
TBPL1_HUMAN/8-92 ICTGATSEEEAKFGARRLARSLQKLG
TBP_PLAFA/138-227 IITGCKSVNKLYTVFQDIYNVLIQYK
TBP_ENTHI/141-227 VLTGAKDEESLNLAYKNIYPILLANR
TBP_TETTH/138-224 VLTGAKTRENINKAFQKIYWVLYNYQ
Q9XG30_GUITH/161-247 VLTGAKQRNDIFQAFSNIYSVLCLYK
TBP_DICDI/112-198 VLTGAKVREYIYEAFENIYPVLSAFK
Q9U7A4_ANTLO/171-257 VLTGAKMRDEIYEAFDNIYPVLTQYK
TBP_SOLTU/110-196 VITGAKVRDETYTAFENIYPVLTEFR
TBP_MOUSE/227-313 VLTGAKVRAEIYEAFENIYPILKGFR
Q07450_ONCVO/92-178 VITGAKYKKDIDDAFNQIYPILKGFK
TBP_DROVI/260-346 VLTGAKVRQEIYDAFDKIFSILKKFK
TRF_DROME/136-222 VFTGAKSRKDIMDCLEAISPILLSFR
TBPC_HALN1/94-181 VITGAKDTETAESAYEYFQSKVQELV
TBP_AERPE/97-182 VITGAKMENEVYDAVKKVARKLKEAD
TBP_PYROC/109-194 VITGAKREEEVYEAVNKIYEKLKKLR
TBP_SULSH/100-185 VITGAKREDEVSKAVKRIFDKLAELD
TBP_THEVO/95-182 VCTGAKEESEIEQAVIKVKKELQKVG
TBPD_HALN1/96-183 VITGGSNPDDAHHALEIIHERLTDLG
TBPB_HALN1/96-183 VITGGQNPDEAEQALAHVQDRLTELG
TBPE_HALN1/96-183 VITGGKEPKDAEHAVDKITSRLEELG
Rechercher : human
Occurrence suivante
Occurrence précédente
Surligner
Respecter la casse
Terminé

NiceSite View of PROSITE: PS00351 - Mozilla Firefox

Fichier Edition Affichage Aller à Marque-pages Outils ?

file:///C:/Documents%20and%20Settings/fcoste/Mes%20documents/Mes%20pr%E9sentations/EcoleChercheursBioinfo/fiche_prosite_TBP.htm

Hotmail Personnaliser les liens Windows Media Windows

ExPASy Home page | Site Map | Search ExPASy | Contact us | PROSITE

Search PROSITE for Go Clear

NiceSite View of PROSITE: PS00351

General information about the entry	
Entry name	**TFIID**
Accession number	**PS00351**
Entry type	PATTERN
Date	NOV-1990 (CREATED); DEC-2004 (DATA UPDATE); SEP-2005 (INFO UPDATE).
PROSITE documentation	PDOC00303

Name and characterization of the entry

Description: Transcription factor TFIID repeat signature

Y-x-[PK]-x(2)-[IF]-x(2)-[LIVM](2)-x-[KRH]-x(3)-P-[RKQ]-x(3)-L-[LIVM]-F-x- [STN]-G-[KR]-[LIVMA]-x(3)-G-[TAGL]-[KR]-x(7)-[AGCS]-x(7)-[LIVMF].

Numerical results

- UniProtKB/Swiss-Prot release number: **48.2**, total number of sequence entries in that release: **195589**.
- Total number of hits in UniProtKB/Swiss-Prot: **116 hits in 67 different sequences**
- Number of hits on proteins that are known to belong to the set under consideration: **116 hits in 67 different sequences**
- Number of hits on proteins that could potentially belong to the set under consideration: **0 hits in 0 different sequences**
- Number of false hits (on unrelated proteins): **0 hits in 0 different sequences**
- Number of known missed hits: **2**
- Number of partial sequences which belong to the set under consideration, but which are not hit by the pattern or profile because they are partial (fragment) sequences: **0**
- Precision (true hits / (true hits + false positives)): **100.00 %**
- Recall (true hits / (true hits + false negatives)): **98.31 %**

Comments

- Taxonomic range: **Archaebacteria, Eukaryotes**
- Maximum known number of repetitions of the pattern in a single protein: **2**
- VERSION: **1**

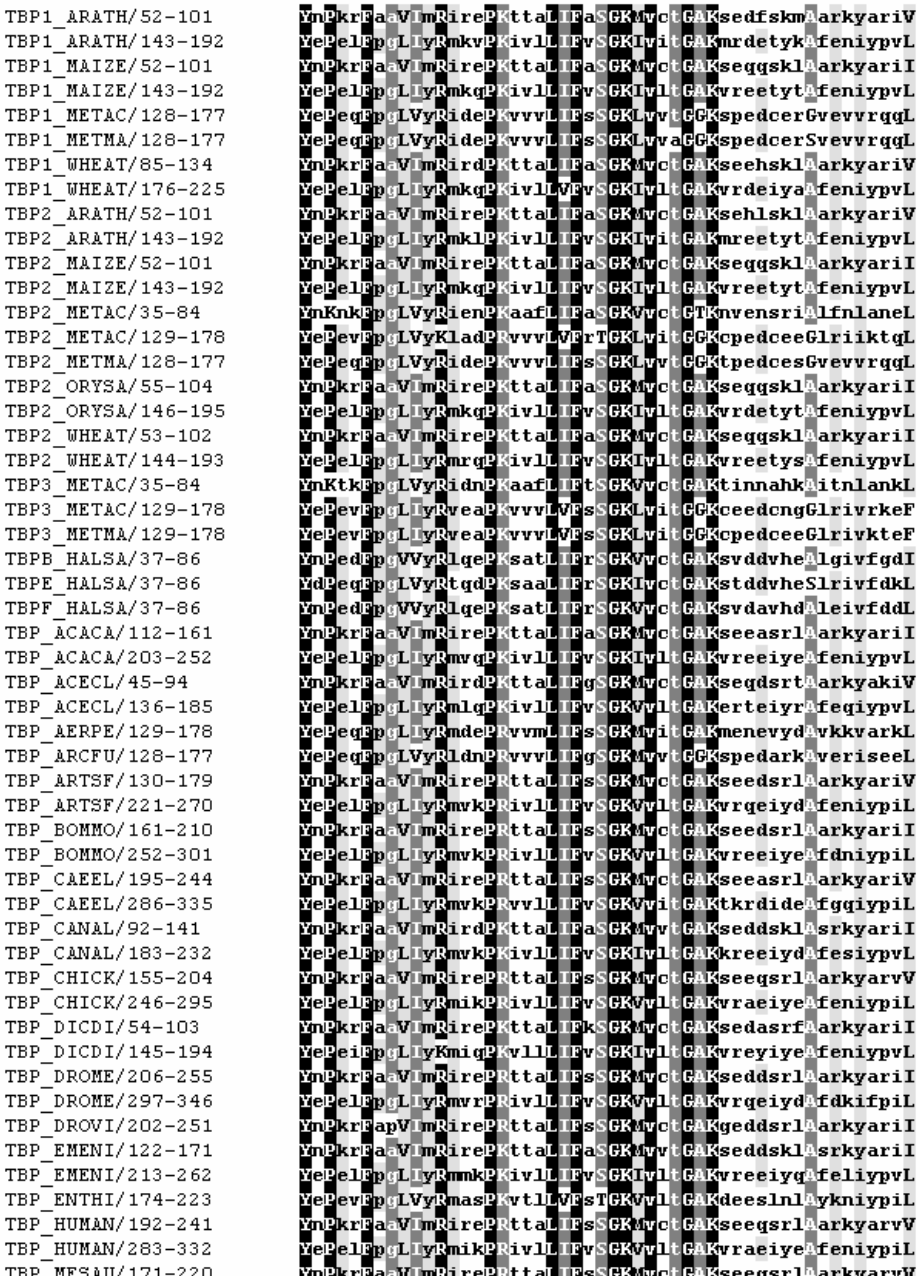
Cross-references

True positive hits:

TBP1_ARATH (P28147), TBP1_MAIZE (P50158), TBP1_METAC (Q8TI26),
TBP1_METMA (Q8PY37), TBP1_WHEAT (P26356), TBP2_ARATH (P28148),
TBP2_MAIZE (P50159), TBP2_METAC (Q8TU94), TBP2_METMA (Q8PY36),
TBP2_ORYSA (Q8W0W4), TBP2_WHEAT (Q02879), TBP3_METAC (Q8TTZ7),
TBP3_METMA (Q8PUZ4), TBPB_HALSA (Q48325), TBPE_HALSA (Q9HN56),
TBPF_HALSA (Q9HHE9), TBP_ACACA (P26354), TBP_ACECL (P46272),
TBP_AERPE (Q9YAT1), TBP_ARCFU (O29874), TBP_ARTSF (O17488),

Y-x-[PK]-x(2)-[IF]-x(2)-[LIVM](2)-x-[KRH]-x(3)-P-[RKQ]-x(3)-L-[LIVM]-

CLUSTAL format alignment



Pfam: TBP - Mozilla Firefox

Fichier Edition Affichage Aller à Marque-pages Outils ?

file:///C:/Documents%20and%20Settings/fcoste/Mes%20documents/Mes%20pr%E9sentations/EcoleChercheursBic

Hotmail Personnaliser les liens Windows Media Windows

Pfam

Protein families database of alignments and HMMs

Wellcome Trust Sanger Institute

TBP

Home Search by Browse by ftp iPfam Help

Figure 1: 1ytf
Complex (transcription regulation/dna)
Yeast tfiia/tbp/dna complex

Key:

Domain	Chain	Start Residue	End Residue
TBP	A	152	238
TBP	A	62	147
TFIIA	B	2	47
TFIIA	C	241	285
TFIIA_gamma_N	D	5	53
TFIIA_gamma_C	D	55	118

The Swissprot/PDB mapping was provided by MSD

1ais Display pdb

***Accession number:* PF00352**

Transcription factor TFIID (or TATA-binding protein, TBP)

Add Annotation

NEW! This family forms **interactions** with other Pfam families, to view them click here

This family forms **structural complexes** with other Pfam families, to view them click here

INTERPRO description (entry IPR000814)

The TATA-box binding protein (TBP) is required for the initiation of transcription by RNA polymerases I, II and III, from promoters with or without a TATA box PUBMED:12782648, PUBMED:10974559. TBP associates with a host of factors, including the general transcription factors TFIIA, -B, -D, -E, and -H, to form huge multi-subunit pre-initiation complexes on the core promoter. Through its association with different transcription factors, TBP can initiate transcription from different RNA polymerases. There are several related TBPs, including TBP-like (TBPL) proteins PUBMED:12878007.

The C-terminal core of TBP (~180 residues) is highly conserved and contains two 77-amino acid repeats that produce a saddle-shaped structure that straddles the DNA; this region binds to the TATA box and interacts with transcription factors and regulatory proteins PUBMED:1436073. By contrast, the N-terminal region varies in both length and sequence.

QuickGO

FUNCTION :	RNA polymerase II transcription factor activity (GO:0003702)
PROCESS :	transcription initiation from RNA polymerase II promoter (GO:0006367)
COMPONENT :	transcription factor TFIID complex (GO:0005669)

For additional annotation, see the PROSITE document PDOC00303 [Expasy | SRS-UK | SRS-USA]

Alignment

Seed (56) Full (364)

Format Coloured alignment

Get alignment View HMM logo

Further alignment options here
Help relating to Pfam alignments here

Domain organisation

View 2 representative architectures
View architectures for 364 proteins

Zoom 0.5 pixels/aa.

View Graphic

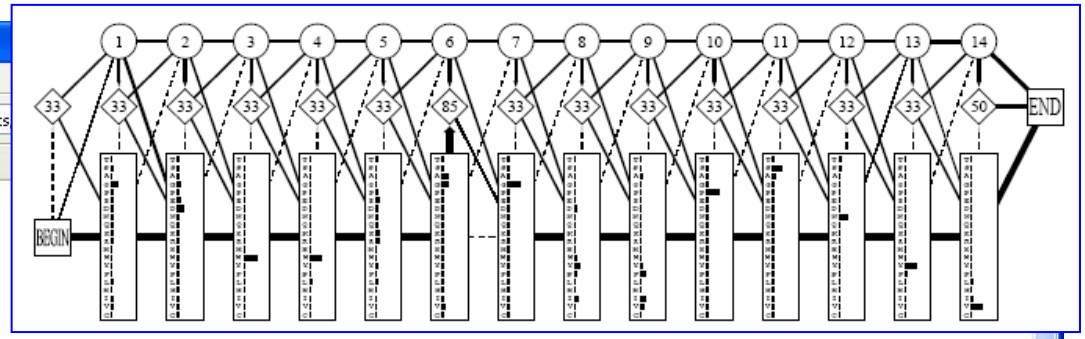
Terminé

Mozilla Firefox

Fichier Edition Affichage Aller à Marque-pages Outils ?

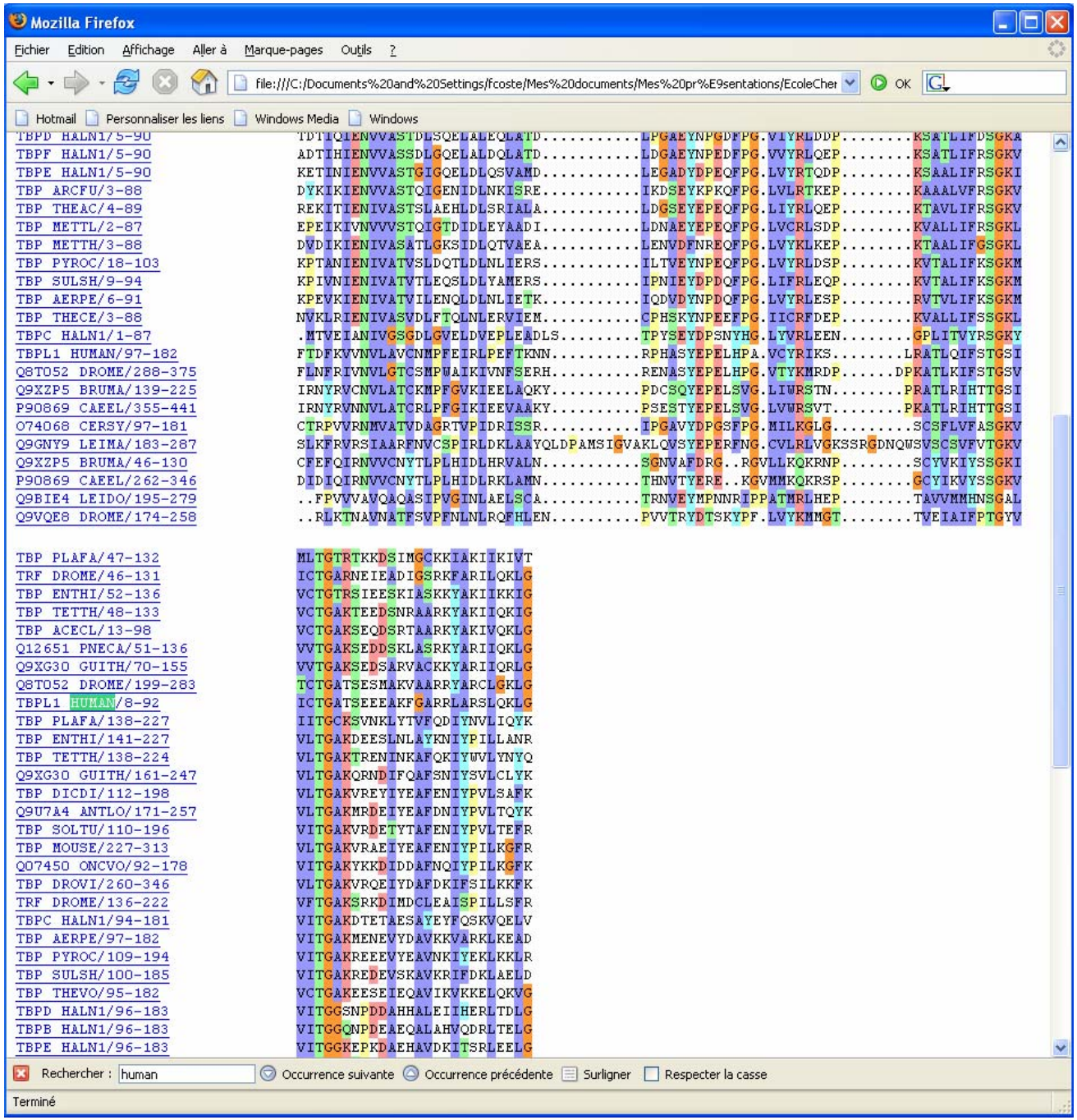
file:///C:/Documents%20and%20Settings/fcoste/Mes%20documents

Hotmail Personnaliser les liens Windows Media Windows



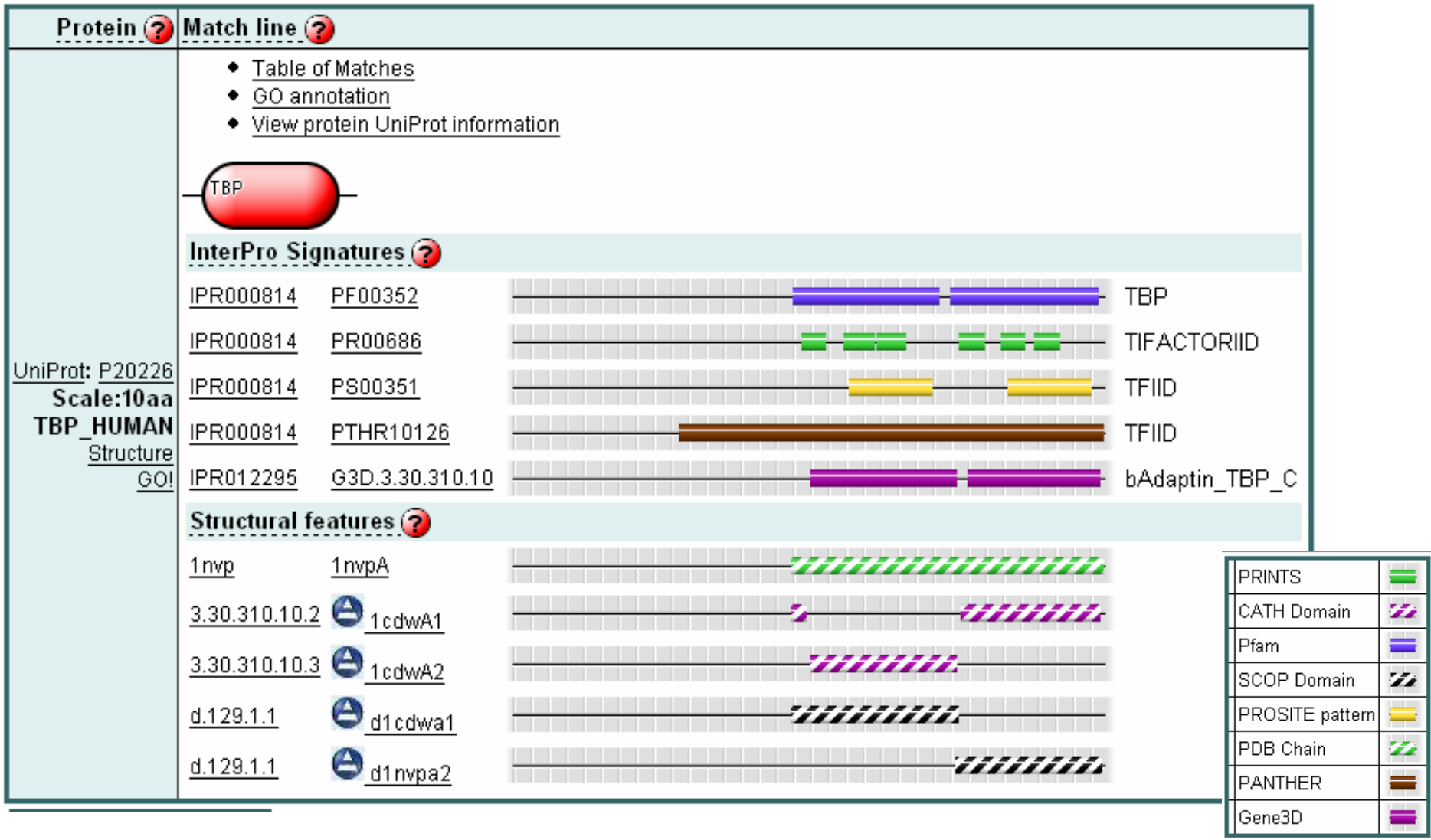

```
HMMER2.0  [2.3.2]
NAME  TBP
ACC   PF00352.11
DESC  Transcription factor TFIID (or TATA-binding protein, TBP)
LENG  89
ALPH  Amino
RF    no
CS    no
MAP   yes
COM   hmmbuild -f -F --wme HMM_fs.ann SEED.ann
COM   hmmcalibrate --seed 0 HMM_fs.ann
NSEQ  56
DATE  Thu Jun 23 13:15:03 2005
CKSUM 8565
GA    25.0 25.0
TC    26.4 25.7
NC    17.9 24.7
XT      -8455     -4  -1000  -1000  -8455     -4  -8455     -4
NULT       -4  -8455
NULE      595  -1558     85    338   -294    453  -1158    197    249    902  -1085   -142    -21   -313     45    531    201    384  -1998   -644
EVD    -8.941253   0.654910
HMM          A       C       D       E       F       G       H       I       K       L       M       N       P       Q       R       S       T       V       W       Y
           m->m    m->i    m->d    i->m    i->i    d->m    d->d    b->m    m->e
            -61       *   -4585
      1     -43    2851     165    -145    1867   -2109   -1585     123     629   -3407   -2488    1264   -3019     553   -1056     311    -691   -3012   -3585   -1081      1
      -    -149    -500     233      43    -381     399     106    -626     210    -466    -720     275     394      45      96     359     117    -369    -294    -249
      -     -11   -9689  -10731    -894   -1115    -701   -1378   -1061   -7459
      2   -2273     482    -414   -1386     420   -3752   -1695     834   -1307     921    1782   -3206    1026   -2895     246   -1793    1738    -335   -2647   -2027      2
      -    -149    -500     233      43    -381     399     106    -626     210    -466    -720     275     394      45      96     359     117    -369    -294    -249
      -     -11   -9833  -10875    -894   -1115    -701   -1378   -7521   -7451
      3   -2263   -3736    1660     775     729    -879      76   -2399     667   -3752   -2825    1138   -1121    -611    1409    -241     820   -2636   -3919   -3070      3
      -    -149    -500     233      43    -381     399     106    -626     210    -466    -720     275     394      45      96     359     117    -369    -294    -249
      -     -10  -10076  -11118    -894   -1115    -701   -1378   -7521   -7443
      4   -3105   -2982   -5565   -4954    2562   -4827   -3655    1783   -4573     866   -1822   -4465    1730   -4140   -4362   -3932   -3148     837   -3383    1021      4
      -    -149    -500     233      43    -381     399     106    -626     210    -466    -720     275     394      45      96     359     117    -369    -294    -249
      -     -10  -10076  -11118    -894   -1115    -701   -1378   -7521   -7435
      5   -2265   -3736   -1160     643   -4056   -3239    -565   -3342    1909   -1275   -2825     995   -3332    1195    1429   -2145     237    1206   -3919   -3237      5
      -    -149    -500     233      43    -381     399     106    -626     210    -466    -720     275     394      45      96     359     117    -369    -294    -249
      -     -10  -10076  -11118    -894   -1115    -701   -1378   -7521   -7426
      6   -4359   -3840   -7061   -6755   -4578   -6929   -7051    3025   -6742   -1626   -3251   -6585   -6655   -6731   -6940   -6320     398    2662   -6392   -5842      6
      -    -149    -500     233      43    -381     399     106    -626     210    -466    -720     275     394      45      96     359     117    -369    -294    -249
      -     -10  -10076  -11118    -894   -1115    -701   -1378   -7521   -7418
      7    -207     -20   -2128    1366   -4023   -3247    1180   -3764   -1490   -3724   -2806    1463   -3340    1637    1749   -2155    -381    1234   -3903   -3228      7
      -    -149    -500     233      43    -381     399     106    -626     210    -466    -720     275     394      45      96     359     117    -369    -294    -249
      -     -11  -10077  -11119    -894   -1115    -701   -1378   -7521   -7410
      8    1097   -3783   -5250   -5624   -6435   -3999   -5490   -6287   -6075   -6522   -5570    4056   -4804   -5492   -5921    -225   -3623   -4978   -6643   -6524      8
      -    -149    -500     233      43    -381     399     106    -626     210    -466    -720     275     394      45      96     359     117    -369    -294    -249
      -     -11  -10076  -11118    -894   -1115    -701   -1378   -7521   -7401
      9   -4392   -3868   -7093   -6787   -4523   -6971   -7092    2597   -6778   -3237    2241   -6628   -6677   -6736   -6966   -6370   -4383    2986   -6365   -5852      9
      -    -149    -500     233      43    -381     399     106    -626     210    -466    -720     275     394      45      96     359     117    -369    -294    -249
      -     -11  -10076  -11118    -894   -1115    -701   -1378   -7521   -7392
     10    -344   -2844   -5407   -4824   -2926   -4743   -3704     590   -4467     449   -2073     674   -4764     859   -4322   -3856   -3026    3040   -3562   -3207     10
```

Terminé


Mozilla Firefox
Fichier Edition Affichage Aller à Marque-pages Outils ?
file:///C:/Documents%20and%20Settings/fcoste/Mes%20documents/Mes%20pr%E9sentations/EcoleCher
OK
Hotmail Personnaliser les liens Windows Media Windows
TBPD_HALN1/5-90 TDTIQIENVVASTDLSQELALEQLATD.....LPGAEYNPGDFPG.VIYRLDDP.....KSATLIFDSGKA
TBPF_HALN1/5-90 ADTIHIENVVASSDLGQELALDQLATD.....LDGAEYNPEDFPG.VVYRLQEP.....KSATLIFRSGKV
TBPE_HALN1/5-90 KETINIENVVASTGIGQELDLQSVAMD.....LEGADYDPEQFPG.LVYRTQDP.....KSAALIFRSGKI
TBP_ARCFU/3-88 DYKIKIENVVASTQIGENIDLNKISRE.....IKDSEYKPKQFPG.LVLRTKEP.....KAAALVFRSGKV
TBP_THEAC/4-89 REKITIENIVASTSLAEHLDLSRIALA.....LDGSEYEPEQFPG.LIYRLQEP.....KTAVLIFRSGKV
TBP_METTL/2-87 EPEIKIVNVVVSTQIGTDIDLEYAADI.....LDNAEYEPEQFPG.LVCRLSDP.....KVALLIFRSGKL
TBP_METTH/3-88 DVDIKIENIVASATLGKSIDLQTVAEA.....LENVDFNREQFPG.LVYKLKEP.....KTAALIFGSGKL
TBP_PYROC/18-103 KPTANIENIVATVSLDQTLDLNLIERS.....ILTVEYNPEQFPG.LVYRLDSP.....KVTALIFKSGKM
TBP_SULSH/9-94 KPIVNIENIVATVTLEQSLDLYAMERS.....IPNIEYDPDQFPG.LIFRLEQP.....KVTALIFKSGKM
TBP_AERPE/6-91 KPEVKIENIVATVILENQLDLNLIETK.....IQDVDYNPDQFPG.LVYRLESP.....RVTVLIFKSGKM
TBP_THECE/3-88 NVKLRIENIVASVDLFTQLNLERVIEM.....CPHSKYNPEEFPG.IICRFDEP.....KVALLIFSSGKL
TBPC_HALN1/1-87 .MTVEIANIVGSGDLGVELDVEPLEADLS.....TPYSEYDPSNYHG.LYVRLEEN.....GPLITVYRSGKY
TBPL1_HUMAN/97-182 FTDFKVVNVLAVCNMPFEIRLPEFTKNN.....RPHASYEPELHPA.VCYRIKS.....LRATLQIFSTGSI
Q8T052_DROME/288-375 FLNFRIVNVLGTCSMPWAIKIVNFSERH.....RENASYEPELHPG.VTYKMRDP.....DPKATLKIFSTGSV
Q9XZP5_BRUMA/139-225 IRNYRVCNVLATCKMPFGVKIEELAQKY.....PDCSQYEPELSVG.LIWRSTN.....PRATLRIHTTGSI
P90869_CAEEL/355-441 IRNYRVNNVLATCRLPFGIKIEEVAAKY.....PSESTYEPELSVG.LVWRSVT.....PKATLRIHTTGSI
O74068_CERSY/97-181 CTRPVVRNMVATVDAGRTVPIDRISSR.....IPGAVYDPGSFPG.MILKGLG.....SCSFLVFASGKV
Q9GNY9_LEIMA/183-287 SLKFRVRSIAARFNVCSPIRLDKLAAYQLDPAMSIGVAKLQVSYEPERFNG.CVLRLVGKSSRGDNQWSVSCSVFVTGKV
Q9XZP5_BRUMA/46-130 CFEFQIRNVVCNYTLPLHIDLHRVALN.....SGNVAFDRG..RGVLLKQKRNP.....SCYVKIYSSGKI
P90869_CAEEL/262-346 DIDIQIRNVVCNYTLPLHIDLRKLAMN.....THNVTYERE..KGVMMKQKRSP.....GCYIKVYSSGKV
Q9BIE4_LEIDO/195-279 ..FPVVVAVQAQASIPVGINLAELSCA.....TRNVEYMPNNRIPPATMRLHEP.....TAVVMMHNSGAL
Q9VQE8_DROME/174-258 ..RLKTNAVNATFSVPFNLNLRQFHLEN.....PVVTRYDTSKYPF.LVYKMMGT.....TVEIAIFPTGYV
TBP_PLAFA/47-132 MLTGTRTKKDSIMGCKKIAKIIKIVT
TRF_DROME/46-131 ICTGARNEIEADIGSRKFARILQKLG
TBP_ENTHI/52-136 VCTGTRSIEESKIASKKYAKIIKKIG
TBP_TETTH/48-133 VCTGAKTEEDSNRAARKYAKIIQKIG
TBP_ACECL/13-98 VCTGAKSEQDSRTAARKYAKIVQKLG
Q12651_PNECA/51-136 VVTGAKSEDDSKLASRKYARIIQKLG
Q9XG30_GUITH/70-155 VVTGAKSEDSARVACKKYARIIQRLG
Q8T052_DROME/199-283 TCTGATSESMAKVAARRYARCLGKLG
TBPL1_HUMAN/8-92 ICTGATSEEEAKFGARRLARSLQKLG
TBP_PLAFA/138-227 IITGCKSVNKLYTVFQDIYNVLIQYK
TBP_ENTHI/141-227 VLTGAKDEESLNLAYKNIYPILLANR
TBP_TETTH/138-224 VLTGAKTRENINKAFQKIYWVLYNYQ
Q9XG30_GUITH/161-247 VLTGAKQRNDIFQAFSNIYSVLCLYK
TBP_DICDI/112-198 VLTGAKVREYIYEAFENIYPVLSAFK
Q9U7A4_ANTLO/171-257 VLTGAKMRDEIYEAFDNIYPVLTQYK
TBP_SOLTU/110-196 VITGAKVRDETYTAFENIYPVLTEFR
TBP_MOUSE/227-313 VLTGAKVRAEIYEAFENIYPILKGFR
Q07450_ONCVO/92-178 VITGAKYKKDIDDAFNQIYPILKGFK
TBP_DROVI/260-346 VLTGAKVRQEIYDAFDKIFSILKKFK
TRF_DROME/136-222 VFTGAKSRKDIMDCLEAISPILLSFR
TBPC_HALN1/94-181 VITGAKDTETAESAYEYFQSKVQELV
TBP_AERPE/97-182 VITGAKMENEVYDAVKKVARKLKEAD
TBP_PYROC/109-194 VITGAKREEEVYEAVNKIYEKLKKLR
TBP_SULSH/100-185 VITGAKREDEVSKAVKRIFDKLAELD
TBP_THEVO/95-182 VCTGAKEESEIEQAVIKVKKELQKVG
TBPD_HALN1/96-183 VITGGSNPDDAHHALEIIHERLTDLG
TBPB_HALN1/96-183 VITGGQNPDEAEQALAHVQDRLTELG
TBPE_HALN1/96-183 VITGGKEPKDAEHAVDKITSRLEELG
Rechercher : human
Occurrence suivante
Occurrence précédente
Surligner
Respecter la casse
Terminé

P20226 Human TATA-box Binding Protein

Vue Interpro :



Banks of motifs : two general references

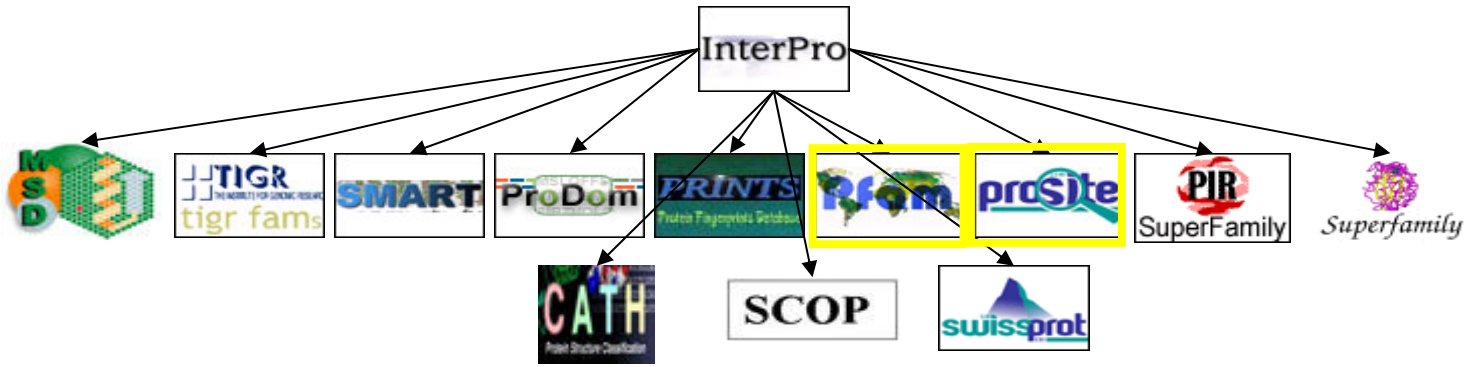
- Regulation patterns : Transfac databases of motifs for binding sites (extended with mutations, composite motifs, commercial now…)

  http://www.gene-regulation.com http://genouest.org

  Origin E. Wingenders Version 6.0 : 6627 sites

  TRANSFAC Patho DB TRANSCompel

- Protein patterns : A unified site for the integration of many banks : Interpro (integrates now also structural data).
  http://www.ebi.ac.uk/interpro/ >80% TrEMBL



Découverte de motifs

A three steps Approach to Pattern Discovery

1. **Choose the language** in which the patterns will be given (*solution space*);
2. **Design the *scoring function*** rating the patterns (from the solution space) with respect to the given data;
3. **Develop an *algorithm*** which given a set of sequences, returns patterns (from the solution space) rating relatively high according to the chosen scoring function.

Brazma, Jonassen, Eidhammer, Gilbert, *J. Comp. Biol,* 1998

Pattern Languages

Deterministic patterns:

- substring patterns (like TATAA)
- regular expression type patterns (like the ones used in PROSITE database

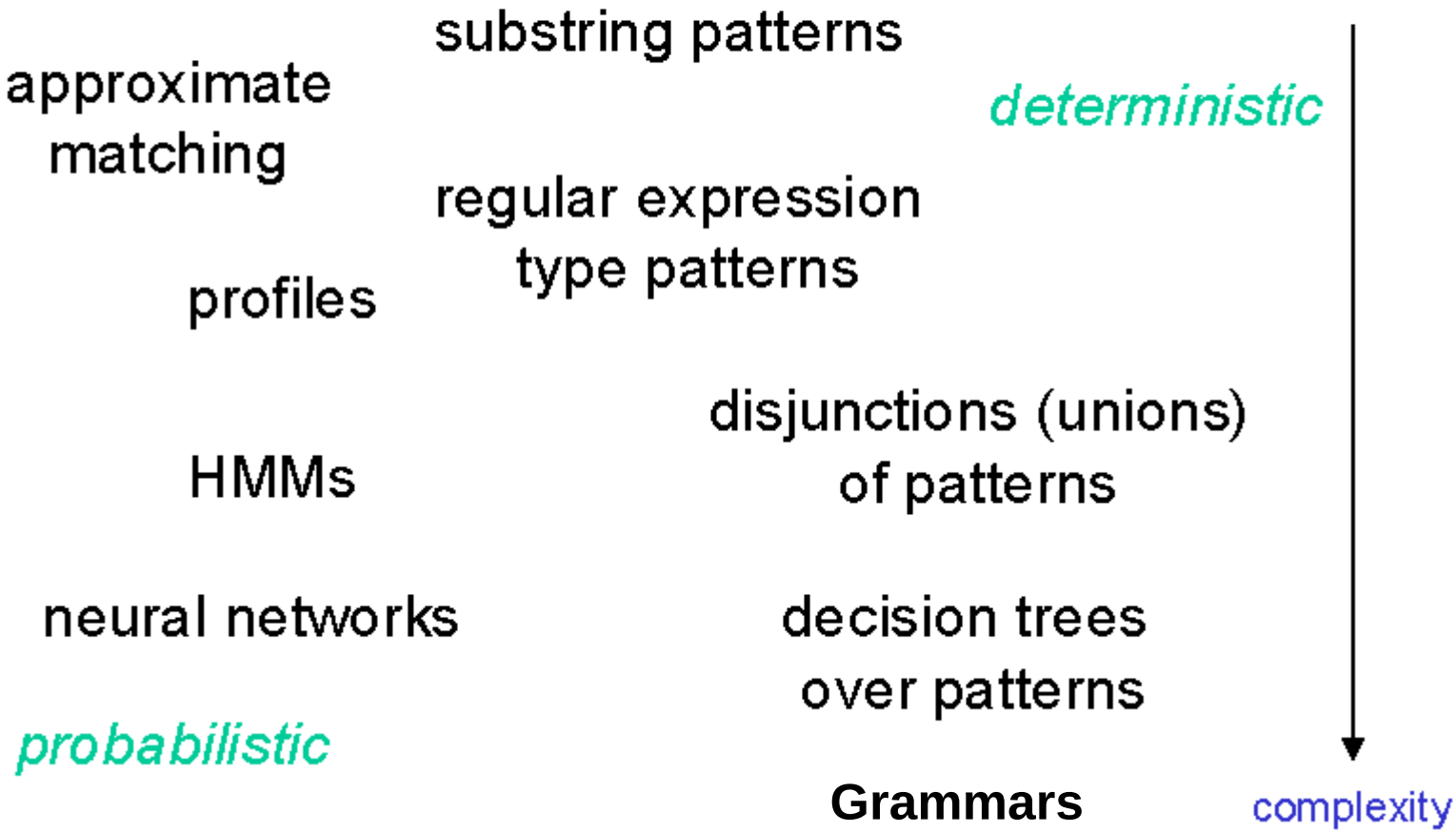
- Probabilistic patterns
  - weight matrices
  - profiles
  - hidden Markov models

Expressivité des motifs

Classe	Exemple	
A	t-c-t-t-g-a	
B	D-R-C-C-x(2)-H-D-x-C	
C	G-G-G-T-F-D-[ILV]-[ST]-[ILV]	
D	V-x-P-x(2)-[RQ]-x(4)-G-x(2)-L-[LM]	
E	G-C-x(1,3)-C-P-x(8,10)-C-C	
F	C-x(2,4)-C-x(3)-[ILVFYC]-x(8)-H-x(3,5)-H	(Prosite, Pratt)
G	G-G-G-T-F-D-*- D-R-C-C-P	
H	G-G-G-T-F-[DE]-*- D-R-C-[PAR]-C	
I	G-G-G-x(2,5)-T-F-[DE]-*-D-x(0,1)-C-[PAR]-C	
J	Expression régulière/Grammaire régulière/Automate	
K	Grammaire algébrique	
M	Grammaire contextuelle	
N	Grammaire à structure de phrase	

D'après Brazma et al

Different description languages



Algorithms for pattern discovery

- **Input:**
  - sequences from the family (positive examples)
  - optionally: sequences not in the family (negative examples)
- **Output:**
  - pattern(s) with high fitness with respect to the input sets - as evaluated by a fitness function.

Approaches to pattern discovery

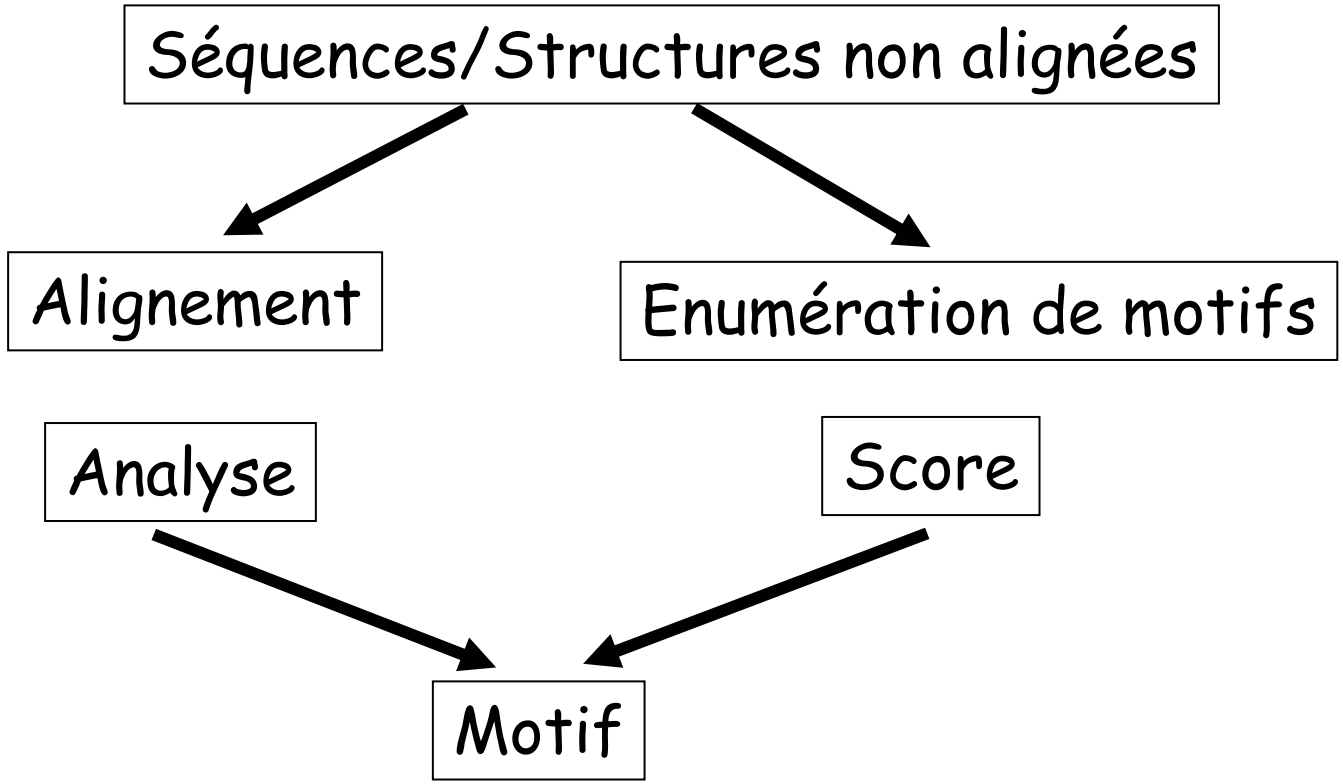
- **Pattern driven:**

  enumerate all (or some) patterns up to certain complexity (length), for each calculate the fitness, and report the best

- **Sequence driven:**

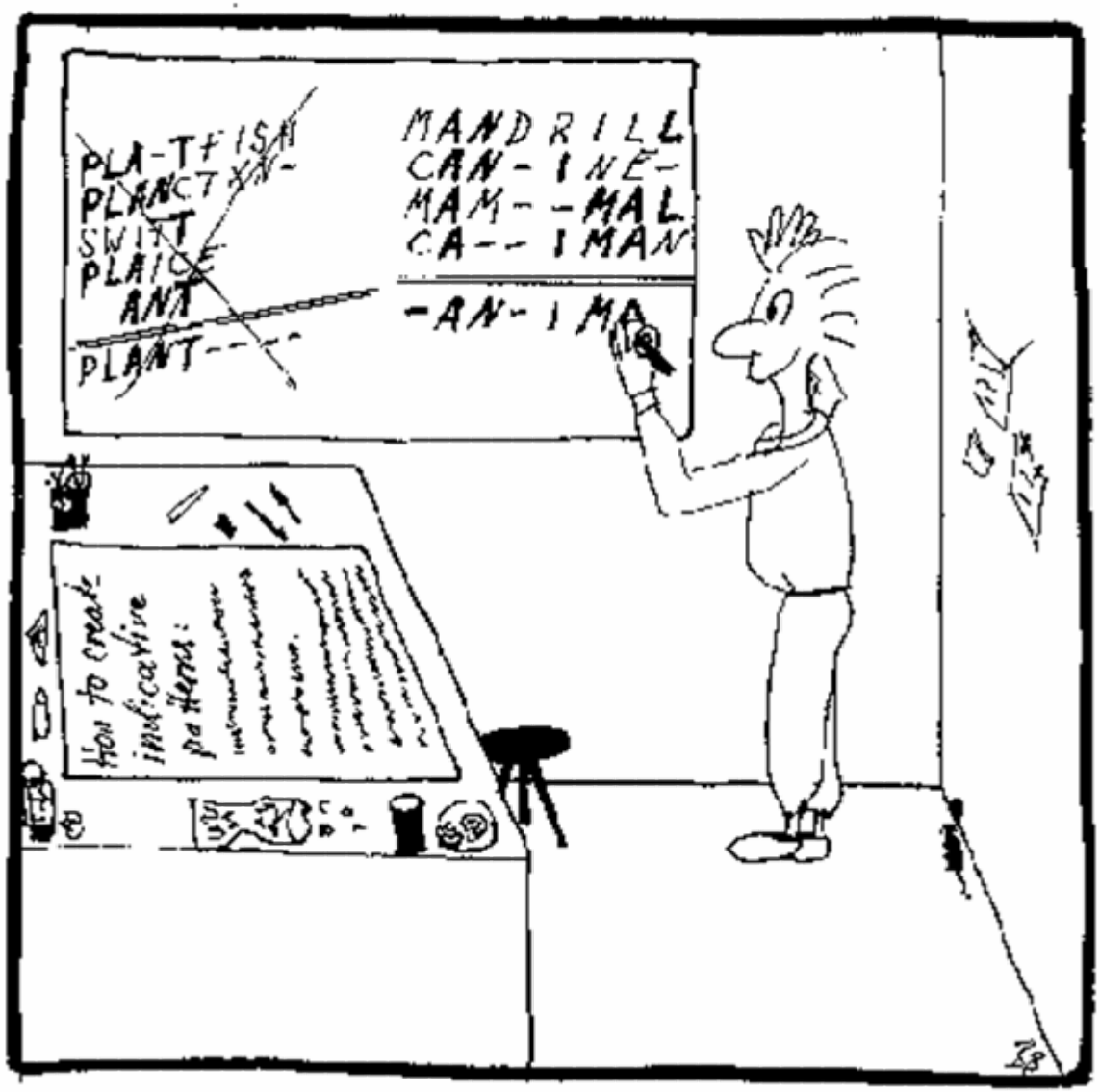
  look for patterns by aligning the given sequences

Séquence driven



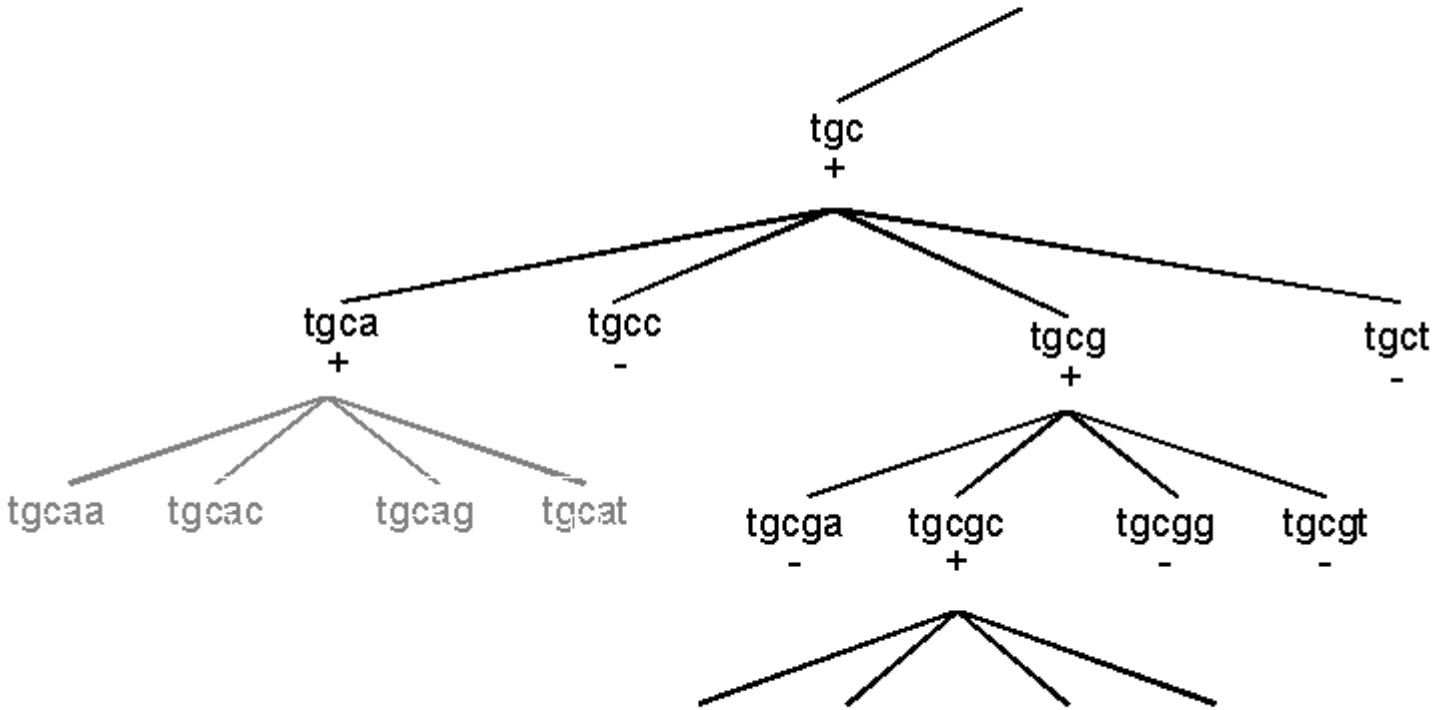
Pattern driven

How we develop Prosite patterns!



Brigitte Boeckmann / 1995

Pattern Driven - pruning the search space

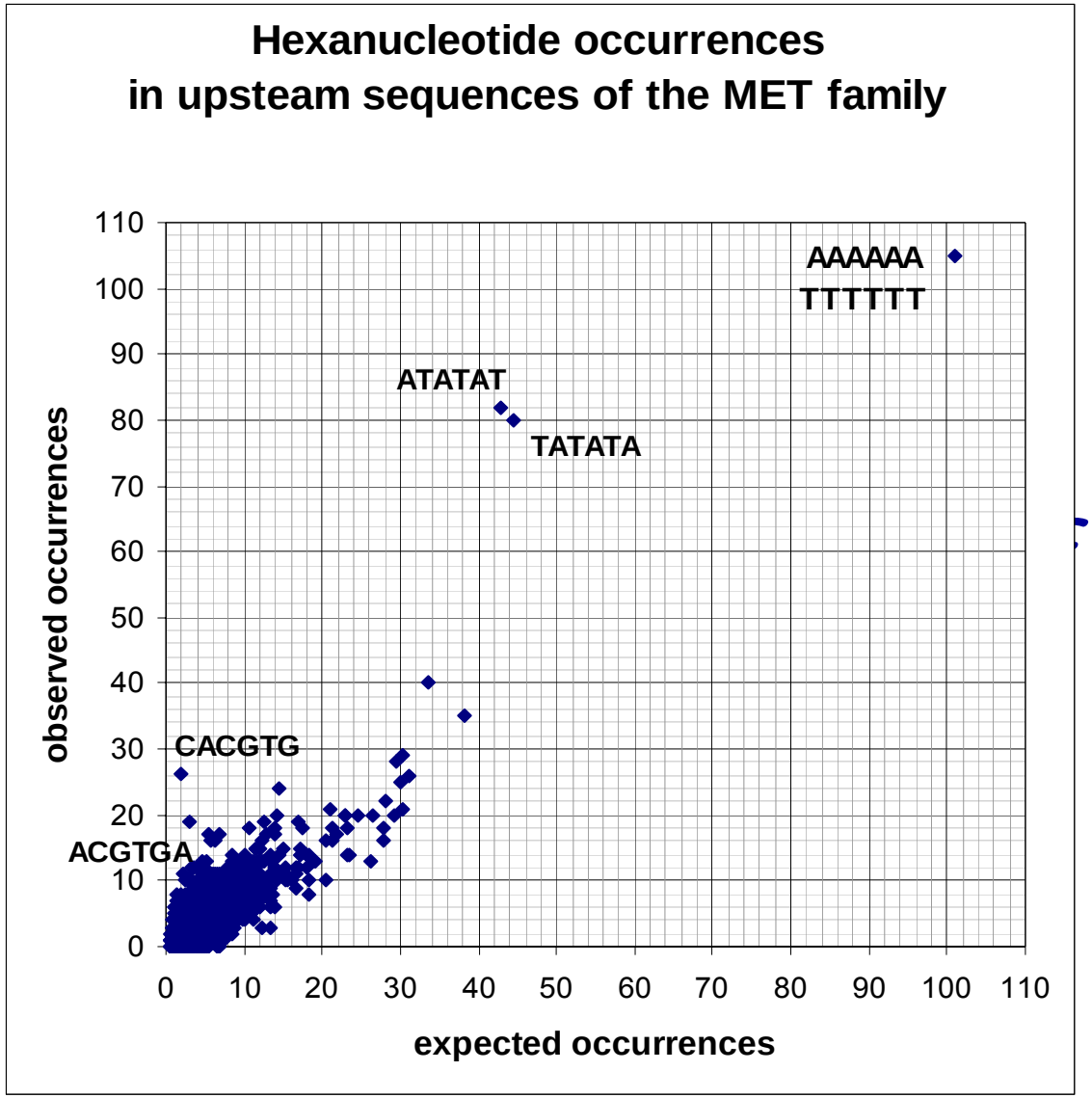


A first useful discovery method : How to discover patterns from scratch?

- Example of study involving a simple combinatorial search and a simple motif evaluation : RSA Tools (Jacques van Helden, Shoshana Wodak UCMB-ULB);

- Search of transcription sites in sequences upstream of families of genes of transcription factors in yeast.

- http://rsat.ulb.ac.be/rsat/



Idea:

Statistical test.
Count hexanucleotides in the family and in the complete genome.
Take as a null hypothesis a binomial law

Space :

2080 patterns

RSA tools : Space of Hypotheses

- Hypotheses space : set of possible motifs.
  Must be chosen with biological relevance.

- Initial idea : motif = words of size k
  4^k possibilities on DNA
  k=6 Size space = 4096 words
- But it is not possible to distinguish the DNA strands :
  one must rather consider pairs
  k=6 Size space = 2080 pairs of words

Test of hypotheses on sequences

Assume n sequences Seq_i and a word w present obs times in these sequences. The size of the space of word hypotheses is NH.

The probability Pw of w is estimated by the frequency on a set of non coding regions of the genome at hand (or a close genome, or…).

$$NW = 2\sum_{i=1}^{n}\left(|Seq_i| - |w| + 1\right)$$ Max number of words of size w (2 strands)

$$\text{NumberOccurrencesPredicted}(w) = NW \cdot Pw$$

$$\Pr(nbobserved(w) = k) = \frac{NW!}{k!\,(NW-k)!}\,Pw^{k}\left(1-Pw\right)^{NW-k}$$

$$\Pr(nbobserved\ (w) \geq k) = \sum_{i=k}^{NW}\Pr(nbobserved\ (w) = i)$$

$$E-value = NH \cdot \Pr(nbobserved\ (w) \geq obs)$$

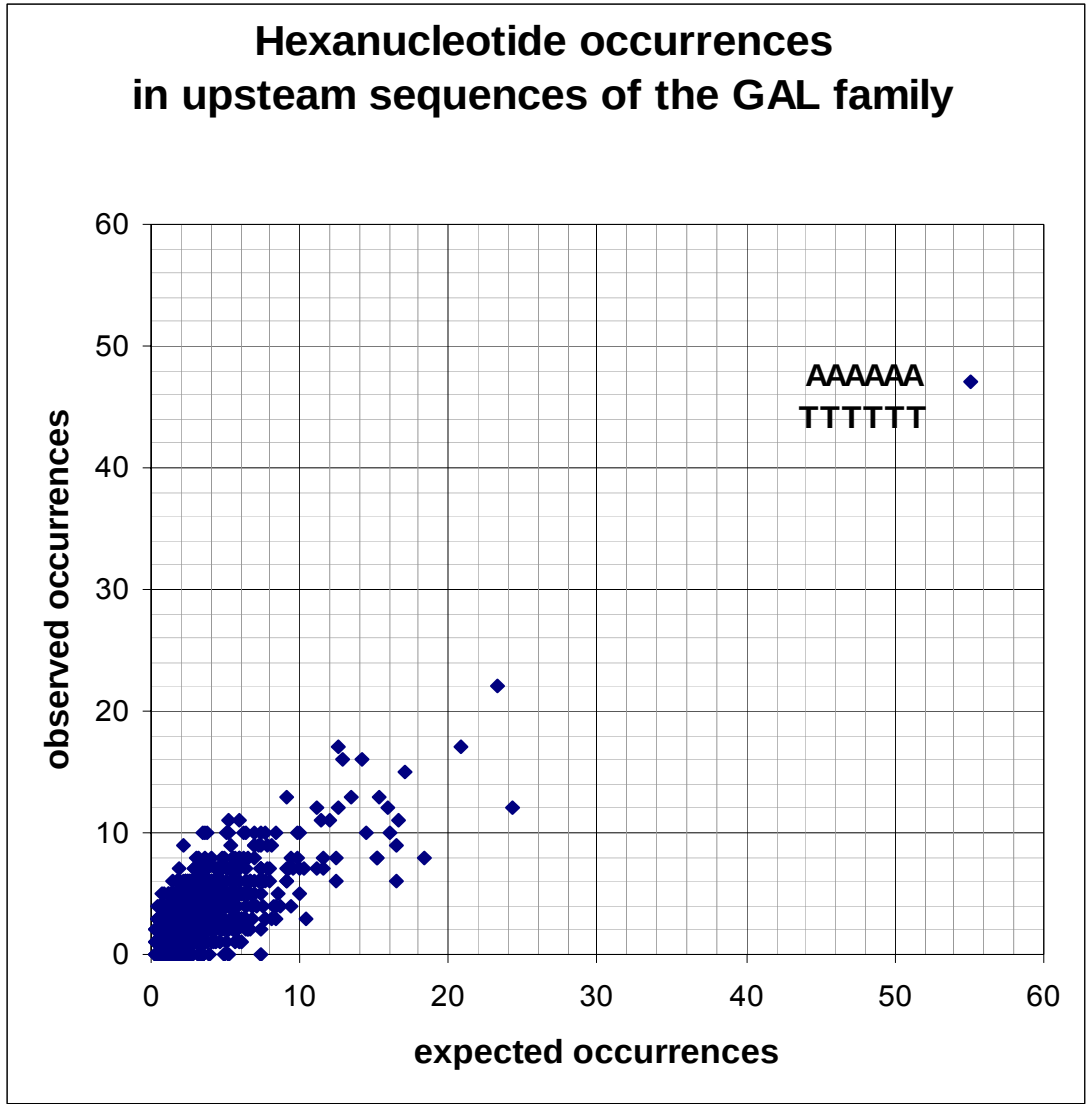
The significativity score is simply - $\log_{10}$(E-value)

Results for MET family

Known pattern	Factors
TCACGTG	Cbf1p/Met4p/Met28p
AAAACTGTGG	Met31p; Met32p

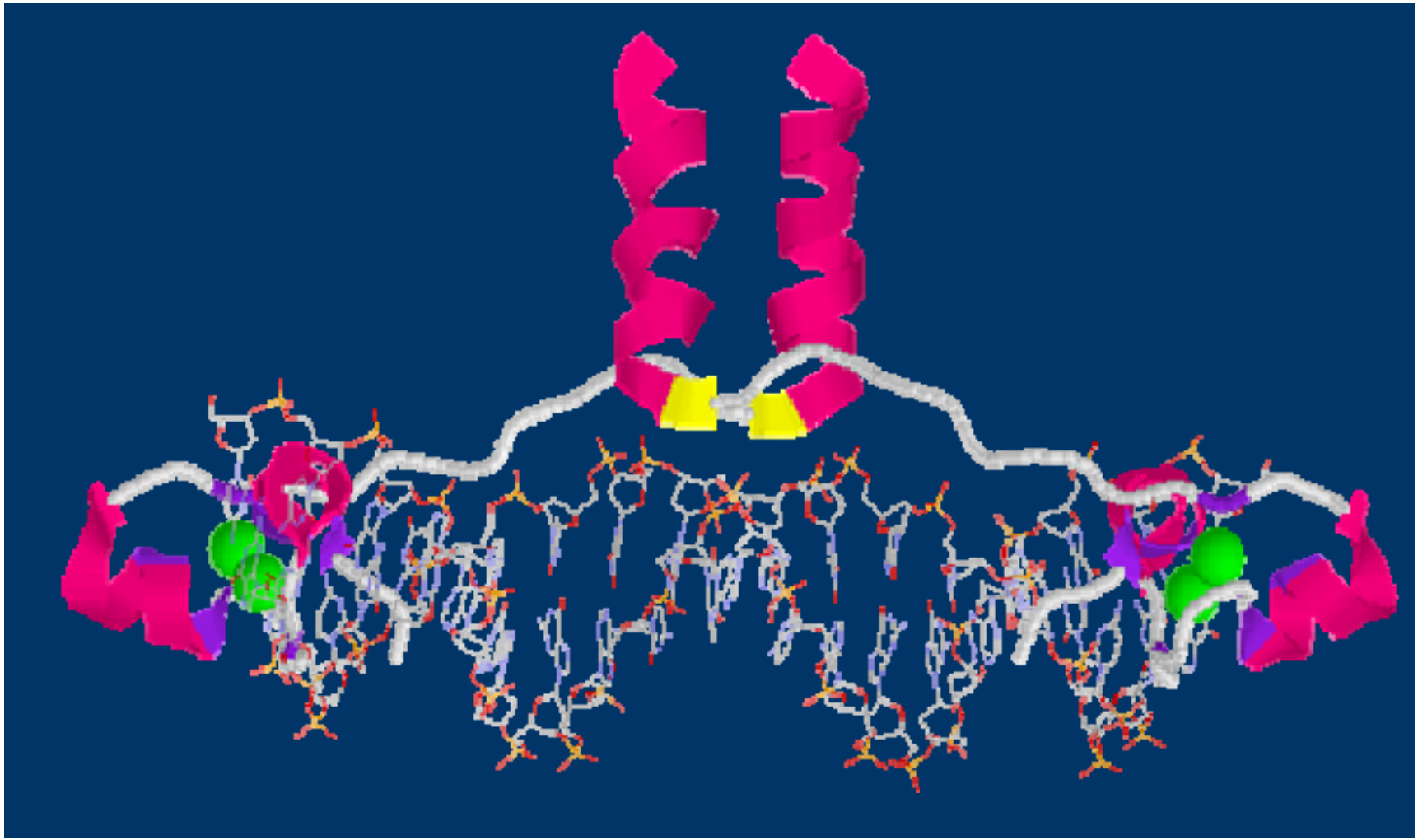
Family	Genes	Patterns	match seq	occ	exp occ	score
MET	MET3	..CACGTG	9	26	2.0	7.0
	MET25	.TCACGT..	9	19	2.9	6.1
	MET2	GTCACG...	6	8	1.4	0.7
	MET16					
	MET19	...TGTGGC	7	10	2.4	0.5
	MET14	..CTGTGG.	8	11	2.1	1.6
	MET5	.ACTGTG..	9	12	3.2	0.6
	MET6	AACTGT...	10	17	5.5	0.9
	SAM1					
	SAM2	.ATATAT	10	82	42.3	0.8
		TATATA.	11	80	43.9	0.2
		GCTTCC	7	12	3.5	0.2

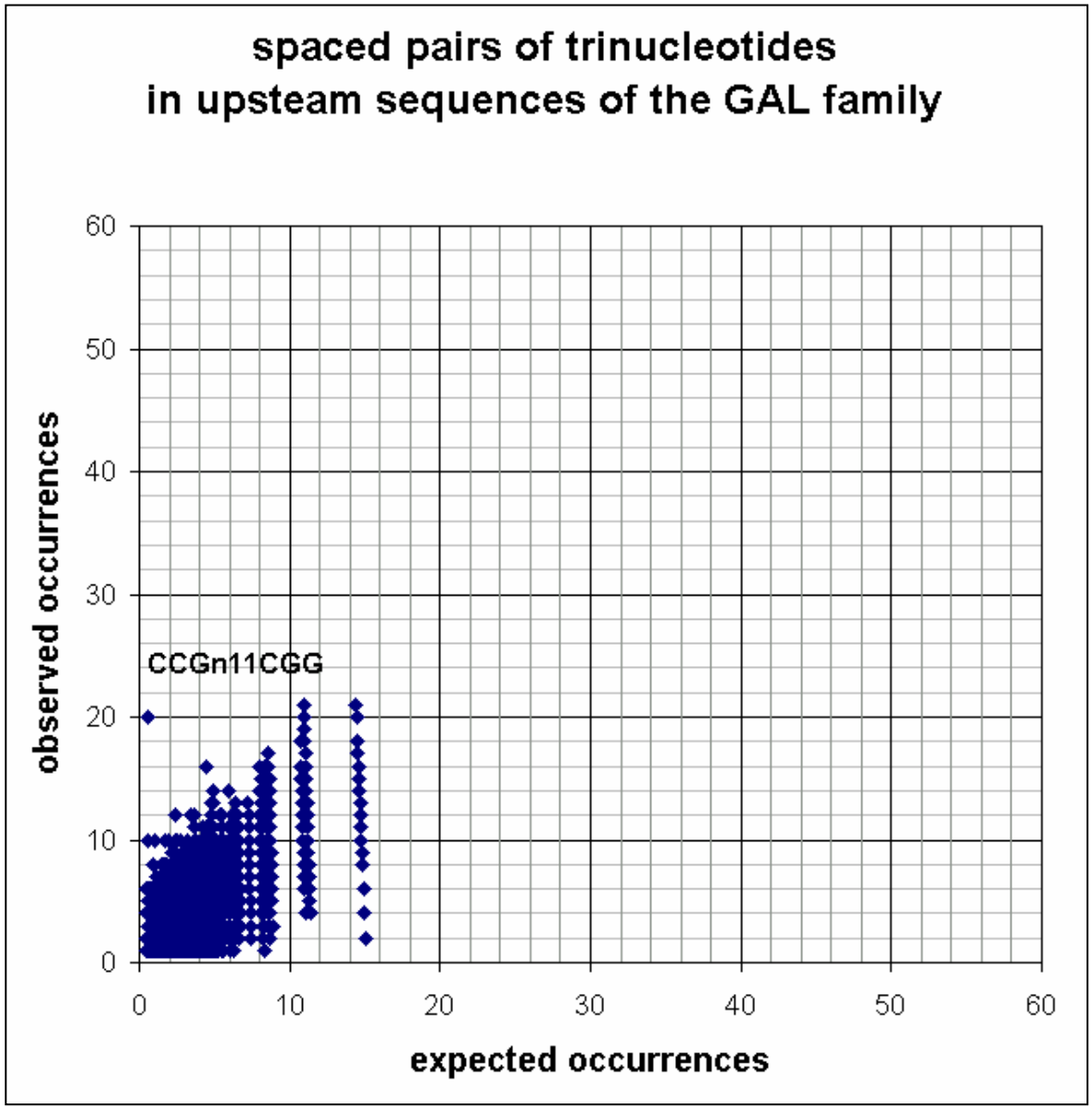
A case that does not work so well...



family GAL (genes expressed in presence of galactose): not even a single significant pattern !

Structure of the interface Gal4p-ADN





Solution : introduction of a gap in the pattern, between 0 and 16

Space :
35360 or 1632 if one takes into account only repeats or palindromes

Pb : the method is not general enough…

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"

The question is: how to model "conservation"?

Source MF Sagot

"Horizontal" conservation measure

With or without "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

Source MF Sagot

“Vertical” 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

Source MF Sagot

"Horizontal" conservation measure

With or without "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

Source MF Sagot

Approaches using a "horizontal" conservation measure

Objective

Given a model (alphabet for the motifs and properties such as quorum and maximum difference rate allowed), find all motifs which satisfy the properties

It is an enumeration problem, which produces in general various (often a great number of) solutions

Algorithm

An exhaustive approach is possible

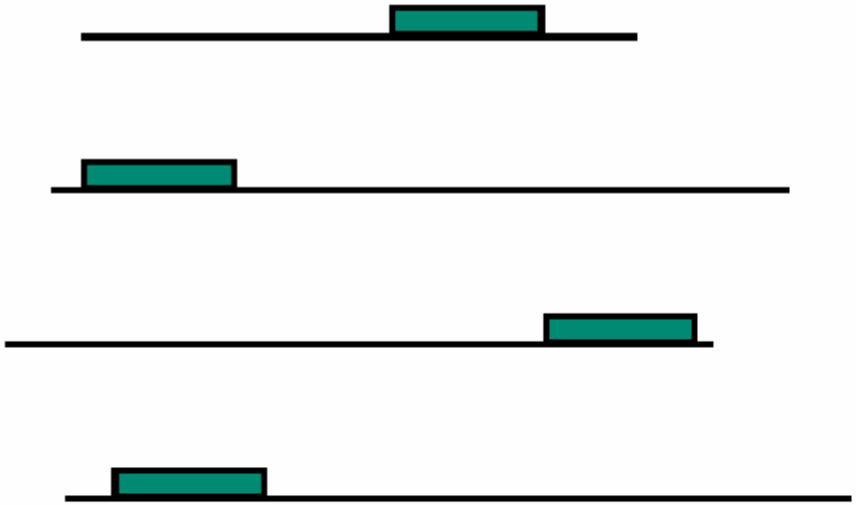
Time complexity depends on properties (linear in sequence length, exponential in number of errors)

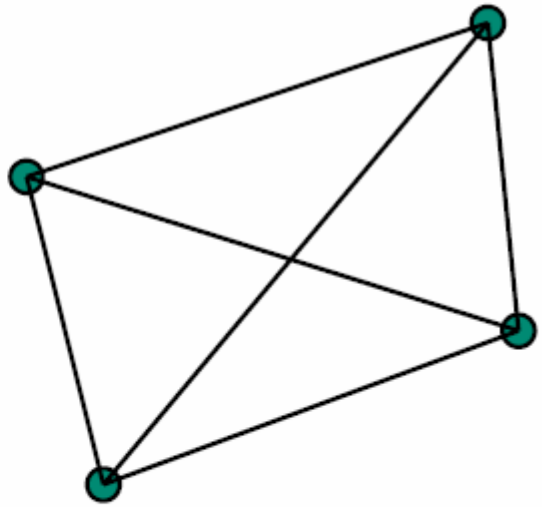
Source MF Sagot

"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)

Source MF Sagot

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:

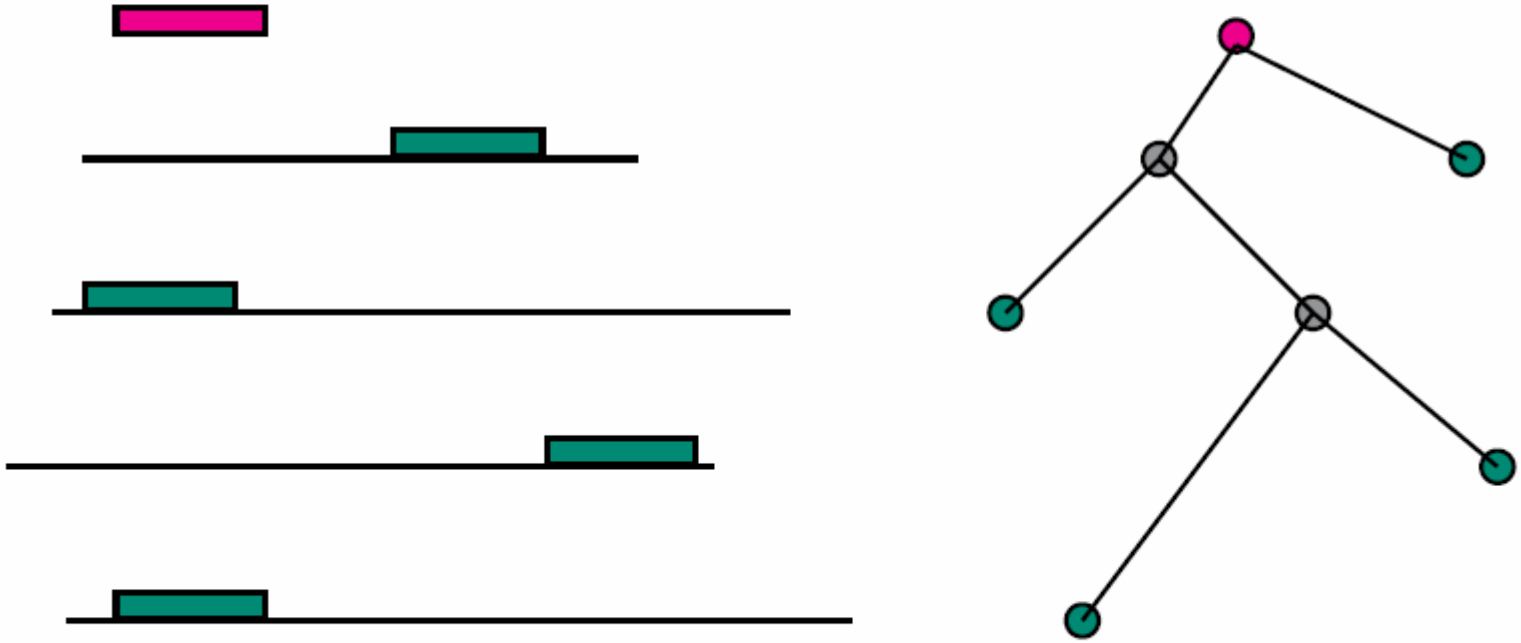
all N-cliques in G

Source MF Sagot

"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)

Source MF Sagot

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:

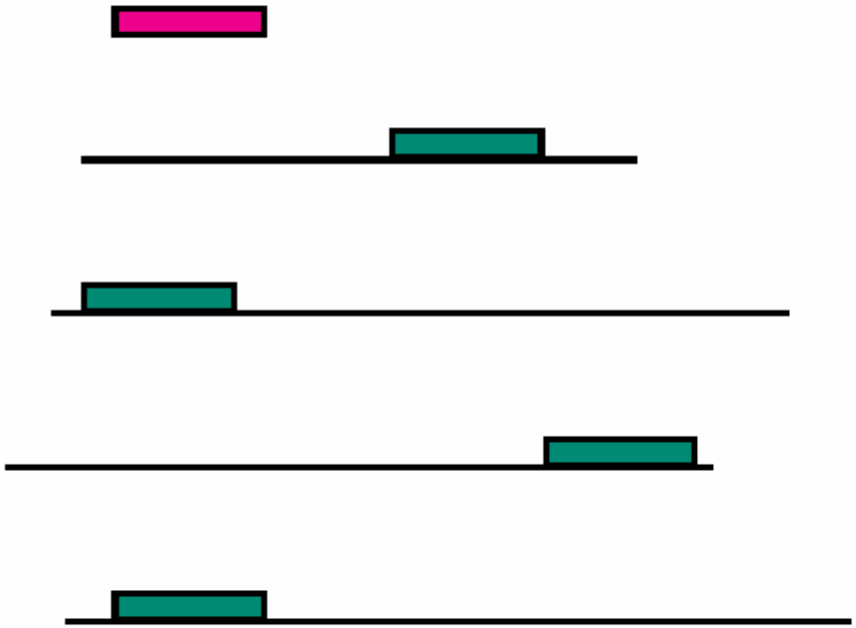
ancestors of all most parsimonious trees whose total length is at most e

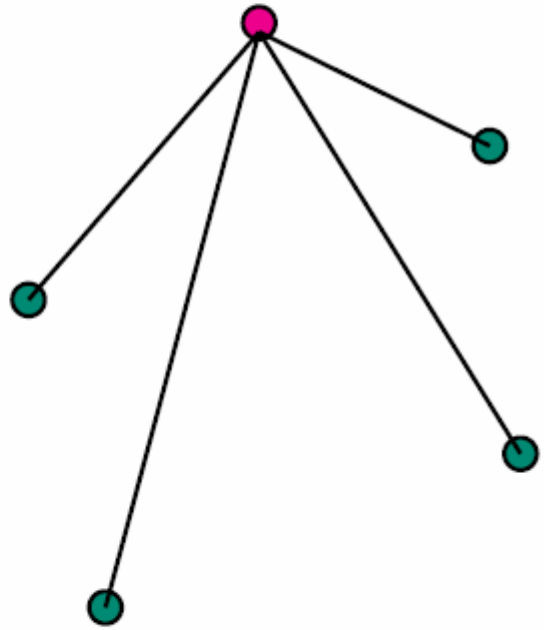
Source MF Sagot

"Conserved" elements: "Models"

Model = Motif + constraints (main one: max. diff. 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)

Source MF Sagot

Motif

"simple pattern"	"limited reg. expression"
TATAAT	TA[AT]N[AT]T
TTGACA	
RFMCP	[ILMV][ASG]XXC[ILMV]H

where N or X is the don't care symbol

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

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

(amino acids)

Source MF Sagot

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)

Source MF Sagot

Un exemple de recherche combinatoire de motifs consensus

Smile [Marsan et Sagot 2000]

Recherche exacte des mots présents :

- dans au moins q séquences (quorum)
- avec au plus e erreurs

La notion de modèle

Modèle

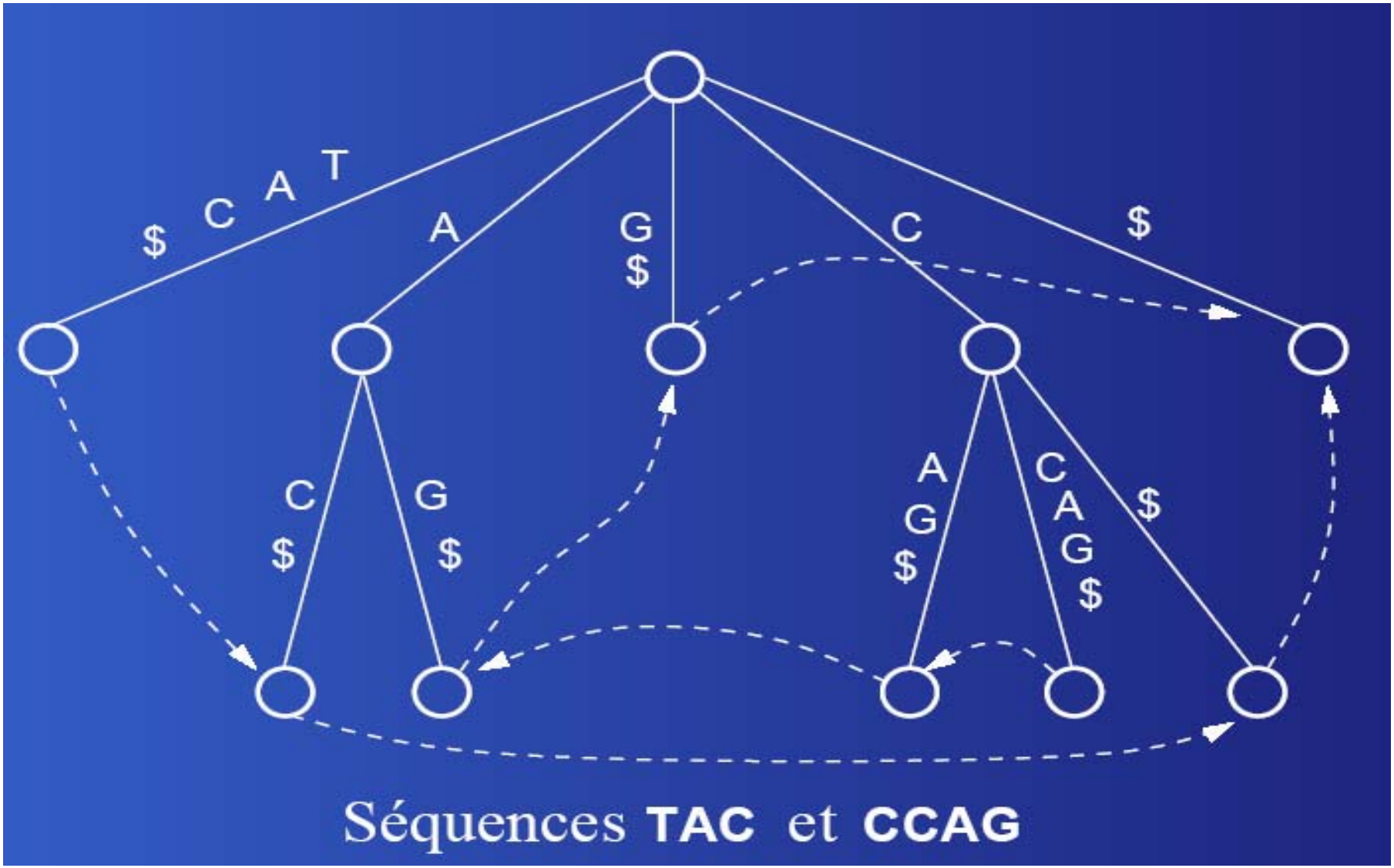
Motif consensus
(distance maximale)
+
Quorum
(en occurrences ou
en séquences)

Propriété

m non valide $\Rightarrow \forall x, y \in \Sigma^*, xmy$ non valide

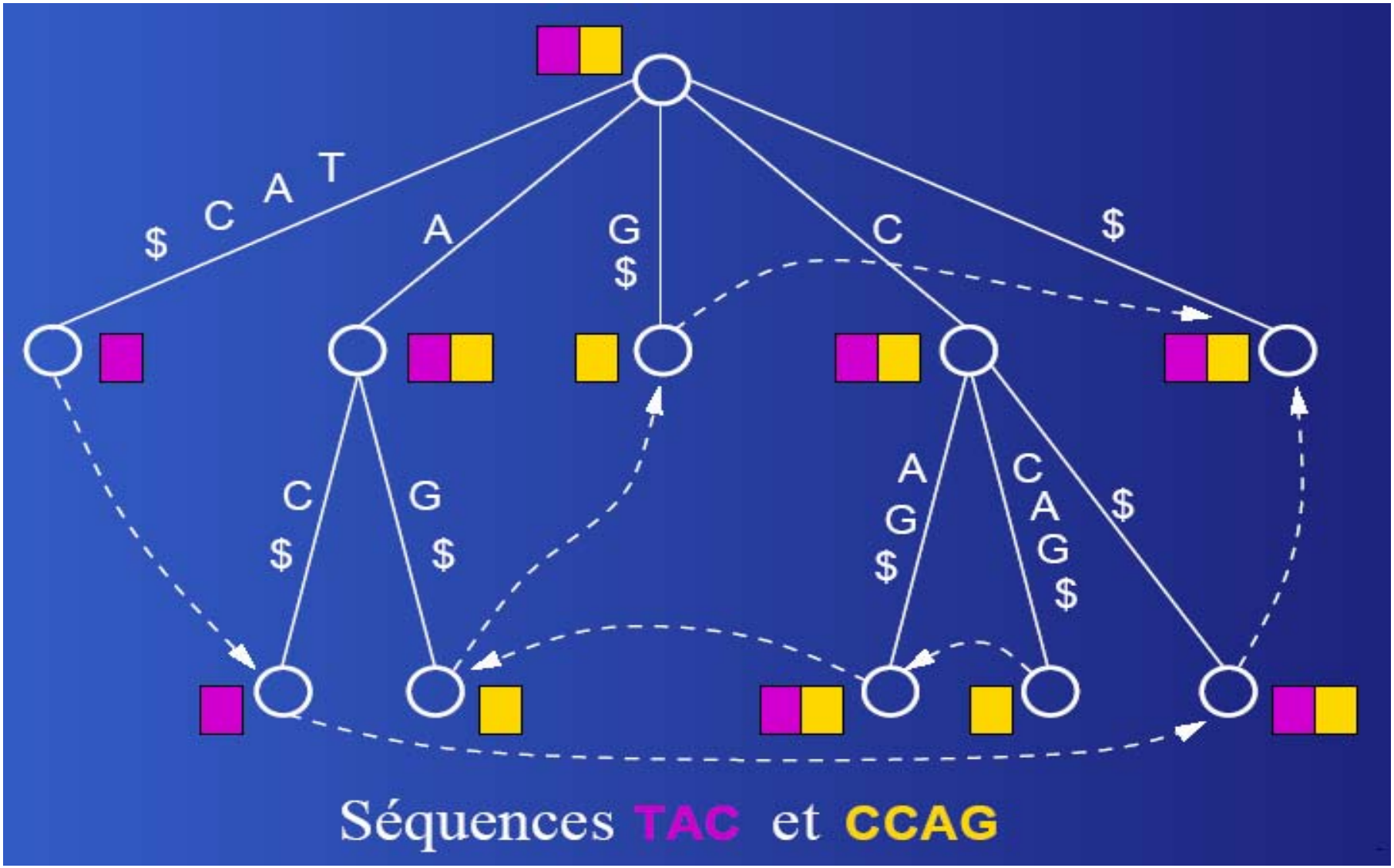
Source Laurent Marsan

Structure de données



Source Laurent Marsan

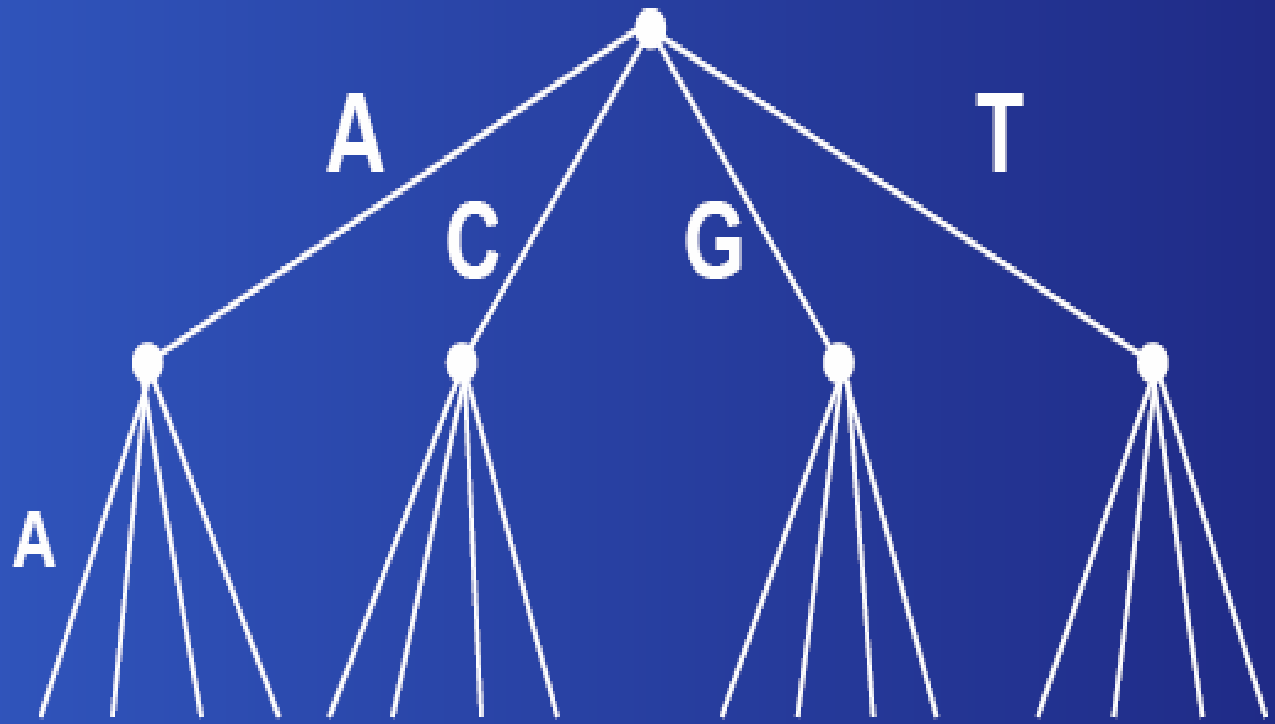
Structure de données



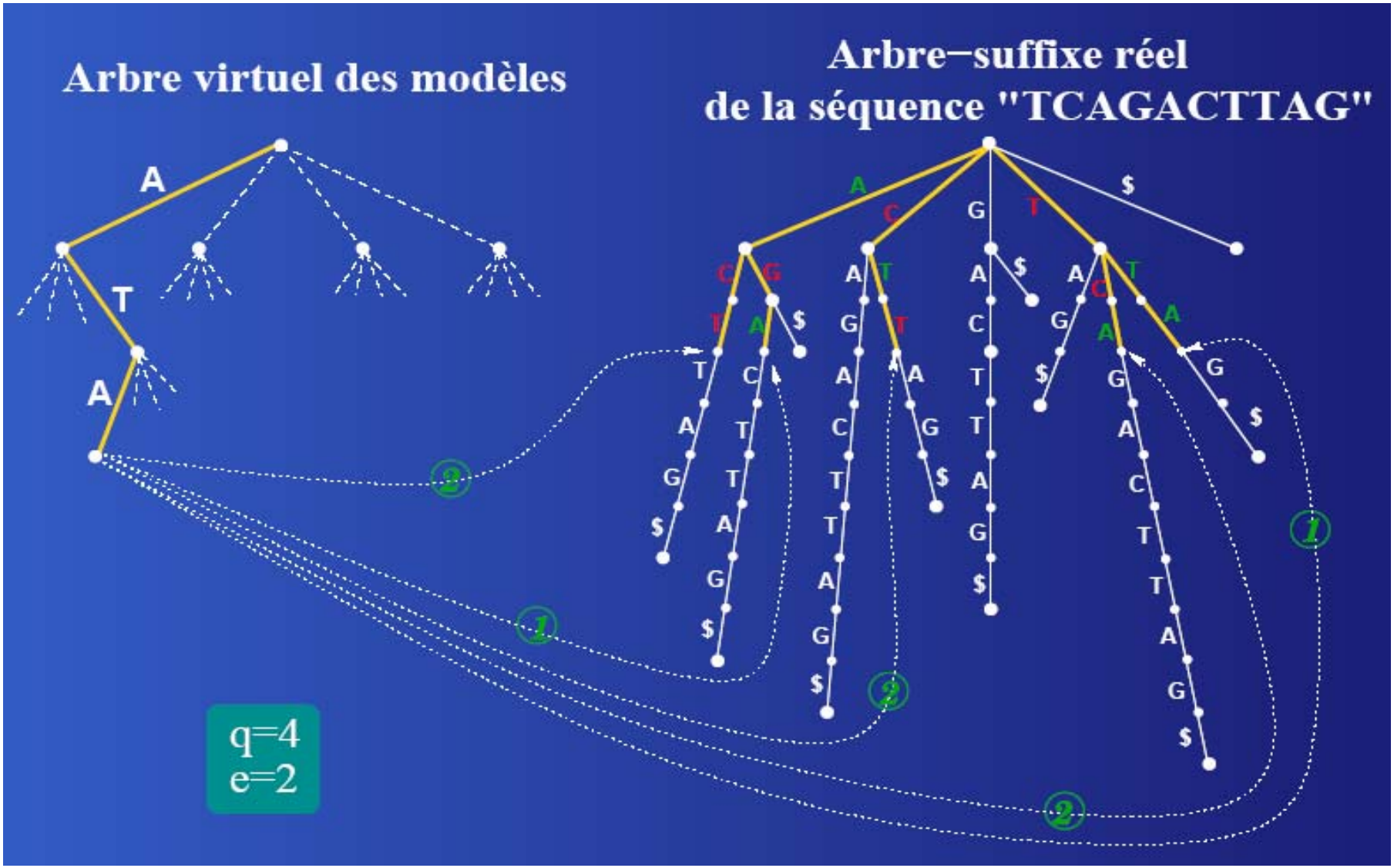
Source Laurent Marsan

Algorithme de base

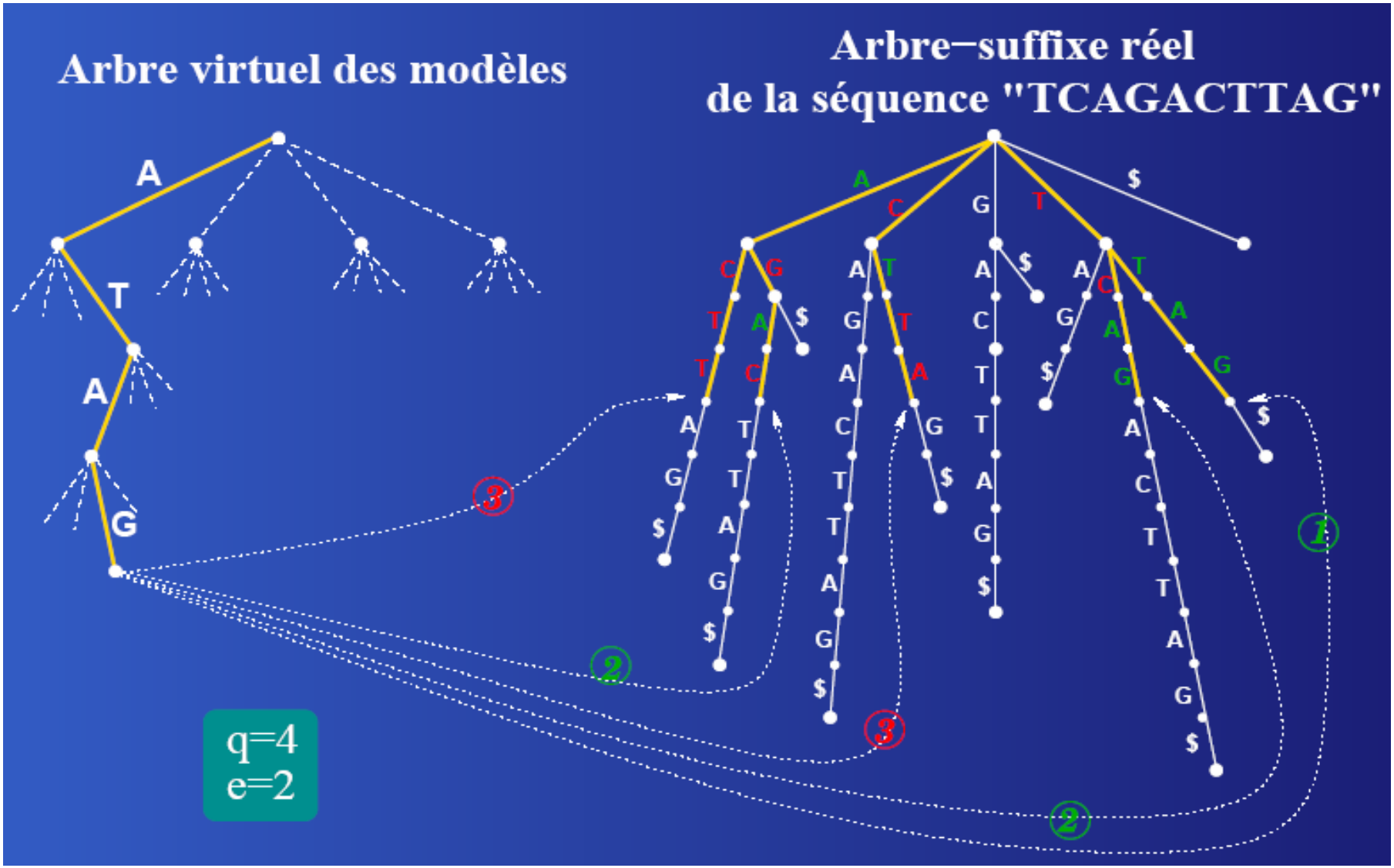
Simuler le parcours prefixe d'un arbre virtuel des modèles



Source Laurent Marsan



Source Laurent Marsan



Source Laurent Marsan

Arrêt de la descente récursive

- quorum non respecté
- longueur maximum atteinte

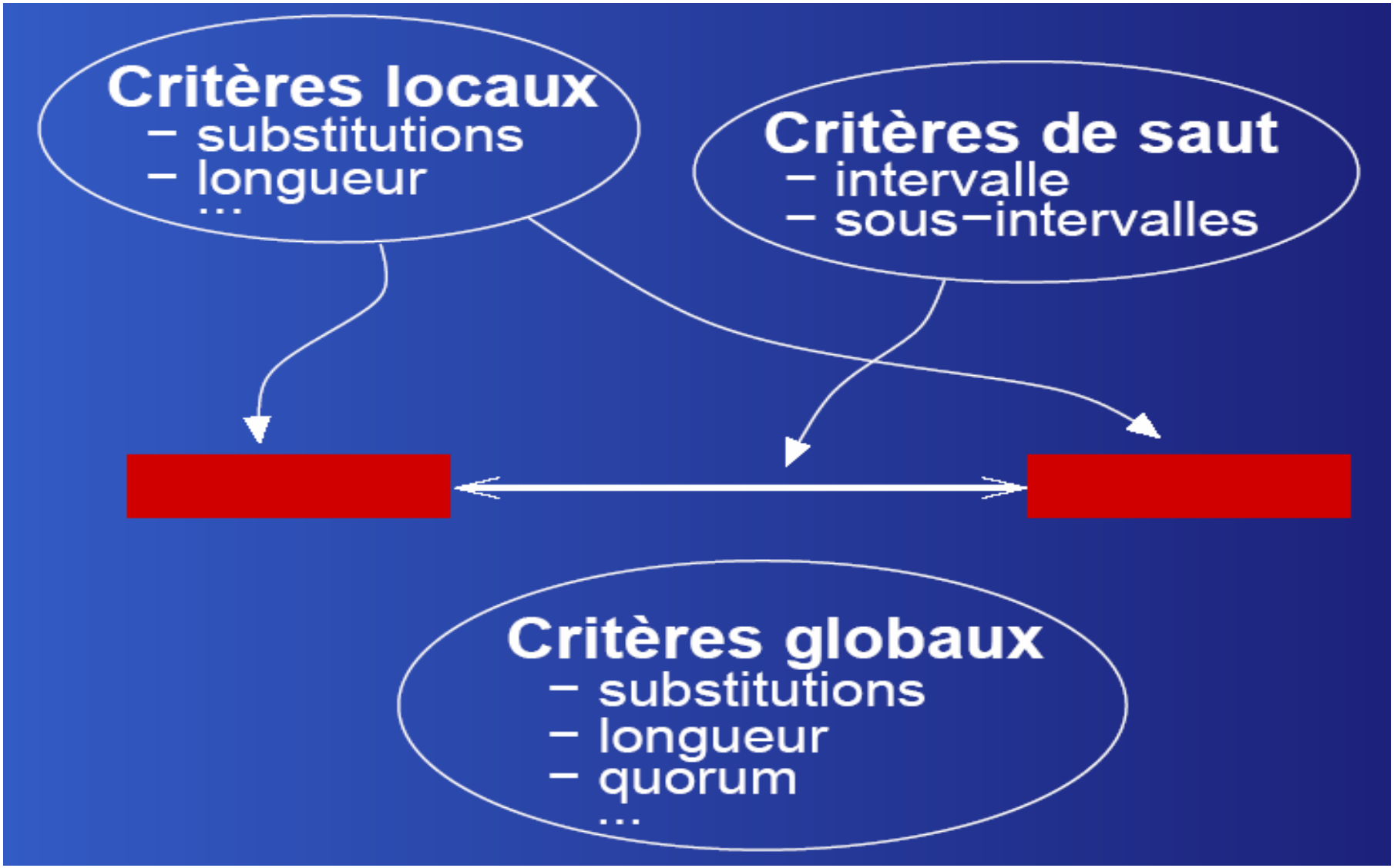
Complexités

Temps : $\mathcal{O}(nN^2\mathcal{V}(e,k))$
Espace : $\mathcal{O}((k+N)nN)$

où $\mathcal{V}(e,k) = \sum_{j=0}^{e} \binom{k}{j} (|\Sigma|-1)^j \leq k^e|\Sigma|^e$

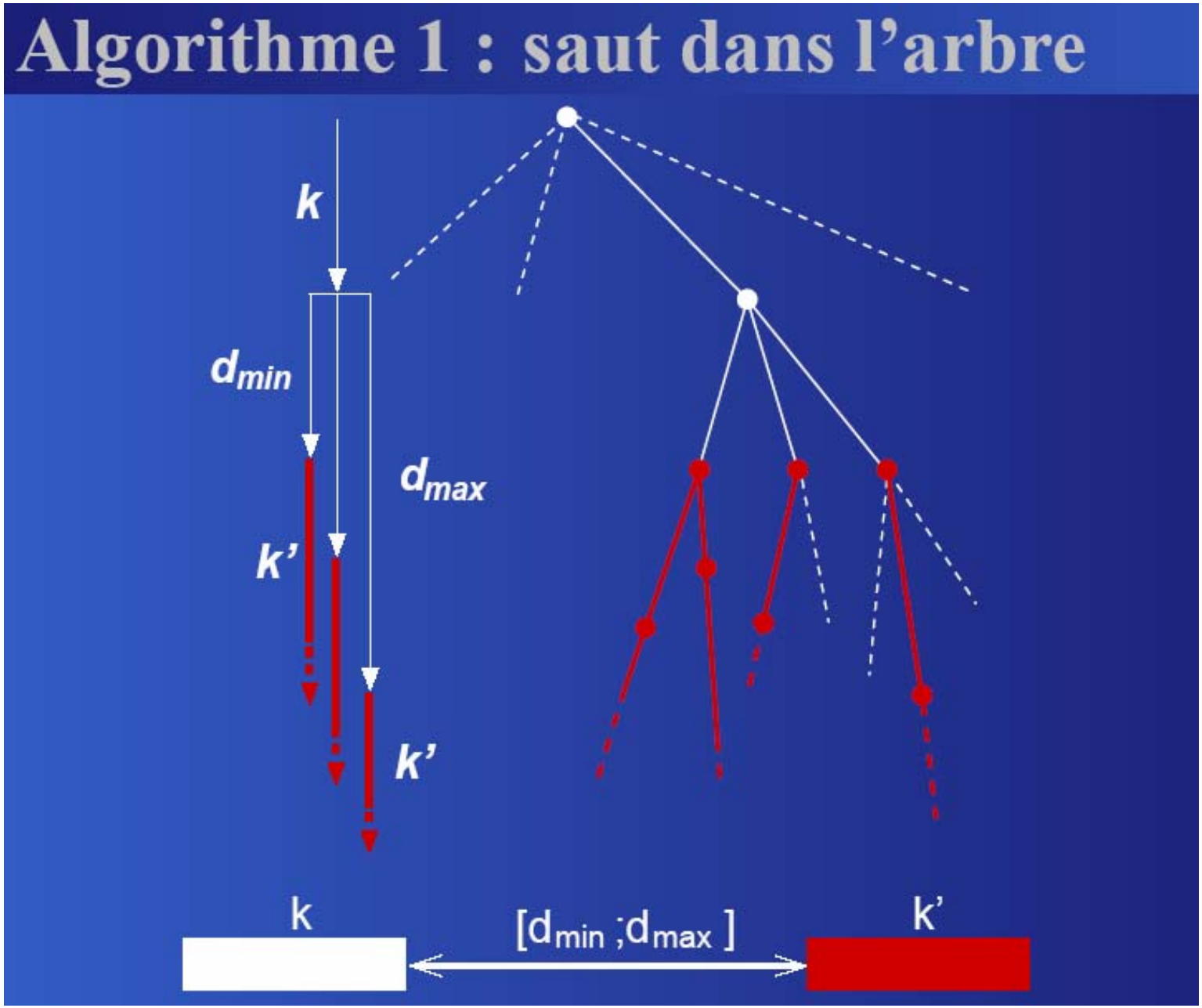
Source Laurent Marsan

Dyades



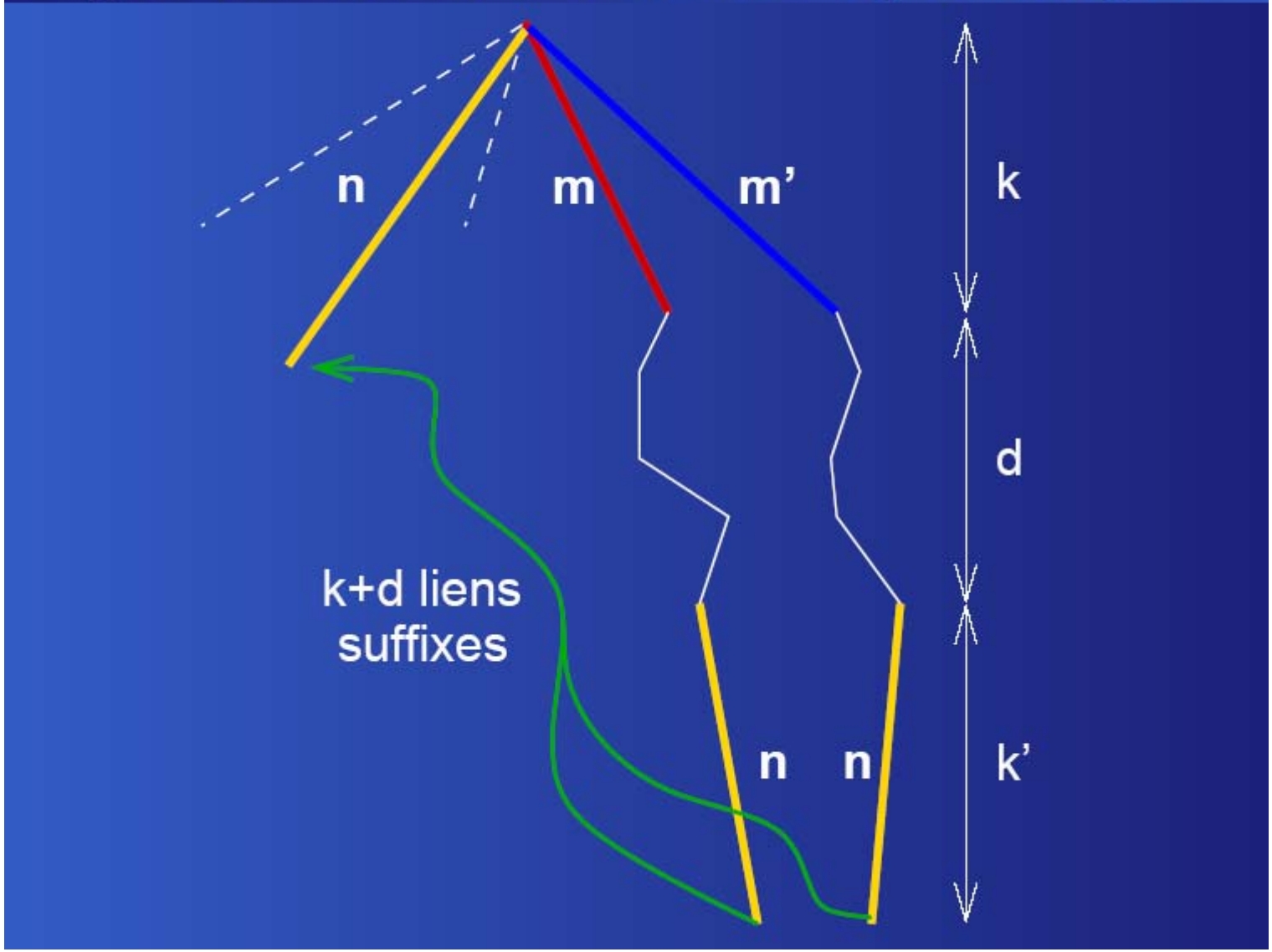
Source Laurent Marsan

Algorithme 1 : saut dans l'arbre

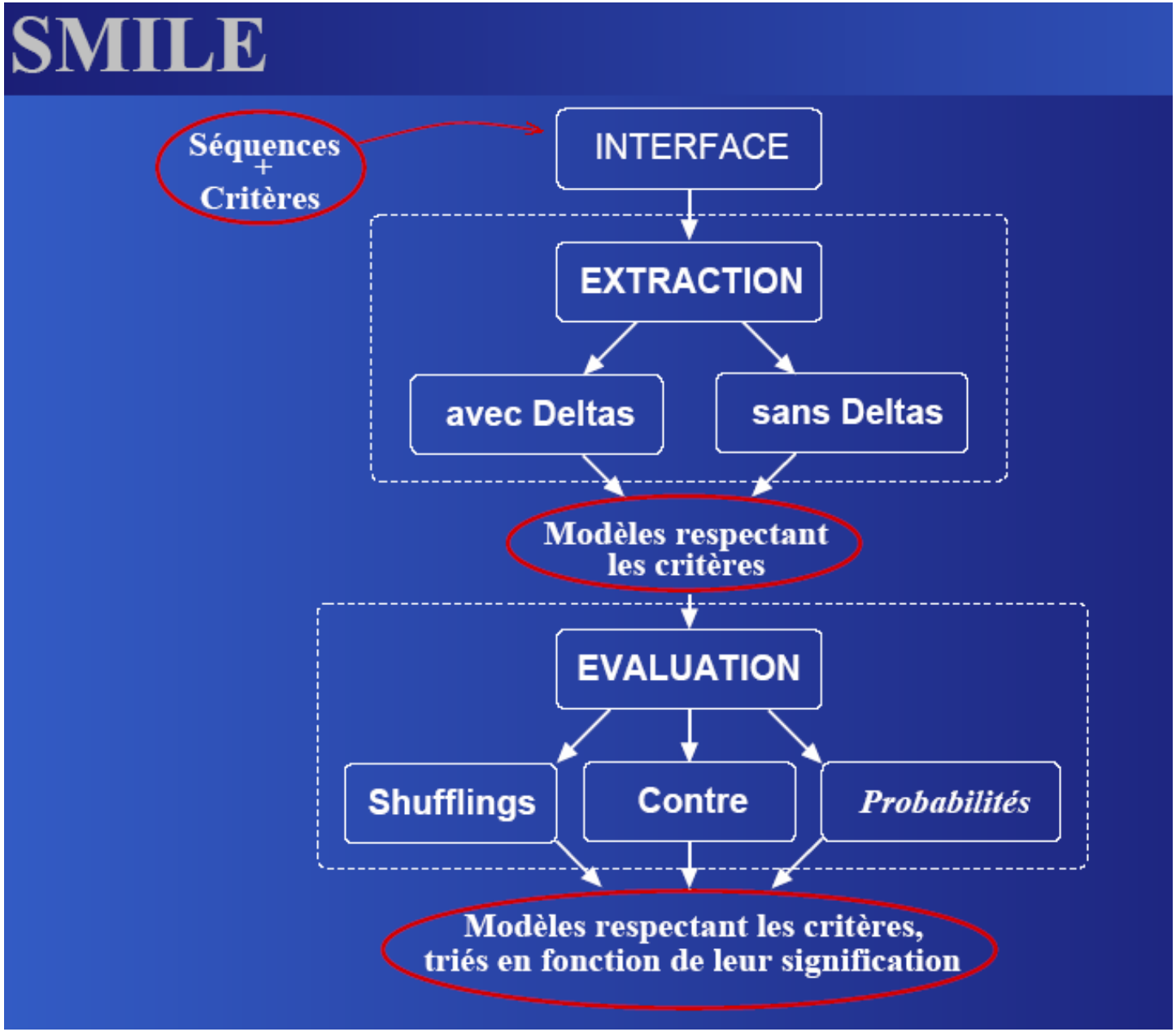


Source Laurent Marsan

Algorithme 2 : un arbre dynamique



Source Laurent Marsan



Source Laurent Marsan

SMILE

Model	%right	#right	%shfl.	#shfl.	Var.	Chi2	Z-score
TTGCCA_TTATAAT	50.38%	66	11.81%	15.47	2.95	45.48	17.14
TTGACA_TATAATA	58.02%	76	17.35%	22.73	3.75	46.12	14.20
TTGACA_GTATAAT	51.15%	67	12.21%	16.00	3.60	45.87	14.18
TTGACT_TATAATA	55.73%	73	15.84%	20.75	4.05	45.35	12.91
TTGACAT_ATAATA	53.44%	70	16.49%	21.60	3.89	39.32	12.43
TTCACA_TATAATA	51.15%	67	15.26%	19.99	3.87	38.03	12.14
ATTGTC_TATAATA	50.38%	66	15.92%	20.85	3.73	35.11	12.09
TTGACAA_ATAATA	53.44%	70	19.05%	24.95	3.75	33.52	12.02
ATTGAC_TATAATA	51.15%	67	16.47%	21.58	3.80	35.19	11.94
TTTACA_TATAATA	61.07%	80	22.22%	29.11	4.27	40.67	11.91

Source Laurent Marsan

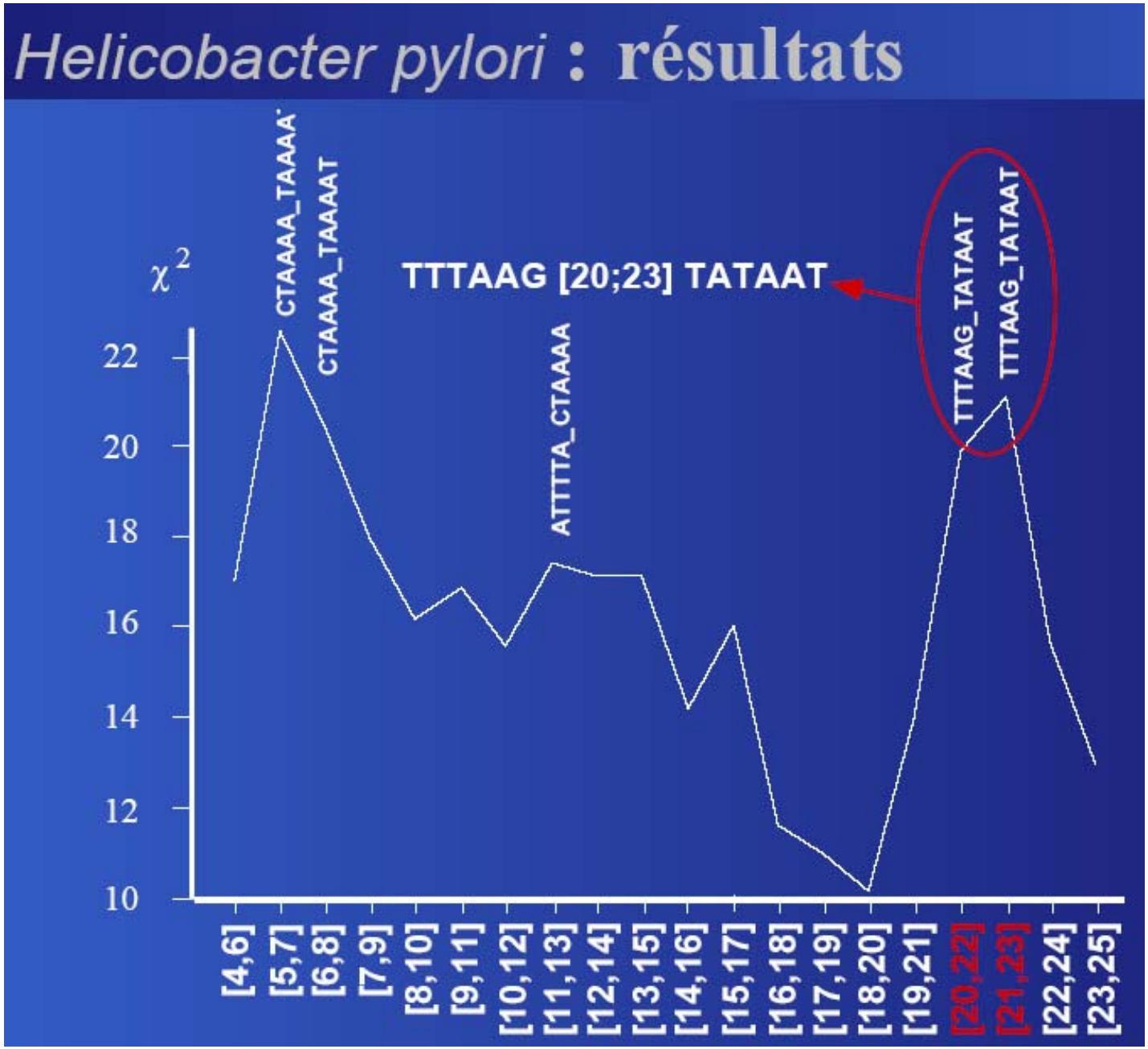
Helicobacter pylori : protocole

Séquencée en 1997 : 1,6Mb, 1590 gènes

- données directement issues des banques
- sélection d'au plus 300 bases en amont des gènes divergents
- 308 séquences, 52Kb
- inférences de deux boîtes avec delta :
  - longueur 6 ou 7 pour chaque boîte
  - une substitution globale
  - quorum 6%
  - intervalle $[4; 25]$
  - sous-intervalles de largeur 2
- *shuffling* conservant les di-nucléotides

Source Laurent Marsan

Helicobacter pylori : résultats



Source Laurent Marsan

Helicobacter pylori : résultats

Intervalle $[20; 23]$, boîtes étendues, un joker N

Model	%right	#right	%shfl.	#shfl.	Var.	Chi2	Z-score
TTTTAAG_GTATNAT	5.52%	17	0.49%	1.52	1.16	13.34	13.36
ATTATAG_NTTAAAA	6.82%	21	1.06%	3.26	1.64	13.50	10.84
TTTTAAG_GTANAAT	6.17%	19	0.69%	2.12	1.62	13.97	10.45
TTTTAAG_NTATAAT	8.12%	25	1.26%	3.89	2.08	16.18	10.16
TATTATA_GNTAAAA	5.52%	17	0.76%	2.35	1.47	11.45	9.95
ATTATAC_NTTAAAA	5.84%	18	0.90%	2.76	1.60	11.58	9.55
TATTATA_NCTTAAA	5.52%	17	0.74%	2.29	1.57	11.58	9.40
AATTATA_CTTAANA	5.52%	17	0.88%	2.72	1.54	10.68	9.25
GTTTTTA_GTTANAA	5.52%	17	0.74%	2.28	1.60	11.60	9.19
TATTCTA_NTTAAAA	6.17%	19	1.09%	3.36	1.74	11.35	9.00

Source Laurent Marsan

Approaches using a "vertical" conservation measure

Objective

Find the set of words that is the "most surprising possible"

It is an optimisation problem, which in general leads to an unique solution

Algorithm

Only approach possible: test all set of words and, for each of them, calculate the value of the formula

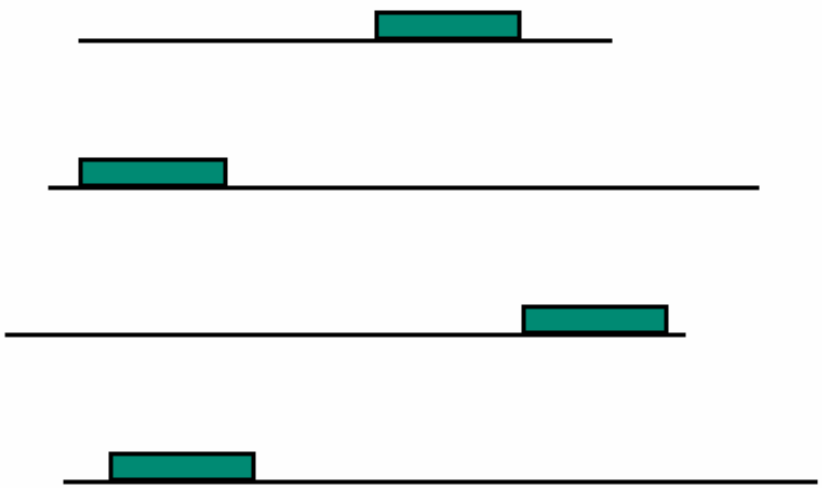
Too time consuming ($O(n^N k)$), one must therefore use heuristics

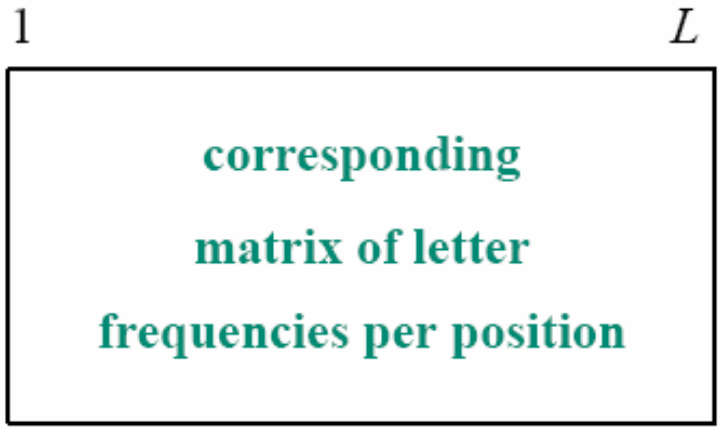
Source MF Sagot

"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)

Source MF Sagot

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:

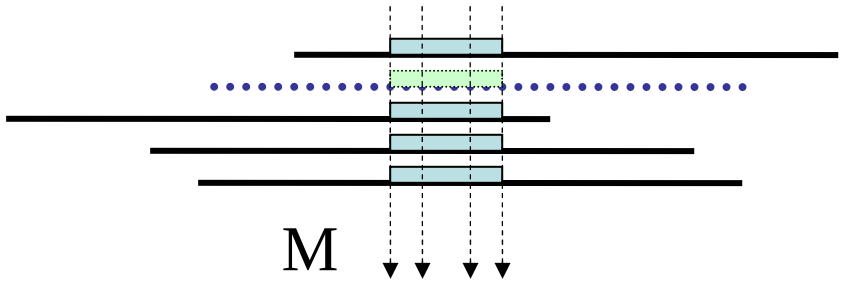
the set(s) of N words of length k, each belonging to a distinct sequence, that has maximum relative entropy

Source MF Sagot

Heuristiques

- Expectation-Maximization
  - MEME, Bailey, 1995
- Gibbs Sampling
  - Lawrence et al, 1993
  - Thijs et al, 2001
- Algorithme glouton
  - (w)consensus, Hertz et al, 1999
- Projection
  - Buhler et al, 2000

Simplified Principle of deterministic algorithm EM



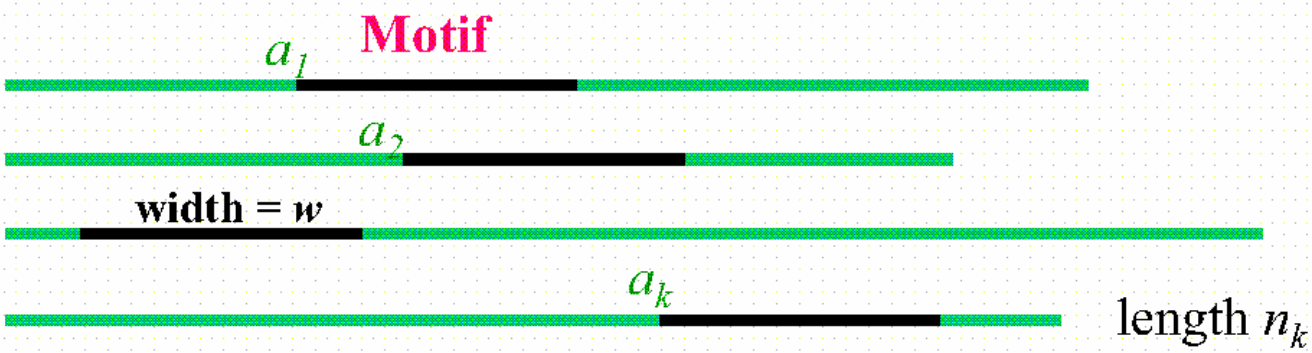
Words of size k

(Used in MM of MEME)

- Initialization : For all sequences, select at random a site (window of size k).
- Do
  - « Expectation » Compute on the set of sites of sequences a model M from the frequency matrix (letters x positions).
  - « Maximization » For each sequence
    Compute a likelihood score for each site (window of size k), based on its probability with respect to M and the Background (vector estimated on a larger sample). Select the site of best likelihood score for the sequence
- Until stability of model M

Gibbs Sampling

Motif Alignment Model



The missing data: Alignment variable: $A=\{a_1, a_2, \ldots, a_k\}$

- Every **non-site positions** follows a common multinomial with $\boldsymbol{p_0}=(p_{0,1}, \ldots, p_{0,20})$
- Every position i in the motif element follows probability distribution $\boldsymbol{p_i}=(p_{i,1}, \ldots, p_{i,20})$

Source: Jun Liu

Gibbs Sampling (cont'd)

The Algorithm

- Initialized by choosing random starting positions

$$a_1^{(0)}, a_2^{(0)}, \ldots\ldots, a_K^{(0)}$$

- Iterate the following steps many times:
  - Randomly or systematically choose a sequence, say, ***sequence k***, to exclude.
  - Carry out the ***predictive-updating*** step to update a_k
- Stop when no more observable changes in likelihood

Source: Jun Liu

Gibbs Sampling Example

- The following slides illustrate Gibbs sampling to discover a motif in yeast DNA sequences.
- This example uses a sequence model that allows multiple sites per sequence.
- Columns are sampled as well as sites.

The Input Data Set

300-600 bp of upstream sequence
per gene are searched in
Saccharomyces cerevisiae.

Source: G.M. Church

The Target Motif

5′- TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATGAAAAATTCATGAG**AAAAGAGTCA**GACATCGAAACATACAT …*HIS7*

5′- ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCG**AAATGACTCA**ACG …*ARO4*

5′- CACATCCAACGAATCACCTCACCGTTATCG**TGACTCACTT**TCTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT …*ILV6*

5′- TGCGAAC**AAAAGAGTCA**TTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC …*THR4*

5′- ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATA**TGACTCATCC**CGAACATGAAA …*ARO1*

5′- ATTGAT**TGACTCATTT**TCCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA …*HOM2*

5′- GGCGCCACAGTCCGCGTTTGGTTATCCGGC**TGACTCATTCTGACTCTTTT**TTGGAAAGTGTGGCATGTGCTTCACACA …*PRO3*

AAAAGAGTCA
AAATGACTCA
AAGTGAGTCA
AAAAGAGTCA
GGATGAGTCA
AAATGAGTCA
GAATGAGTCA
AAAAGAGTCA
****

MAP score = 20.37 (maximum)

Source: G.M. Church

Initial Seeding

5′- TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAA**TGAAAAATTC**ATGAGAAAAGAGTCA**GACATCGAAA**CATACAT …*HIS7*

5′- ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCGAAATGACTCAACG …*ARO4*

5′- CACATCCAACGAATCACCTCACCGTTATCGTGACTCACTTTCTTTCGCATC**GCCGAAGTGC**CATAAAAAATATTTTTT …*ILV6*

5′- TGCGAACAAAA**GAGTCATTAC**AACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC …*THR4*

5′- ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATA**GTAAATTGTC**ATGCATATGACTCATCCCGAACATGAAA …*ARO1*

5′- ATTGATTGACTCATTTTCCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA …*HOM2*

5′- GGCG**CCACAGTCCG**CGTTTGGTTATCCGGCTGACTCATTCTGACTCTTTTTTGGAAAGTGTGGCAT**GTGCTTCACA**CA …*PRO3*

TGAAAAATTC
GACATCGAAA
GCACTTCGGC
GAGTCATTAC
GTAAATTGTC
CCACAGTCCG
TGTGAAGCAC
****

MAP score = -10.0

Source: G.M. Church

Sampling

TGAAAAATTC
GACATCGAAA
GCACTTCGGC
GAGTCATTAC
GTAAATTGTC
CCACAGTCCG
TGTGAAGCAC
****

How much better is the alignment with this site as opposed to without?

TCTCTCTCCA
TGAAAAATTC
GACATCGAAA
GCACTTCGGC
GAGTCATTAC
GTAAATTGTC
CCACAGTCCG
TGTGAAGCAC
****

Source: G.M. Church

Continued Sampling

~~TGAAAAATTC~~
GACATCGAAA
GCACTTCGGC
GAGTCATTAC
GTAAATTGTC
CCACAGTCCG
TGTGAAGCAC
****

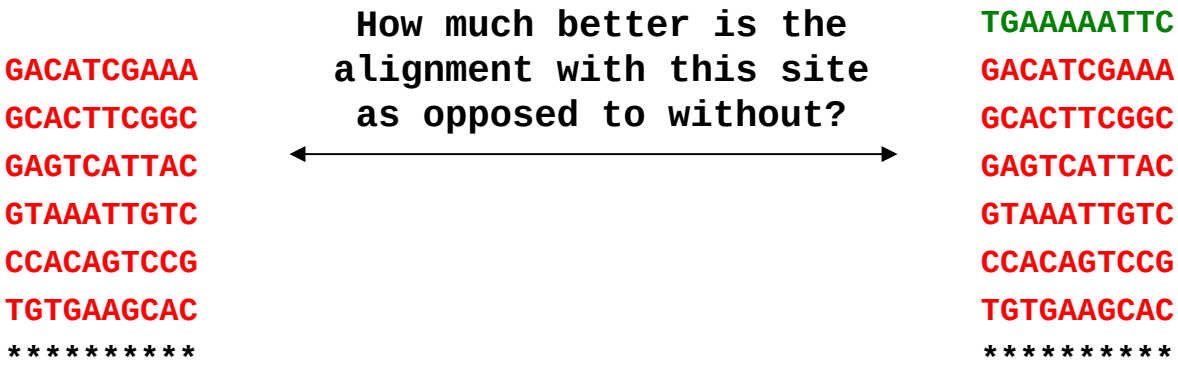
How much better is the alignment with this site as opposed to without?

ATGAAAAAAT
~~TGAAAAATTC~~
GACATCGAAA
GCACTTCGGC
GAGTCATTAC
GTAAATTGTC
CCACAGTCCG
TGTGAAGCAC
****

Source: G.M. Church

Continued Sampling

How much better is the alignment with this site as opposed to without?



Source: G.M. Church

Column Sampling

How much better is the alignment with this new column structure?

```
GACATCGAAA
GCACTTCGGC
GAGTCATTAC
GTAAATTGTC
CCACAGTCCG
TGTGAAGCAC
**********
```

```
GACATCGAAAC
GCACTTCGGCG
GAGTCATTACA
GTAAATTGTCA
CCACAGTCCGC
TGTGAAGCACA
********* *
```

Source: G.M. Church

The Best Motif

5′- TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATGAAAAATTCATGAG**AAAAGAGTCA**GACATCGAAACATACAT …*HIS7*

5′- ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCG**AAATGACTCA**ACG …*ARO4*

5′- CACATCCAACGAATCACCTCACCGTTATCG**TGACTCACTT**TCTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT …*ILV6*

5′- TGCGAAC**AAAAGAGTCA**TTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC …*THR4*

5′- ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATA**TGACTCATCC**CGAACATGAAA …*ARO1*

5′- ATTGAT**TGACTCATTT**TCCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA …*HOM2*

5′- GGCGCCACAGTCCGCGTTTGGTTATCCGGC**TGACTCATTCTGACTCTTTT**TTGGAAAGTGTGGCATGTGCTTCACACA …*PRO3*

AAAAGAGTCA
AAATGACTCA
AAGTGAGTCA
AAAAGAGTCA
GGATGAGTCA
AAATGAGTCA
GAATGAGTCA
AAAAGAGTCA
****

MAP score = 20.37

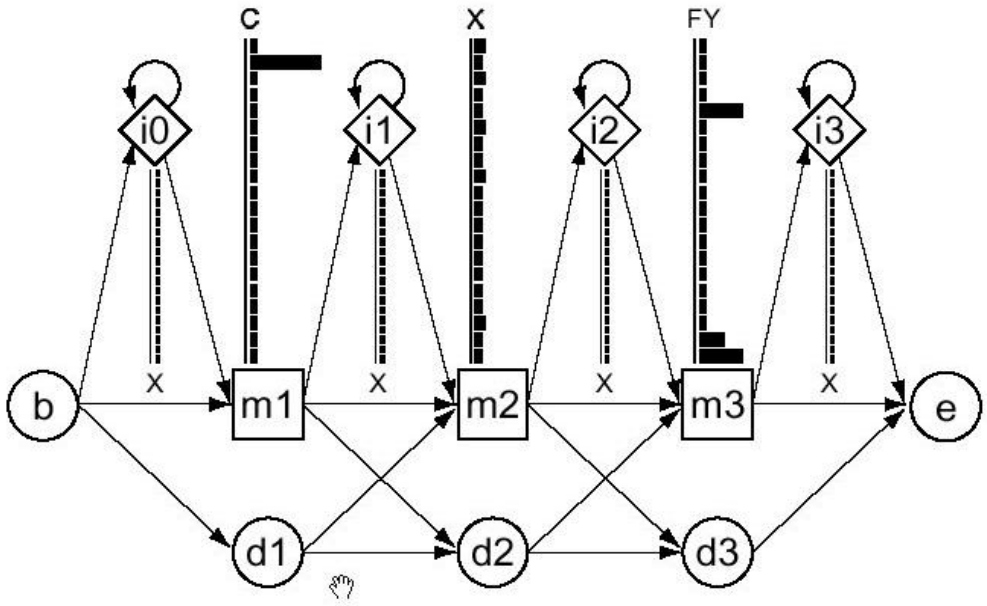
Source: G.M. Church

Profile HMM

- Profile HMMs: were introduced into computational biology in the late 1980's, and for use as profile models since 1994. Profile HMMs and HMM-based genefinders are the most successful HMM applications in computational biology.

Profile HMM

1	2	3
C	A	F
C	G	W
C	D	Y
C	V	F
C	K	Y



A small profile HMM (right) representing a short multiple alignment of five sequences (left) with three consensus columns.

Des HMM simplifiés

Dynamic HMM algorithms: Forward (for scoring) and Viterbi (for alignment) were used. They have a worst case of $O(NM^2)$ in time and $O(NM)$ in space for a sequence of length N and an HMM of M states.

For profile HMMs: that have a constant number of state transitions per state rather than the vector of M transitions per state in fully connected HMMs, both algorithms run in $O(NM)$ in time and $O(NM)$ in space.

Entraînement

Parameters set: an HMM can be built from prealigned (prelabeled) sequences (i.e, where the state paths are assumed to be known). It's simply a matter of converting observed counts of symbol emissions and state transitions into probabilities. In building a profile HMM, an existing multiple alignment is given as input.

HMM training algorithms: BaumWelch expectation maximization or gradient descent algorithms. Gibbs sampling, simulated annealing and genetic algorithm training methods seem better at avoiding spurious local optima in training HMMs and HMM like models.

- The primary advantage of these models over standard methods of sequence search is their ability to characterize an entire family of sequences.

- Thus, each position has a distribution of amino acid, as do *transitions* between states. That is, these linear HMMs have position-dependent character distributions and position-dependent insertion and deletion gap penalties. The alignment of each of a family to a trained model automatically yields a multiple alignment among those sequences.

Building Profile HMM

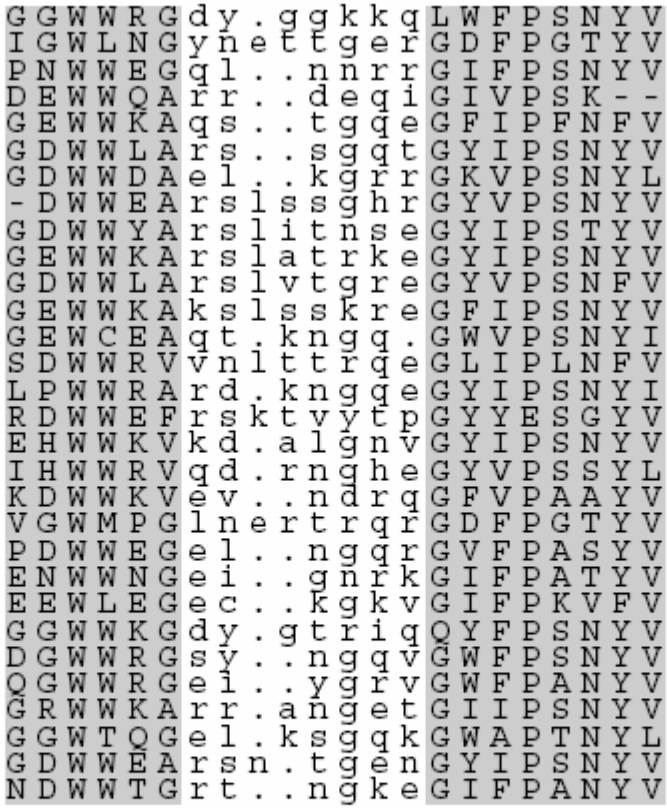
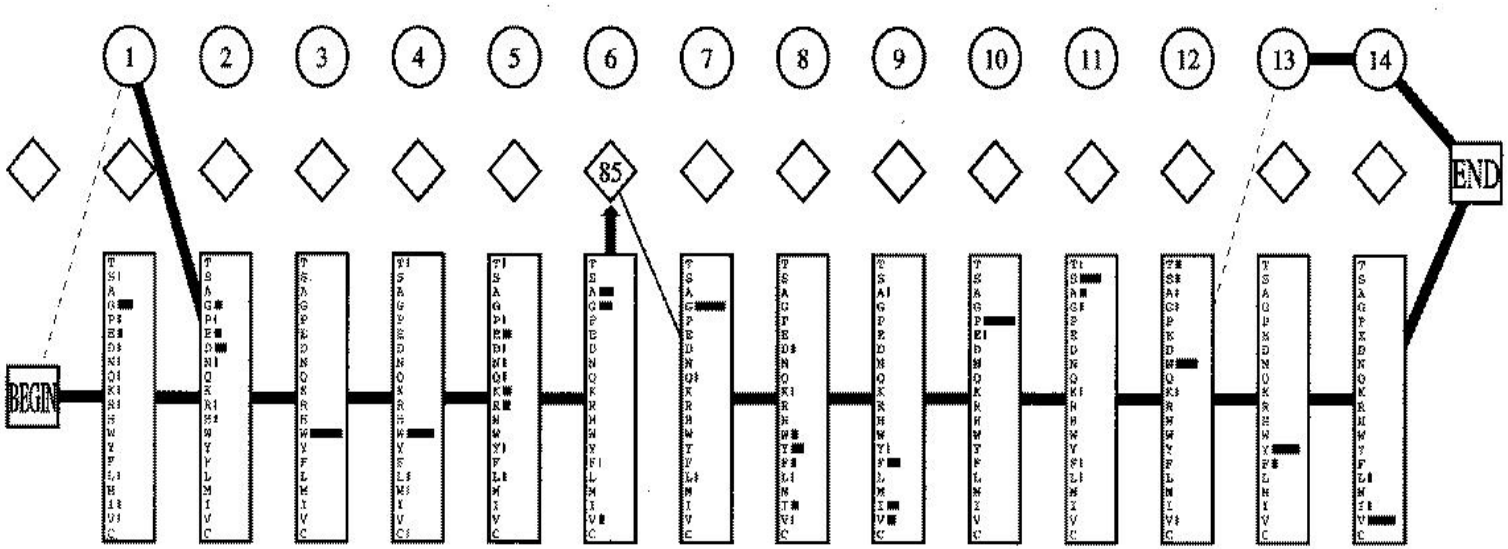


Figure 4.4: An alignment of 30 short amino acid sequences chopped out of a alignment of the SH3 domain. The shaded areas are the most conserved and were chosen to be represented by the main states in the HMM. The unshaded area with lower-case letters was chosen to be represented by an insert state.

Result



Note: transition lines with no arrow head are from left to right. Transitions with probability zero are not shown, and those with very small probability are shown as dashed lines. Transitions from an insert state to itself are not shown; instead the probability times 100 is shown in the diamond. The numbers in the circular delete states are just position numbers. (from SAM package of programs)

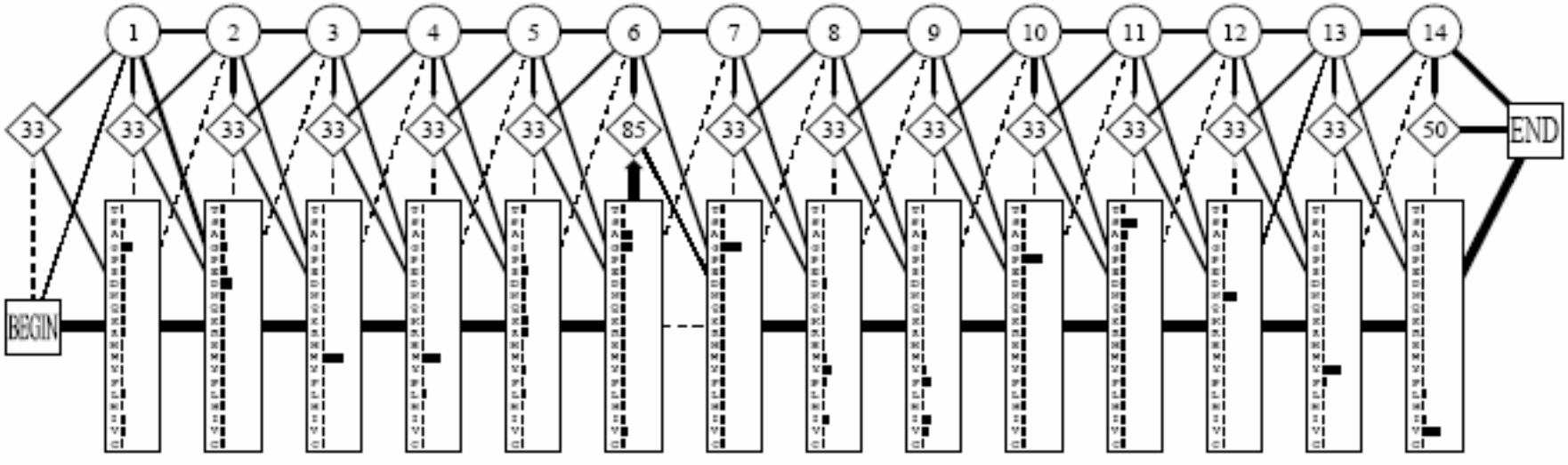
Pseudocounts

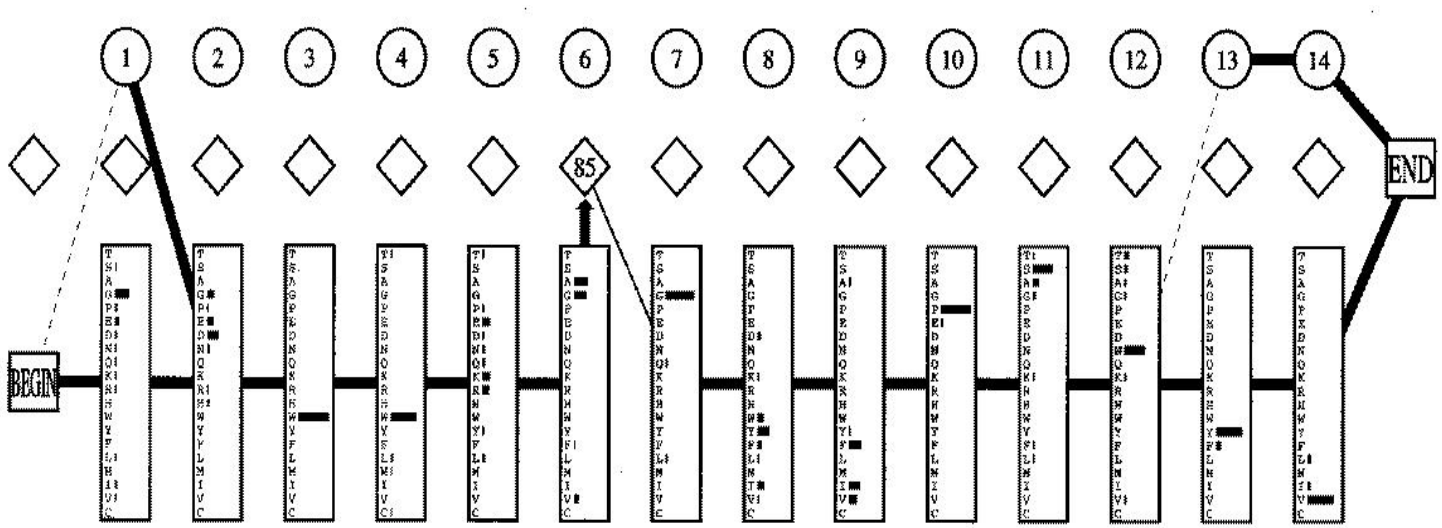
- Adding one to all the counts can be interpreted as assuming a priori that all the amino acids are equally likely. However, there are significant differences in the occurrence of the 20 amino acids in known protein sequences. Therefore, the next step is to use pseudocounts proportional to the observed frequencies of the amino acids instead. This is the minimum level of pseudocounts to be used in any real application of HMMs.

- Because a column in the alignment may contain information about the preferred type of amino acids, it is also possible to use more sophisticated pseudocount strategies. If a column consists predominantly of leucine (as above), one would expect substitutions to other hydrophobic amino acids to be more probable than substitutions to hydrophilic amino acids. One can *e.g.* derive pseudocounts for a given column from substitution matrices.

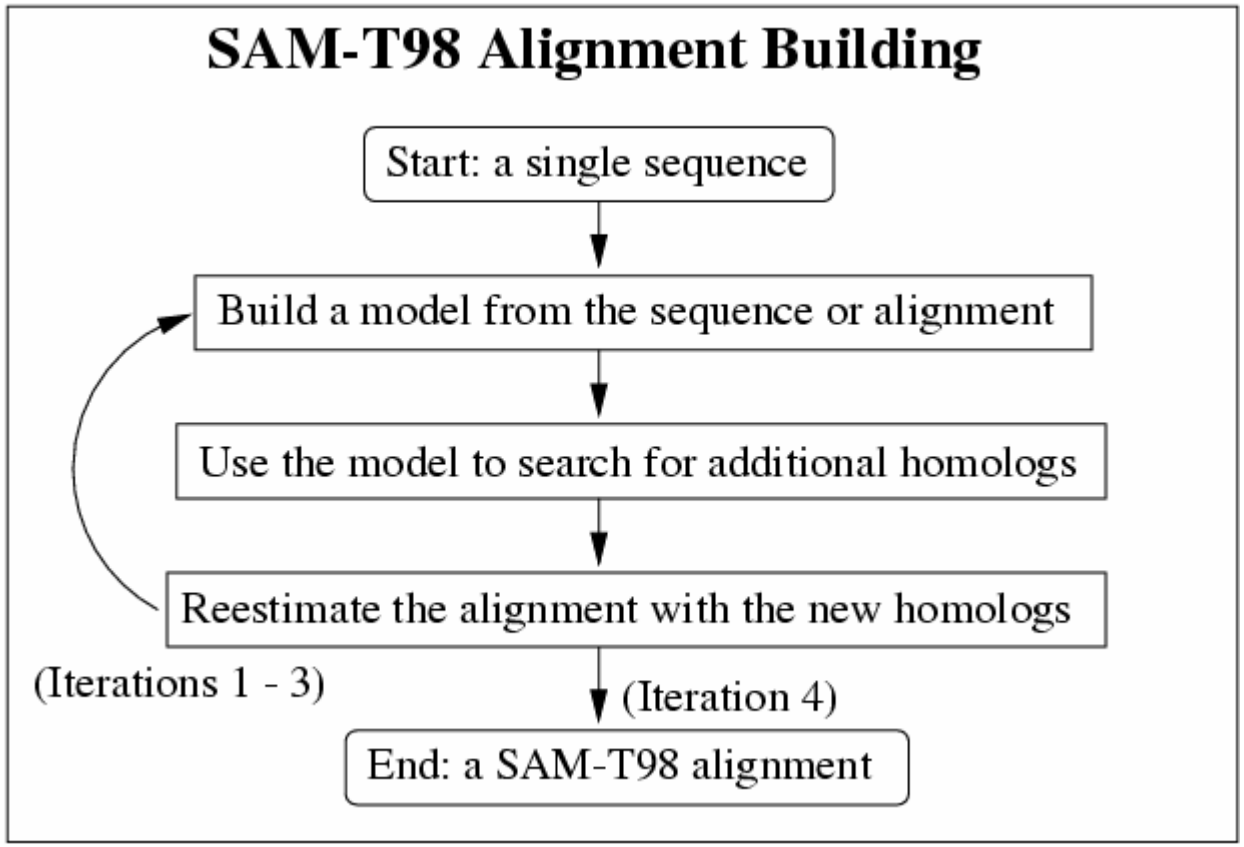
See also SAM Tutorial…

Result + pseudocounts



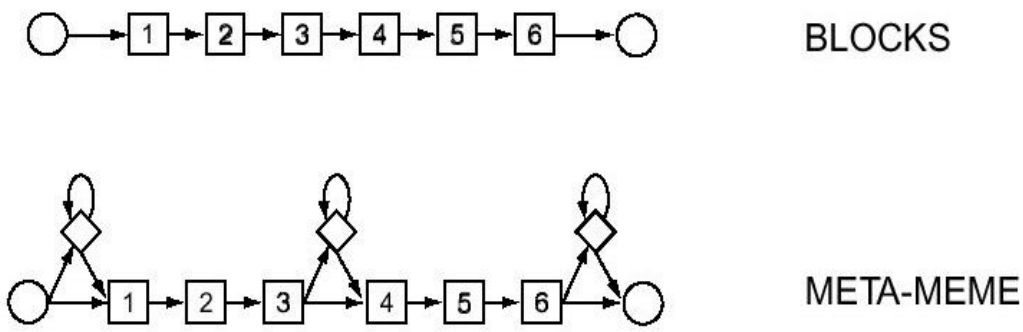


A partir d'une seule séquence



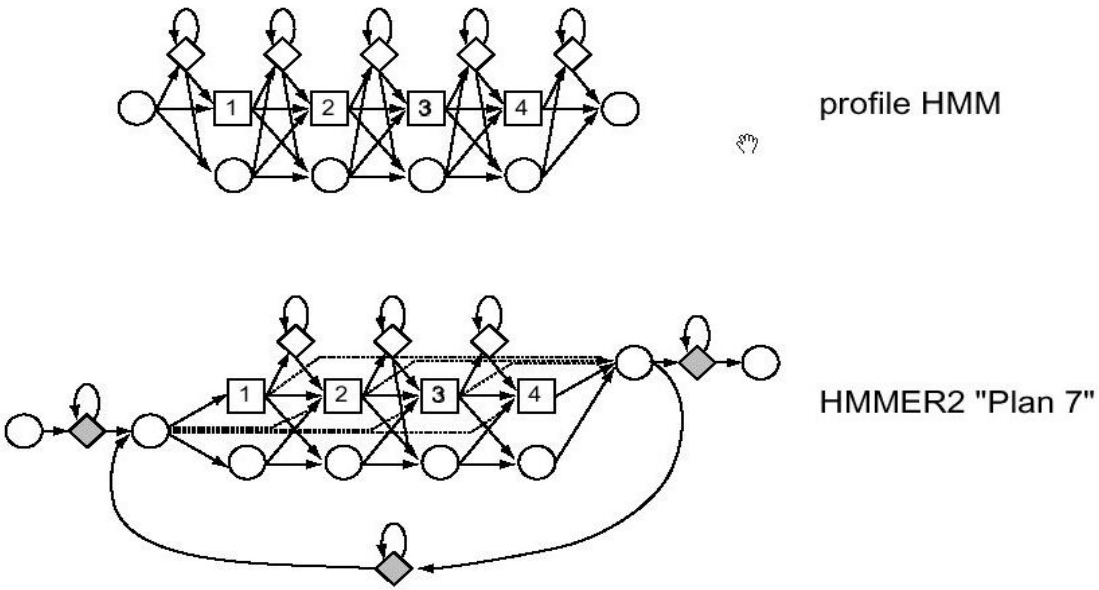
Logiciels

The difference between these software packages is the model architecture they adopt: "profile" models & "motif" models.



Motif model architecture: modeling one or more ungapped blocks of sequence consensus separated by a small number of insert states. Can be viewed as special cases of profile HMMs.

Logiciels



Profile model architecture: models with an insert and delete state associated with each match state, allowing insertion and deletion anywhere in a target sequence.

Conclusions (PHMM)

Three principal advances on Profile HMM methods:

1. Motif based HMMs have been introduced as an alternative to the original Krogh profile HMM architecture.

2. Large libraries of profile HMMs and multiple alignments have become available, as well as compute servers to search query sequences against these resources.

3. There has been an increasing incursion of profile HMM methods into the area of protein structure prediction by fold recognition.

Profile HMM method is a complement to BLAST and FASTA analyses

It will provide a second tier of solid, sensitive, statistically based analysis tools, based on the combination of powerful new HMM software and large sequence alignment databases of conserved protein domains.

Découverte de motifs expressifs sur les protéines

Pratt

An example of combinatorial method

- Aim generally at finding an expression, i.e. a pattern belonging to a language that is user-restricted with various parameters;
- For this purpose, explore the space of possible patterns in an ordered way, following the degree of generality (covering degree) and a fitness score.
- Pratt : Inge Jonassen 1996, program easily available, good expressivity.

Pratt's patterns

Pattern « PROSITE » : A_1 [x(i_k,j_k) Ak], k=2,p

C-x(3)-[ILVFYC]-D-x(8)-H-x(3,5)-H

Identity Component	Ambiguous Component	Fixed Wild-card x(i)	Flexible Wild-card x(i,j)

Limitations :

P	Maximum Number of components	5
L	Maximum Length of a pattern : p+ Σ j_k	23
W	Maximum Length of a Wild-card	8
F	Maximum Flexibility of a Flexible Wild-card : (j-i)	2
N	Maximum Number of Flexible Wild-cards	2
FP	Maximum Value of the product of flexibilities : $\prod (j_k-i_k+1)$)	6

Principes

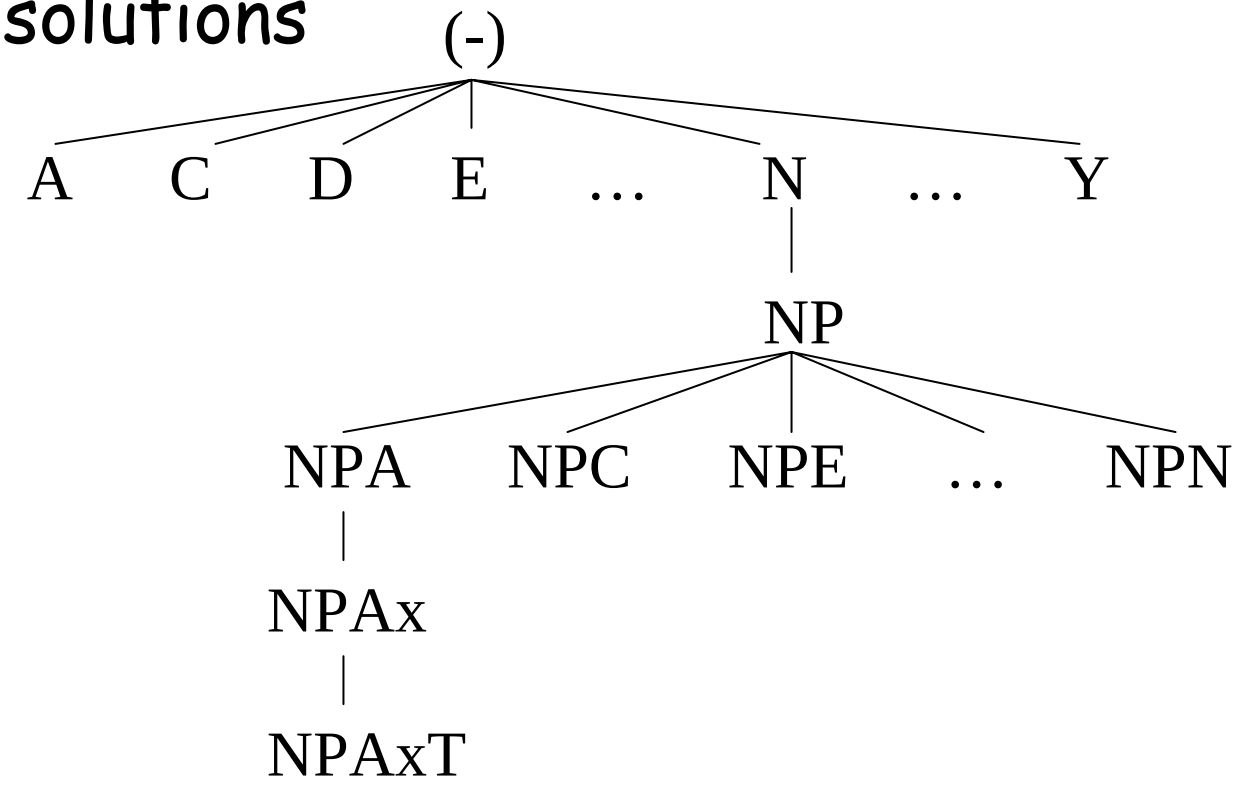
- Pattern Driven basé sur Jonassen et al. 1995
- Partir d'un petit motif (un acide aminé) et l'étendre (avec d 'autres acides aminés) tant que l'on respecte les limites de complexité choisies (longueur du motif, nombre de wildcards, …)

Principes (2)

- On peut distinguer l'utilisation pratique de deux types d'approches
  - Bottom-up
    - trouver des motifs par extension de motifs plus petits
  - Top-down
    - trouver des motifs par intersection entre séquences

Principe (3)

- BU ⇨ Arbre de recherche de l'espace des solutions



Algorithme combinant BU et TD

- Utilisation de l'approche Bottom-up pour faire émerger des motifs candidats
- Positionnement des candidats sur les séquences
- Alignement des séquences par rapport à ces candidats (points d'ancrage)
- Extension à gauche et à droite et évaluation des scores des nouveaux candidats de manière à poursuivre s'il y a augmentation

Points d 'ancrage

- La version 1 de Pratt permet d'orienter la recherche en fonction de Blocks (petits alignements locaux sans gaps de sous-séquences de même taille)
- La version 2 permet de restreindre les motifs à ceux qui valident les séquences tout en respectant un alignement multiple

Paramètres utilisés dans Pratt

- Nombreux paramètres pour définir l'espace de recherche

- Paramètres pour orienter la stratégie de recherche
  - Compromis complexité/exhaustivité
  - Utilisation d'un alignement ou d'une séquence imposée.
- Paramètre de choix du score
  - Quantité d'information
  - MDL (Minimum Description Length)

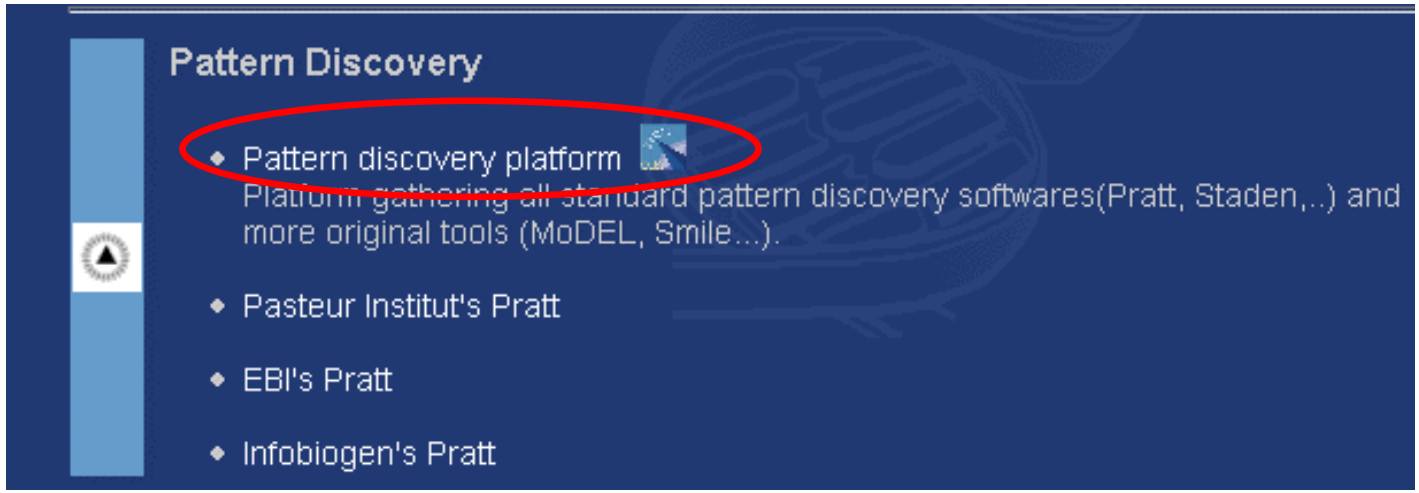
Pratt's algorithm (v2)

- Construction of a pattern graph of allowed patterns ;
- Patterns:={(ε,0,Root)};
- While Patterns ≠∅ #initial search#
  do

  For each (P,Score,Node) in Patterns #initial search#
  do
  - LQ:= add_one_edge(P,Node);
  - LQ':=Generalize2or3(P);
  - For each Q' in LQ'
    If Q covers at least M instances
    Then Patterns := Patterns ∪{(Q', score(Q'))}
- Patterns:= Sort_H_best_scores(Patterns);
- While Patterns ≠∅ #refinment#
  do

  For each P in Patterns
  - LQ:=Specialize _with_ambiguity(P); R:=∅;
  - For each Q in LQ
    If Q covers at least M instances
    Then Patterns := Patterns ∪{(Q, score(Q))}
- Patterns:= Sort_H_best_scores(Patterns)

Pratt Ouest Genopole



Pattern discovery algorithms

Staden algorithm [1]	class A - C
PRATT [2] [3]	class F
or Advanced PRATT	
Meta-Pratt	Pratt on sequences of patterns class I
Landraud algorithm [4]	class G
Smile [5]	class F
Winnover [6]	class A/C
MoDEL [7]	class C'

Doigts de Zinc : paramètres Pratt

```
                    Pratt version 2.1

          Analysing 44 sequences from file /tmp/tmpweb/analseq/

PATTERN CONSERVATION:
   CM: min Nr of Seqs to Match                  40
   C%: min Percentage Seqs to Match           90.9

PATTERN RESTRICTIONS :
   PP: pos in seq [off,complete,start]         off
   PL: max Pattern Length                       50
   PN: max Nr of Pattern Symbols                23
   PX: max Nr of consecutive x's                 5
   FN: max Nr of flexible spacers                2
   FL: max Flexibility                           2
   FP: max Flex.Product                         10
   BI: Input Pattern Symbol File               off
   BN: Nr of Pattern Symbols Initial Search     20

PATTERN SCORING:
   S: Scoring [info,mdl,tree,dist,ppv]        info

SEARCH PARAMETERS:
   G: Pattern Graph from [seq,al,query]        seq
   E: Search Greediness                          3
   R: Pattern Refinement                        on
   RG: Generalise ambiguous symbols            off

OUTPUT:
   OF: Output Filename       /tmp/tmpweb/analseq/pr5047/ou
   OP: PROSITE Pattern Format                   on
   ON: max number patterns                      50
   OA: max number Alignments                    50
   M: Print Patterns in sequences               on
   MR: ratio for printing                       10
   MV: print vertically                        off
```

Doigts de Zinc : motifs Pratt

```
Best Patterns (after refinement phase):
        fitness     hits(seqs)    Pattern
A   1:   16.6802      42(  41)    C-x-H-x(2)-C-x(2)-C
B   2:   15.6802      42(  41)    C-x-H-x(2)-C-x(0,2)-C
C   3:   12.5102      48(  41)    H-x(2)-C-x(2)-C
D   4:   11.5102      99(  40)    L-x(1,2)-S-x(2,3)-S
E   5:   11.5102     209(  40)    S-x(3,4)-S-x(1,2)-S
F   6:   11.5102     191(  40)    F-x(1,2)-F-x(2,4)-C

PATTERN MATCHES:
each . represents 10 sequence symbols
A symbol A-Z,a-z (for example A) in the place of a dot indicates the
starting point of a match to this pattern (in the example; pattern A).

-------------------------------------------------
sw|Q02084|A33_PLEWA: ....w......Fw...HAEo.c..b.m.G.X...dbFRS..u.G.P.....L..mD.w...W.
sw|P35226|BMI1_HUMAN: cC.AS.dG..w.....Wc.b.i..DLEP.L1EL
sw|P25916|BMI1_MOUSE: cC.AS.dG..w.....W..b.i..DwE.vLeD.
sw|P38398|BRC1_HUMAN: ...AC..D...X..EI..D......ML.OF........NDWFLPE..t.i.......rL...
sw|P22681|CBL_HUMAN: ...L...G.mb....UPn....SOm.O.G.....i......AD..N..uhtdLEN.w..O.LL
sw|P22682|CBL_MOUSE: .......G.mb....UP.....SO..O.G.....i....AD..N..uhwdNEN....kE..D.
sw|P43254|COP1_ARATH: U.....A...Pfu...bGt.i...L.ODDW...b..l....NN..FFE...w...uNG.uw.
sw|P23799|ESA8_TRYBB: e....A.......SDbFGc.OHcc.YN...E.TbGQ.HIWDbG.d.NU.mb.JGmbu.dWHR
sw|P26337|ESA8_TRYEQ: ..A.......SDbFGc.OHcPnYN...E.TbGQuHIWDbG.dcNU.mb.JGrb..dWHGL.i
sw|P08393|ICP0_HSV11: ....D..w..T...A...L...G........E...ij...w....L..wwD.v...DEwELU
sw|P28284|ICP0_HSV2H: ...bDQw......mA.GL....G......L........E............EEw..wLEEEE
sw|P29128|ICP0_HSVBJ: m.ACGD..........Wd.....bd...DEE.......Ei....S........E.w.i....
sw|P29836|ICP0_HSVBK: m.ACGD..........Wd.....bd...DEE.......Ei....d.....W..E.w.i....
sw|P28990|ICP0_HSVEB: .NA...FT...WGu....m.EE.el...N...N.EL.LiSLw....i.N.....
sw|P29129|ICP0_PRVIF: ...db.AmGX...FL.k...W...ELLM.ii.F.vEELEE..
sw|P09309|ICP0_VZVD: P.NA.GXFeGDvL.L..N....i..b..iW......c.i.DwREN..
sw|P35227|ME18_HUMAN: cC.AS.DG...FR....W.u.i..hEPELE.WL..
sw|P23798|ME18_MOUSE: cC.AS.DG....b....W.u.i..hEPEL...L..
sw|P23801|PE38_NPVAC: .Si.....P.ACa....P..FFSSG.FDM.L..
sw|P32512|PE38_NPVOP: o.D..mA....c.MF.S..bDOGR.iEN...
sw|O43490|PML1_HUMAN: WQ..WG.....p......wmS.DN..FPDDbEDtb...D.n.TG....HH.c.Fm..DhH..
```

Résultats avancés de Pratt sur les doigts de Zinc

PRATT parameters	
Pattern conservation :	90
Amino Acid regrouping :	Neutral Negative Buried Volume scales Positive Kyte-Doolittle Aliphatic Aromatic Charged
Maximum pattern length :	20
Research stringency :	Medium
Pattern scoring :	info
Max number of pattern symbols :	20
Max number of consecutive indetermination (x) :	6
Max number of flexible gap :	3
Max flexibility of a flexible wildcard :	3
Max flexibility product :	12
Pattern Refinment :	On
Generalise ambiguous symbols :	Off
Max number pattern in output :	20
Results screening	high

Pattern	Occurrence (Occurrence by sequence)	Fitness
☑ C-x-H-x-[CFIMVY]-C-x(2)-C-[ILMVY]-x(3)-[AGILMQTVWY]	40 (40)	21.53
☑ H-x(2,4)-C-x(2)-C-[GILMVWY]-x(3)-[AGHILMQTVWY]	50 (40)	14.12
☑ S-x(5,6)-S-x(0,2)-S-x(4)-[AHILPSTV]	183 (40)	12.25
☑ A-x(4,6)-S-x(2,4)-S-[AFGHILMNPQSTV]-x(4)-[AFGINPQSTV]	132 (40)	12.03
☑ E-x(3,4)-S-x(3,5)-S-x-[AGILNPQSVY]	111 (39)	11.95
☑ S-x(3,4)-S-x(1,3)-S-x-[ACGLNPQSTV]	207 (40)	11.93

Vue graphique des occurrences dans les séquences de doigts de Zinc

BONSAI

*[1] A Machine Discovery from Amino Acid **Sequences** by **Decision Trees** over Regular Patterns* ,
S. Arikawa, S. Kuhara, Y Mukouchi, T. Shinohara
New Generation Computing, pp 361-375, 1993

[2] Knowledge Acquisition from Amino Acid Sequences by Machine Learning System BONSAI,
S. Shimozono, A. Shinohara, T. Shinohara, S. Miyano, S. Kuhara, S. Arikawa
Transactions of Information Processing Society of Japan, 1994

[3] BONSAI Garden: Parallel Knowledge Discovery System for Amino Acid Sequences,
International Conference on Intelligent System for Molecular Biology (ISMB'95),
T. Shoudai, M.Lappe, S. Miyano, A. Shinohara, T. Okazaki, S. Arikawa, T. Uchida, S.Shimozono, T.Shinohara, S.Kuhara
http://www.i.kyushu-u.ac.jp/~shoudai/papers/BONSAI-Garden.html

http://bonsai.ims.u-tokyo.ac.jp/services/services.html (« soon »)

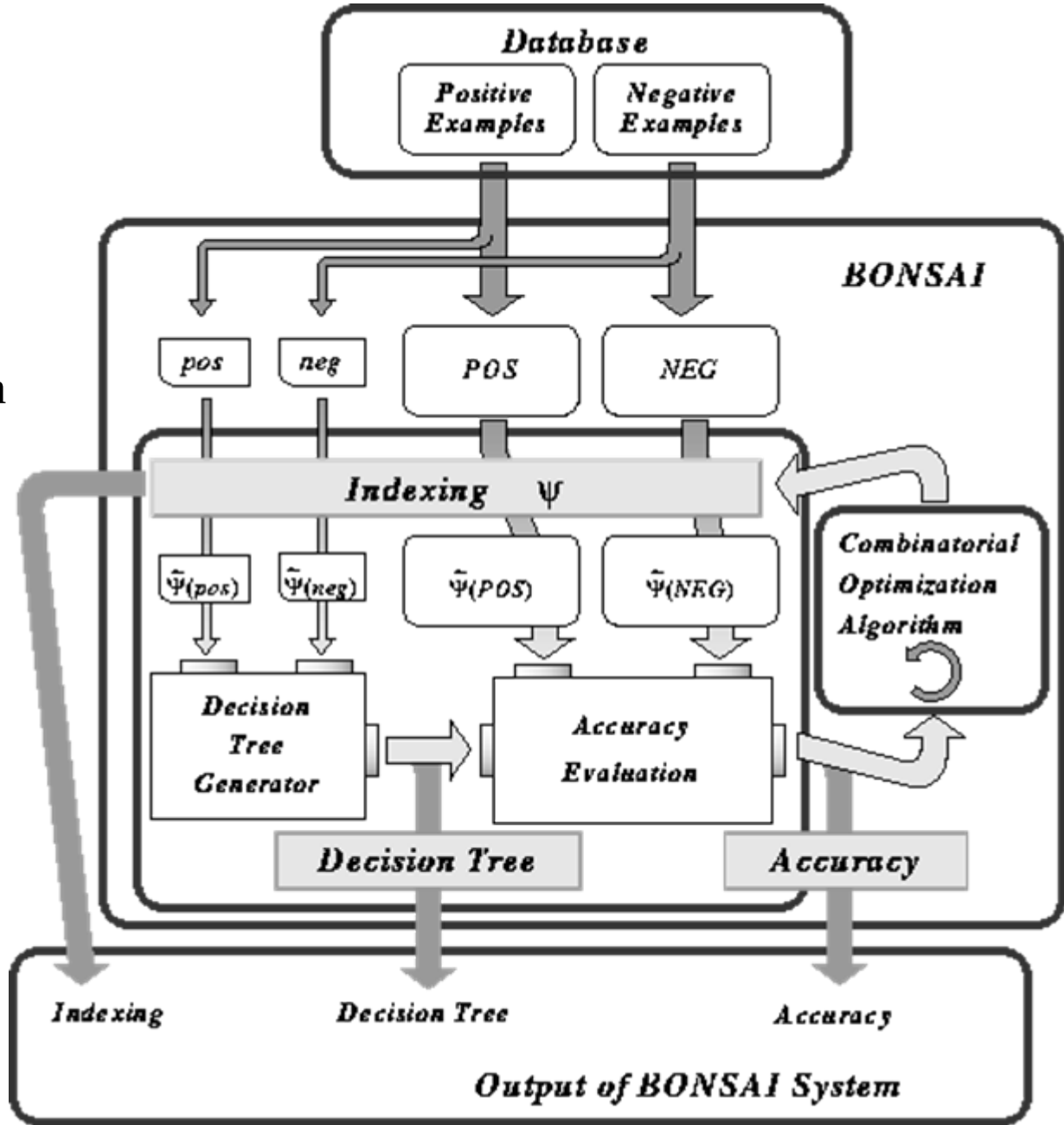
Vue générale de BONSAI

Exemples et contre exemples
(tirés des Bases de Données)

Séparation en échantillon
d'apprentissage et de validation

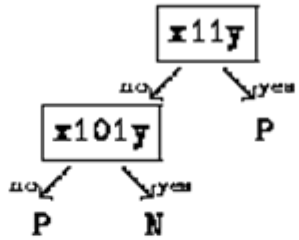
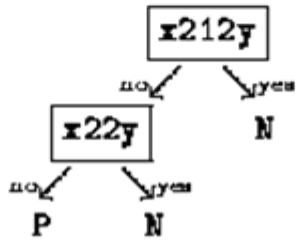
Recodage

Apprentissage et évaluation



Exemple : prédiction de domaines transmembranaires

GLLECCARCLVGAPFASLVATGLCFFGVALFCGCEVEALTGTEKLIETYFSKNYQDYEYL
INVIHAFQYVIYGTASFFFLYGALLLAXGFYTTGAVRQIFGDYKTTICGKGLSATVTGGQ
KGRGSRGQHQAHSLERVCHCLGCWLGHPDKFVGITYALTVVWLLVFACSAVPVYIYFNTW
TTCQSIAAPCKTSASIGTLCADARMYGVLPWNAFPGKVCGSNLLSICKTAEFQMTFHLFI
AAFVGAAATLVSLLTFMIAATYNFAVLKLMGRGTKF

	Alphabet Indexing	Decision tree (regular pattern)	Score
(a)	ACDEFGHIKLMNPQRSTVWY 00110010100101100000		$93.4\% = \sqrt{90.7 \times 96.2}$
(b)	ACDEFGHIKLMNPQRSTVWY 10221120110212100102		$82.1\% = \sqrt{85.5 \times 79.7}$

Construction de l'arbre de décision

```
function MakeTree( P, N : sets of strings ): node;
    begin
        if N = ∅ then
            return( Create("1", null, null) )
        else if P = ∅ then
            return( Create("0", null, null) )
        else begin
            Find a regular pattern π in Π
            minimizing E(π, P, N);
            P1 ← P ∩ L(π);    P0 ← P − P1;
            N1 ← N ∩ L(π);    N0 ← N − N1;
            if (P0 = P and N0 = N) or (P1 = P and N1 = N)
                then return( ( Create("1", null, null) )
            else
                return
                Create(π, MakeTree(P0, N0), MakeTree(P1, N1))
            end
    end
```

Choix de l'attribut

Énumération exhaustive des mots Π présents dans P et N et choix de celui minimisant $E(\Pi,P,N)$:

(b) The objective function (at line 9) to be minimized is defined by

$$E(\pi, P, N) = \frac{p_1 + n_1}{|P| + |N|} I(p_1, n_1) + \frac{p_0 + n_0}{|P| + |N|} I(p_0, n_0),$$

where $p_1 = |P \cap L(\pi)|$, $n_1 = |N \cap L(\pi)|$, $p_0 = |P \cap \overline{L(\pi)}|$, $n_0 = |N \cap \overline{L(\pi)}|$, $\overline{L(\pi)} = \Sigma^* - L(\pi)$, and $I(x, y) = -\frac{x}{x+y} \log \frac{x}{x+y} - \frac{y}{x+y} \log \frac{y}{x+y}$ (if $xy \neq 0$), $I(x, y) = 0$ (if $xy = 0$).

cf. critère ID3

Recodage *(Alphabet indexing)*

Trouver $\psi:\Sigma\rightarrow\Gamma$ tq $\psi(P) \cap \psi(N) = \varnothing$
est un problème NP-Complet

- **Heuristique de recherche locale :**
  1. Choix aléatoire de ψ
  2. Pour chaque voisin ψ' de ψ
     Evaluer le score de ψ'
     (score de l'arbre de décision)
  3. Si meilleur score, $\psi \leftarrow$ meilleur(ψ') et aller en 2.
     Sinon retourner ψ

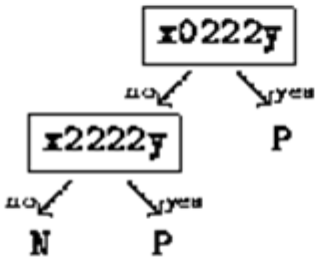
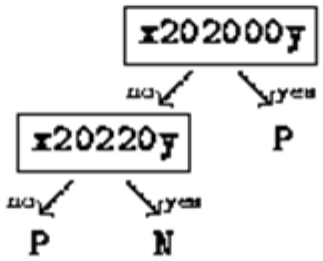
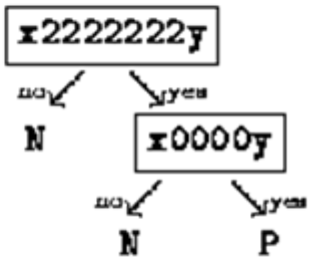
- Approximation en temps polynomial [Shimonozo 1995] ?
- Cluster analysis [Nakakuni et al 1994] ?
- Expérimentations algos génétiques $\rightarrow$ trop long.

Bonsai Garden

- Proposition de plusieurs solutions
  - Bruit
  - Sous classes
- Parallélisme (avec un jardinier…)

Bonsai Garden

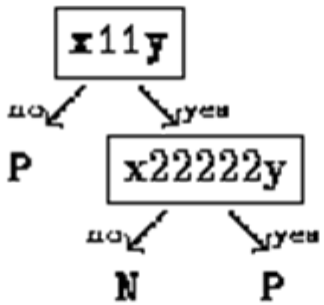
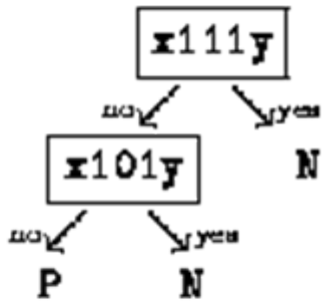
- Prédiction de promoteurs

	BONSAI Garden			
	B_1	B_2	B_3	*Trash*
Indexing	ACGT 1021	ACGT 0222	ACGT 2002	
Decision Tree				
Positive (1,170)	706 (60.3%)	355 (30.3%)	86 (7.4%)	23 (2.0%)
Classified Negative (15,611)	10,564	9,467	10,080	–

CCAAT, GC et TATA box

BONSAI Garden

- Prédiction hélices α

	BONSAI Garden	
	B_1	B_2
Indexing	ACDEFGHIKLMNPQRSTVWY 22020102022112211221	ACDEFGHIKLMNPQRSTVWY 20100100022110011000
Decision Tree		
Positive (6,673)	986 (14.8%)	5,017 (75.2%)
Negative (20,295)	14,426	15,148
The number of classified positive sequences is 6,003 (90.0%)		

Utilisation de motifs plus expressifs

A Practical Algorithm to Find the Best Subsequence Patterns
Masahiro Hirao, Hiromasa Hoshino, Ayumi Shinohara, Masayuki Takeda, Setsuo Arikawa,
DS 2000 et TCS2003

A Practical Algorithm to Find the Best Episode Patterns
Masahiro Hirao, Shunsuke Inenaga, Ayumi Shinohara, Masayuki Takeda, Setsuo Arikawa
DS 2001

Discovering best Variable-Length-Don't-Care Pattern
Shunsuke Inenaga, Hideo Bannai, Ayumi Shinohara, Masayuki Takeda, and Setsuo Arikawa
DS 2002

Transparents de l'exposé disponible sur la page de Shunsuke Inenaga

Subsequence Pattern

Sous mot (*subsequence*) v de w:
mot inclus dans w (en respectant l'ordre)
Exple : w = abbaaabaab
v = bbba

- i.e. v peut être obtenu à partir de w en effaçant certaines lettres
- i.e. en utilisant $*$ pour désigner n'importe quel mot (*Variable-Length-Don't-Care* (VLDC)) :

  $w = * v_1 * v_2 * \ldots * v_m *$

  VLDC pattern $\in (\Sigma \cup *)^*$ $*bb*ba*$

  VLDC subsequence pattern $\in * (\Sigma *)^*$ $*b*b*b*a*$

Basic Algorithm

FindBestVLDC(*S*, *T*)

bestScore = – ∞; *bestVLDC* = ∅;

For all possible pattern *p* **do**

x_p = *numOfMatchedStr*(*p*, *S*);

y_p = *numOfMatchedStr*(*p*, *T*);

score = $f(x_p, y_p, |S|, |T|)$;

if *score* > *bestScore* **then**

bestScore = *score*; *bestSubseq* = *p*;

return *bestVLDC*;

reduce patterns for candidates

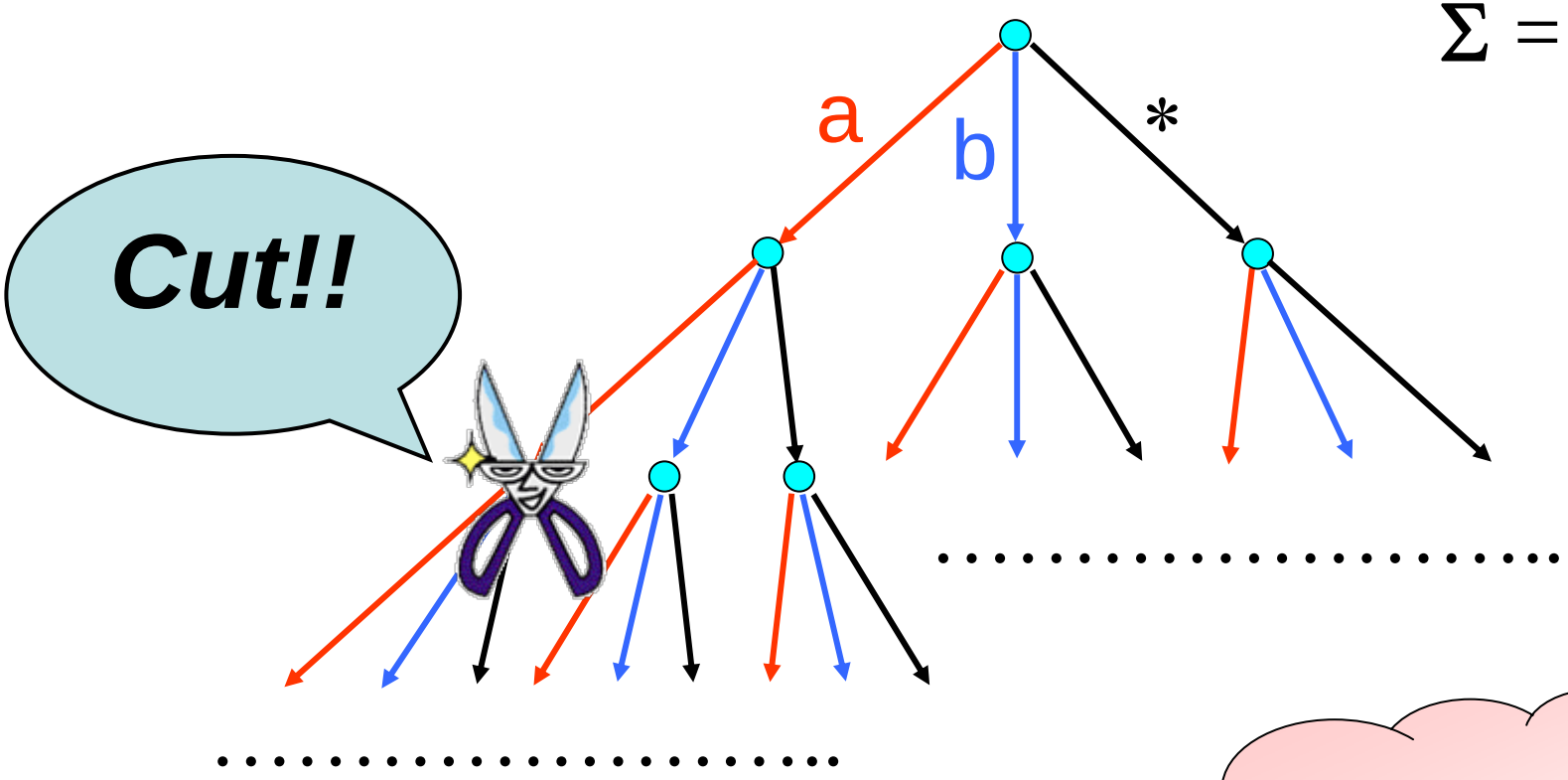
speed-up of these parts

pattern ***p***

Source S. Inenaga

Prune the Search Tree – the first key -

$\Sigma = \{a, b\}$



The search space is exponential...

Source S. Inenaga

Score Function

The "goodness" of pattern *p*

$$good(p, S, T) = f(x_p, y_p, |S|, |T|)$$

conic

S, T : given sets of strings

x_p : num. of strings in S that p matches

y_p : num. of strings in T that p matches

Chi2, Gain d'information, Gini sont coniques

Source S. Inenaga

Conic Function

Function $f(x, y)$ is *conic* if

- fo non-increasing here exists some x_1 such that
  $f(x, y) \geq f(x', y)$ r any $0 \leq x < x' \leq x_1$
  - $f(x, y) \leq f(x', y)$ for any $x_1 \leq x < x' \leq x_{max}$

- for any $0 \leq x \leq x_{max}$, there exists some y_1 such that
  - $f(x, y) \geq f(x, y')$ for any $0 \leq y < y' \leq y_1$
  - $f(x, y) \leq f(x, y')$ for any $y_1 \leq y < y' \leq y_{max}$

Conic Function (Cont.)

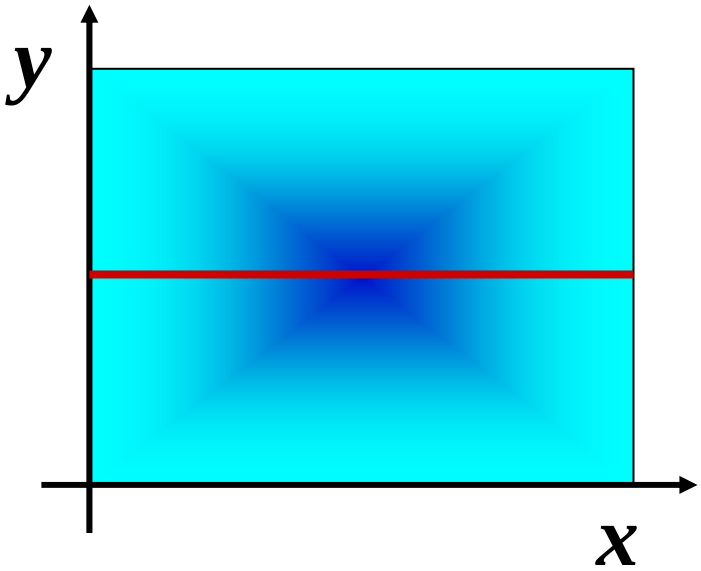
Function $f(x, y)$ is *conic* if

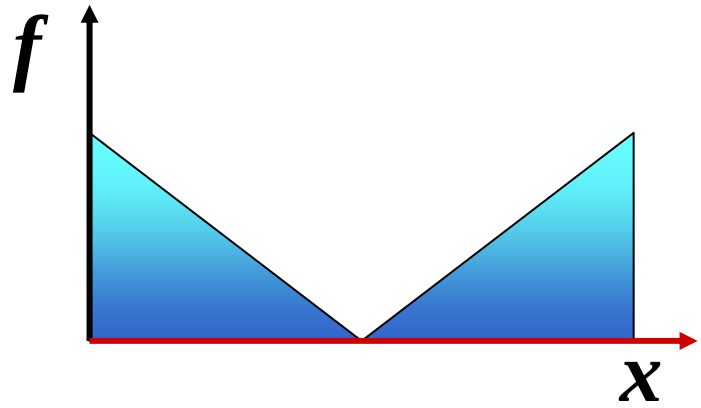
- for any $0 \le y \le y_{max}$, there exists some x_1 such that
  - $f(x, y) \le f(x', y)$ for any $0 \le x < x' \le x_1$
  - $f(x, y) \ge f(x', y)$ for any $x_1 \le x < x' \le x_{max}$

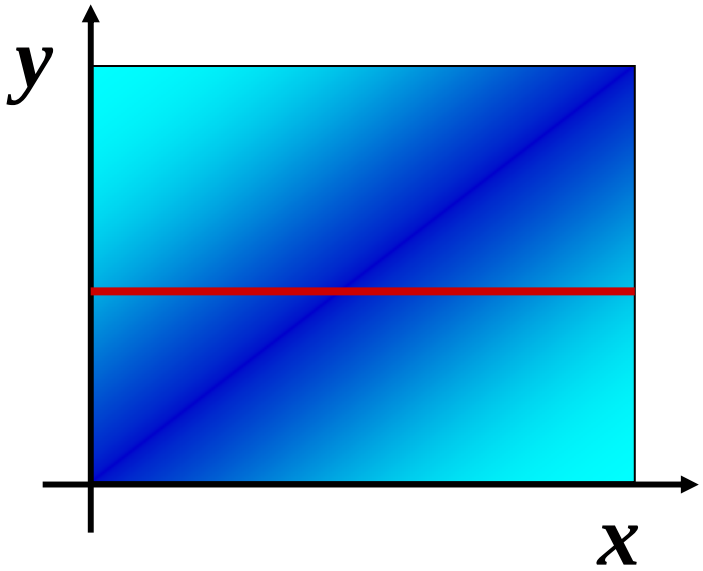
non-decreasing

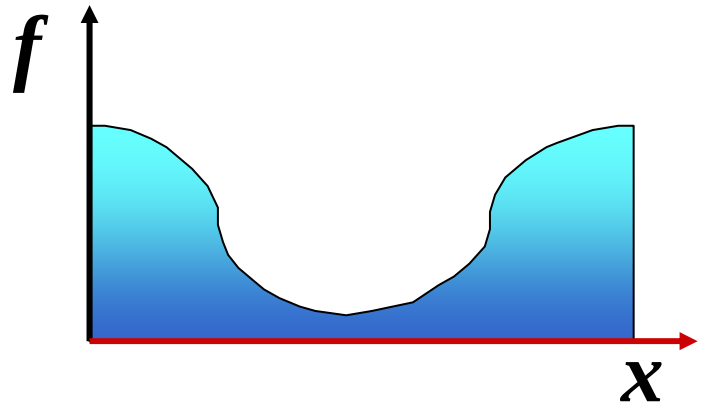
- for any $0 \le x \le x_{max}$, there exists some y_1 such that
  - $f(x, y) \ge f(x, y')$ for any $0 \le y < y' \le y_1$
  - $f(x, y) \le f(x, y')$ for any $y_1 \le y < y' \le y_{max}$

Conic Function (Cont.)



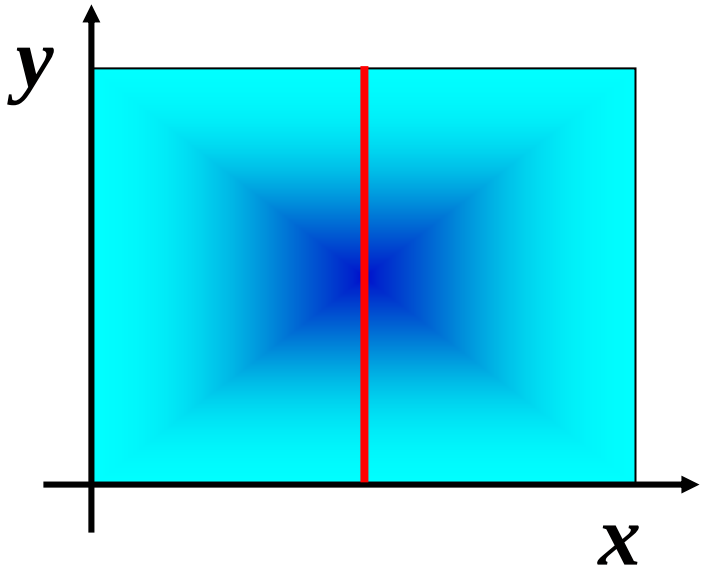


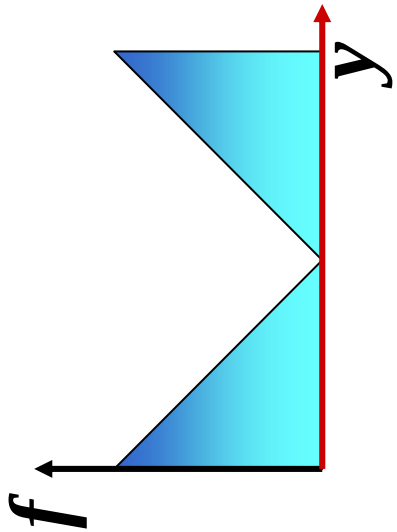


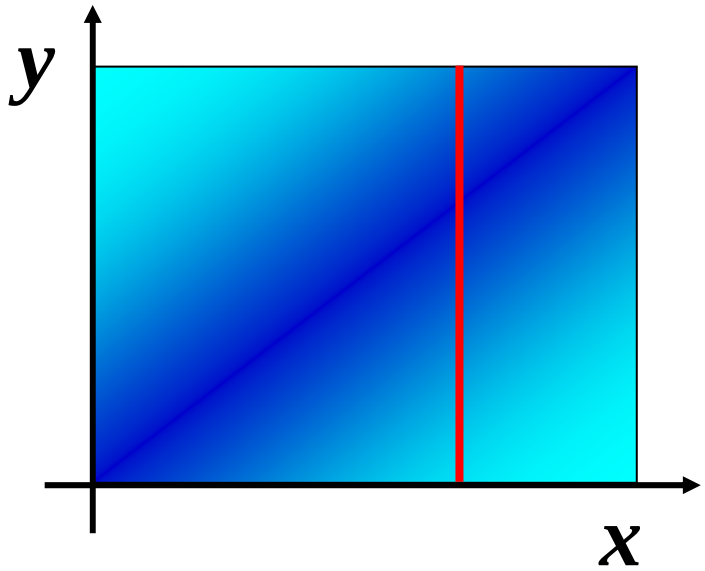


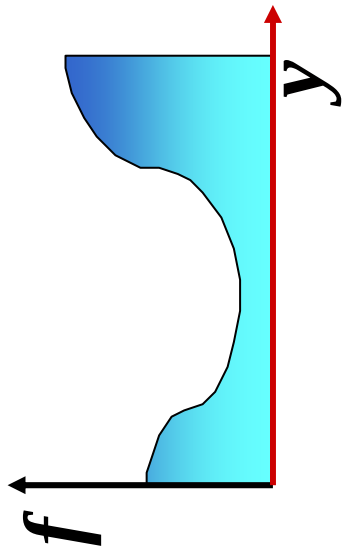
Source S. Inenaga

Conic Function (Cont.)



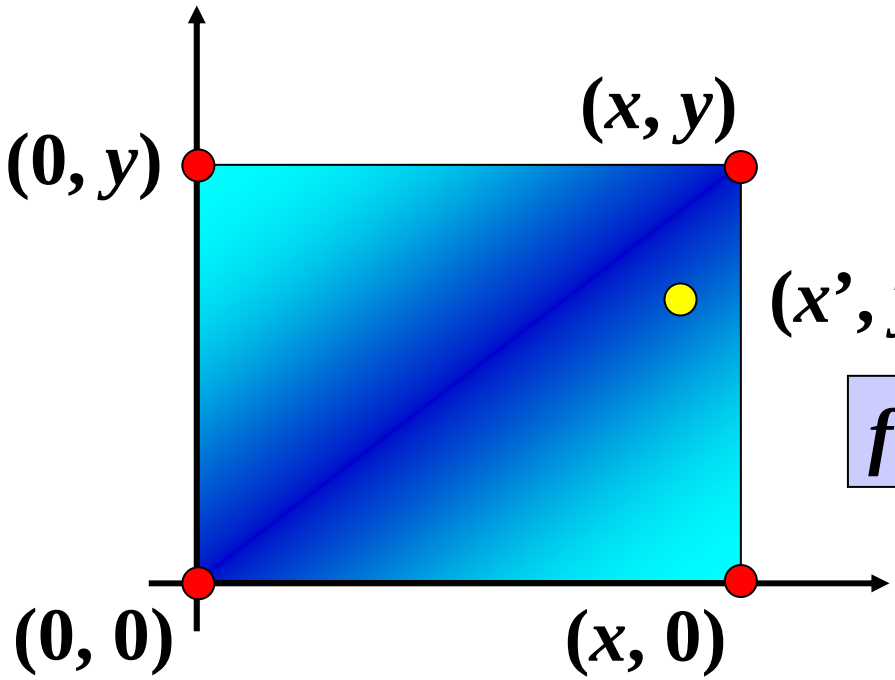






Source S. Inenaga

Property of Conic Function

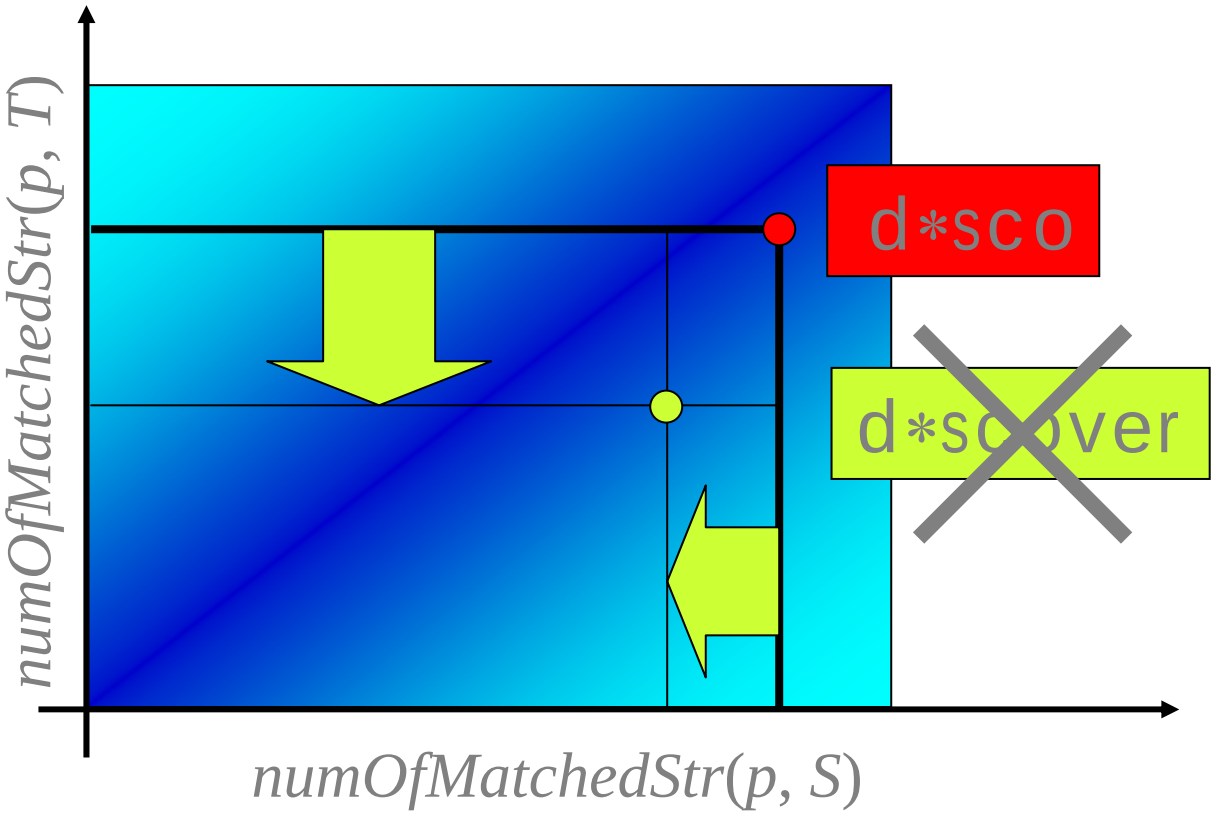


$$f(x', y') \leq upperBound(x, y)$$

$upperBound(x, y)$: the maximum value on the square

$$= max\{f(0, 0), f(x, 0), f(0, y), f(x, y)\}$$

Pruning Heuristics



The *goodness* of d∗scover ≤ The *upperBound* of d∗sco < The current best score

Basic Algorithm

FindBestVLDC(S, T)

bestScore = $-\infty$; *bestVLDC* = **ε**;

For all possible pattern *p* **do**

x_p = *numOfMatchedStr*(*p*, *S*);

y_p = *numOfMatchedStr*(*p*, *T*);

score = $f(x_p, y_p, |S|, |T|)$;

speed-up of these parts

pattern ***p***

if *score* > *bestScore* **then**

bestScore = *score*; *bestSubseq* = *p*;

return *bestVLDC*;

Fast VLDC Pattern Matching – the second key -

1. Using *DFA* for VLDC patterns

   → *Pattern Matching Machine*

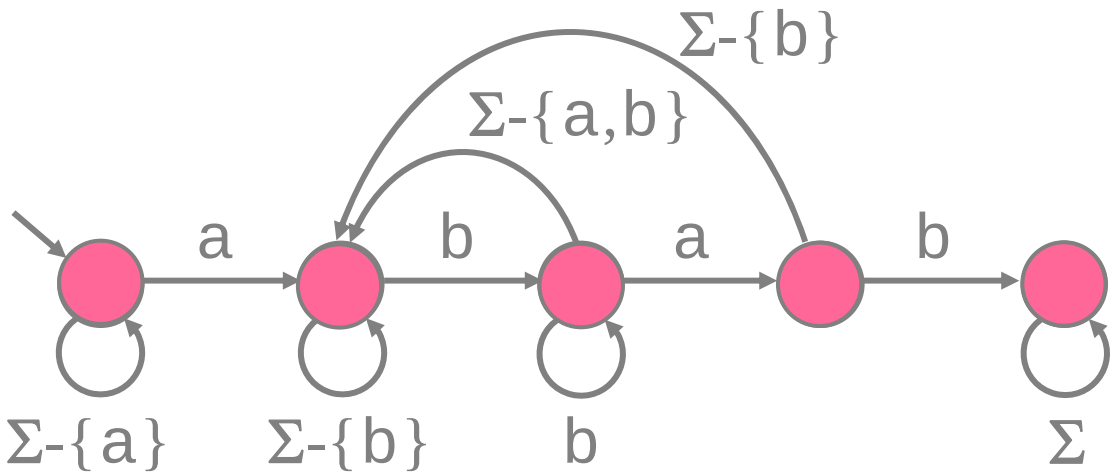
2. Using *Wildcard Directed Acyclic Word Graphs* for text strings

   → *Index Structure*

Computing the Minimum Window Size (Cont.)

1. Using *DFA* for VLDC patterns

p = ∗a∗bab∗



w = aabbab

Source S. Inenaga

Fast VLDC Pattern Matching – the second key - (Cont.)

2 Using *Wildcard Directed Acyclic Word Graphs* for text strings

The *Wildcard Directed Acyclic Word Graph* of a string w, $WDAWG(w)$, is the smallest automaton that recognizes all VLDC patterns which matches w.

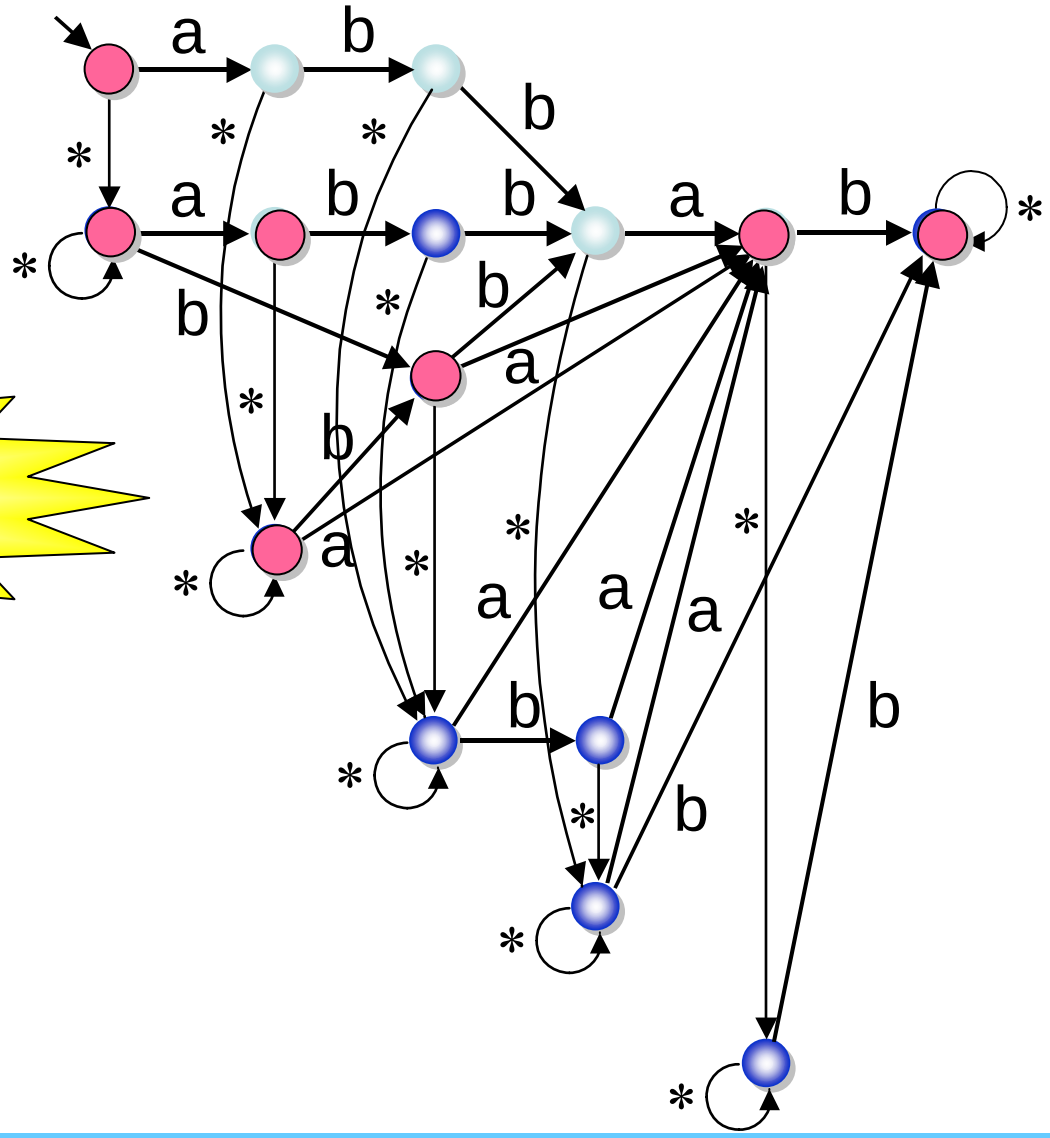
Inenaga *et al*. CPM 2002, MFCS 2002

Fast VLDC Pattern Matching – the second key - (Cont.)

2 Using *WDAWGs*

w = a b b a b

WDAWG(abbab)



p = ∗a∗bab

La suite…

- ARN : plutôt motifs structuraux
- « Junk » DNA
- Protéines, prise en compte du repliement de la protéine (motifs 3D)
- Prise en compte des dépendances entre positions, enchaînement des motifs…
- Apprentissage « croisé » avec :
  - données d'expressions,
  - génomique comparative,
  - topologie des protéines,
  - …

Quelques références bibliographiques générales…

- Motif Discovery on Promoter Sequences, Maximilian Haußler, Jacques Nicolas, 2005
- An Introduction to Hidden Markov Models for Biological Sequences, A. Krogh, 1998
- Finding Patterns in Biological Sequences, Brejová et al, 2000

Pour pratiquer

- RSA tools
- Packages HMM
- Plate-forme découverte de motifs Ouest Genopole

Oups trop loin !