

Analyse des régions régulatrices

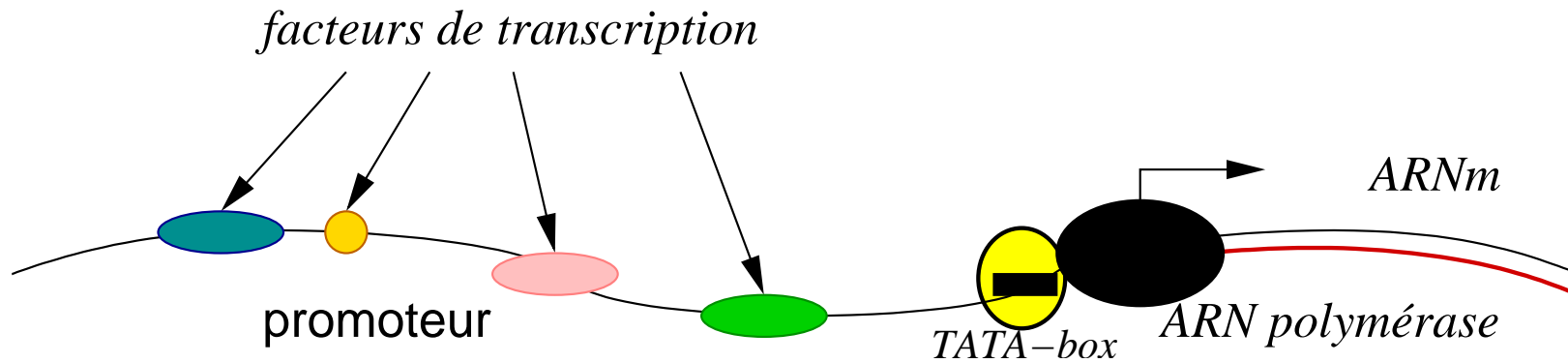
Hélène Touzet



Equipe Bioinfo

LIFL - UMR CNRS 8022

Université des Sciences et Technologies de Lille



- ▷ La fixation d'un facteur de transcription est guidée par la reconnaissance d'un site sur l'ADN. Ces sites sont **courts** et **dégénérés**.
- ▷ D'autres éléments entrent en jeu, et rendent la situation plus complexe : structure de la chromatine, etc.

Actuellement, l'analyse *in silico* des régions régulatrices ne sait prendre en compte que les informations du premier type.

Modélisation des sites de fixation

Point de départ: alignement multiple

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

Sites de fixation du facteur
de transcription *c-Ets-1*
(15 séquences) - Source :TRANSFAC M00032

- ▷ pas d'indels
- ▷ les colonnes sont indépendantes

Représentation par une séquence consensus :

code IUPAC

(International Union of Pure and Applied Chemistry)

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

N C M G G A W G Y N

A	adenine
C	cytosine
G	guanine
T	thymine
U	uracile
R	G A (purine)
Y	T C (pyrimidine)
K	G T (groupe keto)
M	A C (groupe amino)
S	G C (strong)
W	A T (weak)
B	G T C (pas A)
D	G A T (pas C)
H	A C T (pas G)
V	G C A (pas T)
N	A G C T

Représentation d'un motif par une matrice

- ▷ Ligne → position de l'alignement
- ▷ Colonne → acide nucléique (4 colonnes)

G CCGGAAGTG
A CCGGAAGCA
G CCGGATGTA
A CCGGAAGCT
A CCGGATATA
C CCGGAAGTG
A CAGGAAGTC
G CCGGATGCA
T CCGGAAGTA
A CAGGAAGCG
A CAGGATATG
T CCGGAAACC
A CAGGATATC
C AAGGACGAC
T CTGGACCCT

	A	C	G	T	
	7	2	3	3	N
1	14	0	0	0	C
5	9	0	1	0	M
0	0	15	0	0	G
0	0	15	0	0	G
15	0	0	0	0	A
8	2	0	5	0	W
4	1	10	0	0	G
1	6	0	8	0	Y
5	4	4	2	0	N



Position Frequency Matrix

0.47	0.13	0.2	0.2
0.07	0.93	0	0
0.33	0.6	0	0.07
0	0	1	0
0	0	1	0
1	0	0	0
0.53	0.13	0	0.33
0.27	0.07	0.67	0
0.07	0.4	0	0.53
0.33	0.27	0.27	0.13

Des PFM aux PWM

Position Frequency Matrix

0.47	0.13	0.2	0.2
0.07	0.93	0	0
0.33	0.6	0	0.07
0	0	1	0
0	0	1	0
1	0	0	0
0.53	0.13	0	0.33
0.27	0.07	0.67	0
0.07	0.4	0	0.53
0.33	0.27	0.27	0.13

→

Position Weight Matrix

0.91	-0.94	-0.32	-0.32
-1.8	1.9	-2.3	-2.3
0.4	1.26	-2.3	-1.8
-2.3	-2.3	2	-2.3
-2.3	-2.3	2	-2.3
2	-2.3	-2.3	-2.3
1.1	-0.94	-2.3	0.4
0.11	0.07	1.42	-2.3
-1.8	0.4	0	1.1
0.4	0.11	0.11	-0.94

- ▷ Poids positif : les bases qui apparaissent plus que la moyenne
- ▷ Poids négatif : les bases qui apparaissent moins que la moyenne

$$w(x) = \log_2 \left(\frac{f(x)}{0.25} \right)$$

$f(x)$ est la fréquence de x dans la colonne considérée
0.25 suppose que les bases sont équitablement réparties

Calcul du score

Position Weight Matrix

A	C	G	T
0.91	-0.94	-0.32	-0.32
-1.8	1.9	-2.3	-2.3
0.4	1.26	-2.3	-1.8
-2.3	-2.3	2	-2.3
-2.3	-2.3	2	-2.3
2	-2.3	-2.3	-2.3
1.1	-0.94	-2.3	0.4
0.11	0.07	1.42	-2.3
-1.8	0.4	0	1.1
0.4	0.11	0.11	-0.94

**Score de
TACGGATACG**

Calcul du score

Position Weight Matrix

A	C	G	T	
0.91	-0.94	-0.32	-0.32	T
-1.8	1.9	-2.3	-2.3	A
0.4	1.26	-2.3	-1.8	C
-2.3	-2.3	2	-2.3	G
-2.3	-2.3	2	-2.3	G
2	-2.3	-2.3	-2.3	A
1.1	-0.94	-2.3	0.4	T
0.11	0.07	1.42	-2.3	A
-1.8	0.4	0	1.1	C
0.4	0.11	0.11	-0.94	G

Score de TACGGATACG

1. On repère le poids de chaque position dans la PWM

Calcul du score

Position Weight Matrix

A	C	G	T	
0.91	-0.94	-0.32	-0.32	T
-1.8	1.9	-2.3	-2.3	A
0.4	1.26	-2.3	-1.8	C
-2.3	-2.3	2	-2.3	G
-2.3	-2.3	2	-2.3	G
2	-2.3	-2.3	-2.3	A
1.1	-0.94	-2.3	0.4	T
0.11	0.07	1.42	-2.3	A
-1.8	0.4	0	1.1	C
0.4	0.11	0.11	-0.94	G

Score de TACGGATACG

1. On repère le poids de chaque position dans la PWM
2. Le score est la somme des poids

6.16

Sequence LOGO

<http://weblogo.berkeley.edu/logo.cgi>

Exemple pour le site de C-ets-1:

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



- ▷ Permet de visualiser les colonnes qui sont conservées, en faisant apparaître le **contenu informationnel**
- ▷ Mesure d'**incertitude** d'une colonne de l'alignement :

$$H = - \sum_{x \in \{A,C,G,T\}} f(x) \log_2 f(x)$$

Incertitude maximale : les quatre bases ont chacune une fréquence de 25%

$$H = - \sum_1^4 \frac{1}{4} \log_2\left(\frac{1}{4}\right) = \sum_1^4 \frac{1}{4} \log_2(4) = \log_2(4) = 2$$

Incertitude minimale : la conservation est parfaite, il n'y a qu'une seule base.

$$H = - \sum_1^3 0 \log_2(0) + 1 \log_2(1) = 0$$

- ▷ **Conservation** de la colonne: $2 - H$
- ▷ **Contribution du nucléotide** x dans cette colonne: $f(x) \times H$

Banque de données de sites de fixation

- ▷ Transfac

eucaryotes, 5327 facteurs, 674 matrices, construites à partir de 13112 séquences

- ▷ Jaspar (<http://forkhead.cgb.ki.se/JASPAR/>)

- ▷ PlantCare

- ▷ levure : SCPD (<http://cgsigma.cshl.org/jian/>),
YRSA (<http://forkhead2.cgb.ki.se/yrsa/>)

- ▷ ...

Représentation d'un motif par un HMM

HMM : modèle de Markov caché

- ▷ modèle plus fin que les matrices
- ▷ prise en compte des insertions, les délétions
- ▷ prise en compte des dinucléotides, trinucéotides
- ▷ **Profile HMM** : largement utilisé pour la modélisation de motifs protéiques (Pfam)

Ajouter des insertions
des délétions

A - - T G T

A - - G A T

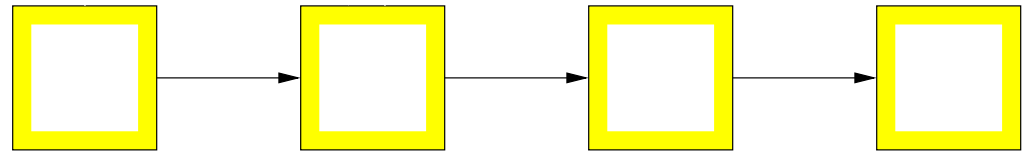
A C T G G T

A - - G - T

A - - C G T

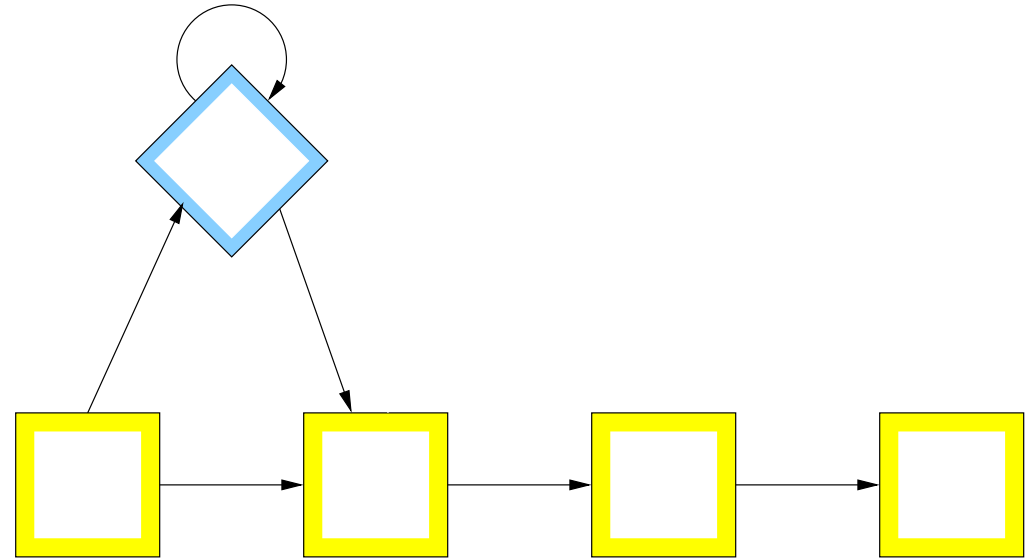
Ajouter des insertions
des délétions

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



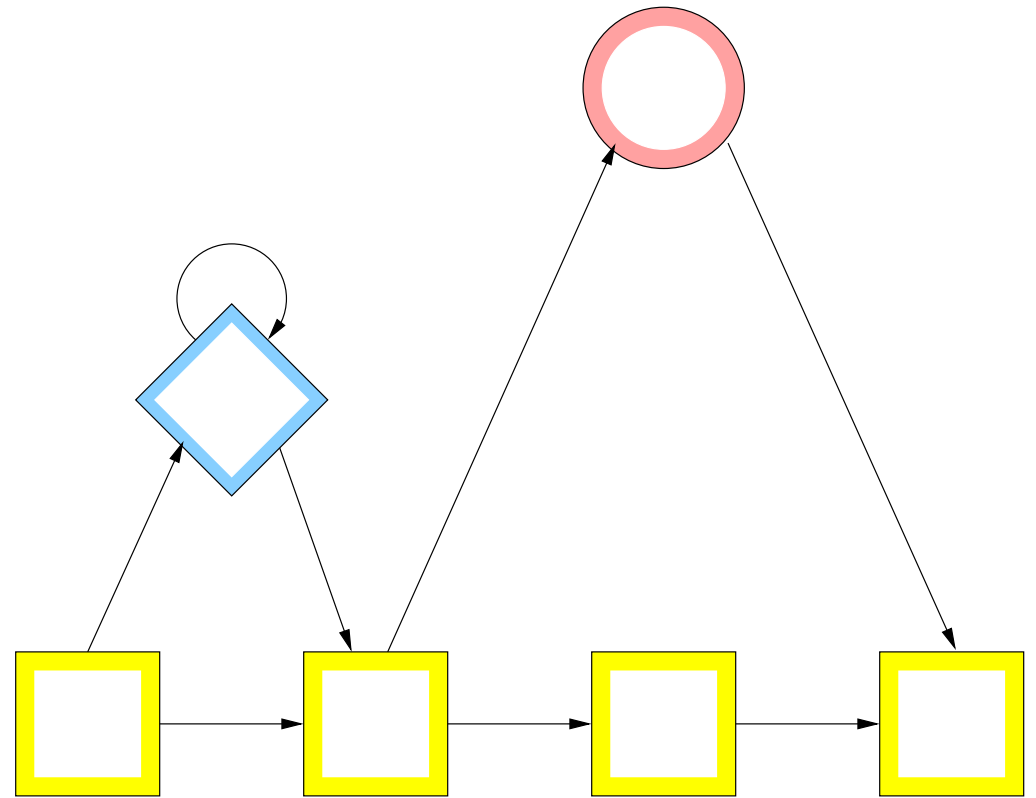
Ajouter des insertions
des délétions

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



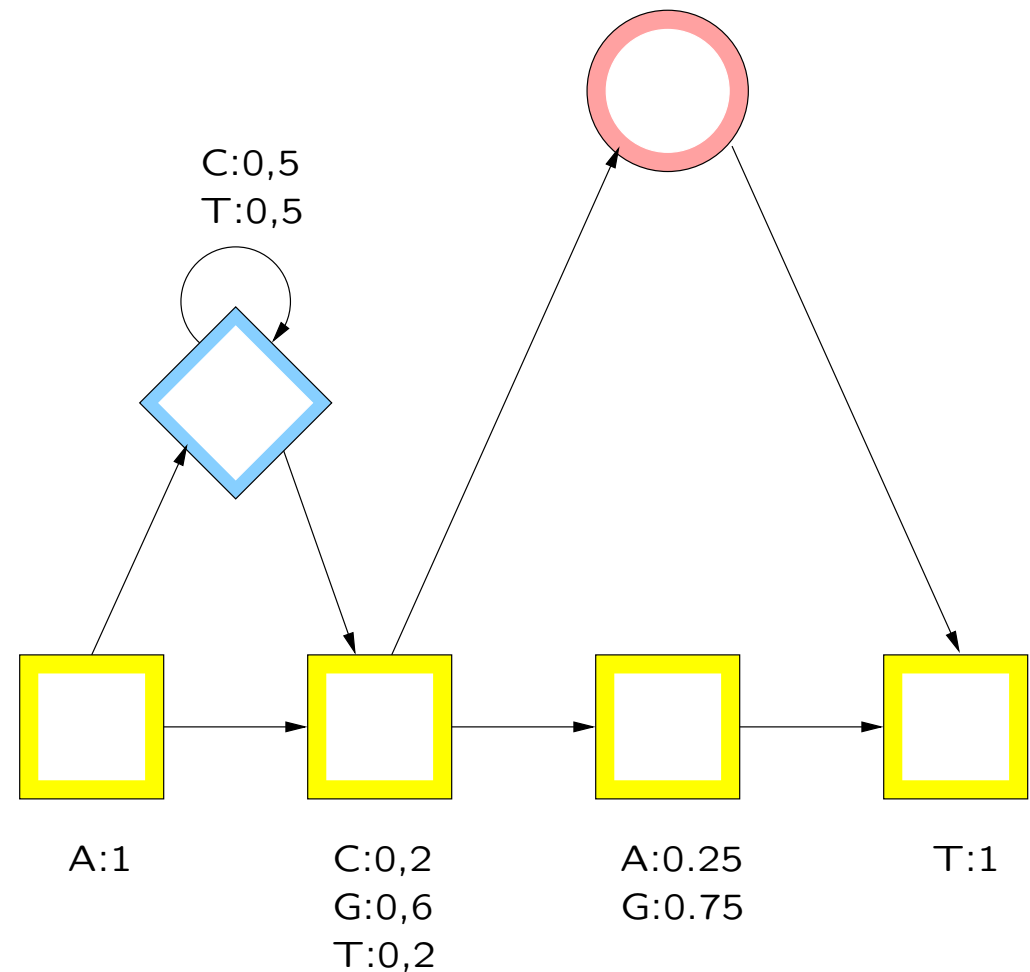
Ajouter des insertions
des délétions

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



Ajouter des insertions
des délétions

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

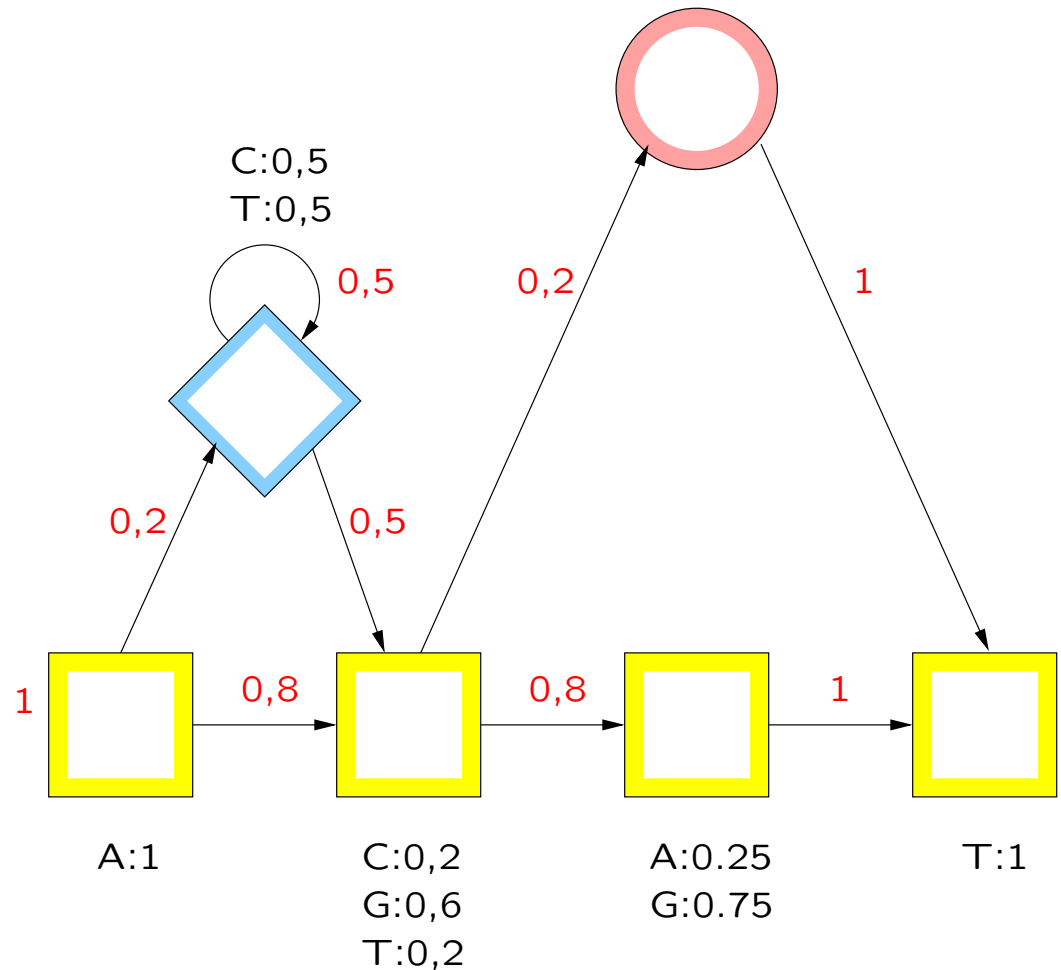


Probabilité d'émission :

fréquence de chaque base dans une colonne

Ajouter des insertions
des délétions

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



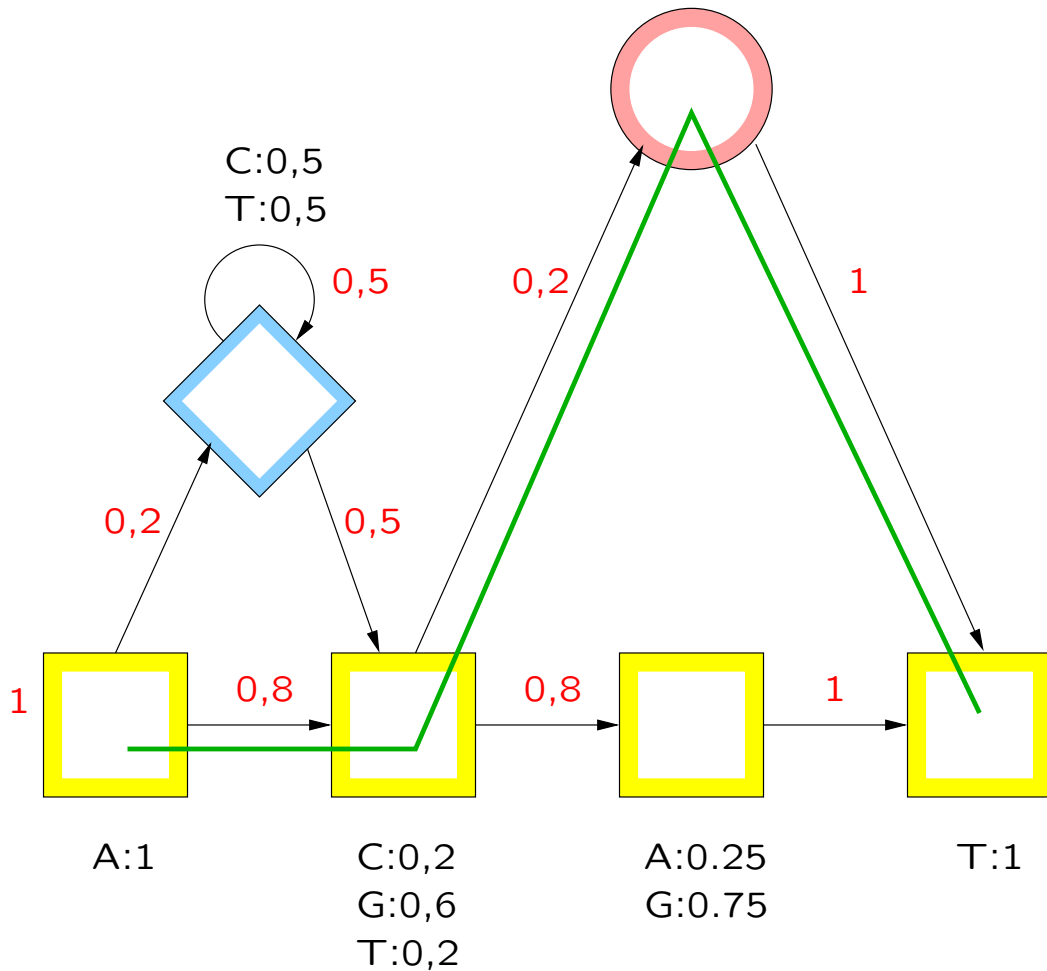
Probabilité d'émission :

fréquence de chaque base dans une colonne

Probabilité de transition :

circulation entre les colonnes

Calcul du score



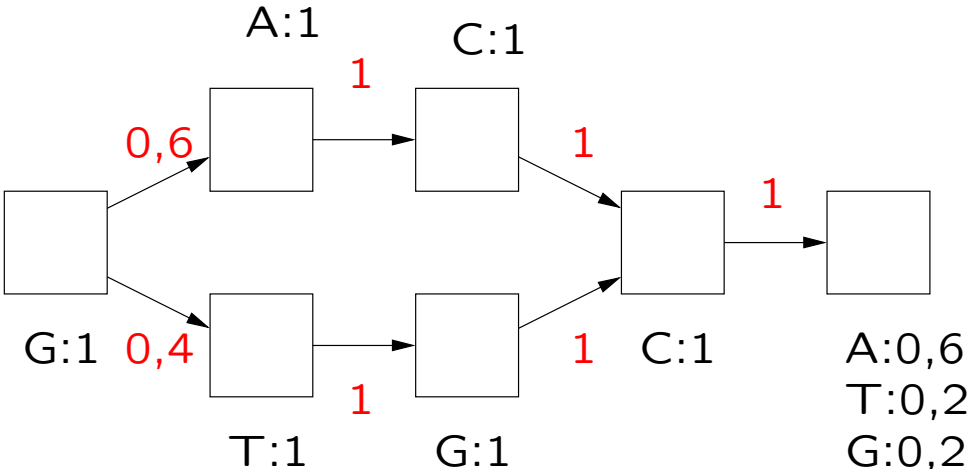
Score de AGT :

Probabilité dans le modèle : $p = 1 \times 1 \times 0,8 \times 0,6 \times 0,2 \times 1 \times 1$

Score : $\log_2(p)$

Ajouter des dépendances entre les positions

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



Bilan sur les matrices

▷ Patser

<http://bioweb.pasteur.fr/seqanal/interfaces/patser.html>

<http://rsat.ulb.ac.be/rsat/>

▷ Matinspector

- Matrices de Transfac
- Élimine les hits chevauchants pour les facteurs d'une même fmaille
- Score en deux parties, avec un cœur de matrice conservé

<http://www.genomatix.de/cgi-bin/matinspector/matinspector.pl>

▷ Contexte d'utilisation sécurisé

- connaissance a priori sur la région concernée : quelques dizaines à quelques centaines de bp
- connaissance a priori sur les facteurs impliqués dans les mécanismes de régulation

La très grande majorité des sites prédits se lie *in vitro*.

Tronche F., Ringeisen F., Blumenfeld M., Yaniv M., Pontoglio M. : *Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome* Journal of Molecular biology, 266, 231245, 1997

Le score reflète la spécificité du site *in vitro*: test sur 7000 sites de fixation potentiels du facteur CTF/NFI.

Roulet E, Busso S, Camargo AA, Simpson AJ, Mermoud N, Bucher P.: *High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites.* Nat Biotechnol. 20(8), 831-5, 2002

▷ Utilisation excessivement **périlleuse** : recherche à l'aveugle



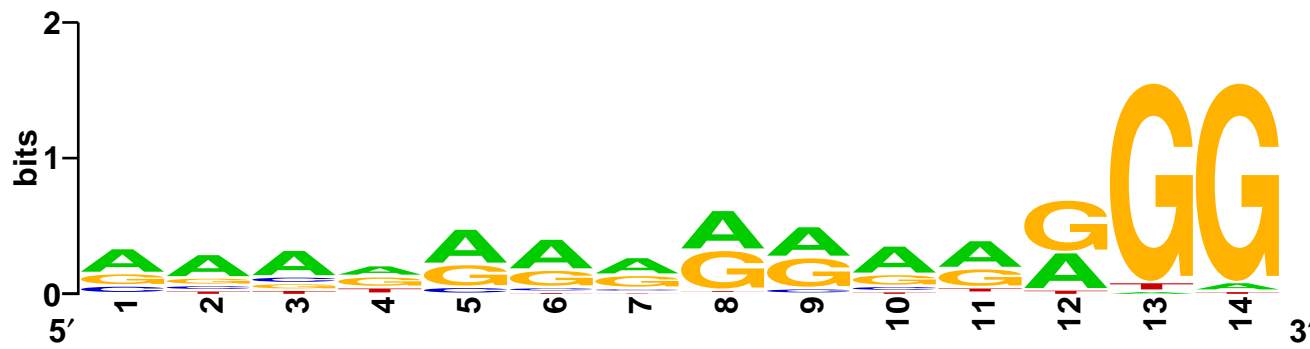
Suivant les matrices, un hit toutes les quelques dizaines de bases
à un hit toutes les quelques centaines de base

Avec l'ensemble des matrices de Transfac :
+90% de faux positifs

- ▷ **Chausse-trappe I** : hétérogénéité de la stringence des matrices

gut-enriched Krueppel-like factor (Transfac M00286 - 49 séquences)

A	26	28	29	18	26	27	23	23	22	28	25	19	1	2
C	8	6	7	2	5	4	5	1	4	5	2	0	0	0
G	13	9	7	18	17	15	17	23	22	12	17	27	46	46
T	2	6	6	11	1	3	4	2	1	4	5	3	2	1



▷ **Chausse-trappe II** : matrices chevauchantes

Sites de fixation pour les boites GC (M00255) et SP1 (M0008)



Chausse-trappes au niveau du génome :

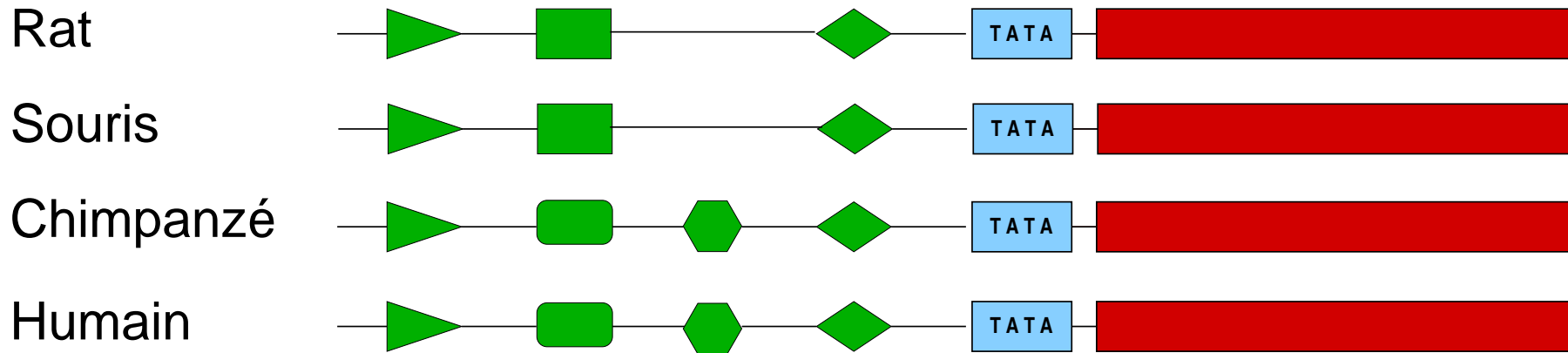
- ▷ biais de composition
- ▷ positionnement exact du TSS
- ▷ éléments de régulation dans les introns, ou dans la région 3'
- ▷ régions répétées

Trois pistes pour s'affranchir des faux positifs

1. empreinte phylogénétique
2. motifs surreprésentés
3. recherche de modules

Piste I : Empreinte phylogénétique

des souris et des hommes



Hypothèses de travail

1. les mécanismes de régulation sont conservés
2. la pression de sélection s'applique aux régions régulatrices (comme aux régions codantes)
3. la distance évolutive est suffisante pour que les zones non impliquées aient divergé

Environ 20% des régions régulatrices sont conservées entre l'homme et la souris.

Oeltjen JC, Malley TM, Muzny DM, Miller W, Gibbs RA, Belmont JW: *Large-scale comparative sequence analysis of the Human and Murine Bruton's tyrosine kinase loci reveals conserved regulatory domains.* Genome Research 1997 (7:315-329)

Hardison RC: *Conserved noncoding sequences are reliable guides to regulatory elements.* Trends Genet 2000 (16:369-372)

Wasserman WW, Palumbo M, thompson W, Fickett JW, Lawrence CE: *Human-mouse genome comparisons to locate regulatory sites.* Nature Genetics 2000 (26:225-228)

Alignement local des Homme et Souris

```
CGCTCC-----CCAA-----CCTCCACCTCC-----CCTCGCTCGGCCTCTATATGCTCCCGGGCTCCCTA
||| ||          ||||          ||||| || ||          || || | |||||  || |||| | |||||
CGCCCAGCAGCTCAAGACCAAGACTCGCCCTCCGCCCCACCCCTACCCCGTGCAGCCTCGGGATACTCCTGGGCTCCC-G

GTGTTGGCTGGAAGTGGGTGACTTAG---AGGCTTAAAGGAG--GGGCGCCTAACC----ACGGACCACGTGTGTGCGGGGGC
|  ||||| ||| | | |||  ||||  ||||| ||| |||  ||  || ||||| ||| | |||||
GCCGTGGCTGGATACGGGCG-CCTAGGGCAGGC----AGGAGGAGGGGGCC---CCCGCTACCGACCACGTGGGCGCGGGGGC

GACAGC---GCCGCCGGGGTGGGGCT-----GAGCGCTGCAAGCCGGGTTCGCCTTGCAGCGCAGGAGTCAGTGGGCGTTGC
||| ||  ||||  |||| || |||          ||||| ||          || ||  || |  ||||  |||| | |||||
GACGGCCGGGCCG--GGGGCGGAGCTTGAGCGAGCGCCGC-----GGCTCTGCTGGGCGCGCTGGAGGCGGTGGGCGTTGC

GCCACGATCTCTCT-----CCTAGCACTATGC-----TC-CCGCCCCACTCACCGCCTTGAAAGTCACAGGAGAAGG
||| ||  ||  ||          | |  || || |||          || ||||| ||          |||          ||
GCCGCGCCTGCCTGGGGAGCGC-GGCGCTGTGCCGCGTGGTTCGCCGCCCA-----TGCC-----GG

-CGGGCTCTAAGACCCAGCAGGCACCATCCTACTGGCGCCTTCG-ATCC-----GAGACCCGTTTGGACACCAGGGGGCG
|| || ||| ||||| ||||| ||          ||||| || | ||          ||| ||  | |||| | |||||
CCGCGCGCTAGGACCCAGCAGGCGCC-----GCGCCGCGCAGCCCGGGGACAGAGGCCGCTCGGACTCTAGGGGGCG

ATGCCGACCCT-----CTATAAAAGCGGTCCCCGCGCGGGCCTGGCCATTCGCGACCCGAAGCTGCGCGGGCGCGAGCCAGTT
| || |  |||          ||||| ||| | || ||||| ||||| ||||| ||| ||||| ||||| || ||
ACGCGG--CCTGCCGGGTATAAAAGCTGGGCCGGCGCGGGCCGGGCCATTCGCGACCCGGAGGTGCGCGGGCGCGGGCGAGCA
```

Alignement local des Homme et Souris

```
CGCTCC-----CCAA-----CCTCCACCTCC-----CCTCGCTCGGCCTCTATATGCTCCCGGGCTCCCTA
||| ||          ||||          ||||| || ||          || || | |||||          || |||| | |||||
CGCCCAGCAGCTCAAGACCAAGACTCGCCCTCCGCCCCACCCCTACCCCGTGCAGCCTCGGGATACTCCTGGGCTCC-G

GTGTTGGCTGGAAGTGGGTGACTTAG---AGGCTTAAAGGAG--GGGCGCCTAACC----ACGGACCACGTGTGTGCGGGGGC
|  ||||| ||| | | |||  ||||  ||| |||  ||  ||  ||||| |  |||||
GCCGTGGCTGGATACGGGCG-CCTAGGGCAGGC----AGGAGGAGGGGGCC---CCCGTACCGACCACGTGGGCGCGGGGGC

GACAGC---GCCGCCGGGGTGGGGCT-----GAGCGCTGCAAGCCGGGTTCGCCTTGCAGCGCAGGAGTCAGTGGGCGTTGC
||| ||  ||||  |||| || |||          ||||| ||          || ||  || |  ||||  |||| |  |||||
GACGGCCGGGCCG--GGGGCGGAGCTTGGAGCGAGCGCCGC-----GGCTCTGCTGGGCGCGCTGGAGGCGGTGGGCGTTGC

GCCACGATCTCTCT-----CCTAGCACTATGC-----TC-CCGCCCCACTCACCGCCTTGAAAGTCACAGGAGAAGG
||| ||  ||  ||          | |  || || |||          ||  |||||          |||          ||
GCCGCGCCTGCCTGGGGAGCGC-GGCGCTGTGCCGCGTGGTTTCGCCGCCCA-----TGCC-----GG

-CGGGCTCTAAGACCCAGCAGGCACCATCCTACTGGCGCCTTCG-ATCC-----GAGACCCGTTTGGACACCAGGGGGCG
|| || ||| ||||| |||  ||          |||||  || | ||          ||| ||  | ||||  | |||||
CCGCGCGCTAGGACCCAGCAGGCGCC-----GCGCCGCGCAGCCCGGGGACAGAGGCCGCTCGGACTCTAGGGGGCG

ATGCCGACCCT-----CTATAAAAGCGGTCCCCGCGCGGGCCTGGCCATTCGCGACCCGAAGCTGCGCGGGCGCGAGCCAGTT
| || |  |||          ||||| ||| |  ||  ||||| ||||| ||||| ||| ||||| ||| |||
ACGCGG--CCTGCCGGGTATAAAAGCTGGGCCGGCGCGGGCCGGGCCATTCGCGACCCGGAGGTGCGCGGGCGCGGGCGAGCA
```

AP2 α (-243/-232)

Boite TATA

Étape 0 : gène d'intérêt

Étape 1

choix des espèces

souris
poulet
fugu
rat

Étape 2

gènes orthologues

Blast
annotation
KOG

Étape 3

extraction des séquences

assemblage

Étape 4

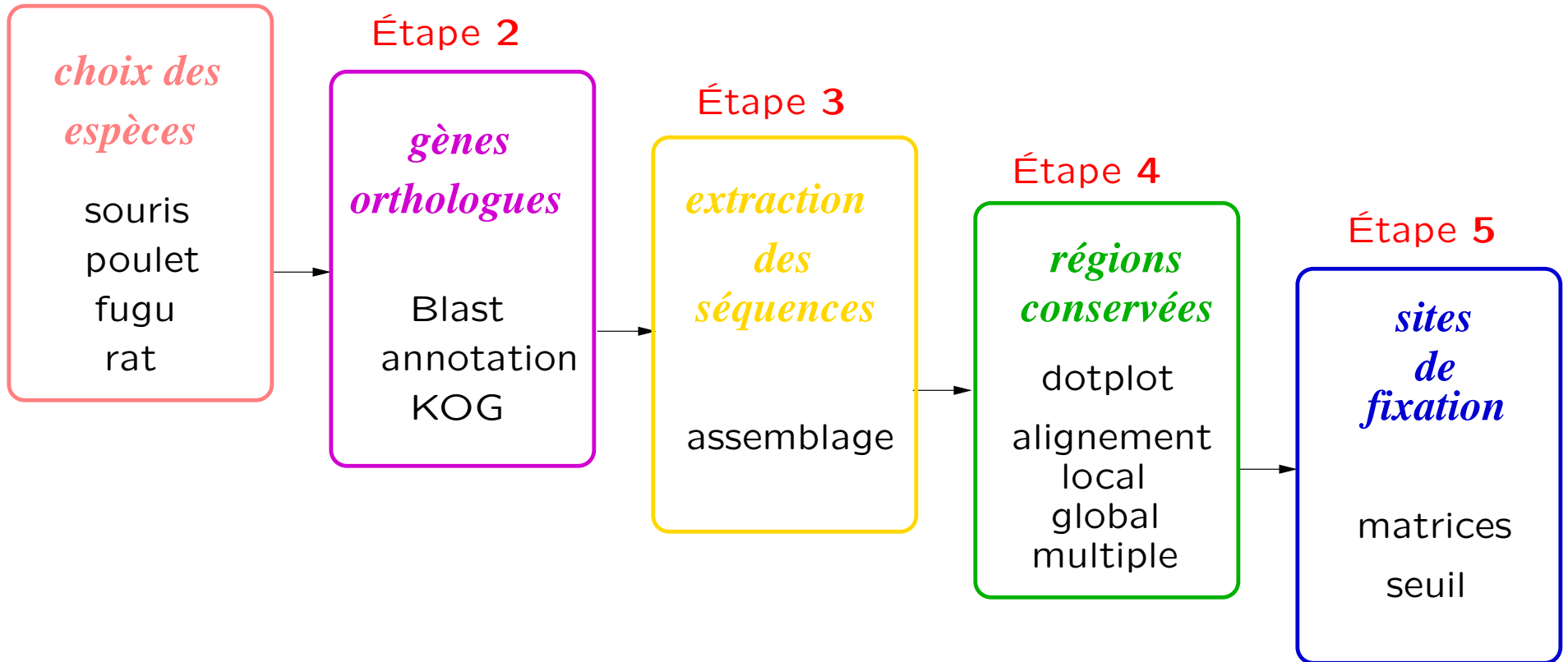
régions conservées

dotplot
alignement
local
global
multiple

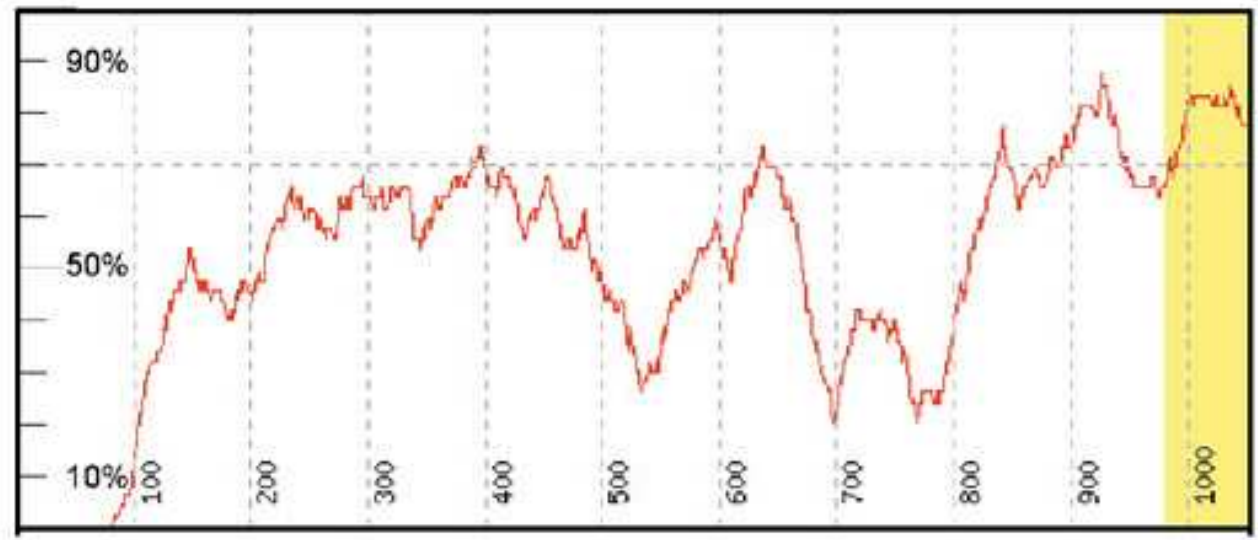
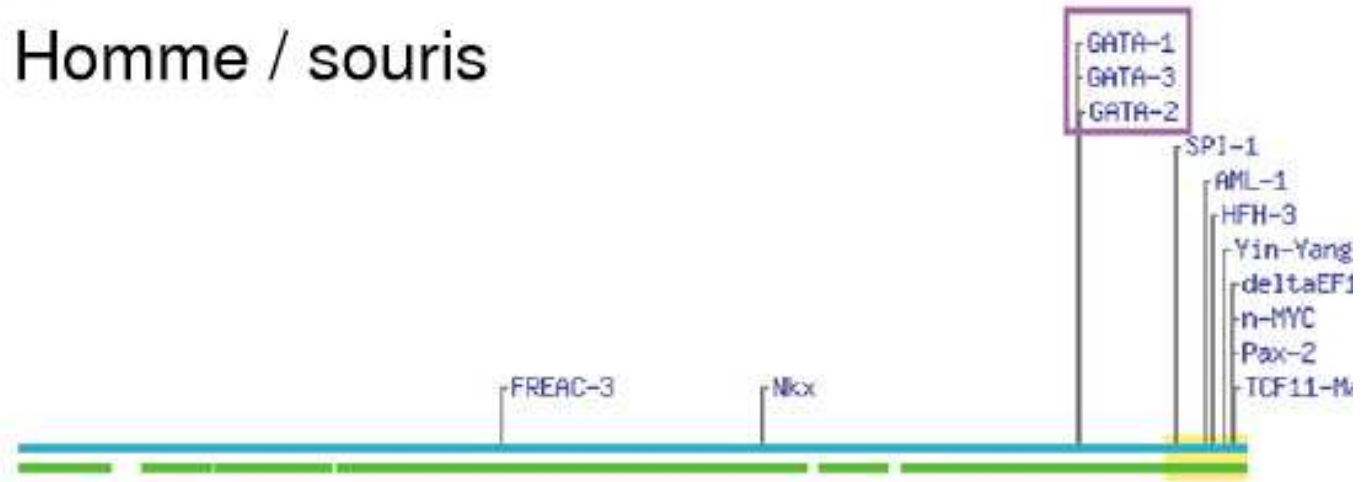
Étape 5

sites de fixation

matrices
seuil

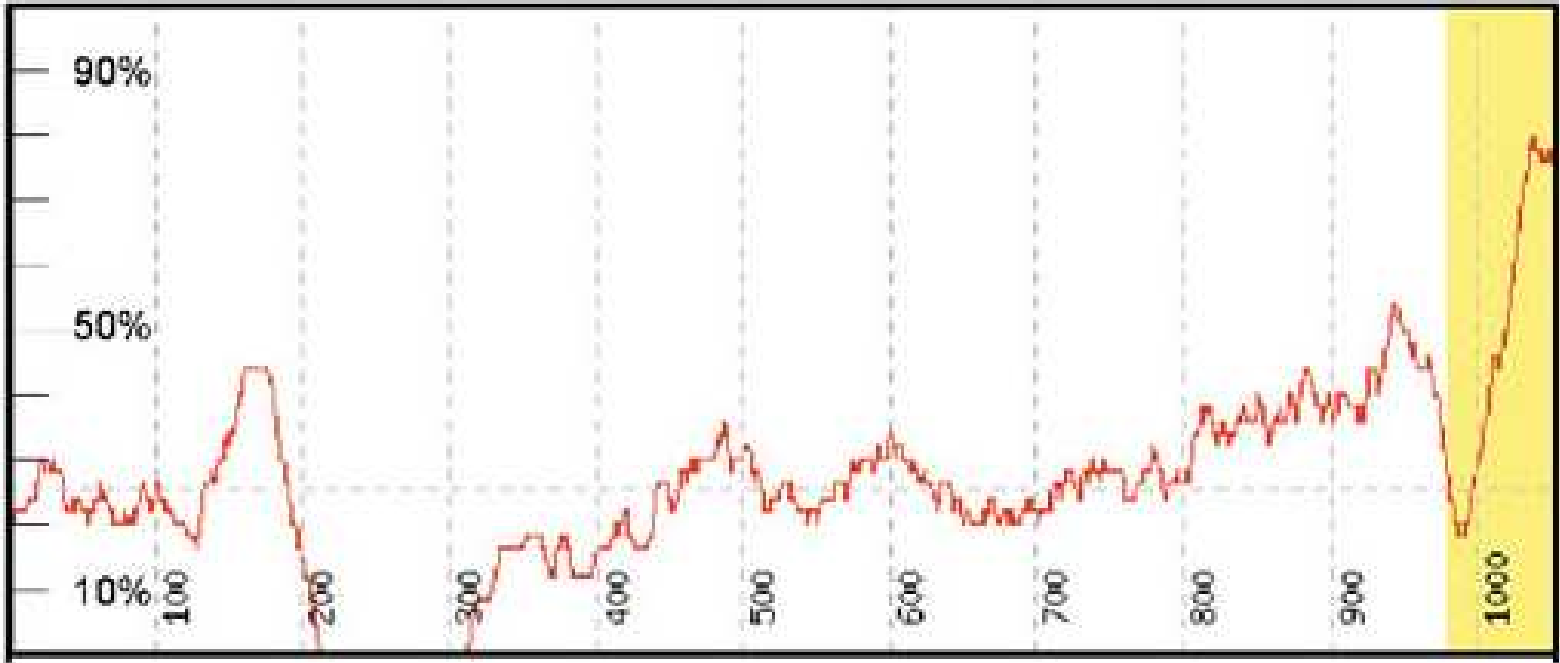


Homme / souris

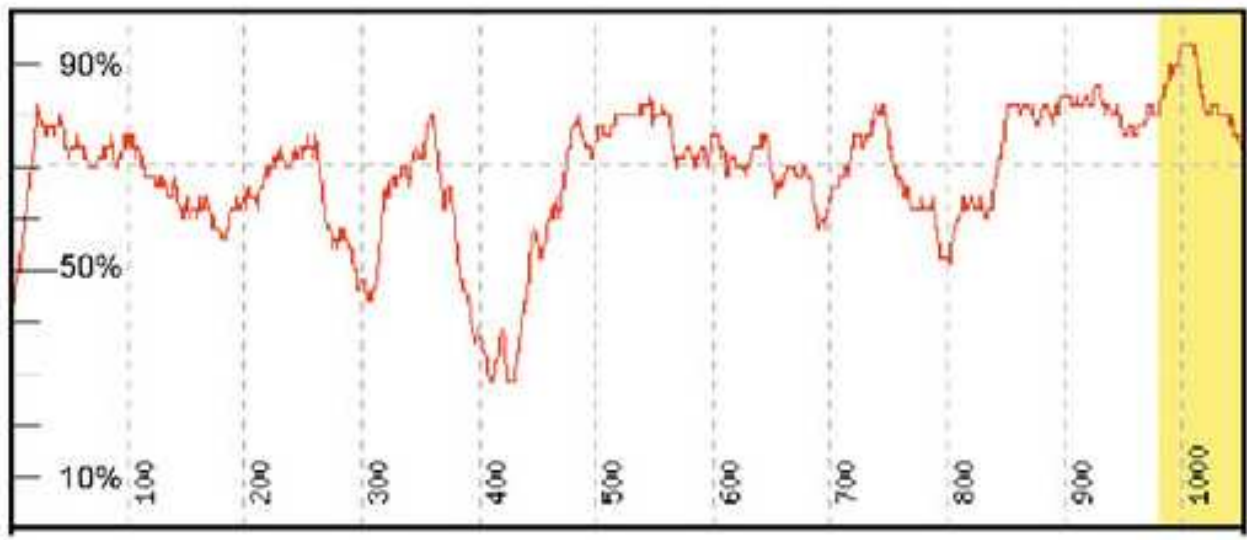
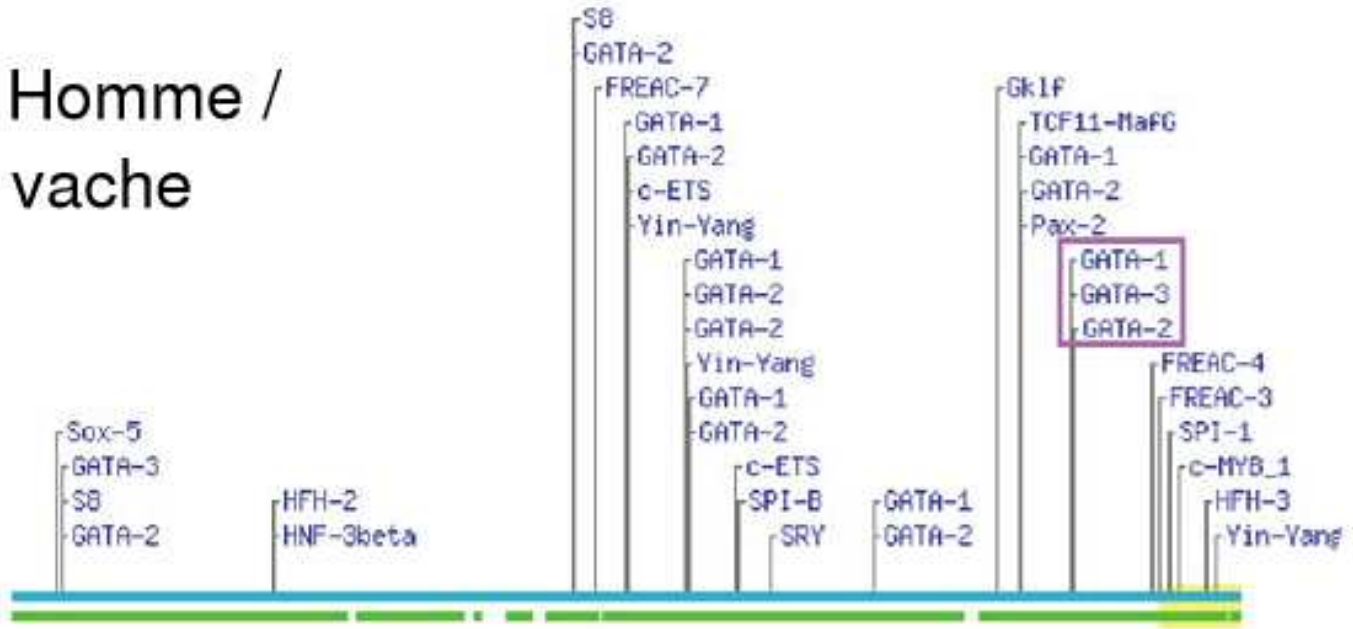


Source : ConSite

Homme /poulet



Homme / vache



Des services “tout en un”

ConSite : <http://www.phylofoot.org>

- ▷ Régions localement conservées combinées en un alignement global
- ▷ Recherche des sites potentiels pour deux séquences, puis confrontation

rVista : <http://rvista.dcode.org/>

- ▷ Nombre quelconque de génomes
- ▷ Recherche des sites potentiels sur la séquence d'intérêt, puis filtre avec les régions conservées

Piste II : Sur-représentation

Données : Ensemble de gènes potentiellement co-régulés

Clusters construits à partir de données d'expression, annotation fonctionnelle, réseaux métaboliques

Hypothèse de travail

Les motifs communs significativement sur-représentés dans les régions amont sont impliqués dans la régulation.

Deux choix stratégiques

- ▷ Modèle de fond
- ▷ Type de motifs : prédiction *de novo* (oligonucléotides, motifs approchés) ou motifs connus

Modèles de fond

Modèles théoriques: loi de probabilité

- ▷ **iid** : toutes les positions sont indépendantes et les bases sont équiprobables
- ▷ **modèle de Bernoulli** : toutes les positions sont indépendantes + %GC
- ▷ **modèle de Markov** : prise en compte des positions précédentes
 - ordre 0 : modèle de Bernoulli
 - ordre 1 : % dinucléotides : AA, AC, AG, AT, CA, etc.
 - ordre 2 : % AAA, AAC, TCA, etc.

Modèles empiriques: base de données de promoteurs, de séquences non codantes

Recherche d'oligonucléotides sur-représentés

Exemple

- ▷ 12 gènes de la levure, régulés par la méthionine
SAM2, MET6, MIUP3, MET30, MET3, MET14, MET1, SAM1, MET17, ZWF1, MET2
- ▷ analyse de la région -800 -1
- ▷ modèle de fond : %GC, régions intergéniques de la levure

Van Helden, J., Andre, B., and Collado-Vides, J.: *Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotides frequencies.* Journal of Molecular Biology 281(5), 827-842, 1998

cacgtg	cacgtg cacgtg	0.000147	13	1.26	1.00e-9	2.1e-6
ccacag	ccacag ctgtgg	0.000259	11	2.22	2.10e-5	4.5e-2
acgtga	acgtga tcacgt	0.000358	13	3.1	2.20e-5	4.6e-2
aactgt	aactgt acagtt	0.000613	17	5.28	3.90e-5	8.0e-2
actgtg	actgtg cacagt	0.000366	12	3.16	0.00011	2.4e-1
gccaca	gccaca tgtggc	0.000299	10	2.59	0.00037	7.6e-1
gcttcc	gcttcc ggaagc	0.000416	12	6.6	0.00037	7.7e-1
séquence	identifiant	fréq. att.	n.o.	n.a.	P-value	E-value

- fréq. att.* : fréquence attendue du motif,
à partir d'un modèle constitué de régions inter-géniques
- n.o.* : nombre d'occurrences observé
- n.a.* : nombre d'occurrences attendu
- P-value* : probabilité du nombre d'occurrences observé
- E-value* : nombre de motifs attendu avec cette probabilité

Alignement des motifs trouvés

cluster # 1		cluster # 2	
tcacgt..	..acgtga	aactgt..	..acagtt
.cacgtg.	.cacgtg.	.actgtg.	.cacagt.
..acgtga	tcacgt..	..ctgtgg	ccacag..
tcacgtga	tcacgtga	aactgtgg	ccacagtt

complexe Met4p/Cbfl/Met28

Met31p

Motifs approchés sur-représentés

Algorithmes probabilistes : part d'un motif tiré au sort dans le texte, et procède par modifications successives jusqu'à obtenir un motif approché surreprésenté stable.

Deux exécutions successives sur les mêmes données ne produisent pas forcément les mêmes résultats.

- ▷ Expectation Maximization : trouve un optimum local pour la vraisemblance des données dans le modèle
- ▷ Gibbs sampling : échappe à ces optimaux locaux en ajoutant des règles de perturbations, au détriment de la garantie d'atteindre un optimum

Paramètres

- ▷ largeur du motif
- ▷ nombre d'erreurs
- ▷ Gibbs sampling : nombre de séquences où le motif est présent

Résultat

- ▷ un ensemble de motifs surreprésentés
- ▷ pour chaque motif, le nombre d'occurrences, les positions, et la probabilité de l'observation (*p-value*).

AlignACE

<http://atlas.med.harvard.edu/>

- ▷ Gibbs sampling
- ▷ bien diffusé
- ▷ infos complémentaires sur la levure (promoteurs connus, comparaison des motifs trouvés, ...)



<http://www.esat.kuleuven.ac.be/~thijs/Work/MotifSampler.html>

- ▷ Gibbs sampling
- ▷ recherche sur 1 ou 2 brins, en option
- ▷ calcul d'une p-value
- ▷ modélisation des séquences avec un modèle de Markov propre à chaque organisme (métazoaires)

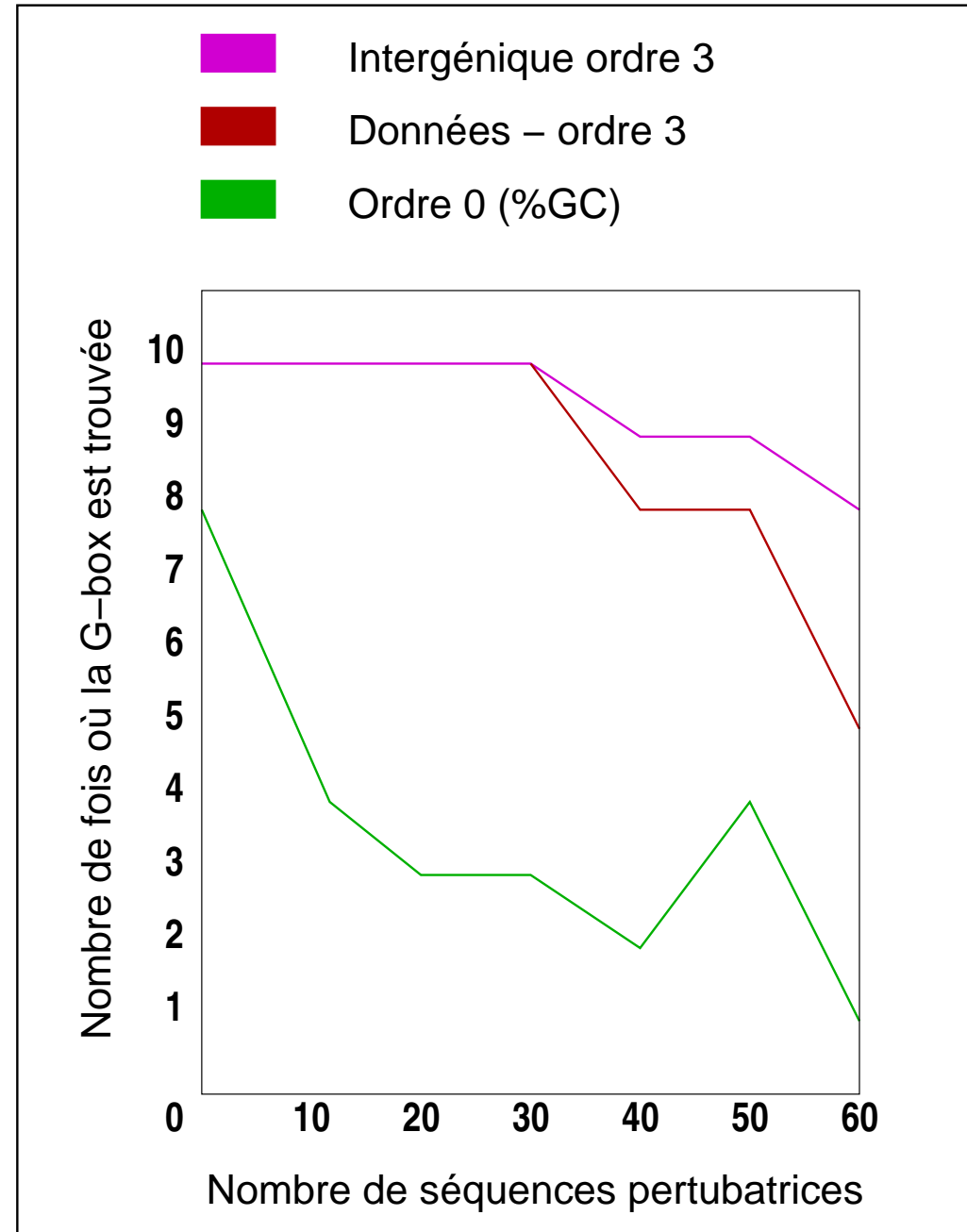
Exemple : gènes de *Arabidopsis thaliana* à boîte GC

- ▷ 33 gènes
- ▷ extraction de 500bp en amont
- ▷ 10 exécutions de **Motifsampler** avec sélection des 10 meilleurs motifs à chaque fois
- ▷ test du meilleur modèle de fond : Markov ordre 0, 1, 2, 3, 4
- ▷ test de la meilleure source de données pour le modèle de fond: données / modèle universel externe
- ▷ test à la résistance au bruit : ajout progressif de séquences intrues, sans boîte GC

Thijs G., Lescot M., Marchal K., Rombauts S., De Moor B., Rouzé P., Moreau Y.: *A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling.* Bioinformatics, 17 - 12, 1113-1122 (2001)

- ▷ modèle de Markov d'ordre **3**
- ▷ modèles basés sur les données, **modèles externes**
- ▷ ajout de séquences perturbatrices

10 runs avec sélection
des 10 meilleurs motifs
à chaque fois





<http://meme.sdsc.edu>

- ▷ Expectation Maximization
- ▷ plusieurs points de départ
- ▷ pas de connaissance a priori du nombre de séquences où le motif est présent

Sensibilité : détecte des surreprésentations jusqu'à 20% des séquences

Timothy L. Bailey and Charles Elkan,: *Fitting a mixture model by expectation maximization to discover motifs in biopolymers.* Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, 1994.

Prediction de novo: bilan

- ▷ pas de garantie
- ▷ fonctionne d'autant mieux que les régions à analyser sont courtes
- ▷ produit beaucoup de faux positifs

Tentatives de filtrage a posteriori

- ▷ comparer avec les matrices connues (à partir des sites)
- ▷ propriétés biochimiques des facteurs de transcription
- ▷ palindromes

Recherche de matrices surreprésentées

Toucan



- ▷ modèle de fond : modèles de Markov (Motifsampler)
- ▷ matrices : TRANSFAC, PlantCare
- ▷ détecte les facteurs sur-représentés pour l'ensemble des séquences

Bons résultats sur des séquences de quelques centaines de bases.
Problème des matrices chevauchantes.

Aerts, S., Thijs, G., Coessens, B., Staes, M., Moreau Y., and De Moor, B.: *TOUCAN: deciphering the cis-regulatory logic of coregulated genes.* Nucleic Acids Research, 31(6), 1753-1764, 2003

Regulatory Sequence Analysis tools

<http://rsat.ulb.ac.be/rsat/>

▷ 149 organismes :

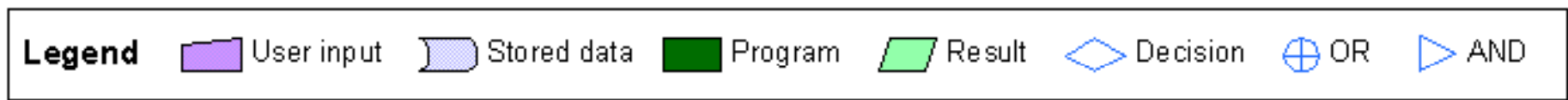
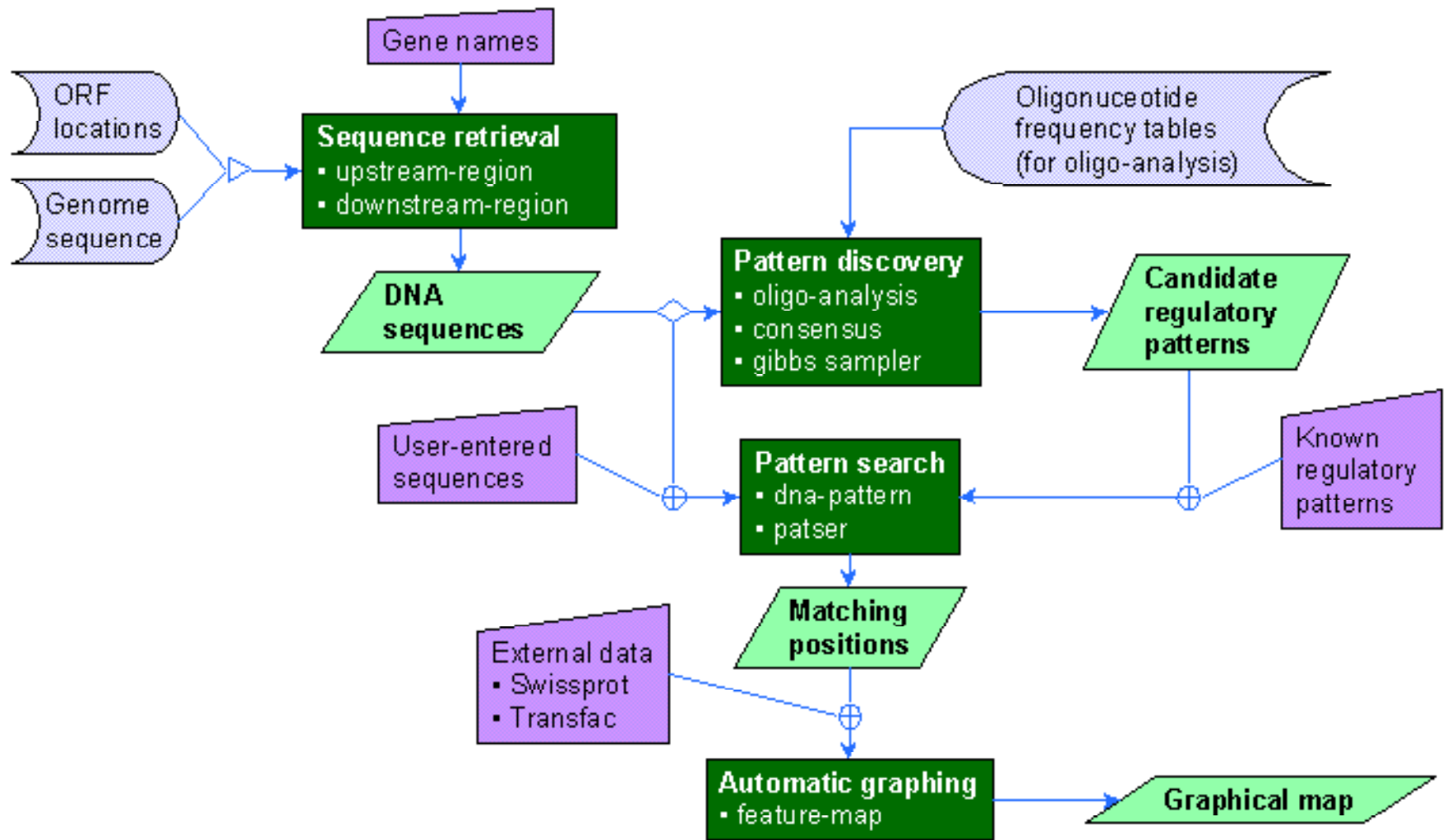
- génomes bactériens intégralement séquencés (GenBank)
- *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*
- **Warning:** les génomes de l'homme et de la souris

▷ extraction des séquences (Ensembl)

▷ oligonucléotides sur-représentés

▷ recherche avec séquences consensus

▷ recherche avec matrices

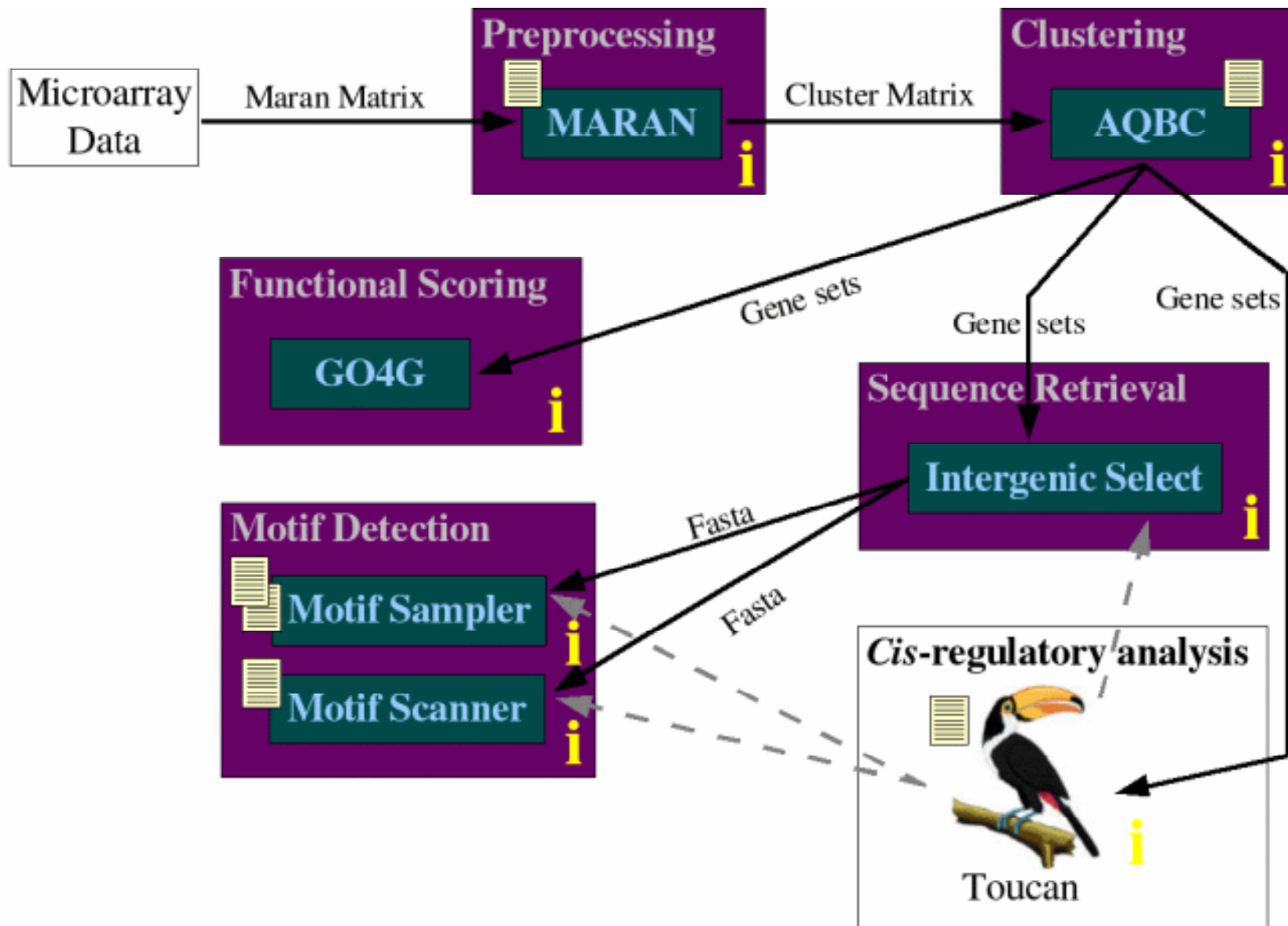


INCLUSive

de l'analyse des microarrays aux régions régulatrices

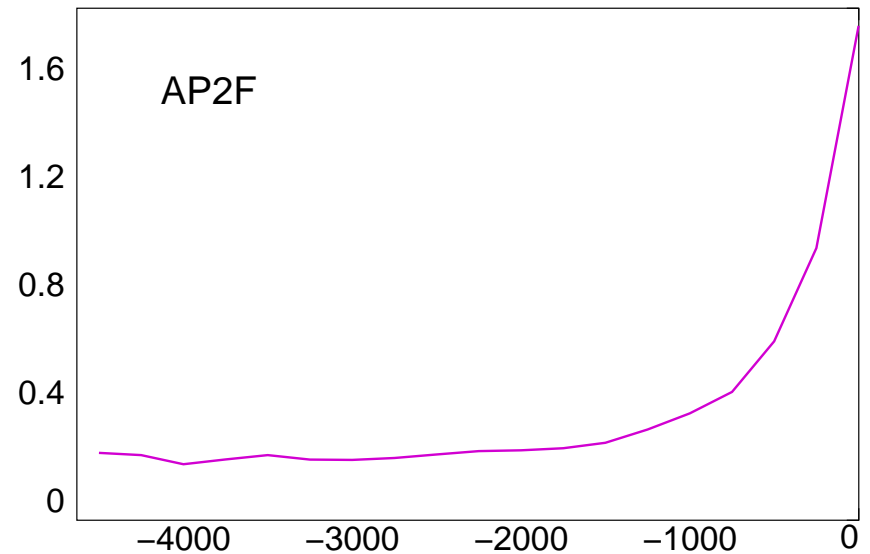
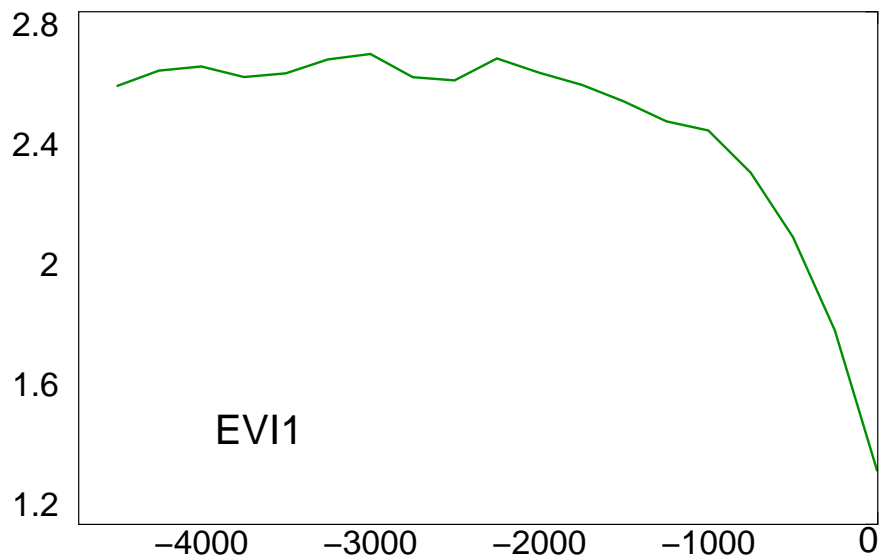
<http://www.esat.kuleuven.ac.be/~dna/BioI/Software.html>

- ▷ analyse des puces : ANOVA
- ▷ clustering
- ▷ analyse des régions régulatrices : MotifSampler, Toucan
- ▷ Gene Ontology



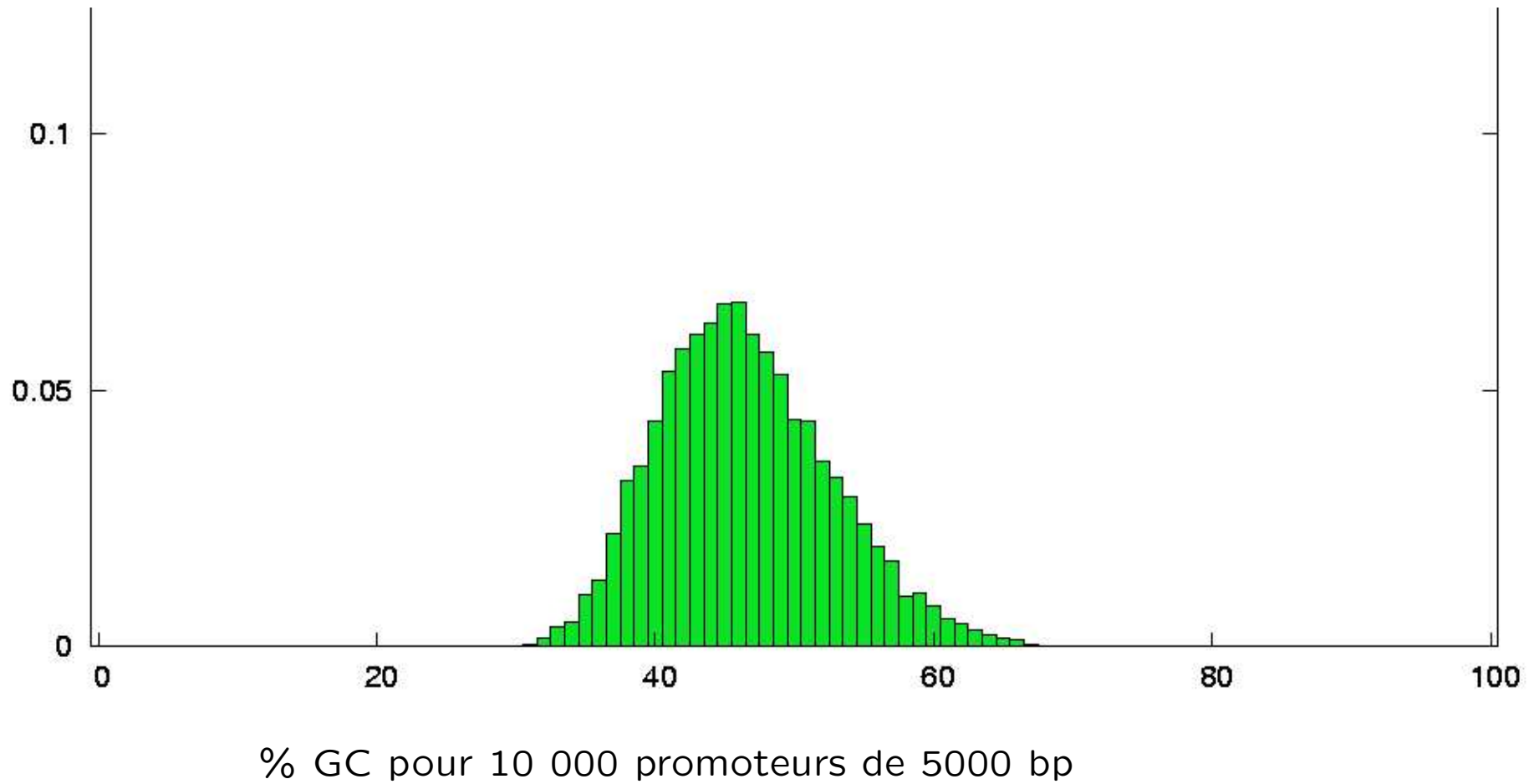
Le cas du génome humain

- ▷ pourcentage en GC
- ▷ pourcentage en dinucléotides
- ▷ distribution des sites potentiels

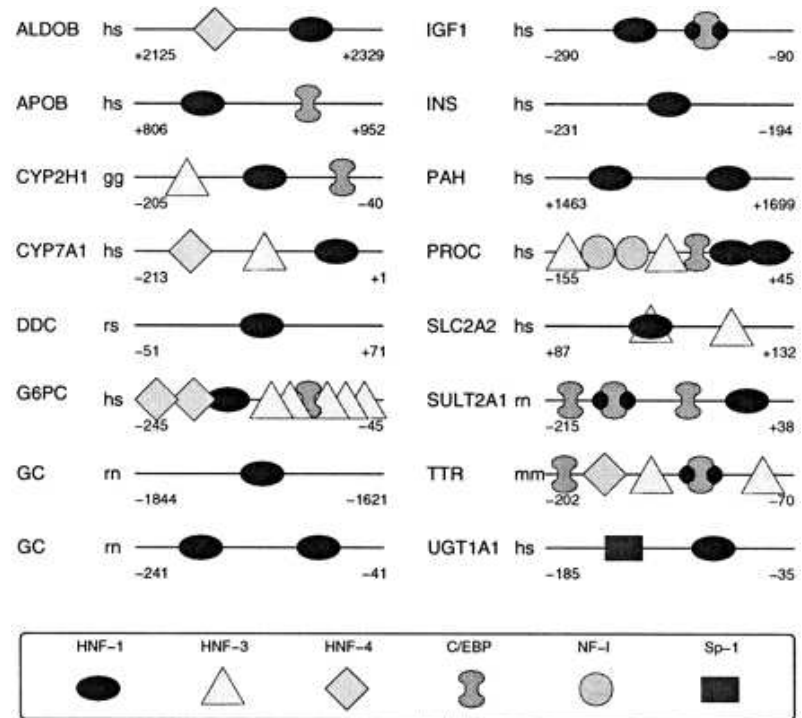


Nombre moyen de sites prédits par fenêtre glissante de 500 bp pour 5000 promoteurs humains. 0 est le TSS.

▷ disparité entre les gènes



Piste III: formation de modules



Source : Krivan & Wasserman

- ▷ nombre d'éléments ?
- ▷ distance et ordre conservés ?
- ▷ portée sur le génome ?

- ▷ **Compel**: base de données de modules (eucaryotes)

<http://compel.bionet.nsc.ru/new/index.html>

- ▷ **GEMS launcher**: recherche d'occurrences de modules définis

http://www.genomatix.de/software_services/software/GEMS_Launcher/GEMS_launcher_stb.htm

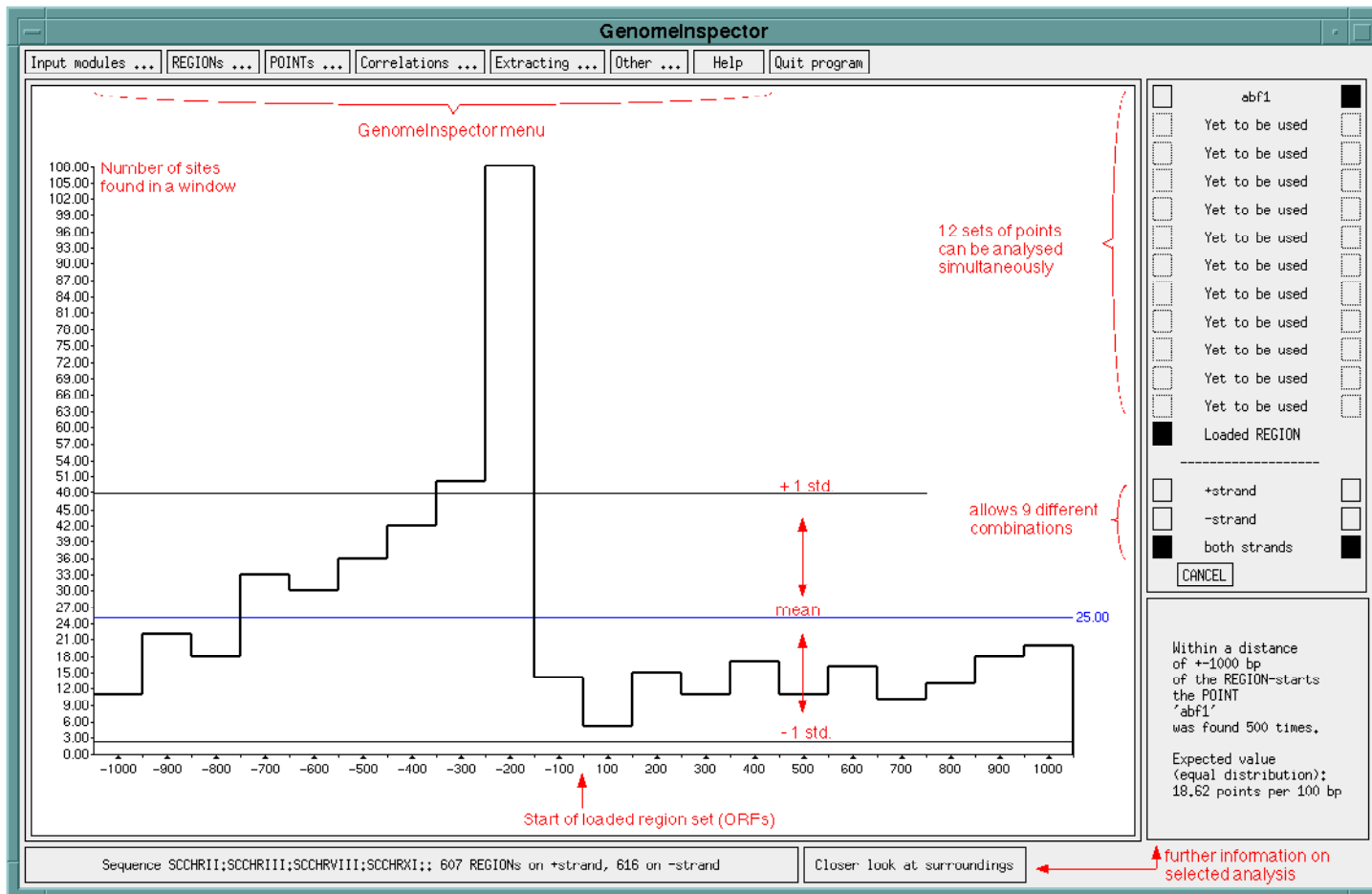
- ▷ **TFCD - Modules dans la levure** : recherche exhaustive de tous les modules présents dans les promoteurs de la base MIPS

Alvis Brazma, Jaak Vilo, Esko Ukkonen, Kimmo Valtonen: *Data Mining for Regulatory Elements in Yeast Genome*. ISMB 1997: 65-74

GenomeInspector :

<http://www.hgmp.mrc.ac.uk/Registered/Option/genomeinspector.html>

Distances et positions conservées entre les éléments du module



ModuleSearcher

<http://www.esat.kuleuven.ac.be/saerts/software/modulesearcher.html>

- ▷ filtre humain /souris
- ▷ nombre d'éléments du module
- ▷ recherche exhaustive de tous les modules dans des fenêtres glissantes (200bp, 100bp)
- ▷ score de significativité des modules

Exemple : 13 gènes co-exprimés avec la cycline B2 durant le cycle cellulaire, module de 4 éléments, fenêtre de longueur 100: CEBPA+STAF+NFY+TCF4

Aerts S., Van Loo P., Thijs G., Moreau Y., De Moor B.: *Computational detection of cis-regulatory modules.* Bioinformatics 19, sup 2, ii5-ii14, 2003

Recherche de gènes spécifiques au muscle

- ▷ facteurs impliqués : Mef-2, Myf, Sp-1, SRF, Tef
- ▷ construction de PWM (à partir de séquence de la littérature)
- ▷ recherche facteur par facteur : 60% des promoteurs de EPD
- ▷ combinaison de facteurs sur des fenêtres de 200 bp + analyse par régression logistique : 60% de l'ensemble d'apprentissage, 4% EPD

Wyeth W. Wasserman and James W. Fickett *Identification of regulatory regions which confer muscle-specific gene expression* Journal of Molecular Biology 278- 1, 167-181, 1998