Analyse des régions régulatrices et transcriptome

Hélène Touzet

touzet@lifl.fr
http://www.lifl.fr/BIOINFO

Équipe Bioinfo — LIFL UMR CNRS 8022 — USTL

October 3, 2006

Analyse bio-informatique des biopuces

conception

expériences

normalisation

formation de clusters

groupes de gènes
co-exprimés

Que faire des groupes de gènes co-régulés ?

Comprendre les mécanismes de contrôle de l'expression des gènes

- **Facteurs de transcription**, premiers acteurs de la régulation transcriptionnelle
- **D**omaine de liaison à l'ADN
- **E**lément cis-régulateur: court motif nucléique conservé
- **L**es séquences non-codantes sont disponibles (2005 : $>$ 250 génomes)

Modélisation des sites de fixations sur l'ADN

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*

Modélisation des sites de fixation sur l'ADN

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*

Représentation d'un site par une matrice

```
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
```

▷

Matrice de Comptage

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

Ligne	position de l'alignement
Colonne	acide nucléique

Représentation d'un site par une matrice

Matrice de Comptage

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

$\triangleright$

$$F_{ij} = \frac{C_{ij} + f_i * pc}{\sum_i C_{ij} + pc}$$

i acide nucléique
j position de l'alignement
f fréquence génomique
pc pseudo-poids

Matrice de Fréquences corrigées

A	C	G	T
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

Matrices de fréquences corrigées

Matrice de Fréquences corrigées

A	C	G	T
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

▷

$$P_{ij} = \log\left(\frac{F_{ij}}{f_i}\right)$$

i acide nucléique
j position de l'alignement
f fréquence génomique

Matrice de Poids

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

- **P**oids positif : les bases plus fréquentes que la moyenne
- **P**oids négatif : les bases moins fréquentes que la moyenne

Localisation des sites de fixation

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

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

Localisation des sites de fixation

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

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

T A C G G A T A C G

Localisation des sites de fixation

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

1 on repère le poids de chaque position dans la matrice

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

T A C G G A T A C G

Localisation des sites de fixation

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

1. on repère le poids de chaque position dans la matrice
2. le score est la somme des poids

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

T A C G G A T A C G score : 6.16

Localisation des sites de fixation

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

On recommence à la position suivante

Score de ACGGATACGT

```
T A C G G A T A C G T T G A C C A T G G T A C C T
T A C G G A T A C G                 score : 6.16
  A C G G A T A C G T               score : -1.86
```

Banque de données de sites de fixation

- **T**ransfac
  `http://www.gene-regulation.com/`
  eucaryotes, 5327 facteurs, 674 matrices, construites à partir de 13112 séquences
- **J**aspar
  `http://forkhead.cgb.ki.se/JASPAR/`
  eucaryotes, matrices *haute qualité*
- **P**lantCare : `http://sphinx.rug.ac.be:8080/PlantCARE/`
- **L**evure : SCPD (`http://cgsigma.cshl.org/jian/`),
  YRSA (`http://forkhead2.cgb.ki.se/yrsa/`)
- etc.

Pertinence de la recherche de sites de fixation

- **Q**ualité du modèle des matrices
  - **L**a très grande majorité des sites prédits se lient *in vitro*
  - **L**e score reflète l'affinité thermodynamique du site
- **P**roblème de prédictions non spécifiques
  - **L**es motifs décrits par les matrices sont courts et dégénérés
  - **N**ombre élévé d'occurrences dues au hasard
  - **Futility theorem**: presqu'aucun des sites prédits *in silico* n'est fonctionnel *in vivo*.

- **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

Comment améliorer la spécificité des prédictions

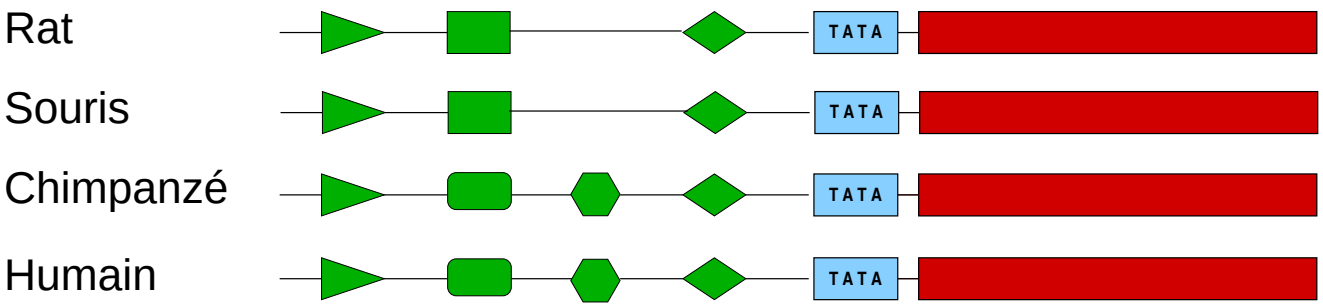
- **Piste I:** génomique comparative, empreinte phylogénétique

  tirer partie de la conservation entre les espèces

- **Piste II:** motifs sur-représentés

  tirer partie du fait que l'on dispose de gènes co-exprimés

Piste I : Empreinte phylogénétique



1. **L**es mécanismes de régulation sont conservés entre les espèces
2. **L**a pression de sélection s'applique aux régions régulatrices (au même titre qu'aux régions codantes)
3. **L**a 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.

Alignement Homme/Souris

```
CGCTCC------------CCAA-------CCTCCACCTCC--------CCTCGCTCGGCCTCTATATGCTCCCGGGCTCCCTA
 ||| ||            ||||       ||||| || ||        || ||  | |||||   || |||| ||||||||
CGCCCCAGCAGCTCAAGACCAAGACTCGCCCTCCGCCCCCCACCCCTACCCCGTGCAGCCTCGGGATACTCCTGGGCTCCC-G
GTGTTGGCTGGAAGTGGGTGACTTAG---AGGCTTAAAGGAG--GGGCGCCTAACC----ACGGACCACGTGTGTGCGGGGGC
|   ||||||||   ||| | | |||   ||||    |||||  ||| |||   ||    || ||||||||| | ||||||||
GCCGTGGCTGGATACGGGCG-CCTAGGGCAGGC----AGGAGGAGGGGGCC---CCCGCTACCGACCACGTGGGCGCGGGGGC
GACAGC---GCCGCCGGGGTGGGGCT------GAGCGCTGCAAGCCGGGTTCGCCTTGCAGCGCAGGAGTCAGTGGGCGTTGC
||| ||   ||||  |||| || |||      |||||| ||      || ||  || |  |||| |||| | ||||||||||
GACGGCCGGGCCG--GGGGCGGAGCTTGGAGCGAGCGCCGC------GGCTCTGCTGGGCGCGCTGGAGGCGGTGGGCGTTGC
GCCACGATCTCTCT-----CCACTAGCACTATGC--------TC-CCGCCCCACTCACCGCCTTGGAAAGTCACAGGAGAAGG
||| ||  ||  ||      | |  || || |||        || ||||||||      |||                  ||
GCCGCGGCCTGCCTGGGGAGCGC-GGCGCTGTGCCGCGTGGTTCGCCGCCCCA-----TGCC-------------------GG
-CGGGCTCTAAGACCCAGCAGGCACCATCCTACTGGCGCCTTCG-ATCC--------GAGACCCGTTTGGACACCAGGGGGCG
 || || ||| ||||||||||| ||         |||||  || | ||        ||| ||   | |||| | ||||||||
CCGCGCGCTAGGACCCAGCAGGCGCC---------GCGCCGCCGCAGCCCGGGGACAGAGGCCGCCTCGGACTCTAGGGGGCG
ATGCCGACCCT-----CTATAAAAGCGGTCCCCGCGCGGGCCTGGCCATTCGCGACCCGAAGCTGCGCGGGCGCGAGCCAGTT
 | || |  |||      ||||||||| |  || ||||||||| |||||||||||||||| || ||||||||||| || ||
ACGCGG--CCTGCCGGGTATAAAAGCTGGGCCGGCGCGGGCCGGGCCATTCGCGACCCGGAGGTGCGCGGGCGCGGGCGAGCA
```

Alignement Homme/Souris

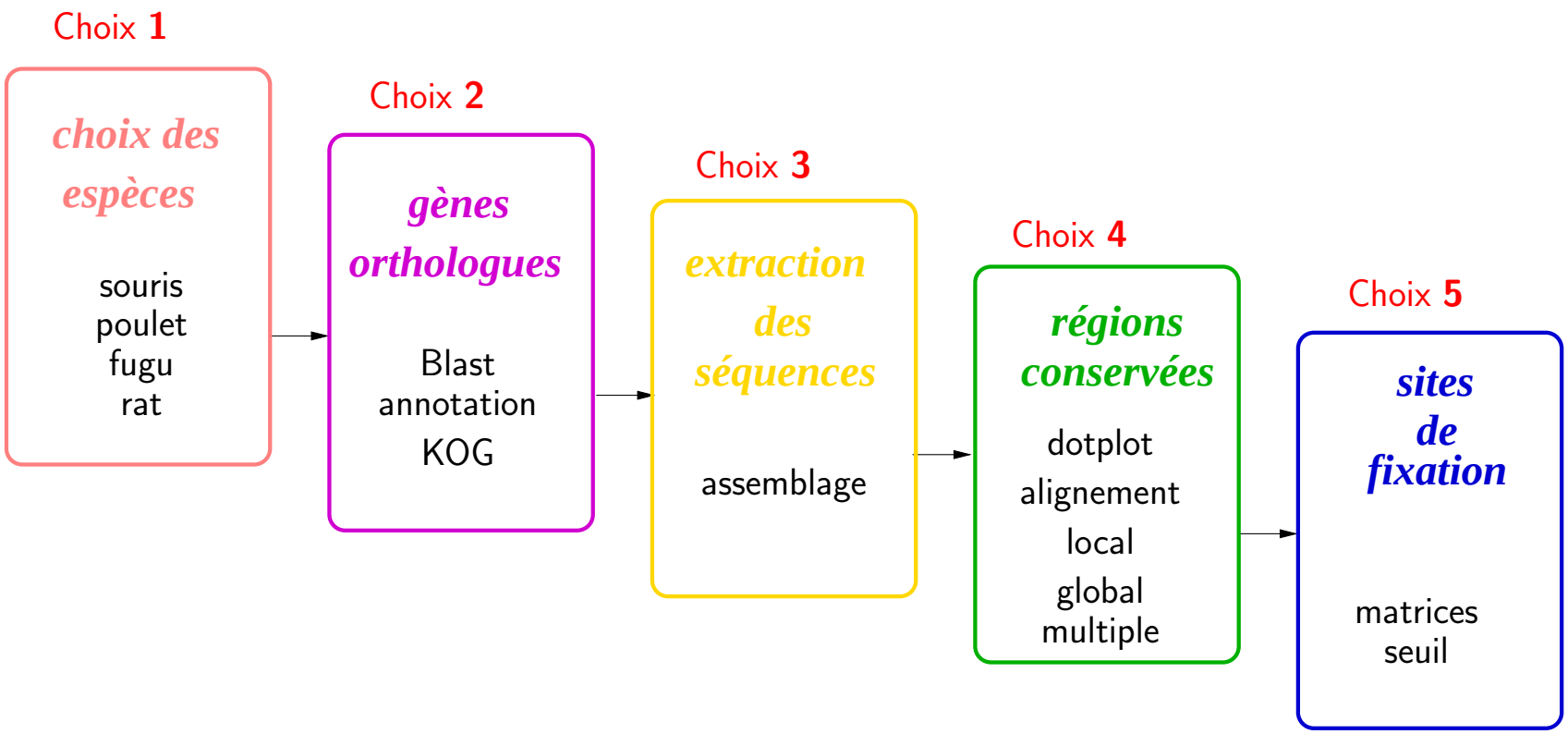
```
CGCTCC------------CCAA-------CCTCCACCTCC--------CCTCGCTCGGCCTCTATATGCTCCCGGGCTCCCTA
 ||| ||            ||||       ||||| || ||        || ||  | |||||   || |||| ||||||||
CGCCCCAGCAGCTCAAGACCAAGACTCGCCCTCCGCCCCCCACCCCTACCCCGTGCAGCCTCGGGATACTCCTGGGCTCCC-G
GTGTTGGCTGGAAGTGGGTGACTTAG---AGGCTTAAAGGAG--GGGCGCCTAACC----ACGGACCACGTGTGTGCGGGGGC
|   ||||||||   ||| | | |||   ||||    |||||  ||| |||   ||    || ||||||||| | ||||||||
GCCGTGGCTGGATACGGGCG-CCTAGGGCAGGC----AGGAGGAGGGGGCC---CCCGCTACCGACCACGTGGGCGCGGGGGC
GACAGC---GCCGCCGGGGTGGGGCT------GAGCGCTGCAAGCCGGGTTCGCCTTGCAGCGCAGGAGTCAGTGGGCGTTGC
||| ||   ||||  |||| || |||      |||||| ||      || ||  || |  |||| |||| | |||||||||||
GACGGCCGGGCCG--GGGGCGGAGCTTGGAGCGAGCGCCGC------GGCTCTGCTGGGCGCGCTGGAGGCGGTGGGCGTTGC
GCCACGATCTCTCT-----CCACTAGCACTATGC--------TC-CCGCCCCACTCACCGCCTTGGAAAGTCACAGGAGAAGG
||| ||  ||  ||      | |  || || |||        || ||||||||      |||                  ||
GCCGCGGCCTGCCTGGGGAGCGC-GGCGCTGTGCCGCGTGGTTCGCCGCCCCA-----TGCC-------------------GG
-CGGGCTCTAAGACCCAGCAGGCACCATCCTACTGGCGCCTTCG-ATCC--------GAGACCCGTTTGGACACCAGGGGGCG
 || || ||| |||||||||||| ||         |||||  || | ||        ||| ||   | |||| | ||||||||
CCGCGCGCTAGGACCCAGCAGGCGCC---------GCGCCGCCGCAGCCCGGGGACAGAGGCCGCCTCGGACTCTAGGGGGCG
ATGCCGACCCT-----CTATAAAAGCGGTCCCCGCGCGGGCCTGGCCATTCGCGACCCGAAGCTGCGCGGGCGCGAGCCAGTT
 | || |  |||      ||||||||| |  || ||||||||| |||||||||||||||| || ||||||||||| || ||
ACGCGG--CCTGCCGGGTATAAAAGCTGGGCCGGCGCGGGCCGGGCCATTCGCGACCCGGAGGTGCGCGGGCGCGGGCGAGCA
```

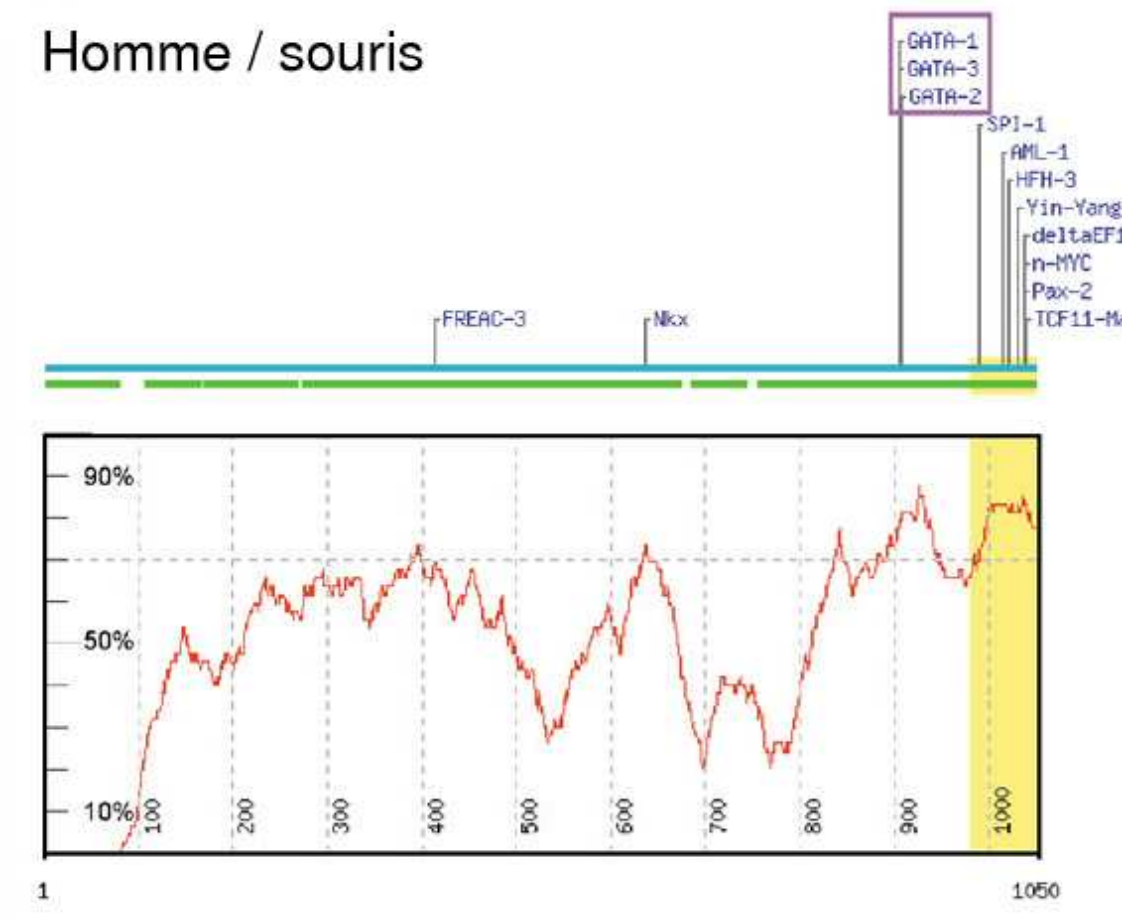
AP2 α (-243/-232)

Boite TATA

Empreinte phylogénétique : mise en œuvre

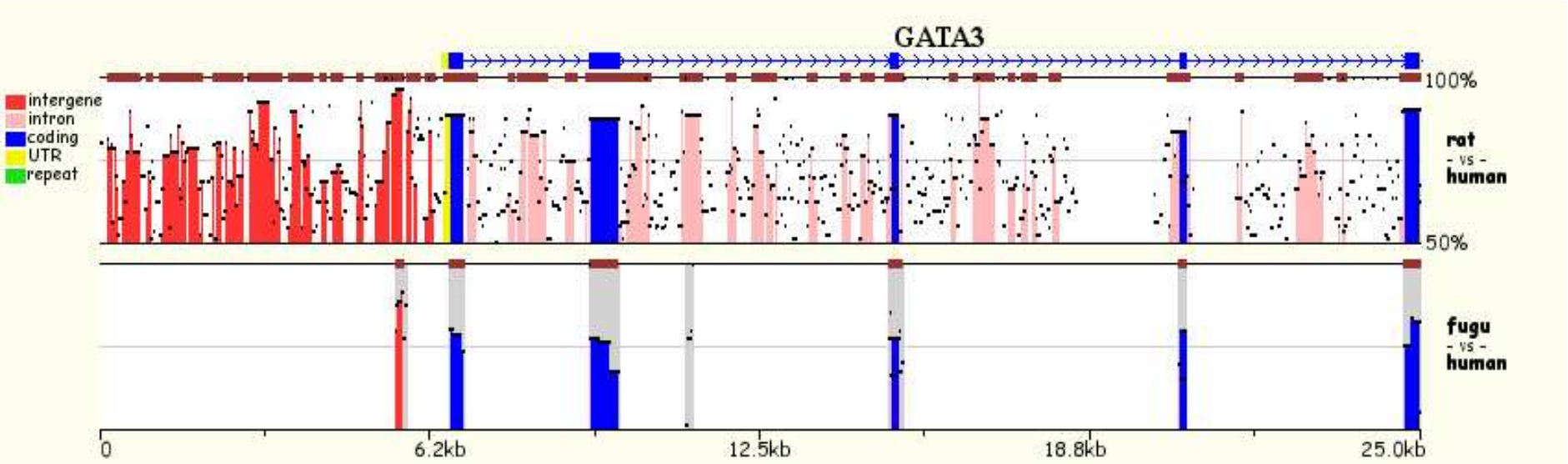
Donnée initiale : gène d'interêt





ConSite : `http://www.phylofoot.org`

- **R**égions localement conservées combinées en un alignement global
- **R**echerche des sites potentiels pour deux séquences, puis confrontation



rVista : `http://rvista.dcode.org/`

- **N**ombre quelconque de génomes
- **R**echerche des sites potentiels sur la séquence d'interêt, puis filtre avec les régions conservées

Piste II : Sur-représentation

- **Données :** Ensemble de gènes potentiellement co-régulés
- **Résultat :** Sites de fixation correspondant à des facteurs impliqués dans le contrôle de la majortité des gènes
- **Hypothèse de travail:**

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

Recherche d'oligonucléotides sur-représentés

- **O**ligonucléotide = court motif exact (6 à 8 bases)
- **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
- **RSA-tools** : `http://rsat.ulb.ac.be/rsat/`
  - 250 génomes (NCBI)
  - recherche d'oligonucléotides sur-représentés
  - recherche de matrices

séquence	2 brins	n.obs.	n.att.	P-value
cacgtg	cacgtg|cacgtg	13	1.26	1.00e-9
ccacag	ccacag|ctgtgg	11	2.22	2.10e-5
acgtga	acgtga|tcacgt	13	3.1	2.20e-5
aactgt	aactgt|acagtt	17	5.28	3.90e-5
actgtg	actgtg|cacagt	12	3.16	0.00011
gccaca	gccaca|tgtggc	10	2.59	0.00037
gcttcc	gcttcc|ggaagc	12	6.6	0.00037

n.obs. : *nombre d'occurrences observé*
n.att. : *nombre d'occurrences attendu*
P-value : *probabilité du nombre d'occurrences observé*

```
     motif 1                                  motif 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

- **M**otif approché = motif avec erreurs
- **P**roblème informatique non-résolu
- **D**es débuts de solution
  - Algorithmes probabilistes : 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 optimums locaux en ajoutant des règles de perturbations, au détriment de la garantie d'atteindre un optimum
- **R**ésultats perfectibles sur des données humaines

Assessing computational tools for the discovery of transcription factor binding sites
Martin Tompa et al. Nature Biotechnology 23, 137 - 144 (2005)

AlignACE

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

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

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

- Gibbs sampling
- modèlisation des séquences avec un modèle de Markov propre à chaque organisme (métazoaires)
- suite logicielle intégrée: des puces (ANOVA) aux gènes

http://meme.sdsc.edu

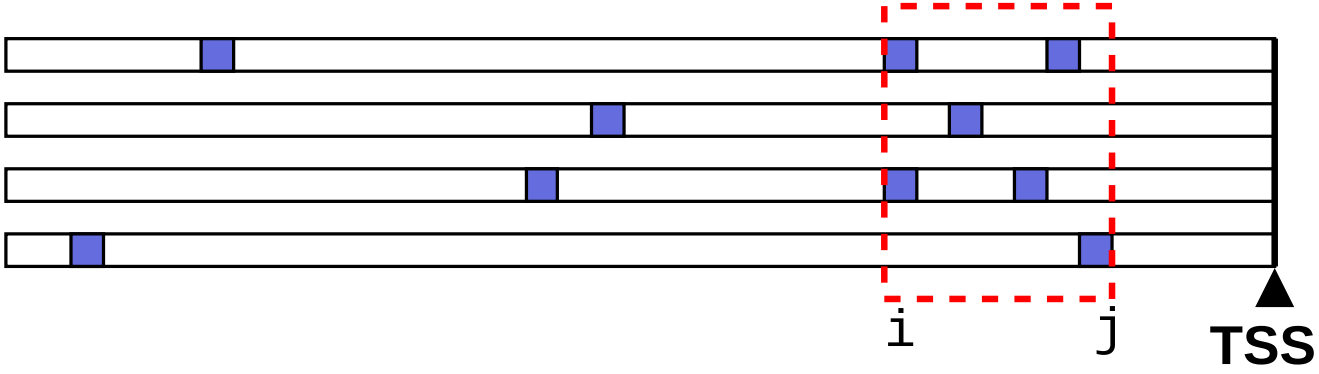
- Expectation Maximization
- pas de connaissance a priori du nombre de séquences où le motif est présent

TFM-Explorer

`http://bioinfo.lifl.fr/TFM-Explorer/`

Predicting transcription factor binding sites using local over-representation and comparative genomics. M. Defrance, H. Touzet, BMC Bioinformatics 2006, 7:396

- **A**pproche hybride: conservation + sites sur-représentés
- **U**tilisation des sites connus: matrices Transfac et Jaspar
- **S**tratégie adaptée aux longues régions régulatrices: recherche locale, qui prend en compte les variations de composition le long du génome
- **M**ulti-organismes: homme/rat/souris



Les gènes cibles des facteurs Rel/NF-κB

(Collaboration avec C. Abbadie - UMR CNRS 8117 - journée thématique B13)

- Puces utilisées: membranes Research Genetics
  - 4000 gènes humains connus
  - radiomarquage
  - deux hybridations successives
- Expériences réalisées: sur-expression de c-Rel

cellules	différentiel	vecteur	cinétique
HeLa	c-Rel/contrôle	plasmide	48h
HeLa	c-Rel/contrôle	plasmide	48h
HeLa	c-Rel/contrôle	adénovirus	18h
Kératino	c-Rel/contrôle	adénovirus	18h
Kératino	c-Rel/contrôle	adénovirus	18h

Sélection de gènes cibles potentiels

- **D**ifférentiel d'expression > 2
- **C**ritère de reproductibilité : deux expériences
- **E**limination des gènes mal identifiés (qualité des clones Image)

69 gènes

. . . avec des doutes sur les données

- *existence de faux négatifs*
- *gènes activés dans HeLa, réprimés dans les kératino et inversement*

- **Q**uels sont les gènes qui sont de bons candidats pour être des gènes cibles directs ?
- **L**es sites sont-ils proximaux, distaux ?
- **Il** y a-t-il d'autres facteurs associés ?

Analyse avec TFM-Explorer

1. identification des numéros RefSeq (`NM_` un identifiant par ARNm)
2. sélection des régions à étudier [-10 000, + 1000] (assemblage du génome humain UCSC)
3. filtrage

facteur	région	significativité	nb sites	nb gènes
NFKAPPAB65_01	-0292:-0020	1.548e-07	69	65%
GC_01	-0111:-0049	1.574e-07	77	74%
CREL_01	-0123:-0020	2.493e-07	47	53%
NFAT_Q6	-0256:-0180	3.614e-07	50	59%
NFY_01	-0117:-0021	6.840e-07	24	31%

Facteurs de la famille Rel/NF-κB
Boîtes GC (région -164/+1)
Nuclear factor of activated T-cells
Nuclear factor Y (Y-box binding factor) (ccaat)

Appliquer la bonne méthode suivant l'organisme

- **G**énomes bactériens, levures
  - RSA-tools, MotifSampler, MEME
- **P**lantes, mammifères, ...
  - longues régions non-codantes hétérogènes
  - TFM-Explorer
- **A**pport croissant de la génomique comparative
- **D**omaine de recherche en pleine évolution

Bibliographie introductive

Applied bioinformatics for the identification of regulatory elements
W.W. Wasserman and A. Sandelin, Nature Reviews Genetics 5(4) 276-287, 2004

Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements
Zhang, Z. and Gerstein, M. J Biol. 2003;2(2):11

DNA binding sites: representation and discovery
G.D. Stormo, Bioinformatics, 16(1) 16-23, 2000