

Motifs biologiques

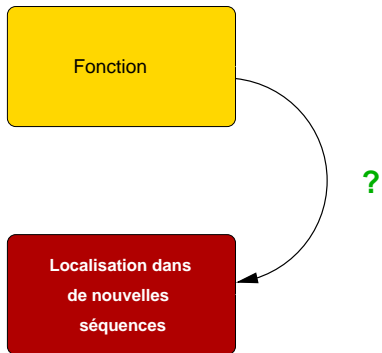
Hélène Touzet

Équipe Bioinfo — LIFL — USTL

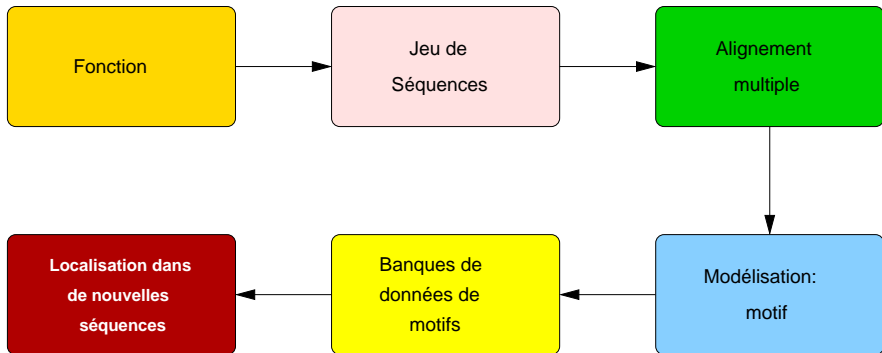
Master recherche informatique

www.lifl.fr/~touzet/masterrecherche.html

Découverte de motifs



Découverte de motifs



- ▶ **Problème 1:** trouver une représentation des motifs à partir des alignements multiples
- ▶ **Problème 2:** concevoir des algorithmes pour localiser les occurrences des motifs dans une nouvelle séquence

Deux applications (pour ce cours)

- ▶ **Motifs protéiques**
 - ▶ Domaines enzymatiques
 - ▶ Sites fonctionnels

- ▶ **Motifs nucléiques:** site de fixation de facteur de transcription

Motifs protéiques

▶ Exemple 1: hormone pancréatique

PMY_PETMA/1-36	PEE..LSKYMLAVRNYINLITRQRY
PPY_LOPAM/1-36	PED..WASYQA AVRHYVNLITRQRY
PAHO_BOVIN/30-65	PEQ..MAQYAAELRRYINMLTRPRY
PAHO_CHICK/26-61	VED..LIRFYNDLQQYLNVTTRHRY
PAHO_ANSAN/1-36	VED..LRFYDNLQQYRLNVFRHRY
NPF_HELAS/4-39	PNE..LRQYLKELNEYAIMGTRRF
NPF_MONEX/1-39	DNKAALRDYLRQINEYFAIIGRPRF

▶ Expression Prosite

[FY]-x(3)-[LIVM]-x(2)-Y-x(3)-[LIVMFY]-x-R-x-R-[YF]

▶ Syntaxe

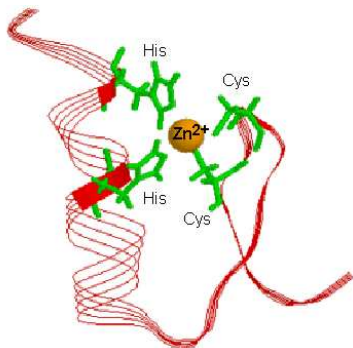
- : séparation des éléments
- x : n'importe quel acide aminé
- (3,5) : nombre d'occurrences (entre 3 et 5)
- [FY] : alternative (N, H ou G)

► Exemple 2 : doigt de zinc

YVCPFDGCN---KKFAQSTNLKSHILT---H
YKCT--VCR---KDISSESRLRTHMFKQ--HH
FQCD--ICK---KTFKNACSVKIHHKN--MH
LKCSVPGCK---RSFRKKRALRIHVSE---H
FECN--MCG---YHSQDRYEFSSHITRG--EH
YTCG--YCTEDSPSFPRPSLLESHISL--MH
YKCEFADCE---KAFSNASDRAKHQNR--TH
YKCN--QCG---IIFSQNSPFIVHQIA---H
FVCHWQDCSRELRPFKAQYMLVVHMRR---H
FRCS--ECS---RSFTHNSDLTAHMRK---H
CKCETENCN---LAFTTASNMRLLHFKR--AH
YRCSYEDCQ---TVSPTWTALQTHLKK---H
FRCV--WCK---QSFPTLEALTTHMKDS--KH
FRCGYKCG---RLYTTAHLKVVHERA---H
YRCPRENCN---RTYTTKFNLKSHILT--FH
YTCPEPHCG---RGFTSATNYKNHVRI---H

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

► Exemple 2 : doigt de zinc



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

Modélisation avec des HMM

HMM = Hidden Markov Model = Modèle de Markov caché

- ▶ **Un ensemble d'états**
- ▶ **Des probabilités de transitions** entre les états
- ▶ **Un ensemble d'observations**
- ▶ **Des probabilités d'émission** qui indiquent pour chaque état la probabilité d'y émettre telle observation

Ici : les observations sont les acides aminés

V E D - - L I R Y

V E D - - L R R Y

P N E - - L R R F

D N K A A L R R F

A E E - - L A - -

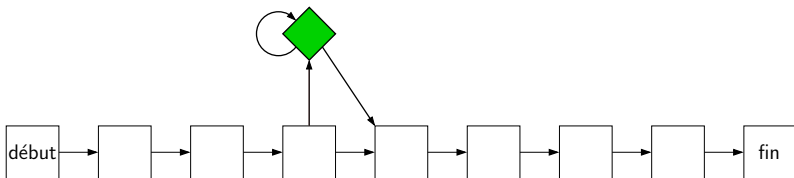
V	E	D	-	-	L	I	R	Y
V	E	D	-	-	L	R	R	Y
P	N	E	-	-	L	R	R	F
D	N	K	A	A	L	R	R	F
A	E	E	-	-	L	A	-	-

Création d'un état par colonne



V E D - - L I R Y
V E D - - L R R Y
P N E - - L R R F
D N K A A L R R F
A E E - - L A - -

Prise en compte des insertions



V E D - - L I R Y

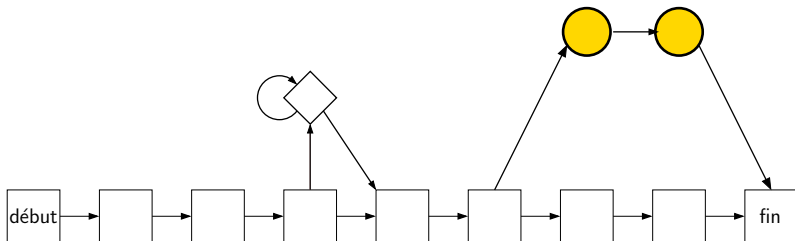
V E D - - L R R Y

P N E - - L R R F

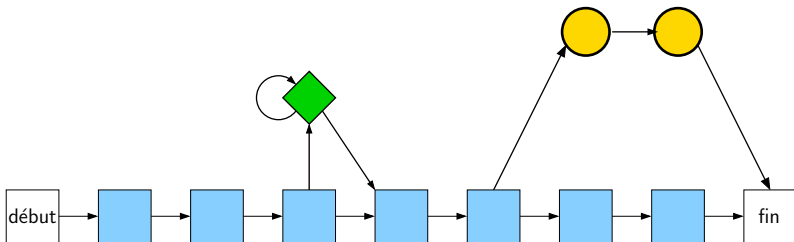
D N K A A L R R F

A E E - - L A - -

Prise en compte des délétions



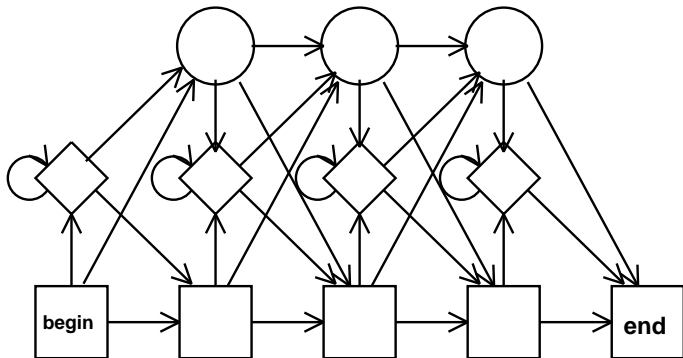
V	E	D	-	-	L	I	R	Y
V	E	D	-	-	L	R	R	Y
P	N	E	-	-	L	R	R	F
D	N	K	A	A	L	R	R	F
A	E	E	-	-	L	A	-	-



En résumé

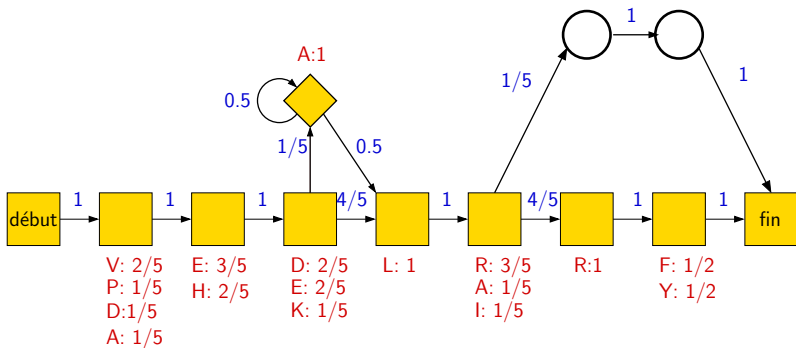
- ▶ **E**tats matchants : colonnes avec moins de 50% de -
- ▶ **E**tats d'insertion : majorité de -
- ▶ **E**tats de délétion : minorité de -
- ▶ **P**robabilités d'émission : on compte le nombre d'occurrences de chaque acide aminé
- ▶ **P**robabilités de transition : on compte le nombre de séquences empruntant la transition
- ▶ **C**orrection avec les pseudo-poids : +1 à chaque compte (loi de Laplace)

Modèle complet :



Recherche avec un profil HMM

- Score : probabilité maximale d'un mot dans le modèle



Score de VHKALARY

$$1 \times \frac{2}{5} \times 1 \times \frac{2}{5} \times 1 \times \frac{1}{5} \times \frac{1}{5} \times 1 \times 0.5 \times 1 \times 1 \times \frac{1}{5} \times 1 \times 1 \times \frac{1}{2} \times 1$$

Recherche avec un profil HMM

- ▶ **Score** : probabilité maximale d'un mot dans le modèle
- ▶ **Algorithme de Viterbi** : trouver le chemin de probabilité maximale
Recherche d'un chemin optimal dans un graphe
- ▶ Seuil d'admission : **E-value**

- ▶ **Exemple** : la séquence NPF_ARTTR contre le HMM de l'hormone pancréatique

Alignments of top-scoring domains:

domain from 3 to 36: score 48.4, E-value = 1.1e-13

```
*->yPskdfPenPGddaspEeelaqYlraLrqYinliTRpRY<-*  
      +++++      P++++s+E++++Ylr++++Yi+l++RpR+  
3    VHLR-----PRSSFSEDEYQIYLRNVSKYIQLYGRPRF    36
```

- ▶ **PFAM** : alignements et HMM pour 7973 familles de protéines (août 2005)



Taux de couverture Pfma-A: 75% (vérifiés)

Taux de couverture Pfam-B:19%

Application 2 : Régulation et Facteurs de transcription

génomé → ARN messagers → protéines

Application 2 : Régulation et Facteurs de transcription

génome → **ARN messagers** → **protéines**

génomique

transcriptome

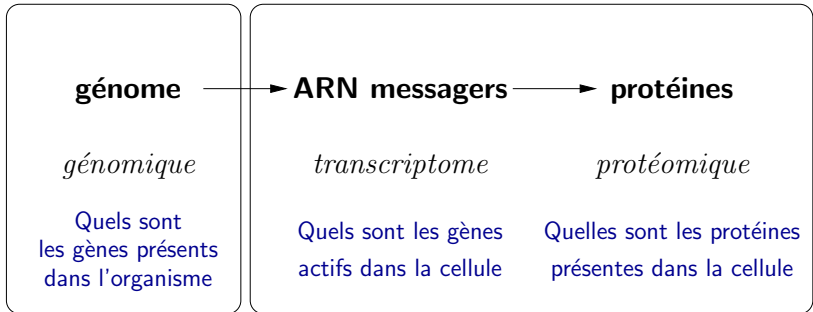
protéomique

Quels sont
les gènes présents
dans l'organisme

Quels sont les gènes
actifs dans la cellule

Quelles sont les protéines
présentes dans la cellule

Application 2 : Régulation et Facteurs de transcription



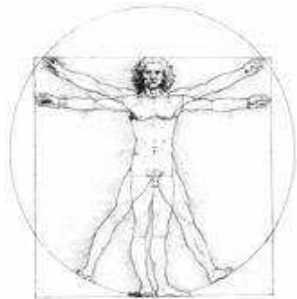
Information identique
dans toutes les
cellules

Information propre à chaque cellule
Expression différentielle du génome

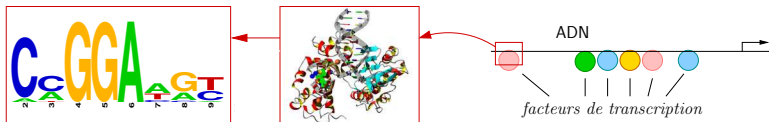
Expression différentielle du génome

L'activité d'un gène dépend

- ▶ du **temps** : étapes du développement
- ▶ du **type cellulaire** : peau, foie, rein, ...
- ▶ du **contexte**: cellule saine/cancéreuse, stimulus extérieur (glucose, stress, lumière, ...)



Comment les gènes sont-ils régulés ?



- ▶ **F**acteurs de transcription (FT): protéines régulatrices
- ▶ **P**remiers acteurs de la régulation transcriptionnelle
- ▶ **D**omaine de liaison à l'ADN : court motif nucléique conservé

Modélisation des sites de fixations des FT

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*

Matrices de comptage

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

Matrices de fréquences corrigées

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



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

- 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 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



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

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

- ▶ **Poids positif** : les bases plus fréquentes que la moyenne
- ▶ **Poids négatif** : les bases moins fréquentes que la moyenne

Recherche de motifs

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

Recherche de motifs

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

Recherche de motifs

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 PWM

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

Recherche de motifs

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 PWM
- 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

Recherche de motifs

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**

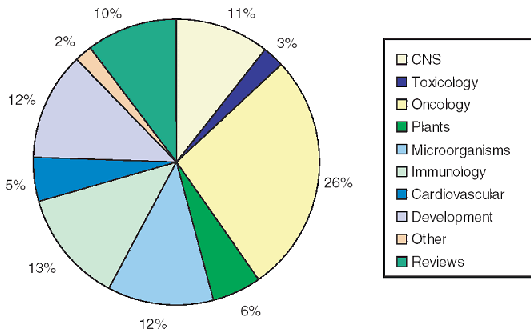
T	A	C	G	G	A	T	A	C	G	T	T	G	A	C	C	A	T	G	G	T	A	C	C	T	
T	A	C	G	G	A	T	A	C	G																
	A	C	G	G	A	T	A	C	G	T															

score : 6.16

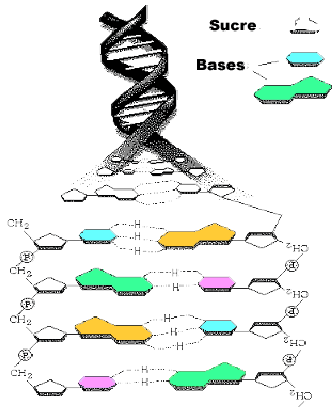
score : -1.86

Transcriptome : la technologie des puces à ADN

- ▶ Permet d'observer l'activité du génome dans une population de cellules, en mesurant le niveau d'expression de chaque gène
mesure la quantité d'ARN messagers
- ▶ Plus de 20 000 publications depuis 1985



Comment attraper les ARN messagers ?



hybridation :

deux brins complémentaires
d'ADN s'apparient suivant la
complémentarité des bases

$A \leftrightarrow T$

$C \leftrightarrow G$

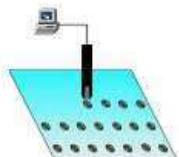
Fabrication des puces à ADN

Lames de verre recouvertes de polylysine



+

6116 ORFs de levure amplifiées par PCR



Spotting (dépôt)

Hybridation

Souche 1



Souche 2



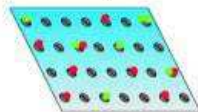
Extraction des ARN

Cy3



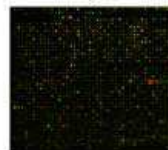
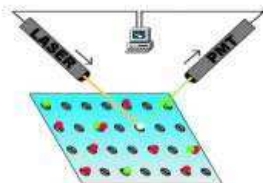
Transcription des ARNm en ADNc

Cy5



Obtention des résultats

Lecture (scanner)

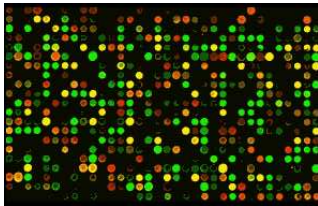


Analyses des résultats

Obtention des résultats

- ▶ **Digitalisation de la puce :**
 - ▶ excitation des éléments fluorescents
 - ▶ lecture
 - ▶ coloration de l'image

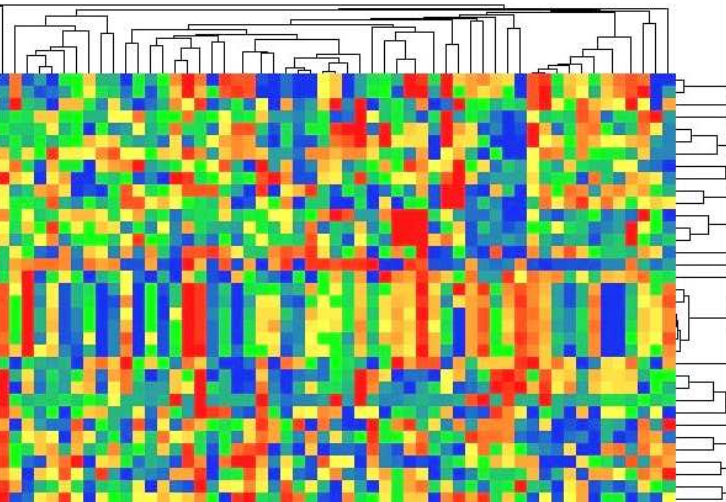
vert : Cy3 > Cy
rouge : Cy5 > Cy3
jaune : Cy5 = Cy3



- ▶ **Analyse de l'image :** matrices d'expression
- ▶ **Fouilles de données, clustering, analyse statistique**

Quels sont les gènes qui s'expriment dans les mêmes conditions?
Quels sont les échantillons similaires?

LE:SK
 CO:COLO205
 CO:HCC-2998
 CO:HCT-15
 CO:HCT-116
 CO:SW-620
 CO:KM12
 CO:HT29
 BR:T-47D
 BR:MCF7
 BR:LOXIM1
 PR:PC-3
 OV:OVCAR-5
 LC:NCI-H322M
 LC:EKVX
 LC:A549/ATCC
 LC:NCI-H460
 OV:SK-OV-3
 OV:TGROV1
 OV:OVCAR-8
 BR:MCF7/ADF-RES
 PR:DU-145
 LC:HOP-62
 RE:CAKI-1
 RE:LUO-31
 RE:ACHN
 RE:786-0
 RE:RXF-393
 RE:TK-10
 RE:A498
 BR:BT-549
 CNS:SNB-75
 CNS:SF-295
 CNS:U251
 CNS:SNB-19
 CNS:SF-539
 OV:OVCAR-4
 OV:OVCAR-3
 BR:H5578T
 CNS:SF-268
 LC:NCI-H226
 BR:MDA-MB-231/A
 LC:HOP-92
 ME:UACC-257
 ME:MALME-3M
 ME:SK-MEL-28
 ME:M14
 ME:SK-MEL-2
 ME:UACC-62
 BR:MDA-N
 BR:MDA-MB-435
 ME:SK-MEL-5
 LC:NCI-H522
 LC:NCI-H23
 RE:SN12C



1-est Homo sapiens e
 2-est ESTs Chr.6 [236
 3-est SID W 486793, I
 4-est SID W 172857, I
 5-est SID 484954, Ch
 6-est SID 301448, ES
 7-est SID 286200, ES
 8-est SID 37330, EST
 9-est SID W 121145, I
 10-est ESTs Chr.10 [4
 11-est SID W 278644
 12-est ESTs Chr.6 [25
 13-est Human fetus b
 14-est Human fetus b
 15-est SID W 131843
 16-protein ? glutamyl
 17-est RPL5 Ribosom
 18-est SID W 487188
 19-est SID W 364351
 20-est SID W 471123
 21-est *Human ferritin
 22-est SID 512268, Hi
 23-est ESTs Chr.22 [4
 24-est SID 485282, T
 25-est Human pre-B c
 26-est *EST A.688811
 27-protein Glutathione
 28-est SID 427845, Hi
 29-est SID 128329, ES
 30-est H factor (comp
 31-est SID W 428225
 32-est ESTs Chr.17 [4
 33-est SID 72214, Hui
 34-est ESTs Chr.1 [42
 35-est ESTs Chr.1 [24

Que faire des familles de gènes co-régulés ?

Comprendre les mécanismes de régulation sous-jacents.

- ▶ **Données** : Ensemble de gènes potentiellement co-régulés
Clusters construits à partir de données d'expression, annotation fonctionnelle, . . .
- ▶ **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 : % trinucleotides : AAA, AAC, TCA, etc.

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

Recherche d'oligonucléotides sur-représentés

▶ Oligonucléotide = 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

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
séquence	2 brins	n.obs.	n.att.	P-value

n.obs. : nombre d'occurrences observé

n.att. : nombre d'occurrences attendu

P-value : probabilité du nombre d'occurrences observé

motif 1

```
tcacgt..  ..acgtga
.cacgtg.  .cacgtg
..acgtga  tcacgt..
```

```
tcacgtga  tcacgtga
```

motif 2

```
aactgt..  ..acagtt
.actgtg.  .cacagt.
..ctgtgg  ccacag..
```

```
aactgtgg  ccacagtt
```

complexe *Met4p/Cbfl/Met28*

Met31p

Comment recenser tous les mots d'une longueur donnée dans un ensemble de séquences

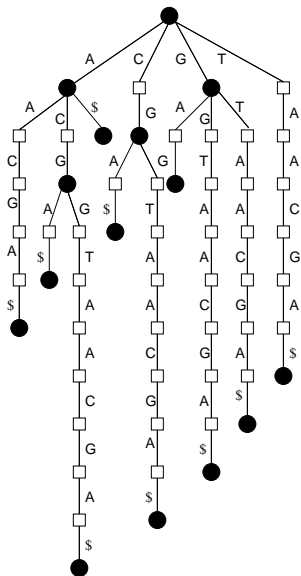


Comment recenser tous les mots de toutes les longueurs

Arbre des suffixes

- ▶ Structure de données qui permet de représenter exactement tous les facteurs d'un texte
- ▶ Une feuille de l'arbre = un suffixe du mot
 - un facteur = un préfixe d'un suffixe
 - un chemin dans l'arbre = un facteur
- ▶ Taille linéaire
 - Chaque nœud interne a au moins deux fils

Exemple : ACGGTAACGA



Deuxième tentative : ajout d'un nouveau caractère \$

une feuille = un suffixe

Algorithmes de construction

Temps et espace linéaire

- ▶ **Weiner** (1973)
Algorithme historique
- ▶ **Mac Creight** (1976)
Construction à l'aide de liens suffixes
- ▶ **Ukkonen** (1995)
Algorithme on-line

Arbre des suffixes généralisés

- ▶ Un ensemble s_1, \dots, s_n de séquences
- ▶ **Arbre des suffixes généralisé**
 - un chemin de la racine à une feuille = un suffixe d'une séquence
- ▶ Construction incrémentale
 - ▶ Arbre des suffixes pour s_1
 - ▶ Ajout des autres séquences s_2, s_3, \dots

Espace et temps linéaire

Recherche de motifs communs

- ▶ Comment déterminer le nombre d'occurrences d'un mot dans un ensemble de séquences?
- ▶ Comment déterminer l'ensemble des mots ayant au moins q occurrences ?



Recherches de motifs communs approchés

Données

- ▶ S : ensemble de séquences
- ▶ e : nombre maximal d'erreurs (substitutions)
- ▶ q : nombre de séquences où le motif doit apparaître (quorum)
- ▶ k : longueur maximale d'un motif

Problème

Trouver tous les mots m de longueur inférieure à k tel qu'il existe au moins q séquences de S contenant un mot présentant au plus e erreurs par rapport à m

Deux arbres

- ▶ **Arbre des motifs** : arbre complet de tous les mots de taille $\leq k$
- ▶ **Arbre du texte** : arbre des suffixes des séquences jusqu'à la profondeur k

Parcours des deux arbres

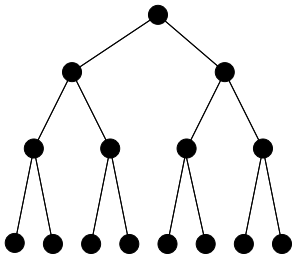
- ▶ **Parcours en profondeur** de l'arbre des motifs : on énumère les motifs
- ▶ **Indication** sur l'arbre du texte du nombre d'erreurs pour le motif en cours pour tous les nœuds de même profondeur
- ▶ **Elagage** de la recherche : quand on a moins de q séquences avec $\leq e$ erreurs

Exemple

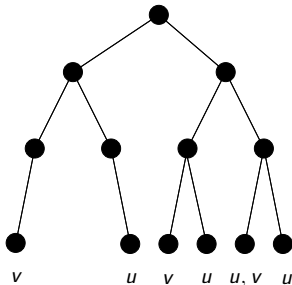
- ▶ **Alphabet** $\{a, t\}$ (t à gauche, a à droite)
- ▶ **Mots** $u = aaaataa$ et $v = aattttt$
- ▶ **Motifs** de longueur ≤ 3 présents dans les deux séquences avec au plus une erreur

Exemple

- ▶ Alphabet $\{a, t\}$ (t à gauche, a à droite)
- ▶ Mots $u = aaaataa$ et $v = aattttt$
- ▶ Motifs de longueur ≤ 3 présents dans les deux séquences avec au plus une erreur



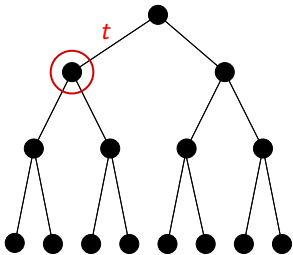
arbre des motifs



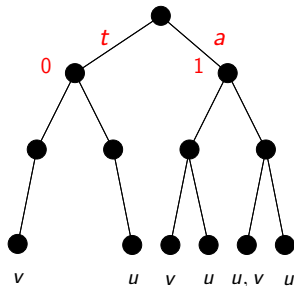
arbre du texte

Exemple

- ▶ Alphabet $\{a, t\}$ (t à gauche, a à droite)
- ▶ Mots $u = aaaataa$ et $v = aattttt$
- ▶ Motifs de longueur ≤ 3 présents dans les deux séquences avec au plus une erreur



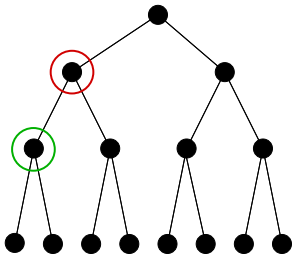
arbre des motifs



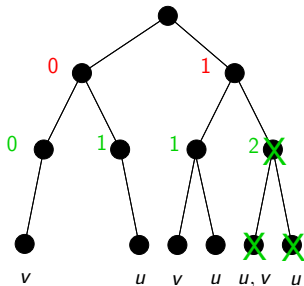
arbre du texte

Exemple

- ▶ Alphabet $\{a, t\}$ (t à gauche, a à droite)
- ▶ Mots $u = aaaataa$ et $v = aattttt$
- ▶ Motifs de longueur ≤ 3 présents dans les deux séquences avec au plus une erreur



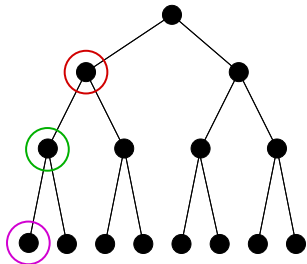
arbre des motifs



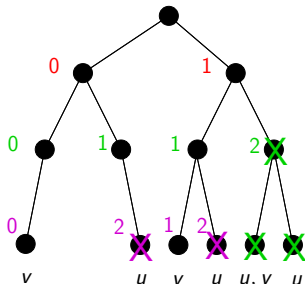
arbre du texte

Exemple

- ▶ Alphabet $\{a, t\}$ (t à gauche, a à droite)
- ▶ Mots $u = aaaataa$ et $v = aattttt$
- ▶ Motifs de longueur ≤ 3 présents dans les deux séquences avec au plus une erreur



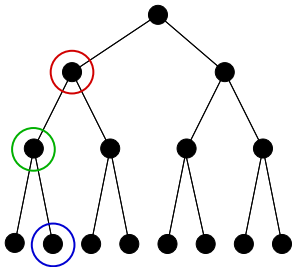
arbre des motifs



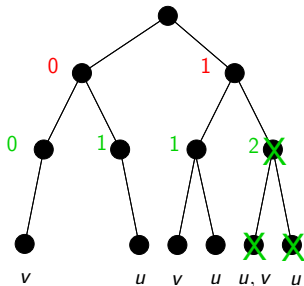
arbre du texte

Exemple

- ▶ Alphabet $\{a, t\}$ (t à gauche, a à droite)
- ▶ Mots $u = aaaataa$ et $v = aattttt$
- ▶ Motifs de longueur ≤ 3 présents dans les deux séquences avec au plus une erreur



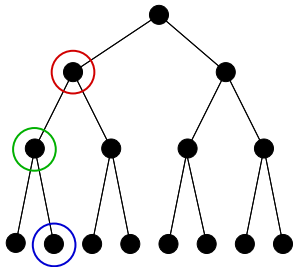
arbre des motifs



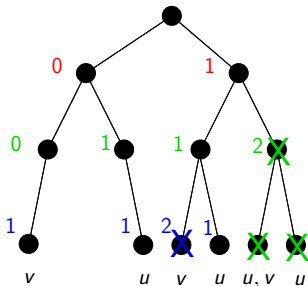
arbre du texte

Exemple

- ▶ Alphabet $\{a, t\}$ (t à gauche, a à droite)
- ▶ Mots $u = aaaataa$ et $v = aattttt$
- ▶ Motifs de longueur ≤ 3 présents dans les deux séquences avec au plus une erreur



arbre des motifs



arbre du texte