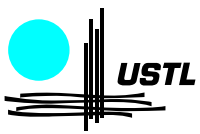
Localisation à grande échelle de motifs nucléiques approchés

décrits par des matrices position-poids

Aude Liefooghe, Hélène Touzet et Jean-Stéphane Varré

Equipe Bioinfo — LIFL UMR CNRS 8022 — USTL

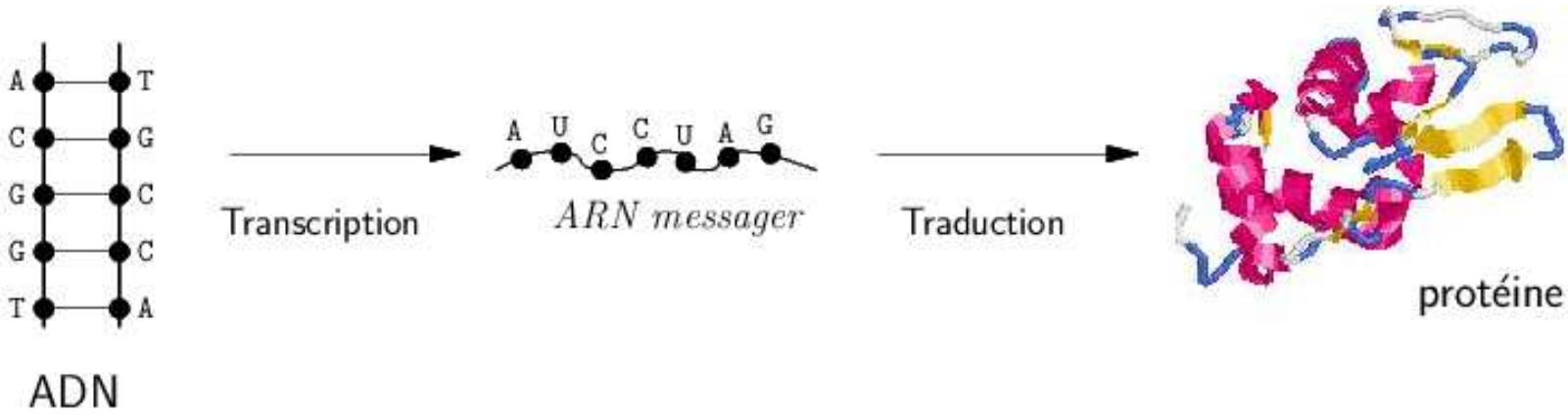
25 novembre 2005



Plan

1. Contexte
   - Sites de fixation des facteurs de transcription
   - Matrices positions-poids
   - Recherche de motifs
2. Accélération 1 : Algorithme exact de localisation de matrices
   - Localisation multiple de matrices
   - Structure d'index
   - Algorithme proposé
3. Accélération 2 : Similitudes entre matrices
   - Analyse générale
   - Classification des matrices
   - Application au problème de localisation

Dogme central



- Niveaux de régulation
- Initialisation de le transcription

Facteurs de transcription

- Premiers acteurs de la régulation transcriptionnelle
- Domaine de liaison à l'ADN : court motif nucléique conservé
- Comment localiser ces sites potentiels ?

Modélisation des sites de fixation des facteurs de transcription par des matrices positions-poids

- Afin de détecter les sites de fixation des facteurs de transcription sur une séquence d'ADN, il faut un modèle
- Les sites sont des motifs courts et variables
- La souplesse des motifs approchés est donc adaptée :
  - Motifs avec erreurs
  - Séquences consensus
  - Expressions régulières
  - Les matrices positions-poids, une description quantitative de l'alignement multiple [Stormo 1998]

Alignement multiple

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*

Alignement multiple

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 C

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

Colonne **i** acide nucléique
Ligne **j** position de l'alignement

Système de score

Obtenir un système de scores

- A valeurs positives pour les acides nucléiques conservés
- A valeurs négatives pour les acides nucléiques non conservés
- Système additifs entre les colonnes
- Sans sur-adaptation

Matrices de fréquences corrigées

Matrice de Comptage C

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

▷

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

i Acide nucléique
j Position de l'alignement
f Fréquence génomique
pp Pseudo-poids

Matrice de Fréquences corrigées F

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 position-poids

Matrice de Fréquences corrigées F

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

▷

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

i Acide nucléique
j Position de l'alignement
f Fréquence génomique

Matrice de Poids P

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

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

Banque de données de matrices

- Transfac `http://www.gene-regulation.com/`
  eucaryotes, 5327 facteurs, 674 matrices, construites à partir de 13112 séquences
- Jaspar `http://forkhead.cgb.ki.se/JASPAR/`
  eucaryotes, 63 matrices *haute qualité*
- PlantCare : `http://sphinx.rug.ac.be:8080/PlantCARE/`
- Levure : SCPD (`http://cgsigma.cshl.org/jian/`),
  YRSA (`http://forkhead2.cgb.ki.se/yrsa/`)
- etc.

Les limites de l'approche traditionnelle de la localisation

Logiciels existants : Patser, Matinspector, etc.

Leur utilisation

- Une seule matrice
- Séquence de taille réduite

Recherche exhaustive des matrices des vertébrés de Transfac sur le génome humain avec Patser $>$ 24h sur un PC de bureau

La nouvelle approche de la localisation

Nouveaux besoins

- Annotation de génome (génomique comparative, gènes co-régulés)
- Passage à l'échelle (recherche dans un génome entier)
- Raisonnement sur des familles de facteurs

Pas d'algorithme efficace de localisation pour ce type de données

Le problème formel

Données de départ

- Longue séquence
- Ensemble de matrices
- Seuil de score (ou p-valeur correspondant à un seuil de score) associé à chaque matrice

Résultat

- Hits des matrices (occurrences de la séquence de score $>$ seuil)

Création d'une structure d'index

Comment accélerer la recherche de motifs ?

- Banques de matrices, données stables
- Prétraiter les matrices
- Factoriser les matrices

Créer une structure d'index de scores précalculés en amont de l'algorithme.

un mot ▷ structure d'index ▷ les scores du mot pour toutes les matrices

Idée de départ : la structure d'index

Structure d'index de scores précalculés.

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

0	...
1	...
...	...
...	...
...	...
...	...
...	...
...	...
813 254	6,16
...	...
...	...
$4^{10} - 1$	...

TACGGATACG score : 6.16
clé : 813 254

Mise en pratique : la structure de sous-index

Structure de sous-index de sous-scores précalculés

On tire partie de l'additivité des scores

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

A	C	G	T
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

0	...
...	...
794	3.14
...	...
$4^5 - 1$	...

0	...
...	...
198	3.02
...	...
$4^5 - 1$	...

TACGG sous-score : 3.14
clé : 794
ATACG sous-score : 3.02
clé : 198
score : 6.16

Elagage dans le parcours de la structure d'index

Pour calculer le score du mot CTGACCATGC, le parcours de tous les sous-index n'est pas nécessaire

sous-index 1

CTGAC

Sous-score : -10.13

sous-index 2

NNNNN

Sous-score maximum : 6.02

Pour un score seuil de 1, quelque soit le deuxième sous-mot, le mot qui commence par CTGAC ne sera pas un hit.

But : anticiper le résultat final à partir d'un résultat partiel grâce aux scores maximaux associés à chaque matrice pour chaque sous-index.

Choix de la structure d'index optimale

Données à fixer

- Nombre de sous-index
- Tailles des sous-index

Objectif

- Minimiser le coût de calcul moyen

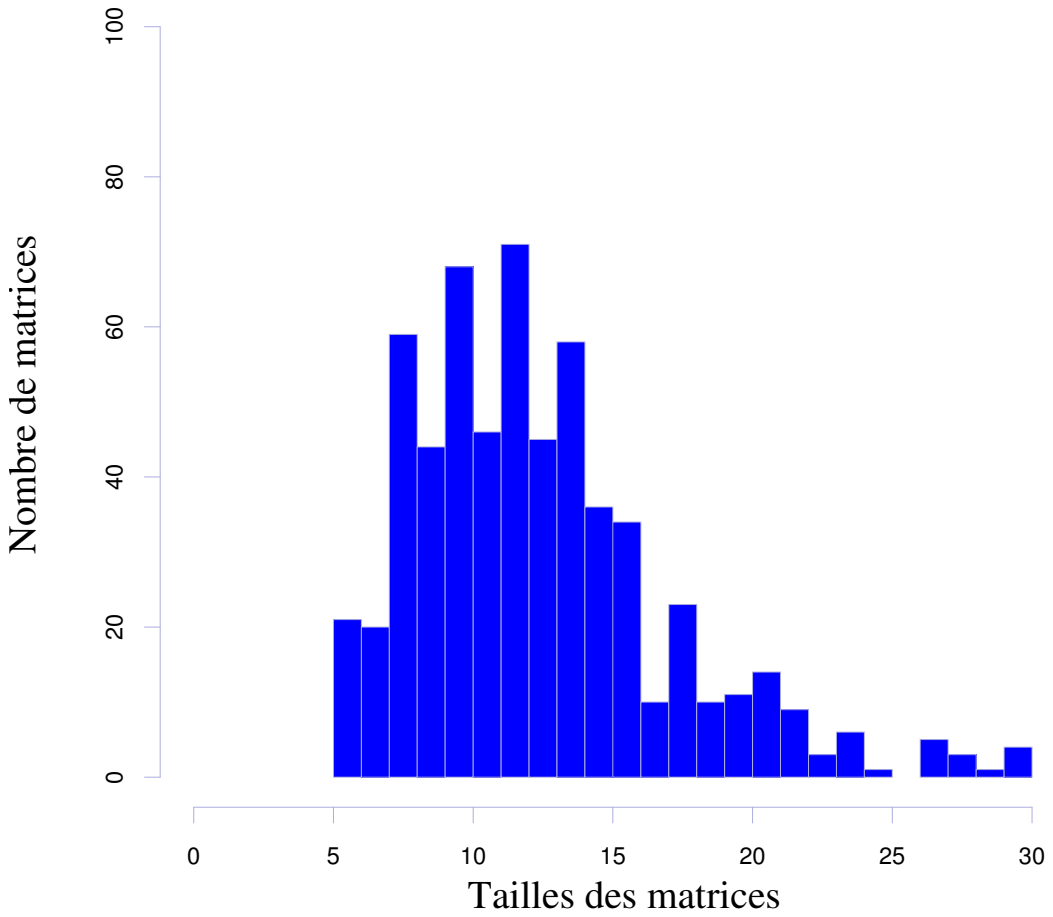
Données de départ

- Répartition de la taille des matrices
- Prédiction de l'élagage dans le parcours de la structure d'index

Contrainte et hypothèse de travail

- Espace mémoire dédié à la structure d'index limité
- La séquence suit un modèle de Bernoulli

Répartition des la taille des matrices Transfac et Jaspar



Nombre moyen d'accès par sous-index

Comment calculer a_{ij} le nombre moyen d'accès au sous-index$[i...j]$?

- M une matrice, α son score seuil, t sa taille
- $p(M, i)$: probabilité que le score de M de la colonne 1 à i soit inférieur à $\alpha - \sum_{k=i+1}^{t} \max_{l=\{a,c,g,t\}} M(k, l)$
- $a_{ij} = \sum_M P(M, j) - P(M, i)$

Le problème d'optimisation

- a_{ij}, nombre moyen d'accès au sous-index$[i...j]$
- h_i, nombre de matrices de longueur i
- $\phi(n, j, e) =$ nombre d'opérations pour une structure optimale
  - constituée de n sous-index,
  - occupant un espace mémoie d'au plus e,
  - traitant les matrices entre les colonnes 1 et j.

$$\phi(n, j, M) \;=\; \inf_{i} \left\{ \phi(n-1, i, e - 4^{j-i} \sum_{k=i+1}^{j} h_k) + a_{ij} \right\}$$

- ϕ est calculé par programmation dynamique
- Choix du n optimal

L'algorithme proposé

En amont

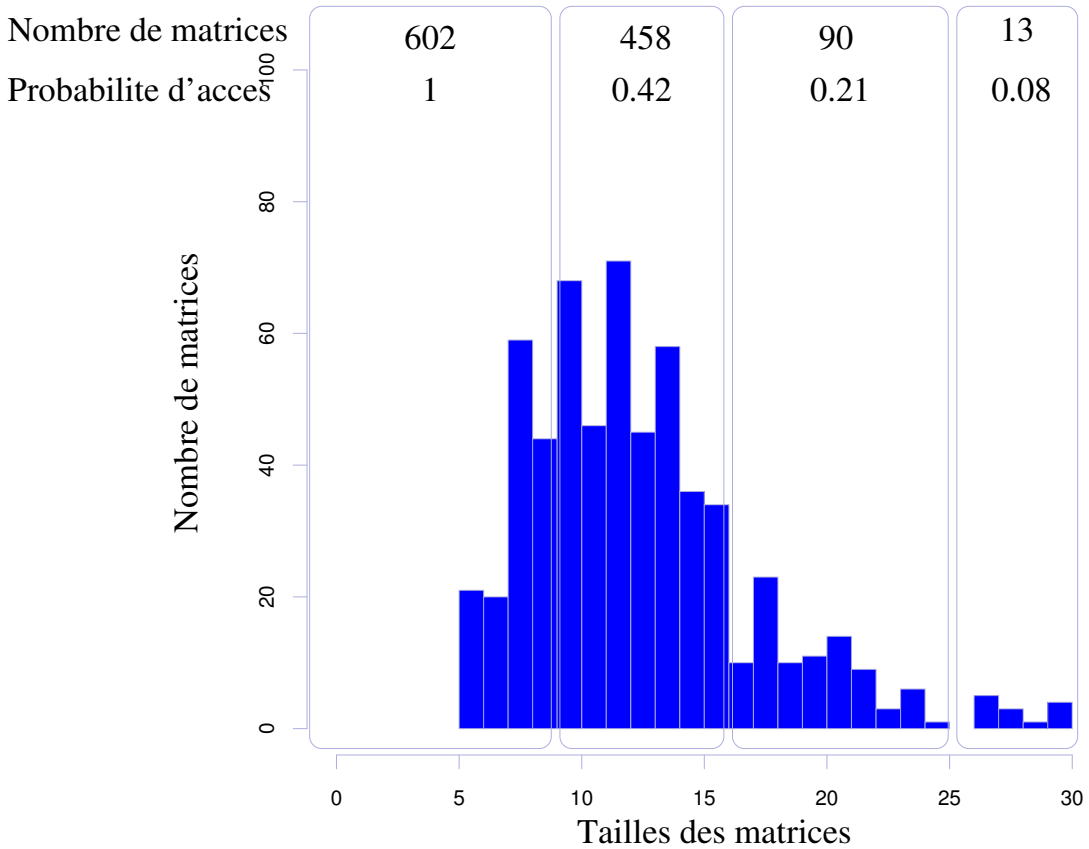
- Création d'une structure de sous-index
  - Calcul du nombre et de la taille des sous-index optimaux
  - Stokage des sous-scores dans la structure d'index

Algorithme

- Découpage du mot d'entrée
- Accès aux sous-scores nécèssaires
- Décision sur le mot : hit ou non.

Une implémentation de l'algorithme proposé

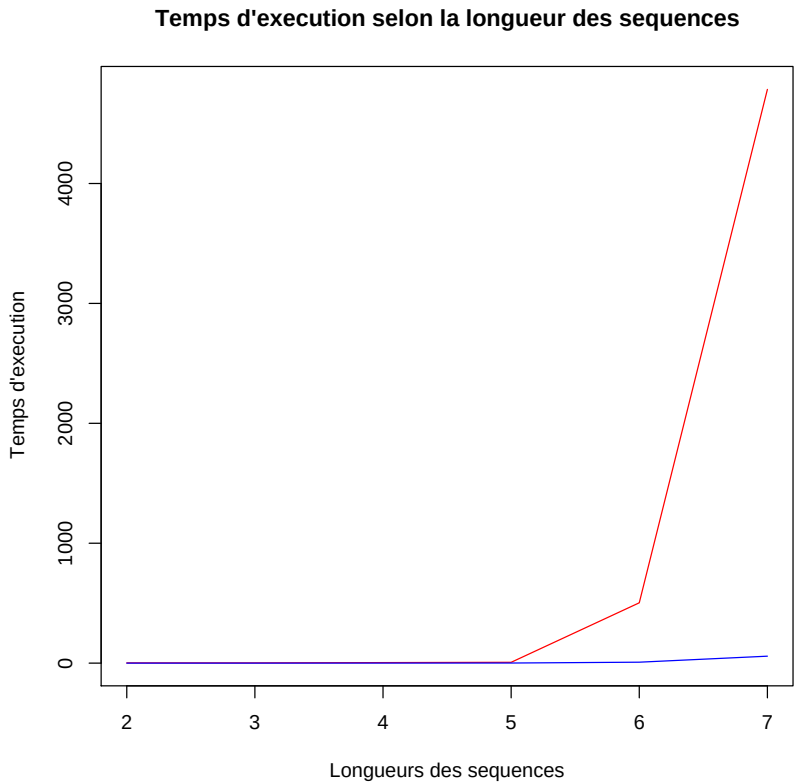
Découpage optimal pour 512 matrices Transfac 8 et 63 matrices Jaspar



4 sous-index de tailles 9, 7, 9, et 5

Une implémentation de l'algorithme proposé

Performances de l'algorithme 10^6



Patser, TFMscan

Observation des redondances entre matrices

GATA

GATA-6

GATA-3

ZN-FINGER, GATA

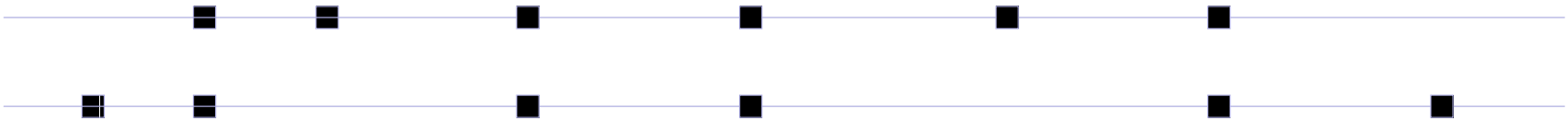
Il existe des similitudes entre les matrices

- Hommogène entre facteur
- Banques de données redondantes

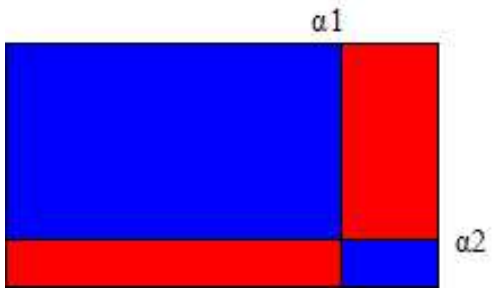
Les matrices se regroupent naturellement en familles

Quantifier la similitude entre les matrices

Calcul des positions communes



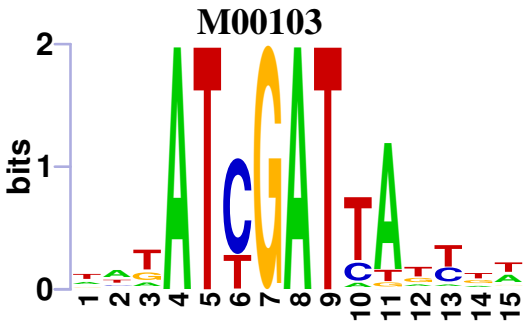
- M_1 et M_2 deux matrices, α_1 et α_2 leur score seuil
- Alignement des matrices (maximiser les hits communs)
- Calcul de la probabilité des positions communes et non communes

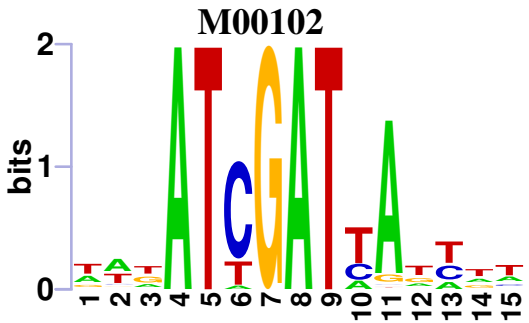


- Modèle de Bernoulli
- Programmation dynamique
- Extention du calcul de la p-valeur en dimension 2.

Exemple de calcul de positions communes entre 2 matrices

Matrices M00102 et M00103 de la base de matrices Transfac





Estimation de la proportion de hits communs entre M00102 et M00103 pour une p-valeur de e^{-5}

	Positions positives	**Positions négatives**
Positions communes	85%	99%
Positions non communes	15%	1%

Classification des matrices

Distance 2 à 2 entre matrices

- Proportion de positions non communes

Algorithme de classification

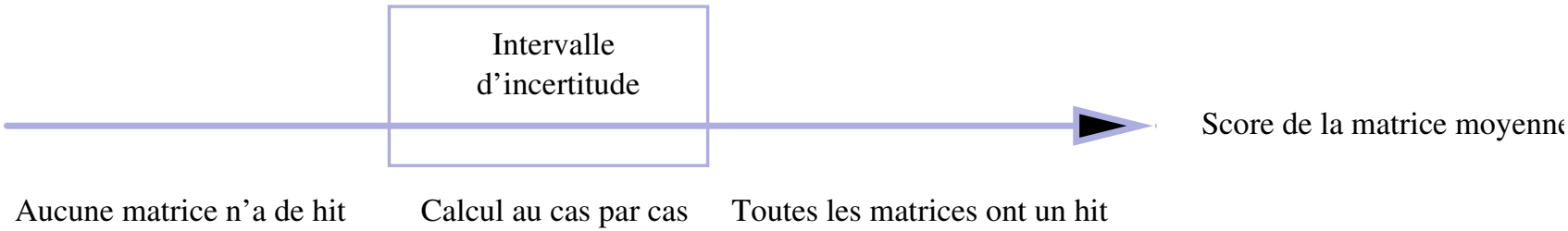
- Hiérarchique ascendant

Matrice représentante d'une classe

- Moyenne entre les matrices de la classe qu'elle représente

Algorithme exact

Première approche de l'exploitation des matrices pour le problème de la localisation des matrices



- Intervalle d'incertitude pour chaque matrice moyenne
- Critère de classification : sélection selon la probabilité de l'intervalle d'incertitude
- Méthode efficace pour un groupe de matrices très homogène

Algorithme approché

Deuxième approche de l'exploitation des matrices pour le problème de la localisation des matrices

Recherche des hits sur les matrices moyennes uniquement, chaque matrice moyenne estime les hits des matrices de sa classe

Critère de classification : sélection des classes selon le taux de positions non communes

Exemple

Classification des matrices de Transfac et Jaspar pour une proportion de hits communs d'au moins 80%

- Nombre de matrices initiales : 602
- Nombre de matrices moyennes : 473

La plupart des positions non communes sont des sites au score tangent au seuil.

Conclusion

Première proposition d'algorithme rapide pour la localisation de matrices

Travaux en cours et perspectives

- Evolution de la composition en acide nucléiques, modèle de fond local
- Méthode approchée qui prend en compte exclusivement le coeur conservé des matrices
- Accélération pour la recherche d'une seule matrice