Groupe Miners

LIMOS

12 FÉVRIER 2015

Stockage, indexation et comparaison d'une grande quantité de données génomiques à l'aide d'algorithmes de traitement d'images dans un environnement d'exécution NoSQL et GPU
(Point d'avancement)

Jocelyn DE GOËR

Directeurs de Thèse :
Myoung-Ah KANG,
Engelbert MEPHU NGUIFO

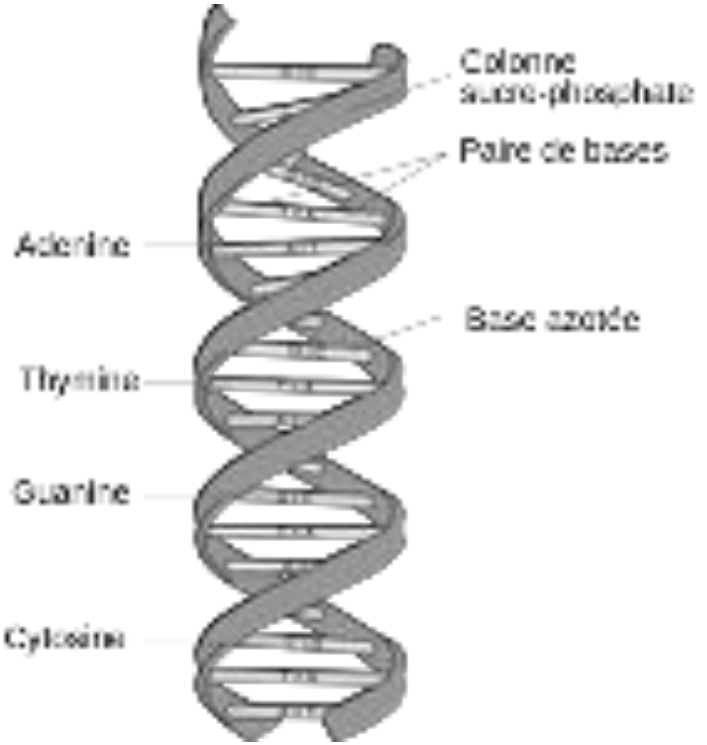
UMR 6158 CNRS

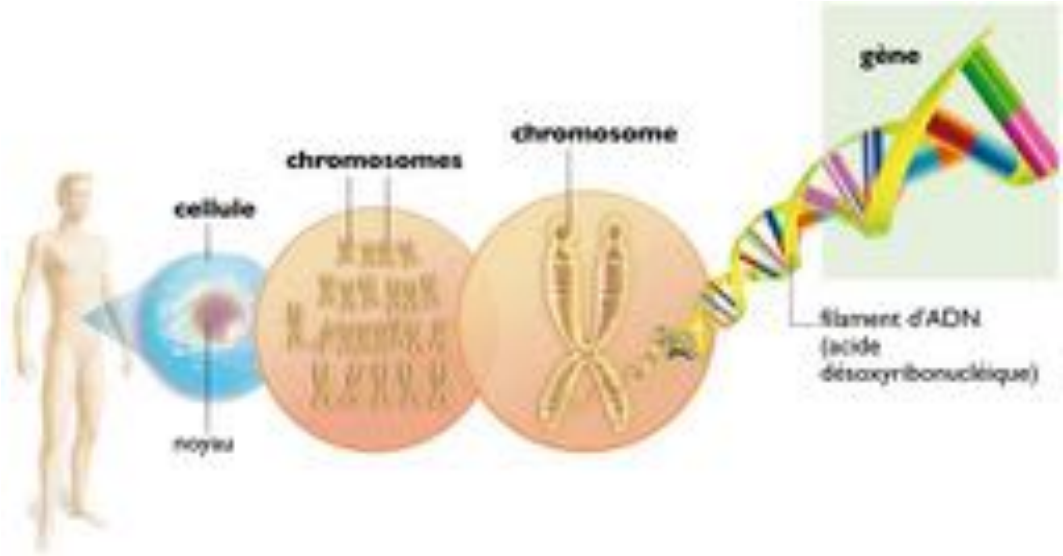
Sommaire

- Contexte biologique
- Indexation et comparaison de séquences d'ADN par hachage perceptuel
- Alignement de séquences d'ADN par corrélation de TCD-CS

- **Le séquençage de l’ADN**

Le séquençage de l'ADN consiste à déterminer l'ordre d'enchaînement des nucléotides pour un fragment d’ADN donné.

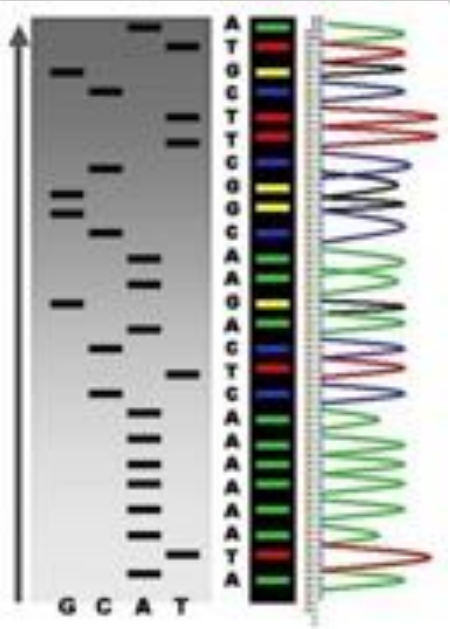




TTTAAAAAACTGTTTGTATAATA
TAATATTATTATATATAATATTA
AGCAACTACTATGATACTAATGA
GTATAGTGCTATTTTATTAATAT
GTAGCGTTAATTTATTTTGTTTT
CAAAATAAATTAACTACTTCTCG
TGGGAATTCCCTAAAGAAGATTA
AATTAAAAAAAAAATAAAAATAG

Un évolution des techniques de séquençage :

- 1977 : Première méthode de séquençage (Pr. Frederick Sanger) – virus phiX174 (5386 nucléotides)
- 2000-2005 : Arrivée des séquenceurs de 2ème génération
  - Aujourd'hui : 80-100 Gpb / j
- 3ème génération (2015-2020) :
  - Débit de plus en plus élevé 500 Gpb / h
  - Arrivée séquenceurs modulables et portatifs

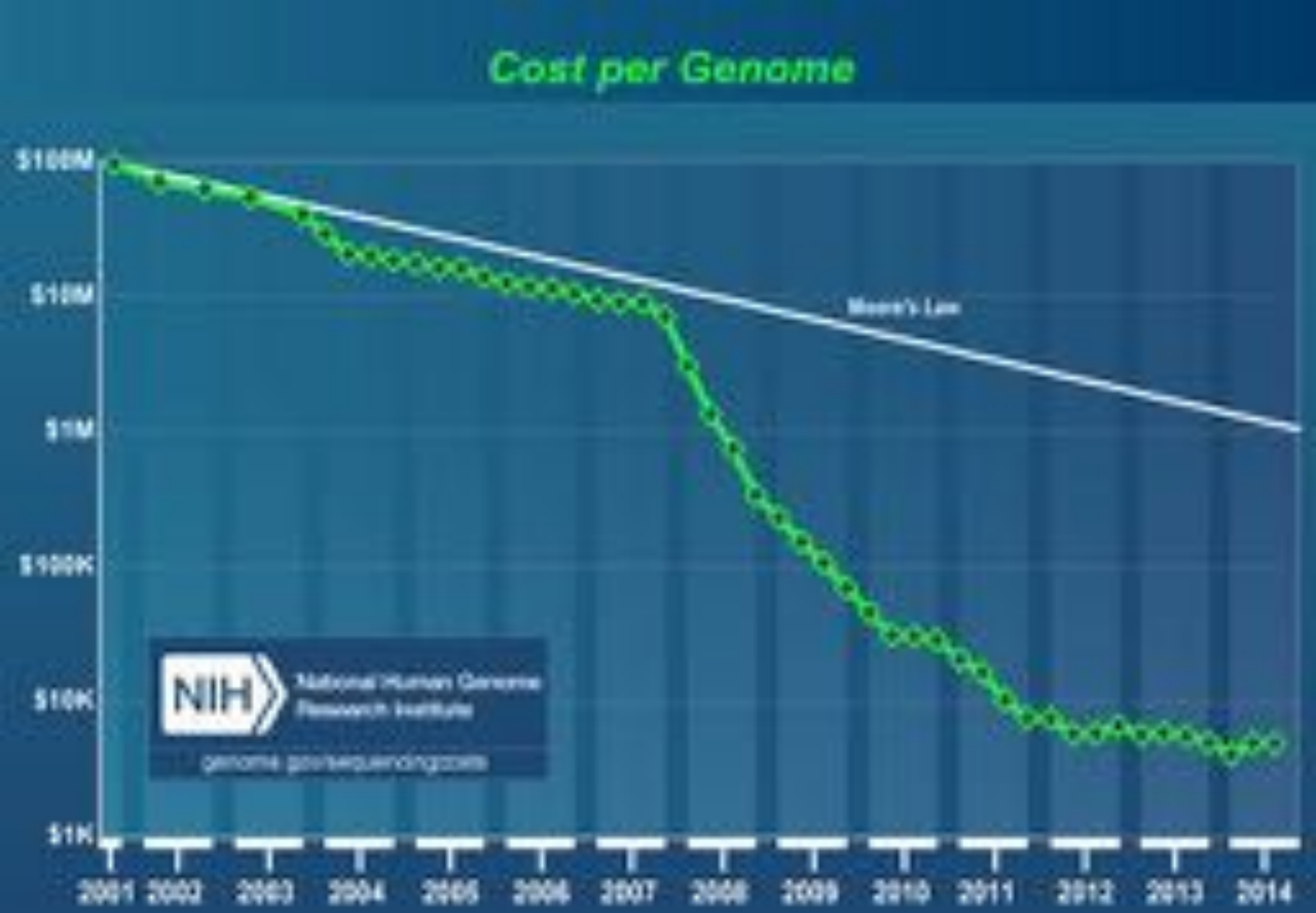


Lecture sur un gel de polyacrylamide


Mini séquenceur portatif


Séquenceur GridIon


Séquenceur Roche 454


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome
Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014

- Problématiques :
  - Identifier le génome auquel appartient une séquence non caractérisée
  - Construire des « séquences consensus » à partir de plusieurs
  - Stockage les données issues du séquençage

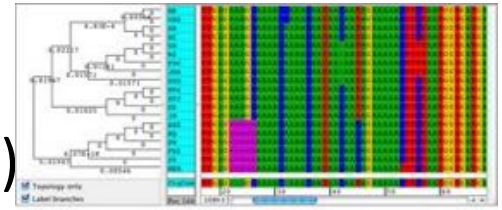
Génome :	Taille :
Grippe	0,013 Mpb
Escherichia coli	4,64 Mpb
Humain	3,2 Gpb
Blé	17 Gpb
Paris Japonica	150 Gpb
Polychaos Dubium	675 Gpb

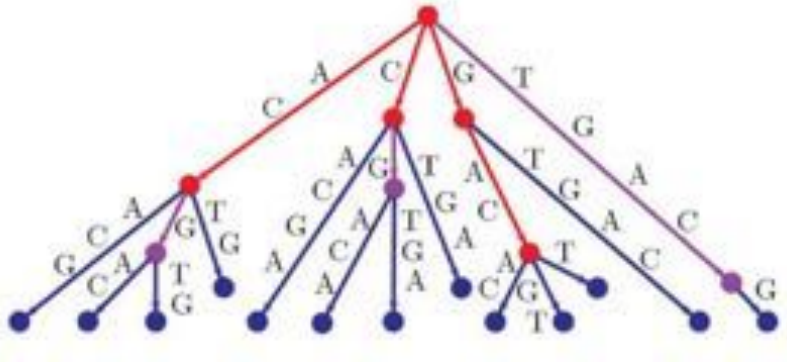
Supports de stockage :	Capacité :
Disquette 3,5’	1,4 Mo
Compact Disk (CD)	650 / 700 Mo
Digital Versatil Disk (DVD)	4,7 Go
Disque dur :	500 Go à 4 To

1 nucléotide = 1 octet

- Meta-génomique
  - Identifier la diversité microbienne d’un échantillon
    - Étude du macrobiote intestinal
    - Diversité microbienne d’un échantillon d’eau de mer

- Techniques de comparaison de textes
  - Alignement de séquences
    - Alignement global (Needleman-Wunsch)
    - Alignement local (BLAST)
  - Indexation
    - Table des « k-mers »
    - Arbre des suffixes





- Algorithmes de moins en moins adaptés au traitement d'une masse importante de données
  - Complexité algorithmique
  - Difficulté de parallélisation
  - Besoins en mémoire vive de plus en plus importants

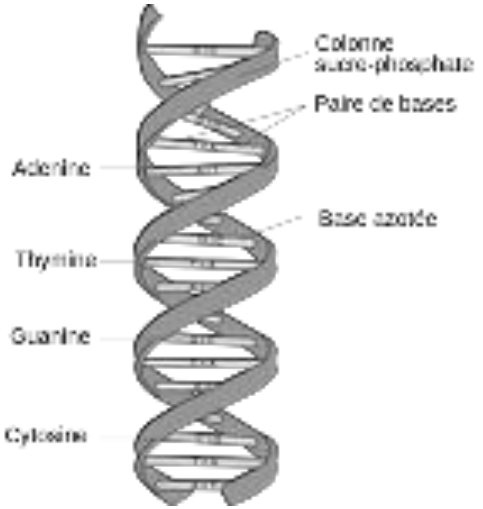
- Sous-ensemble du traitement du signal
  - Premiers travaux à partir de la fin des années 60
  - Applications dans de nombreux domaines :
    - Photo, vidéo, physique, astronomie, géologie, médecine…
  - Aujourd'hui :
    - Reconnaissance de personnes ou d'objets en temps réel
    - Application de filtre (amélioration de l'image, flou, effet artistiques…)
    - Compression d'image ou de vidéo (JPEG, MPEG)
- Une masse de données importante
  - Photo de 10 mégapixels = 3648x2736
  - Vidéo Full HD : 1920x1080 (2 mégapixels) avec 30 images /s

Indexation et comparaison de séquences d'ADN par hachage perceptuel

- **Le hachage perceptuel**
  - Objectif : Pouvoir identifier et comparer rapidement des documents (image, son, vidéo)
  - Principe : calcul d'une emprunte unique (un hache) à partir d'un document
  - Particularité : il n'y a pas « d'effet avalanche », on peut donc calculer une distance entre deux haches
- Exemples :

- **Séquence ADN brute**

```
ATTTTATTGACAAAAATATTAGTTTTTGCTATCATGCATCTAATTTAATAAAGAGAAGTAAAAAGGTGTG
TGATTTAAGAAAAACAAAACTAATAGATAAAATAAGTTCACTAGAACTATACAAATACTCAATATTTTTT
AGAAATTACATTGAAAATGTAGCAGAAGATTGTCTCAAGAACGGACTTATTCTTGAGAGTGCTGCCCACA
ATGTTAGTGAGGTTGAACTTGCTAGGTTAAAGGTACAGCTTAAGAATGCTCTGCTTAATTGTATTATAAG
CTACCGTTTTCATGGGATTGGCTATGTTTTAGTAAAAACCAAAGATACCCTAATAGATCTCGAACAACCC
GTTAATATAGAATTACCTATTGGTTTTGAATACCTTGATTATGAATATGTAAGAGATTTGGGAGTTGATT
TTGATCATATAACCTATAAAGTAAAATCCAACAATAAGAACAATTCTTTAGACGCAGTTAAAATACATAA
AAGTCGACTTATCATATATGAAAACTTTGATTATATCTTAAAAAGATATGTTCCGTGTTATACCGAAAGC
TTTTTACTAGATATTTATTTATTTGAAAAGATATACGTAGAAATAGAAAGACGTATTGAAAACCACAATT
TTTTGTTTTACAAAGATGAATCTTTAGTACAACTACAAGACGCACTTTCTAGTGCAACAACTTCTTTAAG
TGCACTTACTCAGAGTAATAATGATAGGGGAAGTGGCATTTTATCTTCTTTTTTGAGAAAACAAAATTCA
AACAATCATAGTAAAGATATTTCTAATTTAAGAAACCTTAATGACTCATTATCACAGGAGCTTGCTAGGC
TAAAAAGTAATCTAAATAATGAGGGAATGTTTTATACGGCCACCCCTAGTGCTAGTTTAGAGGTTATTAA
ATACGACCTTAGTTACTTAAAGGAGGCTTTAGCATTAATTAAGGCAAAAATTGGTGCAGATACTAAAGAG
CCCTTAACCAGAAGTTTTAACGAACAGGCTAAAGGGCTAGGAAATGATGGTAAAGGGGATAGGAGTAATT
```



- **Conversion en matrice de pixels**

A = 63
T = 127
C = 191
G = 255

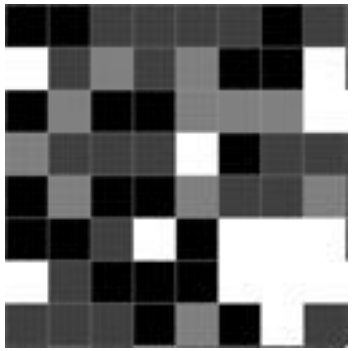


Image 8x8 pixels

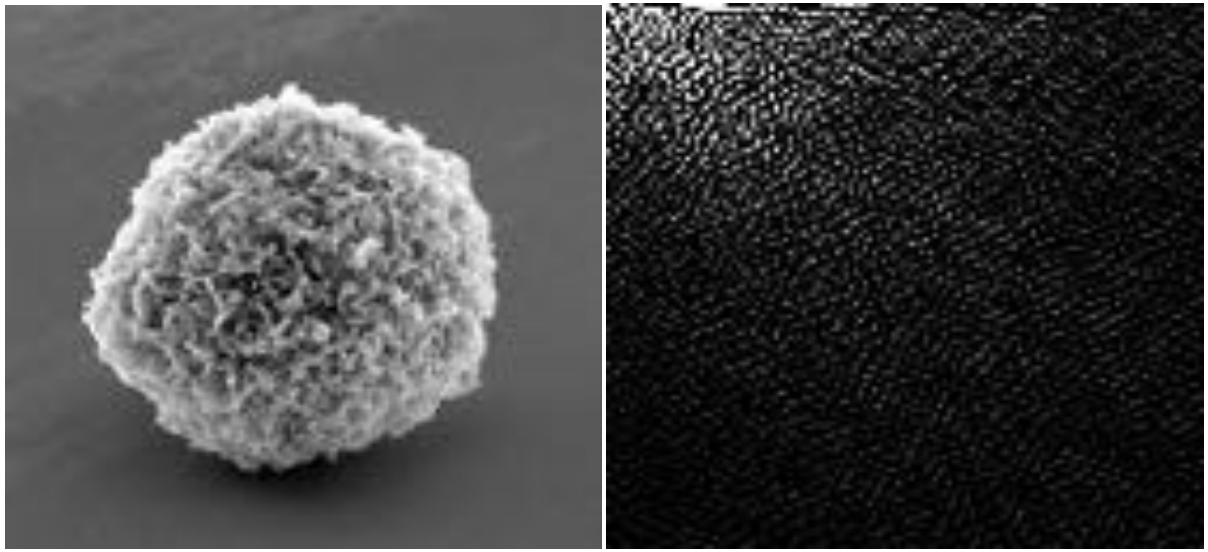
63	63	127	127	127	127	63	127
255	127	191	127	191	63	63	255
63	191	63	63	191	191	191	255
191	127	127	127	255	63	127	127
63	191	63	63	191	127	127	191
63	63	127	255	63	255	255	255
255	127	63	63	63	255	255	255
127	127	127	63	191	63	255	63

- Application d'une TCD
  - TCD (Transformée en Cosinus Discrète)

$$TCD(i,j) = \frac{1}{\sqrt{2}} C(i)C(j) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} pixel(x,y) \cos\left(\frac{(2x+1)i\pi}{2N}\right) \cos\left(\frac{(2y+1)j\pi}{2N}\right)$$

  - TCD : Étape majeure utilisée dans la compression d'images JPEG
  - Domaine spatial -> domaine fréquentiel : Changement de domaine d'étude tout en gardant la même fonction à étudier

- **Domaine fréquentiel**
  - Permet d'avoir une représentation fréquentielle de l'enchaînement des pixels
    - Une fréquence : changement d'intensité
    - Les hautes fréquences : structure de l'image
    - Les basses fréquences : zones homogènes, floues
  - Le passage dans le domaine fréquentiel génère une matrice des fréquences appelée matrice des coefficients

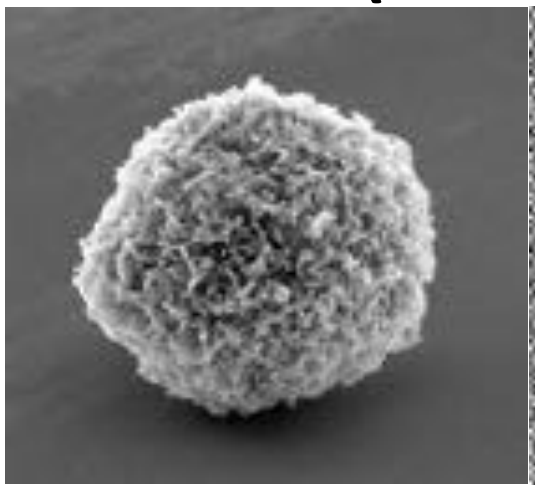


Originale Matrice des coefficients

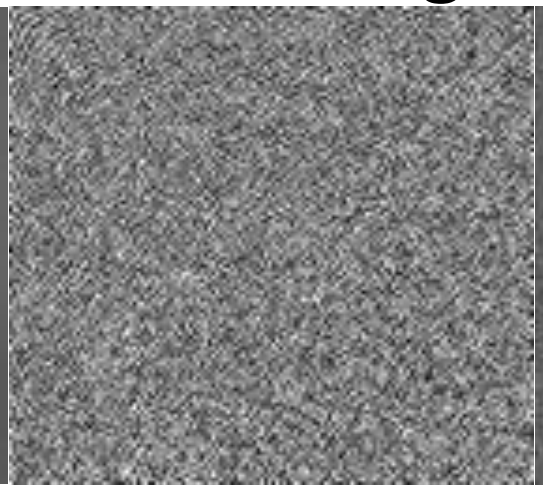
- **Matrice des coefficients**

-2.27214e-07	-1.36317e-11	6.66826e-15	1.77258e-10	1.07809e-11	2.85591e-05	8.20693e-10	-0.00139901	4.58842e-10	-1.29768e-07	-1.08057e-20	1.70999e-10	1.0444e-08	-3.04348e-07	-6.99148e-06	-3.33784e-06
4.58842e-10	-1.29768e-07	-1.08057e-20	1.70999e-10	1.0444e-08	-3.04348e-07	-6.99148e-06	-3.33784e-06	7.12087e-06	-2.70032e-08	2.24238e-07	1.42927e-16	2.9019e-09	-1.06657e-18	-1.2611e-14	-3.05368e-08
7.12087e-06	-2.70032e-08	2.24238e-07	1.42927e-16	2.9019e-09	-1.06657e-18	-1.2611e-14	-3.05368e-08	-3.19239e-16	-2.51033e-10	1.31586e-07	-8.79998e-11	9.7384e-07	9.39668e-08	-4.76745e-09	-4.6195e-07
-3.19239e-16	-2.51033e-10	1.31586e-07	-8.79998e-11	9.7384e-07	9.39668e-08	-4.76745e-09	-4.6195e-07	3.3957e-12	-3.87004e-15	4.41812e-08	-9.25885e-07	1.08594e-08	-5.60283e-10	1.39168e-06	-0.00061398
3.3957e-12	-3.87004e-15	4.41812e-08	-9.25885e-07	1.08594e-08	-5.60283e-10	1.39168e-06	-0.00061398	3.81127e-11	5.59446e-09	3.9315e-07	-1.24428e-07	-1.51669e-14	-3.23477e-07	9.23869e-17	4.42376e-17
3.81127e-11	5.59446e-09	3.9315e-07	-1.24428e-07	-1.51669e-14	-3.23477e-07	9.23869e-17	4.42376e-17	-3.5699e-12	-7.06515e-08	6.54401e-16	4.33939e-07	1.17262e-07	-0.000222753	-2.34755e-09	-6.2929e-10
-3.5699e-12	-7.06515e-08	6.54401e-16	4.33939e-07	1.17262e-07	-0.000222753	-2.34755e-09	-6.2929e-10	-5.88025e-07	3.93649e-11	-1.40832e-14	1.8591e-09	1.14466e-09	4.51959e-13	-1.04752e-05	4.03138e-08
-5.88025e-07	3.93649e-11	-1.40832e-14	1.8591e-09	1.14466e-09	4.51959e-13	-1.04752e-05	4.03138e-08	6.92561e-09	-2.3025e-08	8.96315e-15	0.000109951	-9.80612e-08	1.38381e-05	1.7521e-19	-4.22073e-09
6.92561e-09	-2.3025e-08	8.96315e-15	0.000109951	-9.80612e-08	1.38381e-05	1.7521e-19	-4.22073e-09	7.49855e-13	2.2021e-09	5.89958e-10	8.72603e-12	5.63506e-10	2.48826e-21	-4.17328e-13	3.55109e-05
7.49855e-13	2.2021e-09	5.89958e-10	8.72603e-12	5.63506e-10	2.48826e-21	-4.17328e-13	3.55109e-05	-5.18027e-09	-2.26263e-09	-2.36871e-14	-5.68319e-10	-1.50333e-09	-8.74907e-07	-8.5568e-16	-5.15532e-12
-5.18027e-09	-2.26263e-09	-2.36871e-14	-5.68319e-10	-1.50333e-09	-8.74907e-07	-8.5568e-16	-5.15532e-12	2.54186e-07	3.60654e-11	1.46872e-07	-1.54165e-07	2.6166e-15	3.64129e-07	3.5724e-22	-1.42001e-14
2.54186e-07	3.60654e-11	1.46872e-07	-1.54165e-07	2.6166e-15	3.64129e-07	3.5724e-22	-1.42001e-14	-5.59275e-07	5.13803e-26	-1.9807e-07	7.8816e-05	-3.07882e-10	6.95753e-06	2.63963e-05	-2.67667e-17
-5.59275e-07	5.13803e-26	-1.9807e-07	7.8816e-05	-3.07882e-10	6.95753e-06	2.63963e-05	-2.67667e-17	-1.00883e-06	1.43013e-11	-1.672e-06	-2.10062e-05	9.30256e-10	-2.10472e-19	1.1538e-10	-2.10027e-14
-1.00883e-06	1.43013e-11	-1.672e-06	-2.10062e-05	9.30256e-10	-2.10472e-19	1.1538e-10	-2.10027e-14	-1.6212e-09	4.61035e-05	7.83018e-08	4.17835e-09	-1.11695e-07	-1.57536e-07	-5.31088e-18	-2.38712e-09
-1.6212e-09	4.61035e-05	7.83018e-08	4.17835e-09	-1.11695e-07	-1.57536e-07	-5.31088e-18	-2.38712e-09	3.97071e-24	8.3337e-08	3.698e-07	1.56459e-07	2.50273e-11	-7.11228e-07	1.08954e-07	4.88376e-22
3.97071e-24	8.3337e-08	3.698e-07	1.56459e-07	2.50273e-11	-7.11228e-07	1.08954e-07	4.88376e-22	4.94066e-324	0	0	0	0	0	0	0

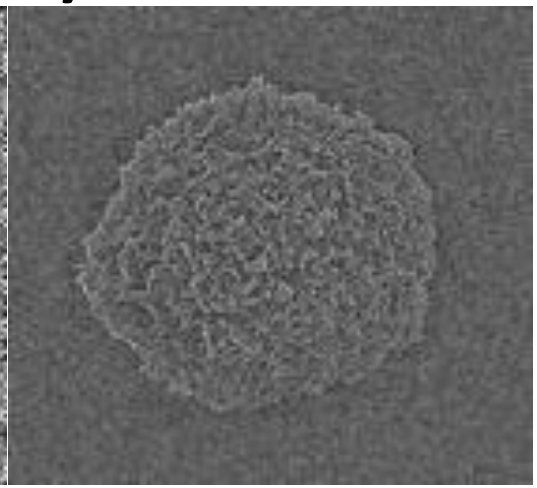
- **La TCD-CS (à Coefficients Signés)**



Originale



Matrice des coefficients binaires



TCD Inverse de la matrice des coefficients

- Calcul de la TCD-CS

$$\mathrm{sgn}(DCT(i,j)) = \begin{cases} 0 & if \quad DCT(i,j) \le 0 \\ 1 & if \quad DCT(i,j) > 0 \end{cases}$$

0	0	1	1	1	1	1	0	1	0	0	1	1	0	0	0
1	0	0	1	1	0	0	0	1	0	1	1	1	0	0	0
1	0	1	1	1	0	0	0	0	0	1	0	1	1	0	0
0	0	1	0	1	1	0	0	1	0	1	0	1	0	1	0
1	0	1	0	1	0	1	0	1	1	1	0	0	0	1	1
1	1	1	0	0	0	1	1	0	0	1	1	1	0	0	0
0	0	1	1	1	0	0	0	0	1	0	1	1	1	0	1
0	1	0	1	1	1	0	1	1	0	1	1	0	1	1	0
1	0	1	1	0	1	1	0	1	1	1	1	1	1	0	1
1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	1	1	0	1	1	1	0
1	1	1	0	1	1	1	0	0	1	0	1	0	1	1	0
0	1	0	1	0	1	1	0	0	1	0	0	1	0	1	0
0	1	0	0	1	0	1	0	0	1	1	1	0	0	0	0
0	1	1	1	0	0	0	0	1	1	1	1	1	0	1	1
1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1

- Génération de la clé de hachage :

Hash 64 bits : **00111110 10011000 10111000 00101100 10101010 11100011 00111000 01011101**

- Comparaison de deux clés de hachage :

Distance de Hamming :

Hash 1 : 00111110 10011000 10111000 00101100 10101010 11100011 00111000 01011101

Hash 2 : 0011**0**110 10011000 1011**1**100 00101100 10101010 1110**11**11 00111000 01011101

- Génération de 6 groupes de simulations

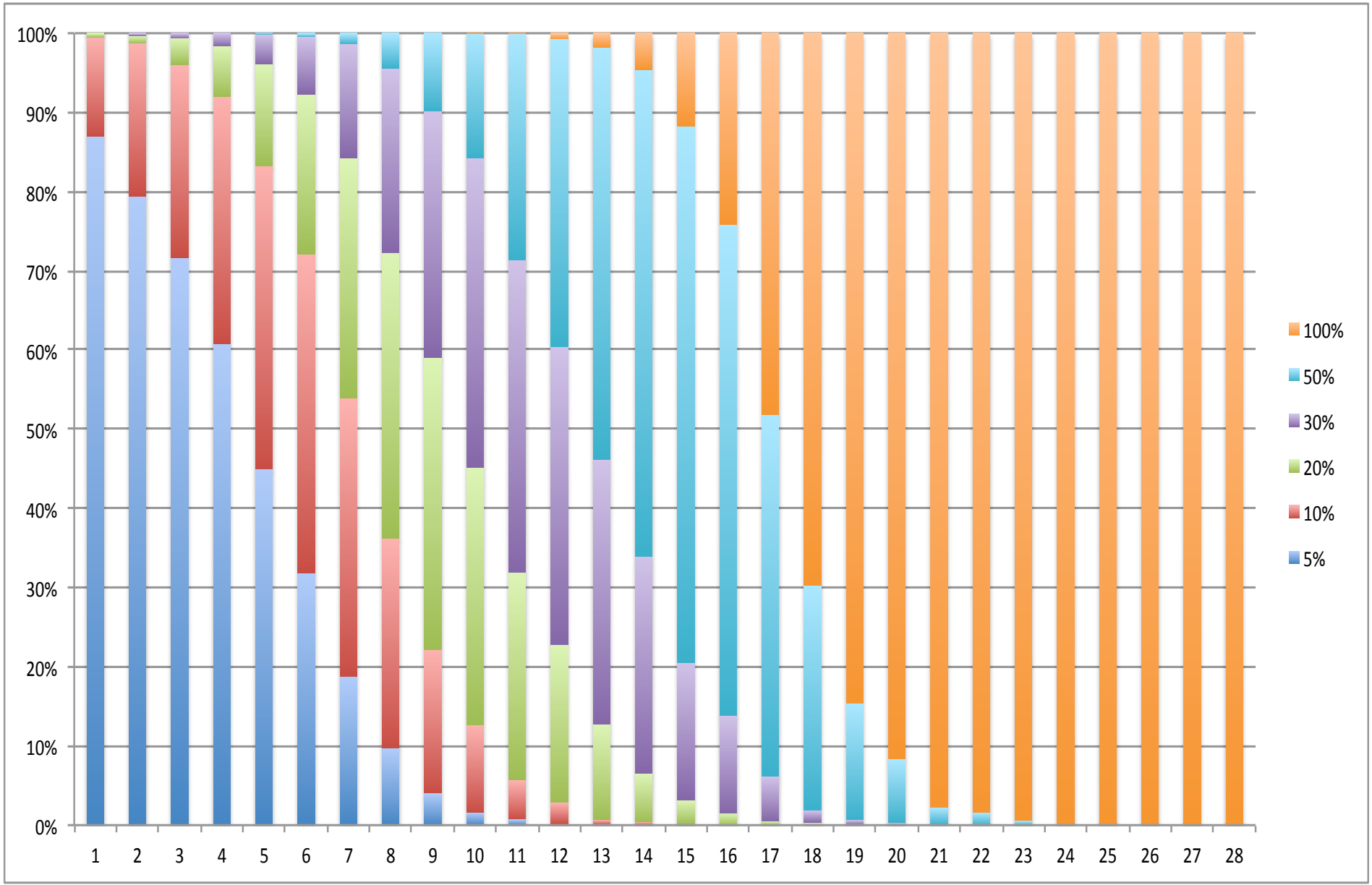
Group	Seq. length	Hash length	Factor reduction	Divergence rate
A	100 pb	4 octets (32 bits)	25x	5%, 10%, 30%, 50%, 100%
B	100 pb	8 octets (64 bits)	12,5x	5%, 10%, 30%, 50%, 100%
C	1000 pb	4 octets (32 bits)	250x	5%, 10%, 30%, 50%, 100%
D	1000 pb	8 octets (64 bits)	125x	5%, 10%, 30%, 50%, 100%
E	10000 pb	4 octets (32 bits)	2500x	5%, 10%, 30%, 50%, 100%
F	10000 pb	8 octets (64 bits)	1250x	5%, 10%, 30%, 50%, 100%

Temps d'exécution :

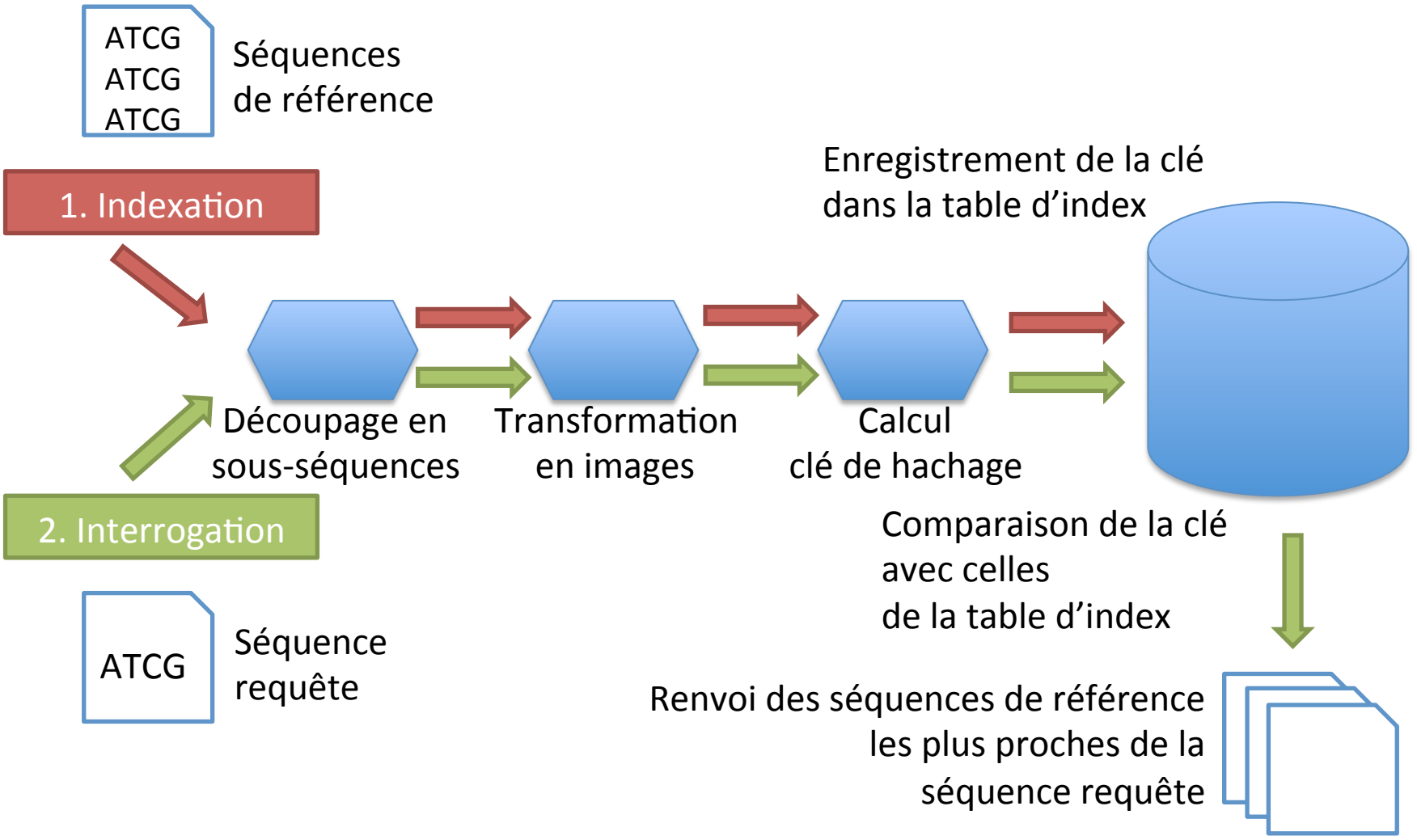
Group	Time	Hashes per second
A	177s	564 515
B	225s	443 037
C	3 686s	27 125
D	3 714s	26 922
E	396 825s	252
F	408 163s	245

Évaluation théorique par simulation :

- Graphique du groupe E (taille 10 000 – clé de 32 Bits)

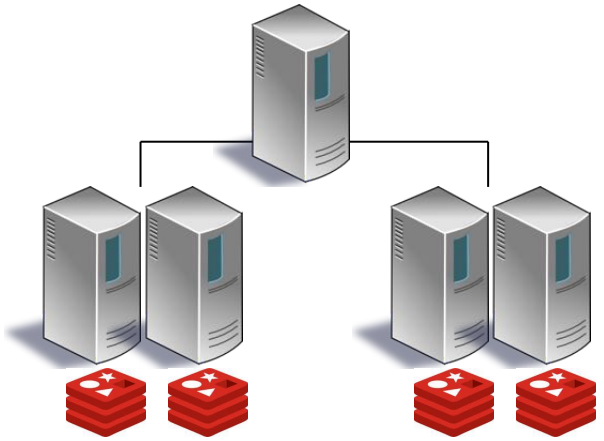


Système d'identification du génome de référence d'une séquence inconnu



- Stockage des clés de hachage
  - Base de données REDIS de type NoSQL
  - Système clé / valeur
  - Base de données « In-Memory »
  - Système de compression des données
  - Système de base de données réparties et dupliquées
- Structure des données

Clé :	Haches (4 ASCII) :
Seq1	1001100010000001 1111101010011001 1011110110101011 0111001100111001
SeqN…	0001011100111000



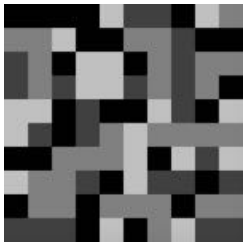
Alignement de séquences ADN par corrélation de TCD-CS

- Article PLOS|ONE : Millaray Curilem Saldías & al. **« Image Correlation Method for DNA Sequence Alignment »**
  - Une méthode d'alignement via une algorithme de corrélation de phase
  - La méthode est validée
  - Temps d'exécution 100x plus lent que BLAST

- **Objectifs :**
  - Optimiser cette méthode
    - Utilisation d'une TCD-CS plutôt qu'une TFD
    - Paralléliser l'algorithme
    - Écrire une implémentation pour GPU sous CUDA

- **La corrélation de phase**
  - Méthode utilisée en Physique Optique
  - Très largement utilisée dans le domaine du traitement d'image
    - Détection de formes, de mouvements...
  - Tolérance au bruit

Méthode d'alignement de séquence par corrélation DCT-CS



Référence

DCT

SO

Domaine fréquentiel

Produit d'Hadamard

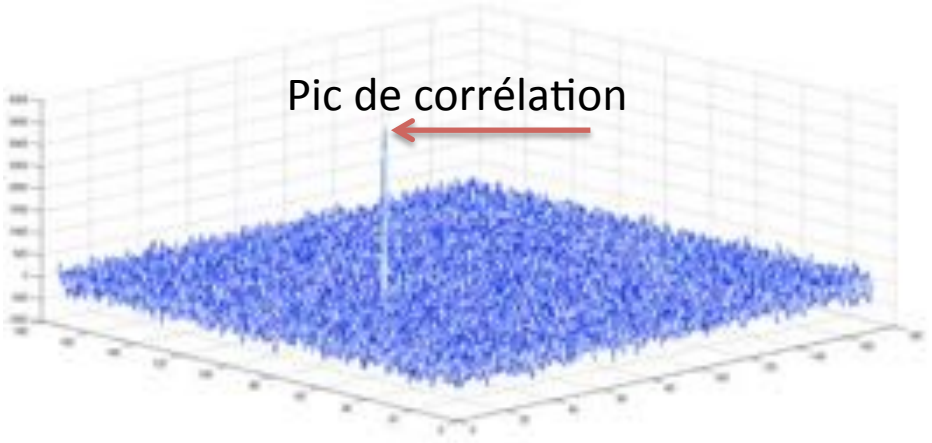
iDCT

SO

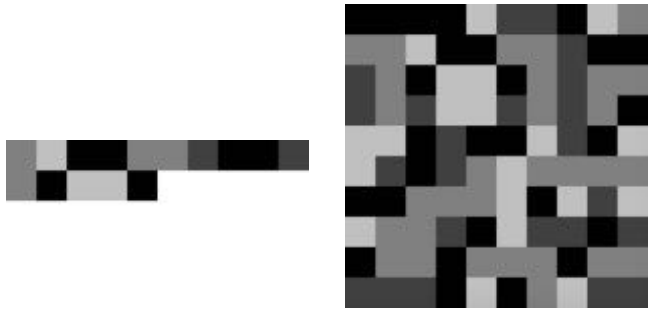
DCT

Requête

Pic de corrélation

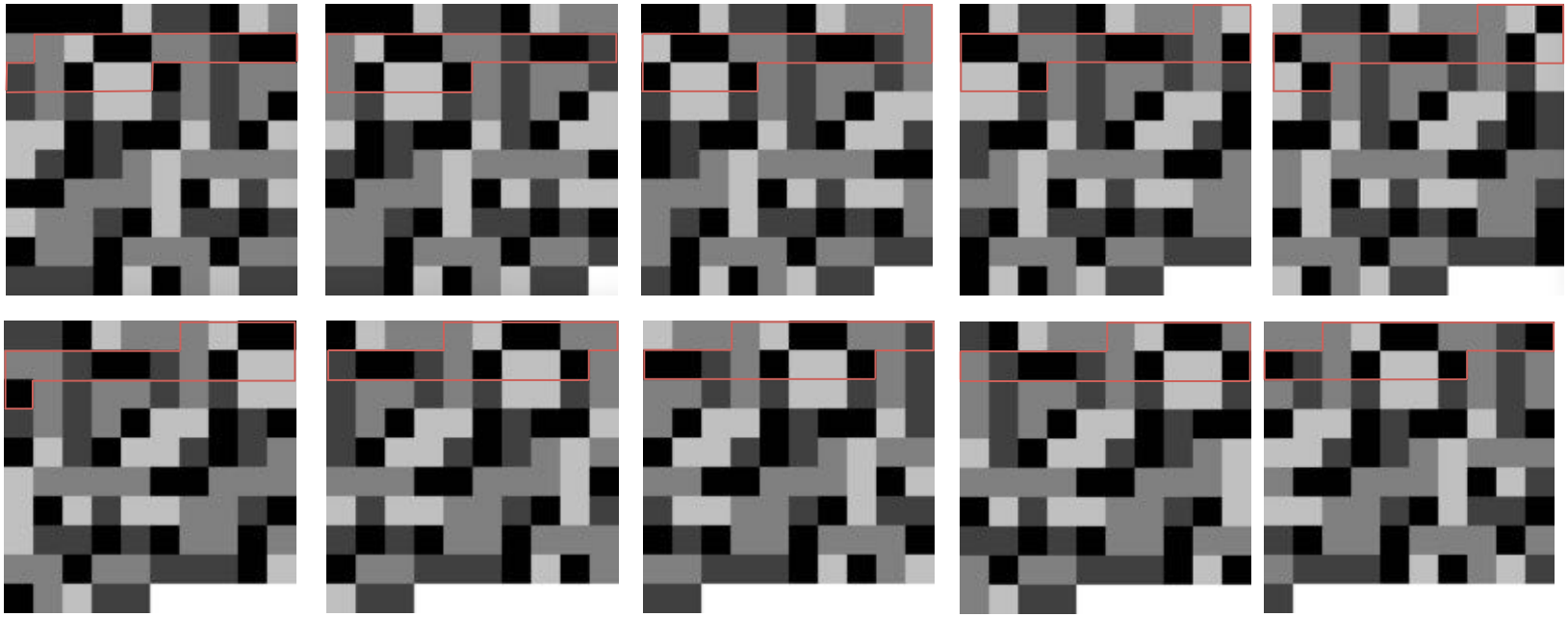


- Afin d'estimer le meilleur alignement, on réalise plusieurs corrélations en opérant des décalages successifs sur toutes la première ligne de la séquence de référence



Séquence requête

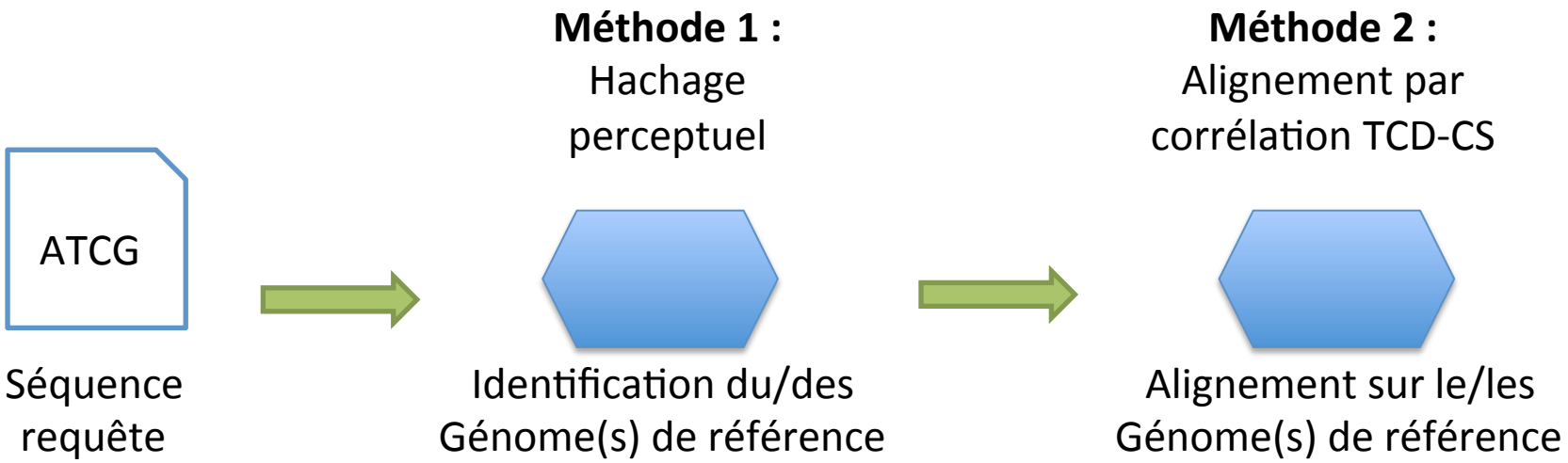
Séquence référence



- **Évaluation par simulation :**
  - 100 000 paires de séquences
    - Séquences primaires d'une taille de 10 000 pb
    - Sous-séquences d'une taille de 100 pb

- **Résultats préliminaires :**
  - Le taux d'alignement > 95%
  - Le temps d'exécution est de 17 min
- **Perspectives :**
  - Poursuivre l'implémentation de la méthode
  - Élaborer de nouveaux tests
  - Développer une version pour GPU
  - Comparer les résultats avec BLAST

- **Système d'information**



- **Technologies :**
  - Développement : langage C++, CUDA
  - Base de données : REDIS