Visualisation interactive des clones V(D)J en fonction des distances d'édition, et suivi de la Leucémie Aigüe Lymphoblastique

Antonin Carette
Tuteur entreprise: **Mathieu Giraud**
Tuteur Universitaire: **Samy Meftali**

LIFL - Équipe Bonsai
Université Lille1

27 Juin 2014

Sommaire

Le **LIFL** (**Laboratoire d'Informatique Fondamentale de Lille**) est un laboratoire Français rattaché à l'Institut des Sciences de l'Information et de leurs Interactions (INS2I) du CNRS (Centre National de la Recherche Scientifique) - basé sur le campus de Lille1.
Dirigé par **Sophie Tison**, il comprend actuellement 10 équipes-projets et équipes du centre de recherche INRIA Lille - Nord Europe.

Le **LIFL** (**Laboratoire d'Informatique Fondamentale de Lille**) est un laboratoire Français rattaché à l'Institut des Sciences de l'Information et de leurs Interactions (INS2I) du CNRS (Centre National de la Recherche Scientifique) - basé sur le campus de Lille1.
Dirigé par **Sophie Tison**, il comprend actuellement 10 équipes-projets et équipes du centre de recherche INRIA Lille - Nord Europe.

L'équipe **Bonsai** est une équipe de bio-informaticiens.
Elle comprend actuellement 20 membres (chercheurs, enseignants chercheurs, thésards, ...), et est actuellement dirigée par **Hélène Touzet**.
Travail avec **Mathieu Giraud**, **Mikaël Salson** et **Marc Duez** sur l'ADN / LAL.

Généralités sur l'ADN

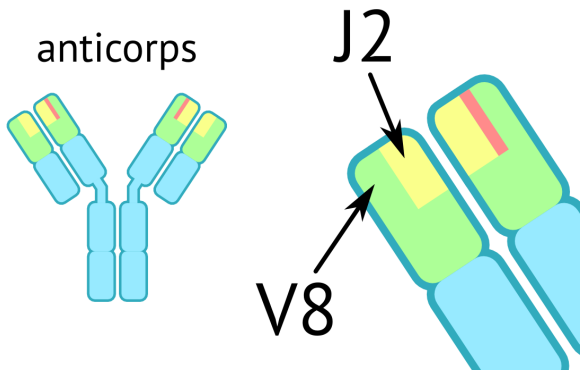
ADN : molécule modélisée en double-hélice, support universel de l'information génétique.
Constituée d'un alphabet de 4 lettres : **A**,**T**,**G**,**C**.
Peut être sujette à des **recombinaisons**, par différents phénomènes biologiques.

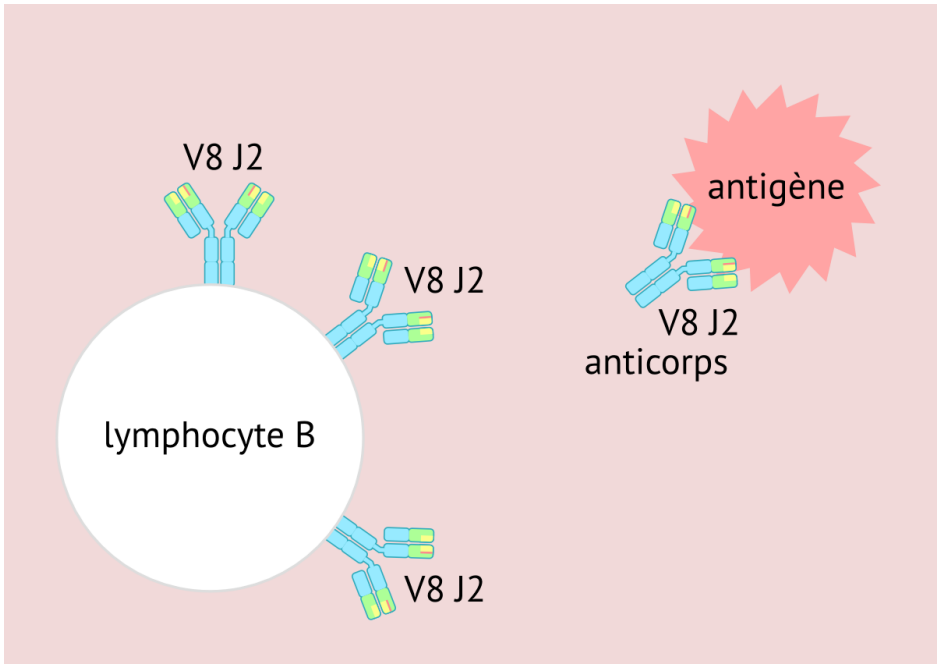
Généralités sur l'ADN

ADN : molécule modélisée en double-hélice, support universel de l'information génétique.
Constituée d'un alphabet de 4 lettres : **A**,**T**,**G**,**C**.
Peut être sujette à des **recombinaisons**, par différents phénomènes biologiques.

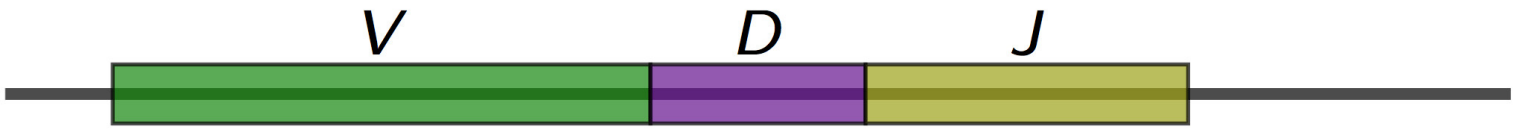
Les Leucocytes : gardiens de notre corps

Leucocyte = globule blanc.
Les **lymphocytes** font partie d'une classe de leucocytes.
Lors de la reconnaissance (spécifique) : activation -> multiplication.
Lymphocytes strictement identiques = **clones V(D)J**.

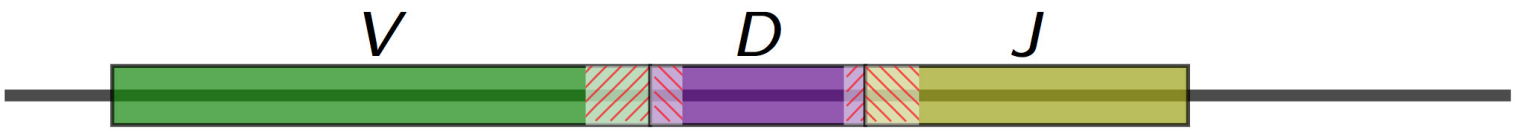

anticorps
J2
V8


V8 J2
V8 J2
lymphocyte B
V8 J2
antigène
V8 J2
anticorps

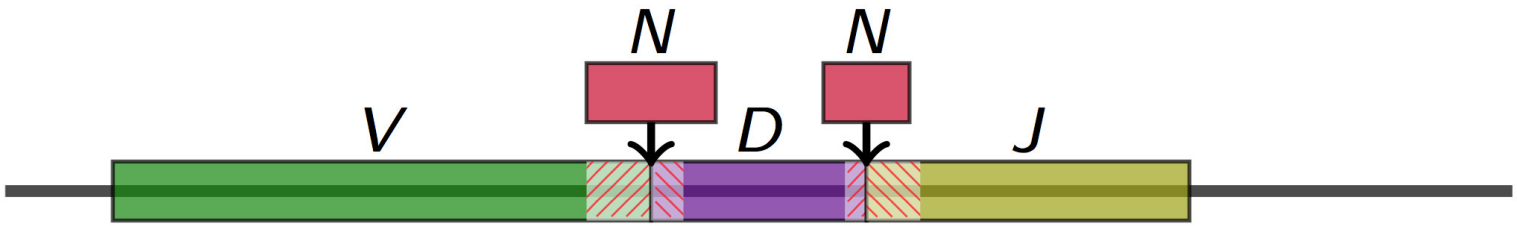
La recombinaison V(D)J



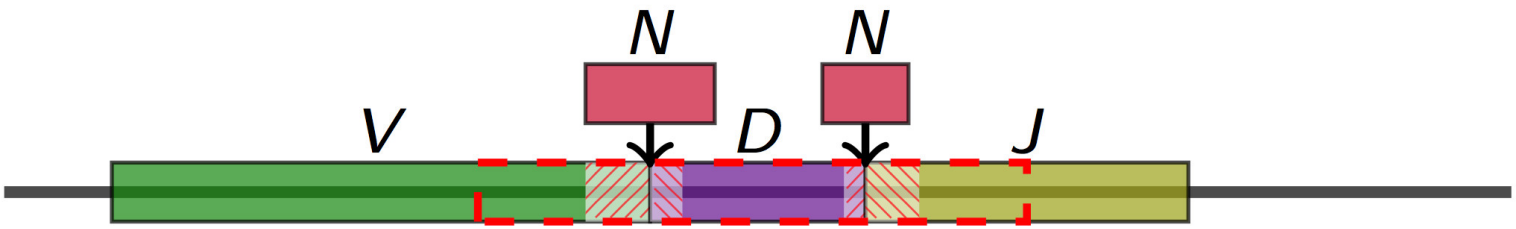
La recombinaison V(D)J



La recombinaison V(D)J



La recombinaison V(D)J



Diversity region

Chaque lymphocyte porte une portion d'ADN V(D)J permettant son identification !

2.10^{12} recombinaisons possibles.

La Leucémie Aigüe Lymphoblastique (LAL)

Cancer liquide affectant majoritairement les enfants.
Surproduction de lymphocytes cancéreux

La Leucémie Aigüe Lymphoblastique (LAL)

Cancer liquide affectant majoritairement les enfants.
Surproduction de lymphocytes cancéreux

Problème

L'accumulation de lymphocytes cancéreux provoque :

- une anémie chez le patient,
- une diminution du nombre de plaquettes,
- une diminution des globules blancs normaux.

Peut amener à la mort, chez le malade.

Vidjil : collaboration entre l'équipe Bonsai et le Centre d'Hématologie de l'Université Lille2. Projet maintenu par **Claude Preudhomme**, **Mathieu Giraud** et **Mikaël Salson**. Ingénieur sur le projet : **Marc Duez**.

Composition du projet

Vidjil est composé de deux programmes :

- un programme en C++ (*algorithme*),
- une interface Web (*interface*) : HTML5, Javascript, Ajax, *frameworks* D3JS et JQuery, Python (web2py)

Vidjil : collaboration entre l'équipe Bonsai et le Centre d'Hématologie de l'Université Lille2. Projet maintenu par **Claude Preudhomme**, **Mathieu Giraud** et **Mikaël Salson**. Ingénieur sur le projet : **Marc Duez**.

Composition du projet

Vidjil est composé de deux programmes :

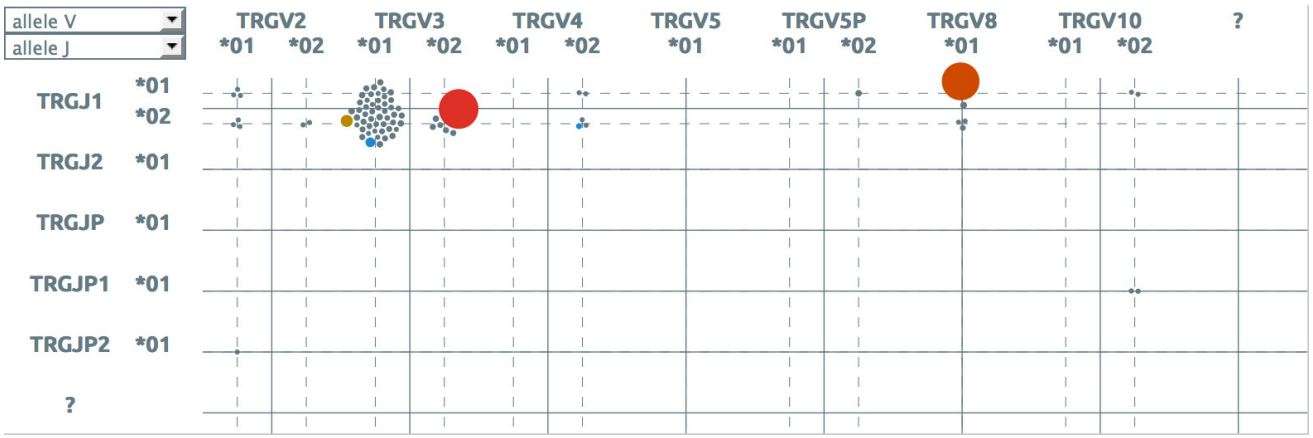
- un programme en C++ (*algorithme*),
- une interface Web (*interface*) : HTML5, Javascript, Ajax, *frameworks* D3JS et JQuery, Python (web2py)

À quoi sert-il ?

- **Détection, dénombrement et suivi des clones V(D)J**.
- **Précision des traitements thérapeutiques**.

Présentation de l'*Interface*

Visualisation selon la distribution **V/J Genes**



Visualisation selon la distribution **V/J Alleles**

Problème

Sur une distribution déjà implanté, comment savoir, dans un groupe de clones, lesquels sont les plus similaires les uns des autres ?

Problème

Sur une distribution déjà implanté, comment savoir, dans un groupe de clones, lesquels sont les plus similaires les uns des autres ?

Solution 1

La solution est de comparer les clones un par un, nucléotide par nucléotide... Ce qui est fastidieux, et pouvant être un source d'erreur !

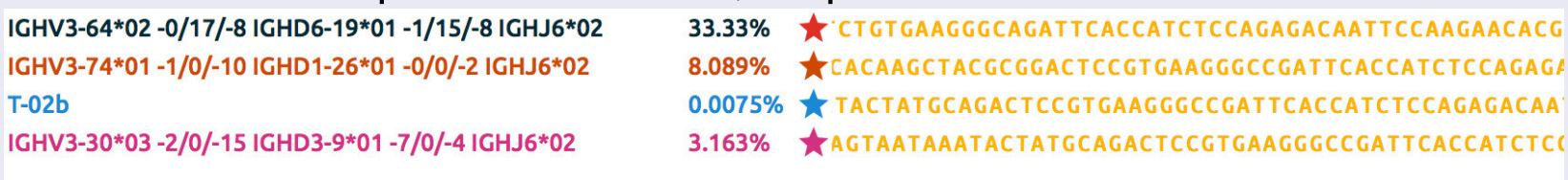
IGHV3-64*02 -0/17/-8 IGHD6-19*01 -1/15/-8 IGHJ6*02	33.33%	CTGTGAAGGGCAGATTCACCATCTCCAGAGACAATTCCAAGAACACG
IGHV3-74*01 -1/0/-10 IGHD1-26*01 -0/0/-2 IGHJ6*02	8.089%	CACAAGCTACGCGGACTCCGTGAAGGGCCGATTCACCATCTCCAGAGA
T-02b	0.0075%	TACTATGCAGACTCCGTGAAGGGCCGATTCACCATCTCCAGAGACAA
IGHV3-30*03 -2/0/-15 IGHD3-9*01 -7/0/-4 IGHJ6*02	3.163%	AGTAATAAATACTATGCAGACTCCGTGAAGGGCCGATTCACCATCTC

Problème

Sur une distribution déjà implanté, comment savoir, dans un groupe de clones, lesquels sont les plus similaires les uns des autres ?

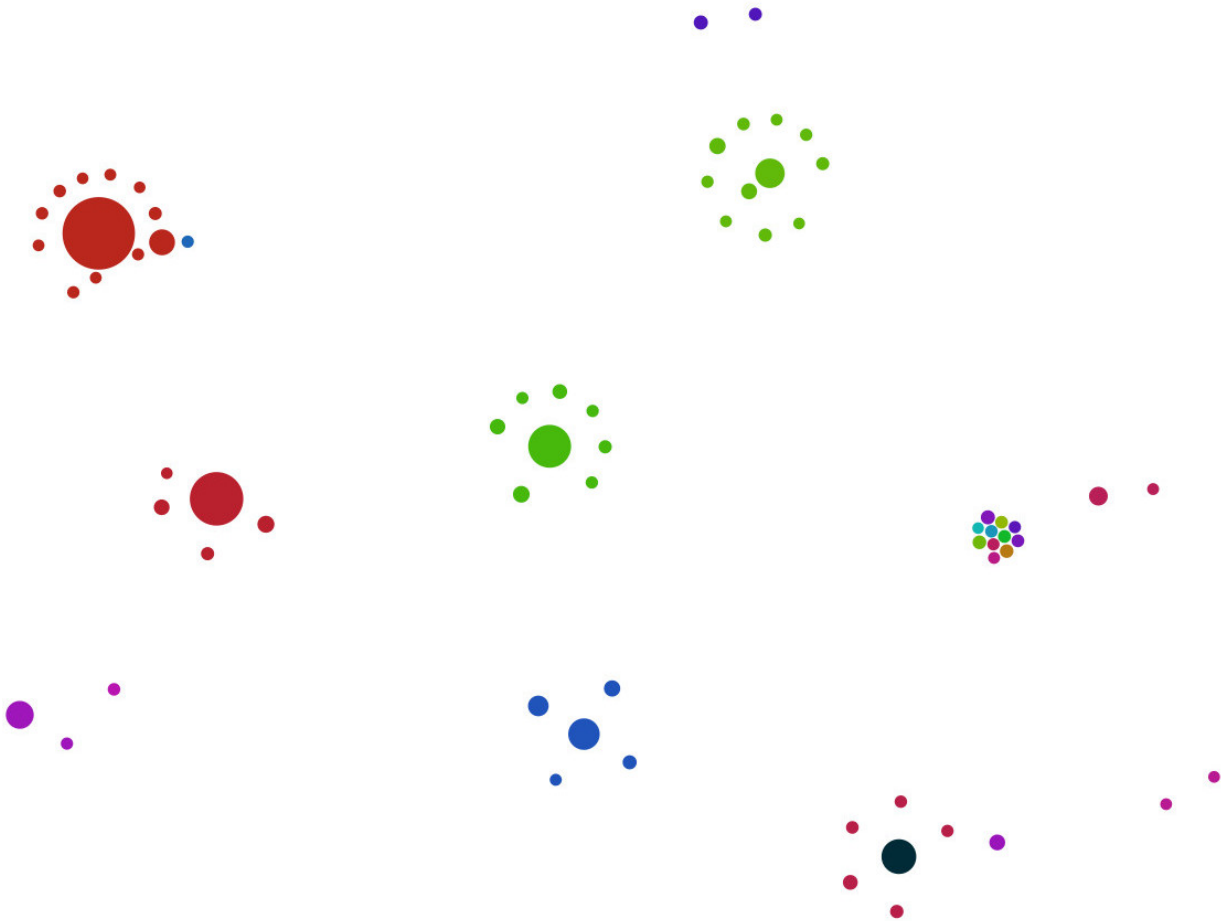
Solution 1

~~Comparer les clones un par un avec l'*interface*, nucléotide par nucléotide... Ce qui est fastidieux, et peut être une source d'erreur !~~



Solution 2

Créer un outil permettant de visualiser directement les clones en fonction de leurs similarités.



Objectifs

Deux objectifs :

- création d'un graphe représentant les distances d'édition entre les clones,
- implantation de l'algorithme de partitionnement DBSCAN.

Objectifs

Deux objectifs :

- création d'un graphe représentant les distances d'édition entre les clones,
- implantation de l'algorithme de partitionnement DBSCAN.

Mais avant...

...il faut d'abord commencer à calculer les distances d'édition !

Sommaire

Qu'est-ce qu'une distance d'édition ?

Une distance d'édition est un nombre de modifications à apporter selon trois opérations, permettant de passer d'une chaîne de caractères à une autre.

3 opérations : insertion/délétion/substitution.

Nous prendrons le poids suivant : **0** pour un *match*, **1** pour insertion/délétion/substitution -> *Levenshtein*.

Qu'est-ce qu'une distance d'édition ?

Une distance d'édition est un nombre de modifications à apporter selon trois opérations, permettant de passer d'une chaîne de caractères à une autre.

3 opérations : insertion/délétion/substitution.

Nous prendrons le poids suivant : **0** pour un *match*, **1** pour insertion/délétion/substitution -> *Levenshtein*.

Exemple 1

Prenons deux chaînes de caractères de même longueur : CHIEN / NICHE.

	C	H	I	E	N
N	-	-	-	-	-
I	-	-	-	-	-
C	-	-	-	-	-
H	-	-	-	-	-
E	-	-	-	-	-

Qu'est-ce qu'une distance d'édition ?

Une distance d'édition est un nombre de modifications à apporter selon trois opérations, permettant de passer d'une chaîne de caractères à une autre.

3 opérations : insertion/délétion/substitution.

Nous prendrons le poids suivant : **0** pour un *match*, **1** pour insertion/délétion/substitution -> *Levenshtein*.

Exemple 1

Prenons deux chaînes de caractères de même longueur : CHIEN / NICHE.

	C	H	I	E	N
N	**1**	2	3	4	4
I	**2**	2	2	3	4
C	**2**	3	3	3	4
H	3	**2**	**3**	4	4
E	4	3	3	**3**	**4**

Qu'est-ce qu'une distance d'édition ?

Une distance d'édition est un nombre de modifications minimal à apporter selon trois opérations, permettant de passer d'une chaîne de caractères à une autre.

3 opérations : insertion/délétion/substitution.

Nous prendrons le poids suivant : **0** pour un *match*, **1** pour insertion/délétion/substitution -> *Levenshtein*.

Exemple 1

Prenons deux chaînes de caractères de même longueur : CHIEN / NICHE.

Le résultat du calcul est : - - **C H** - **E** - :

- **deux insertions** : N, I,
- **deux délétions** : I, N.

Soit une distance d'édition de **4**.

Implantation du calcul dans le programme C++.
Environ 5000 distances retournées pour 100 clones, reprises dans l'*interface*

Mais, sur quoi calcule-t-on ?

Comparaison sur deux **séquences de 40 nucléotides** comprises entre la fin de la région V, et le début de la région J = les ***windows***

Implantation du calcul dans le programme C++.
Environ 5000 distances retournées pour 100 clones, reprises dans l'*interface*

Mais, sur quoi calcule-t-on ?

Comparaison sur deux **séquences de 40 nucléotides** comprises entre la fin de la région V, et le début de la région J = les ***windows***

Pourquoi les *windows ?*

Window = marqueur spécifique d'un lymphocyte.
Avantage par rapport au suivi au cours du temps, en calcul de distances,...

Sommaire

Qu'est-ce qu'un graphe ?

En informatique, un graphe est une structure de données représentée par deux éléments de base :

- les noeuds,
- les arêtes.

Qu'est-ce qu'un graphe ?

En informatique, un graphe est une structure de données représentée par deux éléments de base :

- les noeuds,
- les arêtes.

Idée de la visualisation

- Noeuds : clones,
- Arêtes : distances.

- Visualisation des distances en fonction d'une distance d'édition maximale, dans le *scatterPlot*.
- Permission de faire varier la distance d'édition maximale, de façon interactive (avec l'utilisation d'un *slider*).
- Implémentation en HTML5, Javascript et les frameworks D3JS et JQuery.

Deux idées :

- visualisation en un seul graphe,
- visualisation en plusieurs graphes indépendants.

Représentation d'un graphe complet

Représentation d'un graphe complet

Problème

Faiblesse du moteur physique

Solution

Réduction du nombre d'arêtes inutiles

Représentation d'un graphe complet

Problème

Faiblesse du moteur physique

Solution

Réduction du nombre d'arêtes inutiles

Problème

Représentation non-respectueuse des distances d'édition

Recherche et solution

Semi-échec.
Recherches par rapport à la meilleure visualisation possible avec les outils disponibles : visualisation de graphes indépendants.

Avec plusieurs graphes (indépendants)

Solution retenue.

Avantages

- Rapidité du calcul, et de la représentation
- Visualisation d'un graphe intéressant en particulier
- Respect (dans la grande majorité) des données recueillies par le moteur physique

Avec plusieurs graphes (indépendants)

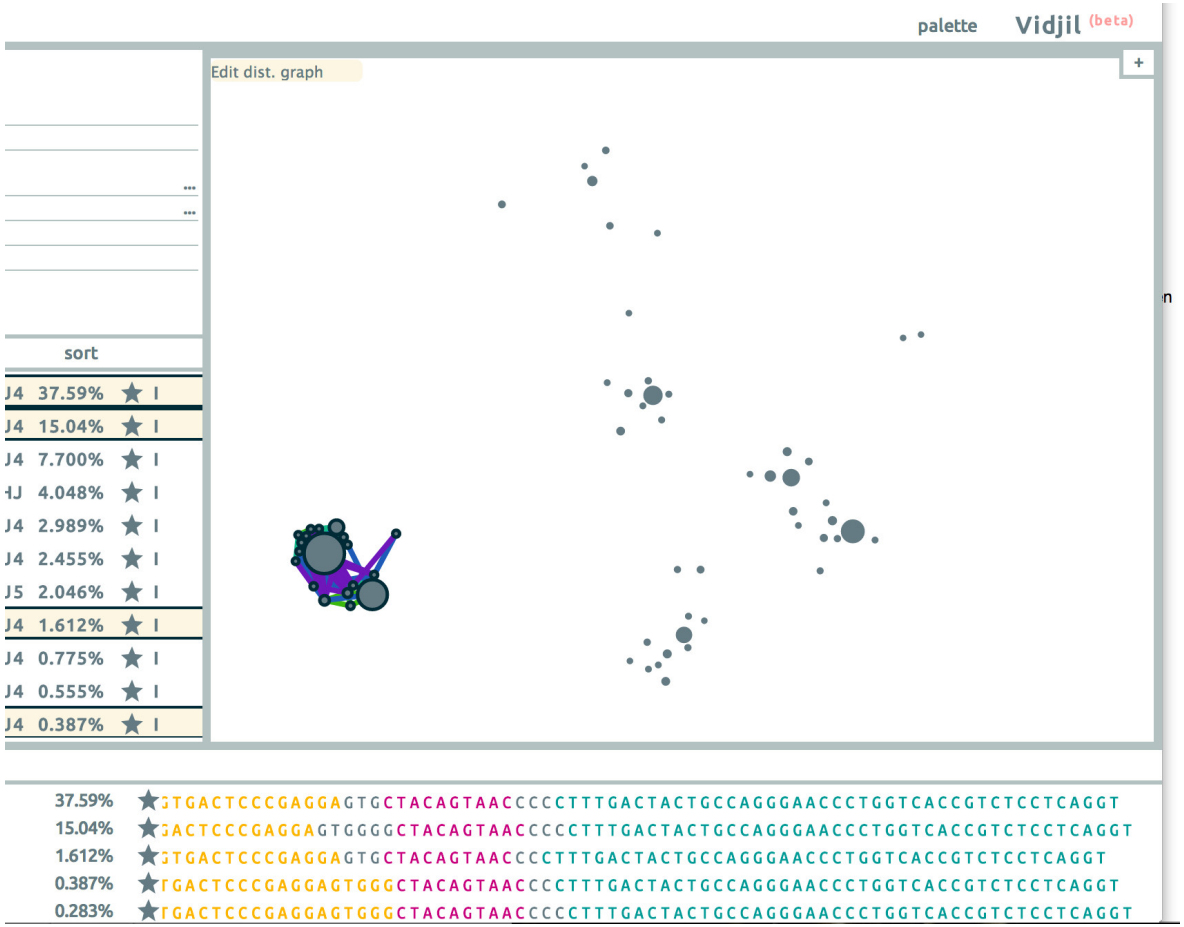
Solution retenue.

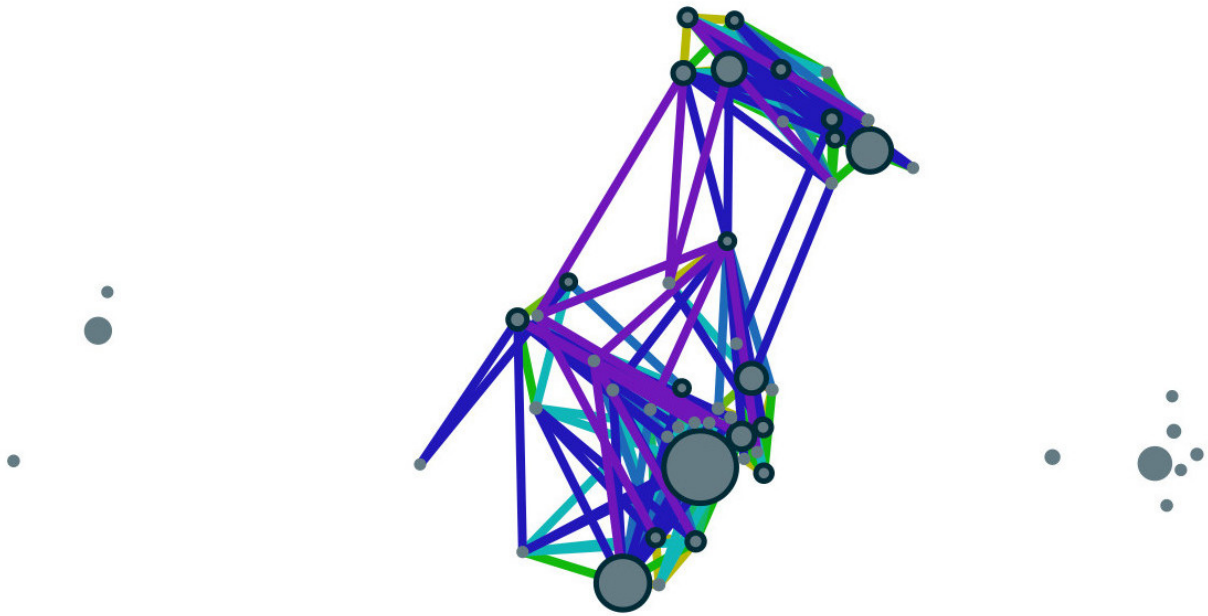
Avantages

- Rapidité du calcul, et de la représentation
- Visualisation d'un graphe intéressant en particulier
- Respect (dans la grande majorité) des données recueillies par le moteur physique

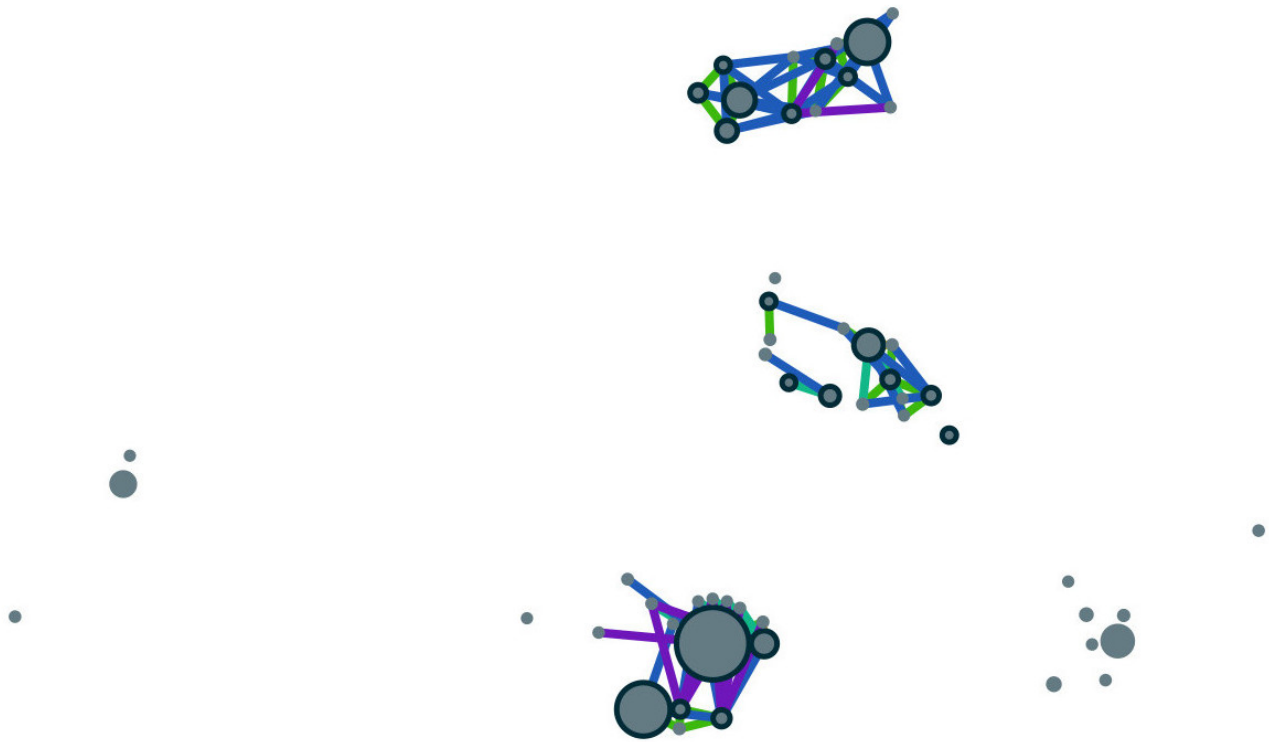
Inconvénients

- Représentation de tous les clones V(D)J entre eux impossible, car prise en compte de 50% de similarité maximum.


palette
Vidjil (beta)
Edit dist. graph
sort
37.59%
15.04%
7.700%
4.048%
2.989%
2.455%
2.046%
1.612%
0.775%
0.555%
0.387%
37.59%
15.04%
1.612%
0.387%
0.283%



Distance d'édition de 9.



Distance d'édition de 5.

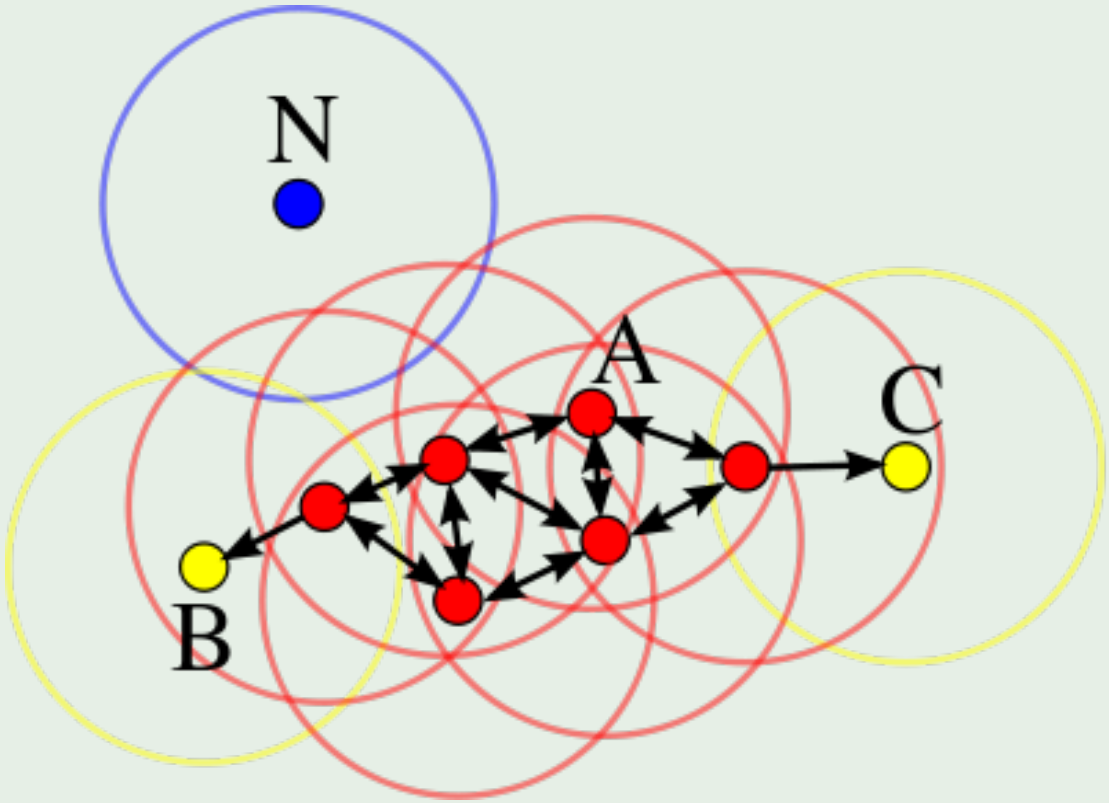
Distance d'édition de 1.

Visualisation basée sur l'**algorithme de clusterisation DBSCAN**

DBSCAN

DBSCAN = Density-Based Spatial Clustering of Applications with Noise
Algorithme publié en 1996 par **Martin Ester**, **Hans-Peter Kriegel**, **Jörg Sander** et **Xiaowei Xu**.
Utilise la densité estimée des clusters (la distance d'édition - ϵ - ainsi que le nombre de voisin minimum - *MinPts* - d'un clone) pour partitionner.

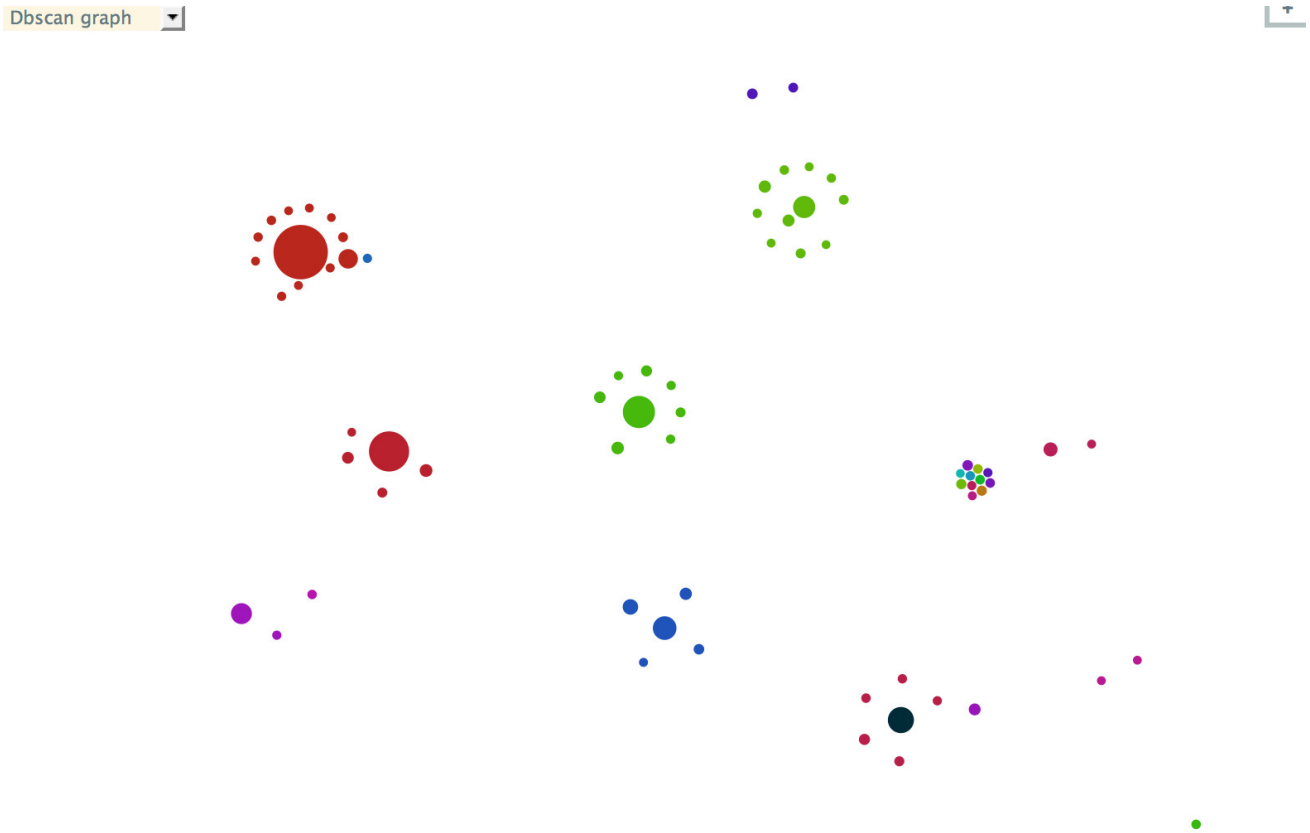
Exemple

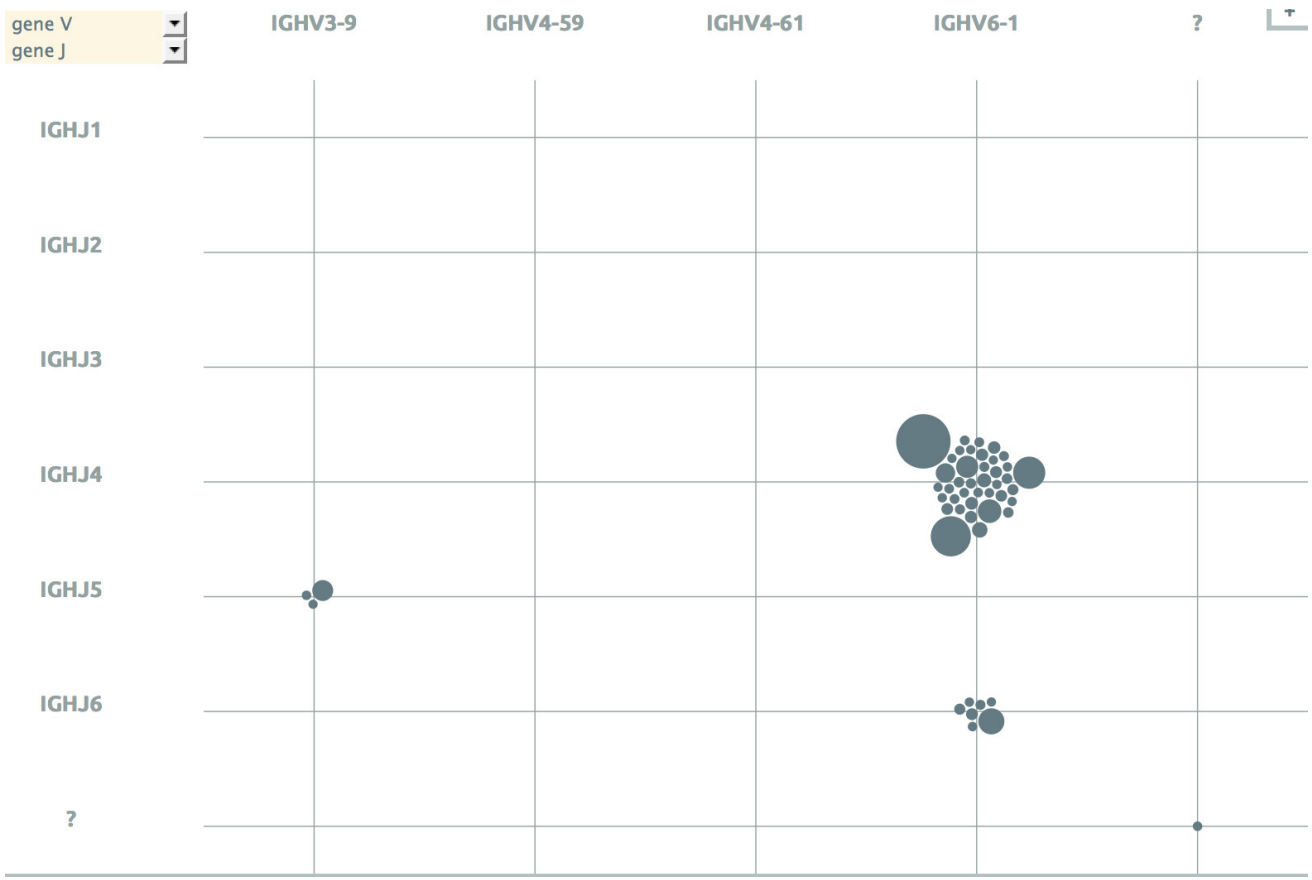


Exemple DBSCAN à *MinPts* = 0 - **en.wikipedia.org**

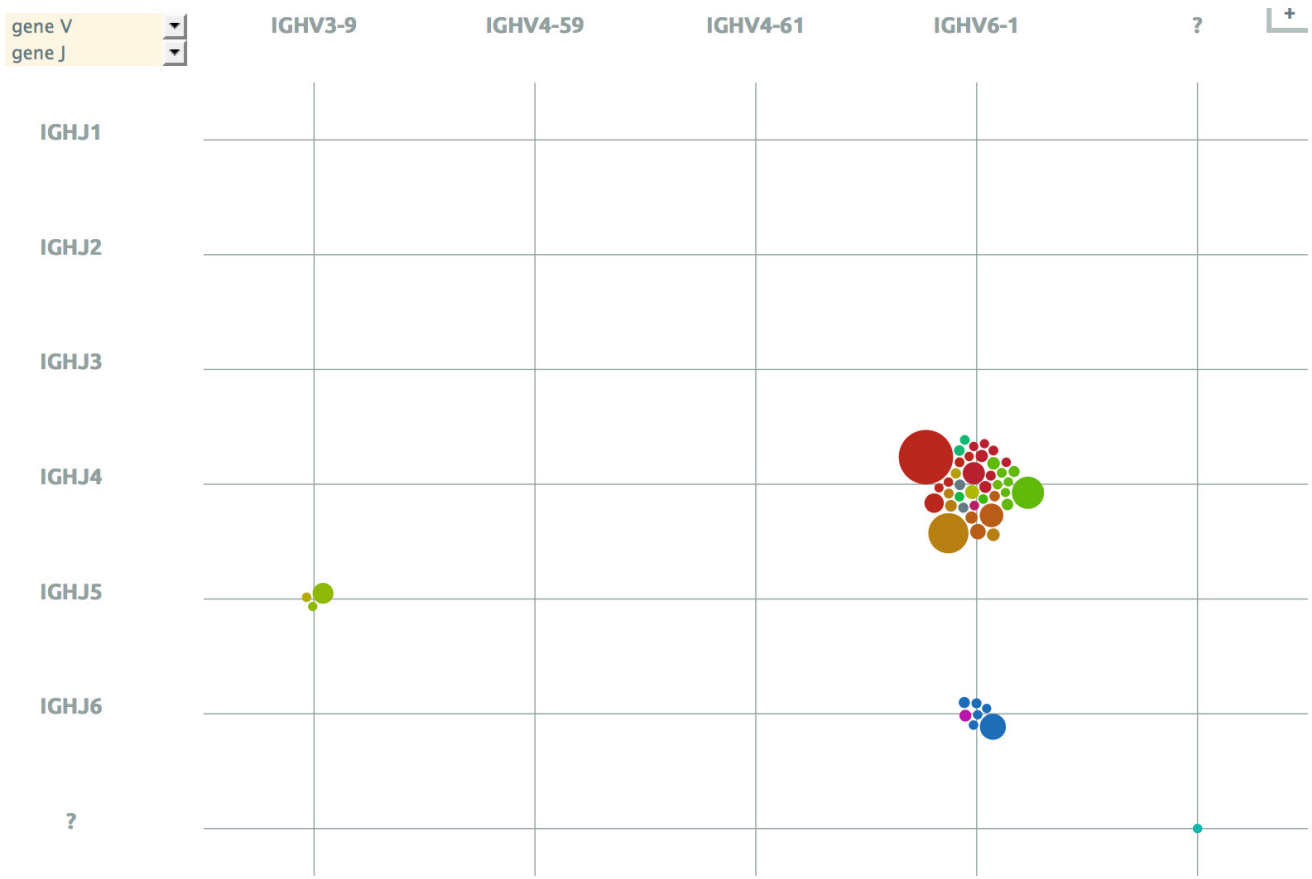
Idée de la visualisation

- Permission de faire varier ϵ ainsi que *MinPts* de façon interactive pour partitionner.
- Instanciation d'un objet DBSCAN à chaque modification d'un/des paramètre(s).
- Cluster = sphère contenant le voisinage, autour du noeud *CORE*.
- Ajout d'une couleur aléatoire pour chaque cluster, pour mieux les différencier sur cette visualisation et les précédentes.
- HTML5, Javascript, et les frameworks D3JS et JQuery - création d'une classe objet spécifique pour faire de l'algorithme un objet + tests unitaires QUnit.


Dbscan graph



Distribution en fonction des gènes, sans colorisation



Même distribution, avec colorisation DBSCAN ($\epsilon = 2$)

Sommaire

Résolution de la problématique

- Problématique résolue via le graphe de distances d'édition et DBSCAN
- Perspectives :
  - implémentation d'un arbre phylogénétique, remplacement au graphe de distribution (**PJI ?**)
  - implémentation des distances d'édition entre clones sur la sortie PDF.

Résolution de la problématique

- Problématique résolue via le graphe de distances d'édition et DBSCAN
- Perspectives :
  - implémentation d'un arbre phylogénétique, remplacement au graphe de distribution (**PJI ?**)
  - implémentation des distances d'édition entre clones sur la sortie PDF.

Bilan personnel

- Intérêt confirmé pour le monde de la recherche
- Apports dans l'organisation et le travail en groupe
- Apports quant à l'ingénierie logicielle ainsi que l'IHM (Intéraction Homme-Machine)

Merci pour votre attention !

Avez-vous des questions ?