e-Biothon : une étape vers l'interopérabilité entre centres nationaux et régionaux et bientôt vers le cloud

A. Franc[1], J.-M. Frigerio[1], P. Blanchard[2], P. Chaumeil[1], O. Coulaud[2], P. Gay[3], F. Rue[2], S. Thérond[4]

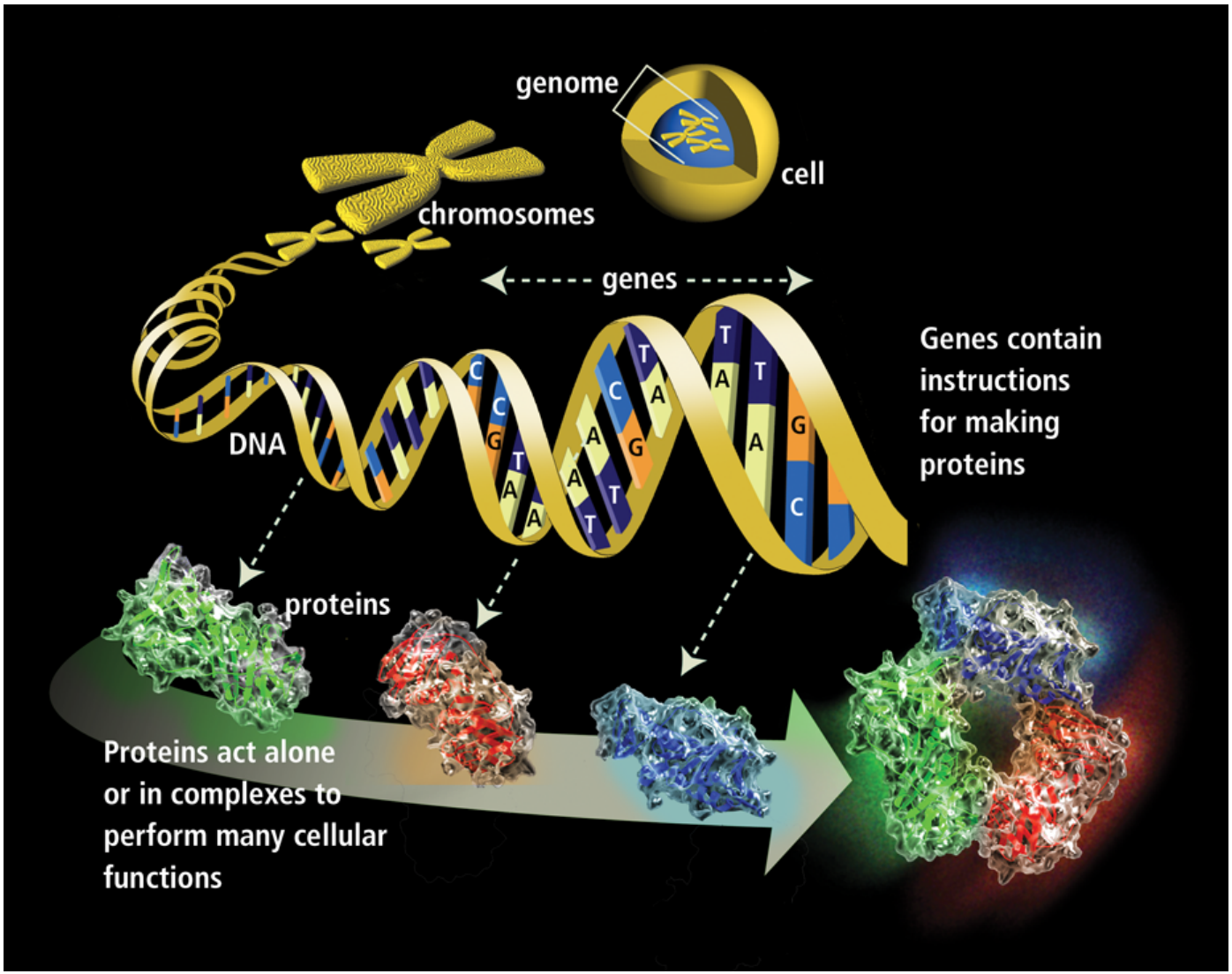
~

[1] INRA BioGeCo & INRIA Pleiades
[2] INRIA HIEPACS
[3] MCIA, Université de Bordeaux
[4] IDRIS, CNRS

Journées SUCCESS ; 23 novembre 2016

La notion de donnée en biologie

Des mots (*ATCGTCGA*...), pas des nombres ...



http://www.councilforresponsiblegenetics.org/geneticprivacy/DNA_sci.html

La diversité comme structure des différences entre textes

Notion de diversité ...

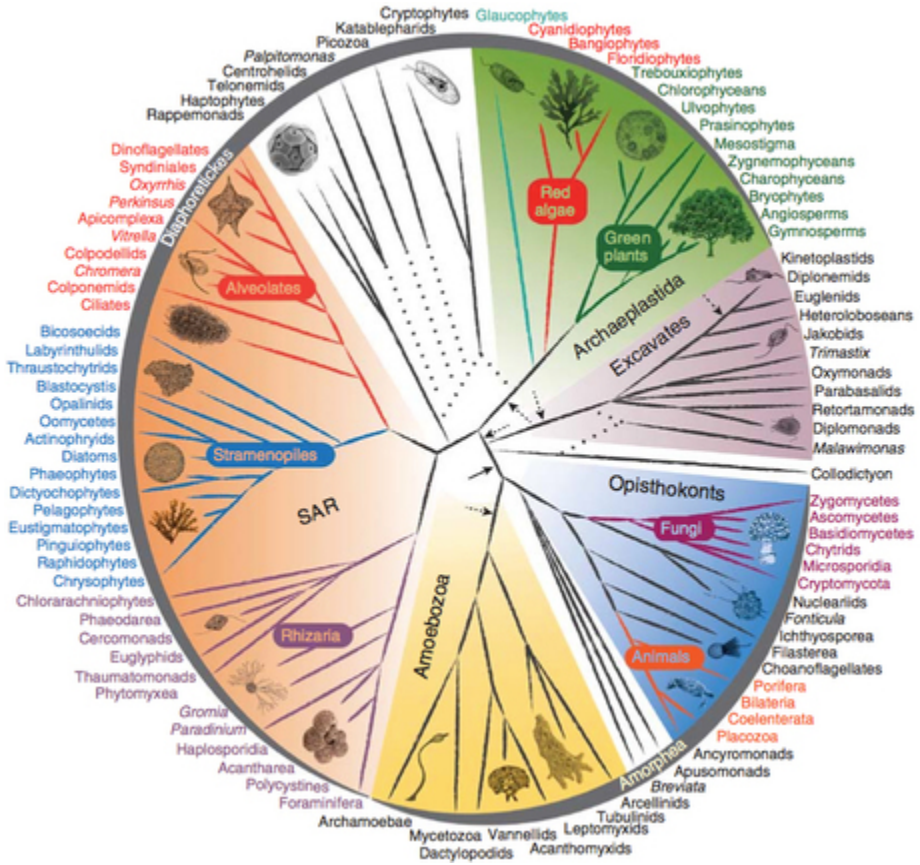
- Rechercher des ressemblances / dissimilarités entre objets différents ;
- la seule "réalité" est l'individu ;
- les catégories : espèce, genre, famille etc. sont des constructions ... issues de classifications

Exemple

```
ATTTCTCGATGTAGCGAGGCATGGCAGT
PARMILESANIMAUXJESUISUNCHIEN
PARMILESANIMAUXJESUISUNCHAT-
PARMILESPLANTESJESUISUNCHE-NE
PARMILESPLANTESJESUISUN-HETRE
```

L'immense diversité des Eucaryotes !

2



The Eukaryotic Tree of Life from a Global Phylogenomic Perspective. Cold Spring Harb Perspect Biol. 2014. 6 :a016147

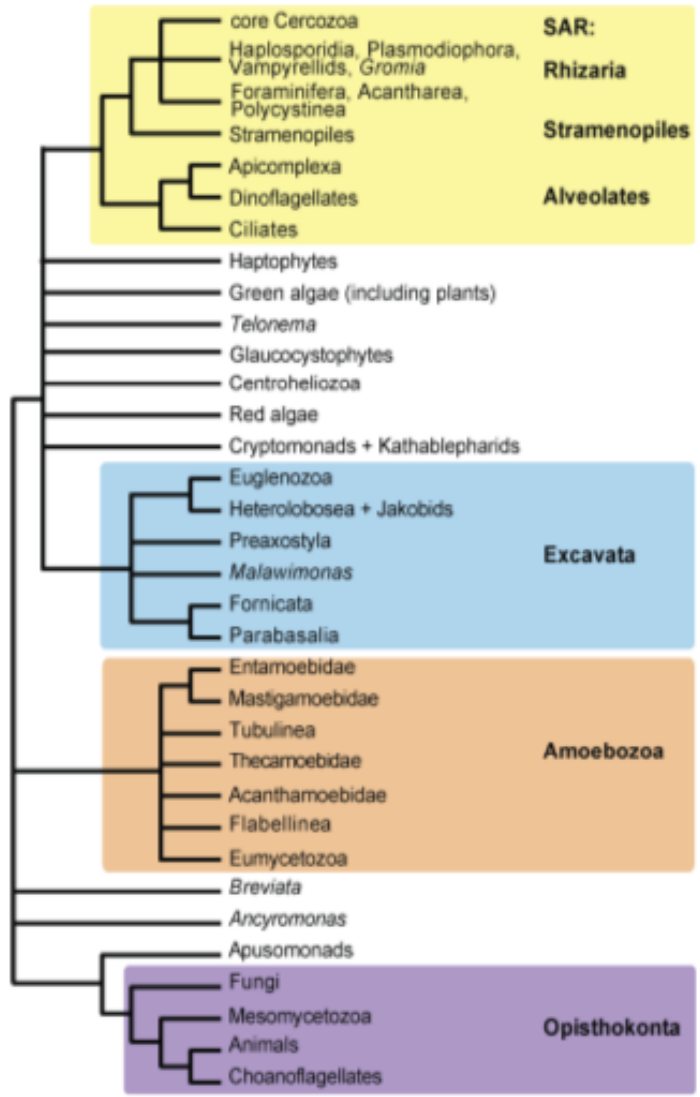


FIGURE 5. Summary of major findings—the evolutionary relationships among major lineages of eukaryotes. Clades have been collapsed into those that we view to be strongly supported. The many polytomies represent uncertainties that remain

Quels outils pour la diversité moléculaire ?

Une typologie des approches bien établies

Phylogénies moléculaires

- parcimonie
- "Neighbor Joining"
- maximum de vraisemblance
- inférence bayésienne

Distances entre séquences

- distance d'édition (Levenstein)
- alignement global (Needleman-Wunsch)
- alignement local (Smith-Waterman)

Deux remarques classiques

1. la distance génétique n'est pas la distance évolutive
2. un arbre est équivalent à l'existence d'une distance ultramétrique

Deux types de classification

Une remarque ...

Deux domaines énormes du **machine learning**

Supervisée

- On connait un motif
- On a une forme inconnue
- On la rattache au motif le plus proche

Exemples

- reconnaissance des caractères
- assignation à une espèce connue

Non supervisée

- On a un nuage de points
- On recherche des clusters

Exemples

- Classification ascendante hiérarchique
- Construire des délimitations d'"espèces" (OTU)

Questions (recherche pour $n \gg 10^3$)

On se donne

- un ensemble de séquences $\{s_i \ ; \ 1 \leq i \leq n\}$
- un tableau de distances deux à deux $D = [d_{ij}]_{i,j=1,\ldots,n}$; $d_{ij} = d(s_i, s_j)$

Question 1

Existe-t-il une dimension $d \in \mathbb{N}$ et un nuage de points
$\mathscr{X} = \{x_1, \ldots, x_n\}, \quad x_i \in \mathbb{R}^d$
tel que $\quad \forall\, i, j, \quad d(x_i, x_j) = d_{ij}$?

Question 2

Si oui, quelle est la forme de ce nuage ?
Elle est une **caractérisation de la biodiversité** (plus qu'un simple indice)
Réponse de la biologie évolutive : un ensemble hiérarchisé de clusters
fondée sur un processus : spéciation et extinction, dérive et sélection ...

Calcul des distances

Complexité du calcul

- il existe un algorithme de calcul (Smith-Waterman) de la distance d'édition entre deux mots
- la complexité du calcul

$$\texttt{disseq} : (w_1, w_2) \longrightarrow d(w_1, w_2)$$

  est en $O(\ell_1 \ell_2)$
- si n séquences, $n(n-1)/2$ distances à calculer
- complexité globale en $O(n^2 \ell^2)$
- si $n = 10^5$ et $\ell = 300bp$, $\kappa \simeq 10^{15}$

Parallélisation du calcul des distances (1/2)

Mise au point dans le cadre du projet e-biothon, sur Babel ; collaboration IDRIS

Projet e-biothon

- le calcul de la matrice D peut se faire en `map` (calcul de chaque valeur `D[i,j]`) / `reduce` (format matriciel)
- peut donc se paralléliser par les données sur un très grand nombre de CPU
- une machine hyperparallèle avec un très grand nombre de CPU est idéale
- projet pilote soumis à l'appel d'offre e-Biothon (machine IBM Blue Gene P, Babel)

Parallélisation du calcul des distances (2/2)

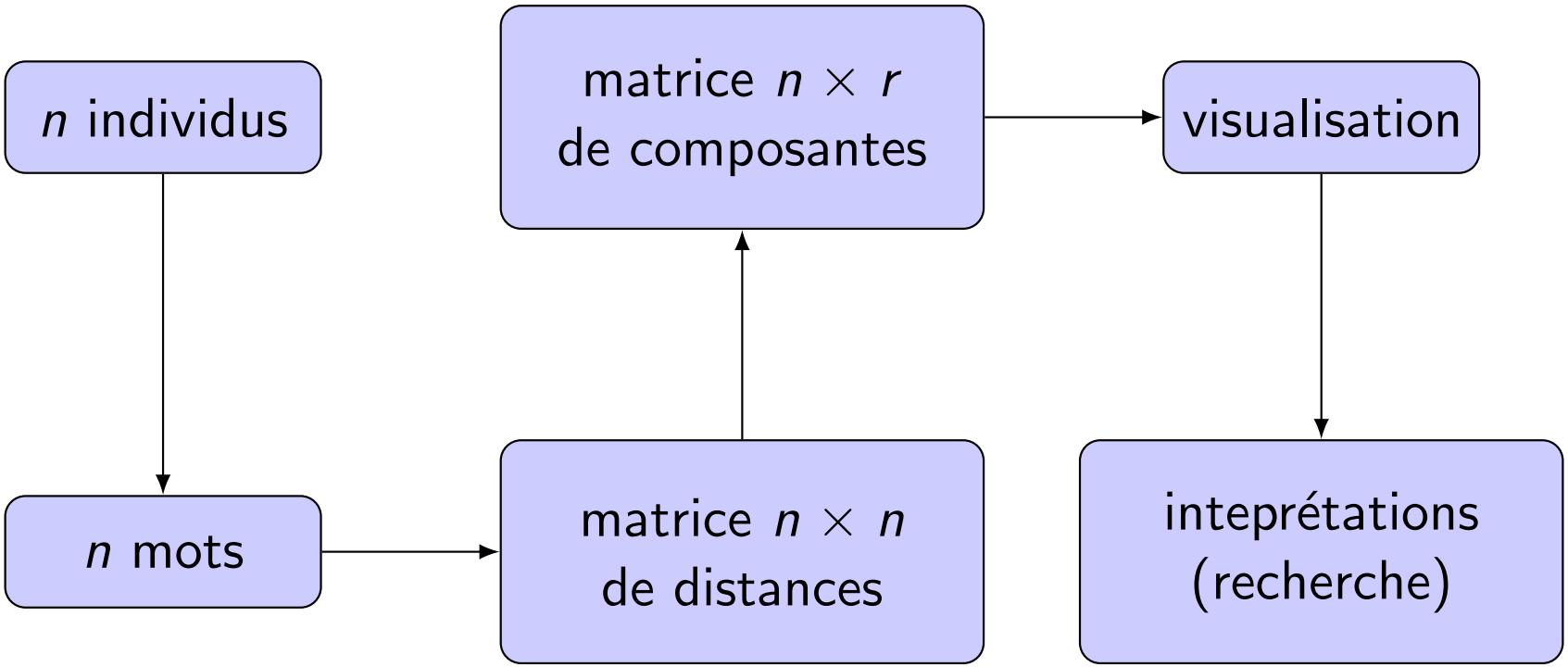
Mise au point dans le cadre du projet e-biothon, sur Babel ; collaboration IDRIS

Mise en œuvre

- insertion de `disseq` dans une procédure `MPI`
- scalabilité parfaite sur un très grand nombre de nœuds
- puis passage (même programme) sur le Blue Gene Q (Turing), avec $2^{14} = 16\,384$ CPU, dans la cadre d'un projet DARI (2015-2016)
- des milliers de matrices calculées, post-traitement en cours

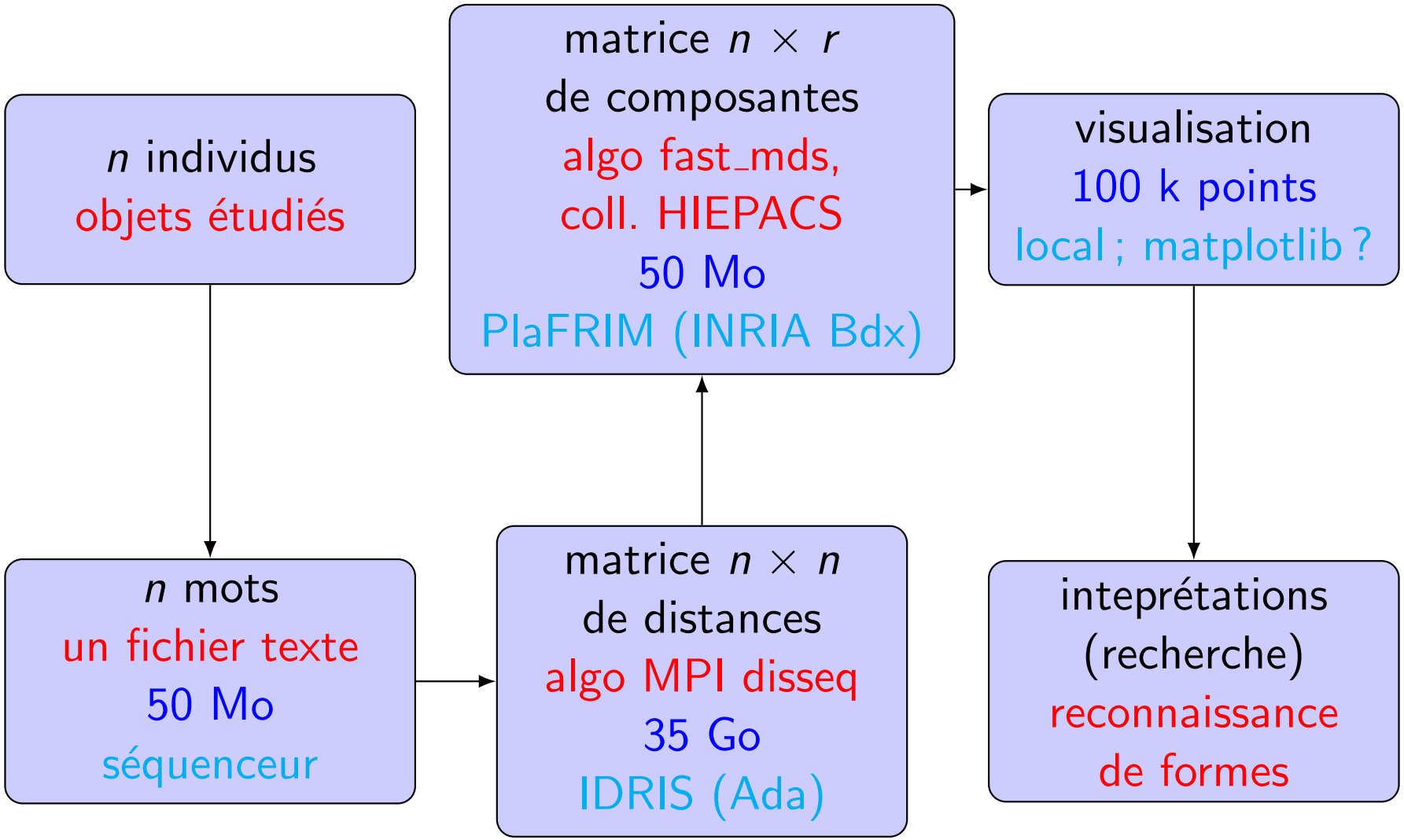
Le parcours d'une donnée ...

A réaliser p fois, pour p expériences, avec $p \gg 1$



Les formats des données ($n \approx 10^5$)

rappel : à réaliser p fois, pour p expériences, avec $p \gg 1$



Interopérabilité

Besoin d'interopérabilité entre

- le lieu de production des données (plateforme de séquençage)
- le centre où le calcul intensif est efficace (machine hyperparallèle Turing, IDRIS)
- le centre où se développe l'algorithme fast_mds (INRIA Bdx, évolution de l'algorithme MDS pour passage à l'échelle, PlaFRIM)
- le site de visualisation des données (nuage de points)

Actuellement

a la mano, fichier par fichier :

- `ssh + tar.bz2 + scp + ...`

Besoin

Processus à "industrialiser" et livrer comme un service

Une boucle IRODS pour gérer les flux/stockage de données

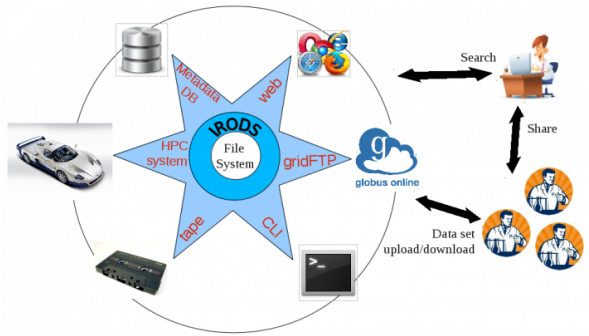
Ce qu'est une boucle iRods

- Un système de gestion distribuée de fichiers
- via un serveur en lien avec une ribambelle d'éléments de stockage
- une partie générique gérée par le logiciel
- une partie modulable spécifique à chaque communauté (interfaces, politiques, procédures)

Utilisation d'iRods dans le projet

Les services de base assurés par iRods

- poser un fichier
- récupérer un fichier
- partager un ensemble de fichiers au sein d'une communauté
- rechercher/archiver/ ... : cycle de la donnée



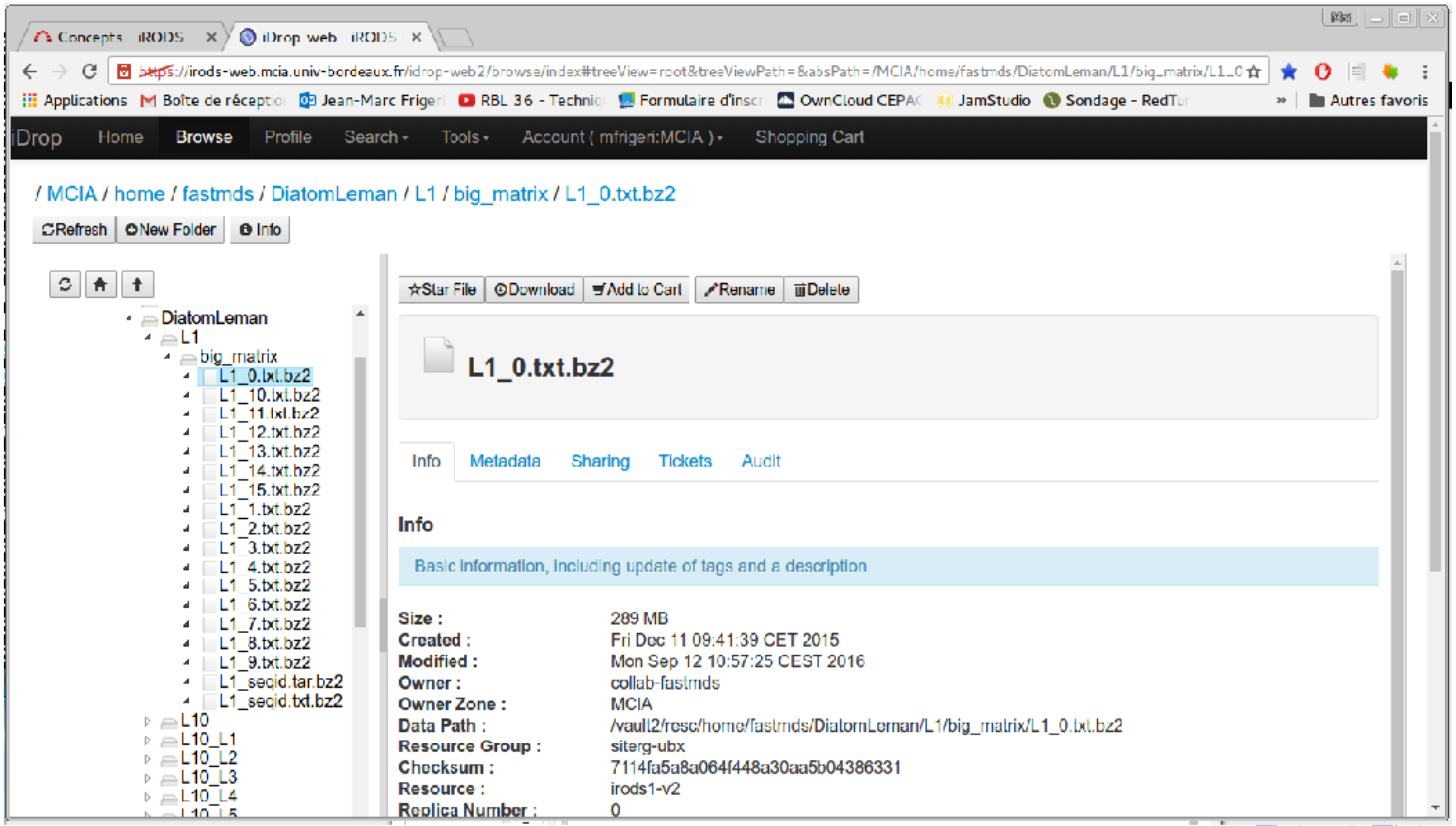
interopérabilité entre centres de calculs (transfert de fichiers)

- iRods comme "gare de triage" des fichiers (pull/push)
- écrire un script en utilisant l'API iRods (développement en cours)

Mise en œuvre

Utilisation de la boucle iRods du MCIA (mésocentre aquitain)

https://irods-web.mcia.univ-bordeaux.fr/idrop-web2/browse/index#treeView=path&treeViewPath=/&absPath=/&browseOptionVal=info



Points de sécurité abordés

Sommes toutes assez standards ...

Dialogue entre responsables des machines connectées

- liste des machines autorisées
- liste des ports autorisés

Authentification

- login
- mot de passe

Remarque

Concernant les flux de données, on a choisi l'option où ADA est connectée avec la boucle IRODS en sortie uniquement
le service peut évoluer en entrées/sorties
mais avec d'autres protocoles de sécurité

Perspective

Vers un service distribué ... ?

Un catalogue de service sur chaque site de calcul là où il est efficace

- calcul des distances : machine hyperparallèle IDRIS ($2^{14} = 16\,384$) cœurs), Turing
- calcul des coordonnées de la MDS : PlaFRIM, librairie dédiée `fast_mds`, versions python et C^{++}
- visualisation : à développer
- grille et cloud EGI : en développement

Une boucle irods associant ces nœuds pour

- le stockage / partage des fichiers
- la gestion des données (duplication, sécurité)
- l'automatisation et optimisation du cycle de calcul (scripts)

Perspective

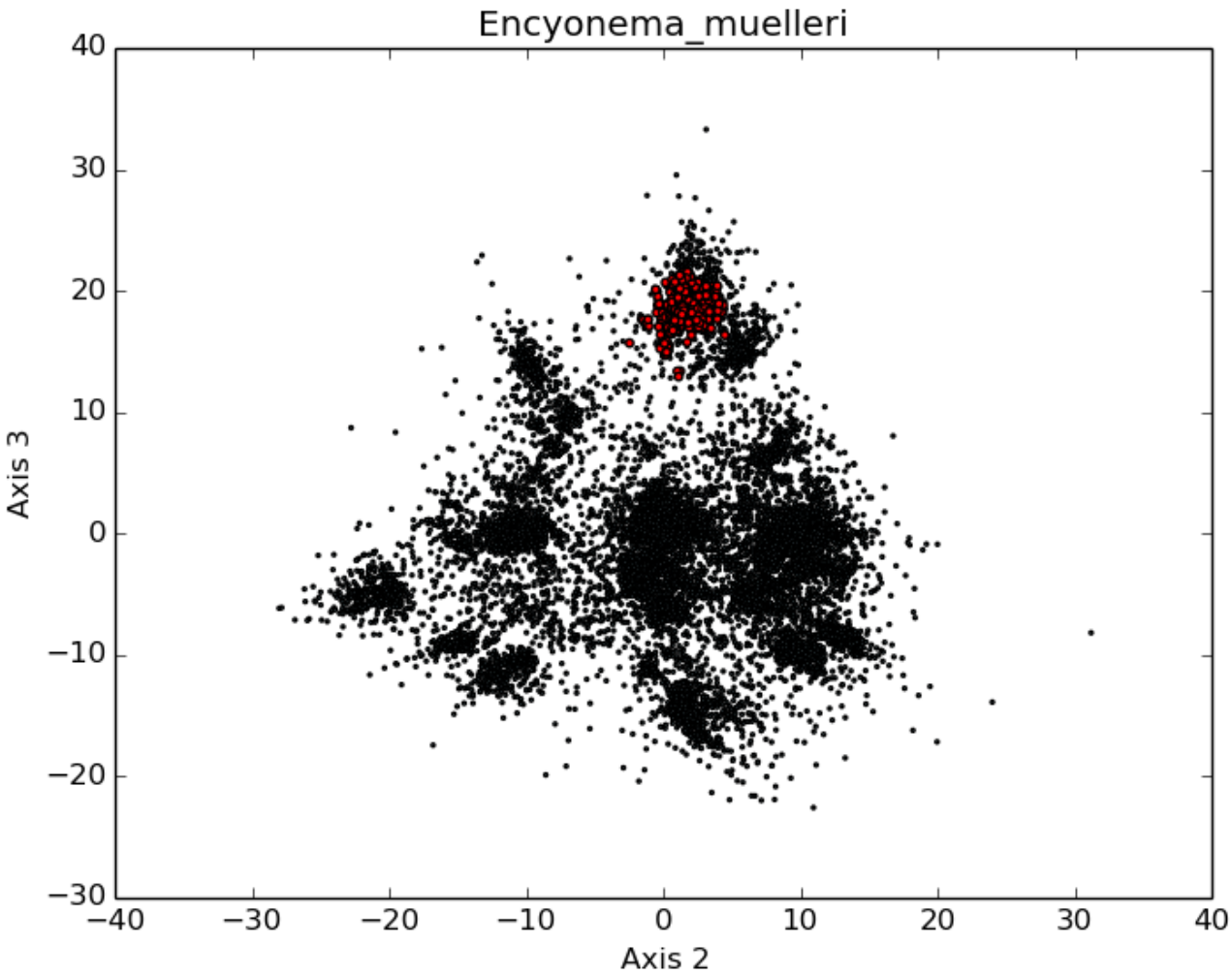
Etat des lieux

Etat des lieux (contribution au projet pilote EOSC, workpackage piloté par IN2P3)

- Collaboration BioGeCo/Pleiade, IDRIS (Sylvie Thérond), PlaFRIM (O. Coulaud, F. Rue), MCIA (P. Gay)
- Est en cours de réalisation : interopérabilité ADA (IDRIS) - PlaFRIM (INRIA Bdx) - laptops
- Reste à développer : interopérabilité avec la grille et le cloud EGI (en dévelopement, coll. J. Pansanel, & I. Blanquer, UPV)
- Ouverture à tout élargissement de la collaboration

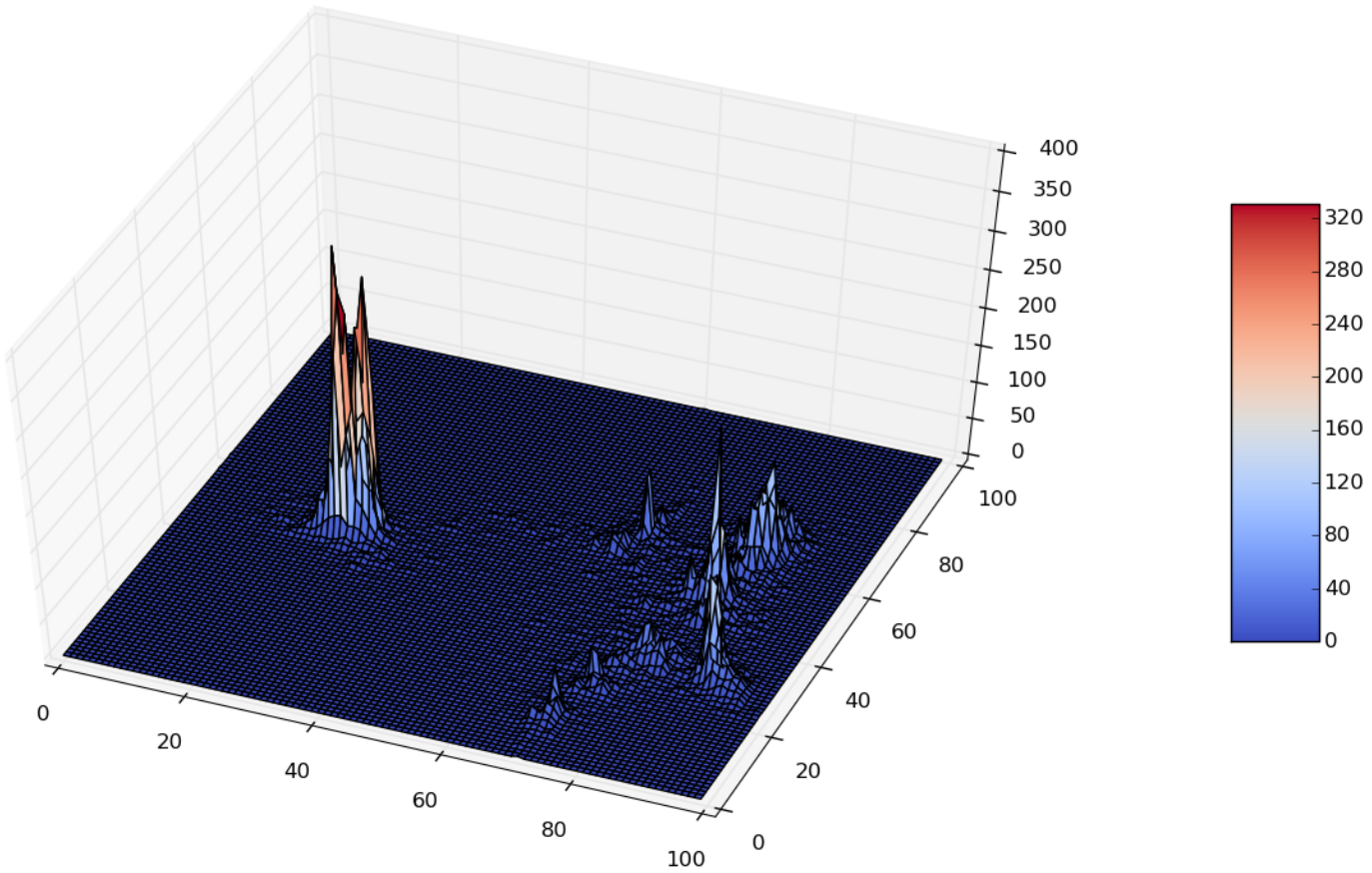
MDS sur un métagénome complet (ou presque)

Diatomées d'eau douce, un échantillon du Lac Léman



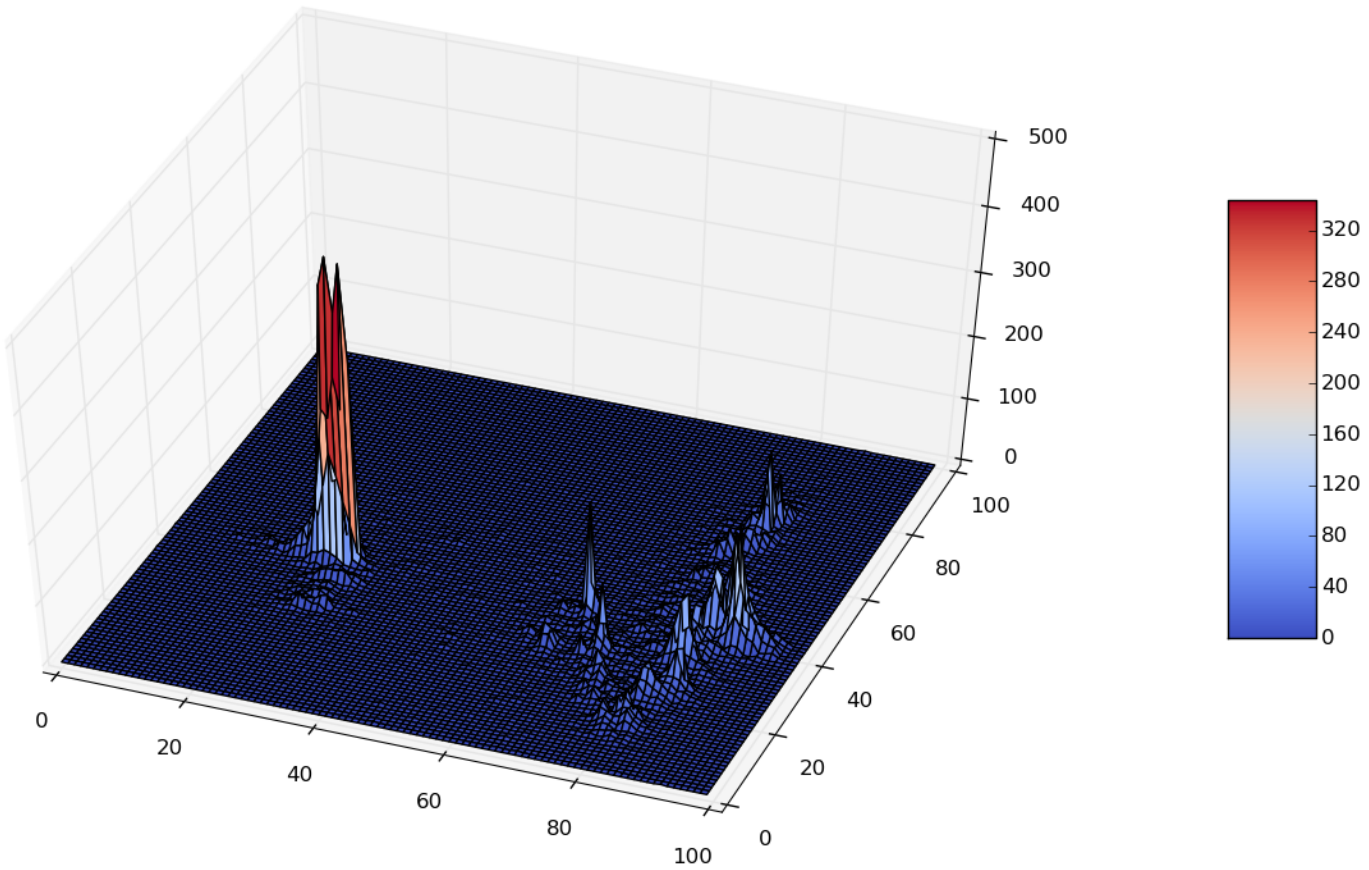
MDS sur un métagénome complet (ou presque)

Diatomées d'eau douce, un échantillon du Lac Léman



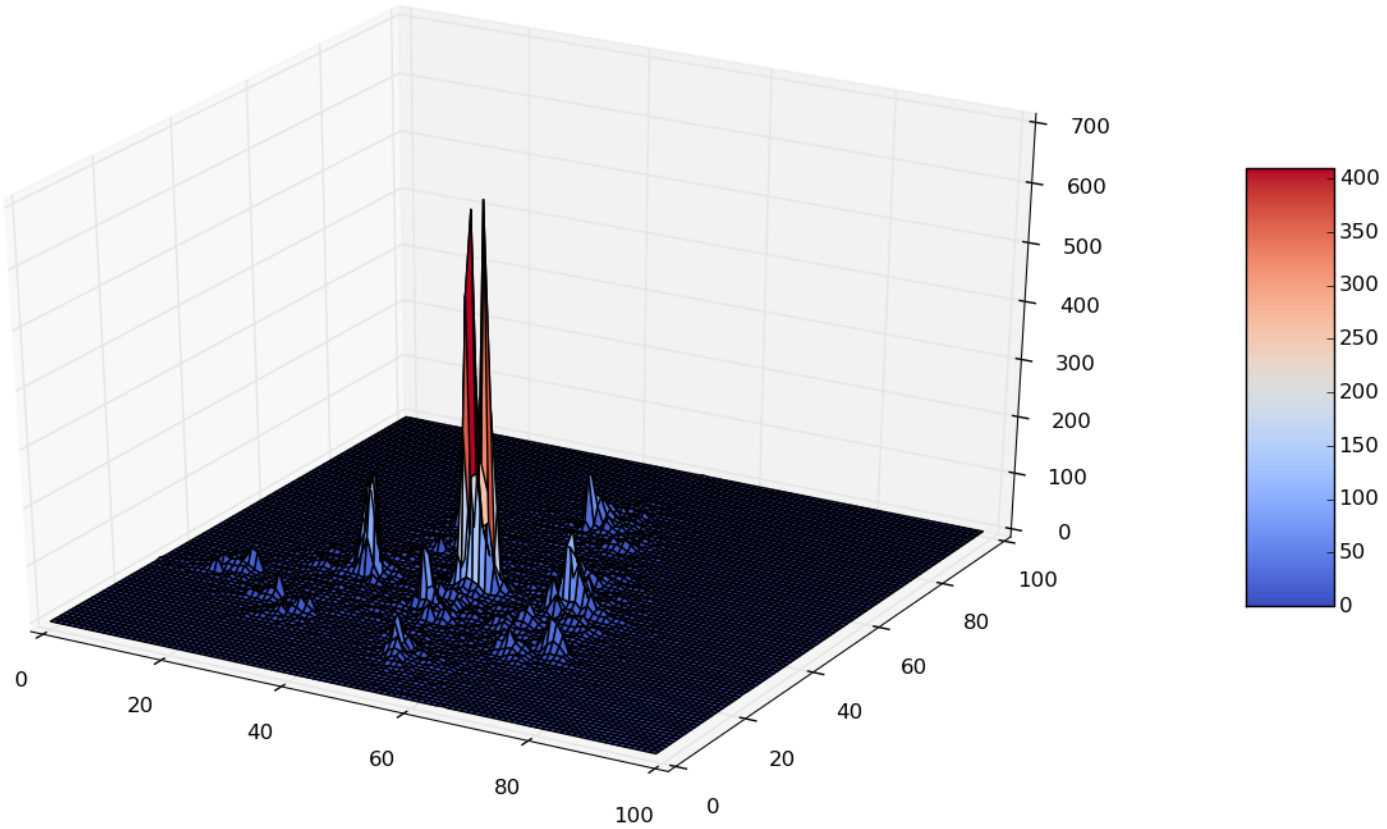
MDS sur un métagénome complet (ou presque)

Diatomées d'eau douce, un échantillon du Lac Léman



MDS sur un métagénome complet (ou presque)

Diatomées d'eau douce, un échantillon du Lac Léman



Remerciements

- Vincent Breton (FG, e-biothon)
- Le projet e-biothon (pied à l'étrier pour le calcul intensif, comme on dit ...)
- Le labex CEBA (diversité des forêt guyanaises)
- Notamment le groupe de l'herbier de Cayenne : Sophie Gonzalez, Jean-François Molino, Daniel Sabatier
- Groupe de Thonon et Uppsala (diversité des diatomées) : Frédéric Rimet, Agnès Bouchez, Maria Kahlert