Pourquoi migrer vers une architecture orientée ‘Big Data’ en bioinformatique pour la génomique ?

jerome.gouzy@inrae.fr

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

INRAE

lipme

cnrs
dépasser les frontières

Bioinformatique au LIPME

Principalement orientée vers l'analyse des séquences, des génomes

- **Développement & Intégration d'interfaces et d'outils d'analyse pour données du sens aux données de séquences**
  - Assemblage (2000-): reconstituer les séquences chromosomiques à partir de fragments d'ADN produit par les robots de séquençage ➔ génération d'un génome
  - Annotation (1994-): localiser les gènes dans le génome (coll. T. Schiex MIAT) puis prédire les fonctions possibles des protéines codées par ces gènes
  - Mesure de l'expression des gènes (2001-): identifier les gènes importants dans un processus biologique (activation, répression)
  - Etc.
- **Des projets de génomique, internes et externes au LIPME, sur de nombreuses espèces**
  - Bactéries: taille génome < 10Mb
  - Champignons: < 100Mb
  - Plantes: 100Mb-15Gb (pic à 3Gb=tournesol cultivé)
- **Une petite équipe**
  - 3 permanents + 0-3 Ingénieurs CDD|Stagiaires
- **Un cluster de calcul linux depuis 2000 (16 CPU vs 1200HT aujourd'hui)**
  - Nos besoins (calcul/stockage) ne sont pas couverts par les PF "ISO" (coûts, limites, techno)
  - Liberté/réactivité par rapport aux évolutions technologiques et aux architectures

Evolution des projets de génomiques 1/2

D'une génomique « statique » à une génomique « dynamique »

- **Accélération continue de la production de génomes « qualité référence »**
  - 1998 – 2001: assemblage/annotation/publication de notre 1er génome de bactérie (6Mb) – consortium international
  - 2003-2011: assemblage/annotation/publication de notre 1er génome de plante (450Mb) – consortium international – draft
  - 2010-2017: assemblage/annotation/publication du génome de tournesol (3Gb = taille génome humain) – plusieurs tentatives/technos mais finale principalement LIPME
  - 2020-2021: assemblage/annotation de 15 génomes >= 3-5Gb (tournesols et apparentés)
  - 2022: assemblage et annotation de 20 génomes >= 3-16Gb
- **Génome de référence centrique ➔ espèce centrique**
  - À multiplier par le nombre d'espèces étudiées
  - ➔2019/20
    - Toutes les données étaient projetées sur LE génome de référence
  - 2020➔
    - Le choix de la référence va dépendre de la question
    - De nombreuses questions impliquent de travailler à l'échelle pan-génomique (diversité) … qui dépend des génomes disponibles … et dont le nombre augmente toujours

Evolution des projets de génomiques 2/2

Réutilisation des données publiques de plus en plus importantes

- **Avant: 1 dataset original = 1 publi**
- **Désormais**
  - Il devient nécessaire d'intégrer beaucoup plus les données publiques de diversité et les résultats publiés pour proposer une vision à l'échelle de l'espèce
  - Le passage à l'échelle de l'espèce implique de collecter le maximum d'information sur les caractéristiques phénotypiques (propriétés) des « individus » séquencés
    - ➔curation couteuse des meta-données de la bibliographie: formats hétérogènes, alias, typo, etc.
    - ➔Doit être maintenu, partagé .. ou refait

Est-ce que la bioinformatique des 15 prochaines années va ressembler à celle des 15 précédentes ?

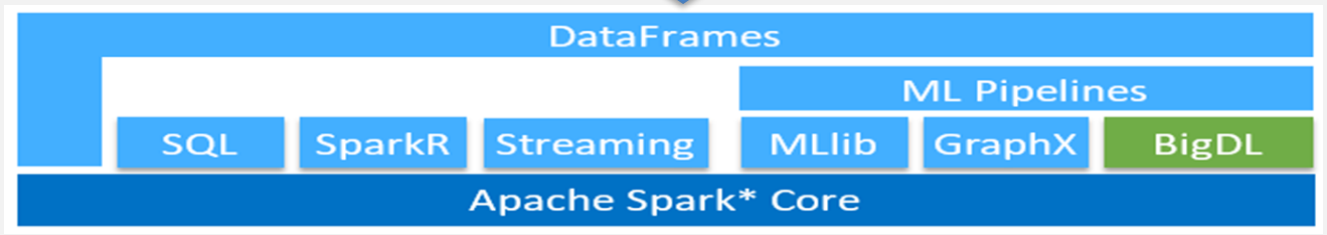
- **Si on pense que OUI**
  - Il ne faut rien changer
  - Il faut continuer à travailler sur des projets indépendants les uns des autres
  - Être prêt à faire et refaire les téléchargements et curation de (meta)données
- **Si on pense que NON**
  - Il faut adapter notre architecture
  - Il faut revisiter notre mode de gestion de projets
    - Toutes les (meta)données de l'espèce d'intérêt, publiques et privées, sont toujours accessibles, requettables, analysables
    - Il en va de même des données curées et des résultats des analyses (données d'analyses ultérieures/intégratives)
    - Un projet n'est plus "standalone", il doit alimenter et utiliser de façon normalisé un repository de (meta)données de l'espèce

Architecture cible

SUNRISE
UNE CULTURE POUR LE FUTUR

Targeted architecture: (Apache Spark) a unified analytics engine for large-scale Sunflower data processing

?



K

INSTITUT CARNOT Plant2Pro

ICSG INTERNATIONAL CONSORTIUM ON SUNFLOWER GENOMICS

SOFIPROTEOL L'engagement durable

ANR · Investissements d'Avenir · GIS bV biotechnologies vertes · syngenta · innolea · INRA SCIENCE & IMPACT · Terres Inovia l'agronomie en mouvement · SOLTIS · RAGT · LEREPS · MAÏSADOUR semences · UPMC SORBONNE UNIVERSITÉS · CAUSSADE semences

Sunflower genomics for the decade

2. Unified infrastructure: analytic engine, data and metadata

Spark ecosystem (storage, computation nodes, algorithms):
today on a « small » spark cluster (50Tb hadoopfs; 608HT; 4Tb RAM) but cloud ready

- MLLIB Machine Learning
- BigDL (deep learning)
- ATLAS in progress
- GWAS to do

« Data Lake »: data in native formats

- Reference genomes (raw, indexed)
- Multiple Sequence Alignments (txt, png)
- Passport data (csv, xml, pdf, …)
- Phenotypic data/metadata (csv, pdf, jpeg,…)

Pourquoi Spark ? 1/2

- **Projet Apache vivant, publication convaincante**
  - Zaharia M. *et al.* 2012 Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing
- **Utilisé dans l'industrie**
  - intéressant pour valoriser l'expérience des CDD
- **Une première analyse d'application encourageante - Spark-ICS (2016-2018)**
  - Exploration de l'application à la bioinformatique de l'architecture « Big Data » Apache Spark

  http://lipm-bioinfo.toulouse.inrae.fr/download/SPARK-ICS/INRA-SPARK-ICS-20180907-1.1.pdf
- **An "Unified Engine for large-scale data analytics"**
  - Un même framework avec de nombreuses méthodes d'analyses de données
    ➔ on peut basculer de l'une à l'autre
    - sans redévelopper les I/O
    - sans se soucier de la parallélisation des méthodes

Pourquoi Spark ? 2/2

- **Scalable et portable**
  - poste de travail (développement) → cluster (exécution) → cloud (diffusion)
- **Multi languages (scala, JAVA, python, R)**
  - Briques d'analyses bioinformatiques de bas niveau en scala (binding C)
  - Ouverture vers un spectre plus large de développeurs-utilisateurs de l'architecture
- **Couplage facile avec des données structurées dans un Data Lake**
  - Éviter les dump/load qui consomment du temps, dupliquent le stockage, et sont sources d'erreurs et de maintenance
- **Fonctionnalités d'analyses automatisées et en quasi temps réel (mon rêve ultime !)**

Les briques nécessaires

- **Identifier les méthodes d'analyses de séquences « élémentaires »**
  - extract, translate, alignement de deux séquences, alignement de plusieurs séquences, intersect, merge, etc..
- **Développer une API « bioinformatique » pour Spark**
  - En scala et/ou via du binding de programmes C/C++
- **Concevoir un Data Lake pour la génomique**
  - Pour structurer au maximum le cycle de vie des (meta-)données/résultats
  - Fournir un environnement d'analyse normalisé aux futurs « data analysts »
- **Développer nos compétences en Machine Learning**
  - Pour extraire des connaissances
  - Pour prendre des décisions automatiquement (ex: sur le paramétrage)
- **Développer nos futurs programmes d'analyse dans cet environnement**
  - Utilisant l'API bioinfo spark
  - Exploitant/alimentant le Data Lake
  - Dans un des langages compatibles Spark (préférence scala)

Etat d'avancement

- ***cf* le retour d'expérience de mercredi matin !**

BIOINFO@LIPME

- **Léo Géré**
- Javad Razavi
- Mael Chiotti
- Axel Verdier

- **Sébastien Carrère**
- Ludovic Cottret ➔ TOXALIM
- Erika Sallet ➔ DISPO
- **Ludovic Legrand**
- **Jérôme Gouzy**
- **+ poste IE ~~mobilité 2022,~~ concours externe 2023**
  - profil technique/méthodo = ce topo!