DATA

hackA tech

Shake science. Shape innovation.

#data
#génome
#biologie
#bioinfo

GATB
Optimiser l'analyse des données génomiques

Inria

CARACTÉRISTIQUES

GATB *(Genome Analysis Toolbox with de-Bruijn graph)* est une solution logicielle liée aux traitements des données génomiques. Elle permet d'analyser des données issues de n'importe quel organisme tel que les bactéries, les plantes, les animaux, ou encore des échantillons complexes (ex. métagénomes).

Pour cela, GATB conçoit des algorithmes qui optimisent les temps d'exécution, qui minimisent l'espace mémoire et qui se déploient sur des calculateurs parallèles.

L'idée est d'exploiter la redondance de séquences des génomes.

Cette technologie assemble de nouveaux génomes dans le but de comparer des organismes spécifiques pour représenter des espèces de référence, ou pour mettre en avant des variations génomiques qui révèlent des propriétés liées aux impacts écologique, agronomique ou clinique.

© Leigh Prather - Fotolia

TRAITEMENT DES DONNÉES

Les données traitées sont des données génomiques en masse. Ce sont généralement des données publiques, basées sur des recherches Académiques. Elles sont bruitées et par conséquent elles contiennent des erreurs.

QUELS AVANTAGES ?

- Optimisation du temps d'exécution
- Minimisation de l'espace mémoire grâce à la structure de données de GATB
- Lecture de données de taille conséquente (grande échelle) grâce à l'application sur des PC multicore.

USE CASES

- **Biologie :** détection de polymorphisme nucléotidique (sNPs), qui est la variation d'une seule paire de bases du génome, entre individus d'une même espèce.
- **Santé :** recherche de marqueurs sur une maladie donnée
- **Écologie :** recherche de marqueurs pour une étude de plante ou pour retrouver des espèces.
- **Agronomie**

FICHE IDENTITÉ

- Licence : Affero GPL version 3
- Langage de programmation : : C++
- OS : : Linux & MacOSX
- Propriété intellectuelle : Inria / IRISA
- Équipe projet : GENSCALE

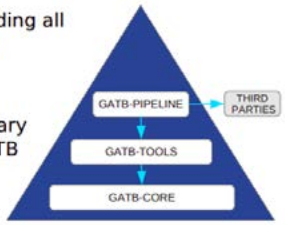
FONCTIONNALITÉS GÉNÉRIQUES

GATB-CORE permet une compilation immédiate (binaire).

- The **GATB** philosophy: a 3-layer construction to analyze NGS datasets

1. **GATB-CORE**: a C++ library holding all the services needed for developing software dedicated to NGS data

2. **GATB-TOOLS**: a set of elementary NGS tools mainly built upon the GATB library (k-mer counter, contiger, scaffolder, variant detection, etc.)

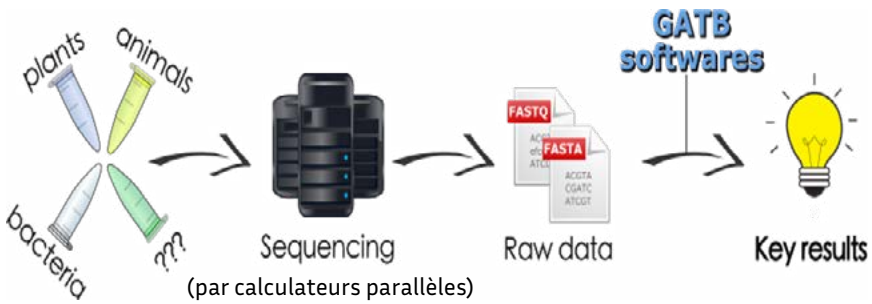


3. **GATB-PIPELINE**: a set of NGS pipeline that links together tools from the previous layer

Pour trouver des marqueurs biologiques, il faut filtrer puis corriger les erreurs.

GATB-CORE transforme un jeu de données d'entrée en un graphe compact de Bruijn sous le format HDF5. Ce graphe peut être utilisé par les outils développés dans la bibliothèque GATB CORE C++.

Des bibliothèques sont disponibles pour des opérations génériques sur des applications de code utilisateur, mais pas sur des séquences génomiques spécifiques.



(par calculateurs parallèles)

CONNAISSANCES MINIMUM REQUISES

- **C++**
- **Connaissances biologiques et génomiques**

READ ME

https://gatb.inria.fr/gatb-programming-tutorial/

Référent : Pierre Peterlongo
* GENSCALE est une équipe-projet commune à Inria, au CNRS, ENS Rennes et Université de Rennes 1.

© P.Peterlongo