Traitement de données bioinformatiques massives (Big Data)

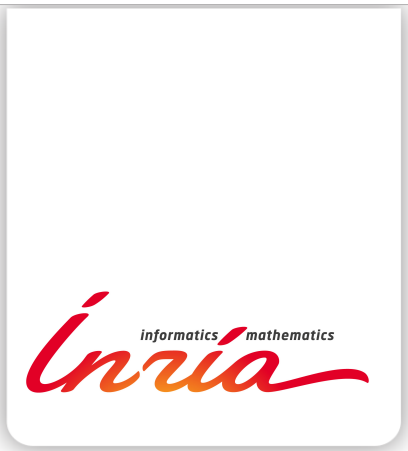
Sarah Cohen-Boulakia, Patrick Valduriez

Traitement de données bioinformatiques massives (Big Data)

Sarah Cohen-Boulakia, Patrick Valduriez

RESEARCH REPORT
N° 8915
May 2016
Project-Teams ZENITH

ISSN 0249-6399 ISRN INRIA/RR--8915--FR+ENG



Traitement de données bioinformatiques massives (Big Data)

Sarah Cohen-Boulakia [*] [†] [‡], Patrick Valduriez [§] [¶] [‡]

Équipes-Projets ZENITH

Rapport de recherche n° 8915 — May 2016 — 8 pages

Résumé : Les volumes des données bioinformatiques disponibles sur le Web sont en constante augmentation. L'accès et l'exploitation conjointe de ces données très réparties (*i.e.*, disponibles dans des sources de données distribuées sur le Web) et fortement hétérogènes (sous forme textuelle ou sous forme de fichiers tabulés, incluant ou non des images, décrites avec différents niveaux de détails et de qualité...), est essentielle pour que les connaissances en biologie puissent progresser. L'objectif de ce rapport est de présenter de façon simple les problèmes posés par l'utilisation conjointe des données bioinformatiques.

Mots-clés : Big Data, données bioinformatiques, recherche d'informations.

[*] Laboratoire de Recherche en Informatique, CNRS UMR 8623
[†] Université Paris-Saclay, Orsay, France
[‡] Institut de Biologie Computationnelle, Montpellier, France
[§] INRIA
[¶] LIRMM

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Bioinformatics big data processing

Abstract: The volumes of bioinformatics data available on the Web are constantly increasing. Access and joint exploitation of these highly distributed data (*i.e*, available in distributed Web data sources) and highly heterogeneous (in text or tabulated files including images, in different formats, described with different levels of detail and different levels of quality ...) is essential for the biological knowledge to progress. The purpose of this short report is to present in a simple way the problems of the joint use of bioinformatics data.

Key-words: Big data, bioinformatics data, information retrieval.

Table des matières

1 Introduction

La compréhension des mécanismes du vivant — la connaissance des mécanismes régissant l'activité de la cellule, la détermination du rôle fonctionnel d'un groupe de protéines ou encore la mise en évidence d'un ensemble de gènes liés à une maladie — dépend étroitement des avancées de domaines multiples : biologie, chimie, physique, électronique, mathématiques et informatique.

Depuis le début des années 90, de nouvelles technologies ont vu le jour comme les techniques d'analyse haut débit. Ces technologies génèrent un nombre extrêmement important de données. Le séquençage de génomes, c'est-à-dire le fait d'identifier les suites de nucléotides (ADN) faisant partie de notre génome, a connu récemment une véritable révolution technologique. Alors que 12 ans ont été nécessaires (lors du "Human Genome Project") pour séquencer le premier génome humain en impliquant des centaines de laboratoires et pour un coût estimé à 10,000 dollars par Megabase [1], les techniques de séquençage permettent en 2016 à une même machine de séquencer 200 génomes humains en une semaine avec un coût de 0,03 dollars par Megabase. Depuis le début des années 2010 de très nombreux laboratoires possèdent ce type de machine de séquençage appelé *séquenceur haut débit* comme indiqué en Figure 1. En conséquence, entre 2010 et 2016, les volumes de données de séquençage générées ont doublés tous les 5 mois.

De très gros volumes de données brutes peuvent aussi être générés dans d'autres contextes émergeants de la bioinformatique. Des plateformes de phenotypage de plantes ont notamment très récemment vu le jour. Elles permettent d'étudier la réaction de différentes *variantes* de génomes (appelés genotypes) de plantes face à diverses conditions environnementales (sècheresse, inondation, présence d'eau de mer dans l'irrigation etc.). L'objectif des expériences menées sur ces plateformes est notamment de mieux sélectionner les futures graines à planter dans des régions aux conditions environnementales particulières pour augmenter le rendement des futurs champs. Les plantes qui sont cultivées dans ces plateformes sont prises en photo sous différents angles chaque jour. Ces photos sont utilisées pour créer des avatars des plantes (leur representation virtuelle) pour étudier et modéliser leur croissance. Des masses de données brutes sont donc générées (plusieurs dizaines de terabytes par an) auxquelles s'ajoutent toutes les données renvoyées par les multiples capteurs de la plateforme (volume exact d'eau versé dans chaque pot, exposition/lumière...).

2 Caractéristiques des données biologiques

L'ensemble des données brutes et des résultats de leurs analyses bioinformatiques sont stockés dans des bases de données biologiques, disponibles (le plus souvent) sur le Web. Le nombre et le contenu de ces bases croissent de façon considérable. Plus de 1 500 bases de données sont recensées et réparties sur le réseau du Web en 2016. La Figure 3 donne un aperçu du réseau formé par ces bases sur le Web, en ne considérant qu'un fragment des bases existantes.

Le contenu de ces bases de données est très hétérogène. Chaque base de données a son propre format de données (un fichier tabulé, un document texte, une image...) et sa propre structure (choix des colonnes d'un fichier tabulé). Mais l'hétérogénéité des données n'est pas uniquement liée à la forme des données. Elle est surtout liée au fait que les données biologiques reflètent des expertises : elles contiennent du texte et des annotations et reflètent donc des avis d'experts qui ne sont pas toujours exprimés avec le même niveau de détail ni fondés sur les mêmes résultats experimentaux. Ces données biologiques ont donc des niveaux de qualité très variables. En conséquence, connaître la *provenance* d'une donnée (comment elle a été produite, par qui elle

1. La taille d'un génome correspond à la quantité d'ADN contenue dans un génome, mesurée en nombre de nucléotides (avec pour unité le Megabase, un million de nucléotides).

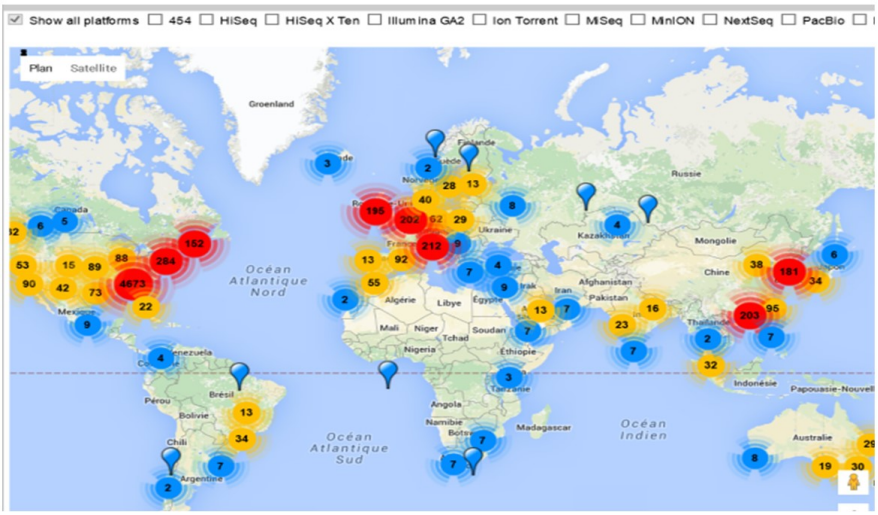


FIGURE 1 – **Cartographie des séquenceurs hauts débits recensés dans le monde (http ://omicsmaps.com).**

FIGURE 2 – **Exemple de la plateforme Phenoarch de l'INRA de Montpellier.**

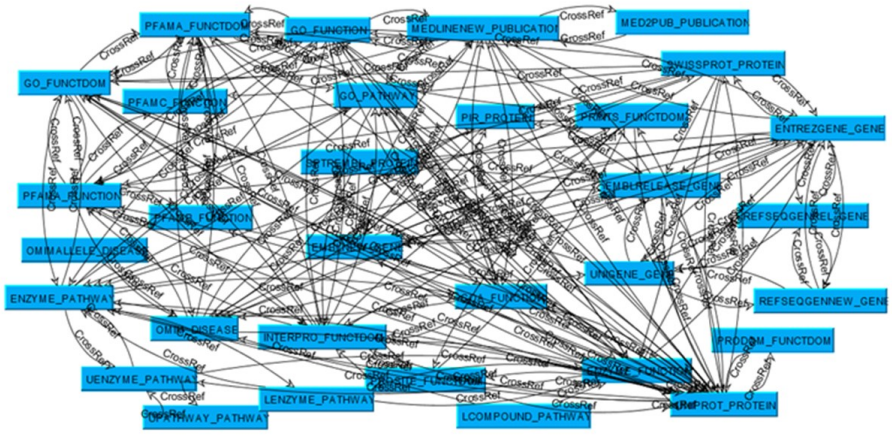


FIGURE 3 – **Fragment du réseau des bases de données bioinformatiques et liens (hypertexte) entre les données.**

a été annotée...) est une information cruciale pour permettre son interpretation correcte.

Les données biologiques sont donc au cœur du paradigme du *Big Data* caractérisé par les "4 V" (https ://fr.wikipedia.org/wiki/Big_data) : **V**olume (les données sont très nombreuses), **V**ariété (leur hétérogénéité est forte), **V**élocité (les données reposent sur des connaissances qui évoluent et changent donc beaucoup dans le temps), et **V**éracité (les données reposent sur des expertises et ont des niveaux de qualité variés).

Pouvoir interroger, comparer et rapprocher les données bioinformatiques est nécessaire pour que les connaissances en biologie et en médecine puissent progresser. Exploiter ce volume et cette diversité d'informations réparties, très fortement hétérogènes, et en constante évolution est un réel défi à relever. Le traitement des données bioinformatiques est au centre de cette problématique et fait l'objet de la prochaine section.

3 Traitement des données

Le passage des données brutes à l'acquisition de nouvelles connaissances biologiques se fait d'abord par le traitement des données brutes, l'analyse de leur contenu, le croisement de ces données avec d'autres données. De nombreux outils (logiciels) sont mis à disposition des bioinformaticiens pour analyser et traiter leurs données. Par exemple, des outils existent pour aider à déterminer les zones du génome (ADN) qui sont différentes entre les cellules saines et les cellules tumorales d'un même individu, d'autres outils peuvent ensuite permettre d'isoler les gènes présents sur ces zones, d'autres encore vont donner accès aux fonctions connues des gènes... Chaque outil permet d'effectuer une étape de traitement des données. Enchaîner ces étapes en combinant l'utilisation d'outils permet de passer des données à l'aquisition de nouvelles connaissances. Souvent, les bioinformaticiens développent eux-mêmes de nouveaux logiciels pour analyser au plus près leurs données. Il faut alors maintenir ces logiciels qui reposent sur des installations précises des machines qui peuvent évoluer (par exemple, changement de la version du système d'exploitation installé sur la machine) et rendre les logiciels inutilisables. En réponse à ces besoins,

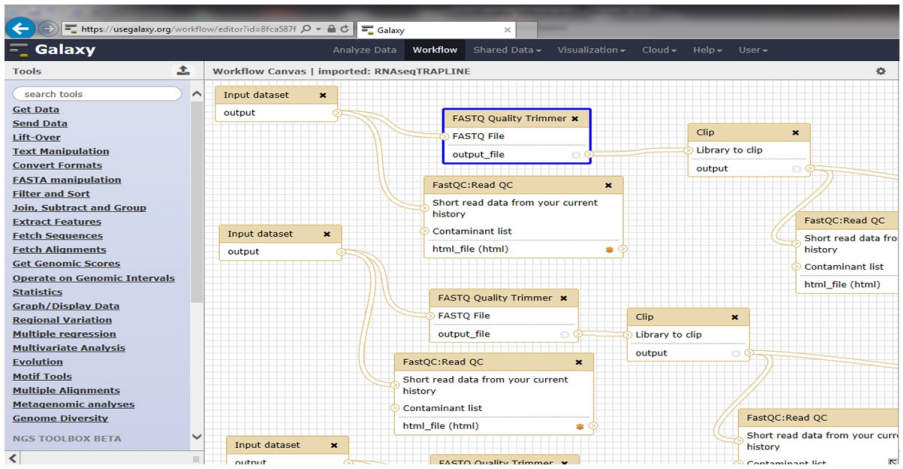


FIGURE 4 – **Interface principale du système Galaxy, à gauche un ensemble d'outils disponibles qui peuvent être déplacés au centre de l'interface et liés les uns aux autres pour former une chaîne de traitement.**

des systèmes spécifiques ont été conçus, appelé *systèmes de gestion de workflows scientifiques*. Ces systèmes offrent aux biologistes et bioinformaticiens une aide pour traiter leurs données. Dans ces systèmes, les biologistes ont accès à un ensemble d'outils (ceux qu'ils ont eux-même développés et ceux qui sont disponibles) qu'ils peuvent combiner pour concevoir des chaînes de traitement (en anglais, des *workflows*) pour analyser leurs données. Parmi les systèmes gratuits et libres (*open source*) les plus utilisés par la communauté bioinformatique, on citera Galaxy [2] et Taverna [3].

De façon très intéressante, les workflows peuvent être partagés et échangés entre utilisateurs. En d'autres termes, un biologiste peut partager avec ses collègues non plus seulement les données qu'il a générées mais aussi les traitements qu'il a effectués. Des portails de workflows scientifiques existent aujourd'hui et donnent accès à des milliers de workflows. C'est le cas de la plateforme myExperiment [4] qui héberge plus de 2 000 workflows scientifiques en 2016.

Parce qu'ils décrivent les outils exacts utilisés dans le traitement de données, les workflows jouent un rôle clé dans le fait d'assurer la reproductibilité des expériences bioinformatiques, enjeu scientifique majeur. Une série d'initiatives a notamment vu le jour (en particulier chez les grands éditeurs scientifiques comme Nature [5]) pour inciter les auteurs d'articles scientifiques à décrire précisément les données et les outils utilisés pour obtenir leurs résultats. L'utilisation de workflows répond donc directement à ce besoin.

2. galaxyproject.org
3. www.taverna.org.uk
4. www.myexperiment.org
5. nature.com/nature/focus/reproducibility

4 Conclusion

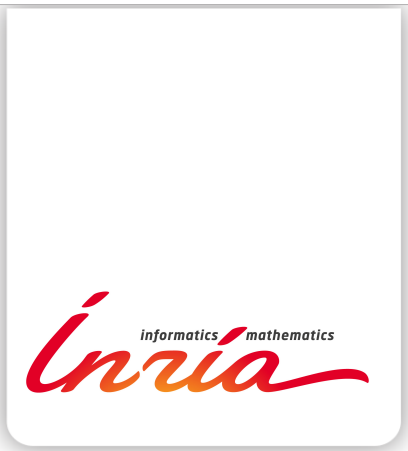
Les données bioinformatiques ont des caractéristiques particulières : elles sont regroupées dans des bases de données réparties, sont hautement hétérogènes, reflètent des expertises, et forment un réseau complexe sur le Web. L'avancée des connaissances en biologie dépend de l'exploitation de la richesse de ces informations fortement complémentaires. Exploiter conjointement un ensemble de données est donc une tâche particulièrement difficile puisqu'il faut faire face à des données possiblement contradictoires et très souvent redondantes. Le traitement de données biologiques par l'utilisation de workflows scientifiques joue un rôle important dans ce domaine puisqu'ils permettent l'échange et le partage des analyses de données et sont un atout fort pour la reproductibilité.

5 Glossaire - Définitions

Mégabase (megabase) : Unité de mesure pour évaluer la taille d'un génome, en milliers de nucléotides.
Workflow scientifique (scientific workflow) : représentation informatique d'une analyse de données (enchaînement d'outils d'analyse).

6 Pour en savoir plus

COHEN-BOULAKIA Sarah et VALDURIEZ Patrick : *Interrogation et gestion de données bioinformatiques pour la biologie moléculaire*, Techniques de l'ingénieur Analyse et mesure en biotechnologie, Editions T.I., 2016.



RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399