A platform to google experiments and boost genomic research

Anna Bernasconi Politecnico di Milano
Dipartimento di Elettronica, Informazione e Bioingegneria
via Ponzio 34/B, 20133, Milano, Italy
Email: anna.bernasconi@polimi.it

***Abstract*—**

I. Introduction

The genome, an organismâĂŹs complete set of genetic information, is the object of study of *genomics*, a quite young field of scientific research. It only began in the late 20^{th} century, originating from the basics on DNA discovered some decades earlier. Latest developments made in genomics research have brought interesting possibilities for applications in other fields, which promise many health and medical benefits.

One of the reasons for such brilliant diffusion of genomics research is the blooming of revolutionary technologies to sequence DNA. We are speaking about Next Generation Sequencing (NGS)—also known as High Throughput Sequencing—which generally describes modern sequencing technologies which allow to determine the exact sequence of nucleotides in a given DNA/RNA molecule at faster rates and lower costs than traditional sequencing techniques. Such speed-up is achieved by means of massively parallel sequencing which enables millions of nucleic acids fragments to be sequenced simultaneously. A single human genome, about 3 billion units of DNA in 23 thousands genes, can now be processed in just a single day and the data storage needed to represent it is around 200 Gigabytes.reference?

Enormous storage and computational infrastructures are needed to handle this massive amount of data which is being produces thanks to NGS. From a challenge for molecular biologists and genomics researchers, this has become a new challenge also for the the Bioinformatics and Applied Computer science community.

In particular, we aim to address these described manca qualche frase per dire effettivamente cosa fa l'informatica per questo campo matters. We are a group of young doctoral students and researchers at Politecnico di Milano, Italy, lead by Professor Stefano Ceri[1], being funded by "Data Driven Genomic Computing" (GeCo), an ambitious Advanced ERC grant 2016-2021, which is driven by the slogan: "data should express high-level properties of DNA regions and samples, high-level data management languages should express biological questions with simple, powerful, orthogonal abstractions."[2]

Computer Science can certainly make this part in life sciences. Projects that GeCo is currently following and supporting involve, for example, understanding the tridimensional organization of DNA and its implications, how the expression of genes and mutations/variations in the nucleotide sequence may provoke tumors, how new drugs can be engineered by joining more molecules in pharmacological networks. GeCo leverages strong computer science knowledge to build infrastructures dor genomic computing.

In particular, a fundamental aspect of this research is represented by the problem of providing infrastructures for storage of data and foster its interoperability. It must be said that nowadays data (even only the open data) are hosted/spread in a huge number of different sites around the world, usually managed by big research world-wide consortia (National Library of Medicine - National Institutes of Health (NLM-NIH []), European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI []), National Human Genome Research Institute (NHGRI []), ENCODE []...). Such organizations maintain the so called *repositories of genomic data*,

An effort that has investigated the problem of locating genomic data to download for research purposes is DNADigest, whose study is documented in [**?**] and in a blog[3].

The difference with their approach is that they try and locate data and provide a means to make datasets available. Instead we want to provide the way to import data inside our model and make it usable from within our platform

Why genomic computing? Technological revolution for DNA Sequencing Availability of huge repositories of open data It is now possible to explain how DNA inheritance and replication cause/influence many diseases, leading to personalized medicine Many biological and clinical problems need data exploration, retrieval and analyisis Genomic datasets are Âńbig dataÂż

The big data issue. Thousands of new genomic experiments are becoming available every day. A study, described in Figure 1, has shown how...

Story before. So the beautiful story that inspired our whole project starts with : NEXT GENERATION SEQUENCING, a revolutionary set of technologies that make reading DNA very fast and not as expensive as it used to be In this way a huge number of genomic datases have been made available How does it work specifically? PRIMARY: Sequencing machines perform the primary data analysis and produce raw

[1] curriculum
[2] http://www.bioinformatics.deib.polimi.it/geco/
[3] https://blog.repositive.io/

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

Fig. 1. Projected annual storage and computing needs of four domains of Big Data in 2025. The phases of acquisition, storage, analysis, and distribution represent all the data lifecycle.

datasets (a single human genome requires about 200 GB). SECONDARY: Computationally expensive pipelines, collectively regarded as secondary data analysis, are then applied to raw data for extracting signals from the genome (such as: mutations, expression levels, peaks of binding enrichment, chromatin states, etc.), thereby producing processed genomic data , which are much smaller in size. TERTIARY: Processed datasets are collected by worldwide consortia, such as TCGA (The Cancer Genome Atlas), ENCODE (the Encyclopedia of DNA Elements), âĂę Rough terms and sizes *Abstract Data* The human DNA sequence is a string of 3.2 billions of base pairs, encoding adenine (A), cytosine (C), guanine (G), and thymine (T); size = 800Mbyte. *Raw/Aligned Data* Data is produced as ÂńreadsÂż, overapping subportions of the genome, and then aligned to a reference genome, with empahsis on quality; size = 200GByte. *Processed Data:* But each of us has ÂńjustÂż 4.1M to 5M mutations, mostly single substitutions/insertions/deletions; size = 125Mbyte We work on process data!!

WHich problems can we solve? (analysing process data) Which cancer types can be explained by disregulation of the tri-dimensional structure of the genome? Which co-occurring (killer) mutations cause the death of a cell in given tumors? Which transcription factors (dimers) always occur together? How can we assign predominant functions to each portion of the genome?

Our repository system aims at invogliare people to analyse data using our data model and out querying language

Research questions Broader picture: Does the Bioinformatics community currently have a tool to locate interesting integrated data to solve biological/clinical questions?

Immediate scope: WouldnâĂŹt the GMQL user be very happy with an integrated, semantically rich, repository to locate the right datasets for his/her research questions?

We want to: -Focus on genomic open data; integration driven by a conceptual model -Conceptual modeling for driving the continuous process of metadata integration and for locating relevant datasets -Disclosure of the semantic properties of the sources; broader spectrum of sources covered, richer set of concepts provided -Integration of subsets of these sources together

This work is included inside the bigger framework of GECO, an ERC advanced grant which aims to...

How would "experiments googling" help genomics researchers?? NIH sponsors donation and participation in Genomics Research, by providing precise non-discriminating policies on how to get involved, how personal privacy is protected, how study eligibility works for everyone[4].

Paper organization.

II. The Genomic Conceptual Model

Data integration (or information integration) is a set of techniques that allow heterogeneous data, coming from different sources, having various structures, formats, unities of measure, languages, to be uniformed into one global agreed representation. Among other techniques, data integration can be based on the use of a *conceptual model*, a composition of concepts which abstract a knowledge area in a compact shape, to make it more understandable and organized.

III. The Integration Procedure

IV. The Search Platform

V. Application in the Research Field

Examples of biological problems

Given three replicas of a Chip.Seq experiment, extract high-confidence regions into one sample, identify which of these regions overlap with given genes, and for each resulting region count ICG mutations and select regions with at least one mutation.

3d structure and tumors

Same/cross gene activity correlations in normal vs tumor cells

Dimers: pairs of TFs that co-regulate genes in rigid and compact pairs.

Super-TADs: clusters of topological domains.

Killer Mutations: pairs of mutations: when both present they cause the death of the cell.

Identification of TFs that co-occur with TEAD4 binding sites.

Detect DNA areas where multiple TFs bind (dense TF binding regions).

DNA Sequencing of Microbioma in Cystic Fibrosis patients who are colonized with mycobacterium abscessus.

[4]NIH maintains a portal for this ckind of information: https://www.genome.gov/27561546/participating-in-genomics-research/

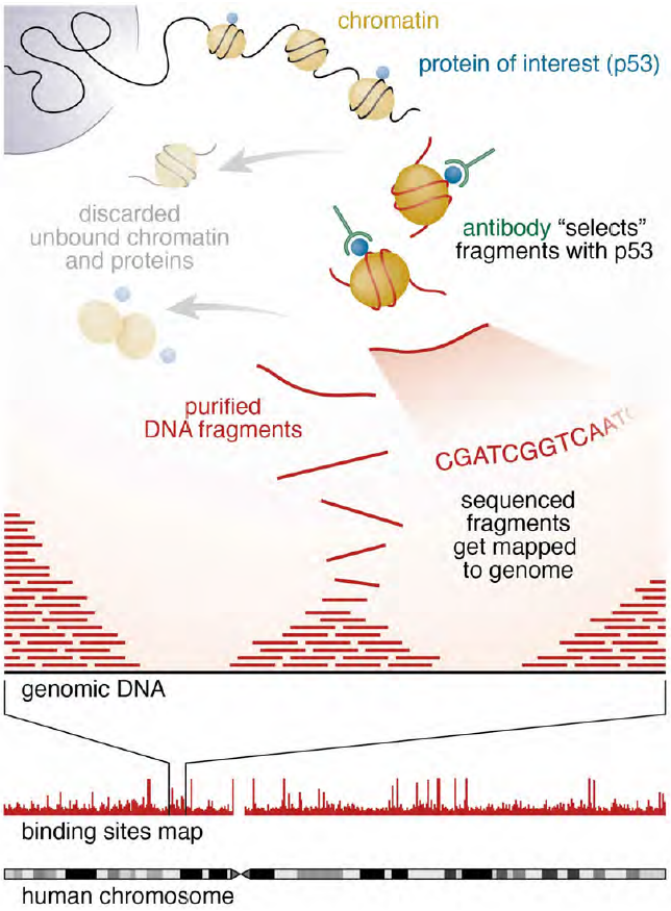

Fig. 2. ...

VI. Related works.

BioKleisli, K2, EnsMart/BioMart, BioStar ...provide integrated access to multiple, heterogeneous sources in the field of biomedical data;
Other works use conceptual models to explain biological entities and their interactions;
DeepBlue is an interesting starting point: hiding of datasource differences to provide easy-to-use interfaces;
Big consortia efforts: BioProject database (from NCBI, EBI and DDBJ), Encode DCC (Data Coordination Center), Tcga GDC (Genomic Data Commons).

VII. Future Work and Conclusions

We are delivering a repository, which will hopefully be appreciated by...
[**?**]
prendere da http://www.bioinformatics.deib.polimi.it/geco/?home e dal mio foglio di esame passaggio anno dottorato
Acknowledgment. This research is funded by the ERC Advanced Grant project GeCo (Data-Driven Genomic Computing), 2016-2021.