

Could you google this experiment?

A platform to 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

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 analysis Genomic datasets are big data

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 datasets have been made available How does it work specifically? PRIMARY: Sequencing machines perform the primary data analysis and produce raw 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), 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 reads, overlapping subportions of the genome, and then aligned to a reference genome, with emphasis on quality; size = 200GByte. *Processed Data*: But each of us has just 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 dysregulation 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

words on GECO project

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

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

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 data-source 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).

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...

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.

II. THE GENOMIC CONCEPTUAL MODEL

III. THE INTEGRATION PROCEDURE

IV. THE TOOL

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.

VI. 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.