Managing the Data Ecosystem in Life Sciences through Bioinformatics

Fotis E. Psomopoulos

Information Processing Laboratory, Dept. Electrical & Computer Engineering, Aristotle University of Thessaloniki (AUTH), GR-57001 Thessaloniki, Greece

Aristotle University of Thessaloniki

Background

Main research areas lay across three major pillars:

(a) Bioinformatics and the evolutionary analysis of genomes based on their protein data,

(b) Data mining methods and techniques in order to analyze the vast amount of data involved, and

(c) Grid and Cloud Computing in order to ensure the necessary computational power and to optimize the developed methods for large-scale approaches.

Particular focus is on the intersection of Bioinformatics and Grid / Cloud Computing in order to gain insight into the specific requirements and bottlenecks expected in large scale computational approaches in Life Sciences.

Case studies

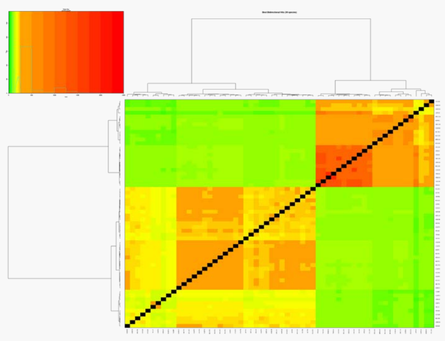
Main goal is to design and implement novel data mining algorithms, tailored to both the specific problem under study as well as to the underlying computational infrastructure.

General approach, applicable to multiple different domains, notably:

A. Pan-genome analysis

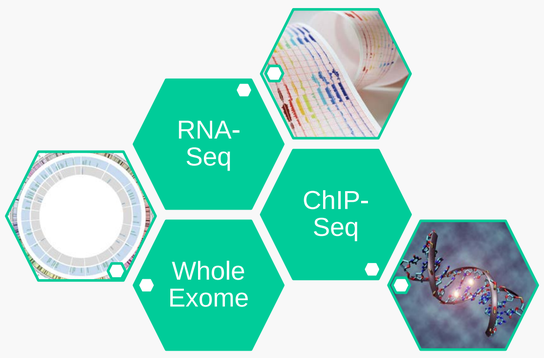
Can provide the framework for estimating the genomic diversity of the dataset at hand and give insights towards the understanding of its observed characteristics.

- Requires a complete information system
- Depends on multiple data sources (different structures, content, etc)



B. NGS Data Integration

A prototype NGS integration pipeline, where data originating from the same physical source (e.g. patient, sample, etc) but sequenced through different mechanisms is combined as a single connected dataset.



Motivation

Information is key in producing knowledge. We undoubtedly live in time of massive information. In Science, there is an abundance of research published and otherwise disseminated daily. **But**:

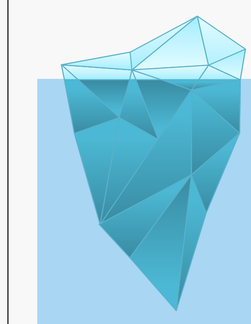
Volume is not always a positive

Useful information can be easily lost within the multitudes of similar, but not necessarily relevant, information

Subscription Walls

Much of the research produced is hidden behind subscription walls, which severely limit access for the wider audience.

This is particularly true in Life Sciences…

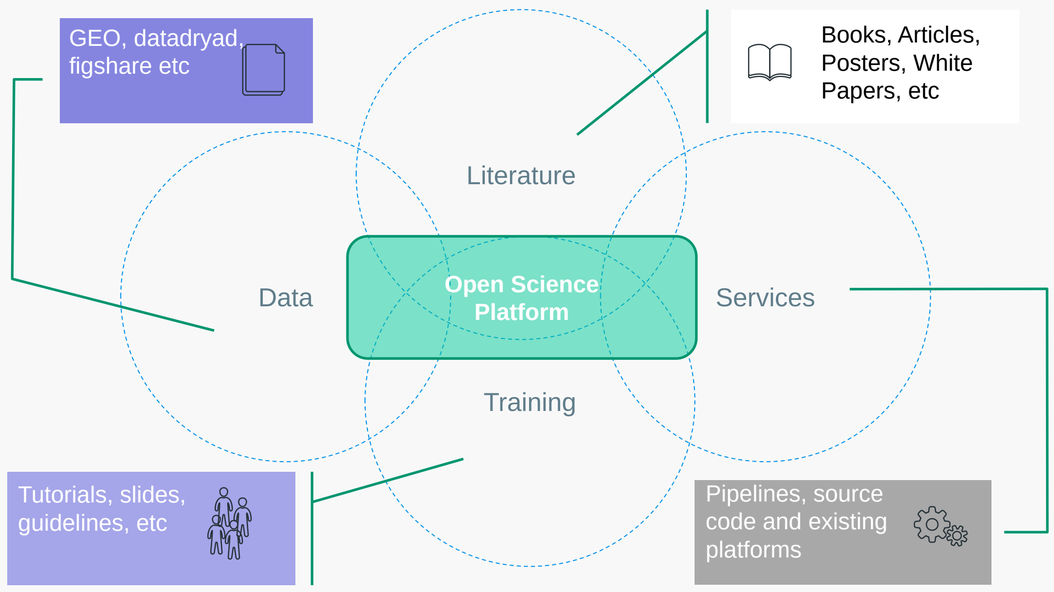


- ✓ **Published data**
  usually downstream analysis, occasionally raw data, almost exclusively low-throughput
- ✓ **Publications**
  just recently a push for open access, but still limited
- ✓ **Tools / Applications**
  available, especially when you know where to look
- ? **Raw data**
  "locked" behind regulations or even through sheer size
- ? **Interconnections**
  data from multiple sources can be connected, but usually never is
- ? **Best practices/guidelines**
  usually through word-of-mouth, "traditional" practices, very few standards

??? *massive loss of raw data and the lack of a system for discovering them, hinders scientific progress* ???

Building a sustainable solution

Creating domain / problem specific platforms can lead to integrated solutions and multiple views of the same issue. Efforts have already started but are in very early stages.



The role of e-Infrastructures

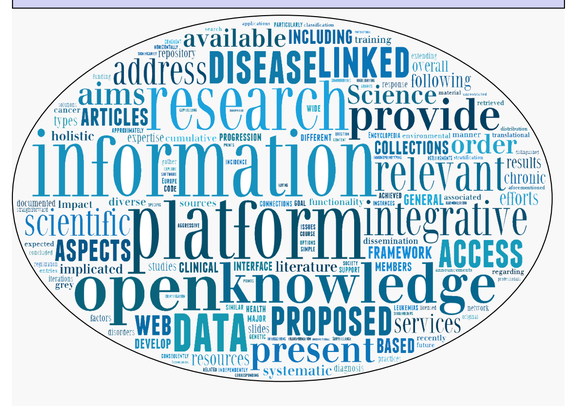
- Transferring the best practices and policies between communities and scientific disciplines
- Building the data infrastructure for bridging different aspects

- Providing the computational infrastructure in the form of Virtual Research Environments (VREs)
- Cloud Computing resources and best practices

- Providing the data services needed for any persistent research environment
- Data retrieval, storage and transfer to execution

- Joining data repositories and open access publication.
- Interconnects all research entities related to scholarly communication

Objectives



Every individual Open Science platform will:

a. act as a nexus of knowledge for the particular domain / problem it addresses
b. provide novel opportunities for systematic integration of different data, thus facilitating horizontal studies

Essentially provide a framework for Bioinformatics analysis that:

1. Enables the re-analysis of old data with new methods
2. Facilitates data remixing and data integration for increased resolution
3. Allows for algorithmic and data-driven hypothesis generation

Take-away message

- Life sciences have become massive in scale; the data we are generating are evidence of our advancing knowledge.
- The simultaneous growth of data collection techniques along with data aggregation and mining algorithms provides an unprecedented opportunity for rapid knowledge discovery.
- Integrated platforms combined with bioinformatics solutions can pave the way towards horizontal studies and novel opportunities.

Relevant Literature

- Voytek B (2016), ***The Virtuous Cycle of a Data Ecosystem***. PLoS Comput Biol 12(8): e1005037. doi:10.1371/journal.pcbi.1005037
- Duarte AMS, Psomopoulos F et. al. (2015), ***Future opportunities and trends for e-infrastructures and life sciences: going beyond the grid to enable life science data analysis***. Front. Genet. 6:197. doi: 10.3389/fgene.2015.00197
- Wicks P et. al. (2011), ***Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm***. Nat Biotechnol.; 29: 411–414. doi: 10.1038/nbt.1837
- Kaye J et. al. (2009), ***Data sharing in genomics—re-shaping scientific practice***. Nat Rev Genet.; 10: 331–335. doi: 10.1038/nrg2573

For more information: fpsom@issel.ee.auth.gr

Thessaloniki , as seen from AUTH