Data intensive analysis approaches in genomics and proteomics: ELIXIR initiatives

(Extended abstract of an invited talk)

© Alexander A. Kanapin
Department of Oncology, University of Oxford, Oxford, UK

alexander.kanapin@oncology.ox.ac.uk

Abstract

Breakthrough in genome sequencing technologies resulted in the unprecedented growth of data volumes in genomics and proteomics. New paradigm of precision medicine signifies wide practical usage of these types of data. ELIXIR, a pan-European bioinformatics consortium meets the challenges arising from production, storage and analysis of massive data collections in genomics and proteomics and proposes several pilot programs, which aim to develop standards and algorithms for the data analysis. The interdisciplinary initiatives of the consortium, such as "BILS-ProteomeXchange integration using EUDAT resources" and "Interoperability of protein resources for drug discovery: Improving Links Between the Human Protein Atlas (HPA) and EMBL-EBI Protein Resources" are of great interest and their successful implementation requires collaboration of researchers and IT engineers. The article also describes general principles of the consortium organization and potential ways of participation in its collaboration projects and programs.

1 Introduction

Biology traditionally was a science based on quantitative observations, and in contrast to physics, it produced relatively small amounts of qualitative data. The situation dramatically changed in a last quarter of XX century. A rapid progress in new technologies of analysis of living systems (cells and organisms) on molecular level resulted in a burst of data, primarily describing features of biological molecules, such as nucleic acids and proteins. A matching appearance of personal computers and global networks facilitated the storage and processing of such information in both small and large scale.

As a result, a new discipline emerged in 1989, when the term "bioinformatics" was mentioned in a title of a scientific paper for the first time [7]. The first databases of primary structures of nucleic acids [3] and proteins [1] were published in 1985 and 1991 respectively. From the very beginning and up to present time, the majority of the data deposited in the biological databanks consists of sequences of biopolymers, namely nucleic acids and proteins.

A successful sequencing of human genome draft in 2001 [6] presented a next big step in the development of bioinformatics and gave a tremendous momentum to creation of new computational engineering solutions and design of novel algorithms for genomic and proteomic data analysis [10].

A progress in biological data acquisition technologies still remains one of major driving forces in bioinformatics. Next Generation Sequencing (NGS) techniques allow to obtain complete genome sequences in a cheap and fast way. This development may change paradigm of traditional medicine towards personal and precise approaches to each of individual patients [11]. However, at the same time it creates new challenges in data intensive analytics for both data storage and manipulation technologies and algorithmic approaches.

The practical solution of such tasks is only possible in a framework of international consortia and collaboration. ELIXIR, a pan-European consortium in bioinformatics opens new opportunities for successful establishment of collaboration in the pilot initiatives of the consortium.

2 Bioinformatics resources

Data management in bioinformatics gradually evolves with the increasing volumes of the biological data. Historically, the protein and nucleic acids databanks delivered the information via CD and other similar media. Later, when networking bandwidth allowed downloading large amounts of data, the databases became available as downloadable flat files. At present time the bioinformatics resources may be classified using the following rough categories:

- Data repositories. The public or commercial data banks containing primary sequences and structures of biopolymers. The repositories also contain tools to analyse data provided by user in a context of the resource. Examples: UniProt, GenBank, RSCB PDB.
- Analytical toolboxes. The complex portals providing exclusive algorithms for user data

Proceedings of the XVII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2015), Obninsk, Russia, October 13 - 16, 2015

analysis.
- Bioinformatics cloud resources.

3 ELIXIR: pan-European collaboration in bioinformatics

The ELIXIR consortium was founded in 2006 by European Laboratory for Molecular Biology (EMBL). The consortium officially started as a fully functional body in December 2013 when the consortium agreement was signed by the first member states. At present it includes 12 full members and 6 observers.

The major goal of ELIXIR is coordination of efforts in quality control and archiving of life sciences data in pan-European scale. The complexity of the data and its heterogeneity calls for creation of infrastructure and system of standards as well as development of proper training programs. ELIXIR will act as a sustainable repository for life science data that has been funded by the public [2].

The consortium is organized as a network of interactions between central hub (Hinxton, UK) and national nodes in each of the member states. The participation in research pilot initiatives is opened to all scientific organizations of the member states.

Currently, ELIXIR is unfolding its activities through series of pilot programs and initiatives. The scientific program of the consortium proposes several research and development avenues along the main directions of future development of data intensive analytics in biological sciences [5].

4 Data intensive analytical programs in proteomics and genomics

4.1 Integrative genomics initiatives in ELIXIR

Comprehensive resources of various data modalities in genomics is essential prerequisite for modern research in biological sciences and translational medicine. EMBL-EBI pioneers the initiative since the creation of one of the first nucleotide sequences database, EMBL-base. Now, as a part of ELIXIR services it provides a diverse spectrum of genomics data, the most outstanding of them are:

- ENA – European nucleotide archive, centred around nucleotide sequencing. The resource contains raw sequencing data, sequence assembly and functional annotation of the data
- EnsEMBL – unique genome annotation resource containing high-quality integrated annotation on vertebrate genomes. The resource comprises data mining interface, BioMart for data retrieval.
- European Variation Archive – a recent development of the novel approach to genomic data, the database contains all types of genetic variation data
- Expression Atlas – RNA-related portal, collecting information about gene expression patterns in different species and various biological conditions.

High quality manual curation and verification of the information in the databases ensures the reliability of the data available. Internal connectivity and integration between the different resources in the Institute allows high level of data integrity and consistency.

4.2 Proteomics in ELIXIR

Proteomics research makes a significant part of the consortium scientific programme. Several protein and protein expression resources have been established in Europe, containing valuable information for biomedical research. Seamless navigation between these resources is an important prerequisite for scientists to make informed decisions about their research into new drug targets and are exploring links between different proteins in healthy and diseased tissues. Swedish national node of ELIXIR plays an important role in this action, working with EMBL-EBI. The consolidated efforts make the Human Protein Atlas interoperable with such proteomic resources as PRIDE, InterPro, and the Gene Expression Atlas.

4.3 BILS - ProteomeXchange

An arrival of tremendous volumes of biological data calls for a need for distributed data storage and replication and reliable and scalable data access interface. One of the ELIXIR pilot initiatives aims to integrate the raw data repositories for mass spectrometry proteomics data run by Bioinformatics Infrastructure for Life Sciences (BILS, Sweden) and ProteomeXchange consortium via the PRIDE database, hosted in EMBL-EBI, UK. The key point in the infrastructure is provided by the European infrastructure EUDAT (http://www.eudat.eu/). The ProteomeXchange consortium facilitates submission and standardization of dissemination practices for proteomics data resources. The main goal of the consortium is to develop a framework to allow standard data submission and dissimentaion pipelines between main proteomic repositories, such as PeptideAtlas, PRIDE and MassIVE. The consortium encompasses 1963 proteomics datasets as of may 2015. PRIDE, one of key participants, stores MS-based proteomics data, such as protein expression data, post-translational modifications, raw MS data and technical metadata.

BILS is a distributed national research infrastructure, supported by the Swedish Research Council, its bioinformatics networks includes 6 nodes in major Swedish universities. Proteios, a multi-user platform for analysis and management of proteomics data was developed as an essential part of the integrative initiatives of BILS.

EUDAT is a pan-European project aiming at building and operating of global collaborative data infrastructure for preserving and exchange of scientific data in various disciplines. Essential components of its software ecosystem, such as B2SAFE and iRODS ensure robust, safe and highly available data access.

B2SAFE software is a key component of the ProteomeXchange data infrastructure.

The initiative could serve as an example of engagement of various types of data storage services in ELIXIR and demonstrate the potential of collaboration among research infrastructures and e-infrastructures to better manage the data deluge.

4.4 Protein resources in drug discovery

Important aspects of many genetic diseases are reflected in potentially different roles of proteins and pathways in diverse cell lineages. Interoperability between databases providing tissue-specificity information and describing expression of genes and proteins in multiple tissues at different stages of development in different diseased conditions becomes critically important for the modern approaches in drug discovery. The heterogeneity of the data representation in these expression resources poses a challenge as they often complement each other and different providers follow different rules to annotate and provide the information. The major goal of the ELIXIR pilot is to define and implement standards and tools to facilitate access and integration of the data for the scientific community. The proteomics and expression resources in the framework include:

- The Human Protein Atlas (HPA) [4], a database of protein expression profiles based on immunohistochemistry.
- The PRoteomics IDEntifications database (PRIDE) [9], a public data repository for protein and peptide identifications.
- The Gene Expression Atlas (GXA) [8], an enriched database of gene expression patterns.

The project proposes the following integration strategies. First, summaries of information from different databases based on a single entry point and on a common format will be created. The approach was successfully introduced before by the EMBL-EBI search portal and includes an amalgamation of service layers on top of a database providing summary data in a standard manner, while the original resources do not change their data or schemas. The non-intrusive approach ensures the independence of the original sources and provides on demand integration. The second approach was adopted by Biosapiens consortium and defines a common terminology and format to describe minimum information for specific data entries. It provides a common language and standard format of the data to integrate and compare protein annotations for 39 databases. The strategy requires an agreement on control vocabularies and changes that might affect data content and annotation process and is therefore more challenging task for the data providers.

Distributed Annotation System (DAS) was used as a communication fabrics to disseminate protein expression summary data and protein sequence annotations, as GXA and PRIDE use DAS to provide expression data. In collaboration with HPA a new DAS service was created to provide expression summaries. Collaboration with other resources, such as UniProt, PDB, pFam, InterPro, PRIDE and IntAct continues, aiming to to create a BioJS component to standardize the visualization of protein features which will be used to represent related expression data such as antibody binding and protein identifications.

One of major challenges for expression information integration among the listed sources is the metadata annotation. The metadata harmonisation implementation is planned as a next step, based on Experimental Factors Ontology (EFO) as a reference system. HPA also proposes XML solution, which is more standardized, and flexible than DAS and might suit better as means of data exchange.

References

[1] Bairoch A, Boeckmann B. The SWISS-PROT protein sequence data bank. Nucl. Ac. Res., v. 19, p. 2247-2249, 1991

[2] Blomberg N. ELIXIR: Data for life. 2014, https://www.elixireurope.org/system/files/ELIXIR_2014_brochure_full.pdf

[3] Burks C, et al. The GenBank nucleic acid sequence database. Comput. Appl. Biosci., v.4, p. 225-233, 1985

[4] Colwill K; Renewable Protein Binder Working Group, Gräslund S. A roadmap to generate renewable protein binders to the human proteome. Nat Methods., v. 15, p.551-558, 2011.

[5] ELIXIR consortium. Scientific programme 2014-2018. Executive summary. 2015, https://www.elixir-europe.org/system/files/ELIXIR-Executive-Summary-2015_Digital.pdf

[6] Lander E. et al. Initial sequencing and analysis of the human genome. Nature, v. 409, p. 860-921, 2001

[7] Masys D. New directions in bioinformatics. J. of Res. Nat. Inst. Stand. and Techn. v.94, p. 59-63, 1989

[8] Petryszak R, et al. Expression Atlas update--a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. Nucleic Acids Res., v. 42, p. 926-932, 2014.

[9] Reisinger F. et al. Introducing the PRIDE Archive RESTful web services. Nucleic Acids Res., pii: gkv382, 2015

[10] Thornton J. The future of bioinformatics. Trends in Biotechn., v. 17, p. 30-31, 1998.

Topol E. Individualized medicine from prewomb to tomb. Cell, v. 157, p.241-253, 2014