Enhancing machine learning application for Rare Disease

Date: *Thursday 8 June, 09:00 (local time, IST)*
INFO: room capacity:
INFO: zoom link
Password: AH2023

Please upload any slides to this folder

Chairs:

Emidio Capriotti
Nuria Queralt-Rosinach

Abstract:

The integration of genomic, phenotypic and clinical data have achieved new paradigms in the diagnosis and care of patients with rare diseases (Saunders et al. Nature Review Genetics 2019). The assembly of large cohorts is essential for the discovery of unknown aetiological variants, the identification of new associations between genes and rare diseases, and important drug target candidates. In this context, the collection of well curated and annotated sets of data and the definition of standard procedures for their analysis is needed for enhancing precision medicine and reducing the cost of the national health systems.

Several ELIXIR communities and focus groups have contributed to different topics related to the collection, annotation and analysis of human genomic data that can be customized for the study of rare diseases. The Rare Disease community organized 3 service bundles dedicated to: i) assess the molecular pathogenicity of the genetic variants collecting procedures and guidelines for their annotation and interpretation, ii) define concept maps enabling the combination tools and methods to create and curate rare disease pathways and networks of genes and iii) promote training events to support the data FAIRification and develop e-learning platform. The Machine Learning Focus group is organized in three subgroups which are collecting and reviewing experimental and synthetic datasets for the assessment of supervised machine learning methods according to the DOME Recommendations (Walsh et al. Nature Methods 2021). The Federated EGA community is contributing to the collection of genomic data in the main European genotype-phenotype archive. The ELIXIR Interoperability platform is generating standard pipelines for the analysis of genomic data.

In this scenario we propose a workshop divided in three sections of 20 minutes each presenting:

1. The activities of the Rare Disease community;
2. The workflow and computing infrastructure for storage, annotation and analysis of data
3. The activity on the collection of experimental and synthetic data for ML

Two slots of 10 minutes each will be dedicated Q&A and the discussion. The main goal of the meeting consists in the definition of a common roadmap for the definition of specific frameworks for the development of machine learning methods in the field of rare disease. We expect collaborations across different ELIXIR focus groups and communities will promote the generation of transnational networks for a European infrastructure for the collection, sharing and analysis of genomic data.

Time (IST)	Subject
9:00 - 9:05	Title: Introduction and aims of the session Lead: Emidio Capriotti Link to slides
9:05 - 9:20	**Machine Learning Focus Group Activities**
	Title: MLFG Task 1 – DOME Registry Speaker: Silvio Tosatto, University of Padova (ELIXIR-IT) Link to slides:
	Title: MLFG Task 2 – Review and assessment of new gold-standard datasets Speaker: Nils Hoffmann, Forschungszentrum Jülich (ELIXIR-DE) Link to slides
	Title: MLFG Task 3 – Development of Synthetic dataset Speaker: Núria Queralt Rosinach, Leiden University Medical Center (ELIXIR-NL) Link to slides
9:20 - 9:40	**Rare Disease Community Activities and Service Bundles**
	Title: Assessing Molecular Pathogenicity for Rare Diseases Speaker: Emidio Capriotti, University of Bologna (ELIXIR-IT) Link to slides
	Title: Systems Biology for Rare Disease Speaker: Friederike Ehrhart, Maastricht University (ELIXIR-NL) Link to slides
	Title: FAIRification and Training activities for Rare Diseases Speaker: Claudio Carta, Istituto Superiore di Sanita' (ELIXIR-IT) Link to slides
9:40 - 10:10	**Relevant efforts in other communities and focus groups**
	Title: The Variomes platform Speaker: Patrick Ruch, Swiss Institute of Bioinformatics (ELIXIR-CH) Link to slides
	Title: OmicsDI and BioModels platforms Speaker: Rahuman Sheriff, European Bioinformatics Institute Link to slides
	Title: Bioschemas for ML datasets Speaker: Leyla Jael Castro (ZB MED, NFDI4DataScience) Link to slides
10:10 - 10:25	Title: Open discussion Moderator: Núria Queralt Rosinach *(please add questions for discussion and help take notes below)*
10:25 - 10:30	Title: Wrap-up Lead: Emidio Capriotti

Questions for discussion

(please add below!)

Data collection and curation

- Where do I have to publish my (data, tool, workflow) resource for the RD community? Should I use a recommended standard or format/profile like RO-Crate or Bioschemas?
- Data quality: Where is a list of datasets used for RD ML analytics? Are there any metrics/scores to check the quality of the data? Are there RD recommended ontologies to use to annotate ML tools for RDs? What file format/data types and its identifiers are recommended/used in analytics?
- Gold standards: Are recognized gold standard datasets for RDs ML? Where could I find them?
- Definition of Bioschimas for the description of variant datasets
- Synthetic data: Where I can find synthetic datasets generated for RD research? Is there a list of tools to generate synthetic data for RDs? What do I have to do to protect the privacy of my patient data?

Machine learning tools and workflow for RD

- Tools/workflows: Where I could find tools/workflows commonly used for RD analytics?
- AI/ML methods: Are there any standard tools or methods to perform (single cell) multi-omics integration? and for diagnosis or drug repurposing ML based analytics for RDs? Are there any interoperability link available among data types and tools?
- How can I reproduce a RD workflow?
- How can I trust a ML-based prediction to diagnose/treat a RD patient?
- Are there best practices for RD ML analytics for example in RDMKit?
- Federated learning: What I have to do to make my RD resource available to be used for federated analytics?
- Secure compute platforms: Are there available secure compute platforms for RD analytics?

Integration with other relevant efforts in the ELIXIR community

- How to reach out other ELIXIR communities for cross-domain analytics
- Cooperation with the hCNV community for the definition of gold standard datasets
- Alignment of MLFG Tasks with RD Service Bundles.
- Curation of dedicated section in OmicsDI
- Contribute to Variomes Platform

Notes (collectively taken)

Introduction
Emidio: Presentation of the aims of the meeting and the 3 different sections.

Section 1

MLFG Task 1 - Silvio Tosatto: Presents the activity of the Task 1 focusing on the presentation of the DOME recommendations and registry. Cooperation with *GigaScience* to adopt DOME recommendations. Presentation of the DOME Registry interface for scoring manuscripts, describing the implementation of machine learning method.

MLFG Task 2 - Nils Hoffmann: Presentation of the activity of the Task 2 which is related to the collection and review of gold standard datasets for training and testing machine learning methods. In particular, Task 2 focus on the collection of human data. Nils describe possible issue on the collection of the data. Presents previous activity in the organization of the BioHackatons in 2022 e the next one in 2023.

MLFG Task 3 - Núria Queralt Rosinach: The lack of data for training machine learning methods is enhancing the generation and collection of synthetic datasets. The Task 3 group analyzed 85 publications describing synthetic datasets and definition of a metrics for the evaluation of the quality of synthetic data. An important aspect is the dissemination of the dataset using FAIR principles to make data easy to retrieve.

Section 2

Service Bundle 1 - Emidio Capriotti: AMP4RD Service bundle aims to identify valuable tools, resources and datasets required for the annotation and interpretation of the genetic variants in the context of Rare Diseases (RD). Resources and tools for RD variant interpretation published. An analysis of integrated ClinVar and Orphanet identified 5000 RD and associated pathogenic variants. Pathogenicity prediction assessment on 4 methods. Bio.tools RD collection (271 tools): scored tools for annotation of pathogenic variants.

Service Bundle 2 - Friederike Ehrhart: Activity of the service bundle 2 in the improvement of WikiPathways. Enhance the use of pathways annotation for studying the mechanism of rare disease. In detail, the service bundle 2 is collecting specific pathways related to rare disease providing information about how to use this information. It was also described the workflow dedicated to the multi-omics analysis of Rare Disease data.

Service Bundle 3 - Claudio Carta: Presents the activity related to the FAIRification of data in the context of EJPRD project. Caludio presented the activity related to the organization of the BYOD meeting, focusing on the FAIRIfication of real data brought by medical doctors in theri clinical practice.

Section 3

Relevant efforts 1- Patrick Ruch: The Variomes platform: Variomes is a high recall search engine t o support the curation of genetic variants. It is recall-oriented, which is good for RDs because prefers recover all the variants more than as much as possible. ML is behind the scene in different components like extracting the ‘size population’ using BERT (potentially useful for the DOME annotations), BUT there is a limitation since BERT needs domain specific such as RD curated data to pre-train the LM

Relevant efforts 2- Rahuman Sheriff: OmicsDI and BioModels platforms: FAIR data and models for machine learning. OMICSDI is a pubmed for omics datasets. Aggregates omics resources. Search

interface has impact metrics. Links to datasets, and links to biostudies that links to datasets and metadata, so FAIR. Biomodels is a repository of curated systems bio models. Making ML models FAIR and reproducible - preprint

Relevant efforts 3- Leyla Jael: Bioschemas for ML datasets. Not done due to lack of time.