IPAD-MD Expert Group Meeting on Data & Resources
Deliverable 7.4 (WP7)

Meeting Report on INFRAFRONTIER / IMPC Machine Learning Hackathon

January 23rd – 24th, 2019
Munich, Germany

Reported by:
Dr. Asrar Ali Khan and Dr. Michael Raess

The IPAD-MD project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 653961

Table of Contents:

Introduction

A hackathon is an event where experts come together to solve a problem or a set of problems for a defined duration. In other words, it's a very focussed, condensed and creative problem-solving session geared towards a specific subject.

The International Mouse Phenotyping Consortium (IMPC) has an impressive 6200 transgenic mouse lines phenotyped with about 70.6 million data points generated. Deep-learning algorithms can be used on these huge annotated data sets like x-ray images and transcriptomic data to generate a predictive tool based on hidden patterns and associations in and between the data.

How can Machine Learning (ML) advance the study of genetic diseases using the enormous resource of mouse phenotyping data from the IMPC and other data sources? In order to find an answer, some of the leading bioinformaticians from Europe and USA came together in Munich for the first INFRAFRONTIER/IMPC Machine Learning Hackathon.

The INFRAFRONTIER / IMPC Machine Learning Hackathon was organised by INFRAFRONTIER as an IPAD-MD activity to incorporate machine learning and deep learning algorithms into the IMPC data analysis pipeline. The Hackathon was attended by 18 data and IT experts from the German Mouse Clinic (GMC), MRC Harwell Institute, Phenomin-ICS and the European Bioinformatics Institute (EMBL-EBI). Prof. Tudor Oprea, from the University of New Mexico and principal investigator of the NIH-funded (Common Fund) Illuminating the Druggable Genome (IDG) programme was also present and gave useful insight and valuable instructions to the participants. The hackathon was organised by INFRAFRONTIER and funded by the IPAD-MD project as a Data & Resource Expert Group Meeting under WP7's task of looking into 'ways to better integrate phenotyping data from individual mouse clinics with the data from the large-scale projects that are accessed via the IMPC portal' (as mentioned in the IPAD-MD grant agreement).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 653961.

Objectives

• Provide means for ideas about applying Machine Learning to be discussed and worked on in a collaborative environment with IMPC members and other experts.

• Build functioning prototypes.

• Create a network of researchers interested in working together on this subject after the hackathon.

Structure

• Participants propose machine learning topics to be worked on at the hackathon via an online google document.

• Attendees then register interest in topic to work on at the hackathon during registration.

• At the hackathon participants will self-organize into topic teams.

• The participants will bring a set of slides and datasets but should be prepared to go off script.

Proposed topics

1. Multi-dimensional analysis of IMPC parameters.
2. Network analysis for identification of pleiotropy using gene clustering.
3. Supervised (deep learning) X-ray image analysis for automated counting.

Agenda

Building 3533, Room 231
Helmholtz Center Munich, Ingolstädter Landstraße 1, D-85764 Neuherberg

23rd January

- 11:00 – 12:00 Arrival
- 12:00 – 13:00 Lunch, introduction to topics and separation into working teams
- 15:00 – 15:30 Coffee break
- 18:00 – End of Day 1
- 18:15 – Taxi to Restaurant
- 18:30 – 20:30 Dinner at Restaurant Osterwaldgarten (http://osterwaldgarten.de)

24th January

- 08:00 – 09:00 Mini-breakfast at Seminar room

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 653961.

- 09:00 – 10:30 Continue work on selected topics and prepare presentations
- 10:30 – 10:45 Short coffee break
- 10:45 – 12:00 Presentations and lessons learned
- 12:00 – 13:00 Lunch

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 653961.

Hackathon Overview

Introductory presentations

Jeremy Mason (EMBL-EBI) started off the hackathon with an introductory presentation. In his presentation, Mr. Mason defined 'Machine Learning' (ML) as computer systems that progressively improve their performance on a specific task or make predictions or decisions without being specifically programmed to perform the task. His talk ended with the overview of the event that includes introductory presentations from the participants, break out into different working groups and outcome/results on the following day.

Dr. Hamed Haseli (EMBL-EBI) presented the idea of using ML to predict essential genes. In other words, will it be possible to predict whether a gene knock out will prove lethal?

Next, Gregor Miller (GMC) introduced several interesting ideas and projects that could use ML approaches. These included large-scale classification / morphometry on histopathological images, prediction of early death / welfare issues based on early-pipeline data, prediction of weight outcome/curve based on early-pipeline data, use high-dimensional phenotype profiles to generate "phenotype families/clusters" (low-dim) and deep learning on X-Ray images. Among these topics, applying deep learning algorithms on X-ray images was further investigated in the hackathon.

Prof. Tudor Oprea (University of New Mexico and IDG) who was invited as external expert in the field, presented the outstanding work being done at the Illuminating the Druggable Genome (IDG) programme in the use of ML for the discovery of new drugs and drug targets. Prof. Oprea gave several examples of the Metapath workflow applications in disease models like Alzheimer's, Schizophrenia and type 2 diabetes. He cautioned that ML models can not verify causality reliably and this will heavily depend on the quality of the input data (good data leads to good models) and stated that the best use of this technology is to identify elusive hidden trends in large amounts of data. He also surmised that ML models are showing promising results in target repurposing for rare diseases.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 653961.

Dr. Anna Swan (MRC Harwell Institute) introduced network analysis for identification of pleiotropy i.e. unexpected co-phenotypes. The search for pleiotropy starts with identification of other phenotypes found in lines along with those of interest, forming networks among phenotypes that are connected found in multiple lines with weighted edges based on the number of lines that have the connected phenotypes. This is followed by cluster analysis to view subsets of the formed networks.

Muñiz Moreno (Phenomin-ICS / IGBMC) presented her graduate work on QUACK, an ML algorithm that can be used to predict new candidate genes in brain dysfunction observed in Down Syndrome. This talk was structured as an interesting case study on biological pathway relationships and functional genomics networks.

Lastly, Dr. Kola Babalola (EMBL-EBI) talk explored the possibility of applying deep learning in IMPC. More specifically, how convoluted neural networks (CNNs) can be used to automatically annotate mouse X-ray images.

After the introductory talks that pitched several interesting ideas for ML applications, the 20 bioinformatics specialists split up into the following working groups to efficiently tackle these issues:

1. Explore integration of IDG's ML method into IMPC
2. Application of deep learning algorithms on mouse X-ray images
3. Clustering networks to identify unexpected phenotypes

Working group overview

WG1: Explore integration of IDG's ML method into IMPC

Aim: Look in depth to the IDG ML method and explore possibilities to apply it to IMPC

Execution:

- Prof. Opera provided access to input and output data from the ML pipeline.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 653961.

- The WG discussed the ML method and pipeline for detecting candidate genes for phenotypes and disease and identified useful cases to explore.

Conclusion/Lessons learnt:

- Received access to input data and results for studying the ML method in detail.
- The exploration of the aforementioned results is ongoing.
- The ML algorithm will be shared by the IDG as a zip file in R script that can be reimplemented in EBI, Harwell and HMGU clusters.
- The implementation of IDG's ML method is extremely complicated with a large amount of input and output data to be analysed. For example, an output file consists of 20,000+ rows and columns (2GB) that can be manipulated to extract the desired features.

Next steps:

The WG identified certain improvements that could contribute to the ML method:

- XGBoost implementation using GPU instead of CPU leading to a 7-9x increased performance, thereby decreasing the time needed to train a model significantly.
- Run the algorithm in parallel on smaller machine rather than on a large dedicated computer.
- Most importantly, understanding and rationalising the input and output from the ML method for its successful reimplementation.

WG2: Application of deep learning algorithms on mouse X-ray images

Aim: To apply deep learning CNNs to annotate mouse X-ray images generated in the IMPC pipeline

Execution:

- The GMC X-ray images data set was used as input which had the following features:
    - "easy" labels: sex, body weight/composition
    - classify normal vs. not normal
    - 3636 IMPC mice with both dorso-ventral and lateral X-Ray
    - 273 genes, 1813/1825 males/females
    - csv file with ID, sex, genes, body weight, equipment ID

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 653961.

- The WG started with automated prediction of sex from the X-ray images. This would be the foundation for automated annotation for other complex traits on different types of X-ray images.
- The CNN Python script from Dr. Babalola (EMBL-EBI) was adapted for this purpose. This script was previously used score degree of damage in histopathology images and uses the VGG16 architecture.
- The WG faced the hurdle of formatting the images into the correct size and orientation.
- In spite of the aforementioned hurdles, an accuracy of 94% was achieved in sex prediction.

Conclusion/Lessons learnt:

- Prediction accuracy was surprisingly high, but also not perfect.
- Orientation of images is important.
- Cropping needs to be accurate.
- Padding or other solutions might bias the model as males might be bigger and need more attention.

Next steps:

- Standardize orientation in whole IMPC data set.
- Label correct body parts and image perspective.
- Consider other labels:
    - Caudal vertebrae
    - Skull shape
    - Pelvis
    - Shape of vertebrae, etc.
- Next promising application is the prediction of body weight.

WG3: Clustering networks to identify unexpected phenotypes

Aims:

- Use network clustering to identify unexpected phenotypes (initially alongside cardio phenotypes) found together in certain IMPC mouse lines.
- Apply this clustering to the whole IMPC dataset.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 653961.

- Provide a p-value for each mammalian phenotype (MP) term to make the network more quantitative.

Execution:

- The WG looked at top level MP terms rather than low level ones to understand pleiotropy on a higher level and to find links between phenotyping procedures that otherwise might be overlooked.
- Group nodes based on phenotyping procedures to highlight connections between procedures rather than in them.
- Discussed how to apply similar methods to behaviour data.
    - Some of the MP terms are very vague like 'abnormal behaviour' which makes clustering with other terms challenging.
- Tested out use of methods of centrality to analyse network topology.
- Two networks were generated:
    - Network 1. Here nodes represent aggregate of MP terms and the edges the proportion of genes shared between them to identify pleiotropic MP terms that are strongly connected.
    - Network 2. Here network edges are the proportion of genes (rather than the actual number of hits). However, this can only be used where the network nodes are parameters, not MP terms.

Next steps:

- Include effect size i.e. increase or decrease of parameter values.
- Data is not consistent across parameters. For example, in the IMPC cardiology phenotyping data set some centres submit data for parameters that other centres do not.
    - Possible solution: apply some kind of ranking by assigning weights or training a neural network.
- Introduce dimensional reduction on the parameters by taking out the positive correlated parameters

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 653961.

Concluding Remarks and Outcome

The IMPC Machine Learning hackathon was held over the course of 2 days, in a noon to noon format and was planned as an intersection between big data, machine learning and mouse phenotyping. During that time, several ideas for possible applications of ML and deep learning in the IMPC were presented. Three such ideas were selected and the participants divided into 3 breakout working groups and focused on a specific aspect or ongoing process of interest. These expert groups led a cooperative and coordinated effort to come up with intelligent and feasible ML-based solutions. Overall, significant headway was made in applying ML in gene cluster, X-ray image and the IMPC data analysis. The overall feeling of the meeting was one of cooperation and coordination of effort and all the attendees were engaged and willing to participate in a hands-on style of problem solving. This meeting brought together members of the IMPC community interested in applying machine learning with the hopes that collaborations and relationships would be the result, and to that end has succeeded. One such possible IMPC collaboration would be with the IDG in their upcoming publication regarding their ML method applied on IMPC data. Substantial interest was shown for another meeting in this style to continue the momentum started here and provide an opportunity to further explore applying ML to the whole IMPC dataset. As IPAD-MD will end by December 2019, the request for future ML hackathons was incorporated in to WP2 of an INFRADEV-03-2018-2019 project proposal submitted by INFRAFRONTIER to the European Commission in March 2019.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 653961.