AAP Contrats doctoraux en Intelligence artificielle

Cofinancé par l'ANR

The application of Artificial Intelligence to support decision-making in drug-based tumour-informed Precision Oncology

1. DESCRIPTION OF THE PHD THESIS PROJECT

This PhD project will investigate the application of Artificial Intelligence (AI), and especially its Machine Learning (ML) component, to Precision Oncology (PO). PO is a form of medicine that uses personal information to prevent, diagnose and treat cancers. Here we focus on the PO topic of how to predict which cancer patients will respond to a given drug treatment from the molecular profiles of their tumours. Importantly most of PO-based strategies have been developed with targeted therapies, a class of agents rationally design to disrupt specific survival and/or proliferation pathways in tumor cells. Yet, conventional cytotoxic agents, which remain a major cornerstone of treatment for a large number of advanced cancers, are most frequently prescribed empirically without molecular selection. Similarly, robust molecular predictors of immunotherapy, an emerging therapeutic approach which has already revolutionized the management and prognosis of an increasing number of malignancies, are still warranted. In this project we will investigate the ability of genomic profiling to predict response to conventional chemotherapeutic agents in addition to targeted therapies (including immunotherapies).

The application of AI to these PO problems is highly promising. Simple one-feature classifiers have been found to be predictive of patient response in a few problem instances (here drug treatment-cancer type binomials). For example, the expression level of the HER2 molecule (feature) to predict which patients with HER2-positive metastatic breast cancer (type) will respond to trastuzumab (drug)[1]. Unfortunately, one-feature classifiers, a common type of PO models also known as single-gene markers, fail to detect many responsive patients and have only been found for a few treatment- type binomials[1]. By contrast, studies using *in vitro* and *in vivo* preclinical models[2,3] have found that ML classifiers are predictive on many problem instances where one-feature classifiers are not. Furthermore, by combining multiple gene alterations, ML classifiers are also able to identify a higher proportion of responsive tumours (i.e. they offer a higher recall)[2,3], including many that could not be matched to a drug by any single-gene marker.

1.1 OBJECTIVES OF THE PROJECT BASED ON THE CURRENT STATE OF THE ART

For each set of drug-treated molecularly-profiled patients, the main objective is to discover which ML model would best predict the response of forthcoming patients to that drug. The ultimately goal is to use these ensemble of models to better guide therapy selection. We also aim at understanding which ML algorithms and feature selection schemes are most suitable for modeling a given molecular profile. This will encompass investigating the effect of various factors affecting generalisation error (e.g. number of tumor samples, number of features characterizing the sample, feature selection bias, sample selection bias, noise, etc.)

1.2 METHODOLOGY

The increasing availability of data with both drug responses and tumour molecular profiles of patients[4] makes now possible to build and evaluate ML classifiers on clinical data at an unprecedented scale. The prospect of applying here sophisticated Deep Neural Network (DNN) algorithms, which have achieved breakthroughs in applications benefitting from large sample sizes[5], is also exciting. However, even the largest clinical data sets for a given treatment and cancer type are of extreme high-dimensionality, meaning that they are of low sample size (few patients) with respect to the number of available features (e.g. the mutational status of each profiled gene in the patient's tumour). Such curse of dimensionality presents an important challenge for any ML technique[6], including DNNs with overfitting-reducing strategies (e.g. Dropout), boosting techniques with tree-based embedded feature selection (e.g. XGBoost) or linear classifiers with regularisation (e.g. Elastic Nets)

The main source of such clinical data sets will be an ongoing Phase-II clinical trial (SAFIR02_Breast; https://clinicaltrials.gov/ct2/show/NCT02299999). We have already been granted access to these data sets and the estimated enrolment of this trial was completed in 2019 (n=1460). In SAFIR02 trial, HER2-negative metastatic breast cancer patients are profiled for somatic mutations and copy-number alterations in 50 disease-relevant genes, while they receive cytotoxic chemotherapy. Patients with stable or responding disease are then randomized between chemotherapy maintenance or targeted therapy matched to specific genomic alterations among one of the 19 considered innovative drug treatments. Thus, this data set will provide a unique occasion for matching sensitivity or resistance to conventional cytotoxic with genomic alterations, allowing extending the PO concept to conventional chemotherapy. In SAFIR02, a sub-study enrolled patients with responding or stable disease but without actionable molecular alterations to be randomized between chemotherapy maintenance or immunotherapy using immune checkpoint inhibitor durvalumab, an anti-PD-L1 monoclonal antibody). Thus, analysis in this subpopulation may allow identifying ML classifiers predictive of immunotherapy maintenance benefit. In addition, NGS of tumor tissues from MOVIE trial - an academic, PHRC2016-funded, currently running phase I/II trial evaluating combination of metronomic chemotherapy with double blockade immune checkpoint inhibitor (durvalumab plus tremelimumab, an anti-CTLA4) in various solid tumors - will be available and will also be examined to identify ML predictors of efficacy with this most innovative strategy.

Training-test partitions of these time-stamped data sets will be carried out for validation purposes. Additional clinical data sets will be assembled from the NCI Genomic Data Commons (GDC)[4], which currently contains over 84,000 patients across 67 cancer tissue types. For example, there are over 100 whole-exome profiled patients with breast invasive carcinoma treated with cyclophosphamide, one of the 19 SAFIR02_Breast drugs. Relevant GDC data sets will be assembled to be used as a second test set or to increase training size.

1. Huang, M., Shen, A., Ding, J. & Geng, M. Molecularly targeted cancer therapy: some lessons from the past decade. *Trends Pharmacol. Sci.* 35, 41–50 (2014).

2. Naulaerts, S. *et al.* Precision and recall oncology: combining multiple gene mutations for improved identification of drug-sensitive tumours. *Oncotarget* 8, 97025–97040 (2017).

3. Nguyen, L. *et al.* Machine learning models to predict in vivo drug response via optimal dimensionality reduction of tumour molecular profiles. *bioRxiv* 277772 (2018). doi:10.1101/277772

4. Jensen, M. A., Ferretti, V., Grossman, R. L. & Staudt, L. M. The NCI Genomic Data Commons as an engine for precision medicine. *Blood* (2017). doi:10.1182/blood-2017-03-735654

5. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks* 61, 85–117 (2015).

6. Marx, V. Machine learning, practically speaking. *Nat. Methods* 16, 463–467 (2019).

1.3 WORK PLAN

Months 1-6: The first task of the project will be to review the literature for ML techniques intended to build classifiers from high-dimensional data. For those techniques with available code, we will benchmark them on synthetic data sets where we can investigate the effect of various factors affecting generalisation error (e.g. number of samples, number of features, feature selection bias, sample selection bias, noise, etc.).

Months 7-12: On the other hand, the Ballester team has integrated univariate feature selection with tree-ensemble ML, which has generally resulted in better performance than tree-ensemble ML techniques alone[3]. An additional advantage of these methods is that they returned a much smaller subset of predictive features, which are in turn important for interpretation (e.g. which molecular factors collectively control patient response to this drug?). The student will build upon this work by investigating the advantages of integrating ML with multivariate feature selection instead. This will be also based on synthetic data sets.

Months 1-12: In parallel, the student will mine SAFIR02_Breast, MOVIE and GDC data to assemble and curate the clinical response and tumor profiling data sets.

Months 13-36: the student will study the application of ML to the SAFIR02_Breast, MOVIE and GDC curated datasets guided by the results of the benchmarking studies on synthetic data sets.

Months 31-36: thesis writing.

1.4 SUPERVISOR AND RESEARCH GROUP DESCRIPTION

The supervisor of this thesis will be Dr Pedro Ballester with Prof Anthony Gonçalves as co-supervisor.

The PhD student will be trained in these AI topics in the Ballester team. Dr Ballester has formal training on AI (he was awarded in 2001 an MSc in *Information Processing and Neural Networks* from the Department of Mathematics of King's College London, UK). Since then, he has mostly worked on the development and application of supervised learning techniques to biomedical problems (including this PO topic). The student will have access to the developed R and Python codes for building ML classifiers for high-dimensional data as well as our experience in mining SAFIR02_Breast, MOVIE and GDC data. We will also provide a dedicated high-performance workstation with 40 CPU threads and 64GB of memory as well as access to a shared cluster with over 500 CPU cores when required. The successful candidate will join Dr Ballester's team, which is an international multi-disciplinary environment of postdocs and PhD students working on the development and application of AI models for healthcare, including other AI for PO projects.

The PhD student will be also trained in the application domain by Prof Anthony Gonçalves, head of the Medical Oncology Department at the Institut Paoli-Calmettes (IPC), a comprehensive cancer center founding-member of the CRCM together with Aix-Marseille University (AMU), CNRS and INSERM. Dr Gonçalves is Professor of Oncology at AMU, and is leading the Breast Cancer program at IPC. IPC is involved in the management of nearly 2000 breast cancer patients per year and has developed an outstanding expertise in both clinical and translational research dedicated to breast cancer, including PO-based approaches. Prof. Gonçalves was involved in SAFIR program since its early initiation and is an active member of its steering committee. He will provide scientific expertise and support for assembling and integrating clinical data, including characterization and classifications of administered chemotherapeutic agents as well as clinical evaluation of tumor response.

2. RECENT PUBLICATIONS

Naulaerts, S., Menden, M.P., Ballester, P.J. (2020) "Concise Polygenic Models for Cancer-Specific Identification of Drug-Sensitive Tumors from Their Multi-Omics Profiles". *Biomolecules* 10 (6), 963.

Bomane, A., Gonçalves, A., Ballester, P.J. (2019) "Paclitaxel response can be predicted with interpretable multi-variate classifiers exploiting DNA-methylation and miRNA data". *Frontiers in Genetics* 10: 1041

Nguyen, L., Naulaerts, S., Bomane, A., Bruna, A., Ghislat, G., Ballester, P.J. "Machine learning models to predict in vivo drug response via optimal dimensionality reduction of tumour molecular profiles". *bioRxiv 277772 (2018). doi:10.1101/277772*

Naulaerts, S., Dang, C., Ballester, P.J. (2017) "Precision and recall oncology: combining multiple gene mutations for improved identification of drug-sensitive tumours". *Oncotarget* 8 (57), 97025.

Bertucci F, Finetti P, Goncalves A, Birnbaum D. (2020) "The therapeutic response of ER+/HER2- breast cancers differs according to the molecular Basal or Luminal subtype". *npj Breast Cancer* (In Press)

Hurvitz SA, Gonçalves A, Rugo HS, Lee KH, Fehrenbacher L, Mina LA, Diab S, Blum JL, Chakrabarti J, Elmeliegy M, DeAnnuntis L, Gauthier E, Czibere A, Tudor IC, Quek RGW, Litton JK, Ettl J. (2020) "Talazoparib in Patients with a Germline BRCA-Mutated Advanced Breast Cancer: Detailed Safety Analyses from the Phase III EMBRACA Trial". *Oncologist* 25 (3), e439-e450

Jabagi MJ, Goncalves A, Vey N, Le Tri T, Zureik M, Dray-Spira R. (2019) "Risk of Hematologic Malignant Neoplasms after Postoperative Treatment of Breast Cancer". *Cancers* 11(10): 1463

3. EXPECTED PROFILE OF THE CANDIDATE

Selection criteria - Essential

- An excellent first and master degree with a major focus on computational analysis of experimental data, preferably in an area directly relevant to the project.
- Skills in the implementation of R or Python scripts for scientific data analysis.
- Ability to communicate effectively in English, both orally and in writing.

Selection criteria - Desirable

- A formal training in machine learning, especially supervised learning from high-dimensional data.
- Master project and/or internship in the application of machine learning to solve real-world problems in the context of biomedical research.
- Experience in writing research for publication in international journals.
- Prior use of computational tools and resources to analyze clinical pharmaco-omic data.
- At least familiarity with the processes of handling, integrating, processing and analyzing molecular profiling data (e.g. DNA-seq, RNA-seq, miRNA-seq or DNA methylation microarrays).
- Comfortable working in linux platforms.

4. SUPERVISORS' PROFILE

The supervisor (Dr Ballester) heads the "Machine Learning for Precision Oncology and Drug Design" team at the Cancer Research Center of Marseille (CRCM; INSERM U1068, CNRS U7258, Aix-Marseille University UM105). Dr Ballester holds an MSc in Information Processing and Neural Networks at King's College London in 2001, and a PhD at Imperial College London on geophysical data mining and inference in 2005. Subsequently, he held postdoctoral positions at University of Oxford, University of Cambridge and EMBL-European Bioinformatics Institute, including a 4-year MRC Methodology Research Fellowship. In October 2014, he left the UK and moved to Marseille to take an INSERM position in the CRCM. He started with first group in February 2015 with an

A*MIDEX Excellence Chair followed by an ANR Tremplin-ERC grant. To date, Dr. Ballester has published 71 peer-reviewed papers (77% as corresponding author, his h-index restricted to papers where he is either first, last or corresponding author is 26). Dr Ballester's research focuses on the development and application of computational methods to predict and analyse the modulation of protein and cell function by small organic molecules. These problems can be tackled by generating predictive models from selected and curated data using ML. Within this area, problems of interest include predicting treatment response of tumours from their molecular profiles for precision oncology, cancer pharmaco-omic modelling for phenotypic drug design, molecular target prediction by bioactivity data mining and target-based drug design.

The co-supervisor (Prof Anthony Gonçalves) is a medical oncologist at the Comprehensive Cancer Center Institute Paoli-Calmettes (IPC), Marseille, France. Fellowship of the Internat des Hôpitaux de Marseille, he obtained his medical degree in 1998 and his PhD in Pharmacology at Aix-Marseille University in 2001, after a stay in the Laboratory of L. Wilson and M.A. Jordan at the University of California- Santa Barbara (CA, USA). He joined the Department of Medical Oncology at IPC in 2001, and became Head of this Department in 2018. At IPC, he is leading the breast cancer program and is also co-heading the early phase trial unit, in charge of developing innovative medical therapeutics for solid tumor patients. At the national level, he is a member of the UNICANCER Breast intergroup. He is a member of La Société Française du Cancer (SFC), ASCO and ESMO. Since 2013, he is Professor of Oncology at the Medical School of La Timone, Marseille, Aix-Marseille University. Prof. Gonçalves has published more than 200 papers in peer-reviewed journals in the domain of basic, translational and clinical oncology. He is also a member of the CRCM and his translational research works have focused on mechanisms of resistance to antimicrotubule agents and on mass spectrometry-based proteomics applied to biomarker identification in cancer patients. He is currently developing clinical trials evaluating rationalized combination of targeted therapeutics in metastatic breast cancer patients.

VISA DU RESPONSABLE DE L'INSTITUT ET DU DIRECTEUR DE LABORATOIRE CONCERNÉS

Visa du responsable de l'institut, NOM Prénom

Jean-Paul Borg

Fait à Marseille,

Signature

Visa du directeur du laboratoire, NOM Prénom

Jean-Paul Borg

Fait à Marseille,

Signature