ECOLE DOCTORALE PIERRE LOUIS DE SANTE PUBLIQUE A PARIS
EPIDEMIOLOGIE ET SCIENCES DE L'INFORMATION BIOMEDICALE

Directeur : Pierre-Yves Boëlle
Responsable pour l'Université Paris Cité : Isabelle Boutron

PROPOSITION DE SUJET DE THESE

SIGLE ET NOM DU LABORATOIRE : Unité de Biologie Fonctionnelle & Adaptative (BFA) - CNRS UMR 8251, Université de Paris.
NOM DE L'EQUIPE : Computational Modeling of Protein Ligand Interactions (ERL U1133)
DIRECTEUR DE THESE : Pr Olivier Taboureau (Université de Paris)
ADRESSE : Université de Paris - Inserm U1133, 35 rue Helene Brion, Bâtiment A, 4ème étage, 75013 Paris

TITRE DE LA THESE :
Artificial intelligence combining transcriptomics and high content imaging data to identify endocrine disruptors as potential breast cancer risk factors (ED-Breast)

CO-ENCADRANT EVENTUEL :
EQUIPE DU CO-ENCADRANT :
LABORATOIRE :

PRESENTATION DU SUJET

2 à 4 pages (références comprises), structurées comme suit :

1. *le contexte scientifique du projet ;*

Humans are daily exposed to diverse hazardous chemicals via skincare products, plastic cups, food, drugs and pesticides to mention but a few sources and the potential adverse effect of these chemicals on human health is a major concern. Many of them have the ability to interfere with the endocrine system and are called endocrine disruptors (ED). They can affect the hormonal regulation of an organism in a wide variety of ways, for example, by mimicking natural hormones, antagonizing their action, or modifying their synthesis, metabolism, and transport through their interference with multiple cellular targets [Gore 2015]. By today, their modes of action (MoA) are not well defined, and national and international initiatives exist to pave these gaps.
One concern about ED is whether exposure to such chemicals, which interfere with the hormonal systems, can increase the risk and progression of breast cancer, the most common invasive cancer in women worldwide. It is known that the majority of breast cancers are hormonally responsive [Colditz 2004] and an increase of estradiol and progesterone synthesis are related to an elevation of breast cancer incidence [Cuzick et al. 2020]. However, others mechanisms can promote the tumor progression. For example, chemicals that alter the activity of enzyme involved in steroidogenesis pathways may be an important risk factor for breast cancer [Cardona 2021]. The mammary tumor etiology is complex and not fully elucidated. Pathways that explained the relationship between cancer and ED are not all known [Sathyamoorthy 2020].

2. *les questions posées ;*

In this context, we propose in ED-breast to combine transcriptomics information with cell-based phenotypic screening, available for a large set of chemicals, in order to identify genes and morphological cell signatures associated to ED and breast cancer. We will take advantage of a recent study that has identified several hundred of chemicals (including pesticides, food additives, drugs among others) that should be considered as potential risk factors for breast cancer as they affect hormonal regulation, notably estradiol and progesterone synthesis [Cardona 2021]. Through our computational analysis, we will connect genes deregulated with cell morphology perturbations features that could be of interest in the assessment of ED related to the risk of breast cancer. Artificial intelligence and deep neural network approaches will be considered to evaluate untested chemicals potentially at risk.

3. *les sources de données qui seront utilisées ;*

To manage this unique project, ED-Breast will compile several sources of omics data performed on chemicals of interest for our project.
About chemicals, we will use an update version of the mammary carcinogen database: http://sciencereview.silentspring.org/mamm_about.cfm?new2019=1 [Rudel 2007]. This database of chemical included 266 rodent mammary carcinogens. In addition, we will use a list of 678 chemicals with relevant endocrine activity. This list includes chemicals that stimulate the production of estradiol and progesterone in the HT-H295R assay [Haggard 2018, Karmaus 2016] as described in Cardona 2021 [Cardona 2021]. It included also an extra 45 reference chemicals tested in 18 in vitro ToxCast assays that measure ER-regulated pathways, including receptor binding and dimerization and cellular proliferation [Judson 2015]. Finally, we will used a list of putative non-carcinogens (PNCs) as chemicals that were tested in a two-year US-National Toxicology Program (NTP) cancer bioassay, with negative results for all tumor sites in all four species-sex (mouse/rat, male/female) combinations, or negative in three and "insufficient evidence" or "not tested" in the other. This set will serve as a negative test set useful in the AI modeling
About omics data, transcriptomics, proteomics and high content imaging data will be used.
For transcriptomics, there are many sources of gene expressions data available. The largest one is the LINCS database (https://lincsproject.org/LINCS/tools/workflows/find-the-best-place-to-obtain-the-lincs-l1000-data) for which more than 11000 chemicals have been tested in vitro for different conditions (times, doses) and different cell lines. In addition, others sources will be considered like gene expression omnibus [Clough 2016] and the cancer genome atlas program (TCGA) [Wang 2016] which collect curated gene expression datasets including those stressed by chemicals/substances. Interestingly, growing amount of gene expression data from RNA sequencing are becoming accessible and will be integrated in this study when possible [Rooney 2020, Harrill 2021].
For high content imaging data - cell morphology perturbation from images (Cell painting) for more than 30000 chemicals providing by the Broad Institute (https://www.broadinstitute.org/imaging/morphological-profiling) will be considered [Bray 2017]. After a preliminary check, we identified already hundreds of chemicals suggested as risk factors for breast cancer and that have morphological perturbations information which give us some confidence about project's objectives. During the project, we might get access to a larger set of compounds (~120 000) that would be released for the scientific community (https://jump-cellpainting.broadinstitute.org/) and so increase the set of chemicals of interest.

Finally, for proteomics, the STITCH (http://stitch.embl.de/) and STRING (https://string-db.org/) databases have collected chemical-protein and protein-protein interactions for around 500 000 chemicals. Such information will be investigated for our set of chemicals that can suggest or validate functional perturbation of proteins related to risk factors for breast cancer.

4. *les méthodes ;*

Once the data collected, bioinformatics analysis will be performed. First, morphological features signatures significantly related to chemicals and their associated endpoints (production of estrogen and progesterone, related to mammary carcinogenicity, tested in a steroidogenic assay) will be selected. To do that, we will apply an innovative method developed internally (protein-protein association network [Taboureau 2020]) that allow to prioritize features co-disturbed similarly by a same set of chemicals.
A similar analysis will be performed with transcriptomics data collected in the aim to obtain genes expressions profiles for chemicals suspected to be potential breast cancer risk factors. Experimental conditions considered will be important on this step. Compounds should (at least) have been tested on the same cell lines, with a relatively close concentrations and incubated in a same time in order to avoid false positive. Pathway's enrichment analysis will be also performed to detect potential signaling pathways highly perturbated by these types of chemicals.
It is important to notice that although the Cell painting chemicals collection is one of the biggest freely accessible libraries with cell morphological perturbation, the experiments have been done on a unique cell line (U2OS). However, it has been reported that it exists some concordance between toxicity endpoints observed in specific cell lines and high content screening [O'Brien 2006]. Therefore, an attention on the concordance between U2OS cell line and breast cancer cell lines will be checked. For example, with the transcriptomics data, the representation of the gene's expression in U2OS versus MCF7 and H295R cell lines will be compared and only genes expressed in both cell lines will be conserved in the analysis in the aim to avoid false positive interpretation. This is a protocol that we have implemented with success recently [Cerisier 2023].

Also, using the information obtained from the bioinformatics analysis, a machine learning model based on AI will be developed. Basically, classification models using the gene signatures and the morphological cell signatures will be performed with a set of chemicals defined as potential/non potential breast cancer risk factors. The models will then be used in a second stage for the prediction of potential breast cancer risk factors for the remaining compounds that have both signatures (~20 000 chemicals). We will use intensive machine learning algorithms such as deep neural network using a hyper optimization grid as developed here [Krishna 2022]. A recent publication has reported a predicting model for chemical-induced liver toxicity using high content imaging phenotypes and chemical descriptors and we will explore the possible advantage to include chemical structure parameters (using physicochemical descriptors or fingerprint) in our predictive model [Chavan 2020].

5. *puissance de l'étude/nombre de sujets ;*

The recent hypothesis that many chemicals should be considered as potential risk factors for breast cancer arise the point to identify these chemicals and to develop new alternatives methods that could provide deep mechanistic understanding of the underlying toxicological effects, enabling adverse outcome prediction and providing a new paradigm for chemical risk assessment. With the development of advanced technologies in high throughput transcriptomics and cell-based phenotypic

screening this study will take the opportunity to compile large sets of omics data for an ensemble of 30000 small molecules and to analyze them using promising artificial intelligence approaches for such large sets of data.

References:

- Bray MA, et al. A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay. *GigaScience*, 2017. 6(12): 1–5.
- Cardona J. and Rudel RA. Application of an in vitro assay to identify chemicals that increase estradiol and progesterone synthesis and are potential breast cancer risk factors. *Environ Health Perspect*.2021. 129 (7): 77003.
- Cerisier N. et al. Linking chemicals, genes and morphological perturbations to diseases. *Toxicol Appl Pharmacol*. 2023, 461, 116407
- Chavan S, et al. Predicting chemical-induced liver toxicity using high-content imaging phenotypes and chemical descriptors: A random forest approach. *Chem Res Toxicol*. 2020, 33: 2261-2275.
- Clough E and Barrett T. The gene expression omnibus database. *Method Mol Biol*. 2016. 1418: 93-110
- Colditz GA, et al. Risk factors for breast cancer according to estrogen and progesterone receptor status. *J Natl Cancer Inst*. 2004. 96(3) :218–228
- Cuzick J, et al. Use of anastrozole for breast cancer prevention (IBIS-II): long-term results of a randomised controlled trial. *Lancet* 2020. 395(10218):117–122.
- Gore AC, et al. EDC-2: The Endocrine Society's Second Scientific Statement on Endocrine-Disrupting Chemicals. *Endocr. Rev.* 2015. 36: E1–E150.
- Harrill JA, et al. High-Throughput transcriptomics platform for screening environmental chemicals. *Toxicol Sci*. 2021. 181(1): 68-89.
- Judson RS, et al. Integrated model of chemical perturbations of a biological pathway using 18 in vitro high-throughput screening assays for the estrogen receptor. *Toxicol Sci*. 2015. 148:137-154.
- Haggard DE, et al. High-throughput h295r steroidogenesis assay: Utility as an alternative and a statistical approach to characterize effects on steroidogenesis. *Toxicol Sci*. 2018. 162:509-534.
- Krishna, S, et al. High-Throughput Chemical Screening and Structure-Based Models to Predict hERG Inhibition. *Biology*, 2022. *11*(2): 209.
- O'Brien PJ, et al. High concordance of drug-induced human hepatotoxicity with in vitro cytotoxicity measured in a novel cell-based model using high content screening. *Arch. Toxicol*. 2006, 80:580−604.
- Rooney et al. Mining a human transcriptome database for chemical modulators of NRF2. *PLoS One* 2020, 15(9): e0239367.
- Rudel RA, et al. Chemicals causing mammary gland tumors in animals signal new directions for epidemiology, chemicals testing, and risk assessment for breast cancer prevention. *Cancer*. 2007. 109: 2635-2666.
- Sathyamoorthy N and Lange CA. Progesterone and breast cancer: an NCI workshop report. *Horm Cancer*. 2020. 11(1):1–12.
- Taboureau O. et al. Integrative systems toxicology to predict human biological systems affected by exposure to environmental chemical. *Toxicol Appl Pharmacol*. 2020. 405, 115210.
- Wang Z, et al. A practical guide to the cancer genome atlas (TCGA). *Methods Mol Biol*. 2016. 1418, 111-41.

6. *le calendrier prévisionnel (présenté sous forme d'un échéancier semestriel, doit être suffisamment précis pour constituer un document de référence sans pour autant traiter du détail. Le calendrier doit inclure la période de rédaction et celle d'examen par les rapporteurs) ;*

Gantt Chart						
	6 months	**12 months**	**18 months**	**24 months**	**30 months**	**36 months**
Task1: Data collection						
Task2: Analysis of cell morphology data						
Task3: Analysis of transcriptomics data						
Task4: Analysis of proteomics data						
Task5: Development of IA models						
Task 6: manuscript ad thesis writing						

7. *le thème de chacun des articles prévus. Une proposition de sujet de thèse doit comporter au moins deux articles originaux.*

- One article will describe an analysis on cell morphology perturbations which are highly observed for chemicals considered as potential risk factors for breast cancer. Such analysis has not yet been reported in the literature

- One article will describe the development of an artificial intelligence model combining different source of omics data and able to identify new chemicals as risk factors for breast cancer. In this article, the relation between transcriptomics, proteomics, cell imaging will be also investigated in the aim to suggest the signatures that are specific on each of these data.

PRÉREQUIS, FORMATION : MASTER IN BIOINFORMATICS, IN SILICO DRUG DESIGN, DATA SCIENCES

CONTACT POUR CE SUJET : OLIVIER TABOUREAU
EMAIL : OLIVIER.TABOUREAU@U-PARIS.FR
TELEPHONE : 01 57 27 83 88

SPECIALITE DE LA THESE

Santé publique - Epidémiologie	☐
Santé publique - Epidémiologie clinique	☐
Santé publique - Epidémiologie sociale	☐
Santé publique - Epidémiologie génétique	☐
Santé publique - Biostatistique	☒
Santé publique - Biomathématiques	☐
Santé publique - Biostatistique et Biomathématiques	☐
Santé publique - Informatique médicale	☐
Santé publique - Imagerie biomédicale	☐
Santé publique - Bioinformatique	☒
Santé publique - Recherches sur les services de santé	☐
Santé publique - Economie de la santé	☐
Santé publique - Science des données	☐
Santé publique – Prévention et promotion de la santé	☐

ECOLE DOCTORALE PIERRE LOUIS DE SANTE PUBLIQUE A PARIS
EPIDEMIOLOGIE ET SCIENCES DE L'INFORMATION BIOMEDICALE

Directeur : Pierre-Yves Boëlle
Responsable pour l'Université Paris Cité : Isabelle Boutron

SIGNATURE DU. DE LA DIRECTEUR.TRICE DE THESE

[redacted]

VISA DU.DE LA DIRECTEUR.TRICE DU LABORATOIRE
(DEROGATION DE SIGNATURE NON ACCEPTEE)

AVIS FAVORABLE ☐

[redacted]

Ecole Doctorale 393
Centre Biomédical des Cordeliers
15, rue de l'Ecole de Médecine 75006 Paris
https://ed393.sorbonne-universite.fr/

Contact : ed393@sorbonne-universite.fr/Téléphone : 01.44.27.24.35