Machine Learning approaches for the identification of therapeutic targets against “triple negative” breast cancers

P. Dhalluin

Mines ParisTech – paul.dhalluin@mines-paristech.fr

Key words

Bioinformatics; Machine Learning; Triple Negative Breast Cancer

Abstract

Given the cost and time required for drug screening and the criticality of the problem, not only of breast cancer, but particularly of triple negative breast cancer, original and innovative methods seem to be necessary to make rapid progress in this field. In order to grasp fully how triple-negative breast cancer is expressed and what the mechanisms of action of drugs against this cancer are, chemogenomic approaches based on Machine Learning algorithms capable of identifying therapeutic targets for drugs have been developed.

CBIO – Centre for Computational Biology

1. Introduction

1.1 Triple negative breast cancer

Breast cancer is the most common form of female cancer. It affects approximately 54,000 women a year (it represents 33.5% of all new cancer cases). In France, one woman out of nine will be affected by this cancer during her lifetime.
Although the overall 5-year survival rate is 85%, breast cancer is responsible for approximately 11,000 deaths per year, and the need for new therapeutic strategies for the most invasive forms remains.

There are actually, not one, but four types of breast cancer, sharing certain global characteristics such as the ability to spread, but with different biological characteristics, possibly different aggressiveness, but above all different therapeutic arsenals.
These four classes of breast cancers are distinguished according to the proteins expressed on the surface of the tumor cells which serve as "biological markers" to attribute its class to each cancer.

- ER+ cancers express the estrogen receptor
- PR+ cancers express the progesterone receptor
- HER2+ cancers express HER2, an epidermal growth factor receptor
- Triple negative cancers (TNBC), which are classified by default and do not express any of the three markers above

A tumor may express none, one, or several of the ER, PR, and HER2 markers.

Type of breast cancer	Biomarker
ER+	Estrogen receptor
PR+	Progesterone receptor
HER2+	EGF receptor
Triple negative	None of the three above

***Fig 1.** The four types of breast cancer*

Most breast cancers are treated by surgery, radiotherapy and chemotherapy. Cytotoxic chemotherapy aims at killing rapidly dividing cells, especially tumor cells. In addition, ER+, PR+ or HER2+ cancers can benefit from specific therapies: hormonal type for ER+ and PR+, and antibody type for HER2+.
These targeted therapies explain why these cancers have a better survival prognosis than triple negative breast cancers.

Because the "triple negative" class is affected by default, it is likely that it does not constitute a homogeneous set of tumors, but that it may contain different subclasses with different markers (thus different therapeutic targets) not yet discovered.

1.2 Screening of TNBC drugs

The team of Fabien Reyal from the Institut Curie has screened 1000 drug candidate molecules in order to identify drugs capable of killing triple negative breast cancer cells, but not the healthy cells (which would be too toxic to constitute a drug). 80 molecules were thus selected, for which certain targets are known.

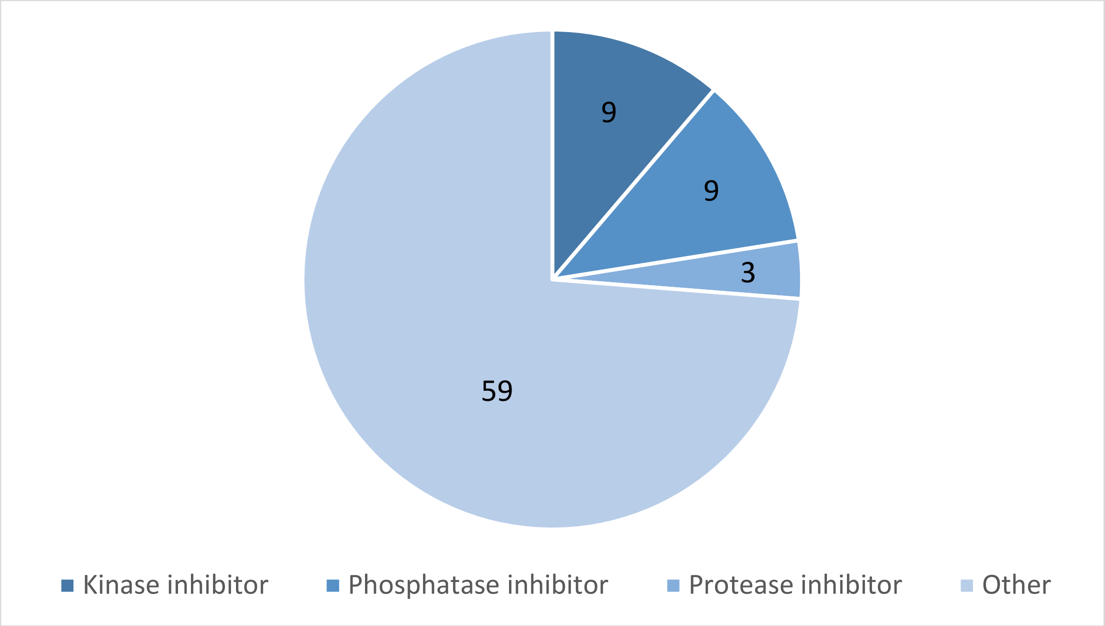


Fig 2. *Repartition of the 80 hits screened*

1.3 Subject

The research work consisted in predicting all the potential targets for these 80 molecules in order to identify specific markers of TNBC and their mechanisms of action. It actually appears to be a reverse problem. Traditionally, the aim of such research subjects is to find new drugs to fight cancer. In this case, the drugs have been identified. They are known to be effective against tumor cells. We aim at understanding their ways of working in order to get a clearer comprehension of how TNBC is expressed and how to tackle it.

The screening of drugs is a long and complicated process. The originality of this work is the method employed to get around this issue. The number of targets per hit is indeed very limited and does not provide enough perspective of biological analysis. We want to add more proteins to the target profiles of hit molecules. This is the reason why the digital prediction of interactions between drugs and human proteins, easier and more flexible, is very interesting. It enables to expand the spectrum of potential targets, in order to make the analysis simpler and more robust. It is based on a simple hypothesis: similar molecules will interact with similar proteins.

This work was conducted with the CBIO, which has developed chemo-genomic approaches based on Machine Learning algorithms and capable of predicting large-scale protein-ligand interactions in the space of molecules and proteins.

Ultimately, this project will contribute to better define the class of TNBC cancers, by distinguishing new subtypes sharing similar biological characteristics, better understand the underlying cancer processes, and propose targeted therapeutic alternatives for these cancers.

2. Method and Materials

2.1 Prediction of drug-target interaction

The idea of the computational method used to determine the existence or the absence of interaction between drugs and proteins is to separate the couples drug-protein interacting from the ones which do not interact. Because this separation is not trivial, data must first be transformed into a space where the prediction problem is easier to solve. Similarity based methods make it possible. We remind that this work is based on the idea that similar compounds interact with similar targets.

2.1.1 Numerical encoding of molecules and proteins

With the computational manipulation and analysis of molecules and proteins naturally comes the problem of their representation. Numerous descriptors and similarity measures exist in order to numerically represent chemical compounds. The best suited descriptor both provides usable data for the chosen method and is not too complex to manipulate. For example, 3D descriptors do exist, but are heavy to handle. In this case, where the prediction of interaction does not come from a docking method, there is no need to operate on such data. Moreover, our goal is to make large scale predictions, which would be laborious with 3D descriptors.

The encodings of molecules and proteins are different. Molecular compounds are mostly represented by string notations or by graphical representations of their Lewis formula, whereas proteins are mainly represented by their amino acid or genetic sequences.

In our case, drugs were encoded with their SMILES (Simplified Molecular Input Line Entry Specification), which is a line string notation for describing the structure of chemical species and proteins with their genetic sequences.

Here is an example of the SMILES of a molecule in comparison with its skeletal formula.

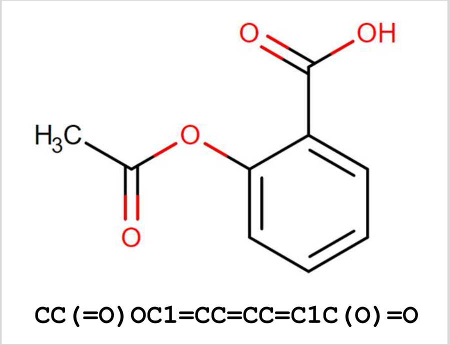


Fig 3. *An example of SMILES: Aspirin*

2.1.2 Dataset

Several bio-activity databases exist and are freely available online. Some provide quantitative measurements while others, like the one used during this work, provide binary information about interactions between molecules and their targets.

The data used to train the algorithm comes from the DrugBank, which is a free-to-access online database. It contains information on drugs, such as chemical and pharmaceutical detail, and on drugs' targets, such as the protein sequence and structure.

In the latest version of DrugBank (released 2021-01-03) can be found 14,470 drugs linked to 5,251 protein sequences.

Nevertheless, all the information available on the DrugBank was not used in this work. A smaller number of proteins and drugs were used to train the algorithm. The selection was made depending on the characteristics of compounds and targets. Only drug-like molecules and human proteins were kept. This selection led to the training set described below.

Training set of interactions	
Proteins (genetic sequences)	2670
FDA-approved molecules (SMILES)	5 070
Interactions	17 000

Fig 4. *Description of the dataset*

2.1.3 Similarity measures

Similarity based methods provide an efficient strategy to solve problems of data representation. The general idea is to create a new space where classification of data becomes possible. This is the purpose of the Kernel method used in the pipeline developed by the CBIO. The classifier is based on data provided by such method.

A kernel K is a symmetric positive definite function: $X * X \to R$ that can be viewed as a similarity measure. It compares protein and molecule pairs and returns a real number.

For a dataset of n molecules and p proteins, the results of Kernel applications are two matrixes:

- $K_{molecule}$ which is the $n * n$ matrix of similarity for the drugs
- $K_{protein}$ which is the $p * p$ matrix of similarity for the targets

For each matrix, $K_{i,j}$ is equal to the similarity between the molecule i and j determined by the Kernel function.

Given molecule and protein Kernels, we can define the Kronecker Kernel K_{pair} by:

$$K_{pair}\big((m,p),(m',p')\big) = K_{molecule}(m,m') \times K_{protein}(p,p')$$

with m for molecule and p for protein

K_{pair} is a $np * np$ matrix that captures the interactions existing between the features of the drugs and the features of the targets. It is the final and most interesting spatial representation of the status of molecule-target interaction.

2.1.4 Classification – Support Vector Machine

Support Vector Machine (SVM) is a supervised learning model. It analyzes data for classification and regression analysis. In this case, this method is meant to determine the optimal hyperplane that separates the pairs $(m,\, p)$ that interact from the ones which do not interact. This hyperplane is identified with the Kronecker Kernel K_{pair} constructed with the training dataset previously presented.

Given one drug and one protein, the prediction of interaction is calculated with the distance and the position of the pair (drug, protein) regarding the separating hyperplane.

2.2 Pathway enrichment analysis

The prediction of the hits' targets does not mark the end of the work. Indeed, in addition of identifying TNBC markers, the aim of this study is also to determine the mechanisms of action of the drugs in order to understand how TNBC is expressed and how to develop adapted specific treatments. Moreover, the number and the diversity of the proteins predicted with high scores (high probabilities of interaction with hits) do not allow straightforward biological interpretation.

This is the reason why we proposed to interpret the results with a method of pathway enrichment analysis. A biological pathway is a series of interactions between the molecules in a cell. This series leads to certain products or changes in the cell, that can be for example assembling new molecules, turning genes on and off, or making a cell move.

2.2.1 Principle

The method of pathway enrichment analysis consists in a statistical test that identifies the pathways that are more enriched in a list of proteins than they would be by chance. Put in other words, we searched whether the proteins targeted by the hits screened preferentially belonged to certain biological pathways which would be critical for TNBC survival.

2.2.2 Analysis tools and database

We used the public web server g:Profiler to realize the analysis. It is a server that has been created for characterizing and manipulating gene lists. It is currently available for more than 400 species, including mammals, plants, fungi and insects.

g:Profiler offers numerous tools. Among them, the g:GOSt module in particular was particularly suited for our analysis. It is actually the core of the server and performs enrichment analysis on gene lists that the user provides.

The analysis can be based on several databases of biological pathways. In our case, we used the database Reactome. This is an online, free-access pathway database that is hierarchized in sub-pathways. This is an interesting feature of the database because it enables to assemble proteins that are not directly linked to one another in one specific pathway, but that are belonging to pathways intervening in the same global area of action.

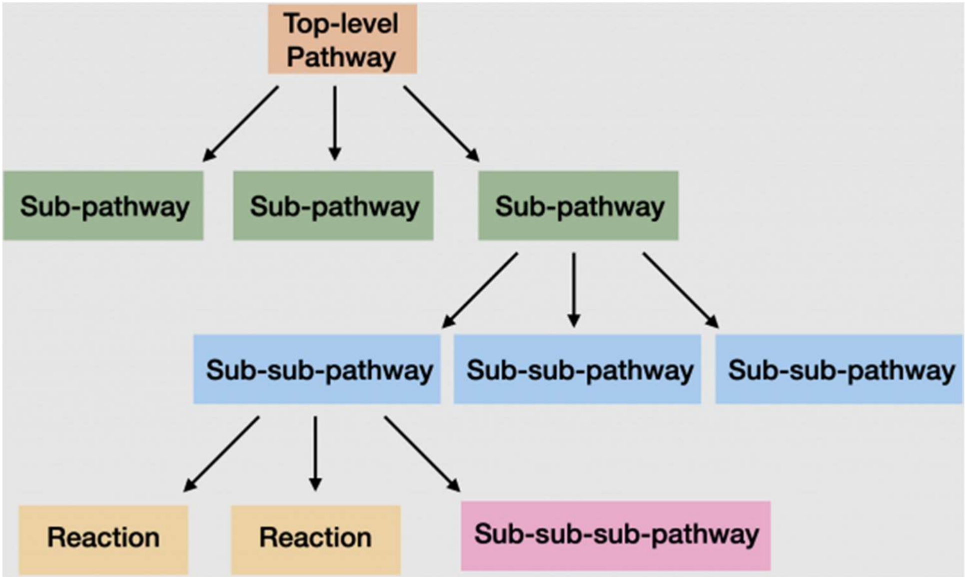


Fig 4. *Basic structure of the Reactome database*

3. Implementation and Results

3.1 Use of the CBIO pipeline

3.1.1 Difficulties

The CBIO pipeline, whose mode of operation was previously explained, has been under development for several years. It is the fruit of the work of different PhD students and was therefore, although absolutely necessary for the work, really difficult to understand. This was the first and main issue at the beginning of the trimester: take in hand one's code and use it.

Furthermore, given the scale of the data at stake and the methods employed in the pipeline, the use of the Mines cluster was indispensable, as soon as the project had started. This was the second issue encountered. The use of the cluster is not trivial and slowed down the advancements in the first weeks.

3.1.2 Prediction of targets

Given one drug at its entry, the CBIO pipeline predicts the probabilities of interactions between this drug and the entire dataset of proteins, which represents 2670 predictions. For each and every one of these potential targets, the probability of interaction is calculated as explained in 2.1.

There are two possibilities of usage of the pipeline in order to predict the targets for one specific drug:

- The drug already belongs to the training dataset:
  It is not usual to make prediction for data that is already in the training set. But the idea here is to expand the target spectrum for every drug of the hit list. Therefore, we have got to consider the interactions already existing.
  The implementation of the predictions in this case is simple. The algorithm already has access to the molecule's SMILES.
- The drug does not belong to the training dataset:
  The molecule is called "orphan". There is no known interaction with any protein.
  In this case, we have to provide the pipeline with the specific SMILES of the drug to enable the predictions.

3.2 Results

3.2.1 Results of the predictions

The results of the predictions are presented in the form of tables.

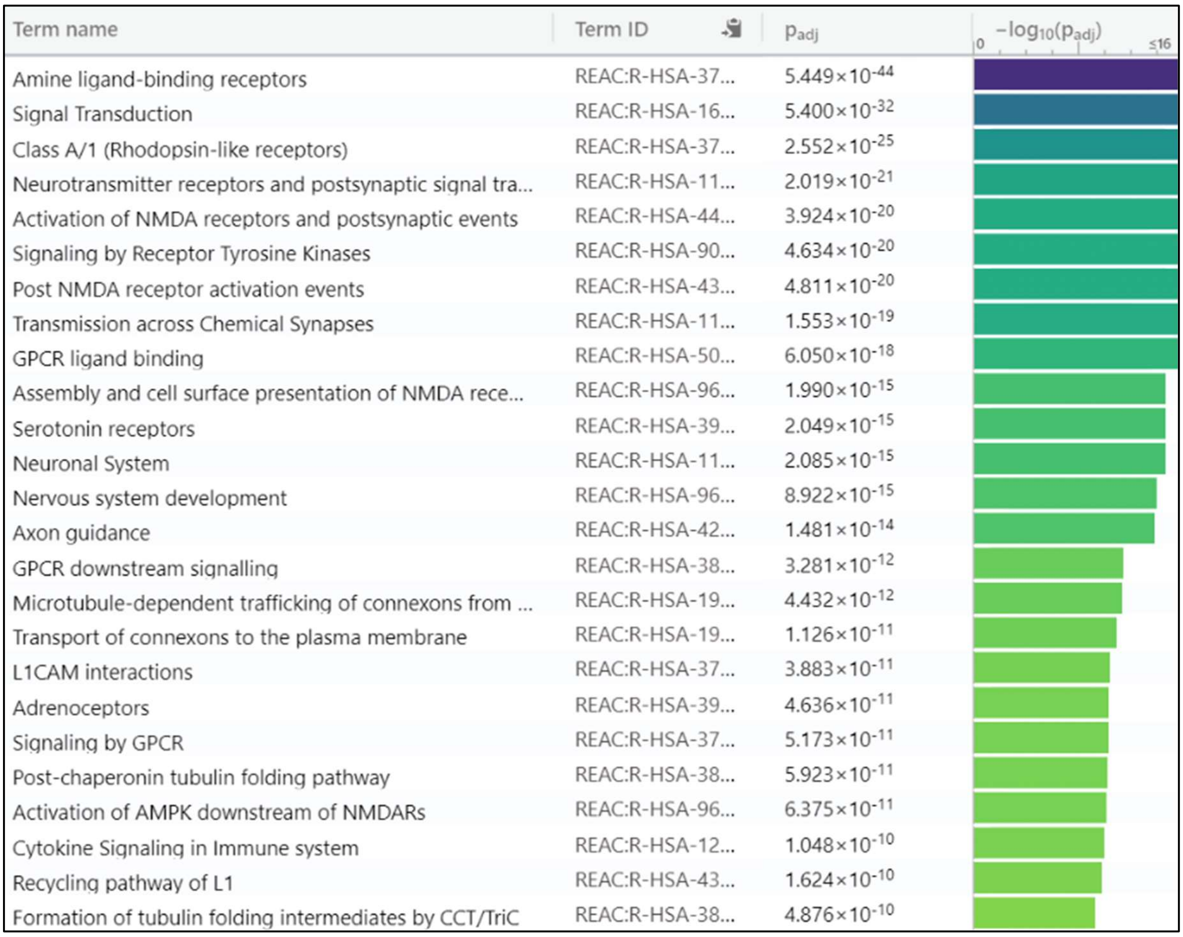
Uniprot ID	Gene	Name	Prediction	New DTI
Q96LZ3	PPP3R2	Calcineurin subunit B type 2	0.990354753	False
P49069	CAMLG	Calcium signal-modulating cyclophilin ligand	0.990353105	False
P62937	PPIA	Peptidyl-prolyl cis-trans isomerase A	0.990351146	False
P30405	PPIF	Peptidyl-prolyl cis-trans isomerase F	0.990345544	False
P63098	PPP3R1	Calcineurin subunit B type 1	0.989843502	True
O43447	PPIH	Peptidyl-prolyl cis-trans isomerase H	0.902588651	True

Fig 5. *Top predictions for Cyclosporine (hit)*

Given one drug, the table of prediction returned by the algorithm is similar to the Fig 5. It contains 2670 rows. Indeed, each row corresponds to one human protein of the dataset. The proteins are ranked by decreasing probability of interaction. They are characterized by their names, their IDs, their genetic sequences and by whether or not they are known to be interacting with the given drug. The last column in indeed a binary label that distinguishes the known existing interactions and the ones entirely predicted. A protein does not correspond to a new Drug-Target Interaction, and is therefore "False", if the interaction has entirely been predicted. And vice-versa. It is reassuring to see that the proteins that are indeed known to interact are predicted with the best probabilities. It is also reassuring to see that there is no tremendous gap between the prediction of real and virtual interactions. This is a hint that the algorithm does not suffer from over-fitting and that the high-predicted results are robust and suitable for analysis.

3.2.2 Results of the pathway enrichment analysis

After having predicted the potential targets for all the molecules in the hit list, we decided to only keep the best results, meaning all the proteins that had the best chances to interact with at least one of the hits. This selection was enabled by the use of a threshold that we set at 0.9 (probability of interaction). This led to a list of 440 proteins that was food for g:Profiler analysis.



Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$ 0 ≤16
Amine ligand-binding receptors	REAC:R-HSA-37...	5.449×10^{-44}	
Signal Transduction	REAC:R-HSA-16...	5.400×10^{-32}	
Class A/1 (Rhodopsin-like receptors)	REAC:R-HSA-37...	2.552×10^{-25}	
Neurotransmitter receptors and postsynaptic signal tra...	REAC:R-HSA-11...	2.019×10^{-21}	
Activation of NMDA receptors and postsynaptic events	REAC:R-HSA-44...	3.924×10^{-20}	
Signaling by Receptor Tyrosine Kinases	REAC:R-HSA-90...	4.634×10^{-20}	
Post NMDA receptor activation events	REAC:R-HSA-43...	4.811×10^{-20}	
Transmission across Chemical Synapses	REAC:R-HSA-11...	1.553×10^{-19}	
GPCR ligand binding	REAC:R-HSA-50...	6.050×10^{-18}	
Assembly and cell surface presentation of NMDA rece...	REAC:R-HSA-96...	1.990×10^{-15}	
Serotonin receptors	REAC:R-HSA-39...	2.049×10^{-15}	
Neuronal System	REAC:R-HSA-11...	2.085×10^{-15}	
Nervous system development	REAC:R-HSA-96...	8.922×10^{-15}	
Axon guidance	REAC:R-HSA-42...	1.481×10^{-14}	
GPCR downstream signalling	REAC:R-HSA-38...	3.281×10^{-12}	
Microtubule-dependent trafficking of connexons from ...	REAC:R-HSA-19...	4.432×10^{-12}	
Transport of connexons to the plasma membrane	REAC:R-HSA-19...	1.126×10^{-11}	
L1CAM interactions	REAC:R-HSA-37...	3.883×10^{-11}	
Adrenoceptors	REAC:R-HSA-39...	4.636×10^{-11}	
Signaling by GPCR	REAC:R-HSA-37...	5.173×10^{-11}	
Post-chaperonin tubulin folding pathway	REAC:R-HSA-38...	5.923×10^{-11}	
Activation of AMPK downstream of NMDARs	REAC:R-HSA-96...	6.375×10^{-11}	
Cytokine Signaling in Immune system	REAC:R-HSA-12...	1.048×10^{-10}	
Recycling pathway of L1	REAC:R-HSA-43...	1.624×10^{-10}	
Formation of tubulin folding intermediates by CCT/TriC	REAC:R-HSA-38...	4.876×10^{-10}	

Fig 6. *Top predictions for enriched pathways*

The Fig 6 represents the top predictions for the analysis run on g:Profiler. The pathways are ranked by their probability of being enriched (and therefore, by the result p of the statistical test run in pathway enrichment analysis) with no regard toward the levels of the pathways in the hierarchy of Reactome. Moreover, this last feature of the ranking is the main reason why this representation of the results still is not easily readable.

4. Analysis and Perspectives

4.1 Analysis of the results

As observed in 3.2.2, the mere fact of having generated the ranked list of enriched pathways does not enable straightforward biological analysis. Indeed, pathways, sub-pathways, sub-sub-pathways (and so on) are put on the same level and so, create a large amount of correlated data. The simple solution that we chose in order to reduce the number and the diversity of the outcomes was to regroup pathways related to the same top-level pathway.

4.1.1 Expected pathways

Two of the most represented top-level pathways among the outcomes of the pathway enrichment analysis are expected for proteins involved in cancer. Those are:

- Signal Transduction:

  Signal transduction is the mechanism by which a cell responds to the information it receives from its environment via receptors on its surface. This mechanism controls a cascade of secondary signals, internal to the cell or external (e.g. actions on other cell types), and internal cellular processes (metabolism, cell cycle). This allows the cell to live in context.

  A tumor cell has difficulty living in context. Its signaling pathways are disrupted. The cell does not respond or responds badly to its environment. It is therefore not surprising that the “Signal Transduction” pathway is at work in the expression of cancer.
- Developmental Biology:

  Developmental biology refers to all the processes by which organisms grow and develop. It includes in particular the genetic control of cell growth and cell differentiation.

  Yet, a particularity of the tumor cell is that it loses its differentiation by dividing in an uncontrolled, very rapid and disorderly manner. While a healthy cell keeps its characteristics, a cancer cell tends to lose a part of them. This is why developmental biology is relevant to the study of cancer.

It is reassuring to observe such pathways. This means that some of the hits have known or predicted targets involved in pathways relevant to cancer in general.

4.1.2 Unexpected pathways

Unexpected pathways are actually the best research tracks such analysis can lead to, because it widens the classic spectrum of research in the field. We remind that the purpose of this work is to better describe how TNBC is expressed. Indeed, describing it better means understanding it better and therefore treating it better. This is why original, specific markers and pathways are very much more interesting: the top-level pathway “Neuronal System” was very less expected than the two pathways presented in 4.1.1. It means that some TNBC drugs’ targets could be related to pathways normally active in the neuronal system.

Proteins from this system could be expressed in TNBC, and could play a role in the deregulation of the cells. Therefore, proteins in these neuronal pathways could be therapeutic targets to treat TNBC. These proteins in question may or may not be targets of the molecules in the hit list. Due to the fact that a pathway is a series of interactions between molecules, there are potentially several levels in the pathway at which treatments could aim, not only the levels identified as targets for hits.

4.2 Perspectives

4.2.1 Continuation of the work

The next step of this work would be to verify whether proteins intervening in neuronal pathways are expressed in TNBC, and not in corresponding healthy cells.

As mentioned before, we may have identified some targets for the hits. But they might not be the most interesting proteins to target in order to treat cancer. Therefore, we should look for other key proteins in the pathways. The interest of proteins that have a role in the neuronal system must be evaluated. Their activity could play a role in cell proliferation, which is a main feature of cancer expression. And if key proteins were to already be well known, active molecules and treatments may already exist on the market.

This could lead to drug repositioning, which is a much faster and less expensive process than the development of a new molecule. However, treatments targeting the neuronal system will certainly have an effect on the brain. There is naturally a benefit-risk balance to be found.

We still need to understand how drugs work in TNBC, but in the first instance it is enough that it works. The study of optimized therapies comes only afterwards. An idea for example would be to design molecules that would treat cancer in the breast without crossing the blood-brain barrier and acting on the brain.

4.2.2 Suggested improvements

The algorithm from The CBIO has been developed and improved for several years now by qualified researchers. Therefore, none of the suggested improvements will be about the very core of the method.

A first track to work on is the development of the training set. Indeed, the algorithm is much more efficient in predicting the interaction of "non-orphan" drugs. Numerous databases exist similar to the Drugbank that was used during this trimester. They would allow to have a larger and more suitable dataset on which to train the pipeline.

A second track of reflection concerns the processing of prediction results obtained through the CBIO pipeline. The selection of proteins via a threshold has admittedly allowed us to obtain results at the end of the search for enriched pathways, but it can certainly be improved. Indeed, the basis of this work is based on the postulate that similar molecules interact with similar proteins. If we take this reflection further, similar molecules are likely to interact with targets that have similar functions in similar pathways. The search for enriched pathways would probably be more refined if it were based on the target proteins of coherent groups of molecules rather than on the target proteins of the whole set of hits. The issue remains, however, to define and form these coherent groups.

A first idea could be to gather molecules belonging to the same large families of molecules. This was tried during this study, by testing pathway enrichments for the targets of kinase, phosphatase and protease inhibitors. Unfortunately, the number of molecules in each family was too small and it did not allow to obtain satisfactory results.

A second idea could be to perform clustering. Interaction clusters could indeed be identified (via similarity calculations). Coherent groups of molecules would then be the hits belonging to the different clusters. In addition, clustering may allow the identification of different cancer types within triple negative breast cancer, which remains a goal of this study.

5. Conclusion

Given the cost and time required for drug screening and the criticality of the problem, not only of breast cancer, but particularly of triple negative breast cancer, original and innovative methods seem to be necessary to make rapid progress in this field. It is in this context that this 3-month work was carried out. Indeed, the originality of the approach used in this specific case must be emphasized. First of all, we recall the inverse nature of the proposed problem resolution. We know how to treat the cancer in question since hits have been identified. But understanding how this cancer is expressed allows us to develop more adapted treatments, which we could not necessarily foresee beforehand. However, the originality of the work does not only lie here, but also in the very innovative methodology. The virtual prediction of interactions in order to determine the targets of active molecules is a very clever approach to circumvent the problem of laboratory tests, by proposing a flexibility allowed by chemogenomic approaches based on Machine Learning algorithms. However, the originality does not stop there. Since expanding the spectrum of target proteins for hits is not enough to perform a biological analysis, statistical tests (pathway enrichment analysis) are then performed on results already obtained by other mathematical methods.

The results must therefore obviously be analyzed with considerable hindsight. However, many observations corroborate a coherent functioning of the whole pipeline and thus a robust prediction of the pathways targeted by the hits, thus the place of expression of the TNBC. We can mention in particular the presence of pathways known to be involved in cancer in the final predictions.

This work opens up new avenues for further research in the treatment of TNBC, both in the laboratory and in bioinformatics. In addition to proposing interesting and promising results (in particular the discovery of the involvement of neuronal system pathways), this study has above all demonstrated the robustness of the methodology used, although it could seem original.

May the continuation of this work allow great advances in the understanding of triple negative breast cancer and its treatment.

Acknowledgments

I would like to thank the people with whom I was in contact during the whole trimester, who helped me, as much in the biological understanding of the problem and the results as in the understanding of the bioinformatics method used, and who knew how to guide me throughout my work so that it would be as prolific as possible. I would like to thank especially Matthieu Najm, Pr. Véronique Stoven and Dr. Chloé-Agathe Azencott.

I would also like to thank the people who organized this R-SANTE research trimester and the members of the CBIO who made this project an incredible experience to discover the research environment.

References

[1] Playe, Benoit. *Machine learning approaches for drug virtual screening*. Diss. Université Paris sciences et lettres, 2019.

[2] Sadacca, Benjamin. *Pharmacogenomic and High-Throughput Data Analysis to Overcome Triple Negative Breast Cancers Drug Resistance*. Diss. Université Paris-Saclay, 2017.