Evaluation of the public databases' relevance as comprehensive tools for pharmacological mechanisms

Pierre Beaudier

Internship report

2018/2019

Evaluation of the public databases' relevance as comprehensive tools for pharmacological mechanisms

Supervisors:

Olivier DAMERON

Adeline DUCHENE

Intern:

Pierre BEAUDIER

UMR CNRS 6290
Institut de Génétique et Développement de Rennes (IGDR)

Annabelle MONNIER
Equipe Oncogénomique translationnelle

UNIVERSITÉ DE RENNES 1

ENGAGEMENT DE NON PLAGIAT

Je, soussigné (e) Pierre Beaudier
Etudiant (e) en Master Bioinformatique

Déclare être pleinement informé (e) que le plagiat de documents ou d'une partie de documents publiés sous toute forme de support (y compris l'internet), constitue une violation des droits d'auteur ainsi qu'une fraude caractérisée.

En conséquence, je m'engage à citer toutes les sources que j'ai utilisées pour la rédaction de ce document.

Signature

Laboratoire de Génomique Médicale
BMT-HC - CHU Pontchaillou

2 rue Henri le Guilloux
35033 Rennes Cedex
FRANCE

Annabelle MONNIER
annabelle.monnier@univ-rennes1.fr
TÉL. 33 (0)2 99 28 92 54

Acknowledgements

I thank Olivier DAMERON and Adeline DUCHENE for their help, advice and supervising during this internship.

I thank the GenOuest bioinformatics core facility (http://www.genouest.org) for providing access to their cluster

Table of contents

I) Introduction

1. Central nervous system diseases and glia

Central nervous system (CNS) diseases are neurological disorders affecting the function or the structure of the brain and the spinal cord. There is a great number of different types of CNS diseases, spanning from addiction to neurodegenerative disorders. The treatments for those diseases often consist in medication. However, drugs affecting the brain usually come with important and frequent adverse effects or lack of response, varying between individuals [12].

A combination of a psychotropic and a secondary drug meant to either remove the adverse effects or enhance the psychotropics' efficacy can be considered as a potential solution for this problem. The choice of the secondary drug requires to find the target that is responsible for the psychotropic's limitations. Over recent years, some psychotropic have been determined to have secondary effects on the glia [13], potentially outlining a new type of combination between a neuronal target drug and a glial target drug.

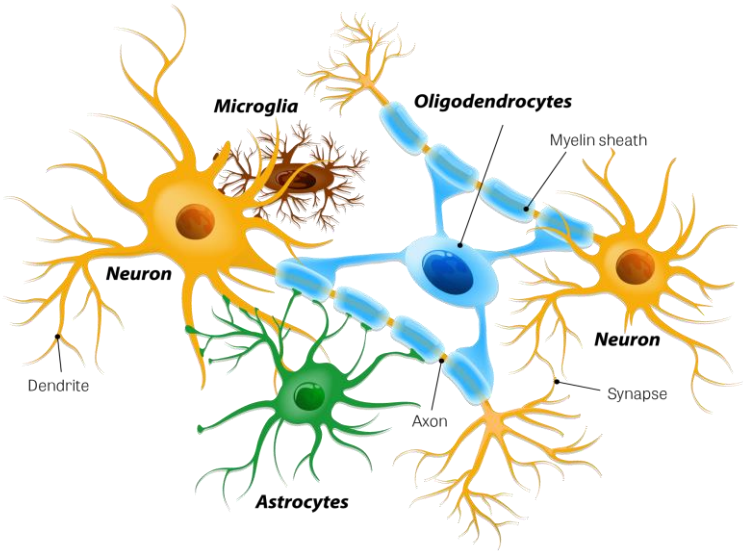

Figure 2. Glial cells – neurons interactions. [14]

Glial cells are the neuron's environmental cells. They constitute around 50% of the brain volume. They are mainly divided in 3 different types: microglia, astrocytes and oligodendrocytes [15]. While they were originally thought to only support neurons passively,

the recent decades have proven that assumption to be false. The glial cells have been found to control multiple aspects of the development and wiring of the brain, as well as providing the metabolic support to neurons and regulating cell-cell interactions [16]. Despite these many discoveries, many of the mechanisms behind their function have yet to be discovered.

2. Drug Repurposing/Repositioning

The development of a new drug is an expensive, complex and lengthy process costing up to billions of dollars and taking up to decades for a single drug [1]. Furthermore, developing a drug is always a risky venture as the success rate for the FDA (USA) evaluations is only around 10% [2] and once the drug is fully developed and marketable, there is still a risk of commercial failure.

Drug development, while it has the potential to be highly profitable, can fail on many steps of the way and is never guaranteed success [3].

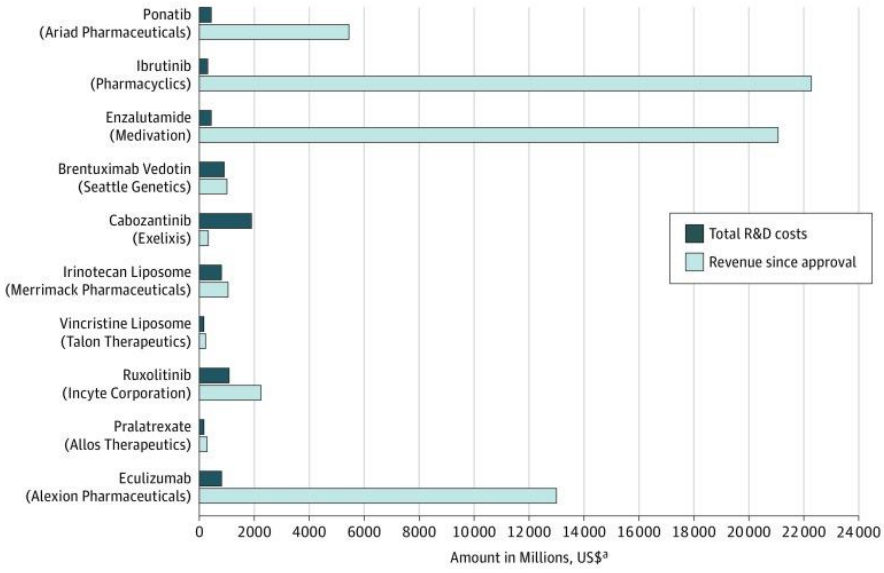

Figure 1. Comparison of Drug Development Costs with Revenue Earned After Approval. [4]

In an effort to reduce the risks and the costs, drug repurposing (also called drug re-tasking, drug repositioning or drug reprofiling) offers an alternative to the classic approach of drug development [5]. Drug repurposing is the process in which already registered and studied drugs will be addressed to another indication. This method allows drug companies to bypass

most of development costs as well as having more confidence for safety studies as those have already been conducted for the first commercialization of the compound [1].

This process started by empirical discoveries of drugs effects on pathologies they were not supposed to affect. It has now evolved into the use of bioinformatics tools to determine new potential substances usages [6].

Those bioinformatics tools use computational approaches to predict novel uses by exploiting various data such as gene expression profiles, protein targets, pharmaceutical mechanisms, etc [7] [8]. The information is exploited by creating networks of similarities between drugs thus determining which drugs could potentially find new uses [9]. However, those tools are not foolproof. Most of the data generated will be predicted but not proven, therefore the results always have to be considered with caution and be considered only as potential ways to explore via future experiments.

3. Drugs combination

The combination of multiple drugs is an old practice that can be traced back to millennia ago. Those drugs were developed to target a single disease at first but it has recently begun to be used to treat multiple diseases at once [10].

The many discoveries in biology these past decades allow for a better understanding of diseases and of the drug used as treatment, allowing science to progress from empiric discoveries to theoretical possibilities. It is now possible to determine if a disease can be better treated by targeting one or multiple targets and the drugs secondary effects can be modulated as well [11]. This opens the opportunity for drugs combination which offers multiple benefits compared to simple drugs such as:

- Increased efficacy: one drug enhancing another one by blocking the elimination process.

- Increased efficacy: one drug enhancing the properties of the other without modulating its PK parameters. The only interaction would be pharmaco-dynamic (additional effect).

- Decreased toxicity: one drug counter-acting the negative secondary effects of another drug.

4. Rationale and Objective

The aim of this internship is to develop a tool to predict potential drug combinations for brain pathologies.

An important proportion of the drugs used to treat central nervous system diseases target neurotransmission while also impacting the glia. We make the hypothesis that this secondary effect is be detrimental but can be modulated by a secondary drug that could be combined with the psychotropic.

This tool would look for a selected neuronal drug's secondary effects on the glia by exploring the multiple databases available online and extracting the recorded drug interactions. From those interactions their pathways' contents are selected and filtered to obtain as output the name of the potential glial targets that could be modulated through a second drug.

The first part of the internship consisted in finding databases containing potentially relevant data to our objective to then gather data from the different sources chosen in RDF format. For some of those databases, the data are freely accessible online and need no further work but other databases require to download their data and re-create the database locally, sometimes needing to be converted to the appropriate RDF format as well. The second part consisted in exploring the gathered databases with SPARQL queries and creating an outline of all the information that needs to be extracted to make an appropriate selection. This selection will then remove irrelevant data to our study to retain only glia results.

We focused on Modafinil, a drug used to treat excessive daytime sleepiness in narcoleptic patients, as a proof of concept. An interaction between Modafinil and two glial targets, the connexins 30 and 43, has been discovered in the past years [17] but is not indicated in drugs or targets databases. As such, finding the connexin in the results, while not a foolproof validation, would serve to add credit to this method.

II) Materials and methods

1. RDF/SPARQL

Resource Description Framework (RDF) is a model of graph designed to store data and metadata. It was developed by the World Wide Web Consortium (W3C) and is the Semantic Web's main language. It is not formally considered a format but a data model. There are multiple RDF formats such as RDF/XML, Turtle, JSON-LD, OWL, etc.

RDF is based on the concept of a triple: subject, predicate and object. The predicate declares a relationship between the subject and the object thus creating a statement. A collection of these statements represents a directed multi-graph. The data stored under the RDF model can be accessed in two ways: locally or remotely. The remote access is done through endpoints, these endpoints are links to access the data stored within.

SPARQL is a query language used to retrieve and manipulate data stored in an RDF format. The name is a recursive acronym: **S**PARQL **P**rotocol **A**nd **R**DF **Q**uery Language. It is one of the major technologies of the Semantic Web which is constituted of hundreds of SPARQL services offering access to huge amounts of data, growing continuously [18].

Virtuoso is a query service allowing for SPARQL queries on multiple remote data endpoints at once. This type of query going through multiple databases is called a federated query. However, for those federated queries to work, it is necessary for the multiple endpoints used to possess common identifiers for compounds as to be able to connect them.

Unfortunately, this is not the case for all the databases that were used for this internship. The solution consisted in using Python to convert queries' results to make them useable by other databases with which there is no common identifiers.

Two Python packages were used to execute the SPARQL queries:

- RDFLib [19] is a library containing an RDF/XML parser and is used on local RDF files that could not be stored in endpoints.
- SPARQLWrapper is a wrapper around a SPARQL service. It is used to send queries on endpoints directly from Python and return the results in a easily useable format.

AskOmics [20] is a visual SPARQL query interface supporting both intuitive data integration and querying while shielding the user from most of the technical difficulties underlying RDF and SPARQL. In this project, it has been be used to convert databases that are not available in RDF but that are available in tsv/csv format. This software will convert files in such format into a RDF format.

AskOmics also offer the possibility to load RDF data into a local private endpoint.

2. Data used

We looked mainly for two types of databases:

- Drug databases from which we could extract the drugs recorded interactions as well as the sources documenting them.

- Target databases containing proteins, genes and pathways as to link the drugs' databases interactions and connect them to the genome and proteome.

3. SPARQL data

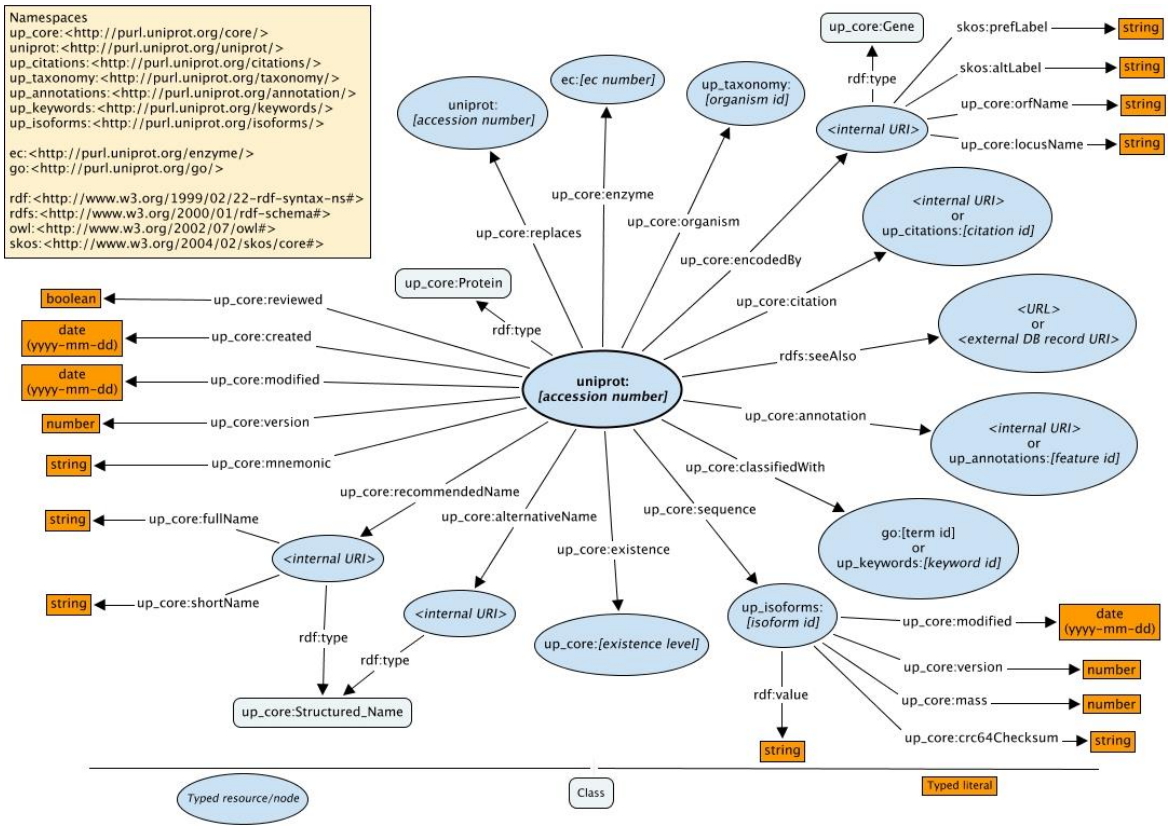

Figure 3. Structure of the Uniprot SPARQL endpoint. [21]

On this figure is the example of a classic RDF structure, in this case the Uniprot endpoint. From the Uniprot accession number, the entire data associated is accessible through the corresponding relations. If the accession number is not known, it is possible to obtain it if enough of the associated data is known as the triples in the RDF model work both ways.

For example, the entry P17302 is the entry for the connexin 43. You can obtain the mnemonic name with the following relation:

```
VALUES (?entry){up_core:P17302}

?entry     up_core:mnemonic     ?mnemonic
Subject        Predicate            Object
```

If the mnemonic name is known but no the entry, this relation can be used as well.

```
VALUES (?mnemonic){up_core:CXA1_HUMAN}

?entry     up_core:mnemonic     ?mnemonic
```

This is however one of the simpler structures as, like biopax3 which is used by the Ensembl endpoint. some can prove more complicated

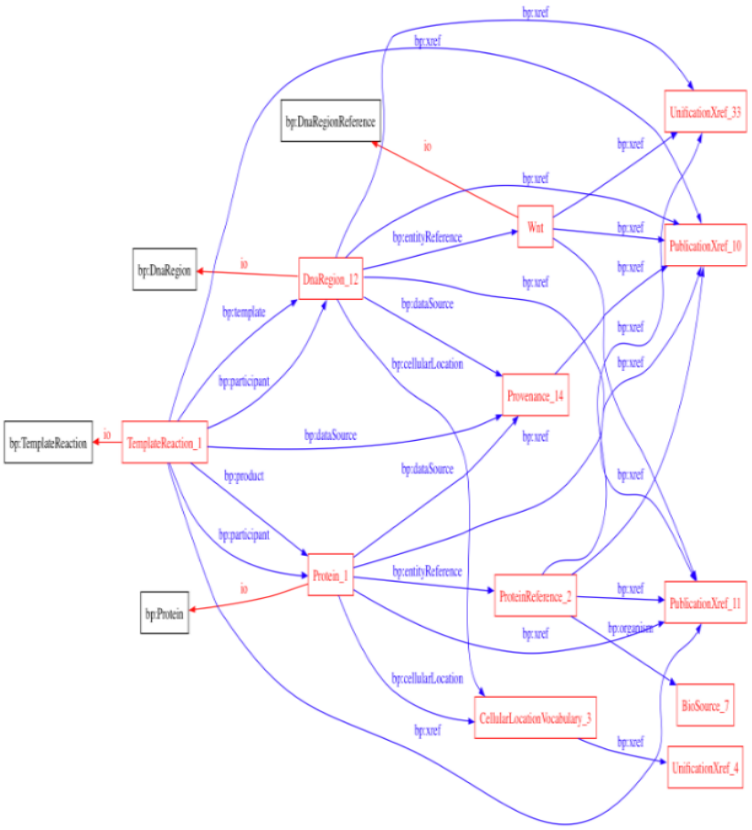

Figure 4. Example of data relationships using the biopax3 model. [22]

III) Results

1. Databases

a. Drug databases

The input of the pipeline being a drug, the first step is to find drug databases. The IDs from other databases as well as the known interactions (and their sources) will be extracted from those. The interactions are mostly genes and proteins interactions.

8 databases were considered:

Database	Description	Data of interest
DrugBank [23]	Freely accessible database, it contains information on drugs and drugs targets. There are 13 338 drug entries as of April 2019. It is one of the most widely used drug databases and naturally came up when we began searching.	Proteins interactions
PubChem [24]	Database of chemical molecules and their activities against biological assays. There are 97 million compounds entries.	Centralized data from multiple databases
Harmonizome [25]	Collection of information about genes and proteins from 114 datasets provided by 66 online resources.	Proteins and genes interactions
PharmGKB [26]	Pharmacogenics knowledge resource that encompasses clinical information including clinical guidelines and drug labels, potentially clinically actionable gene-drugs associations and genotype-phenotype relationships.	Proteins and genes interactions
KEGG [27]	Database resource for understanding high-level functions and utilities of the biological systems.	Proteins and genes interactions
ChEMBL [28]	Manually curated database of bioactive molecules with drug-like properties. It brings together chemical, bioactivity and genomic data to aid the translation of genomic information into effective new drugs.	Proteins and genes interactions

Wikidata	Free and open knowledge base that can be read and edited by both humans and machines. Wikidata acts as central storage for the structured data of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wiktionary, Wikisource, and others.	IDs for drugs and interactions
DGIdb [29]	Drug-gene interaction database.	Genes interactions

Out of those 8 databases, 2 were not used:

KEGG is one of the most complete databases available online and could have brought plenty of data on both drugs and targets. Unfortunately, it does not offer programmatic access and is not available to download for free. Therefore, it was taken out of the list of drugs databases.

PubChem gathers an important part of all other drugs databases as well as providing bioassays which could have brought new possibilities for target identification. However, the API offers only partial access to the database. It is available for download in RDF, and it is possible to re-create the database locally but it required a 500 Go SSD with 64 Gb of memory. The possibility of using the GenOuest platform to upload the database in the cloud was considered, however due to the short duration of the internship, this would have been a complicated and time-consuming process for only a few weeks of use.

It should be noted that this choice of database is not exhaustive, the main point of focus for this selection, after their availability, are the drug's interactions. The objective in gathering the maximum of databases is to have a maximum of known interactions, as those differ from database to database. Since there is no centralized database that gathers all the information from all the available sources, we are left with the only option of exploring as many databases as possible. With more time, more databases would have been selected and integrated.

b. Target databases

Next are the target databases, the objective there is to gather enough data on the most commons targets, being genes and proteins.

Three databases were selected:

Database	Description	Data of interest
Uniprot [21]	Freely accessible database of protein sequences and functional information	Proteins
Ensembl endpoint [30]	Genome browser for vertebrate genomes	Proteins/Genes
Wikipathway [31]	Database of biological pathways	Pathways

c. Database Availability

Database	Status
DrugBank	Downloadable in XML
PharmGKB	Downloadable in TSV
DGIdb	Downloadable in TSV
Wikidata	Endpoint freely accessible
Harmonizome	Outdated API
Uniprot	Endpoint freely accessible
Ensembl/ChEMBL	Endpoint freely accessible
Wikipathway	Endpoint freely accessible

PharmGKB and DGIdb were available to download freely, not in RDF but instead in TSV. Some small modifications were necessary, mostly in the columns' names but also in some of the data which contained unreadable symbols, to make the files readable by AskOmics. The software then converted the two databases in RDF. These were then accessible through the local endpoint created by the software.

DrugBank was downloadable in one XML file of 1.7 Gb. I developed a Python script using the minidom package, which contains an XML parser, was used to convert in an RDF format: Turtle.

```xml
<drugbank>
    <drug>
        <name>DrugA</name>
        <id>DB00001</id>
        <interactions>
            <drugs>
                <drug>DrugB</drug>
            </drugs>
            <proteins>
                <protein>ProteinA</protein>
            </proteins>
        </interactions>
    </drug>
</drugbank>
```

XML

```
@prefix drugbank: <https://www.irisa.fr/test#>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

drugbank:DrugA a rdf:type drugbank:Drug;
    drugbank:idIs 'DB00001'^^xsd:string;
    drugbank:interactsWithDrug 'DrugB'^^xsd:string;
    drugbank:interactsWithProtein 'ProteinA'^^xsd:string.
```

RDF (Turtle)

Figure 4. Example of XML and Turtle files containing the same information.

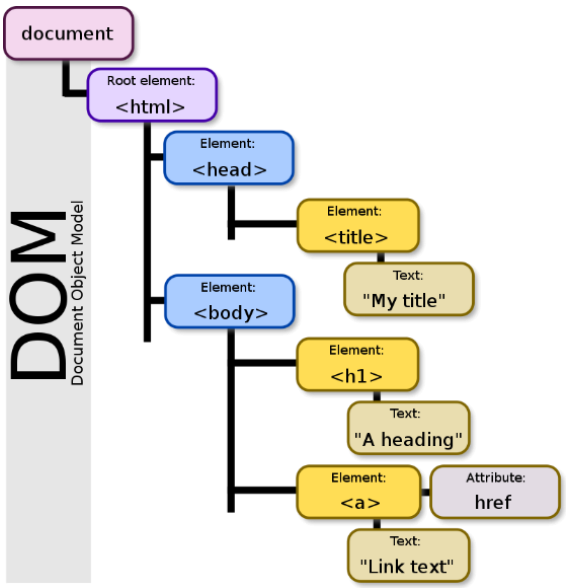

Figure 5. Structure of a DOM tree.

The parser will scan the entire file and extract a DOM tree on which it will be possible to select elements by name. The major inconvenient with this parsing method is that the file is being loaded up entirely and re-created in DOM format. While this is not a problem with small files, the DrugBank database weights 1.7 Gb which can be a problem depending on the computer's power.

From there, data of interest can be extracted with the function getElementsByTagName(). However, this function is recursive, meaning it will search for the given tag name in every subsection of the tree. The parser cannot make the difference between a drug tag referencing a Drug entity and a Drug-Drug interaction. The workaround here is to use the attributes provided in the tags. For example, a Drug entity tag will add information such as:

<drug type="biotech" created="01/01/2000">

Not all of the data provided was kept for the conversion as some of it was not necessary and would have slowed the conversion and led to heavier files which would have slowed the SPARQL queries

The following data was kept:

Identification: Name / DrugBank ID / CAS ID / ATC ID / External IDs

Interactions: Targets / Drug-Drugs interactions / Drugs-Proteins interactions / GO terms / PubMed References

The database was converted in two ways: one file for the entire database and one file per drug (15 000 files in total). The objective was to find which solution was optimal. The conversion took the same amount of time for both methods.

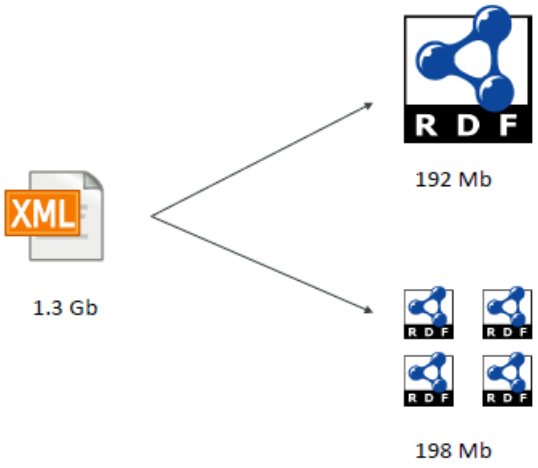

Figure 6. The two ways considered to convert DrugBank from XML to Turtle.

Converting into one file amounted to a smaller size since the prefixes at the beginning of each turtle file was only necessary once but the difference was only of 6 Mb (around 3%).

As for queries, they are much slower when looking for data on a single drug on the whole database rather than on one drug file (around 10 minutes against 6 seconds). This means the multiple files method is much more practical and was thus selected. However, the drawback is that since drugs are within separate files, it is not possible to execute a query on multiple drugs at once, instead requiring multiple queries whereas only one would have been necessary with the whole database in one file.

Harmonizome has an API that is programmable through Python, but the documentation is outdated making the API unusable. However, on each of the database's entity webpage, it is possible to retrieve the full page's data in JSON format. This was used by creating a Python script with the urllib package to query the URI leading to this JSON formatted data which could then be extracted.

At this point, all databases are accessible and connected. The SPARQL queries are carried out through the Python package SPARQLWrapper (with the exception of DrugBank for which the RDFlib package is used since the files are not located in an endpoint).

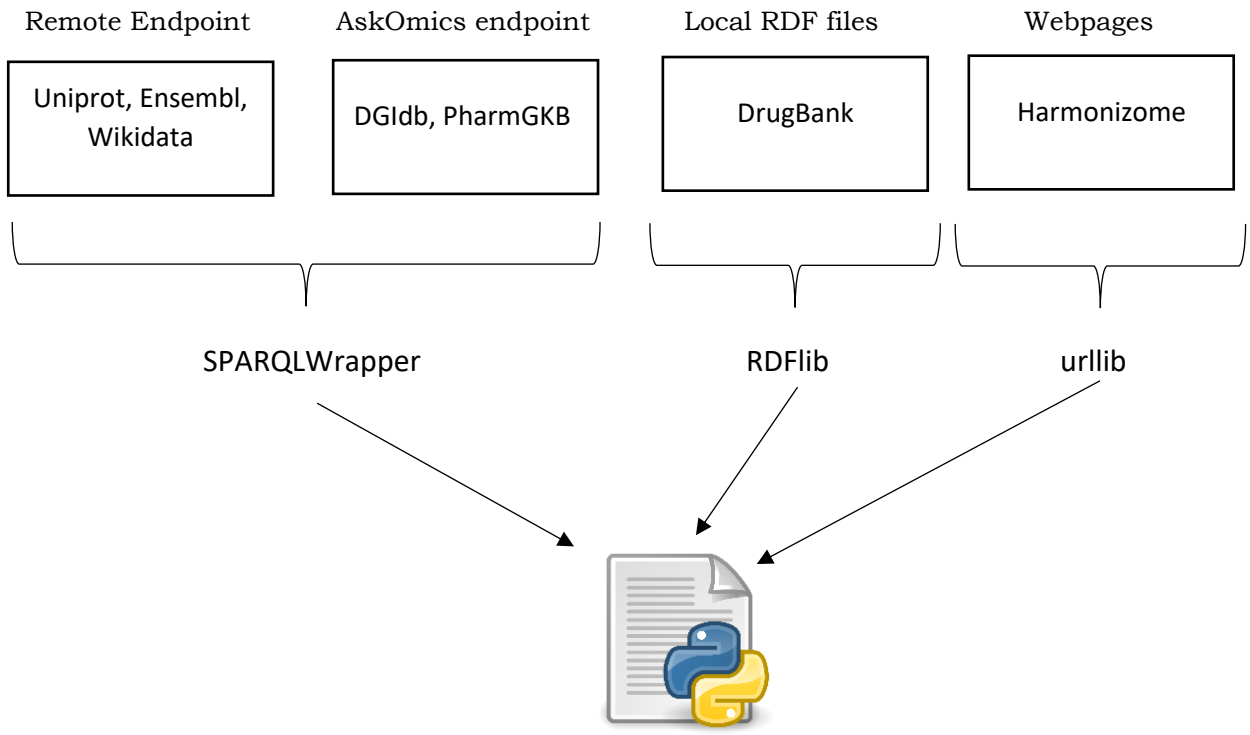

Figure 7. How the data was extracted

Python is used here for multiple reasons:

-It allows to execute multiple SPARQL queries at once

-Connecting some of the databases can prove tedious due to differences in the way data is stored. Python allows us to convert results from a query to make them readable to another database. The vast amount of packages also ensures that there is a way to use the data in whatever format it is.

-The access to Harmonizome was done through the designed Python API at first and some elements of this API have been recycled to be used with the URL parser.

Throughout the internship, there was no noticeable difference between the query time of SPARQLWrapper on Virtuoso and Virtuoso directly.

2. Connections between Databases

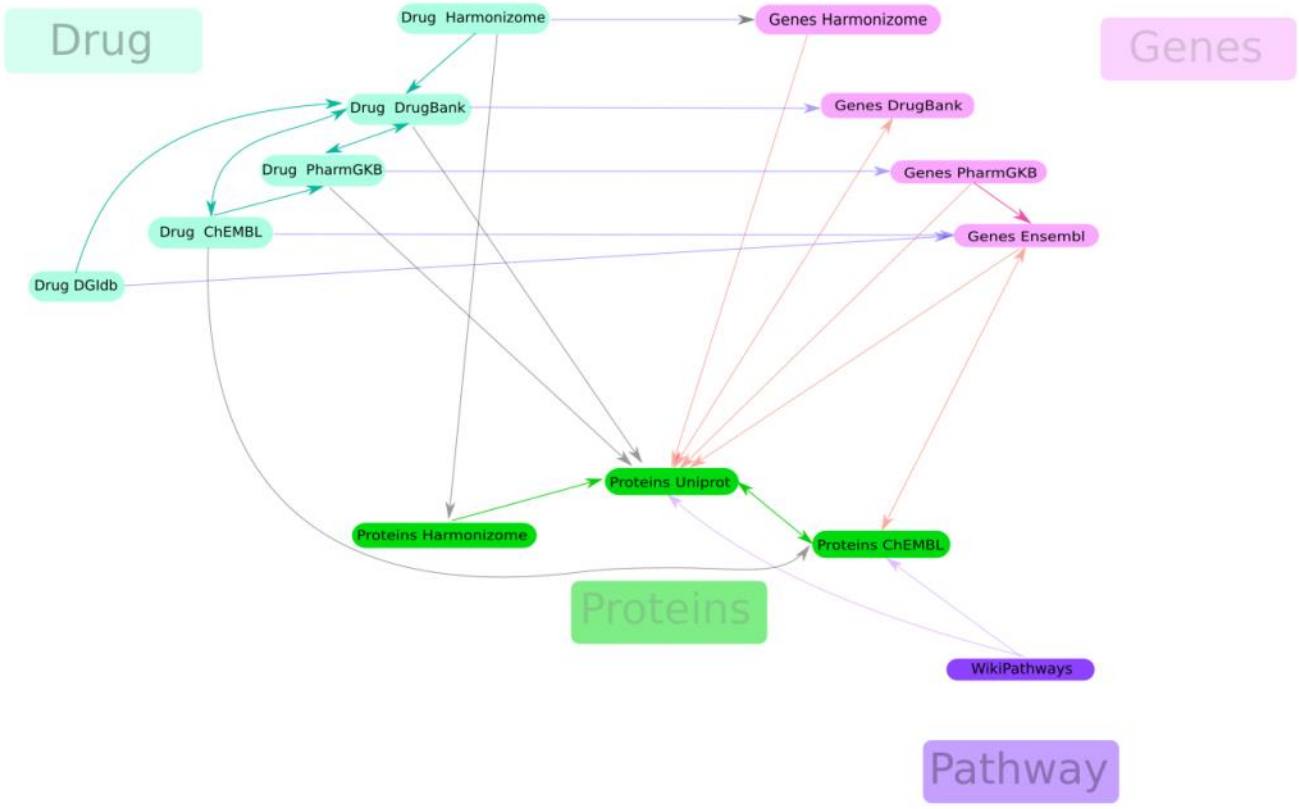

Figure 8. Connections between our list of databases.

A connection in the graph, means that it is possible to obtain from DatabaseA a matching ID for an element in DatabaseB. Technically, it is possible to go from almost any database to another simply with the element name but what is shown in the figure are the direct links between them.

Here we can observe that our set of databases is well connected, with the exception of the two smaller databases DGIdb and Harmonizome.

3. Pipeline with Modafinil

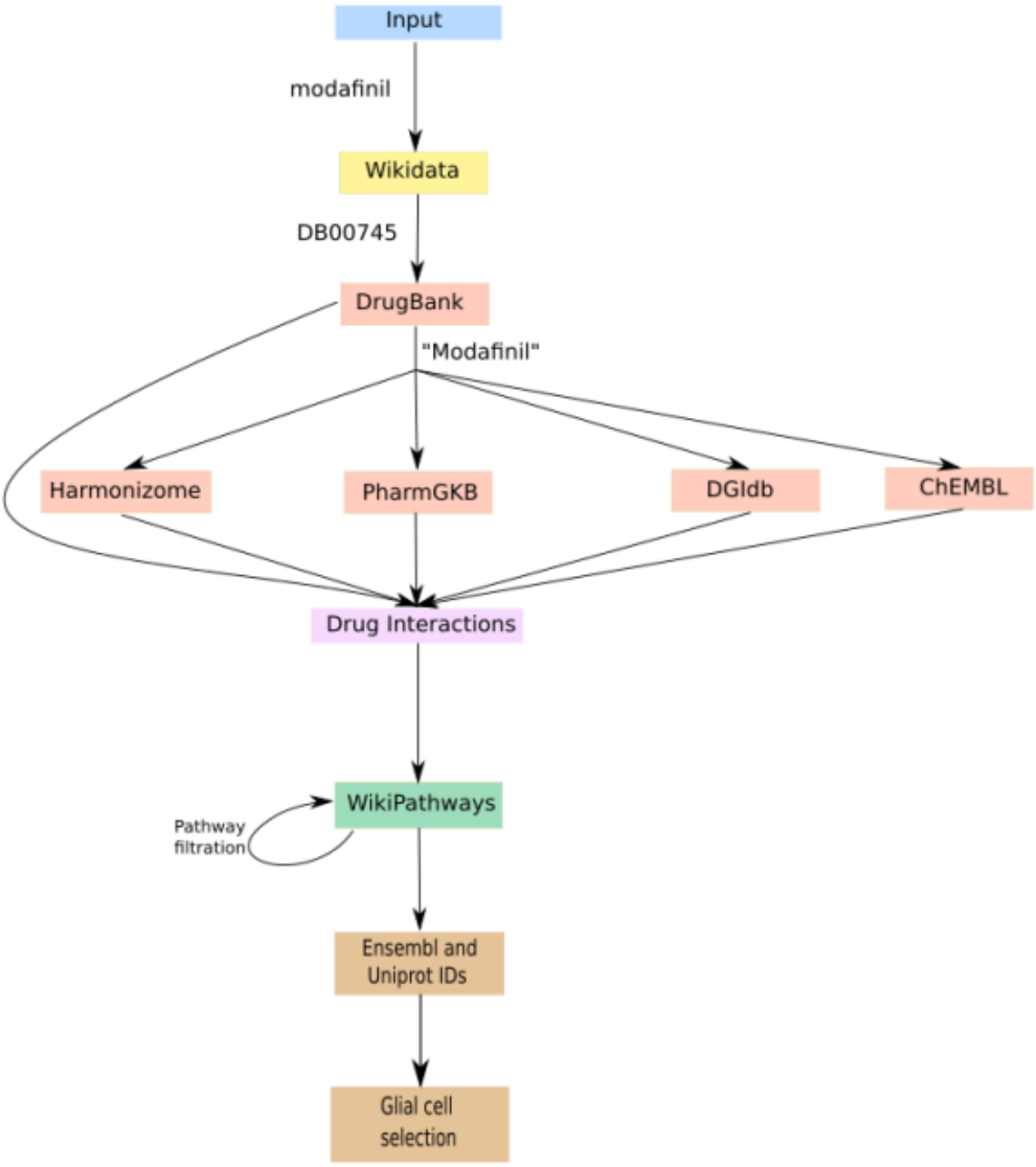

Figure 9. Pipeline's path through selected databases when extracting the potential proteins interactions of Modafinil

The pipeline input is the name of the drug, in this case the Modafinil. With this, a query is made on Wikidata to extract the DrugBank ID. This query is necessary as the DrugBank database is the only one of those selected from which data cannot be extracted merely with the drug name as the files are named with the drug's DrugBank ID.

Queries are then sent to the drugs databases, the Drugs-Target interactions and the sources documenting those are extracted. Three kind of interactions are obtained: proteins, genes and genes variants, this last category is not used for the rest of the pipeline as there is no gene variants database in our list.

WikiPathway is then used to extract all the pathways which contain the genes and proteins interacting with the drug. The content of those pathways is then stored.

However, all the results returned by WikiPathways are not relevant since we are looking for glial results. In an effort to make the results more relevant and to reduce their number, some selections will be made:

- The pathways that do not take place in the brain are removed. A threshold is set where a given percentage of the pathway's proteins have to be present in the brain. This percentage is not 100% because the databases are not exhaustive and the ontology for some of the proteins might not be complete leading to some pathways being relevant but removed because of a few elements.

- The pathways are given a score. This score is computed on the size of the pathway and the amount of recorded interactions occurring in this pathway.

- The proteins are given a score. This score is computed on the amount of times the protein appears in the results and on the score given to the pathway from where it's coming from.

At the time of writing, the computation is still in progress and will not be displayed in the results.

4. Output

Out of all the drugs databases, a total of 30 interactions have been extracted. They are constituted of 16 genes, 11 proteins and 3 variants. As it has been said before, there is no variant database in our list at the moment so those interactions are not used.

After searching in the pathways containing the interactions, we extracted a total of 9358 human proteins. A filtration is made with Uniprot to select only the proteins localized in the

brain. This filtration is not made on the glia yet because Uniprot does not have that level of precision, the purpose is to reduce the amount of results and gin time for the next step. This selection is not done with one of the databases from our list but with a list of glial genes coming from an external source. After this selection, we are left with 1016 results. In these results, we do find the connexins 30 and 43 that were found to interact with Modafinil. [17]

IV) Discussion

1. Databases

There is plenty of information available on the many online databases publicly available. However, this information is dispersed. The recorded drug interactions for example, differ greatly from one database to the other even if some elements are constant.

Few of those many databases were directly available through RDF, whether it is because they are not stored this way or because they chose not to give full access to their data. Some of the most important databases (Uniprot and Ensembl for example) provide a freely accessible RDF endpoint via Virtuoso which make them very easy to access and eventually connect to other databases. It should be noted that those two endpoints are already linked, with the documentation in both of those indicating how to do so. Unfortunately, there are no major drugs databases offering that same service at the exception of ChEMBL which is included in Ensembl. In spite of this, we have established that it is possible to integrate them in a RDF network of databases but it is a precarious answer. Since the data is downloaded, it is not kept to date with the live database and gets no update, potentially meaning that an error in the data could stay until the next download. It also means it is necessary to re-download each time a database is updated, which is a doable but inconvenient process.

Furthermore, the files available for download sometimes present errors or discrepancies such as names miswritten or written differently (chemical name instead of usual name). In TSV files, some of the information is sometimes placed in the wrong column.

What is surprising about this state of things is that plenty of databases either used to be available through a public endpoint created by the database administrators or were part of a

project to gather them inside a global endpoint. Those endpoints are not activated anymore but can still be found on Bio2RDF, a biological database designed to gather RDF databases.

The availability of the database is not the only blocking step to establish a network of connections. One important issue is the structure of the information. In our case, there is no problem with DGIdb, DrugBank or PharmGKB because there is a direct access to the files and we can control how to set up the data inside the local AskOmics endpoint or in the Turtle files. Because of this, we have full knowledge on the way the data is contained and can easily create SPARQL queries. However, this is not the case for the public endpoints of Uniprot and Ensembl. Those are both huge databases containing many different types of information which need to be stored and linked together. This results in a complex data network which design' is either specific to the database or following a set of rules. In our case, Uniprot has it's own specific structure while Ensembl is constructed under the biopax3 model. In both cases, the understanding of the structure can be tedious especially when the documentation provided is incomplete, erroneous or imprecise which can be the case for both those databases. For example, some of the prefixes and relationships between entities indicated in the Uniprot documentation either do not exist or are wrongly flagged. As for Ensembl, the documentation for the biopax3 model is very basic, despite it being deep and complex, which means it is necessary to search for a more thorough documentation elsewhere. Fortunately, example queries are provided for the two endpoints which helps as a good introduction for both simple and more complex queries but there is few of them and they do not reflect the vast amount of information available.

2. Results relevance

There are several factors that need to be taken in account in order to consider our results. Firstly, every result obtained is entirely dependent of the databases' validity. Because we are gathering data from multiple sources to gather as much information as possible instead of cross-examining it, any erroneous result from the database could not be filtered. If an error in a drug data happened to slip by in a drug database, there would be very little way for us to detect it. The solution here lies solely on the choice of the databases and the careful verification of sources from the results. This verification, however, would only remove false

positive results but not false negative ones since they would be automatically removed from our results before we could manually detect the error.

Secondly, the design of the pipeline is debatable as well, exploring the pathways is but one of the many possibilities for exploring the available databases. Multiple options were considered and this one was selected because it was considered pertinent and practical. An interaction between a drug and a gene or a protein could logically impact the levels of a molecule from that pathway, but not necessarily. The results obtained are solely predictions and not confirmed interactions, as those could only be obtained via experiments.

3. Results

We have a total of 1016 proteins as results. This is a number too big to draw conclusions on the relevance of this method, however the two targets that were supposed to be found (connexins 30 and 43) do come up in our list which is an encouraging sign. The pathway filtration has yet to be implemented and may be the decisive step to determine if those results came up by chance or not. If the proteins are ranked by score, the important amount of proteins is not going to be a problem since we will be able to know which ones are truly relevant and which are not.

4. Generalization

Aside from the functions currently being implemented (pathway and proteins statistical selection as well as glial filtration), there is a lot of room for improvement of the pipeline. For starters, more databases could be used to potentially gather interactions that were not recorded within our own set of databases. A variant database would also be necessary to utilize the variants interactions.

But overall, there is an almost endless amount of possibilities when accessing such large databases. The list of databases chosen, now that they're gathered and linked together, can be used as basis for an entirely different focus just as it could be used to complement the current work. For example, the PubChem database, which wasn't included because of a lack

of time, possess Bioassays which could be used to compute similarities between multiple drugs. This way, we could potentially predict common genes and proteins interactions between drugs. Overall, each database added to the list from this point could bring a new potential type of analysis.

IV) Conclusion

There is a huge amount of publicly available information online, whether it is under a RDF format or not. With the exception of a few databases making the deliberate choice of not providing their data, this information is easily accessible, mostly through downloadable files and some databases even provide a free direct programmatic access.

As for the objective of this internship, we obtained a total of 1016 proteins which is too much to consider it achieved at this point. However, the proteins used as validation are present in our list and a ranking score should be implemented in the coming weeks.

Bibliography

[1] Morgan, Steve, Paul Grootendorst, Joel Lexchin, Colleen Cunningham, et Devon Greyson. « The Cost of Drug Development: A Systematic Review ». Health Policy 100, no 1 (avril 2011): 4-17. https://doi.org/10.1016/j.healthpol.2010.12.002.

[2] Weng, Liming, Li Zhang, Yan Peng, et R Stephanie Huang. « Pharmacogenetics and pharmacogenomics: a bridge to individualized cancer therapy ». Pharmacogenomics 14, no 3 (février 2013): 315-24. https://doi.org/10.2217/pgs.12.213.

[3] Cha, Y., T. Erez, I. J. Reynolds, D. Kumar, J. Ross, G. Koytiger, R. Kusko, et al. « Drug Repurposing from the Perspective of Pharmaceutical Companies ». British Journal of Pharmacology 175, no 2 (2018): 168-80. https://doi.org/10.1111/bph.13798.

[4] Prasad, Vinay, et Sham Mailankody. « Research and Development Spending to Bring a Single Cancer Drug to Market and Revenues After Approval ». JAMA Internal Medicine 177, no 11 (novembre 2017): 1569-75. https://doi.org/10.1001/jamainternmed.2017.3601.

[5]Pushpakom, Sudeep, Francesco Iorio, Patrick A. Eyers, K. Jane Escott, Shirley Hopper, Andrew Wells, Andrew Doig, et al. « Drug Repurposing: Progress, Challenges and Recommendations ». Nature Reviews. Drug Discovery, 12 octobre 2018. https://doi.org/10.1038/nrd.2018.168.

[6] Talevi, Alan. « Drug repositioning: current approaches and their implications in the precision medicine era ». Expert Review of Precision Medicine and Drug Development 3, no 1 (2 janvier 2018): 49-61. https://doi.org/10.1080/23808993.2018.1424535.

[7] Ozdemir, E. Sila, Farideh Halakou, Ruth Nussinov, Attila Gursoy, et Ozlem Keskin. « Methods for Discovering and Targeting Druggable Protein-Protein Interfaces and Their Application to Repurposing ». In Computational Methods for Drug Repurposing, édité par Quentin Vanhaelen, 1903:1-21. New York, NY: Springer New York, 2019. https://doi.org/10.1007/978-1-4939-8955-3_1.

[8]Chou, Ting-Chao. « Theoretical Basis, Experimental Design, and Computerized Simulation of Synergism and Antagonism in Drug Combination Studies ». Pharmacological Reviews 58, no 3 (septembre 2006): 621-81. https://doi.org/10.1124/pr.58.3.10.

[9] Li, Shuang, Yuze Cao, Lei Li, Huixue Zhang, Xiaoyu Lu, Chunrui Bo, Xiaotong Kong, et al. « Building the Drug-GO Function Network to Screen Significant Candidate Drugs for Myasthenia Gravis ». PloS One 14, no 4 (2019): e0214857. https://doi.org/10.1371/journal.pone.0214857.

[10] Foucquier, Julie, et Mickael Guedj. « Analysis of drug combinations: current methodological landscape ». Pharmacology Research & Perspectives 3, no 3 (juin 2015). https://doi.org/10.1002/prp2.149.

[11] Huang, Hui, Ping Zhang, Xiaoyan A. Qu, Philippe Sanseau, et Lun Yang. « Systematic prediction of drug combinations based on clinical side-effects ». Scientific Reports 4 (24 novembre 2014). https://doi.org/10.1038/srep07160.

[12] Health Quality Ontario. « Pharmacogenomic Testing for Psychotropic Medication Selection: A Systematic Review of the Assurex GeneSight Psychotropic Test ». *Ontario Health Technology Assessment Series* 17, nº 4 (2017): 1-39.

[13] Liu, Xinhe, Jean-Marie Petit, Pascal Ezan, Joël Gyger, Pierre Magistretti, et Christian Giaume. « The Psychostimulant Modafinil Enhances Gap Junctional Communication in Cortical Astrocytes ». *Neuropharmacology* 75 (décembre 2013): 533-38. https://doi.org/10.1016/j.neuropharm.2013.04.019.

[14] Brookshire, Bethany. "Scientists Say: Glia." Science News for Students, www.sciencenewsforstudents.org/blog/scientists-say/scientists-say-glia.

[15] Jäkel, Sarah, et Leda Dimou. « Glial Cells and Their Function in the Adult Brain: A Journey through the History of Their Ablation ». Frontiers in Cellular Neuroscience 11 (13 février 2017). https://doi.org/10.3389/fncel.2017.00024.

[16] Zuchero, J. Bradley, et Ben A. Barres. « Glia in Mammalian Development and Disease ». Development (Cambridge, England) 142, no 22 (15 novembre 2015): 3805-9. https://doi.org/10.1242/dev.129304.

[17] Duchêne, Adeline, Magali Perier, Yan Zhao, Xinhe Liu, Julien Thomasson, Frédéric Chauveau, Christophe Piérard, et al. « Impact of Astroglial Connexins on Modafinil Pharmacological Properties ». *Sleep* 39, nº 6 (01 2016): 1283-92. https://doi.org/10.5665/sleep.5854.

[18] David C. Faye, Olivier Curé, Guillaume Blin. A survey of RDF storage approaches. Revue Africainede la Recherche en Informatique et Mathématiques Appliquées, INRIA, 2012, 15, pp.11-35. hal-01299496

[19] Carl Boettiger. (2018). rdflib: A high level wrapper around the redland package for common rdf applications (Version 0.1.0). Zenodo. <https://doi.org/10.5281/zenodo.1098478>

[20] Charles Bettembourg, Olivier Dameron, Anthony Bretaudeau, Fabrice Legeai. AskOmics : Intégration et interrogation de réseaux de régulation génomique et post-génomique. *IN OVIVE (INtégration de sources/masses de données hétérogènes et Ontologies, dans le domaine des sciences du VIVant et de l'Environnement)*, Jun 2015, Rennes, France. pp.7. ⟨hal-01184903⟩

[21] The UniProt Consortium, UniProt: a worldwide hub of protein knowledge, *Nucleic Acids Research*, Volume 47, Issue D1, 08 January 2019, Pages D506–D515, https://doi.org/10.1093/nar/gky1049

[22] Demir, Emek, Michael P Cary, Suzanne Paley, Ken Fukuda, Christian Lemer, Imre Vastrik, Guanming Wu, et al. « The BioPAX community standard for pathway data sharing ». *Nature Biotechnology* 28 (9 septembre 2010): 935.

[23] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2017 Nov 8. doi: 10.1093/nar/gkx1037.

[24] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. 2019 Jan 8; 47(D1):D1102-1109. doi:10.1093/nar/gky1033. [PubMed PMID: 30371825]

[25] Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, Ma'ayan A. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database (Oxford). 2016 Jul 3;2016. pii: baw100.

[26] M. Whirl-Carrillo, E.M. McDonagh, J. M. Hebert, L. Gong, K. Sangkuhl, C.F. Thorn, R.B. Altman and T.E. Klein. "Pharmacogenomics Knowledge for Personalized Medicine" *Clinical Pharmacology & Therapeutics* (2012) 92(4): 414-417.

[27] Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., and Tanabe, M.; New approach for understanding genome variations in KEGG. Nucleic Acids Res. 47, D590-D595 (2019). [pubmed] [doi]

[28] Gaulton A, Bellis LJ, Bento AP, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res*. 2012;40(Database issue):D1100–D1107. doi:10.1093/nar/gkr777

[29] Kelsy C Cotto, Alex H Wagner, Yang-Yang Feng, Susanna Kiwala, Adam C Coffman, Gregory Spies, Alex Wollam, Nicholas C Spies, Obi L Griffith, Malachi Griffith, DGIdb 3.0: a redesign and expansion of the drug–gene interaction database, *Nucleic Acids Research*, Volume 46, Issue D1, 4 January 2018, Pages D1068–D1073, https://doi.org/10.1093/nar/gkx1143

[30] P.J. Kersey, J.E. Allen, A. Allot, M. Barba, S. Boddu, B.J. Bolt, D. Carvalho-Silva, M. Christensen, P. Davis, C. Grabmueller, N. Kumar, Z. Liu, T. Maurel, B. Moore, M. D. McDowall, U. Maheswari, G. Naamati, V. Newman, C.K. Ong, D.M. Bolser., N. De Silva, K.L. Howe, N. Langridge, G. Maslen, D.M. Staines,A.Yates**.** Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Research* 2018 46(D1)D802–D808 https://doi.org/10.1093/nar/gkx1011

[31] Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, Mélius J, Cirillo E, Coort SL, Digles D, Ehrhart F, Giesbertz P, Kalafati M, Martens M, Miller R, Nishida K, Rieswijk L, Waagmeester A, Eijssen LMT, Evelo CT, Pico AR, Willighagen EL. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research *Nucleic Acids Research*, (2017) doi.org/10.1093/nar/gkx1064 PMC5753270