Analyzing Drug-Gene Interactions Using Prolog

Lama Khalil *khali121@uwindsor.ca*
Ahmed Shafeek *ashafeek@uwindsor.ca*
Carson Dickie *dickiec@uwindsor.ca*

December 7, 2020

Github Repository

Contents

1 Abstract

Advances in DNA sequencing have lead to a greater understanding of the interplay between genetics and the effectiveness of drug therapies. It is now clear that a person's unique set of alleles plays a key role in how they will respond to a given therapy. The rapidly increasing pool of research on gene-drug relationships has created a need for new tools capable of aggregating and analyzing the data produced by these studies. This project proposes a Prolog database as a means to accomplish these tasks. We created a set of Prolog facts that capture the gene-drug interactions for a small subset of drugs related primarily to the treatment of mental health conditions. We then developed several rules (algorithms) capable of analyzing and making recommendations based on the data. The results suggest that with more time and effort, a Prolog database could become a powerful and useful pharmacogenetic tool.

2 Keywords

- Personalized medicine
- pharmacogenomics
- Prolog
- Pharmacokinetics
- Mental Health

3 Introduction

Drug therapies are used to treat many diseases and disorders. With each treatment however, there are risks of side effects and other unintended consequences. Advances in our understanding of the human genome have made it clear that genetics play a significant role in the way in which we process and react to drugs[1]. The emergence of studies describing the interactions between genetics and pharmacology has led to a need for tools that can aggregate, analyze and draw conclusions from this new data. Prolog provides a strong framework for analyzing existing relationships and identifying new ones. This paper aims to demonstrate how Prolog can be used to draw conclusions and make recommendations based on a set of gene-drug associations.

4 General Concepts

4.1 Genes

All living organisms contain genetic material in the form of deoxyribonucleic acid (DNA). DNA is a double stranded helix of organic molecules known as nucleotides bases. The four bases are Adenosine (A), Thymine (T), Guanine (G), and Cytosine(C). Under normal circumstances the two strands that make up the DNA are linked through the binding of Adenosines to Thymines and Guanines to Cytosines[2].

In humans, DNA is bundled into 23 pairs of chromosomes within the nucleus of each cell. Each chromosome stores the genetic information for many genes. Genes, in turn, determine an individual's physical characteristics[3]. The DNA that comprises a gene is first transcribed into ribonucleic acid (RNA) and then that RNA is translated into a series of Amino Acids which join to form proteins. Proteins, in turn, determine everything from the colour of a person's hair to the way in which they will respond to a medication. An individual's set of genes is known as their genotype, while the physical expression of these traits is known as their phenotype[4].

An individual inherits 23 chromosomes from each of their parents through a process known as meiosis during sexual reproduction. Thus, an offspring's genotype will differ from both of their parent's. Additionally, each time a new cell is created in the body the genetic material is copied through a process known as mitosis. These processes are imperfect and flaws in replication and other natural processes can lead to mutations – alterations in the nucleotide sequence of the genome. These processes, combined, have led to a wealth of genetic diversity within the human population[5].

One of the most common mutations is the single-nucleotide polymorphism (SNP). SNPs result from the substitution of a single nucleotide at a specific position in the genome. Over time, the inheritance of SNPs and other mutations will lead a gene to exist in several different forms across a population[6]. These variants are known as alleles. Humans carry two copies of most genes (one from each parent). Consequently, a person might carry two identical alleles (meaning they are homozygous for that trait) or two different alleles (meaning they are heterozygous for that trait) for each gene.

When single nucleotide polymorphisms are identified they are given unique identifiers known as RS Numbers or RSids, which stands for Reference SNP cluster ID. For example, rs2235067T is representative of an SNP in the MDR1 gene, where Cytosine (which occurs in 87% of the population) is replaced with Thymine. The Cytosine variant of the gene is known as the wild type – the type of allele occurring most frequently in the population. It should be noted that the family of genes responsible for encoding Cytochrome P450 enzymes (proteins) use a unique but widely accepted "star" nomenclature for identifying alleles. Under this nomenclature, each allele is given a "star number" such as *1/*3. The two numbers represent the two copies of the allele carried by an individual, and *1/*1 is often (but not always) considered the wild type[7]. P450 enzymes are of interest because of their involvement in the clearance of drugs and other compounds.

4.2 Genetic Sequencing

Genetic sequencing is the process of determining the nucleic acid sequence in DNA. The human genome was sequenced for the first time in 2003. The process was originally very slow, but the technology has since evolved, and a person's genetic code can now be analyzed rapidly[8]. These breakthroughs have allowed scientists to study the exact sequence of nucleotide bases in a strand of DNA. Thus, nucleotide sequences can now be linked to genes and genes to proteins. Scientists can now take a sample of DNA from an individual and identify which alleles they carry (either by matching them to existing alleles or documenting new ones).

4.3 Pharmacogenomics

Drugs – substances that cause a change in an organism's physiology or psychology – are used to treat, cure, or prevent diseases and other negative conditions. Drugs work by interacting with receptors and enzymes (both of which are proteins) within the body. For a drug to take effect it must be present in the bloodstream in sufficient concentration for a sufficient time. Thus, for a drug to be effective it must be administered in the proper dose and at the right interval. The dose rate describes the ratio of dose rate (usually in milligrams) to dose interval (usually in hours). Pharmacokinetics is the study of how the body metabolizes, distributes and absorbs drugs[9].

Two people will often respond very differently to the same drug. There is now a strong body of evidence suggesting that genetic variation plays a key role in determining how a given individual will process and react to a drug. Our ability to rapidly sequence DNA has created new opportunities for understanding these gene-drug interactions. The study of these interactions is known as pharmacogenomics (or pharmacogenetics)[10] and the application of pharmacogenomics towards developing prescription protocols tailored to our unique genetics is known as personalized medicine[11].

4.4 Prolog

Prolog is a logic programming language with an important role in artificial intelligence. In Prolog, logic is expressed as relations - a combination of facts and rules. Facts are known pieces of information, while rules represent if-then clauses: if the facts on the right hand side of the rule are found to be true, then the rule is true. Prolog is especially well suited for problems that involve objects - in particular, structured objects - and relations between them[12].

5 Objectives

Phamacogenetics is still in its infancy, and while are now many studies documenting gene-drug interactions, their results exist as disjoints sets. Organizations, such as PharmGKB, are working to catalogue and unify the wealth of new information into relational databases. Centralized storage, however, is only the first of many steps towards a fully understanding of gene-drug interactions. The next step will be to develop algorithms to aggregate, analyze and draw conclusions from the data.

The purpose of this research project is to design and implement a logical database by which we will record, structure and eventually quantify gene-drug associations and how variations in each gene affect these associations. We hope that by structuring and synthesizing the data we will be able to identify useful relationships and interactions. Ideally, we will develop one or more algorithms to identify the right drug and/or dosing regimen for a given condition and set of genetic variants.

Considering the limited time and manpower available to us, the scope of this project will be limited to proof of concept and not a complete work. With more time and effort, however, this information should benefit both researchers and physicians..

6 Results

We catalogued roughly 200 gene-drug associations within a Prolog-based logical database. These associations span 12 drugs and roughly 15 genes, each with many variants. Associations are recorded as facts in the following format:

```
association('GENE', 'ALLELE', 'DRUG', 'PMID', 'MAG/DIR', 'EFFECT', 'GROUP').
```

Where the parameters are as follows (for more information on each parameter see the appendices):

- GENE – the name of the gene (example: MDR1).
- ALLELE – the genetic variant, given as either an RsID or in “star” nomenclature.
- DRUG – The chemical compound identified in the association (example: Amitriptyline).
- PUBMED ID – A unique number identifying the paper in which the association was published.
- MAGNITUDE/DIRECTION – The magnitude and direction of the association. If the magnitude was not given, then it is recorded as 1 (representing an increase) or -1 (representing a decrease).
- EFFECT – The effect of the association (example: clearance).
- POPULATION UNDER STUDY – The group of people in which the association was identified (example: depression).

We also wrote rules defining each gene, drug, effect and population group as follows:

- gene(’GENE’).
- drug(’DRUG’).
- condition(’CONDITION’).

- effect('EFFECT').
- positive_effect('EFFECT').
- negative_effect('EFFECT').

Where positive effects are effects that are clearly positive (such as chance of remission) and negative effects are effects that are clearly negative (such as increased risk of suicidal ideation).

Our initial dataset did not include any weighted associations (meaning each association was described as simply an increase or a decrease). This is because many studies do not quantify the associations they identify, and those that do often quantify them using a variety of different metrics that are hard to unify. With the time and manpower available to us, we simply did not have time to determine the magnitude of these relationships. We did, however, include a small subset of weighted associations recording both direction and magnitude. These were used as a proof of concept for additional algorithms.

We designed 2 functions (or rules) to aggregate and quantify the associations. They are each overloaded to accept either a drug and a genetic variant in the form of a RsID or a drug and a gene with its corresponding variant in "star" notation as parameters.

Listing 1: List Effects

```
list_effects(Drug, RsID, PM, NM):-
    findall(
            (ID,P),
            (positive_effect(P),
             association(_,RsID,Drug,ID,1,P,_)),
           PM
    ),
    findall(
            (ID,N),
            (negative_effect(N),
             association(_,RsID,Drug,ID,1,N,_)),
           NM
    ).
```

list_effects() takes a drug and an allele as parameters. It returns lists of both the positive effects and the negative effects documented for the given combination of drug and allele, along with PUBMED IDs of the papers where these effects were reported.

Listing 2: Effect Score

```
effect_score(Drug, RsN, PC, NC):-
    list_effects(Drug, RsN, PM, NM),
    length(PM, PC),
    length(NM, NC).
```

effect_score () takes a drug and an allele as parameters. It returns two scores – a positive effect score and a negative effect score. The scores are obtained by adding up the number of positive effects and negative effects, respectively.

Our third algorithm recommends whether to prescribe a medication to a patient with a given allele:

Listing 3: Recommendation

```
recommendation(Drug, RsN, X):-
      effect_score(Drug, RsN, PC, NC),
      (NC > 0, X = 'no';
       PC > 0, X = 'yes';
       NC == 0, PC == 0, X = 'neutral').
```

The user provides a drug and an allele, and the algorithm returns a recommendation. The recommendation is negative if any negative effects are associated with the given gene-alle combination. The recommendation is positive if there are no negative effects and one or more positive effects are associated with the given gene-allele combination. The recommendation is neutral otherwise.

The final algorithm focuses on associations that affect drug clearance rate. Drug clearance, the rate at which a drug is cleared from the body, is directly proportional to the dose rate where the dose rate is a ratio of the dose size (in miligrams) to the dose interval (in hours):

$$DoseRate \propto ClearanceRate \tag{1}$$

$$DoseRate = \frac{DoseSize}{DoseInterval} \tag{2}$$

Listing 4: Dose Rate Multiplier

```
dose_rate_multiplier(Drug, Gene, Allele, Rate):-
      association(Gene, Allele, Drug, _, X, 'clearance', _),
      Rate is 1 * (1 + X).
```

The algorithm asks the user to provide an allele and a drug and outputs the corresponding increase or decrease to the dose rate (expressed as a multiplier). The multiplier can then be applied to the maintenance dose (the fraction of the drug dose that was eliminated from the body during the previous dosing interval)[9].

7 Conclusion

pharmacogenomics and Pharmacokinetics are closely linked. As the pool of research studying the relationship between genes and drug processing grows, so too will the need for tools that can aggregate, analyze and draw conclusions from this data. The Prolog database designed for this project is one such tool. The database, while limited in its capabilities, provides a host of useful features for both researchers and medical practitioners. Researchers can quickly obtain all pertinent associations for a given drug-genetic variant combination and medical practitioners can easily determine what negative and positive effects might be associated with prescribing a medication to a patient with a given set of alleles. The work serves as a proof of concepts and provides a strong case for continuing to build on the work done thus far. The next section outlines the steps that can be taken to further this work and enhance the power and value of this tool.

8 Future Work

This work, while valuable as a proof of concept, is far from complete. The first step in advancing the value and power of the Prolog database would be to populate it with more data. Currently, the database contains only a handful of drugs and genes. A wider range of drugs would broaden its use and increase its predictive power. Unfortunately, populating the database is currently very time consuming. We initially ruled out the possibility of using a parsing algorithm to populate the data due to the qualitative nature of many recorded associations. A well-designed algorithm could, however, at least reduce the amount of manual input required.

The next step would be to quantify all the associations within the database. Our initial plan was to quantify all relationships, but this proved surprisingly difficult due to the wide range of measuring, naming and documenting techniques used in pharmacogenetic studies. One study might report their findings in terms of an effect on dose, while another might report findings in terms of clearance rate and yet another might refer to rate of metabolism. Further investigation might reveal that the first study observed that an increase in dose was required because the drug in question was being cleared more rapidly from the patient's systems. Thus, all these studies are describing the same type of association, but each has given it a different name, unit of measurement, etc. This means that quantifying associations will often be a difficult task and may not even be possible without the input from an expert in the field. We recommend that scientists involved in pharmacogenetic research establish an agreed upon method of naming, measuring and categorizing these relationships.

As the quantity and quality of the associations contained within the database grows, more complex and powerful algorithms can be developed. The current algorithms, for example, only draw conclusions for one drug-allele association at a time. Future iterations might include a map-reduce like feature that would be applied to a drug and a set of alleles. The recommendation algorithm would first be mapped to all drug-allele pairs. These results could then be reduced into one final analysis or recommendation based on the sum of the parts.

Finally, an intuitive and user-friendly interface would broaden the appeal of the Prolog database. Running commands in Prolog is an arduous task, where each parameter must be specified regardless of whether or not some default values will be used. Researchers would benefit from an interface that provided quick access to features and simplified the input requirements. Medical practitioners hoping to use such a tool in their practice would require a graphical user interface.

9 Acknowledgement

The concept for this project was originally proposed by Dr. Abedalrhman Alkhateeb. We found the scientific article Drug-Gene Interactions: Inherent Variability in Drug Maintenance

Dose Requirements, published in the Pharmacy and Therapeutics Journal, invaluable in our initial understanding of gene-drug interactions.

9.1 Special Thanks

We'd like to thank Dr. Abedalrhman Alkhateeb for suggesting introducing us to pharmacogenomics and for supporting us through this endeavour.

9.2 Contributions

Each group member contributed at or above the expectations laid out in our initial agreement to work together. Expectations were reviewed in regular team meetings, and there were never any concerns with respect to a team member performing below expectations.

10 Appendices

10.1 Appendix A - Software Details

This Prolog application was designed using GNU prolog (version gprolog-1.4.5).

10.2 Appendix B - Listing and Description of Prolog facts

10.2.1 Genes

A list of the genes currently represented within the Prolog database. Note that genes begining with 'CYP' are part of the Cytochrome P450 family. These genes use 'star nomenclature' to identify variants (alleles). Other genes use the RsID.

- gene('CYP2D6').
- gene('CYP2A6').
- gene('CYP2B6').
- gene('CYP3A4').
- gene('CYP1A2').
- gene('CYP2C9').
- gene('CYP2C19').
- gene('CYP1B1').
- gene('CYP3A5').
- gene('MDR1').
- gene('CYP2D9').

- gene('HTR2A').
- gene('FKBP5').
- gene('HTR1B').
- gene('COMT').
- gene('HTR7').
- gene('GABRQ').
- gene('SLC6A4').
- gene('TPH2').
- gene('GRIA1/3').
- gene('GRIA3').
- gene('SLC6A2').

10.2.2 Drugs

A list of drugs intended for the database. Note that some of these drugs are not yet represented, but it is our intention to include them in time.

- drug(amitriptyline).
- drug(amphetamine).
- drug(aripiprazole).
- drug(atomoxetine).
- drug(brexpiprazole).
- drug(bupropion).
- drug(chlordiazepoxide).
- drug(citalopram).
- drug(clomipramine).
- drug(clopidogrel).
- drug(clozapine).
- drug(desipramine).
- drug(diazepam).

- drug(doxepin).
- drug(escitalopram).
- drug(fluoxetine).
- drug(fluvoxamine).
- drug(haloperidol).
- drug(iloperidone).
- drug(imipramine).
- drug(irbesartan).
- drug(metoprolol).
- drug(mirtazapine).
- drug(modafinil).
- drug(nefazodone).
- drug(nortriptyline).
- drug(olanzapine).
- drug(olanzapine).
- drug(omeprazole).
- drug(paroxetine).
- drug(perphenazine).
- drug(phenytoin).
- drug(pimozide).
- drug(protriptyline).
- drug(risperidone).
- drug(sertraline).
- drug(thioridazine).
- drug(tramadol).
- drug(trimipramine).
- drug(venlafaxine).

- drug(vortioxetine).
- drug(warfarin).
- drug(zuclopenthixol).

10.2.3 Conditions

A list of treatment groups to which the gene-drug associations apply.

- condition('NA'). - treatment group not identified within the study.
- condition('depression').
- condition('mood disorder').
- condition('anxiety').
- condition('OCD'). - obsessive compulsive disorder.
- condition('schizophrenia').
- condition('healthy'). - healthy individuals.
- condition('elderly'). - individuals over the age of 65.

10.2.4 Effects

A list of possible, measured effects arising from gene-drug associations. Where possible, effects are divided into positive and negative subcategories. Note that a drug-gene association can have either increase (positive) or decrease (negative) the likelihood/ occurrence of a given effect.

- effect('adverse effects'). - the likelihood of adverse effects occurring.
- effect('blood concentration'). - the concentration of a drug in the blood.
- effect('clearance'). - the rate of clearance of a drug from the body.
- effect('dose'). - the dose required for a drug.
- effect('HAM-A reduction'). - a measured decrease in anxiety levels.
- effect('improvement'). - the likelihood of improvement after a set period of time.
- effect('levels'). - the levels of a drug in the blood.
- effect('metabolism'). - the metabolism of a drug.
- effect('side effects'). - the likelihood of side effects.
- effect('response'). - a measure of the effectiveness of a drug.

- effect('suicide'). - the likelihood of suicidal thoughts.
- effect('toxicity'). - the toxicity of a drug.
- effect('discontinuation'). - the likelihood of discontinuing a treatment.
- effect('plasma nortriptyline'). - the concentration of nortriptyline in the plasma.
- effect('remission'). - the likelihood of remission.
- effect('amitriptyline-nortriptyline ratio'). - the ratio of amitriptyline to nortriptyline
- effect('plasma levels'). - the concentration of a drug in the plasma.
- effect('Cmax and AUC'). - the peak serum concentration achieved by a drug.

The effects condsidered positive:

- positive_effect('improvement').
- positive_effect('remission').
- positive_effect('response').
- positive_effect('HAM-A reduction').

The effects considered negative:

- negative_effect('adverse effects').
- negative_effect('side effects').
- negative_effect('suicide').
- negative_effect('toxicity').

References

[1] Sayer I Portelli M. Genetic basis for personalized medicine in ashtma. *Expert Review of Respiratory Medicine*, 6(2):223–236, 2012.

[2] Spencer M. The stereochemistry of deoxyribonucleic acid. ii. hydrogen-bonded pairs of bases. *Acta Crystallographica*, 12(1):66–71, 1959.

[3] Pearson H. Genetics: what is a gene? *Nature*, 441(7092):398–401, 2006.

[4] Richard Dawkins. Replicator selection and the extended phenotype. *Ethiology*, 47(1):61–76, 1978.

[5] Alan F Wright. Genetic variation: Polymorphisms and mutations. *Encyclopedia of Life Sciences*, 2005.

[6] Liu JS Chen T Waterman MS Sun F Zhang K, Qin ZS. Haplotype block partitioning and tag snp selection using genotype data and their applications to association studies. *Genome Research*, 14(5):908–916, 2004.

[7] Talbot JN Sprague JE Kisor DF, Kane MD. *Pharmacogenetics, Kinetics, and Dynamics for Personalized Medicine.* Jones Bartlett Learning, Burlington, Massachusetts, 2014.

[8] M. Ronaghi; S. Karamohamed; B. Pettersson; M. Uhlen P. Nyren. Real-time dna sequencing using detection of pyrophosphate release. *Analytical Biochemistry*, 242(1):84–89, 1996.

[9] Tozer TN Rowland M. *Clinical Pharmacokinetics and Pharmacodynamics.* Lippincott Williams Wilkins, Baltimore, Maryland, 2011.

[10] Technical Report. Pharmacogenetics: potential for individualized drug therapy through genetics. *Trends in Genetics*, 19(11):660–666, 2003.

[11] Johnson JA. Stratified, personalised or p4 medicine: a new direction for placing the patient at the centre of healthcare and health education. *Academy of Medical Sciences*, 2015.

[12] Ivan Bratko. *Prolog Programming for Artificial Intelligence.* Person Education Canada, 2011.