Final Report

By Tasnia Tahsin

Mentor: Graciela Gonzalez

Abstract: As the severity and frequency of Adverse Drug Reactions (ADRs) continues to increase, developing effective and efficient algorithms for early detection of ADRs has become a critical issue. Extensive research is being conducted in this relatively new area of bioinformatics and a variety of methodologies have been proposed and implemented for this purpose. This paper suggests a novel approach for studying ADRs using an online system called GeneRanker.

1. Introduction

Drugs are designed to cure diseases but unfortunately they also tend to cause a multitude of undesired effects in the human body called adverse drug reactions (ADRs). Formally defined as a "response to a drug which is noxious and unintended and which occurs at doses normally used in man for prophylaxis, diagnosis, or therapy of disease or for the modification of physiologic function", ADRs have recently given rise to widespread public concern. Every year approximately 2 million patients in the United States are estimated to be affected by a severe ADR resulting in roughly 100,000 fatalities. This makes ADR the fourth leading cause of death in the U.S, right after cancer and heart diseases [6].

The incidence of several serious and unexpected ADRs led to the withdrawal of 19 widely used marketed drugs over the last decade making ADR a major commercial concern for pharmaceutical companies. Moreover it is estimated that $136 billion is spent on treating ADRs in the U.S. every year and other nations worldwide are also facing similar difficulties [6]. Thus, developing effective mechanisms for detecting potential ADRs before marketing drugs is now critically important. The tools for doing this are currently very limited and extensive research needs to be done to make this task more efficient, inexpensive and reliable.

One promising means for predicting the adverse effects of a drug involves the study of its molecular interactions within the human body. However, humans are complex machines and adverse drug reactions are caused by intricate, often unknown biochemical processes. There are principally five different ways in which a drug may produce an adverse reaction - on-target interaction, off-target interaction, hypersensitivity, activation to toxic metabolites and genetic variation. On-target toxicities are toxicities that are produced when a drug modifies its intended therapeutic targets in desired or undesired cells. Off-target toxicities are caused by the interaction of a drug with unintended targets. Hypersensitivity is an immune mediated drug response[3].

Drugs may also produce an adverse effect if they are activated to toxic metabolites during the drug metabolism process. Moreover, variations in genetic predispositions may cause different people to react differently to the same drug. Some may show no adverse reaction at all while some may be severely

affected. For instance, genetic mutation may alter the activity of enzymes involved in drug metabolism, thereby affecting the rate of drug absorption, distribution and excretion. A decrease in the rate of drug excretion may cause the drug to build up in the body and produce toxic effects [3].

Due to the complexity of the processes involved, it is difficult and not always possible to determine the exact cause of an ADR using currently available methods. Also, the precise manner in which a drug works inside the body is yet unknown. We are still not fully aware of all the targets a drug may have, how exactly it may be affecting these targets and what metabolites its activation may result in. Predicting all the adverse reactions of a given drug by looking at the molecular level is therefore not a trivial matter. ADRs caused by genetic variation are probably some of the most difficult to detect and tend to be more severe in nature.

In this paper we propose a novel method for exploring the genetic activity behind ADRs by using an online system called GeneRanker. GeneRanker presents a ranked list of genes potentially related to a given toxicity or a given list of genes by mining disease-gene and gene-gene interactions from biomedical abstracts [1]. By exploiting GeneRanker to obtain this ranked list of genes for drugs and their suspected ADRs, we may enhance and speed up the process of finding gene polymorphisms associated with an ADR. This information in turn will greatly assist research directed towards development of personalized medicine.

The rest of the paper is organized as follows: Section 2 summarizes related work in this field. Section 3 describes the technique we applied. Section 4 presents the results obtained. Finally Section 5 discusses the results and concludes the paper

1. Related Work

The use of computational techniques for analyzing drug activity at the molecular level has recently spurred a lot of interest. Researchers have been trying a wide variety of such methods but owing to the complexity of this task none has yet been successful enough to allow ADR prediction. In [4] Ma'ayan et al studied the bipartite network of drugs and their molecular targets and discovered that this network is scale free. [5] introduced the concept of 'metabolic drug scope'. This represents the set of all compounds and reactions that a drug can potentially influence. Analysis of the metabolic scopes of 276 human approved drugs revealed that drugs may be grouped into distinct categories based on their scopes and certain therapeutic properties of these drugs can also be linked to their scopes.

[2] used an unique approach to discover off-targets of drugs by using side-effect similarity and was fairly successful. 1018 drug-drug connections were found among 746 marketed drugs by using this measure of drug similarity. Testing 20 unexpected drug pairs via in vitro binding assays confirmed 13 of the implied drug-target interactions. 11 had binding affinities strong enough to cause ADR. [6] analyzed correlations between ADRs and metabolic pathways. First a set of drugs that produce identical ADRs were retrieved from the PharmaPendium database. The protein targets for these drugs were then predicted using chemical fingerprints for the compounds. Finally the pathways associated with these targets were

extracted from Metabase (GeneGo) and ranked according to their likelihood of being linked to the ADR. Some of the pathways thus found were intuitive while some not easily predictable.

2. Method

GeneRanker is a freely available online application for mining disease-gene relationships by extracting data from the CBioC database. The CBioC database contains over 1 million protein-protein interactions and over 300,000 gene-disease associations extracted from 1.6 million biomedical abstracts using natural language processing. CBioC also integrates almost 380,000 protein-protein interactions from IntAct, MINT, BIND, and DIP. The NLP engine behind CBioC, IntEx, has been found to be 65% accurate for protein-protein extractions from text and 77% accurate for gene-disease association extractions [7].

Given a disease, GeneRanker first extracts all the genes directly associated with it from the CBioC database. This initial set of genes, called the seed set, is then expanded by retrieving from the database all protein-protein interactions that involve these genes. A network is thus built by drawing edges between genes that are reported to interact and a two level scoring formula based on the network is applied to rank these genes. The first part of the formula, called the seed measure, counts the number of interactions of each gene in the extended set with the genes in the seed set and thus provides a measure of existing evidence. The second part, called the clustering coefficient, calculates the ratio of the actual edges in the neighborhood of a gene to the maximum number of edges that can exist in the neighborhood and thus measures the importance of a gene in keeping its neighbors connected. The harmonic mean of the two scores is then found for each gene and used as the basis for ranking. Mathematically this can be described as follows [1]:

Let G be the gene network built as described before, and let g be a gene in G. Then we define:

- c(g,g') = 1 if g and g' are directly connected, 0 otherwise
- n(g) = {g' in G | c(g,g') = 1}, i.e. the set of neighbors of g
- l(g) = 0 if g is obtained during the seed process, or
  l(g) = 1 + min {l(g') | g' in n(g)} otherwise

In other words l(g) represents the level at which g is added to the network, in which case the seed genes have an assigned level value of 0.

The score associated to a gene is then defined as:

$$s(g) = 2 / (s_s(g)^{-1} + s_c(g)^{-1})$$

where

$$s_s(g) = \sum_{g' \in n(g)} 1 - l(g')$$

and

$$s_c(g) = \frac{2 * \sum_{g_1 g_2 \in n(g)} c(g_1, g_2)}{|n(g)| * (|n(g)| - 1)}$$

In this project we used GeneRanker for mining ADR-gene and drug-gene relationships. The former was simple since GeneRanker is capable of producing the ranked list of genes for most adverse drug reactions given the name of the reaction. For the latter, we first obtained the set of genes known to be directly associated with the given drug from PharmGKB, a publicly available online tool for aiding research on pharmacogenetics. These genes include the targets of the drugs, drug metabolizing enzymes as well as genes related to the drug via pathways and publications. We used this set of genes as the seed set in GeneRanker and obtained the corresponding ranked list of genes.

Next we filtered out from the list of genes for the ADRs, the ones that were associated with over 50% of the diseases mentioned in the CbioC database. This was done to reduce noise. Finally we found all the genes that were present in the overlap between the ranked list of genes obtained for a drug and its known adverse reaction. When finding overlaps we used only a certain number of the top ranked genes in the ADR gene lists but all the ones in the drug gene lists. This is because we were generally able to obtain a very small number of seed genes for the drugs. As mentioned earlier, an important part of the final score of a gene is its seed measure. Due to the small size of the seed sets of the drugs, this was not a very reliable means of judging the relevance of a gene in the network and thus the final ranking could not be trusted. Moreover, the gene network for the drugs tended to be relatively small and so we decided that using the entire list should not introduce too much noise.

The cut-off value for the ranked gene list obtained for a given ADR was decided by looking at the scatter graph of the score of a gene versus its rank and finding the rank at which it levels out. For instance the corresponding graph for tardive dyskinesia is shown in figure 1.

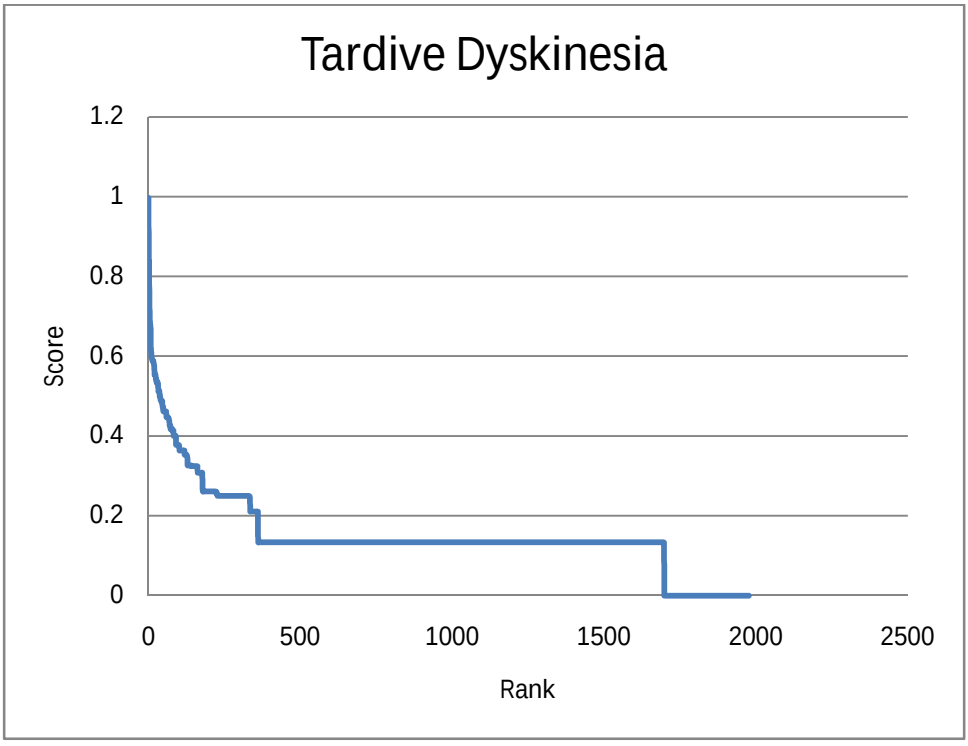

Figure 1: The score versus the rank of genes listed by GeneRanker for tardive dyskinesia

It is thus clear that the top 400 genes in this list will tend to be more relevant. We hypothesize that the genes involved in the mechanism through which the drug causes the adverse reaction are likely to be present in the overlap. Adverse effects caused by drug-drug interactions were also studied in a similar manner. In this case the relevant genes are assumed to be present in the overlap between all the drugs and the adverse effect.

3. Results

We mainly analyzed our results for drugs associated with tardive dyskinesia (TD) and long QT Syndrome (LQTS). TD is a drug induced hyperkinetic movement disorder characterized by purposeless, repetitive movements. About 20%–30% of the patients receiving prolonged antipsychotic treatment are seen to suffer from TD and genetic mutation is believed to play an important role in its development. Several genes whose polymorphisms are currently suspected to be related to antipsychotic induced TD include DRD2, DRD3, DRD4, GSTM1, GSTT1, COMT, CYP2D6, CYP1A2 and HTR2A. Studies have been conducted in various ethnic populations to determine the probability of different variants of these genes being linked to TD. Some of the results were contradictory. For instance several researchers found positive correlation between Ser9GLy DRD3 polymorphisms and development of TD while others found a negative correlation [9, 10, 11, 12].

The overlap of the ranked gene list of TD with that of the antipsychotic drugs ziprasidone, chlorpromazine and quetiapine had 119, 182 and 137 genes respectively. We tested to see whether genes whose polymorphisms are currently suspected to be related to antipsychotic drug induced TD were present in this overlap or not. DRD2, DRD4, COMT and HTR2A were present in all the overlaps. DRD3 was actually present in all of the four lists of genes but did not show up in any of the overlaps. This is because it is ranked 1103 in the gene network for dyskinesia and we are only considering the top 400. CYP2D6 and CYP1A2 were not present in the ziprasidone gene network but were ranked highly for quetiapine and chlorpromazine. Again they did not show up in the overlap for any of the drugs since their rank in dyskinesia was over 1000. GSTM1 and GSTT1 were present in the td-chlorpromazine overlap and td-quetiapine overlap but was absent in the td-ziprasidone overlap. These results are summarized in Table 1.

Gene	TD/chlorpromazine overlap	TD/quetiapine overlap	TD/ziprasidone overlap
DRD2	Present	Present	Present
DRD3	Absent (missed only due to low ranking in dyskinesia)	Absent (missed only due to low ranking in dyskinesia)	Absent (missed only due to low ranking in dyskinesia)
DRD4	Present	Present	Present
HTR2A	Present	Present	Present
COMT	Present	Present	Present
CYP1A2	Absent (missed only due to low ranking in dyskinesia)	Absent (missed only due to low ranking in dyskinesia)	Absent
CYP1D6	Absent (missed only due to low ranking in dyskinesia)	Absent (missed only due to low ranking in dyskinesia)	Absent
GSTM1	Present	Present	Absent

GSTT1	Present	Present	Absent

Table 1: Table showing results from the analysis of tardive dyskinesia and the antipsychotic drugs chlorpromazine, quetiapine and ziprasidone.

Cisapride and Clarithromycin are two drugs that are known to interact and cause Long QT syndrome, a ventricular arrhythmia characterized by a prolonged QT interval on the surface electrocardiogram [8]. 11 genes were found in the overlap of cisapride, clarithromycin and LQTS. We again evaluated this list as before. According to [8] drug induced LQTS can be a result of the mutation of 6 different genes- KCNJ2, KCNE1, KCNE2, KCNQ1, KCNH2 and SCN5A. We found the last three of these genes in the overlap of cisapride, clarithromycin and LQTS. All of these genes apart from KCNJ2 were present in the overlap between clarithromycin and LQTS which had a total of 14 genes. These results are shown in Figure 2 and Table 2.

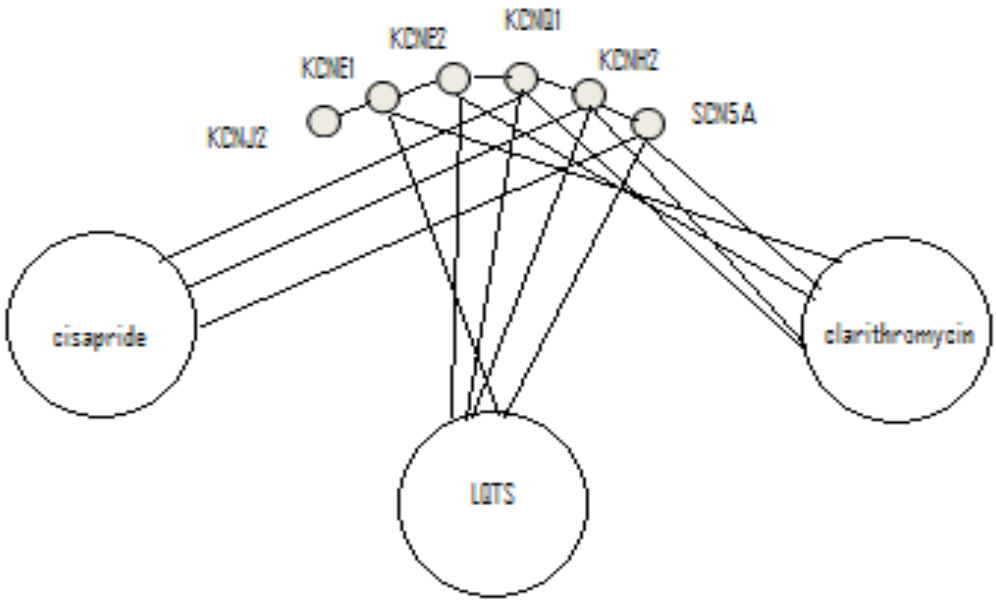

Figure 2: The genes whose mutations are said to be associated with drug induced LQTS are given at the top. An edge is drawn between each gene and each network which contains the gene.

Gene	Present in LQTS - Clarithromycin - Cisapride overlap?	Present in LQTS - Clarithromycin overlap?	Present in LQTS – Cisapride overlap?
KCNJ2	No	No	No
KCNE1	No	Yes	No
KCNE2	No	Yes	No
KCNQ1	Yes	Yes	Yes
KCNH2	Yes	Yes	Yes

SCN5A	Yes	Yes	Yes

Table 2: Table showing the results from the analysis of LQTS and the two drugs known to cause it: cisapride and clarithromycin.

4. Discussion and Conclusion

Our results indicate that GeneRanker can prove to be an important tool in pharmacogenetics and hence has the potential to aid development of personalized medicine. The fact that several of the genes whose mutations were linked to an adverse effect were present in the overlap between the gene list of the adverse effect and that of the drug suspected of causing it, suggests that studying the other genes in the overlap may prove to be useful as well. One interesting observation was the difference between the number of genes whose polymorphisms are associated with LQTS in the LQTS/Clarithromycin/Cisapride overlap as opposed to LQTS/Clarithromycin overlap. Since the latter contains more genes, it is possible that clarithromycin plays a more active role in the development of LQTS. Also, compared to the other antipsychotic drugs, ziprasidone had fewer of the suspected TD-linked genes in its overlap. This may indicate that ziprasidone induced TD is related to fewer gene polymorphisms.

Some of the genes whose variants are suspected to cause TD and LQTS did not show up in any of the relevant networks. This shows that GeneRanker, although quite effective, is not perfect. There are several reasons for this. First of all the databases used by GeneRanker are not complete and may not contain all possible protein-protein interactions and gene disease relationships. Thus GeneRanker may be missing several genes that actually are linked to the seed genes or may be ranking several of the important genes very low since interactions involving these genes are absent from the database. Also, GeneRanker simply retrieves all genes correlated with the ones in the seed set. It does not take into account how the genes are exactly related. It is possible for genes to be mentioned together in text without them being directly linked. Thus, some of the interactions retrieved by GeneRanker may also be false positives and this may be the reason we obtained such a large number of genes in the overlap between the gene network of drugs and that of their adverse effects.

Several of the genes suspected of causing TD did not show up in its overlap with the tested antipsychotic drugs as they were ranked below 400 in the TD gene network. This is an example of the well known tradeoff between precision and recall. Increasing our cut off level would have increased the number of relevant genes found in the overlap but would have also introduced approximately 100 more genes in the overlap, many of which may not have been useful.

Simply finding all the genes present in the overlap between the gene list of drugs and that of their adverse reactions however may not always be very successful and efficient. It is likely to be a lot more effective if we compared the networks obtained for drugs and their adverse reactions from GeneRanker. Several applications for comparing networks are available online. These include Cytoscape, NeAT, MetNetAligner, Graemlin etc. We are currently concentrating on using Cytoscape to take in the gene networks for drugs and ADRs obtained from GeneRanker and using it to find intersections between networks. However, due to the complexity of biological processes, data mining techniques still need to

develop considerably before they can produce any significant contribution towards biomedical discoveries.

1. Gonzalez G, Uribe JC, Tari L, Brophy C, Baral C. Mining Gene-Disease relationships from Biomedical Literature: Incorporating Interactions, Connectivity, Confidence, and Context Measures. Pacific Symposium in Biocomputing; 2007; Maui, Hawaii; 2007
2. Campillos et al. "Drug target identification using side-effect similarity." Science. 2008 321(5886):263-266.
3. Liebler and Guengerich. "Elucidating Mechanisms of drug-induced toxicity." Nature Reviews Drug Discovery. May 2005 4: 410-420.
4. Ma'ayan A et al. "Network analysis of FDA approved drugs and their targets."Mt Sinai J Med. 2007 74:27-32.
5. Nacher J. and Schwartz J. "Local and global modes of drug action in biochemical networks." BMC Chemical Biology. 2009 9:4
6. Scheiber et al. "Gaining Insight into Off-Target Mediated Effects of Drug Candidates with a Comprehensive Systems Chemical Biology Analysis." Journal of Chemical Information and Modeling. Feb 2009 49(2):308-317
7. Gonzalez G. , Uribe JC, Armstrong B., McDonough W., Berens M.E. "GeneRanker: An Online System for Predicting Gene-Disease Associations for Translational Research."
8. Aerssens J. and Paulussen A.D.C. "Risk factor for drug induced long QT syndrome." Neth Heart J. February 2005. 13(2):47-56.
9. Lattuada E. et al. "Tardive dyskinesia and DRD2, DRD3, DRD4, 5-HT2A variants in schizophrenia: an association study with repeated assessment." Int. J Neuropsychopharmacol. Dec 2004. 7(4): 489-93.
10. Lerer B. et al. "Combined analysis of 635 patients confirms an age-related association of the serotonin 2A receptor gene with tardive dyskinesia and specificity for the non-orofacial subtype." Int. J Neuropsychopharmacol. Sep 2005. 8(3):411-25.
11. Zai C. et al. "Association study of tardive dyskinesia and twelve DRD2 polymorphisms in schizophrenia patients". Int. J Neuropsychopharmacol. 2007. 10:639-651.
12. Foster A. et al. "Pharmacogenetics of antipsychotic adverse effects: Case studies and a literature review for clinicians". Neuropsychiats Dis Treat. Dec 2007 3(6):965-973