PREDICTING PREVENTABLE ADVERSE EVENTS USING INTEGRATED SYSTEMS PHARMACOLOGY

GUY HASKIN FERNALD[1], DORNA KASHEF[2], NICHOLAS P. TATONETTI[1]

Center for Biomedical Informatics Research[1], Department of Computer Science[2], Stanford University, Stanford, CA 94305, USA

Preventable adverse drug events are a major cause of morbidity and mortality worldwide. In many cases, these events could have been predicted in advance using current biological and pharmacological knowledge. Previous attempts to predict adverse events have been limited in scope and ability to predict off-target effects accurately. Recently, public data sources detailing adverse events and biological effects of chemicals have grown sufficiently robust to allow latent patterns to be revealed using automated predictive methods. We selected six severe adverse events reported in the Food and Drug Administration's Adverse Event Reporting System (AERS) and used the drugs and diagnoses for the reports to predict which adverse events had occurred. We designed feature vectors containing weights for the gene ontology categories that were perturbed by the drugs and diseases present in each adverse event report. We trained and compared the predictive capabilities of three machine learning algorithms: naïve Bayes, logistic regression, and support vector machines. Logistic regression and support vector machines performed best and ablative analysis of the features demonstrated which gene ontology categories are most predictive of the adverse events. Finally, by removing drugs from the training data we were able to demonstrate robust performance of our predictions on drugs unknown to our predictive model.

1. Introduction

Each year in the United States serious adverse drug events cause over 100,000 deaths (1). Recent advances in our understanding of the pharmacogenetics of drug activity present the opportunity to integrate biological data and adverse event reports together to predict off target drug side effects. In the United States, adverse events are reported and stored in the U.S. Food and Drug Administration's Adverse Events Reporting System (AERS), which contains nearly 3 million voluntary reports from 2004-2008 alone (2). Each report details the drugs involved, characteristics of the patients, and the observed adverse events. Unfortunately, the database remains underutilized for predicting adverse events. We hypothesized that by integrating knowledge and applying machine learning algorithms we could accurately predict adverse events.

Drugs are often designed with a goal of targeting just one gene or genetic pathway, however, most commonly the drug will also interact with many more unintended targets, which can lead to adverse events. In limited cases the

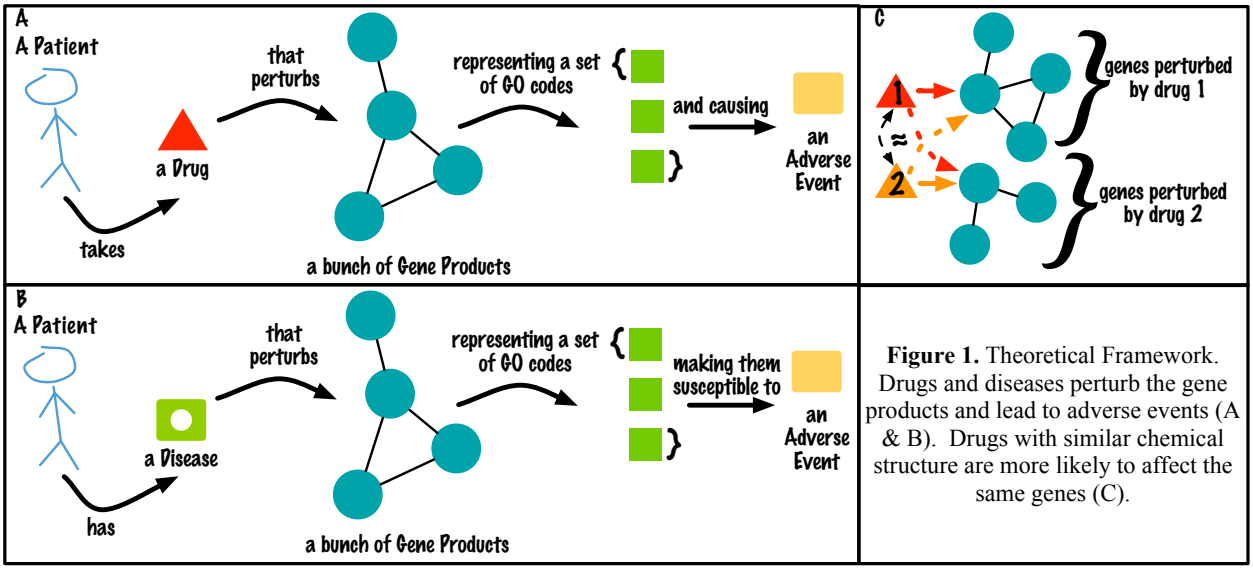

Figure 1. Theoretical Framework. Drugs and diseases perturb the gene products and lead to adverse events (A & B). Drugs with similar chemical structure are more likely to affect the same genes (C).

observed adverse events can be predicted as a consequence of known actions of the drug. However, for most drugs it is very difficult to accurately predict which side effects will be observed(3). The most recent evidence suggests that to accurately predict off-target drug effects, a systems approach is required.

Our approach is based on a theoretical framework that assumes that perturbations in gene products will lead to phenotypic changes in the state of the patient. In the case of a patient with an adverse event there are two main sources of perturbations in gene products: 1) the drugs that patient is taking; and, 2) the disease or diagnosis of the patient. When a patient takes a drug the action of the drug will ultimately alter the effects of gene products in the body or cause the body to initiate new biochemical processes. These changes can lead to adverse events.

Additionally, a patient has been diagnosed with one or more diseases which may have already altered the biochemical state of the patient and made the more susceptible to adverse events.

Using this framework we can view a patient with an adverse event as having a set of genes that have been perturbed by both the drugs administered and the diseases present (figure 1 A,B). In many cases there can be thousands of genes which are thought to be affected by either the drugs or diseases of the patient. For the purposes of prediction and classification this would require that each patient be represented by feature vectors with thousands of genes. For many classification algorithms this is impractical. The computation may be intractable or the model may over-fit the data. To avoid this problem we hypothesized that we could represent the perturbed genes for each event report as terms in the Gene Ontology(4). By choosing a higher level abstraction to represent the affected genes we reduced our feature set to a size that is both tractable and avoids over-fitting.

Drugs exact their biochemical effects by binding to compounds in the body. For example, often a substructure of the drug will bind to an active site on a protein and change the behavior of that protein, resulting in a change in behavior. In previous work it has been shown that drugs that are structurally similar are more likely to share genetics(5). To incorporate this concept we hypothesized that one drug would influence the same genes as a similar drug with a probability proportional to the similarity score (figure 1C). A drug that is identical will have a similarity score of 1 and will influence the same genes. Conversely, a drug with a similarity score of 0 can not be assumed to effect the same genes.

2. Methods

A report in the AERS database is comprised of both the drugs the patient was prescribed and the diseases for which the patient was receiving treatment. It is necessary to translate the reports of drugs and diseases into a single vector. We accomplish this by mapping both drugs and diseases to genes and then map those genes to Gene Ontology (GO) terms. Once the drugs and diseases have both been mapped to a common set of features they can be combined to generate a final feature vector for the report.

The expected baseline performance of the methods was established by constructing feature vectors based on only presence or absence of each drug in the report. The feature vector for each report was a binary vector where the indices of the report's drugs were set to 1. The positive set of reports for each event was comprised of all of the positive examples for that report and a balanced negative set of reports was randomly chosen from the remaining reports. The number of positive reports in the AERS for each adverse event was 10,613, 7,021, 3,462, 3,283, 2,917, 1,595 for rhabdomyolysis, Stevens-Johnsons Syndrome, QT prolongation, cholestasis, deafness, and Torsade de Pointes, respectively.

AERS and PharmGKB use different terminologies to describe drugs and diseases. We used the Open Biomedical Annotator to map terms used by the Adverse Event Reporting System to those used by the Pharmacogenomics Knowledge Base. This mapping allows us to integrate the data across the two data sources effectively and efficiently.

Genes are the major drivers of biology and therefore any perturbation of the homeostasis of the genetic network may result in side effects. Drugs are designed with specific genetic targets in mind so that the genetic network will be perturbed in a predictable way. However, drugs often bind to unintended targets and can cause unexpected perturbations to the genetic network(6). Hypothesizing that unintended drug-gene interactions may be the cause of side effects we relate drugs to their pharmacogenes (the genes that interact with the drug) using the Pharmacogenomics Knowledge Base (PharmGKB)(7). While the PharmGKB contains very high quality relationships between drugs and their pharmacogenes the data is very sparse, and so very little underlying pharmacogenetics may be known about the small set of drugs that are associated with one report from the AERS. To remedy this we use a technique, developed by our lab, for predicting the pharmacogenes for a drug. In previous work we showed that drugs that were structurally similar were much more likely to share a pharmacogene(5). First, we defined a drug-drug similarity matrix by calculating the pairwise structural similarity of each drug reported in the AERS. We then use the *max-dot product* to multiply the drug-drug similarity matrix by the drug-gene pharmacogene matrix derived from PharmGKB. Where the *max-dot product* for two matrices, A, a m by n matrix, and B, a n by r matrix, is defined by equation 1.

$$(A\ m \cdot B)_{ij} \quad = \quad max(A_{ik} \cdot B_{kj}) \ \forall \ k \ \in \ \{1, \ldots, n\} \qquad (1)$$

This matrix multiplication of the drug-drug similarity matrix by the drug-gene matrix allows us to project all drugs (including those for which no known pharmacogenetics exists) into gene space. While this projection provides insight into the biology of action of the drug it also projects the drug vectors into a higher dimensional space. To reduce this feature space we employed a biological feature reduction method. We did so by constructing an interaction matrix between genes and Gene Ontology terms. We chose to use 283 high level ontology concepts representing general biological processes and functions. We combine our drug-gene matrix with this gene-go matrix by performing another *max-dot product*. The result of these two matrix multiplications is a drug by GO matrix which

represents the projection of the drugs into gene ontology space (figure 2A). As the gene ontology terms represent high level biological concepts this projection represents the function of the drug and provides a tractable sized matrix from which we can derive our feature vectors. Equation 2 shows the matrix multiplications that result in the drug-go matrix.

$$DrugGO \quad = \quad DrugDrug \cdot DrugGene\ m \cdot GeneGO \qquad (2)$$

Each report in the AERS also contains the diseases for which the patient was being treated. In order to effectively integrate these features with the drug feature matrix we have defined above [Eq (2)] we mapped the indications to the 283 gene ontology terms in the same manner. We defined a disease-gene matrix from the curated relationships found in PharmGKB and performed a *max-dot product* with the gene-go matrix [Eq (3)]. This results in an indication-GO matrix which now takes on the same dimension as the drug-GO matrix (figure 2B).

$$IndicationGO \quad = \quad IndicationGene\ m \cdot GeneGO \qquad (3)$$

We used the drug-GO and indication-GO matrices defined in the previous two sections to construct the feature matrices for each report in the AERS. For each report we extract the rows in the drug-GO matrix that correspond to the drugs listed in the report. We then take the max value of this matrix for each column. This results in a vector of 283 values which indicate the maximum projection of the drugs of the report into gene ontology space. We repeat this method for the indication-GO matrix as well to construct a second vector of 283 values which represents the maximum projection of the indications of the report into gene ontology space (figure 2D). Finally we combine these two vectors by simply adding their components (figure 2E). This results in our final feature vector for this report which summarizes the overall projection of the report into gene ontology term space. Using this method we created one feature vector for each of the 70,000 reports.

We used the Weka, Matlab, SAS, and SVMLite machine learning libraries to train naïve Bayes, logistic regression, and support vector machine learning algorithms on the features and validated the predictions using 10-fold cross validation. We compared the performance of the algorithms using the following summary statistics: Area Under the Receiver Operating Characteristic Curve (AUROC) and Area Under the Precision-Recall Curve (AUPR).

We investigated the power of our methods for predicting adverse events for novel drugs. A novel drug is one where the pharmacogenomics of the drug are unknown and for which no side effects for the drug are known. To simulate this situation we randomly chose a set of 20 drugs and removed any pharmacogenomic relationships for those drugs from the drug-gene matrix. We also removed all of the reports which contained any of the drugs from the training data. We then trained the machine learning algorithms and assessed the ability of the classifiers to predict the adverse events for each novel drug report.

Additionally, we ran 283 separate logistic regression classifiers by removing the gene ontology terms one at a time. By comparing the difference between the entire model (283 GO terms) AUROC values for these classifiers (282 GO terms) on 10-fold cross validation we generated a Δ AUROC value for each feature, which gives an indication of how much each GO term feature contributes to the overall predictive power.

3. Results

To establish an expected performance for our classifier we used the binary feature vectors for drug presence to construct a positive and negative training set. We trained a naïve Bayes classifier and validated it with 10-fold cross validation. We plotted six ROC curves representing the average performance of the classifier on each adverse events, (figure 3). The average AUROC for these classifiers was 0.74.

Table 1. AUROC Comparison.

	RH	SJ	DF	CH	TP	QT
Naïve Bayes	0.79	0.76	0.71	0.70	0.80	0.75
Logistic Regression	0.87	0.85	0.81	0.80	0.87	0.83
SVM	0.78	0.79	0.72	0.71	0.78	0.77

Using the methods described above and illustrated in figure 2A we derived the drug-Gene Ontology "drug-GO" matrix. First the drug-gene matrix (597 x 1696) was multiplied using the *max-dot product* by the gene-GO (1696 x 283) matrix to yield a drug-GO matrix (597 x 283). Then a drug-drug similarity matrix (1271 x 597) was multiplied using the *max-dot product* by the drug-GO matrix (597 x 283) to yield a final drug-GO matrix (1271 x 283). Similarly we derived the indication-Gene Ontology "indication-GO" matrix by taking the *max-dot product* between an indication-gene matrix (672 x 2762) and a gene-go matrix (1696 x 283) to yield a indication-GO matrix (672 x 283). For each report we extracted the drug and indication rows from the drug-GO and indication-GO matrices and then took the maximum column-wise to yield two penultimate feature vectors. These vectors were then combined using element-wise addition to yield the final feature vector for the report. 28,236 reports contained both indications and drug which could be mapped and the remaining reports were dropped from the analysis.

We trained and cross-validated naïve Bayes, logistic regression, and support vector machine learning algorithms on the training data and compared performances using the AUROC (table 1). The logistic regression classifiers had

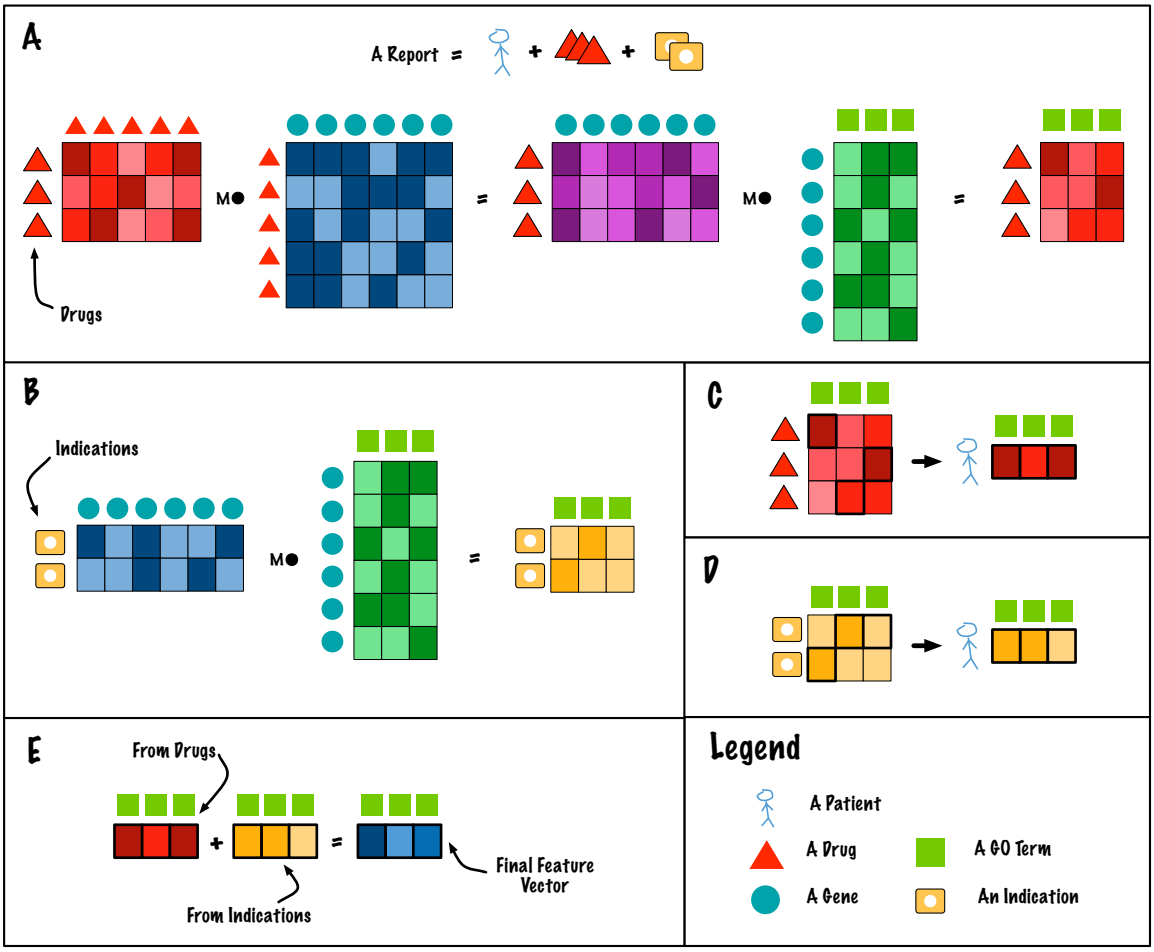

Figure 2. (A) Derivation of the drug by Gene Ontology matrix. A drug-drug similarity matrix is multiplied by a drug-gene matrix using the *max-dot product* to yeild a drug-gene matrix. The drug-gene matrix is then multiplied by a gene-GO matrix using the *max-dot product* to yeild a drug-GO matrix. (B) Derivation of the Indication-Gene Ontology matrix. A disease-gene matrix is multiplied by a gene-GO matrix using the *max-dot product* to yield an indication-GO matrix. (C) The max of each column of the patient-drugs by GO matrix is taken to construct a drug feature vector for the patient. This vector represents the cumulative projection of the drugs listed in the patient's report into GO space. (D) The max of each column of the patient-indications by GO matrix was taken to construct an indication feature vector for the patient. This vector represents the cumulative projection of the diseases of the patient into GO space. (E) The patients drug feature vector and indication feature vector are added element-wise to create the final feature vector for the record. This entire process is repeated to create a final feature vector for reach of the 70,000 reports.

the best overall performance; 0.87, 0.85, 0.81, 0.80, 0.87, and 0.83 for rhabdomyolysis, Stevens-Johnsons sysndrome, deafness, cholestasis, Torsade de Pointes, and QT prolongation, respectively. For those same adverse events the AUPRs were 0.87, 0.85, 0.79, 0.77, 0.85, and 0.79 (figure 4).

Under ablative analysis each of the GO term features did not contribute a substantial amount to the AUROC for rhabdomyolysis. The largest Δ AUROC values were 0.02, and the top three GO terms returned by the analysis were transcription activator activity, protein binding, and transport respectively.

To assess the methods ability to predict side effects of novel drug structures we dropped all data about a randomly chosen set of 20 drugs and constructed a training set of the remaining reports. The reports which contained any of the randomly selected 20 drugs were placed in the test set. The logistic regression classifier successfully predicted the adverse events for these novel structures with AUROC's of 0.77, 0.75, 0.63, 0.80, 0.79, and 0.80 for rhabdomyolysis, Stevens-Johnson's syndrome, deafness, cholestasis, Torsade de Pointes, and QT prolongation, respectively.

4. Discussion

In this work we showed that the FDA's Adverse Event Reporting System can be used as a rich source of data for the training of machine learning algorithms to predict adverse events. The reports in the AERS contain information

about the drugs that a patient was prescribed, the diagnoses for that patient, and the adverse events which occurred. We first presented a novel method of data integration which takes advantage of the *max-dot product*, a matrix multiplication technique appropriate for biological data. We used this integration technique to integrate a drug-drug similarity matrix, a drug-gene pharmacogene matrix, and a gene-GO feature reduction matrix to yield a drug-GO matrix (figure 2A). We then similarly integrated a disease-gene matrix with the gene-GO feature reduction matrix to yield a disease-GO matrix (figure 2B). Finally we present an algorithm for utilizing these matrices to construct feature vectors for each patient record in the AERS (figure 2E). Finally, we used the matrix of feature vectors to train naïve Bayes, logistic regression, and support vector machine learning algorithms and validated their performance. We showed that given a patient, which we represent as simply set of drugs and a set of diseases, we can accurately predict which adverse events are most likely to occur (average AUROC = 0.838). We demonstrated that the algorithm achieves significant performance for each of the six adverse eve nts (figure 4).

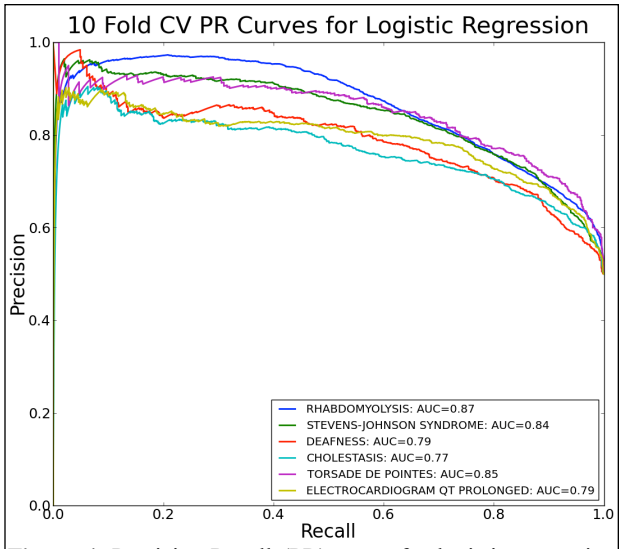

Figure 4. Precicion-Recall (PR) curve for logistic regression on each of the six adverse events.

We use the *max-dot product*, which is more biologically relevant than the dot product as it is much less susceptible to noise. This becomes evident in the case where we integrate a drug-drug similarity matrix with the drug-gene pharmacogenomics matrix. Consider an unknown drug A and a set of drugs S which all bind to a gene B. In the regular dot product drug A may look only slightly similar to all of the drugs in S, however, the result of the dot product would indicate a very high likelihood of interaction between drug A and gene B, especially as the size of S increases. In the *max-dot product* the likelihood of drug A interacting with gene B can only be as great as the highest pairwise similarity between drug A and the drugs in set S. In previous work our lab demonstrated that this method is a better estimate for this potential interaction and avoids introducing false positive interactions (5).

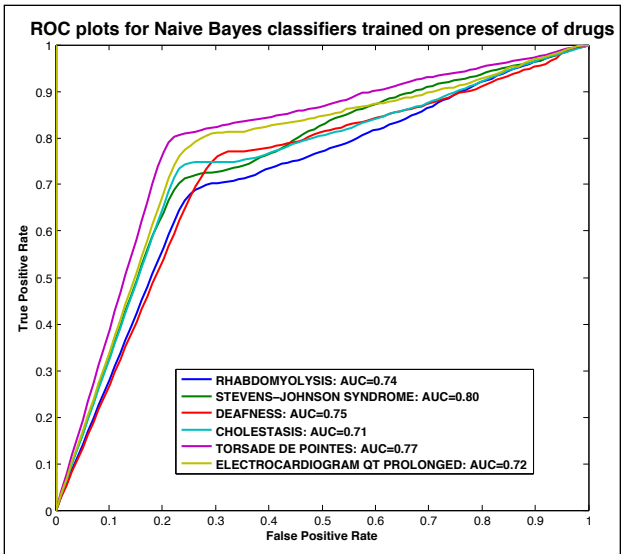

Figure 3. Receiver Operating Characteristic curves for each of the 6 adverse events trained on the binary drug presence feature data using the naïve Bayes classifier.

Our work demonstrates the utility of the AERS as a data source for training machine learning algorithms to predict the likely side effects for patients and novel drugs. The methods we present here are general so that other other adverse event reporting systems, such as the Canadian or British systems, can also be integrated to improve the prediction power of the algorithms. The methods and system for predicting adverse events will be a vital tool for medical researchers as well as pharmaceutical companies, both of which have a vested interest in anticipating side effects of the drugs they are researching.

References

1. Scheiber, J., et al., Journal of chemical information and modeling, 2009. **49**(2): p. 308-17.
2. Pratt, L.A. and P.N. Danese, Nat Biotechnol, 2009. **27**(7): p. 601-2.
3. Tatonetti, N., T. Liu, and R. Altman, Genome Biol, 2009. **10**(9): p. 238.
4. Müller, H.M., E.E. Kenny, and P.W. Sternberg, PLoS Biol, 2004. **2**(11): p. e309.
5. Hansen, N.T., S. Brunak, and R.B. Altman, Clin Pharmacol Ther, 2009. **86**(2): p. 183-9.
6. Adams, J.C., et al., PLoS Comput Biol, 2009. **5**(8): p. e1000474.
7. Klein, T.E. and R.B. Altman The Pharmacogenomics Journal, 2004.