Network models to improve drug-ADE prediction

Anirban Chatterjee,Shabaz Basheer Patel,Himanshu Bhandoh
Stanford University

Email: achatter@stanford.edu,shabaz@stanford.edu,hbhandoh@stanford.edu

***Abstract*—This project aims to investigate drug-protein-ADE causation relations to identify target proteins associated with the reported ADEs. We aim to investigate new drug vs existing drug target protein commonalities to predict likely new drug - ADE relationships based on existing known target protein-ADE relationships. This would help optimize clinical trials for focused testing on these predicted ADEs for new drugs, and would drastically increase likelihood of ADE detection in the clinical trial stage versus ADE detection post drug release for treatment, thus averting serious uncontrolled and often fatal adverse effects. We approach this problem by using genomic network models and as our results show, this gives a better performance than the present state of art which uses logistic regression.**

***Index Terms*—Adverse Drug Events (ADE), Support Vector Machine (SVM), Naive Bayes, Neural Networks**

I. Introduction

Accurately identifying adverse drug events early is an increasing concern in the medical industry. Medical errors have been the cause behind death and injury of over one million patients in the US alone. ADEs contribute to about fifth of that number (IOM,1999). Even though there are a number of surveillance techniques in practice that monitor ADE, studies indicate that the best review technique is chart monitoring[1], which however, is time-expensive. Hence, several automated learning techniques have been employed to tackle this issue. The most impactful research in this field has been the work by Cami et. al.[2], where pharmaco-safety networks(PPNs) have been employed. In PPNs, known drug-ADE relationships on specific drugs are used to predict likely unknown ADEs. The crux of this predictive approach relies on leveraging existing, contextual drug safety information, thereby having the potential to identify certain ADEs very early. By training a logistic regression model, predictions for several ADEs were made which were not listed in 2005 database. The findings suggested that predictive network methods can be useful for predicting unknown ADEs.

We investigated drug-target-ADE causation relations to identify target proteins associated with the reported ADEs. Thereby, we investigated new drug vs existing drug target protein commonalities to predict likely new drug - ADE relationships based on existing known target protein-ADE relationships. Similarity between drugs are often computed using pharmacogenomic approach [3] which uses adverse drug events vectors and computes a cosine correlation coefficient. Another popular approach is the chemogenomic approach [4] which computes drug-drug similarity by evaluating chemical structural similarities between drugs. In our project we have adopted a genomic approach to the problem where drug-drug similarity is computed by comparing the target proteins.

II. Plan of Approach

To formalize the problem, we represent each drug as a 1-by-n dimensional vector with binary entries. The i^{th} entry in this vector indicates whether the drug targets the ith protein. With m such vectors corresponding to each of the m drugs in our database, we can construct an m-by-n dimensional matrix. Given a particular ADE, we construct a mx1 vector y such that the j^{th} entry indicates whether the j^{th} drug has at least one mention of the ADE in our adverse reports database.

Table 1: Drug/ADE matrix structure:

Drug/ADE	hsa:3030	.....	hsa:2025	hsa:5054
D00066	1	.....	0	1
.....				
D00144	0	.....	1	0

The modeling as done in the above described method incorporated the effect of target proteins that were known for certain to be hit by drugs. However, drugs often have unknown/hidden targets that are not reported by clinical trials, and often these targets majorly contribute to ADEs. After analysis through the previous model, we decided to modify the drug-target matrix to include possible (hidden) targets given our knowledge of the targets that are known for sure to be hit.

We hypothesized that genomic sequence similarity of two targets might be a useful indicator of how closely related (w.r.t genomics) two targets are. Hence, with the prior knowledge that a drug hits a certain target, it might be possible to predict the probability that the drug will also hit another (hidden) target given the genomic similarity between the two targets. To ascertain the genomic similarity between two targets, we computed the Smith-Waterman score, due to Smith and Waterman [5]. This score uses local alignment to compute commonalities between two genomic sequences. The average score that the scoring system would yield for a random sequence is the output expectation score.

We obtained the Smith-Waterman similarity coefficients between all our target-target pairs as calculated by Yamanishi. With these coefficients, we computed the likelihood for a drug

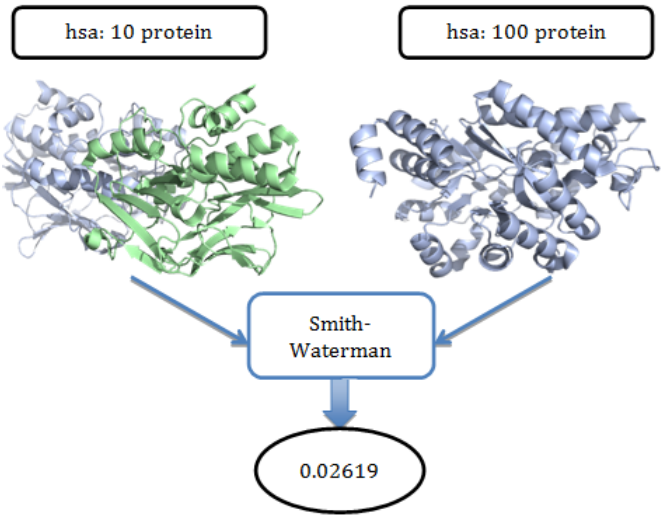

Fig. 1: Data Collection Lists

to target a hidden protein by updating the probability of being present as the previously calculated probability multiplied by the drug target similarity coefficient. This allowed us to modify table 1 to include probability values for drug-target interaction for all drug-target pairs. We proceeded with the prediction algorithm with this new modified matrix (Table 2).

Table 2: Drug/ADE matrix structure:

Drug/ADE	hsa:3030	.....	hsa:2025	hsa:5054
D00066	1	.....	0.34	1
.....				
D00144	0.78	.....	1	0.27

y-vector Structure:

Dyspnonia	1	.....	1	0

III. Data Collection

Two drug databases were consulted to gather the required data. Kegg(Kyoto Encyclopedia of genes and genomes) lists drugs by their database entries(D numbers) and lists chemical and structural properties of the drugs along-with their commercial names(TN, USP,etc.). AERS(now FAERS) lists information on adverse drug events and medication reports.

From the Kegg database, drug-target protein datasets for known drugs were obtained alongwith their commercial names. The drug-target protein matrix is a set of 1s (representing whether a particular drug targets a specific target protein) and 0s (representing that a drug does not target a protein).

The most recent dataset from the AERS database was collected and drug-ADE reports were registered.The drug-ADE database lists out the report numbers and ADEs pertaining to these reports for the 4th quarter of 2012. We obtained the database tying the reports to drugs, and tied

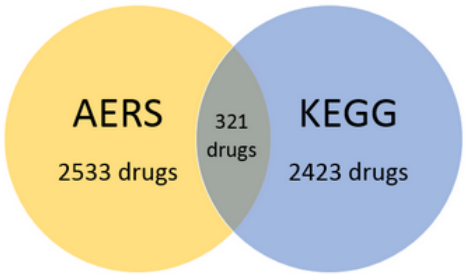

Fig. 2: Data Collection Lists

back all the drugs to the respective ADEs. We used this data to create an output vector for each ADE, with 1s and 0s representing whether or not a particular drug caused a specific ADE respectively. Then, in order to improve the performance of the prediction we decided to modify the drug-target matrix to include hidden targets given our prior knowledge of the targets that are sure to be hit. We have assumed that any adverse event report for a drug claiming that a drug caused an adverse event is true, and that the drug causes this adverse event.

We mapped the drug v/s target proteins matrix listed in Kegg with the AERS drug event listing for drugs common to both Kegg and AERS, and were thus able to map a list of 321 drugs corresponding to 3965 adverse events (listed in AERS) with the corresponding target/non-target data for 436 target proteins for the 321 drugs. In order to achieve the modified drug-target interaction matrix, we obtained the Smith-Waterman similarity coefficients between all our target-target pairs as calculated by Yamanishi. With these coefficients in hand, we computed the likelihood for a drug to target a hidden protein. This allowed us to modify the previous drug-target interaction matrix to include a probability value for all the hidden drug-target interactions. We proceeded with the prediction algorithm with this new modified matrix.

We created a network mapping the drugs to their target proteins thereby creating a m-by-n dimensional matrix that is sparse. The network shown in figure 3 depicts the drug and target protein interaction network. This is the conventional approach followed in order to predict an event. We also proceed with the modified drug-target interaction matrix to obtain better performance.

IV. Training and Testing

Given the nature of our problem definition, there are multiple learning algorithms that can be employed. We considered a single ADE first and decided to train the data initially using Naive Bayes and SVM. The error in testing for differents algorithms for the five ADEs for both the

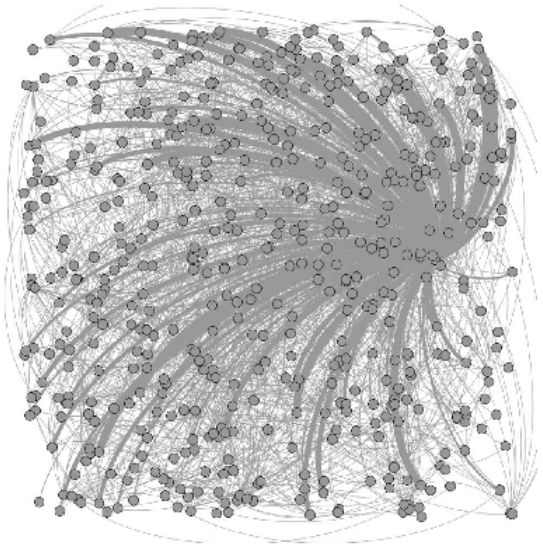

Fig. 3: Network Model of Drugs vs Targets

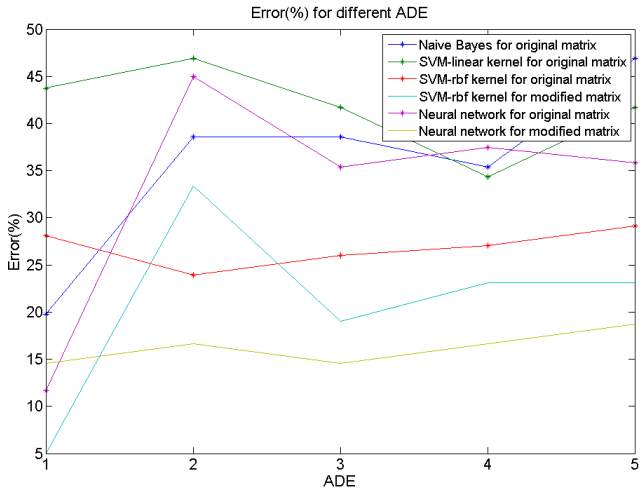

Fig. 4: Events Vs Error% for various methods

Drug-ADE interaction matrices are tabulated below:

Table 3: Drug/ADE Prediction Error through Naive Bayes using the original Drug-Target matrix:

ADE	Naive Bayes(Error%)
1.Diarrhoea	19.79
2.Pneumonia	38.59
3.Sinusitis	38.59
4.Alopecia	35.41
5.Arrhythmia	46.88

Table 4: Drug/ADE Prediction Error through SVM using the original matrix:

ADE	SVM-Linear kernel (Error %)
1.Diarrhoea	43.75
2.Pneumonia	46.88
3.Sinusitis	41.67
4.Alopecia	34.38
5.Arrhythmia	41.67

Table 5: Prediction error through SVM using rbf kernel for both the matrices:

ADE	Original Matrix(%)	Modified Matrix(%)
1.Diarrhoea	28.13	5
2.Pneumonia	23.96	33.34
3.Sinusitis	26.04	19.05
4.Alopecia	27.08	23.09
5.Arrhythmia	29.17	23.09

We proceeded on by using classifier such as neural networks for the present data set. For five ADE s we trained a neural network with 45 hidden nodes. We ran the algorithm on both the original and modified matrix and noted the error percentages in table 6. As can be inferred from the error percentages, there was a significant reduction in the average error percentage:

Table 6: Prediction error through Neural Network Classifier for both the matrices:

ADE	Original Matrix(%)	Modified Matrix(%)
1.Diarrhoea	11.66	14.58
2.Pneumonia	44.99	16.66
3.Sinusitis	35.414	14.58
4.Alopecia	37.49	16.67
5.Arrhythmia	35.83	18.75

We noted that calculating the hidden target protein matrix helped improve accuracy of our predictions considerably. This may be attributed to the property of promiscuous drugs being prone to hitting similar targets. Therefore, the drug is susceptible to binding to an adverse event-causing pocket due to the similarity in target protein structure with a protein that was originally intended as target.

V. Conclusion

We noted that just by considering drug-target interactions alone without additional variables, we were successful in predicting ADE with an accuracy of up to 86% using neural networks. We found that drug-ADE prediction accuracy improved considerably by calculating the protein target similarities to calculate hidden target protein likelihood. The effect can be attributed to the fact that ADEs are caused by drugs attacking hidden target proteins in addition to intended targets based on target-target genomic structure similarity. Errors in prediction can be attributed to more complex drug-drug interrelations, wherein two or more drugs may interact to attack different target proteins. More intensive and complex algorithms may help cater to drug-drug inter-effects to improve ADE prediction. We noted that the small size of our training dataset may be an impediment in the accuracy of ADE prediction. Collection

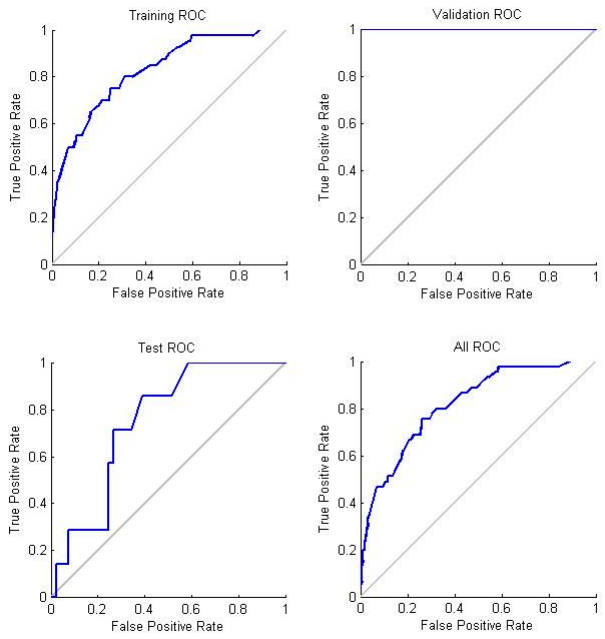

Fig. 5: ROC Curves for neural network classifier using modified matrix

of more data by combining information from multiple public databases i.e. AERS, SIDER, JAPIC etc. might help improve the accuracy of event prediction.

Additionally, considerations around ethnicity, gender, medical history and economic background etc may be crucial factors in improving predictability of ADEs. The dearth of availability of data for such parameters may make the pursuit a challenging one. Considering higher order drug-drug interactions may also help improve ADE predictability, but is a computational and time intensive task. Also, it wont work well in the case of immune-mediated adverse events.

VI. Acknowledgement

Authors acknowledge and extend heartfelt gratitude to Hamsa Bastani for her assistance and guidance in the project.

References

[1] Jha, Ashish K., et al. ”Identifying adverse drug events development of a computer-based monitor and comparison with chart review and stimulated voluntary report.” Journal of the American Medical Informatics Association 5.3 (1998): 305-314.

[2] Cami, Aurel, et al. ”Predicting adverse drug events using pharmacological network models.” Sci Transl Med 3.114 (2011): 114ra127.

[3] Takarabe, Masataka, et al. ”Drug target prediction using adverse event report systems: a pharmacogenomic approach.” Bioinformatics 28.18 (2012): i611-i618.

[4] Yamanishi, Yoshihiro. ”Chemogenomic Approaches to Infer DrugTarget Interaction Networks.” Data Mining for Systems Biology. Humana Press, 2013. 97-113.

[5] Smith TF, Waterman MS., Identification of common molecular subsequences, J Mol Biol. 1981 Mar 25;147(1):195-7