COMPARISON OF COMPUTATIONAL MODELS IN THE PREDICTION OF ADVERSE DRUG REACTION

Jumoke Soyemi[1], Oyelade Jelilli[2] and Adebiyi Ezekiel[3]

[1]Department of Computer Science
The Federal Polytechnic, Ilaro.
Ogun State, Nigeria**.**

[1,2,3]Department of Computer and Information Sciences
Bioinformatics Research Cluster
Covenant University, Ota.
Ogun State, Nigeria.

Abstract

Computational pharmacology is the application of bioinformatics and computational biology with relevance to pharmacology, including understanding of drug action, adverse drug reaction, identification of drug targets and drug design. Early and accurate identification of adverse drug reactions (ADR) is critically important for drug development and clinical safety. Often times the adverse effect of drugs are not discovered until years later after the drugs' release to the market. The post hoc analysis is usually unable to detect rare or delayed on-set ADR until clinical evidence accumulates. The process of drug development and ADRs discovery takes years, meaning that a lot of harm would have been caused to lives before evidences are accumulated, therefore developing a computational pharmacology model that can be used to make informed decisions so as to reduce the rate of attrition in drugs under development and increase the number of drugs with an acceptable benefit/risk ratio is paramount. This paper reviews the computational methods that have been used so far to address this issue and also compare them..

Keywords: Adverse drug reactions, Computational pharmacology, Computational methods.

Introduction

Serious and apparently unpredictable adverse drug reactions continue to be a major public health problem. Studies of toxicity and unintended side effects can lead to improved drug safety and efficacy. Also early and accurate identification of adverse drug effect is critical to public health. The post hoc analysis is usually unable to detect rare or delayed on-set ADR until clinical evidence accumulates. This process can take years. Computational method will enable the identification of adverse drug reactions (ADRs) outside of the drug development laboratory. This method will help researchers to accurately pinpoint certain adverse events based on the interactions between drug proteins and drug compounds.

There is need to identify side effects of drugs earlier in the drug development cycle to save lives and reduce costs. There have been cases in which medications with off-target protein side effects have reached the marketplace and later been recalled. There is need to determine

the safety of such therapies before they reach patients. Given the long drug development timeline, high development costs, and low chance of a new drug in the early stages of testing ever reaching the pharmaceutical market, the process could provide a less expensive option for certain types of drug testing.

During the drug discovery process, candidate drug compounds are typically mixed with proteins associated with a specific disease to assess drug efficacy and toxicity. The method can identify side effects with several types of target proteins; however, drug compounds binding to off-target proteins can result in unwanted side effects or adverse events and the sheer amount of potential off-target bindings makes testing cost prohibitive for companies. As a result, certain ADRs may remain undetected in clinical trials, only to be discovered once the drug reaches patients. Hence computational method of drug discovery should be expanded and more research to be encouraged in this wise since a wrong drug released into the market can claim millions of lives at once. This is the primary motivation of this paper and to expose researcher more to this area of need.

The paper is divided into seven sections; section one introduces the paper, Pharmacology, Drug and Adverse Drug Reactions are explained in section two. Section three looks at some Computational Methods of drug-ADR prediction while section four compares the methods. The benefits of using Computational Methods over Conventional method were looked into in section five and finally section six concludes the paper.

Pharmacology, Drug and Adverse Drug Reactions (ADR)

Pharmacology

Pharmacology is the science that studies drugs, their interaction on the biological system with the purpose of understanding the drug properties, the actions, interactions of drug molecules and effect of such interactions [1]. Pharmacology studies include the treatment and prevention of major diseases using drug therapy. Although a lot of development has been recorded in producing new drugs, a major challenge is side effect produced by several of such drugs. Some side effect could be minor and when the effect becomes adverse, it then becomes a serious concern since it could threaten the life of individuals.

Pharmacology is subdivided into Pharmacokinetic and Pharmacodynamics. Pharmacokinetics is the study of movement of drugs into, within and out of the body. It is simply what the body does to the drug or medicine that is the changes in drug concentration as the drug moves through the different compartments of the body. Pharmacodynamics is the study of drug effect and the mechanism of action. In the simplest term it is what the drug does to the body that is drug toxicity in the body. Pharmacology models are categorized into Pharmacokinetic and Pharmacodynamics. Pharmacokinetic models predict the time dependence of a drug's concentration in the body fluids following its administration while Pharmacodynamic models deal with the action of the drug once it reaches its target organ. A complete pharmacology model would integrate both such that each type is independently useful for different purposes.

Pharmacology and toxicology are related areas that demands understanding of certain properties and actions. Nevertheless, for pharmacology, the emphasis is on the effect of drug therapy while toxicology emphasis is on adverse drug reaction of drug and the associated risks

Drug

In the perspective of system biology, drugs are molecules that cause perturbations to biological systems with several molecular interactions including protein-protein interactions, metabolic pathways and signal transduction pathways, leading to the observed adverse effect [2]. The response of the body to drug shows both the expected effect of target interaction as well as combination of the entire effect off target interaction. The fact that a drug has a strong attraction for its target does not imply that it doesn't bind to other protein pockets with changing affinities that result in possible side effects. The concept is demonstrated through comparing pathways upset by toxic compounds which create associations between drug side effects and biological pathways [3]. Apart from the fact that in vitro could be used to determine side effect by testing compounds using biochemical and cellular assays, the experimental discovery of drug adverse effect is very high in respect to cost and efficiency [4]. This calls for the computational method to predict drug adverse effect on time before getting to the clinical stages to minimize resources such as fund and time used.

Adverse Drug Reaction (ADR)

Toxicology is the study of adverse drug reaction of drugs on the body and how to check such effect. The outcome of medicine on various systems of the body may result in what is known as Adverse Drug Reaction. Adverse drug reaction is unsafe and poses risk to the health of individual and in most cases such drugs are withdrawn from the market. Most serious ADR can be classified as either; type A, where the underlying mechanism is dose dependent, or type B or idiosyncratic, where the event is not predictable from the normal pharmacology of the drug and is generally independent of dose [5][6]. ADR has been reported in [7] to cause 10-30% hospital admissions, 30-150 billion dollars annual costs, 180,000-life threatening or fatal ADRs annually, 50% of these could have been prevented. Also a clinical trial of a drug uses 1000s of patients. It is therefore critical that prompt and exact identification of ADRs be done and on time before they get to the market for the safety and development of public health.

Traditional Method of drug development and ADR Detection

Drugs ADR are first studied in animals through extensively series of laboratories before being tested in human. The preclinical evaluation of drug therefore involves in vitro and in vivo studies in animal to look out for unintended pharmacological and toxic effects [8]. Also the approval to market a drug requires the obligation by sponsor to conduct additional studies among which is the cohort studies to examine the benefits and potential ADR of the new drug in different populations.

Clinical testing of drug investigation is in three phases. In phase I, clinical trials are done to establish Safety. The new drug is tested on close medical surveillance in a small number of healthy volunteers to determine the safety and tolerability and provide information on how new drug behaves in the human body. In phase II, clinical trials are carried out to establish Efficacy. Here, the new drug is tested for the first time in patients to determine its effectiveness. This trial generates information on the optimum dose, dose schedule and route of administration. Phase III is a clinical trials to establish clinical benefits. The new drug is tested identifying dose in large number of patients, usually at clinical trial centers around the world. The aim is to demonstrate that the new drug is more effective and the best currently available treatment. For maximum objectivity, this trial is usually double blinded, which means neither the patient nor the physician knows which patient is receiving the new drug or which patient is receiving the standard care. The last phase which is phase IV is a post marketing studies and surveillance. After the drug has been released to the market, it is still closely monitored; reports of adverse effect are recorded in central registries. Ability to know the potential harm (ADR) that a drug could cause is critical to human health. The aim of drug development and post marketing evaluation is to give information that enables medical personnel and patients to make an educated decision about potential benefits of drug as well as harms such drug could pose.

This traditional method takes years and does not predict the ADR on time. The argument in this paper is that computational method of drug prediction should be explored to save enormous time spent waiting to accumulate ADR presence and also the amount spent on drug ADR detection could be reduced.

Computational Methods of ADR Prediction

Presently there are various computational approaches proposed for use in ADR prediction. The methods are pathway-based approaches, chemical structure approaches, and Machine learning approaches.

Pathway-based Methods

Pathway-based method associates drug ADR to distressed biological pathways since the pathways deals with proteins aimed by the drug. The work done by Campillo [9] argues that drugs with the same ADR are likely to share the same profile of protein target. This study used this feature to predict drug target for established drugs using ADR similarities. Another study by Fukuzaki et al [10] suggested a method for associating ADR to common pathways sharing correlated changes of gene expression profiles given the drug of interest. The method nevertheless needs gene expression data experimented under chemical perturbation of drug. Further study by Xie et al [11] established a method to determine off targets for drugs through drug docking into proteins binding pocket that are related to the main target. Here drug to protein interactions that has maximum docking scores are merged to identify

biological pathways, thus permitting the detection of probable off target binding networks for the drug. This methods performance determined by the accessibity of protein 3D structures and established biological pathways, thus limiting the wide application.

Chemical Structure-based Methods

This principle associate drug ADR to chemical structure make up. Scheiber et al [11] established a method that finds chemical substructures relating to ADR. The limitation in this work is that it did not make provision for integrated framework that can predict ADR for drug molecule. Another study by Yamanishi et al [12] suggested a method that predict pharmacological and ADR information by chemical structures. This is used to suggest drug target interaction. This method unfortunately cannot predict high dimensional ADR profiles.

Machine Learning Methods

Machine learning (ML) is the science of building systems that automatically learn from data. ML is a data-driven approach to problem solving, in which the hidden patterns and trends present within data are first detected and then leveraged to help make better decisions. This process of automatically learning from data and in turn using that acquired knowledge to inform future decisions is extremely powerful. Indeed, machine learning is rapidly becoming the engine which powers the modern data-driven economy [13]. The prominent types of machine learning are discussed.

Nearest neighbour (NN)

This method of prediction requires locating a pre-defined quantity of training samples that are nearest in distance to the current point thus predicting the label from there. The sample may be constant defined by a user and sometimes it may differ in term of local density of points. Generally, the distance may be a metric measure or standard Euclidean distance. The latter is most preferred. Neighbours based method is a non-generalising machine learning methods since they have ability to recall all its training data. It is one of the most straightforward approaches to predict drug-ADR.

The k-NN algorithm can also categorize drug–drug pairs using the closest training examples in the four drug–drug pair similarity spaces. Here, a hamming distance matrix can be used to measure a nearness and implemented the standard protocol of 3-NN using three steps: (i) to calculate the distances between an unknown drug–drug pair and all drug–drug pairs in the training set; (ii) to select three drug–drug pairs that are most similar to the drug–drug pair y from the training set based on the calculated hamming distances; and (iii) to categorize drug–drug pair y into the group to which the majority of the three drug–drug pairs belong.

Support Vector Machine (SVM)

Support vector machine belong to supervised learning and is one of the recent development in machine learning. Its performance is superb in various application especially high value applications. Maximum use of SVM is limited because it is resource intensive to compute.

SVM basics have a maximum margin decision boundary that has been proved mathematically to produce robust and predictable implementation as well as sound performance of several real world applications [14]. It has thus become a well-known classification method in bioinformatics as well as chemoinformatics because of its excellent delivery in prediction [15][16]. SVM is also able to test various kernel functions of linear, Gaussian RBF using different width parameters and also polynomial using different degree parameters. Nevertheless, the implementation requires the construction of **q** individual SVM classifiers for **r** ADR implying a huge computational complexity since **q** is large in real application.

Regularized Least-Squares classifier (RLS)

The Regularized Least-Squares classifier denoted as RLS [16, 17] is a basic supervised learning algorithm. If an appropriate kernel has been chosen for RLS, the accuracy of RLS will be similar to support vector machine (SVM), whereas the computation complexity of RLS is much less than SVM. The RLS algorithm can be divided into three separate sub algorithms for defining the kernel matrix: RLS-KP, RLS-KS and RLS-avg. Here, KP and KS are short for Kronecker Product [18,19] and Kronecker Sum [19], respectively. RLS machine is good for solving classification issues by using a linear system that is equivalent the number of features or the number of training examples. Also with RLS machine, it is possible to get the precise measure of the Loo error with training [20]. RLS classifier is a good substitute for SVM classifier because it is simple and requires a low computational burden. It again illustrates ability to generalize when compared to SVM classifier.

Semi-supervised Link Prediction classifier (SLP)

The link prediction problem is usually described as a task to predict how likely a link exists between an arbitrary *pair* of nodes. Semi-supervised Link Prediction is a new semi-supervised learning method for link prediction problems, where the task is to predict unknown parts of the network structure by using auxiliary information such as node similarities. This method can fill in missing parts of tensors and is applicable to multi-relational domains, allowing us to handle multiple types of links simultaneously.

Semi-supervised Link Prediction classifier denoted as SLP is a semi-supervised learning algorithm [21,22], and the basic assumption of SLP is ‘‘Two node pairs that are similar to each other are likely to have the same link strength’’ [22]. SLP also can be divided into three independent sub algorithms for defining Laplacian matrix L: SLP-KP, SLP-KS and SLP-avg. The overall handling of SLP-avg is similar to RLS-avg.

Naïve Bayes (NB)

NB is another machine learning method that has in existence over 50 years ago and applied to biomedical informatics. It is a widely used method of classification and prediction in different domains including bioinformatics. The underlying probability model of the Naïve Bayes classification method is best described as a model with statistically independent characteristics. A naïve Bayesian classifier supposes that the existence of a characteristic for a given class is independent from the existence of the other characteristics for that class. A computationally efficient method that also performs classification perfectively well even compared to complex methods. Nevertheless NB, when dealing with large number of features, it is mis-calibrated and predicts with subsequent likelihoods that are too close to 0 and 1[23]. Some of the merits of this method are the ability to rapidly construct from data, ability to quickly make inferences, compatibility in term of space and good performance in practice. Although the condition in which NB is based is rarely true in real lrfrfife yet its competitive performance in classification is great.

Comparison of Computational Methods

Methods	Concept	Merits	Demerits
Pathway based method	Campillo's method associates drug ADR to distressed biological pathways since the pathways deals with proteins aimed by the drug. Drugs with the same ADR are likely to share the same profile of protein target. Fukuzaki et al's method associates ADR to common pathways sharing correlated changes of gene expression profiles given the drug of interest.		The method needs gene expression data experimented under chemical perturbation of drug.
Chemical based method	This method associate drug ADR to chemical structure make up and finds chemical substructures relating to ADR		The method cannot predict high dimensional ADR profiles
Nearest Neighbour	It predicts a given drug **d** to have the similar side-effects as those of the drug (in a training set) whose chemical substructure profile is the	The training is fast, It learns complex, non-linear functions easily and also keeps all information	The method is quite slow at prediction; it needs lots of storage space and is easily fooled by irrelevant attributes.

	most similar. In the case of individual query, the **k** nearest neighbours is sourced out for, and if *k'* of *k* has a side-effect, the prediction score of k'/k are assigned to the query drug. The procedure is repeated for *q* side-effects.		
Support Vector Machine	The implementation requires the construction of **q** individual SVM classifiers for **r** ADR implying a huge computational complexity since **q** is large in real application.	It has become a well-known classification method in bioinformatics as well as chemoinformatics because of its excellent delivery in prediction. The quality of generalization and ease of training of SVM is far beyond the capacities of the traditional methods. SVM performs well on data sets that have many attributes, even if there are very few cases on which to train the model. It performs well where large amount of unlabelled data and a small amount of labelled data is available.	Maximum use of SVM is limited because it is resource intensive to compute. The main disadvantage of the SVM algorithm is that it has several key parameters that need to be set correctly to achieve the best classification results for any given problem.
Regularized Least Square Method (RLS)	The RLS is a very basic classification method which is used for regression. When a good kernel is used with least squares, it has classification accuracy similar to that of SVM. RLS method is good for solving classification issues by using a linear system that is equivalent to the number of features or the number of training examples.	RLS is conceptually simple and yields high accuracy. It is a good substitute for SVM classifier because it is simple and requires a low computational burden. It again illustrates ability to generalize when compared to SVM classifier. Training RLSC requires only the solution of a single system of linear equations, which is conceptually much simpler unlike SVM that requires solving convex quadratic program	For large problems, SVM can solve problem in less time than it takes to compute every entry in K for RLS. Furthermore if $O(l^2)$ memory is not available to store K, then the RLSC problem cannot be solved at all via direct methods (It can only be solved using conjugate gradient methods, re-computing the kernel matrix K at every iteration, but this will be intractably slow for large problems.
Semi-Supervised Prediction Method	Two node pairs that are similar to each other are likely to have the same	It performs well where large amount of unlabelled data and a	The method if not carefully designed can results in over-fitting

	link strength. SLP therefore predicts unknown ADR by using auxiliary information such as drug similarities. This method improves the effectiveness of learning method by using unlabeled data with the same feature.	small amount of labelled data is available just like SVM.	
Naïve Bayes	The underlying probability model of the Naïve Bayes classification method is best described as a model with statistically independent characteristics. A naïve Bayesian classifier supposes that the existence of a characteristic for a given class is independent from the existence of the other characteristics for that class.	It has the ability to rapidly construct from data, ability to quickly make inferences, compatibility in term of space and good performance in practice. Its competitive performance in classification is great. The method is super simple	NB, when dealing with large number of features, is mis-calibrated and predicts with subsequent likelihoods that are too close to 0 and 1

Benefits of Computational Methods

Some of the most prominent advantages of using computational methods are as follows;

Accurate: Computational methods such as ML use data to discover the optimal decision-making engine for problem. As more data is collected, the accuracy can increase automatically. Thus accurately predicting the expected ADR associated with a particular drug given the right data. The accurate predictions help in drug development, reducing the rate of drug failure and within a limited time period.

Automated: As new data come in, the output interest can be automatically estimated using computational method.

Fast: Computational method can generate answers in a matter of milliseconds as new data stream in, allowing systems to react in real time.

Customizable: A large number of data-driven problems can be addressed with computational methods. Example, ML models are custom built from individual data, and can be configured to optimize whatever metric drives individual task.

Scalable: As business grows, computational method such as ML easily scales to handle increased data rates. Most ML processes are parallelizable, enabling infinite scalability [24].

Conclusion

Drug safety is of great importance to public health. The detrimental effects of drugs do not only limit their application but also cause suffering in individual patients and evoke distrust of pharmacotherapy. Also early and accurate identification of adverse drug effect is critical to public health. Since the post hoc analysis is usually unable to detect rare or delayed on-set ADR until clinical evidence accumulates. The process takes years. Also the rate of disease resistance to drugs is at the increase and therapies are needed early on time. This paper has been able to look at some of the most important computational methods that could be employed in drug-ADR prediction and expose researchers to the need to explore the computational approach to drug ADR prediction.

References

[1] Seth I. Berger and Ravi Iyengar, Role of systems pharmacology in understanding drug adverse events. Systems Biology and Medicine 2011.

[2] Atonetti N, Liu T, Altman R: Predicting drug side-effects by chemical systems biology. **G***enome Biol* 2009, **10:**238

[3] Scheiber J, Chen B, Milik M, Sukuru S, Bender A, Mikhailov D, Whitebread S, Hamon J, Azzaoui K, Urban L, Glick M, Davies J, Jenkins J: Gaining insight into off-target mediated effects of drug candidates with a comprehensive systems chemical biology analysis. Chem Inf. Model 2009, 49(2): 308-317

[4] Whitebread S, Hamon J, Bojanic D, Urban L: Keynote review: in vitro safety pharmacology profiling: an essential tool for successful drug development. Drug DiscoVery Today 2005, 10(21): 1421-1433

[5] Jane P.F. Bai and Darrell R. Abernethy. Systems Pharmacology to Predict Drug Toxicity: Integration Across Levels of Biological Organization. Annu. Rev. Pharmacol. Toxicol. 2013. 53:451–73.

[6] Peter L. Anderson, Pharm.D. The ABCs of Pharmacokinetics. Winter 2005. Available: http://www.thebody.com/content/art875.html. Accessed 3rd August, 2014.

[7] Armijo JA. 2003. *Farmacocinética: Absorción, Distribución y Eliminación de los Fármacos*. En: Flórez J, Armijo JA, Mediavilla A, *Farmacología Humana*, 4ta edición. Masson. Barcelona. pp: 51-79.

[8] Jesse A. B., Susan C. G. and Susan S. E. (2008), Adverse Event Detection in Drug Development: Recommendations and obligations. Am J Public Health.

[9] Campillos M, Kuhn M, Gavin A, Jensen L, Bork P: Drug target identification using side-effect similarity. *Science* 2008, **321**(5886)**:**263-6.

[10] Fukuzaki M, Seki M, Kashima H, Sese J: Side Effect Prediction using Cooperative Pathways. *IEEE International Conference on Bioinformatics and Biomedicine 2009 IEEE BIBM 2009)* 2009, 142-147.

[11] Xie L, Li J, Xie L, Bourne P: Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors**.** *PLoS Comput Biol* 2009, 5**:**e1000387.

[12] Yamanishi Y, Kotera M, Kanehisa M, Goto S: Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* 2010, **26:**i246-i254.

[13] Henrik B. and Joseph W. R. (2014) Real-World Machine Learning. Manning Publications.

[14] Furey T, Cristianini N, Duffy N, Bednarski D, Schummer M, Haussler D: Support vector machine classification and validation of cancer tissue samples using microarray expression data. Toxicology 2000, 16(10): 906-914

[15] Kramer S, Frank E, Helma C**:** Fragment generation and support vector machines for inducing SARs**.** *SAR QSAR Environ Res* 2002, **13**(5)**:**509-523.

[16] Schölkopf B, Tsuda K, Vert J: *Kernel Methods in Computational Biology*. MIT Press; 2004.

[17] Rifkin R, Yeo G, Poggio T (2003) Regularized least-squares classification. Nato Science Series Sub Series III Computer and Systems Sciences 190: 131–154.

[18] Wang YC, Deng N, Chen S, Wang Y (2013) Computational Study of Drugs by Integrating Omics Data with Kernel Methods. Molecular Informatics 32: 930–941.

[19] Laub AJ (2005) Matrix analysis for scientists and engineers: Siam

[20] Ancona, N, Maglietta R., D'Addabbo A., Liuni S. and Pesole G. (2005) Regularized Least Square Cancer from DNA microarray data. BMC Bioinformatics 1:6

[21] Xia Z., Wu L. Y., Zhou X., and Wong S. (2010) Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. BMC Systems Biology 4.

[22] Raymond R., and Kashima H. (2010) Fast and Scalable Algorithm for Semi-supervised Link Prediction on static and Dynamic Graphs. In: Balcazar J.L., Bonchi K., Gionis A., Sebag M. editors. Machine Learning and Knowledge Discovery in Databases. Pp.131-147.

[23] Wei W., Shyam V. and Gregory F. C. (2011) The application of naïve Bayes model averaging to predict Alzheomer's disease from genome-wide data. J Am Med Inform Assoc. 18(4) : 370 -375

[24] Henrik B. and Joseph W. R. (2014) Real-World Machine Learning. Manning Publications.