Prediction of biological activity using heterogeneous information sources

PhD thesis

Ádám Arany

Semmelweis University

Doctoral School of Pharmaceutical Sciences

Supervisor: Dr. Péter Mátyus, DSc

Official reviewers: Dr. Gábor Horváth, Ph.D
Dr. Tóthfalusi László, Ph.D

Head of the Final Examination Commitee: Dr. Imre Klebovich, DSc

Members of the Final Examination Commitee: Dr. László Őrfi, Ph.D
Dr. Tamás Paál, CSc

Budapest

2016

1 Introduction

As the productivity of the pharmaceutical research and development is lagging behind the sharply increasing costs, the pharmaceutical industry is continuously searching for new approaches in drug discovery. These problems are aggravated also by the price pressure caused by expiring patents, and the ever complicated regulatory procedures. In my doctoral research I developed and applied computational methods related to two topics, which revolutionized the pharmaceutical industry to ameliorate the effect of the dropping effectiveness of the research and development pipeline: drug repositioning and personalized medicine.

Drug repositioning or repurposing is a cost-effective and risk-reducing straightforward strategy, which aims at reusing already approved drugs in new therapeutic indications. From the machine learning perspective the main distinctive feature of drug repositioning compared to de novo drug discovery is the availability of a wide range of information sources. While conducting the research my primary goal was to develop computational methods to

harness these information sources in drug repositioning. As a first step I created a benchmark dataset containing six different information sources (three chemical structure descriptors, two side effect based descriptors and a target profile), and a drug-indication gold standard set. The goal of my first computational experiment was to compare a novel data fusion methodology, called Kernel Fusion Repositioning (KFR), with a baseline method. My contribution primarily concerned the design and implementation of the KFR framework as well as the application of the KFR framework on the problem of repositioning for Parkinson's disease. As one of the authors of a novel multi-target prediction method I also applied this method to the repositioning benchmark, and analysed the effect of multi-target learning on accuracy.

My second topic was related to personalized medication, which facilitates the optimal therapy for the patient and is also favourable for the researcher interested in drug development. Predicting the patient-by-patient variability of the pharmacokinetics can help the investigator adjust the doses in a personalized way in order to maximize

efficacy and minimize side effects and toxicity. I participated in researching the interpersonal variability of methotrexate pharmacokinetics at high dose levels, developed new clinical descriptors bridging patient and treatment levels, and investigated their usage by applying a novel Bayesian multivariate statistical technique to identify predictive genetic variants. Moreover, I compared the results against already existing ones based on frequentist statistics.

2 Objectives

The objectives of my doctoral thesis are:

To develop a novel data fusion method for the prediction of the biological effects of small-molecular drugs by integrating heterogeneous information sources.

- To apply the data fusion method for finding Parkinson's disease related drugs, and to evaluate the ability of this method to enhance drug discovery, especially drug repositioning.
- To develop and evaluate a novel matrix factorization based method capable of predicting multiple activities simultaneously, and to compare it with a single target baseline method.
- To adapt and apply a novel Bayesian multivariate statistical technique to identify genetic variants predictive of the interpersonal variability of methotrexate pharmacokinetics at high dose levels.

3 Methods

At the start of the research information sources describing compounds were constructed: Molecular Access Keys (MACCS); molecular connectivity, shape and electrotopological fingerprint (MOLCONN-Z); 3D pharmacophore based fingerprint; side effect occurrences and frequencies; and known drug-target interactions. We defined the vector representation of the compounds for each information source. Also similarity metrics was identified to compute pairwise similarity kernels from the features for the methods requiring similarities. The Tanimoto similarity was used for every information source with binary features, whereas the cosine similarity was applied for sources based on real valued features.

To assess the common information content of these data sources, that is to evaluate their complementarity, we computed the Spearman correlations of all pairwise similarities.

I participated in the development of the novel Kernel Fusion Repositioning (KFR), which method uses the one-class SVM framework and serves as a reference model

class in the comparison of other data fusion methods. In the late fusion setup we computed prioritization based on different data sources separately, and fused the ranking with the Borda protocol. AUC[ROC], AUC[CROC(exp)], BEDROC and fixed threshold *sensitivity* and *specificity* measures were used to evaluate predictive performance. The early discovery focus was α=20.0. Two thresholds were introduced for both the sensitivity and the specificity: top25 and top100. ATC, a widely accepted classification system was utilized to compare the predictive performance of the different ranking methods.

As the research group at the Department of Organic Chemistry has interest in Parkinson's disease therapies, we applied the data fusion based methodology to prioritize repositioning candidates for Parkinson's disease (PD). Nevertheless, the developed methodology can be applied to a wide range of repositioning projects in general.

I also participated in the development of the Bayesian matrix factorization method Macau for the drug-indication prioritization task. An important aim of the present work is to make the method applicable for settings without

negative samples. The probability that a missing association does not hold is much higher than the probability that it exists but it has not been verified yet. In the research a well-established strategy was selected to randomly choose a subset of the missing associations identified as the negative set. To validate that the multi-task effect between ATC level 4 classes can improve our results we used a column-wise ridge regression (a form of regularized OLS regression) as a benchmark.

The second major topic of my thesis is pertaining to personalized medicine. As a member of the research group I analysed the effect of 29 preselected single nucleotide polymorphisms (SNP) from the genes ABCB1, ABCC1, ABCC2, ABCC3, ABCC10, ABCG2, GGH, SLC19A1, NR1I2. In gene selection we significantly relied on the literature and on relevant scientific findings as well. When estimating functionality we relied on the classification of the polymorphism and its localization. The isolation of the genetic material from blood was carried out by using Qiagen isolation kits (QIAmp DNA Blood Maxi Kit / QIAmp DNA Blood Midi Kit; Qiagen, Hilden, Germany).

For sequencing GenomeLab SNPstream genotyping platform (Beckman Coulter) was used.

As the clinical data contain 551 methotrexate blocks for 59 patients with a variable number of blocks per patient, the aggregation of the block level variables was necessary. I recommended using the median value over blocks for every variable as a patient level value, a convention used both in the frequentist and in the Bayesian analysis. All pharmacokinetics and toxicity measurements were recorded at block level; the median was therefore computed. To apply the BN-BMLA method the discretization of the continuous variables is necessary. I suggested the discretization based on median values to get a balanced dataset with an equal number of samples for different values. In case of multinomial variables we applied binning to binary variables to reach lower model complexity. The original hepatotoxicity and myelotoxicity variable was multinomial with four possible values. Based on the balanced dataset criterion the team binned the myelotoxicity as grade 1 vs. grade 2-4, and the hepatotoxicity as grade 1-2 vs. grade 3-4.

4 Results

The aim of the research we conducted into computational drug repositioning was to compare the predictive performance of the newly developed Kernel Fusion Repositioning (KFR) method as an *intermediate* fusion method and a standard *late fusion* method, the Borda protocol based fusion via using one-class support vector machines as the model class. The Level 4 ATC classes were used as prediction tasks. We have found that in all cases, primarily underpinned by the early discovery measures, the intermediate data fusion has better predictive performance.

To explore the advantages of the developed Kernel Fusion Repositioning (KFR) method, we calculated the Spearman correlation of the ordering based on the given single data source and the output orderings of the two fusion methods to compare their behaviour. A notable feature of these results, also a key result of my work, is that the relative contributions of the different data sources are quite stable across the different drug categories in case of the Borda method, while the kernel

fusion based method shows adaptive, query-specific properties.

We suggested a criterion on query compactness to define an acceptable training set for prioritization. The proposed solution relies on the use of the intraset similarity (ISS), the average of all pairwise similarities of the elements in the training set. We normalized it with the average of all similarities in the full set of drugs. The measure ISS/UAS shows a good correlation with AUC values: all classes which have higher than one ISS/UAS value, have at least 0.5 AUC.

We analysed the result given to four Parkinson's disease (PD) related queries composed of neuroprotective agents, dopaminergic agents, muscarinic agents and NMDA antagonists by the KFR system. We applied a filter based on the 2013 and 2016 PubMed abstract database, to remove compounds from the resulted ordering without co-occurrences with PD using the following PubMed search query: *(„Parkinson" OR „Parkinson's Disease" OR „PD") AND INN.* From a prospective point of view, which is the most reliable evaluation, it is interesting to note that

the co-occurrence number for some of the highly prioritized compounds increased significantly, while the relative increase was less significant for others or there was no change at all. These three groups show good correspondence with the following groups: possible repositioning candidates, already known drugs and false positives.

The performance of the newly developed matrix factorization method Macau on the ATC class prediction task was also evaluated to show that in case of all information sources Macau has a considerably higher predictive performance than ridge regression.

In the personalized medicine related work I adapted BN-BMLA, a novel multivariate Bayesian technique to identify genetic polymorphisms predictive of the inter-individual differences of the pharmacokinetics and toxicity of methotrexate.

As all pharmacokinetics and toxicity measurements were recorded for every treatment blocks, we suggested a median aggregation for these variables at patient level. This convention was used both in the frequentist and the

Bayesian analysis. To apply the BN-BMLA method the discretization of the continuous variables is necessary. The discretization based on median values was suggested to get a balanced dataset with an equal number of samples for different values. In case of the NR1I2 gene BN-BMLA identified a single polymorphism from a set of linked polymorphisms, which shows that BN-BMLA methodology can distinguish between direct relations and transitive relations. While some SNPs are identified by both frameworks (two SNPs in case of AUC, and one in case of myelotoxicity), there are weaker candidates suggested by both the frequentist and the Bayesian methodology. Our results suggest that using the consensus of different methods for robust detection of association is an appropriate pragmatic approach to be followed.

5 Conclusions

The results of my research allows for making the following conclusions and statements:

- While conducting my research I significantly contributed to and participated in the development process of a novel *intermediate* data fusion method, the Kernel Fusion Repositioning (KFR) framework. Our research evaluations showed that KFR has a superior performance compared to the *late fusion* baseline Borda protocol as justified by the AUC measure, and especially verified by all applied early discovery measures in terms of a drug repositioning benchmark problem.

- In order to examine the behaviour of the methods I analysed the Spearman's rank correlation of the single data source based prioritization results with the data fusion based prioritization results and we found that KFR shows adaptive, query-driven properties. This property is an important advantage, which makes the method applicable in

a wide range of pharmacological groups and different chemical spaces.

- The experiments showed an anomalous behaviour in case of extremely high query heterogeneity and we witnessed that the query compounds are not ranked high in the resulted ordering. In this case the predictive power of the method can be really poor. We suggested a criterion measuring the average pairwise similarity of the query compounds to filter these cases, and showed that this criterion can identify the queries resulting in poor predictive performance.

- The KFR framework was applied to identify potential repositioning candidates in Parkinson's disease therapy and compounds showing high co-occurrence with those in the literature were retrieved. All results were validated further in a prospective evaluation. Also, steps of a novel computational route for drug repositioning candidate identification were outlined.

- I participated in the development process of Macau, a novel Bayesian matrix factorization method capable of predicting multiple targets simultaneously. While conducting the research I compared Macau to a single target method (Ridge regression) and found it superior in the case of all information sources.

- My research justifies successful adaptation and application of BN-BMLA, a novel multivariate Bayesian method, in complementing and confirming the already existing frequentist results in a study conducted into the pharmacokinetics of high dose methotrexate therapy. The results suggest that the effective combination of the Bayesian and frequentist methods in the field of the robust detection of association is an appropriate strategy, whereas the BN-BMLA method is more beneficial in the case of investigating interactions or redundancies.

List of own publications

Hegyi M, Arany A, Semsei AF, Csordas K, Eipel O, Gezsi A, Kutszegi N, Csoka M, Muller J, Erdelyi DJ, Antal P, Szalai C, and Kovacs GT, *Pharmacogenetic analysis of high-dose methotrexate treatment in children with osteosarcoma.* Oncotarget, 2016. **IF = 5.008***

Temesi G, Bolgar B, Arany A, Szalai C, Antal P, and Matyus P, *Early repositioning through compound set enrichment analysis: a knowledge-recycling strategy.* Future Med Chem, 2014. **6**(5): p. 563-75. IF = **3.744**

Arany A, Bolgar B, Balogh B, Antal P, and Matyus P, *Multi-aspect candidates for repositioning: data fusion methods using heterogeneous information sources.* Curr Med Chem, 2013. **20**(1): p. 95-107. **IF = 3.715**

Bolgar B, Arany A, Temesi G, Balogh B, Antal P, and Matyus P, *Drug repositioning for treatment of movement disorders: from serendipity to rational discovery strategies.* Curr Top Med Chem, 2013. **13**(18): p. 2337-63. **IF = 3.453**

Cumulative impact factor: 15.920