Automatic discovery of transferable patterns in protein-ligand interaction networks

Aida Mrzic [1, 2], **Dries Van Rompaey** [2, 3], **Stefan Naulaerts** [1, 2, 4], **Hans De Winter** [3], **Wim Vanden Berghe** [5], **Pieter Meysman** [1, 2], **Kris Laukens** [Corresp. 1, 2]

[1] Adrem Data Lab, University of Antwerp, Antwerp, Belgium

[2] Biomedical Informatics Network Antwerp (biomina), University of Antwerp, Antwerp, Belgium

[3] Laboratory of Medicinal Chemistry, University of Antwerp, Wilrijk, Belgium

[4] Computational Biology and Drug Design (CBDD), CRCM (INSERM U1068), F-13009 Marseille, France; Institut Paoli-Calmettes, F-13009 Marseille, France; AMU, F-13284 Marseille, France; CNRS (UMR7258), F-13009 Marseille, France, Marseille, France

[5] Laboratory of Protein Chemistry, Proteomics and Epigenetic Signaling (PPES), University of Antwerp, Wilrijk, Belgium

Corresponding Author: Kris Laukens
Email address: kris.laukens@uantwerpen.be

In recent years, the pharmaceutical industry has been confronted with rising R&D costs paired with decreasing productivity. Attrition rates for new molecules are tremendous, with a substantial number of molecules failing in an advanced stage of development. Repositioning previously approved drugs for new indications can mitigate these issues by reducing both risk and cost of development. Computational methods have been developed to allow for the prediction of drug-target interactions, but it remains difficult to branch out into new areas of application where information is scarce.

Here, we present a proof-of-concept for discovering patterns in protein-ligand data using frequent itemset mining. Two key advantages of our method are the transferability of our patterns to different application domains and the facile interpretation of our recommendations. Starting from a set of known protein-ligand relationships, we identify patterns of molecular substructures and protein domains that lie at the basis of these interactions. We show that these same patterns also underpin metabolic pathways in humans. We further demonstrate how association rules mined from human protein-ligand interaction patterns can be used to predict antibiotics susceptible to bacterial resistance mechanisms.

Automatic discovery of transferable patterns in protein-ligand interaction networks.

Aida Mrzic[A,B*], Dries Van Rompaey[B,C*], Stefan Naulaerts[A,B,D], Hans De Winter[C], Wim Vanden Berghe[E], Pieter Meysman[A,B], and Kris Laukens[A,B,‡]

[*] ***Authors contributed equally***

[‡] ***Corresponding author: kris.laukens@uantwerpen.be***

[A] ***Adrem Data Lab, Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium***

[B] ***Biomedical Informatics Network Antwerp (biomina), University of Antwerp, Antwerp, Belgium***

[C] ***Laboratory of Medicinal Chemistry, University of Antwerp, Wilrijk, Belgium***

[D] ***Cancer Research Center of Marseille, INSERM U1068, F-13009 Marseille, France; Institut Paoli-Calmettes, F-13009 Marseille, France; Aix-Marseille Université, F-13284 Marseille, France; and CNRS UMR7258, F-13009 Marseille, France***

[E] ***Laboratory of Protein Chemistry, Proteomics and Epigenetic Signaling (PPES), University of Antwerp, Wilrijk, Belgium***

ABSTRACT

In recent years, the pharmaceutical industry has been confronted with rising R&D costs paired with decreasing productivity. Attrition rates for new molecules are tremendous, with a substantial number of molecules failing in an advanced stage of development. Repositioning previously approved drugs for new indications can mitigate these issues by reducing both risk and cost of development. Computational methods have been developed to allow for the prediction of drug-target interactions, but it remains difficult to branch out into new areas of application where information is scarce.

Here, we present a proof-of-concept for discovering patterns in protein-ligand data using frequent itemset mining. Two key advantages of our method are the transferability of our patterns to different application domains and the facile interpretation of our recommendations. Starting from a set of known protein-ligand relationships, we identify patterns of molecular substructures and protein domains that lie at the basis of these interactions. We show that these same patterns also underpin metabolic pathways in humans. We further demonstrate how association rules mined from human protein-ligand interaction patterns can be used to predict antibiotics susceptible to bacterial resistance mechanisms.

1 INTRODUCTION

The pharmaceutical industry has been confronted with a decline in R&D productivity. Indeed, the industry has been said to face a productivity crisis. [1] The drug development process is an expensive and time-consuming endeavor, with estimated costs for new drugs reaching up to 2.6 billion USD and a time-to-approval ranging from 10 to 17 years. [2] Drug development programs have tremendous attrition rates, with only a select few candidates making it to the market. An attractive alternative to this laborious process is identifying new applications for drugs that are already on the market, an approach known as *drug repositioning* or *drug repurposing*. Drug repositioning lowers the risk, time and cost involved with developing new drugs, as their toxicity, clinical safety and pharmacokinetics

have already been established. Preclinical toxicity for instance remains an important driver of the attrition of drug candidates. [3] The accurate identification of drug-target interactions (DTI) is thus of tremendous value. The applications of these techniques are not limited to drug repurposing, as they can also be used to identify small molecules for which no interacting proteins have been described to open up new avenues for drug discovery. [2, 4]

Interactions between drugs and their targets may be identified experimentally through various screening methods. However, screening every possible combination of known drugs and targets is prohibitively expensive. The low cost and high throughput of computational screening approaches renders them an interesting alternative. Following the classification described by Ezzat et al., computational approaches towards this problem can broadly be categorized into three classes. [5] The first class consists of ligand-based approaches, which is based on the concept that similar drugs tend to have similar targets. The second class is docking, where the three-dimensional structures of the ligand and the target protein are used to predict a possible binding mode and assign an energy score. A major drawback of docking is its reliance on the three-dimensional structure, which is not available for the majority of proteins. The third class, chemogenomic approaches, combines protein and drug data to discover novel DTIs. This type of approach can be further divided into two broad categories: feature-based methods and similarity-based methods [5–7].

Feature-based methods derive feature vectors for both drugs and targets. An example of these features might be hydrophobicity or amino acid composition for proteins, and molecular fingerprints or geometric descriptors for drugs. These features vectors are used to train machine learning models, which may then be used to identify novel DTIs. Similarity-based methods rely on similarities between drugs and targets to predict novel DTIs. These may further be divided into four separate categories [5]: (i) neighborhood methods that predicts novel interactions for drug (protein) based on a nearest neighbor; (ii) bipartite local methods that predict interactions for drugs and proteins separately, and then combine results for the final prediction; (iii) network diffusion methods which use graph-based techniques for DTI prediction; and finally (iv) matrix factorization methods that learn feature matrices from the DTI matrix and use these for novel DTI predictions.

While a great deal of progress has been made in the prediction of interactions between drugs and their targets, it remains difficult to predict interactions for new application areas, where data may not be so readily available. New methods which capture the interactions between proteins and ligands in a general manner may therefore be invaluable. In this work, we present a method for discovering patterns underlying interactions between proteins and ligands through frequent itemset mining. Frequent itemset mining was first conceptualized to investigate customer behavior in grocery shopping. [8] Transactions of customers could be analyzed to identify frequently co-occuring purchases, for instance the combination of milk, bread and butter. Such associations can be mined to identify a rule, for instance when a customer purchases milk and bread, he will also purchase butter. These rules could then be used to guide marketing decision making.

In recent years, frequent itemset mining has also been applied to a number of problems in bioinformatics, such as the identification of metabolites from mass spectral data. [9, 10] In this work, we use frequent itemset mining to identify patterns governing the interaction of ligands with their target proteins. Two key advantages of our method are the transferability of our patterns to different application areas and the facile interpretation of our recommendations. More complex machine learning techniques such as deep learning or random forest approaches are often more powerful, but this comes at the expense of interpretability. These approaches tend to be black boxes, where it is difficult to gain insight into the inner workings of the predictions. In contrast, as frequent itemset mining produces an explicit list of patterns and recommendation rules, the interpretation is straightforward. Furthermore, frequent itemset mining may be used as part of a pipeline to select features for use in more advanced machine learning models.

Starting from known protein-ligand relationships, we uncover patterns consisting of molecular substructures and protein domains that underlie these relationships. We demonstrate how these

patterns can be used to explain metabolic pathway data and we further show how this approach can be used to predict antibiotic resistance.

2 MATERIALS AND METHODS

2.1 Problem description

Our goal is to obtain a set of patterns from the transactional dataset containing molecular fingerprint keys for the ligands and domains for the proteins. To this end, we will use frequent itemset mining to discover which chemical structure elements and domains frequently co-occur. The method is illustrated in figure 1.

2.2 Frequent itemset mining

Frequent itemset mining discovers frequently co-occurring items in a transactional data set. In this type of data set, each transaction represents a set of items (i.e. *itemset*). Here, we created a transactional data set starting from known protein-ligand interactions. As ligands are represented by their substructures and targets by their protein domains, each item is either a chemical substructure or a protein domain. A transaction consists of all chemical substructures and protein domains describing a single protein-ligand interaction. We define the support of an itemset as the number of appearances in the data set, where itemset is frequent if its support is higher than a predefined threshold. Here, we mined for frequent itemsets of the following form.

$$\{molecular\ fingerprint, protein\ domain\}$$

Having obtained these frequent patterns, we can then mine these for association rules. An association rule is an implication in the form $x \Rightarrow y$. The left hand side, *body*, or *antecedent* is an item x present in the dataset and the right hand side, *head*, or *consequent* is an item y which is frequently associated with x. The support of an association rule $x \Rightarrow y$ is equal to the support of items in its body and head, i.e. $x \cup y$. Given that many rules are produced in this step and the most frequent rules are not necessarily the most interesting ones, we can further prune them using additional interestingness measures, confidence and lift. The *confidence* in a given rule is the frequency with which the rule was found to be correct. The *lift* for a given rule is defined as the frequencies for both items occurring together divided by the frequency by which either item occurs.

To mine the association rules we used the R package *arules* [11]. The mining algorithm of choice was *apriori* [12]. It searches for frequent itemsets in breadth-first manner: it identifies all frequent itemsets of size k, then uses them to create all candidate itemsets of size $k + 1$. Once all frequent itemsets have been found, association rules are created. The support, confidence and lift thresholds used herein were 0.1%, 10% and 1, respectively. We mined for association rules in the following form:

$$protein\ domain\ d \Rightarrow molecular\ fingerprint\ fp$$

2.3 Data

Protein-ligand information was downloaded from STITCH (Search Tool for Interacting Chemicals), a database of known and predicted interactions between chemicals and proteins [13]. The current incarnation, STITCH 5, covers 1.6 billion interactions between almost 10 million proteins across 2000 organisms and half a million chemicals. All non-human chemical-protein interactions were filtered out, as well as protein-protein interactions where present. This resulted in a simple protein-ligand network for *Homo sapiens*, containing 14,987,535 interactions between 19,182 proteins and 781,250 ligands. The molecular structure of the ligands were obtained from STITCH 5 under the form of SMILES strings. These were used to calculate a substructure-key based fingerprint for each molecule, a vector where each bit encodes the presence of a certain structural property of the molecule. We elected to use the MACCS fingerprint, because of its small length of 166 bits, which

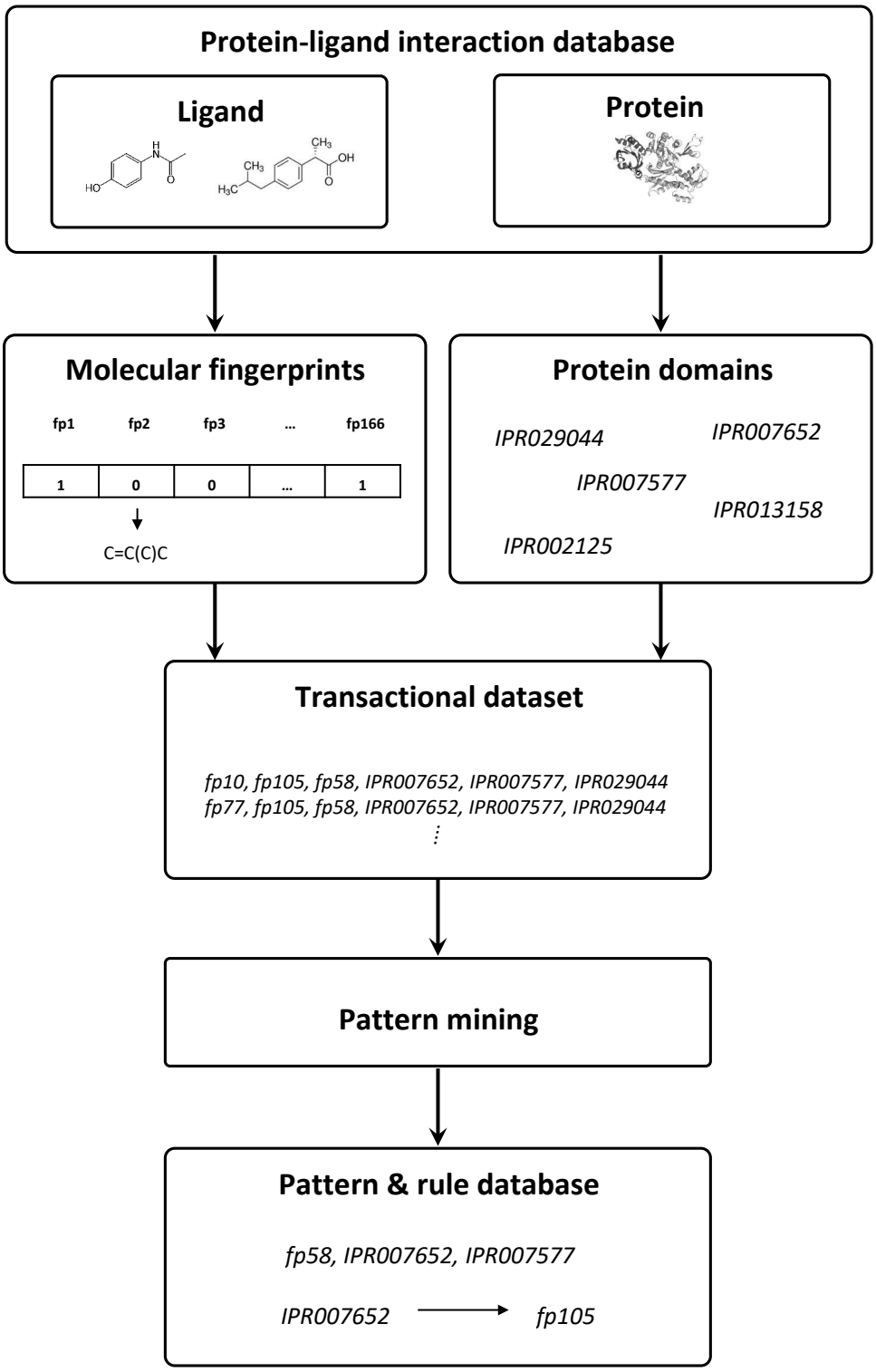

Figure 1. Starting from protein-ligand data, a transactional dataset was created consisting of fingerprint keys of the ligands and the domains of the proteins. We mined for frequent itemsets, retaining only those itemsets with at least one molecular fingerprint key and one domain. These frequent patterns were then minded for association rules of the form: *protein domain d is associated with molecular fingerprint key fp.*

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27002v1 | CC BY 4.0 Open Access | rec: 22 Jun 2018, publ: 22 Jun 2018

reduces the dimensionality of our mining, and its availability across many different cheminformatics packages. [14] It should be noted that the first MACCS key is not defined in RDKit, resulting in a total of 165 possible fingerprints. Each of these MACCS keys was considered as a separate item and all 165 fingerprint keys were identified in our dataset. Fingerprinting was performed using the RDKit cheminformatics package. [15] The Interpro [16] protein domains were downloaded from UniProt [17], retaining only high-quality entries curated by SwissProt and discarding unreviewed, predicted entries. Each protein was represented by at least one protein domain, resulting in a total of 16,254 unique protein domains.

We then sought to investigate if these patterns are generalizable across different areas of application. We have therefore opted to use two diverse datasets as our validation: ConsensusPathDB [18], a general database consisting of independent small molecule-protein data, including metabolic pathways, and the Comprehensive Antibiotic Resistance Database (CARD), which contains data on antimicrobial resistance (AMR) [19], including the interactions between antibiotics and the bacterial antibiotic resistance proteins. A list of interactions between metabolites and enzymes was then downloaded from ConsensusPathDB, which contains a total of 3527 relationships. The interactions between antibiotics and antibiotic resistance proteins were then downloaded from CARD, resulting in a total of 7,444 relationships.

2.4 Protein-ligand patterns

Starting from the protein-ligand data originating from STITCH 5 as described in section 2.3, we created a transactional dataset consisting of structural information, encoded as structural features corresponding to the MACCS fingerprint, and protein information, encoded as proteins domains. After filtering out any transactions present in the ConsensusPathDB validation set [18], 17,064 transactions were retained.

These transactions were then mined for frequently co-occurring items. We mined for frequent itemsets with a minimum prevalence in the dataset of 0.001, corresponding to a *support* higher than 17, thus retaining only those patterns present in at least 17 transactions. Itemsets were furthermore required to contain at least one fingerprint and one domain. For reasons of computational tractability, we restricted the size of our itemsets to three. The following example illustrates the form of the frequent patterns. This pattern describes the co-occurrence between a sulfotransferase domain and the NS and S=O substructures.

$$\{\textit{molecular fingerprint}, \textit{protein domain}\}$$
$$\textit{f60 [S=O]}, \textit{f33 [NS]}, \textit{IPR000863 [Sulfotransferase domain]}$$

These patterns provide insight into which items frequently co-occur. In section 3.2 we compare the patterns mined from the STITCH database to the patterns governing the interactions in an independent metabolite-protein dataset.

After obtaining frequent patterns, we mined them for association rules. We retain only those rules that contain one or more protein domain(s) in the body and a molecular fingerprint in the head. This step filters uninteresting itemsets that do not contain a combination of both domain and structural information. Due to the restriction to the size of the itemset to three, we only consider rules that contain either one or two protein domains in its body and one molecular fingerprint key in its head. The following example shows a rule stating that proteins with a sulfotransferase domain will frequently interact with an SO_3 substructure.

$$\textit{protein domain } d \Rightarrow \textit{molecular fingerprint } fp$$
$$\textit{IPR000863 [Sulfotransferase domain]} \Rightarrow \textit{f39 [}SO_3\textit{]}$$

In order to select *interesting* rules, we will further filter them based on two metrics describing the performance of the rule in its original dataset - confidence and lift. Rules which meet the given criteria will be used to predict the interactions between antibiotics and antibiotic resistance proteins in section 3.3.

	Pattern present in transaction	Pattern absent in transaction
Pattern present in STITCH	$ps \cap px$	$ps \setminus px$
Pattern absent in STITCH	$px \setminus ps$	$pn \setminus (px \cup ps)$

Table 1. Contingency table for Fischer's exact test. The set of possible combinations of the MACCS keys and protein domains in transaction x is denoted as px. The set of possible combinations of MACCS keys and protein domains for the entire dataset is denoted as pn. The set of patterns derived from STITCH is denoted as ps.

3 RESULTS

3.1 Mining the STITCH database for molecular interaction patterns

Mining for frequent itemsets resulted in 5,765,302 relationships between ligand structural features represented as fingerprint keys and the proteins domains that interact with them. Subsequent association rule mining resulted in 183,222 association rules. The frequent patterns we identified contain 490 unique protein domains, while the association rules contain 487 unique protein domains.

3.2 Similar molecular patterns describe metabolic pathways

Having identified a set of patterns in a ligand-protein dataset, we then sought to investigate whether similar patterns also describe metabolic pathways in humans. Starting from the pathway-metabolite data (3,527 pathways in total), we mined all present metabolite structural fingerprint-domain patterns. We then compared the patterns we mined from the protein-ligand dataset to the patterns mined from the metabolite dataset. Fischer's exact test was then used to determine whether the patterns derived from the STITCH database correlate well with the patterns derived from ConsensusPathDB. A contingency table for our patterns is given in Table 1. The p-value of the Fischer's exact test is the probability of observing a set of values at least as extreme as these (or more extreme values) by chance alone, which can be calculated using the hypergeometric distribution. A low p-value thus indicates that these patterns are unlikely to be the result of chance and that the two categorical statements are thus likely correlated.

A p-value is calculated for each transaction x. Figure 2 shows the histogram of the p-values for this test, indicating that our method is able to identify protein domain - substructure relationships for many of the documented pathways. Figure 3 shows the ratio of patterns mined from the STITCH database to the patterns mined from the metabolite dataset. For instance, the enzyme CYP4F2 catalyzes alpha-tocopherol-omega-hydroxylation, a key step in the degradation of vitamin E. For this transaction, the ratio of metabolites and protein domains is equal to one. This means that every metabolite substructure and protein domain combination that can be identified in this transaction corresponds to one of the relationships that was mined out of the STITCH dataset. In other words, the entirety of the molecular interactions within this pathway can be inferred from the patterns mined from STITCH. Figure 4 shows the pathway with each of the substructures identified through pattern mining shown in colour.

3.3 Predicting antibiotic resistance patterns using association rules

Antibiotic resistance is one of the major challenges for global health care. More and more bacteria are growing resistant to antibiotics used in the clinic, highlighting the need for an improved understanding of these mechanisms. To demonstrate the utility of the association rules derived from the STITCH dataset, we used our set of rules to predict which antibiotics may be affected by a certain resistance mechanism. Our validation dataset consists of the CARD database, which provides a list of proteins and the antibiotics to which they confer resistance, for a total of 7,444 relationships composed of 877 unique proteins and 151 unique antibiotics. Protein domains were extracted for each protein, while each antibiotic was converted to a series of molecular fingerprints. In order to predict antibiotic

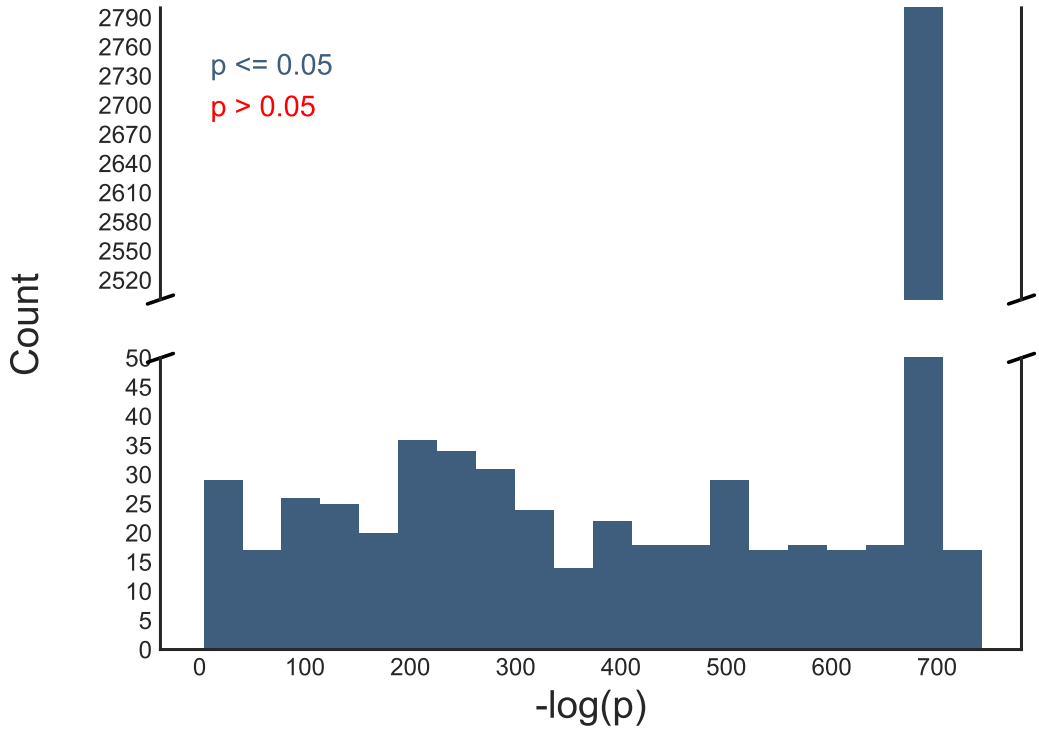

Figure 2. Patterns identified in the STITCH dataset match patterns in a metabolic pathway dataset. This figure shows the logarithm of the p-values for the Fischer's exact test determining how well the patterns mined from the STITCH dataset match patterns mined from a metabolic pathway dataset for each of the 3,527 metabolite - protein transactions. Higher -log(p) values indicate more significant enrichment. Significantly enriched transactions are shown in blue, non-significantly enriched transactions are shown in red.

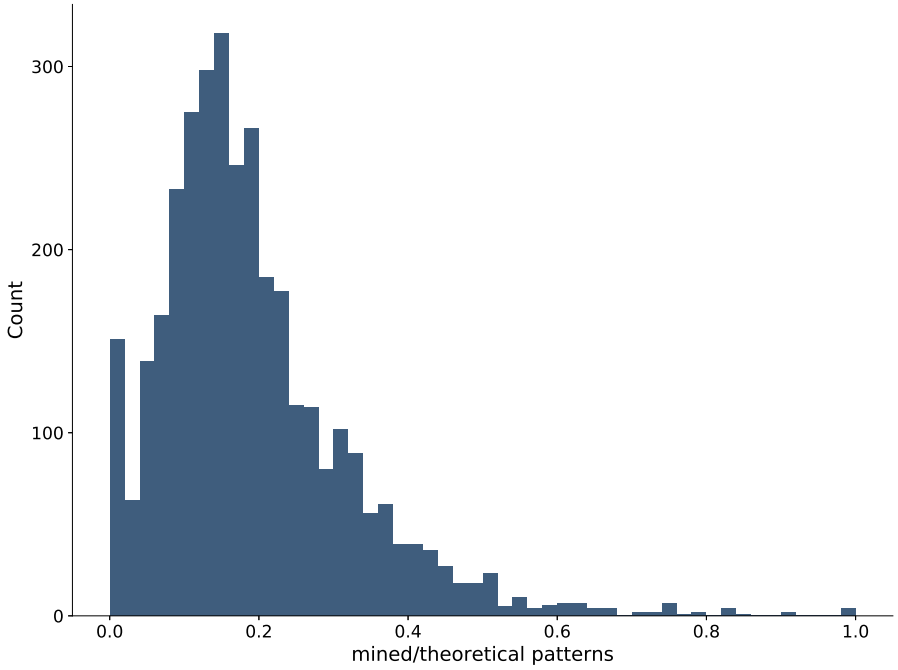

Figure 3. The ratio of patterns mined from STITCH to those present in the transaction for each of the 3,527 metabolite - protein transactions present in the ConsensusPath database. For a number of pathways this ratio was equal to 1, indicating that every substructure-domain combination present in this reaction corresponds to one of the relationships that was mined from the STITCH dataset.

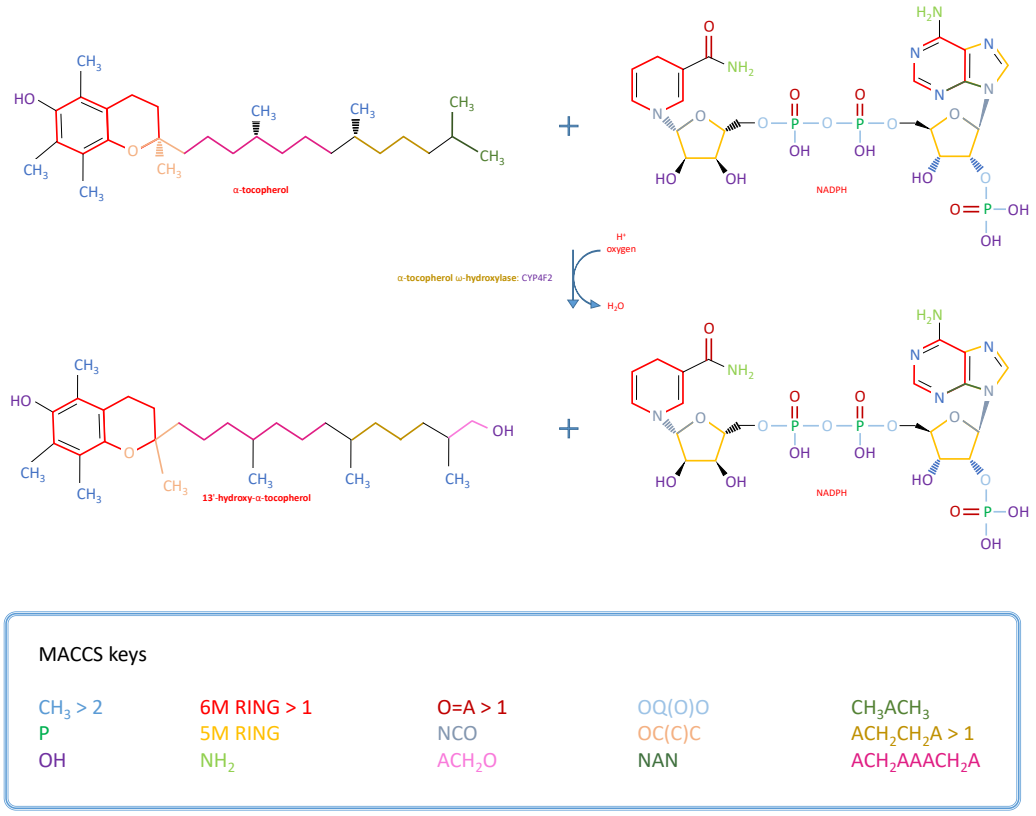

Figure 4. Patterns mined from STITCH explain alpha-tocopherol-omega-hydroxylation. The ratio of patterns mined from STITCH to patterns present in the transaction was equal to one for the alpha-tocopherol-omega-hydroxylation reaction catalyzed by CYP4F2. Every metabolite substructure and protein domain combination present in this reaction thus matches one of the relationships obtained by mining the STITCH dataset. The colour of each substructure corresponds to one of the MACCS keys shown under the figure.

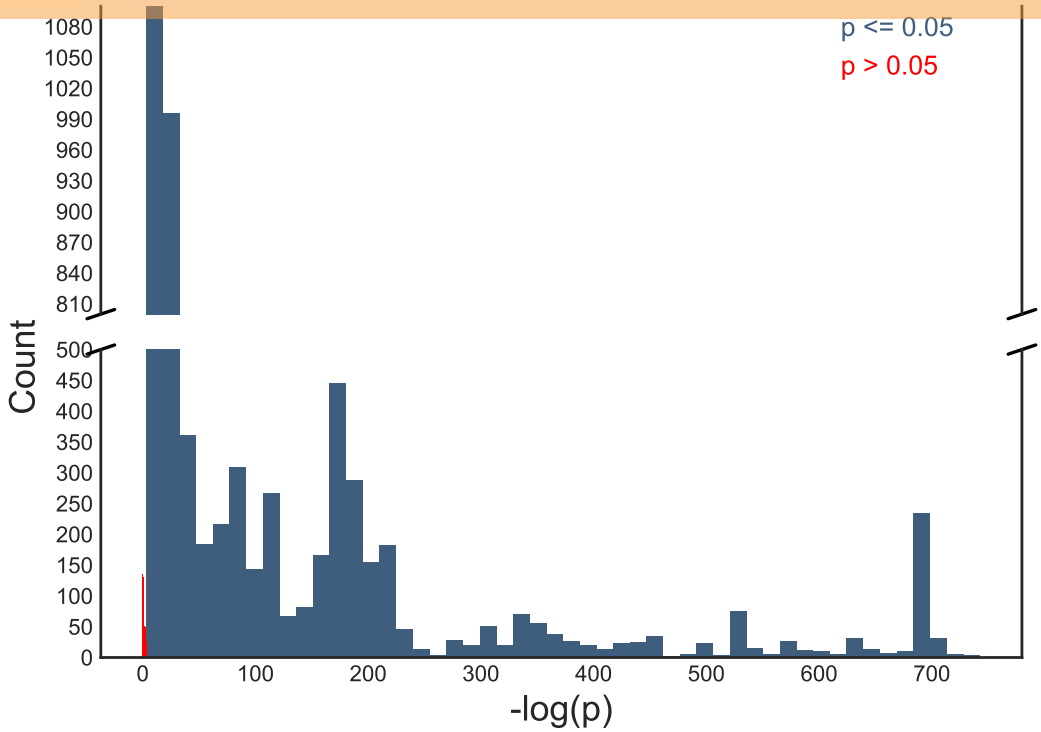

Figure 5. Association rules can recommend patterns for unrelated datasets. This figure shows the logarithm of the p-values for the Fischer's exact test determining how well the patterns proposed by our association rules (derived from STITCH) match patterns mined from a metabolic pathway dataset for of the 7,444 antibiotic - antibiotic resistance protein transactions present in the CARD database. Higher -log(p) values indicate more significant enrichment. Significantly enriched transactions are shown in blue, non-significantly enriched transactions are shown in red.

resistance, we used our set of association rules in the following fashion: for every protein from CARD, represented by protein domains, we identified the set of rules containing those protein domains in the rule body. These rules were then used to recommend substructures for the protein, sorted by the mean confidence of the rule recommending them. In order to determine whether these recommended substructures are statistically superior to randomly assigning substructures to protein domains, we used a Fisher's exact test in the same manner as previously described, here comparing our recommended patterns to the patterns mined from the resistance protein - antibiotic transactions. Figure 5 shows the p-values for this test, which indicates that our method is able to provide relevant recommendations.

We furthermore calculate a receiver operator characteristic (ROC) curve for these recommendations (Figure 6). The ROC curve plots the true positive rate (TPR), the predicted substructures which are actually present in the antibiotics to which the protein confers resistance, as a function of the false positive rate (FPR), or the substructures predicted by our method which are *not* present in the antibiotics to which the protein confers resistance. These results demonstrate that our method can accurately identify substructures of antibiotics which are sensitive to drug resistance proteins, based on the average confidence of the method for each of the recommendations.

The fingerprint recommendations we have generated for each antibiotic resistance protein were then used to rank all 151 antibiotics by the likelihood of being affected by this resistance mechanism. The results are summed up in Table 2. While the mean rank of the true hit was low (68), at least one correct antibiotic was ranked within the top fifteen for 28% of the proteins.

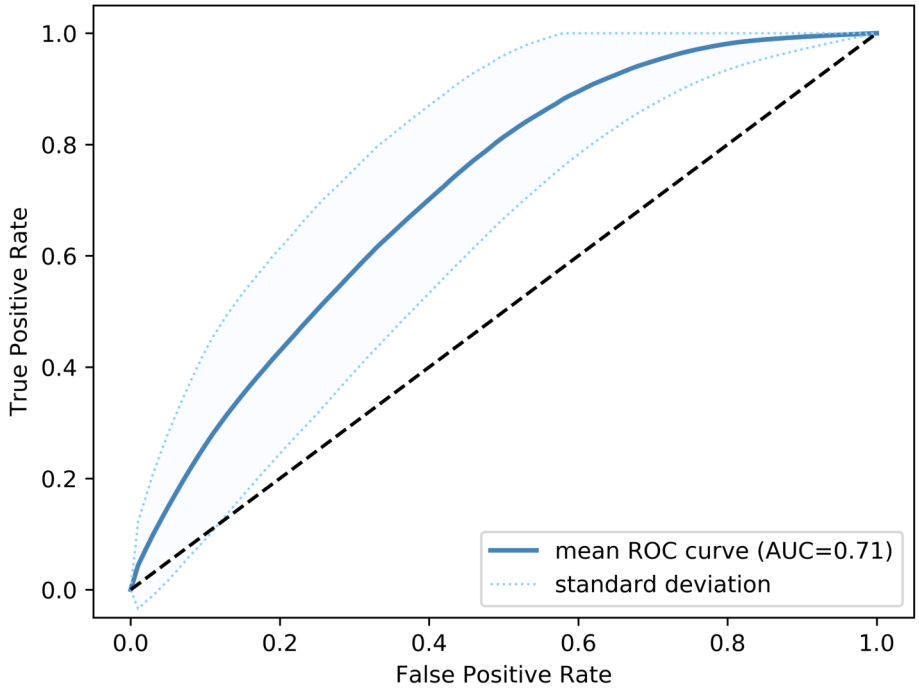

Figure 6. Association rules can be used to predict drug resistance. The ROC curve plots the true positive rate (TPR), the predicted substructures which are actually present in the antibiotics to which the protein confers resistance, as a function of the false positive rate (FPR), or the substructures predicted by our method which are *not* present in the antibiotics to which the protein confers resistance. The mean ROC curve shown here was obtained by averaging over the ROC curves for all antibiotic resistance proteins.

#unique proteins	877
#unique antibiotics	151
mean rank of true positive	68
#true positive ranked in top15	2048 (28%)

Table 2. Summary of the results for ranking antibiotics susceptible to antibiotic resistance proteins based on association rules.

4 CONCLUSION

The prediction of interactions between drugs and their targets is central to the field of cheminformatics. Such methods have tremendous application potential, for instance in the development of new drugs or the predication of side effects. Numerous methods have been developed allowing for such predictions, but it remains difficult to transfer knowledge to new application areas where information about binding is scarce.

We present a proof-of-concept showing that a conceptually elegant frequent itemset mining approach is capable of elucidating the molecular patterns governing drug-target interactions. By mining databases for frequently occurring interactions between molecular substructures and protein domains, patterns may be identified which capture these molecular interactions. We mine patterns from a protein-ligand interaction dataset and show that similar patterns also underlie an orthogonal dataset of metabolic pathways. A set of association rules which may be used to recommend substructures for given protein domains was generated based on the patterns identified in a human protein-ligand database. For a given bacterial antibiotic resistance protein, these rules were able to recommend substructures present in susceptible antibiotics. The utility of these rules was further demonstrated by using them to rank antibiotics by their likelihood for interaction with a given bacterial resistance protein. Our results show that this method is able to identify and extract patterns from one dataset and then utilize them in diverse settings.

The itemset mining approach we use here is conceptually elegant and provides easy to understand recommendations. Another key advantage is that it is highly flexible, allowing for the inclusion of a variety of discrete features. In future work, the itemsets examined here may be extended to include additional features of the protein such as post-translational modifications or amino acid mutations. More elaborate substructure key based fingerprints may also be used to further augment this method. Finally, the features derived using this method may be used to train supervised machine learning models in order to further augment predictive performance.

In conclusion, we show that general patterns for molecular interactions may be identified through frequent itemset mining, and that this method may be used to transfer insights mined from these patterns to diverse application areas.

REFERENCES

[1] Fabio Pammolli, Laura Magazzini, and Massimo Riccaboni. The productivity crisis in pharmaceutical r&d. *Nature reviews Drug discovery*, 10(6):428, 2011.

[2] Joseph A DiMasi, Henry G Grabowski, and Ronald W Hansen. Innovation in the pharmaceutical industry: new estimates of r&d costs. *Journal of health economics*, 47:20–33, 2016.

[3] Michael J Waring, John Arrowsmith, Andrew R Leach, Paul D Leeson, Sam Mandrell, Robert M Owen, Garry Pairaudeau, William D Pennie, Stephen D Pickett, Jibo Wang, et al. An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature reviews Drug discovery*, 14(7):475, 2015.

[4] Ted T Ashburn and Karl B Thor. Drug repositioning: identifying and developing new uses for existing drugs. *Nature reviews Drug discovery*, 3(8):673, 2004.

[5] Ali Ezzat, Min Wu, Xiao-Li Li, and Chee-Keong Kwoh. Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. *Briefings in Bioinformatics*, page bby002, 2018.

[6] Hao Ding, Ichigaku Takigawa, Hiroshi Mamitsuka, and Shanfeng Zhu. Similarity-based machine learning methods for predicting drug–target interactions: a brief review. *Briefings in Bioinformatics*, 15(5):734–747, 2014.

[7] Zaynab Mousavian and Ali Masoudi-Nejad. Drug–target interaction prediction via chemogenomic space: learning-based methods. *Expert Opinion on Drug Metabolism & Toxicology*, 10(9):1273–1287, 2014. PMID: 25112457.

[8] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM, 1993.

[9] Stefan Naulaerts, Pieter Meysman, Wout Bittremieux, Trung Nghia Vu, Wim Vanden Berghe, Bart Goethals, and Kris Laukens. A primer to frequent itemset mining for bioinformatics. *Briefings in bioinformatics*, 16(2):216–231, 2013.

[10] Aida Mrzic, Pieter Meysman, Wout Bittremieux, and Kris Laukens. Automated recommendation of metabolite substructures from mass spectra using frequent pattern mining. *bioRxiv*, page 134189, 2017.

[11] Michael Hahsler, Christian Buchta, Bettina Gruen, and Kurt Hornik. *arules: Mining Association Rules and Frequent Itemsets*, 2018. R package version 1.6-1.

[12] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

[13] Damian Szklarczyk, Alberto Santos, Christian von Mering, Lars Juhl Jensen, Peer Bork, and Michael Kuhn. STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic Acids Research*, 44(Database issue):D380–D384, 2016.

[14] Adria Cereto-Massague, Maria Jose Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallve, and Gerard Pujadas. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63, 2015.

[15] Rdkit: Open-source cheminformatics.

[16] Robert D Finn, Teresa K Attwood, Patricia C Babbitt, Alex Bateman, Peer Bork, Alan J Bridge, Hsin-Yu Chang, Zsuzsanna Dosztányi, Sara El-Gebali, Matthew Fraser, et al. Interpro in 2017—beyond protein family and domain annotations. *Nucleic acids research*, 45(D1):D190–D199, 2016.

[17] The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, 2017.

[18] Atanas Kamburov, Ulrich Stelzl, Hans Lehrach, and Ralf Herwig. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Research*, 41(D1):D793, 2013.

[19] Baofeng Jia, Amogelang R. Raphenya, Brian Alcock, Nicholas Waglechner, Peiyao Guo, Kara K. Tsang, Briony A. Lago, Biren M. Dave, Sheldon Pereira, Arjun N. Sharma, Sachin Doshi, Mélanie Courtot, Raymond Lo, Laura E. Williams, Jonathan G. Frye, Tariq Elsayegh, Daim Sardar, Erin L. Westman, Andrew C. Pawlowski, Timothy A. Johnson, Fiona S.L. Brinkman, Gerard D. Wright, and Andrew G. McArthur. Card 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Research*, 45(D1):D566–D573, 2017.