Linking Biochemical Pathways and Networks to Adverse Drug Reactions

Huiru Zheng, *Member, IEEE*, Haiying Wang, Hua Xu, Yonghui Wu, Zhongming Zhao, Francisco Azuaje

***Abstract*—There is growing interest in investigating the biochemical pathways involved in cellular responses to drugs. Here we propose new methods to explore the relationships between drugs, biochemical pathways and adverse drug reactions (ADRs) at a large scale. Using sparse canonical correlation analysis of 832 drugs characterized by 173 pathways and 1385 ADRs profiles, we identified 30 highly correlated sets of drugs, pathways and ADRs. This included known and potentially novel associations. To evaluate the predictive performance of our method, the extracted correlated components were used to predict known ADR profiles from drug pathway profiles. A relatively high prediction performance (AUC: 0.894) was achieved. To further investigate their association, we developed a network-based approach to extract potentially significant modules of pathway-ADR associations. Five statistically significant modules were extracted. We found that most of the nodes contained in the modules are either pathways linked to a very limited number of drugs or rare ADRs. The work provides a foundation for future investigations of ADRs in the context of biochemical pathways under different clinical conditions. Our method and resulting datasets will aid in: a. the systematic prediction of ADRs, and b. the characterization of novel mechanisms of action for existing drugs. This merits additional research to further assess its potential in improving personalised drug safety monitoring, as well as for the repositioning of drugs in the longer-term**

***Index Terms*— Biological pathways; adverse drug reactions; sparse canonical correlation analysis, pharmacogenetics**

I. Introduction

According to the World Health Organisation [1], an adverse drug reaction (ADR) is defined as "a response to a drug which is noxious and unintended, and which occurs at doses used in humans for the prophylaxis, diagnosis or therapy of disease, or for the modification of physiological function". It has been well recognised that ADRs are a significant cause of morbidity and mortality, resulting in a significant burden on the healthcare service across the world [2], [3]. For example, it has been estimated that ADRs would account for 6.5% of all UK hospital admissions, which costs the National Health Service (NHS) up to £466 million annually [4]. Thus, during early phases of drug development, identification of potential ADRs is critical for successful drug development.

The recognised significance has triggered huge efforts from industry and scientific communities to develop various computational models for predict potential ADRs at large scale. Bender et al. [5] explored the chemical space and made the first attempt to predict ADRs across hundreds of ADR categories from chemical structure alone, achieving 92% classification accuracy. Cami et al. [6] developed a computational network-based method for predicting ADRs. They constructed a network representation of the associations between drugs and adverse drug events (ADE) using 809 drugs and 852 ADEs collected since 2005, and then trained a logistic regression model to predict unknown side effects of drugs in the network. Liu et al. [7] proposed a machine-learning-based approach for ADR prediction by integrating chemical structure information, drug related biological properties, such as protein targets and pathway information, and drug phenotypic characteristics. They found that drug phenotypic information such as the drug indication is the most informative feature of ADR prediction. The model successfully predicted the ADRs that are associated with the withdrawal of rofecoxib and cerivastatin. More recently, Harpaz et al. [8] proposed a "signal-detection strategy" that combines the adverse event reporting system (AERS) of the USA Food and Drug Administration (FDA) and electronic health records (EHRs) for detection of ADRs. Finally, Lui et al. [9] proposed a causality analysis model based on structure learning to identify important factors that contribute significantly to specific drug ADRs. After applying the causal features captured by the proposed model to a traditional support vector machine classifier, a significant increase in performance was reported

We have entered big data era. There are massive amounts of pharmacogenetic and related data already available, and growth rate of such data is expected to be even higher in the

Manuscript received April 7, 2014.

H. Zheng is with the Computer Science Research Institute, University of Ulster, Jordanstown Campus, Shore Road, Newtownabbey BT37 0QB, United Kingdom (e-mail: h.zheng@ulster.ac.uk).

*H. Wang is with the Computer Science Research Institute, University of Ulster, Jordanstown Campus, Shore Road, Newtownabbey BT37 0QB, United Kingdom (phone: 44-28-90366981, fax: 44-28-90366068, e-mail: hy.wang@ulster.ac.uk).

H. Hua is with the School of Biomedical Informatics, University of Texas, Houston, Texas, USA (e-mail: Hua.Xu@uth.tmc.edu)

Y. Wu is with the School of Biomedical Informatics, University of Texas, Houston, Texas, USA (e-mail: Yonghui.Wu@uth.tmc.edu)

Z. Zhong is with the Department of Biomedical Informatics, Vanderbilt University, Tennessee, USA (e-mail: zhongming.zhao@Vanderbilt.Edu)

F. Azuaje is with the NorLux Neuro-Oncology Laboratory, Public Research Centre for Health (CRP-Santé), Luxembourg (e-mail: francisco.azuaje@crp-sante.lu).

.

next few years. Therefore, it is important for us to identify important molecular signals underlying the pharmacogenetic data. Specifically, it is promising to investigate biochemical pathways involved in cellular response to drugs because drug targets are often involved in important pathways. Wallach et al. [10] highlighted that understanding the biological processes behind the occurrence of ADRs may have significant applications and implications in the life sciences and pharmaceutical industries. This may lead to the development of safer and more effective drugs, the discovery of new bio-markers, and the identification of new uses for existing drugs (drug repositioning). Silberberg et al. [11] argued that uncovering drug-induced signaling pathways is an important step in understanding a drugs' mode of action and inferring drug properties such as ADRs. In an integrative analysis using human protein-protein and protein-DNA interactions, as well as drug targets and drug-induced gene expression data, they identified 428 drug-specific signalling sub-networks and 99 putative signalling pathways. In another study by Chen et al. [12], the authors hypothesized that a portion of a pathway (i.e., a sub-graph of the pathway) might be more sensitive in drug response to a particular biological condition than the whole pathway. This hypothesis is valid because canonical pathways, like those annotated in the Kyoto Encyclopaedia of Genes and Genomes (KEGG) database [13], might be too large and complicated while only subpart(s) of a pathway are under regulation in a cellular condition or in the response to environmental changes. Correspondingly, Chen et al. developed a computational framework for searching primary subnetwork(s) of drug responses by effectively utilizing the sub-pathway information.

By extending our preliminary analysis [14], here, we further investigated the relationship between biochemical pathways and ADRs at a large scale using computational approaches. Through our computational analyses, we aimed to answer the following questions: (1) Can we identify correlated sets of pathways and ADRs by a computational approach? (2) Can we predict a drugs' ADR based on its pathway activity profiles? And (3) How to effectively measure the association between pathways and ADRs using the data from knowledge base?

The rest of this paper is organized as follows. Section II briefly describes the method, including the datasets and prediction algorithms. The results are presented in Section III. The discussion and conclusions, together with future research directions, are given in Section IV.

II. METHODOLOGY

A. Datasets

The dataset was obtained from a study by Liu et al. [7]. It contains 832 drugs, and each drug was represented by the following two high-dimensional profiles.

1) A 1385 binary vector whose elements encode for the presence or absence of an ADR by "1" or "0", respectively. The associations between drugs and ADRs were extracted from SIDER [16].

2) A 173 binary vector in which "1" indicates the association between a drug and a corresponding pathway. The relationship between drugs and pathways was constructed by mapping protein targets extracted from DrugBank [17] to the corresponding KEGG biological pathways [18], [19] through their protein-coding gene symbols.

In total, 2182 links between 832 drugs and 173 KEGG pathways and 59,205 associations between the drugs and 1385 ADRs were identified. While each drug has a relatively large number of ADRs with a mean of 42.7 and a standard deviation of 87.0, the number of pathways linked to each drug is relatively small (12.6 on average with a standard deviation of 22.2). Among them, the drug arsenic trioxide was found to be associated with the largest number of pathways (51) derived from its protein targets, and pregabalin was found to have the largest number of ADRs (453)

B. Canonical correlation analysis

Developed by H. Hotelling [20], canonical correlation analysis (CCA) aims to quantify the associations among two sets of features (pathways and ADR in our case) on the same set of samples, i.e. drugs in this study. It has become a well-known tool in statistical analysis and has attracted growing attention over the past years [21], [23] .

Let each drug be represented by a pathway feature vector $\mathbf{x} = \{x_1, x_2, \cdots, x_p\}^T$ and an ADR feature vector $\mathbf{y} = \{y_1, y_2, \cdots, y_q\}^T$ where p and q stand for the number of pathways and ADRs under study respectively. Ordinary CCA (OCCA) seeks to find two weight vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ for $\mathbf{x}$ and $\mathbf{y}$, i.e. $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \cdots, \alpha_p\}^T$ and $\boldsymbol{\beta} = \{\beta_1, \beta_2, \cdots, \beta_q\}^T$ such that the following correlation coefficient is maximized.

$$corr(u, v) = \frac{\sum_{i=1}^{n} \alpha^T x_i \cdot \beta^T y_i}{\sqrt{\sum_{i=1}^{n} (\boldsymbol{\alpha}^T \boldsymbol{x}_i)^2} \sqrt{\sum_{i=1}^{n} (\boldsymbol{\beta}^T \boldsymbol{y}_i)^2}} \quad (1)$$

, where n is the total number of drugs under consideration. $\mathbf{u} = \boldsymbol{\alpha}^T \mathbf{x}$ and $\mathbf{v} = \boldsymbol{\beta}^T \mathbf{y}$ are called canonical components (CCs). In the matrix form, the above optimization problem can be rewritten as follows:

$$\max\{\boldsymbol{\alpha}^T \boldsymbol{X}^T \boldsymbol{Y} \boldsymbol{\beta}\} \quad \text{subject to} \quad \|\boldsymbol{\alpha}\|_2^2 \le 1, \quad \|\boldsymbol{\beta}\|_2^2 \le 1 \quad (2)$$

, where $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_n}]^T$ and $\mathbf{Y} = [\mathbf{y_1}, \mathbf{y_2}, \cdots, \mathbf{y_n}]^T$ denote the $n \times p$ and $n \times q$ matrices, respectively.

It has been shown that normally vectors $\mathbf{u}$ and $\mathbf{v}$ derived from OCCA are not sparse, making the interpretation of results quite difficult. In an attempt to impose the sparsity on weight vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to yield interpretable factors, we applied the sparse version of CCA (SCCA) based on a penalized matrix decomposition (PMD) technique introduced by Witten et al. [21]. The idea is to impose additional constraints to the elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, i.e.

$$\max\{\boldsymbol{\alpha}^T \boldsymbol{X}^T \boldsymbol{Y} \boldsymbol{\beta}\} \quad \text{subject to} \quad \|\boldsymbol{\alpha}\|_2^2 \le 1, \quad \|\boldsymbol{\beta}\|_2^2 \le 1, P_1(u) \le c_1, P_2(v) \le c_2 \quad (3)$$

, where $c_1 \in (0,1]$ and $c_2 \in (0,1]$ are parameters used to control the sparsity. P_1 and P_2 are convex penalty functions,

which can take on a variety of forms [21]. In this study, the SCCA was implemented using (the R package) PMA [22]. A lasso penalty was used to obtain the corresponding CCs.

In order to obtain multiple CCs, a deflation manipulation was carried out recursively. The criterion expressed in (3) was implemented repeatedly each time by using the $Z = X^T Y$ matrix as the residuals obtained by subtracting the previous found factors from the matrix. As a result, m pairs of weight vectors, in which high scoring in both sets are extracted as correlated sets, will be obtained. The reader is referred to [21] and [23] for a detailed description of the implementation.

III. Results

A. Associations between ADRs and pathways: a statistical analysis

There is no direct link between the number of linked pathways and the number of associated ADRs for each drug as shown in Fig. 1 with the Pearson correlation coefficient being close to zero (0.074). For example, the drug *pregabalin*, a drug used for neuropathic pain, has the largest number of ADRs (453). However, its protein targets were mapped only to 2 KEGG pathways (hsa04614 and hsa01040). Interestingly, there is no KEGG pathway found to be associated with the drug venlafaxine, a drug used for the treatment of major depressive disorder, yet it has 319 ADRs. The number of associated ADRs for the top 10 drugs that have the largest number of pathways varies substantially, ranging from 19 to 244. These results highlight the limited amount of available knowledge about mechanism of action of clinically approved drugs.

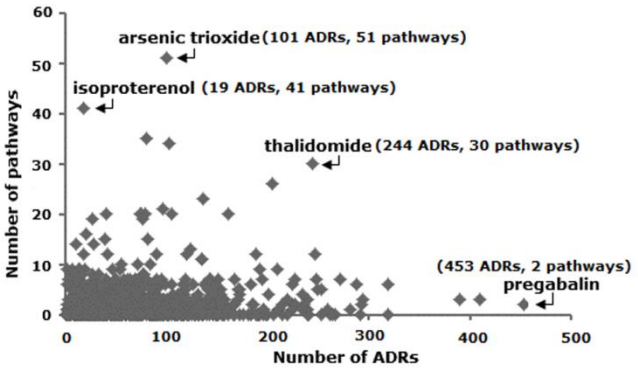

Fig. 1 The correlation between the number of pathways linked to each drug and the number of ADRs.

B. Associations between ADRs and pathways: SCCA-based analysis

In order to extract correlated sets of pathways and ADRs, we applied SCCA to the dataset. We then evaluated the predictive performance of the method by recovering known ADRs from the extracted drug pathway profiles. The system was implemented within the R framework [24]. The best performance was achieved with $c_1 = c_2 = 0.1$ and $m = 30$.

1)Extraction of KEGG pathway-ADR associations: The SCCA-based analysis provides us with 30 CCs, each containing a limited number of correlated, high scoring pathways and ADRs. To gain a global view of pathway-ADR associations, we merged the results for all derived components and represented them as a network, in which pathways and ADRs are connected if they are found in the same component (Fig.2). For simplicity of visulisation, we focused on pathways and ADRs whose weights are greater than 0.1. Accordingly, this network has a total of 353 nodes, including 296 ADRs and 57 pathways, and 755 connections.

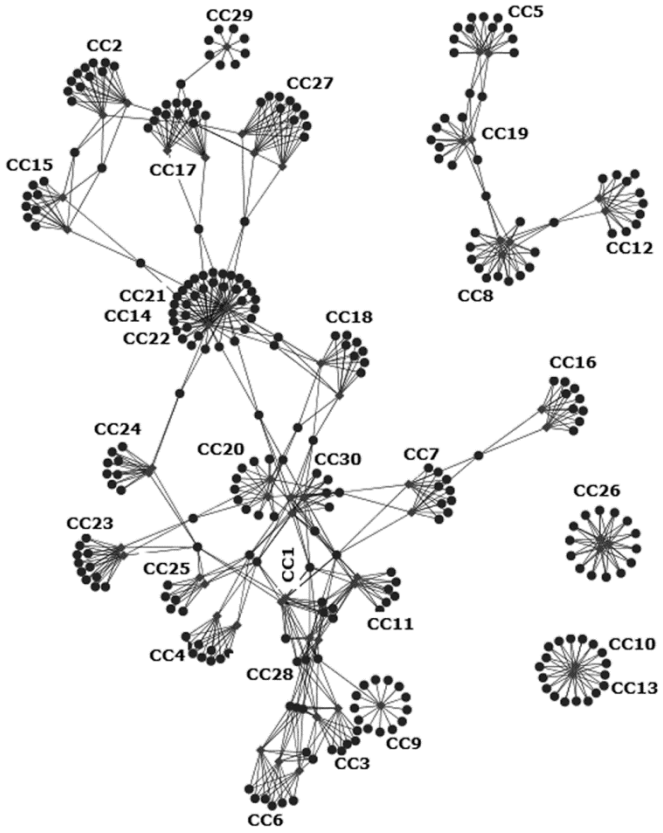

Fig. 2 An illustration of the network of pathways and ADRs using the extracted 30 CCs. Pathways (light grey rectangles) and ADRs (dark circles) are connected if they are found in the same extracted CC

The network shows a modular structure, where links between ADR and pathway nodes are much denser within each CC than between CCs. While CC14, CC21, and CC22 share the same set of pathways, i.e. taste transduction (hsa04742) and type II diabetes mellitus (hsa04930), the ADRs having a high score in these three components are very different with those having a distinct set of ADRs (11 in CC14, 16 in CC21 and 17 in CC22). A similar observation can be made when examining the association between pathways and ADRs in CC10 and CC13, in which the same pathways, i.e. oocyte meiosis (hsa04114) and progesterone-mediated oocyte maturation (hsa04914), were found.

A closer look at the degree distribution reveals that the distribution over ADR and pathway nodes is different. All the pathways are associated with at least 8 ADRs while other pathways: taste transduction (hsa04742) and type II diabetes mellitus (hsa04930) are connected to the largest number of ADR (41). On the other hand, more than 70% of ADRs are connected to less than 3 pathways. Out of 296 ADRs, only 17 are found to be associated with more than 5 pathways with the ADR parapsoriasis (C0030491) linked to the highest number of pathways (11).

For each component, the associated canonical correlation coefficient was estimated. We observed that the components with high correlation tend to contain pathways related to very few drugs and rare ADRs mainly observed in these drugs. For example, CC9 has a highest canonical correlaction close to 1.0 (0.985). The only pathway found in this component with a score greater than 0.1 is proteasome pathway (hsa03050), which is only related to bortezomib, the first therapeutic proteasome inhibitor to be tested in humans. Intererstingly, all 14 ADRs contained in the components with a score higher than 0.1 are associated with this drug. They are : C0004030, C0015544, C0018775, C0032768, C0025309, C0155773, C0019357, C0040558, C0002726, C0235329, C0259749, C0155919, C0162323, and C0085077. The top pathway found in CC6 with the canonical correlation 0.942 is inositol phosphate metabolism (hsa00562). The only drug whose target protein is mapped to this pathway is *lithium*, which affects the flow of sodium through nerve and muscle cells in the body. The top two ADRs, i.e. nontoxic goiter (C0221777) and toxic goiter (C0600086) contained in this component are found to be associated with this drug only.

Our analyses provide two lists of drugs for each extracted component: those with a high score for the associated pathways and ADRs respectively. Interestingly, most drugs that have high scores for pathways in a component with a high correlation are found to have high scores for the ADRs in the same component. This is consistent with the idea that, in principle, the more we know about the mechanism of action of a drug, the more we can know about its potential adverse effects. For example, the drug lithium has the highest scores for both the pathways and ADRs contained in CC6.

2) *Performance evaluation:* We tested the assumption that the extracted correlated sets are predictive of ADRs. To do this, we evaluated the performance of the method by using the extracted CCs and drug pathway profiles to detect known ADR profiles extracted from the SIDER database [16].

A 5-fold cross-validation was applied, i.e., the entire dataset is randomly partitioned into 5 subsets of approximately equal size and each subset in turn is used as the test set while the remaining 4 subsets are used as training data. The goal of the classification posed in this application is to predict ADRs associated with each drug based on its pathway information. The performance was assessed by a receiver operating characteristics (ROC) curve, which is a plot of true positive rate (the percentage of actual positives correctly identified) against false positive rate (the fraction of false positives out of the negatives) at various prediction score thresholds. Any predicted ADR with a prediction score greater than a given threshold is considered as positive and negative otherwise.

The area under the ROC curve (AUC) was estimated to summarize the prediction performance, as illustrated in Fig.3, where the prediction scores for all the ADRs were merged and a global ROC curve was obtained. Afterwards, we estimated the performance by changing the sparsity parameter, i.e. c_1 and c_2, from 0 to 1 with 0.1 increments and the number of CC from 10 to 100 with 10 increments. The optimal performance was derived with $c_1 = c_2 = 0.1$ and $m = 30$. Both SCCA and OCCA achieved fairly good results with SCCA having a slightly better performance (AUC: 0.894 for SCCA and 0.888 for OCCA, well above the performance based on random assignment). This suggests that the extracted pathway-ADR associations are indeed useful for drug ADR prediction. In comparisons with other studies, the results derived from our method are at a competitive level with models based on chemical and biological features [7], [26].

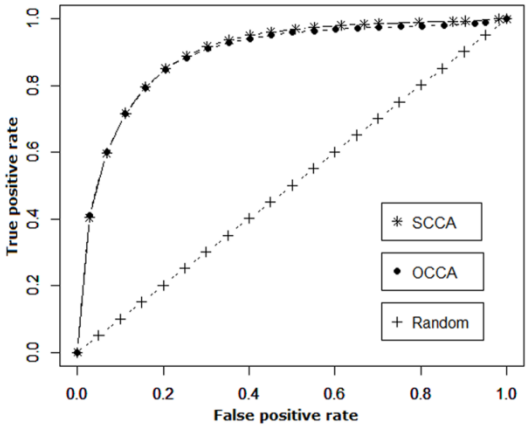

Fig. 3 ROC curves based on a 5-fold cross validation. Comparison of the performance between SCCA, OCCA, and random

The predictive power of the proposed method can be further demonstrated by examining the prediction accuracy of the predicted ADRs for each drug. We checked the predicted ADRs with high prediction scores against the known ADRs reported in the SIDER database [16]**Error! Reference source not found.**. For example, for the drug *pramipexole,* the ADR ranked highest in the prediction score is *nasal polyps* (C0027430), a known ADR for pramipexole [16]. Among the top 10 high scoring ADRs, 6 are the known ADRs linked to *pramipexole*. Similarly, the top predicted ADR for the drug *ropinirole* is one of the known ADRs listed in the SIDER database, and 9 out of top 15 ADRs with high prediction scores are known ADRs for this drug [16].

Turning to biological interpretability, we found that the proposed SCCA method has the advantage over other machine learning techniques, such as Support Vector Machine and Naïve Bayesian, which do not provide direct biological interpretation clues. As shown in Fig. 4, each correlated set derived by SCCA has only a few dominant elements whose weight is far greater than the average. Most of the elements in the weight vectors associated with each component are zero or close to zero in each component, suggesting that SCCA has the ability to select a small number of features as informative pathway and ADRs. In contrast, almost all elements contained in the weight vectors derived from OCCA are non-zero and there is no clear dominant element found in most of components. Interpreting such a weight vector may prove to be rather difficult in practice.

C. Associations between ADRs and pathways: Network-based analysis

In this section we explore the feasibility of using a network-based approach to extract associations between pathways and ADRs. We first estimated the similarity between a pathway and an ADR in terms of their drug profiles. Let a pathway, p_i, be represented by a binary vector in which each

element indicates that the corresponding drug is known to impact the pathway or not, i.e. $\boldsymbol{p_i} = \{d_1, d_2, \cdots, d_n\}$, $(n = 1, 3, \cdots, 832)$.

$$d_i = \begin{cases} 1, & i^{th}\ drug\ is\ involved\ in\ the\ given\ pathway \\ 0, & otherwise \end{cases} \quad (4)$$

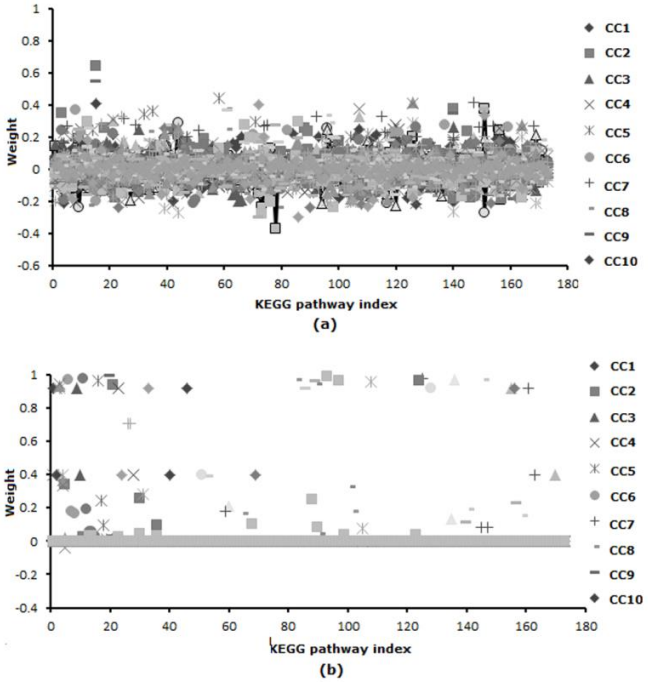

Fig. 4 The distribution of weight vectors over pathways in (a) OCCA and (b) SCCA. The first ten canonical components are shown.

Similarly, let an ADR, i.e., s_j, be represented by a binary vector whose elements encode whether the ADR is associated with the corresponding drug, $\boldsymbol{s_j} = \{d_1, d_2, \cdots, d_n\}$, $(n = 1, 3, \cdots, 832)$.

$$d_i = \begin{cases} 1, & the\ ADR\ is\ observed\ in\ the\ i^{th}\ drug \\ 0, & otherwise \end{cases} \quad (5)$$

The association between pathway, p_i, and ADR, $\boldsymbol{s_j}$, can be estimated using the Jaccard similarity coefficient, i.e.

$$sim(p_i, s_j) = \frac{|p_i \cap s_j|}{|p_i \cup s_j|} \quad (6)$$

The values generated by (6) vary between 0 and 1 where "1" implies that the ADR is associated with the same set of drugs that impact the given pathway.

We then constructed a pathway-ADR network, in which the edge between pathway and ADR nodes is represented by the Jaccard similarity estimated using Equation (6) as shown in Fig. 5. Unlike the unweighted network depicted in Fig. 2, in which a pathway and a ADR is connected if they are found in the same extracted CC, the weighted network illustrated in Fig. 5 is based on the proportion of drugs associated with both the ADR and the pathway. For a better visualization, here we focused on the analysis of the association with the similarity greater than 0.1. The resulting network including 724 nodes (160 pathways and 564 ADRs) and 1744 weighted edges is characterized by a small number of nodes having a high degree accompanied by a large number of nodes whose degree is less than 3 as depicted in Fig. 5. The top 10 most connected nodes are all pathway nodes and the top 3, i.e. gap junction (hsa04540), calcium signaling pathway (hsa04020), neuroactive ligand-receptor interaction (hsa04080) connect to more than 100 ADRs. Interestingly, the cell adhesion molecules pathway (hsa04514) is associated with only 2 drugs, i.e. *glatiramer acetate* used to treat multiple sclerosis and *lenalidomide* used to treat patient with myeloma yet it connects to 33 ADRs, suggesting that these two drugs may in reality have a large number of ADRs. The actual numbers of known ADRs associated with these 2 drugs are 234 and 234 respectively. A similar observation can be made when we examine the Proteasome pathway (hsa03050), which is only linked to one drug (*bortezomib*) but it has connections with 22 ADRs.

The most connected ADR node is C0085786, i.e. alveolitis fibrosing, while nearly of ADR nodes are connected to less than 3 pathways. Unlike pathway nodes whose degree is strongly correlated with the number of associated drugs (Pearson correlation coefficient: 0.828), there is virtually no correlation between the degree and the number of associated drugs for ADR nodes (Pearson correlation coefficient: 0.0337). For example the ADR C0027497, i.e.nausea, is observed in more than 700 drugs, while it is found to link to 4 pathways, i.e. gap junction (hsa04540), neuroactive ligand-receptor interaction (hsa04080), calcium signaling pathway (hsa04020) and salivary secretion (hsa04970).

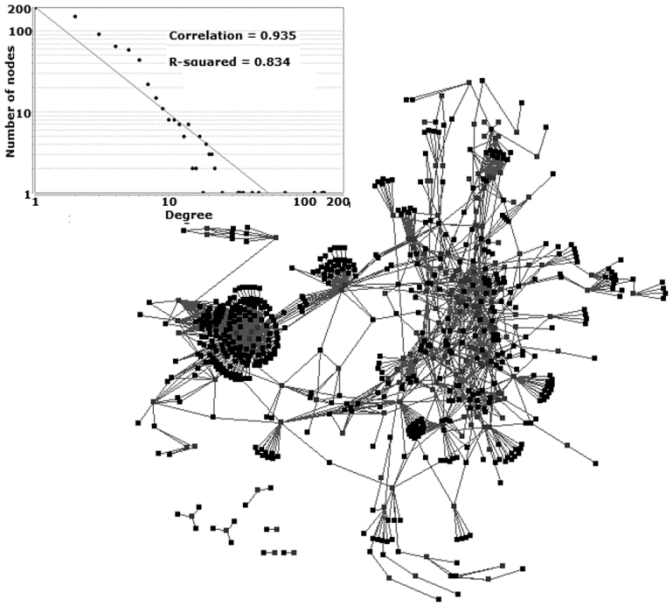

Fig. 5 An illustration of a pathway-ADR network, in which the weighted edges between pathway and ADR nodes reflect the proportion of drugs associated with both the ADR and the pathway estimated using Equation (6).

Next we applied a recently published network clustering algorithm, ClusterOne [27], to extract potentially significant modules of pathway-ADR associations. A total of 5 modules were identified ($p < 0.01$) as shown in Fig. 6. The p-value was established by using a one-sided Mann-Whitney U test performed on the in-weights (The sum of the weights of all the

edges both of whose endpoints lie in the cluster) and out-weights (The sum of the weights of the edges having one endpoint in the cluster and the other outside) of the vertices [27]. A closer look at these modules reveals that most of nodes contained are either pathways linked to a very limited number of drugs or rare ADRs. For example, the pathway proteasome (hsa03050) in Module 1 is found to be only connected to one small molecule drug, i.e. *bortezomib*, which is the first proteasome inhibitor to be approved for the treatment of relapsed multiple myeloma and mantle cell lymphoma. All the other nodes in Module 1 are rare ADRs, which are observed in less than 3 drugs including *bortezomib*. They are aspergillosis (C0004030), toxoplasmosis (C0040558), hearing loss bilateral (C0018775), keratitis herpetic (C0019357), meningoencephalitis (C0025309), failure to thrive (C0015544), portal vein thrombosis (C0155773), and post herpetic neuralgia (C0032768). Another example is Module 4 which consists of 6 pathways and 10 ADRs. The average number of drugs associated with these nodes is 3.25. The ADR benign neoplasm of skin (C0004998) is linked to the largest number of drugs (10). The top 2 drugs shared by these nodes, i.e. *glatiramer acetate* (ATC code: L03AX13) and *thalidomide* (ATC code: L04AX02) belong to antineoplastic and immunomodulating agents. The drug *glatiramer acetate* used for reduced frequency of relapses in relapsing-remitting multiple sclerosis interacts with all 6 pathways, i.e. intestinal immune network for IgA production (hsa04672), type I diabetes mellitus (hsa04940), asthma (hsa05310), allograft rejection (hsa05330), graft-versus-host disease (hsa05332), and autoimmune thyroid disease (hsa05320) and has 4 ADRs listed in the module, i.e. benign neoplasm of skin (C0004998), systolic murmur (C0232257), xanthoma (C0302314), and Cervix carcinoma stage 0 (C0851140). The small molecule drug *thalidomide* used for a number of immunological and inflammatory disorders is linked to 4 pathways, i.e. type I diabetes mellitus (hsa04940), asthma (hsa05310), allograft rejection (hsa05330), and graft-versus-host disease (hsa05332) and has 6 ADRs contained in the module, i.e. benign neoplasm of skin (C0004998), causalgia (C0007462), uterine cervical erosion (C0007869), chronic myeloid leukaemia (C0023473), lichen unspecified (C0023643), phocomelia (C0031575) and microcytic anaemia (C0085576).

The nodes in Module 3 are linked to more than 25 drugs on an average with the ADR hyperkalaemia (C0020461) being observed in more than 90 drugs. However, the proportion of drugs shared by at least two nodes in this module is relatively high. For example, out of 13 drugs found to have the ADR pemphigus (an autoimmune blistering skin disorder, C0030807), 10 are linked to all 3 pathways in the module, i.e. hypertrophic cardiomyopathy (hsa05410), chagas disease (hsa05142), and renin-angiotensin system (hsa04614). Interestingly, these drugs act on the cardiovascular system and are annotated with the same Anatomical Therapeutic Chemical Classification (ATC) code at the third level, i.e. C09AA (ACE inhibitors). A similar pattern was observed when examining the drugs shared by the ADR hyperkalaemia (C0020461) and the three pathways. Despite that a wide range of ATC codes are used to annotate the 92 drugs having the ADR hyperkalaemia (C0020461), the ATC codes for the drug set shared by C0020461 (hyperkalaemia), hsa05410 (hypertrophic cardiomyopathy), hsa05142 (chagas disease), and hsa04614 (renin-angiotensin system) are exactly the same at the first 3 levels, i.e. (ACE inhibitors).

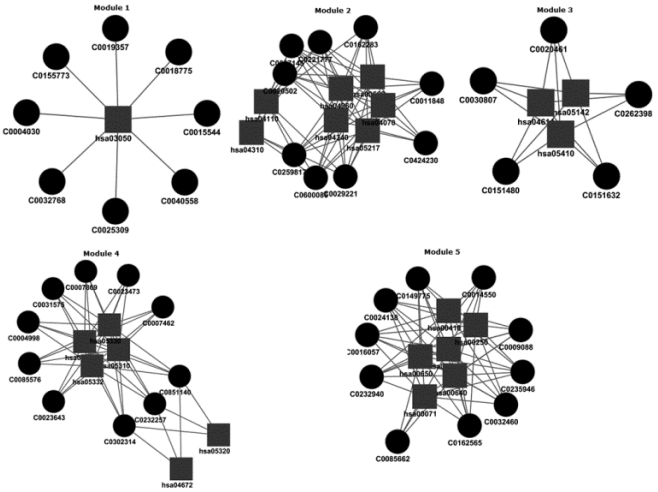

Fig. 6 Five statistically significant modules identified by ClusterOne. Rectangle nodes denote biochemical pathways and circle nodes represent ADRs. The p-values associated with each module are 0.002, 0.0001, 0.001, 0.004, and 0.000004 respectively.

IV. Conclusion

This investigation proposes a new method to study the relationship between biochemical pathways and ADRs at a large scale. Using sparse canonical correlation analysis of 832 drugs with two profiles for 173 pathways and 1385 ADRs, a total of 30 correlated sets of pathways and ADRs were extracted. To evaluate the performance of the method, the extracted correlated components were used to identify known ADR profiles from drug pathway profiles using a 5-fold cross validation. A relatively high prediction performance (AUC: 0.894) was achieved. To have a global view of pathway-ADR associations, we represented all the components through a network, in which pathways and ADRs are connected if they are found in the same correlated set. We found that nearly half of ADRs were associated with only a few biochemical pathways. To further investigate the association between pathways and ADRs, we developed a network-based approach, in which the association between a pathway and an ADR was estimated by using the Jaccard similarity coefficient. A network consisting of 160 pathways, 564 ADRs and 1744 weighted edges was constructed and 5 statistically significant modules were extracted. We found that most of the nodes contained in the modules are either pathways linked to a very limited number of drugs or rare ADRs. At one level, this corroborates the limitations of available knowledge about ADRs and drug action mechanisms. But it also highlights the opportunities for improving such knowledge through systematic, network-based prediction approaches.

To assess global prediction performance across all ADRs, we followed the approach adopted by Pauwels et al. [26], i.e.

drawing a global ROC based on combining the prediction scores for all ADRs. It is worth noting that some ADRs are observed in only a few drugs. For example, it has been observed that the ADR acanthosis nigricans (C0000889) is only associated with one drug in the dataset, i.e. nicotinic acid used to treat high levels of cholesterol and triglycerides. The resulting imbalanced dataset may have a significant impact on the estimation of the prediction performance, which would be an important part of our future research. Another limitation of our study is the constrained sample of drug-pathway associations, which is typically based on the notion of single drug-single target relationships. However, the resulting pathway-centric models go beyond this classical notion of drug-induced perturbation, and can expand our view of drug-target interactions. We will further explore the drug-pathway associations in a dynamic cellular systems, though such data is currently still a limitation. In comparison to other computational approaches such as support vector machines and *k*-nearest neighbours [7], the SCCA-based approach used in this study has a clear advantage in terms of interpretability of results [26]. Nevertheless, the comparison with other related methods, such as those introduced in [5], and the examination of the potential clinical implications of novel extracted associations between KEGG pathways and ADRs deserve further investigation. One possible approach to investigate this in a prospective way is to build prediction models based on information available until a particular year, EndYear, followed by an independent validation on data generated after EndYear.

In summary, our method and resulting datasets will aid in: a) the systematic prediction of ADRs, and b) the characterization of novel mechanisms of action for existing drugs. The predictions that our method generate are both testable and biologically interpretable. We believe that the combination of these advantages into a single prediction strategy opens new research opportunities for improving personalised drug safety monitoring, as well as for the repositioning of drugs in the long-term.

ACKNOWLEDGMENT

We thank Jingchun Sun at the University of Texas for his constructive comments on the paper

REFERENCES

[1] World Health Organisation , “International drug monitoring-the role of the hospital-A WHO report,” Drug Intell Clin Pharm 4, 1970, pp.101-110.

[2] G. Onder, C. Pedone, F. Landi, M. Cesari, C. Della Vedova, R. Bernabei, G. Gambassi, “Adverse drug reactions as cause of hospital admissions: results from the Italian Group of Pharmacoepidemiology in the Elderly (GIFA),” J Am Geriatr Soc. 2002, 50(12), pp.1962-1968.

[3] E.J. Phillips, S. A. Mallal, “Pharmacogenetics of drug hypersensitivity,” Pharmacogenomics. 2010,11(7):973-987.

[4] M, Pirmohamed, S, James, S. Meakin, C. Green, A.K.Scott et al. “Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients,” BMJ 329, 2004, pp. 15-19.

[5] A. Bender, J. Scheiber, M. Glick, J.W. Davies, K. Azzaoui, J. Hamon, et al., “Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure,” ChemMedChem. 2007, 2(6), pp. 861-873.

[6] A. Cami, A. Arnold, S. Manzi, and B. Reis, “Predicting adverse drug events using pharmacological network models,” Sci Transl Med, 2011, 21;3 (114):114ra127.

[7] M. Liu, Y. Wu, Y. Chen, J. Sun, Z. Zhao, X.W. Chen, et al. “Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs,” J Am Med Inform Assoc. 2011, 19(e1), pp.e28-35.

[8] R. Harpaz, S. Vilar, W. Dumouchel, H. Salmasian, K. Haerian, N. H Shah, et al. “Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions,” J Am Med Inform Assoc. 2013, 20, pp.413-419.

[9] M. Liu, R. Cai, Y. Hu, M.E. Matheny, J. Sun, J. Hu, H. Xu, “Determining molecular predictors of adverse drug reactions with causality analysis based on structure learning,” J Am Med Inform Assoc., 2014, 21(2), pp.245-251

[10] I. Wallach, N. Jaitly, and R. Lilien, “A Structure-Based Approach for Mapping Adverse Drug Reactions to the Perturbation of Underlying Biological Pathways,” 2010, PLoS ONE 5(8): e12063.

[11] Y. Silberberg, A. Gottlieb, M. Kupiec, E. Ruppin, R. Sharan, “Large-Scale Elucidation of Drug Response Pathways in Humans,” Journal of Computational Biology, 2012, vol. 19, no. 2. pp. 163-174.

[12] X. Chen, J. Xu, B. Huang, J. Li, X. Wu, L. Ma, et al. “A sub-pathway-based approach for identifying drug response principal network,” Bioinformatics, 2011, 27(5), pp.649-654.

[13] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, “Data, information, knowledge and principle: back to metabolism in KEGG,” Nucleic Acids Res., 2014, 42, D199–D205.

[14] H. Zheng, H. Wang, H. Xu, Z. Zhao, F. Azuaje. “Correlating adverse drug reactions with biochemical pathways in humans,” Proceedings of 2013 IEEE International Conference on Bioinformatics and Biomedicine, Shanghai, China, pp. 197-200, 2013.

[15] L. Xie, J. Li, L. Xie, P.E. Bourne, “Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors,” PLoS Comput Biol. 5(5):e1000387, 2009. doi: 10.1371/journal.pcbi.1000387

[16] M. Kuhn, M. Campillos, I. Letunic, et al., “A side effect resource to capture phenotypic effects of drugs,” Mol Syst Biol., 2010,6:343.

[17] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, et al., “DrugBank 3.0: a comprehensive resource for 'omics' research on drugs,” Nucleic Acids Res. 2011,39(Database issue), pp.D1035-1041.

[18] M. Kanehisa,S. Goto, M. Furumichi et al, “KEGG for representation and analysis of molecular networks involving diseases and drugs,” Nucleic Acids Res 2010;38 (Database issue):D355e60.

[19] M. Kanehisa, S. Goto, M. Hattori et al. “From genomics to chemical genomics: new developments in KEGG,” Nucleic Acids Res 2006;34(Database issue):D354e7.

[20] H. Hotelling, “Relations between two sets of variates,” Biometrika, 1936, 28, pp.321– 377.

[21] D. Witten, R. Tibshirani and T. Hastie, “A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis,” Biostatistics, 2009, 10(3), pp. 515-534.

[22] D. Witten, R. Tibshirani, S. Gross and B. Narasimhan, “PMA: Penalized Multivariate Analysis,” 2013, R package version 1.0.9. http://CRAN.R-project.org/package=PMA

[23] Y. Yamanishi, E. Pauwels, H. Saigo, V. Stoven, “Extracting Sets of Chemical Substructures and Protein Domains Governing Drug-Target Interactions,” Journal of Chemical Information and Modeling, 2011, 51, pp.1183-1194.

[24] R Core Team, “R: A language and environment for statistical computing,” R Foundation for Statistical Computing, Vienna, Austria, 2013, URL http://www.R-project.org/.

[25] S. Mizutani, E. Pauwels, V. Stoven, S. Goto, and Y. Yamanishi, “Relating drug–protein interaction network with drug side effects,” Bioinformatics, 28 (18), pp. i522-i528, 2012

[26] E. Pauwels, V. Stoven, and Y. Yamanishi, “Predicting drug side-effect profiles: a chemical fragment-based approach,” BMC Bioinformatics, 12:169, 2011.

[27] T. Nepusz, H. Yu, and A. Paccanaro, “Detecting overlapping protein complexes in protein-protein interaction networks,” Nature Methods, vol. 9, pp. 471-472, 2012

D. Chen, M. Frezza, S. Schmitt, J. Kanwar, Q. P. Dou, “Bortezomib as the first proteasome inhibitor anticancer drug: current status and future perspectives,” Curr Cancer Drug Targets. 2011,11(3), pp.239-253