A Framework of Knowledge Integration and Discovery for Supporting Pharmacogenomics Target Predication of Adverse Drug Events: A Case Study of Drug-Induced Long QT Syndrome

Guoqian Jiang, MD, PhD, Chen Wang, PhD, Qian Zhu, PhD,
Christopher G. Chute, MD, Dr.PH
Department of Health Science Research, Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN

Abstract

Knowledge-driven text mining is becoming an important research area for identifying pharmacogenomics target genes. However, few of such studies have been focused on the pharmacogenomics targets of adverse drug events (ADEs). The objective of the present study is to build a framework of knowledge integration and discovery that aims to support pharmacogenomics target predication of ADEs. We integrate a semantically annotated literature corpus Semantic MEDLINE with a semantically coded ADE knowledgebase known as ADEpedia using a semantic web based framework. We developed a knowledge discovery approach combining a network analysis of a protein-protein interaction (PPI) network and a gene functional classification approach. We performed a case study of drug-induced long QT syndrome for demonstrating the usefulness of the framework in predicting potential pharmacogenomics targets of ADEs.

1 Introduction

Adverse drug events (ADEs) have been well recognized as a cause of patient morbidity and increased health care costs in the United States. With rapid developments in human genomics, the genetic component of ADEs is being considered as one of significant contribution factors for drug response variability and drug toxicity, thus representing a major component of the movement to pharmacogenomics and individualized medicine [1-2].

Text mining of published literature resources such as MEDLINE for identifying pharmacogenomics target genes and/or pathways is considered as an important research area that is complementary to the human-based curation approach as used in the PharmGKB [3-5]. In particular, knowledge-driven text mining that leverages existing pharmacogenomics knowledge is a promising direction. For example, Pakhomov, et al used PharmGKB to train text mining approaches for identifying potential gene targets for pharmacogenomic studies, demonstrating the capability of finding new gene targets [6]. Xu, et al developed a knowledge-driven conditional approach to extract pharmacogenomics specific drug-gene relationships from MEDLINE abstracts [7]. In these studies, the information/knowledge extraction is mainly focused on the binary relations between drugs and their gene targets. In addition, few of such studies have been focused on the pharmacogenomics targets of ADEs. We hypothesize that it would be helpful to use a ternary relation domain model among drugs, ADEs and their associated gene targets for guiding the knowledge integration and discovery.

The objective of the present study is to build a framework of knowledge integration and discovery that aims to support pharmacogenomics target predication of ADEs. Specifically, we leverage a semantically coded ADE knowledgebase known as ADEpedia [8] and a semantically annotated literature corpus Semantic MEDLINE [9] and integrate them in a semantic web based framework. The ADEpedia (http://adepedia.org), developed in our previous and ongoing studies, is a standardized knowledgebase of ADEs that intends to integrate existing known ADE knowledge for drug safety surveillance from disparate resources such as the FDA Structured Product Labeling (SPL), the FDA Adverse Event Reporting System (AERS) and the Unified Medical Language System (UMLS). Semantic MEDLINE is a recent development by the National Library of Medicine that integrates document retrieval, advanced natural language processing (NLP), and automatic summarization and visualization to support more effective biomedical information management [9]. Semantic MEDLINE identifies genes noted in biomedical text as associated with a disease process and can potentially simplify secondary database curation [10].

Based on the integration, we first retrieve the genetic associations of drugs and ADEs using SPARQL-based semantic query services. We then develop a knowledge discovery model for predicting potential pharmacogenemics targets of an ADE. To demonstrate the usefulness of the framework, we perform a case study on long QT syndrome induced by tricyclic antidepressive agents. Long QT syndrome is a heart condition in which delayed repolarization of the heart following a heartbeat causes prolongation of the QT interval, and increases the risk of torsades de pointes, ventricular fibrillation and sudden cardiac death. Drug induced QT prolongation, is an increasing public

health problem [11]. While many of the drugs known to prolong the QT interval were antiarryhythmics (e.g. quinidine), many non-cardiac drugs such as tricyclic antidepressants have also been reported to cause QT prolongation. At the cellular level, the blockade of rapid outward potassium current by these drugs is responsible for their pro-arrhythmic effect.

2 Materials and Methods

2.1 Materials

2.1.1 Semantic MEDLINE in RDF graphs

In our previous study [12], we have converted the Semantic MEDLINE in a relational database into six RDF graphs using a Semantic Web RDF transformation tool called D2R server (http://d2rq.org/d2r-server). RDF is a W3C standard that specifies a graph-based data model to represent Semantic Web data that enables powerful data integration of heterogeneous data sets (http://www.w3.org/TR/2004/REC-rdf-mt-20040210/). In the present study, we utilized two of the six RDF graphs: the disease-gene graph and the drug-gene graph.

2.1.2 ADEpedia: A Standardized Knowledgebase of ADEs

As mentioned above, the ADEpedia intends to integrate existing known ADE knowledge from disparate resources to achieve a comprehensive ADE knowledgebase [8,13]. In the ADEpedia, the drugs and the ADEs are normalized using the UMLS Concept Unique Identifiers (CUIs). In the present study, we represent the normalized drug-ADE knowledge from the ADEpedia in RDF data model for the integration.

2.1.3 Human Protein Reference Database

Protein-protein interaction (PPI) information comprising 9,303 proteins and 35,000 protein-protein interactions is collected from the Human Protein Reference Database (HPRD) [14], which contains manually curated physical interactions among proteins. It has been known that proteins, the end product of genes, usually perform molecular function as a group, by physically interacting with each other. This interaction relationship is important to the Drug-ADE gene context, since it may highlight interactions and related pathway that mediate occurrence of adverse-effect. As the alternations of protein interactions could contribute to diseases onset or progression, PPI has been used to investigate disease biomarkers [15], and could also have potential in pharmacogenomics study of ADEs.

2.2 Methods

2.2.1 Knowledge Representation and Integration Using a Semantic Web Based Approach

Figure 1 shows our system architecture of knowledge integration for pharmacogenomics knowledge discovery applications in a semantic web-based framework. In the Semantic Normalization layer, 1) we transform and represent the ADE knowledge in a RDF based data model; 2) we utilize two RDF graphs in a Semantic MEDLINE RDF store developed in our previous study. The two RDF graphs represent the domain patterns for the associations of disease-gene and drug-gene.

In the Semantic Integration Layer, 1) we extract the severe ADE knowledge from the ADEpedia based on the severity information of each ADE. The severe ADE knowledge base was developed using an approach described in a separate paper [16]. The severity definition is based on the Common Terminology Criteria for Adverse Events (CTCAE) 5-scale grading system [17]. We assert that the grade >=3 is considered as "Severe". 2) As we have normalized both the drugs and the ADEs using the UMLS CUIs, and the Semantic MEDLINE uses UMLS CUIs for the semantic annotations, so we use the CUIs as the anchor to extract the genetic association information of both drugs and ADEs for all severe ADEs from the RDF graphs of Semantic MEDLINE. 3) We integrate the genetic associations of severe ADEs into a separate RDF graph that is loaded into a 4store-based RDF store (http://4store.org). We build a SPARQL endpoint against the RDF store for providing standard semantic query services.

In the Knowledge Discovery Layer, the pharmacogenomics knowledge discovery applications can invoke the semantic query services through the SPARQL endpoint to extract their associated genes annotated in Semantic MEDLINE, the related PubMed IDs for target

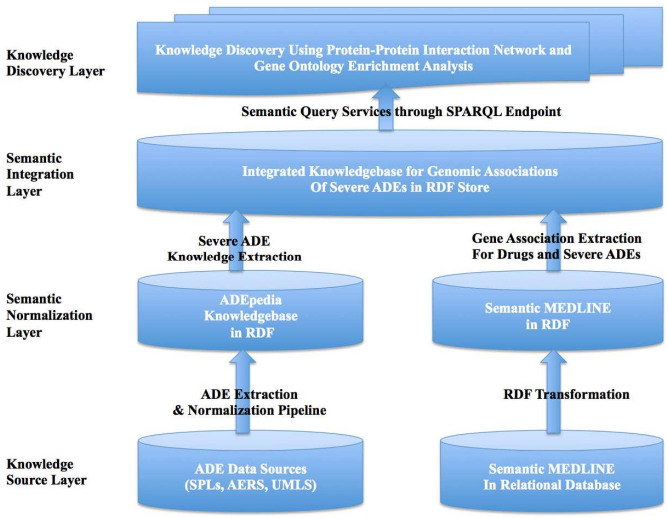

Figure 1. System architecture

drugs and severe ADEs. In this study, we developed a knowledge discovery utilizing the HPRD PPI network and Gene Functional Classification and Enrichment Analysis.

2.2.2 Pharmacogenomics Knowledge Discovery Using a Network Analysis Approach

To predicate potential pharmacogenomics targets of ADEs, we developed a network analysis approach utilizing the HPRD protein-protein interaction (PPI) network. The approach is described as follows.

Given a drug-gene list with *M* genes $D = \{d_i\}_{i=1,\cdots,M}$ and a ADE-gene list of *N* genes $A = \{a_j\}_{j=1,\cdots,N}$, our aim is to prioritize drug-gene closely related to ADE-gene list, and ADE-gene closely related to Drug-gene list, in the domain of PPI network. PPI network is defined as a graph $G = (V, E)$, where V is vertex set of genes, including drug-genes and ADE-genes: $d_i, a_j \in V$, and E is edge set indicating experimentally validated interactions between two genes.

The closeness of one ADE-gene to drug-gene list is defined as an average distance between this gene and all the genes in drug-gene list:

$$c(a_j, D) = \frac{1}{M}\sum_{i=1}^{M} dist_G(a_j, d_i),$$

where $dist_G(\cdot)$ is the shortest distance between two vertices on graph G. Similarly, we could define closeness of one drug-gene to ADE-gene list: $c(d_i, A) = \frac{1}{N}\sum_{j=1}^{N} dist_G(d_i, a_j)$.

In order to prioritize closely related genes, the next question is how to assess if observed closeness is statistically significant comparing to random cases. Take $c(a_j, D)$ as an example, we design a hypothesis-testing scheme by generating a large number (*P*=50,000 in this study) of false gene-lists $\left\{\tilde{D}_p\right\}_{p=1,\cdots,P}$ with same size of true drug-gene list D, and then compute p-value by following equation:

$$\text{p-value}(a_j) = \frac{\text{number of c}(a_j, D) \le \text{c}(a_j, \tilde{D}_p)}{P}.$$

This p-value is more comparable across different genes than $c(a_j, D)$, since the relative topological importance of each gene is also controlled through random sampling.

To explore the functional groups of the prioritized drug- and ADE- genes, we performed Gene Functional Classification and Enrichment Analysis using an online bioinformatics application known as DAVID developed by the National Institute of Allergy and Infectious Diseases (NIAID), NIH [18]. For the validation of prioritized target genes, we also manually reviewed the relevant PubMed abstracts.

3 A Case Study: Long QT Syndrome Induced by Tricyclic Antidepressants

We retrieved 265 subclasses of the tricyclic antidepressants class represented by the UMLS CUI "Antidepressive Agents, Tricyclic|C0003290". Using a SPARQL query against the knowledgebase in a RDF store, we extracted 218 records for the target drug class "Antidepressive Agents, Tricyclic|C0023976", covering 77 unique genes associated with the drug class and its 15 descendant drugs (including Amitriptyline, Clomipramine, Desipramine, and Fluoxetine, etc.). For the target ADE "Long QT Syndrome| C0023976", we extracted 205 records, covering 11 unique gene IDs that are associated with the disorder.

For the use of HPRD PPI network, we converted the gene IDs associated with the drug and the ADE into the HPRD IDs using an online ID conversion application called bioDBNet:db2db [19]. As a small portion of gene IDs does not have corresponding HPRD IDs, we got 71 HPRD IDs for those 77 drug genes and 11 HPRD IDs for those 11 ADE genes.

Table 1 shows the network analysis results, generating a list of genes (comprising 15 drug-genes and 8 ADE genes) that are statistically significant ($p<0.05$) based on PPI network analysis. The results indicate that the significant drug-genes (or ADE-genes) are more closely related to the ADE-genes (or drug-genes) and should be prioritized for further consideration in predicating pharmacogenomics study.

We explored the functional groups of the prioritized genes using a gene functional classification tool. We identified two functional gene clusters based on the GO enrichment score (see Table 2). We found that the first cluster contains both drug-associated genes and ADE-associated genes whereas the second cluster contains the

genes only from those drug-associated genes. We consider that the first cluster implies a shared genetic mechanism between the drugs from the drug class tricyclic antidepressive agents and the ADE long QT syndrome. The GO enrichment analysis of the genes in the first cluster indicates that the genes are functionally enriched in "metal ion transport", "potassium ion transport", and "regulation of heart contraction", etc.

Table 1. The lists of prioritized drug- and ADE- genes that are statistically significant ($p<0.05$) based on PPI network analysis for the use case "Long QT Syndrome induced by Tricyclic Antidepressive Agents".

Prioritized Drug-Associated Genes (Antidepressive Agents, Tricyclic|C0023976)				**Prioritized ADE-Associated Genes** (Long QT Syndrome|C0023976)			
Gene ID	**Gene Symbol**	**Average Distance**	**p-value**	**Gene ID**	**Gene Symbol**	**Average Distance**	**p-value**
146	ADRA1D	3.125	0.0000	6331	SCN5A	3.4909	0.0001
3757	KCNH2	2.125	0.0000	3753	KCNE1	3.6	0.0019
22953	P2RX2	3.625	0.0009	3751	KCND2	3.8364	0.003
2890	GRIA1	2.625	0.0011	9992	KCNE2	4.4727	0.0083
28954	REM1	3	0.0034	50488	MINK1	3.4909	0.014
3358	HTR2C	3.375	0.0043	6640	SNTA1	3.3636	0.02
928	CD9	3	0.0059	3757	KCNH2	3.6	0.0203
41	ACCN2	3.25	0.0072	3784	KCNQ1	4	0.022
2904	GRIN2B	2.75	0.0136				
3777	KCNK3	3	0.014				
760	CA2	3.375	0.02				
3763	KCNJ6	3.625	0.0211				
3350	HTR1A	4	0.0318				
3351	HTR1B	5	0.032				
3356	HTR2A	3.25	0.0452				

Table 2. The results of gene functional classification of prioritized drug- and ADE- genes.

Gene Cluster	**Gene ID**	**Gene Symbol**	**Gene Name**	**Category**
Gene Cluster 1 - Enrichment Score: 7.75238507542537	3751	KCND2	potassium voltage-gated channel, Shal-related subfamily, member 2	ADE-Gene
	3753	KCNE1	potassium voltage-gated channel, Isk-related family, member 1	ADE-Gene
	9992	KCNE2	potassium voltage-gated channel, Isk-related family, member 2	ADE-Gene
	3777	**KCNK3**	**potassium channel, subfamily K, member 3**	**Drug-Gene**
	3784	KCNQ1	potassium voltage-gated channel, KQT-like subfamily, member 1	ADE-Gene
	41	**ACCN2**	**amiloride-sensitive cation channel 2, neuronal**	**Drug-Gene**
	6331	SCN5A	sodium channel, voltage-gated, type V, alpha subunit	ADE-Gene
	3763	**KCNJ6**	**potassium inwardly-rectifying channel, subfamily J, member 6**	**Drug-Gene**
	3757	**KCNH2**	**potassium voltage-gated channel, subfamily H (eag-related), member 2**	ADE-Gene **Drug-Gene**
Gene Cluster 2 - Enrichment Score: 4.571038563001669	3350	HTR1A	5-hydroxytryptamine (serotonin) receptor 1A	Drug-Gene
	3351	HTR1B	5-hydroxytryptamine (serotonin) receptor 1B	Drug-Gene
	3358	HTR2C	5-hydroxytryptamine (serotonin) receptor 2C	Drug-Gene
	3356	HTR2A	5-hydroxytryptamine (serotonin) receptor 2A	Drug-Gene
	146	ADRA1D	adrenergic, alpha-1D-, receptor	Drug-Gene

We retrieved the PubMed IDs linked with the drug genes (KCNK3, ACCN2, KCNJ6 and KCNH2) in the first cluster and manually reviewed all the original abstracts. Among 16 PubMed IDs retrieved, 12 abstracts are true positive (75%), i.e. correctly reflecting the association between a tricyclic antidepressant and a target gene. Of the 12 abstracts, 8 abstracts (linked with KCNH2) mentioned of the target drug genes that are related to long QT syndrome whereas 4 abstracts (linked with ACCN2 and KCNJ6) did not mention of. The results indicate that KCNH2 is a well-studied gene across the target drug and the target ADE while ACCN2 and KCNJ6 are potential candidates for the pharmacogenomics-target predication.

4 Discussions and Concluding Remarks

In this study, we have built a framework of knowledge integration and discovery for supporting pharmacogenomics-target predication of ADEs. We utilized a ternary relation domain model to guide the knowledge integration for genetic associations of severe ADEs. In the current prototype implementation, we successfully integrated a semantically annotated literature corpus Semantic MEDLINE with a normalized ADE knowledgebase ADEpedia using a semantic web-based data integration approach. The semantic web-based approach is increasingly used for drug-related data integration studies. For instance, Bio2RDF (http://bio2rdf.org/) integrated a number of life science datasets (including DrugBank - http://www.drugbank.ca/) using a semantic web linked open data approach for biological knowledge discovery. For another instance, there is an effort to convert AERS dataset into RDF-based linked data (http://aers.data2semantics.org/). Comparing with these efforts, our ADEpedia project mainly focused on the standardization of ADE knowledge using standard drug and ADE terminologies (e.g., RxNorm, MedDRA, SNOMED CT and UMLS). We found that the normalization of the drugs and the ADEs using the UMLS is extremely important for both data integration and aggregation. For example, the UMLS enables us to retrieve all the descendants of the drug class "Antidepressive Agents, Tricyclic|C0003290", which provided the aggregation power for collecting genetic associations of the drug class. In addition, we added a module in our framework to extract the severe ADE knowledge since the clinical applications of pharmacogenomics on ADEs are usually focused on the clinically severe ADEs. However, we found that identifying the severity information of an ADE remains a challenging task. We are exploring a semi-automatic approach for identifying the severity information of ADEs utilizing a number of existing knowledge resources such as CTCAE, FDA structured product labels and FDA AERS reporting data [16].

For the knowledge discovery, we have developed a network analysis-based approach to prioritize target genes related to ADEs. As gene lists derived from text mining could be very general and contain many false-positives, we proposed to apply network analysis to filter out less-relevant genes through additional evidence. The prioritization of closely related genes could shed some lights on potential mechanism regarding how drug- and ADE-associated genes affect each other. In addition, the gene functional classification and enrichment analysis provided further evidence for the prioritized genes identified from the network analysis, which, we believe, was validated by a manual review of related PubMed abstracts. In the future, we will explore how to represent and aggregate all these evidence from various knowledge resources in order to improve the performance of the knowledge discovery model for predicting the pharmacogenomics-targets of severe ADEs.

Acknowledgements: This work was supported in part by the Pharmacogenomic Research Network (NIH/NIGMS - U19 GM61388) and the SHARP Area 4: Secondary Use of EHR Data (90TR000201).

References

[1] Wang L. Pharmacogenomics: a systems approach. Wiley Interdiscip Rev Syst Biol Med. 2010 Jan-Feb;2(1):3-22. Review.
[2] Ross CJ, Visscher H, Sistonen J, Brunham LR, Pussegoda K, Loo TT, Rieder MJ, Koren G, Carleton BC, Hayden MR; CPNDS Consortium. The Canadian Pharmacogenomics Network for Drug Safety: a model for safety pharmacology. Thyroid. 2010 Jul;20(7):681-7. Review.
[3] Thorn CF, Klein TE, Altman RB. Pharmacogenomics and bioinformatics: PharmGKB. Pharmacogenomics. 2010 Apr;11(4):501-5.
[4] Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, Altman RB, Klein TE. Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther. 2012 Oct;92(4):414-7. doi: 10.1038/clpt.2012.96.
[5] Garten Y, Coulet A, Altman R. Recent progress in automatically extracting information from the pharmacogenomic literature. Pharmacogenomics, 11 (2010), pp. 1467–1489
[6] Pakhomov S, McInnes BT, Lamba J, Liu Y, Melton GB, Ghodke Y, Bhise N, Lamba V, Birnbaum AK. Using PharmGKB to train text mining approaches for identifying potential gene targets for pharmacogenomic studies.J Biomed Inform. 2012 Oct;45(5):862-9. Epub 2012 May 4.
[7] Xu R, Wang Q. A knowledge-driven conditional approach to extract pharmacogenomics specific drug-gene relationships from free text.J Biomed Inform. 2012 Oct;45(5):827-34. Epub 2012 Apr 27.
[8] Jiang G, Solbrig HR, Chute CG. ADEpedia: a scalable and standardized knowledge base of Adverse Drug Events using semantic web technology. AMIA Annu Symp Proc. 2011;2011:607-16. Epub 2011 Oct 22.
[9] Rindflesch, T.C., Kilicoglu, H., Fiszman, M., Rosemblat, G., Shin, D. Semantic MEDLINE: An advanced information management application for biomedicine. Information Services and Use, Volume 31, Issue 1-2, 2011, Pages 15-21.
[10] Workman TE, Fiszman M, Hurdle JF, Rindflesch TC. Biomedical text summarization to support genetic database curation: using Semantic MEDLINE to create a secondary database of genetic information. J Med Libr Assoc. 2010 Oct;98(4):273-81.
[11] Yap YG, Camm AJ. Drug induced QT prolongation and torsades de pointes. Heart. 2003 Nov;89(11):1363-72. Review.
[12] Tao C, Zhang Y, Jiang G, Chute CG. Optimizing Semantic MEDLINE for translational science studies using semantic web technologies. Proceedings of MIXHS'2012 Workshop, October 29, 2012. (In press)
[13] Jiang G, Liu H, Solbrig HR, Chute CG. ADEpedia 2.0: Integration of Normalized Adverse Drug Events (ADEs) Knowledge from the UMLS. AMIA CRI Summit 2013. (In press).
[14] Keshava Prasad, T. S., R. Goel, et al. (2009). "Human Protein Reference Database--2009 update." Nucleic Acids Res .37(Database issue): D767-772.
[15] Chen Wang, Sook Ha, et al. (2012). "Computational analysis of muscular dystrophy sub-types using a novel integrative scheme". Neurocomputing Volume 92, September, 2012, Pages 9-17.
[16] Jiang G, Wang L, Liu HF, Chute CG. Building a knowledge base of severe adverse drug events based on AERS reporting data using semantic web technologies. MedInfo 2013. (In submission)
[17] CTCAE URL: http://ctep.cancer.gov/protocolDevelopment/electronic_applications/ctc.htm; last visited at September 24, 2012.
[18] DAVID Online Application: http://david.abcc.ncifcrf.gov/home.jsp; last visited at September 24, 2012.
[19] BioDBNet: db2db conversion application : http://biodbnet.abcc.ncifcrf.gov/db/db2db.php; last visited at September 24, 2012.