Scientometrics (2021) 126:3723–3748
https://doi.org/10.1007/s11192-021-03892-4

Revealing potential drug-disease-gene association patterns for precision medicine

Xuefeng Wang[1] · Shuo Zhang[1] · Yao Wu[1] · Xuemei Yang[2]

Received: 4 December 2019 / Accepted: 2 February 2021 / Published online: 6 March 2021

Abstract

Precision medicine means giving patients the right treatment at the right dose at the right time with minimum ill consequences and maximum efficacy. It is medicine personalized to the individual's genes, environment, and lifestyle and, ultimately, its widespread use will require a deep understanding of the genomic variations that create predispositions or resistances to various diseases. Some of the links between genes and diseases are already known, and more are being discovered every day. Similarly, much is known about which drugs are efficacious for treating which diseases, but there is still more to learn. The issue now is how to extract this information from the biomedical literature in way that can keep pace with today's rapid discoveries in medical research. Efforts to assemble an organized database of such knowledge to data have focused on mathematical statistic methods, computer-aided methods, etc. Success has been mixed as previous methods usually result in false positive or depend on training sample sets, lacking of generality in different research fields, which have choked advancements in precision medicine. To break through this bottleneck, we need novel methods that can extract and leverage the valuable information locked within the constraints of the data we have. Hence, in this paper, we present a new text-based computational framework for extracting full three-way drug-disease-gene triplet information related to colorectal cancer from biomedical texts. The framework consists of two main steps. The first is to construct an integrated drug-disease-gene network by extracting pair-wise associations between diseases, drugs, and genes, and then store unique drug-disease-gene triplets for further analysis. Since the constructed network is highly likely to be too sparse, the next step is to complete the incomplete links in the network, i.e., to predict novel links from genes to diseases to drugs. To validate our framework, we conducted a case study on colorectal cancer, mining the literature for drug-disease and disease-gene associations. An analysis of the subsequent inferences drawn between the two shows that this approach can help to inform novel research hypotheses and identify new knowledge triplets about various diseases, both of which are significant for the advancement and implementation of precision medicine.

Keywords Precision medicine · Subject-action-object (SAO) · Drug-disease-gene linkage · Association rules

✉ Xuefeng Wang
wxf5122@bit.edu.cn

Extended author information available on the last page of the article

Introduction

The concept of precision medicine is to provide prevention and treatment strategies that take the individual into account—their genes, their environment, their lifestyle. However, advancing the field of precision medicine depends on establishing tools and frameworks for regulating, compiling, and interpreting the influx of information, and at a pace that can keep up with rapid scientific developments. These frameworks might be drug discovery systems, gene sequencing techniques, health care devices, etc. (Mirnezami et al. 2012). Currently, research into precision medicine is proceeding on two main frontiers: a near-term focus on cancers and a longer-term aim to generate knowledge that applies to a whole range of diseases and health issues (Collins and Varmus 2015).

As we know, cancers are fast becoming the world's leading cause of death. Researchers have already revealed many of the molecular lesions that can cause cancer, showing that each kind of cancer has its own genomic signature. Although cancers are largely a consequence of accumulating genomic damage over one's life, inherited genetic variations and epigenetics variations do contribute to cancer risk—sometimes profoundly (Egger et al. 2004; Cheung and Liu 2009). Hence, recent findings from oncogenic mechanisms have begun to influence cancer risk assessments, diagnostic categories, and therapeutic strategies, with the increasing use of drugs and antibodies designed to counter the influence of specific molecular drivers. A recent study, using a panel of commonly implicated genes, suggested that a genomic alteration could be identified in 96% of undiagnosed primary tumors. And, in 85% of those cases, the tumor was potentially treatable by a known drug (Ross et al. 2015). Studies such as this demonstrate that comprehensive association patterns do exist between drugs, diseases, and genes, and, if these drug-disease-gene patterns could be discovered and profiled, we may be able to identify novel treatment paradigms for genetic-based diseases, especially cancers. These are the types of advancements needed to promote the development of precision medicine.

With the explosion in biomedical texts, much scholarly effort has been expended in developing approaches to mine the relationships discovered between biomedical entities, e.g., drugs to treat diseases, genes linked to proteins. These associations are scattered across the literature and, while not always easy to find and extract, they are a valuable source of supplementary data for domain knowledge discovery. Moreover, the ability to systematically analyze the heterogeneous data, would provide biomedical researchers with unprecedented opportunities to infer novel associations among different biomedical entities in the context of precision medicine and translational research studies. The majority of the current approaches focus on relationships between only two kinds of entites, such as drug-drug interactions (Duke et al. 2012; Bui et al. 2014), protein–protein interactions (Mason and Verwoerd 2007), protein-gene relations (Fundel et al. 2007), disease candidate genes (Hristovski et al. 2005, Ozgur et al. 2008), and drug repositioning (Christos et al. 2011). Usually these analysis methods can be divided into two types: mathematical statistic method and computer-aided method. Almost all of the mathematical statistic methods follow a similar paradigm, but their methods for identifying biomedical entities in a text and extracting the relationships between them are diverse. Another method depends on computer techniques, such as natural language processing (NLP), machine learning, deep learning, text mining and Bayesian statistics. The mathematical statistic method based on criteria like word co-occurrence or word frequency, which frequently results in false positives. Computer-aided approach is heavily reliant on a good training sample set, and most models cannot be generalized to different research fields. Additionally, both these approaches depend on existing datasets, and neither considers the semantic relationships between entities.

Given these shortcomings, what is needed now is a broad research program to encourage creative approaches to precision medicine with a focus on novel ways to extract effective domain knowledge from the plethora of data we have available. This knowledge gained, in the form of definitive relationships between biomedical entities, must then be used to build the evidence base needed to guide clinical practice. This is the goal of the second research frontier. Ultimately, precision medicine should ensure that patients get the right treatment at the right dose at the right time, with minimum ill consequences and maximum efficacy.

In this paper, we demonstrate how to fully integrate our prior knowledge on drugs, diseases, and genes, and then how to use that knowledge in a systematic framework to infer the incomplete links between them through association rules. To showcase the framework, we used it to analyze the biomedical literature for drug-disease-gene links associated with three diseases—ulcerative colitis, Chron’s disease, and ileitis, then verified our findings with a manual review of the relevant texts. The results show that the framework has the potential to: (1) identify potential disease relationships; (2) prioritize candidate disease genes; (3) predict novel options for drug repurposing; (4) provide insights that could help to formulate novel research hypotheses; and (5) identify new triplet associations for various diseases. Each of these contributions is significant to the implementation and advancement of precision medicine.

The rest of the paper is organized as follows: "Literature review" section introduces the related work. "Methods and data" section presents the research methodology and data sources. "Results" section contains the case analyses and results. "Conclusion" section concludes the paper with a discussion on the limitations of this study and opportunities for future works.

Literature review

Text-based knowledge discovery

Many genetic mutations predispose individuals to disease (Greenman et al. 2007). The practice of precision medicine involves identifying such mutations in patients and modifying patient treatments to reflect each person’s different physiology risks (Collins and Varmus 2015). Databases of drug-disease-gene relationships play an important role in this process by acting as a reference for providers to refer to determine the significance of their patient’s mutations. From this information, practitioners can prescribe the optimal drug to treat the individual (Ashley et al. 2010; Dewey et al. 2014). However, there are many more associations scattered across the biomedical literature that have not yet been included in these databases, and, as the pace of medical discoveries increases, it is becoming harder and harder to keep these databases up-to-date.

Hence, many researchers are turning to data analytics as a relationship mining tool. For instance, Ozgur et al. (2008) collected an initial set of known disease-related genes and introduced an automatic approach based on text mining and network analysis to predict gene-disease associations. They used the degree, eigenvector, betweenness, and closeness centrality metrics to rank the genes in the network, all based on the assumption that the central genes in that disease-specific network were likely to be related to the disease. Hu and Agarwal (2009) constructed a large-scale disease-drug network for drug repositioning as well as a drug target/pathway identification system based on disease and drug expression profiles using GEO datasets. Finally, they extracted 170,027 significant interactions from 7000 publicly-available transcriptomic profiles, including 645 disease-disease, 5008 disease-drug, and 164,374 drug-drug relationships. Zhou and Fu (2018) integrated the MeSH database with term weights and co-occurrence methods to predict gene-disease associations based on the cosine similarity between gene vectors and disease vectors. They evaluated the performance of cosine similarity in predicting the links between genes and disease by using the gene-disease association data in the OMIM database as golden standard. In the research of Roy et al. (2019), graph theory was utilized for quantitative analysis of the epigenetic network of hepato-cellular carcinoma (HCC). They evaluated the the essentiality of the node in the epigenetic network by using topological parameters like clustering coefficient, eccentricity, degree, etc. and the important vertices represented the genes involved in the epigenetic mechanism of HCC. To systematically analyze drug-disease-gene relationships, Simone et al. (2012) integrated data from structural and chemical datasets and created a drug-target-disease network for 147 promiscuous drugs, 553 protein targets, and 44 disease indications. The key contribution of their research is that novel links from drugs to targets and diseases can be predicted by completing incomplete bi-cliques.[1] Zhang et al. (2014) proposed a novel network-based method to identify statistically over-expressed subnetwork patterns (network motifs) in an integrated disease-drug-gene network extracted from Semantic MEDLINE. Out of the heterogeneous networks, they constructed association data on FDA-approved drugs and analyzed five significant network motifs. Sun et al. (2016) introduced a new data fusion model based on *n*-cluster editing as a novel multi-source triangulation strategy, which was further combined with semantic literature mining. They also confirmed that utilizing drug-disease-gene triangulation coupled with sophisticated text analysis is a robust approach for identifying new candidates for drug repurposing.

However, there are three common limitations with the above approaches. These are: (a) Word-based mathematical statistic approaches, such as word co-occurrence, frequently result in false positives because the semantic relationships between entities are not taken into consideration; (b) Most existing computer-aided methods for predicting causal disease genes rely on a specific type of evidence and are therefore limited in their applicability (Natarajan and Dhillon 2014); (c) None of the above approaches explicitly focus on extracting three-way relationships from texts, e.g., drug-disease-gene, for specific diseases. There is work that captures links from drugs to diseases or diseases to genes but not directly among all three. Additionally, studies by Simone et al. (2012), Zhang et al. (2014), and Sun et al. (2016) involve building integrated disease-drug-gene networks, but their analysis is still limited to a series of relationship pairs—drug-disease, drug-gene/protein, drug-drug associations, etc.

An investigation of all pair-wise plus three-way associations among these entities is necessary to understand the complexity of these interplays and to infer possible interactions within the context of the whole knowledge. Yet developing an efficient, robust, and flexible approach to extract a drug-disease-gene triplet from free text is still problematic for several reasons. First, correctly mining complex bio-entities from biomedical literature has been a long-standing challenge. Second, mining three-way relationships is obviously exponentially more complicated than mining two-way relationships (Singhal et al. 2016). Third, the

[1] A bi-clique is a network motif in which two sets of nodes all mutually interact with each other (Simone et al. 2012).

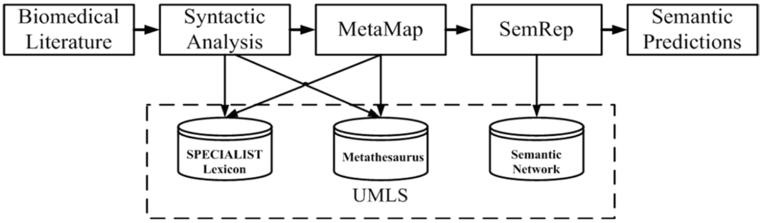


Fig. 1 General overview of SemRep to the extraction of semantic predictions

associations among different entities are typically very sparse, giving rise to cold-start and other problems (Zhang et al. 2014).

Natural language processing in the biomedical domain

Several research groups are developing and applying NLP methodologies in biomedical informatics. The complexity of natural language dictates that semantic interpretation be focused in scope, typically by the domain of discourse. The majority of this work is knowledge-based, and the specific domain guides the type and amount of knowledge used. Often this is drawn from existing resources, such as the Unified Medical Language System (UMLS), but several systems rely solely on locally-developed knowledge bases. One example is SemRep, which is a semantic interpreter that uses underspecified syntactic analysis and UMLS knowledge sources to provide a partial semantic interpretation of the biomedical research literature (Rindflesch et al 2000a, b, c). Specifically, UMLS consists of three modules: the Metathesaurus, Semantic Network,[2] and SPECIALIST Lexicon. The results of these text-driven assertions is a UMLS semantic network relationship, expressed as a subject-action /predicate/verb-object triple (SAO), in which the action is the relation (Rindflesch and Fiszman 2003). The subject and object arguments are drawn from the UMLS Metathesaurus, where each argument is assigned a semantic type according to its properties. This module comprises over 100 controlled vocabularies, such as MESH and SNOMED-CT. Combined with the UMLS Semantic Network, all concepts contained in Metathesaurus, including synonyms, are assigned a semantic type according to their properties, e.g., Clinical Drug (clnd), Disease or Syndrome (dsyn), and Gene or Genome (gngm). In addition, the UMLS Semantic Network contains a range of semantic relations, such as TREATS, PART OF, CANSES, among others. From Fig. 1, we can see that underspecified syntactic analysis relies on the UMLS SPECIALIST Lexicon. After input and tokenization, the text is submitted to an underspecified parser. Part-of-speech ambiguities are resolved with the Xerox part-of-speech tagger (Cutting et al. 1992) and a parser that identifies simple noun phrases, verbs, and appositives are selected from the text. MetaMap then maps these noun phrases to concepts in the UMLS Metathesaurus. To be interpreted as a semantic prediction, the semantic types of the UMLS Metathesaurus concepts that the syntactic arguments are mapped to must match the semantic types allowed in the Semantic Network. The semantic information defined in the UMLS can be further leveraged to

[2] UMLS Semantic Network has 54 semantic types and 133 semantic relations.

extract associations in specific domains and to identify domain patterns for specific studies through advanced computational methods. Researchers can also choose different semantic types to meet the specific needs of their research. For instance, for the purposes of this research, we selected clnd, dsyn, and gngm.

As an example to more clearly illustrate how SemRep works, consider the following sentences:

1. Association between the interleukin 23 receptor and ankylosing spondylitis is confirmed by a new UK case–control study and meta-analysis of published series (Karaderi et al. 2009);
2. One is SNP rs11209026 in exon 9 of IL23R for association with Crohn's disease, which is predicted to be probably damaging by PolyPhen2 (Huang et al. 2012);

From these sentences, SemRep suggested the following disease-gene relations:

1. C1537403|IL23R gene|gngm, aapp|gngm|ASSOCIATED_WITH|C0038013|Ankylosing spondylitis|dsyn|dsyn||.
2. C1537403|IL23R gene|gngm, aapp|gngm|ASSOCIATED_WITH|C0010346|Crohn Disease|dsyn|dsyn||.

Our proposal for a novel computational framework to extract drug-disease-gene triplets from biomedical literature leverages SemRep's semantic predictions of drug-disease and disease-gene but goes a step further to complete the incomplete links between these pair-wise relationships; the end results are drug-disease-gene triplets. Notably, with this approach, false positives and non-universal cease to be a problem.

In summary, the main contributions of this work are as follows:

1. A novel computational framework for extracting full three-way drug-disease-gene triplet information from biomedical texts.
2. Predictions of novel links from drugs to diseases and genes based on completing the incomplete links in network from potential associations between diseases.
3. A corpus containing 11,889 drug-disease-gene triplets related to colorectal cancer with their corresponding CUIs. The corpus may be used by relevant researchers, providing new ideas for future researches.

Methods and data

Data

Noncommunicable diseases (NCDs) are now responsible for the majority of global deaths, and cancer is expected to rank as the leading cause of death and the single most important barrier to increasing life expectancy in every country of the world in the twenty-first century. Bray et al.'s (2018) status report on the global burden of cancer forecast an estimated 18.1 million new cancer cases and 9.6 million cancer deaths in 2018 (17.0 m/9.5 m excluding nonmelanoma skin cancer). For both sexes combined, lung cancer is the most commonly diagnosed cancer, accounting for the most cancer deaths at 18.4%. This incidence of cancer is closely followed by female breast cancer (11.6%), colorectal cancer (10.2%), and prostate cancer (7.1%)

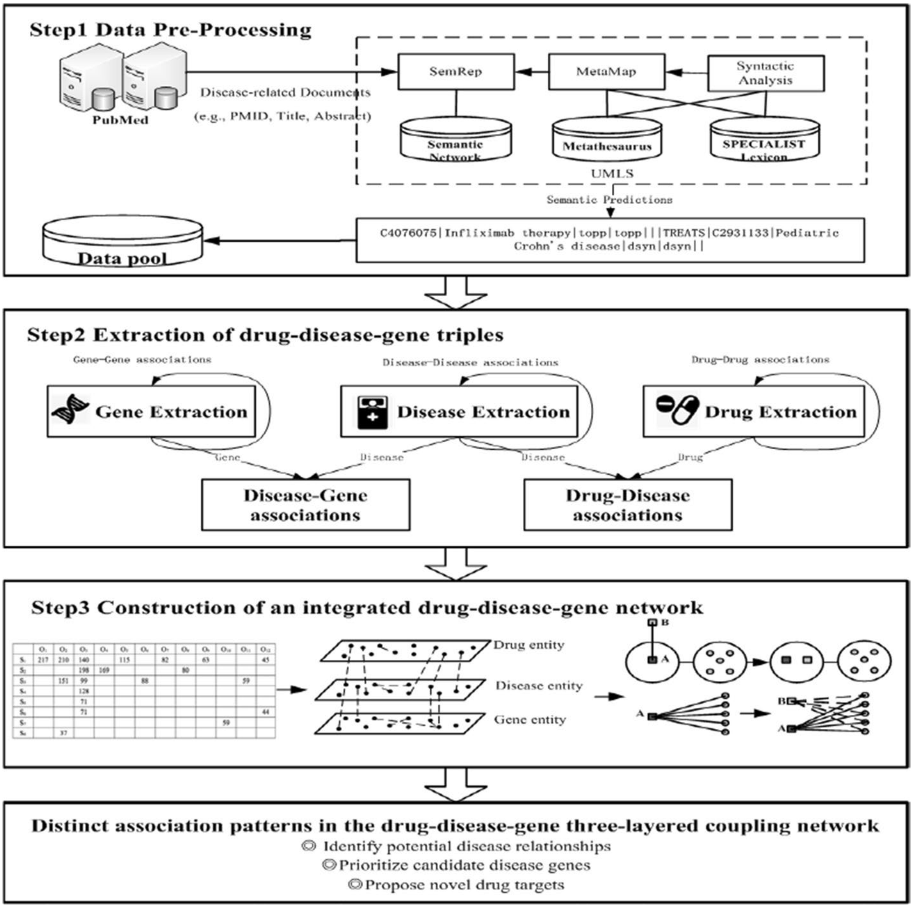
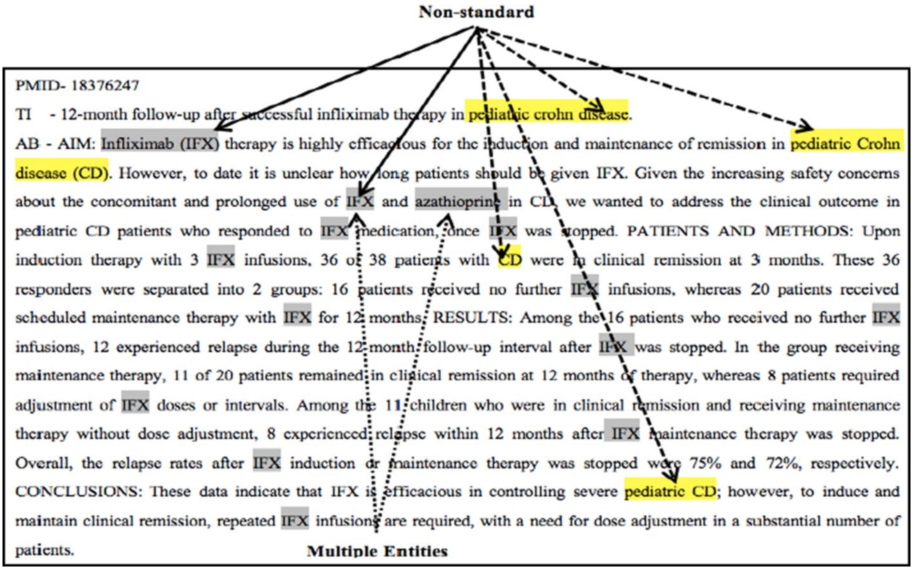

Step1 Data Pre-Processing
PubMed
Disease-related Documents
(e.g., PMID, Title, Abstract)
SemRep
MetaMap
Syntactic Analysis
Semantic Network
Metathesaurus
SPECIALIST Lexicon
UMLS
Semantic Predictions
C4076075|Infliximab therapy|topp|topp|||TREATS|C2931133|Pediatric Crohn's disease|dsyn|dsyn||
Data pool
Step2 Extraction of drug-disease-gene triples
Gene-Gene associations
Disease-Disease associations
Drug-Drug associations
Gene Extraction
Disease Extraction
Drug Extraction
Gene
Disease
Disease
Drug
Disease-Gene associations
Drug-Disease associations
Step3 Construction of an integrated drug-disease-gene network
Drug entity
Disease entity
Gene entity
A
B
Distinct association patterns in the drug-disease-gene three-layered coupling network
◎ Identify potential disease relationships
◎ Prioritize candidate disease genes
◎ Propose novel drug targets

Fig. 2 Framework for the research

and, for death, by colorectal cancer (9.2%), stomach cancer (8.2%), and liver cancer (8.2%). However, the incidence of colorectal cancer is increasing (Ahnen et al. 2014), especially in China where it is threatening the lives and health of many (Zhu et al. 2017). Although our framework could be applied to any disease, this need makes colorectal cancer a worthy test case to analyze.

We assembled our corpus by collecting biomedical literature from PubMed using the following query: “intestinal diseases"[MeSH Terms] OR ("intestinal"[All Fields] AND "diseases"[All Fields]) OR ("intestinal diseases"[All Fields]) AND ("1900/01/01"[PDAT]:"2019/07/29"[PDAT]) AND "humans"[MeSH Terms] AND English[lang]”.

The initial search returned 422,621 relevant biomedical texts, for which we collected the PubMed PMID, the title and abstract as our local dataset.


Non-standard
Multiple Entities

PMID- 18376247

TI - 12-month follow-up after successful infliximab therapy in pediatric crohn disease.

AB - AIM: Infliximab (IFX) therapy is highly efficacious for the induction and maintenance of remission in pediatric Crohn disease (CD). However, to date it is unclear how long patients should be given IFX. Given the increasing safety concerns about the concomitant and prolonged use of IFX and azathioprine in CD, we wanted to address the clinical outcome in pediatric CD patients who responded to IFX medication, once IFX was stopped. PATIENTS AND METHODS: Upon induction therapy with 3 IFX infusions, 36 of 38 patients with CD were in clinical remission at 3 months. These 36 responders were separated into 2 groups: 16 patients received no further IFX infusions, whereas 20 patients received scheduled maintenance therapy with IFX for 12 months. RESULTS: Among the 16 patients who received no further IFX infusions, 12 experienced relapse during the 12-month follow-up interval after IFX was stopped. In the group receiving maintenance therapy, 11 of 20 patients remained in clinical remission at 12 months of therapy, whereas 8 patients required adjustment of IFX doses or intervals. Among the 11 children who were in clinical remission and receiving maintenance therapy without dose adjustment, 8 experienced relapse within 12 months after IFX maintenance therapy was stopped. Overall, the relapse rates after IFX induction or maintenance therapy was stopped were 75% and 72%, respectively. CONCLUSIONS: These data indicate that IFX is efficacious in controlling severe pediatric CD; however, to induce and maintain clinical remission, repeated IFX infusions are required, with a need for dose adjustment in a substantial number of patients.

Fig. 3 An example showing the complexity of mining triplet information from titles and abstracts

Our computational framework for extracting drug-disease-gene triplets

The broad research framework is illustrated in Fig. 2.

Step 1: Data pre-processing

The challenge in extracting the relationships between biomedical entities with NLP is heightened due to several factors. As shown in Fig. 3, a single title and abstract contain references to multiple entities; naming conventions for the various entities are complex; those conventions tend not be standard; and so on.

Therefore, the purpose of this step is to remove meaningless data and retrieve (only) relevant information. Using SemRep (as outlined in "Natural language processing in the biomedical domain" section) to provide semantic interpretations between the different biomedical entities, we retrieve 2,336,540 SAO structures in the following format:

e.g., PMID—18,376,247

1. C4076075|Infliximab therapy|topp|topp|||TREATS|C2931133|Pediatric Crohn's disease|dsyn|dsyn||.
2. C0004482|Azathioprine|hops,orch,phsu|hops|||ASSOCIATED_WITH|C0010346|Crohn Disease|dsyn|dsyn||.
3. C0004482|Azathioprine|hops,orch,phsu|phsu|||TREATS(INFER)|C2931133|Pediatric Crohn's disease|dsyn|dsyn||.

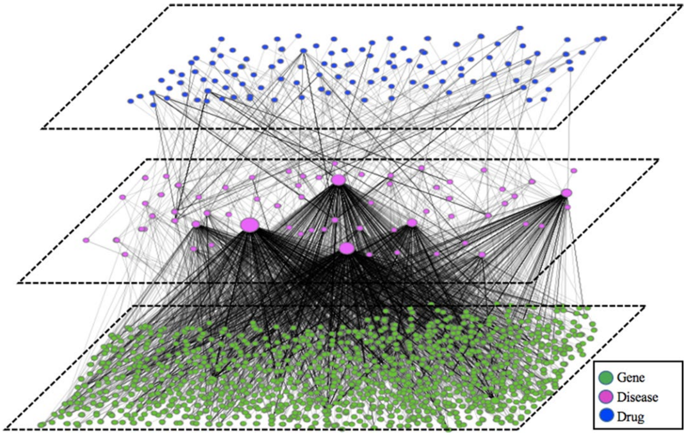


Fig. 4 An integrated drug-disease-gene network

Step 2: Drug-disease-gene triplet extraction

This step further narrows the structures according to type. For our purposes, these were clnd (drug), dsyn (disease), and gngm (gene). Additionally, we limited the associations to drug-drug, drug-disease, drug-gene, disease-disease, disease-gene, and gene–gene. For simplicity, all associations are considered to be non-directional. In other words, as long as there is an association between two entities, we considered there to be an edge between them.

To facilitate the many traversal needs of the local dataset, data should be stored in the following format *(Drug, Semantic relation, Drug), (Drug, Semantic relation, Disease), (Disease, Semantic relation, Disease), (Disease, Semantic relation, Gene), (Gene, Semantic relation, Gene), (Drug, Semantic relation, Gene).*

Step 3: Construction of an integrated drug-disease-gene network

Constructing this network occurs in two stages:

1. Disease-Gene links: First, from over 4081 diseases in the dataset, we selected 688 diseases which related to genes. From these, given the SAO structures, we added 1582 genes and 7753 disease-gene links to the network.
2. Drug-Disease links: Second, we traversed the local dataset again to mine drug-disease links. From this, we added 110 diseases, 116 drugs, and 538 relationships between them to the network.

The final result was an integrated three-layer network with 1,321 nodes (105 drugs/69 diseases/1,148 genes) and 5562 edges, as shown in Fig. 4, along with 11,832 drug-disease-gene triplets related to colorectal cancer and its associated complications. Those wishing to access the full results can visit https://pan.baidu.com/s/1OTdRXjBi2y7WWCkVkppfaw (Password: 4mr4).

Table 1 Statistics of the three extracted biomedical entities and six association types

Association type	Record	Unique association	Number of unique entities
Disease-gene	7753	3952	688 (Disease)/1582 (Gene)
Disease-drug	538	276	110 (Disease)/116 (Drug)
Drug-gene	3	3	3 (Drug)/3 (Gene)
Disease-disease	49,965	20,129	3991 (Disease)
Gene–gene	9840	6821	3816 (Gene)
Drug-drug	114	58	43 (Drug)
Total	68,229	31,252	4458 (Gene)/4081(Disease)/131 (Drug)

Table 2 Statistical indicators of the six kinds of complex networks

			Node	
	Data Source	Average degree (k)	(N)	($\ln N$)
Disease-Disease	dataset 1	10.08	3991	8.29
Disease-gene	dataset 2	3.48	2270	7.73
Disease-drug	dataset 3	2.44	226	5.42
Drug-gene	dataset 4	1.00	6	1.79
Drug-drug	dataset 5	2.70	43	3.76
Gene–gene	dataset 6	3.58	3816	8.24

With the exception of the disease-disease network, each individual network was, unsurprisingly, very sparse ($k< <\ln N< <N$; drug-drug, drug-disease, drug-gene, disease-gene, gene–gene) (Arenas et al. 2006; Lü et al. 2009), but also too complex to extract valuable information. The descriptive statistics are shown in Tables 1, 2.

One method of overcoming this problem is to identify the most likely potential associations in the network for further analysis. Therefore, following Agrawal et al. (1994), we introduced association rules for the disease entities, according to the research conclusions in Zhang et al. (2014). As some examples, one rule is: "Diseases that are associated with each other are more likely to associate with a group of common genes." Another is: "Similar diseases can be treated by same drugs." In addition, disease-disease network performance is better than the others, which could provide more value information (Table 2).

The top left of Fig. 5 shows a part of the disease-gene network, which contains one disease (Disease A) and five genes, all of which mutually interact with each other. Thus, if can we can prove that there is a potential association between disease A and B, these five genes can be regarded as candidate genes of Disease B. For drug-disease network, the calculation principle is the same as we mentioned above.[3] Then, we combined the computed results from these two parts to obtain the novel links from drugs to diseases and genes. Finally, we obtained 498 association rules between diseases. Table 3 lists a few examples, and the 49

[3] To confirm predicated links from the results, the literature is mined for disease-gene and drug-disease links respectively.

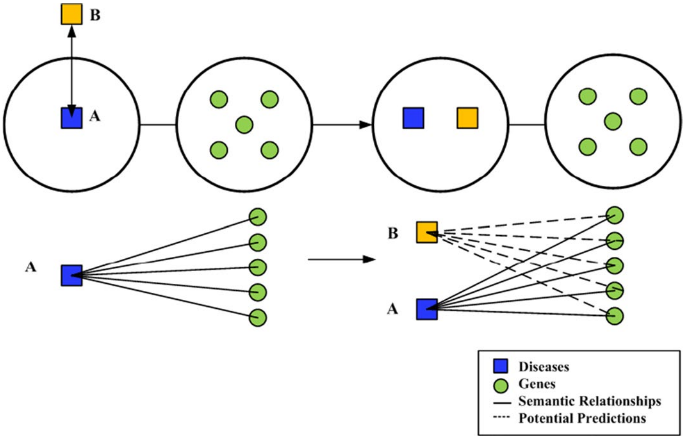


Fig. 5 An example of some incomplete networks Adding an A-B edge according to a set of association rules complete the links providing a more complete picture of the potential associations between diseases

Table 3 A sample of the novel candidate associations predicted by the framework

CUI_D	CUI_D	Support	Confidence	Lift
C0012634	C0267116	0.19	0.000962052	3.20
C0012634	C0030920	0.19	0.018920363	1.16
C0012634	C0017168	0.19	0.011330839	1.34
C0012634	C0267288	0.19	0.000106895	5.34
C0012634	C0520459	0.19	0.01111705	1.30
C0012634	C0392164	0.19	0.000320684	4.01
C0012634	C0400837	0.19	0.000213789	1.07
C0012634	C2674218	0.19	0.000320684	2.67
C0012634	C0017097	0.19	0.002030999	1.54
C0012634	C0005859	0.19	0.000106895	5.34

(min_support ≧ 0.01, lift > 1)

associations with the highest confidence levels are included in the Appendix. The results are discussed in more detail in the next section.

Results

Genomic sequencing can be used as a molecular microscope to classify tumors according to their specific but abnormal biology. Identifying and targeting diseased pathways expressed in a tumor, rather than classifying tumors according to their histological or anatomical tissue of origin, is a revolution in cancer therapeutics that is well underway. As Dulbecco (1986) mentions in his research, cancer seems to be locked to the expression of some viral genes. If we wish to learn more about their "hit-and-run" attack strategy, concentrating on cellular genomes is essential.

According Bray et al. (2018), there will be an estimated 18.1 million new cancer cases and 9.6 million cancer deaths in 2018 (17.0 m/9.5 m excluding nonmelanoma skin cancer).

Lung cancer is the most commonly diagnosed cancer across both sexes, and accounts for the most cancer deaths at 18.4%. Female breast cancer closely follows for incidence (11.6%), then colorectal cancer (10.2%) and prostate cancer (7.1%). For mortality, colorectal cancer leads (9.2%) followed by, stomach cancer (8.2%) and liver cancer (8.2%). However, the incidence of colorectal cancer is increasing (Ahnen et al. 2014), especially in China where it is threatening the lives and health of many (Zhu et al. 2017). Thus, using the computational framework we proposed in this paper, the most relevant complications of colorectal cancer (ulcerative colitis, ileitis, Crohn's disease) are selected from the top 49 association rules (*min_support≧0.01, lift > 1*) after consulting with medical experts.

The next three sections discuss each disease, in turn, beginning with a brief summary of its presentation and common symptoms. The links between diseases and genes mined from the literature represent potential pathogenic genes. The links between diseases and drugs represent candidates for drug repurposing, i.e., drugs that are currently being used to treat one disease that may be efficacious for treating another. These discussions conclude each section.

Ulcerative colitis [C0009324]

Background

Ulcerative colitis (UC) is a chronic inflammatory bowel disease characterized by symptoms of bloody diarrhea, abdominal cramps, and fatigue. The association between UC and colorectal cancer has been documented, and depends on the extent and duration of UC (Eaden et al. 2000). Patients are younger in cases of UC-associated colorectal cancer. They also more frequently have multiple cancerous lesions, and histologically show mucinous or signet ring cell carcinomas. The prevalence of colorectal cancer with UC is different in various geographic regions (Laszlo et al. 2006), and the risk begins to increase 8 or 10 years after the diagnosis of UC.

Using our framework, we extracted 14 drugs and 871 genes related to UC, which we cross-checked in Semantic MEDLINE. These 14 drugs can be divided into four types:

Aminosalicylic acid Mesalamine enema [C1246845]; Mesalamine eEnema [Rowasa] [C0307525]; Sulfasalazine enema [C1248060]; Mesalamine in rectal dosage form [C0360081]

*Anti-inhibitor*Anti-inhibitor [C4284262]; Sodium cromoglycate in oral dosage form [C0360197]; Nicotine transdermal patch [C0358855]; Nicotine chewing gum [C0599654]

*Corticosteroid*Prednisolone enema [C1247637]; Hydrocortisone enema [C1246471]; Budesonide 3 MG [C1128974]; Budesonide 9 mg [C3531316]; Prednisolone rectal foam [C1247642]

*Natural therapy*Aloe vera gel [C0974143]

Discovered disease-gene candidates

Using the association rules, the framework extracted 11 candidate Disease-Genes links from the literature. These appear in Table 4, followed by a brief summary of the main findings from each article.

Steffen et al., (2014) believe that inflammatory bowel disease (IBD) is caused by a combination of environmental factors and susceptible genes. Using a candidate gene

Table 4 Ulcerative colitis—Disease-Gene candidates

CUI_Disease	CUI_Gene	Gene	References	Confidence
C0009324	C1334877	NFKBIA	Steffen et al. (2014)	High
	C1422392	CERS2	Oertel et al. (2017)	High
	C1335093	OSM	West et al. (2017)	High
	C1332700	CCR5	Matsuzaki et al. (2003)	High
	C1970017	CCDC88B	Fodil et al. (2017)	High
	C1413387	CHI3L1	Chen et al. (2011)	High
	C1416796	LANCL2	https://www.businesswire.com/news/home/20170921005188/en/	High
	C0879468	CSF1R	Huynh et al. (2009)	High
	C1540188	ADIPOR1	Obeid et al. (2014)	Low
	C1334470	SMAD7	Chen et al. (2015)	Low
	C1423062	SIRT1	Ma et al. (2018)	Low

8 have direct evidence (high confidence level); 3 have indirect evidence (low confidence)

approach, this group assessed 39 mainly functional single nucleotide polymorphisms (SNPs) in 26 genes that regulate inflammation in a clinically homogeneous group of severely diseased patients. The results show that NFKBIA *[CUI: C1334877]* is associated with risk of UC. Like UC, the loss of intestinal barrier function is a hallmark of IBD. The molecular mechanisms are not well understood but likely involve dysregulation of membrane composition, fluidity, and permeability, which are all essentially regulated by sphingolipids, including ceramides of different chain lengths and saturation. CERS2 *[CUI:C1422392]* is crucial for maintaining colon barrier function and epithelial integrity. In this vein, Oertel et al. (2017) find several factors that may weaken endogenous defenses against endogenous microbiomes: an increase in long-chain ceramides/(dh)-ceramides, sphinganine in the colon, and CERS2 knockdown and its associated changes in several sphingolipids, such as a drop in very long-chain ceramides/(dh)-ceramides.

West et al. (2017) find genetic deletion and/or pharmacological blockades of OSM significantly attenuate colitis. Further, high pre-treatment OSM expression is strongly associated with the failure of anti-tumor necrosis factor (TNF) therapy. OSM is thus a potential biomarker and therapeutic target for UC, with particular relevance for anti-TNF resistant patients. Fodil et al. (2017) reports that CCDC88B *[CUI:C1970017]* inactivation in T-cells may prevent colitis. Further, patients with Crohn’s disease or UC usually present with high levels of CCDC88B, CHI3L1 *[CUI:C1413387]* (Chen et al., 2011), and CCR5 *[CUI:C1332700]* (Matsuzaki et al. 2003). Subsequent studies have provided further evidence that LANCL2[4] [CUI:C1416796] may be a new molecular target in preventing and treating UC-associated colorectal cancer, and CSF1R [CUI:C0879468] hyper-stimulation could be involved in hyperproliferative disorders of the small intestine, such as Crohn’s disease and UC (Huynh et al. 2009).

In previous research, APN null mice expressed an increase in the APN receptor ADIPOR1 [CUI:C1540188] at both the protein and RNA level, and knocking down ADIPOR1

[4] https://www.businesswire.com/news/home/20170921005188/en/.

Table 5 Ulcerative colitis—Candidates for drug repurposing

CUI_Disease	CUI_Drug	Drug	References	Confidence
C0009324	C1123173	Ciprofloxacin 500 mg	Turunen et al. (1998)	High
	C4034144	Methotrexate injection	Nathan et al. (2010)	High
	C4019255	Adalimumab injection	Lee et al. (2017)	High

in vitro in the presence of dextran sulphate sodium (DSS) hindered the ameliorating effects of APN with respect to proliferative, apoptotic, and inflammatory markers (Obeid et al. 2014). Some researchers have also shown that an imbalance between pro- and anti-inflammation is an important mechanism of steroid resistance in UC, and that miRNAs may be involved in this process. In vivo miRNA profiles of serum samples have shown that con-miR-195 is the most obvious influence factor (SMAD7 mRNA *[CUI: C1334470]*, which is a potential target of miR-195). Decreases in miR-195 lead to an increase in SMAD7 expression and a corresponding up-regulation of p65 and the AP-1 (activator protein 1) pathway, which may explain cases of steroid resistance in UC patients (Chen et al. 2015).

From a study on the pathogenesis of lung injury in rats with UC, Ma et al. (2018) find that a lower-expression of SIRT1[CUI:C1423062] in lung tissue is closely related to oxidative stress and inflammatory injury, which may be the molecular mechanisms of lung injury in UC.

Discovered disease-drug candidates

With the links between diseases-genes established, the next relationship in the triple is from diseases to drugs. The links extracted for UC are shown in Table 5.

Methotrexate (MTX) [CUI:C4034144] is used as a second-line immunomodulator in patients with IBD when purine analogs are not tolerated or lack efficacy. High-level evidence indicates the efficacy of MTX administered in intramuscular form with Crohn's disease, but there are few reports of experiments with subcutaneous delivery. Of these, Nathan et al. (2010) studied 45 patients with Crohn's disease and 23 with UC (median age, 46 years; range, 20–80 years; 54% men), each with an intolerance (69%) or resistance (31%) to purine analogs. MTX was initiated in 74% of patients in doses of 25 mg (33) or 20 mg, administered by subcutaneous self-injection in 90% of subjects. Subcutaneously administered MTX showed apparent efficacy, acceptance, tolerance, and safety in patients with Crohn's disease or UC who were steroid-dependent and where purine analogs had been ineffective or intolerable.

Lee et al. (2017) conducted research about whether recalcitrant pyoderma gangrenosum (PG) with UC can be treated by adalimumab injection *[CUI:C4019255]*. In the research, they reported a case of a patient with UC with recalcitrant PG who failed numerous trials of immunosuppressive agents and etanercept but dramatically responded to adalimumab. The successful treatment of PG in their patient suggests that adalimumab may be a valuable therapeutic option for patients with PG and UC.

Turunen et al. (1998) evaluated the role of ciprofloxacin *[CUI:C1123173]* in the induction and maintenance of remission in UC in patients responding poorly to conventional therapy with steroids and mesalamine. During the first 6 months, the treatment-failure rate was 21% in the ciprofloxacin-treated group and 44% in the placebo group

Table 6 Crohn's disease—Disease-Gene candidates

CUI_Disease	CUI_Gene	Gene	References	Confidence
C0010346	C1417495	MUC3A	Kyo et al.(2001)	High
	C1416961	MADCAM1	Bachmann et al. (2006)	High
	C1708427	IL17 wt allele	McGovern et al. (2010)	High
	C1413243	CD86	Chen et al. (2009)	High
	C1415629	HNF4A	Marcil et al. (2012)	High
	C1704807	MLH1 wt allele	Pokornyet al.(1997)	Low
	C1419786	S100A4	Cunningham et al. (2010)	Low
	C1426212	SOCS3	Li et al. (2018)	Low

5 have direct evidence (high confidence); 3 have indirect evidence (low confidence)

($P=0.02$). Endoscopic and histological findings were used as secondary end points and showed better results in the ciprofloxacin group at 3 months but not at 6 months. The addition of a 6-month ciprofloxacin treatment for UC improved the results of conventional therapy with mesalamine and prednisone.

Crohn's disease [C0010346]

Background

Crohn's disease is a chronic and debilitating inflammatory condition of the gastrointestinal tract. Peak incidence is in early adult life, although any age can be affected, and a majority of affected individuals progress to relapsing and chronic disease (Stappenbeck et al. 2011). Some early studies indicate that patients with IBD, especially those with long-standing and extensive UC, have an increased risk of colorectal cancer. Moreover, other researchers have suggested that patients with Crohn's disease also have a higher risk of colorectal cancer (Freeman 2001). However, part of this increased risk in patients may be related to the presence of a rectal stump, rather than to Crohn's disease per se.

We extracted 8 drugs and 360 genes related to Crohn's disease, which we cross-checked in Semantic MEDLINE. The drugs can be divided into four types:

*Aminosalicylic acid*Mesalamine snema [Rowasa] [C0307525];

*Anti-inhibitor*Anti-inhibitor [C4284262]; Adalimumab injection [C4019255]; Sodium cromoglycate (oral) [C0360197]; Methotrexate injection [C4034144]

*Corticosteroid*Budesonide 9 mg [C3531316];

*Others*Methylene blue injection [C4081241]; Ciprofloxacin 500 mg [C1123173]

Discovered disease-gene candidates

Eight Disease-Genes candidates were extracted for Crohn's disease as shown in Table 6.

Kyo et al. (2001) report evidence that MUC3[*CUI:C1417495*] consists of two genes, MUC3A and MUC3B. Additionally, they analyzed SNPs in exonic sequences of the 3′

portions of these two genes to investigate whether sequence variations in those regions could result in differences in IBD susceptibility from person-to-person. Their results show that non-synonymous SNPs of the MUC3A gene involving a tyrosine residue could mean a genetic predisposition to Crohn's disease ($P = 0.0132$). Notably, it has been suggested that tyrosine residue may have a role in cell signaling. Their findings suggest that variants of MUC3A may have a distinct involvement in the occurrence of both UC and Crohn's.

Further, the mucosal addressin cell adhesion molecule-1 (MADCAM1) [*CUI:C1416961*] is selectively expressed in the endothelial cells of intestinal mucosa and gut-associated lymphoid tissue. Engaging MADCAM1 to its ligand, integrin alpha4beta7, on lymphocytes is associated with the homing of gut-associated lymphocytes to normal gastrointestinal tract and inflammation sites. Bachmann et al. (2006) was able to explain the differences between Crohn's and UC from the expression patterns of MADCAM1, with the results indicating a more extensive expression of MADCAM1 in Crohn's, which can not only contribute to mucosal inflammation but also to transmural inflammation.

McGovern et al. (2010) results indicate that the IL23/IL17 [*CUI:C1708427*] pathway is pivotal to the development of chronic mucosal inflammation seen in Crohn's. In their study, patients with both active and inactive Crohn's disease had higher numbers of IL-4-, IL-17-, and IL-23(p19)-positive cells in the lamina propria than in the controls. They therefore conclude that activation of the IL-23/IL-17 axis is fundamentally connected to the etiology of Crohn's disease, and that the increasing sensitivity of epithelium to microbial LPS may be the basis for the relapsing nature of the disease (Veera et al. 2008).

Chen et al. (2009) investigated the expression of the co-stimulatory molecule CD86 [*CUI:C1413243*] and the inducible co-stimulator (ICOS) in the intestinal mucosa of Crohn's disease to explore its pathologic significance. Their results show an increased amount of enterocytes and CD86- or ICOS-positive LPMC in Crohn's patients, suggesting that co-stimulatory molecules may play a role in its pathogenesis. The enterocytes may act as non-specific antigen that presents in cells during the process of cellular immunity activation. Marcil et al. (2012) evaluated the association between genetic variants of HNF4α [*CUI:C1415629*] and Crohn's in two distinct pediatric cohorts in Canada. This is the first report to show that the HNF4A locus may be a common genetic determinant of childhood-onset Crohn's.

A significant correlation was found between Crohn's disease, UC, and MLH1 [*CUI:C1704807*] ($p = 0.037$) in a study by Pokorny et al. (1997) comparing MLH1 exon 15/D3S1611 haplotypes of Crohn's colitis in patients with UC. These are novel genetic and clinical associations between MLH1 and IBD. Cunningham et al. (2010) examined the expression profile of S100A4 in the resected ileum of patients with fibrostenosing Crohn's disease. The results from knockdown experiments indicate a potential role for S100A4 [*CUI:C1419786*] in mediating intestinal fibroblast migration. In addition, identified risk polymorphisms affecting the Jak-STAT3 pathway in patients with Crohn disease could affect TGF-β1 and collagen I expression and in the pathway's negative regulator, SOCS3 [*CUI:C1426212*]. Experiments by Li et al. (2018) show that two factors cause sustained Jak-STAT3 activity in muscle cells in patients with fibrostenotic Crohn's disease, along with excess TGF-β1 production, Collagen I production, and fibrosis. These are autocrine IL-6 production in mesenchymal cells and subepithelial myofibroblasts (SEMF). Paradoxically, there are lower levels of SOCS3in these cells. From these results, they conclude that decreased SOCS3 protein levels are unique to fibrostenotic patients.

Table 7 Crohn's disease—Candidates for drug repurposing

CUI_Disease	CUI_Drug	Drug	References	Confidence
C0010346	C0974143	Aloe vera gel	Vemu et al. (2015)	High

Discovered disease-drug candidates

Only one Disease-Drug candidate was found for Crohn's disease, as shown in Table 7.

Emu oil is an animal product used by the Aborigines of Australia to treat inflammation, burns, and other similar conditions. In other parts of the world, aloe vera is used in a similar way. Given that Crohn's is an inflammatory disease and the relevant therapeutic properties of these two substances, (Vemu et al. (2015) conducted a study to evaluate the efficacy of aloe vera and emu oil alone and in combination as an alternative to sulfasalazine (an allopathic drug) for treating Crohn's disease. The histomorphological changes indicated that the combination of aloe vera and emu oil resulted in better protection than sulfasalazine by suppressing the oxidative ($P < 0.05$).

Ileitis [C0020877]

Ileitis is related to the above diseases. For instance, UC patients with pancolitis and backwash ileitis, an extension of the inflammatory process into the terminal ileum, may be at increased risk of colorectal carcinoma (Yamaguchi et al. 2010). Also, narrowing or constriction of the abdomen in cross-sectional imaging at the time of a terminal ileitis diagnosis has been correlated to the eventual onset of Crohn's disease. In turn, this increases the risk of colorectal cancer. However, no significant correlation has been found between clinical symptoms, endoscopic features, laboratory testing, NSAID use, smoking history, or family history of IBD.

According to research by Sundaram et al. (2003), genetic alterations may be one of the causes of IBD. In patients with IBD, nutrient absorption is inhibited in the intestine leading to the most common and disabling symptoms of this disorder: diarrhea, malnutrition, weight loss, abdominal pain, and eventually a failure to thrive. However, current medical therapy has important limitations. Aminosalicylates are only modestly effective (Sutherland et al. 2006); corticosteroids (*e.g.,* glucocorticoids) can cause unacceptable adverse events and do not provide a benefit as maintenance therapy; and TNF antagonists, although efficacious (Sandborn et al. 2005), predispose patients to serious infection (Keane et al. 2001). Thus, new treatment strategies are needed.

We extracted 0 drugs and 4 genes related to Ileitis using our framework, which we cross-checked in Semantic MEDLINE.

Discovered disease-gene candidates

Table 8 lists the Disease-Gene candidates for Ileitis.

Table 8 Ileitis—Disease-Gene candidates

CUI_Disease	CUI_Gene	Gene	References	Confidence
C0020877	C1367307	STAT3	Mitsuyama et al. (2006)	High
	C1332688	CCL25	Rivera-Nieves et al. (2006)	High
	C1413191	CCR7	Mcnamee et al. (2013)	High
	C1417494	MUC2	Sovran et al. (2013)	High
	C1705969	CTLA4 wt (allele)	Assi and Wilson (2013), Venditti et al. (2015)	High
	C1334098	IL10	Zhou et al. (2014)	High
	C1336636	TLR4	Narimatsu et al. (2015)	High
	C1336637	TLR5	Lopetuso et al. (2017)	High

8 have direct evidence (high confidence)

Mitsuyama et al. (2006) demonstrates that the signal transducer and activator of transcription STAT3 [*CUI: C1367307*] suppresses the cytokine signaling SOCS3 pathway, which is pivotal in human IBD. Subsequent research on whether STAT3 activation contributes to ileitis shows that STAT3 signaling is critical in the development of intestinal inflammation in SAMP1/ Yit mice, and therefore STAT3 blockade may have a therapeutic effect.

Rivera-Nieves et al. (2006) investigated the expression of CCL25 [*CUI: C1332688*] and CCR9 as a function of disease progression in a spontaneous murine model of chronic ileitis (SAMP1/YitFc) using flow cytometry, real-time reverse-transcription polymerase chain reactions, an enzyme-linked immunosorbent assay, and immunohistochemistry. They believe these molecules are most influential during the early stages of chronic murine ileitis. CCR7 [*CUI: C1413191*] also acts as a chemokine receptor, and an immunoblockade of CCR7 will result in further effector T-cell retention, which exacerbates ileitis (Mcnamee et al. 2013). Research by Sovran et al. (2013) shows an association between the MUC2 gene [*CUI: C1417494*] and ileitis. Moreover, homeostatic mechanisms can prevent ileitis in mice that have deficient MUC2 production.

New antitumor immunotherapy strategies for Stage IV metastatic melanoma include ipilimumab, which is a monoclonal antibody against CTLA4 [*CUI: C1705969*]. Assi and Wilson (2013) presented two cases of long-duration immune-related responses with ipilimumab in a phase II trial. A 66-year-old woman with multiple lung metastases from a primary scalp melanoma received 4 doses of ipilimumab with a mixed clinical response. However, after the first maintenance dose, she developed severe ileitis and colitis that responded to steroid therapy. Venditti et al. (2015) also finds that ipilimumab and immune-mediated adverse events could lead to anti-CTLA4 induced ileitis.

Zhou et al. (2014) explored the change and significance of IL8, IL4, and IL10 [*CUI: C1334098*] in the pathogenesis of terminal ileitis in rata. The results confirm that IL10 and IL4 can inhibit the inflammatory reaction of terminal ileum and, conversely, that IL8 can induce the inflammatory reaction in terminal ileitis and chemokines aggregation and mediate inflammatory reaction by mediating other inflammatory factors; as a proinflammatory cytokine, IL8 can inhibit IL10, which is a key anti-inflammatory cytokine produced by activated immune cells and plays a critical role in the control of immune responses.

The toll-like receptor TLR4 [*CUI: C1336636*] and aberrant leukocyte migration to the intestinal mucosa are reported to be involved in the pathology of intestinal enteropathy, and

Table 9 Ileitis—Candidates for drug repurposing

CUI_Disease	CUI_Drug	Drug Name	References	Confidence
C0020877	C3531316	Budesonide 9 mg	Lombardi et al. 2010; Boyd et al. (1995)	High
	C1123173	Ciprofloxacin 500 mg	McLaughlin et al. (2008)	High

TLR 2 agonists have been found to evoke hyposensitivity to TLR 4 stimulation in vitro. Further experiments by Narimatsu et al. (2015) on toll-like receptor TLR2 agonists show that they could ameliorate indomethacin-induced murine ileitis by suppressing TLR4 signaling. Lopetuso et al. (2017) also provide evidence that aberrant, elevated TLR5 expression is present in the ileal epithelium of SAMP mice, which is augmented in the presence of gut microbiomes, and that TLR5 activation in response to bacterial flagellin results in an inability to maintain appropriate epithelial barrier integrity. Together, these findings represent a potential mechanistic pathway that can exacerbate and perpetuate chronic gut inflammation in ileitis and, possibly, in patients with Crohn's disease.

Discovered disease-drug candidates

The two Disease-Drug candidates found for Ileitis are listed in Table 9.

Budesonide [*CUI:C3531316*] is used to treat Crohn's disease. However, in experiments by Lombardi et al. (2010), oral budesonide was used to successfully treat localized eosinophilic ileitis with mastocytosis. Boyd et al. (1995) compared the effects of plain and controlled-ileal-release (CIR) formulations of budesonide on intestinal inflammation, with the results suggesting that CIR budesonide is significantly more effective in reducing intestinal inflammation than plain budesonide. Additionally, the site of delivery influences its effectiveness, and the local (topical), rather than systemic, action of this compound is primarily responsible for its anti-inflammatory effect. Ciprofloxacin [*CUI:C1123173*] is also used for the treatment of Crohn's disease, but McLaughlin et al. (2008) shows that T1313 combined with ciprofloxacin and metronidazole is highly effective for treating of pre-pouch Ileitis following a restorative proctocolectomy.

Conclusion

Many genetic mutations predispose individuals to disease (Greenman et al. 2007). The practice of precision medicine involves identifying such mutations in patients and modifying patient treatments to reflect each person's different physiological risks (Collins and Varmus 2015). A corpus of drug-disease-gene relationships plays an important role in this process by acting as a reference for providers to help determine the significance of their patient's mutations and optimize the drugs prescribed on an individual basis (Ashley et al. 2010; Dewey et al. 2014). However, prescribing a precision course of treatment with full knowledge of the medical literature requires an investigation into all known pair-wise and three-way associations among bio-entities. Further, the complexity of these existing associations must be understood if we are to infer novel associations between these entities going forward. Many studies have explored pair-wise associations, with much knowledge gained. However, with this study, we go part of the way to overcoming the challenges associated with identifying the three-way associations, which have historically been much harder to ascertain.

Hence, in this paper, we present a framework for how to integrate prior knowledge regarding drugs, diseases, and genes, and how to use this in a systems approach to complete the incomplete links between them. We also show that introducing association rules among disease entities can help to infer new relationships between drugs, diseases, and genes. We validated the links predicted from the results with a manual literature review, and the results indicate that the proposed computational framework has the potential to: (1) identify potential disease relationships (see Table 10 in the Appendix); (2) prioritize candidate disease genes (see Tables 4, 6, and 8); (3) predict novel options for drug repurposing (see Tables 5, 7, and 9); (4) provide insights that could help to formulate novel research hypotheses; and (5) identify new triplet associations for various diseases. Each of these contributions is significant to the implementation and advancement of precision medicine.

The major limitation of the method is its requirement for manual external validation. Further, additional relevant information might be mined from the full text or supplementary material, which cannot be found in the title and abstract alone. Overcoming these limitations we leave to future work as the latter limitation in particular has been shown to be an important source of biomedical information (Jimeno-Yepes and Verspoor 2014).

Appendix

See Tables 10 and 11.

Table 10 New associations between diseases

CUI_D	CUI_D	Support	Confidence	Lift
C0012634	C0267116	0.19	0.000962052	3.2
C0012634	C0030920	0.19	0.018920363	1.16
C0012634	C0017168	0.19	0.011330839	1.34
C0012634	C0267288	0.19	0.000106895	5.34
C0012634	C0520459	0.19	0.01111705	1.3
C0012634	C0392164	0.19	0.000320684	4.01
C0012634	C0400837	0.19	0.000213789	1.07
C0012634	C2674218	0.19	0.000320684	2.67
C0012634	C0017097	0.19	0.002030999	1.54
C0012634	C0005859	0.19	0.000106895	5.34
C0012634	C0679362	0.19	0.000106895	1.34
C0012634	C0041296	0.19	0.004810262	1.01
C0012634	C0022104	0.19	0.039337253	1.31
C0012634	C1142110	0.19	0.000962052	1.72
C0012634	C0080276	0.19	0.000106895	5.34
C0012634	C2931180	0.19	0.000213789	2.14
C0012634	C0003615	0.19	0.019134153	1.28
C0012634	C0342895	0.19	0.001389631	4.96
C0012634	C0024899	0.19	0.000427579	1.53
C0012634	C4284413	0.19	0.000106895	1.34
C0012634	C0012813	0.19	0.005558525	1.04
C0012634	C0034888	0.19	0.002993052	1.02
C0012634	C0008780	0.19	0.000106895	1.78
C0012634	C0018081	0.19	0.000320684	1.07
C0010346	C0020877	0.12	0.00331675	2.15
C0010346	C1290884	0.12	0.033001658	2.01
C0010346	C0949272	0.12	0.015091211	6.39
C0010346	C0267841	0.12	0.000165837	1.38
C0010346	C0030524	0.12	0.013764511	6.14
C0010346	C0025007	0.12	0.001824212	3.14
C0010346	C0030785	0.12	0.0013267	3.01
C0010346	C0000833	0.12	0.006965174	1.63
C0010346	C0029453	0.12	0.005804312	1.61
C0010346	C1290864	0.12	0.00199005	1.88
C0010346	C0003509	0.12	0.00066335	2.37
C0010346	C0243001	0.12	0.005804312	3.05
C0010346	C1960526	0.12	0.000165837	8.29
C0010346	C0009324	0.12	0.089220564	1.16
C0010346	C0042870	0.12	0.003814262	1.51
C0010346	C0020875	0.12	0.070480929	7.83
C0010346	C0341268	0.12	0.108457711	7.46
C0010346	C0021390	0.12	0.143283582	1.42
C0021390	C0406549	0.1	0.000197824	2.47
C0021390	C0263445	0.1	0.000197824	3.29
C0021390	C0010054	0.1	0.002769535	1.41
C0021390	C0009319	0.1	0.029475767	1.63

Table 10 (continued)

CUI_D	CUI_D	Support	Confidence	Lift
C0021390	C0009324	0.1	0.137289812	1.78
C0021390	C0010346	0.1	0.170919881	1.42
C0021390	C3495919	0.1	0.004549951	1.06

Table 11 Drug-disease-gene triples

CUI_Drug	Drug	CUI_Disease	Disease	CUI_Gene	Gene
C3531316	Budesonide 9 mg	C0020877	Ileitis	C1367307	STAT3
				C1332688	CCL25
				C1413191	CCR7
				C1417494	MUC2
				C1705969	CTLA4 wt (allele)
				C1334098	IL10
				C1336636	TLR4
				C1336637	TLR5
C1123173	Ciprofloxacin 500 mg	C0020877	Ileitis	C1367307	STAT3
				C1332688	CCL25
				C1413191	CCR7
				C1417494	MUC2
				C1705969	CTLA4 wt (allele)
				C1334098	IL10
				C1336636	TLR4
				C1336637	TLR5
C1123173	Ciprofloxacin 500 mg	C0009324	Ulcerative colitis	C1334877	NFKBIA
				C1422392	CERS2
				C1335093	OSM
				C1332700	CCR5
				C1970017	CCDC88B
				C1413387	CHI3L1
				C1416796	LANCL2
				C0879468	CSF1R
				C1540188	ADIPOR1
				C1334470	SMAD7
				C1423062	SIRT1
C4034144	Methotrexate injection	C0009324	Ulcerative colitis	C1334877	NFKBIA
				C1422392	CERS2
				C1335093	OSM
				C1332700	CCR5
				C1970017	CCDC88B
				C1413387	CHI3L1
				C1416796	LANCL2
				C0879468	CSF1R
				C1540188	ADIPOR1
				C1334470	SMAD7
				C1423062	SIRT1

Table 11 (continued)

CUI_Drug	Drug	CUI_Disease	Disease	CUI_Gene	Gene
C4019255	Adalimumab injection	C0009324	Ulcerative colitis	C1334877	NFKBIA
				C1422392	CERS2
				C1335093	OSM
				C1332700	CCR5
				C1970017	CCDC88B
				C1413387	CHI3L1
				C1416796	LANCL2
				C0879468	CSF1R
				C1540188	ADIPOR1
				C1334470	SMAD7
				C1423062	SIRT1
C0974143	Aloe vera gel	C0010346	Crohn's disease	C1417495	MUC3A
				C1416961	MADCAM1
				C1708427	IL17 wt (allele)
				C1413243	CD86
				C1415629	HNF4A
				C1704807	MLH1 wt (allele)
				C1419786	S100A4
				C1426212	SOCS3

Acknowledgments This work is partly supported by the General Program of the National Natural Science Foundation of China (Grant No.71774012, 71673024) and the strategic research project of the Development Planning Bureau of the Chinese Academy of Sciences (Grant No.GHJ-ZLZX-2019-42). The findings and observations present in this paper are those of the authors and do not necessarily reflect the views of the supporters or the sponsors. The authors would like to thank the anonymous reviewers for their constructive input into this paper.

References

Ahnen, D. J., et al. (2014). The increasing incidence of young-onset colorectal cancer: a call to action. *Mayo Clinic Proceedings, 89*(2), 216–224.

Agrawal, R., Carey, M., Faloutsos, C., Ghosh, S., & Swami, A. (1994). Quest: A project on database mining. *Acm Sigmod Record, 23*(2), 514.

Arenas, A., Diaz-Guilera, A., & Pérez-Vicente, C. J. (2006). Synchronization reveals topological scales in complex networks. *Physical review letters, 96*(11), 114102.

Ashley, E. A., et al. (2010). Clinical assessment incorporating a personal genome. *The Lancet, 375*(9725), 1525–1535.

Assi, H., & Wilson, K. S. (2013). Immune toxicities and long remission duration after ipilimumab therapy for metastatic melanoma: two illustrative cases. *Current Oncology, 20*(2), 165–169.

Bachmann, C., et al. (2006). Targeting mucosal addressin cellular adhesion molecule (MAdCAM)-1 to noninvasively image experimental Crohn's Disease. *Gastroenterology, 130*(1), 8–16.

Boyd, A. J., Sherman, I. A., & Saibil, F. G. (1995). Effects of plain and controlled-ileal-release budesonide formulations in experimental ileitis. *Scandinavian Journal of Gastroenterology, 30*(10), 974–981.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA A Cancer Journal For Clinicians, 68*(6), 394–424.

Bui, Q. C., Sloot, P. M., Van Mulligen, E. M., & Kors, J. A. (2014). A novel feature-based approach to extract drug–drug interactions from biomedical text. *Bioinformatics, 30*(23), 3365–3371.

Chen, A. J., Li, F., Zhang, Y., Gong, E. C., & Shi, X. Y. (2009). Expression of co-stimulatory molecule CD86 and its inducible co-stimulator in Crohn disease and their pathologic significance. *Journal of Peking University. Health Sciences, 41*(6), 620–624.

Chen, C. C., et al. (2011). Chitinase 3-like-1 expression in colonic epithelial cells as a potentially novel marker for colitis-associated neoplasia. *American Journal of Pathology, 179*(3), 1494–1503.

Chen, G., Cao, S., Liu, F., & Liu, Y. (2015). Mir-195 plays a role in steroid resistance of ulcerative colitis by targeting smad7. *Biochemical Journal, 471*(3), 357–367.

Cheung, W. Y., & Liu, G. (2009). Genetic variations in esophageal cancer risk and prognosis. *Gastroenterology Clinics of North America, 38*(1), 75–91.

Christos, A., Anuj, S., Vassilis, V., Spyros, D., & Aris, P. (2011). Literature mining, ontologies and information visualization for drug repurposing. *Briefings in Bioinformatics, 12*(4), 357–368.

Collins, F. S., & Varmus, H. (2015). A new initiative on precision medicine. *New England Journal of Medicine, 372*(9), 793–795.

Cunningham, M. F., Docherty, N. G., Burke, J. P., & O'Connell, P. R. (2010). S100A4 expression is increased in stricture fibroblasts from patients with fibrostenosing Crohn's disease and promotes intestinal fibroblast migration. *American Journal of Physiology Gastrointestinal & Liver Physiology, 299*(2), G457.

Cutting, D., Kupiec, J., Pedersen, J., & Sibun, P A. (1992). A practical part of speech tagger. In Third conference on applied natural language processing, 133-140.

Dewey, F. E., et al. (2014). Clinical interpretation and implications of Whole-Genome Sequencing. *Journal of the American Medical Association, 311*(10), 1035–1045.

Duke, J. D., et al. (2012). Literature based drug Interaction prediction with clinical assessment using electronic medical records: novel myopathy associated drug interactions. *Plos Computational Biology, 8*(8), e1002614.

Dulbecco, R. (1986). A turning point in cancer research: equencing the human genome. *Science, 231*(4742), 1055–1056.

Eaden, J. A., Abrams, K., Ekbom, A., Jackson, E., & Mayberry, J. (2000). Colorectal cancer prevention in ulcerative colitis: a case-control study. *Alimentary Pharmacology and Therapeutics, 14*(2), 145–153.

Egger, G., Liang, G., Aparicio, A., & Jones, P. A. (2004). Epigenetics in human disease and prospects for epigenetic therapy. *Nature, 429,* 457–463.

Fodil, N., et al. (2017). CCDC88B is required for pathogenesis of inflammatory bowel disease. *Nature Communications, 8*(1), 1–12.

Freeman, H. J. (2001). Colorectal cancer complicating Crohn's disease. *Canadian Journal of Gastroenterology, 15*(4), 231–236.

Fundel, K., Küffner, R., Zimmer, R. (2007). RelEx—Relation extraction using dependency parse trees. *Bioinformatics, 23*(3), 365–371.

Greenman, C., et al. (2007). Patterns of somatic mutation in human cancer genomes. *Nature, 446*(7132), 153–158.

Hristovski, D., Peterlin, B., Mitchell, J. A., & Humphrey, S. M. (2005). Using literature-based discovery to identify disease candidate genes. *International Journal of Medical Informatics, 74*(2/4), 289–298.

Hu, G., & Agarwal, P. (2009). Human disease-drug network based on genomic expression profiles. *PLoS ONE, 4*(8), e6536.

Huang, J., Ellinghaus, D., Franke, A., Howie, B., & Li, Y. (2012). 1000 genomes-based imputation identifies novel and refined associations for the wellcome trust case control consortium phase 1 data. *European Journal of Human Genetics, 20*(7), 801–805.

Huynh, D., et al. (2009). CSF-1 Dependence of paneth cell development in the mouse small intestine. *Gastroenterology, 137*(1), 136–144.

Jimeno-Yepes, A., & Verspoor, K. (2014). Literature mining of genetic variants for curation quantifying the importance of supplementary material. *Database The Journal of Biological Database, 3.*

Karaderi, T., et al. (2009). Association between the interleukin 23 receptor and ankylosing spondylitis is confirmed by a new UK case-control study and meta-analysis of published series. *Rheumatology, 48*(4), 386–389.

Keane, J., et al. (2001). Tuberculosis associated with infliximab, a tumor necrosis factor alpha-neutralizing agent. *New England Journal of Medicine, 345*(15), 1098–1104.

Kyo, K., Muto, T., Nagawa, H., Lathrop, G. M., & Nakamura, Y. (2001). Associations of distinct variants of the intestinal mucin gene MUC3A with ulcerative colitis and Crohn's disease. *Journal of Human Genetics, 46*(1), 5–20.

Laszlo, L., et al. (2006). Risk factors for ulcerative colitis-associated colorectal cancer in a hungarian cohort of patients with ulcerative colitis: results of a population-based study. *Inflammatory Bowel Diseases, 12*(3), 205–211.

Lee, J. H., et al. (2017). Treatment of recalcitrant pyoderma gangrenosum with ulcerative colitis by adalimumab injection. *Annals of Dermatology, 29*(2), 260–262.

Lombardi, C., Salmi, A., Savio, A., & Passalacqua, G. (2010). Localized eosinophilic ileitis with mastocytosis successfully treated with oral budesonide. *Allergy, 62*(11), 1343–1345.

Lopetuso, L. R., Jia, R., Wang, X. M., Jia, L. G., Petito, V., Goodman, W. A., & Pizarro, T. T. (2017). Epithelial-specific Toll-like receptor (TLR) 5 activation mediates barrier dysfunction in experimental ileitis. *Inflammatory bowel diseases, 23*(3), 392–403.

Lü, L., Jin, C. H., & Zhou, T. (2009). Similarity index based on local paths for link prediction of complex networks. *Physical Review E, 80*(4), 046122.

Ma, S., Sun, Q., & Guo, B. (2018). Expression of Sirt1 in the lung tissue of rats with ulcerative colitis and its relationship with oxidative stress and inflammatory reaction. *Anatomy Research, 40*(6), 489–493.

Marcil, V., et al. (2012). Association between genetic variants in the HNF4A gene and childhood-onset Crohn's disease. *Genes & Immunity, 13*(7), 556–565.

Mason, O., & Verwoerd, M. (2007). Graph theory and networks in biology. *IET systems biology, 1*(2), 89–119.

Matsuzaki, K., Hokari, R., Kato, S., Tsuzuki, Y., Tanaka, H., & Kurihara, C., et al. (2003). Differential expression of CCR 5 and CRTH 2 on infiltrated cells in colonic mucosa of patients with ulcerative colitis. *Journal of Gastroenterology & Hepatology, 18*(9), 1081–1088.

McGovern, D. P. B., et al. (2010). Genetic epistasis of IL23/IL17 pathway genes in Crohn's disease. *Inflammatory Bowel Diseases, 15*(6), 883–889.

McLaughlin, S. D., Bell, A. J., Clark, S. K., Tekkis, P. P., Ciclitira, P. J., & Nicholls, R. J. (2008). T1313 combined ciprofloxacin and metronidazole is highly effective for the treatment of pre-pouch ileitis following restorative proctocolectomy. *Gastroenterology, 4*(134), 529.

Mcnamee, E. N., Masterson, J. C., Jedlicka, P., Collins, C. B., Williams, I. R., & Rivera-Nieves, J. (2013). Ectopic lymphoid tissue alters the chemokine gradient, increases lymphocyte retention and exacerbates murine ileitis. *Gut, 62*(1), 53–62.

Mirnezami, R., Nicholson, J., & Darzi, A. (2012). Preparing for precision medicine. *New England Journal of Medicine, 366*(6), 489–491.

Mitsuyama, K. (2006). STAT3 activation via interleukin 6 trans-signalling contributes to ileitis in SAMP1/Yit mice. *Gut, 55*(9), 1263–1269.

Narimatsu, K., et al. (2015). Toll-like receptor (TLR) 2 agonists ameliorate indomethacin-induced murine ileitis by suppressing the TLR4 signaling. *Journal of Gastroenterology & Hepatology, 30*(11), 1610–1617.

Natarajan, N., & Dhillon, I. S. (2014). Inductive matrix completion for predicting gene-disease associations. *Bioinformatics, 30*(12), 60–68.

Nathan, D. M., Iser, J. H., & Gibson, P. R. (2010). A single center experience of methotrexate in the treatment of crohn's disease and ulcerative colitis: a case for subcutaneous administration. *Journal of Gastroenterology & Hepatology, 23*(6), 954–958.

Obeid, S., George, J., & Hebbard, L. (2014). P-188 the role of adiponectin in inflammatory bowel disease. *Inflammatory Bowel Diseases, 20*(1), S102.

Oertel, S., et al. (2017). Ceramide synthase 2 deficiency aggravates AOM-DSS-induced colitis in mice: role of colon barrier integrity. *Cellular & Molecular Life Sciences, 74*(16), 3039–3055.

Ozgür, A., Vu, T., Erkan, G., & Radev, D. R. (2008). Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics, 24*(13), 277–285.

Pokorny, R. M., Hofmeister, A., Galandiuk, S., Dietz, A. B., & Neibergs, H. L. (1997). Crohn's disease and ulcerative colitis are associated with the DNA repair gene MLH1. *Annals of Surgery, 225*(6), 718–723.

Rindflesch, T. C., Bean, C. A., & Sneiderman, C. A. (2000). Argument identification for arterial branching predications asserted in cardiac catheterization reports. In *Proceedings of the AMIA Symposium* (pp. 704–708). American Medical Informatics Association.

Rindflesch, T. C., & Fiszman, M. (2003). The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics, 36*(6), 462–477.

Rindflesch, T. C., Rajan, J. V., & Hunter, L. (2000). Extracting molecular binding relationships from biomedical text. In *Sixth Applied Natural Language Processing Conference* (pp. 188–195).

Rindflesch, T.C. , Tanabe, L. ,Weinstein, J. N. , & Hunter, L. (2000). EDGAR: Extraction of drugs, genes and relations from the biomedical literature. Pacific Symposium on Biocomputing, 517–528.

Rivera-Nieves, J., et al. (2006). Antibody blockade of CCL25/CCR9 ameliorates early but not late chronic murine ileitis. *Gastroenterology, 131*(5), 1518–1529.

Ross, J. S., et al. (2015). Comprehensive genomic profiling of carcinoma of unknown primary site. *Jama Oncology, 1*(1), 40–49.

Roy, N., Raj, R., Rai, S., & Varadwaj, P. K. (2019). Deciphering the novel target genes involved in the epigenetics of hepatocellular carcinoma using graph theory approach. *Current Genomics, 20*(8), 545–555.

Sandborn, W. J., Reinisch, W., Rachmilewitz, D., Hanauer, S. B., & Colombel, J. F. (2005). Infliximab induction and maintenance therapy for ulcerative colitis: the ACT2 trial. *Ztschrift Für Gastroenterologie, 43*(5), A104–A105.

Simone, D., Joachim, H. V., Matthias, R., & Michael, S. (2012). Drug repositioning through incomplete bicliques in an integrated drug–target–disease network. *Integrative Biology, 4*(7), 778–788.

Singhal, A., Simmons, M., & Lu, Z. (2016). Text mining for precision medicine: automating diseasemutation relationship extraction from biomedical literature. *Journal of the American Medical Association, 23*(4), 766–772.

Sovran, B., et al. (2013). Mo1813 homeostatic mechanisms preventing ileitis in mice with absent or deficient MUC2 production. *Gastroenterology, 144*(5), S-669.

Stappenbeck, T. S., et al. (2011). Crohn disease: a current perspective on genetics, autophagy and immunity. *Autophagy, 7*(4), 355–374.

Steffen, B., et al. (2014). Polymorphisms in the inflammatory pathway genes TLR2, TLR4, TLR9, LY96, NFKBIA, NFKB1, TNFA, TNFRSF1A, IL6R, IL10, IL23R, PTPN22, and PPARG are associated with susceptibility of inflammatory bowel disease in a danish cohort. *PLoS ONE, 9*(6), e98815.

Sun, P., Guo, J., Winnenburg, R., & Baumbach, J. (2016). Drug repurposing by integrated literature mining and drug-gene-disease triangulation. *Drug Discovery Today, 22*(4), 615–619.

Sundaram, U., et al. (2003). Rabbit chronic ileitis leads to up-regulation of adenosine A1/A3 gene products, oxidative stress, and immune modulation. *Biochemical Pharmacology, 65*(9), 1529–1538.

Sutherland, L. R., & MacDonald, J. K. (2006). Oral 5-aminosalicylic acid for maintenance of remission in ulcerative colitis. *Cochrane Database of Systematic Reviews*, (2), CD000544.

Turunen, U. M., et al. (1998). Long-term treatment of ulcerative colitis with ciprofloxacin: A prospective, double-blind, placebo-controlled study. *Gastroenterology, 115*(5), 1072–1078.

Veera, H., et al. (2008). IL-23/IL-17 immunity as a hallmark of Crohn's disease. *Inflammatory Bowel Diseases, 14*(9), 1175–1184.

Vemu, B., Selvasubramanian, S., & Pandiyan, V. (2015). Emu oil offers protection in Crohn's disease model in rats. *Bmc Complementary & Alternative Medicine, 16*(1), 1–9.

Venditti, O., et al. (2015). Ipilimumab and immune-mediated adverse events: A case report of anti- CTLA4 induced ileitis. *Bmc Cancer, 15*(1), 87.

West, N. R., et al. (2017). Erratum: oncostatin m drives intestinal inflammation and predicts response to tumor necrosis factor-neutralizing therapy in patients with inflammatory bowel disease. *Nature Medicine, 23*(6), 788.

Yamaguchi, N., Isomoto, H., Shikuwa, S., Ohnita, K., & Nakao, K. (2010). Proximal extension of backwash ileitis in ulcerative-colitis - associated colon cancer. *Medical Science Monitor International Medical Journal of Experimental And Clinical Research, 16*(7), 87–91.

Zhang, Y., Tao, C., Jiang, G., Nair, A. A., Su, J., Chute, C. G., & Liu, H. (2014). Network-based analysis reveals distinct association patterns in a semantic MEDLINE-based drug-disease-gene network. *Journal of Biomedical Semantics, 5*(1), 33.

Zhou, H. Y., Yan, J., Fang, L., Zhang, H., Su, L. G., & Zhou, G. H. (2014). Change and significance of IL-8, IL-4, and IL-10 in the pathogenesis of terminal ileitis in sd rat. *Cell Biochemistry & Biophysics, 69*(2), 327–331.

Zhou, J., & Fu, B. (2018). The research on gene-disease association based on textmining of PubMed. *BMC Bioinformatics, 19*(1), 37.

Zhu, J., Tan, Z., Hollis-Hansen, K., Zhang, Y., Yu, C., & Li, Y. (2017). Epidemiological trends in colorectal cancer in China: an ecological study. *Digestive Diseases and Sciences, 62*(1), 235–243.

Authors and Affiliations

Xuefeng Wang[1] · Shuo Zhang[1] · Yao Wu[1] · Xuemei Yang[2]

[1] School of Management and Economics, Beijing Institute of Technology, Beijing 100081, China

[2] Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100005, China