Journal of
Biomedical Semantics

SOFTWARE

Open Access

Literature evidence in open targets - a target validation platform

Şenay Kafkas[1,2*], Ian Dunham[1,2] and Johanna McEntyre[1,2]

Abstract

Background: We present the Europe PMC literature component of Open Targets - a target validation platform that integrates various evidence to aid drug target identification and validation. The component identifies target-disease associations in documents and ranks the documents based on their confidence from the Europe PMC literature database, by using rules utilising expert-provided heuristic information. The confidence score of a given document represents how valuable the document is in the scope of target validation for a given target-disease association by taking into account the credibility of the association based on the properties of the text. The component serves the platform regularly with the up-to-date data since December, 2015.

Results: Currently, there are a total number of 1168365 distinct target-disease associations text mined from >26 million PubMed abstracts and >1.2 million Open Access full text articles. Our comparative analyses on the current available evidence data in the platform revealed that 850179 of these associations are exclusively identified by literature mining.

Conclusions: This component helps the platform's users by providing the most relevant literature hits for a given target and disease. The text mining evidence along with the other types of evidence can be explored visually through https://www.targetvalidation.org and all the evidence data is available for download in json format from https://www.targetvalidation.org/downloads/data.

Keywords: Target validation, Text mining, Target-disease associations, Document ranking, Information retrieval

Background

Understanding the underlying mechanisms of diseases is crucial in translational research. Discovering the association between drug target and disease has become a main focus for scientists since it is key for developing new drugs or repurposing them. Scientists gather various evidence representing different aspects of target-disease associations such as gene expression changes and the role of genetic variations to increase understanding. Such evidence can be stored in structured databases and requires integration to obtain complete and comprehensive knowledge in target validation studies.

Motivated by this, the Target Validation Platform (https://targetvalidation.org) [1] integrates different evidence from various resources with the aim of assisting scientists to identify and prioritise drug targets (proteins and their genes) associated with diseases and phenotypes. The evidence includes common disease genetic evidence based on GWAS study results from GWAS Catalog [2], rare Mendelian disease evidence based on ClinVar [3] clinical variant information from EVA and text mined target-disease associations from the Europe PMC (https://europepmc.org/) literature database [4] (see Table 3 for a complete list of evidence types).

Europe PMC contains over 33 million records and expands at a rate of over a million articles per year—one article every two minutes as scientists publish their findings continuously. Text mining target-disease associations is crucial for an integrated platform like the Target Validation Platform, since it provides a high volume of complementary and up-to-date data to the other type of evidences, otherwise the knowledge would stay hidden in millions of documents.

In this study, we present the Europe PMC Open Targets literature component that identifies target-disease associations in documents and ranks the documents

* Correspondence: kafkas@ebi.ac.uk
[1] European Molecular Biology Laboratory (EMBL-EBI), European Bioinformatics Institute, Wellcome Genome Campus, Hinxton CB10 1SD, UK
[2] Open Targets, Wellcome Genome Campus, Hinxton CB10 1SD, UK

according to their confidence based on rules utilising expert-provided heuristic information. Our main aim is to provide a scalable, robust and continuous text-mining service to the community for a real-world and very important application—target validation. Many of the previous studies focused on extracting gene-disease association from the literature [5–7]. However, only a few of them specifically focused on developing methods for integrated resources; DisGeNET [8] and DISEASES [9] for example cover various types of evidence for target validation. These two systems provide confidence scores for target-disease associations extracted from Medline abstracts for a given disease or target and don't provide very regular updates to the data. In DisGeNET, the target-disease text mining method is based on a machine learning approach while in DISEASES, target-disease associations are extracted based on scoring their co-occurrences according to their confidence. In comparison to DisGeNET and DISEASES, our system operates on full text articles in addition to abstracts, and ranks documents according to the confidence for a given target-disease association rather than ranking the associations extracted from the whole set of Medline abstracts. More specifically, we calculate a document confidence score for each given (article, target, disease) triple which represents how valuable the document is in the scope of target validation for the given target-disease association (see "Document scoring" section). However, the confidence score of a given target-disease association is handled at the platform level and calculated based on all the evidence data in the platform by using a harmonic sum approach (see [1] for the details). This confidence score at the association level represents the overall credibility of the evidence for a given target-disease association. Our approach to target-disease extraction differs from these systems, and probably many other traditional text-mining studies, in that we rely on heuristic information from experts/users for developing the system. The platform was first launched in December, 2015 and is publicly available at https://targetvalidation.org. Since then, our system has served the platform regularly (monthly) with up-to-date data.

Implementation

Resources used

The literature source that we used in the study is the Europe PMC database. Europe PMC is one of the largest biomedical literature databases in the world which provides public access to >30.4 million abstracts and >3.3 million full text articles from PubMed and PubMed Central. In our analyses, we used the latest version of the Open Access full text articles (http://europepmc.org/ftp/archive/v.2016.06/) (~1.2 Million), and all of the PubMed abstracts (~26 Million) from the database.

Two comprehensive resources, UniProt and the Experimental Factor Ontology (EFO) are used to identify target and disease names in text, respectively. These two resources are chosen as the reference resources by Open Targets. The data providers of the platform are asked to ground their target and disease entities in to these reference resources so as to integrate the evidence in the platform. Therefore, two dictionaries are generated and refined from the human part of the SwissProt Database (the annotated part of UniProt, Release 2015_10) (http://www.uniprot.org/) and disease and phenotype parts of EFO (http://www.ebi.ac.uk/efo/) (Release 2.74) before applying text mining. In the refining process, we filtered out the terms that would introduce potentially very high numbers of false positives. These are the terms having character length < 3 (e.g. "A" is a gene name) and terms that are ambiguous with common English words (e.g. "Large" is a protein name as well). In addition, we generated term variations by replacing the widely used Greek letters in gene/disease names with their symbols (e.g. replacing "alpha" with α). The final target and disease dictionaries consisted of a total of 104,434 and 29,846 terms respectively. These dictionaries are available from ftp://ftp.ebi.ac.uk/pub/databases/pmc/otar/.

Target and disease name annotation

We used the Europe PMC text-mining pipeline, which is based on Whatizit [10], to annotate target and disease names in text with the two dictionaries described above. Although we reduce a very high level of ambiguity by applying the dictionary refinement process before text mining the documents, some target and disease name abbreviations could still be ambiguous with some other names. For example, ALS which is an abbreviation used for "Amyotrophic Lateral Sclerosis", is ambiguous with "Advanced Life Support" in some articles (e.g. see PMID:26811420). Therefore, we implemented and used a disease and target name abbreviation filter for screening out the potential false positive abbreviations introduced during the annotation process. Our tool differs from the available abbreviation finders, such as [11] since it behaves rather as a filter specifically for potential false positive target and disease name abbreviations annotated based on our dictionaries.

The abbreviation filter operates based on several rules using heuristic information. Regular expressions are used for identifying the text sequences in the form of "**X.... Y.... Z.... (XYZ)**". The text in parentheses (i.e. (XYZ)) is identified as a gene/disease name abbreviation candidate if it is in the uppercase form, has length <6 (the length was decided by manually analysing a random subset of the Uniprot and EFO dictionaries) and annotated by the system either as a disease or a gene name, whereas, the text located immediately before the parentheses is identified as the potential long form. For example, in the following sentence from the article having PMID:26811420; "The guidelines form the basis for all levels of resuscitation training, now from first aid to *advanced life support (ALS)*," the italicised text matches with our pattern defined above. "ALS" would be the abbreviation candidate and "advanced life support" would be the potential long form. Documents matching the pattern above are analysed manually by an expert to come up with heuristics that we can apply in filtering the ambiguous abbreviation. Abbreviation candidates satisfying one of the following rules are kept as true target/disease abbreviations, otherwise, they are filtered out:

For disease name abbreviation candidates:

- If any of the EFO long forms of the abbreviation candidate exists in the document
- If the long form extracted from the text contains any of the keywords (disease, disorder, syndrome, defect, etc.) that can be used to describe a disease

For gene or protein name abbreviation candidates:

- If (XYZ) appears more than 3 times in the document body (this rule applies to OA full text documents only)
- If the long form matches any of the terms from SwissProt or Enzymes (http://enzyme.expasy.org/)
- If the long form ends with (-ase/-ases) OR it contains any of the keywords (factor, receptor, gene, protein etc.) that can be used to describe a target name
- If at least 3 sentences for full text and at least 2 sentences for abstracts contain one of the keywords: "mutation, SNP, variation, gene, inhibit, variation, variant, polymorphism, mutant, isoform, protein, enzyme, activate, antibody, transcription, tumor suppressor, express, overexpress, regulator, receptor, oncogene" along with the protein name abbreviation candidate and a disease name.

Target-disease association identification

Our association extraction method is based on identification of target-disease co-occurrences at the sentence level and applying several filtering rules to reduce noise possibly introduced by the high sensitivity, low specificity co-occurrence method. Our filtering rules utilise heuristic information from a careful manual analysis of the text data to filter out potential false positive associations. More specifically, the manual analyses are conducted iteratively by analysing a randomly selected set of results and identifying the reasons behind the false positives in the results so that we could formulate them as filtering rules to tune our system.

The system applies the following filtering rules:

1. Filter out all type of articles except "Research" articles (e.g. Reviews, Case Reports).
2. Filter out target-disease associations appearing in the following sections: Methods, References, Acknowledgement & Funding, Competing Interests, Author Contribution and Supplementary Material.
3. Filter out target-disease associations that appear only once in the body of a given article but not in the article's title or abstract.

Sections of a given document are identified by using our Section Tagger [12] tool that we developed previously.

Document scoring

A document scoring algorithm is implemented and integrated in to the system to assign each document a confidence score for a given target-disease association. Document confidence score for a given target-disease association represents how valuable the document is in the scope of target validation by taking credibility of the given association into account. Document confidence scores are used to rank all the documents relevant to a given target-disease association. The algorithm is based on weighting document sections and sentence locations differently for full text articles and abstracts respectively (see Table 1 and Table 2). The weighting approach is often used in text mining tasks for assigning confidence scores. For example in [9] different weights are assigned to the different features for calculating the confidence scores of the identified associations. In our study, we assign weights from the range of [1–10] which is wide enough to pick different weights for different sections based on their potential confidence. The following formulas, CS_1 and CS_2 are used to calculate the confidence scores for abstracts and full text articles respectively:

$$\begin{aligned} S_1\left(PMID_x, Gene_y, Disease_z\right) = \sum\nolimits_{i=First}^{Last} & Sentence\ Location\ Weight_i \\ & * \#association\left(Gene_y, Disease_z\right)\ in\ Sentence\ Location_i \\ & + Gene_y\ in\,abstract * 0.2 \end{aligned}$$

$$S_2\left(PMID_x, Gene_y, Disease_z\right) = \sum_{i=intro}^{Other} Section\ Weight_i * \#association\left(Gene_y, Disease_z\right) \\ in\ Section_i$$

$$Boosting\,Up\,Factor = Median\ of\ all\ OA\ article\ body\ scores, \\ ie.\ S_2\left(PMID_x, Gena_y, Disease_z\right)$$

Table 1 Sentence location weights in abstracts

Sentence Location	Weight
First or second	2
Last	5
Other	3

Table 2 Section weights in full text articles

Section	Weight
Title	10
Abstract	See Table 1
Results, Figure, Table	5
Discussion, Conclusion	2
Introduction, Case Study, Appendix, Other	1

$$\begin{aligned} CS_1\left(PMID_x, Gene_y, Disease_z\right) &= Section\ Weight_{Title} * \#association\left(Gene_y, Disease_z\right) \\ &\quad + S_1\left(PMID_x, Gena_y, Disease_z\right) \\ &\quad + Boosting\,Up\,Factor \end{aligned}$$

$$\begin{aligned} CS_2\left(PMID_x, Gene_y, Disease_z\right) &= S_1\left(PMID_x, Gene_y, Disease_z\right) \\ &\quad + S_2\left(PMID_x, Gene_y, Disease_z\right) \end{aligned}$$

The weights are selected based on heuristic information and our goal is to identify associations that are the subject of the given paper, rather than instances that are reviewing prior knowledge. Therefore, we assign the highest weight, which is 10, to “Title”, since an article title would contain the most confident information and highlight the main finding of the study. The lowest weight (1), is assigned to “Introduction”, since well-known associations are often reported here while a higher weight (5) is assigned to the “Results”, “Figures” and “Tables” sections where the new findings are generally reported.

The sentence location weights that are used for abstract scoring are determined based on a sentence level concept analysis by using CoreSC [13]. CoreSC is a text-mining tool which assigns each sentence one of its 11 pre-defined concepts such as “Results” and “Background”. Our concept analysis performed on randomly selected 360 MEDLINE abstracts revealed that most of the time, the last sentence of a given abstract is a “Results” sentence, while the first/second one is generally an introductory sentence (“Background”) (CoreSC analysis results are available at ftp://ftp.ebi.ac.uk/pub/databases/pmc/otar/). We further verified our finding by manually checking some of the abstracts from this set. Hence, we assign the highest weight (5) to the last sentence and lower weights to the first/second and other sentences accordingly.

Results & discussion

Performance evaluation

The ultimate goal of this study is to provide a scalable, robust and continuous service to the biomedical community for target validation, by using text mining methods. Therefore, we took a different approach from many traditional text mining studies and benchmarked the system based on expert perspective—expert satisfaction and feedback are the most valuable parameters for us to judge on the system’s performance. Our service has been up and running since December 2005 and we continuously improve our algorithms as we receive user feedback. Nevertheless, as a case study, we estimated the overall performance of the system on two randomly selected samples by using Mean Average Precision (MAP) which is a commonly used metric in evaluation of ranking system performance. MAP takes into account the relative order of the documents retrieved by the system and gives more weight to the documents returned at higher ranks [14]. We manually estimated the MAP for abstracts only as 89% and for full text articles as 90% on the top 25 documents of the two randomly selected gene-disease associations which were IGF1—Diabetes and NOD2—Inflammatory Bowel Disease. We also estimated the correlation coefficients between the abstract only and full text article scores as 0.82 and 0.94 for IGF1—Diabetes and NOD2—Inflammatory Bowel Disease respectively. Obtaining almost the same MAP values for both abstracts only and full text articles as well as high correlation coefficients between the scores are promising for our heuristic score adjustment.

The individual performances of the components used are as follows: The target and disease names are identified based on Whatizit by using SwissProt and EFO as terminological resources. The target (gene/protein) name tagging method of Whatizit is compared against some other existing methods on different gold standard datasets previously [15]. Results reveal that Whatizit delivers gene/protein name annotations (grounded in Swissprot) at the state-of-the-art level (~60% F-score values are obtained on different gold standard corpora). The results show that there is still some room for improving the performance and in future we will explore on expanding our Swissport dictionary with the other available resources (e.g. Entrez Gene Database). We evaluated our disease name tagger which is based on EFO on randomly selected 50 abstracts manually. Our tool achieves a recall of 83.67%, a precision of 97.61% and an F-score of 90.10%. Results show that there is still some room for the EFO’s coverage improvement. Indeed, one of the considerations of Open Target is the EFO’s coverage as EFO is being used as the reference dataset for diseases/phenotypes in the platform. Therefore, we previously analysed the coverage of EFO against other 5 major disease/phenotype resources (ORDO, UMLS, MP, HDO and HP) [16]. Based on our previous finding, which is in line with the current finding, Open Targets is currently working on developing methods to expand EFO’s coverage. The abbreviation name filtering performance alone was estimated to have an F-Score value of 92.3% by evaluating randomly selected 50 sentences from the Open Access articles reporting on target-disease associations.

The Section tagger's performance was previously estimated manually on 100 full text articles as an F-score of 98.02% [12].

In the near future, we plan to organise a hackathon that would allow us to form a gold standard dataset and also conduct extensive usability test. The gold standard dataset as well as the user feedback, would allow us to carry out extensive evaluations on our design strategies, and improve them if necessary.

User experience

Since the first release of the Europe PMC Open Targets component, we iteratively improved our text mining algorithm and the visualisation of the text mining evidence in the Target Validation Platform based on user feedback. Initial user testing showed that the incorporation of the text mining evidence in to the platform filled in perceived gaps in evidence caused by limitations in coverage by the other direct evidence sources. The users also valued the reinforcement of other evidence when complementary text mining evidence was available. Feedback from users of incorrect associations predominantly from false positive entity recognition assisted us in improving our filters.

Added value from the literature mined target-disease associations

The Target Validation Platform currently covers evidence from literature mining, genetic associations, somatic mutations, known drugs, gene expression, affected pathways and animal models. (Please refer to [1] for further information about how the other types of evidence data are gathered.) In the current release (release 1.2) of the platform, there are a total number of 2,485,000 distinct target-disease associations. Table 3 shows a comparison of the target-disease association data currently available in the platform. The literature evidence constitutes the largest amount of data compared to the other type of evidence (such as gene expression and animal models). Currently, there are more than 1.1 million (47% of the whole evidence data) distinct target-disease associations extracted from ~26 million PubMed abstracts and ~1.2 million open access full text articles. Other large amounts of evidence data are provided from the gene expression (~900 K) and animal models (~600 K) sources. The analysis shows that 21.75% (197,943) of gene expression, 43.31% (56,228) of genetic associations, 69.36% (2506) of affected pathways, 16.55% (99,836) of animal models, 33.59% (19,801) of somatic mutations and 34.56% (19,811) of known drugs evidence data overlap with the literature mining data. The majority of the distinct associations in the platform are identified exclusively through literature mining (~850 K, 34.21%) showing the added value from text mining.

The discrepancy between the literature mining data and the other type of evidence data is due to the fact that each evidence data is gathered by using different methods as well as resources. For example, gene expression data is gathered from Expression Atlas (https://www.ebi.ac.uk/gxa/home), the scope of which is microarray or RNA-Seq experiments. Other evidence data such as genetic associations and known drugs are gathered through manual curation of the literature by experts and from DailyMed (https://dailymed.nlm.nih.gov/dailymed/). Our approach is based on computationally extracting evidence data from the literature. In many of the curated studies, which may report associations between many targets and several diseases, it is unusual to highlight the individual association results in a way that is detectable by the sentence co-occurrence approach and often these associations are confined to a supplementary data table. Indeed, previous studies focusing on text mining supplementary material revealed that there are many more data in supplementary material compared to abstract and full text [17, 18]. Although text mining and manual curation both use the biomedical

Table 3 Comparison on the target-disease association data in the Target Validation Platform (release 1.2)

Evidence Type	Total number of distinct target-disease associations	Overlapping target-disease association						Total number of exclusively identified associations
		Gene Expression	Genetic Associations	Affected Pathways	Animal Models	Somatic Mutations	Known Drugs	
Literature Mining	1,168,365	197,943	56,228	2506	99,836	19,801	19,811	850,179
Gene Expression	909,960	X	18,945	901	35,616	32,795	9913	669,330
Genetic Associations	129,826	X	X	1912	26,504	3626	2133	62,999
Affected Pathways	3613	X	X	X	1045	310	163	714
Animal Models	602,995	X	X	X	X	2965	4421	486,167
Somatic Mutations	58,941	X	X	X	X	X	1845	16,197
Known Drugs	57,319	X	X	X	X	X	X	33,005

Total number of distinct target-disease associations in the platform is 2,485,000 literature as a resource, the coverage of the methods is different and complementary. In fact, in our early work with users the text-mining approach was highly valued precisely because it accesses evidence from papers that do not contribute to the curated databases. One further reason for any discrepancy originates from the licencing restrictions on the reuse of full text content. We can only text mine the full text of Open Access publications (and all MEDLINE abstracts), while experts can curate evidence from the non-open access publications, accessed for reading via journal subscriptions.

We further analysed the contribution of text mining based on the associations by disease and associations by target in Table 4 and Table 5 respectively. Table 4 shows comparison of the associations by disease in the platform. Currently, there are a total number of 9426 associations by disease in the platform. The majority of these diseases are provided from genetic associations (5912), literature mining (5801) and animal models (4942). Our analysis shows that 56.02% (405) of gene expression, 59.98% (3546) of genetic associations, 88.89% (504) of affected pathways, 68.86% (3403) of animal models, 53.75% (494) of somatic mutations and 82.72% (1489) of known drugs provided target associated diseases overlap with the literature mining data. The majority of the distinct associations by disease in the platform are identified exclusively through genetic associations (1336, 14.17%) and literature mining (1304, 13.83%).

Table 5 shows comparison of the associations by target in the platform. Currently, there are a total number of 30592 associations by target in the platform. The majority of these targets are provided from gene expression (29,842), literature mining (14,728) and genetic associations (10,200). Our analysis shows that 47.64% (14,217) of gene expression, 85% (8670) of genetic associations, 96.23% (664) of affected pathways, 94.36% (5187) of animal models, 94.32% (3903) of somatic mutations and 97.35% (736) of known drugs provided disease associated targets overlap with the literature mining data. The majority of the distinct associations by target in the platform are identified exclusively through gene expression (14,148, 46.25%) which is understandable given the comprehensive gene coverage in gene expression experiments such as RNA-seq.

Altogether, our analysis shows that literature mining suggests many more new target-disease associations (850,179, see Table 3) rather than new diseases (1304, see Table 4) or targets (321, see Table 5) involved in associations.

Examples of target-disease associations exclusively identified by literature mining

Our analysis reveals that there are a total number of 850,179 target-disease associations exclusively identified by literature mining. One such example is the CTGF gene and male breast carcinoma association (Fig. 1) (https://www.targetvalidation.org/evidence/ENSG00000118523/EFO_0006861). Currently, there is evidence for the association of 101 different targets with male breast carcinoma. All of these targets are identified through literature mining and only 4 of them are also supported by the known drugs evidence.

Another example is the ST3GAL4 and diabetes mellitus association. There are 1572 different publications potentially reporting this association (Fig. 2).

(https://www.targetvalidation.org/evidence/ENSG00000110080/EFO_0000400). Currently, there is evidence for the association of 5017 different targets with diabetes mellitus. 3670 of these targets are identified through literature mining.

Conclusions

Here, we present the Europe PMC Open Targets component, a new service for analysing and visualising target-disease associations from the literature within Open Targets. The aim of this component is to help users by providing the most relevant literature hits for

Table 4 Comparison of the associations by disease in the Target Validation Platform (release 1.2)

Evidence Type	Total number of distinct associations by disease	Overlapping associations by disease						Total number of exclusively identified associations by disease
		Gene Expression	Genetic Associations	Affected Pathways	Animal Models	Somatic Mutations	Known Drugs	
Literature Mining	5801	405	3546	504	3403	494	1489	1304
Gene Expression	723	X	520	196	309	460	328	25
Genetic Associations	5912	X	X	527	3725	530	1193	1336
Pathways	567	X	X	X	443	168	310	9
Animal Models	4 942	X	X	X	X	281	752	811
Somatic Mutations	919	X	X	X	X	X	354	113
Known Drugs	1800	X	X	X	X	X	X	179

Total number of distinct associations by disease in the platform is 9426

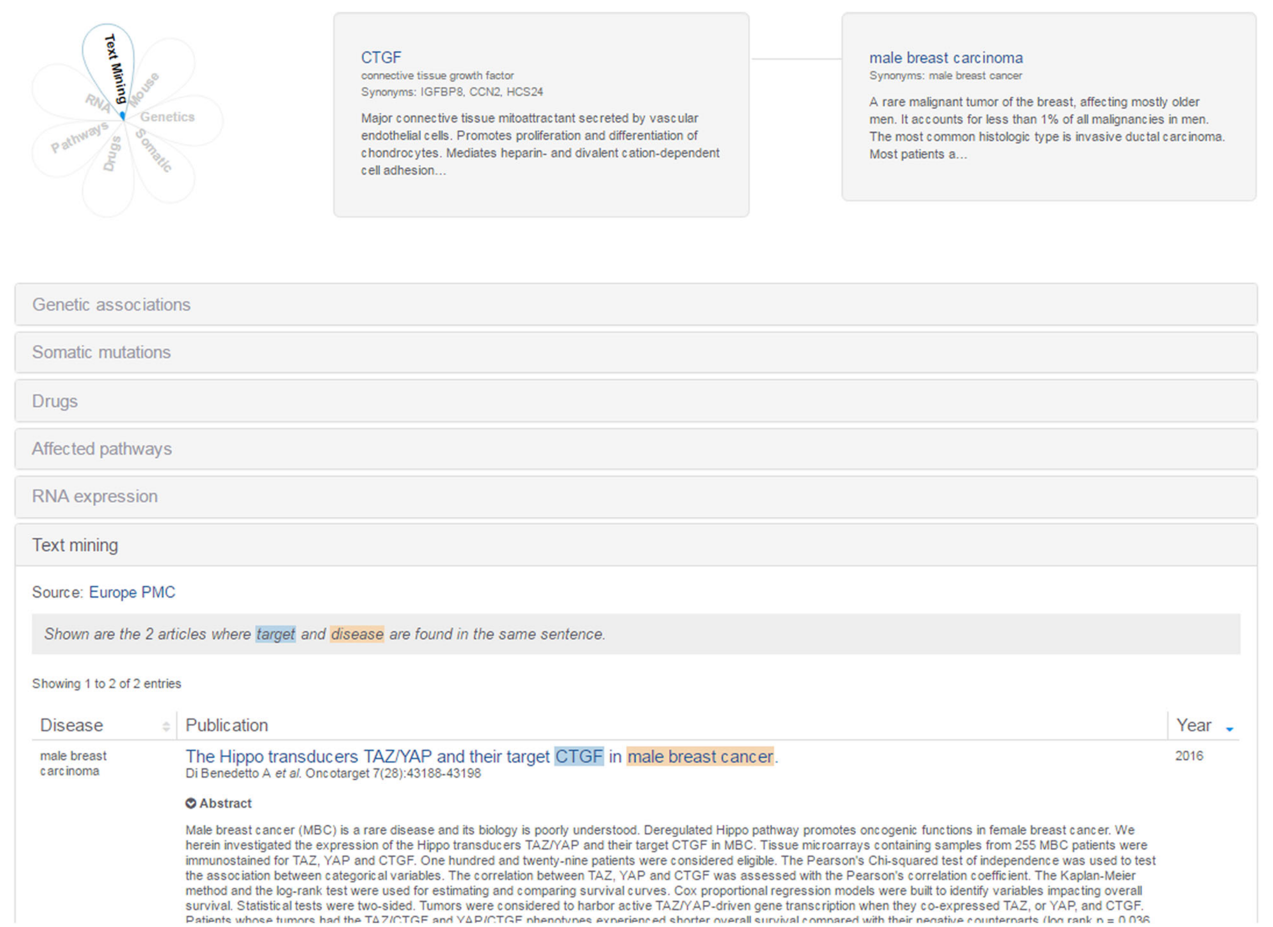
Table 5 Comparison of the associations by target data in the Target Validation Platform (release 1.2)

Evidence Type	Total number of associations by target	Overlapping associations by target						Total number of exclusively identified associations by target
		Gene Expression	Genetic Associations	Affected Pathways	Animal Models	Somatic Mutations	Known Drugs	
Literature Mining	14,728	14,217	8670	664	5187	3903	736	321
Gene Expression	29,842	X	9817	671	5449	4125	743	14,148
Genetic Associations	10,200	X	X	561	4072	3165	569	217
Pathways	690	X	X	X	379	324	70	4
Animal Models	5497	X	X	X	X	3744	484	8
Somatic Mutations	4138	X	X	X	X	X	330	2
Known Drugs	756	X	X	X	X	X	X	1

Total number of distinct associations by target in the platform is 30,592

a given target and disease. The platform users reported that the text mining evidence helped Open Targets to become more complete and a given association is more credible when it is supported not only by text mining but also by the other types of evidence. Our text mining algorithm and visualisation of the text mining evidence are improved iteratively based on user feedback.

Currently, we are analysing the EFO coverage by comparing it against the other existing disease/



CTGF
connective tissue growth factor
Synonyms: IGFBP8, CCN2, HCS24

Major connective tissue mitoattractant secreted by vascular endothelial cells. Promotes proliferation and differentiation of chondrocytes. Mediates heparin- and divalent cation-dependent cell adhesion...

male breast carcinoma
Synonyms: male breast cancer

A rare malignant tumor of the breast, affecting mostly older men. It accounts for less than 1% of all malignancies in men. The most common histologic type is invasive ductal carcinoma. Most patients a...

Genetic associations

Somatic mutations

Drugs

Affected pathways

RNA expression

Text mining

Source: Europe PMC

Shown are the 2 articles where target and disease are found in the same sentence.

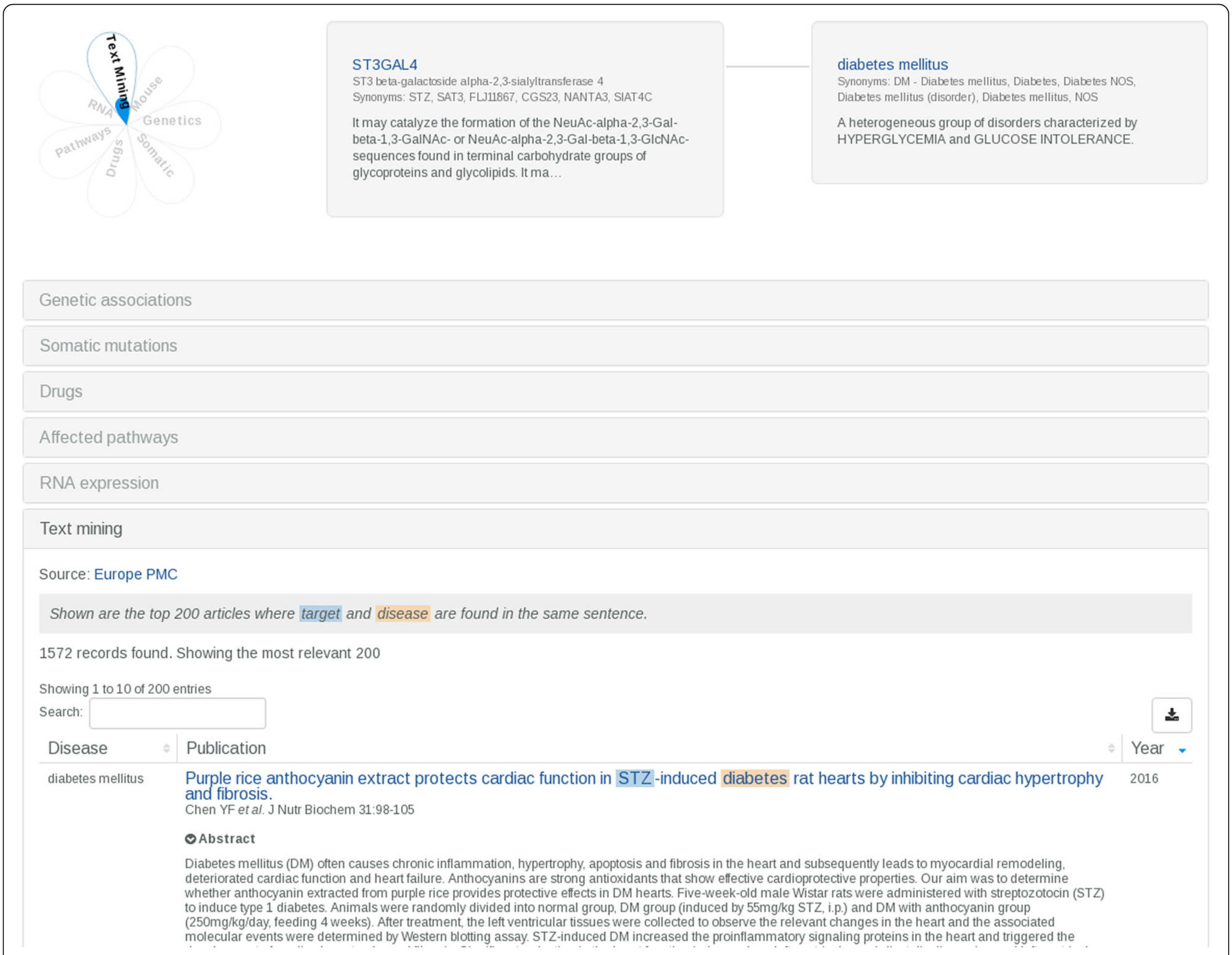
Showing 1 to 2 of 2 entries

Disease	Publication	Year
male breast carcinoma	The Hippo transducers TAZ/YAP and their target CTGF in male breast cancer. Di Benedetto A *et al.* Oncotarget 7(28):43188-43198	2016

Abstract

Male breast cancer (MBC) is a rare disease and its biology is poorly understood. Deregulated Hippo pathway promotes oncogenic functions in female breast cancer. We herein investigated the expression of the Hippo transducers TAZ/YAP and their target CTGF in MBC. Tissue microarrays containing samples from 255 MBC patients were immunostained for TAZ, YAP and CTGF. One hundred and twenty-nine patients were considered eligible. The Pearson's Chi-squared test of independence was used to test the association between categorical variables. The correlation between TAZ, YAP and CTGF was assessed with the Pearson's correlation coefficient. The Kaplan-Meier method and the log-rank test were used for estimating and comparing survival curves. Cox proportional regression models were built to identify variables impacting overall survival. Statistical tests were two-sided. Tumors were considered to harbor active TAZ/YAP-driven gene transcription when they co-expressed TAZ, or YAP, and CTGF. Patients whose tumors had the TAZ/CTGF and YAP/CTGF phenotypes experienced shorter overall survival compared with their negative counterparts (log rank p = 0.036

Fig. 1 The CTGF and male breast carcinoma association



ST3GAL4
ST3 beta-galactoside alpha-2,3-sialyltransferase 4
Synonyms: STZ, SAT3, FLJ11867, CGS23, NANTA3, SIAT4C

It may catalyze the formation of the NeuAc-alpha-2,3-Gal-beta-1,3-GalNAc- or NeuAc-alpha-2,3-Gal-beta-1,3-GlcNAc-sequences found in terminal carbohydrate groups of glycoproteins and glycolipids. It ma…

diabetes mellitus
Synonyms: DM - Diabetes mellitus, Diabetes, Diabetes NOS, Diabetes mellitus (disorder), Diabetes mellitus, NOS

A heterogeneous group of disorders characterized by HYPERGLYCEMIA and GLUCOSE INTOLERANCE.

Genetic associations

Somatic mutations

Drugs

Affected pathways

RNA expression

Text mining

Source: Europe PMC

Shown are the top 200 articles where target and disease are found in the same sentence.

1572 records found. Showing the most relevant 200

Showing 1 to 10 of 200 entries
Search:

Disease	Publication	Year
diabetes mellitus	Purple rice anthocyanin extract protects cardiac function in STZ-induced diabetes rat hearts by inhibiting cardiac hypertrophy and fibrosis. Chen YF *et al.* J Nutr Biochem 31:98-105	2016

Abstract

Diabetes mellitus (DM) often causes chronic inflammation, hypertrophy, apoptosis and fibrosis in the heart and subsequently leads to myocardial remodeling, deteriorated cardiac function and heart failure. Anthocyanins are strong antioxidants that show effective cardioprotective properties. Our aim was to determine whether anthocyanin extracted from purple rice provides protective effects in DM hearts. Five-week-old male Wistar rats were administered with streptozotocin (STZ) to induce type 1 diabetes. Animals were randomly divided into normal group, DM group (induced by 55mg/kg STZ, i.p.) and DM with anthocyanin group (250mg/kg/day, feeding 4 weeks). After treatment, the left ventricular tissues were collected to observe the relevant changes in the heart and the associated molecular events were determined by Western blotting assay. STZ-induced DM increased the proinflammatory signaling proteins in the heart and triggered the

Fig. 2 The ST3GAL4 and diabetes mellitus association

phenotype resources such as Disease Ontology (http://disease-ontology.org/) and Unified Medical Language System (https://www.nlm.nih.gov/research/umls/). In future, we plan to expand the EFO's coverage based on our findings. We also work on classifying articles based on the available evidence types in the platform such as genetic variations and RNA expression. This would provide users with a better understanding and more insight on the weight of individual target-disease associations.

Availability and requirements

All target-disease data is available for download from https://www.targetvalidation.org/downloads/data as compressed json files.

The compiled target and disease dictionaries as well the dataset used in MAP estimation are available from ftp://ftp.ebi.ac.uk/pub/databases/pmc/otar/ for download.

The source code is available from the contact author upon request. The code runs on linux system.

Abbreviations

CS: Confidence score; EFO: Experimental Factor Ontology; Europe PMC: Europe PubMed Central; GSK: GlaxoSmithKline; MAP: Mean average precision; PMID: PubMed Identifier; RNA: Ribonucleic acids

Acknowledgements

We thank Maria Liakata for her help with running CoreSC on a sample set of abstracts. This paper is an extended version of a conference paper by the same name, presented at Phenotype Day@ ISMB 2016.

Funding

This work was funded by Open Targets.

Authors' contributions

ŞK, ID and JM conceived of the study. ŞK implemented the software and performed the experiments. ID provided the heuristic information and performed the manual evaluations. All authors evaluated the results and contributed to the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate
Not applicable.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 24 October 2016 Accepted: 31 May 2017
Published online: 06 June 2017

References

1. Koscielny G, et al. Open targets: a platform for therapeutic target identification and validation. Nucleic Acids Res. 2016;45(D1):D985–94.
2. Welter D, Macarthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014; 42(Database issue):D1001–6.
3. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42(Database issue):D980–5.
4. Europe PMC Consortium. Europe PMC: a full-text literature database for the life sciences and platform for innovation. Nucleic Acids Res. 2015; 43(Database issue):D1042–8.
5. Özgür A, Vu T, Erkan G, Radev DR. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. Bioinformatics. 2008;24(13):i277–85.
6. Al-Mubaid H, Singh RK. A text-mining technique for extracting gene-disease associations from the biomedical literature. Int J Bioinform Res Appl. 2010;6(3):270–86.
7. Hou W-J, Kuo B-Y. Discovery of Gene-disease Associations from Biomedical Texts. Electron J Comput Sci Inf Technol. 2016;4(1):1–8.
8. Piñero J, Queralt-Rosinach N, Bravo À, Deu-Pons J, Bauer-Mehren A, Baron M, Sanz F, Furlong LI. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database (Oxford). 2015. doi: 10.1093/database/bav028.
9. Pletscher-Frankild S, Pallejà A, Tsafou K, Binder JX, Jensen LJ. DISEASES: text mining and data integration of disease-gene associations. Methods. 2015;74:83–9.
10. Rebholz-Schuhmann D, Arregui M, Gaudan S, Kirsch H, Jimeno A. Text processing through Web services: calling Whatizit. Bioinformatics. 2008; 24(2):296–8.
11. Doğan RI, Comeau DC, Yeganova L, and Wilbur WJ. Finding abbreviations in biomedical literature: three BioC-compatible modules and four BioC-formatted corpora. Database (Oxford) 2014; 2014: bau044, doi: 10.1093/database/bau044.
12. Kafkas Ş, Pi X, Marinos N, Talo' F, Morrison A, McEntyre J. Section level search functionality in Europe PMC. J Biomed Semant. 2015. doi: 10.1186/s13326-015-0003-7.
13. Liakata M, Saha S, Dobnik S, Batchelor C, Rebholz-Schuhmann D. Automatic recognition of conceptualization zones in scientific articles and two life science applications. Bioinformatics. 2012;28:991–1000.
14. Manning CD, Raghavan P and Schütze H. Introduction to Information Retrieval. NY: Cambridge University Press; 2008.
15. Rebholz-Schuhmann Dietrich, Kafkas Ş, Kim J-H, Li C, Jimeno Yepes A, Hoehndorf R, Backofen R, Lewin I. Evaluating gold standard corpora against gene/protein tagging solutions and lexical resources. J Biomed Semant. 2013;4:28. doi: 10.1186/2041-1480-4-28 .
16. Kafkas Ş, Dunham I, Parkinson H, and McEntyre J. Use of text mining for Experimental Factor Ontology coverage expansion in the scope of target validation. Proceedings of the Joint International Conference on Biological Ontology and BioCreative, Corvallis, Oregon, United States, August 1–4, 2016.
17. Yepes AJ and Karin Verspoor. Literature mining of genetic variants for curation: quantifying the importance of supplementary material. Database (Oxford). 2014: bau003. doi:https://doi.org/10.1093/database/bau003.
18. Kafkas Ş, Kim J-H, Pi X, Mcentyre J. Database citation in supplementary data linked to Europe PubMed Central full text biomedical articles. J Biomed Semant. 2015;6:1. doi:10.1186/2041-1480-6-1.