BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 5 2012, pages 709–714
doi:10.1093/bioinformatics/bts026

Data and text mining

Advance Access publication January 13, 2012

Mining and evaluation of molecular relationships in literature

Christian Senger[1,†], Björn A. Grüning[1,†], Anika Erxleben[1], Kersten Döring[1], Hitesh Patel[2], Stephan Flemming[1], Irmgard Merfort[2] and Stefan Günther[1,*]

[1]Pharmaceutical Bioinformatics, Institute of Pharmaceutical Sciences, Albert-Ludwigs-University, Hermann-Herder-Str. 9 and [2]Pharmaceutical Biology and Biotechnology, Institute of Pharmaceutical Sciences, Albert-Ludwigs-University, Stefan-Meier-Str. 19, D-79104 Freiburg, Germany

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Specific information on newly discovered proteins is often difficult to find in literature. Particularly if only sequences and no common names of proteins or genes are available, preceding sequence similarity searches can be crucial for the process of information collection. In drug research, it is important to know whether a small molecule targets only one specific protein or whether similar or homologous proteins are also influenced that may account for possible side effects.

Results: *prolific* (protein-literature investigation for interacting compounds) provides a one-step solution to investigate available information on given protein names, sequences, similar proteins or sequences on the gene level. Co-occurrences of UniProtKB/Swiss-Prot proteins and PubChem compounds in all PubMed abstracts are retrievable. Concise 'heat-maps' and tables display frequencies of co-occurrences. They provide links to processed literature with highlighted found protein and compound synonyms. Evaluation with manually curated drug–protein relationships showed that up to 69% could be discovered by automatic text-processing. Examples are presented to demonstrate the capabilities of *prolific*.

Availability: The web-application is available at http://prolific.pharmaceutical-bioinformatics.de and a web service at http://www.pharmaceutical-bioinformatics.de/prolific/soap/prolific.wsdl.

Contact: stefan.guenther@pharmazie.uni-freiburg.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 12, 2011; revised on December 16, 2011; accepted on January 9, 2012

1 INTRODUCTION

Collecting information on proteins in the large and fast growing body of available biomedical information can be an elaborate task, particularly if proteins are only rarely investigated, newly predicted or sequenced, or if no common names are available for proteins and genes. Researchers may want to search for potential lead structures for given targets, determine the function of the protein, or analyse side effects of drugs with regard to protein similarity. To gather needed information for proteins as possible targets for chemical compounds, there are a variety of databases available providing manually curated information (Aranda *et al.*, 2010; Günther *et al.*, 2008; Knox *et al.*, 2011; Kuhn *et al.*, 2010; Sharman *et al.*, 2011; Zhu *et al.*, 2010a). However, a need for information beyond the scope of the existing databases often requires researchers to find information in biomedical texts by searching PubMed (http://www.ncbi.nlm.nih.gov/pubmed) with MeSH terms. Although the texts are rather unstructured, PubMed is widely accepted as the most comprehensive and up-to-date source for biomedical information (Berardi *et al.*, 2008). While subsequent analysis of information found in texts seems inevitable, text- or data-mining tools greatly enhance the process of information collection (Alex *et al.*, 2008). A variety of tools can be found via the Critical Assessment of Information Extraction systems in Biology initiative (BioCreAtIvE, Krallinger *et al.*, 2008). Several tools focus on the recognition of proteins and genes in texts (Hur *et al.*, 2009) or extraction of information of relationships between biological entities (Rebholz-Schuhmann *et al.*, 2007; Zhu *et al.*, 2010b). Others focus on a convenient provision of mined data of several types like proteins, drugs and diseases via web service (Rebholz-Schuhmann *et al.*, 2008) or the co-occurrence of biological entities like EXCERPT (Mewes *et al.*, 2011), CoPub (Fleuren *et al.*, 2011) and CIL (Grüning *et al.*, 2011). Finding co-occurences of biological entities is a well-accepted method for finding associations of any type that may be hidden in context or implicit (Andronis *et al.*, 2011) and supports on-going research in Natural Language Processing. CIL was recently published, enabling users to search for compound structures, similar structures and names in all PubMed abstracts. Proteins referred to in those texts are related to the compounds. The relationships and their frequency are displayed as results.

prolific (protein-literature investigation for interacting compounds) closes the gap between protein information in literature and sequence information on proteins with a one-step solution. Thus, it complements CIL. It enables researchers to collect a variety of information on proteins without identifiers or synonyms. Similar sequences are searched and all compounds mentioned in the same context are identified. It provides all information including abstracts with highlighted found entities in a pre-calculated, swiftly accessible database visualizing data in a 'heat-map' with colours representing displayed values. Thus, *prolific* has a protein-centric point of view, in contrast to CIL with its compound-centric view, allowing to obtain new information on proteins, compounds and their relationships. However, this requires specialized data structures allowing for high speed data retrieval for the large amount of data. The interconnection of both applications results in an additional benefit, enabling users to switch between protein and compound literature research using result compounds

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

in CIL and result proteins in *prolific* as query compounds and proteins with minimal effort.

Possible fields of application of *prolific* are depicted with three paradigm scenarios: the literature search for new drugs or potential lead structures based on known drug–target interactions, the analyses of drugs' side effects, and the discovery of the function of newly sequenced genes.

2 DATA AND METHODS

2.1 Databases

All titles, abstracts, keywords and substance terms of articles contained in PubMed possessing abstracts (11.7 million out of 20.6 million) were made locally available in *prolific*. Furthermore, the local database contains all PubChem (Bolton *et al.*, 2008) compounds (35.4 million) with 606.3 million (partially overlapping) synonyms. Those are connected to their 28.3 million annotated parents (i.e. basic compounds without ions, etc.) and all UniProtKB/Swiss-Prot (Magrane and Consortium, 2011) protein IDs (i.e. from the manually annotated section of UniProt) with 2.0 million synonyms. Gene Ontology Annotation (GOA) terms (Barrel *et al.*, 2009) and their 'gene symbols' were attached to protein IDs. PubChem parents and GOA gene symbols were used to pool very similar compounds (e.g. diclofenac sodium and diclofenac potassium) or homologues of different organisms, respectively.

2.2 Finding similar proteins

By default, protein synonym searches in the database are conducted using the unmodified user input. If users decide to use an approximate search, wildcards are used as suffix or spelling correction is used to find misspelled names (Hunspell, http://hunspell.sourceforge.net). Sequence similarity is determined for protein and nucleotide sequences utilising BLAST+ (Camacho *et al*., 2009).

2.3 Finding biological entities in texts

Protein synonyms were searched using the Whatizit web services (Rebholz-Schuhmann *et al.*, 2008). Abstracts were retrieved from the local database and passed to Whatizit. Protein synonyms in abstracts and titles were annotated with protein IDs by Whatizit. Annotated texts were stored in the local database and parsed to store protein–article relationships after filtering with a protein 'stop word list'. It is composed of natural language words and words with a high frequency of occurrence which have to be filtered out because of unspecific meanings (e.g. 'And' for the 'Calmodulin-related protein 97A' or 'ANOVA' usually used in statistical context but also as abbreviation for 'RNA-binding protein Nova-2'). The list was generated after analysing synonyms of found compounds and their frequencies of occurrences in all PubMed articles.

Compound synonyms were searched in all PubMed articles with available abstracts following the synonym processing rules described in Hettne *et al.* (2009). Titles, abstracts, MeSH terms and substance names were indexed and searched with all compound synonyms provided by PubChem. Found article–compound relationships were stored in the database, after filtering with a compound stop word list analogous to the protein stop word list.

Protein–article–compound relationships were grouped in four grading classes: (i) co-occurrence of protein and compound in the abstract, (ii) co-occurrence in the same sentence, (iiia) co-occurrence in the same sentence enclosing a 'functional process' or 'molecular function' derived from the Gene Ontology (Ashburner *et al.*, 2000) and (iiib) co-occurrences in a sentence enclosing curated 'relationship' verbs, derived by analysing approximately 2 million randomly selected abstracts for verbs occurring with compounds and proteins in the same sentence. Those relationships were assembled and stored de-normalized in a NoSQL database, allowing fast access to the large amount of data.

To obtain precision, recall, and F-Score, a test set of 120 articles was randomly selected from the SuperTarget database (Hecker *et al.*, 2012). SuperTarget provides manually curated compound–target relations with associated PubMed references. To include all compounds (also non-drugs) and all proteins (also non-targets) the abstracts were re-annotated by four Biologists and Bioinformaticians.

Those abstracts were searched in the *prolific* database. Precision and recall were calculated with the resulting found proteins and compounds and the annotations of the abstracts.

2.4 Quality assessment

To evaluate the results of *prolific*, DrugBank target proteins, drugs and drug–target relationships were used. DrugBank provides about 1500 FDA approved drugs (and additional 4000 experimental drugs), 3800 targets and about 1400 drug–target relationships with approved drugs. *prolific* was queried using all target proteins from DrugBank. Subsequently, result data were analysed to see if drug–target relationships of DrugBank are contained and at which rank they are placed in the result list. Reasons for missed drug-targets were assessed quantitatively where possible, or qualitatively otherwise.

Hyperlinks to all downloaded data, synonym lists and stop word lists used in *prolific* are provided in the Supplementary Material. Text indexing and search took place using Xapian (Xapian 1.2.0, http://xapian.org). Data was stored in a PostgreSQL database (PostgreSQL 8.4.8, PostgreSQL Global Development Group). De-normalized data was stored in a NoSQL database (MongoDB 1.8.1, 10gen, Inc., New York, USA). A fixed update schedule of the de-normalized relation will be performed every 6 months.

3 RESULTS AND DISCUSSION

3.1 Workflow

The *prolific* workflow and the search process is not visible for the user, but can be controlled by adjusting search parameters (Fig. 1). A query can be started with names, IDs, or sequences. If the exact name is unknown for a search with names, an approximate search can be triggered. In the case of ID or name searches, the database will be queried for corresponding sequences. A similarity search is conducted with query sequences or sequences obtained from the database within the set of protein sequences available in the database. The database is queried for all PubMed articles containing found proteins. Compounds mentioned in those abstracts found by *prolific* are retrieved from the pre-calculated database. Subsequently, results will be evaluated, counted and visualized. An option to read the texts with highlighted found entities is provided to the user. If too few or too many results are found, search parameters can be adjusted for e.g. lower similarity search thresholds or to narrow searches for co-occurrences with protein and compound in the same sentence, enclosing a 'functional' GO term or 'relationship' verb.

3.2 Relationships of protein synonyms, compound synonyms and articles

In all abstracts, 8.8 million relationships could be found between protein synonyms and articles. Between compounds and articles 12.5 million relationships were found. After assembling, 309.5 million protein–article–compound relationships were available. Besides the web -interface, data can be accessed via web service.

Precision and recall were calculated as 94 and 72% (F-Score: 82%).

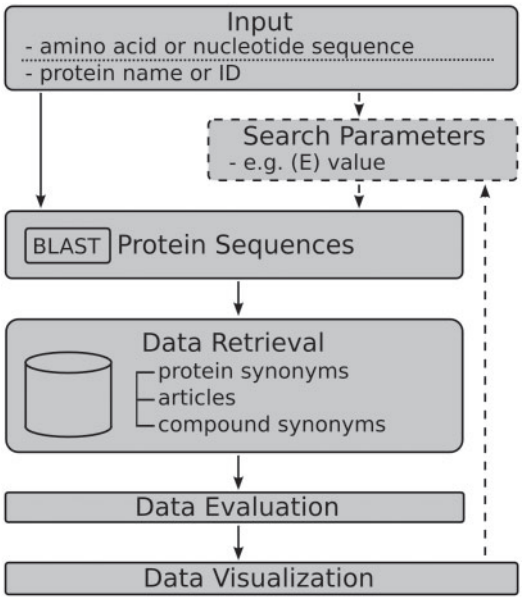

Fig. 1. Flowchart of searches with *prolific*. Dashed arrows indicate the possibility to adapt parameters, e.g. if no results are found with default settings or to narrow searches for co-occurrences of protein and compound in the same sentence.

3.3 Data validation

Using the targets of all drug–target relationships of DrugBank as query proteins in *prolific*, 56% of all DrugBank relationships could be found as protein–compound relationships (Table 1). Considering that *prolific* does not include biologicals (e.g. antibodies and nucleic acids) and some proteins or drugs could not be found in any abstract, 69% of the relationships were identified. Qualitative analyses of 50 out of the 357 relationships not found with *prolific* revealed that in 132 analysed full articles 78 abstracts contained a protein synonym deviating too much from UniProt synonyms, 48 contained target synonyms only in the article body, and six targets were DNA or RNA. The 102 protein or compound synonyms not found in the abstracts separately were e.g. too exact ('gamma-aminobutyric acid receptor subunit rho-1' instead of 'GABA') or had only IDs (i.e. alpha-numeric codes) in the respective synonym database. If the relationships were found, 70% of the drugs were found as compounds in the first 100 places of the *prolific* result order (Table 2). Of all compounds found with *prolific* for which the query proteins are available in both databases, on average 23% (Median, interquartile range 17–27, FDA approved drugs only) and 32% (Median, interquartile range 27–40, FDA approved and experimental drugs) were contained in DrugBank.

3.4 Visualization

Results of *prolific* queries are presented in a heat-map, with colours indicating the frequencies at which protein synonyms co-occur with compound synonyms in abstracts (Fig. 2). A 'noise filter' is applied to the table to display only those relationships occurring more frequently than a given threshold. The threshold can be adjusted in the parameter settings. Columns represent proteins found with the query sequence, name or ID and similar proteins found via BLAST. Proteins are sorted by similarity. Furthermore, proteins are pooled by their gene symbol inherited from GOA. This allows for a more compact overview and merges proteins found independently of organisms with identical results. Rows represent compounds (i.e. their synonyms) found in the abstracts assigned to found proteins. The order is given by the row's sum of abstracts. Compounds are pooled by their parents for a better overview and because synonyms of some derivatives may largely overlap and thus yield equal results.

Synonyms and sketches of compounds as well as (E) values and synonyms of proteins are provided by hovering over row or column headers with the mouse, hyperlinks lead to UniProt and PubChem. Table cells are linked to article lists where users can inspect title, abstract, keywords and substance terms in which found entities are highlighted, and which are linked to UniProt, PubChem and PubMed (Fig. 3). Moreover, the article lists show the numbers for restricted searches e.g. co-occurrences in the same sentence. In case a user decides for a conventional result table, an alternative view can be selected on the results page.

Table 1. DrugBank drug-target relationships in *prolific* searched with the target as query protein

Category	Hits, *N*(%)
Found relationships	795 (56)
Protein–compound combination not found in abstracts	357 (25)
Drug of relationship is a biological[a]	157 (11)
Protein of relationship not found in any abstract	65 (5)
Drug of relationship not found in any abstract	37 (3)
All drug-target relationships[b]	1411 (100)

[a]Protein or nucleic acid.
[b]With FDA-approved drugs.

Table 2. Result table placement of DrugBank drug-target relationships found with *prolific* searched with the target as query protein

Rank	Hits, *N*(%)
1	63 (8)
1–10	268 (34)
1–100	553 (70)
1–1000	755 (95)
1–∞	795 (100)

3.5 prolific-CIL interplay

Compounds mentioned in *prolific* results can be clicked to initiate subsequent CIL searches. Vice versa, CIL has been enhanced and now allows for querying *prolific* with proteins from CIL's result set. Additionally, menus were added to the heat-maps' column headers enabling users to query *prolific* repeatedly with homologues of *prolific*'s results and analogously CIL with similar compounds of CIL's results. CIL has been updated with the newly generated data with regard to stop word lists, current compound synonyms, protein synonyms and abstracts.

3.6 Result evaluation

The results of *prolific* reflect a large part of the DrugBank drug-targets. Displaying the first 100 (default) or 1000 (optional) found

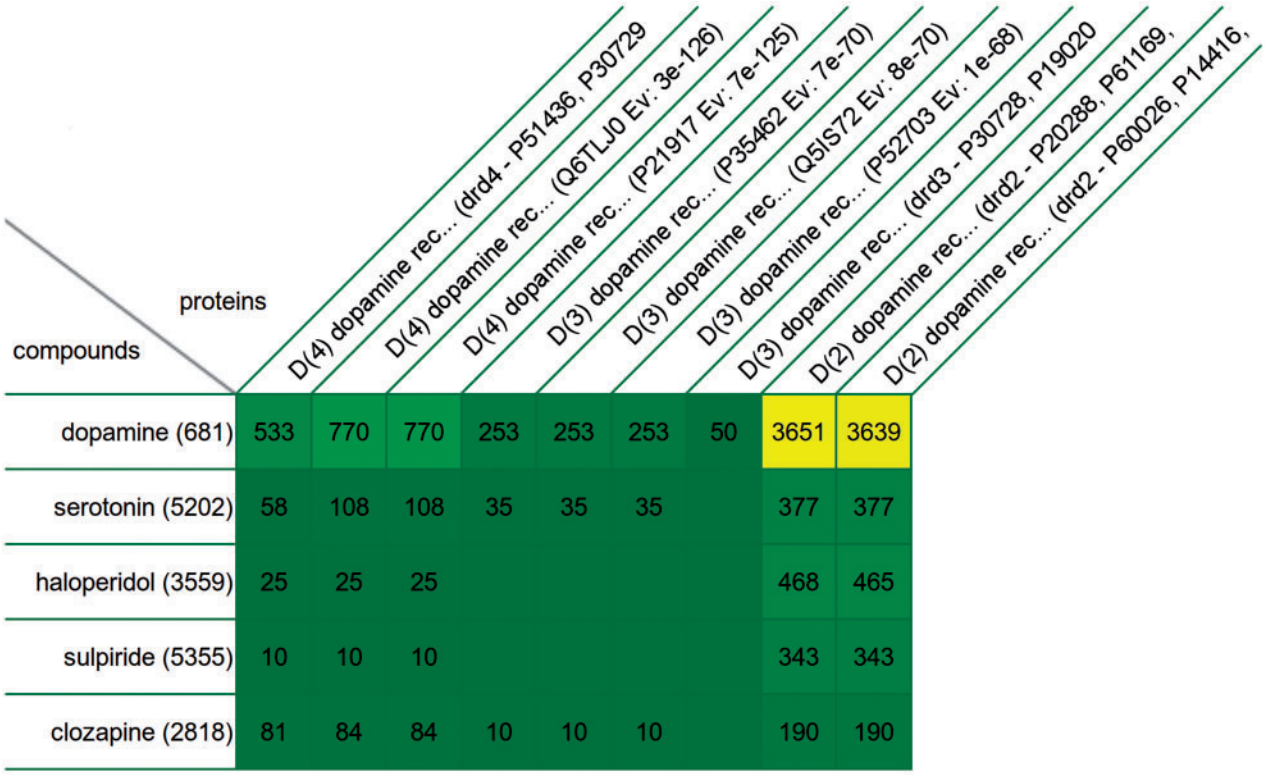

Fig. 2. Result visualization of a search with 'D(4) dopamine receptor' with a heat-map in *prolific*.

compounds ordered by frequency of occurrence covers most of the drug–target relationships, still allowing for a comprehensive overview. Missing relationships are mainly caused by synonyms strongly deviating in texts from those in databases and relationships mentioned in the article body only, not in the abstract. The latter could be addressed in the future by using full texts increasingly available via 'Open Access' initiatives, while the former one points out the need for good vocabularies not only containing scientific or common names but also e.g. common abbreviations, which in turn must be specific enough. Almost a quarter up to a third of the compounds found with *prolific* is a drug in DrugBank with described drug–target relationships. Further investigation should analyse in detail the remaining part of found compounds. However, those results highlight the capabilities of *prolific* in finding and visualizing literature data containing protein–drug as well as other protein–compound co-occurrences.

3.7 Paradigm scenarios

To demonstrate the capabilities of the database, three paradigm scenarios were analysed in which *prolific* supports and expedites research.

Potential lead structure search: Brd4 is a new cancer-related drug-target (Blobel *et al.*, 2011). Using the sequence of Brd4 as query, *prolific* shows all related proteins containing a bromodomain and the related known inhibitors, e.g. Brd2 and doxorubicin. These compounds could be considered as possible lead compounds for the screening of specific Brd4 inhibitors. Direct links are provided to related literature.

Protein function prediction: Q813T7 is a predicted protein of *Bacillus cereus* with some evidence at protein level. Using the sequence as query in *prolific* shows that ATP and cob(I)alamin are the most mentioned compounds in literature related to homologous proteins. Estimating the function on the basis of sequence similarity with InterPro (Hunter *et al.*, 2009), the protein is predicted as a cob(I)alamin adenosyltransferase using ATP as a co-substrate. *prolific* highlights the related literature that describes the putative molecular function in a single step.

Side effect prediction: the dopamine-4-receptor is a main target for drugs against schizophrenia. Searching for related compounds with *prolific* (Fig. 2) reveals that dopamine, serotonin and clozapine are most mentioned in the same context of the receptor. Furthermore, many abstracts exist that describe the relationship of the compounds to the dopamine-2-receptor and the dopamine-3-receptor. Found literature shows that those additional target receptors could be considered for side effects of the anti-schizophrenia drug clozapine.

3.8 Limitations

The presented approach has also some limitations. First, quality of the data strongly depends of the quality of PubChem and UniProt synonyms, which may vary with each update. However, even if the main focus is not for text processing purposes, vocabulary extension and curation for protein and compound synonyms is ongoing and vocabularies can be applied subsequently in *prolific*, further enhancing results. Second, the data in general only represents co-occurrences of protein synonyms, compound synonyms and GO

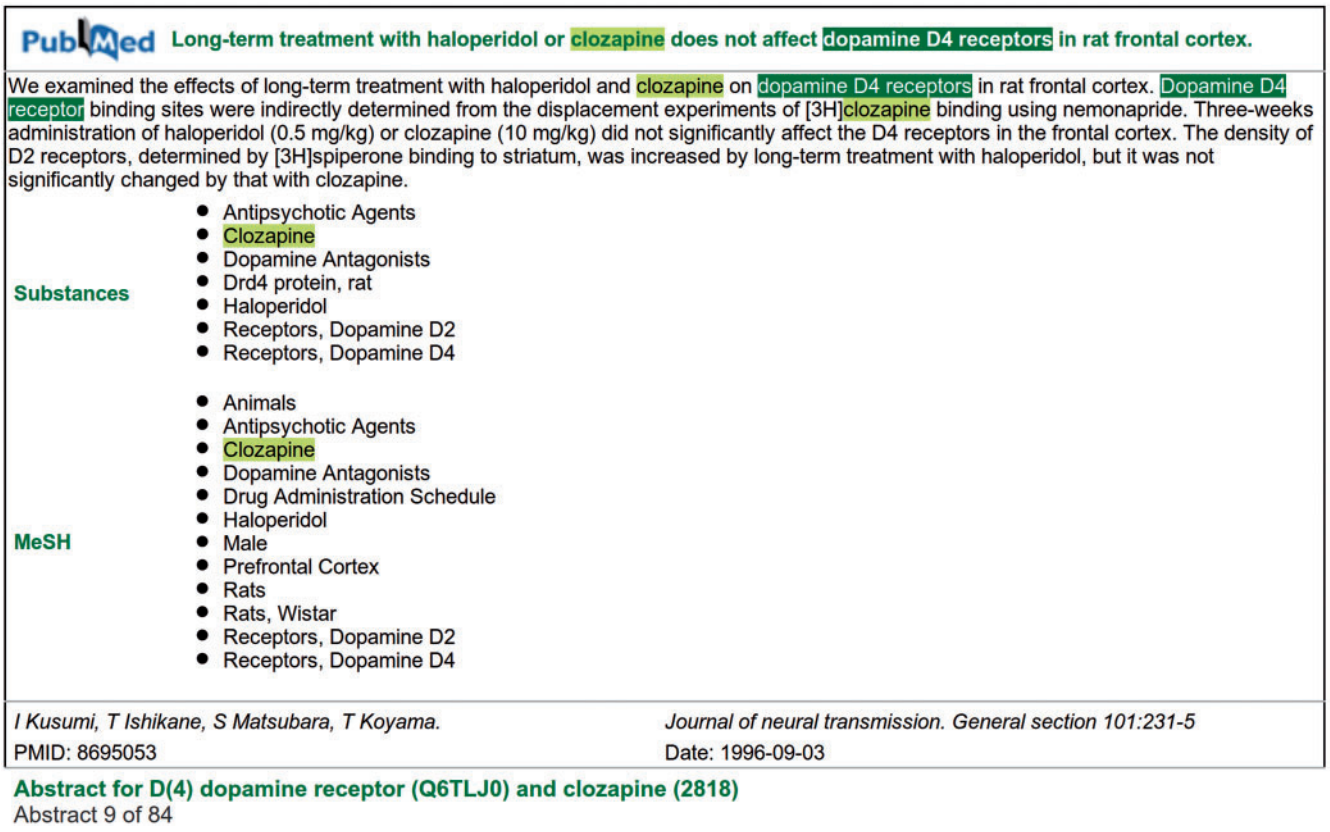

Fig. 3. Visualization and highlighting of found articles in *prolific*.

terms or curated verbs in titles, abstracts or sentences, keywords and substances. Thus, a subsequent curation of result data is necessary. Furthermore, entities in the body of articles are not discovered, yet.

3.9 Future developments

Further developments include protein–protein relationships and the processing of full articles which are freely available. Furthermore, extensions are in preparation to provide tools for interactive curation, graphical mapping of data to metabolic pathways and statistical analyses. Future statistical analyses may also cover trends in publications related to protein–compound relationships.

4 CONCLUSIONS

prolific supports researchers looking for potential drug targets, side effects and protein functions by providing a comprehensive overview of proteins, homologues and compounds mentioned in the same texts. Possibly, the combined use of other available independent tools for protein-similarity searches, text-mining, filtering and co-occurrence detection could provide similar results. However, this would mean researchers would have to transform data, reapply several tools a number of times, and prepare visualizations on their own. Those shortcomings are eliminated by *prolific* with a one-step solution and a comprehensive result overview, while maintaining the advantages of well-defined interfaces by supplying a web service.

Funding: Excellence Initiative of the German Federal and State Governments (ZUK 43); and by the German National Research Foundation (DFG, LIS 45).

Conflict of Interest: none declared.

REFERENCES

Alex,B. *et al.* (2008) Assisted curation: does text mining really help? *Pac. Symp. Biocomput.*, 556–567.

Andronis,C. *et al.* (2011) Literature mining, ontologies and information visualization for drug repurposing. Brief. Bioinform., **12**, 357–368.

Aranda,B. *et al.* (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38** (Database issue), D525–D531.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Barrell,D. (2009) The GOA database in 2009-an integrated Gene Ontology Annotation resource. *Nucleic Acids Res.*, **37** (Database issue), D396–D403.

Berardi,M. *et al.* (2008) Biomedical literature mining for biological databases annotation. In Giannopoulou,E.G. (ed), *Data Mining in Medical and Biological Research*, InTech – Open Access Publisher, University Campus STeP Ri, Slavka Krautzeka 83/A, 51000 Rijeka, Croatia, pp. 267–290.

Blobel,G.A. *et al.* (2011) Short hairpin RNA screen reveals bromodomain proteins as novel targets in acute myeloid leukemia. *Cancer Cell*, **20**, 287–288.

Bolton,E.E. *et al.* (2008) Chapter 12 - PubChem: Integrated platform of small molecules and biological activities. *Ann. Rep. Comput. Chem.*, **4**, 217–241.

Camacho,C. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Fleuren,W.W.M. *et al.* (2011) CoPub update: CoPub 5.0 a text mining system to answer biological questions. *Nucleic Acids Res.*, **39** (Web Server issue), W450–W454.

Grüning,B.A. *et al.* (2011) Compounds in literature (CIL): screening for compounds and relatives in PubMed. *Bioinformatics*, **27**, 1341–1342.

Günther,S. *et al.* (2008) SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res.*, **36** (Database issue), D919–D922.

Hecker,N. *et al.* (2012) SuperTarget goes quantitative: update on drug-target interactions. *Nucleic Acids Res*., **40,** D1113–D1117.

Hettne,K.M. *et al.* (2009) A dictionary to identify small molecules and drugs in free text. *Bioinformatics*, **25**, 2983–2991.

Hunter,S. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res*, **37** (Database issue), D211–D215.

Hur,J. *et al.* (2009) SciMiner: web-based literature mining tool for target identification and functional enrichment analysis. *Bioinformatics*, **25**, 838–840.

Knox,C. *et al.* (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.*, **39** (Database issue), D1035–D1041.

Krallinger,M. *et al.* (2008) Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. *Genome Biol*., **9** (Suppl. 2) S1.

Kuhn,M. *et al.* (2010) STITCH 2: an interaction network database for small molecules and proteins. *Nucleic Acids Res.*, **38** (Database issue), D552–D556.

Magrane,M. and Consortium,U. (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database,* **2011,** bar009; doi:10.1093/database/bar009.

Mewes,H.W. *et al.* (2011) MIPS: curated databases and comprehensive secondary data resources in 2010. *Nucleic Acids Res*, **39** (Database issue), D220–D224.

Rebholz-Schuhmann,D. *et al.* (2008) Text processing through web services: calling Whatizit. *Bioinformatics*, **24**, 296–298.

Rebholz-Schuhmann,D. *et al.* (2007) EBIMed-text crunching to gather facts for proteins from Medline. *Bioinformatics*, **23**, e237–e244.

Sharman,J.L. *et al.* (2011) IUPHAR-DB: new receptors and tools for easy searching and visualization of pharmacological data. *Nucleic Acids Res.*, **39** (Database issue), D534–D538.

Zhu,F. *et al.* (2010a) Update of TTD: Therapeutic Target Database. *Nucleic Acids Res.*, **38** (Database issue), D787–D791.

Zhu,Q. *et al.* (2010b) WENDI: A tool for finding non-obvious relationships between compounds and biological properties, genes, diseases and scholarly publications. *J. Cheminform.*, **2**, 6.