W422–W429 *Nucleic Acids Research, 2014, Vol. 42, Web Server issue* *Published online 16 May 2014*
doi: 10.1093/nar/gku432

Alkemio: association of chemicals with biomedical topics by text and data mining

José A. Gijón-Correas, Miguel A. Andrade-Navarro and Jean F. Fontaine*

Computational Biology and Data Mining, Max Delbrück Center for Molecular Medicine, Berlin, 13125 Berlin, Germany

Received January 10, 2014; Revised April 29, 2014; Accepted May 2, 2014

ABSTRACT

The PubMed® database of biomedical citations allows the retrieval of scientific articles studying the function of chemicals in biology and medicine. Mining millions of available citations to search reported associations between chemicals and topics of interest would require substantial human time. We have implemented the Alkemio text mining web tool and SOAP web service to help in this task. The tool uses biomedical articles discussing chemicals (including drugs), predicts their relatedness to the query topic with a naïve Bayesian classifier and ranks all chemicals by *P*-values computed from random simulations. Benchmarks on seven human pathways showed good retrieval performance (areas under the receiver operating characteristic curves ranged from 73.6 to 94.5%). Comparison with existing tools to retrieve chemicals associated to eight diseases showed the higher precision and recall of Alkemio when considering the top 10 candidate chemicals. Alkemio is a high performing web tool ranking chemicals for any biomedical topics and it is free to non-commercial users. Availability: http://cbdm.mdc-berlin.de/~medlineranker/cms/alkemio.

INTRODUCTION

Associating chemicals with biological functions (e.g. disease, metabolism or proliferation) allows biomedical researchers to understand or predict roles of chemicals in biological pathways and diseases. As PubMed (1), the main database of biomedical citations, contains millions of entries, it might not be possible to manually derive such associations. Automatically mining the literature using text mining algorithms offers comprehensive and predictive searches.

Using predictive text mining algorithms through web interfaces or web services could help biomedical scientists in various applications. For example, it could be useful in defining a list of thousands of chemicals for a screening experiment, in extending available data on molecular pathways, in retrieving the most relevant chemicals related to a disease or in predicting chemicals for poorly studied diseases. Available text mining tools allow computing association of chemicals with defined topics. However, query topics could be restricted to given biological dictionaries (e.g. CoPub (2)), and tools handling free-text queries could be limited by computing resources (e.g. PolySearch (3)) or may not rank candidate chemicals (e.g. EBIMed (4)).

Based on gene-citation links, it was demonstrated previously that fast and accurate ranking of all genes of a species for given biomedical topics can be achieved by the Génie text mining algorithm (5). Génie associates genes to biomedical topics by classifying the gene-related PubMed citations using the MedlineRanker algorithm (6). The MedlineRanker document classification algorithm implements a naïve Bayesian classifier that models biases in word usage in a set of PubMed citations related to a topic of interest. Applying a similar approach than Génie to rank chemicals may produce accurate results and be of broad interest to biomedical researchers because query topics are not limited and simply defined as sets of related citations.

Based on Génie, MedlineRanker and chemical annotations of PubMed citations, we have implemented the Alkemio web tool that ranks all annotated chemicals (including drugs) in the literature for any biomedical topics. The method was benchmarked using known molecular pathway-related chemicals and compared to existing tools using disease-related chemicals. Alkemio is freely available to non-commercial users.

MATERIALS AND METHODS

Implementation

The Alkemio algorithm uses PubMed citations with English abstracts. PubMed citations are manually annotated using the Medical Subject Headings thesaurus (MeSH®) (1). MeSH provides a controlled vocabulary organized as a tree with 16 branches under the root level such as anatomy, organisms, diseases, and chemicals and drugs. For each PubMed citation, two sets of MeSH terms were extracted: the main MeSH terms used to characterize the content of the citation (from tags MeshHeadingList) and the chemical MeSH terms with non-zero registry numbers (from tags

*To whom correspondence should be addressed. Tel: +49 30 9406 4307; Fax: +49 30 9406 4240; Email: jean-fred.fontaine@mdc-berlin.de

ChemicalList). The latter is used to define a list of chemical names to be ranked; the former and the latter are used to select a set of citations defining the topic of interest as an alternative to a free-text query to PubMed. The following types of chemicals were not used: enzymes and coenzymes (MeSH tree number D08), proteins (MeSH tree number D12.776) and multiprotein complexes (MeSH tree number D05.500). As of 3 April 2014, we could extract a total of 8 877 369 chemical–PubMed links for 50 394 chemicals.

Alkemio retrieves PubMed citations related to the input chemical list using their annotation by chemical MeSH terms and classifies them using MedlineRanker as related to the input topic or not. Results of Alkemio are summarized by chemical and sorted by false discovery rates (FDRs). *P*-values and FDRs are computed from a simulation on 10 000 random citations. The approach is similar to the ranking of genes implemented in the Génie algorithm.

Alkemio uses the MedlineRanker document classification algorithm that was previously published (see for details (6,7)). Briefly, nouns are extracted by a part-of-speech processor from titles and abstracts downloaded from the PubMed XML files (tags ArticleTitle and AbstractText). A stop word list is used to filter out common and non-informative nouns. Provided a training set composed of citations related to a topic of interest and a background set composed of random citations, MedlineRanker produces a weighted list of discriminative nouns. Weights are defined by naïve Bayesian statistics. This weighted list of nouns is then used to score other citations; each noun matching one in the list would add its weight (only once) to the total score. Scores are converted into *P*-values thanks to random simulations and a *P*-value cutoff is used to decide if a citation is related to the topic of interest or not. If the query topic is defined by a MeSH term and the same MeSH term is found in the manual annotations of a citation, this citation will be automatically set as relevant to the topic by associating to it a *P*-value equal to zero.

Web pages were built with WordPress 3.8 or programmed using HTML 4, JavaScript, PHP and Perl 5.10.1. Data were stored in a MySQL 5.1.41 database. Web pages were tested using various web browsers (Firefox 26, Chrome 31, Internet Explorer 10 and Safari 7) and operating systems (Ubuntu 12.04 and 12.10, Windows XP and Seven and Mac OS X 10.9).

Pathway benchmark

To benchmark Alkemio, data about all ($n = 7$) human pathways involving at least 10 chemicals identified by a Chemical Abstracts Service (CAS) registry number were downloaded from WikiPathways (8) on 11 March 2013. Alkemio ranked all chemicals for each pathway with default parameters (using as input the result of a query to PubMed, using the quoted pathway name, ranking chemicals from citations from the last 3 years, using a *P*-value cutoff for abstract selection equal to 0.01 and a FDR cutoff equal to 0.001). Alkemio ranks were transformed into scores using the following formula: score = sqrt(1/rank). Alkemio returned a total of 10 356 candidate chemicals. Registry numbers were either converted to CAS numbers if provided as FDA SRS's UNIIs (File UNIIs 18Nov2013 Records.txt from http://fdasis.nlm.nih.gov/srs/download/srs/UNII_Data.zip) ($n = 1803$) or removed from the candidate list if provided as Enzyme Commission numbers ($n = 28$). Because there were very few known positives in comparison to the number of candidates, the ranking performance was evaluated with the QiSampler tool (9). QiSampler parameters were set to 1000 repetitions and a sampling rate of 50%. Note that due to the low number of positive cases, control curves may not start as theoretically expected at coordinates (0,0) and end at (1,1).

Comparison

In order to compare Alkemio with existing tools (FACTA (10) and PolySearch (3)), we used data from the Comparative Toxicogenomics Database (CTD) as gold standard (11). FACTA and PolySearch were chosen because they accept any topic as input and they rank the candidate chemicals. The CTD data were downloaded on 4 September 2013, and contained a total of 56 907 chemicals with a CAS number and 67 817 manually set associations between chemicals and diseases. As input to query the tools, we selected eight different diseases represented by the following MeSH terms: Alzheimer's disease (AD) (72 associations with chemicals), anemia (215 associations), arthritis rheumatoid (77 associations), diabetes mellitus (57 associations), meningitis (40 associations), pancreatitis (83 associations), porphyrias (12 associations) and purpura (37 associations). Due to the low minimal number of known associations for a selected disease (minimum = 12) and the variable amount of candidates returned by the different tools (minimum = 33), we focused the comparison to the top 10 candidates retrieved from each tool, although we also analyzed the top 100 if enough candidates were available.

The CAS numbers were used for mapping the candidate chemicals to CTD. The tools did not report only CAS numbers but also FDA SRS's UNIIs (Alkemio), KEGG (FACTA), DrugBank (FACTA and PolySearch) and HMDB identifiers (PolySearch). Conversion to CAS was done using data from FDA SRS, DrugBank (http://www.drugbank.ca/system/downloads/current/drugbank.txt.zip and http://www.drugbank.ca/system/downloads/current/drug_links.csv.zip) (12), HMDB (http://www.hmdb.ca/downloads/hmdb_metabolites.zip) (13), and results of a manual query to the Chemical Translation Service (14) to convert KEGG identifiers. The percentage of candidate chemicals associated to a CAS number was 97.7 ($n = 20\,277$) for Alkemio, 90.7 ($n = 18\,978$) for FACTA and 87.5 ($n = 1101$) for PolySearch. Then, we (i) removed chemicals not mapped to one of the 56 907 chemicals having a CAS registry number in CTD from the lists of candidates, (ii) removed duplicate candidates for each tool and disease keeping the one with the maximal score and (iii) identified those candidates that agreed with CTD chemical–disease associations as true positives.

Tools' parameters were set to retrieve as many candidates as possible in a reasonable time. Alkemio was queried on 2 April 2014, using its SOAP web service to rank chemicals found in articles published during the last 3 years with a *P*-value cutoff for abstract selection set to 0.01 and a FDR cutoff set to 1. Alkemio's FDR was transformed into a score

as follows: score = sqrt(1/rank). FACTA was queried on 2 April 2014, manually to retrieve all drugs and compounds associated to the query sorted by Pointwise Mutual Information. FACTA's scores were transformed to be only positive, keeping the same ordering of the candidates (i.e. the absolute value of the minimal score + 1 was added to all scores). PolySearch was queried on 5 November 2013, manually to retrieve all drugs and metabolites associated to a query disease. Polysearch's automated disease synonym list was used; searches were limited to a maximum of 10 000 abstracts from the past 1 year of the PubMed database and the minimum number of citations was set to 1. The PolySearch's rscore was used as the main score for the benchmark. Due to very slow computations or technical problems, we had to limit queries to PolySearch to using <10 000 abstracts, and drug candidates for AD and arthritis rheumatoid used the default parameters (all available PubMed abstracts, but retrieved candidates limited to 2000).

RESULTS

Graphical user interface

The Alkemio algorithm that ranks chemicals for biomedical topics was implemented as a web server freely available to non-commercial users. It is composed of an input form, documentation and information pages. The topic of interest is defined by a list of biomedical citations internally represented by their PubMed identifiers (PMIDs). The web interface provides a facility to define this list either from a list of exact MeSH terms (e.g. 'Alzheimer Disease', 'Phosphorylation' or 'Stem Cells') or from a free-text query to PubMed (e.g. 'early onset Alzheimer's disease'). All chemicals linked to PubMed citations published during the last year can be ranked. As of on 2 April 2014, this set of citations represents 8073 chemicals and 116 458 chemical–PMID links. The search can be extended to 3 years to rank 16 017 chemicals using 828 992 chemical–PMID links (computing time: 2–3 min). Alternatively, a smaller list of chemicals to be ranked can be defined with exact names from the MeSH database. In this case, all their chemical–PMID links are used (no limitation to the last 3 years). PubMed citations are downloaded and updated daily, while chemical–PMID links are computed monthly.

Two cutoff values can be defined by the user. First, the FDR cutoff limits the displayed chemicals by significance. Second, the *P*-value cutoff for the selection of abstracts is used by a naïve Bayesian classifier to decide if a PubMed citation linked to a chemical is relevant to the topic of interest. The lower this value, the higher the precision, but the lower the recall of relevant citations. A default value is set to 0.01, but different ways to define it automatically are proposed optimizing the precision or the *F* score from internal statistical simulations.

The output page contains mainly the ranked list of chemicals that shows for each candidate a registry number, a chemical name linked to an appropriate external MeSH entry page, the number of associated PMIDs, the number of PMIDs classified as relevant to the topic (hits), the FDR and a list of 10 top PMIDs having the lowest classification *P*-values (Figure 1). Discriminative words and weights used by the text mining classifier are shown in a table. A file download section at the bottom provides results of the analysis as text files with tab-separated values. Displayed PMIDs are linked to a new window showing related title, abstract, journal, authors, links to PubMed or full texts and MeSH annotations. Discriminative words found in this window and names of the target chemical are highlighted in brown or rose, respectively.

Web service

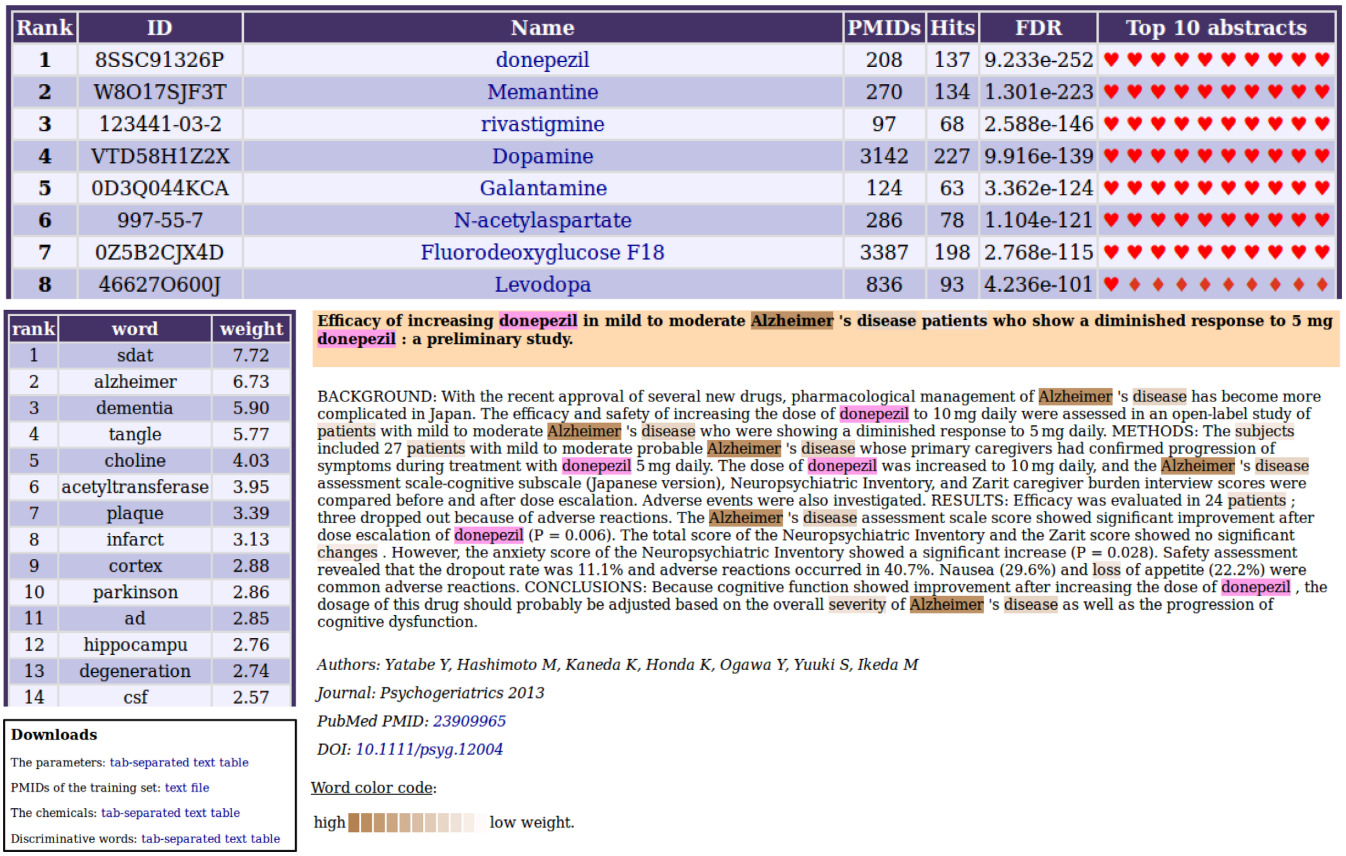
The Alkemio's SOAP web service is designed for programmatic access over the network. For example, computations from Alkemio could be retrieved from another program or a user could run batch queries. Using common programming languages, a programmer can connect to the service using a WSDL descriptor (http://cbdm.mdc-berlin.de/tools/medlineranker/soap/wsdl/AlkemioSOAPbeta1.wsdl) and call the main function with appropriate parameters (see documentation on the web page). The output of Alkemio is returned as XML data, which is valid against the Alkemio's DTD (http://cbdm.mdc-berlin.de/tools/medlineranker/soap/dtd/alkemio_beta1.dtd). The SOAP web service can be used without programming with a client Perl script as command line interface accepting files as inputs (e.g. for topics or for a list of chemicals).

Benchmarks

Alkemio results for seven human pathways were compared to chemicals known to have a role in these pathways in WikiPathways (Table 1). Although few chemicals were reported in each pathway (between 12 and 14), Alkemio returned many candidates (between 545 and 2431). Thus, the relevance of scores given to true positives was evaluated by areas under receiver operating characteristic (ROC) curves using a sampling strategy (9). Areas ranged from 73.6 to 94.5% (Table 1 and Figure 2).

We have also compared the performance of Alkemio with FACTA and PolySearch when retrieving manually annotated known associations of chemicals and eight diseases from the CTD database. In this comparison, Alkemio queried using the PubMed option retrieved in average 897 candidates for a disease (minimum = 517 and maximum = 1561), Alkemio queried using the MeSH option retrieved in average 1281 candidates for a disease (minimum = 684 and maximum = 3155), FACTA 1474 (minimum = 674 and maximum = 2448) and PolySearch 109 (minimum = 33 and maximum = 374). We compared the precision in the top 10 candidates for each tool and disease (Figure 3). Alkemio's precision ranged from 0.20 to 0.70 or 0.10 to 0.80 when it was queried using the PubMed or the MeSH option, respectively. Compared to FACTA and PolySearch, it had the best reported precision in all eight (100%) cases (six times alone and two times together with another tool) or in seven (87.5%) cases (six times alone and one time together with another tool) when it was queried using the PubMed or the MeSH option, respectively.

Then, we compared the precision in the top 100 candidates (Supplementary Figure S1). Compared to FACTA and PolySearch, Alkemio had the best precision five or four times when it was queried using the PubMed or the MeSH



Rank	ID	Name	PMIDs	Hits	FDR	Top 10 abstracts
1	8SSC91326P	donepezil	208	137	9.233e-252	♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥
2	W8O17SJF3T	Memantine	270	134	1.301e-223	♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥
3	123441-03-2	rivastigmine	97	68	2.588e-146	♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥
4	VTD58H1Z2X	Dopamine	3142	227	9.916e-139	♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥
5	0D3Q044KCA	Galantamine	124	63	3.362e-124	♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥
6	997-55-7	N-acetylaspartate	286	78	1.104e-121	♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥
7	0Z5B2CJX4D	Fluorodeoxyglucose F18	3387	198	2.768e-115	♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥ ♥
8	46627O600J	Levodopa	836	93	4.236e-101	♥ ♦ ♦ ♦ ♦ ♦ ♦ ♦ ♦ ♦

rank	word	weight
1	sdat	7.72
2	alzheimer	6.73
3	dementia	5.90
4	tangle	5.77
5	choline	4.03
6	acetyltransferase	3.95
7	plaque	3.39
8	infarct	3.13
9	cortex	2.88
10	parkinson	2.86
11	ad	2.85
12	hippocampu	2.76
13	degeneration	2.74
14	csf	2.57

Downloads

The parameters: tab-separated text table

PMIDs of the training set: text file

The chemicals: tab-separated text table

Discriminative words: tab-separated text table

Efficacy of increasing donepezil in mild to moderate Alzheimer 's disease patients who show a diminished response to 5 mg donepezil : a preliminary study.

BACKGROUND: With the recent approval of several new drugs, pharmacological management of Alzheimer 's disease has become more complicated in Japan. The efficacy and safety of increasing the dose of donepezil to 10 mg daily were assessed in an open-label study of patients with mild to moderate Alzheimer 's disease who were showing a diminished response to 5 mg daily. METHODS: The subjects included 27 patients with mild to moderate probable Alzheimer 's disease whose primary caregivers had confirmed progression of symptoms during treatment with donepezil 5 mg daily. The dose of donepezil was increased to 10 mg daily, and the Alzheimer 's disease assessment scale-cognitive subscale (Japanese version), Neuropsychiatric Inventory, and Zarit caregiver burden interview scores were compared before and after dose escalation. Adverse events were also investigated. RESULTS: Efficacy was evaluated in 24 patients ; three dropped out because of adverse reactions. The Alzheimer 's disease assessment scale score showed significant improvement after dose escalation of donepezil (P = 0.006). The total score of the Neuropsychiatric Inventory and the Zarit score showed no significant changes . However, the anxiety score of the Neuropsychiatric Inventory showed a significant increase (P = 0.028). Safety assessment revealed that the dropout rate was 11.1% and adverse reactions occurred in 40.7%. Nausea (29.6%) and loss of appetite (22.2%) were common adverse reactions. CONCLUSIONS: Because cognitive function showed improvement after increasing the dose of donepezil , the dosage of this drug should probably be adjusted based on the overall severity of Alzheimer 's disease as well as the progression of cognitive dysfunction.

Authors: Yatabe Y, Hashimoto M, Kaneda K, Honda K, Ogawa Y, Yuuki S, Ikeda M

Journal: Psychogeriatrics 2013

PubMed PMID: 23909965

DOI: 10.1111/psyg.12004

Word color code:

high low weight.

Figure 1. Alkemio's output. As example, Alkemio was queried to retrieve AD chemicals on 31 March 2014. The topic was queried as a MeSH term (Alzheimer Disease), citations from the last 3 years were used, the *P*-value cutoff for abstract selection equaled 0.01 and the FDR cutoff equaled 0.001. The main result shown at the top of this figure is a table of ranked candidate chemicals with registry numbers (ID), names as MeSH terms, number of related PubMed citations (PMIDs), number of PubMed citations classified as relevant to AD (hits), FDR and links to the 10 best PubMed citations (top 10 abstracts). PubMed citation links are displayed by level of confidence: highest level of confidence (manual validation; red heart symbols), high precision (90% precision from random simulations; red diamonds), good precision (70% precision from random simulations; black spades) and others passing the cutoff (black clubs). By clicking on a PubMed citation link, detailed information will be displayed in a new window (bottom right corner), including abstract, MeSH terms, discriminative words (brown color gradient) and the target chemical name (rose highlighting). The output also contains the list of discriminative words used by the document classifier (left-hand side), and a download section to retrieve the data as text tables (bottom left corner).

Table 1. Molecular pathway benchmark

Pathway	Known chemicals	Candidates	True positives	Recall	Area under ROC curve
Biogenic amine synthesis	14	545	8	0.57	0.862
Cholesterol biosynthesis	14	1740	10	0.71	0.862
Monoamine transport	12	1563	9	0.75	0.945
Nucleotide metabolism	14	1952	11	0.79	0.926
One-carbon metabolism	12	1443	7	0.58	0.856
Tryptophan metabolism	14	2431	9	0.64	0.736
Urea cycle and metabolism of amino groups	13	682	10	0.77	0.914

option, respectively. As PolySearch did not identify enough candidates for six diseases, Alkemio was mostly compared to FACTA. As we set the same number of candidates for each tool, comparing other statistics such as the recall and the *F*1 score leads to the same result (Supplementary Figure S1).

Use case

We have illustrated the use of Alkemio with a query focused on the study of the chemicals relevant to the topic of protein aggregate formation in the brain, an effect usually associated to neurodegenerative disease. To do this analysis, we used Alkemio with default parameters (using as input the result of a query to PubMed, ranking chemicals from citations from the last 3 years, using a *P*-value cutoff for abstract selection equal to 0.01 and a FDR cutoff equal to

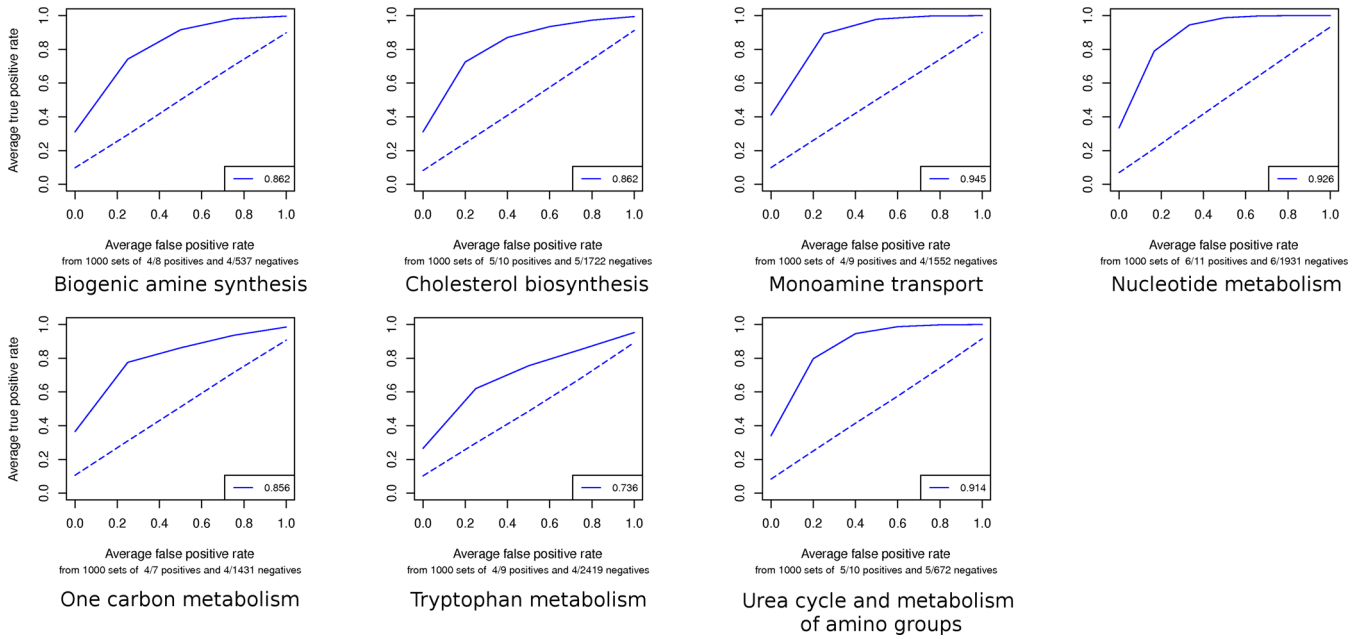


Figure 2. Molecular pathway benchmark. Alkemio was queried to retrieve chemicals related to seven molecular pathways. Pathways were selected from the WikiPathways database if associated with >10 chemicals. Due to the low number of known chemicals in these pathways (from 12 to 14) and to the high number of candidates returned by Alkemio (between 545 and 2431), we evaluated the retrieval performance using a random sampling strategy with the QiSampler tool (9). The figure shows ROC curves (blue lines) and control curves from random simulations (dashed lines) produced by the QiSampler tool when selecting 1000 repetitions and 50% of sampling rate. Legends show area under the curve.

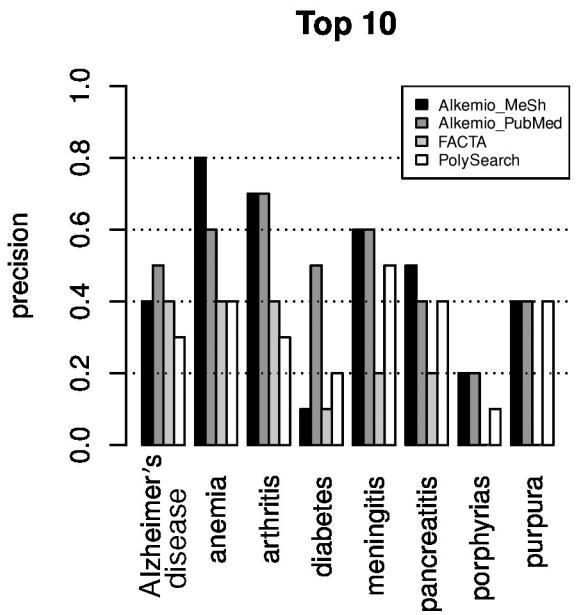


Figure 3. Comparison with existing tools. Precision in the top 10 candidate chemicals retrieved by Alkemio, FACTA and PolySearch according to manual associations between chemicals and diseases from the CTD database. As the CTD data are not comprehensive, many true positives are not automatically identified and the observed precision underestimates the real precision.

0.001) using as input a query with the words protein, aggregates and brain. This loosely defined query, nevertheless, selected abstracts related to the intended topic as indicated by the top ranked words that resulted: aggregate, tauopathie(s) (a class of degenerative diseases associated to the aggregation of the protein tau), synuclein (a family of proteins of which α-synuclein is known to aggregate in AD) and huntingtin (the protein mutated in Huntington's disease, a neurodegenerative disease).

The list of chemicals included many strong associations (Table 2). The reasons why these chemicals are relevant to the topic of protein aggregates in neurodegeneration can be easily examined by following the links to the abstracts, whose words are colored according to their discriminative power. Examination of the top ranked chemicals illustrates the different types of implications they can have in the context of the study of a disease.

The first ranked chemical was polyglutamine in reference to polyglutamine tracts, a protein feature whose pathological expansion results in several neurodegenerative diseases, probably due to the alteration of protein interactions modulated by these tracts. The second chemical was dopamine, which is depleted in the brains of Parkinson's disease patients. Others were oxidopamine (6-hydroxydopamine), 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine, 1-methyl-4-phenylpyridinium, rotenone and glutamate, which are neurotoxic compounds, often used to induce a Parkinson's disease-like phenotype in rats and mice. The following chemicals were also returned: thioflavin T used as a fluorescent marker in experiments, levodopa (L-DOPA) used to treat Parkinson's disease and 3-nitropropionic acid used to cause mitochondrial dysfunction and neurodegeneration in rodent models of Huntington's disease.

Relevant bibliography is selected by Alkemio and can be accessed by running this query (representative PubMed records are indicated in Table 2). In summary, the chemicals

Table 2. Top chemicals related to protein, aggregates and brain

Rank	Chemical	Representative Alkemio selected PMID	Relation to topic
1	Polyglutamine	22433867	Mechanism
2	Dopamine	22967820	Marker
3	Oxidopamine	22016808	Toxin
4	1-Methyl-4-phenyl-1,2,3,6-tetrahydropyridine	21448659	Toxin
5	Thioflavin T	21464905	Technical
6	Rotenone	21736921	Toxin
7	1-Methyl-4-phenylpyridinium	23754278	Toxin
8	Glutamic acid	22560595	Toxin
9	Levodopa	22764226	Protector
10	3-Nitropropionic acid	21448659	Toxin

obtained are relevant to the topic and their relation to neurodegenerative disease (technical, part of the natural mechanism, marker, toxin or protector) can be easily understood by examination of the top ranked references to PubMed.

If a focus on a particular category of chemicals is desired, one may try to build more specific queries, but these have to be constructed specifically for each topic. However, this may not be too difficult after examining the bibliography obtained in the first less specific step. For example, after compiling the previous example, it became obvious that the term 'protective' is used in the field of neurodegeneration to define chemicals that are used for a beneficial effect as opposed to toxins. Therefore, we illustrated the selection of chemicals in a second step using the same parameters with a query containing the following words: protein, aggregates, brain and (additionally to the previous query) protective.

As before, the words were generally related to neurodegeneration with top weighed words such as aggregate, huntingtin, htt and synuclein. However, the top ranked chemicals included compounds such as davunetide and clioquinol (rank positions 3 and 6, respectively), used as therapeutic agents, or marker 24-hydroxycholesterol (rank 5), after polyglutamine and oxidopamine, and only one toxin (trimethyltin, rank 4).

In summary, Alkemio allows to quickly produce queries of increased precision through the selection of links to relevant PubMed records. Usually, examining one or two abstracts suffices to understand the evidence behind the selection of a chemical in the ranked list.

DISCUSSION

The implementation of Alkemio as a web tool allows users to rank tens of thousands of chemicals discussed in the recent literature for any topic of interest. The SOAP web service allows the programmatic usage of Alkemio (e.g. for batch queries) from common programming languages or from the provided client script. Building large lists of scored candidates manually is impractical. Alkemio can provide in a short time a list of hundreds or thousands of candidate chemicals for a given topic. These candidates are scored by the words from their related literature. Such large lists of topic-related chemicals can have multiple uses such as in supporting the analysis of molecular pathways or diseases.

A task for a biomedical scientist could be to find chemicals related to molecular pathways. Pathway data are accessible from dedicated databases (e.g. WikiPathways), but it may not be up to date to the current literature. Here we have shown that Alkemio can retrieve thousands of chemicals associated to particular pathways, although only 12–14 were known in WikiPathways. The good ranking of those known chemicals as assessed by ROC curves suggests that the scored lists provided by Alkemio are appropriate for the task.

Alkemio has better performance than existing tools (FACTA (10) and PolySearch (3)) to retrieve disease-associated chemicals. On the one hand, the tools tried to retrieve all existing relevant chemicals from the literature; on the other hand, the gold standards contained only a limited selection of chemicals relevant for their purpose (i.e. description of only key molecules involved in pathways for WikiPathways, or environmental chemicals involved in diseases for marker, mechanism or therapeutic role for CTD). Consequently, benchmarking the tools using those gold standards resulted in underestimated performances, although they allowed an unbiased comparison.

For example, in the Alkemio's top 10 candidates for AD queried by a MeSH term, the following chemicals not automatically identified as true positives were found manually to be true positives: creatine (CAS: 57-00-1), which is used to define markers in combination with other molecules (15), oxidopamine (CAS: 1199-18-4), which was studied in AD (16) and may have a mechanism in common with the amyloid-β peptide involved in AD (17), choline (CAS: 62-49-7), which had positive results on the disease as nutrient (18), dopamine (CAS: 51-61-6), which is linked to non-cognitive aspects of dementia and target of known AD drugs (19) and thioflavin T (CAS: 2390-54-7), which is used in many studies in fluorescence assays to follow aggregation or fibril formation (see e.g. (20)). Therefore, Alkemio's precision for the top 10 ranked chemicals is actually 100% and not 40% as automatically computed using CTD data.

The differences in performance between the tools may be explained by the fact that, contrary to FACTA and PolySearch, Alkemio does not rely on information extraction techniques to get associations between chemicals and PubMed citations. Such techniques have limited accuracy and are often used prior to manual validations (21,22). Alkemio relies on manually set MeSH (15) annotations that may not be comprehensive, but that are of the highest quality. We note that the differing scoring schemes used by the tools compared possibly had an impact on the results. In

addition, it would be interesting to expand the benchmark with more diseases and with other topics. As we have used existing tools and their web interfaces (HTML pages), such expansion would require considerably more manual work. On the contrary, Alkemio could be automatically queried using the client script to its SOAP web service.

Notably, results are sensitive to updates. The system is updated daily and results may differ with time (e.g. new citations could be linked to chemicals or citations may be deleted from PubMed). Parameters have also an impact on the ranking results. Defining the topic using a PubMed query would model the topic using the most recent related citations. Topics may be discussed differently over time. For example, novel methods may be first technically described and later just used in standard protocols. Defining the topic using MeSH terms would model the topic more generally because internally a random selection of related citations is performed. The automatic selection of chemicals for ranking can be done using citations from the last 1 or 3 years. The former choice performs in a smaller dataset and, therefore, is faster, but will produce results biased toward recent research trends; it could be useful for preliminary searches. In general, the bigger the dataset, the more powerful the statistics. Thus, we would recommend using citations from the last 3 years to rank many chemicals with Alkemio. Also, chemicals having numerous citations may be favored in their ranking simply because related statistics would be defined from big numbers.

For the future, we plan to reduce Alkemio's computing time by running parallel computations. We would also try to use citation metadata in addition to words in abstracts as classification features. In conclusion, the Alkemio tool accessible from its web interface or SOAP web service ranks with high performance chemicals related to a biomedical topic of interest in practical time. Alkemio is freely accessible to non-commercial users at the following URL: http://cbdm.mdc-berlin.de/~medlineranker/cms/alkemio.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENT

We thank Dr Nancy Mah for meaningful discussions and critical reading of the manuscript.

FUNDING

Funding for open access charge: Max Delbrück Center for Molecular Medicine (Berlin, Germany).
Conflict of interest statement. None declared.

REFERENCES

1. NCBI Resource Coordinators. (2014) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **42**, D7–D17.
2. Frijters,R., Heupers,B., van Beek,P., Bouwhuis,M., van Schaik,R., de Vlieg,J., Polman,J. and Alkema,W. (2008) CoPub: a literature-based keyword enrichment tool for microarray data analysis. *Nucleic Acids Res.*, **36**, W406–W410.
3. Cheng,D., Knox,C., Young,N., Stothard,P., Damaraju,S. and Wishart,D.S. (2008) PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res.*, **36**, W399–W405.
4. Rebholz-Schuhmann,D., Kirsch,H., Arregui,M., Gaudan,S., Riethoven,M. and Stoehr,P. (2007) EBIMed–text crunching to gather facts for proteins from Medline. *Bioinformatics*, **23**, e237–e244.
5. Fontaine,J.F., Priller,F., Barbosa-Silva,A. and Andrade-Navarro,M.A. (2011) Genie: literature-based gene prioritization at multi genomic scale. *Nucleic Acids Res.*, **39**, W455–W461.
6. Fontaine,J.F., Barbosa-Silva,A., Schaefer,M., Huska,M.R., Muro,E.M. and Andrade-Navarro,M.A. (2009) MedlineRanker: flexible ranking of biomedical literature. *Nucleic Acids Res.*, **37**, W141–W146.
7. Ortuno,F.M., Rojas,I., Andrade-Navarro,M.A. and Fontaine,J.F. (2013) Using cited references to improve the retrieval of related biomedical documents. *BMC Bioinformatics*, **14**, 113.
8. Kelder,T., van Iersel,M.P., Hanspers,K., Kutmon,M., Conklin,B.R., Evelo,C.T. and Pico,A.R. (2012) WikiPathways: building research communities on biological pathways. *Nucleic Acids Res.*, **40**, D1301–D1307.
9. Fontaine,J.F., Suter,B. and Andrade-Navarro,M.A. (2011) QiSampler: evaluation of scoring schemes for high-throughput datasets using a repetitive sampling strategy on gold standards. *BMC Res. Notes*, **4**, 57.
10. Tsuruoka,Y., Tsujii,J. and Ananiadou,S. (2008) FACTA: a text search engine for finding associated biomedical concepts. *Bioinformatics*, **24**, 2559–2560.
11. Davis,A.P., Murphy,C.G., Johnson,R., Lay,J.M., Lennon-Hopkins,K., Saraceni-Richards,C., Sciaky,D., King,B.L., Rosenstein,M.C., Wiegers,T.C. *et al.* (2013) The Comparative Toxicogenomics Database: update 2013. *Nucleic Acids Res.*, **41**, D1104–D1114.
12. Law,V., Knox,C., Djoumbou,Y., Jewison,T., Guo,A.C., Liu,Y., Maciejewski,A., Arndt,D., Wilson,M., Neveu,V. *et al.* (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.*, **42**, D1091–D1097.
13. Wishart,D.S., Jewison,T., Guo,A.C., Wilson,M., Knox,C., Liu,Y., Djoumbou,Y., Mandal,R., Aziat,F., Dong,E. *et al.* (2013) HMDB 3.0—the Human Metabolome Database in 2013. *Nucleic Acids Res.*, **41**, D801–D807.
14. Wohlgemuth,G., Haldiya,P.K., Willighagen,E., Kind,T. and Fiehn,O. (2010) The Chemical Translation Service—a web-based tool to improve standardization of metabolomic reports. *Bioinformatics*, **26**, 2647–2648.
15. Ashford,J.W., Adamson,M., Beale,T., La,D., Hernandez,B., Noda,A., Rosen,A., O'Hara,R., Fairchild,J.K., Spielman,D. *et al.* (2011) MR spectroscopy for assessment of memantine treatment in mild to moderate Alzheimer dementia. *J. Alzheimers. Dis.*, **26**(Suppl. 3), 331–336.
16. Streltsov,V.A. and Varghese,J.N. (2008) Substrate mediated reduction of copper-amyloid-beta complex in Alzheimer's disease. *Chem. Commun. (Camb.)*, 27, 3169–3171.
17. Mazziotti,M. and Perlmutter,D.H. (1998) Resistance to the apoptotic effect of aggregated amyloid-beta peptide in several different cell types including neuronal- and hepatoma-derived cell lines. *Biochem. J.*, **332**(Pt 2), 517–524.
18. Wurtman,R.J., Cansev,M., Sakamoto,T. and Ulus,I.H. (2009) Use of phosphatide precursors to promote synaptogenesis. *Annu. Rev. Nutr.*, **29**, 59–87.
19. Zhang,L., Zhou,F.M. and Dani,J.A. (2004) Cholinergic drugs for Alzheimer's disease enhance in vitro dopamine release. *Mol. Pharmacol.*, **66**, 538–544.
20. Itkin,A., Dupres,V., Dufrene,Y.F., Bechinger,B., Ruysschaert,J.M. and Raussens,V. (2011) Calcium ions promote formation of amyloid beta-peptide (1–40) oligomers causally implicated in neuronal toxicity of Alzheimer's disease. *PLoS One*, **6**, e18250.
21. Davis,A.P., Wiegers,T.C., Johnson,R.J., Lay,J.M., Lennon-Hopkins,K., Saraceni-Richards,C., Sciaky,D., Murphy,C.G. and Mattingly,C.J. (2013) Text mining effectively scores and ranks the literature for improving chemical-gene-disease curation at the comparative toxicogenomics database. *PLoS One*, **8**, e58201.

22. Neves,M., Damaschun,A., Mah,N., Lekschas,F., Seltmann,S., Stachelscheid,H., Fontaine,J.F., Kurtz,A. and Leser,U. (2013) Preliminary evaluation of the CellFinder literature curation pipeline for gene expression in kidney cells and anatomical parts. *Database (Oxford)*, **2013**, bat020.