Database, Vol. 2012, Article ID bas037, doi:10.1093/database/bas037

Original article

Collaborative biocuration—text-mining development task for document prioritization for curation

Thomas C. Wiegers*, Allan Peter Davis and Carolyn J. Mattingly

Department of Biology, North Carolina State University, Raleigh, NC 27695-7617, USA

***Corresponding author:** Tel: +1 207 288 9880; Fax: +1 207 288 2130; Email: tcwieger@ncsu.edu

Submitted 5 June 2012; Revised 11 September 2012; Accepted 2 October 2012

The Critical Assessment of Information Extraction systems in Biology (BioCreAtIvE) challenge evaluation is a community-wide effort for evaluating text mining and information extraction systems for the biological domain. The 'BioCreative Workshop 2012' subcommittee identified three areas, or tracks, that comprised independent, but complementary aspects of data curation in which they sought community input: literature triage (Track I); curation workflow (Track II) and text mining/natural language processing (NLP) systems (Track III). Track I participants were invited to develop tools or systems that would effectively triage and prioritize articles for curation and present results in a prototype web interface. Training and test datasets were derived from the Comparative Toxicogenomics Database (CTD; http://ctdbase.org) and consisted of manuscripts from which chemical–gene–disease data were manually curated. A total of seven groups participated in Track I. For the triage component, the effectiveness of participant systems was measured by aggregate gene, disease and chemical 'named-entity recognition' (NER) across articles; the effectiveness of 'information retrieval' (IR) was also measured based on 'mean average precision' (MAP). Top recall scores for gene, disease and chemical NER were 49, 65 and 82%, respectively; the top MAP score was 80%. Each participating group also developed a prototype web interface; these interfaces were evaluated based on functionality and ease-of-use by CTD's biocuration project manager. In this article, we present a detailed description of the challenge and a summary of the results.

Introduction

The Comparative Toxicogenomics Database (CTD; http://ctdbase.org) is a publicly available resource that aims to promote understanding about the mechanisms by which drugs and environmental chemicals influence the function of biological processes and human health (1). CTD data are manually curated by a team of PhD-level biocurators. Articles are typically prioritized by chemicals of interest and distributed to biocurators, who then capture relevant data using our first-generation web-based curation application (2). Curated data include chemical–gene/protein interactions, chemical–disease relationships and gene–disease relationships. These data are integrated with select external datasets to facilitate development of novel hypotheses about chemical–gene–disease networks (3).

All manually curated data are captured using freely available controlled vocabularies. Chemicals are represented using terms from the Chemicals and Drugs subset of the National Library of Medicine's Medical Subject Headings (MeSH) vocabulary (4); genes and proteins are represented using the Entrez Gene vocabulary (5); diseases are represented using CTD's novel disease vocabulary MEDIC (6) that merges OMIM and the Disease subset of the MeSH vocabulary (4,7), and chemical–gene/protein interactions are captured using CTD's action vocabulary (1). The implementation of a web-based curation application has had many positive effects on the CTD curation process, including increasing the efficiency of curation, enhancing the flexibility of biocurator location, introducing real-time quality control, and easing data management and storage (2). Research has demonstrated that further

Database, Vol. 2012, Article ID bas037, doi:10.1093/database/bas037

enhancement of the curation process for CTD, as well as for many manually curated biomedical resources, would be achieved by improving: (i) the triage and prioritization of data-rich relevant articles and (ii) the identification of curatable content within these articles (8). The 'BioCreative Workshop 2012' subcommittee dedicated a focus area, or track (Track I), to development of systems that would address these important, yet unmet needs of the biocuration community.

The CTD project was chosen by the subcommittee as a source for the project data because it possesses a large and high quality set of manually curated information that contains elements that are of broad interest and relevance to the biomedical research community, specifically chemicals, genes/proteins and diseases. In addition, CTD, with its own fully automated text-mining pipeline, has significant experience in text mining research and development (8).

During September 2011, Track I issued an open invitation to text-mining teams to develop a system to assist biocurators in the selection and prioritization of relevant articles for curation for CTD (http://www.biocreative.org/events/bc-workshop-2012/CFP/#track1). The participants formed their own teams, sometimes across multiple institutions, and registered for the competition via the BioCreative web site. Although there were open communications between CTD staff and participants, there was no formal collaboration or interaction between the participants themselves; in fact, the participating teams were not announced by organizers until after the competition was completed.

Participants were asked to provide two major deliverables that included: (i) prioritization of relevant articles, as well as NER result sets and (ii) a prototype web interface that would present a biocurator with these articles and the relevant information highlighted using integrated NER tools. CTD staff then evaluated each group's results based on document ranking effectiveness and pre-determined entity recognition metrics, as well as a qualitative review of the web interface.

Methods

Training phase

In order for participants to effectively rank articles and identify relevant data, it was critical for them to gain an understanding of the CTD curation process. To facilitate this understanding, a detailed document entitled, 'Summary of Curation Details for the Comparative Toxicogenomics Database', was distributed to participants (http://www.biocreative.org/tasks/bc-workshop-2012/Triage/). In addition, a training dataset was made available to participants that consisted of 1725 articles that had been previously triaged and curated by CTD biocurators. The data were presented in a series of input files that included all associated curated data for eight target chemicals (raloxifene, aniline, amasacrine, doxorubicin, aspartame, quercetin, 2-acetylaminofluorene and indomethacin). It is important to note that all text mining associated with Track I was limited to the PubMed abstract; full text was not text mined.

In January 2012, the 'BioCreative Track I File Upload Facility' web site was released (Figure 1). This web site enabled participants to upload their benchmarking files. The web site in turn produced a report containing detailed information regarding their benchmarking performance and aggregate statistics. Specifically, a report was generated that calculated the aggregate 'mean average precision' [MAP; (9)] score, as well as the recall scores for each data type curated (chemicals, genes, diseases and action terms). Additional details were provided that enabled participants to understand how these scores where calculated (Figure 1).

It is important to note that the standard text-mining metric, precision, was not appropriate for Track I. The gold standard data were comprised of curated—rather than cited—gene, disease and chemical actors within each abstract. There are many instances where cited actors are not actually involved in the types of interactions captured by CTD curators; furthermore, there are instances where curated actors are found only in the full text of the article. Consequently, the complete universe of valid and cited actors specifically resident within each abstract is not recorded by CTD curators. Recall scores were calculated by simply dividing the number of distinct curated actors identified by the text-mining tools—either by a synonym to the term or by the term itself—by the total number of distinct curated actors. Micro-averaging was used for aggregate recall scores.

Recall scores were provided for each data category (chemicals, genes, diseases and action term) within each article. Three fields were provided for each data category, on a 'per article' basis and included:

- 'Curated Terms'—This field listed the terms, if any, that a CTD biocurator previously curated for each data category.
- 'Text Mined Terms'—This field listed the text-mined terms, if any, that a participant provided for each data category.
- 'Match Explanation'—This field provided an explanation of how matches between the curated and text-mined terms were determined. Because providing synonyms to curated terms are counted as matches, the notation of *CYP1→CYP1A1*, for example, indicated that the term *CYP1* was text mined, which is a valid synonym for the actual underlying curated term *CYP1A1*; alternatively, *FZR1→FZR1* indicated that the text-mined term of *FZR1* exactly matched the curated term.

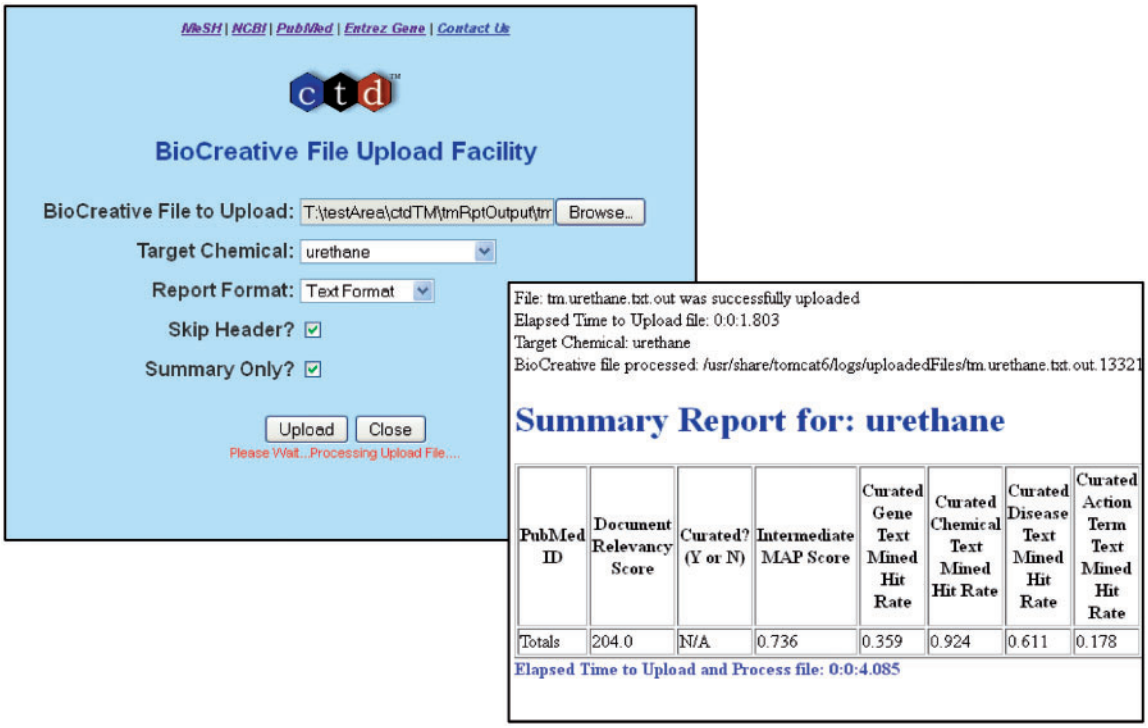

Figure 1. The BioCreative Track I File Upload Facility. A web interface was developed to allow participants to upload their results (back panel). Following successful uploads, a report was generated and returned to each participant that contained summary or detailed information for each dataset; a summary report is shown.

In all, the following information was provided for each article submitted in the form of a post-submission report:

- PubMed ID
- Curated (Y or N)?
- Intermediate MAP Score
- Curated Gene Hit Rate
- Curated Chemical Hit Rate
- Curated Disease Hit Rate
- Curated Action Hit Rate
- Text Mined Genes
- Curated Genes
- Gene Match Explanation
- Text Mined Chemicals
- Curated Chemicals
- Chemical Match Explanation
- Text Mined Diseases
- Curated Diseases
- Disease Match Explanation
- Text Mined Action Terms
- Curated Action Terms
- Action Term Match Explanation
- Curated Interaction(s), i.e. the interactions associated with the PubMed ID, as captured by the curator, e.g.: 'zinc affects the expression of ABL1 protein'.

The final line of the report provided the aggregate MAP and recall scores in each category. The reports were provided in both HTML and text formats; summary versions were also provided at the participant's discretion that contained solely the aggregate statistics. Figure 1 provides an example of the summary version of the report.

Test phase

On 6 February 2012, a Track I Test Dataset was released to participants. The purpose of this dataset was to evaluate the performance of the participants' text-mining pipeline without their prior knowledge of the curated results. The Track I Test Dataset comprised 444 articles that were previously manually curated by CTD biocurators and contained information about three additional target chemicals (urethane, phenacetin and cyclophosphamide). Table 1 provides an overview of both the Training and Test Datasets. Unlike the comprehensive curated data provided in the Training Dataset, the Test Dataset contained only the basic identification information for each article (PubMed ID, Title, Abstract, Journal Name and Date). Each participant was asked to process the Test Dataset using their text-mining pipeline, and provide the following information for each article/target chemical combination:

- PubMed ID
- Title
- Abstract
- Journal

Database, Vol. 2012, Article ID bas037, doi:10.1093/database/bas037

Table 1. BioCreative corpus overview

Target chemical	No. of references	No. of curatable / uncuratable	No. of interactions	No. of distinct actors	Curated chemical mean ± standard deviation	No. of distinct gene actors	Curated gene mean ± standard deviation	No. of distinct disease actors	Curated disease mean ± standard deviation	No. of distinct action terms
Cyclophosphamide	154	107/47	526	150	1.40 ± 0.78	351	3.28 ± 8.78	79	0.74 ± 0.78	192
Phenacetin	86	66/20	740	321	4.86 ± 5.13	271	4.11 ± 4.53	6	0.09 ± 0.34	149
Urethane	204	107/97	720	210	1.96 ± 1.37	351	3.28 ± 8.33	91	0.85 ± 0.81	238
Aspartame	156	46/110	132	86	1.87 ± 1.63	51	1.11 ± 1.23	25	0.54 ± 0.78	60
2-Acetylaminofluorene	178	81/97	508	142	1.75 ± 1.17	340	4.20 ± 9.63	19	0.23 ± 0.55	179
Indomethacin	85	76/9	681	157	2.07 ± 1.60	447	5.88 ± 18.07	40	0.53 ± 0.60	213
Aniline	226	100/126	650	271	2.71 ± 2.26	280	2.80 ± 2.56	21	0.21 ± 0.78	264
Raloxifene	270	163/107	1897	417	2.56 ± 2.41	887	5.44 ± 16.78	72	0.44 ± 0.71	388
Amsacrine	69	38/31	243	119	3.13 ± 3.72	73	1.92 ± 4.43	6	0.16 ± 0.37	64
Doxorubicin	199	138/61	1487	236	1.71 ± 0.92	1183	8.57 ± 64.07	58	0.42 ± 0.60	374
Quercetin	542	392/150	4158	1291	3.29 ± 2.71	1719	4.39 ± 9.43	72	0.18 ± 0.43	1197
Totals	2169	1314/855	11742	3400	2.59 ± 2.51	5953	4.53 ± 23.04	489	0.37 ± 0.65	3318

The BioCreative text-mining corpus was comprised of a total of 11 chemicals, 3 of which represented the Track I Test Dataset. The chemicals associated with the Track I Test Dataset were cyclophosphamide, phenacetin and urethane; the remaining 8 chemicals comprised the Track I Learning Dataset.

- Cited Gene Actor Terms, as identified by the NER tools as being referenced in the abstract.
- Cited Chemical Actor Terms, as identified by the NER tools as being referenced in the abstract.
- Cited Disease Actor Terms, as identified by the NER tools as being referenced in the abstract.
- Marked-up HTML of abstract with tagged links back to CTD for all corresponding terms.
- Document Relevancy Score.
- Optional: Marked-up HTML of relevant sentences/phrases extracted with tagged links back to CTD for all actors and terms.
- Optional: Cited Action Terms.
- Optional: Cited Interactions, e.g.: 'zinc affects the expression of ABL1 protein'.

Table 2 provides an example of the reports provided by the participants.

The benchmarking results and associated documentation were due on 20 February 2012. Upon receipt of the benchmarking data from the participants, CTD staff evaluated the results by calculating the following metrics for each participant:

- MAP score
- Curated Gene Term Recall Score
- Curated Chemical Term Recall Score
- Curated Disease Term Recall Score
- Curated CTD Action Term Recall Score

Results

A total of seven groups participated:

- BiTeM Group; Division of Medical Information Sciences, University Hospitals of Geneva and University of Geneva; Information Science Department, University of Applied Science; Geneva, Switzerland.
- Department of Computer Science and Information Engineering, National Cheng Kung University; Department of Information Engineering, Kun Shan University, Tainan, Taiwan.
- Institute of Computational Linguistics, University of Zurich.
- Two groups from the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan.
- Department Of Computer Science, East China Normal University.
- National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA.

To maintain anonymity, each group was randomly assigned a coded identification number.

Table 2. Example of participant reporting requirements

PubMed ID	17368022
Title	Analogs of the marine alkaloid makaluvamines: synthesis, topoisomerase II inhibition, and anticancer activity.
Abstract	Twelve analogs of makaluvamines have been synthesized. These compounds were evaluated for their ability to inhibit the enzyme topoisomerase II. Five compounds were shown to inhibit topoisomerase catalytic activity comparable to two known topoisomerase II targeting control drugs, etoposide and m-AMSA. Their cytotoxicity against human colon cancer cell line HCT-116 and human breast cancer cell lines MCF-7 and MDA-MB-468 has been evaluated. Four makaluvamine analogs exhibited better IC(50) values against HCT-116 as compared to control drug etoposide. One analog exhibited better IC(50) value against HCT-116 as compared to m-AMSA. All 12 of the makaluvamine analogs exhibited better IC(50) values against MCF-7 and MDA-MB-468 as compared to etoposide as well as m-AMSA.
Journal	Bioorg Med Chem Lett
Cited Gene Actors	TOP2A
Cited Chemical Actors	AMSACRINE ETOPOSIDE
Cited Disease Actors	COLONIC NEOPLASMS BREAST NEOPLASMS
Marked-up HTML of Abstract	Twelve analogs of makaluvamines have been synthesized. These compounds were evaluated for their ability to inhibit the enzyme TOPOISOMERASE II. Five compounds were shown to inhibit TOPOISOMERASE catalytic activity comparable to two known TOPOISOMERASE II targeting control drugs, ETOPOSIDE and M-AMSA...
Document Relevancy score	0.5
Marked-up HTML of relevant sentences/phrases	1.) These compounds were evaluated for their ability to inhibit the enzyme TOPOISOMERASE II. 2) Five compounds were shown to inhibit TOPOISOMERASE catalytic activity comparable to two known TOPOISOMERASE II targeting control drugs, ETOPOSIDE and M-AMSA...
Cited Action Terms	Decreases activity
Cited Interactions	Etoposide results in decreased activity of TOP2A protein Amsacrine results in decreased activity of TOP2A protein

Each participating team was asked to process the test dataset using their text-mining pipeline and provide associated data to CTD staff for each article/target chemical combination. Table 2 provides an example of the reporting requirements. Note: The formatting has been slightly modified in the example for clarity of presentation.

Database, Vol. 2012, Article ID bas037, doi:10.1093/database/bas037

MAP. For MAP (9) score calculations, an article was counted as relevant if it had one or more associated curated interactions. Across the groups, MAP scores were fairly high and consistent, ranging from 71% to 80% (Figure 2).

Curated term recall. The results for recall scores were significantly more mixed than the MAP scores. The aforementioned standard text-mining metrics, recall and precision, were not appropriate for Track I. The recall score for each gene, chemical and disease term was calculated by comparing the list of text-mined terms with the list of curated terms for each article and in each respective data category. As indicated above, if the curated term, or a synonym for the curated term (as defined by the corresponding CTD controlled vocabulary), was found in the text-mined list, it was counted as a match.

Gene recall ranged from 2% to 49% (Figure 3). Chemical recall ranged from 5% to 82% (Figure 4). Disease recall ranged from <1% to 65% (Figure 5). Note that four of the seven participants scored near zero in disease recall; although it is unclear precisely why these four groups performed poorly, three of the groups used tools developed in-house. With respect to the optional data fields, only one metric was measured: curated CTD action term recall rate. Group 139 successfully identified 30% of the curated action terms; none of the remaining groups was able to successfully identify curated action terms. The results of MAP scores, and chemical, gene, disease and action term recall scores, were also aggregated onto a single bar graph for each participating group (Figure 6).

Aggregate benchmarking results summary. Table 3 provides a summary of each team's approach to NER and IR; CTD's pipeline is also described in Table 3. Two of the groups clearly distinguished themselves with respect to aggregate benchmarking results. Group 121 held the highest *MAP* score (80%), while also delivering strong recall scores in the three major recall categories (chemicals, genes and diseases). Group 116 delivered the highest recall scores in two of the three major data categories (i.e. gene and disease recall). Three other groups (120, 139 and 130) had respectable recall scores in most, if not all, of the major data categories.

The groups were also asked to provide a system description, all of which were reasonably clear and well-written.

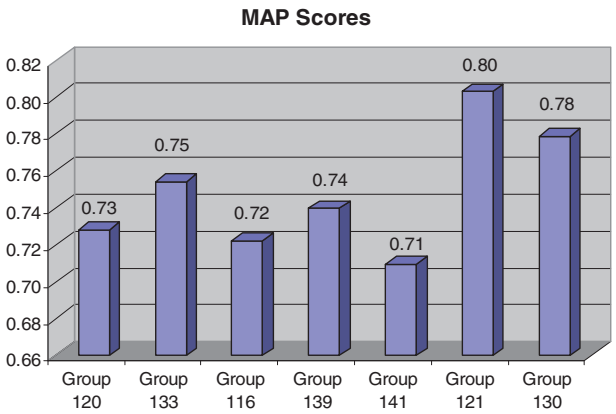

Figure 2. MAP (9) score results for each participating group. For MAP score calculations, an article was counted as relevant if it had one or more associated curated interactions. Across the groups, MAP scores were fairly high and consistent, ranging from 71% to 80%.

Prototype web interface. The participants were asked to deliver a prototype web interface to Track I organizers by 1 March 2012. All seven groups that participated in the benchmarking portion of the challenge also submitted a prototype web interface. Each interface was then evaluated based on functionality and ease-of-use by CTD's biocuration project manager.

Of the seven entries, six provided very sophisticated functionality. (Please note that the web interfaces described below that tag gene, chemical and disease terms were only as effective at doing so as their benchmarking results suggest; the same is true for those web interfaces that provided a ranked list of PubMeds: their ranking effectiveness is reflected in their benchmarking MAP scores).

Group 121. The biocurator accesses the system by clicking the 'Login' link. Once login is complete, the user is presented with a list of chemicals for curation. Clicking on one of the chemicals takes the biocurator to a ranked list of articles associated with the chemical with the following information:

- Title,
- Author(s),
- Journal name, date and page numbers,
- PubMed ID,
- Related citations hyperlink,
- Abstract hyperlink.

The biocurator may remove an article from the list by simply clicking on a single 'delete' hyperlink [e.g. Delete from (BC2012-test-urethane)]. Clicking on the 'Abstract' hyperlink causes an expansion of the screen to include the complete abstract text (Figure 7a). All genes, chemicals and disease actors contained in the title or abstract and identified by the text-mining tool are color-coded and hyperlinked back to the CTD web interface.

Clicking on the title causes a detail page to be displayed (Figure 7b). The detail page contains most of the same information as the main page, but also includes a list of text-mined chemical, gene and disease actors, each of

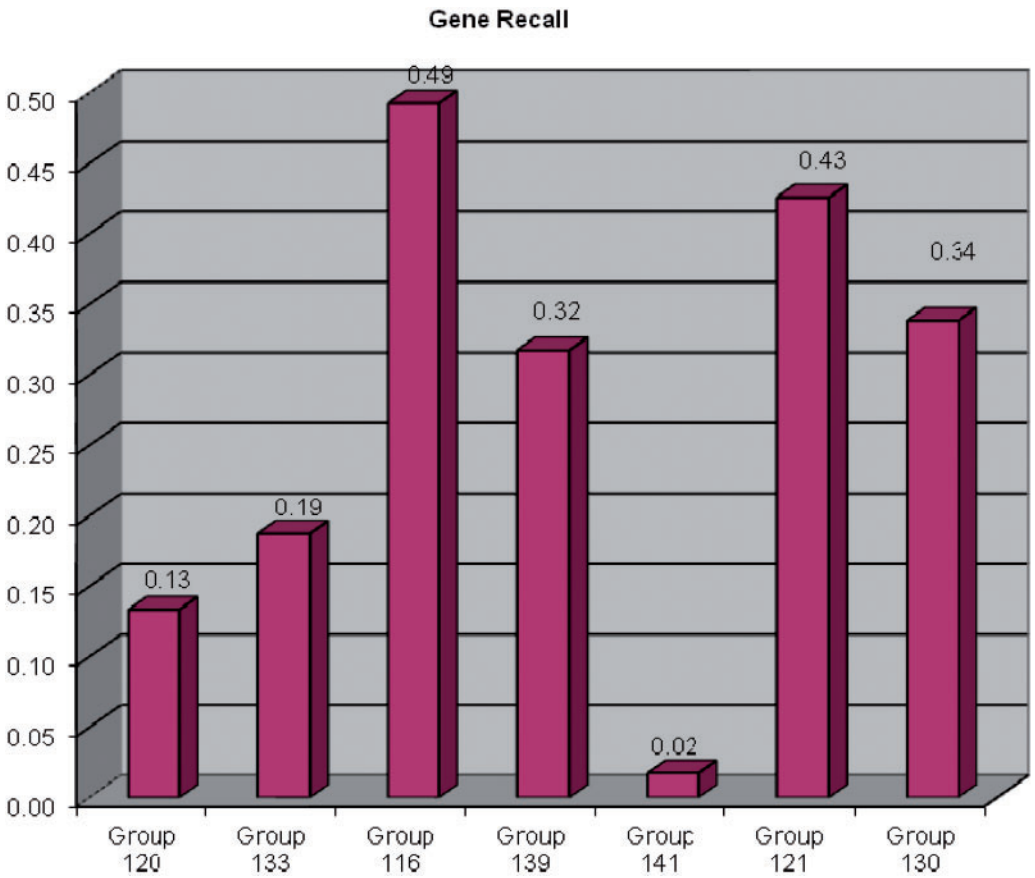

Figure 3. Gene recall results for each participating group. The ability for text-mining tools to recognize curated genes was measured; terms and synonyms to terms were counted as matches. Gene recall ranged from 2% to 49%.

which is hyperlinked back to CTD. The interface enables the user to save new annotations, as well as confirm and/or reset existing entries. Although this particular feature as currently implemented does not appear to be of direct application to CTD, it certainly has interesting long-term implications.

The interface includes several additional options. A list of text-mined target chemicals is displayed on the main target chemical screen, enabling the biocurator to easily jump from one chemical list to another for curation. The ranked list of articles can be re-sorted based on date or relevancy score. Clicking on a chemical, gene or disease checkbox on the main target chemical screen or on the detail page causes these actors to either be highlighted and hyperlinked or made simply plain text. Because there is sometimes an overlap in chemical, gene and disease names, there is a feature that enables the user to correct and save a text-mined actor designation to another category. Finally, there are 'Display Management' screens that enable biocurators to select their highlighting preferences in the interface (Figure 7c). For example, a user can specify whether or not to display chemicals, genes and diseases by default, as well as to set the colors of the display.

Group 116. The biocurator is presented with a list of target chemicals to curate. After clicking on a target chemical hyperlink, the user is presented with a ranked list of articles by their PubMed ID, relevancy score and title. Clicking a PubMed ID presents detailed information in a new tab. The new tab displays a split screen; on the left hand side is the 'Document' panel that displays the title and abstract text, along with all of the MeSH terms associated with the paper; the right side of the screen is the 'Annotation' panel.

The 'Annotation' panel initially consists of two tabs: 'Concepts' and 'Interactions'; a 'Terms' tab may also be displayed if the user selects it from the toolbar. The 'Concepts' tab (Figure 8a) lists the chemical, disease and gene terms identified during the text-mining process, including the accession, term name, frequency of appearance in the abstract and type of term (i.e. chemical, disease or gene); the 'Concepts' tab contains an entry for each concept identified in at least one term in the document. Each of the concepts is also scored using an algorithm developed by the team. If concept rows are expanded by clicking the plus button, a hyperlink to the relevant Web page of the CTD site appears. The 'Interactions' tab (Figure 8b) displays

Database, Vol. 2012, Article ID bas037, doi:10.1093/database/bas037

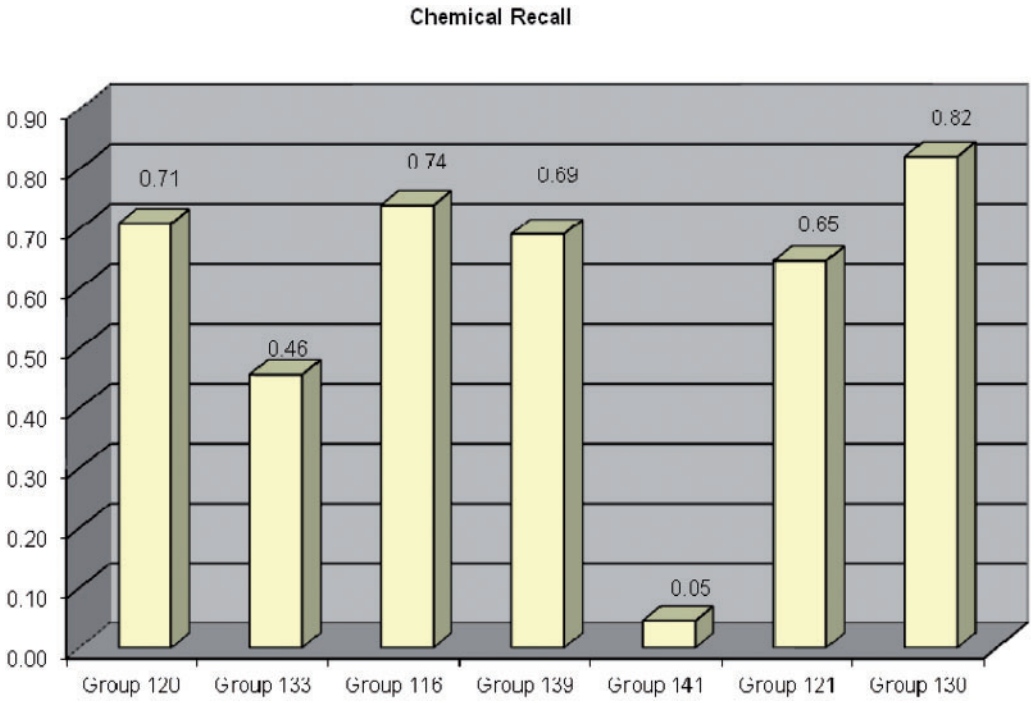

Figure 4. Chemical recall results for each participating group. The ability for text-mining tools to recognize curated chemicals was measured; terms and synonyms to terms were counted as matches. Chemical recall ranged from 5% to 82%.

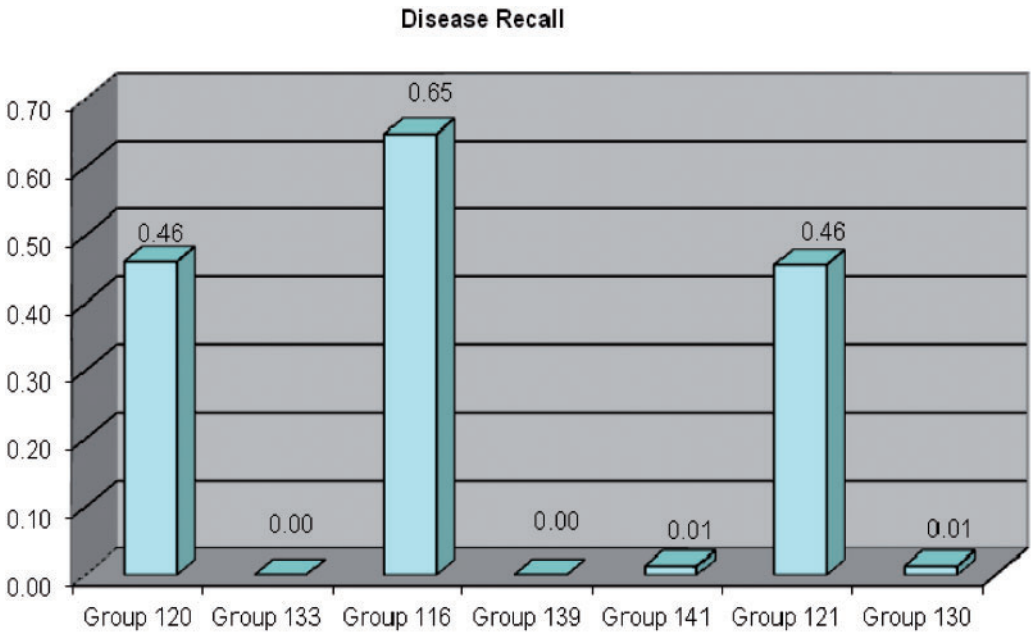

Figure 5. Disease recall results for each participating group. The ability for text-mining tools to recognize curated diseases was measured; terms and synonyms to terms were counted as matches. Disease recall ranged from <1% to 65%.

potential interactions contained within the abstract; these interactions are also derived using a scoring algorithm developed by the team. For each potential interaction, a confidence score is displayed, along with the type and name of each chemical, disease and gene actor. In the 'Interactions' tab, clicking on the name of a participating concept opens the relevant CTD web page. The 'Terms' (Figure 8c) tab contains an entry for each stretch of text

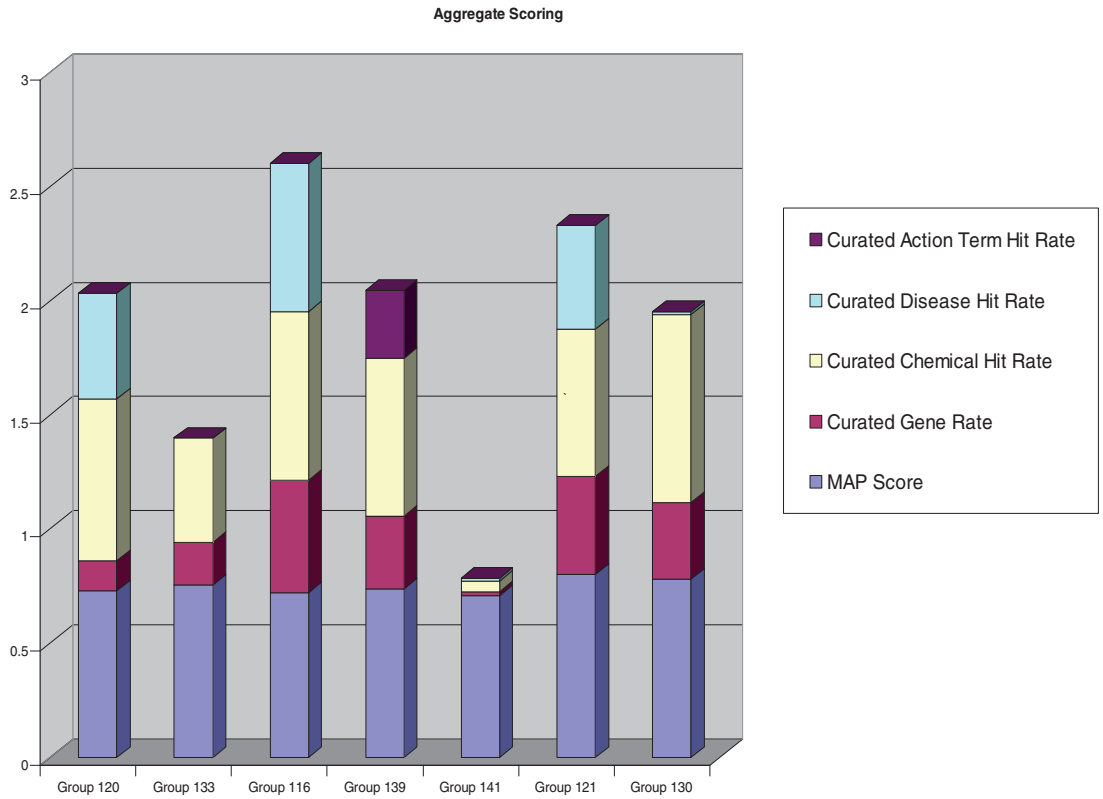

Figure 6. Aggregate metrics for each participating group. The results of MAP (9) scores and chemical, gene, disease and action term recall scores are aggregated onto a single bar graph for each participating group. Two of the groups clearly distinguished themselves with respect to aggregate benchmarking results. Group 121 held the highest MAP score (80%) while also delivering strong recall scores in the three major recall categories (chemicals, genes and diseases). Group 116 delivered the highest recall scores in two of the three major data categories (i.e. gene and disease recall). Three other groups (120, 139 and 130) had respectable recall scores in most, if not all, of the major data categories.

considered as a technical term. However, no concept disambiguation is made, i.e. a term can contain references to more than one concept, even of different types (e.g. genes, chemicals, etc.). One of the more interesting features of the 'Annotation' panel is that check boxes are displayed next to each interaction and concept; clicking these check boxes will cause the associated text-mined data to be highlighted and hyperlinked within the abstract text or alternatively, simply plain text. All of this is done without a screen refresh, so it is extremely fast.

The interface included several additional and very convenient options. The user may remove concepts, interactions or terms from the 'Annotation' tab by simply selecting an associated checkbox and clicking the 'Remove Selected' button. One may also highlight a term and add it to the concepts' list by simply double clicking on it and completing the necessary data, including term, term-type, concept values, comments and search databases (i.e. CTD or Entrez), in the 'Inspectors' tab. Mousing over a term/concept causes the term-type and associated accession IDs to be displayed. The user may dump all the information associated with an article into XML format by selecting the option from the menu. The curation actions taken upon a document are logged into the document itself and/or into a separate database.

Group 133. The user is initially presented with a selection of chemicals to curate. The biocurator selects a chemical, clicks the 'Submit' button and is presented with a split screen (Figure 9). On the left hand side of the screen is an ordered list of ranked articles with associated information, including relevancy score, PubMed ID, article title, journal name and abbreviated abstract.

The biocurator may begin curating from the list. Clicking one of the ranked articles causes a 'Detail Info' frame containing detailed information to be displayed on the right hand side of the split screen. More specifically, each of the data elements described above is provided, along with the complete abstract text. The title and the abstract contain highlighted genes and chemicals within the text, as well as lists of each beneath the abstract; the lists hyperlink each text-mined actor back to the CTD web interface. A link is also provided to view the PubMed at NCBI on a separate tab.

Database, Vol. 2012, Article ID bas037, doi:10.1093/database/bas037

Table 3. NER and IR tools summary

Group number	NER	IR
116	Gene, chemical and disease: proprietary algorithms supplemented by PubMed metadata	*Lucene* (10) with customization
120	Gene: *NormaGene* (11) Diseases and chemicals: Ad-hoc keyword recognizer based on the controlled vocabularies provided by CTD	*EAGLi* (12)
121	Gene, chemical and disease: *SemCat* (13) coupled with proprietary vector space-based algorithm	Support vector machine-based proprietary algorithms.
130	Gene: *AllAGMT* (14). Chemical: conditional random fields with training patterns extracted from CTD Disease: proprietary dictionary-based algorithms coupled with *MEDIC* (6)	Co-occurrence network-based proprietary algorithms
133	Gene: *Banner* (15) Chemical: *OSCAR4* (16) Disease: *MEDIC* (6)	Rules-based proprietary algorithms
139	Gene: *Banner* (15) Chemical: *OSCAR4* (16) Disease: *MEDIC* (6) Action Term: CTD Action Term vocabulary coupled with proprietary algorithms	Term frequency-inverse document frequency-based proprietary algorithms
141	Gene, Chemical, and Disease: *MetaMap* (17)	Rules-based proprietary algorithms
CTD (8)	Gene: *Abner* (18), *MetaMap* (17), In-house gene normalizer Chemical: *OSCAR3* (19), *MetaMap* (17) Disease: *MetaMap* (17) Action Term: CTD Action Term vocabulary coupled with proprietary algorithms	Rules-based proprietary algorithms

A brief summary of each participating team's, as well as CTD's, NER and IR tools. There was a large variance in the tools employed by the participants.

Group 120. In order to begin curation, the biocurator enters a chemical, as well as a list of associated PubMed IDs separated by tab or new line characters. The list of PubMed IDs is text mined and processed on a real-time basis. Once the text mining is complete, the biocurator is presented with a list of ranked and relevancy score-sorted PubMeds, including the following information (Figure 10):

- PubMed ID
- Article title
- Journal name
- Text-mined genes
- Text-mined chemicals
- Texted-mined diseases
- Relevancy score

To the left of each PubMed ID, a +/− button either expands or contracts the display of the PubMed's abstract. The abstract contains highlighted genes, chemicals and diseases within the text, each of which is hyperlinked back to the CTD web interface.

Group 139. The web interface provided was very similar to Group 133's prototype; only subtle differences were apparent (Figure 11).

Group 130. Clicking on the 'System Demo' link presents the user with the main curation screen. Portions of the main screen are apparently under construction and are not currently functional. However, selecting a chemical from the 'Data set' field and clicking the 'Submit' button presents the biocurator with a list of ranked PubMeds that are associated with the selected target chemical. For each PubMed, its numeric sequential rank is provided, along with the article's title, abstract and a list of text-mined chemical, gene and disease actors (Figure 12). Each of the text-mined actors is highlighted within the title and the abstract text. Each of the actors provided in the respective lists beneath the abstract are hyperlinked back to CTD, although the hyperlinks may or may not actually link to a CTD actor, i.e. the actors do not appear to have been mapped to actual CTD terms.

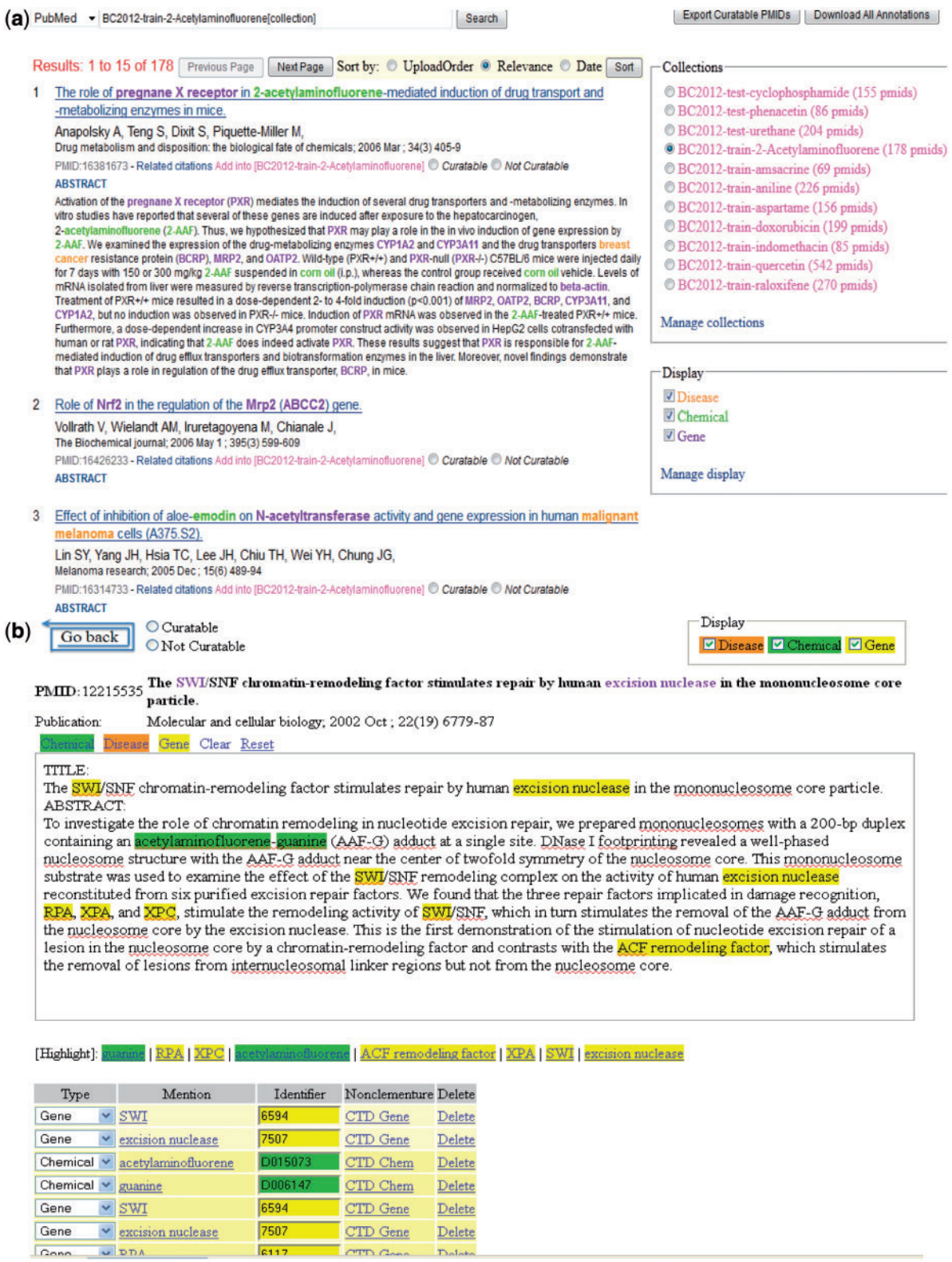

Figure 7. (**a**) Group 121 web interface. A screenshot of Group 121's ranked list of chemicals for curation in their web interface. (**b**) A screenshot of Group 121's curation detail page in their web interface. (**c**) Screenshots of two of Group 121's data management-related pages in their web interface.

(continued)

Database, Vol. 2012, Article ID bas037, doi:10.1093/database/bas037

Figure 7. Continued.

Group 141. The biocurator is presented with two options for curation:

- 'Single Mode'—Allows a user to enter a single PubMed ID for text mining.
- 'Batch Mode'—Allows a user to load a file containing one or more PubMed IDs; the file must contain one PubMed ID on each line without a blank line, including the last line.

In 'Single Mode', entering a single PubMed ID and pressing 'Submit' resulted in a report being displayed, providing the PubMed ID entered and a relevancy score (Figure 13). There were columns available for text-mined gene, chemical, disease and action terms. The report provided no further functionality.

In 'Batch Mode', uploading an input file will cause a new page to be opened, providing information associated with the upload along with a link to a results file. Clicking on the link causes TAB-delimited records to be displayed, one for each PubMed ID in the input file. Each TAB-delimited record contains basic information about the PubMed, as well as a relevancy score.

In conclusion, six of the seven submissions for the web interface component of the Track I challenge effectively presented the ranked and highlighted data. Of the six submissions, however, the products developed by groups 121 and 116 provided exceptional functionality and were deemed very user-friendly with potential for future expansion and application.

Conclusions

The Track I project was a very involved assignment. Development of effective ranking and recognition tools, as well as a prototype web interface that conveyed these results in a user-friendly manner required a high degree of systems development and integration.

Of the seven groups, five performed very well in virtually every category.

Apart from an interest in furthering text mining research, CTD's motivation in designing and administering Track I was to determine if participants might present solutions that could potentially improve the existing CTD text mining pipeline and/or CTD's web-based curation tool.

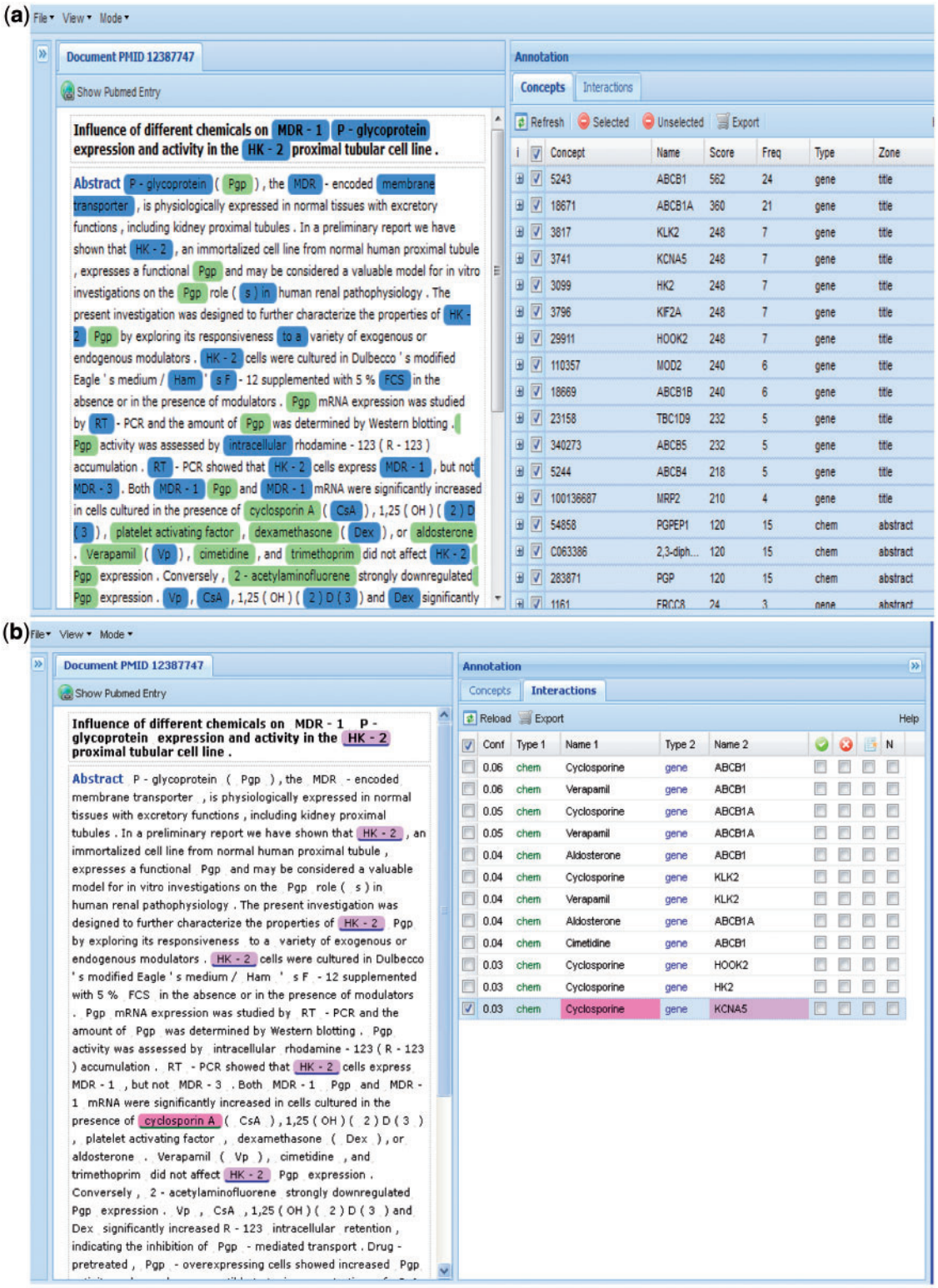

Figure 8. (**a**) Group 116 web interface. A screenshot of Group 116's Concepts tab in their web interface. (**b**) A screenshot of Group 116's Interactions tab in their web interface. (**c**) A screenshot of Group 116's Terms tab in their web interface.

(continued)

Database, Vol. 2012, Article ID bas037, doi:10.1093/database/bas037

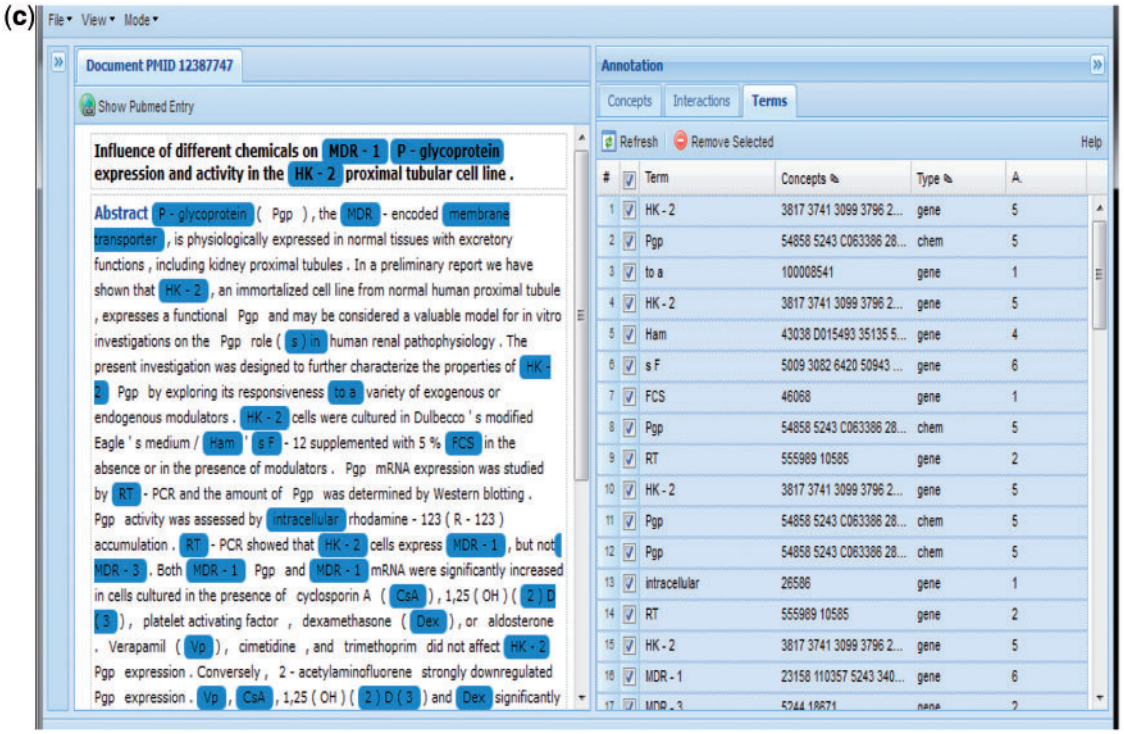

Figure 8. Continued.

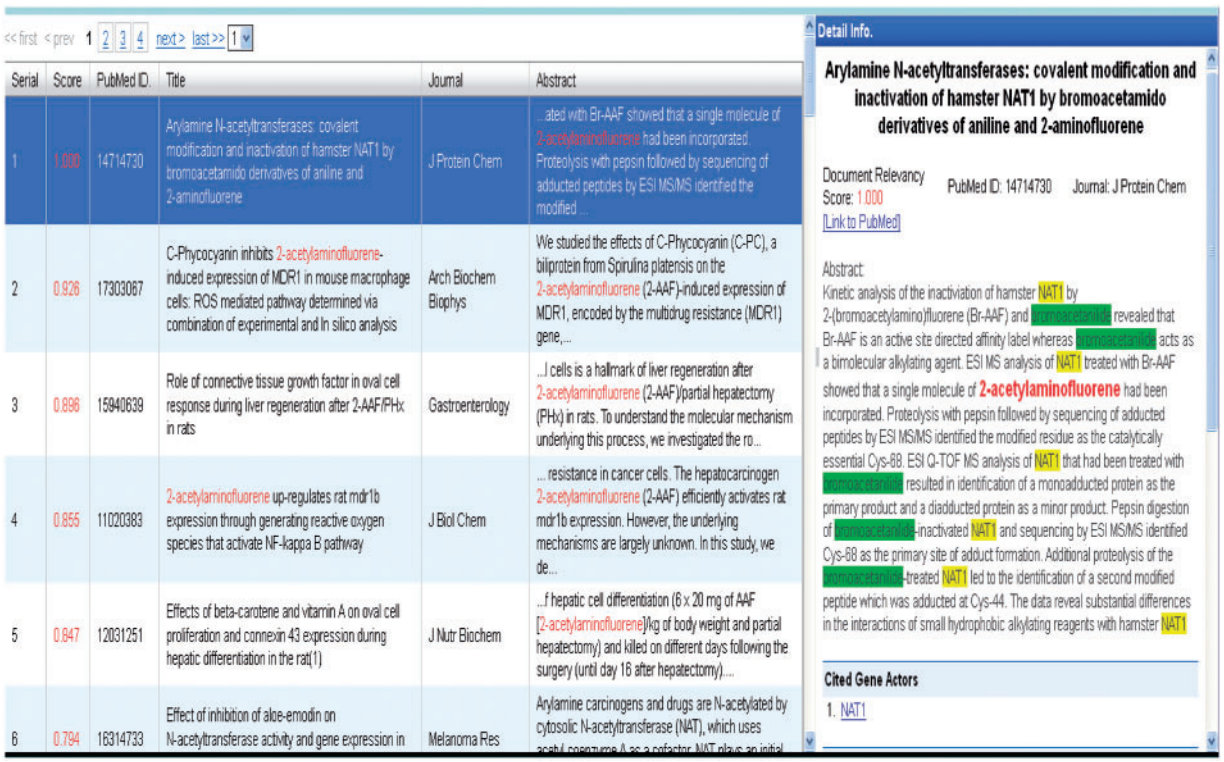

Figure 9. Group 133 web interface. A screenshot of Group 133's web interface.

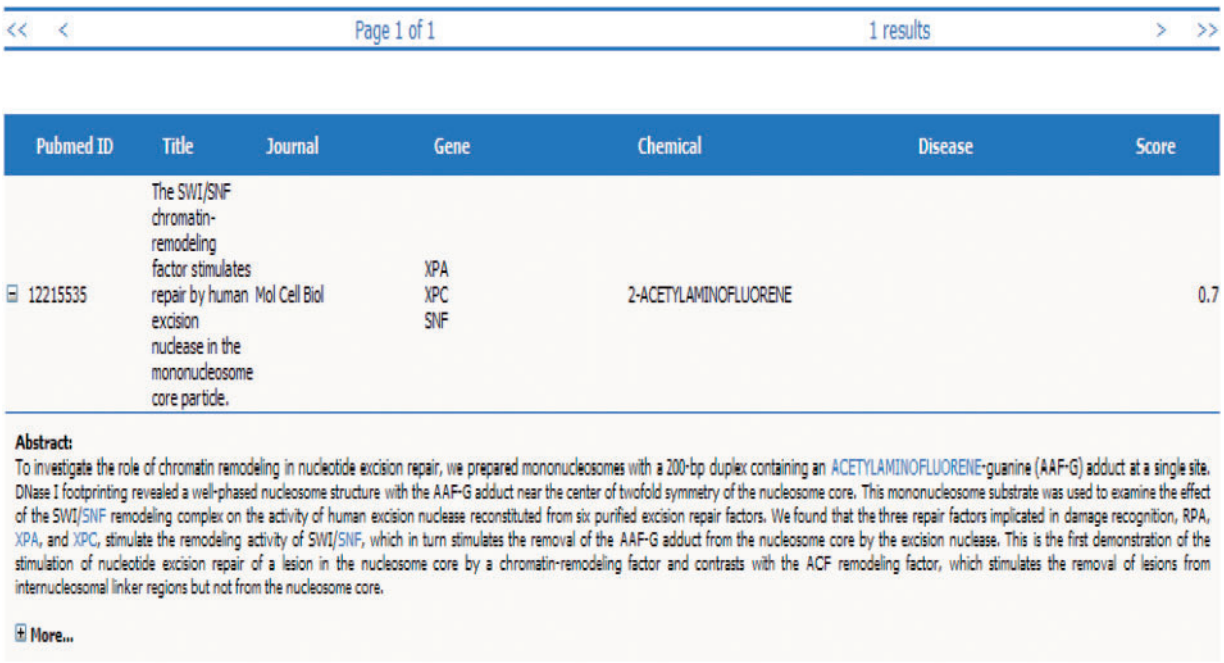

Figure 10. Group 120 web interface. A screenshot of Group 120's web interface.

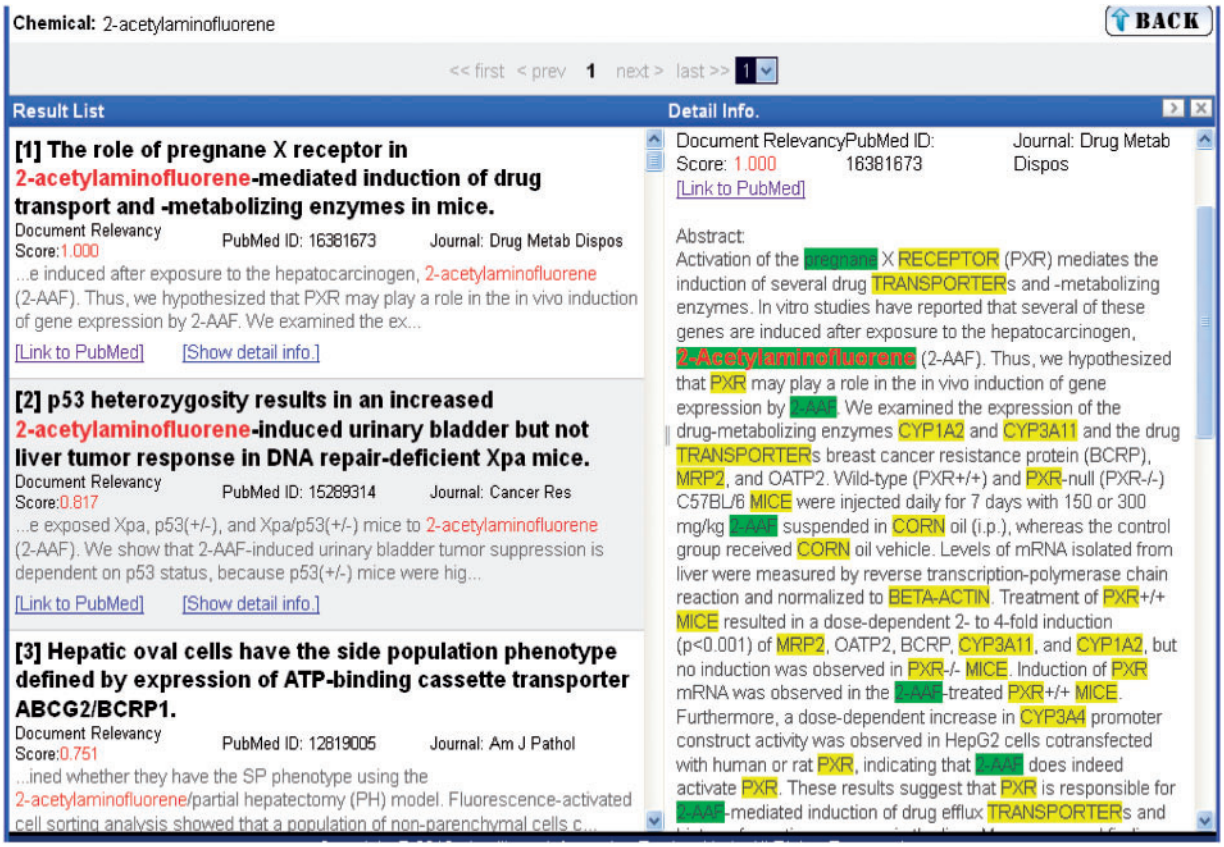

Figure 11. Group 139 web interface. A screenshot of Group 139's web interface.

Database, Vol. 2012, Article ID bas037, doi:10.1093/database/bas037

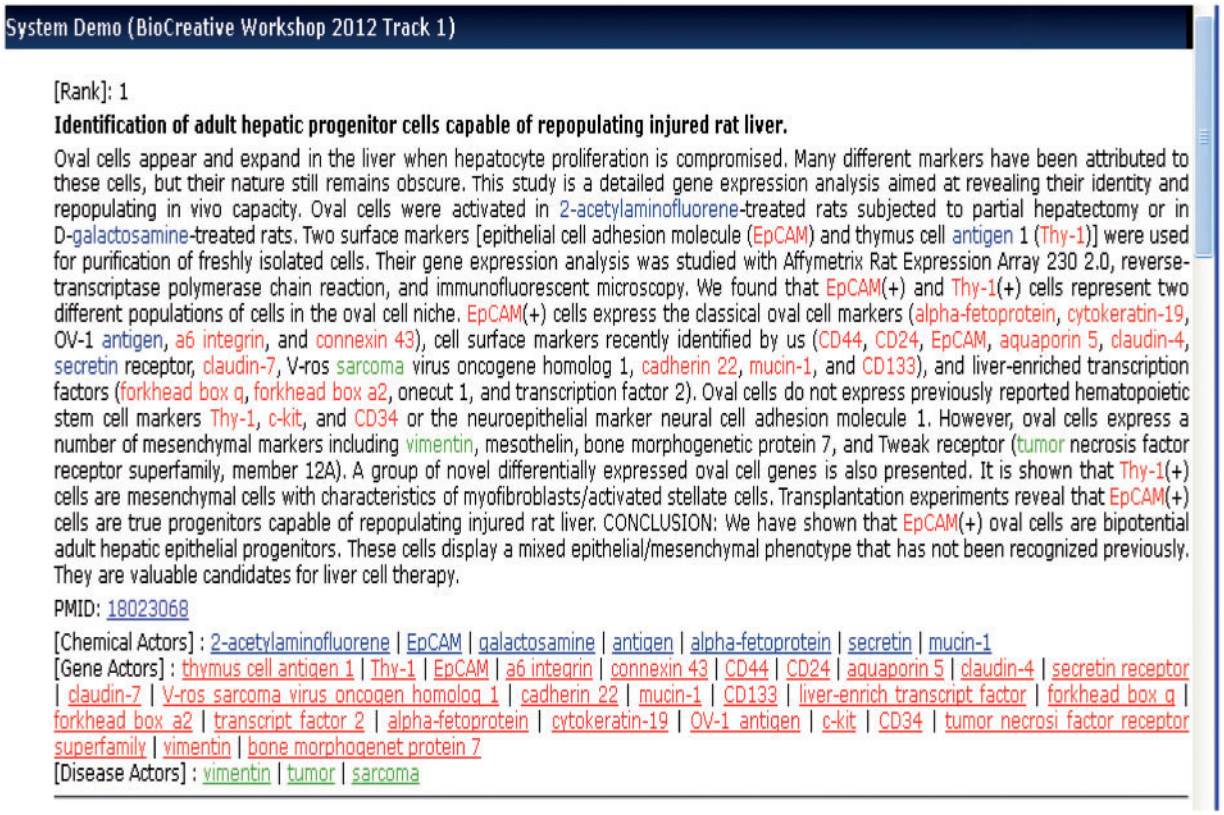

Figure 12. Group 130 web interface. A screenshot of Group 130's web interface.

Processed Succesfully

Test Result					
PubMed	Actor Gene	Actor Chmecal	Actor Disease	Action Term	Document Relevancy Score
18772604	EDNRB1	pigment epithelium-derived factor		RSE results in increased expression of LEVEL	0.9

Figure 13. Group 141 web interface. A screenshot of Group 141's web interface.

The potential benefit of the collective results to CTD are component dependent.

The existing CTD text-mining pipeline was run against the test cases and CTD's tools outperformed all the participating systems in nearly every individual benchmarking category, including MAP score. However, Group 116 outperformed CTD's pipeline in disease recall and Group 139 outperformed CTD's pipeline in action term recall; CTD placed second in both cases. Although collaboration is planned with Group 116 to explore the feasibility of disease recognition tool integration, CTD's text-mining pipeline will remain largely intact for the foreseeable future. The results of the Track I benchmarking component was very beneficial to CTD in that it confirmed the high quality of CTD's existing text-mining pipeline. The superiority of CTD's text-mining pipeline is not altogether unexpected; staff understanding of the CTD domain is obviously extensive, as has been the experimentation with text-mining tool integration (8). But confirmation of the pipeline's overall effectiveness is very helpful.

The benefit to CTD for participation in Track I is more obvious for the web interface component. CTD staff has not yet fully integrated its text-mining pipeline into its curation tool (2). Although none of the web interfaces

developed in conjunction with Track I could be directly integrated into CTD's curation tool as a result of the complexity imposed by the tool's technical infrastructure, certainly some of the features could have direct application to CTD.

CTD will remain involved in BioCreative and plans to design and administer a track for BioCreative 2013. One of the issues of interest to CTD is systems integration and interoperability. Tools developed by Track I participants were written using a wide variety of technologies and within technical infrastructures that would not necessarily easily integrate into CTD's existing text-mining pipeline. Initial plans for CTD involvement in BioCreative 2013 called for participants to build interoperable tools that could be accessed remotely by batch-oriented CTD text-mining processes using technologies such as 'Web services'; this approach, if effective, could serve to decouple CTD's technical infrastructure from each participating team's potentially disparate technical infrastructure.

In conclusion, the groups far surpassed expectations and are to be congratulated on their efforts and accomplishments in a short period of time. In addition to the successful generation of systems that may have long-term application for either CTD or other curated database groups, the success of the Track I program underscores the enhanced benefits that result from collaborative efforts among otherwise disparate biological and computational groups.

Funding

This program is supported by funds from the National Institute of Environmental Health Sciences (ES014065).

Conflict of interest. None declared.

References

1. Davis,A.P., King,B.L., Mockus,S., Murphy,C.G., Saraceni-Richards,C., Rosenstein,M., Wiegers,T. and Mattingly,C.J. (2011) The Comparative Toxicogenomics Database: update 2011. *Nucleic Acids Res.*, **39**, D1067–D1072.
2. Davis,A.P., Wiegers,T.C., Murphy,C.G. and Mattingly,C.J. (2011) The curation paradigm and application tool used for manual curation of the scientific literature at the Comparative Toxicogenomics Database. *Database*, **2011**, bar034.
3. Davis,A.P., Murphy,C.G., Saraceni-Richards,C., Rosenstein,M., Wiegers,T. and Mattingly,C.J. (2009) The Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. *Nucleic Acids Res.*, **37**, D786–792.
4. Coletti,M.H. and Bleich,H.L. (2001) Medical Subject Headings used to search the biomedical literature. *J. Am. Med. Inform. Assoc.*, **8**, 317–323.
5. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2011) Entrez gene: Gene-centered information at ncbi. *Nucleic Acids Res.*, **39**, D52–D57.
6. Davis,A.P., Wiegers,T.C., Rosenstein,M.C. and Mattingly,C.J. (2012) MEDIC: a practical disease vocabulary used at the comparative toxicogenomics database. *Database*, **2012**, bar057.
7. Amberger,J., Bocchini,C. and Hamosh,A. (2011) A new face and new challenges for online mendelian inheritance in man (omim(r)). *Hum. Mutat.*, **32**, 564–567.
8. Wiegers,T.C., Davis,A.P., Cohen,K.B., Hirschman,L. and Mattingly,C.J. (2009) Text mining and manual curation of chemical-gene-disease networks for the comparative toxicogenomics database (CTD). *BMC Bioinformatics*, **10**, 326.
9. Voorhees,E.M. and Harman,D.K. (2005) TREC: Experiment and Evaluation in Information Retrieval. MIT Press, Cambridge, Massachusetts, USA.
10. Gospodnetic,O. and Hatcher,E. (2004) *Lucene in Action*. Manning Publications, Greenwich.
11. Heckmann,L.H., Sorensen,P.B., Krogh,P.H. and Sorensen,J.G. (2011) NORMA-Gene: A simple and robust method for qPCR normalization based on target gene data. *BMC Bioinformatics*, **12**, 250.
12. Gobeill,J., Patsche,E., Theodoro,D., Veuthey,A.-L., Lovis,C. and Ruch,P. (2009) Question answering for biology and medicine. *Inform. Technol. Appl. Biomed.*, **2009**, 1–5.
13. Tanabe,L., Thom,L.H., Matten,W., Comeau,D.C. and Wilbur,W.J. (2006) SemCat: semantically categorized entities for genomics. *AMIA Annu. Symp. Proc.*, **2006**, 754–758.
14. Hsu,C.N., Chang,Y.M., Kuo,C.J., Lin,Y.S., Huang,H.S. and Chung,I.F. (2008) Integrating high dimensional bi-directional parsing models for gene mention tagging. *Bioinformatics*, **24**, 286–294.
15. Leaman,R. and Gonzalez,G. (2008) BANNER: An executable survey of advances in biomedical named entity recognition. *Pac. Symp. Biocomput.*, 652–663.
16. Jessop,D.M., Adams,S.E., Willighagen,E.L., Hawizy,L. and Murray,P. (2011) OSCAR4: a flexible architecture for chemical text-mining. *J. Cheminform.*, **3**, 41.
17. Aronson,A.R. (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc. AMIA Symp.*, **2001**, 17–21.
18. Settles,B. (2005) ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, **21**, 3191–3192.
19. Corbett,P. and Copestake,A. (2008) Cascaded classifiers for confidence-based chemical named entity recognition. *BMC Bioinformatics*, **9 (**Suppl 11), S4.