Data and text mining

Query Chem: a Google-powered web search combining text and chemical structures

Justin Klekota2,3,*, Frederick P. Roth4 and Stuart L. Schreiber1,2,3

1Howard Hughes Medical Institute, 12 Oxford Street, Cambridge, MA 02138, 2Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA 02142, 3Harvard University, University Hall, Cambridge, MA 02138 and 4Harvard Medical School, Department of Biological Chemistry and Molecular Pharmacology, 250 Longwood Avenue, SGM-322, Boston, MA 02115

Received on February 6, 2006; revised on March 28, 2006; accepted on April 19, 2006
Advance Access publication May 3, 2006
Associate Editor: Alfonso Valencia

ABSTRACT
Summary: Query Chem (www.QueryChem.com) is a Web program that integrates chemical structure and text-based searching using publicly available chemical databases and Google’s Web Application Program Interface (API). Query Chem makes it possible to search the Web for information about chemical structures without knowing their common names or identifiers. Furthermore, a structure can be combined with textual query terms to further restrict searches. Query Chem’s search results can retrieve many interesting structure–property relationships of biomolecules on the Web.

Contact: Klekota@gmail.com

INTRODUCTION
The recent proliferation of Web-based chemical databases has made information about millions of compound structures and their biological properties publicly available. Chembank (Strausberg, and Schreiber, 2003), ZINC (Irwin and Shoichet, 2005), Pubchem (Wheeler et al., 2005), PubChem (Chen et al., 2005) and ChemMine (Girke et al., 2005) are among the increasing number of chemical databases. Web-based tools, including Chmoogle (Gubernator, 2005) and Chemfinder’s online tool (Chemfinder.com, Cambridge-soft), search chemical databases for information associated with a compound structure. Despite the valuable information these tools provide, they capture only a fraction of the chemical information stored as text in millions of Web pages with compounds referenced by diverse names (Wilbur et al., 1999) (not as chemical structures). None of this internet-based information is wholly captured by any existing database or search engine (Banville, 2006). Furthermore, no existing tool combines both text search terms and chemical structures in its searches, allowing search results for a compound to be prioritized by relevance to user-specified chemical properties or biological activities.

In order to bridge the gap between chemical databases and the rest of the internet, Query Chem, Web-based software that integrates structure and text-based Web searches, has been developed. Given a compound structure and text search terms, Query Chem searches for similar structures using PubChem, Chembank or Chmoogle and then performs a series of Google searches of the entire Web combining the user-specified text search terms and the names of the matching structures. Query Chem prioritizes Web search results by their relevance to user-specified chemical structure and text search terms, such as ‘toxicity’ or ‘protein target’. Query Chem’s searches can retrieve known structure–property relationships of many biomolecules, often correctly identifying a compound’s known protein target, metabolism, and usage in the first page of search results (Fig. 1).

PROCEDURE
Query Chem, located at www.QueryChem.com, is a Perl/Common Gateway Interface (CGI) Web program which uses the Google Web API (www.Google.com/APIs) (Fig. 2). Users enter compounds in the form of text string representations of chemical structures called SMILES (www.Daylight.com) or enter compound structures by drawing them using the JME Structure Editor (www.Molinspiration.com/JME/). If a compound is known by a name only, links are provided to various chemical databases so that the SMILES string of that named compound can be located and copied into Query Chem. Users can also specify whether a full-structure or a substructure search should be performed. In addition to compound information, the user may enter textual search terms. Users may select either Chembank, PubChem or Chmoogle (technically a search engine) as the database searched to identify compounds matching the query SMILES string. This database is searched to retrieve up to 10 matching compounds. For example, many chemical databases, including Chembank, identify matching compounds through the use of binary fingerprints. These fingerprints may encode the SMILES strings of each compound’s substructures, thereby performing SMILES string matching quickly using bit operations on the fingerprints. Compound database searching issues have been described more fully elsewhere (www.Daylight.com).

If a full-structure search is performed, a measure of the structural similarity between the queried compound structure and each matching structure, e.g. the Tanimoto coefficient, is listed next to each matching compound. If a substructure search is performed, the compounds containing the queried substructure are retrieved. The names of matching compounds are retrieved from the chemical database and combined with the query text to perform a Web search using Google for each matching compound. If a compound name comprises multiple words, then each word in the name is connected...
by dashes so that Google will search the entire compound name as a single phrase, e.g. ‘Ethyl-alcohol’. Separate search results are returned for each matching compound, with links to the complete search results in the selected chemical database, in Google, and to the scientific literature via Google Scholar (Fig. 1). Since the Google API limits the frequency of searches for a given Google API key (1000 Google searches per day), frequent Query Chem users are encouraged to obtain their own Google API keys at www.google.com/apis for more searches.

RESULTS

We have confirmed that Query Chem retrieves many known relationships between compound structures and their properties, e.g. protein targets, metabolites, clinical usage or pollution sources. Three examples of full-structure searching, using the default Tanimoto chemical similarity threshold of 0.85, are discussed below. We have performed searches with a structurally diverse representation of queried compounds ranging from the complex 31-membered ring system of Rapamycin to the simplest 6-membered ring Benzene. We have tested Query Chem with example text search terms that demonstrate relationships between compounds, relationships between compounds and proteins, and the environmental prevalence of a compound. These examples show that Query Chem can retrieve information and chemical relationships of interest to chemists, biomedical researchers and students, as well as the lay community. There are no limits on allowed types of text search terms, so the following examples should not limit the reader’s imagination.

Protein targets of Rapamycin

Rapamycin (Fig. 3A) is a natural product used as an immunosuppressant and as a potential anti-cancer agent. Rapamycin has two protein targets, TOR and FKBP12, which in complex with Rapamycin lead to a cascade of molecular events that induce cell-cycle arrest at the G1/S transition (Sabers et al., 1995).

We used Query Chem with the PubChem database to search the structure of Rapamycin in combination with the search term ‘protein target’. In total, 87 similar compounds were identified from the PubChem compound database. Among these, Rapamycin and its close structural homologues (including Everolimus, Temsirolimus and Demethoxyrapamycin) yielded multiple search results. Although many other compounds did not yield search results (because they had no names or synonyms listed in PubChem), the search results of Rapamycin and its derivatives correctly identified TOR and FKBP12 as protein targets.

Metabolite of Trandolapril

Trandolapril (Fig. 3B) is an FDA-approved drug used to treat hypertension and some forms of chronic heart disease that acts by inhibiting the angiotensin-converting enzyme. Trandolapril is a prodrug which only becomes fully activated upon metabolism by the liver where it is converted to its most active form Trandolaprilat (Chevillard et al., 1994). Using Query Chem with Chembank to search the structure of Trandolapril and the search term ‘metabolite’. The four matching structures, their Tanimoto coefficients and their Google search results are all conveniently displayed in one page. Trandolapril’s metabolite, Trandolaprilat, is identified in the first search result.
term ‘metabolite’, the known metabolite of Trandolapril was correctly identified as Trandolaprilat. Repeating the same search using the structure of Trandolaprilat produced similar results. In total, Query Chem identified four similar compounds all of which had numerous search results on Google. The exact match Trandolapril was identified as well as its close structural homologues Enalapril, Lisinopril and Ramipril, which have similar clinical applications (Chevillard et al., 1994). The known metabolites of Enalapril and Ramipril were found among the search results, as was the fact that Lisinopril is not a prodrug.

Sources of Benzene pollution

Benzene (Fig. 3C) is a ubiquitous molecule occurring in nature and used in a variety of industrial settings. Benzene is also a common pollutant and a known carcinogen (Hrelia et al., 2004). Using Query Chem with Chmoogle selected to search the structure of Benzene in combination with the search terms ‘source of pollution’, numerous sources of Benzene pollution were cited in the results. Among the retrieved sources of Benzene pollution were automobiles, volcanoes, forest fires, chemical plants and smoking. No other matching compounds had search results.

CONCLUSIONS

Query Chem significantly enhances searching for chemical information on the Web and in the scientific literature (via Google Scholar) by integrating structure-based and text-based searching. This spares manual construction of lengthy literature and eliminates the requirement that all (or any) names or synonyms be known for a compound or that there be prior knowledge about similar chemical structures. Furthermore, combining chemical names with other textual search terms helps address the ambiguity of some chemical names (Wilbur et al., 1999; Banville, 2006) including Megaphone (CAS 64332-37-2), Moronic Acid and Broken Windowpane (aka Fenestrane). These chemical names are shared by non-chemical entities and turn up non-chemical search results in Google, but return chemically relevant search results when combined with user-specified text search terms, e.g. ‘chemical composition’ and ‘chemical synthesis’. One limitation of Query Chem is that Google only allows a maximum of 32 words in its searches, so that some compound synonyms will not be searched in a single step. Users are encouraged to refine searches via links to a Google search page with the initial query as a starting point.

Query Chem—which retrieves a broad array of information about chemical structures from the Web—has functionality that is distinct from Chmoogle (Gubernator, 2005), a search engine which searches for non-textual structure files (such as SD files) on the Web. Query Chem’s interface with Chmoogle demonstrates the complementarity of these two approaches, by providing quick retrieval of relevant
structures in Chmoogle followed by a search for literature about compound structures and their user-specified properties via Google. There is enormous potential to identify structure–property relationships of biomolecules on the Web using integrated search engines. The search capabilities of Query Chem should increase over time as more chemical names and synonyms are added to the publicly available chemical databases. Finally, we note that this application illustrates a more general principle—that Web search tools can be greatly improved and extended by allowing non-textual data objects to be used as search terms.

ACKNOWLEDGEMENTS

The authors thank Gabriel Berriz of the Harvard Medical School for his consultations. The authors also thank the National Cancer Institute’s Initiative for Chemical Genetics, which supports this informatic research and ChemBank. S.L.S. is an Investigator at the Howard Hughes Medical Institute at the Department of Chemistry and Chemical Biology, Harvard University. F.P.R. was supported in part by the Keck Foundation and by NIH grants R01 HG0017115, R01 HG003224, and U01 HL81341.

Conflict of Interest: none declared.

REFERENCES


