

SCRIPDB: a portal for easy access to syntheses, chemicals and reactions in patents

Abraham Heifets^{1,2} and Igor Jurisica^{1,2,3,*}

¹Department of Computer Science, University of Toronto, Toronto, Ontario, M5G 1L7, ²Ontario Cancer Institute, Princess Margaret Hospital, University Health Network, and the Campbell Family Institute for Cancer Research, Toronto, Ontario, M5G 2M9 and ³Department of Medical Biophysics, University of Toronto, Toronto, Ontario, M5G 1L7, Canada

Received August 18, 2011; Revised October 7, 2011; Accepted October 8, 2011

ABSTRACT

The patent literature is a rich catalog of biologically relevant chemicals; many public and commercial molecular databases contain the structures disclosed in patent claims. However, patents are an equally rich source of metadata about bioactive molecules, including mechanism of action, disease class, homologous experimental series, structural alternatives, or the synthetic pathways used to produce molecules of interest. Unfortunately, this metadata is discarded when chemical structures are deposited separately in databases. SCRIPDB is a chemical structure database designed to make this metadata accessible. SCRIPDB provides the full original patent text, reactions and relationships described within any individual patent, in addition to the molecular files common to structural databases. We discuss how such information is valuable in medical text mining, chemical image analysis, reaction extraction and *in silico* pharmaceutical lead optimization. SCRIPDB may be searched by exact chemical structure, substructure or molecular similarity and the results may be restricted to patents describing synthetic routes. SCRIPDB is available at <http://dcv.uhnres.utoronto.ca/SCRIPDB>.

INTRODUCTION

US patent information is in the public domain and describes innovations in the medical, biological, chemical and agricultural fields. Such relevant and accessible material is ideal for scientific analysis and, indeed, databases such as PubChem (1) and ChEBI (2,3) contain chemical structures disclosed by patents. While such

databases are highly useful, structural databases are insufficient for a number of scientific investigations.

The extraction of component structures from a patent discards information about chemical relationships. These relationships can be explicitly labelled, such as a molecule's role as reagent or product in a chemical synthesis. Relationships may also be implicitly embedded in the context of the complete patent. For example, molecules that co-occur in drug patent claims are likely to have similar biological behavior.

These exemplar relationships have been valuable in statistical analyses for automated reaction extraction (4,5) and bioisostere discovery (6). Reaction extraction characterizes the molecular transformations that occur within a set of syntheses. Bioisostere discovery catalogs molecular substituents that participate in similar biological interactions. Such analyses require access to large data sets, which are often unavailable, proprietary, or expensive.

In addition to the relationships among a patent's molecular structures, a patent's data files are directly useful. One use of patent files is the creation of data sets for optical structure recognition. The task in optical structure recognition is to parse a chemical image and recover the depicted molecular structure. The training and tests sets therefore require correctly matched pairs of images and molecular structure files. As a final example, a patent's written contents can serve as a target for text analytics. Patent descriptions and claims constitute a large corpus of biomedical text, coarsely annotated by, e.g., patent classification and drug–disease pairs.

To address these gaps in available metadata, we have created SCRIPDB. While providing uncomplicated searching of the patent literature, we have been careful not to eliminate underlying information. Users of the database may download the full text of the patent as well as molecular structure files and images. Additional summary files were generated and augmented, rather than

*To whom correspondence should be addressed. Tel: 416 581 7437; Fax: 416 946 4619; Email: juris@ai.toronto.edu

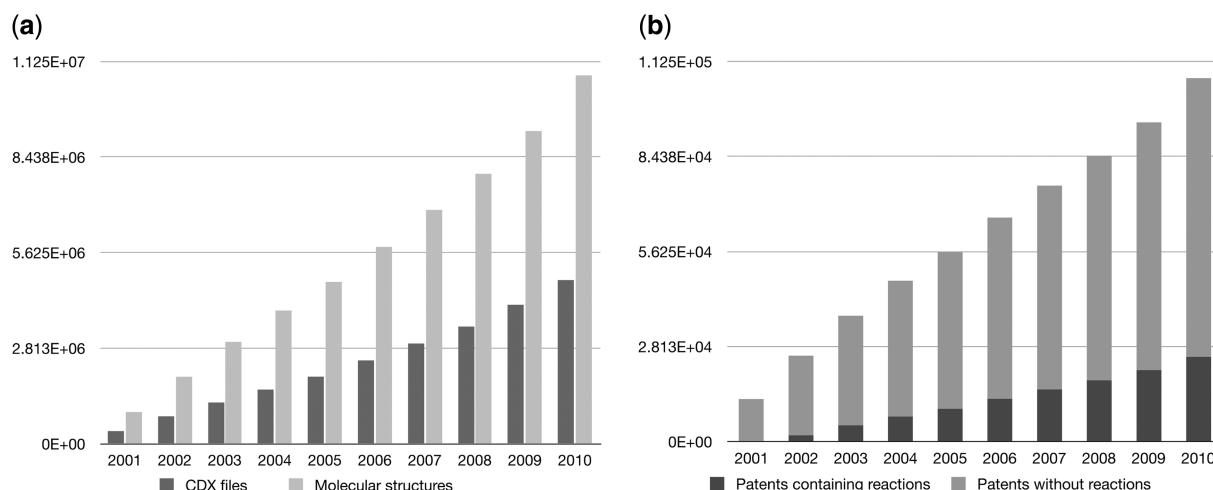


Figure 1. Cumulative SCRIPDB content. Although SCRIPDB includes patents from 2011, we show data through 2010, the last complete year. **(a)** shows the number of ChemDraw CDX structure files, and the structures described therein, available in SCRIPDB for various years. SCRIPDB contains 4 814 913 CDX files from 2001 through 2010, comprising 10 840 646 molecules. Duplicate molecules were filtered from each patent but not across patents, as described in the text. **(b)** shows the number of patents and details the subset containing reactions. For 2001 through 2010, SCRIPDB contains 107 560 patents, of which 25 048 contain synthetic reactions.

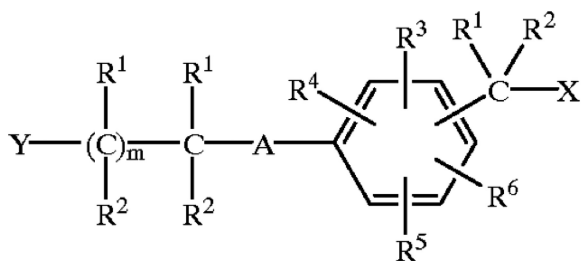


Figure 2. Example Markush structure from US Patent 6268 504 (9) defining a chemical class via substituent, positional and frequency variations.

replace, the original data. Thus, SCRIPDB permits researchers to effectively access the full information contained in the patent literature.

MATERIALS AND METHODS

Data collection and processing

The United States Patent and Trademark Office (USPTO) provides full patent text, drawings, and, since 2001, chemical structures in complex work units. This data became available as a free bulk download, hosted on Google servers, in June 2010. The raw data files comprise every granted patent, numbering several thousand per week, and totaling over 10 terabytes of data. However, most patents are not relevant to the biological, chemical or medical domains.

Since 2001, patented chemical structures can be described using standard molecular file formats. The USPTO makes disclosed molecules available as either MDL Molfiles (MOL) or ChemDraw binary CDX files. We used the presence of these chemical structure files to identify patents of potential interest. Figure 1 shows the

amount of data in SCRIPDB, as measured by the quantity of CDX files (Figure 1a), individual structures (Figure 1a), and patents (Figure 1b). These numbers are consistent with previous analyses of patent data (7).

A particular structure will often be described with both a CDX and Molfile. However, the information contained in such parallel files is not always identical. For example, a CDX file can label a molecule's role in a reaction as reagent or product, whereas a Molfile cannot. We provide both CDX and Molfile data formats, as well as original TIFF and generated SVG images. For additional convenience, we collated the individual structure files into one structure-data file (SDF) per patent. During collation, the generated SDF files were filtered to remove duplicate molecules. The filtering ensures that duplicated structures are removed from individual patents' SDF files but not across patents, as shared molecules may be used to uncover legitimate relationships among the patents.

Patents often describe sets of molecules rather than, or in addition to, specific structures. Markush structure notation is commonly used to succinctly describe large (or infinite) molecular classes by choosing constituent molecular fragments from alternative substituents, positions, frequencies or homologies. The enumeration and comparison of Markush classes are significant challenges for cheminformatics systems (8). For example, Markush structures typically depict variable substituent placement as bonds that cross rings, as in Figure 2. These bonds frequently appear in Molfiles as separate propane molecules laid over the core scaffold. SCRIPDB provides basic Markush handling by extracting the molecular core and canonicalizing variable substituents, which permits Markush structures to be found via substructure searches. The original Markush structure is then retrievable from the original patent's structure or image files.

JURISICA LAB
IBM Life Sciences Discovery Center

We found 77 relevant molecules. Download their patents. Download all patents.

Page 1 of 1. Showing 77 molecules.

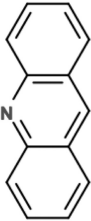
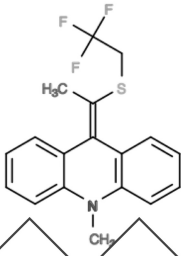
	<ul style="list-style-type: none"> Read US Patent 07842941 Download US Patent 07842941 SMILES: <chem>c1ccc2c(c1)nc1c(c2)cccc1</chem> Download file: .CDX .MOL .TIF .SVG
	<ul style="list-style-type: none"> Read US Patent 06270695 Download US Patent 06270695 SMILES: <chem>C/C(=c/c1c2cccc2n(c2c1cccc2)C)/SCC(F)(F)F</chem> Download file: .CDX .MOL .TIF .SVG

Figure 3. Sample of search results for molecules containing an acridine substructure.

Implementation

Raw patent data was downloaded from Google's bulk download of USPTO granted patents with embedded image data (<http://www.google.com/googlebooks/uspto-patents-redbook.html>). Filtering, collation and de-duplication were performed on an IBM CL1350 cluster with 1344 cores over 168 Infiniband-connected HS21-XM BladeServers and a DCS9550 storage system. The cluster runs the CentOS operating system, version 5.1, and manages coarse-grained parallelism with the Portable Batch System. SDF files were generated and duplicate structures were removed using OpenBabel, SVN revision 4487 (10). OpenBabel was also used to generate SVG files and compute canonical SMILES strings (<http://www.daylight.com/smiles/>) for display of retrieved search results, as shown in Figure 3.

The web interface to SCRIPDB was implemented in Python 2.6 using the Django web application framework, version 1.1 (11). Chemistry-specific search functionality, such as substructure searches or structural similarity using the Tanimoto coefficient (12) of OpenBabel FP2 linear structural fingerprints, is provided via integration with Pybel (13). Search structures may be specified via SMARTS queries, uploaded Molfiles, or via the interactive ChemWriter molecular editor (<http://metamolecular.com/chemwriter/>). ChemWriter is implemented in pure Javascript and permits SCRIPDB to be accessed from any major browser for desktop or iPad without the installation of external plugins.

RESULTS

Structures

New patents are granted each week, providing a steady stream of additional data for SCRIPDB. Here, we report results to the end of 2010, which is the last complete year of patent data. At the end of 2010, SCRIPDB contained 107 560 patents, including 4 814 913 non-redundant CDX structure files. Many structure files describe multiple molecules which, after de-duplication of molecules within a patent, yield a total of 10 840 646 molecules. Not only does this constitute a significant amount of total data (7) but the rate of structure disclosure appears to be growing. Both 2009 and 2010 had record numbers of molecules disclosed, at 1 259 097 and 1 639 522 structures, respectively. SCRIPDB data statistics are summarized in Figure 1.

However, focusing solely on the chemical structures ignores valuable chemical information. For many applications, molecular relationships need to be analyzed at various levels of granularity; for example, within a single CDX file, a particular patent, a group of patents that refer to a specific disease, or within the entirety of the database. As seen in Figure 1, patents typically contain multiple structural files which, in turn, contain multiple structures. For example, US Patent 6 884 815 contains 8 187 structure files (14).

Figure 4 shows the distribution of molecules per patent. The largest group contains patents that describe relatively few molecules (specifically, 10 or fewer). However, most patents catalog larger structural series. Two-thirds

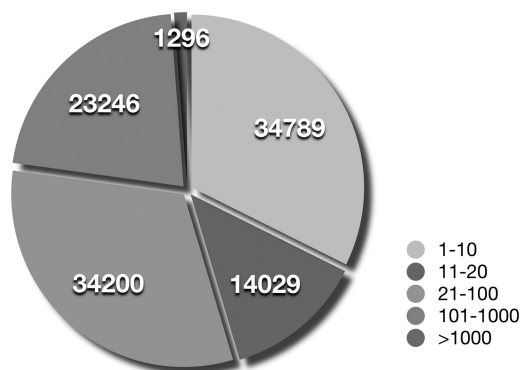


Figure 4. Structures per patent in SCRIpDB. While 34,789 patents contain ten or fewer structures, two-thirds of patents contain more than ten and 1,296 patents contain more than a thousand structures.

of patents contain more than 10 structures, over half of patents contain more than 20 structures, and almost 1 300 patents describe more than 1 000 structures each. A manual examination of a sample of patents shows that these structures often describe molecular analogs produced in the context of a medicinal chemistry optimization program.

Synthetic reactions

An important relationship described in CDX files is chemical synthesis. Figure 1b details the subset of patents that contain reactions. These are the patents that contain a CDX file with a ReactionStep object (<http://www.cambridgesoft.com/services/documentation/sdk/chemdraw/cdx/ReactionStep.htm>). Of the 107 560 patents from 2001 to 2010, SCRIpDB contains 25 048 patents that describe syntheses.

Fundamental molecular roles, such as reagent and product, are lost when a synthesis is split into separate structures. These relationships may be difficult to rederive, especially in multistep syntheses. Figure 5, which shows the number of reaction steps in SCRIpDB's syntheses, demonstrates that synthetic pathways requiring multiple reaction steps occur frequently in the patent literature. The 25 048 synthesis-containing patents describe a total of 341 764 individual reaction steps. While the most common are single-step syntheses, 52 462 syntheses (27.6%) have at least two reaction steps.

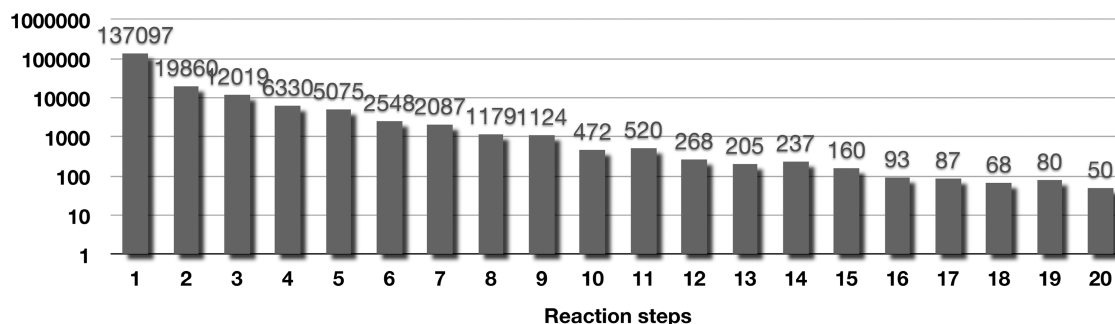


Figure 5. Number of CDX data files that describe syntheses of various lengths.

DISCUSSION

In addition to their direct value in intellectual property licensing (15) and competitive business analysis (16), patents can serve as a useful resource for a variety of academic research. We examine the datasets used in previous research and show that the data available in SCRIpDB is sufficient, quantitatively and qualitatively, to provide value for future investigations. Additionally, the information in patents is complementary to data available in the conventional academic literature (17), suggesting additional insight may be derived from integration with existing datasets. Specifically, we survey patent data as a tool for the development of chemical image parsing, biological text mining, reaction extraction, and bioisostere discovery.

Patents as a source of chemical images

The problem of optical structure recognition is to automatically extract molecular structure information from images. A necessary resource in the training and evaluation of such systems is a set of molecular images for which the true molecular structure is already known. Ideally, the images are representative of chemical images as commonly used in practice.

Validation sets comprising 6 185 images (18) and 454 images (19) were recently reported in the literature. By comparison, SCRIpDB provides millions of CDX structure and TIFF image files. This comprises a validation set several orders of magnitude larger than previously reported, even after eliminating possibly-confounding complex synthetic schemes. Furthermore, SCRIpDB contains SVG images produced by OpenBabel in addition to the original patent's TIFF images. This redundancy permits testing structure recognition algorithms for robustness to the idiosyncrasies of alternative image generation tools.

Patents as biomedical literature

Statistical analysis of the biomedical literature requires large quantities of freely accessible documents. Much of the biological text mining research has therefore used PubMed abstracts of journal articles while more recent full-text analyses have focused on Open Access journals (20). A large corpus of 162 259 full-text journal articles in HTML were used in the TREC Genomics track (21). The

107 560 patents in SCRIPDB constitute a document collection of comparable size.

Additionally, patents are interestingly complex. Patents are inherently semistructured and are partially annotated by their embedded chemical structures, gene sequences (7), mathematical equations, and data tables. Patents also contain citations to related papers (22) and patents, permitting co-citation analysis. Patent relationships may be derived and analyzed based on shared terms in patent text, shared molecules, patent assignee (6) or ontological categorization (<http://www.uspto.gov/web/offices/opc/documents/classescombined.pdf>).

Patents as a reaction database

Libraries of chemical transforms are used for combinatorial library design (23), static synthetic feasibility (24), and full retrosynthetic analysis (5). Although such libraries can be constructed by hand (25), automated extraction is an effective and labor-saving alternative (4,5,26,27). Such systems determine molecular substructures that are modified in the same manner across many different syntheses. These modifications can then be treated as putative reactions, predicting that if the same molecular substructure is found in a new molecule, it can be changed in the same way. Automated transform extraction therefore requires a large corpus of example syntheses within which to find consistent molecular changes.

In this capacity, SCRIPDB with its 341 764 reactions compares favorably to the 42 333 reactions in the Methods in Organic Synthesis database recently used for reaction extraction (5) or the 30 530 CCR reactions used to characterize functional group reactivity (28). While smaller than commercial databases containing millions of reactions, SCRIPDB has the virtue of being freely accessible.

Patents as a bioisostere catalog

After finding an initial lead molecule, medicinal chemists will typically create and test a large series of analogous compounds, seeking to increase binding affinity; improve absorption, distribution, metabolism, excretion and toxicity profiles; or avoid the intellectual property restrictions of competitors' patents. One systematic approach for exploring chemical space is to substitute bioisosteres, which are molecular fragments that have similar shape and chemical properties. For example, hydrogen and fluorine have similar van der Waals radii and the same valence. Exchanging a hydrogen with a fluorine permits the medicinal chemist to maintain the same molecular shape while optimizing the molecule's charge distribution.

There is strong interest both in techniques to determine bioisosteres (29,30) and in idea-generation tools that propose alternative molecules *en masse* by modifying a lead molecule *in silico* using the same approach as a medicinal chemist (31). Southall and Ajay (6) investigated bioisosteric replacements in kinase patents by analyzing 116 550 compounds. The maximum common substructure was computed for pairs of compounds and the remainder of the molecules were identified as exchangeable chemical replacements.

Southall and Ajay (6) were interested in the research strategies of drug companies, so only compared compounds found in patents assigned to different companies. A similar analysis can be performed for the compound series within a single patent (32). Each patent defines a set of bioisosteric replacements that were determined to be reasonable, interesting and synthetically feasible by a medicinal chemist. Figure 4 demonstrates that SCRIPDB contains sufficient patents with large chemical series to extract sensible chemical replacements.

CONCLUSION AND FUTURE WORK

The impetus of intellectual property protection creates a deluge of patents that carries enormous quantities of chemical and biological information. While patented molecules are accessible via chemical databases, the extraction of component structures from a patent needlessly removes information about chemical relationships. SCRIPDB is designed to make such metadata broadly accessible. We examined the information used in medical text mining, chemical image parsing, reaction extraction and the development of computational tools for lead optimization. We demonstrated that the quantity and quality of SCRIPDB's data compares favorably with existing commercial and free data sets. In many cases, it is complementary to existing data.

In the future, we plan to reduce the manual intervention necessary for the incorporation of new patents. Automatic patent processing is critical for maintaining the completeness of SCRIPDB, since new patents are released weekly. Automated updating will also support deposition of non-redundant molecules into PubChem (1). In addition, we wish to pursue integration with other value-adding databases, such as ChEMBL (33) and CDIP (<http://ophid.utoronto.ca/cdip>), and provide programmatic access to SCRIPDB via RESTful web services.

Currently SCRIPDB incorporates patents only from the United States, because the USPTO provides distinct structure files. However, valuable patent data is available from other countries' patent offices. Robust optical recognition of chemical structures would permit future integration with patent offices that provide molecules as images, such as the European Patent Office, the Japanese Patent Office, World Intellectual Property Organization, the UK Intellectual Property Office and the Canadian Intellectual Property Office.

ACKNOWLEDGEMENTS

We thank Kristen Fortney, Max Kotlyar, Izhar Wallach, and the anonymous referees for their valuable comments on earlier drafts. The views expressed do not necessarily reflect those of the Ontario Ministry of Health and Long Term Care.

FUNDING

Computational analysis was supported in part by Canada Foundation for Innovation (CFI #12301 and

CFI #203383); Ontario Research Fund (GL2-01-030); Canada Research Chair Program (to I.J., in part); Ontario Ministry of Health and Long Term Care (in part). Funding for open access charge: Ontario Research Fund (GL2-01-030).

Conflict of interest statement. None declared.

REFERENCES

- Wang, Y., Xiao, J., Suzek, T.O., Zhang, J., Wang, J. and Bryant, S.H. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, **37**, W623–W633.
- Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M. and Ashburner, M. (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, **36**, D344–D350.
- de Matos, P., Alcántara, R., Dekker, A., Ennis, M., Hastings, J., Haug, K., Spiteri, I., Turner, S. and Steinbeck, C. (2010) Chemical entities of biological interest: an update. *Nucleic Acids Res.*, **38**, D249–D254.
- Wilcox, C.S. and Levinson, R.A. (1986) *A Self-Organized Knowledge Base for Recall, Design, and Discovery in Organic Chemistry*, Vol. 306, American Chemical Society, pp. 209–230, ACS Symposium Series, chapter 18.
- Law, J., Zsoldos, Z., Simon, A., Reid, D., Liu, Y., Khew, S.Y., Johnson, A.P., Major, S., Wade, R.A. and Ando, H.Y. (2009) Route designer: a retrosynthetic analysis tool utilizing automated retrosynthetic rule generation. *J. Chem. Inf. Model.*, **49**, 593–602.
- Southall, N.T. and Ajay. (2006) Kinase patent space visualization using chemical replacements. *J. Med. Chem.*, **49**, 2103–2109.
- Li, W., McWilliam, H., de la Torre, A.R., Grodowski, A., Benediktovich, I., Goujon, M., Nauche, S. and Lopez, R. (2010) Non-redundant patent sequence databases with value-added annotations at two levels. *Nucleic Acids Res.*, **38**, D52–D56.
- Barnard, J.M. and Wright, P.M. (2009) Towards in-house searching of Markush structures from patents. *World Patent Inform.*, **31**, 97–103.
- Raveendranath, P., Zeldis, J., Vid, G., Potoski, J.R., Ren, J. and Iera, S. (1999) Aryloxy-alkyl-dialkylamines. *U.S. Patent Trademark Office*, US 6,268,504.
- Guha, R., Howard, M.T., Hutchison, G.R., Murray-Rust, P., Rzepa, H., Steinbeck, C., Wegner, J. and Willighagen, E.L. (2006) The Blue Obelisk – interoperability in chemical informatics. *J. Chem. Inf. Model.*, **46**, 991–998.
- Forcier, J., Bissex, P. and Chun, W. (2008) Python Web Development with Django. 1 edition, Addison-Wesley Professional, Boston, MA.
- Tanimoto, T.T. (1958) An elementary mathematical theory of classification and prediction. IBM, *Technical report*.
- O’Boyle, N., Morley, C. and Hutchison, G. (2008) Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chemistry Central Journal*, **2**, 5.
- Thurkauf, A., shu He, X., Zhao, H., Peterson, J., Zhang, X., Brodbeck, R., Krause, J., Maynard, G. and Hutchison, A. (2003) High affinity small molecule C5a receptor modulators. *U.S. Patent Trademark Office*, US 6,884,815.
- Chen, Y., Spangler, S., Kreulen, J., Boyer, S., Griffin, T.D., Alba, A., Behal, A., He, B., Kato, L., Lelescu, A. et al. (2009) SIMPLE: a strategic information mining platform for licensing and execution. *International Conference on Data Mining Workshops*, 270–275.
- Hattori, K., Wakabayashi, H. and Tamaki, K. (2008) Predicting key example compounds in competitors’ patent applications using structural information alone. *J. Chem. Inf. Model.*, **48**, 135–142.
- Thangaraj, H. (2007) Information from patent office could aid replication. *Nature*, **447**, 638–638.
- Filippov, I.V. and Nicklaus, M.C. (2009) Optical structure recognition software to recover chemical information: OSRA, an open source solution. *J. Chem. Info. Model.*, **49**, 740–743.
- Valko, A.T. and Johnson, A.P. (2009) CLiDE Pro: the latest generation of CLiDE, a tool for optical chemical structure recognition. *J. Chem. Info. Model.*, **49**, 780–787.
- Verspoor, K., Cohen, K.B. and Hunter, L. (2009) The textual characteristics of traditional and Open Access scientific journals are similar. *BMC Bioinformatics*, **10**, 183.
- Hersh, W. and Voorhees, E. (2009) TREC genomics special issue overview. *Inform. Retrieval*, **12**, 1–15.
- Griffin, T.D., Boyer, S.K. and Council, I.G. (2011) Annotating Patents with Medline MeSH Codes via Citation Mapping. *Advances in Experimental Medicine and Biology*, Vol. 680. Springer, New York, pp. 737–744.
- Pirok, G., Máté, N., Varga, J., Szegezdi, J., Vargyas, M., Dóránt, S. and Csizmadia, F. (2006) Making ‘Real’ molecules in virtual space. *J. Chem. Inf. Model.*, **46**, 563–568.
- Podolyan, Y., Walters, M.A. and Karypis, G. (2010) Assessing synthetic accessibility of chemical compounds using machine learning methods. *J. Chem. Inf. Model.*, **50**, 979–991.
- Corey, E.J. and Wipke, W.T. (1969) Computer-assisted design of complex organic syntheses. *Science*, **166**, 178–192.
- Satoh, K. and Funatsu, K. (1999) A novel approach to retrosynthetic analysis using knowledge bases derived from reaction databases. *J. Chem. Inf. Comp. Sci.*, **39**, 316–325.
- Wang, K., Wang, L., Yuan, Q., Luo, S., Yao, J., Yuan, S., Zheng, C. and Brandt, J. (2001) Construction of a generic reaction knowledge base by reaction data mining. *J. Mol. Graph. Model.*, **19**, 427–433.
- Tanaka, A., Okamoto, H. and Bersohn, M. (2010) Construction of functional group reactivity database under various reaction conditions automatically extracted from reaction database in a synthesis design system. *J. Chem. Inf. Model.*, **50**, 327–338.
- Sheridan, R.P. (2002) The most common chemical replacements in drug-like compounds. *J. Chem. Inf. Comp. Sci.*, **42**, 103–108.
- Langdon, S.R., Ertl, P. and Brown, N. (2010) Bioisosteric replacement and Scaffold Hopping in lead generation and optimization. *Mol. Inform.*, **29**, 366–385.
- Stewart, K.D., Shiroda, M. and James, C.A. (2006) Drug Guru: A computer software program for drug design using medicinal chemistry rules. *Bioorgan. Med. Chem.*, **14**, 7011–7022.
- Heifets, A. and Jurisica, I. (2011) Diversity as a first-class ranking measure in automated bioisosteric replacement. *Abstracts of Papers*. 242nd ACS National Meeting & Exposition, Denver, Colorado, United States, August 28–September 1, 2011; American Chemical Society, Washington, DC, COMP-279.
- Warr, W. (2009) ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). *J. Comput. Aid. Mol. Des.*, **23**, 195–198.