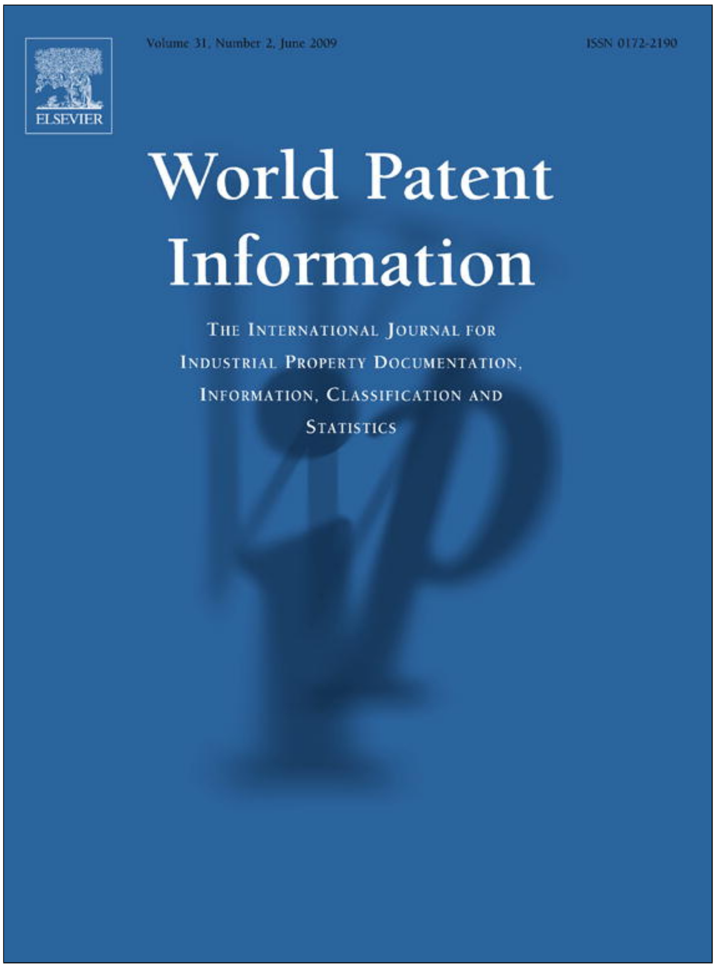


Volume 31, Number 2, June 2009 ISSN 0172-2190

World Patent Information

The International Journal for Industrial Property Documentation, Information, Classification and Statistics

World Patent Information 31 (2009) 97–103

Contents lists available at ScienceDirect

World Patent Information

journal homepage: www.elsevier.com/locate/worpatin

Towards in-house searching of Markush structures from patents ☆

John M. Barnard *, P. Matthew Wright

Digital Chemistry Ltd, The Iron Shed, Harewood House Estate, Harewood, Leeds LS17 9LF, UK

ARTICLE INFO

Keywords:
Markush structure
Chemical structure search
Oracle
Combinatorial library
Drug design
Patents
In-house searching

ABSTRACT

Most large pharmaceutical and biotechnology companies now use Oracle RDBMS chemistry data cartridges to manage databases of individual molecules for chemical structure searching. These systems are often linked to processes for new drug discovery and provide a common interface to a diverse range of specific structure databases. Recently some of these cartridges have been extended to handle Markush representations of un-enumerated combinatorial libraries alongside discrete molecules. An obvious extension would be to enable them to handle the Markush structures from chemical patents, though these have features and complexities not required for the representation of combinatorial libraries. The existing publicly available systems for handling patent Markush structures have changed little in the past 15 years and cannot easily be integrated with in-house systems; in-house access to chemical structures from patents is thus restricted at present to databases of specific molecules. A number of technical issues need to be tackled to enable the existing Markush-capable Oracle cartridges to handle data from patents, and several options are available for obtaining appropriate Markush structure databases for use with them. A demonstration system has been developed, using data from Thomson Reuters' World Patents Index Markush File, and Digital Chemistry's Oracle cartridge Torus. In-house access to patent Markush data could provide improved informatics support to the drug discovery process, both to enable patentability criteria to be added to computer-assisted drug design, and to expand the techniques available for data-mining in the patent literature.

1. Oracle chemistry cartridges

Most pharmaceutical and biotechnology companies maintain in-house database management systems, allowing access to information on chemical compounds in which they are interested [1]. The databases used in these systems may include internal compound registries (compounds synthesised by the company in question), chemical suppliers' catalogues, and publicly available chemical databases (either free or commercial). The systems not only permit storing and searching of the chemical structures and associated data, but may also include modules for prediction of physico-chemical properties, and for various sorts of analysis and data-mining.

The industry-standard relational database management system is Oracle [2] which, in its standard form, offers text and numeric searching only. Chemical structure searching is provided by third-party chemistry "cartridges", which "plug into" Oracle, and allow chemical structure-based queries to be formulated using Oracle's standard query language, SQL, and integrated with text and numeric searches. There are about a dozen commercially available Oracle chemistry cartridges, from a variety of vendors. Table 1 lists those believed to be available in March 2008. A few user companies have developed their own chemistry cartridges (e.g. [3]), in some cases using commercially available tools to actually perform the substructure searching and other operations.

The exact capabilities for chemical structure searching differ between cartridges, but generally they all offer:

- full structure searching (find a specified molecule in the database),
- substructure searching (find molecules in the database that contain a specified substructure),
- similarity searching (find molecules in a database which are structurally similar to a specified target molecule).

Various analysis tools may also be associated with the chemistry cartridge, for example to calculate estimates of physico-chemical property values for the molecules in a database. Frequently-calculated properties include the well-known Lipinksi "Rule of Five" properties (molecular weight, counts of hydrogen bond acceptors and donors and rotatable bonds in the molecule, and the octanol-water partition coefficient). Lipinski et al. [4] have suggested that molecules whose values for these properties lie outside

☆ The substance of this paper was presented orally at IPI-Confex, Seville, Spain, 3–5 March 2008 and as a poster at the 8th International Conference on Chemical Structures, Noordwijkerhout, The Netherlands, 1–5 June 2008.

* Corresponding author. Tel.: +44(0)114 233 3170.

E-mail addresses: john.barnard@digitalchemistry.co.uk (J.M. Barnard), matthew.wright@digitalchemistry.co.uk (P.M. Wright).

0172-2190/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wpi.2008.09.012

Table 1
Commercially available Oracle chemistry cartridges (websites visited 9 June 2008).

Supplier	Cartridge name	Website
Accelrys, Inc.	Accord Chemistry Cartridge	http://accelrys.com/products/accord/tools-components/accord-chemistry-cartridge.html
CambridgeSoft Corporation	Oracle Cartridge	http://www.cambridgesoft.com/solutions/details/?es=2&esv=5
ChemAxon Kft.	JChem Cartridge	http://www.chemaxon.com/product/jc_cart.html
ChemNavigator.com, Inc.	ChemMatrix	http://www.chemnavigator.com
Daylight CIS Inc.	DayCart	http://www.daylight.com/products/daycart.html
Digital Chemistry Ltd	Torus	http://www.digitalchemistry.co.uk/prod_torus.html
Dotmatics Limited	pinpoint	http://www.dotmatics.com/products_pinpoint.jsp
ID Business Solutions Ltd.	ActivityBase Chemistry	http://www.idbs.com/activitybase/chemistry/
InfoChem GmbH	ICCartridge	http://infochem.de/en/products/software/iccartridge.shtml
Symyx Technologies Inc.	MDL Direct	http://www.mdl.com/products/framework/rel_chemistry_server/index.jsp
Tripos	AUSPYX	http://www.tripos.com/index.php?family=modules,SimplePage,,,&page=auspyx

certain specified ranges are unlikely to be successful drugs. Other analysis tools may include ones to cluster the molecules into groups [5]. This allows selection of representative molecules from a database (one molecule from each cluster) for synthesis and testing; if an "active" molecule is found, then other molecules from the same cluster may also be synthesised and tested.

The currently available Oracle chemistry cartridges are generally only able to handle databases of individual specific molecules, each of which forms a single row of the Oracle table. The chemistry search interfaces may have some "Markush-like" features (e.g. variable atom lists, or R-groups), but these relate to the structure queries only, and not to the chemical structures in the databases being searched. A few recently introduced cartridges do have some limited capabilities for handling Markush structures in the database. These are Digital Chemistry's Torus, ChemAxon's JChem, and Accelrys's Accord Markush Extension for AEI (see Table 1) though the last of these in fact uses some features of the Torus search engine, under licence from Digital Chemistry. Search capabilities may include full structure, substructure and similarity search (all without the need for enumeration of the individual molecules covered by the Markush structure) and the identification of the "overlap" between two Markush structures.

A major motivation behind the development of Markush structure-handling capabilities in Oracle chemistry cartridges has been the need to handle extremely large virtual combinatorial libraries of molecules [6]. These libraries represent large sets of structurally related compounds, often including billions or trillions of individual molecules, which *might* be synthesised as part of a drug discovery programme, and it is clearly impractical to store and search them individually in a conventional Oracle cartridge. In silico analysis of the library can be used to identify small subsets of its members that would be *worth* synthesising and testing as potential drug leads. Such analysis can be based on property estimations, similarity to known active compounds, or elimination of molecules containing substructures associated with toxicity or other problems. An obvious extension to this type of analysis might be the elimination of molecules claimed or disclosed by competitors' patents.

2. Markush structures

Strictly speaking, a "Markush Claim" is a patent claim made under a particular US legislative precedent, though in general usage "Markush" has come to mean a chemical structure with variable parts, conventionally shown as "R-groups" etc. For combinatorial libraries, the Markush structures are usually quite simple, involving two or three R-groups at fixed positions around a central scaffold (see Fig. 1). Each R-group may have up to several thousand alternative members, giving a library that includes anything from a few tens to many billions of molecules. In some cases there may be quite complex "nesting" of the R-groups, especially where the library is based on a multi-step synthesis scheme.

In contrast, the Markush structures found in chemical patents tend to be a lot more complicated, with a small or vestigial scaffold, many more R-groups, often deeply nested, and substituent groups attached at variable positions (see Fig. 2).

Other features of the Markush structures typically encountered in chemical patents include substituent groups occurring a variable number of times in different positions, head-to-tail concatenation of repeating groups (with variable repetition counts) and the use of generic nomenclature to define R-groups (e.g. from Fig. 2: "an optionally substituted five-, six- or seven-membered saturated, unsaturated or partially saturated heterocycle or bicyclic heterocycle containing up to two heteroatoms…"). The use of this type of expression means that patent Markush structures often cover infinite numbers of individual molecules, making enumeration infeasible.

Two commercial systems have been available since the late 1980s for chemical structure-based searching of the patent literature, each with its own proprietary database and dedicated search software, mounted on its own online host system. These are the Markush DARC Merged Markush Service (MMS) available though the online host Questel, and Chemical Abstracts MARPAT system, available through the online host STN. These systems have evolved relatively little (especially with respect to their user interfaces) during the last 15 years, and are discussed and compared in several literature articles [7–9].

3. In-House Access to Patent Chemistry Data

Pharmaceutical and biotechnology companies are becoming increasingly interested in accessing chemical structure data from patents using databases and systems inside their corporate Internet "firewalls". To do so allows them to integrate patent data with other in-house data sources, so that they can all be accessed using simple interfaces available to all chemists. This integration allows structural information from patents to be used more effectively in the early stages of drug discovery (i.e. lead generation) not only to alert chemists to the existence of competitors' patents in the area of chemical space they are exploring but also to allow some of the data mining tools that have been developed in recent years for analysis of chemical structure databases [10] to be applied to patent literature as well. A further benefit of in-house searching is confidentiality, since structure queries do not need to pass over a public network.

At present, in-house access to chemical structure data from patents is largely confined to specific compounds mentioned or claimed in patent documents; Markush structures are not accessible. Several databases of chemical structures are commercially available, of which the best-known is probably Elsevier's CrossFire Patent Chemistry Database [11], which allows searching of exemplified (and some prophetic) compounds claimed in World, European and US patents, along with display of relevant Markush

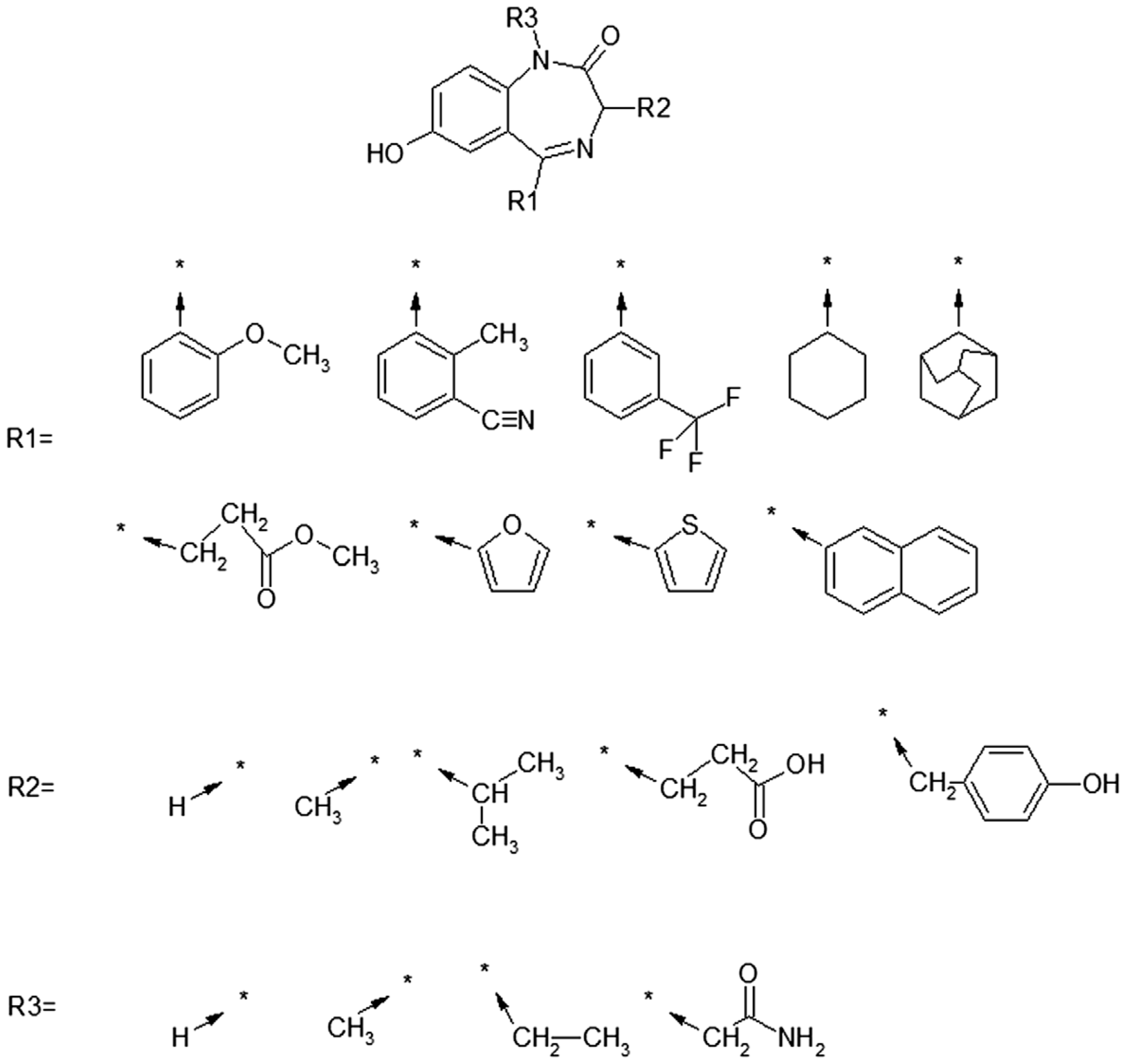
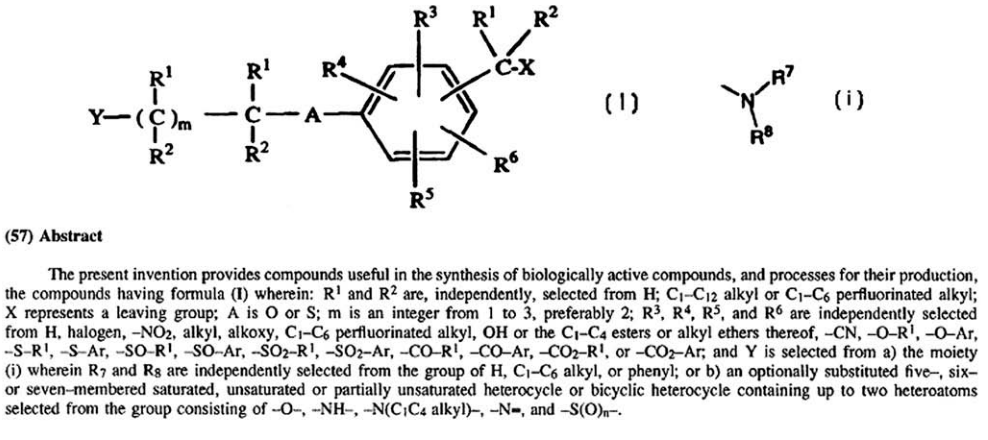


Fig. 1. A Markush structure representing a combinatorial library.


R3
R1
R2
R4
C-X
R1
R1
Y—(C)m
C
A
R2
R2
R6
R5
(I)
N
R7
R8
(i)

(57) Abstract

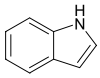
The present invention provides compounds useful in the synthesis of biologically active compounds, and processes for their production, the compounds having formula (I) wherein: R^1 and R^2 are, independently, selected from H; C_1–C_{12} alkyl or C_1–C_6 perfluorinated alkyl; X represents a leaving group; A is O or S; m is an integer from 1 to 3, preferably 2; R^3, R^4, R^5, and R^6 are independently selected from H, halogen, $-NO_2$, alkyl, alkoxy, C_1–C_6 perfluorinated alkyl, OH or the C_1–C_4 esters or alkyl ethers thereof, –CN, $-O-R^1$, –O–Ar, $-S-R^1$, –S–Ar, $-SO-R^1$, –SO–Ar, $-SO_2-R^1$, $-SO_2-Ar$, $-CO-R^1$, –CO–Ar, $-CO_2-R^1$, or $-CO_2-Ar$; and Y is selected from a) the moiety (i) wherein R_7 and R_8 are independently selected from the group of H, C_1–C_6 alkyl, or phenyl; or b) an optionally substituted five–, six– or seven–membered saturated, unsaturated or partially unsaturated heterocycle or bicyclic heterocycle containing up to two heteroatoms selected from the group consisting of –O–, –NH–, $-N(C_1C_4$ alkyl)–, –N=, and $-S(O)_n-$.

Fig. 2. Part of a Markush structure from a patent specification.

structures (though these are not searchable). Thomson Reuters' Chemistry Resource [12] provides searchable specific chemical structures from more than one million patents covered by the Derwent World Patents Index, and is available both online on various hosts, and via an FTP data feed for in-house use.

Amongst other available databases, the SURECHEM database, produced by ReelTwo Ltd. [13] is built by automatic analysis of patent text, identifying and extracting chemical names, and converting them to structure-searchable records using chemical name-translation software. There have been significant improvements in the accuracy of automatic chemical name identification and conversion software in recent years [14], though Walsh [15] has commented that, based on his experience with using this type of data in a large pharmaceutical company, there remain significant problems with transferring structural data from nomenclature in free-text documents to structure-searchable databases, and that "curated" data remains the "gold standard". In support of this, a recent study by Emmerich [16] has compared the retrieval performance of patent full text files with value-added databases, when searching for patents concerned with a particular drug, and found that value-added databases provided more than 30% of the unique references, and that many of the free-text results were only marginally relevant.

In addition to these commercial databases, a few companies have been building their own databases containing chemical structures from patents of particular interest; in at least one case known to the authors the structure records include some limited Markush features, though no details have been revealed.

Clearly, if those Oracle chemistry cartridges that are currently capable of handling limited forms of Markush structure (e.g. as representations of combinatorial libraries) are to be able to search chemical patent databases, they will require some extension. In particular, they will need to be able to represent and appropriately search the additional forms of structural variability that are characteristic of Markush structures from patents, but are not found in combinatorial libraries. In the classification proposed by Dethlefsen et al. [17], these are: *p*-variability (variability in the position of attachment of a substituent group), *f*-variability (variability in the frequency of occurrence of a group, either in head-to-tail concatenation, or as multiple substituents in variable positions) and *h*-variability (variability based on specification of a homologous series of groups, such as "alkyl" or "heteroaryl"). The first two of these can, in principle at least, be internally "disguised" as simple *s*-variability (variability produced by listing a series of alternative substituent groups). The third requires a more radical approach, which allows matching between specific atom-bond chemical structures, such as  and *h*-variant expressions such as "fused heteroaromatic", a feature sometimes described as "translation". Appropriate representations and search algorithms have previously been developed for this, and are implemented in the existing commercial patent Markush search systems [7], using devices such as "superatoms" (Markush DARC) and "generic group nodes" (MARPAT). Other approaches may also be appropriate, and may need to take into account the need to implement features beyond the simple structure searching available in the existing systems, e.g. for similarity search, property calculation, diversity analysis and data mining. Enumeration (either complete or selective) of structures and calculated properties of individual molecules is often useful in analysis of combinatorial libraries, but for Markush structures from patents it may be infeasible, especially where *h*-variation occurs. Thus the enumeration capabilities in Markush-capable Oracle cartridges may need to be restricted for some types of Markush.

An important consideration for in-house access to Markush structure data is the source of the database(s) to be searched. The most obvious option is for the producers of the existing published Markush patent databases to make them available for in-house use, but this will demand their willingness to make appropriate commercial agreements with software producers and end-users, and the relevant Oracle cartridges will need to be able to read their data formats. Building of new Markush databases is clearly an alternative, and may prove suitable for companies interested only in a small number of patents, for example in a particular therapeutic area. Appropriate graphical input software will also be needed, as existing chemical editors do not have sufficient sophistication to draw the complex Markush structures found in patents and generate searchable representations of them. A third possibility is (semi-)automatic processing of original machine-readable patent documents. These are now readily available from patent offices, but tend to treat the chemical structure diagrams purely as images, and do not establish direct links between the symbols used in the diagrams (e.g. "R1") and their definitions, which are shown as a mixture of text (including linear formulae, systematic and trivial nomenclature) and structure diagrams. Such processing would thus require accurate structure diagram recognition and chemical name identification and translation software, along with sophisticated natural-language processing able to make the appropriate semantic links. The progress in all these areas in recent years has already been mentioned [14], and in the long term this is likely to prove the most satisfactory approach to building Markush structure databases (or indeed, any sort of patent database) for searching. However, as also noted above, the conclusions of Walsh [15] and Emmerich [16] concerning the current superiority (in terms of retrieval performance) of curated or value-added databases, suggest that a substantial element of human involvement in the extraction of Markush data from patent documents is likely to be needed for some time to come. In principle it may also be possible for Patent Offices to be involved in establishing an electronic submission format for Markush structure data.

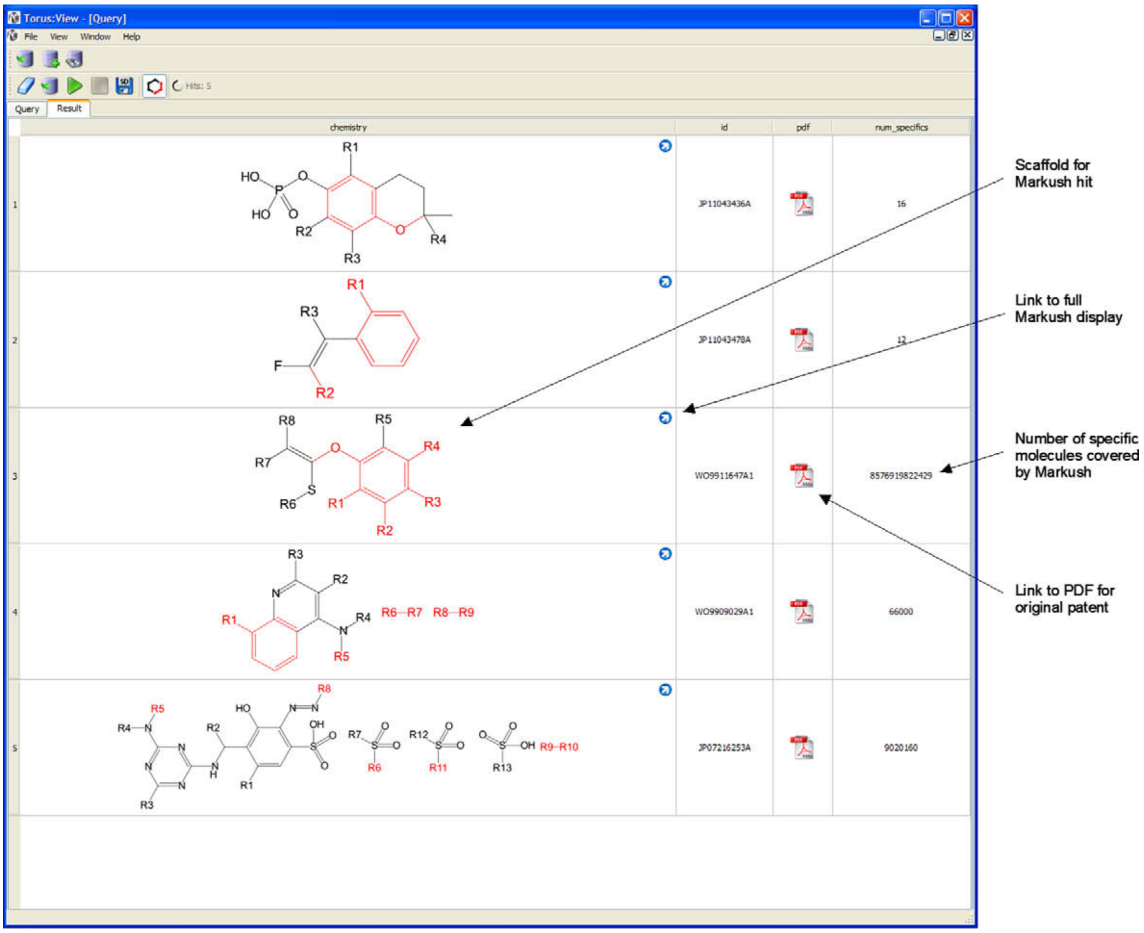
4. A demonstration of Markush patent searching

Digital Chemistry Ltd and the Scientific Business of Thomson Reuters have recently collaborated to demonstrate how a patent Markush structure database might be searched using an Oracle chemistry cartridge. A demonstration system was put together in which a small number of records from the Derwent World Patents Index Markush file (part of the Merged Markush Service), in Markush DARC format, were read into an Oracle data table using the Torus Oracle cartridge, and appropriate search records built in order to allow structure searches to be performed on them. Torus has been developed to allow Markush structures representing combinatorial libraries to be stored and searched alongside individual specific molecules within Oracle, and this clearly limits the current capabilities of the demonstration system, though extensions to Torus are planned, which will allow it to handle the more complex features found in Markush structures from patents. In particular, Torus not yet able to handle the *h*-variant expressions represented by Markush DARC superatoms, and so these were converted to heavy metal atoms unlikely to occur elsewhere in the structures.

Torus allows queries to be specified either as specific molecules (full structure search) or as substructures. Though the Torus client application, Torus:View, has interfaces to standard structure and substructure drawing programs, such as MDL Draw [18], the queries are converted and passed to the search engine as SMILES or SMARTS text strings [19], which are line notations commonly used for compact representation of chemical structures and substructure queries. Fig. 3 illustrates the display of the results table from a simple substructure search; Fig. 4 shows the display of the full Markush structure for one of the hits.

There are a number of points to note concerning the displays in Fig. 3 and 4.

- The results table display in Fig. 3 shows just the "core" of each matching Markush structure with a link to the display of the full Markush structure.
- The results table display also has a link to display the PDF file for the original patent document.
- The number of specific molecules shown in the final column of the results table is calculated by appropriately multiplying



	chemistry	id	pdf	num_specifics
1		JP11043436A		16
2		JP11043478A		12
3		WO9911647A1		8576919822429
4		WO9909029A1		66000
5		JP07216253A		9020160

Scaffold for Markush hit

Link to full Markush display

Number of specific molecules covered by Markush

Link to PDF for original patent

Fig. 3. Display screen for results of a substructure search in Digital Chemistry's Torus Oracle Cartridge, using a file of Markush structures from the Merged Markush Service file. The query used for this search was the SMARTS pattern Oc1ccccc1, which represents a phenol group substituted at any position. In the structure display for each hit, the matching substructure (or any R-group in which it, or part of it, occurs) is highlighted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

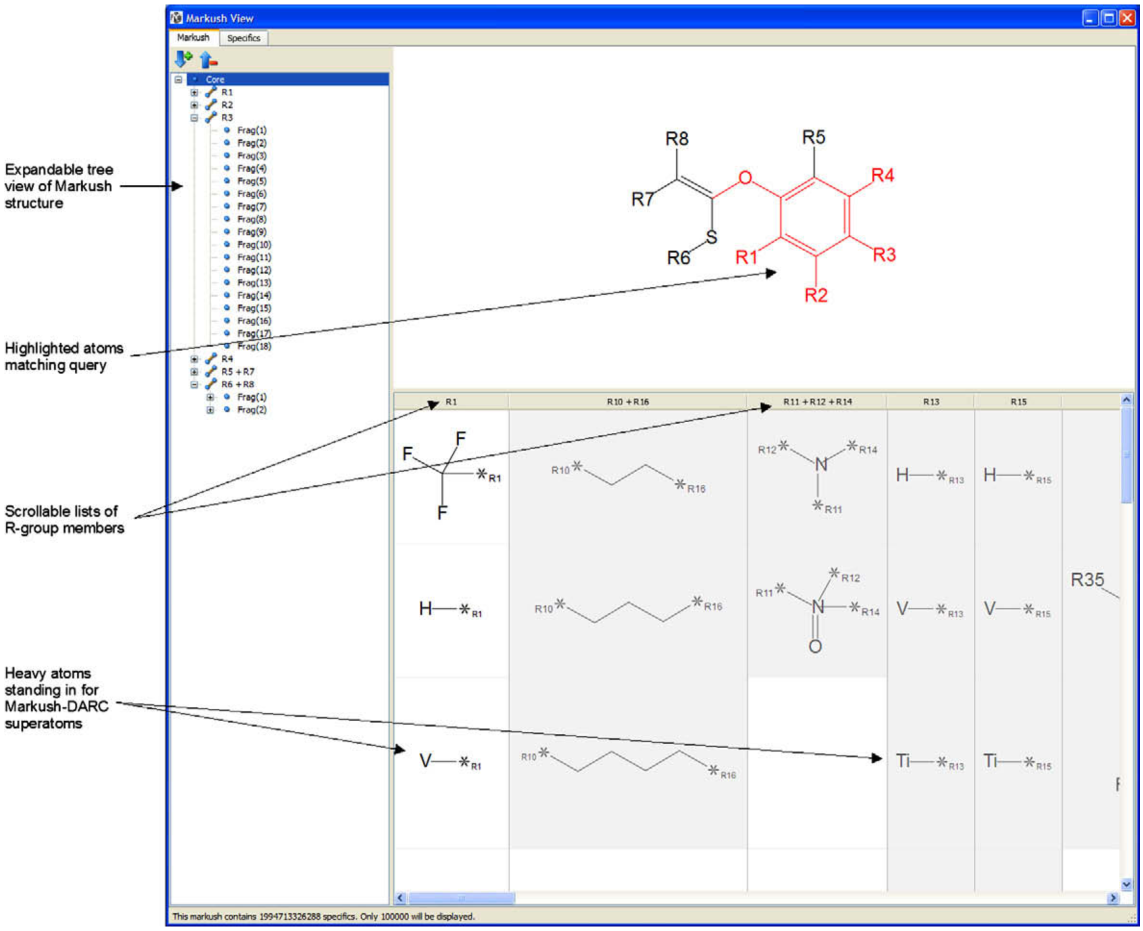
together the number of members for each R-group (taking any nesting of R-groups into account). It does not consider the number of different specific alternatives that may be covered by single Markush-DARC superatoms such as CHK and HET.

- In this demonstration, heavy atoms such as V and Ti stand in for the superatoms, so as to allow the data to be searched using the current capabilities of Torus.
- Not all the scrollable columns of R-group members can be shown on a single screen in the full Markush display (Fig. 4) because of the size and complexity of the structure involved; the horizontal scroll-bar at the bottom of the screen allows these to be accessed.
- Columns are shown for some R-groups that do not appear in the Markush core; these represent the members of R-groups that are nested within the member values of other R-groups.
- Each "Rn" symbol represents a single attachment for an R-group; in some cases several "Rn" attachments are defined in combination in a single column.

The search engine in Torus also has the capability of identifying the "overlap" between two combinatorial libraries represented by Markush structures – that is, the specific molecules that are common to both. This set of molecules can also be displayed as a Markush structure, though in the small number of examples from the Derwent WPIM database that were used for this demonstration, no overlaps were in fact found.

Several aspects of the demonstration system require further work if an operational system based on it is to be brought to full commercial development. The need to be able to handle Markush DARC superatoms, allowing "translation" between specific (atom-bond) and generic (superatom) representations was mentioned in Section 3. Various intermediate stages, providing sub-optimal but usable search features, can be identified for this, and we are currently considering the most appropriate development path. Full handling will require extension to the structure query capabilities of standard SMARTS [19], in order to allow the user to search for superatoms (or other equivalent generic chemical descriptions) and to provide user control over the extent to which specific and generic groups should be allowed to match in any particular instance.

The process by which Markush structures are input to the Merged Markush Service database (largely dictated by the constraints of the Markush DARC system) results in some reorganization of the structure diagrams and thus the diagrams displayed in Torus (and for that matter in the online Markush DARC system) sometimes look very different to those appearing in the original patent. In addition, nomenclatural terms used in the patent are replaced by structure diagrams, even for trivial groups like methyl and chloro. (In contrast the display in MARPAT makes extensive use of nomenclatural terms, where appropriate). There is clearly a great deal of scope for various forms of automatic analysis of the Markush DARC data (perhaps in conjunction with analysis of the original patent document) in order to try to bring the display



This markush contains 1994713326288 specifics. Only 100000 will be displayed.

Fig. 4. Display of the full Markush structure for one of the "hits" from Fig. 3. Further discussion in text.

of search results within Torus closer to the original. This type of processing could also form a useful part of an operational system based on the demonstration.

5. Uses of in-house access to patent Markush structures

The availability of in-house access to databases of Markush structures from patents could bring useful informatics support to drug discovery in pharmaceutical and biotechnology companies. It would permit the integration of patents with existing databases and search software, and enable patentability criteria to be used in drug design. It could also provide a mechanism to support data mining of the chemical data in patents, expanding the use of this data for structure/activity analyses, competitive intelligence and the identification of un-patented "gaps" in chemical space. It may also have a role as an adjunct to the existing public systems for novelty and infringement searches in the patent literature.

Calcagno [20] has recently discussed the use of the Derwent WPI chemical fragment codes to cluster together patents claiming similar chemical structures. Oracle chemistry cartridges are already able to generate chemical structure "fingerprints" for Markush structures, based on the occurrence of substructure fragments, and these thus provide an equivalent to the Derwent fragment codes. Different fingerprints can be designed, which are optimised for use either in similarity calculation or in search screening. It is possible to distinguish between fragments that occur in all molecules covered by the Markush, and those that occur only in certain molecules. "Modal" fingerprints [21] can also be generated, which are based on the proportion of molecules from the Markush in which each fragment appears; this type of fingerprint could be used for more sophisticated similarity calculations between Markush structures.

Acknowledgements

We are grateful to Tim Miller and Linda Clark (Thomson Reuters) for making the Markush DARC records available to us, and for helpful discussions. We also acknowledge the contributions made by our colleagues at Digital Chemistry Ltd. to the development of Torus, and the demonstration patent search system described above, in particular Tony Cook, Martin Gronow, Richard Hampshire, Annette von Scholley-Pfab, and Dan Thomas.

References

[1] Miller MA. Chemical database techniques in drug discovery. Nat Rev Drug Discov 2002;1(3):220–7.
[2] Oracle Corporation, 500 Oracle Parkway, Redwood Shores, CA 94065, USA. www.oracle.com.
[3] Watson P, Verdonk M, Hartshorn MJ. A web-based platform for virtual screening. J Mol Graph Model 2003;22:71–82.
[4] Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliver Rev 2001;46(1):3–26.
[5] Downs GM, Barnard JM. Clustering methods and their uses in computational chemistry. Rev Comp Chem 2002;18:1–40.
[6] Green DVS. Virtual screening of virtual libraries. Prog Med Chem 2003;41: 61–97.
[7] Barnard JM. A comparison of different approaches to Markush structure handling. J Chem Inf Comput Sci 1991;31:64–8.
[8] Berks AH. Current state of the art of Markush topological search systems. World Patent Inform 2001;23:5–13.

[9] Simmons ES. Markush structure searching over the years. World Patent Inform 2003;25:195–202.

[10] Weaver DC. Applying data mining techniques to library design, lead generation and lead optimization. Curr Opin Chem Biol 2004;8:264–70.

[11] Available from: http://info.patentchemistrydatabase.com/ (accessed 10 June 2008).

[12] Available from: http://scientific.thomsonreuters.com/products/chemistryresource/ (accessed 10 June 2008).

[13] Available from: http://surechem.reeltwo.com/ (accessed 10 June 2008).

[14] Banville D. Mining chemical structural information from the drug literature. Drug Discov Today 2006;11:35–42.

[15] Walsh D. Integration of chemical structures from patents for pharmaceutical discovery: a Pfizer perspective. Presented at IPI-Confex, Seville, Spain, 2–5 March 2008. Available from: (http://www.ipi-confex.com/).

[16] Emmerich C. Comparing first-level patent data with value-added patent information: a case study in the pharmaceutical field. World Patent Inform 2008 (in press) 10.1016/j.wpi.2008.06.003.

[17] Dethlefsen W, Lynch MF, Gillet VJ, Downs GM, Holliday JD, Barnard JM. Computer storage and retrieval of generic structures in patents. 11. Theoretical aspects of the use of structure languages in a retrieval system. J Chem Inf Comput Sci 1991;31:233–53.

[18] MDL, 2440 Camino Ramon, Suite 300 San Ramon, CA 94583, USA. Available from: http://www.mdli.com/products/framework/mdl_draw/index.jsp.

[19] Available from: http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.

[20] Calcagno M. An investigation into analyzing patents by chemical structure using Thomson's Derwent World Patent Index codes. World Patent Inform 2008;30:188–98.

[21] Shemetulskis NE, Weininger D, Blankley CJ, Yang JJ, Humblet C. Stigmata: an algorithm to determine structural commonalities in diverse datasets. J Chem Inf Comput Sci 1996;36:862–71.

John Barnard has been working with Markush structures since he started his PhD work on the machine-readable representation of them at Sheffield University in 1979. He remained with the Sheffield Markush project team until 1985, when he founded Barnard Chemical Information Ltd. The company provided consultancy and software development services to many of the world's leading pharmaceutical companies and chemical information providers, and in 2005 it was acquired by the newly-formed Digital Chemistry Ltd., of which he is now Scientific Director.

Matthew Wright obtained his PhD in Computational Chemistry from Sheffield University in 1996. He has worked in various roles for a number of leading chem-infomatics companies, and joined Barnard Chemical Information in 2002, before going on to co-found Digital Chemistry in 2005. In his present role as Product Director, he is responsible for the Torus product line for chemical searching and analysis, which is currently being extended to allow patent Markush structures to be stored and retrieved.