Searching for Structural Information in Patents and Using it for Drug Discovery

Dr John M. Barnard

Scientific Director

Digital Chemistry Ltd., UK

Presented at Optibrium Consultants' Day

Cambridge, 27th November 2012

digital chemistry

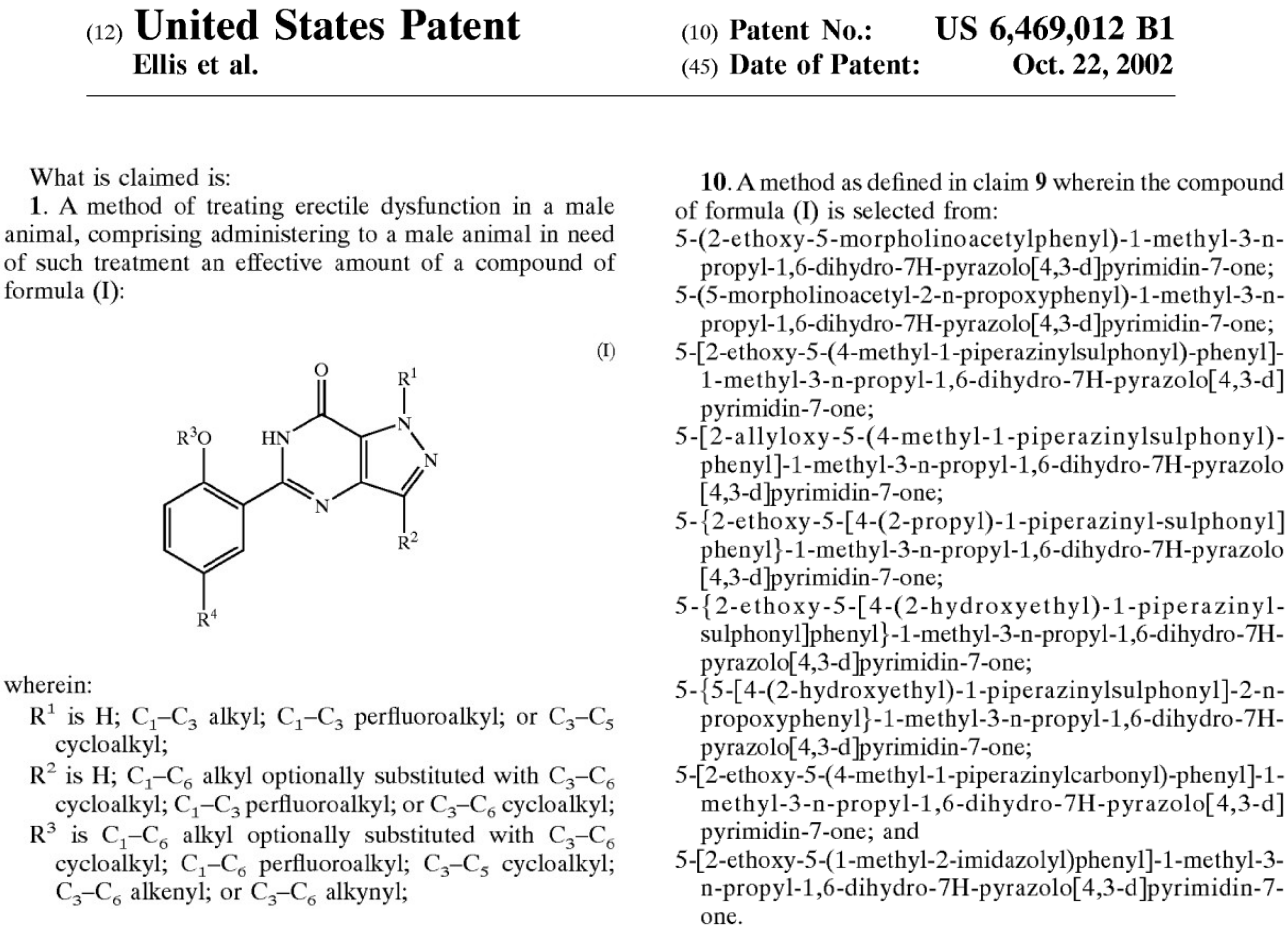
What are patents for?

- Contract between government and inventor to encourage innovation
  - inventor reveals nature of invention
  - government grants monopoly over exploitation for limited period
- Invention must be novel, useful and non-obvious
  - may include new compounds, new uses for existing ones, new synthesis methods, formulations, etc.
  - also non-obvious advantages of subsets of known compounds
- Essential to traditional business model of pharmaceutical industry

Structural information in patents

Markush claim

Claimed example compounds



(12) **United States Patent**
Ellis et al.

(10) **Patent No.:** **US 6,469,012 B1**
(45) **Date of Patent:** **Oct. 22, 2002**

What is claimed is:

1. A method of treating erectile dysfunction in a male animal, comprising administering to a male animal in need of such treatment an effective amount of a compound of formula (I):

(I)

O
R1
R3O
HN
N
N
N
R2
R4

wherein:

R^1 is H; C_1–C_3 alkyl; C_1–C_3 perfluoroalkyl; or C_3–C_5 cycloalkyl;

R^2 is H; C_1–C_6 alkyl optionally substituted with C_3–C_6 cycloalkyl; C_1–C_3 perfluoroalkyl; or C_3–C_6 cycloalkyl;

R^3 is C_1–C_6 alkyl optionally substituted with C_3–C_6 cycloalkyl; C_1–C_6 perfluoroalkyl; C_3–C_5 cycloalkyl; C_3–C_6 alkenyl; or C_3–C_6 alkynyl;

10. A method as defined in claim **9** wherein the compound of formula (I) is selected from:

5-(2-ethoxy-5-morpholinoacetylphenyl)-1-methyl-3-n-propyl-1,6-dihydro-7H-pyrazolo[4,3-d]pyrimidin-7-one;

5-(5-morpholinoacetyl-2-n-propoxyphenyl)-1-methyl-3-n-propyl-1,6-dihydro-7H-pyrazolo[4,3-d]pyrimidin-7-one;

5-[2-ethoxy-5-(4-methyl-1-piperazinylsulphonyl)-phenyl]-1-methyl-3-n-propyl-1,6-dihydro-7H-pyrazolo[4,3-d]pyrimidin-7-one;

5-[2-allyloxy-5-(4-methyl-1-piperazinylsulphonyl)-phenyl]-1-methyl-3-n-propyl-1,6-dihydro-7H-pyrazolo[4,3-d]pyrimidin-7-one;

5-{2-ethoxy-5-[4-(2-propyl)-1-piperazinyl-sulphonyl]phenyl}-1-methyl-3-n-propyl-1,6-dihydro-7H-pyrazolo[4,3-d]pyrimidin-7-one;

5-{2-ethoxy-5-[4-(2-hydroxyethyl)-1-piperazinyl-sulphonyl]phenyl}-1-methyl-3-n-propyl-1,6-dihydro-7H-pyrazolo[4,3-d]pyrimidin-7-one;

5-{5-[4-(2-hydroxyethyl)-1-piperazinylsulphonyl]-2-n-propoxyphenyl}-1-methyl-3-n-propyl-1,6-dihydro-7H-pyrazolo[4,3-d]pyrimidin-7-one;

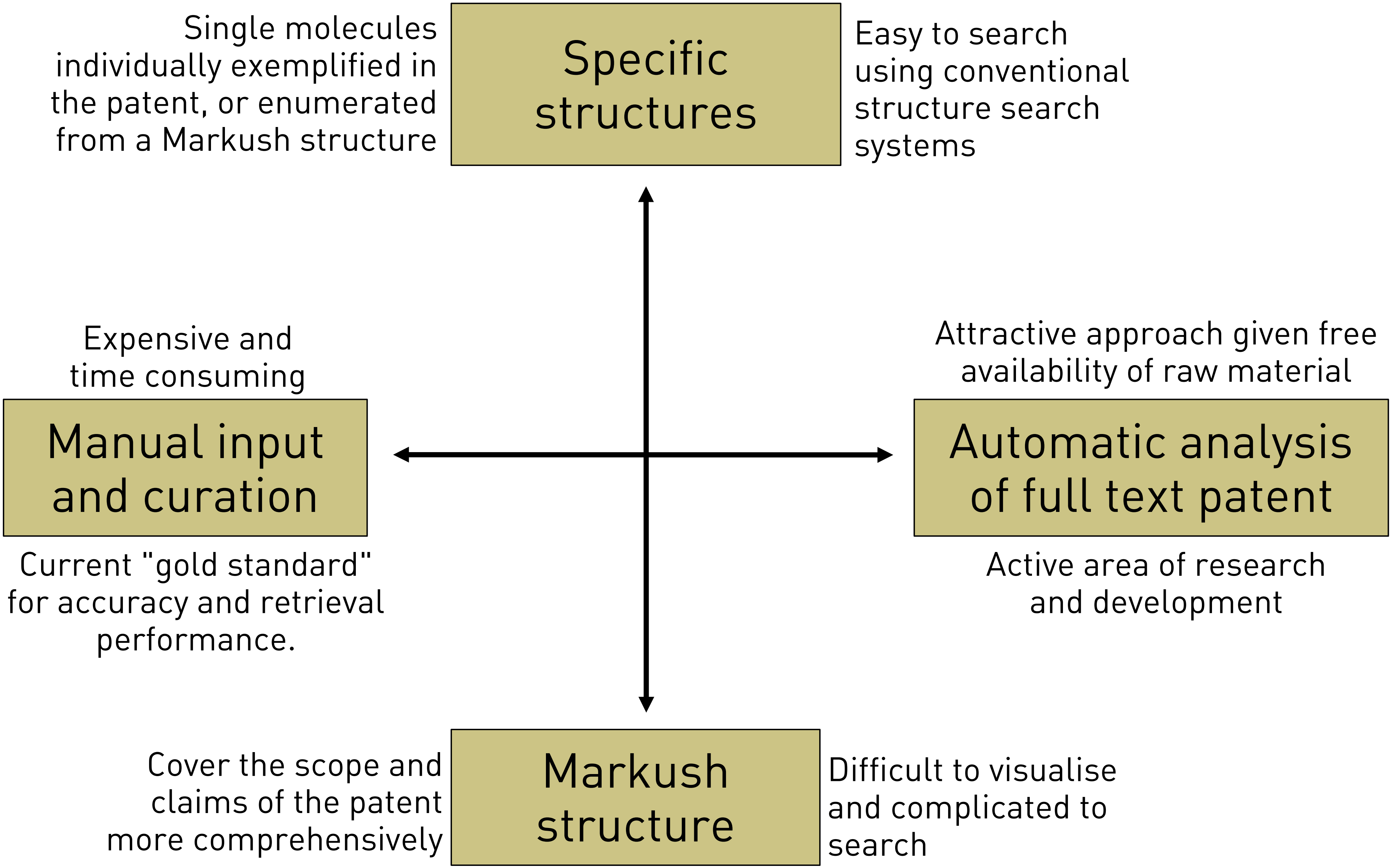
5-[2-ethoxy-5-(4-methyl-1-piperazinylcarbonyl)-phenyl]-1-methyl-3-n-propyl-1,6-dihydro-7H-pyrazolo[4,3-d]pyrimidin-7-one; and

5-[2-ethoxy-5-(1-methyl-2-imidazolyl)phenyl]-1-methyl-3-n-propyl-1,6-dihydro-7H-pyrazolo[4,3-d]pyrimidin-7-one.

Structure search in patents

- Established commercial specific structure databases
- Established commercial Markush systems
  - Markush DARC (MMS) and MARPAT
  - launched in late 1980s but little changed since then
- Much recent and current activity
  - data mining of patent text
  - new databases of specific structures from patents
  - new search software
  - integration of patent data into drug discovery informatics

Chemical patent databases



Single molecules individually exemplified in the patent, or enumerated from a Markush structure

Specific structures

Easy to search using conventional structure search systems

Expensive and time consuming

Manual input and curation

Current "gold standard" for accuracy and retrieval performance.

Attractive approach given free availability of raw material

Automatic analysis of full text patent

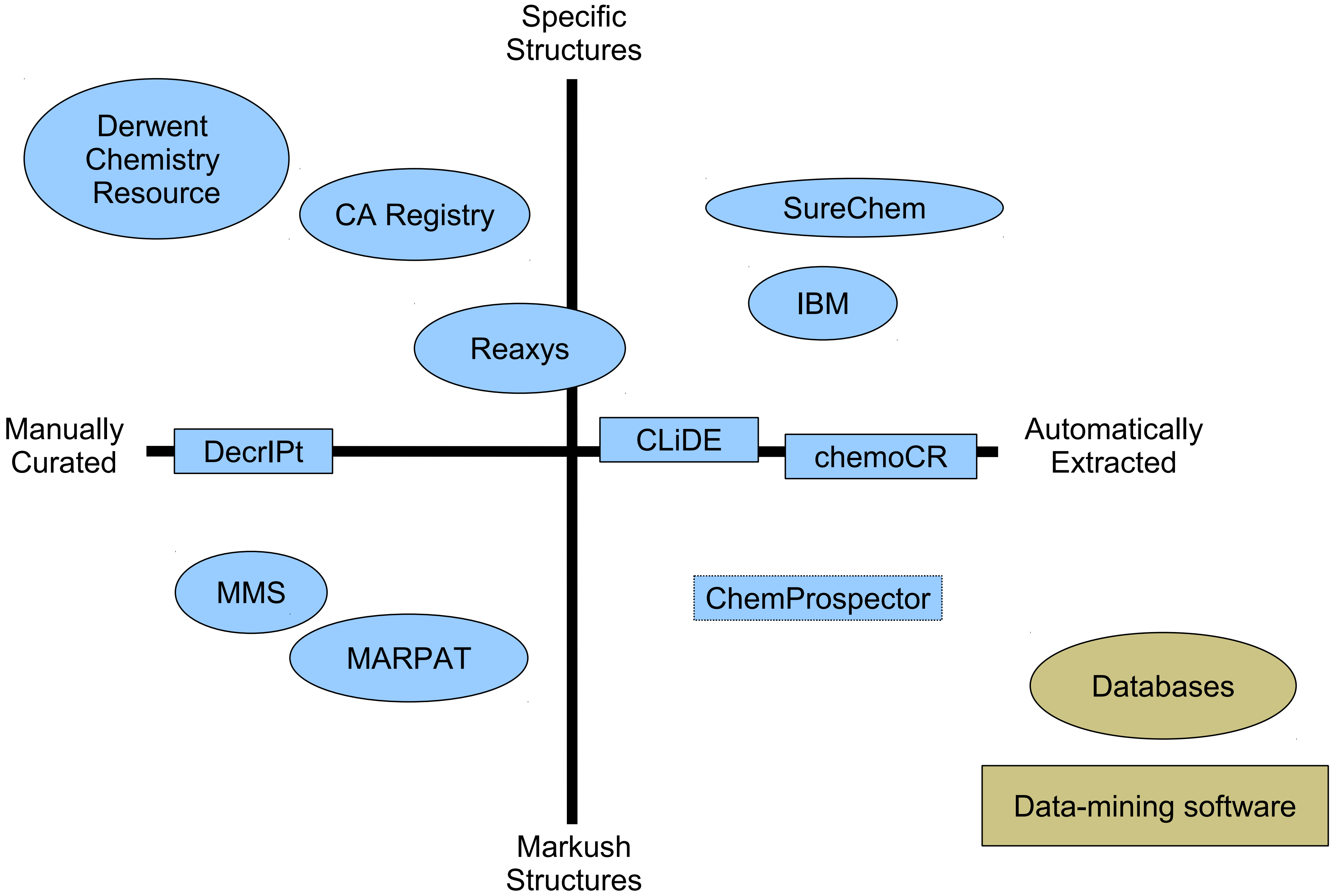
Active area of research and development

Cover the scope and claims of the patent more comprehensively

Markush structure

Difficult to visualise and complicated to search

Databases and data mining software



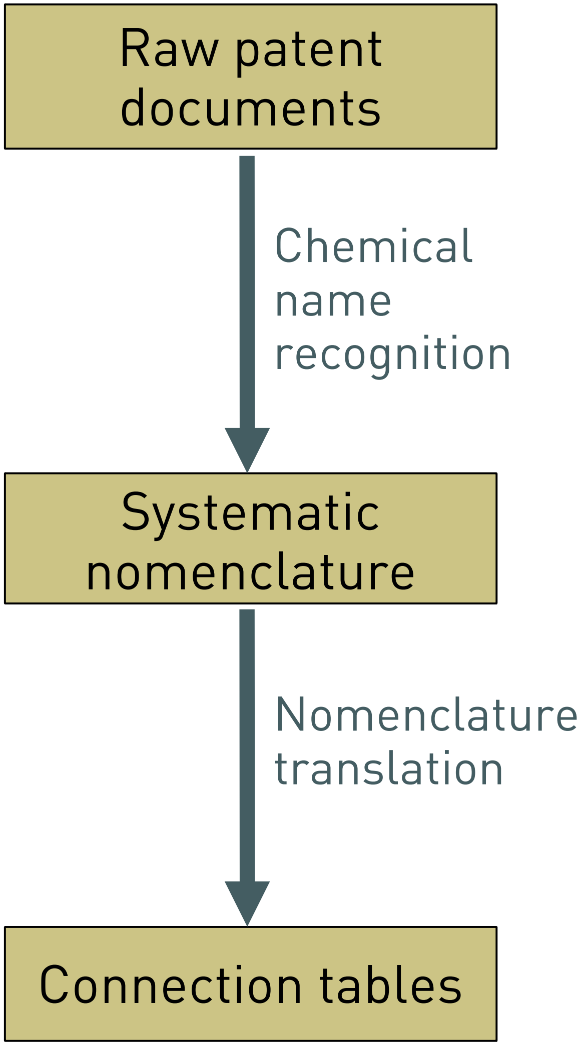
Specific structure databases

- Manually-curated databases of specific structures from patents have a long history
  - Chemical Abstracts Registry (from 1907)
  - Derwent Chemistry Resource (linked to WPI)
  - etc.
- Some newer databases use combination of automatic extraction and manual curation
  - Elsevier Reaxys database incorporates former MDL Patent Chemistry Database (patents from 1976)
- Can be searched by conventional full structure and substructure search systems

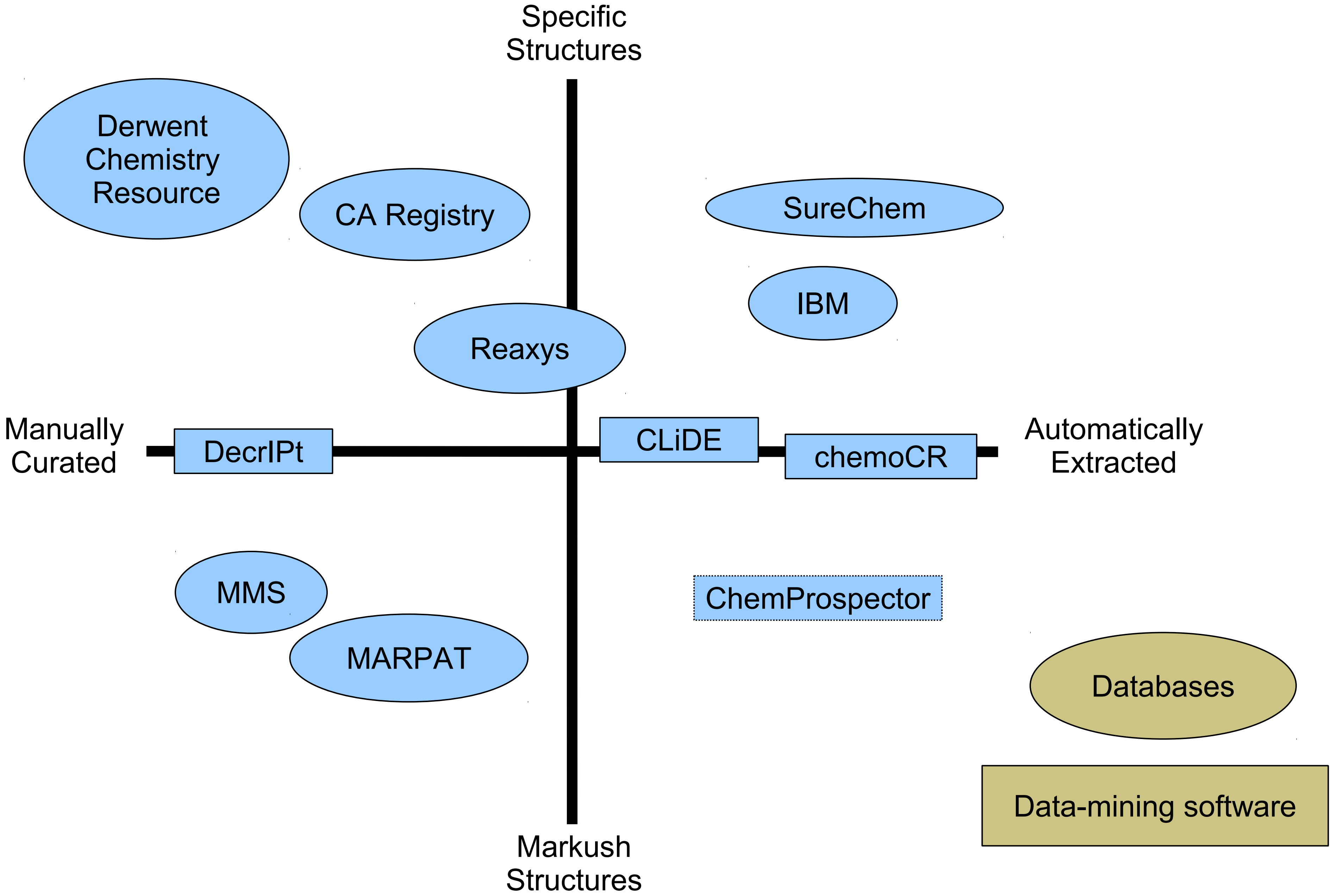
Mining patents for specific structures

digital chemistry

- Free availability of full-text patent documents since 1990s has encouraged data mining to extract specific structures
- Commercial and open-source software used for both steps
  - use of multiple NT programs can improve quality
- Recent work in both academic and commercial environments
  - Cambridge University (OSCAR, OPSIN)
  - SURECHEM database (Macmillan/Digital Science)
  - IBM Patent database



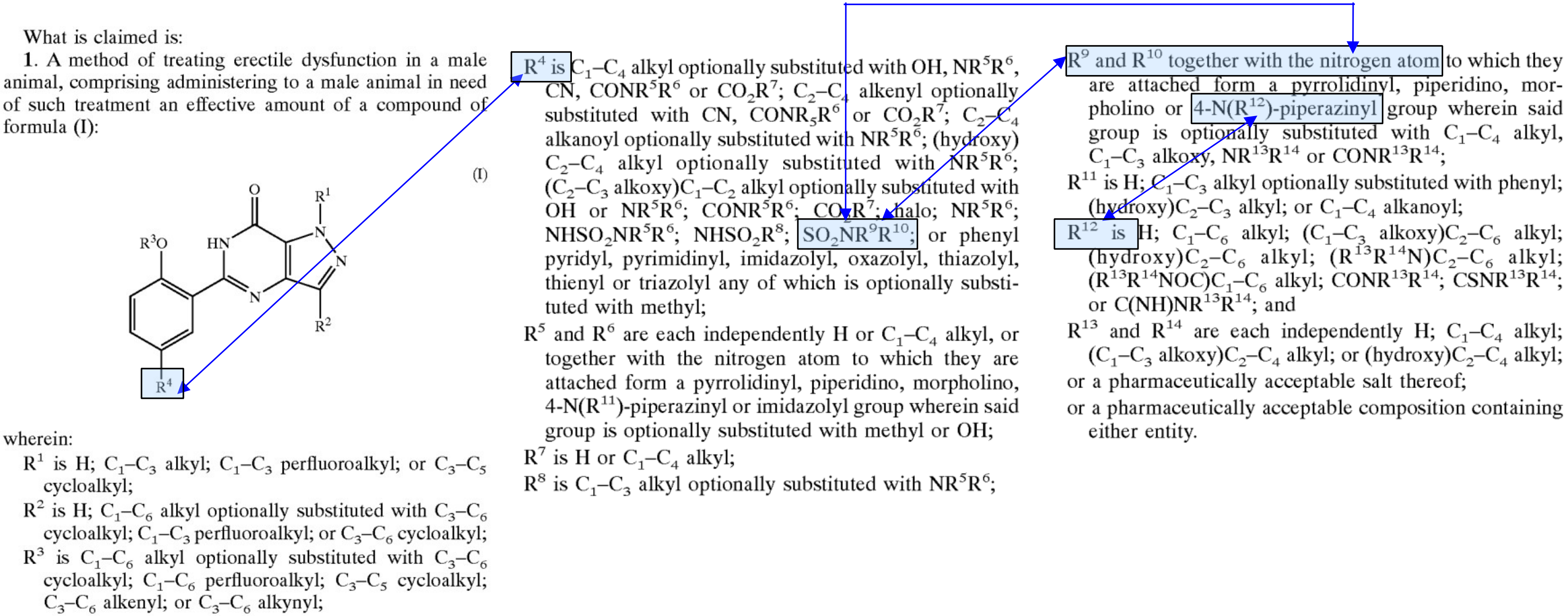
Databases and data mining software



Mining patents for Markush structures

digital chemistry

- much more complicated than for specific structures
- requires analysis of both structure diagrams and text, and of semantic relationships between them



What is claimed is:

1. A method of treating erectile dysfunction in a male animal, comprising administering to a male animal in need of such treatment an effective amount of a compound of formula (I):

(I)

O
R^1
R^3O
HN
N
N
N
R^2
R^4

wherein:

R^1 is H; C_1–C_3 alkyl; C_1–C_3 perfluoroalkyl; or C_3–C_5 cycloalkyl;

R^2 is H; C_1–C_6 alkyl optionally substituted with C_3–C_6 cycloalkyl; C_1–C_3 perfluoroalkyl; or C_3–C_6 cycloalkyl;

R^3 is C_1–C_6 alkyl optionally substituted with C_3–C_6 cycloalkyl; C_1–C_6 perfluoroalkyl; C_3–C_5 cycloalkyl; C_3–C_6 alkenyl; or C_3–C_6 alkynyl;

R^4 is C_1–C_4 alkyl optionally substituted with OH, NR^5R^6, CN, $CONR^5R^6$ or CO_2R^7; C_2–C_4 alkenyl optionally substituted with CN, $CONR_5R^6$ or CO_2R^7; C_2–C_4 alkanoyl optionally substituted with NR^5R^6; (hydroxy)C_2–C_4 alkyl optionally substituted with NR^5R^6; (C_2–C_3 alkoxy)C_1–C_2 alkyl optionally substituted with OH or NR^5R^6; $CONR^5R^6$; CO_2R^7; halo; NR^5R^6; $NHSO_2NR^5R^6$; $NHSO_2R^8$; $SO_2NR^9R^{10}$; or phenyl pyridyl, pyrimidinyl, imidazolyl, oxazolyl, thiazolyl, thienyl or triazolyl any of which is optionally substituted with methyl;

R^5 and R^6 are each independently H or C_1–C_4 alkyl, or together with the nitrogen atom to which they are attached form a pyrrolidinyl, piperidino, morpholino, 4-N(R^{11})-piperazinyl or imidazolyl group wherein said group is optionally substituted with methyl or OH;

R^7 is H or C_1–C_4 alkyl;

R^8 is C_1–C_3 alkyl optionally substituted with NR^5R^6;

R^9 and R^{10} together with the nitrogen atom to which they are attached form a pyrrolidinyl, piperidino, morpholino or 4-N(R^{12})-piperazinyl group wherein said group is optionally substituted with C_1–C_4 alkyl, C_1–C_3 alkoxy, $NR^{13}R^{14}$ or $CONR^{13}R^{14}$;

R^{11} is H; C_1–C_3 alkyl optionally substituted with phenyl; (hydroxy)C_2–C_3 alkyl; or C_1–C_4 alkanoyl;

R^{12} is H; C_1–C_6 alkyl; (C_1–C_3 alkoxy)C_2–C_6 alkyl; (hydroxy)C_2–C_6 alkyl; ($R^{13}R^{14}N$)C_2–C_6 alkyl; ($R^{13}R^{14}NOC$)C_1–C_6 alkyl; $CONR^{13}R^{14}$; $CSNR^{13}R^{14}$; or $C(NH)NR^{13}R^{14}$; and

R^{13} and R^{14} are each independently H; C_1–C_4 alkyl; (C_1–C_3 alkoxy)C_2–C_4 alkyl; or (hydroxy)C_2–C_4 alkyl;

or a pharmaceutically acceptable salt thereof;

or a pharmaceutically acceptable composition containing either entity.

Mining patents for Markush structures

- Seminal work at Sheffield University (mid-1990s)
  - based on analysis of Derwent Abstracts
- Now a very active area of research and development

CLiDE Pro

Leeds University, KeyModule Ltd.

Extension of original chemical OCR software, based on identifying and separating graphical regions from text

Fraunhofer SCAI

Work based on **ChemoCR** program for analysis of documents with images, extended to reconstruct Markush structures from patents, using an extended SMILES notation. Limited success in initial results.

ChemProspector

InfoChem GmbH (under THESEUS program, funded by German government)

Involved development of image to structure converter, annotator to extract text data, and semantic parser. Markush structures extracted to a proprietary format with some success, though many examples failed to fully analyse the Markush.

Prospects for automatic analysis

- Can automatically-extracted databases supplant manually curated ones?
  - manually-curated databases remain “gold standard”
  - automatically extracted ones are gaining ground, especially for specific structures
    - nomenclature translation software has improved
    - source nomenclature quality remains an issue
- Markush structure extraction still has a long way to go
  - issues of generic nomenclature translation
  - likely to remain a need for manual intervention to resolve ambiguities
    - but automation could do much of the “donkey work”
    - might also assist with transcription and other errors that creep in during manual curation

Searching chemical patents

- Specific structure databases can be searched by conventional (sub)structure and similarity
  - limited to those specific structures extracted from patent
- Main Markush search systems now showing their age
  - only available online
  - clunky interfaces
  - difficult to visualise complete structures
- New systems in the pipeline
  - improved visualisation
  - in-house deployment
  - new search algorithms

- Thomson Reuters making MMS database available for in-house use
- Chemical Abstracts Service retaining control over MARPAT database
- Potential for new (automatically-extracted?) databases

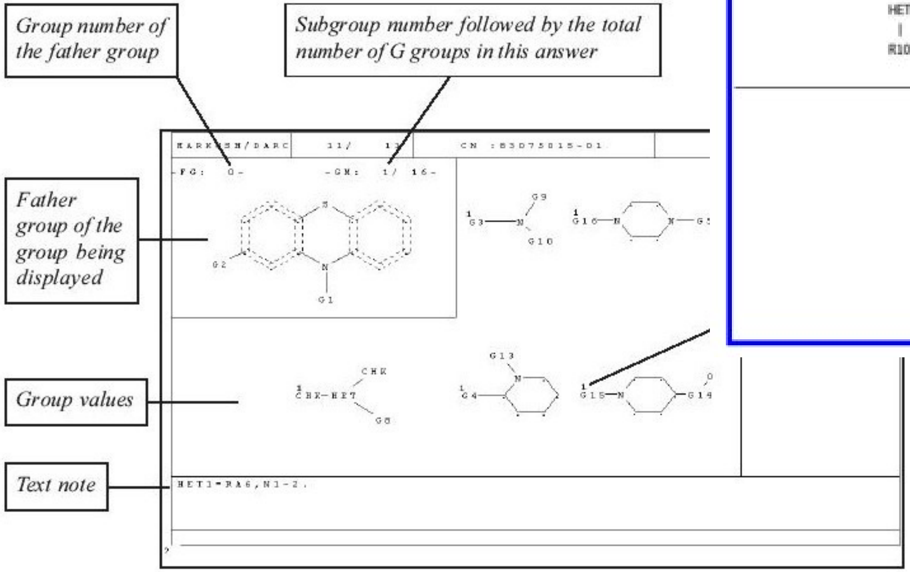
In-house Markush searching

- New commercial software for substructure search of Markush databases
  - ChemAxon, Digital Chemistry
  - able to handle homology variation ("R1= alkyl, heteroaryl" etc.)
  - intended to enable Thomson MMS database to be searched in-house, rather than online via Questel
- Advantages
  - improved visualisation of Markush
  - partial enumeration of specific molecules covered
  - end-user chemist access to patent databases
  - integration with drug discovery informatics systems

} rather than separate patent search department

Roche Marvis System

Interactive desktop application for visualization of MMS structures, developed at Roche using Pipeline Pilot



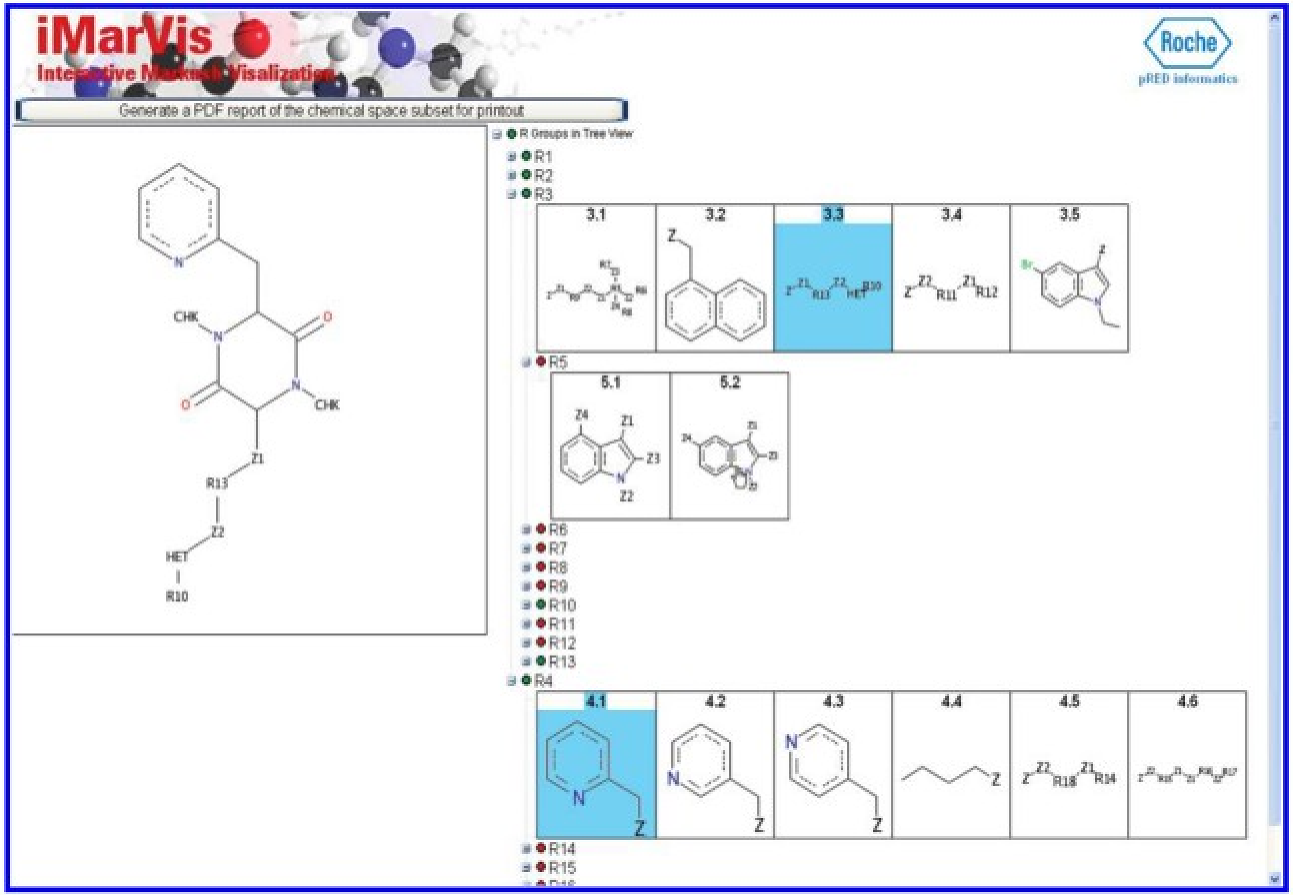
Group number of the father group

Subgroup number followed by the total number of G groups in this answer

Father group of the group being displayed

Group values

Text note



Deng *et al.*, *J. Chem, Inf. Model.* **2011**, *51*, 511

Deng *et al.*, *World Patent Inf.*, **2012**, *34*, 128

Astra Zeneca Periscope system

- Markush patent system recently developed for in-house use (Cosgrove *et al.*, *J. Chem. Inf. Model.* **2012**, *52*, 1936)
- Three main components
  - XML-based Markush Input Language (MIL)
  - graphical input program (MENGUIN)
  - "is it in a markush?" (i3am) search program
- Applications
  - Free-Wilson structure-activity analysis based on Markush structure and specific examples with activity data
  - monitoring controlled substances – legislation often uses Markush-like designations
  - searching virtual libraries – library represented as Markush, against which sets of specific molecules can be matched

Astra Zeneca Periscope system

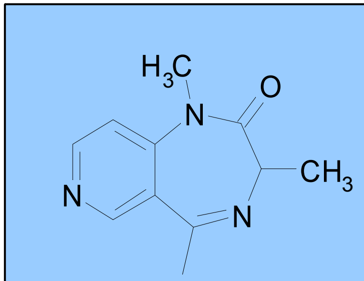
- Markush Input Language
  - “exact” R-groups (SMARTS) and “inexact” (element, bond, ring counts etc.)
  - attachment information etc.
  - input using MENGUIN Rich Internet Application in browser, with JDraw applet and Java back-end
  - “typical Markush can be encoded in a few hours”
  - authors advocate use for “encoding open searchable archive of Markush structures from patents”
- i3am search application
  - implemented using OpenEye OEChem toolkit
  - determines whether a specific-molecule query is covered by a Markush
  - not a comprehensive Markush search system

Astra Zeneca Chemistry Connect

- Large-scale integrated application to facilitate data mining
  - Muresan *et al.*, *Drug Discovery Today* **2011**, *16*, 1019
  - Tyrchan *et. al.*, *J. Chem. Inf. Model.* **2012**, *52*, 1480
- Integrates SAR data from literature, patents and other public sources
  - includes IBM patent data (specific molecules)
- Large number of small applications for analysis, including patent “key compound” prediction
  - specific molecules used to generate Markush by R-group decomposition based on maximum common substructures
  - “key compounds” lie at intersection of highly-populated R-groups
  - authors suggest that better results could be obtained by using original Markush core for identification of R-groups

Alternative approaches to search

- Structure search of Markush databases presents several challenges
  - lack of suitable available databases
  - absence of standard exchange formats
  - complexities of matching specific structures against homology-variant groups



vs.

R84 is a substituted or unsubstituted, mono-, di- or polycyclic, aromatic or non-aromatic, carbocyclic or heterocyclic ring system, or …

- Several more or less “wacky” attempts made to “finesse” the search, usually using some sort of similarity search, but none has achieved conspicuous success

Practicalities of searching

- Established systems generally regarded as insufficient to cope with the complexity of modern patents
  - in many cases high recall is required
  - searchers have to put up with poor precision

Recall and precision are not usually relevant to chemical structure search

- deterministic isomorphism algorithms give 100% recall and 100% precision

Situation with Markush structures is more ambiguous

- some parts of Markush may be more important than others ("what the patent teaches") meaning that hits may have degrees of relevance
- ranked output might help bring most relevant hits to the top of long lists

Retrieval system evaluation

- Systematic evaluation of performance of search systems may be useful
- TREC-CHEM track started in 2010 under auspices of long-running Text-REtrieval Conferences
  - multi-year experiment using standard set of patents and queries with relevance judgements
  - not run in 2012, following demise of Information Retrieval Facility (IRF) which provided much support
  - commercial databases not included
  - most participants are academic groups using e.g. nomenclature identification and translation software
- Evaluation of retrieval systems with graded relevance judgements is in its infancy

Markush structure representation

- Established formats are proprietary, complex, limited and/or unfriendly
- AstraZeneca's XML-based MIL has some possibilities, though is rather tied to their Periscope system
- InChI working party has looked at extending InChI standard and software to handle generic structures
  - based on canonicalising individual R-group values, with assembly into Markush structure
  - InChI Trust has approved proposals from Digital Chemistry Ltd for staged implementation
  - awaiting allocation of funding

Conclusions and Prospects

- Data mining of specific structures from patents increasingly proving useful
- Automatic mining of Markush descriptions still in its infancy
  - likely to be used ultimately in semi-automatic curation systems
- New generation of Markush search software appearing, using existing curated databases
  - may be deployed in-house and online, with improved visualisation
  - may feature relevance ranking of hits
- In-house applications being developed by individual companies to support analysis of structural information in patents
  - integration of Markush and specific compound data likely to be important

Contact details

Dr John M. Barnard

Scientific Director, Digital Chemistry Ltd.
46 Uppergate Road, Sheffield S6 6BX, UK

john.barnard@digitalchemistry.co.uk
+44 (0)114 233 3170

G. M. Downs and J. M. Barnard, "Chemical patent information systems," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, **2011**, *1* (5), 727-741 (DOI: 10.1002/wcms.41)