Mining chemical structural information from the drug literature

Debra L. Banville

AstraZeneca Pharmaceuticals, 1800 Concord Pike, Wilmington, DE 19850, USA

It is easier to find too many documents on a life science topic than to find the right information inside these documents. With the application of text data mining to biological documents, it is no surprise that researchers are starting to look at applications that mine out chemical information. The mining of chemical entities – names and structures – brings with it some unique challenges, which commercial and academic efforts are beginning to address. Ultimately, life science text data mining applications need to focus on the marriage of biological and chemical information.

Imagine a pharmaceutical researcher at the start of a new project involving a G-protein-coupled receptor (GPCR), who needs to find compounds that target that receptor. Before 1988, a researcher new to the GPCR area would have nine journal articles to read based on a cumulative search across Chemical Abstracts Plus, Embase and Medline with the removal of duplicates. By 1992, there would be a manageable 256 articles (and ~34 patents) to read. But by 2005 – and over 14,000 publications later – the researcher would decide that knowing everything about GPCRs is not a practical goal. The researcher would settle for 'some' information on the subject, focus on their particular target of interest, and hope to fill in any gaps in knowledge as time goes on.

This illustrates the paradigm shift (from one of information gathering to one of information mining) that the scientific community is facing [1]. Scientists can easily find an overload of information on many areas of research (information gathering); the current problem is how the researcher can find the important information (information mining) without unknowingly filtering out something essential. How can a researcher leave room for discovery while infusing their knowledge and experience into the process of a literature research? Finally, how can this information be managed over time?

Early demonstration of text data mining can be found in the medical literature of the mid-1980s [2,3]. However, not until the last six years has the focus shifted to the development of commercial tools, for example, ClearForest, OmniViz, TEMIS and Linguamatics, to name a few [4–6] (see press releases from Clear Forest, 4 August 2005, and Omniviz, 10 January 2004). Life science applications for these technologies primarily focus on relating biological information (e.g. genes, proteins and pathway analysis) to disease area [7–13] (Kankar, P. *et al*. MedMeSH Summarizer: text mining for gene cluster. *Proc. SIAM Conf. In Data Mining*, SDM, April 2002, Arlington, VA, USA pp. 548–565). In the pharmaceutical field, it is ideally the marriage of biological and chemical information that needs to be the ultimate focus of text data mining applications. Early mentions of chemical information 'mining' started around 1988, following interest in improving searching for chemical reactions [14]. However, the actual concept of indexing chemical reactions for searching is substantially older (see 'Chemical name recognition and extraction' section below). Interest in analytical chemical information extraction [15] and optical chemical structure recognition tools soon followed [16–18]. Since 1998, the nascent awareness of mining chemically related text has started to evolve to the more recent need to incorporate chemical structure mining of the literature [19–20] (Hehenberger, M. Text mining of chemical and patent literature *Proc. Internatl. Chem. Inform. Conf*. 25–28 October 1999, Annecy, France, pp. 83–94). The reasons for mining chemical entities from the literature include:

- The lack of a universal publication standards for identifying each unique chemical entity.

Corresponding author: Banville, D.L. (Debra.Banville@AstraZeneca.com)

1359-6446/06/$ – see front matter ©2006 Elsevier Ltd. All rights reserved. PII: S1359-6446(05)03682-2

TABLE 1

Multiple methods for expressing chemical structural information in the literature

Method	Descriptions and examples[a]
1) Systematic chemical names	IUPAC, IUPAC-like, non-IUPAC.
	Sulfuric acid is also known as hydrogen sulfate and ferrosulfate; sulfate can also be spelled as sulphate.
	Special symbols and fonts; commas, periods, hyphens, parentheses, apostrophes, plusses, minuses and Greek symbols.
2) Common or generic names	Aspirin, camphor, water and alcohol.
3) Trade Names	Seroquel = quietapine.
4) Company codes	ZD5077 = ICI204636 = ZM204636.
5) Abbreviations	DMS for dimethyl sulfate
6) Index and reference numbers	From CAS registry numbers, EINECS, Beilstein registry numbers, etc.
7) Anaphors	Compounds are named earlier in the text but co-referenced to a shorter name, the anaphor, later in the text. For example, a compound number is an anaphor where '…bioactivity is found in compounds [1–10] listed in Table 5…'.
8) Generic and fragmented descriptors	Compounds are inferred by generalized names or descriptors. For example, 'Preparation of 2-Aziridinemethanols… the corresponding 3-phenyl-substituted derivative…'; 'R = CH3, CH2OH, Ph' or molecular formula.
9) Chemical Structures	Explicit and implicit structures e.g. Markush structures, where R1 = CH3, COOH, etc…

[a] Abbreviations: CAS, Chemical Abstracts Service; IUPAC, International Union of Pure and Applied Chemistry.

- Selective (versus comprehensive) indexing policies of bibliographic chemical databases [21,22] that prevent researchers from relying solely on these sources for definitive chemical entity retrieval.
- The need to understand how chemical structures link to biological processes.

Areas of research that would benefit from structure–biological links include drug safety and drug discovery [23–25]. The information extraction processes to make these links are fairly manual, requiring many hours of reading to extract minor but essential details embedded within numerous full text documents. For example, detailed information on compounds not found to have an adverse effect or potential biomarker information linked to adverse or efficacious events would highly benefit drug safety research.

Overall, pharmaceutical companies are driven by the need to reduce the failure rate of compounds that go into the clinic, thereby reducing costs. This is considered crucial for the future growth of this sector [26,27]. In general, most researchers currently take a minimalistic approach to facilitate their decision-making. There is a heavy reliance on highly filtered searches, bibliographic database indexing, citation searching and random discovery to select key articles, patents, meeting abstracts and reports. Overall, the ability and experience of researchers, good information-science support and serendipity are involved in assembling this information. Furthermore, the ability to manage the information over time is very difficult, with a high likelihood of missing or losing track of important structural–biological links.

Although there is a limited number of automated chemical data mining tools available for addressing this growing problem, there is a clear sense in the scientific community that the development of these technologies is going to grow exponentially over the next five years.

The problems

No true standard for representing chemical structural information

Chemical structural information is difficult to mine from the literature, largely due to the diverse ways in which the information embedded in documents is summarized (Table 1).

Methods 1–6 in Table 1 generally involve the ways in which compounds are named. Even if the literature adhered to one standard for naming compounds, such as the Union of Pure and Applied Chemistry (IUPAC) standard, there would still be problems caused by large variations in how IUPAC standards can be applied. In other cases, the ability to associate name with structure might be proprietary, for example, a company trade name for a compound might be in an article without a corresponding structure.

Methods 7–8 in Table 1 describe the ways in which compounds can be inferred as a class of compounds but not explicitly named (e.g. benzodiazapines). Method 9 describes the use of chemical structures – rather than textual naming – to represent chemical compounds (both explicitly and implicitly as Markush structures).

Ultimately, it would be easier if all authors had to explicitly name compounds and annotate their documents with one structural standard, such as chemical coordinate tables (e.g. Chime [28]) or simplified molecular input line entry system strings (Daylight SMILEs [29]). In the absence of one standard, true chemical structural mining of the literature has to apply a multifaceted approach to identify the diverse representations of chemical information within the literature.

Documents can be found in textual and/or image formats

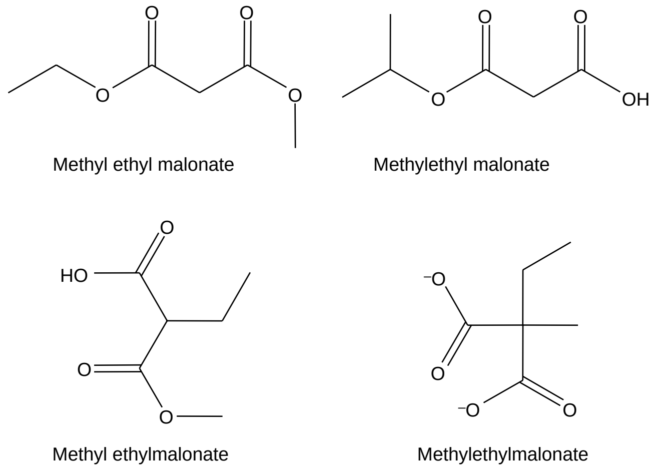
Different document formats further complicate the problems of mining chemical information from the literature. Although documents exist in two basic format types of 'text' or 'image', each type has many variations (e.g. text, rich text format, word document, hypertext markup language, portable document format, or graphical interface file and tagged image file formats). Journal articles are generally found in a text format embedded with images corresponding to figures and/or tables. Chemical entity recognition will require the ability to extract information form the text and images in all multiple variations

To maintain the integrity of intellectual property, patent documents are usually available as images of text documents (e.g. .tiff or .pdf). Most full text patent documents are actually optical character recognition (OCR) documents converted from image to text with little or no error checking [30,31] [see also full text information

TABLE 2

Chemical entity recognition applications

Application	**Methods covered (from Table 1)**	**Document type**
Name identification		
a) CambridgeSoft, ChemFinder/Text™ [41]	1	Text
b) ReelTwo, SureChem™ (http://surechem.reeltwo.com)	1–2, 5, 8	Text and web service for Medline
c) IBM, Chemical Name Spotter and UIMA [42]	1–6	Data warehouse including full-text US patent documents, 25 years, >8 million structures
d) MDL Beilstein, Chemical Reader (product to be released in late 2005 as the CER Skill Cartridge™ with MDL and TEMIS) [44]	1–8	Text
e) Merck, TIMI [45]	2–3	Text
f) SCAI Fraunhofer, ProMiner™ (www.scai.fhg.de)	2	Text
Structure identification		
f) SimBioSys, CLiDE™ [19,20,48] (www.simbiosys.ca)	9	Text with images
g) SCAI Fraunhofer Institute, CSR™ [47] (www.scai.fhg.de) search on 'CSR' for updates	9	Text with images



Drug Discovery Today

FIGURE 1

Chemical name recognition can be a challenging problem. In this simple example, the placement or misplacement of spaces between methyl, ethyl and malonate can result in four different chemical structures that correspond to: methyl ethyl malonate, methylethyl malonate, methyl ethylmalonate and methylethylmalonate.

at the European Patent Office (http://ep.espacenet.com) and the Micropatent web page, a fee-for-service provider (www.micropat.com)]. OCR documents are frequently riddled with translation errors and the absence of all graphical images, including chemical structures. For example, in one patent 'methylpropyl-amine' was recognized as 'methylpropvll-amine', and 'EXAMPLE 22. Amino-3,4' as 'EXAMPLE 22-Amino-3, 4'. Hence, comprehensive chemical name and structure recognition applications would ideally handle images and text with extensive error recognition. A recent application of this can be seen at the United States Patent and Trade Office (USPTO, www.uspto.gov/patft/index.html) [32].

Publishers' restrictions on full text data mining

Accessibility to full text documents is necessary to perform extensive full text mining of the literature. With the advent of full-text data mining and the demand for open access to the literature, there is an increased concern by electronic journal publishers that control of their one asset – the literature to which they sell access – will be lost [33,34]. As researchers find more ways to exploit tools for accessing and managing their information needs, publishers are clearly starting to explore new business models for delivering literature (see also a series of articles highlighting different facets of this debate on the Nature website, www.nature.com/nature/focus/accessdebate/index.html).

Partial solutions

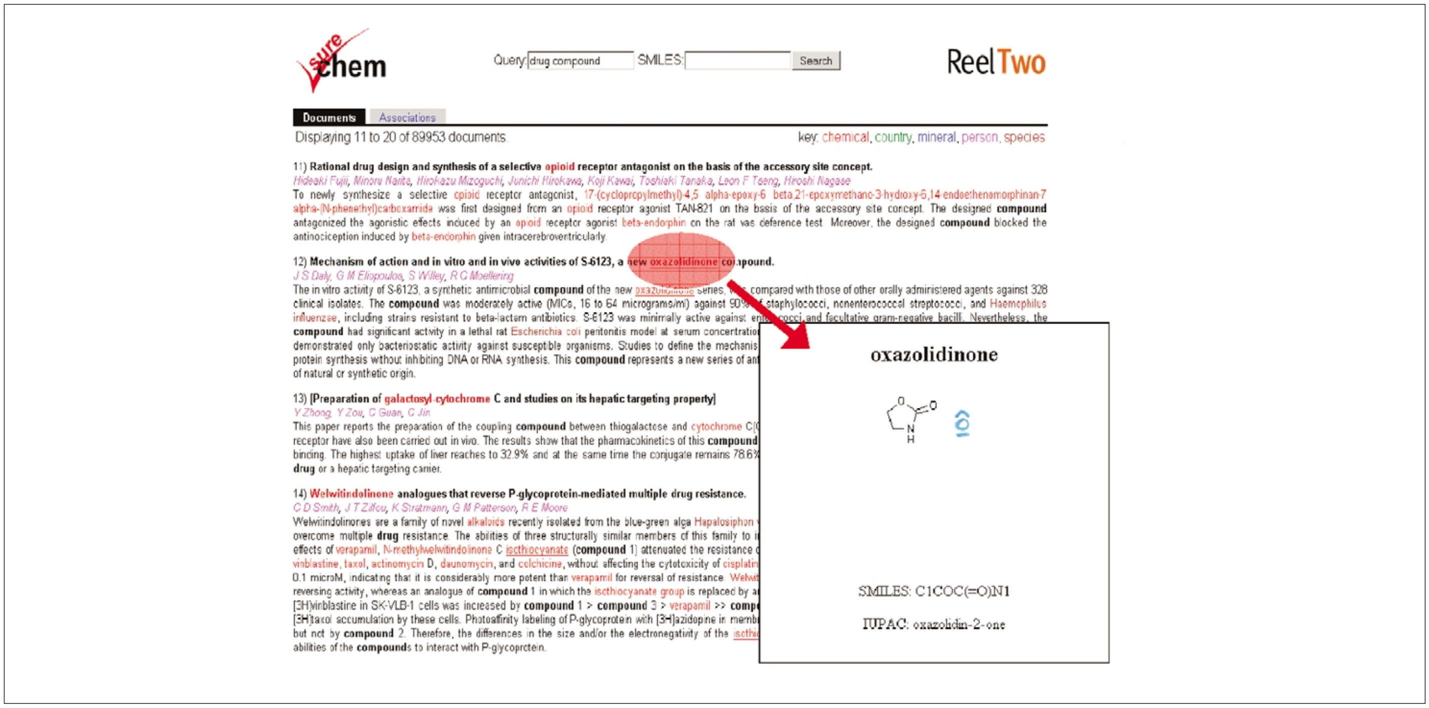
One solution would involve a chemical-reading capability that includes:

- Recognition of each chemical entity;
- Extraction of each chemical entity;
- Conversion of the chemical entity into a structure-searchable format saved in a structure database;
- Annotation of the text with a link to the structure in a searchable database.

In this scenario, structure-searching in the database will result in retrieval of documents containing the chemical entity and its exact location within the text being highlighted.

To explore issues of chemical structure mining, vendors, academics and pharmaceutical companies have addressed a variety of partial solutions for chemical name recognition that collectively cover the first six methods of explicit naming (Table 1, Table 2) [35] (N. Goncharoff and H. Grotz, personal communications; Chemical Text Data Mining panel discussion, MDL User Conference, May 2004, Boston, MA, USA). Each depends on the ability of the software to recognize and convert the chemical entity to a structure.

When a compound has been recognized, it needs to be tagged or extracted for conversion to a structure-searchable form. Software tools that perform chemical name recognition are usually rule-based so that patterns are identified that detect likely chemical names within the text [36]. Identification of a chemical name within the text is a complicated process not only because of the huge variation in naming methods (Table 1), but also the variations in expressing chemical names within each method. For example, a systematic name can include multiple variations in how hyphens



Query: drug compound SMILES:

Displaying 11 to 20 of 89953 documents

key: chemical, country, mineral, person, species

11) **Rational drug design and synthesis of a selective opioid receptor antagonist on the basis of the accessory site concept.**
Hideaki Fujii, Minoru Narita, Hirokazu Mizoguchi, Junichi Hirokawa, Koji Kawai, Toshiaki Tanaka, Leon F Tseng, Hiroshi Nagase
To newly synthesize a selective opioid receptor antagonist, 17-(cyclopropylmethyl)-4,5 alpha-epoxy-6 beta,21-epoxymethano-3-hydroxy-6,14-endoethenomorphinan-7 alpha-(N-phenethyl)carboxamide was first designed from an opioid receptor agonist TAN-821 on the basis of the accessory site concept. The designed **compound** antagonized the agoristic effects induced by an opioid receptor agorist beta-endorphin on the rat vas deference test. Moreover, the designed **compound** blocked the antinociception induced by beta-endorphin given intracerebroventricularly.

12) **Mechanism of action and in vitro and in vivo activities of S-6123, a new oxazolidinone compound.**
J S Daly, G M Eliopoulos, S Willey, R C Moellering
The in vitro activity of S-6123, a synthetic antimicrobial **compound** of the new oxazolidinone series, w compared with those of other orally administered agents against 328 clinical isolates. The **compound** was moderately active (MICs, 16 to 64 micrograms/ml) against 90% f staphylococci, nonenterococcal streptococci, and Haemophilus influenzae, including strains resistant to beta-lactam antibiotics. S-6123 was minimally active against ent cocci and facultative gram-negative bacilli. Nevertheless, the **compound** had significant activity in a lethal rat Escherichia coli peritonitis model at serum concentratio demonstrated only bacteriostatic activity against susceptible organisms. Studies to define the mechanis protein synthesis without inhibiting DNA or RNA synthesis. This **compound** represents a new series of ant of natural or synthetic origin.

13) **[Preparation of galactosyl-cytochrome C and studies on its hepatic targeting property]**
Y Zhong, Y Zou, C Guan, C Jin
This paper reports the preparation of the coupling **compound** between thiogalactose and cytochrome C[receptor have also been carried out in vivo. The results show that the pharmacokinetics of this **compound** binding. The highest uptake of liver reaches to 32.9% and at the same time the conjugate remains 78.6% **drug** or a hepatic targeting carier.

14) **Welwitindolinone analogues that reverse P-glycoprotein-mediated multiple drug resistance.**
C D Smith, J T Zilfou, K Stratmann, G M Patterson, R E Moore
Welwitindolinones are a family of novel alkaloids recently isolated from the blue-green alga Hapalosiphon overcome multiple **drug** resistance. The abilities of three structurally similar members of this family to i effects of verapamil, N-methylwelwitindolinone C iscthiocyanate (**compound** 1) attenuated the resistance vinblastine, taxol, actinomycin D, daunomycin, and colchicine, without affecting the cytotoxicity of cisplatin 0.1 microM, indicating that it is considerably more potent than verapamil for reversal of resistance. Welwit reversing activity, whereas an analogue of **compound** 1 in which the iscthiocyanate group is replaced by a [3H]vinblastine in SK-VLB-1 cells was increased by **compound** 1 > **compound** 3 > verapamil >> **comp** [3H]taxol accumulation by these cells. Photoaffinity labeling of P-glycoprotein with [3H]azidopine in memb but nct by **compound** 2. Therefore, the differences in the size and/or the electronegativity of the iscthi abilities of the **compound**s to interact with P-glycoprctein.

oxazolidinone

SMILES: C1COC(=O)N1

IUPAC: oxazolidin-2-one

FIGURE 2

The chemical name recognition capabilities available in the SureChem™ application. A query on 'drug compounds' against a Medline, or other unstructured text, returns all citations containing chemical entities with each chemical name highlighted in *red*, and ready for manual inspection. Structure queries can also be run using SMILEs strings, and associations between chemical compounds, and other entities (e.g. country, mineral, person or species) can be made visually where each entity is marked up in the text with color coded highlighting. Reproduced with permission from Nicko Goncharoff of ReelTwo.

and dashes (e.g. 1,1- versus 11- versus 1–1-) are used [37]. In another example (Figure 1), the placement of spaces on three components of a name: methyl, ethyl, and malonate, results in four different structures.

Chemical name recognition and extraction

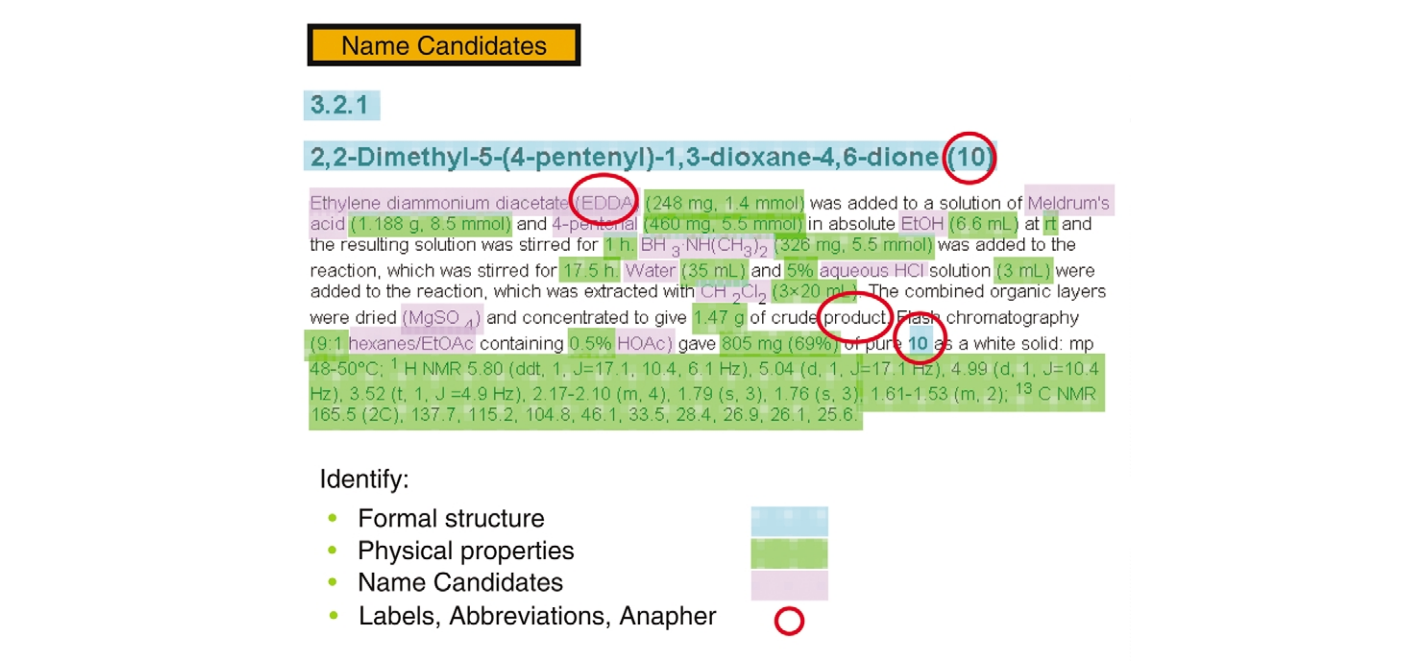
Chemical name recognition has evolved in the past 48 years. In 1958, Eugene Garfield [38,39] first recognized that a systematic chemical name could be algorithmically converted into a molecular formula and then to line notation. This work quickly led to formation of the Index Chemicus in 1959, followed by a variety of name-to-structure (e.g. the connection table) algorithms from other groups [40–47]. All of these methods focused on the conversion of chemical names previously extracted from text.

By 1989, 31 years after Garfield's work, Hodge *et al.* [48,49] described methods for extracting chemical names from text fields including bibliographic titles and supplementary terms to assign Chemical Abstract Service (CAS) registry numbers. The application of these types of methods to full text using natural language processing (NLP), lexical, syntactic and semantic methods was initially addressed in the mid-1980s by Zamora and colleagues [50–53] to build chemical reaction databases. In the early 1990s, Chowdhury and Lynch [54,55] worked on semi-automated processes for mining abstracts. They applied NLP techniques to extract generic structure descriptions from patent abstract text provided by documentation abstracts. These abstracts are rewritten for brevity and conform to well-developed conventions. Owing to the highly consistent and rule-based structure of these abstracts, an automated extraction application was successfully devised using these rigorous conventions. Mining of highly structured textual data (sometimes called template mining) eliminates some of the difficulties of text-mining free or unstructured text.

In the absence of highly structured text, the important driving force for free text mining of entire documents (like patents) is the recognition that indexing terms, titles, keywords or abstracts do not always contain the necessary details. In many instances, only full-text searching can provide the information required to build a chemical reaction database or to map structure–activity relationships. The challenge lies in the ability to take text written without any prescribed set of rules and develop an automated process capable of extracting meaningful information for evaluation by a knowledgeable end user.

In the late 1990s, Kemp and Lynch [56,57] described a method that uses segmentation algorithms and statistics to identify chemical names within full text patent documents. After being identified, the chemical name is tagged using standard generalized markup language (SGML) and made available for additional processing. Subsequent work at the US National Library of Medicine (NLM) by Wilbur *et al.* [58] focused on comparing three methods of chemical name recognition – a segmentation approach and two Bayesian approaches. The authors successfully demonstrated the value of each approach, and concluded that a combination approach would enhance chemical name recognition in future work.

Recent developments demonstrate the possibility of commercially available systematic chemical name recognition and extraction algorithms (Table 2) (S. Boyer, N. Goncharoff and H. Grotz, personal communications; Chemical Text Data Mining panel discussion, MDL User Conference, May 2004, Boston, MA, USA). These commercial algorithms are capable of identifying and converting various systematic chemical names to structures, and annotating the text (Figure 2).



Name Candidates

3.2.1

2,2-Dimethyl-5-(4-pentenyl)-1,3-dioxane-4,6-dione (10)

Ethylene diammonium diacetate (EDDA) (248 mg, 1.4 mmol) was added to a solution of Meldrum's acid (1.188 g, 8.5 mmol) and 4-pentenal (460 mg, 5.5 mmol) in absolute EtOH (6.6 mL) at rt and the resulting solution was stirred for 1 h. $BH_3 \cdot NH(CH_3)_2$ (326 mg, 5.5 mmol) was added to the reaction, which was stirred for 17.5 h. Water (35 mL) and 5% aqueous HCl solution (3 mL) were added to the reaction, which was extracted with CH_2Cl_2 (3×20 mL). The combined organic layers were dried ($MgSO_4$) and concentrated to give 1.47 g of crude product. Flash chromatography (9:1 hexanes/EtOAc containing 0.5% HOAc) gave 805 mg (69%) of pure 10 as a white solid: mp 48-50°C; ^{1}H NMR 5.80 (ddt, 1, J=17.1, 10.4, 6.1 Hz), 5.04 (d, 1, J=17.1 Hz), 4.99 (d, 1, J=10.4 Hz), 3.52 (t, 1, J =4.9 Hz), 2.17-2.10 (m, 4), 1.79 (s, 3), 1.76 (s, 3), 1.61-1.53 (m, 2); ^{13}C NMR 165.5 (2C), 137.7, 115.2, 104.8, 46.1, 33.5, 28.4, 26.9, 26.1, 25.6.

Identify:

- Formal structure 🟦
- Physical properties 🟩
- Name Candidates 🟪
- Labels, Abbreviations, Anapher ○

FIGURE 3

An illustration of MDL's in-house Reading Machine application. The illustration shows a similar mark up/highlighting capability as shown in Figure 2. In this application, chemical entities are categorized as being either 'formal structures' (blue), in other words, the main focus of the paragraph, or as 'named candidates' (violet). Note that the 'formal structures' designation, in this figure, has highlighted the entire section header (i.e. from the articles section subheading, 3.2.1, the chemical name and the anaphor designation for the name). The program is capable of parsing out the chemical name from the other section header entities. Physical properties (green) and labels, abbreviations and anaphors (circled in red) are also identified and highlighted thereby facilitating easier manual inspection of these documents. This example also emphasizes the need to recognize and associate anaphors to their intended compound names, i.e. both '10' and '(10)' are synonyms for the compound name (2,2-Dimethyl-5-(4-pentenyl)-1,3-dioxane-4,6-dione). Reproduced with permission from Helmut Grotz of MDL Beilstein.

Common and generic names, trade names, company code names, common abbreviations and index names require a database, or lookup functionality, to link structures to names (Table 1). Work published on text influenced molecular indexing (TIMI) describes such a system using their chemical knowledge base (CKB) to look up and associate the name to the structure [59]. [The CKB is a compilation of commercially available resources (e.g. the Merck Index).] A system demonstrated by IBM also includes lookup functionality in the name extraction. Access to various free databases (e.g. PubChem, ChemIS and Chemistry WebBook) has the potential to create a lookup feature. Indexes like the CAS registry numbers (RNs) and Beilstein RNs are not generally used in full text literature sources, and have not always been a reliable way to link to structure because various data sources frequently mistype and/or misapply these numbers. Common abbreviations are also difficult to deal with in the literature because an abbreviation in one discipline can have a different meaning in another related discipline, or might simply be redefined within the context of the document.

Nonstandard abbreviations and anaphors require a more robust text data mining approach to identify and relate these identifiers to structures accurately [36]. (Anaphors are substitutes for a preceding word or group of words, for example, where compounds are co-referenced to compound numbers, abbreviations, or pronouns.) In these cases, the context in which these identifiers are used has to be understood and ambiguities need to be removed by a text mining approach. In most cases, these approaches currently require some manual intervention to validate the links between names and structures. MDL rely on NLP tools to create the Beilstein database. NLP tools assist in the extraction and disambiguation processes (Figure 3) (Chemical Text Data Mining panel discussion, MDL User Conference, May 2004, Boston, MA, USA). First, the title compound name (in Figure 3) is systematic with the number ('10') preceding it. A mining application has to disambiguate the name and the number; the number is an anaphor or label for the compound (shown in red in Figure 3). The application has to then associate the number 10, used later in the paragraph without the brackets, as the title compound. The passage in Figure 3 is a typical example of the chemical literature where there is a mix of systematic and common names throughout.

Chemical image extraction

In addition to the challenges of extracting and interpreting chemical names, there are challenges in extracting and interpreting chemical structures. Chemical structures are usually images, regardless of the document format. Interest in the ability to recognize chemical images then convert these images to structure-searchable images dates back to the early 1990s [16,17]. Chemical literature data extraction (CLiDE™), a commercially available application, provides [17,18]: segmentation of the document [to identify sections of each page as text, tables, figures (Figure 4)]; structure-image identification [to identify chemical structures within each segment, including chemical reactions within the text, tables or figures of scanned (.bmp or .tiff) or .pdf documents]; and extraction and conversion of images into a structure-searchable form.

Once identified and extracted, the entity (text name, image name or image structure) needs to be converted to a structure, and

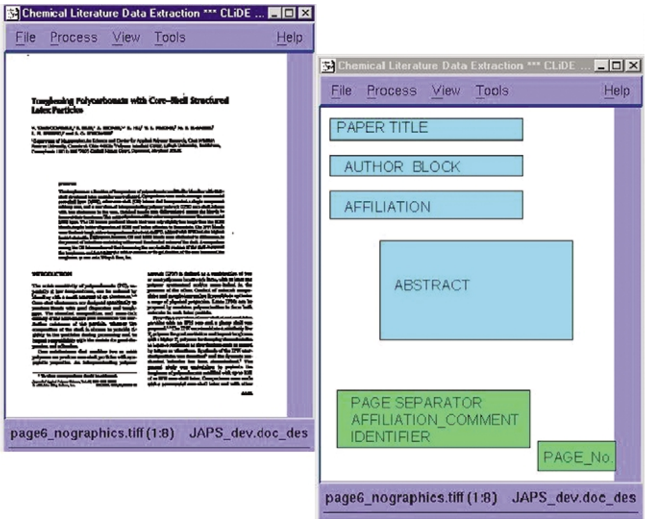

Chemical Literature Data Extraction *** CLiDE ...
File Process View Tools Help
Toughening Polycarbonate with Core–Shell Structured Latex Particles
page6_nographics.tiff (1:8) JAPS_dev.doc_des
Chemical Literature Data Extraction *** CLiDE ...
File Process View Tools Help
PAPER TITLE
AUTHOR BLOCK
AFFILIATION
ABSTRACT
PAGE SEPARATOR
AFFILIATION_COMMENT
IDENTIFIER
PAGE_No.
page6_nographics.tiff (1:8) JAPS_dev.doc_des

FIGURE 4

Chemical literature data extraction (CLiDE™). CLiDE™ is capable of segmenting documents into their various sections (text, figures, tables, etc…) and identifying the chemical structures within each segment. The ability to parse this information into segments is essential for providing a more targeted contextual search. For example, the ability to search against figure captions alone can be very powerful when attempting to target the main observations and points of the paper, points not necessarily stated in the abstract. This type of caption search would complement a typical keywords or abstract search. Conversely, a search in the 'discussion' section of an article could provide the ability to see more speculative author comments. Because these author comments are unlikely to make it into the abstract or title of a paper, there is generally no way to see these articles without manually reading every full text article on the subject. Reproduced with permission from Peter A. Johnson of SimBioSys and the University of Leeds, School of Chemistry.

the paper can then be indexed and annotated with that structure. Conversion from a structural image to a computer readable structure is a challenging problem requiring extensive error checking. This fact combined with the drive to develop more comprehensive life science text-mining capabilities has resulted in renewed interest in improving the accuracy of image to structure conversion (Personal communications with M. Hofmann from the SCAI Fraunhofer Institute, Germany, on their soon to be released Chemical Structure Reconstruction (CSR™) tool, and with P.A. Johnson from SymBioSys on CLiDE™ error checking)

Rational for business use

Business value has been demonstrated at several leading pharmaceutical companies for using chemical structure mining [57]. At AstraZeneca, for example, chemical names are recognized from a collection of key project-related patents. The compound names are converted to structures (using CambridgeSoft's Name=Struct) and imported into a relational database (using Oracle) with associated target and patent information (e.g. patent assignees, patent number and filing date). This data can be analyzed according to structural attributes and related to the associated target and patent text (using MDL ISIS/Base). This capability has provided significant value to this company.

Because a typical project at AstraZeneca can involve 50–100 key patents, covering thousands of chemical structures, having structural information in a relational database significantly reduces the information into a manageable form. This relational structural database, with links from each structure to target and patent information, provides very powerful capabilities. Researchers have the ability to map competitive landscapes using the structures for each of their target/disease areas. These landscapes can be used to identify new areas of potential research. In the absence of this capability, researchers have previously had had to manage this information in their heads. Furthermore, once this information is available electronically, it can be connected with other applications and databases, for example, data can be exported to predictive-computation applications for further analysis. Although this tool is not comprehensive in its ability to extract chemical structures from documents, it works with the previous manual methods and allows researchers more time to apply their experience to analysis of the data when making crucial decisions. This is an essential point from this work – chemical text-data mining algorithms do not have to be 100% comprehensive, but they do have to be easy to use and facilitate manual processes to show significant business value.

Overall, the creation of a patent literature database of structures has provided a more efficient and effective way to manage information and drive key decisions. Without this application, it would not have been possible to make an adequate business case and build sponsorship for the next steps around contextual information.

Contextual information: the next step

Once a chemical entity is identified in the literature and available as a searchable structure, the next step is to extract a contextual understanding. The ability to build contextual associations between chemical and biological data is achievable with text mining (Figure 3) but still requires a great deal of manual intervention and validation.

Capabilities like document segmentation are also necessary because the location of the information (e.g. figures or tables versus abstracts versus titles etc.) together with the related text provides context. For example, is the structure of interest embedded in a table with the title 'Compounds found to have adverse effects' or labeled 'Top biologically active compounds'?

Identifying the 'right' sources and source types

In the GPCR example in the introduction to this review, the researcher was increasingly faced with an information overload. In such a situation, the chemical and biological mining processes outlined here would be immensely valuable; however, some potential problems need to be tackled when deciding on the chemical literature set needed. Are the right databases being searched? Are these databases bibliographic or full text, and are they with indexing (e.g. Medline, Chemical Abstracts or Embase) or without indexing (e.g. electronic Journals)? Are full text sources needed (e.g. Scirus, HighWire Press, OVID or internal documents)? If patent information is needed, are the correct patent authorities being considered, and is full-text searching capability necessary? Some key patent authorities include the USPTO, the World Intellectual Property Organization , the European Patent Office, the UK patent Office, and the Japan Patent Office. Are the needed sets available electronically? Database selection is a very important part of a mining process and one frequently overlooked.

A typical pharmaceutical researcher would need to mine a broad set of information sources. The ability to search both bibliographic and full text databases would address drug safety issues. The ability to search key patent authorities is a requirement in most cases. The ideal chemical mining application would require a broad set of information. However, an understanding of the limitations of scope and coverage is necessary when performing any analysis and should always be considered.

Conclusion

Defining a vision

The mining of chemical entities to facilitate good decision-making by life science researchers is clearly in the early stages of development. However, some of the key components to achieving this capability are starting to fall in place:

- Increased access to electronic textual documents and 'cleaner' OCR documents (e.g. patents from the USPTO) versus raw OCR documents.
- Better methodologies for identifying and converting chemical names and/or structure images of many types into standard structures.
- Better methodologies for extracting contextual relationships, especially between chemical and biological information.
- Better methodologies for annotating and linking structures in a database to the related articles of interest.

In the future, the researcher's ideal search experience will not be to retrieve over 14,000 citations on GPCRs but a summary of these papers in the form of:

- Highlighted compounds in the text that are searchable by structure and of most interest.
- Summaries of these compounds with relevant links to data (e.g. biological effects and measures, and synthetic preparations).
- Computationally analyzed summaries of these compounds using predictive software, drug safety flags and comparisons with other known compounds.
- Clustered or categorized information based on topical areas.

The ability to organize over 14,000 citations into a form that is meaningful to the end user will increase the likelihood of the researcher finding and focusing on the most relevant literature. It will also provide the researcher with a broader contextual perspective by relating each paper back to other information.

Acknowledgements

I thank several of my colleagues at AstraZeneca who supported my work in this area and provided feedback on this article: James Rosamond, James Damewood and Bob Stumpo.

References

1 Claus, B.L. and Underwood, D.J. (2002) Discovery informatics: its evolving role in drug discovery. *Drug Discov. Today* 7, 957–966

2 Swanson, D.R. (1986) Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect. Biol. Med*. 30, 7–18

3 Swanson, D.R. (1987) Two medical literatures that are logically but not bibliographically connected. *J. Am. Soc. Inf. Sci.* 38, 228–233

4 Stensmo, M. and Thorson, M. (2003) Unstructured Information management report. *Infosphere* March report 1-110

5 Redmond, L. (2002) Mining for meaning? The joy of text. Euromap 17Dec. article

6 Kontostathis, A. *et al.* (2003) A survey of emerging trend detection in textual data mining. In *Survey of text mining: cluster classification and retrieval (*Berry, W.M., ed.), Chapter 9, Springer-Verlag

7 Mack, R. and Hehenberger, M. (2002) Text-based knowledge discovery: search and mining of life-sciences documents. *Drug Discov. Today* 7 (*11 Suppl*), S89–S98

8 Krallinger, M. *et al.* (2005) Text-mining approaches in molecular biology and biomedicine. *Drug Discov. Today* 10, 439–445

9 Tanabe, L. *et al.* (1999) MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques* 27, 1210–1214, 1216–1217

10 Rzhetsky, A. *et al.* (2000) A knowledge model for analysis and simulation of regulatory networks in bioinformatics studies aiming at disease gene discovery. *Bioinf.* 16, 1120–1128

11 Hahn, U. *et al.* (2002) Creating knowledge repositories from biomedical reports: the MEDSYNDIKATE text mining system. *Pac. Symp. Biocomput.* 7, 338–349

12 Weber, M. *et al.* (2001) Using concepts in literature-based discovery: simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. *J. Am. Soc. Inf. Sci. Tech.* 52, 548–557

13 Mack, R. *et al.* (2004) Text analytics for life science using the unstructured information management architecture. *IBM Sys. J.* 43, 490–515

14 Borkent, J.H. *et al.* (1988) Chemical searching compared in REACCS, SYNLIB and ORAC. *J. Chem. Inf. Comput. Sci.* 28, 148–150

15 Postma, G.J. *et al.* (1996) Automatic extraction of analytical chemical information. System description, inventory of tasks and problems, and preliminary results. *J. Chem. Inf. Comput. Sci.* 36, 770–785

16 McDaniel, J.R. and Balmuth J.R. (1992) Kekulé: OCR—Optical chemical (structure) recognition. *J. Chem. Inf. Comput. Sci.* 32, 373–378

17 Ibison, P. *et al.* (1993) Chemical literature data extraction: The CliDE project. *J. Chem. Inf. Comput. Sci.* 33, 338–344

18 Simon, A. and Johnson, A.P. (1997) recent advances in the CLiDE project: Logical layout analysis of chemical documents. *J. Chem. Inf. Comput. Sci.* 37, 109–116

19 Shabrang, M. *et al.* (2003) An application of text mining in chemistry? *226th ACS National Meeting,* 7-11 September 2003, New York, NY, U. S. A. (Abstract CINF-021)

20 Hauser, W.C. *et al.* (1999) Discovery of chemical relationships in patents-a combination of natural searching, with clustering and visualizations. *217th ACS national Meeting,* 21-25 March 1999, Anaheim, CA, U. S. A. (Abstract CINF-008)

21 *STNews* (2001) CA Lexicon on STN. *Chemical Abstract Services Newsletter* January-February 2001 feature article

22 Brueggemann, R. and Voight, K. (1995) An evaluation of online databases by methods of lattice theory. *Chemosphere* 31, 3585–3594

23 Richard, A.M. *et al.* (2003) Standardization and structural annotation of public toxicity databases: Improving SAR capabilities and linkage to -omics data. *226th ACS National Meeting*, 7-11 September 2003, New York, NY, USA (Abstract TOXI-024)

24 Helma, C. *et al.* (2000) Data Quality in Predictive Toxicology: Identification of Chemical Structures and Calculation of Chemical Properties. *Environ. Health Perspect.* 108, 1029–1033

25 Russell, J. (2005) Pfizer Researcher Paints Industry Pain Points on HBS Panel. *Biol.-IT World,* Feb. 4

26 Caldwell, G.W. *et al.* (2001) The new pre-clinical paradigm: compound optimization in early and late phase drug discovery. *Curr. Top. Med. Chem.* 1, 353–366

27 Lakings, D.B. (2000) Non-clinical drug development: pharmacology, drug metabolism, and toxicology. *New Drug Approv.* 100, 17–54

28 Dorland, L. (2002) News from online: What's new with Chime? *J. Chem. Educ.* 79, 778–782

29 Weininger, D. (1994) A distributed chemical information database system. *Special Publication – Roy. Soc. Chem* 142, 67–74

30 Jonckheere, C. (1997) Intranet for patent searching. *Proc.Int.Chem.Info.Conf.,* 19-22 October 1997, Nimes, FR (pp. 63–69)

31 Hearle, E.M. (1993) Today's information, tomorrow's technology. *Proc. Montreux Int. Chem. Inf. Conf.* (pp. 84–87)

32 Thomson, M.A. (1998) Patents on the web: The impact on "traditional" patent searching. *215th ACS National Meeting*, March 29-April 2, 1998, Dallas, TX, USA (Abstract CINF-002)

33 Rovner, S.L. (2005) Opening access *C&E News* 16 May 2005, 40–44

34 Wolpert, A.J. (2005) Movement toward open access: Why new models of research communication are inevitable. *229th ACS National Meeting*, 13-17 March 2005, San Diego, CA, USA (Abstract CINF-036)

35 Cooper, J.W. *et al.* (2005) Automatic discovery and annotation of organic chemical names in patents. 229th ACS National Meeting, 13-17 March, San Diego, CA, USA (Abstract COMP-316)

36 Jackson, P. and Moulinier, I. (2002) *Natural language processing for online applications: Text retrieval, extraction and categorization,* John Benjamins Pub. Co.

37 Brecher, J. (1999) Name=Struct: A practical approach to the sorry state of real-life chemical nomenclature. *J. Chem. Inf. Comput. Sci.* 39, 943–950

38 Garfield, E. (1962) An algorithm for translating chemical names to molecular formulas. *J. Chem. Doc.* 2, 177–179

39 Garfield, E. (2001) From laboratory to information explosions…the evolution of

chemical information services at ISI *J. Info. Sci.* 27, 119–125

40 Vander Stouw, G.G. *et al.* (1967) Procedures for converting systematic names of organic compounds into atom-bond connection tables. *J. Chem. Doc.* 7, 165–169

41 Cooke-Fox, D.I. *et al.* (1989) Computer Translation of Systematic Organic Chemical Nomenclature, Part 1. Introduction and background to a Grammar-based Approach. *J. Chem. Inf. Comput. Sci.* 29, 101–105

42 Cooke-Fox, D.I. *et al.* (1989) Computer translation of systematic organic chemical nomenclature. 2. Development of a formal grammar. *J. Chem. Inf. Comput. Sci.* 29, 106–112

43 Cooke-Fox, D.I. *et al.* (1989) Computer translation of systematic organic chemical nomenclature. 3. Syntax analysis and semantic processing. *Chem. Inf. Comput. Sci.* 29, 112–118

44 Cooke-Fox, D.I. *et al.* (1990) Computer translation of systematic organic chemical nomenclature. 4. Concise connection tables to structure diagrams. *J. Chem. Inf. Comput. Sci.* 30, 122–127

45 Cooke-Fox, D.I. *et al.* (1990) Computer translation of systematic organic chemical nomenclature. 5. Steroid nomenclature. *J. Chem. Inf. Comput. Sci.* 30, 128–132

46 Cooke-Fox, D.I. *et al.* (1991) Computer translation of systematic organic chemical nomenclature. 6. (Semi)automatic name correction. *J. Chem. Inf. Comput. Sci.* 31, 153–160

47 Wisniewski, J.L. (1993) *AUTONOM-a Chemist's Dream: System for (Micro)computer Generation of IUPAC-compatible Names from Structural Input In Chemical Structures 2* (Warr, W. A., ed.), pp 55-64, Springer-Verlag

48 Hodge, G.M. *et al.* (1989) Automatic Recognition Of Chemical Names In Natural-Language Texts. *Abstracts of Papers of the American Chemical Society*. *197*, April 1989 (Abstract CINF-17)

49 Hodge, G.M. (1989) Enhanced Chemical Name Identification Algorithm. *Abstracts of Papers of the American Chemical Society 202*, August 1989 (Abstract CINF-41)

50 Zamora, E. and Blower, P.E. (1984) Extraction of chemical reaction information from primary journal text using computational linguistics techniques. 1. Lexical and syntactic phases. *J. Chem. Inf. Comput. Sci.* 24, 176–181

51 Zamora, E. and Blower, P.E. (1984) Extraction of chemical reaction information from primary journal text using computational linguistics techniques. 2. Semantic phase. *J. Chem. Inf. Comput. Sci.* 24, 181–188

52 Ai, C.S. *et al.* (1990) Extraction of chemical reaction information from primary journal text. *J. Chem. Inf. Comput. Sci.* 30, 163–169

53 Smeaton, A.F. (1992) Progress in the application of natural language processing to information retrieval tasks. *Comp. J.* 35, 268–278

54 Chowdhury, G.G. and Lynch, M.F. (1992) Automatic interpretation of the texts of chemical patent abstracts. 1. Lexical analysis and categorization. *J. Chem. Inf. Comput. Sci.* 32, 463–467

55 Chowdhury, G.G. and Lynch, M.F. (1992) Automatic interpretation of the texts of chemical patent abstracts. 2. Processing and results. *J. Chem. Inf. Comput. Sci.* 32, 468–473

56 Kemp, N. and Lynch, M. (1998) Extraction of information from the text of chemical patents. 1. Identification of specific chemical names. *J. Chem. Inf. Comput. Sci.* 38, 544–551

57 Goldfarb, C. (1990) *The SGML handbook*, Clarendon

58 Wilbur, W.J. *et al.* (1999) Analysis of biomedical text for chemical names: A comparison of three methods. *Proc. AMIA Symp.* 1999, 176–180

59 Singh, S.B. *et al.* (2003) Text influenced molecular indexing (TIMI): A literature database mining approach that handles text and chemistry. *J. Chem. Inf. Comput. Sci.* 43, 743–752

Forthcoming articles

Chromatin control and cancer drug discovery: realising the promise
by Adam G. Inche and Nicholas B. La Thangue

***Drosophila* models pioneer a new approach to drug discovery for Parkinson's disease**
by Alexander J. Whitworth, Paul D. Wes and Leo J. Pallanck

Role of pharmacologically active metabolites in drug discovery and development
by Aberra Fura

Emerging chemical and biological approaches for the preparation of discovery libraries
by Grant E. Boldt, Tobin J. Dickerson and Kim D. Janda

Oncology exploration: charting cancer medicinal chemistry space
by David G. Lloyd, Georgia Golfis, Andrew J.S. Knox, Darren Fayne, Mary J. Meegan and Tudor I. Oprea

Selective optimization of side activities
by Camille Wermuth

The impact of microwave-assisted organic synthesis in drug discovery
by Farah Mavandadi and Åke Pilotti

Optimizing the use of open-source software applications in drug discovery
by Werner J. Geldenhuys, Kevin E. Gaasch, Mark Watson, David D. Allen and Cornelis J. Van der Schyf

SAGE and related approaches for cancer target identification
by Dale Porter, Jun Yao, and Kornelia Polyak

Improving the hit-to-lead process: data-driven assessment of drug-like and lead-like screening hits
by Tobias Wunberg, Martin Hendrix, Alexander Hillisch, Mario Lobell, Heinrich Meier, Carsten Schmeck, Hanno Wild and Berthold Hinzen