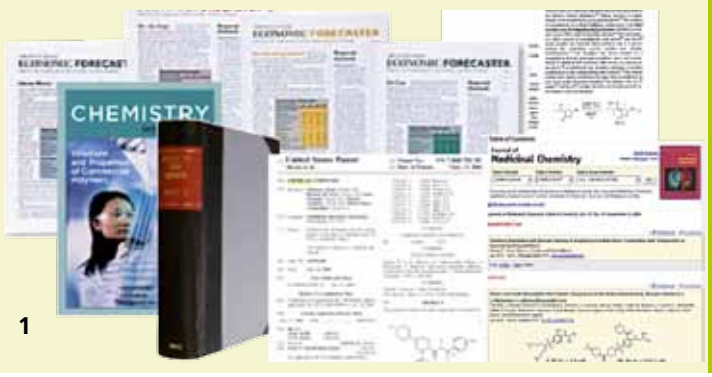
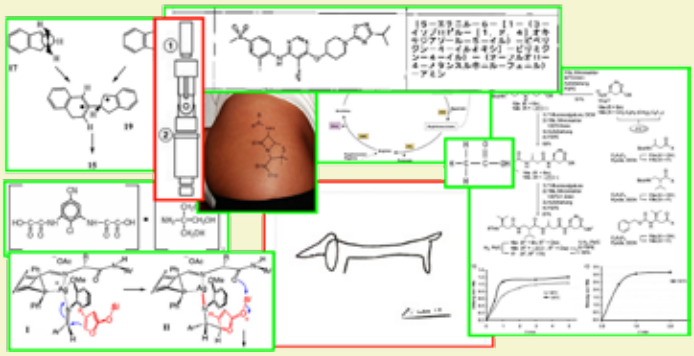
Fraunhofer
SCAI

FRAUNHOFER INSTITUTE FOR ALGORITHMS AND SCIENTIFIC COMPUTING SCAI



1



2

1 Chemical depictions can be found in all kinds of documents like patents, journals, theses, etc.
2 Not all images in chemical sources contain structures.

ChemoCR® – TOOL FOR CHEMICAL COMPOUND RECONSTRUCTION

Fraunhofer Institute for Algorithms and Scientific Computing SCAI

Schloss Birlinghoven 1
53757 Sankt Augustin
Germany

Contact
Dr. Marc Jacobs
phone +49 2241 14-4013
marc.jacobs@scai.fraunhofer.de
www.scai.fraunhofer.de/chemocr

Distribution
scapos AG
phone +49 2241 14-4400
www.scapos.com

chemocr
chemical compound reconstruction

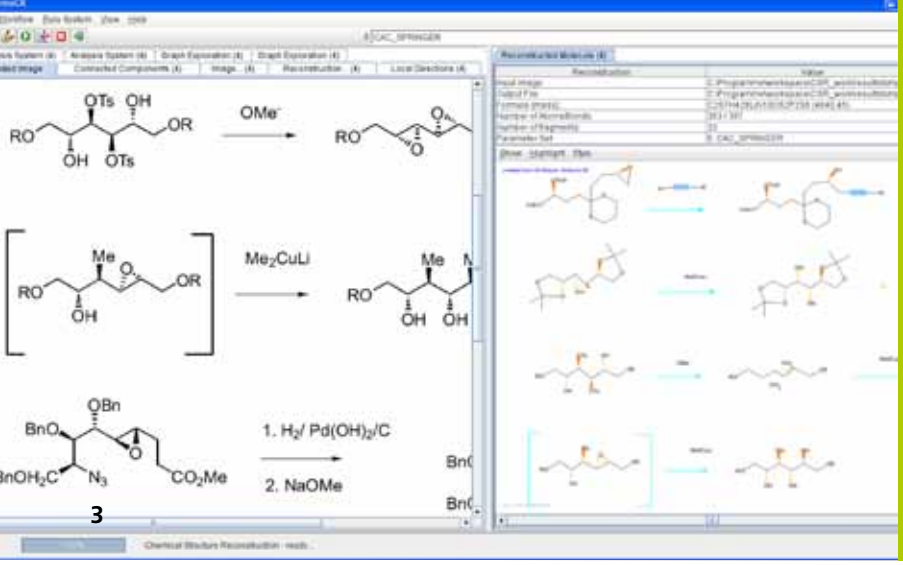
Situation

SCAI Bioinformatics has a longtime working experience on the automated extraction of information from text in the biomedical literature. While established techniques exist for text mining, chemical structure extraction is a key feature and presents a significant technological challenge. In contrast to general purpose search engines such as GOOGLE the extraction of the so called connection table from images allows chemical relevant queries. Using GOOGLE image search a user could ask: "Find all documents which are showing aspirin in an image". The resulting images are identified by searching for the term aspirin in the caption or surrounding text.

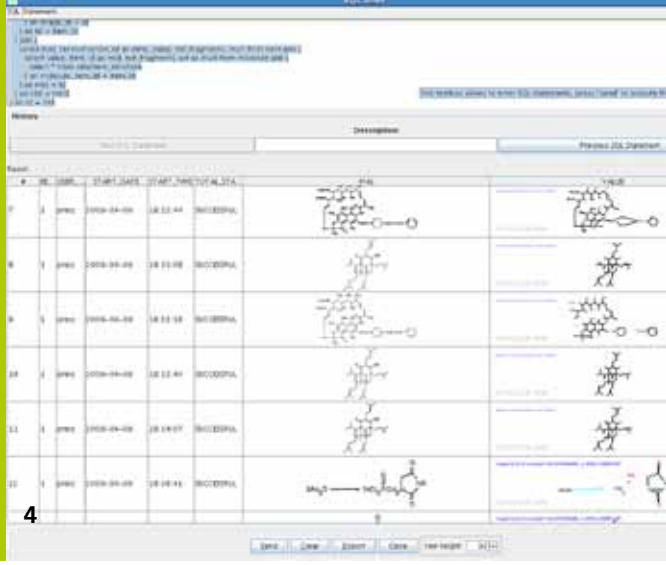
A chemical structure extraction allows for semantically rich queries like "Find chemical similar structures of aspirin", "What core structures have been patented for COX-2-inhibitors?", "Show reaction schemes for the synthesis of aspirin." This search is not relying on the text and therefore even works for Asian patents without translating them.

We are focusing the scope of our research on chemical entity recognition from images in collaboration with our strategic partner InfoChem GmbH. Depictions can be found as images in nearly all electronic sources of chemical information (e.g. journals, reports, patents, and web interfaces of chemical databases) (cf. figure 1).

Nowadays, these images are generated with special drawing programs, either automatically from computer-readable file formats or by the chemist through a graphical user interface. Although drawing programs can produce and store the information in a computer-readable format, chemical structure depictions are published



3



4

as bitmap images. To make published chemical structure information available in a computer-readable format, images representing chemical structures have to be manually converted by redrawing every structure. This is a time-consuming and error-prone process.

chemoCR is a software which supports this process by automatic extraction and conversion of chemical depictions.

Technology

In order to solve the problem of recognizing and learning chemical structures in image documents, our chemoCR system combines pattern recognition techniques with a chemical rule based expert system.

The method is based on the idea of identifying the most significant fragments of small molecules from depictions.

The workflow consists of three phases: image vectorization, chemical entity extraction and molecule reconstruction.

The main features of the software are:

- Conversion of PDF documents into full page scans
- Segmentation of full page scans into chemical schemes
- Classification and rejection of non-chemical depictions
- Conversion of various bitmap images (e.g. BMP, TIF, PNG) into chemical file formats (e.g. SMILES, SDF)
- Reconstruction of the full bond information (single, double, triple, chiral bonds)
- Recognition of superatoms and their conversion into structural representation
- Scoring scheme for the reconstruction process based on known chemical scaffolds
- Matching of the reconstructed structure against given input structures
- Fully automatic batch processing mode (can be distributed over a cluster)
- The whole process and the result can be logged in a database or the properties section of SDF
- GUI for manual curation (cf. figure 3)
- Training ability for the OCR process (e.g. fused letters) and teaching new superatoms
- Customization via simple manipulation of XML parameter files
- Chemical intelligence (e.g. filling free valences)
- Recognition of R-groups and reaction symbols but not including Markush structures

Expanding Application Fields

In this highly interdisciplinary domain, interesting information is often presented as a combination of text and graphics.

Combining textual IE methods with chemoCR for the multimodal information extraction of Markush structures from patents has not yet been addressed.
This functionality will be part of a future solution. At the moment we are extending the chemical rule system for complex reaction schemes.

Technical Specification

The chemoCR core functionality is based on platform-independent JAVA libraries. It has been extensively tested on UNIX operating systems (Fedora Linux, Sun Solaris) and on Windows XP and VISTA. Users may apply our software interactively by a graphical user interface or run it distributed in batch processing mode in a grid enabled hardware environment.

Offering

We are offering document corpora evaluation and conversion projects. We have established an in-house workflow of manual selection of representative pages from your document collection, fully manual abstraction of a Gold standard. We are doing optimization and evaluation of our automatic methods. All results are stored in a web based retrieval prototype. We provide a detailed report on the evaluation. Based on the results we will jointly develop a strategy for the conversion of large collections – reducing the manual extraction effort.

3 *ChemoCR graphical user interface: The left panel shows the input image and the intermediate reconstruction results. On the right the resulting molecule is drawn.*

4 *All extracted information can be stored in a database retrieval system.*