

Automated Extraction of Chemical Information from Chemical Structure Depictions

a report by

Marc Zimmermann¹ and Martin Hofmann-Apitius²

Head of: 1. Computational Chemistry Group; 2. Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, Sankt Augustin

Chemical information comprises chemical entities (named entities and structure depictions) and relationships between them (temporal, spatial or semantic relationships). Chemists communicate chemical information through printed and electronic media and, consequently, chemical entities can appear in scientific text as brand names, assigned catalogue names or International Union of Pure and Applied Chemistry (IUPAC) names. However, the preferred representation of chemical entities is often a 2D depiction of the chemical structure (see *Figure 1*). 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). For example, entering the search term 'chemical structure' into Google's image search retrieves 1.37 million websites, with approximately 50% of them displaying true chemical structure depictions (CSDs) (search performed in August 2007).

Nowadays, these images are generated with special drawing programs, either automatically from connection table file formats, or by the chemist through a graphic user interface. Although drawing programs can produce and store the information in a computer-readable format (e.g. structure data (SD) files), CSDs are still published as bitmap (BMP) images (e.g. graphic interchange format (GIF) for web interfaces, or BMP for text documents). Consequently, the structure information can no longer be applied to chemical analysis software packages. To make published chemical structure information available in a computer-readable format (connection tables), images representing chemical structures have to be converted by human experts redrawing every structure. This is a time-consuming, error-prone process and, although this extraction of chemical information is often performed in countries where expert labour is cheap, it can be costly.

It seems likely the mode for publishing chemical information will change, and CSD files (connection tables) will hopefully be made available as embedded files attached to structure depictions. However, over 100 years of chemical structure drawing has resulted in myriad CSDs in

...images representing chemical structures have to be converted by human experts redrawing every structure.

electronic archives. This archived information is highly relevant to all industries and institutions dealing with chemistry, but as the potential benefit of this information cannot easily be predicted, people involved in chemical data management are waiting for cost-effective, automated systems that help them to reconstruct chemical information from all sorts of electronic media (ranging from scanned documents to fully structured digital libraries). This article will discuss matters pertinent to chemical structure reconstruction (chemoCR®).

Challenges Associated with Computational Approaches to Chemical Structure Reconstruction

The general methodology of interpreting a CSD consists of the following three stages:

1. Pre-processing:

- identification of a CSD in a document;
- segmentation of the document and importation of the CSD; and
- normalisation of greyscale and size of the image.

2. Reconstruction of chemical information:

- vectorisation of graphical elements;
- interpretation of different classes of graphical elements, e.g. dashed lines, dotted lines, wedges, etc.;
- character recognition and interpretation of symbols; and
- graph compilation.

3. Post-processing:

- export of connection table;
- quality assessment of reconstruction and display of results; and
- interaction with the user/teaching of the system.



Marc Zimmermann is Head of the Computational Chemistry Group at the Fraunhofer Institute for Algorithms and Scientific Computing (SCAI). In the Department of Bioinformatics at Fraunhofer SCAI he has been instrumental in the development of the chemoCR® software for the reconstruction of chemical information from chemical structure depictions. He studied computer science at the University of Bonn, and in his PhD thesis he developed new algorithms for the analysis of high-throughput screening (HTS) data.



Martin Hofmann-Apitius is Head of the Department of Bioinformatics at Fraunhofer Institute for Algorithms and Scientific Computing. In July 2006 he was appointed Professor for Applied Life Science Informatics at the University of Bonn. His professional background is in both the pharma and biotech industries, as well as in academic research. His special interest is in information extraction technology for the life sciences. He studied biology at the University of Tuebingen.

Pre-processing

The first problem for the automated execution of the workflow is to evaluate whether an image contains chemical information. For automated analysis of archival data, it would be desirable to be able to identify CSDs in documents based on features of the image itself. A classification system that is able to do this is currently under research in our group, but so far there is no working solution to this problem. Consequently, segmentation of documents and import of images has to be carried out after manual selection of images, even though some of the search capabilities in, for example, patent collections allow significant enrichment of documents comprising chemical structure depictions. Normalisation of heterogeneous structure depictions is also a largely unsolved problem. This is particularly relevant when dealing with heterogeneous archival information, but also when structure depictions in patents are to be analysed.

Reconstruction of Chemical Information

To a computer, most of the CSDs are just a collection of pixels with a different greyscale. Of course, CSDs can also appear as images with colours; frequently we do also find images that contain only black and white. The computer has to find out where the lines (representing bonds) are, which pixels belong to wedges and other non-standardised graphical elements and which elements in the image represent characters. In computer graphics, lines can be described as vectors with a start- and end-point. Whether a series of pixels with a certain greyscale forms a vector is tested by a 'vectoriser', which is an algorithm that establishes a vector from any non-random distribution of pixels. Vectorisers analysing CSDs have to be able to deal with the entire spectrum of possible peculiarities of chemical drawing (e.g. double bonds that are represented by one thick line instead of two clearly distinguishable lines representing bonds), and thus vectorisers that have been generated for the reconstruction of technical drawings are not *a priori* suited for chemoCR.

Other elements that have to be identified by the vectoriser are stereo bonds, represented by arrowhead-like drawings (wedges). These elements vary quite significantly in CSDs, and wedges represented by dashed lines have to be distinguished from those represented by one thick arrowhead-like element. Dotted lines and dashed lines have to be interpreted and impose a significant challenge to vectoriser algorithms.

Atom symbols represented by characters in CSDs have to be dealt with separately. They can be identified by their special features, such as the ratio between the lengths of vectors arranged in well-defined angles. The character interpretation can be performed using off-the-shelf optical character recognition (OCR) tools,¹ but the presence of characters being isolated in chemical drawings tends to mislead generic OCR tools and better results can be achieved using dedicated character OCR developments (Santiago Akle Serrano, unpublished observation). Examples of special characters and graphical elements as they appear in CSDs can be seen in *Figure 2*.

Once the vectors representing chemical bonds have been defined by the vectoriser algorithm(s), the reconstruction software has to identify connections between individual vectors. This is done in a step looking for 'connected components'. At the graphic level, connected components can be defined as vectors that share one end-point, but depending on the quality of the input image it may be necessary to 'repair' gaps that have been introduced in connected components due to image handling

Figure 1: Example of Complex Chemical Structure Depiction

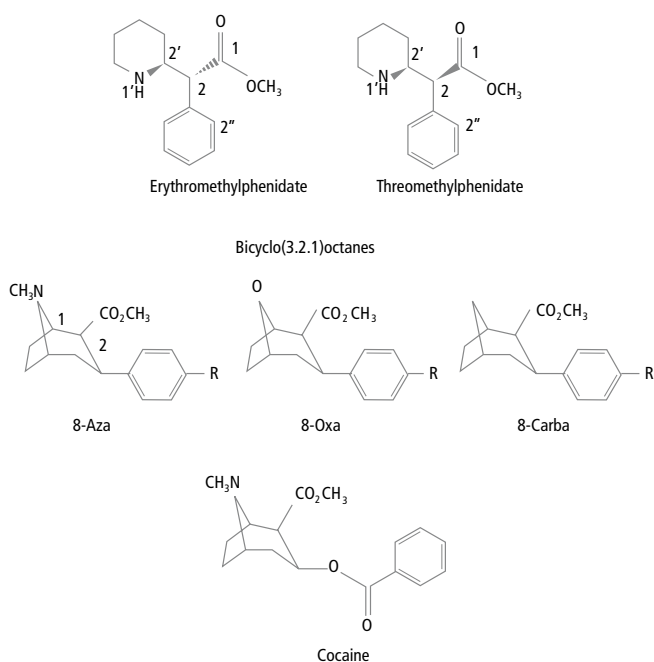


Figure 2: Examples of Difficult Characters or Graphical Elements

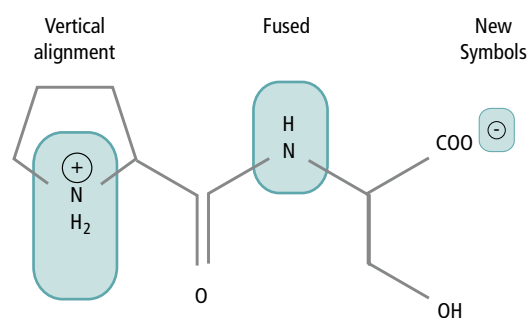
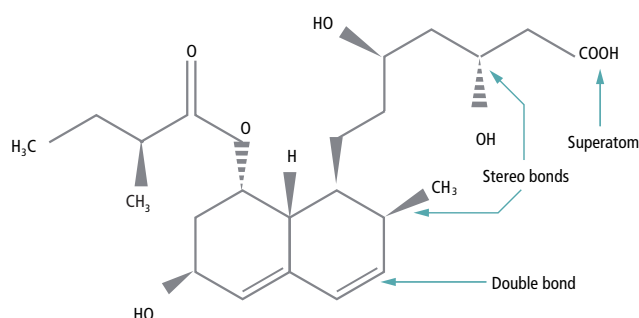


Figure 3: Molecule Annotated to Illustrate Chemical Semantics



or bad printing quality. Of course, the decision as to whether a gap represents an intended separation of two graphical elements or is in fact a printing/drawing error that needs to be fixed (and the components need to be connected) is very often a non-trivial one. Decision-making can be supported by rule-based approaches, as well as by learned strategies for well-defined classes of chemical drawings (e.g. adaptation of parameters for certain drawing tools).

Finally, the reconstruction of graphical elements and characters has to be interpreted as a chemical molecule. This means that the interpretation

has to make sense of the vectors ('bonds') and characters (e.g. superatoms) in the image. Chemical semantics are needed to provide an accurate interpretation of the CSD, and in the ideal scenario the system provides a sort of confidence value for this interpretation of the elements in the CSD (see *Figure 3*).

Post-processing

Following the compilation of a molecule structure, the connection table representing the chemical structure is exported to either a database or a user interface. If the software system performs an assessment of the reliability of the reconstruction process, the user can interact with the system and either correct the reconstructed graph (this would mean an editing function) or teach the system how to deal with a defined

The ability to adapt to specific reconstruction challenges is of utmost importance, and adaptive approaches towards chemoCR will decide how far automated systems for chemoCR will be usable...

problem (e.g. the expansion of a superatom, or the interpretation of very short wedges representing stereo bonds). The ability to adapt to specific reconstruction challenges is of utmost importance, and adaptive approaches towards chemoCR will decide how far automated systems for chemoCR will be usable for industrial-scale information content production.

Tools for Chemical Structure Reconstruction – An Overview

Computerised approaches to the reconstruction of chemical information from CSDs go back to 1992, when IBM's Steven Boyer issued a patent on computer-based chemical structure recognition.² We learned that IBM is offering chemical structure information generated from US Patent and Trademark Office (USPTO) patents to the pharmaceutical industry, but so far we have not been able to find the official offering from IBM. The patent has not been renewed and expired recently.

Also in 1992, the Kekulé program was published,³ which to our knowledge was the first program of its kind. The original paper outlined the complete workflow, from scanning of images to editing and expert correction of the reconstructed chemical structure. Kekulé is able to export connection tables in various formats (e.g. SD file), and the authors report on initial benchmark testing with more than 400 compounds.³ The examples demonstrated in the publication fall into the class of clearly drawn structures as they can be found in high-quality chemical journals. The system includes the user in the workflow as an editor of reconstructed structures, but does not seem to learn from erroneous reconstructions. Active development of Kekulé seems to have stopped a long time ago, and to our knowledge the program is not freely available to interested scientists.

In the early 1990s, a project performed at the University of Leeds led to the development of the chemical literature data extraction (CLiDE)

project, a program that was further developed into the first commercial tool for chemoCR.⁴ CLiDE essentially followed the strategy for chemoCR outlined above. In our hands, CLiDE reconstructs about 50% of the blockbuster drug structures of 2004 drawn using ChemDraw.⁵ However, CLiDE does not allow teaching the system to learn from reconstruction errors and, therefore, the system is not adaptive. Furthermore, CLiDE cannot be started from the command line and thus cannot easily be integrated into more complex workflows, for example those based on the unstructured information management architecture (UIMA) framework. Nonetheless, for quite some time CLiDE was the only tool available on the market for chemoCR, and as such set the standard in this area. The authors of this review have no information on large-scale application of CLiDE in the publishing industry or in the pharmaceutical industry. The program itself and initial applications were published between 1992 and 1997, but we did not find any reference for CLiDE in the scientific literature after 1997.^{6,7,8}

Recently, the matter of chemical information reconstruction from structure depictions has been addressed again by two teams, in the US and in Europe. Igor Filippov's team at the National Cancer Institute (NCI) in the US has introduced optical structure recognition analysis (OSRA). Our team at the Fraunhofer Institute for Algorithms and Scientific Computing (SCAI) has developed chemoCR over the last three years.

OSRA is an academic application that can be freely downloaded or tested online.⁹ The tool is part of a set of chemoinformatics tools available from the Computer-Aided Drug Design Group at the NCI-Frederick, Maryland. OSRA follows the generic CSD reconstruction workflow outlined above. It has been reported to accept as input over 90 graphic formats parseable by ImageMagick: GIF, Joint Photographic Experts Group (JPEG), portable network graphics (PNG), tagged image file format (TIFF), portable document format (PDF), PostScript (PS) and others. As a method of output, OSRA generates a simplified molecular input line entry

In the early 1990s, a project performed at the University of Leeds led to the development of the chemical literature data extraction (CLiDE) project, a program that was further developed into the first commercial tool for chemoCR.

specification (SMILES) of the molecular structure images used as input. The system has recently been tested by Antony Williams, who reported his experiences in the ChemSpider blog.¹⁰ OSRA seems to work reasonably well on clean structure depictions; however, Williams reports on problems with stereo bonds, double bonds and metallo-organic structures. In conclusion, OSRA is rated as 'work in progress', but the general approach and the idea of a tool freely available to the chemist community is appreciated widely in chemistry blogs and scientific forums.

The approach taken by our team at Fraunhofer SCAI does not significantly deviate from the basic workflow outlined above.¹¹ However, after testing 'off-the-shelf' open-source vectorisers and different character OCR tools, we decided to develop more dedicated systems for vectorisation and

character OCR, as most of the major problems were obviously linked to the 'general purpose' nature of open-source vectorisers and OCR tools. This is not really surprising because vectorisers had initially been developed for vectorisation of technical drawings and OCR tools assume that they work on lines of text, not on isolated characters in an unknown graphic with unusual lines and graphic elements.

...scientific challenges behind 'chemical optimal character recognition' are substantial and the vision of large-scale automated reconstruction of chemical information from large digital archives will require sustained research and development over years.

In order to solve the problem of recognising and learning chemical structures in image documents, our chemoCR system combines evidence-based learning techniques with supervised machine-learning concepts (i.e. support vector machines) and a rule-based expert system. The method is based on the idea of identifying from depictions the most significant semantic entities (e.g. chiral bonds, superatoms, reaction arrows). All steps in the process make use of chemical knowledge in order to detect and fix errors. The system can be adapted to different sets of input images, a feature that pays tribute to the heterogeneity of CSD sources. Among the dedicated developments for chemoCR are:

- a new vectorisation algorithm based on textures;¹²
- a new OCR tool for chemical characters combining evidence-based learning and support vector machines;
- a new expert system for the extraction of chemical entities by combining graphical primitives and chemical knowledge; and
- a scoring module for the reconstruction validation.

The tool accepts various bitmap images (e.g. BMP, GIF, PNG) as input, but does not yet perform document segmentation autonomously. Depictions with multiple molecules (including complex synthesis reaction schemata) can be handled, but full-page scans are not segmented yet. New graphical elements and new characters (e.g. fused atom symbols) can be learned by the system if the expert user corrects erroneous reconstructions. Consequently, the system becomes smarter the longer it is in use.

Future Challenges

From numerous discussions with scientists in the chemical and molecular biology sciences, as well as from discussions we had with

scientists working in the pharmaceutical and publishing industries, we know that the attempts to automate the reconstruction of chemical information from CSDs are highly appreciated. There are a few challenges that lie ahead of us on the road towards large-scale production-ready systems for chemoCR:

- automated recognition of images representing chemistry in documents (document segmentation);
- automated pre-processing of images and improved automated optimisation of parameters for chemical OCR;
- generation of large, well-annotated corpora of chemical images to train adaptive components in chemical OCR systems; and
- use of annotated corpora for benchmarking purposes, after the "critical assessment of text mining systems in biology" stance advocated by BioCreative.¹³

Conclusion

The perspectives of automated chemical information reconstruction from CSDs are quite attractive and, consequently, this field has attracted computer scientists working on this problem since the early 1990s. However, the various scientific challenges behind 'chemical OCR' are substantial and the vision of large-scale automated reconstruction of chemical information from large digital archives will require sustained research and development over years. Nonetheless, in the mid-term we expect first reports on automated workflows

The next really big scientific challenge lies in the co-ordinated, multimodal information extraction from documents containing both chemical images and chemical (or biological) entities in text.

extracting relevant information from large document collections, e.g. patents or large digital archives. Moreover, we anticipate that the international research community will initiate public benchmarking activities and join forces for the generation of well-annotated corpora. Besides the perspectives on the reconstruction of chemical information from archives, we also see an opportunity for greatly enhanced metadata generation for web-based chemical information. The next really big scientific challenge, however, lies in the co-ordinated, multimodal information extraction from documents containing both chemical images and chemical (or biological) entities in text. ■

1. <http://www.gocr.de>
2. US Patent and Trademark Office patent number 5345516.
3. McDaniel JR, Balmuth JR, Kekule: OCR-optical chemical (structure) recognition, *J Chem Inf Comput Sci*, 1992;32:373–8.
4. Ibison P, Kam F, Simpson RW, et al., Chemical Structure Recognition and Generic Text Interpretation in the CLiDE project, Proceedings on Online Information 92 meeting, London, England, 1992.
5. Zimmermann M, Bui Thi LT, Hofmann M, Combating Illiteracy in Chemistry: Towards Computer-Based Chemical Structure Reconstruction, *ERCIM News No. 60*, January 2005.
6. Kam F, Simpson RW, Tonnelier T, et al., Chemical Literature Data Extraction. Bond Crossing in Single and Multiple Structures, Proceedings of the 1992 Chemical Information Conference, Annecy, France, 1992.
7. Ibison P, Jacquot M, Kam F, et al., Chemical Literature Data Extraction: The CLiDE Project, *J Chem Inform Comput Sci*, 1993; 33(3):338–44.
8. Simon A, Johnson AP, Recent Advances in the CLiDE Project: Logical Layout Analysis of Chemical Documents, *J Chem Inform Comput Sci*, 1997;37(1):109–16.
9. <http://cactus.nci.nih.gov/cgi-bin/osra/index.cgi>
10. <http://www.chemspider.com/blog/?p=83>
11. <http://www.scai.fraunhofer.de/chemocr.html>
12. Algorri M-E, Zimmermann M, Friedrich CM, et al., Reconstruction of Chemical Molecules from Images, Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society EMBC 2007, accepted for publication.
13. <http://biocreative.sourceforge.net/>