Chemical Structure Reconstruction with chemoCR

Dr. Marc Zimmermann

Fraunhofer-Institute for Algorithms and Scientific Computing (SCAI)
Schloss Birlinghoven
D-53754 Sankt Augustin
E-mail: CSR-team@scai.fraunhofer.de
Internet: http://www.scai.fraunhofer.de/chemocr.html

Abstract

chemoCR makes chemical information contained in depictions of chemical structures accessible as connection table for computer programs. In order to solve the problem of recognizing and translating chemical structures in image documents, our chemoCR system combines pattern recognition techniques with supervised machine learning concepts. The method is based on the idea of identifying from structural formulas the most significant semantic entities. Semantic entities are for example chiral bonds, superatoms and reaction arrows. The workflow consists of three phases: image preprocessing, semantic entity recognition, and molecule reconstruction plus validation of the result. All steps of the process make use of chemical knowledge in order to detect and fix errors. The system can be trained and adapted to different sources of input images. The reconstructed connection table can be used by all chemical software.

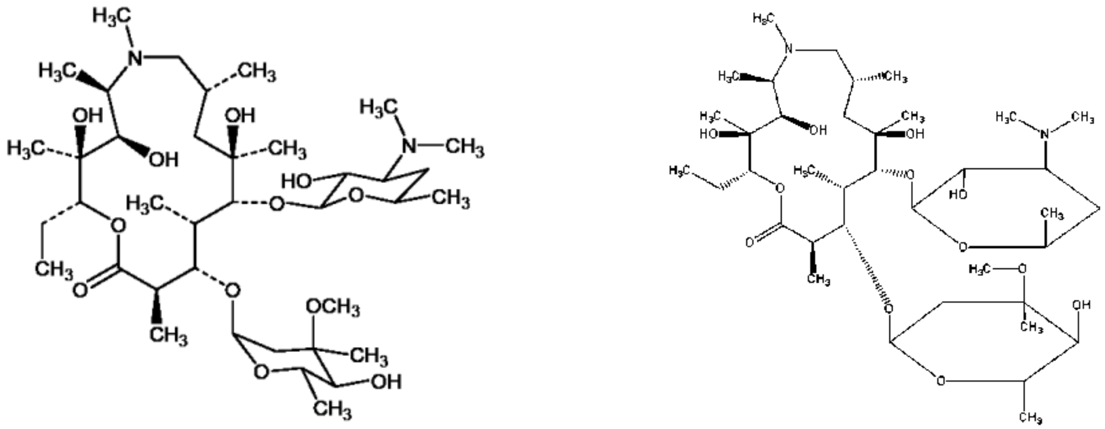


Figure 1: This figure shows the drug azithromycin drawn with two different structural editors. The drawn molecules are identical but the images are quite different.

1. Problem Description

Chemical entities can appear in scientific texts as trivial and brand names, assigned catalogue names, or IUPAC names. However, the preferred representation of chemical entities is often a two-dimensional depiction of the chemical structure. 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). Nowadays these images are generated with special drawing programs (e.g. ISISdraw [10], cf. **Figure 1**), either automatically from connection table 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 as bitmap images (e.g. GIF for web interfaces or BMP for text documents). As a consequence, the structure information can no longer be used as input to chemical analysis software packages. 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 [13, 15].

2. Methods

Several articles are focusing on the chemical compound reconstruction problem (CSR) [3, 6, 8]. In order to solve the problem of recognizing and learning chemical structures in image documents, our chemoCR system combines pattern recognition techniques with supervised machine-learning concepts. The method is based on the idea of identifying from depictions the most significant semantic entities (e.g. chiral bonds, superatoms, reaction arrows...). The workflow consists of three phases: image preprocessing, semantic entity recognition, and molecule reconstruction plus validation of the result. All steps of 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. The following algorithms have been developed and integrated into the system:

- a new vectorization algorithm based on textures
- a new OCR tool for chemical characters using machine learning
- a new page segmentation algorithm which makes use of classification of chemical objects in mixed diagrams
- a new expert system for the extraction of chemical entities by combining graphical primitives and chemical knowledge
- a scoring module for the reconstruction validation

The chemoCR core functionality is based on platform-independent JAVA libraries. Details can be found at [1, 2, 5, 7, 13, 12, 14].

For the Image-to-Structure Task (I2S) we used chemoCR v0.93 running in batch mode using the image set as input and producing a set of SDFiles. As first step the correct parameter set has to be selected. A parameter set is a collection of parameter settings which influence the quality of the reconstruction process. As the shape of lines in pixel images tend to be not generally uniform chemoCR introduced a lot of parameters to adapt to different drawing styles. By now there are 12 different sets of parameters each of which just goes fine with a special sort of input images (e.g. different drawing tools, different journal styles, patents, textbooks). A lot of experience and knowledge went along in optimizing them.

A representative set of example images from the training set and the topic set have been manually processed in the graphical user interface and the results have been carefully checked against the provided MOL files. The provided images have been quite uniform using the same drawing style and almost the same image resolution. Therefore we decided to select a single parameter set and apply it to the whole image set. The preconfigured parameter set “Houben-Weyl” has been selected – showing the best results. The Houben-Weyl set has been developed for the book series “Houben-Weyl” which are published by Thieme and contain the description of organic chemistry reactions over a large period of time. Thus the depictions vary in the shown in different chemical scaffolds and drawing style, but the quality of the images is high. The same applies to the provided images from the USPTO.

From the manual experimentation a standard workflow has been chosen. A workflow consists of several independent workflow modules which are executed in a predetermined order like a pipeline. This pipeline can be configured by switching modules on and off as necessary.

The predefined workflow from chemoCRSettings.xml will binarize the input image, analyze the connected components, extract and convert letters in the image, detect chiral bonds, vectorize simple bonds and bond sets (i.e. ring systems and chains) and finally combine everything into a molecule. If there are several molecules in the image they can be splitted and written in a multi molecule SDF [9]. The result of the whole process is validated and possible reconstruction errors. We give a short overview on the used workflow modules in the next section. Two preliminary steps (page segmentation and PDF processing) have not been applied as the provided images are already extracted from the patent documents. But they become necessary if one wants to process the original documents.

The PDF Conversion Task

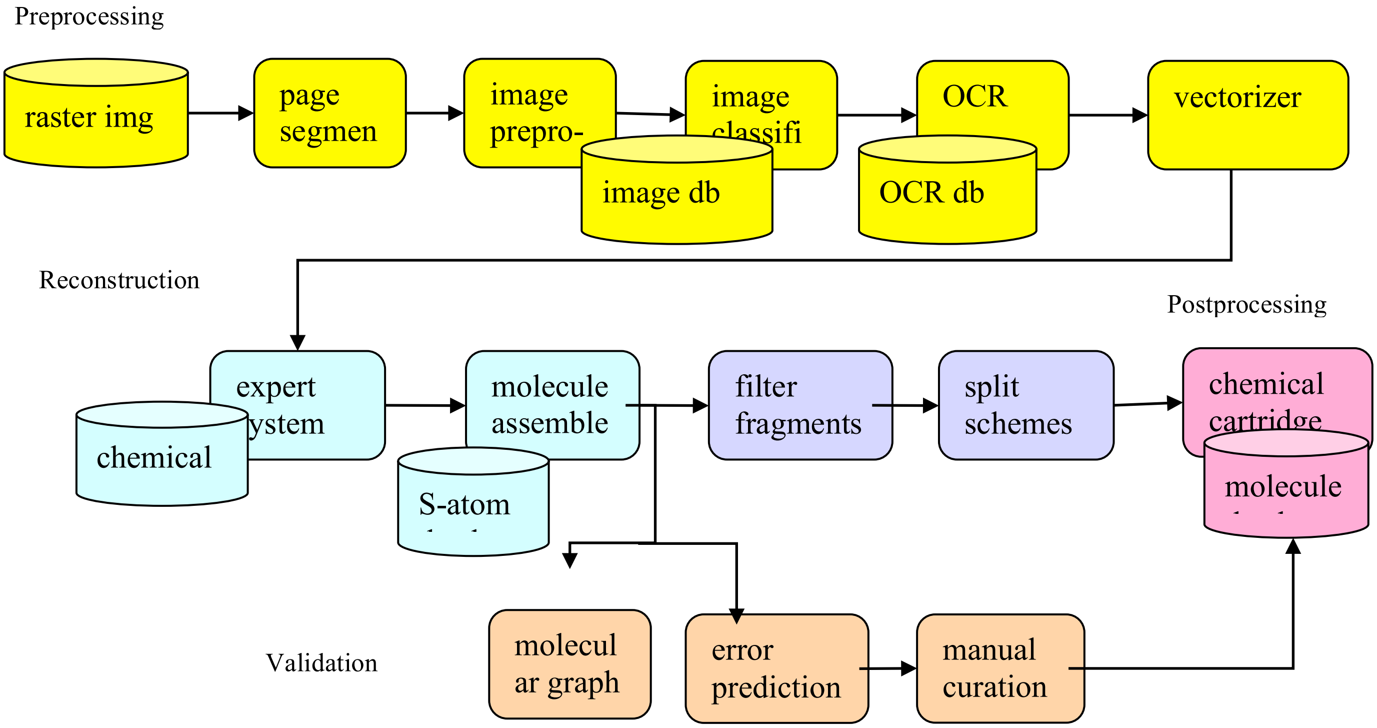
Chemical structure depictions are often packed in documents which explain the content of the depiction. These documents are often saved in Adobe Portable Document Format (PDF). To extract the information from the documents chemoCR has a PDF Conversion module, which splits the document to its single pages and saves them as a bitmap image. After this conversion it is recommended to start the Page Segmentation to get the depictions of chemical structures from the set of single pages.

The Page Segmentation Task for Chemistry Journals

Scientific literature contains a lot of information which is found not only in text but also in depictions, tables and diagrams. Even though chemistry journals are already digitally available (e.g. as PDF); however for extracting chemical related information like structural formulas from these sources it is necessary to identify and separate chemical objects and their context from the rest of the document. This is a quite complicated task due to the fact that the layout of scientific articles is quite complicated and differs from normal text sources. In the second step the separated chemistry objects are processed by name-to-structure or image-to-structure tools like chemoCR.

The main Workflow Description

chemoCR organizes its work in a pipeline of subtasks: the workflow modules. They are grouped into four sections: Preprocessing Section (yellow boxes), Reconstruction Section (light blue boxes), Postprocessing Section (dark blue boxes), and Validation Section (orange boxes). The different modules of the modules of chemoCR will be described in the next sections.



Preprocessing Section

The following modules belong to the Preprocessing Section. These will analyze and convert the bitmap image into higher abstraction level objects like connected components, vectors and letters.

Vaporizer Unit

chemoCR's competence in the fields of the reconstruction of chemical structure depictions works best with input images just consisting of structure diagrams. The Vaporizer Unit now serves as an eraser for those parts of the image that are presumably not structure diagrams - but text or other human readable information. The vaporizer just erases any misleading information in the image. But the deleted parts are kept in the output for further processing.

Image Preprocessing

This module can be used to normalize the input images. The input image can be scaled, rotated, sharpened, eroded (thinning of lines), dilated (thickening of lines) and skeletonized (center line). The effects of the preprocessing can be studied by using the wizard and turn on the live preview. All operations are performed on a copy of the original image.

Connected Components

Necessary Step. It extracts the connected components from the image. In detail here will be grouped all foreground pixels which are 8-connected into a component.

Tag Text

Necessary Step. It identifies connected components that map to text areas.

OCR (Optical Character Recognition)

Necessary Step. It converts the bitmap letters which are tagged into characters using OCR software. The used software depends on the configuration of the INI file. The internal OCR module is used by default. Some letters may be rejected by the OCR as non-characters.

Compute Local Directions

Necessary Step. It computes the local directions of segment clusters, i.e. searches for ascending, descending, horizontal and vertical lines in the bitmap image. This is a necessary step for the vectorizer.

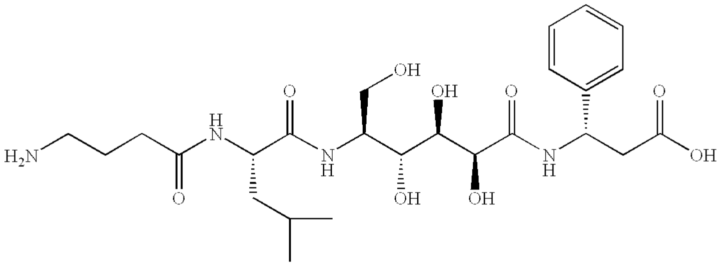


Figure 2: Example image from the USPTO set. In this image 3 thick chirals will be tagged.

Tag Thick Chirals In Image

Necessary Step. It identifies chirals which are drawn as thick wedges. This module also computes the orientation of the chiral. It does not check for the chemical correctness of the assigned chiral center.

Vectorize Bitmap Image

Necessary Step. It converts the bitmap image into a set of vectors. This is the main algorithm and relies on the Compute Local Directions module.

Reconstruction Section

The Reconstruction Section is split in two parts: the Expert System Section part and the part Assembling of the Molecule. Both modules rely on the Chemical Rule System.

The reconstruction of the image implies that chemical components like bonds or chirals are identified from the pixel image. The information that a specific line is in fact a bond or a double bond is derived by applying so-called rules. A rule inspects a component within its neighborhood and looks at its characteristics to decide if this component belongs to that sort of chemical component that the rule is looking for. There is of course more than one rule necessary to identify all the sorts and shapes of chemical components. The rules together form the rule set. A rule set is an XML file that is open to changes or extensions by the user.

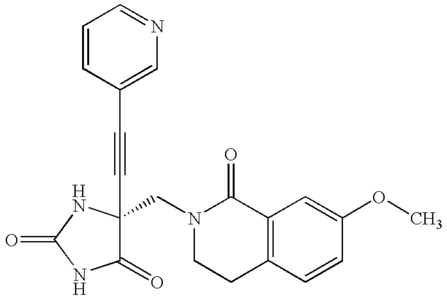


Figure 3; Example image from the USPTO set. In this image the triple bond is drawn in the same orientation as the neighboring single bonds. This is a typical case where a new rule is to be added to the expert system. As this is not a standard way in drawing them.

The expected result of applying a rule is the classification of the specific component. The classification makes use of special keywords: one specific keyword for each classification. These keywords tell chemoCR how to reconstruct the underlying component. Currently there are 14 different chemical classes which are assigned to graphical objects from the image by the rule system: BOND, DOUBLEBOND, TRIPLEBOND, BONDSET, DOTTED CHIRAL, STRINGASSOCIATION, DOT, RADICAL, REACTION, REACTION ARROW, REACTION, REACTION PLUS, CHARGE, and UNKNOWN.

Expert System Section

The workflow modules listed here do consult the Expert System for reconstruction. Chemical knowledge now is applied.

Build the Analysis System

Necessary Step. It initializes the analysis system for the graph exploration algorithm. This module collects all information on vectors, connected components and letters and their properties like average length, shape, color etc.

Orientation Graph Exploration

Necessary Step. It uses a graph constraint exploration algorithm to assign the annotations. Each connected component is visited and suitable rules are evaluated for this component. The successful rule with the highest priority value will define the annotation.

Assembling of the Molecule

The following modules use the annotations of the annotation container for building the chemical graph. All parts are assembled into semantic objects. E.g. the two vectors which are assigned as double bond are now combined into a single vector with label double bond. Each assembly step can be turned off individually. The according objects will be ignored and not be transferred into the molecule.

Assemble Bond Sets
Necessary Step. It selects each bondset and adds all vectors of this bondset as bonds to the molecule. Bondsets normally represent ringsystems and aliphatic chains.

Assemble Dotted Chirals
Necessary Step. It selects each set of vectors belonging to a dotted chiral. All these vectors will lead to a single bond in the molecule. Depending on the representation of the dotted chiral the bond will be oriented. This module checks not for valid chiral centers.

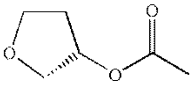


Figure 4: Example image from the USPTO set. In this image a dotted chiral is identified in the five-membered ring.

Assemble Simple Bonds
Necessary Step. It selects each single vector and tries to group them into single bonds, double bonds and triple bonds.

Assemble Atom Types
Necessary Step. It selects character groups from the image. It checks for valid identifiers from the periodic system or for atom symbols enriched by hydrogens, e.g. NH2. It tries to connect the atom group to the next bond which is close enough and pointing to this group.

Crossed Bonds Necessary Step
It selects vectors that are intersecting each other. This will create four bonds for the two vectors, having a carbon at its center (i.e. the crossing point).

Assemble Bridged Bonds
Necessary Step. It selects broken bonds which indicate bridged ring systems. This will fuse two existing short bonds being intersected by another bond into a single bond if drawn as an intersection.

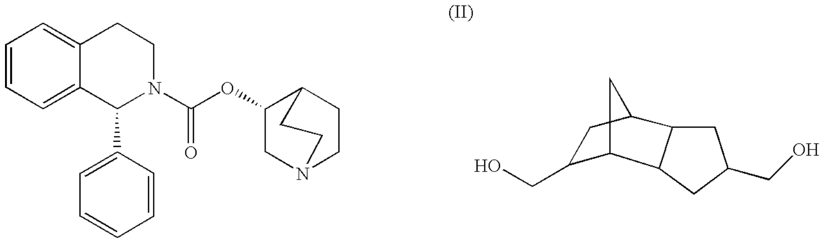


Figure 5: Example images from the USPTO set showing different drawing styles for bridged ring systems.

Assemble Superatoms

Necessary Step. It selects groups of characters and tests if they are known as superatoms. If there is a suitable synonym in the loaded superatom database it will add the appropriate bonds and atoms to the molecule. Superatoms can be extended and visualized in the molecule panel.

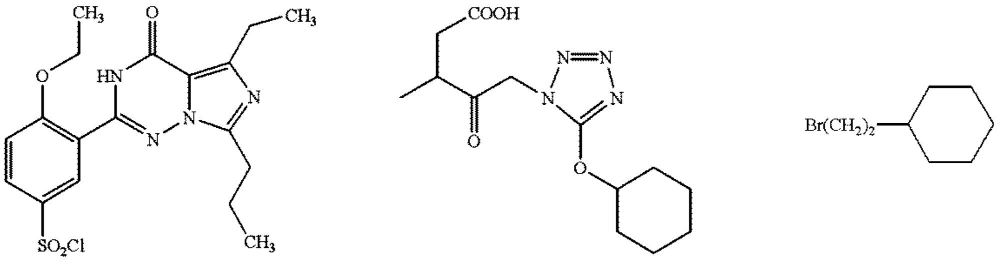


Figure 6; Example images from the USPTO set. In these images different superatoms like SO2Cl, COOH or BR(CH2)2 have to be recognized.

Assemble Reaction Symbols

Necessary Step. It looks for reaction arrows, plus and brackets. These objects will only be added to the molecule if the selected molecule file format supports them.

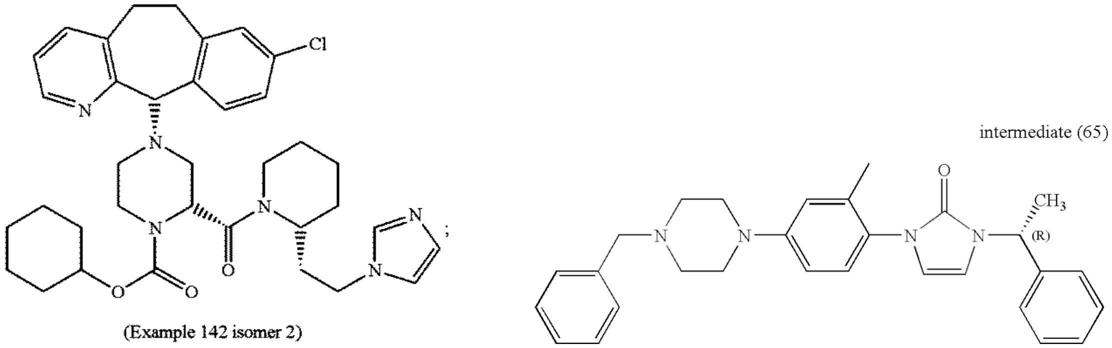


Figure 7: Example images from the USPTO set. In these images captions like Example or Intermediate have to be recognized.

Assemble Captions

Necessary Step. It selects groups of characters which can be captions of the image or the molecules. These captions will only be added to the molecule if the selected molecule file format supports them.

Assemble Charges

Necessary Step. It selects negative and positive charges and tries to add them to the appropriate atom. Delocalized charges will only be added to the molecule if the selected molecule file format supports them.

Create Molecule

Necessary Step. Finally the resulting molecule will be assembled from the different parts. The following workflow modules will modify this molecule further on.

Postprocessing Section

The following modules belong to the Postprocessing Section which will mostly start external programs; Calculate Physico-chemical Properties, Start 3D-Viewer, Start WWW Query.

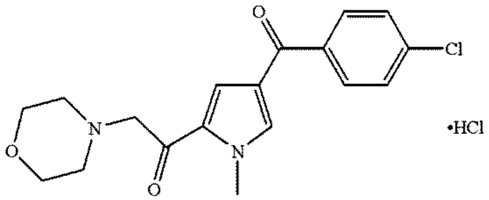


Figure 8: Example image from the USPTO set. In this image HCL is removed or written in another molecule by the split fragments or the delete small fragments module.

Delete Small Fragments

It removes small fragments from the molecule by deleting them. This will remove e.g. Methyl groups resulting from noise in the image or removes salts and counter ions.

Split Fragments

It creates from each fragment a new molecule. All generated molecules will become a collection which is written to a multi mol SDF file. The collection is shown in a new GUI in the results panel.

Validation Section

The following modules belong to the Validation Section which will help to analyze the reconstruction result.

Validate Reconstruction

It analyzes the resulting molecule for existing and possible chemical errors. This module tests valences, bond lengths and angles, typical atom types and fragments. The result is a score between 0 (certainly an error) to 1 (the reconstruction should be perfect).

All images have been processed by switching on the validation of the result and cleaning of small fragments, no image preprocessing has been used, and some letters have been trained for the OCR.

3. Problems elements in the USPTO images

Several problems within the image set that have been identified but not tackled:

Drawing defects

The input quality of the scanned image plays a great role for the image recognition. Typical problems are image noise (speckles, disrupted lines), grey level images with antialiasing, watermarks, different backgrounds and overlapping or fused objects. In the training and test sets occur only overlapping objects (cf. figure 9). All images are already binarized.

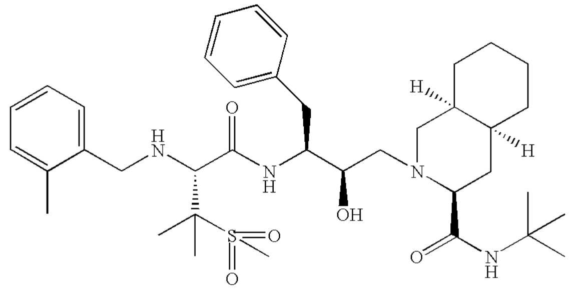


Figure 9; In this example image the S of the Sulfur is overlapping with the double bond.

Unclear semantics

For the typical bonds (like single bonds, double bonds, …) all drawing tools use similar drawing primitives (simple thin lines). Some times dotted lines, thick lines or wavy lines are used to display e.g. hydrogen bonds, metal bonds, attachment points or aromatic bonds. There is no convention in the drawing style. Some examples can be found in the next figure.

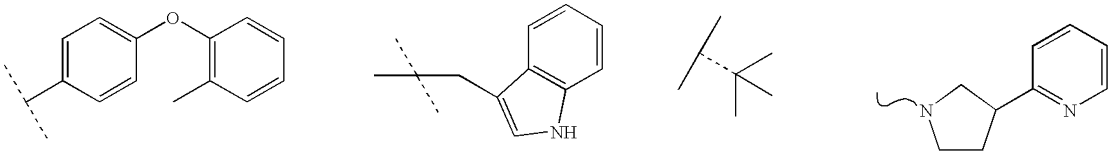


Figure 10: Example images showing dotted or wavy lines with unclear semantics.

Problematic in chemical depictions are the semantic of arrows and dots. Arrows can be interpreted as reaction arrows, highlighting parts of the molecule, attachment points, electron shifts, etc. … Dots can be interpreted as radicals, chiral carbon centers, …

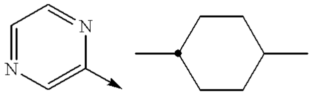


Figure 11: Example images with unclear semantic elements.

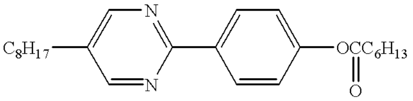


Figure 12: Example with unclear superatom elements.

Superatoms can be displayed as an acronym (e.g. Boc Butyloxycarbonyl group), a sum formula (C8H17), some kind of line notation (e.g. COOH). It is not always clear how the atoms are connected. Especially if the superatom has more than one attachment point (e.g. amino acids) the correct orientation can only be guessed. chemoCR has only a limited number of predefined superatoms, which will be replaced in an arbitrary position as long as the valences are correct.

Major problems for image reconstruction tools are Markush elements like repetition elements, variable attachment points, variable ring sizes and R-group elements. There is no standard file format that can represent all kind of Markush elements that can be found in patents. There is a first prototype of chemoCR that can identify some Markush elements. This has not been used in TREC because the I2S task was defined not to contain them.

(6)

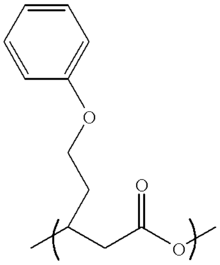


Figure 13; Images containing Markush elements like repetition units.

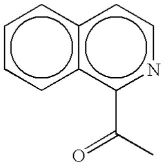


Figure 14; Examples containing aromatic ringsystems. It is unclear where to place the double bonds.

The vectorizer of chemoCR can't recognize curves and circles; it can only convert objects into straight lines. Therefore a special kind of aromatic rings cannot be detected (cf. figure 14). A rule has been added to detect the circles by looking for the letter "O" and then disambiguate it in the context of the letter. A large "O" in the center of a ringsystem can be interpreted as an aromatic ringsystem. Sending all objects to the OCR module can introduce more noise recognizing letters instead of bonds. For the I2S task the aromatic ring rule has been switched off.

4. Results

From the 1000 total unique structures 656 structures have been recalled as a perfect match.

Bibliography

[1] M.E. Algorri, M. Zimmermann, C. Friedrich, S. Akle, and M. Hofmann-Apitius. Reconstruction of chemical molecules from images. In Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society EMBC, 2007.

[2] M.E. Algorri, M. Zimmermann, and M. Hofmann-Apitius. Automatic recognition of chemical images. In Eighth Mexican International Conference on Current Trends in Computer Science (ENC 2007), pages 41–46, 2007.

[3] S. Boyer. Optical recognition of chemical graphics. In Document Analysis and Recognition, Proceedings of the Second International Conference on Publication, pages 627–631, 20–22 Oct 1993.

[4] ChemAxon Ltd., M´aramaros k¨oz 3/a, Budapest, 1037 Hungary. Marvin Developer's Guide.

[5] C. Haupt. Entwicklung und Implementierung einer Protokoll-Datenbank zur Automatisierung elementarer Funktionen von chemoCR. bachelor thesis, Fachhochschule Bonn-Rhein-Sieg, Fachbereich Informatik, 2007.

[6] P. Ibison, M. Jacquot, F. Kam, A. G. Neville, R.W. Simpson, C. Tonnelier, T. Venczel, and A.P. Johnson. Chemical literature data extraction: The CLiDE project. JCICS, 33(3):338–344, 1993.

[7] P. Kral. Chemical structure recognition via an expert system guided graph exploration. Diploma thesis in bioinformatics, Ludwig-Maximilians- University, Munich, 2007.

[8] R. McDaniel and J.R. Balmuth. Kekule: Ocr-optical chemical (structure) recognition. JCICS, 32(4):373–378, 1992.

[9] MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, California, USA. MACCS II Manual.

[10] MDL Information Systems, Inc.: 1990-1996, http://www.mdli.com. ISIS/Draw 2.1.

[11] L.T. Bui Thi. Graph-Rekonstruktion im Rahmen chemischer Strukturrepräsentationen. Diplomarbeit, Bayerische Julius-Maximilians-Universität, Würzburg, 2005.

[12] M. Zimmermann, J. Fluck, L. T. Bui Thi, C. Kolrik, K. Kumpf, and M. Hofmann. Information extraction in the life sciences: Perspectives for medicinal chemistry, pharmacology and toxicology. Current Topics in Medicinal Chemistry, 5:785–796, 2005.

[13] M. Zimmermann, C. M. Friedrich, and M. E. Algorri. Combating illiteracy in chemistry: Towards computer-based chemical structure reconstruction. In 1st German Conference on Chemoinformatics, Goslar, 2005.

[14] M. Zimmermann andM. Hofmann-Apitius. Automated extraction of chemical information from chemical structure depictions. touch briefings, 2007.

[15] M. Zimmermann, L. T. Bui Thi, and M. Hofmann. Combating illiteracy in chemistry: Towards computer-based chemical structure reconstruction. ERCIM News, 60:40–41, 2005.