bioinformatics

situation

SCAI Bioinformatics has a longtime working experience on the automated extraction of information from biomedical literature. Based on our experience in the field of biological information extraction (cf. BER¹ and ProMiner²), we recently extended the scope of our research towards chemical entity recognition together with our strategic partner the InfoChem GmbH. 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, 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 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.

To make chemical information contained in drawings of chemical structures accessible for computer programs, two projects have been started in the past: Kekulé³ and CLiDE⁴ and recently OSRA⁵. Of those three projects, only the CLiDE project led to software commercially available. However, CLiDE is a tool that is unable to learn from human expert intervention and it will repeat the same error again and again even if erroneous recognition of structures by CLiDE has been corrected by a chemist.

contact

Dr. Marc Zimmermann
Fraunhofer Institute SCAI
Schloss Birlinghoven
53754 Sankt Augustin
Germany

phone: +49 (0) 2241 - 14 - 2276
fax: +49 (0) 2241 - 14 - 2656
marc.zimmermann@scai.fraunhofer.de

www.scai.fraunhofer.de/chemocr.html

Published September 2007
In order to solve the problem of recognizing and learning chemical structures in image documents, our chemo system combines pattern recognition techniques with a chemical rule based expert system. The method is based on the idea of identifying from depictions the most significant fragments of small molecules. The workflow consists of three phases: image vectorization, chemical entity extraction and molecule reconstruction.

The main features of the software (v. 0.7) are:

- Conversion of various bitmap images (e.g. BMP, TIF, PNG) into chemical file formats (e.g. SMILES, SDF) but no full page segmentation
- 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 2)
- Training ability for the OCR process (e.g. fused letters) and teaching new superatoms
- Customization via easy 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 chemo for the multimodal information extraction of Markush structures from patents has not yet been addressed (cf. Figure 1). This functionality will be part of a future solution. At the moment we are extending the chemical rule system for new semantic entities seen in patents.

Technical Specification

The chemo 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™. 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.

Figure 2: Snapshot of the chemo graphical user interface. In the left panel the input image and the intermediate reconstruction results are shown. To the right the resulting molecule has been drawn, additional information as the mass, number of ringsystems, InChI identifiers can be computed and stored in the SDF.