

NMRShiftDB

A free information system for organic molecules and their spectral data

Christoph Steinbeck

Introduction

Identification and structure elucidation of unknown natural products is an important aspect of current fields like drug discovery, metabolomics or chemical ecology. In a process known as dereplication, a scientist would record molecular fingerprint spectra and search spectral databases to check whether the compound at hand is already known (Figure 1). Only if this search is unsuccessful it is reasonable to reach for one of the more sophisticated ab-initio tools for computer-assisted structure elucidation [1, 2].

The work described here aims to use free software, the easy access provided by the World Wide Web and the collaborative potential of the Open-Source movement to build a completely transparent structure-property database "NMRShiftDB" for storage and retrieval of small organic molecules and their NMR chemical shift data [3]. The software has reached a stable state and has successfully been operating over the last few months (<http://www.nmrshiftdb.org>). It is intended to grow into a general spectroscopic information system by extending it to store other types of spectroscopic data, like mass and infra red spectra.

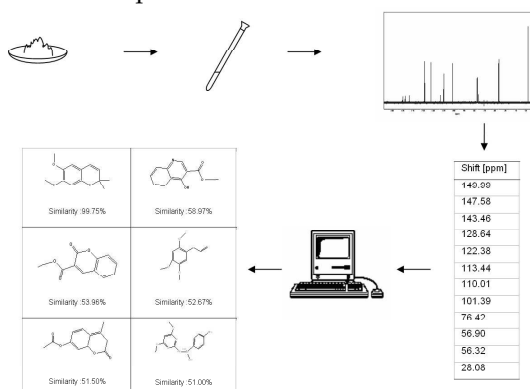


Figure 1: Dereplication as a first step in the computer assisted structure elucidation of a natural product

Features

NMRShiftDB provides the following major functionality:

- Spectra and subspectra similarity searches
- Structure-, substructure- and structural similarity searches

- Prediction of NMR spectra based on HOSE codes and the database material
- Interface for user registration and administration (necessary for logging submissions)
- Interface for peer-reviewing submitted data for quality assurance

It also offers various other, non-spectrum related search facilities, like chemical name, formula, molecular mass etc. A simplified depiction of NMRShiftDB's datastructure is shown in Figure 2. A full entity relationship (ER) diagram can be found in [3].

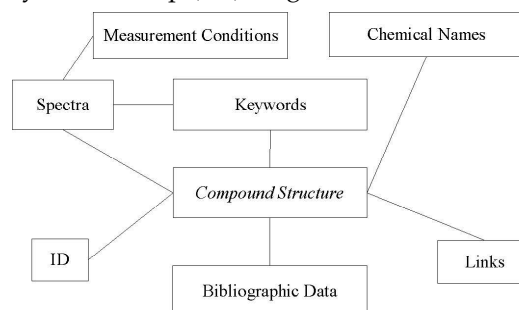


Figure 2: Simplified structure of a NMRShiftDB dataset

Looking for Collaborations

Openness of both source code and content is a fundamental principle of the NMRShiftDB. Its software is published under GNU General Public License (GPL) [4]. Database content as well as derived data fall under Open Content License (OPL) [5]. Software and data get published regularly on <http://www.sourceforge.net> and are archived there. A replication of the database by collaborating institutions is explicitly encouraged. We are envisioning an extended mirror system to achieve a high availability of the system. Currently a system of four mirrors in three different geographic locations is at work. The mirroring system will increase availability and enable participating institutions to control responsiveness of their server directly.

Since new datasets can be added by the user community in an open submission process, NMRShiftDB needs to ensure quality of its data systematically. Each submitted dataset is subjected to an automatic quality control followed by a peer review process in order to secure a uniformly good database quality. Data in the database is also checked against itself regularly.

Results

The first stable release of NMRShiftDB was released in November 2003 on <http://www.nmrshiftdb.org>. At the time of writing, the system's extent is characterized by 8239 structures, 8971 spectra (most of them ¹³C spectra, some ¹H and a few ³¹P) and almost 200 registered contributors. A standalone client is being developed to aid in collecting data locally and contributing them to NMRShiftDB later. This tool will allow for easier contribution and a more convenient assembly of private, specialized or in-house collections.

Acknowledgments

We would like to thank Dr. Willy von der Lieth, Dr. Wilhelm Hull (German Cancer Research Center (Deutsches Krebsforschungszentrum, DKFZ)) and Dr. Heinz Kolshorn (University of Mainz, Germany) for generously contributing datasets from in-house databases. Thanks also go to Dr. Thomas Kämpchen, University of Marburg, for organizing the installation of the Marburg mirror of NMRShiftDB. We are also grateful to the Max-Planck-Institute of Chemical Ecology, Jena, Germany, for hosting our main server hardware, for providing the technical support, and for constantly adding data to the database. The

NMRShiftDB project is funded by the German Research Council (Deutsche Forschungsgemeinschaft, DFG)

Dr. habil. Christoph Steinbeck
Cologne University Bioinformatics Center
c.steinbeck@uni-koeln.de

Bibliography

- [1] C. Steinbeck. The automation of natural product structure elucidation. *Current Opinion in Drug Discovery and Development*, 4(3):338–342, 2001.
- [2] C. Steinbeck. Computer-assisted structure elucidation. In Johann Gasteiger, editor, *Handbook on Chemoinformatics.*, volume 2, pages 1378–1406. Wiley-VCH, Weinheim, 2003.
- [3] C. Steinbeck, S. Kuhn, and S. Krause. NMRShiftDB - Constructing a Chemical Information System with Open Source Components. *Journal of Chemical Information and Computer Sciences*, 43(6):1733 – 1739, 2003.
- [4] The Free Software Foundation. The GNU General Public License, 1991.
- [5] The Open Content Movement. The Open Content License, 1998.

What's 2004 going to bring?

A force field, QSAR, substructure search, and much more.

by Egon Willighagen

2004

CDK is nearing its fourth anniversary (September 2004) and is growing faster each year: both the developer and user communities are getting larger each month, as well as the number of products based on the CDK library. This article tries to give an idea of what can be expected from the CDK project later this year.

Force Field

One new feature that will be added this year is a force field. The group of Christoph Steinbeck has worked on a Java implementation of the MM2 force field[1]. Such force fields are used to calculate the energy of a 3D molecular structure, and in combination with an

optimization method, it can be used to optimize the 3D geometry of that molecule. It provides a faster alternative to a more accurate quantum mechanical calculation.

Independently, a CDK plugin is being written that interfaces with Ghemical [2] using a web interface. Such interoperation between CDK and other programs and libraries is expected to show up in CDK more often in the future.

QSAR

Another field which is likely to get much more attention is quantitative structure-activity/property relationships (QSAR/QSPR). QSAR models are made by correlating molecular descriptors with activities or properties of the set of molecules being modeled. Thousands of descriptors have been proposed (see [3], and still more are proposed every day. It is expected that CDK will implement a subset of these later this year.

Very recently, a new SourceForge project has been started to address this specific field of research af-