

## 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 <sup>13</sup>C spectra, some <sup>1</sup>H and a few <sup>31</sup>P) 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](mailto: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-

ter a discussion on the `cdk-devel@lists.sf.net` mailing list: <http://qsar.sf.net/>. This project aims to bring together open source developers from many projects and develop a Java GUI program that interfaces with all aspects of QSAR(-like) research: setting up a data set, descriptor calculation, model building, up to model validation.

The CDK project is expected to contribute to this project by providing implementations of several of these components.

## SMARTS

Recently, the `UniversalIsomorphismTester` was adapted to allow for custom `Atom-Atom` and `Bond-Bond` matching. Prior to this change `Atoms` were matched based only on element symbol. As a result it was not possible to distinguish an  $sp^2$  and  $sp^3$  carbon. In addition, it was not possible to match an atom to any halogen. This shortcoming has been fixed now.

The next step is to write a SMARTS [4] parser and editor that can create `SMILESAtoms` that can match real atoms based on the given query. This subproject has been started recently, but the full query language is not implemented yet. A basic example has been implemented (see `cdk.test.isomorphism.SMARTSTest`). In this example the SMARTS query `'C=*` is used, thus a carbon double bonded to any atom. This is the source code that implements this:

```
SmilesParser sp = new SmilesParser();
AtomContainer atomContainer = sp.
    parseSmiles("CC(=O)OC(=O)C");
// acetic acid anhydride
QueryAtomContainer query =
    new QueryAtomContainer();
SMARTSAtom atom1 = new SMARTSAtom();
atom1.setLabel("*");
SMARTSAtom atom2 = new SMARTSAtom();
atom2.setSymbol("C");
query.addAtom(atom1);
query.addAtom(atom2);
```

## Literature

"Literature" is a recurrent column describing recently published articles that have in some way to do with CDK.

by Egon Willighagen

This column intends to give an overview of recently published articles that have some relation to CDK: they might describe algorithms implemented

```
query.addBond(
    new OrderQueryBond(atom1, atom2, 2)
);
boolean isSubstructure =
    UniversalIsomorphismTester.
    isSubgraph(atomContainer, query);
```

The `SMARTSAtom.match()` method only implements the `* atom` and much needs to be done before this fully works. Feel free to browse the source code in the `cdk.smiles.smarts` package.

## More

These three things are not the only ongoing development of CDK, but show three very interesting new features. People are encouraged to read the *CDK ChangeLog* which will appear in each issue. But here are some keywords: more reactions, tighter CML support, partial atomic charges and more CDK plugins.

Egon Willighagen  
University of Nijmegen, The Netherlands  
egonw@sci.kun.nl

## Bibliography

- [1] M.L Allinger. MM2. A Hydrocarbon Force Field Utilizing V1 and V2 Torsional Terms. *J. Am. Chem. Soc.*, 99, 1977.
- [2] Ghemical. <http://ghemical.sf.net/>, April 2004.
- [3] R. Todeschini and V. Consonni. *The Handbook of Molecular Descriptors*, volume 11 of *Methods and Principles in Medicinal Chemistry*. Wiley-VCH, Weinheim, Germany, 2000.
- [4] Daylight website. <http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>, April 2004.

in CDK, use CDK in research, or describe software that uses CDK. Normally, this article will discuss articles published since the previous issue, but since this is the first issue, it will describe all CDK related publications that appeared so far.

The articles will be described in the order in which they appeared, but I'll take the liberty to start with the CDK article itself.