*Published online 12 November 2008* *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D295–D299***
*doi:10.1093/nar/gkn850*

# SuperToxic: a comprehensive database of toxic compounds

**Ulrike Schmidt[1], Swantje Struck[1], Bjoern Gruening[1], Julia Hossbach[1], Ines S. Jaeger[1,2], Roza Parol[1], Ulrike Lindequist[3], Eberhard Teuscher[3] and Robert Preissner[1,*]**

[1]Structural Bioinformatics Group, Institute of Molecular Biology and Bioinformatics, Charité – University Medicine Berlin, Arnimallee 22, 14195 Berlin, [2]Department of Cardiology and Angiology, Charité - University Medicine Berlin, Schumannstr. 20/21, 10117 Berlin and [3]Institute of Pharmacy, Ernst-Moritz-Arndt-University Greifswald, Friedrich-Ludwig-Jahn-Strasse 17, 17489 Greifswald, Germany

Received August 15, 2008; Revised September 23, 2008; Accepted October 16, 2008

## ABSTRACT

**Within our everyday life, we are confronted with a variety of toxic substances of natural or artificial origin. Toxins are already used, e.g. in medicine, but there is still an increasing number of toxic compounds, representing a tremendous potential to extract new substances. Since predictive toxicology gains in importance, the careful and extensive investigation of known toxins is the basis to assess the properties of unknown substances. In order to achieve this aim, we have collected toxic compounds from literature and web sources in the database SuperToxic. The current version of this database compiles about 60 000 compounds and their structures. These molecules are classified according to their toxicity, based on more than 2 million measurements. The SuperToxic database provides a variety of search options like name, CASRN, molecular weight and measured values of toxicity. With the aid of implemented similarity searches, information about possible biological interactions can be gained. Furthermore, connections to the Protein Data Bank, UniProt and the KEGG database are available, to allow the identification of targets and those pathways, the searched compounds are involved in. This database is available online at: http://bioinformatics.charite.de/supertoxic.**

## INTRODUCTION

Toxins are hazardous substances, causing illness or damage to an exposed organism if inhaled, swallowed or absorbed through the skin. They can be found all over in nature and are widely used as drugs in medicine, as toxicity strongly depends on concentration.

In nature, animals and plants use toxic substances as protection from predators. For example, poisonous mushrooms or plants use toxins to protect themselves against herbivores. A lot of snakes, scorpions or spiders produce poison to guard themselves from other animals. A number of these substances, originally used by animals or plants to poison their enemies, have become valuable within medicine. In cancer treatment, Paclitaxel, a toxin from the Yew tree (*Taxaceae*) (1), has been applied successfully in the treatment of breast cancer. Vinorelbin, an alkaloid from *Cataranthus roseus*, shows good results in the therapy of different carcinomas (2). Very successful in the fight against infection diseases are the toxins of a variety of fungi, the antibiotics (3). These substances, originally produced from the mushroom to protect themselves against bacterial infections, depict a great breakthrough in medicine, as an impressive amount of medical conditions can now be cured.

There are different measurements to estimate toxicity: LD50 and LC50 (lethal dose or concentration at which 50% of a population dies) are widely established but also TGI (total growth inhibition), NOEL (no observable effects limit) or LOEL (lowest observable effects level) are used.

The wide use of toxins proves the scientific importance, and confronts researchers with the question for the nature of toxicity. What makes a compound toxic? How can toxicity be detected for unknown compounds? To answer these questions, a close investigation of toxic compounds is inescapable, making it necessary to provide a collection of toxins.

Databases like Mvir (4) or SCORPION (5) are excellent sources for detailed information, for example, compounds

*To whom correspondence should be addressed. Tel: +49 30 8445 1649; Fax: +49 30 8445 1551; Email: robert.preissner@charite.de

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

from *Scorpiones* or DNA and protein sequence analysis. Admittedly, these databases are restrictively applicable for complete consideration of the structural and chemical properties of toxic compounds, within science and industry.

To solve this problem, we established the database SuperToxic, which provides a comprehensive collection of toxins from different sources (animals, plants, synthetic, etc.), combined with chemical features, as well as information about commercial availability. This dataset enables a detailed investigation of the correlations between chemical, functional and structural properties of toxins.

Furthermore, these data can be used to evaluate the risk of use for compounds within medicine or industry, and give valuable insight into the mechanisms of toxicity. While certain toxins affect many types of cell lines, some toxic compounds only interfere with defined cell types leading to a specific toxicity. Cytostatic drugs, which are often used in chemotherapy, affect the cell cycle or the DNA replication mechanisms (6) and are therefore toxic to all living cells, although tumorigenic cells are more influenced due to their high rate of cell proliferation. In contrast, omeprazole, a drug for the treatment of gastric or duodenal ulcer, only affects cells in the stomach, as it needs an acidic environment to become active (7). Since it is of great interest to figure out, whether a compound interacts specifically with particular cells, SuperToxic is a distinguished tool for the search of such information. Additionally, the toxicity of an unknown substance can be estimated by comparison with structurally similar compounds with known toxicity. Another application of this database is the estimation of health hazards for a variety of chemicals, especially with respect to the new European chemical management system REACH (registration, evaluation, authorization of chemicals) (8), a regulation for the registration, evaluation and confinement of chemicals. This regulation is necessary, as in many chemical production processes, for example, fabrication of colors, varnishes or leather, the usage of toxic compounds is almost inevitable. Therefore, it is essential to assess potential hazardous effects, to safeguard such substances during transportation, usage and storage. The vast usage of chemicals and new chemical registration programs, like REACH, demands alternatives to experimental validation. All producers or importers, who introduce more than one ton per annum of a substance to the European Union, must evaluate the chemical regarding its toxicity, according to REACH. In order to reduce cost-intensive animal testing of toxic compounds, the promotion of data exchange (9–11) and predictive toxicology gains importance and acceptance (12). There are several theoretical approaches, besides QSAR (quantitative structure–activity relationship) (13), which describe the relationship between toxicity and physicochemical properties of compounds. For such predictions, high quality, comprehensive and well-structured databases are essential. The toxicity values and chemical information given by SuperToxic provide such basis for hazard assessment.

## THE DATABASE

SuperToxic comprises data from publicly available databases and scientific literature, assembling a vast amount of toxic compounds. Currently, there are about 60 000 structures with corresponding properties stored in the database. Additionally, properties like the number of hydrogen bond (H-bond) donors and acceptors, molecular weight or the octanol–water partition coefficient logP, which allow the evaluation of the Lipinski's Rule of Five (14), can be found within the database. The web interface (available at http://bioinformatics.charite.de/supertoxic) provides different options to access the data:

- Within the 'Toxin Search' option (Figure 1, I-A), it is possible to perform a distinct search, given the name of the compound, the Chemical Abstracts Services Registry Number (CASRN) or NSC number, an identifier of the National Cancer Institute database. A search for a certain pathway lists all compounds associated with this pathway.
- The 'Structure Search' option (Figure 1, I-B) allows a structure upload via InChI (IUPAC International Chemical Identifier), SMILES (Simplified Molecular Input Line Entry System) or MOL file. Additionally, the structure can be drawn, using a built-in molecule editor. The provided structure can either be used as input for a similarity screening or a substructure search.
- Another way to search the database is implemented in the 'Property Search' (Figure 1, I-C). Here, the definition of value ranges for certain attributes (e.g. the molecular weight, the logP, the number of rings, H-bond donors or acceptors) provides a list of all database entries fulfilling the conditions.
- To browse the whole database, the user can choose an alphabetic character or numbers, to display all database entries starting with the selection. Alternatively, all CASRN or NSC numbers, which are available in the database, can be listed.

To demonstrate the functionality of the database, the toxicity and the chemical characterization of a new potential drug candidate (Hit1) is exemplarily evaluated: the structure can either be uploaded as MOL file or be drawn, using the molecule editor. The similarity search results in a list of matches to the query compound ordered by similarity (Figure 1 II, column 1–6). The following information are given:

- links to resources, external to SuperToxic:
  - order information for more than 60 different suppliers (Figure 1, II-C)
  - pathway information (Figure 1, II-D)
  - ligand information
  - target information
- synonyms including CASRN and NSC number
- the empirical formula

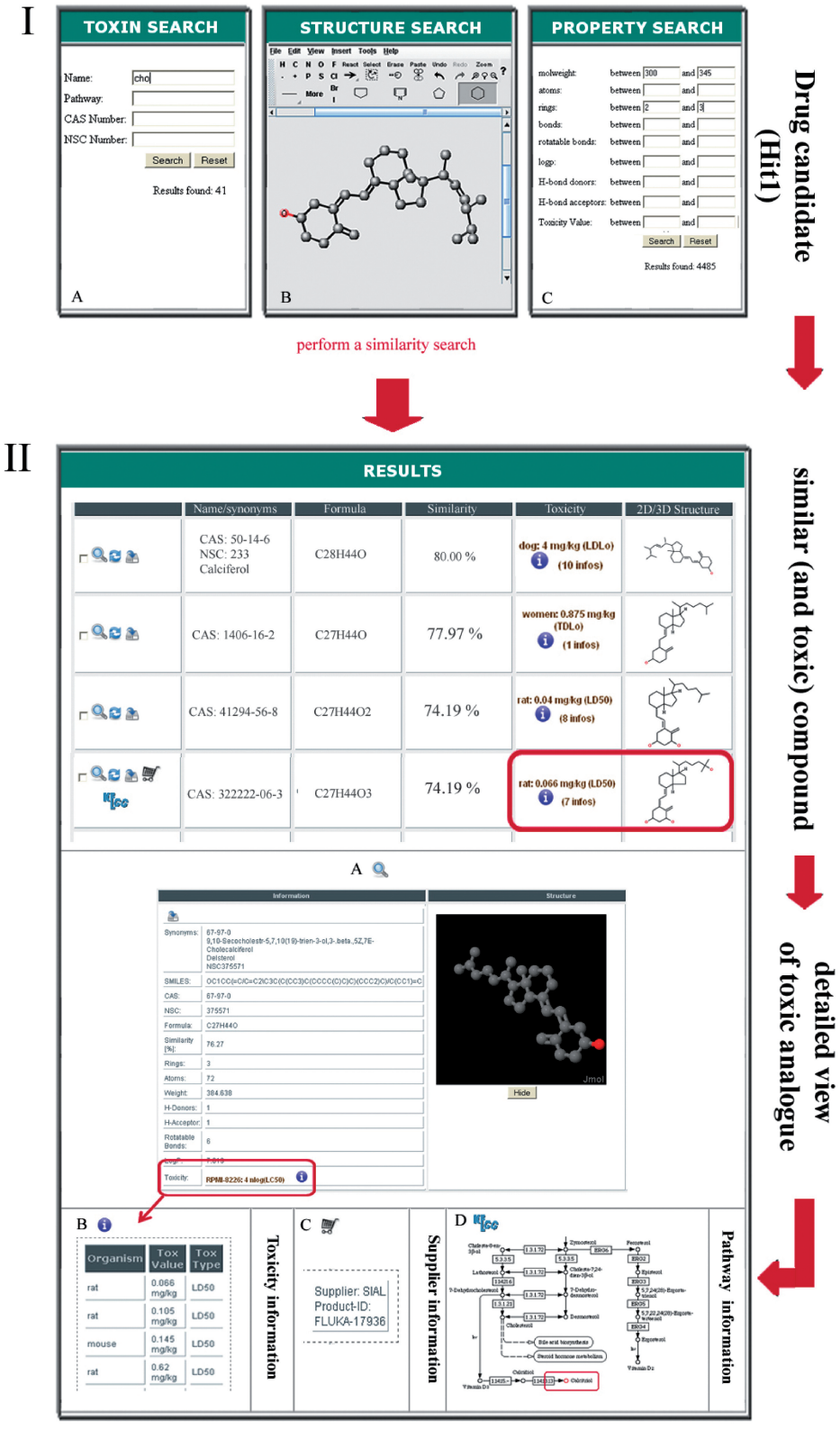


**Figure 1.** Flow chart of queries and search results in SuperToxic. (**I**) Search option of the web interface. (**A**) Search for toxic compounds via name, CASRN or NSC number and pathway. (**B**) Structure search: upload a MOL file, enter SMILES or InChI code or dray the structure, e.g. of a potential drug candidate (Hit1). Based on this query structure a similarity or substructure search can be performed. (**C**) Search for toxic compounds via properties, like molecular weight, number of atoms or rings, H-bond donors or acceptors, LogP or the toxicity value. (**II**) Result table of a similarity search (Hit1 a query structure), showing a summary for each compound. (**A**) A detailed view shows all properties. (**B**) All toxicity values for a compound. (**C**) Supplier information. (**D**) Link to KEGG pathways.

- toxicity information (Figure 1, II-B):
  - dose
  - testing type (e.g. LC50)
  - organism or cell line, the compound was tested on
- a two-dimensional structure
- a three-dimensional visualization of the structure

For each compound in the result table, a separate 'Detailed View' window (Figure 1, II-A) is provided, displaying further structural and chemical properties, such as number of rings, atoms, H-bond donors and acceptors, rotatable bonds, SMILES, logP and molecular weight. All these information can be downloaded in PDF format, together with the atomic coordinates in MOL format.

Some of the similar compounds, found for Hit1, like Alfacalcidol (Figure 1 II, red boxes), for example, are very toxic. This finding suggests that despite all the other similar compounds found, which are only slightly toxic, the candidate itself might be tested for toxicity starting with very low concentrations.

Another useful feature of SuperToxic, is the user upload interface, which allows the scientific community to contribute to the database. There are several data required for the upload: the structure, toxicological information (type of toxicity, dose, unit, organism or cell line) and an email address for further correspondence. After manual curation, the database will be updated with the new compound.

## METHODS

### Data mining

SuperToxic was established on the basis of data from the publicly available databases PubMed, DSSTox (15) and NCI60 (16). Furthermore, the book 'Biogene Gifte' (17), was manually surveyed, making the data available online for the first time. Toxicity data were collected from literature by extensive text mining. A searchable index from the PubMed database was built. In the next step, the index was filtered for toxicity-related keywords and various patterns, like units or IUPAC names. Finally, all relevant text passages were manually curated in regard to the presence of any toxicity information.

The data from all sources were merged to eliminate duplicated compounds. The database currently contains about 60 000 compounds. Also included, are references to the origin of the compounds and about 600 entries in the Kyoto Encyclopedia of Genes and Genomes (KEGG) (18). To detect potential targets in biochemical pathways approximately 400 compounds were detected in the Protein Data Bank (PDB) (19,20) and the corresponding targets were linked via more than 800 KEGG and 3600 UniProt (21) entries.

### Calculation of chemical properties

The calculations of chemical properties, e.g. molecular weight and number of H-bond donors and acceptors, were performed with functions from OpenBabel 2.1, an open source chemical toolbox (22) (http://openbabel.sourceforge.net/). To compute the properties for the structures, the MyChem extension (an implementation of the OpenBabel 2.1 library for MySQL) was used (http://mychem.sourceforge.net/).

### Analysis of chemical and structural properties

For the complete dataset, the distributions of molecular weight, LogP and H-bond donors and acceptors were analyzed, whereas drugs (23) and natural compounds (24) served as reference groups.

A reduced dataset, derived from the NCI60 panel, was subdivided into three toxicity groups, represented by $-\log$ (LC50): the slightly toxic, medium toxic and highly toxic compounds.

For each group, the distribution of chemical properties was calculated separately, to reveal interdependencies regarding the toxicity. The results are shown under 'Statistics' on the SuperToxic website.

### Structural fingerprint

The similarity search is performed, using so-called structural fingerprints, a binary string with a length of 1536 bits, which encodes for the chemical characteristics of a compound. Within this database, a combination of two fingerprints was used: (i) a 1024 bit fingerprint based on MDL; (ii) a 512 bit fingerprint encoding for 317 structural properties defined as SMARTS pattern (http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html), provided by OpenBabel (http://openbabel.org/wiki/FP2). The first one generates a fingerprint for each chemical structure, the second provides for exact structural patterns. The application of this combined fingerprint leads to an improvement of the similarity and substructure searches and yields in more detailed results. The fingerprints of all compounds were precalculated and stored in the database.

### Similarity search

During the similarity search, the fingerprint of the input structure is built and compared with the fingerprints of the database entries using the Tanimoto coefficient. It is a similarity index and defined as:

$$T = \frac{N_{ab}}{N_a + N_b - N_{ab}}$$

$N_a$ and $N_b$ describe the number of bits, set to 1 in the fingerprint, of compound $a$ and $b$, respectively. $N_{ab}$ is the number of bit positions set to 1 in both fingerprints. A molecule with a Tanimoto coefficient $\geqslant 0.85$ to an active compound is assumed to be biologically active itself (25). The Tanimoto calculations are performed using MyChem.

### Substructure search

For the substructure search, the database entries are filtered, according to the number of atoms and rings. Thus, structures with less atoms or rings, compared with the query molecule, are not considered narrowing down the search space. Afterwards, the fingerprint of the query

structure is compared with the remaining database entries' fingerprints. If all bits set to one coincide, the query structure is a substructure of this database compound.

### Server

SuperToxic is designed as a relational database, which is implemented in a MySQL server. For chemical functionality, the MyChem/OpenBabel package is added. The web access is enabled via an Apache Webserver 2.2. The web site is built in PHP5 and HTML. For the molecule drawing and uploading function, the tool MarvinSketch and for visual inspection of compounds, Jmol (26) is implemented.

## CONCLUSION AND FUTURE DIRECTIONS

SuperToxic is a rich source of toxicological data, combining structural, functional and chemical information, along with corresponding toxicity values. Features, like similarity screening and substructure search, enable to characterize and estimate the potential toxicity of substances which have not been validated yet.

The application of SuperToxic might help to reduce the amount of animal testing, e.g. for the risk assessment of new drugs, or to fulfill the new EU REACH requirements. Additionally, this database represents a valuable support during toxin research, as the information about the compounds will facilitate experimental design. The range of toxicity, possible targets, mode of action and chemical modification to lower toxicity can be retrieved.

SuperToxic is planned for further enlargement, data concerning peptides and proteins will be added, and ecotoxicological information will be considered. In addition to that, an upload function enables the scientific community to contribute by adding new compounds or supplementary information. In order to provide up-to-date information the database will be updated twice a year.

## FUNDING

Deutsche Forschungsgemeinschaft (SFB 449); IRTG Berlin-Boston-Kyoto; Investitionsbank Berlin (IBB); Deutsche Krebshilfe. Funding for the open access charge: Deutsche Forschungsgemeinschaft (SFB-449).

*Conflict of interest statement*. None declared.

## REFERENCES

1. Crown,J., O'Leary,M. and Ooi,W.S. (2004) Docetaxel and paclitaxel in the treatment of breast cancer: a review of clinical experience. *Oncologist*, **9 (Suppl. 2)**, 24–32.
2. Conti,F. and Vici,P. (1998) [Vinorelbin in the treatment of breast cancer: current status and prospectives for the future]. *Clin. Ter.*, **149**, 61–74.
3. Solomkin,J.S. and Miyagawa,C.I. (1994) Principles of antibiotic therapy. *Surg. Clin. North Am.*, **74**, 497–517.
4. Zhou,C.E., Smith,J., Lam,M., Zemla,A., Dyer,M.D. and Slezak,T. (2007) MvirDB—a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Res.*, **35**, D391–D394.
5. Srinivasan,K.N., Gopalakrishnakone,P., Tan,P.T., Chew,K.C., Cheng,B., Kini,R.M., Koh,J.L., Seah,S.H. and Brusic,V. (2002) SCORPION, a molecular database of scorpion toxins. *Toxicon*, **40**, 23–31.
6. Eckhardt,S. (2002) Recent progress in the development of anticancer agents. *Curr. Med. Chem. Anticancer Agents*, **2**, 419–439.
7. Langtry,H.D. and Wilde,M.I. (1998) Omeprazole. a review of its use in Helicobacter pylori infection, gastro-oesophageal reflux disease and peptic ulcers induced by nonsteroidal anti-inflammatory drugs. *Drugs*, **56**, 447–486.
8. Commission of the European Communities (2003) Proposal concerning the registration, evaluation, authorisation and restriction of chemicals (REACH). (COM(2003)644Final), Bruxelles, EU.
9. Wullenweber,A., Kroner,O., Kohrman,M., Maier,A., Dourson,M., Rak,A., Wexler,P. and Tomljanovic,C. (2008) Resources for global risk assessment: the International Toxicity Estimates for Risk (ITER) and Risk Information Exchange (RiskIE) databases. *Toxicol. Appl. Pharmacol.*, doi: 10.1016/j.taap.2007.12.035.
10. Fostel,J.M. (2008) Towards standards for data exchange and integration and their impact on a public database such as CEBS (Chemical Effects in Biological Systems). *Toxicol. Appl. Pharmacol.*, doi:10.1016/j.taap.2008.06.015.
11. Waters,M., Stasiewicz,S., Merrick,B.A., Tomer,K., Bushel,P., Paules,R., Stegman,N., Nehls,G., Yost,K.J., Johnson,C.H. *et al.* (2008) CEBS—Chemical Effects in Biological Systems: a public data repository integrating study design and toxicity data with microarray and proteomics data. *Nucleic Acids Res.*, **36**, D892–D900.
12. Benigni,R., Netzeva,T.I., Benfenati,E., Bossa,C., Franke,R., Helma,C., Hulzebos,E., Marchant,C., Richard,A., Woo,Y.T. *et al.* (2007) The expanding role of predictive toxicology: an update on the (Q)SAR models for mutagens and carcinogens. *J. Environ. Sci. Health C Environ. Carcinog. Ecotoxicol. Rev.*, **25**, 53–97.
13. Benigni,R. and Bossa,C. (2008) Predictivity of QSAR. *J. Chem. Inf. Model*, **48**, 971–980.
14. Lipinski,C.A., Lombardo,F., Dominy,B.W. and Feeney,P.J. (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Del. Rev.*, **23**, 3–25.
15. Richard,A.M. (2004) DSSTox update & future plans. *QSAR and ModelingSociety Newsletter*, **15**, 34–36.
16. Shoemaker,R.H. (2006) The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer*, **6**, 813–823.
17. Teuscher, E. and Lindequist, U. (1994) *Biogene Gifte*. 2. Aufl. Gustav Fischer Verlag, Stuttgart, 3. Aufl. in press, Wiss. Verlagsgesellschaft Stuttgart, Germany.
18. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
19. Deshpande,N., Addess,K.J., Bluhm,W.F., Merino-Ott,J.C., Townsend-Merino,W., Zhang,Q., Knezevich,C., Xie,L., Chen,L., Feng,Z. *et al.* (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.*, **33**, D233–D237.
20. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
21. The UniProt (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190–D195.
22. Guha,R., Howard,M.T., Hutchison,G.R., Murray-Rust,P., Rzepa,H., Steinbeck,C., Wegner,J. and Willighagen,E.L. (2006) The Blue Obelisk-interoperability in chemical informatics. *J. Chem. Inf. Model*, **46**, 991–998.
23. Goede,A., Dunkel,M., Mester,N., Frommel,C. and Preissner,R. (2005) SuperDrug: a conformational drug database. *Bioinformatics*, **21**, 1751–1753.
24. Dunkel,M., Fullbeck,M., Neumann,S. and Preissner,R. (2006) SuperNatural: a searchable database of available natural compounds. *Nucleic Acids Res.*, **34**, D678–D683.
25. Martin,Y.C., Kofron,J.L. and Traphagen,L.M. (2002) Do structurally similar molecules have similar biological activity? *J. Med. Chem.*, **45**, 4350–4358.
26. Herráez,A. (2006) Biomolecules in the computer: Jmol to the rescue. *Biochem. Mol. Biol. Educ.*, **34**, 255–261.