Data Quality in Predictive Toxicology: Identification of Chemical Structures and Calculation of Chemical Properties

Christoph Helma,[1,2,3] Stefan Kramer,[1] Bernhard Pfahringer,[4] and Eva Gottmann[2,3]

[1]Institute for Computer Science, University Freiburg, Freiburg, Germany; [2]Institute for Cancer Research, University of Vienna, Vienna, Austria; [3]Institute for Environmental Hygiene, University of Vienna, Vienna, Austria; [4]Austrian Research Institute for Artificial Intelligence, Vienna, Austria

Every technique for toxicity prediction and for the detection of structure–activity relationships relies on the accurate estimation and representation of chemical and toxicologic properties. In this paper we discuss the potential sources of errors associated with the identification of compounds, the representation of their structures, and the calculation of chemical descriptors. It is based on a case study where machine learning techniques were applied to data from noncongeneric compounds and a complex toxicologic end point (carcinogenicity). We propose methods applicable to the routine quality control of large chemical datasets, but our main intention is to raise awareness about this topic and to open a discussion about quality assurance in predictive toxicology. The accuracy and reproducibility of toxicity data will be reported in another paper. *Key words*: **carcinogenicity, knowledge discovery, machine learning, predictive toxicology, quality assurance, structure–activity relationships.** *Environ Health Perspect* **108:1029–1033 (2000). [Online 10 October 2000]**

http://ehpnet1.niehs.nih.gov/docs/2000/108p1029-1033helma/abstract.html

In industry, government, and scientific research, structure–activity relationships (SARs) play an expanding role in estimating the potential toxicity of chemicals. Traditionally, human experts use experience and expertise to identify structural features responsible for toxic action (e.g., structural alerts) (*1*). Recent developments in artificial intelligence research and the improvement of computational resources have led to efficient data mining methods that can automatically extract SARs from toxicity databases with structurally diverse (noncongeneric) compounds. They are, in our opinion, better suited for toxicologic problems than classical regression-based SAR methods because they can use structural information very efficiently and the resulting models are easier to interpret for toxicologic and chemical experts than regression models. An overview about techniques and applications of data mining in toxicology was published by Helma et al. (*2*).

The basic procedure is to submit the existing experimental data (the learning set) to a machine learning program that detects relationships between chemical structures and toxic effects. This SAR model can then be used to predict the toxicity of untested compounds (Figure 1).

An advantage of this data-driven approach in predictive toxicology is that SAR models can be derived in an unbiased way and that it is possible to make new discoveries in already existing data (*3–8*). The predictive-toxicology evaluation (PTE) projects (*8,9*) demonstrated that predictions from data-driven methods are at least as accurate as those from human experts and expert systems. They rely heavily on the quality and representation of chemical and toxicologic data in the training and testing sets.

Although the importance of high-quality learning sets is generally accepted, we found no papers in the literature that address this topic. Therefore, we suspect that, in many cases, existing datasets were used for SAR models without the performance of a detailed inspection. Our intention in this study and in the companion study (*10*) is to present a case study that indicates problematic areas. At present we are not able to give general guidelines similar to good laboratory practices, but we hope to open a discussion about quality assurance of toxicologic databases and their relevance to the development of SAR models.

In particular, we present our experience from the application of machine learning techniques to the Carcinogenic Potency Database (CPDB) (*11*), the largest publicly available database with results from long-term rodent carcinogenicity experiments. It is a secondary source for carcinogenicity data and includes detailed (species-, sex-, strain-, and organ-specific) information for 1,298 chemical agents. The CPDB includes data from the National Toxicology Program (NTP; Research Triangle Park, NC, USA), which lists the results of highly standardized long-term rodent carcinogenicity assays (393 compounds) (*12,13*). The CPDB also includes a compilation of carcinogenicity assays (1,028 compounds) from the general literature that meet a predefined set of quality criteria (*11*). A total of 1,298 compounds have been studied (there is a partial overlap between compounds studied by the NTP and those reported in the literature).

In this paper we will cover the identification of chemical structures and the estimation of chemical properties; in our upcoming report (*10*), we will discuss the reliability of toxicologic information. All data discussed in this paper [with the exception of the CPDB (*14*)] is available from the authors (*15*).

Identification of Compounds and Retrieval of Structures

In most cases, toxicity databases do not contain chemical structures. The identification and representation of the chemical structures in the toxicity database is therefore the first step towards a data set for SAR studies. The correct identification of structures is crucial because all following operations rely on it. Because the databases are usually too large for a detailed inspection of each compound, this step is less trivial than it might seem at a first glance.

Identification by Chemical Abstracts Registry numbers (CAS). Because chemical structures are not provided in most toxicity databases, it is necessary to identify each compound and to retrieve structures from external databases. In most cases, the CAS number is the only common identifier.

Although the CAS was designed to identify chemicals, it is often not an ideal identifier for toxicologic purposes. General problems (e.g., typing errors, mismatched CAS numbers, popular nomenclature, etc.) associated with CAS and nomenclature have been previously reported (*16*). We had surprisingly few problems with typing errors, as indicated by the internal Check Digit Verification of CAS Registry Numbers (*17*). In the original files, CAS numbers for 62 compounds (excluding mixtures) were missing, but we were able to assign a CAS number to a almost all compounds using search engines

Address correspondence to C. Helma, Institute for Computer Science, Machine Learning Lab, University Freiburg, Am Flughafen 17, D-79110 Freiburg/Br, Germany. Telephone: 49-761-203-8013. Fax: 49-761-203-8007 E-mail: helma@informatik.uni-freiburg.de

This project was funded by the Austrian Federal Ministry of Science and Transport (GZ 70.017/2-Pr/4/87). Partial support was also provided by the "Jubiläumsfond der Österreichischen Nationalbank" under grant 6930. The Austrian Federal Ministry of Science and Transport provides general financial support for the Austrian Research Institute for Artificial Intelligence.

Received 2 February 2000; accepted 31 May 2000.

such as ChemFinder (*18*), ChemID (*19*), and the Beilstein Database (*20*). The search failed for four compounds because we were unable to identify their structure based on their popular names in the CPDB.

The main problem with using CAS numbers as identifiers for toxicologic purposes is that toxicologically irrelevant differences (e.g., crystal water) between structures lead to different CAS numbers for otherwise similar structures. As the CAS number does not indicate structural similarities, it may be difficult to retrieve a structure if the CAS number is different in the structural database from that listed in the toxicity database. If, on the other hand, the same structure is identified by more than one CAS number, the example will gain too much weight in the machine learning process.

In the CPDB, 15 compounds were described as mixtures (instead of by a CAS number). Because an SAR relies on exactly defined chemical structures, we excluded these compounds from further searches and did not include them in our experiments. Further mixtures and compounds with impurities were detected by text searches for items such as "mixture" and "pure" in the accompanying files of the CPDB. These were also excluded from the search for structures, but many mixtures and compounds with undefined composition (e.g., polymers) were detected only at a later stage when the structures were already available or not retrievable.

Retrieval and representation of 2-dimensional structures. Due to budget restrictions, we had to search for chemical structures predominately in public sources. Table 1 summarizes the databases that we used in our work (*21–26*).

Because the structures were obtained in different formats, we had to convert them into a common format. We used SMILES strings (*27*) because they are compact and intuitive representations of two-dimensional (2-D) structures and are the most commonly accepted input format for computational chemistry programs. The majority of quality checks was performed on this representation.

The first trial with some computational chemistry programs revealed that the most common errors were associated with three types of problems.

Presence of disconnected structures. Most computational chemistry programs accept only one structure as input. It is therefore impossible to calculate chemical properties for compounds with separate entities (e.g., mixtures, organic salts, presence of H_2O, HCl, etc.). For this reason we decided to remove toxicologically irrelevant entities (e.g., multiple instances of the same entity, H_2O, HCl, etc.), represent salts with covalent bonds, and remove compounds with several larger entities (e.g., mixtures, organic quarternary amines). The positive side effect was that we were able to detect many mixtures and improperly defined compounds which were not marked as mixtures in the original CPDB. Especially in structures from the NTP database, we found that organic salts did not contain any counterions (this is probably another way to deal with the problem of disconnected structures, but in our experience, it may lead to incorrect calculations).

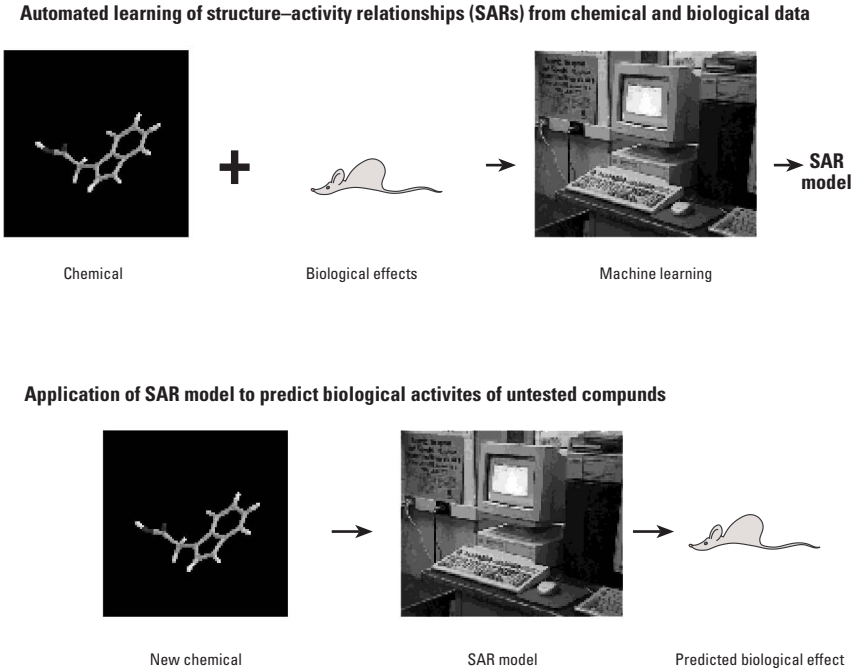

Figure 1. Visualization of the machine learning process.

Representation of charges. In many databases, structures are represented without charges. The disadvantage of this is that every program that performs valency checks will indicate wrong valences, especially in compounds with nitro and other nitrogen-containing groups. We chose to represent nitrogen-containing groups in their charged form {e.g., [N+](=O)[O-] for nitro groups} to keep the correct valences and ensure correct calculations.

Presence of hydrogens. In SMILES strings, hydrogen atoms are normally not specified explicitly. Instead their presence is inferred from normal valence assumptions. This is usually not a problem because most computational chemistry programs are designed to accept SMILES strings as input and attach hydrogens automatically. Problems arise however when SMILES are transformed into another representation using, for example, BABEL (*28*), which does not attach hydrogens, or CORINA (*29*), in which the output of hydrogens has to be requested explicitly. If incomplete molecular structure is used as input for a quantum mechanical program like MOPAC (*30*), the results of the calculation are, of course, useless. This mistake tends to remain undetected because many visualization programs do not render hydrogens and chemists are accustomed to viewing chemical structures without hydrogens.

After the correction of the most common systematic errors, we tested our structural database with seven different computational chemistry programs [KOWWIN (*31,32*), BABEL (*28*), TSAR (*33*), CSBR (*34*), DEPICT (*35*), and CORINA (*29*)], recorded their error messages, and changed the structures accordingly. We made additional corrections after browsing through the structures manually. Finally, all structures were converted to unique (canonical) SMILES (*36*) to detect and remove duplettes and then submitted to the test programs for a final check. This examination indicated that all of the structures were syntactically correct, and remaining error messages were caused by limitations of the test programs. We recorded all changes of the original structures in logfiles for further examination and used revision control extensively.

For many chemists this approach may seem to be too pragmatic and may seem that too little effort was used to represent the "correct" structure of each compound, but in projects with large, noncongeneric learning datasets and limited *a priori* knowledge of toxic mechanisms, it is impossible to estimate the biologically active structure. At present, SAR models for complex toxicologic effects are still to a large extent "black box models" because of the complexity of molecular

mechanisms involved. It is therefore very important to aim for a consistent representation of chemical structures. Ideally, the representation should be accepted by computational chemistry programs, give accurate results (as determined by a comparison with experimentally derived values, where available), and enable the automated detection of structural features (e.g., functional groups, structural alerts).

Calculation of Chemical Descriptors

In the development and application of SARs it is essential to identify the structural or chemical properties that predict the end point of interest. Presently, the choice of structural and property descriptors for complex toxicologic effects relies strongly on the intuition of the individual researcher, but the majority of SAR systems are limited to certain types of descriptors and allow little or no experimentation with new descriptors. We use programs based on inductive logic programming for our machine learning experiments. Within this framework it is possible to choose chemical descriptors very flexibly and to use propositional data [data that can be represented in tabular form (e.g., properties of a whole molecule); more precisely, each example is represented as a single tuple in a single relation] as well as relational data [data that cannot be represented in tabular form (e.g., chemical structures); more precisely, each example is represented as a set of tuples in multiple relations].

Figure 2 depicts the procedure we currently employ for the calculation of chemical descriptors. All format conversions were done by BABEL (*28*).

Representation of the two-dimensional structure. For the machine learning programs used in our experiments, we converted the two-dimensional structures into Prolog facts (*38*) using a Prolog program written by B. Pfahringer. To check the plausibility of the conversions we used another program that translates Prolog facts back to SMILES. We converted original and reverse-translated SMILES to unique SMILES and compared them.

Because chemists are accustomed to thinking in terms of functional groups composing a molecule, we are currently developing a higher level representation of molecules based on functional groups (*39*). We are continuing this study and, at present, we have no automated procedure to check the representation. Therefore, we still rely on the manual inspection of randomly selected examples (*40*).

Estimation of lipophilicity. The octanol–water partition coefficient is a physical property used extensively to describe the lipophilic or hydrophobic properties of a chemical. It is the ratio of the concentration of a chemical in the octanol phase to its concentration in the aqueous phase of a two-phase system at equilibrium. Because measured values range between at least 12 orders of magnitude, the logarithm (logP) is commonly used to characterize its value.

We used KOWWIN (*31,32*) to calculate the logP because a recent comparison of lipophilicity (*41*) algorithms showed that it is accurate and suitable for different types of structures. For 13,058 compounds with reliable, experimental logP values, KOWWIN estimates had a standard deviation of 0.436, an absolute mean error (absolute deviation) of 0.316, and a correlation coefficient (r^2) of 0.95.

When we applied KOWWIN on our dataset, we found that the representation of structures had a large influence on the outcome of the calculations. We made the majority of adjustments described in the previous sections (removal of disconnected structures, representation of charges, etc.) to ensure correct logP values.

Calculation of three-dimensional structures. It is generally accepted that the interaction of chemicals with biological macromolecules is to a large extent determined by their three-dimensional shape. In SAR studies with noncongeneric compound this aspect is usually neglected, partially because it is time consuming to optimize three-dimensional (3-D) structures, partially because 3-D information is difficult to use when there is no common substructure for an alignment of molecules (e.g., as in comparative molecular field analysis). However, an optimized 3-D structure is also the prerequisite for the reliable estimation of electronic properties indicating chemical reactivity.

Initial 3-D structures were calculated by CORINA (*29,42*), a rule-based system for the generation of molecular geometries. For flexible molecules, up to three geometries were generated to provide different starting points for the following calculations. CORINA was able to provide 3-D structures for all CPDB compounds with SMILES structures.

The PM3 algorithm implemented in MOPAC (*30*) was used to calculate the final 3-D structures. More than 1,500 structures (multiple conformations for flexible molecules) were optimized on a 200MHz Dual PentiumPro running Linux 2.2.5-15smp (Red Hat, Inc., Research Triangle Park, NC, USA). Unfortunately, MOPAC is not designed for the batch processing of such a large number of compounds, and considerable time was spent identifying the best conditions for batch processing. We recommend using the PM3 algorithm [more elements parameterized than in AM1 (*30*)] and performing the optimization in cartesian coordinates (key word: XYZ) instead of polar coordinates. With this setting we were able to obtain geometries for 1,498 (98%) of 1,535 initial structures.

Semiempirical molecular orbital methods (such as PM3) require that for each element some quantities are estimated by fitting experimental data. At present, not all elements are parameterized; therefore, calculations for 36 structures failed if the structure contained one

Table 1. Summary of the CPDB.

Database[a]	No. of compounds	Format	Remarks
NTP special reports[b]	464	MDL Molfiles[c]	Only structures from the NTP program, WWW searchable
NCI database[d]	237,771	SMILES[e]	Compressed text file
ChemFinder[f]	> 75,000[g]	ChemDraw Binaries,[h] usable only on Windows and Macintosh platforms	WWW searchable by names, CAS numbers, and structures
Beilstein Crossfire[i]	> 7,000,000[g]	MDL Molfiles[c]	Commercial, hardly usable for batch searching, clients for Windows or Macintosh platforms only

[a]Additional structural databases are available from ChemID*plus* (*21*). [b]Data available from the NTP (*22*). [c]Description available from CTfile Formats (*23*). [d]Data available from the National Cancer Institute (*24*). [e]Description available from the SMILES Home Page (*25*). [f]Data available from ChemFinder (*26*). [g]Structures are not always available. [h]No specifications are available. [i]Data available from Beilstein (*20*).

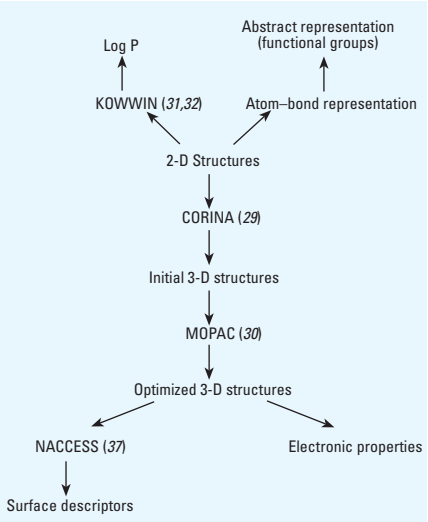

Figure 2. Calculation of chemical descriptors.

of the following elements: boron, calcium, titanium, vanadium, chromium, manganese, iron, nickel, copper, zirconium, niobium, barium, tungsten, or osmium. The geometry optimization was performed by an iterative minimization of the molecule's energy until a self-consistent field was achieved. In eight cases it was necessary to override geometry safety checks with the key word GEO-OK; these calculations should be treated with caution. One compound was too large to be treated by our version of MOPAC.

We obtained the following molecular properties from MOPAC output files:

- Dipole (asymmetricity of charge distribution)
- Electronic energy (potential energy of the electrons in a molecule)
- Electronegativity (tendency to attract electrons)
- Heat of formation (energy difference between the molecule and the elements in their standard state)
- HOMO (energy of the highest occupied molecular orbital, indicates ability to donate electron)
- HOMO–LUMO [(energy difference between the HOMO and the lowest unoccupied molecular orbital (LUMO)]
- Hybridization dipole (dipole contribution of hybridized bonds)
- Ionization potential (energy needed to remove an electron)
- LUMO (indicates the ability to accept electrons)
- Largest interatomic distance (largest distance between atoms, indicates molecular size)
- Molecular weight
- Point-charge dipole (dipole contribution of atomic charges)
- Total energy (potential energy of all possible interactions between electrons and nuclei of the molecule).

These properties are to a large extent descriptors for the electronic nature of the molecules, indicating their reactivity for biological macromolecules. Because we are not aware of a database with a sufficient number of experimentally derived properties (and some properties such as HOMO and LUMO are not experimentally observable), we were not able to verify these calculations. To reduce the run time of these calculations, we are currently comparing the MOPAC results with calculations from PETRA (*43*), a program based on empirical algorithms. To account for steric effects, we used NACCESS (*37*) to extract solvent-accessible surfaces (total, nonpolar, polar) from the final geometries. These calculations were successful for all compounds.

The assessment of 3-D structures and electronic properties remains an open problem for our data set because experimentally measured values are not available for sufficient compounds. At this point the only chance to judge the performance of different algorithms for the optimization of 3-D structures and chemical properties is to compare the resulting SAR models, but this procedure is indirect and suffers from the influence of other variables (e.g., learning algorithms).

Conclusion

We are currently far from defining a general procedure for the quality assurance of chemical data in SAR studies. As a starting point for further discussions, we include some recommendations for the retrieval of structures from external databases and the calculation of chemical descriptors:

- For each structural database it is necessary to determine if the representation of the structures is suited for the particular problem
- In a first trial, all structures should be submitted to the computational chemistry programs used in the particular study to determine *a*) if they are accepted and *b*) if they give reasonable results (calculated properties should be compared to experimental data when available)
- All encountered problems should be logged to detect systematic problems within the database and the applied algorithms
- Based on this experience, rules for the consistent representation of chemical structures should be formulated, documented, and applied to the database
- All changes to the original data should be documented (applying, for example, a version control system) to make errors traceable and allow later revisions
- Finally, the plausibility of each database and of calculated results should be checked by human experts by drawing randomly selected samples based on statistical criteria (*40*).

After summarizing our experiences with the quality assurance of chemical data in predictive toxicology, we conclude that the currently available databases and computational chemistry programs are too faulty to be trusted without further inspection. The development of reliable quality control procedures definitely needs more discussion, exchange of experience, and research activity. In this sense, we hope that we will raise some awareness in regard to data quality issues and quality assurance in predictive toxicology.

References and Notes

1. Ashby J, Paton D. The influence of chemical structure on the extent and sites of carcinogenesis for 522 rodent carcinogens and 55 different human carcinogen exposures. Mutat Res 286:3–74 (1993).
2. Helma C, Gottmann E, Kramer S. Knowledge discovery and data mining in toxicology. Statistical Methods in Med Res (in press).
3. Klopman G, Rosenkranz HS. Approaches to SAR in carcinogenesis and mutagenesis. Prediction of carcinogenicity/mutagenicity using MULTI-CASE. Mutat Res 305:33–46 (1994).
4. King R, Sternberg M, Srinivasan A. Relating chemical activity to structure: an examination of ILP successes. N Generation Comput 13:411–433 (1995).
5. King RD, Srinivasan A. Prediction of rodent carcinogenicity bioassays from molecular structure using inductive logic programming. Environ Health Perspect 104(suppl 5):1031–1040 (1996).
6. Fayyad U, Uthurusamy R. Data mining and knowledge discovery in databases. Commun ACM 39(11):24–26 (1996).
7. Kramer S, Pfahringer B, Helma C. Mining for causes of cancer: machine learning experiments at various levels of detail. In: Proceedings of the Third International Conference on Knowledge Discovery and Data Mining. (KDD-97). Menlo Park, CA:AAAI Press, 1997.
8. Srinivasan A, King RD, Bristol DW. An assessment of submissions made to the predictive toxicology evaluation challenge. In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence. (IJCAI-99). San Francisco, CA:Morgan Kaufmann, 1999;270–275.
9. Bristol DW, Wachsman JT, Greenwell A. The NIEHS predictive-toxicology evaluation project. Environ Health Perspect 104(suppl 5):1001–1010 (1996).
10. Gottmann E, Kramer S, Pfahringer B, Helma C. Unpublished data.
11. Gold LS, Zeiger E. Handbook of Carcinogenic Potency and Genotoxicity Databases. Cleveland, OH:CRC Press, 1997.
12. Selkirk JK, Soward SM, eds. Compendium of abstracts from long-term cancer studies reported by the National Toxicology Program of the National Institute of Environmental Health Sciences from 1976 to 1992. Environ Health Perspect 101(suppl 1):1–281 (1993).
13. Results from NTP studies. Available: http://ntp-server.niehs.nih.gov/main_pages/NTP_ALL_STDY_PG.html [cited 28 January 2000].
14. CPDB Homepage. Available: http:/potency.berkeley.edu/cpdb.html [cited 28 January 2000].
15. Available: ftp://helma.informatik.uni-freiburg.de/pub/cpdb/ [updated 18 May 2000].
16. Chemical errors found on WWW sites. Available: http://www.chemfinder.com/errorsfound.asp [cited 28 January 2000].
17. Check Digit Verification of CAS Registry Numbers. Available: http://www.cas.org/EO/checkdig.html [cited 28 January 2000].
18. ChemFinder. Available: http://www.chemfinder.com/ [cited 28 January 2000].
19. ChemID. Available: http://igm.nlm.nih.gov/ [cited 28 January 2000]
20. Beilstein Database. Available: http://www.beilstein.com/ [cited 28 January 2000].
21. ChemID*plus* Chemical Search Input Page. Available: http://chem.sis.nlm.nih.gov/chemidplus/ [cited 28 January 2000].
22. Search Chemical Structures from NTP TR's. Available: http://ntp-db.niehs.nih.gov/Main_Pages/pub-Structures.html [cited 28 January 2000].
23. CTfile Formats. Available: http://www.mdli.com/downloads/literature/ctfile.pdf [cited 28 January 2000].
24. Smiles Strings for NCI structures. Available: http://dtp.nci.nih.gov/docs/3d_database/structural_information/smiles_strings.html [cited 20 January 1999].
25. SMILES Home Page. Available: http://www.daylight.com/dayhtml/smiles/ [cited 28 January 2000].
26. ChemFinder.Com Database and Internet Searching. Available: http://chemfinder.camsoft.com [cited 28 January 2000].
27. Weininger D. SMILES, a chemical language and information system 1. Introduction and encoding rules. J Chem Inf Comput Sci 28:31–36 (1988).
28. Babel - A Molecular Structure Information Interchange Hub. Available: http://www.chem.ohiou.edu/~dolata/babel.html [cited 28 January 2000].
29. CORINA. Fast and Efficient Generation of High-Quality 3D Molecular Models. Available: http://www2.ccc.uni-erlangen.de/software/corina/ [cited 28 January 2000].
30. MOPAC 7. Available: ftp://esca.atomki.hu/mopac7/LINUX/ [cited 28 January 2000].
31. Meylan WM, Howard PH. Atom/fragment contribution method for estimating octanol-water partition coefficients. J Pharm Sci 84:83–92 (1995).
32. SRC Logkow Program. Available: http://

esc_plaza.syrres.com/interkow/logkow.htm [cited 28 January 2000].
33. TSAR. Available: http://www.oxmol.com/software/tsar/ [cited 28 January 2000].
34. CACTVS System Homepate. Available: http://www2.chemie.uni-erlangen.de/software/cactvs/index.html [cited 28 January 2000].
35. DEPICT. Available: http://www.daylight.comdaycgi/depict [cited 28 January 2000].
36. Weininger D, Weininger A, Weininger JL. SMILES. 2. Algorithm for generation of unique SMILES notation. J Chem Inf Comput Sci 29:97–101 (1989).
37. Hubbard S.J, Thornton JM. NACCESS. Computer Program, 1993. London: University College, 1993.
38 Sterling L, Shapiro E. The Art of Prolog: Advanced Programming Techniques. Cambridge, MA:MIT Press, 1986.
39. Pfahringer B, Gottmann E, Helma C, Kramer S. Efficiency/representational issues in toxicological knowledge discovery. In: Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools. Menlo Park, CA: AAAI Press,1999;86–89
40. Grant EL, Leavenworth RS. Statistical Quality Control. New York:McGraw Hill, 1996.
41. Mannhold R, Dross K. Calculation procedures for molecular lipophilicity: a comparative study. Quant Struct-Act Relat 15:403–409 (1996).
42. Gasteiger J, Rudolph C, Sadowski J. Automatic generation of 3D-atomic coordinates for organic molecules. Tetrahedron Comp Method 3:537–547 (1990).
43. Gasteiger J. Empirical methods for the calculation of physicochemical data of organic compounds. In: Physical Property Prediction in Organic Chemistry. Heidelberg:Springer Verlag, 1988;119–138.