Chimia 52 (1998) 673–677

ISSN 0009–4293

QSAR Analysis through the World-Wide Web

Peter Ertl*

Abstract. A WWW-based system for structure-activity analysis developed and used in *Novartis Crop Protection* in Basel is presented. The system enables easy calculation of important hydrophobic, electronic, and steric molecular properties, as well as interactive QSAR analysis. Based on generated models, predictions can be made with regard to the biological activities and environmental characteristics of new, as yet unsynthesized molecules. By bringing sophisticated and easy-to-use tools enabling analysis of structure-activity relationships directly to the desk of synthetic organic chemists, the system supports the efficient design of new, active, and environmentally acceptable agrochemicals.

1. Introduction

Quantitative structure-activity analysis (QSAR) introduced by *Hansch* and *Fujita* [1][2] proved to be one of the major breakthroughs in the understanding of the relationships between the physicochemical properties of molecules and their biological activity. This methodology is based on an assumption that bioactivity for a series of congeneric molecules may be expressed as a function of hydrophobic, electronic, and steric molecular characteristics. During a QSAR analysis, such functions are identified by various sophisticated statistical techniques. A calculated model allows factors to be identified which are important for biological activity, and supports the rational design of new structures with improved potency.

Since its introduction more than three decades ago, the application of the *Hansch-Fujita* QSAR analysis has been described in more than 5000 scientific papers. Numerous industrial applications have led to the design of several commercial drugs and pesticides. Despite the advent of other approaches, such as 3D QSAR or pharmacophore modelling, classical QSAR still remains one of the main working tools in areas such as design of pesticides, where information on the 3D structure of target sites is sparse, as well as in material design and environmental chemistry.

Several commercial tools are available enabling scientists to perform sophisticated QSAR analyses. These programs, however, usually run on UNIX workstations, and their application requires a sound knowledge of computer software. Together with their complex interfaces and the need for a relatively long training period, this has meant that, in most chemical companies, such tools are used mainly by specialized molecular modellers. This situation is not very satisfactory. Synthetic chemists should be involved much more in the direct QSAR work, since their project-specific knowledge is crucial for a creative structural design. Although most are quite interested in doing so, they tend to be put off by factors such as the need to remember UNIX commands, to master the complicated interface and command set for the modelling application, as well as other support programs (such as quantum chemical packages).

Recently, however, a possible solution to this problem has emerged – namely, the World-Wide Web. The enormous popularity of Web technology is due to its three great advantages – platform independence, ease of use, and a high degree of interactivity. On a company or university network hosting various types of computers with different operating systems, the possibility of connecting up all these machines in a user-friendly way is very important. And the Web provides this possibility. Web-based tools are very easy to interact with, since all use the same, simple interface. And finally, various newly emerging technologies such as Java, sophisticated Web scripting, VRML (Virtual Reality Modelling Language), or chemical markup language [3] have added new functionality to the Web and made it a fully dynamic environment which is ideal for the development of user-friendly chemical applications [4][5].

*Correspondence: Dr. P. Ertl
Novartis Crop Protection AG
Lead Discovery
WRO-1060.7.20
CH–4002 Basel
E-Mail: peter.ertl@cp.novartis.com

2. WWW-Based Molecular Modelling System

In *Novartis Crop Protection* in Basel, we have been using Web technology to deliver powerful and easy-to-use modelling tools directly to the desk of synthetic organic chemists since 1995. A Web-based chemical information and molecular modelling system [6] developed in-house and currently used by more than 200 chemists enables:

- easy retrieval of molecules from the corporate database,
- creation and editing of molecules by using a Web molecular editor written in Java,
- sophisticated visualization of molecules and their surface properties,
- automatic generation of hydrophobic, electronic, and steric molecular descriptors,
- interface to quantum chemical calculations and visualization of results,
- interactive QSAR analysis,
- molecular and substituent similarity searches.

The system utilizes a client-server Web architecture. Users interact with it through their Web browsers (clients) which are installed mainly on *IBM* PC-compatible computers. All the 'heavy processing' is done far away from these desktop machines, on a *Silicon Graphics Origin 200* server. Java applets (small graphic programs incorporated directly into the Web page) are used for the interactive manipulation of molecules.

The core of the system has been outlined in our previous paper [6]. In the current article, the modules for calculating molecular physicochemical properties and for the structure-activity analysis are described. Both these tools assist *Novartis* chemists in the design of new, potent, and at the same time environmentally acceptable agrochemicals.

2.1. *Calculation of Molecular Properties*

In QSAR analysis, hydrophobic, electronic, and steric descriptors are used to quantify the properties of molecules under study. The Web system enables easy calculations of these properties to be made

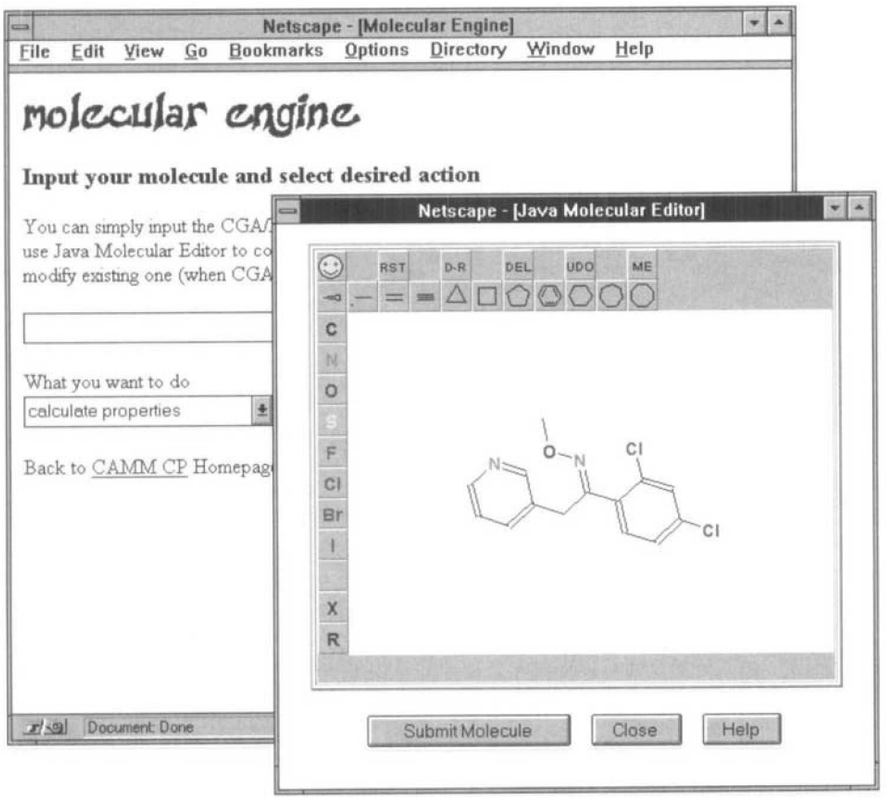


Fig. 1. *Input of a molecule with help of Java editor*

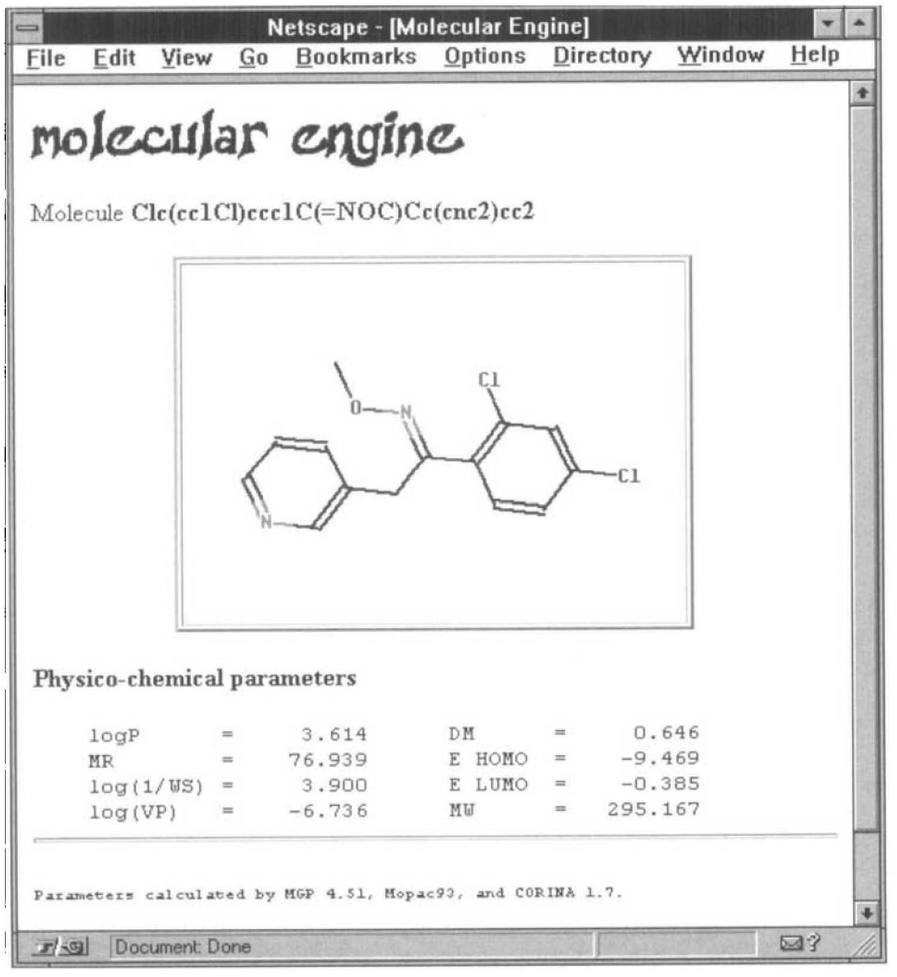


Fig. 2. *Calculation of molecular physicochemical properties*

for single molecules, as well as batch calculations for the whole set of molecules.

Hydrophobic properties determine the ability of a molecule to be transported in the environment and in an organism to interact with biological membranes, and to bind to a receptor by hydrophobic forces. Hydrophobic properties which are calculated by our system are:

- logP (logarithm of the octanol-water partition coefficient),
- MR – molar refractivity,
- log(1/WS) – water solubility,
- log(VP) – vapor pressure.

In-house programs based on published theories [7–10] are used for the calculation of hydrophobic parameters.

Electronic properties characterize the electronic distribution within the molecule. They account for the ability of a molecule to react, as well as for the electronic interaction with a receptor. Electronic properties are calculated by the AM1 [11] semiempirical quantum chemical method. Available electronic properties are:

- dipole moment,
- energy of the HOMO (highest occupied molecular orbital),
- energy of the LUMO (lowest unoccupied molecular orbital).

Molecules for which data are to be calculated may be entered into the system simply in the form of their company test number, or SMILES string [12], or drawn with the help of a simple molecular editor written in Java [13] (*Fig. 1*). Once a job is submitted, a relatively complex chain of processes is started. Various programs are launched, including the CORINA 3D builder [14] which creates a 3D molecular geometry, the quantum chemical Mopac93 package [15] which calculates electronic parameters, and several in-house programs which calculate hydrophobic properties. Thanks to the client-server Web architecture, the end-user does not see anything of this complicated processing, which takes place far from his/her PC on the central 'number crunching' server. All these programs run automatically, and use the appropriate default parameters set by expert modellers. Despite the complex processing, response time is short, and the results are delivered within 4–5 seconds (*Fig. 2*).

Since the calculation of parameters for a series of molecules one-by-one would not be very convenient, it is also possible to submit a whole list of molecules for processing at once. In addition to the standard descriptors mentioned above, other parameters, such as various surface or steric properties, are also calculated. The

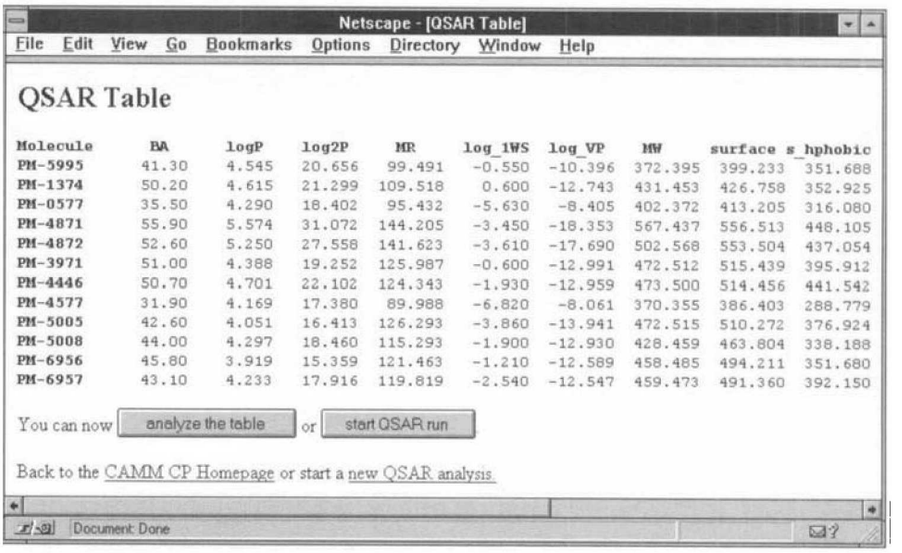
resulting QSAR table (*Fig. 3*) may be examined graphically by using an interactive Java applet, or may be submitted for further statistical processing.

2.2. *Calculation of Substituent Parameters*

For data sets consisting of molecules which have the same core skeleton and differ only in substituent pattern, the molecular properties which are used in the structure-activity analysis are usually characterized by various experimentally determined substituent constants. In the actual QSAR calculations, these constants are extracted manually from various data tables. This approach, however, has numerous disadvantages, most notably an unavailability of data for many important functional groups and a low quality of parameters for uncommon substituents. Our Web-based system solves this problem by enabling an easy interactive calculation of important substituent parameters for any organic functional group [16]. Hydrophobic properties are represented by the estimated octanol-water partition coefficient and the molar refractivity [7][8]. The electron-donating and -withdrawing power of substituents is characterized by parameters compatible with the *Hammett* σ constants. These are calculated according to the method developed in-house [17] from simple quantum chemical parameters. To illustrate the quality of data generated in this way, a plot of calculated *vs.* experimental σ constants for 77 common organic substituents is shown in *Fig. 4*. Steric substituent properties are represented simply by the topological size (number of nonhydrogen atoms) and maximum topological length of the substituent.

Users interact with the system through the simple Web interface. In the entry part of the program, the substituent for which data should be calculated is created with the help of our molecular editor, the calculation is launched, and after a short processing time, the results are displayed (*Fig. 5*).

The processing engine behind this module may also be called up directly (without the graphic interface) just by referencing to the Web address of the processing script with the SMILES code of the substituent as a parameter. In this way, it is possible to calculate data for a large number of substituents in 'batch' mode. By means of this technique, *e.g.*, a database of more than 80000 functional groups with calculated properties has been generated, which is used in the design of targeted combinatorial libraries.



Molecule	BA	logP	log2P	MR	log_1WS	log_VP	MW	surface	s_hphobic
PM-5995	41.30	4.545	20.656	99.491	-0.550	-10.396	372.395	399.233	351.688
PM-1374	50.20	4.615	21.299	109.518	0.600	-12.743	431.453	426.758	352.925
PM-0577	35.50	4.290	18.402	95.432	-5.630	-8.405	402.372	413.205	316.080
PM-4871	55.90	5.574	31.072	144.205	-3.450	-18.353	567.437	556.513	448.105
PM-4872	52.60	5.250	27.558	141.623	-3.610	-17.690	502.568	553.504	437.054
PM-3971	51.00	4.388	19.252	125.987	-0.600	-12.991	472.512	515.439	395.912
PM-4446	50.70	4.701	22.102	124.343	-1.930	-12.959	473.500	514.456	441.542
PM-4577	31.90	4.169	17.380	89.988	-6.820	-8.061	370.355	386.403	288.779
PM-5005	42.60	4.051	16.413	126.293	-3.860	-13.941	472.515	510.272	376.924
PM-5008	44.00	4.297	18.460	115.293	-1.900	-12.930	428.459	463.804	338.188
PM-6956	45.80	3.919	15.359	121.463	-1.210	-12.589	458.485	494.211	351.680
PM-6957	43.10	4.233	17.916	119.819	-2.540	-12.547	459.473	491.360	392.150

Fig. 3. *QSAR table with calculated properties*

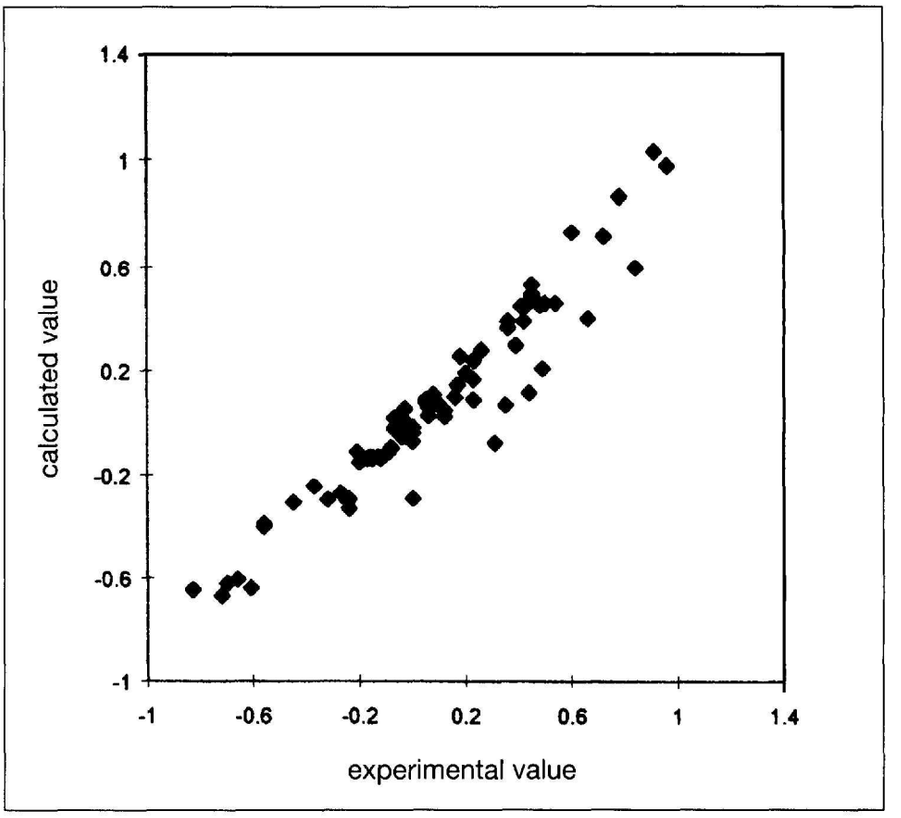


Fig. 4. *Calculated* vs. *experimental* σ *constants for 77 common substituents*

2.3. *Visualization of Surface Properties*

The examination of various properties displayed on the molecular surface, such as the electrostatic potential (MEP) or the lipophilicity potential (MLP) [18], can provide useful hints concerning the influence of various factors on the biological activity. The MEP identifies those parts of the molecule which act as preferred target sites for electrophilic attack, or most favorable hydrogen-bond donors and acceptors (*Fig. 6*). The surface lipophilicity potential can reveal those parts of the molecule which are involved in hydrophobic interactions with a receptor. Particularly useful is a visual comparison of potentials for series of similar molecules (using the best-known image processing machine, namely the human brain). By this approach, it is possible to see the 3D effects, which would not be possible to identify just by standard methods.

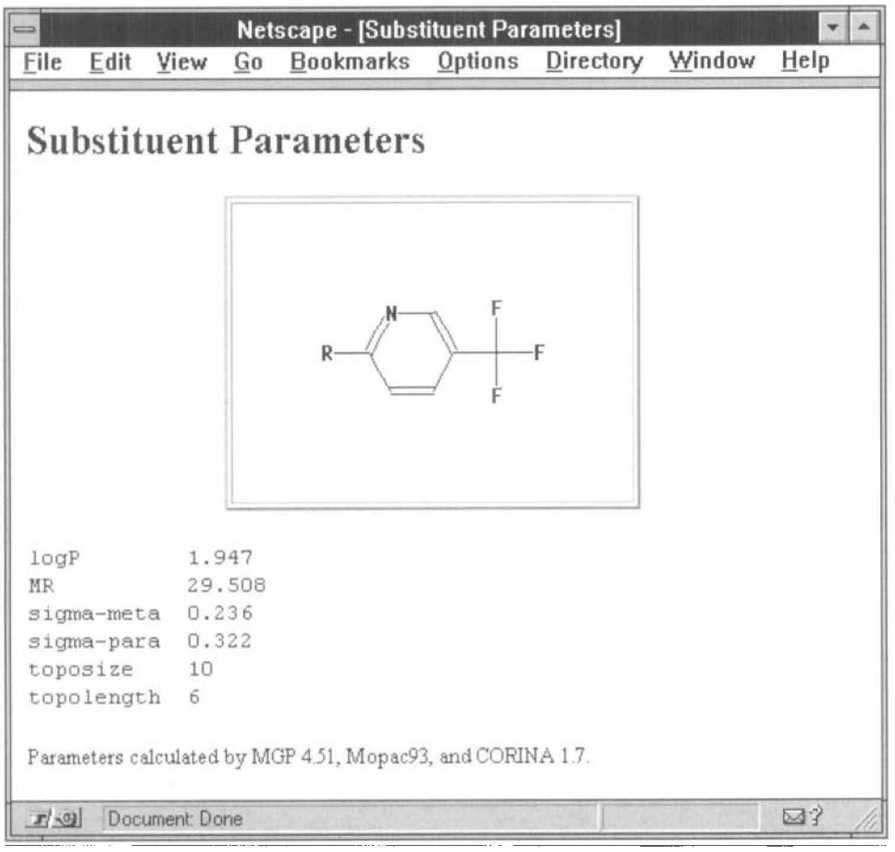


Fig. 5. *Calculation of substituent parameters*

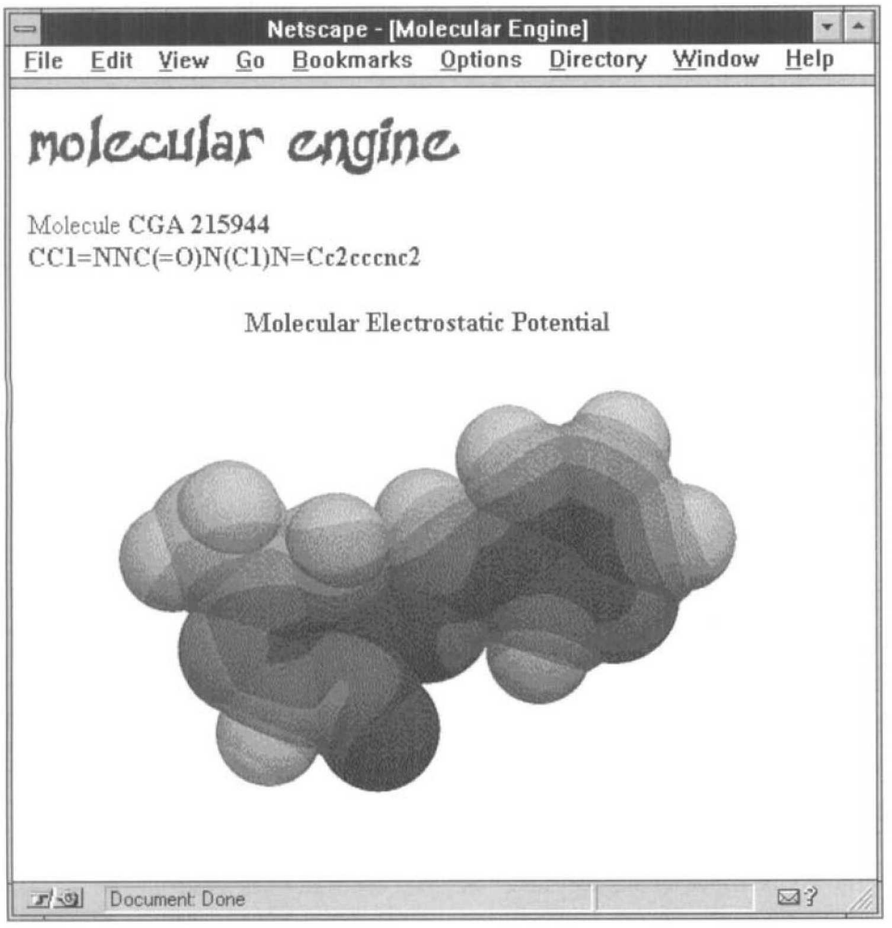


Fig. 6. *Display of molecular surface properties*

2.4. *Tools for QSAR Analysis*

The goal of QSAR analysis is to find some functional relationship between biological activity and various molecular descriptors. The generation of a descriptor table has been described in the previous sections. In a QSAR module, these parameters are correlated with the dependent variables supplied by the user, most often biological activities, but possibly also other data, such as drug penetration or toxicity, may be used.

An expert system automatically identifies the best statistical method to be used. For smaller data sets, all possible combinations of up to three parameters are checked systematically to find the equation which best explains the variation in biological activities. For larger data sets, where the systematic approach would be too time-consuming, the best equations are identified by a procedure based on a genetic algorithm [19]. The QSAR module is optimized not to find the best fit (often with only limited predictive power), but the most robust equations with good predictivity. Selection of the 'best' model is therefore based exclusively on the cross-validation procedure, and the number of parameters in an equation is strictly limited to prevent overfitting.

The resulting equations (*Fig. 7*) may be viewed and examined graphically (quality of fit, identification of outliers) using the interactive Java applet (*Fig. 8*). The model may be used for the prediction of target parameters for new, as yet unsynthesized molecules.

3. Conclusions

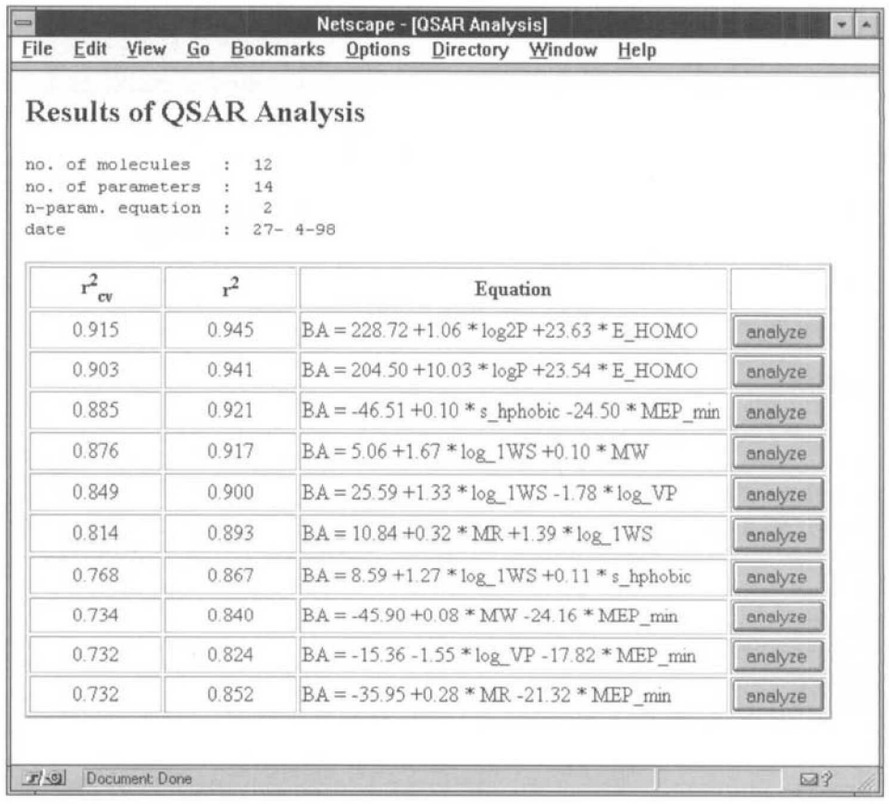
The Web-based system for structure-activity analyses described here has been used in the Crop Protection Sector of *Novartis* for about three years. Thanks to the user- friendliness of Web technology, it was possible to introduce it without any special training. The acceptance of the system is very good, and it is currently accessed by more than 200 users, mostly bench chemists, not only from Basel headquarters, but also from other *Novartis* research sites. Most of the computational modules have been developed in-house, which assures easy maintenance and constant upgradability. Other not inconsiderable advantages are zero license costs and no limitations concerning the number of users. The only commercial programs used are the CORINA 3D geometry builder [14] and the Mopac93 quantum chemical package [15].

The Web-based system for analyzing structure-activity relationships contributes to the goal of the Crop Protection Sector of *Novartis* – namely, to develop innovative products which deliver efficient and effective solutions to agricultural production problems with an optimum benefit/risk ratio.

I thank Dr. *D. Poppinger* and Dr. *T. Maetzke* for critical reading of the manuscript and for helpful comments.

Received: September 11, 1998

[1] C. Hansch, A. Leo, 'Exploring QSAR, Fundamentals and Applications in Chemistry and Biology', ACS, Washington DC, 1995.
[2] C. Hansch, T. Fujita, Eds., 'Classical and Three-Dimensional QSAR in Agrochemistry', ACS, Washington DC, 1995, pp. 13–44.
[3] P. Murray-Rust (http://www.vsms.nottingham.ac.uk/vsms/talks).
[4] S.M. Bachrach, P. Murray-Rust, H.S. Rzepa, B.J. Whitaker, Network Science, 1996 (http://www.awod.com/netsci/Issues/Mar96/feature4.html).
[5] W.-D. Ihlenfeldt, J. Gasteiger, *Chemie in Unserer Zeit* **1995**, *29*, 249.
[6] P. Ertl, O. Jacob, *THEOCHEM* **1997**, *419*, 113; see also P. Ertl, http://www.elsevier.com/homepage/saa/eccc3/paper6
[7] A.K. Ghose, A. Pritchett, G.M. Crippen, *J. Comput. Chem.* **1988**, *9*, 80.
[8] V.N. Viswanadhan, A.K. Ghose, G.R. Revankar, R.K. Robins, *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163.
[9] K. Wakita, M. Yoshimoto, S. Miyamato, H. Watanabe, *Chem. Pharm. Bull.* **1986**, *34*, 4663.
[10] W.J. Lyman, W.F. Reehl, D.H.R. Rosenblath, Eds., 'Handbook of Chemical Property Estimation Methods', ACS, Washington DC, 1990.
[11] M.J.S. Dewar, E.G. Zoebisch, A.F. Healy, J.J.P. Stewart, *J. Am. Chem. Soc.* **1985**, *107*, 3902.
[12] D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.
[13] JavaMolecularEditor© applet developed in *Novartis* enables easy construction and modification of organic molecules directly within the HTML page and generation of SMILES of the processed molecule. Editor is freely available for noncommercial use. Interested parties should contact the author.
[14] J. Sadowski, J. Gasteiger, *Chemical Reviews* **1993**, *93*, 2567.
[15] MOPAC 93, J.J.P. Stewart, *Fujitsu Ltd.*, Tokyo, Japan, 1993. Available from Quantum Chemistry Program Exchange, University of Indiana, Bloomington, Indiana.
[16] P. Ertl, *J. Mol. Graphics. Modell.* **1998**, *16*, 11.
[17] P. Ertl, *Quant. Struct.-Act. Relat.* **1997**, *16*, 377.
[18] F. Croizet, M.H. Langlois, J.P. Dubost, P. Braquet, E. Audry, P. Dallet, J.C. Colleter, *J. Mol. Graphics* **1990**, *8*, 153.
[19] D. Rogers, A.J. Hopfinger, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854.



r^2_{cv}	r^2	Equation	
0.915	0.945	BA = 228.72 +1.06 * log2P +23.63 * E_HOMO	analyze
0.903	0.941	BA = 204.50 +10.03 * logP +23.54 * E_HOMO	analyze
0.885	0.921	BA = -46.51 +0.10 * s_hphobic -24.50 * MEP_min	analyze
0.876	0.917	BA = 5.06 +1.67 * log_1WS +0.10 * MW	analyze
0.849	0.900	BA = 25.59 +1.33 * log_1WS -1.78 * log_VP	analyze
0.814	0.893	BA = 10.84 +0.32 * MR +1.39 * log_1WS	analyze
0.768	0.867	BA = 8.59 +1.27 * log_1WS +0.11 * s_hphobic	analyze
0.734	0.840	BA = -45.90 +0.08 * MW -24.16 * MEP_min	analyze
0.732	0.824	BA = -15.36 -1.55 * log_VP -17.82 * MEP_min	analyze
0.732	0.852	BA = -35.95 +0.28 * MR -21.32 * MEP_min	analyze

Fig. 7. *Structure-activity equations generated by the QSAR module*

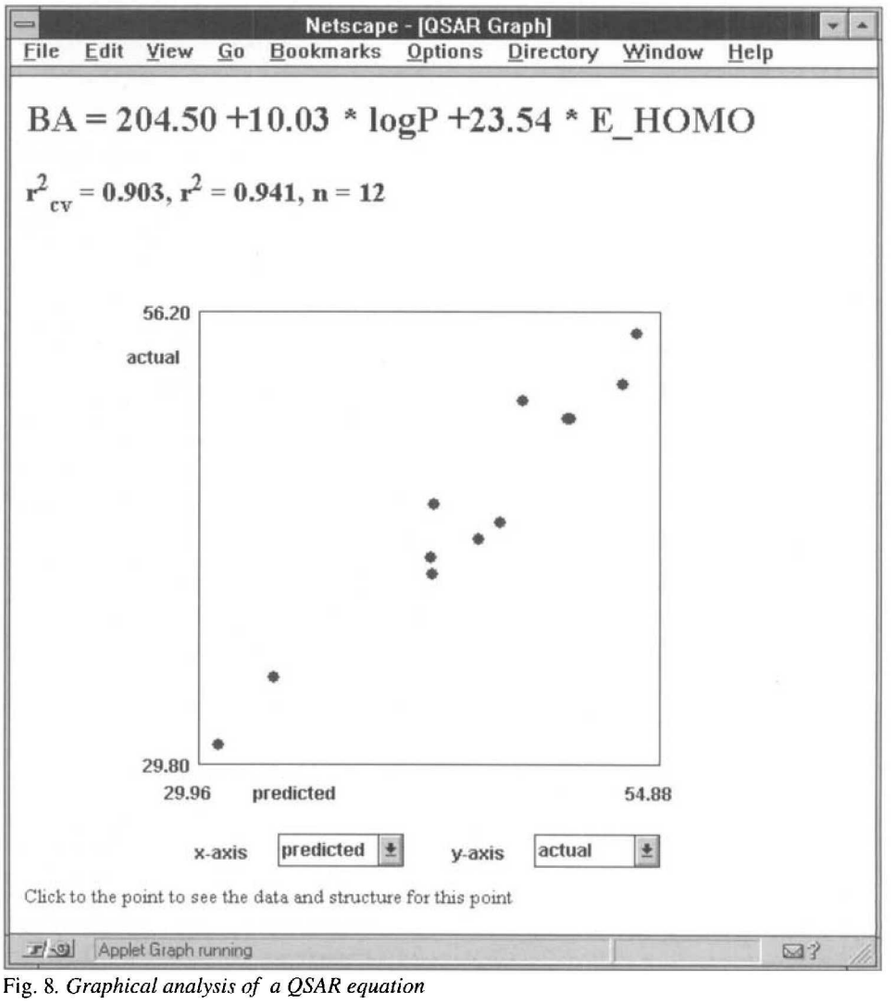


Fig. 8. *Graphical analysis of a QSAR equation*