ELSEVIER

Journal of Molecular Structure (Theochem) 622 (2003) 53–62

THEO
CHEM

www.elsevier.com/locate/theochem

Rule interpreter: a chemical language for structure-based screening

Stoyan Karabunarliev[a,b,*], Nina Nikolova[c], Nikolay Nikolov[d], Ovanes Mekenyan[a]

[a]*Laboratory of Mathematical Chemistry, University 'Assen Zlatarov', 8010 Bourgas, Bulgaria*
[b]*Department of Chemistry, University of Houston, Houston, TX 77204-5003, USA*
[c]*Laboratory of Parallel and Distributed Processing, Bulgarian Academy of Sciences, 1756 Sofia, Bulgaria*
[d]*Center For Biomedical Engineering, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria*

Abstract

A chemical language for definition and use of logical rules in screening of chemical databases is described. The rules are based on user-defined screens, which combine substructure matching with constraints on molecular descriptors, stereochemical configurations and mutual 3D placements of chemical groups. Screens are written in extended SMILES notation with the option to define variant chemical groups and constraints in a single entry. Rules are Boolean logic expressions comprised of screens and preceding rules. Arbitrary decision trees can be constructed by using nested and conditional statements referring to the rules defined. The language was used in a database-integrated QSAR expert system for aquatic toxicity, which exploits the concept of toxicochemical analogues. Another example of its usage addresses the prediction of androgen receptor binding affinity.
© 2002 Elsevier Science B.V. All rights reserved.

Keywords: Rule description language; Rule interpreter; Database screening; Substructure search; Stereoelectronic constraints; Stereochemical screens

1. Introduction

In the last decades considerable efforts have been invested in methods for preliminary assessment of toxicological hazards from chemical structure. Especially challenging is the development of structure-activity relationships (SARs) to screen large data sets of diverse chemical structures for toxicological activity in a technically sound manner. Two SAR approaches can be outlined in this field [1–4]. The first one is focused on the toxicodynamics of biological interactions and addresses toxicochemical differentiation of chemicals. It uses pattern recognition techniques to identify these common features in molecular and electronic structure, which result in a similar toxic action. The approach typically operates with toxicophores—the chemical groups responsible for specific mechanisms of action—or their steroelectronic images. Once noncongeneric chemicals are toxicochemically differentiated, a second approach could be used to further assess quantitatively the toxic potencies within groups of chemicals with a common mechanism of action. This approach, loosely named correlative SAR, typically accounts for the toxicokinetic factors. Here various molecular descriptors, ranging from measurable properties to quantum-chemical quantities, are used to explain the quantitative variation of a given kind of toxic potency.

[*] Corresponding author. Address: Laboratory of Mathematical Chemistry, University 'Assen Zlatarov', 8010 Bourgas, Bulgaria.
E-mail addresses: karabunarliev@uh.edu (S. Karabunarliev), omekenya@btu.bg (O. Mekenyan).

0166-1280/03/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0166-1280(02)00617-6

There is no strict borderline between these two approaches in terms of objectives or techniques. For instance, multi-criteria pattern recognition SAR has been used for quantitative assessment of ligand binding strengths [5]. On the other hand, it has been shown that electrophilic mode of toxic action is elicited in aquatic organisms in pronounced correlation with certain quantitative molecular descriptors [6–8]. The two problem and approaches have been viewed as complementary when assessing the environmental hazards of industrial chemicals [4]. Whereas pattern recognition techniques are used to classify the chemicals in terms of likely mechanism of action, more accurate quantitative assessments of potencies by correlative SARs are found adequate for some of the resulting toxicochemical classes. Computer software has proved indispensable for implementation of these two methods, particularly when large quantities of chemical and toxicity data must be encompassed.

Receptor-site mapping models are typical for highly specific molecular interactions and assume 3D complementarity between ligands and receptor. Rather than advancing from an explicit model of receptor-site structure, such models merely focus on the molecular shape or electrostatic potential of the test chemicals. Thereby various techniques are used to implicitly reflect the uniqueness of the receptor, e.g. the Comparative Molecular Field Analysis (CoMFA) by Cramer et al. [9], the GRID method by Goldstein et al. [10], and the active analogue approach by Marshall [11]. Recently we used the so-called COmmon REactivity PAttern (COREPA) approach to develop rules for the propensity of chemicals to bind to hormone receptors [5,12]. Operating with a limited number of stereoelectronic descriptors, the method is robust enough to serve for screening of very large chemical databases.

In order to advance and use pattern recognition SARs for chemical screening, we developed a language for formulation and application of rules, which are selective to chemical, 3D and electronic structure in parallel. Provided that toxicochemical knowledge is at hand, the language allows the definition of complex screening criteria. We used it to improve and particularize such criteria by operating with large chemical datasets. Herein we report the syntax and structure of the language. The software, which supports simultaneously SAR-development environment, database screening, and endpoint assessment, is outlined next. We finally summarize two SAR studies, which were made possible by the computer methodology described.

2. General description

The Rule Interpreter (RI) is a computer program for development and execution of rule scripts. Rule scripts are written in the so-called Rule Description Language (RDL). They define various mechanistic rules and implement branched decision schemes based on them. By means of the RI, rule scripts operate on chemicals as represented in data files. Currently, RI supports two chemical file formats, SMILES and CMP. In the former case chemical structure is encoded in SMILES notation developed by Weininger [13]. Conventional SMILES provides basic description of molecules in terms of chemical graphs. Such a representation is often called two-dimensional (2D), although no coordinates are included at all. For SMILES files, the rule program script is restricted to that subset of RDL, which addresses 2D chemical structure only. The CMP file format encompasses richer chemical information and was designed within the OASIS computer system [14,15]. Chemicals are represented in separate logical records of the CMP file. Apart from molecular connectivity, CMP records reflect 3D molecular and electronic structure obtained from quantum-chemical molecular-orbital (MO) computations. Relevant physical-chemical and toxicological data from tests or assessments are also included when available. The numerical values are structured in designated data fields within each record. Hereafter, we call all these quantities *descriptors* regardless of their particular type or origin. Descriptors will be, however, classified into *molecular* and *site-specific*, depending on whether they pertain to the whole molecule or to a distinct part of it, e.g. atomic site, covalent bond, and functional group. On output, the RI produces a copy of the input file plus the results given either in free text format (SMILES files) or internal binary representation (CMP files). It can for instance assign to each chemical from the file an integer number, indicating the putative toxicochemical class it belongs to. According to the script, the RI also calculates or selects certain descriptors, which are deemed toxicity-relevant within the assigned toxicochemical class. Finally, the RI may by itself generate toxic potency estimates, either for tested or untested chemicals.

3. The rule description language

The RDL combines SMILES chemical language with some extensions and structures borrowed from programming languages. A formal description of RDL is given in Appendix A. The RDL script is divided into three parts called freely 'define', 'rule' and 'apply' sections. The first two sections contain various definitions, whereas the last section can be viewed as the body of the rule program.

3.1. Define section

This first section contains definitions of *screens*. Basically, screens are written in SMILES and function as queries for substructure search in the chemical graphs. For instance, the screen 'c1ccccc1N=(O)=O' implies the presence of a nitrobenzene moiety in the chemical. Screens including *qualifiers* impose additional requirements on the molecular fragment(s) that are matched on 2D structural level. Qualifiers are substrings enclosed in curly brackets within a simple screen definition. Like descriptors, qualifiers can be site-specific when associated with distinct atoms or covalent bonds. Such qualifiers succeed the corresponding atom or bond entries within the SMILES notation. An atom or bond entry may have several successive qualifiers. Reserved qualifiers are used to denote for a given atom its ionic state, hybridization, participation in rings, number of adjacent protons, and chiral parity. For instance 'c1c{H}c(C{sp3})c{H}cc1N=(O)=O' specifies the presence of a *para*-nitrotoluene moiety and excludes any *ortho*-substituents. Similarly, reserved bond qualifiers may ask for certain bond configuration, e.g. *cis* or *trans* for a double bond. The screen 'c1ccccc1–C = {t}C–c1ccccc1' means, for instance, *trans*-stilbene. A general qualifier type, called hereafter descriptor qualifier, imposes restraints on numeric descriptors. The descriptor qualifier contains the descriptor name and, optionally, the acceptable numeric limits for the actual descriptor value. Numeric constraints appear in the forms of a numeric range, an upper limit, or a lower limit. The descriptor referred by a qualifier can be molecular or site-specific. In the latter case, the descriptor pertains to this particular atom or bond entry, which precedes the qualifier. The placement of qualifiers pertaining to molecular (or global) descriptors within the screen definition is immaterial. Some examples of site-specific descriptors used in conjunction with qualifiers are given hereafter. The screen 'c1ccccc1N{H2}{$-0.3 < q$}' implies an aniline moiety with the charge on the aniline nitrogen $q > -0.3$ a.u. Screens can be very general. For instance, 'C{ar}{e_lumo < 0.5}' will select chemicals with at least one π-conjugated C-atom and energy of LUMO (e_lumo) lower than -0.5 eV.

Apart from the permanently stored molecular or site-specific descriptors, which can be addressed by qualifiers, reserved names are used for some variable (or dynamic) descriptors that depend on both placement and screen context. In particular, a reserved descriptor name 'enumerate' denotes the number of times a simple screen is encountered. Upon detection of such a descriptor, the program counts how many times the preceding screen is encountered in the molecule in question, and the number obtained is checked against the limits specified. For instance, the screen 'C{H}=O{1 < enumerate}' will select only chemicals with more than one aldehyde function.

A string can define only a joint chemical subgraph in terms of standard SMILES. A screen may, however, ask for two or more disjoint fragments, that must be all present in molecule. This is achieved through the use of the delimiter '_' that formally joins two or more simple (or component) screens into one. Component screens are combined on a logical and basis with regard to substructure matching. A chemical matches a *joint screen*, if it matches all the component screens, and the corresponding fragments are not overlapping or directly bound. Thus, in terms of SMILES, the delimiter '_' is rather a label for the absence of a chemical bond. However, being interpreted as a formal SMILES bond entry, it may carry dynamic descriptor qualifiers of a special type. Namely, reserved descriptor names are used to denote the geometric distance between the disjoint molecular fragments that match the component screens. A descriptor named 'distance' denotes the distance between their geometric centers according to the current, fixed 3D molecular structure. The screen 'O{H}_{1.5 < distance < 2.6}C{sp2}=O' implies the presence of hydroxy and carbonyl groups within the 1.5–2.6 Å distance range. Another qualifier with the reserved descriptor name 'tweak' invokes a directed conformational search aiming to render the distance between the fragments into the specified limits. The 'tweak' procedure is of practical importance when screening relatively flexible molecules that are represented by a single 3D conformation in the data file. The conformational space explored is restricted to the torsional degrees of freedom of noncyclic single bonds. Conformations with non-bonded contacts closer than van der Waals radii are avoided. Among the several search methods implemented we use most frequently the linear, steepest-descend method of Hurst [16], which is not rigorous but efficient enough for screening of large chemicals inventories.

Several screens delimited by comas are united in a group on a logical or basis with respect to substructure matching and form *composite screen*. A chemical complies to a composite screen, if it complies to at least one of its component screens. Any screen, simple, joint, or composite is assigned to a *screen identifier*. Once assigned, screen identifiers can serve as entries in succeeding screen definitions. The level of implicit nestling of screen identifiers in a screen definition is unlimited. In terms of SMILES syntax, predefined screen identifiers can be viewed as valid atom entries. Practically, when parsing a screen definition for the next atom entry, the longest substring that coincides with a predefined screen identifier is first sought. If no such identifier is found, the substring is considered an explicit atom entry and checked against standard atom labels. Unlike usual program variables, screen identifiers can be assigned only once. The excerpt from a program script, which is given below, illustrates composite screens and the use of screen identifiers. Text enclosed in "('..')" is perceived as comments.

Xhalo:F,Cl,Br,I ('composite screen Xhalo comprised of halogen atoms')

Benz:c1ccccc1 ('screen for chemicals containing a benzene ring')

Polyarom:Benz_C{ar} ('chemicals with a benzene ring and another π-conjugated moiety')

Cunsat:C{sp1},C{sp2},C{ar} ('unsaturated C-site; the component screen C{ar} is redundant since C{ar} is a subset of C{sp2}')

ABunsat:Cunsat—C{sp3}{acy}-Xhalo ('α,β-unsaturated halides; the qualifier {acy} requires additionally that the halo-carbon does not participate in a ring'}

PolyarSide:Benz_C{sp2}{sc} ('chemicals with a benzene ring and another sp^2 carbon belonging to a side chain')

3.2. Rule section

This section contains definitions of *rules*. Rules handle separate chemicals from the data file and play the role of Boolean logic functions whose result is a logical *true* or *false*. A simple rule is a screen identifier enclosed in quotation marks. It returns a logical *true* result for those chemicals that match the screen. Rules are organized in expressions by means of the standard logical operators *and*, *or*, and *not*. Enclosing brackets are used to specify a priority of operations other than the standard one. Any simple rule or rule expression is assigned to a rule identifier, which, on its turn, can participate in the succeeding rule expressions. Below is given an example, which defines the rule 'Rbenz'.

Rbenz: 'Benz' and not 'Polyarom' ('any benzene ring containing chemical that has no other π-conjugated moiety')

3.3. Apply section

This section implements the actual decision scheme and can be viewed as the rule program's body. Herein, screens and rules defined in previous sections are given implicitly by their identifiers. The data for a chemical added or modified on output is addressed by molecular descriptor names. The decision scheme is constructed of *statements* that can be of three types: *assignment*, *conditional*, and *compound*. The assignment statement is like the one of programming languages. On the left hand stands the name of the molecular descriptor to be assigned; on the right hand is the value. The value is either given by a numeric constant, or implicitly specified by a screen identifier. Integer constants are typically used for toxicochemical class assignment, but they may also quantify, for instance, the probabilities of certain biological interactions. Screen identifiers in assignment statements, in difference, serve to store in dedicated fields toxicophore-specific descriptor data for the chemical. This feature is of practical importance when such descriptors are found or expected to correlate with toxic potencies. Thus, a screen identifier carries a numeric value only under the following circumstances. First, the chemical complies with the screen. Second, the screen contains at least one descriptor qualifier. In such a case, the screen identifier returns the actual descriptor value of the descriptor qualifier in the matching screen. For instance, the aforementioned screen 'C{H}=O{1 < enumerate}' returns the number of aldehyde groups, if more than one. Depending on qualifier context, screens may denote the lowest/highest among several descriptor values. Thus the screen 'c1ccccc1{0 < accept_dlc}' will select the highest acceptor delocalizably value (accept_dlc) for any of the benzene carbons, rather than the first encountered one. Similarly, 'C{sp2}{q < 10}' will return the charge of the most electronegative sp^2 hybridized C-site.

Several statements can be united in a single one. Statements, which are delimited by semicolons and enclosed in the keywords **begin** and **end**, form a compound statement. The statements within a compound statement are executed successively. The whole apply section is virtually a compound statement that has a starting keyword **apply** instead of **begin**.

Conditional statements furnish the branching of the decision tree. They have the form: **if** *rule* **then** *statement1* **else** *statement2*. *Rule* is the clause of branching given by means of a rule identifier, whereas *statement1* and *statement2* are statements of any type, including compound ones. The execution flow is passed either to *statement1* or *statement2* depending on whether *rule* is true or false for the current chemical. A conditional statement can be truncated by omitting the '**else** *statement2*' part.

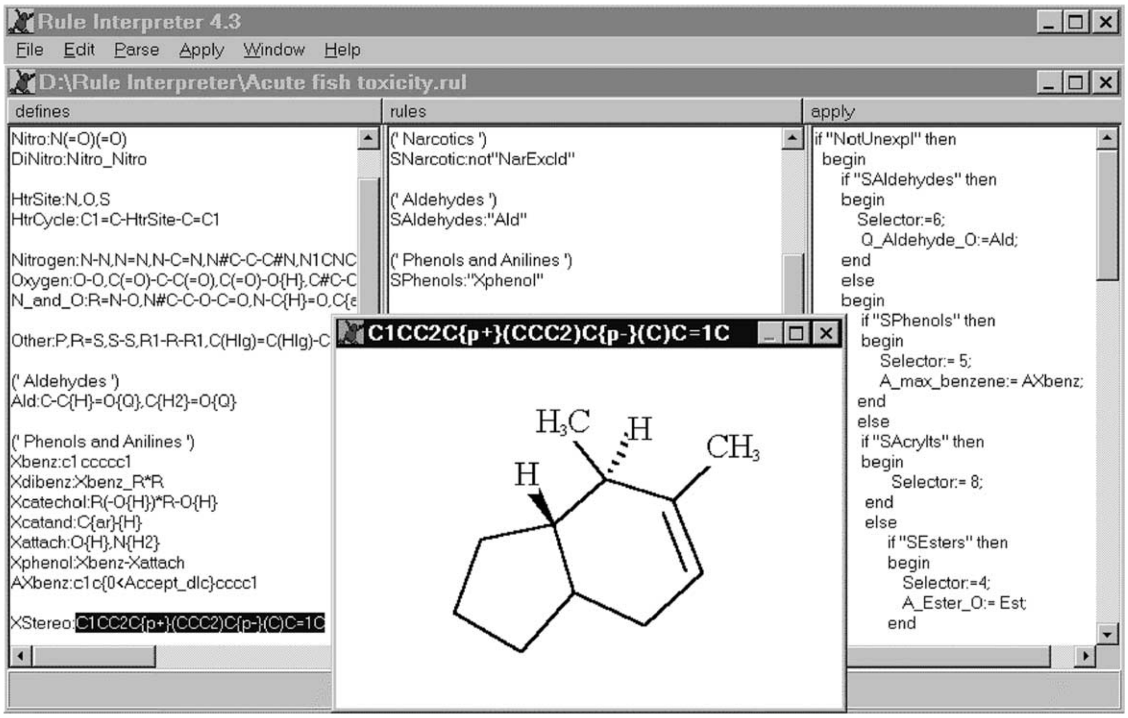
4. Program and implementation

The RI was written in Delphi code (Borland Int.) under MS Windows. Many data structures and low-level modules were inherited from the OASIS SAR system [14,15]. This applies particularly to the machine representation, file storage and basic manipulation of chemical graphs. The RI supports a text editor, which is customized to handle rule program script files (the default filename extension is 'rul'). Within the text editor, the script is divided into three windows for each program section. Apart from the typical functions of a text editor, the program can 'parse' and apply the script. In the former case, the script is checked syntactically and internal structures are updated. Syntax errors, if any, are traced back in the text and messaged by the type of the interpretation problem encountered. If the script is syntactically correct, the data structures relating identifiers to their context are updated. Contents errors are most likely in the first script section where screens are defined. To enhance their entry, the software incorporates a 2D chemical model builder from the SMILES-like notations. Practically, the conventional chemical depiction can be generated for any string in the text pane, which represents a valid screen construction. Thereby screen identifiers are recursively expanded to their chemical contents. The depicting software reflects as well the optional stereochemical notations within the string by rendering the so-called 2.5D chemical representation (see Fig. 1). The apply function is enabled only after the whole script is fully validated. Prior to running the apply section of the script, a file dialogue for the input and output CMP or SMILES data files is invoked. Upon completion the list of chemicals is displayed together with those descriptor values that have been newly generated or assigned as result of the screening processes.

5. Applications

5.1. Acute fish toxicity

The rule interpreter was used to develop a database-integrated QSAR expert system for acute toxicities to fish [17]. The expert system employs


Rule Interpreter 4.3
File Edit Parse Apply Window Help
D:\Rule Interpreter\Acute fish toxicity.rul
defines
rules
apply
C1CC2C{p+}(CCC2)C{p-}(C)C=1C
H₃C
H
H
CH₃

```
Nitro:N(=O)(=O)
DiNitro:Nitro_Nitro

HtrSite:N,O,S
HtrCycle:C1=C-HtrSite-C=C1

Nitrogen:N-N,N=N,N-C=N,N#C-C-C#N,N1CNC
Oxygen:O-O,C(=O)-C-C(=O),C(=O)-O{H},C#C-C
N_and_O:R=N-O,N#C-C-O-C=O,N-C{H}=O,C{ε

Other:P,R=S,S-S,R1-R-R1,C(Hlg)=C(Hlg)-C

(' Aldehydes ')
Ald:C-C{H}=O{Q},C{H2}=O{Q}

(' Phenols and Anilines ')
Xbenz:c1 ccccc1
Xdibenz:Xbenz_R*R
Xcatechol:R(-O{H})*R-O{H}
Xcatand:C{ar}{H}
Xattach:O{H},N{H2}
Xphenol:Xbenz-Xattach
AXbenz:c1c{0<Accept_dlc}cccc1

XStereo:C1CC2C{p+}(CCC2)C{p-}(C)C=1C
```

```
(' Narcotics ')
SNarcotic:not"NarExcld"

(' Aldehydes ')
SAldehydes:"Ald"

(' Phenols and Anilines ')
SPhenols:"Xphenol"
```

```
if "NotUnexpl" then
 begin
   if "SAldehydes" then
   begin
     Selector:=6;
      Q_Aldehyde_O:=Ald;
   end
   else
   begin
     if "SPhenols" then
     begin
       Selector:= 5;
       A_max_benzene:= AXbenz;
    end
    else
     if "SAcrylts" then
     begin
        Selector:= 8;
     end
     else
        if "SEsters" then
        begin
          Selector:=4;
           A_Ester_O:= Est;
        end
```

Fig. 1. The RI interface with excerpts from the RDL script providing toxicochemical differentiation for acute fish toxicities. The screen 'XStereo' is additionally inserted to illustrate the 2D chemical model builder from SMILES.

the two-step approach outlined in the introduction. The assessment of acute fish toxicities was chosen as a testing ground for the approach not only because these are very essential toxicity endpoints of environmental concern. Over the last decades, concordant toxicochemical knowledge has been gained about modes and mechanisms of toxic action to fish [18–21]. Two similar toxicodynamic classification schemes have been independently developed by Verhaar et al. [22,23] and Russom et al. [24] Furthermore, the fathead minnows acute toxicity database [25] that could be used still represents one of the largest and most congruous toxicity collections today. It contains LC_{50} values for about 660 industrial chemicals and, for most of them, the primary modes of toxic action determined in supplementary tests [18,19]. The database was provided by the US Environmental Protection Agency at the National Health and Environmental Effects Research Laboratory Mid-Continent Ecology Division in Duluth, Minnesota and further augmented with 3D molecular structures and electronic descriptors from quantum-chemical MO computations [26]. The database's integration in an expert system is multifunctional. First, the database served as a training set in the development of the toxicochemical classification scheme. Secondly, same-class tested chemicals formed correlation samples in the derivation of QSARs for toxic potencies, log $1/LC_{50}$. Practically, these class-specific correlative QSARs, whenever attainable, are not stored explicitly in the expert system, but defined only in terms of the several significant descriptors related to toxicity. The free terms in the actual linear equations are computed each time automatically, so as to provide best fits for the current training sets in the database that can be enlarged. Finally, the toxicity assessments of untested chemicals, if at all feasible, are always linked to corresponding correlation clusters of their tested toxicochemical analogues.

The implemented classification is based on two major toxicochemical categories: narcotics and reactive chemicals [20,21]. The first category includes baseline narcotics [27,28] and other nonspecifically acting chemicals causing their effect by noncovalent interactions. The reactive chemicals are subdivided, in part, into smaller classes encompassing molecules possessing well-defined reactive groups. The first step in the classification procedure is the separation of potentially reactive chemicals. Thereby we adopt the general criteria of Verhaar et al. [22,23], and augment them by some combinations of neighboring dipolar groups assumed to be reactive, either. Practically, potentially reactive chemicals are initially resolved by means of a composite screen including all likely reactive functions or combinations of functions. Next, more specific screens for electrophilic/proelectrophilic toxicophores associated with several relatively well-known mechanisms of toxicity [21,26] are applied. These involve, for instance, aldehydes, α-unsaturated halides, acrylates, and allyl and propargyl alcohols. Potentially reactive chemicals, which are not exactly assigned specific molecular mechanism, are automatically dropped out of further treatment. Thus $\sim$20% of the tested chemicals in the database remain 'unexplained' from a toxicochemical perspective. From the putative narcosis-acting chemicals, baseline narcotics are sought out by a restrictive rule, eliminating various chemical groups assumed to elicit toxic action other than nonpolar narcosis [18,24]. No attempts were undertaken to further subdivide the rest chemicals, which were loosely qualified as nonspecifically acting.

The toxic potency of baseline narcotics is exclusively dictated by toxicokinetics. The linear relation of log $1/LC_{50}$ and octanol-water partition coefficient log K_{ow}, revealed by Veith et al. [27] has been subsequently assumed to represent the minimal toxicities chemicals may exhibit (or baseline) [28]. We adopt this linear toxicity model directly, but slope and intercept are left dependent on the actual sample of tested chemicals in the database. In difference to baseline narcotics, groups of nonspecific-acting chemicals typically require an additional descriptor to quantify variations in electrophilicity. Thereby the energy of LUMO and the maximal acceptor delocalizability have been shown to perform equally well [6,7]. We choose the former merely for convenience in interpretation. Correlative models for groups of reactive electrophiles, albeit not so reliable because of small-sized training sets available, use in addition to log K_{ow} one or two site-specific electronic descriptors pertaining to the corresponding toxicophores [26]. The classification of chemicals is achieved exclusively by means of 2D-substructure screens, without reference to 3D structure or descriptors. However, toxicophore-selective screens for reactive chemicals incorporate descriptor qualifiers, which retrieve the relevant site-specific descriptors that are employed eventually in the multi-varied QSARs.

5.2. Androgen receptor binding affinity

Besides for computerized toxicochemical classification of noncongeneric chemicals, the development of RDL was influenced by demands to screen large inventories of industrial chemicals for potential hormone receptor binding agents. Such screening was based on preliminary developed rules reflecting the structural similarity of active receptor ligands. Rules were derived with the help of the aforementioned COREPA approach, which searches low-dimensional projections in the descriptor space that provide a satisfactory separation of active and passive xenobiotics. Since molecular flexibility contributes fuzziness to all conformation-dependant descriptors (including geometric ones and electronic ones, to a lesser extent) they have been addressed by RDL screens already when low-dimensional separation into 'active' and 'passive' areas was sought. We used the approach to model the androgen receptor (AR) binding affinity of 21 steroidal and nonsteroidal ligands, whose binding affinities (pK_i) were measured in a competitive binding assay. The tested chemicals were divided in 3 groups: highly active (p$K_i \geq 0.7$), active ($-2.0 <$ p$K_i < 0.7$), and inactive ones (p$K_i \leq -2.0$). The integral distributions (the so-called common reactivity patterns) were obtained as the products of the probability distributions for the six highly active and eight inactive molecules, taken separately. The best 1D separation of these two subsets was found to be by a dynamic descriptor for the interatomic distance of two nonbonded heteroatoms, denoted further by X. Below the corresponding screen definitions are given in terms of RDL (Section 3.1).

```
X:O,N,Cl,F
Xactive:X_{10.4 < distance < 11.1}
X{−0.322 < q < −0.312}
```

Xinactive: X_{2.0 < distance < 9.0}
X{−0.322 < q < −0.312}

In the last two joint screen definitions, the second heteroatom is required to possess electric charge between −0.322 and −0.312 a.u. For conformers of highly active ligands, the interatomic distance of these two active sites fell predominantly in the 10.4–11.1 Å range (screen Xactive), whereas the range of 2.0–9.0 Å (screen Xinactive) was mostly populated by conformers of the nonactive chemicals. The line

Ractive: 'Xactive' and not 'Xinactive'

from the *Rule* section sets up the rule, which most closely corresponds to the $pK_i = 0.7$ borderline in activity. Highly active chemicals have all one or more conformations, which obey rule 'Ractive'. In contrast, none of the low-energy conformers of the rest 15 molecules from training set meets the screening criterion. The actual assignment of activity rank (descriptor AR_binding) is achieved in the *Apply* section by the following conditional statement.

```
if 'Ractive' then AR_Binding := 1
   else AR_Binding := 0;
end.
```

The screening criterion has been checked for another 7 compounds, which were subsequently tested. Within the validation set of 64 different low-energy conformers in total, all less-than- highly active chemicals ($pK_i < 0.7$) were successfully recognized.

6. Summary

RI provides chemists and toxicologists with a robust and flexible tool for screening of chemical databases and inventories. The variety of admissive substructure screens is practically unrestricted in terms of complexity and size, so that screening problems of different nature can be handled. Furthermore RDL combines substructure search with descriptor-oriented selection, incorporates Boolean logic and allows unlimited branching in a tree-like decision structure. The software was particularly developed in close relation with the problem of large-scale toxicochemical differentiation and assessment for noncongeneric chemicals. Beyond doubt, prediction of toxic mechanisms and mode of actions from chemical structure will still remain a highly intellectual expert task. However, even when chemical rules identifying probable toxicochemical analogues have been approximately formulated, their computer implementation is still challenged by unresolved details and minor uncertainties. Moreover, for any practical application, additional validation and adjustment of the approach from existent toxicity information are still needed. By developing methodology and software we were able to largely enhance that process.

Acknowledgements

We acknowledge support from the US EPA, Cooperative Agreement (CR822306-01-0), and the European Commission, EU Project IC20-CT98-0114 (EDAEP). We thank Drs S.P. Bradbury and C.R. Russom from the US Environmental Protection Agency, National Health and Environmental Effects Research Laboratory Mid-Continent Ecology Division, Duluth, MN USA for providing us the database in electronic form.

Appendix A. Formal description of RDL in Bacus-Naur metalanguage style

〈ruleprogram〉 :: = define«〈screenlist〉 « rules « 〈rulelist〉 « apply « 〈statementlist〉 « end.
〈screenlist〉 :: = 〈screendefinition〉[«〈screendefinition〉]
〈rulelist〉 :: = 〈ruledefinition〉[«〈ruledefinition〉]
〈statementlist〉 :: = 〈statement〉[;〈statement〉]
〈screendefinition〉 :: = 〈screen〉:〈screenstr〉
〈ruledefinition〉 :: = 〈rule〉:〈booleanexpr〉
〈booleanexpr〉 :: = "〈screen〉" |"〈rule〉"|(〈booleanexpr〉)|not〈booleanexpr〉|〈booleanexpr〉 or 〈booleanexpr〉|〈booleanexpr〉 and 〈booleanexpr〉
〈statement〉 :: = 〈assignmentst〉|〈conditionalst〉|〈compoundst〉

⟨assignmentst⟩ ∷ = ⟨moldescriptor⟩ ≔ ⟨realnum-
⟩|⟨screen⟩
⟨moldescriptor⟩* is a name from a predefined molecular descriptor list
⟨conditionalst⟩ ∷ = **if** ⟨rule⟩ **then** ⟨statement⟩
‘**else** ⟨statement⟩’
⟨compoundst⟩ ∷ = **begin** ⟨statementlist⟩ **end**
⟨screen⟩ ∷ = ⟨letter⟩[⟨alphanum⟩]
⟨rule⟩ ∷ = ⟨letter⟩[⟨alphanum⟩]
⟨screenstr⟩ ∷ = ⟨atomentry⟩|⟨screen⟩|(-
⟨screenstr⟩)|⟨screenstr⟩‘⟨bondentry⟩’⟨screenstr⟩|
screenstr⟩,⟨screenstr⟩
⟨atomentry⟩ ∷ = ⟨atomlabel⟩[⟨atomqualifier⟩][-
⟨num⟩]
⟨num⟩ is the SMILES notations of ring-closure bonds. Notation-specific consistency requirements apply.
⟨atomlabel⟩ ∷ = C|c|N|n|O|o|S|s|B|P|F|Cl|Br-
|I|R|⟨otheratom⟩
⟨otheratom⟩ is any chemical element label enclosed in [ ]

R denotes any atom in substructure search.

The default chemical bond for lowercase atom entries is aromatic, otherwise it is single.

⟨atomqualifier⟩ ∷ = {⟨adlabel⟩}|{‘⟨realnum-
⟩⟨‘⟨atomdescriptor⟩‘⟨⟨realnum⟩’}|{‘⟨realnum⟩⟨’-
moldescriptor⟩‘⟨⟨realnum⟩’}
⟨adlabel⟩* ∷ = acy|scy|dcy|sk|sc|sp1|sp2-
|sp3|ar|2 + |2 − |p + |p − |h|h2|h3|h4| + | − |.

Meaning of some: acy—not belonging to a ring; scy—belonging to one ring; dcy—belonging to more than one rings; sk—belonging to the skeletal part; sc—belonging to a side chain; ar—*π*—conjugated;h2—at least two protons attached; p + —positive chiral parity.

⟨bondentry⟩ ∷ = ⟨bondlabel⟩[⟨bondqualifier⟩]
⟨bondlabel⟩ ∷ = –| = |#|*|.|_

Different bond labels correspond to single, double, triple, aromatic, ionic bond, and no bond

⟨bondqualifier⟩ ∷ = {⟨bdlabel⟩}|{‘⟨realnum⟩⟨’-
⟨bondescriptor⟩‘⟨⟨realnum⟩’}
⟨bdlabel⟩* ∷ = t|c|g|g + |g −

Meaning: t—*trans*; c—*cis*; g—*gauche*; g + —*gauche* clockwise; g − —*gauche* counter clockwise.

⟨atomdescriptor⟩* and ⟨bondescriptor⟩* are names from predefined descriptor lists.
⟨alphanum⟩ ∷ = ⟨letter⟩|⟨num⟩
⟨letter⟩ ∷ = a|b|···|z|A|B|···|Z
⟨num⟩ ∷ = 0|1|2|···|9
⟨realnum⟩ ∷ = ‘—’⟨num⟩[⟨num⟩]‘[num]’

Captions

⟨⟩	enclose categories
|	disjunctive or
∷	is defined as
[ ]	enclose terms that may be omitted or repeated any number of times
‘ ’	enclose terms that may be omitted
«	line feed as delimiter
*	case-insensitive

References

[1] A.M. Richard, Mutat. Res. 305 (1994) 73–97.

[2] A.M. Richard, Toxicol. Lett. 103 (1998) 611–616.

[3] N. Greene, J. Chem. Inf. Comput. Sci. 37 (1997) 148–150.

[4] W. Karcher, S. Karabunarliev, J. Chem. Inf. Comput. Sci. 36 (1996) 672–677.

[5] O.G. Mekenyan, et al., Environ. Sci. Technol. 31 (1997) 3702–3711.

[6] G.D. Veith, O.G. Mekenyan, Quant. Struct.-Act. Relat. 12 (1993) 349–356.

[7] O.G. Mekenyan, G.D. Veith, SAR QSAR Environ. Res. 1 (1993) 335–344.

[8] S. Karabunarliev, O.G. Mekenyan, W. Karcher, C.L. Russom, S.P. Bradbury,Quant. Struc.-Act. Relat. 15 (1996) 311–320.

[9] R.D. Cramer III, D.E. Patterson, J.D. Bunce, J. Am. Chem. Soc. 110 (1988) 5959–5967.

[10] R.A. Goldstein, J.A. Katzenellenbogen, Z.A. Luthey-Schulten, D.A. Seielstad, P.G. Wolynes, Proc. Natl. Acad. Sci. USA (Biochemistry) 90 (1993) 9949–9953.

[11] G.R. Marshall, in: H. Kubinyi (Ed.), 3D QSAR in Drug Design: Theory, Methods and Applications, Escom, Leiden, 1993, pp. 80–116.

[12] O.G. Mekenyan, R.A. Goldstein, J.A. Katzenellenbogen, Z.A. Luthey-Schulten, D.A. Seielstad, P.G. Wolynes, Quant. Struct.-Act. Relat. 18 (1999) 139–153.

[13] D. Weininger, J. Chem. Inf. Comput. Sci. 28 (1988) 31–36.

[14] O.G. Mekenyan, S. Karabunarliev, D. Bonchev, Comput. Chem. 14 (1990) 193–200.

[15] O.G. Mekenyan, et al., Comput. Chem. 18 (1994) 173–187.

[16] T. Hurst, J. Chem. Inf. Comp. Sci. 34 (1994) 190–196.

[17] S. Karabunarliev, S.D. Dimitrov, N. Nikolova, O.G. Mekenyan, in: J. Walker (Ed.), Proceedings of Eighth Workshop on QSAR in the Environmental Sciences, Baltimore, USA, May 1998, SETAC, Pensacola, 2002, in press.

[18] S.P. Bradbury, T.R. Henry, R.W. Carlson, in: W. Karcher, J. Devillers (Eds.), Practical Applications of QSARs in Environmental Chemistry and Toxicology, Kluwer, Dordrecht, 1990, pp. 295–315.

[19] S.J. Broderius, M.D. Kahl, M.D. Hoglund, Environ. Toxicol. Chem. 9 (1995) 1591–1605.

[20] J.L.M. Hermens, Environ. Health Perspect. 87 (1990) 219–225.

[21] R.L. Lipnick, Sci. Total Environ. 109/110 (1991) 131–153.

[22] H.J.M. Verhaar, C.J. van Leeuwen, J.L.M. Hermens, Chemosphere 25 (1992) 471–491.

[23] H.J.M. Verhaar, E.U. Ramos, J.L.M. Hermens, J. Chemometrics 10 (1996) 149–162.

[24] C.L. Russom, et al., Environ. Toxicol. Chem. 16 (1997) 948–967.

[25] D.L. Geiger, L.T. Brooke, D.J. Call (Eds.), Center for Lake Superior Environmental Studies. Acute Toxicities of Organic Chemicals to Fathead Minnows (*Pimephales promelas*), vol. 5, University of Wisconsin—Superior, Superior, WI, 1990, pp. 1–4, see also vols. 1–4.

[26] S. Karabunarliev, et al., Quant. Struc.-Act. Relat. 15 (1996) 302–310.

[27] G.D. Veith, D.J. Call, L.T. Brooke, Can. J. Fish. Aquat. Sci. 40 (1983) 743–748.

[28] R.L Lipnick, W. Karcher, J. Devillers (Eds.), Practical Applications of QSARs in Environmental Chemistry and Toxicology, Kluwer, Dordrecht, 1990, pp. 281–294.