J Comput Aided Mol Des (2006) 20:361–373
DOI 10.1007/s10822-006-9058-y

ORIGINAL RESEARCH PAPER

Representation of molecular structure using quantum topology with inductive logic programming in structure–activity relationships

Bård Buttingsrud · Einar Ryeng · Ross D. King · Bjørn K. Alsberg

Received: 15 March 2006 / Accepted: 20 July 2006 / Published online: 13 October 2006

Abstract The requirement of aligning each individual molecule in a data set severely limits the type of molecules which can be analysed with traditional structure activity relationship (SAR) methods. A method which solves this problem by using relations between objects is inductive logic programming (ILP). Another advantage of this methodology is its ability to include background knowledge as 1st-order logic. However, previous molecular ILP representations have not been effective in describing the electronic structure of molecules. We present a more unified and comprehensive representation based on Richard Bader's quantum topological atoms in molecules (AIM) theory where critical points in the electron density are connected through a network. AIM theory provides a wealth of chemical information about individual atoms and their bond connections enabling a more flexible and chemically relevant representation. To obtain even more relevant rules with higher coverage, we apply manual postprocessing and interpretation of ILP rules. We have tested the usefulness of the new representation in SAR modelling on classifying compounds of low/high mutagenicity and on a set of factor Xa inhibitors of high and low affinity.

Keywords Structure representation using quantum topology (StruQT) · Atoms in molecules (AIM) · Bader theory · Inductive logic programming (ILP) · Structure–activity relationship (SAR) · Quantitative structure–activity relationship (QSAR)

B. Buttingsrud · E. Ryeng · B. K. Alsberg (✉)
Chemometrics and Bioinformatics Group, Department of chemistry, Norwegian University of Science and Technology, Trondheim, Norway
e-mail: bjorn.alsberg@nt.ntnu.no

R. D. King
Computational Biology Group, Department of Computer Science, University of Wales Aberystwyth, Wales, United Kingdom

Introduction

A structure–activity relationship (SAR) relates biological or therapeutic activity of a drug to its chemical structure. SAR modelling has been widely applied in elucidating biological processes and in the development of new drugs. In addition to being reliable, a model should also be comprehensible and provide a better understanding of the chemistry and biology behind the problem. Any SAR methodology consists of a representation to describe chemical structure and a learning algorithm which relates a compound's activity to its structure.

Most learning algorithms employed in SAR problems require an *attribute-based representation* where each molecule is described as a list of properties. An early example of this was the use of global molecular properties, e.g. hydrophobicity, molecular refractivity [1, 2]. Attempts to incorporate attributes describing local properties within a molecule include the use of topological indices, which characterise molecular structure as a single scalar [3]. A more elegant way to represent local knowledge is to describe the substructures and their associations directly, e.g. ''a benzene

ring connected to a nitro group''. However, this representation is *relational* instead of attribute-based and conventional learning algorithms cannot make effective use of this. It requires a description of the *relationship* between two attributes.

Comparative field analysis (CoMFA) is an example of a widely used attribute-based method for describing the local properties of molecules [4]. The electrostatic potential or similar distributions are estimated by placing each molecule in a 3D grid and calculating the interaction between a probe atom at each grid point and the molecule. When the molecules are properly aligned in a common reference frame, each point in space becomes comparable and can be assigned an attribute such that attribute-based learning methods can be used. However, CoMFA fails to provide accurate results when the lack of a common skeleton prevents a reasonable alignment. The need for alignment is a result of the attribute-based description of the problem.

The only method fully capable of handling relational data directly without relying on alignment of molecules is inductive logic programming (ILP) [5]. King et al. [6] introduced a SAR method based on ILP which described chemical structure as a collection of atoms connected through chemical bonds. ILP enables the inclusion of background knowledge by defining many high-level chemical concepts, e.g. benzene ring, methyl group and the three topologically distinct ways to connect three benzene rings. They generated rules which are easy to understand such as: ''A compound is highly mutagenic if it has an aliphatic carbon atom attached by a single bond to a carbon atom which is in a six-membered aromatic ring''. Much work has been done to improve ILPs ability to solve SAR problems; generation of indicator variables to provide quantitative estimates of the activity [7, 8], building pharmacophore models [9, 10], dealing naturally with multiple conformations [10], performing structure-based drug design [11] and improvements in algorithms to reduce search space [12, 13]. However, no non-trivial improvements have been applied to the original atom/bond representation. It is desirable to replace this representation with one which is richer and more able to describe well-known chemical concepts such as conjugation, hyperconjugation, delocalisation effects, aromaticity, electrophilicity, nucleophilicity, covalent and ionic bonds in a quantum mechanical setting.

Richard Bader's atoms in molecules (AIM) theory provides a link between quantum and classical chemistry which enables a natural framework for explaining chemical concepts and phenomena [14]. Compact descriptors based on bond properties from AIM theory have been successfully applied to various attribute-based SAR problems [15–24]. We have previously presented the idea of combining an AIM-based representation named *structure representation using quantum topology* (StruQT) [15, 16] with inductive logic programming [25]. This approach combines a powerful and chemically interpretable representation with a learning algorithm able to describe the true relational nature of molecular structure. In this paper, we investigate the new method using a classical SAR problem relating to the prediction of mutagenicity and to the prediction of the affinity of a set of factor Xa inhibitors.

Theory

Inductive logic programming

Learning algorithms based on propositional logic are able to learn *if-then* rules such as ''If a molecule has a benzene ring then it is biologically active'', given that the precoded descriptor benzene ring exists. Propositional learners require that molecules are represented as a vector with one element for each attribute, and hence they are referred to as attribute-based methods. The output is a set of rules or a decision tree that classifies the data using these attributes. A propositional algorithm cannot learn a more complex rule like ''If a molecule has a benzene ring connected to a nitro group then it is active'' without having precoded a benzene ring connected to a nitro group as an attribute where the *relationship* between two attributes is used. Furthermore, this limitation prevents propositional learning algorithms to take fully advantage of the molecular structure hypothesis—that a molecule is a collection of atoms connected by bonds—which is perhaps the most central hypothesis in chemistry.

The incorporation of relations enables us to naturally describe atomic connections in molecules. For instance, the bond between atoms A and B might be represented using a bond *predicate* bond (A, B) where a predicate here defines a relationship between two objects (atoms). Inclusion of predicates and therefore the capability of handling relational data replaces propositional logic with first-order logic. The class of learning algorithms which generates rules expressed in first-order logic are known as inductive logic programming (ILP).

The classical atom/bond representation used to solve SAR problems [6] is based on the molecular structure hypothesis. Atoms are represented in the form

```
atom(127,127_1,c,22,0.191).
```

stating that the first atom in compound 127 is a carbon atom of type 22 (aromatic) with a positive charge of 0.191.

Similarly,

```
bond(127,127_1,127_6,7)
```

states that there is a type 7 bond (here aromatic) between the first and sixth atom in compound 127.

Another advantage with a learning algorithm formulated in first-order logic is the possibility of including background knowledge in the form of computer programs. Since chemists often study molecules in terms of molecular groups, the atom/bond representation was extended with programs that define such high-level chemical concepts. Contrary to propositional algorithms, ILP can learn rules which use structural combinations of these concepts without having to explicitly include the combinations as new attributes. Another example of useful programs used as background knowledge is the definition of a distance measure. This allows a three-dimensional representation with rules in the form: ''A molecule is active if it has a benzene ring and a nitro group separated by a distance of 4 ± 0.5 Å''. It should be noted that the inclusion of distance measures does not require an alignment of the molecules. It is also straightforward to include more than one conformation of each compound which allows the consideration of conformational flexibility which is often a major drawback by conventional QSAR/SAR methodologies.

The input to an ILP algorithm is the background knowledge about the problem (descriptions of atoms, bonds and programs defining high-level concepts) and a set of positive/active and negative/inactive molecules. The task is then formulated as finding a set of rules which explains as many positive examples as possible while covering the fewest number of negative examples.

When the program searches for a new rule, it selects a single positive example to guide the search. It then tries to build a new rule that covers this example and as many other positive examples as possible. The first attempt to build the new rule uses only a single predicate such as benzene(A, B) stating that a molecule A is active if it has a benzene ring B. A set of possible extension to the best of these candidate rules are generated and tested to find the optimal combination. The default criteria maximises the difference between positive and negative examples covered by the rule. This process is repeated as long as the best of these candidate rules have a higher positive coverage and lower negative coverage than the shorter candidates. After the candidate rules have reached a certain length, the inclusion of more predicates will no longer improve the rules because the next predicate removes more true than false positives leading to a lower overall coverage.

All the positive examples that are covered by the new rule are removed from the data set before a new positive example is chosen and the entire rule generation process is repeated. The negative examples that are erroneously classified as positives by the rule are kept in the data set. Otherwise the next candidate rule that classifies them incorrectly would not be punished for this. Eventually, no more significant rules can be found, or all the positive examples have been classified correctly, and the set of rules is considered complete. This set of rules, often referred to as the theory explaining the problem under study, is returned as the output from the ILP algorithm. For a more technical and mathematical description of the rule-making process in ILP, the reader is referred to the review article by Muggleton and De Raedt [26].

A major limitation of ILP in learning quantitative SAR models (QSAR) is how to meld first-order logic with probability theory. This is arguably the most important theoretical area in machine learning [27]. One very successful practical approach to this problem is to apply ILP as a pre-processing step to learn suitable propositional descriptors (attributes), and then to use standard QSAR approaches [7, 8]. Recent work has also taken the alternative approach of using statistical procedures as background knowledge [7].

Bader's atoms in molecules quantum theory

The molecular structure hypothesis is central to most aspects of chemical intuition and knowledge. However, the concept of an atom is not explicitly defined in traditional quantum mechanics. It is only concerned with particles moving in potential fields. This poses a representational dilemma; we wish to make use of the rigour and physical correctness of quantum mechanics, but we also require that our models are easy to understand and related to conventional chemical intuition. A quantum theory which meets both requirements has been presented by Bader [14]. A unique definition of the concept of atoms and bonds results from a study of the topology of a molecule's electron distribution. The theory provides a link between quantum mechanics and many standard chemical concepts.

Central to AIM theory is an analysis of the topological properties of the electron density distribution ρ. It is hard to find the relevant information in the

electron density directly since most of the variation is located around the nuclei. The topological characteristics of a scalar field may be summarised with the properties of its critical points and this forms the basis of our electronic structure representation. The four different types of critical points in the electron density in a stable molecule are associated with atomic nuclei, bonds, rings and cages:

Nuclear attractor (NA) is a critical point with maximum electron density along all three dimensions. The locations of the maxima are for all practical purposes equal to the nuclei positions.

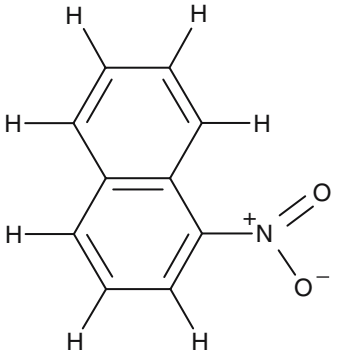
Bond critical point (BCP) is a point with maxima along two dimensions and a minimum along the third. A BCP is always located between two bonded atoms. It has minimum electronic density along the bond path between the two atoms. However, it has a maximum along the plane perpendicular to the bond path.

Ring critical point (RCP) is a point with minima along two dimensions and a maximum along a third dimension in the electron density. The atoms surrounding a ring have maximum density, and it decreases along the path towards the RCP. However, the point is a maximum along the path perpendicular to the ring surface.

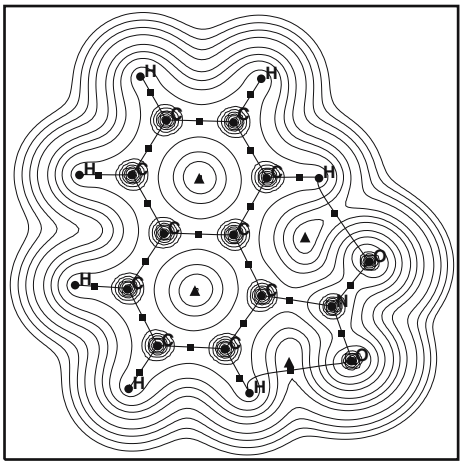
Cage critical point (CCP) is a point with minimum electron density along all three dimensions. Cages are formed when at least two rings bound a region with minimum electron density.

The electron density contour map of the molecular plane in 1-nitronaphthalene (see Fig. 1(a)) is illustrated in Fig. 1(b). Three different critical points are observed: nuclear attractors (circles), bond critical points (square) and ring critical points (triangle). We see that according to AIM theory, the oxygen atoms in the nitro group are bonded to the neighbouring hydrogen atoms. These bond paths create new ring structures which are not usually explicit in our chemical knowledge. Each bonded pair of atoms is connected by a line which goes through a BCP. This path is called an *atomic interaction line* (AIL) and is related to another topological property of the electron density field: The gradient vector field and its gradient paths.

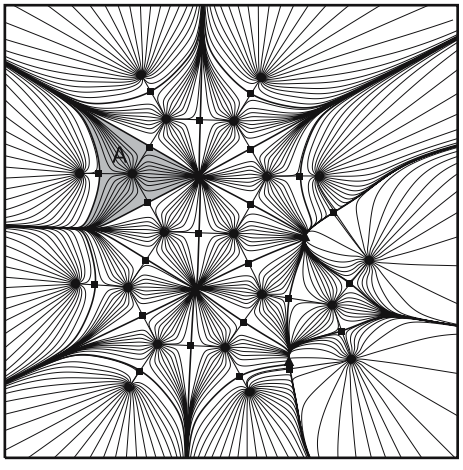
The gradient vector denoted $\nabla\rho$ is defined as the direction of steepest ascent, and by tracing the path with infinitesimal small steps, we obtain a gradient path. These paths never intersect, and they always terminate at a critical point. Most gradient paths terminate at nuclear attractors, and the complete set of paths ending in a specific attractor is known as its *basin*. They naturally partition the molecule into regions which are identified as the AIM definition of an atom. Figure 1(c) shows selected gradient paths for



(a) Structural formula



(b) Contour plot of ρ



(c) Gradient vector field of ρ

Fig. 1 Illustration of electron density contour plot, critical points, atomic interaction lines, gradient paths, interatomic surfaces and definition of atoms. The structural formula of 1-nitronaphthalene is shown (**a**). Figure (**b**) and (**c**) shows respectively the contour plot and gradient paths for the molecular plane. Squares indicate nuclear attractors, circles are bond critical points while ring critical points are shown as triangles. Thick lines in (**c**) are the interatomic surfaces

1-nitronaphthalene where the thin lines are paths terminating at the nuclear attractors. They either originate at infinity or at another critical point. The thick lines terminating at a BCP defines the interatomic surfaces separating the different atoms from each other. One example of applying the definition of an atom is the grayed region in Fig. 1(c) marked A which is an aromatic carbon atom bounded to a hydrogen atom. We see that atoms in molecules are not spherical.

The properties of a critical point can be further characterised using the Hessian matrix. It describes the second derivatives or curvature at the point with respect to the position coordinates. A diagonalisation of the matrix gives eigenvalues λ_i describing the curvature at the critical points in a representation independent of the choice of molecular coordinate system. Hence, the eigenvalues may be used to describe the general properties of critical points. For instance, a BCP always has two negative and one positive eigenvalue. As AIM theory defines a chemical bond as a point in the electron density distribution, it is straightforward to describe the nature of a bond with quantitative measures. For instance, the following properties are very useful [23]:

The electron density (ρ_b) is related to bond order and bond energy.

The laplacian ($\nabla^2\rho_b$) can be computed as the sum of the eigenvalues of the Hessian matrix ($\lambda_1 + \lambda_2 + \lambda_3$). It measures the degree of charge concentration at the bond. Negative values are associated with shared interaction present in covalent bonds. On the other hand, positive values are called close-shell interaction and are present in ionic, hydrogen and van der Waals bonds. While covalent and ionic bonds are usually used as qualitative concepts within chemistry, AIM theory provides a quantitative description with more detailed information. This clearly illustrates the advantage of using AIM theory for SAR problems.

The ellipticity (ϵ) is defined as $\lambda_1/\lambda_2 - 1$ and measures the deviation of the electron density distribution from being symmetrical in the plane orthogonal to the bond path. This relates to the π character of a chemical bond and AIM theory again provides a quantitative generalisation of classical chemical concepts. It also measures the susceptibility of ring bonds to rupture.

The electron density at the critical point along with its laplacian is also used to characterise the nature of RCPs.

StruQT-ILP

It is advantageous to replace the original atom/bond representation with one based on AIM theory for the following reasons: Many chemical concepts are explained and quantified by the theory leading to more relevant and comprehensive rules. The atom and bond typing used in the original representation are based on *ad-hoc* definitions from the original modelling software while a representation using AIM theory provides a generic and principled way to form types. Previous work on AIM descriptors in modelling attribute-based QSARs have only used properties of the bond critical points [15–24]: the value of the electron density, the ellipticity and the laplacian. We expand this set by including properties of the ring critical points. These quantitative properties, the critical points themselves and their connectivities in addition to high-level background programs form the foundation of the current version of our StruQT-ILP representation.

Similar to the atom predicate in the atom/bond representation, we have defined a na predicate. For example,

```
na(mol162, na162_1).
element(mol162, na162_1, c).
```

states that in compound 162, the critical point named *na162_1* is a nuclear attractor corresponding to a carbon atom.

Instead of a bond predicate, our new quantum topological representation uses the related concept of an atomic interaction line:

```
ail(mol162, bcp162_2, na162_2, na162_3).
```

stating that the second and third atom in molecule 162 is connected through the second BCP.

In addition to atom/bond predicates, King et al. [6] also included high-level chemical structures such as the definition of a benzene ring and a nitro group. We have added similar predicates with some minor improvements. For instance, predicates are included to describe how two rings may be connected. The first predicate (fused) defines two rings bonded with two common atoms. Another possibility is two rings connected through a single bond between an atom in each ring. Table 1 gives the complete list of predicates used in this article.

Some comments about the definition of these predicates are necessary. When King et al. defined their predicates, they used additional data from a molecular modelling software. For instance, the concept of aromatic atoms were already defined by the software. However, The StruQT-ILP representation starts with a set of critical points, their mathematical properties and connections. *Any chemical concepts or structure must be defined from these fundamental properties*. For instance, we define an aromatic bond using the quantitative

Table 1 Chemical predicates defined in StruQT representation

Basic critical point predicates	
na	A critical point is a nuclear attractor
BCP	A critical point is a bond critical point
rcp	A critical point is a ring critical point
element	Specifies the element type of an atom
ail	Defines an atomic interaction line
Chemical structure predicates	
ring_size_5	A ring with five atoms
ring_size_6	A ring with six atoms
chem_ring	Ring where all atoms are connected through strong bonds
not_chem_ring	Ring with at least one weak (e.g. hydrogen) bond
carbon_5_ring	A chem_ring with five carbon atoms
carbon_6_ring	A chem_ring with five carbon atoms
arom_ring	A chem_ring with only aromatic bonds
not_arom_ring	A chem_ring which are not an arom_ring
carbon_5_aromatic_ring	An arom_ring with five carbon atoms
benzene	An arom_ring with six carbon atoms
anthracene	Three benzene rings fused linearly
phenanthrene	Three benzene rings fused in a curve
ball3	Three benzene rings fused in a ball
hetero_aromatic_5_ring	A heterogeneous arom_ring withfive atoms
hetero_aromatic_6_ring	A heterogeneous arom_ring withsix atoms
nitro	A nitro group
methyl	A methyl group
Relationship predicates	
fused	Two rings are connected with two common atoms
connected_rings_unfused	Two rings are connected with a single bond
connected_ring_group	A ring is connected to a chemical group
ringmember	An atom is a part of a ring
atommember	An atom is a part of a chemical group
Numerical predicates	
electron_density_lteq	Electron density is less than or equal a value
ellicticity_lteq	The ellipticity is less than or equal a value
laplacian_lteq	The laplacian is less than or equal a value
electron_density_gteq	Electron density is larger than or equal a value
ellicticity_gteq	The ellipticity is larger than or equal a value
laplacian_gteq	The laplacian is larger than or equal a value

properties of the BCP. We also define a not_chem_ring predicate where there is not a strong, covalent bond between all atoms in a ring. There may for instance be a hydrogen bond between a hydrogen and nitrogen atom leading to a ring structure different in nature than ordinary rings (here denoted chemical rings).

Our present definitions are based on the empirical data for the problem of predicting mutagenic activity, and better, more robust definitions are required for further work. However, the definitions proved to be applicable also for the other problem studied in this work.

Materials

Data

Mutagenesis data set

Debnath et al. [28] studied the problem of predicting whether aromatic and heteroaromatic nitro compounds cause mutagenic changes measured by Ames test using *Salmonella typhimurium* TA98. The data set includes diverse molecules such as 2,4-dinitrophenylhydrazine, nitrocoronene, mono-, di-, and tetranitroarenes, nitroindoles, nitroindazoles, nitrofurans and nitrodiazines. As the compounds are heterogeneous and cannot be superimposed onto a common template, attribute-based methods have problems finding comparable structural attributes. The representation used in the original article by Debnath et al. [28] consisted of the energy of the lowest unoccupied molecular orbital (LUMO) and the partition coefficient between octanol and water in addition to two binary-valued indicator variables (I_1 and I_2). The descriptor I_1 was set equal to 1 for all compounds containing three or more fused rings and I_a was set to 1 for five examples of acenthrylenes as they showed lower activity than expected from their molecular properties. These descriptors were selected manually by experienced chemists after detailed inspection of this particular data set. All these descriptors are global

in nature, and thus the problem was solvable by attribute-based methods such as ordinary least square regression. The data set was further divided into 188 so-called ''regression-friendly'' compounds and 42 which are not readily modelled by attribute-based regression methods. Even though this division of the data set is artificial and somewhat misleading, we will keep it to allow a relevant comparison with the results obtained from previously published results. This division and data set have been used as a benchmark data set in machine learning [28] to test various methods such as linear regression, back-propagation neural network, CART and the atom-bond representation with ILP [6]. This benchmark data set is the most well-document problem studied using the atom/bond representation. If we are able to prove that the new methodology gives additional information about this problem, we have rather strong evidence that the introduction of a more flexible representation is well-founded.

A compound was classified as active or highly mutagenic when the logarithm of revertants/nmol produced by the mutagen was above zero. Of the 188 compounds, 125 are classified as active and 63 as inactive while similar values for the 42 compounds are 12 and 30, respectively.

Factor Xa data set

Fontaine et al. [29] recently published their new methodology anchor-GRIND on a data set discriminating between factor Xa inhibitors of high and low affinity. It is a modification of the *gr*id-*in*dependent *d*escriptors (GRIND) method [30, 31] which uses a mathematical transformation to provide alignment-free descriptors from molecular interaction fields (MIF). The user is required to define a single common position in the structure of all compounds in a series of molecules which is named an *anchor point*. This point is used as a common reference which enables a more precise geometrical description. They conclude that ''the descriptors obtained are far more specific and therefore produce better models, which are easier to interpret'' than in the original GRIND methodology. The data set contains a wide series of compounds which all share an amidine group which is often observed in factor Xa inhibitors and binds in a well-defined pocket. The inhibitors were used for binary classification between low (K_i < 10 nM) and very high (K_i > 1 μM) activity. This resulted in 156 low-activity and 279 high activity compounds in the data set. It was divided using random assignment into a calibration and test set with 290 and 145 compounds, respectively.

We should at this point note that whereas the anchor-GRIND method requires an anchor point, this is not an requirement of StruQT-ILP. We could have compared our method with the original GRIND method, but the authors have unfortunately not reported the prediction accuracy for this method in their article. They do however report the corresponding value for another data set which shows that the standard GRIND method produces a model with a lower cross-validated correlation coefficient and which is harder to interpret than anchor-GRIND [29]. Hence, we aim at comparing our new methodology with a method specifically developed to handle a special, although very common situation in building SAR models.

Computational details

All molecules were optimised at density functional theory (DFT) level using the B3LYP functional with a 6-31G** basis set computed with the Gaussian 98 program [32]. DFT was chosen as it is able to include the effects of electron correlation at a low computational cost. A recent study of nitroaromatic compounds also gave good results with DFT [33]. Among DFT functionals B3LYP gives the best estimates of BCP properties [24]. Even though the absolute values of BCP properties are highly basis-set dependent, they preserve the trends even at small basis sets. A large basis set is therefore not crucial to model a reliable SAR because only relative differences between molecules are important.

One of the active compounds in both the 188 mutagenesis data set and the Factor Xa calibration set were removed from computation since they contain an iodine atom which is not parametrised at the selected basis set. All AIM properties were computed using the AIM analysis program MORPHY98 [34]. For the Factor Xa data set, MORPHY98 failed to find cage critical points in 10 molecules. We therefore did not use cage critical points in the rule-making process. One molecule was also removed from the test set since the number of molecular orbitals was too high for MORPHY98.

Data analysis

We used the ILP system Aleph [35] which is implemented in the computer language Prolog. A sequential covering algorithm is usually employed in ILP to learn the rules (induce command in Aleph). One positive/active example is selected and a rule is induced which

ensures it is correctly classified. All examples covered by this rule are removed and the process is repeated. One problem with this approach is that the rules and predictive ability depend on the order of the active compounds. A better option is to select more than one example and only include the best rule generated into the theory. The examples covered by this rule are as usual removed, and new active molecules are selected to explain the remaining examples. This is achieved in Aleph using the samplesize option which specifies the number of examples selected randomly at each step. We selected this number equal to the number of active molecules which means that the most optimal rule given the specified background knowledge and other settings is found. This also has the advantage that all molecules are selected, and the algorithm is no longer stochastic, but deterministic. After removing the examples covered by the selected rule, the best rules explaining the remaining examples are found in a similar way. Another advantage of this approach is that the best rule is the one which explains the highest number of active compounds not yet explained by another theory. Hence, the rules in the obtained theory are smaller in number and better fit to explain the diversity among the active compounds in a few rules.

Pruning was used to remove irrelevant rules in order to reduce the size of the search space. For instance, there is a large number of rules containing the predicates

```
na(A, B), na(A, C), na(A, D)
```

stating that molecule A has three atoms (which may or may not be the same). Combinations of such predicates severely slows down the program, and consequently only one na is allowed in a rule. Other atoms are included through the use of for instance ail predicates. Similar pruning rules were also included.

The prediction accuracy was evaluated using leave-one-out crossvalidation for the two mutagenesis data sets and using a test set for the factor Xa data set. This difference reflects the choices made previously in the literature [6, 29].

Manual postprocessing of rules

In addition to looking only at the prediction accuracy, it is also important to study the relevancy of the obtained rule. Relevancy of rules is a relative concept, but it is obvious that rules with high coverage are preferable. A theory with many rules implies that the model consists of many sub-models which are not relevant for the entire data set. Page and Srinivasan [36] emphasised the importance of the interaction between ILP and human experts in further development of the methodology. We propose here a simple postprocessing to obtain more relevant rules with higher coverage. The most interesting rules from ILP are selected and manual refinements are suggested by studying the molecules covered by the rule. One of the main reasons for performing postprocessing is that there are many ways of describing the same concept, but some may be more preferable to human experts than others. For instance, different rules explaining similar concepts may be combined to give a better rule with higher coverage. We used Prolog which is a computer language using first-order logic, to interactively modify the rules and study the new ones.

A methodological problem with postprocessing relates to the cross-validated prediction accuracy. We should perform the postprocessing within each cross-validation segment, however this often requires too much manual work. Fortunately, ILP finds rules which have high coverage, and the rules are often not influenced by single molecules being removed in cross-validation. The task which ILP tries to solve may be specified as finding a few rules which explains the variation in the data set. From a biological point of view this approach is appealing since there is often a limited number of ways that a compound can bind to a binding site. The aim is to find this set of possibilities. Removing a few samples using either cross-validation or an external test set should not change these rules. Hence, the rules with highest coverage often appear in each cross-validation loop and the postprocessing may be performed outside the loop. It should be pointed out that we are using postprocessing mainly as a step to find rules that are more interpretable, and not as a measure to improve the prediction ability.

Results

Mutagenesis data set

Table 2 gives cross-validated prediction results for both the 188 and the 42 mutagenesis dataset. We used a slightly different parameter setting (*noise=2*, *minpos=5*) compared to the original work on the atom/bond representation [6] to reduce the risk of overfitting. The samplesize option explained earlier was also not used by the original authors leading to suboptimal results. This partly explains the difference between the results reported here (84%) and in the original article (81%). The StruQT-ILP representation gives slightly

Table 2 Cross-validated prediction accuracy for classical atom/bond and StruQT-ILP representation

Data set	Representation	Accuracy (%)
Mutagenesis 188	Atom/bond	84
	StruQT-ILP	86
Mutagenesis 42	Atom/bond	91
	StruQT-ILP	91
Factor Xa	One-block anchor-GRIND	88
	Two-block anchor-GRIND	84
	StruQT-ILP	88

McNemar's test [39] was used to compare the prediction methods for the mutagenesis data set which showed that the results are not significantly different

higher prediction accuracy (86%) than the original atom/bond representation (84%), but the difference is not significant. This is not very surprising since the mutagenesis data set is the best known example where the original representation gives excellent results. It is also important to realise that our new representation resembles the original atom/bond representation in many ways. When quantitative bond properties are replaced by a suitable qualitative scale and some of the flexibility in the StruQT-ILP representation is removed, the two representations are equivalent. While the atom/bond representation uses *ad-hoc* typing of chemical properties from the original modelling software, StruQT-ILP is a generic and principled way to form these types. There is another difference in the current implementation: The atom/bond representation includes atomic partial charges which we have not presently included into our representation. In fact, most rules obtained with the atom/bond representation involves predicates using the partial charges.

Even though prediction accuracy provides an objective measure for model performance, other aspects are equally important. A model should be comprehensive and allow the domain expert to draw relevant conclusions. In this context, we continue by studying the obtained rules, their coverage and the importance of postprocessing.

The raw ILP rules for both the atom/bond and StruQT representation are given in Fig. 2. Both approaches result in five different rules. The rules are reported in the Prolog language and require some explanation. For instance, the first ILP-StruQT rule is reported as active(A):– chem_ring(A, B), laplacian_lteq(A, B, 0.14672) which translates into english as ''A molecule named A is active if it has a chemical ring B with a laplacian value larger than or equal 0.14672''. The number of reported digits should not be stressed to much by the reader as standard ILP is rather limited in its ability to deal with numerical values. It must use discretisation algorithms to handle them, and the default handling is rather simplistic as it uses the numerical values present in the active compound used as seed example. Hence, the value 0.14672 is the laplacian value at the ring in the seed example. This handling of numerical values results in a large search space with a subsequent increase in the risk of overfitting. Even though our new representation includes more numerical values, the results are improved since the included descriptors are relevant for prediction. The fact that ILP generates rules which are easily translated into plain English is an important advantage. Chemists are often interested in structural features of a compound which cause activity, and ILP creates such rules directly. Other alignment-free SAR methods often have much larger problems when it comes to model interpretation.

The first ILP-StruQT rule which involves a numerical AIM descriptor requires a more detailed explanation. We are going to use this rule both as an example of the interpretation of AIM theory and how to improve the rules using postprocessing. The rule states that ''a compound is active if it has a chemical ring with a laplacian value lower than 0.147''. The laplacian value measures the curvature of the electron density. Aromatic rings typically have a low curvature since the electrons are delocalised around the ring. The laplacian value is even lower than usually observed in aromatic rings which indicates that it is surrounded by chemical structure allowing further delocalisation effects. Based on the chemical structures present in the mutagenesis data set, one possible explanation is that the ring is bonded to two benzene rings. We then propose a new rule stating that a molecule is active if it has a chemical ring bonded to two benzene rings in the way illustrated in Fig. 3. The relevance of this rule is supported by Debnath et al. [28] where an indicator variable related to the presence of three or more fused rings improved their regression model. Four examples of compounds with high mutagenicity which are explained by this rule are shown in Fig. 4. Previous ILP studies have as far as the authors know not found similar rules despite the fact that this feature was detected by chemical experts as important in the original article [28]. Finding this rule using the simpler atom/bond representation requires rules of very long length which are prohibited due to efficiency considerations. It is also encouraging to observe that this rule has a much higher coverage than any other rule reported in the literature. Even though the postprocessed rule does not contain any AIM descriptors, they have been important in finding the rule. Hence, we conclude that our new representation are able to find

Fig. 2 The ILP rules obtained with both the atom/bond and StruQT representation. The number of positive and negative molecules which each of rule covers are given in parenthesis

Atom/bond representation		StruQT representation	
Rule 1:	(61/2)	Rule 1:	(55/2)

`active(A) :-`
`atm(A,B,c,27,C),`
`lteq(C,-0.076).`
		`active(A) :-`
		`chem_ring(A,B),`
		`laplacian_lteq(A,B,0.14672).`	

Rule 2:	(41/2)	Rule 2:	(41/0)

`active(A) :-`
`atm(A,B,c,29,C),`
`gteq(C,0.008).`
		`active(A) :-`
		`carbon_5_ring(A,B).`	

Rule 3:	(15/1)	Rule 3:	(14/1)

`active(A) :-`
`atm(A,B,o,40,C),`
`lteq(C,-0.411).`
		`active(A) :-`
		`benzene(A,B),`
		`electron_density_gteq(A,B,0.02064).`	

Rule 4:	(17/0)	Rule 4:	(41/1)

`active(A) :-`
`atm(A,B,c,10,C),`
`atm(A,D,c,10,C),`
`bond(A,B,D,1).`
		`active(A) :-`
		`benzene(A,B),`
		`laplacian_lteq(A,B,0.14933),`
		`electron_density_gteq(A,B,0.01901).`	

Rule 5:	(6/1)	Rule 5:	(5/0)

`active(A) :-`
`atm(A,B,n,38,C),`
`C=0.817.`
		`active(A) :-`
		`hetero_aromatic_5_ring(A,B),`
		`ring_size_5(A,C),`
		`laplacian_lteq(A,C,0.38479).`	

new rules and knowledge not provided by the more simple atom/bond representation which therefore justify the new methodology. It is also interesting to see that ring critical points may provide a compact description of a ring structure along with its surrounding. The atom/bond description does not have a corresponding structural feature.

Another interesting rule cover 42 positive examples and states that ''a molecule is active if it has a five-membered carbon ring''. From a chemical point of view this rule is strange since this structure is sometimes located in the interior and sometimes in the exterior of the molecule. However, the high coverage clearly indicates that this structure is of importance. From a study of the compounds covered by the rule, we find that they also have a benzene ring fused to the five-membered carbon ring. This illustrates that post-processing of the rules is important in building relevant and reliable rules. These two rules alone have a prediction accuracy of 84.5% which is impressive since King and collegues [37] needed between 10 and 14 rules to obtain similar prediction accuracy. The latter fact also illustrates another important lesson from our work: The usage of samplesize is able to reduce the number of rules considerably as our theory with the atom/bond representation contain only five rules. A theory consisting of fewer rules is clearly more comprehensible.

The 42 mutagenesis data set consists of only 12 active compounds[1] and the single rule reported in [6] which is given in Fig. 5(a) covers 8 of these. This leads to a prediction accuracy of 90.5% and we do not expect to obtain better results using our new representation. The rule obtained from the StruQT-ILP representation which is presented in Fig. 5(b) does not contain AIM descriptors, but it is still different. The structures are equal only when the ring atoms are carbon, the double bond is between a nitrogen and oxygen atom and another oxygen atom is added to structure (a). This illustrates the problem with the generality of ILP rules: ILP is designed to find the most general rules consistent with accuracy and coverage. However, the StruQT-ILP rule is preferable since it is more related to chemical structures actually present in the mutagenesis data set. There is no evidence in the data that the ring contains anything except carbon atoms and the less general rule is often preferred by chemists. It also specifies that the ring is bonded to a nitro group instead of the more general concept that it is two atoms Y and Z which have a bond of order 2. When several equally general rules exist, the choice made by ILP is arbitrary and not always the one preferred by domain experts.

[1] King et al. [6] wrongly classified one of the inactive compounds as active resulting in 13 active compounds. This explains the higher accuracy reported here.

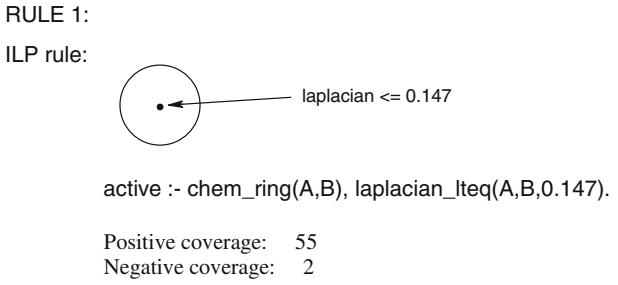


Fig. 3 Illustration of postprocessing of rules for the 188 data set

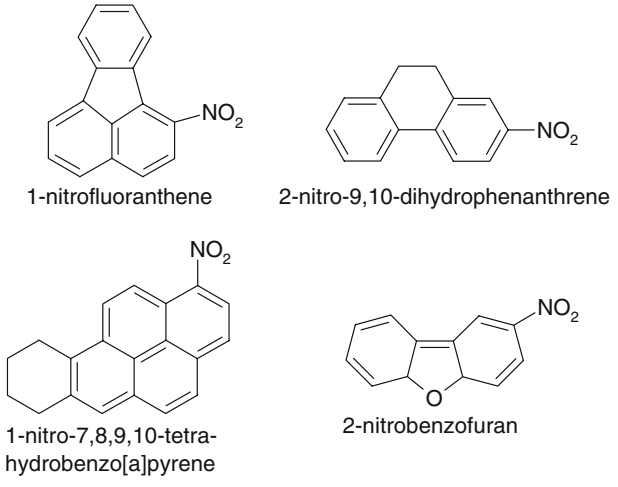


Fig. 4 Four compounds of high mutagenicity explained by the first postprocessed rule using the StruQT-ILP representation

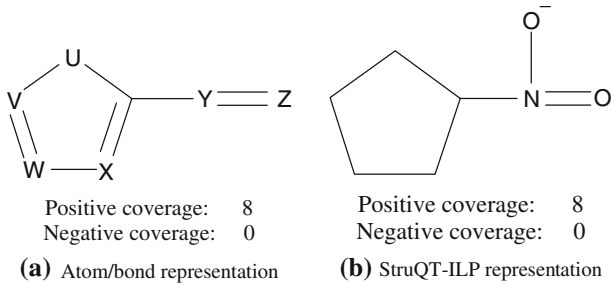


Fig. 5 The ILP rules learned on the 42 dataset. (**a**) Structure found using original atom/bond representation where atoms U-Z are not necessarily carbon. (**b**) Structure found using StruQT-ILP representation

Even though postprocessing of rules was not necessary in this situation, similar situations may be resolved using this strategy.

Factor Xa data set

The aim of using the Factor Xa data set was to compare the StruQT-ILP methodology with another alignment-free methodology, anchor-GRIND. Prediction accuracies for both methods are given in Table 2. The one-block anchor-GRIND model describes only the interaction between the anchor point and specific positions of the molecular interaction fields named R group which are chemical groups susceptible for modifications. They are specific to one or a subset of compounds present in the studied series of molecules. The two-block model also includes descriptors of the interactions within each R group. For the Factor Xa problem, the two-block model does not provide new information not already described by the one-block model, and it introduces only noise leading to a worse prediction accuracy. Our new methodology gives equal prediction accuracy as the best of the anchor-GRIND models. The ILP rules have the advantage that they are easily translated into plain english and presented in a form suitable for chemists and other experts.

A graphical presentation of theories with varying number of rules are given in Fig. 6. The largest theory which is the one reported in Aleph by default has seven rules. Beyond this point, the added rules only explain one more active example, and are therefore not important. Both the calibration and test set prediction accuracy increase gradually with the number of rules. We see that the prediction accuracy is almost 80% after only two rules. This is an important point for interpretation. We also see from the same figure how the number of active compounds which remains to be

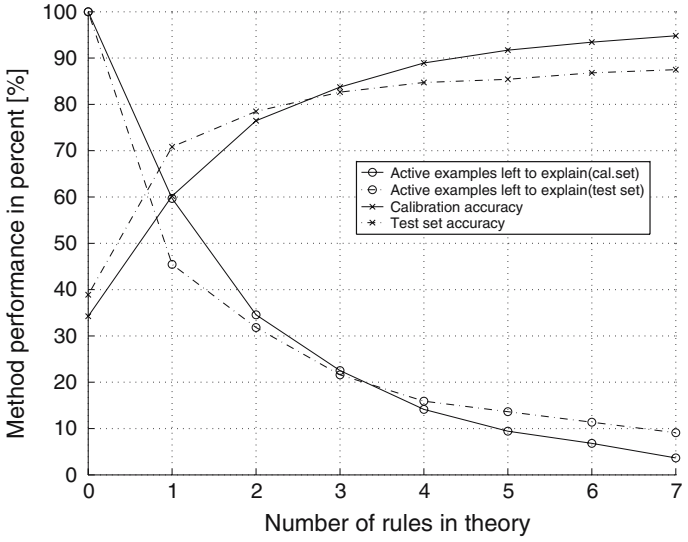


Fig. 6 The prediction accuracy of the calibration and test set at varying number of rules for the factor Xa data set. The percent-wise number of active compounds which remains to be covered by the theory are also given

explained decreases gradually with increasing number of rules in the theory.

Discussion

To solve SAR problems one needs both a representation which is comprehensible and easily interpretable and a data analysis method that can make use of the knowledge in an effective way. Although SAR problems are relational they have traditionally been handled using attribute-based methods. This has been possible through either the corresponding use of global descriptors such as hydrophobicity, and the loss of ability to describe structure in detail; or through fixed reference frames and sophisticated statistical methods to try and recapture relational information made inaccessible in processing (e.g. CoMFA).

The ability of ILP to handle relational problems directly allows more relevant and flexible representations. They do not require the extra coding which is often necessary for attribute-based methods to capture relational data. Even though the atom/bond representation was an important step in this direction, it neglects much of the chemical knowledge gained after the postulation of the molecular structure hypothesis. Our new method is the first which is both firmly rooted in quantum mechanics, encoding the molecular electronic structure effectively and related to classical chemical concepts. Straightforward extensions of the representation as presented here are possible: Integration of various properties over the atomic basin allows the computation of atomic properties such as atomic charge, dipole moment and volume. Cartesian coordinates of all critical points takes our representation into the realm of 3D QSAR. Including more knowledge about critical points such as the bonding radius, more detailed data from the Hessian matrix and energy values can be used to construct more flexible models. We have only discussed AIM theory as a topological analysis of the electron density, but AIM theory includes much more than this. Another field which can be subjected to topological analysis is the laplacian showing where the electron density is locally concentrated or depleted. Its critical points explain the concept of an electron pair which has been central to chemistry for the last 90 years and the well-known valence shell electron pair repulsion (VSEPR) model. Other concepts such as Lewis acids and bases are also related to the critical points of the laplacian field.

Even though ILP has valuable advantages over attribute-based learning algorithms, it is important to be aware of its limitations and drawbacks such as its difficulty in dealing directly with numbers. The development to include probability theory into ILP has not come sufficiently far to allow efficient handling of uncertainty. However, this problem is being investigated in the literature [38].

We want to emphasise the importance of combining ILP with chemical intuition and interpretation. We have presented two ways of doing this: Our new representation provides knowledge about chemical concepts which may be important for understanding a variety of problems. Postprocessing of rules is also important to find the ones with highest relevance. Following this approach produced rules with significantly higher coverage at comparable prediction errors. The advantage of our new representation is expected to increase as ILP methodology is further developed to better handle numerical data. We intend

to study more novel problems to investigate the ability of our StruQT-ILP representation to provide chemical knowledge about problems formulated as SAR problems.

Acknowledgement This work was supported by The Norwegian Research Council (grant no. 154265/V40).

References

1. Hansch C (1969) Acc Chem Res 2:232
2. Hansch C, Dunn WJ III (1964) J Am Chem Soc 86:1616
3. Hall LH, Kier LB (1991) In: Lipkowitz KB, Boyd DB (eds) Reviews in computational chemistry, vol 2. VCH Publishers, New York, pp 367–422
4. Cramer RD III, Patterson DE, Bunce JD (1988) J Am Chem Soc 110:5959
5. Nienhuys-Cheng SH, de Wolf R (1997) Foundations of inductiv logic programming, volume 1228 of Lecture notes in artificial intelligence. Springer-Verlag, Berlin
6. King RD, Muggleton SH, Srinivasan A, Sternberg JE (1996) Proc Natl Acad Sci USA 93:438
7. Srinivasan A, Page D, Camacho R, King RD (2006) Mach Learn
8. Srinivasan A, King RD (1999) Data Min Knowl Disc 3:37
9. Finn P, Muggleton S, Page D, Srinivasan A (1998) Mach Learn 30:241
10. Marchant-Geneste N, Watson KA, Alsberg BK, King RD (2002) J Med Chem 45:399
11. Enot DP, King RD (2003) Lecture Notes in Artificial Intelligence 2838:156
12. Nattee C, Sinthupinyo S, Numao M, Okada T (2005) In Lecture notes in artificial intelligence vol 3430, pp 92–111. Springer-Verlag, Berlin
13. Srinivasan A, King RD, Bain ME (2003) J Mach Learn Res 4:369
14. Bader RFW (1990) Atoms in molecules: A quantum theory. Number 22 in International series of monographs on chemistry. Clarendon Press, Oxford
15. Alsberg BK, Marchand-Geneste N, King RD (2000) Chemometr Intell Lab 54:75
16. Alsberg BK, Marchand-Geneste N, King RD (2001) Anal Chim Acta 446:3
17. Chaudry UA, Popelier PLA (2004) J Org Chem 69:233
18. Smith PJ, Popelier PLA (2004) J Comput Aid Mol Des 18:135
19. Chaudry UA, Popelier PLA (2003) J Phys Chem A 107:4578
20. O'Brian SE, Popelier PLA (2002) J Chem Soc Perkin Trans 2:478
21. Popelier PLA, Chaudry UA, Smith PJ (2002) J Chem Soc Perkin Trans 2:1231
22. O'Brian SE, Popelier PLA (2001) J Chem Inf Comput Sci 41:764
23. Popelier PLA (1999) J Phys Chem A 103:2883
24. O'Brian SE, Popelier PLA (1999) Can J Chem 77:28
25. King RD, Marchand-Geneste N, Alsberg BK (2001) Linköping electronic articles in Computer and Information Science 6
26. Muggleton S, De Raedt L (1994) J Logic Programming 20:629
27. Kersting K, De Raedt L (2002) Basic principles of learning bayesian logic programs
28. Debnath AK, Lopez de Compadre RL, Debnath G, Shusterman AJ, Hansch C (1991) Anal Chim Acta 34:786
29. Fontaine F, Pastor M, Zamora I, Sanz F (2005) J Med Chem 48:2687
30. Pastor M, Cruciani G, McLay I, Pickett S, Clementi S (2000) J Med Chem 43:3233
31. Fontaine F, Pastor M, Sanz F (2005) J Med Chem 47:2805
32. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Zakrzewski VG, Montgomery JA, Stratmann RE, Burant JC, Dapprich S, Millam JM, Daniels AD, Kudin KN, Strain MC, Farkas O, Tomasi J, Barone V, Cossi M, Cammi R, Mennucci B, Pomelli C, Adamo C, Clifford S, Ochterski J, Petersson GA, Ayala PY, Cui Q, Morokuma K, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Cioslowski J, Ortiz JV, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Gomperts R, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Gonzalez C, Challacombe M, Gill PMW, Johnson BG, Chen W, Wong MW, Andres JL, Head-Gordon M, Replogle ES, Pople JA (1998) Gaussian 98 (Revision A1). Gaussian Inc., Pittsburgh PA
33. Onchoke KK, Hadad CM, Dutta PK (2004) Polycycl Aromat Compd 24:37
34. MORPHY98 – A program written by P.L.A. Popelier with a contribution from R.G.A. Bone. UMIST, Manchester, England
35. Srinivasan A ALEPH: A learning engine for proposing hypothesis. http://www.web.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/aleph.pl.
36. Page D, Srinivasan A (2003) J Mach Learn Res 4:415
37. Srinivasan A, King RD, Muggleton SH (1999) The role of background knowledge: using a problem from chemistry to examine the performance of an ILP program. Technical Report PRG-TR-08-99, Oxford Univsersity Computing Laboratory, Oxford
38. De Raedt L, Kersting K (2004) Lecture Notes in Artificial Intelligence 3244:19
39. McNemar Q (1947) Psychometrika 12:153