Structure Activity Relationships (SAR) and Pharmacophore Discovery Using Inductive Logic Programming (ILP)

Michael J. E. Sternberg[a]***and Stephen H. Muggleton**[b]

[a] Structural Bioinformatics Group, Centre for Bioinformatics, Department of Biological Sciences, Imperial College London, South Kensington, London SW7 2AZ, UK
[b] Computational Bioinformatics Laboratory, Department of Computing, Imperial College London, South Kensington, London SW7 2AZ, UK

Abstract

The application of Inductive Logic Programming (ILP), a form of machine learning, to derive structure activity relationships (SAR) and to discover pharmacophores is reported. The ILP approach was initially applied to model 1D SARs in terms of the attributes of the molecules. Subsequently 2D ILP SARs were developed describing chemical connectivity. Finally ILP has been used to model 3D SARs in which the conformation of the pharmacophore can be described. ILP has advantages over many other widely used methods as it can reason with relations and hence discover chemical substructures and 3D features without these aspects having been explicitly encoded prior to learning. In particular, there is no requirement for a structural superposition. Additionally, the results of ILP provide chemical descriptions that can readily be understood by a medicinal chemist. In several trials, ILP-based SARs have been shown to be significantly more accurate than widely-used methods.

1 Introduction

The derivation of structure activity relationships (SARs) is central to molecular modelling. SARs are widely used in the systematic design and refinement of pharmaceutical agents and in the identification of structural alerts of toxicity and mutagenicity. Because of their widespread importance both in fundamental and commercial research, many methodologies have been developed (see other articles in this volume and for example [1–7]. In this article we describe the application of logic-based reasoning (using inductive logic programming, ILP) to SAR [8, 9]. We will show how this method naturally enables one to encode and reason with chemical connectivity, 3D stereochemistry, include information from positive (i.e. active) and negative compounds, does not require molecular superposition, and can express the resultant rules in a form readily understandable to chemists.

* To whom correspondence should be addressed: e-mail: m.sternberg@imperial.ac.uk, Tel.: +44-(0)20-7594-5212, Fax: +44-(0)20-7594-5264

Key words: ▭

Abbreviations: ■

2 Some Major SAR Methodologies

We begin by describing a few major approaches for SAR that have been widely used. This account will inevitably include some generalisations but the strategies described have been selected as they have been, or still are, widely used approaches. In several studies ILP-based SAR were compared against their performance on test data sets. We need to identify limitations in these approaches to highlight where the ILP can provide an alternative methodology. However we acknowledge that every machine learning strategy, including ILP, has its strengths and limitations and today often a combination of approaches provides the best strategy to develop a SAR.

A widely-used early approach, extensively explored by Hansch, employed regression on global attributes of the compounds such as hydrophobicity, molecular refractivity and other properties e.g. [10]. This approach is effectively 1D taking no account of the chemical structure of the molecules. However it was extended to 2D by including stereochemical based indicator variables that identified the presence or absence of substructures. The choice of these substructures was by inspection; typically outliers from the 1D SAR would be identified and common chemical sub structures identified. Subsequent work on 1D and 2D SAR employed neural networks instead of regression e.g. [11]. The limitations of these types of approaches are:

Dateiname: Q10005.3D
Pfad: P:/verlage/wiley/qsar/SPALTEN/
Setzer: Segeritz
TrennProgramm: English
Status: 1. AK von Seite 1 bis 6
Pagina: 1
Seite: 1 te von 6
Umfang (Seiten): 6
Auftragsnummer: 03-1434
3B2-Version: 6.05d

- The attribute set, in particular the stereochemical features, have to be defined and new attributes will not be learnt.
- The resultant predictive algorithm can be difficult to understand.
- The strategy only works well when there is a common molecular scaffold
- 3D information is not included

Alongside these methods, 2D structure based pharmacophore analysis was used, for example using graph searches for sub-structures [12–14]. This approach can identify key chemical substructures amongst a diverse set of molecules but is unable to include global properties, such as hydrophobicity, in the analysis.

Recently attention has focussed on describing the 3D properties of the molecules. A widely used approach is CoMFA [3, 4] (comparative molecular field analysis) (see the dominance of these methodologies in SAR papers published in J. Med. Chem. over the last few years). Typically the lowest energy conformer of each molecule is identified and superposed in 3D. Molecular properties such as steric repulsion, hydrophobicity, hydrogen-bonding acceptor/donor potential are mapped on to an enclosing grid. Partial least squares analysis is then used to derive the SAR. The correlation results can then be mapped back onto the molecular structure and inspected visually to provide stereochemical insight into possible ligand/receptor interactions. A major limitation of this approach is that alignment rules are required to superpose the different structures. The CoMFA fields that are generated are highly dependent on the choice of atoms that were superposed. Another recent approach is the CATALYST package [15] (from Accelrys, San Diego, CA, USA). This quantitatively superposes point pharmacophores to derive a QSAR. The method does not reason with the internal chemical substructures and has difficulties in scaling to large data sets (100s of compounds).

3 Inductive Logic Programming (ILP)

ILP [16, 17] is a subfield of Machine Learning (ML) [18]. ML involves the automatic construction of high-level knowledge from low-level data. Subfields of ML are separated largely according to the way in which the learned knowledge is represented (e.g. Hidden Markov Models, Bayes' nets, decision trees and Logic Programs). ILP uses Logic Programs (a computationally efficient fragment of Mathematical Logic) for representing example data, background knowledge and hypotheses. Mathematical Logic is one of the longest-standing and most versatile approaches to representation of scientific knowledge. It is extensively used throughout Computer Science. The ability of ILP to make use of explicit background knowledge within an expressive representation language is particularly powerful in complex scientific applications.

We now explain the basis of inductive logic programming as a machine learning method. The approach learns from known examples or observations (i.e. it employs the reasoning known as induction). The observations, the background knowledge and the resultant rules are expressed as first order logic programs, such as compound no 21 contains atom no 12. A typical ILP procedure is illustrated in Figure 1. The observations form the examples and in ILP both positive examples (e.g. active molecules) and negative examples (inactive molecules) can be used. The background knowledge describes features of these examples and can encode properties that are 1D (such as hydrophobicity), or 2D (chemical connectivity) or 3D (spatial relationships). The learning engine then employs an algorithm that identifies which combinations of the background knowledge best cover as many of the positive examples whilst covering the fewest number of negative examples. The resultant rule is then output by the learning algorithm. This rule can be stored and then a further cycle of learning undertaken to identify rules for positive examples not previously covered by the first rule. Alternatively the learnt rule can be added back into the background knowledge. The learning algorithm employs a measure known as compression (c) rule that describes the power of the rule in maximising the number of positive examples covered by the rule (p) whilst having the fewest number of negative examples also predicted (n).

$$c = p - n.$$

In addition, one often includes a measure that considers the simplicity of the learnt rule – how many pieces of background knowledge are included in the rule (b) and the objective function becomes

$$c = p - n - b.$$

The implementation of this procedure of course depends on the precise algorithm. Our work used programs developed by Muggleton and coworkers. We initially used GOLEM [19] and subsequently used the algorithm PROGOL [20]. Both programs encode the examples, the background knowledge and the resultant rules using the language PROLOG. The learning algorithm can be in any language

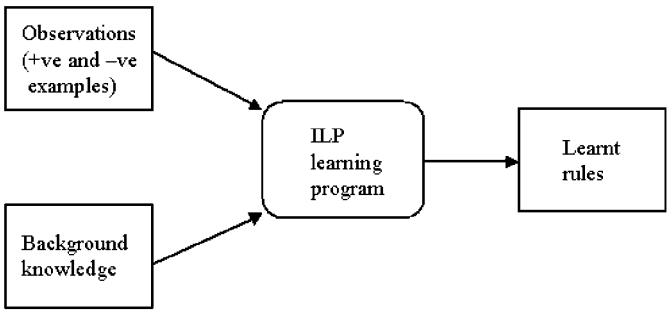

Figure 1. Flow diagram of Inductive Logic Programming (ILP)

– currently PROGOL has been written in C and in PROLOG.

4 Potential Advantages of ILP

There are several potential advantages of ILP in developing a SAR

1. ILP uses relationships rather than attributes and logic can infer new relationships. One encodes that atom A is bonded to atom B and that atom B is bonded to atom C. The program can infer that atom A is connected to atom C via atom B without this having to be encoded. Thus new stereochemical features required for accurate modelling can be learnt rather than having to be explicitly encoded as attributes. Attribute-based approaches, including regression, neural networks, CoMFA and CoMSIA, are unable to learn new attributes being only able to model the inter-relationships of existing attributes.
2. Logic-based methods learn from both the active and inactive set of molecules. Many superposition approaches, including graph-based searching for common pharmacophores, focus only on the active set of molecules.
3. Both CoMFA and CoMSIA require molecular superposition of the series of molecules and the choice of atoms to be superposed is often ambiguous and markedly affects the entire subsequent SAR. When there is a common scaffold superpositions can be made relatively easily, but with a diverse set, such as those obtained via high throughput screens, superposition is a major problem. Logic-based SAR do not require superposition as they can encode the internal atom-atom distances and reason with geometry.
4. The resultant rules are expressed as logic programs. These can readily be converted into descriptions readily understandable to a chemist (e.g. the active molecules contain atom A bonded to B and B is then bonded to E).

ILP has certain limitations. The logical representation does not handle numerical calculations readily. Strategies can easily be developed to model interatomic distances (see below). However the quantitative activity of each molecule cannot be input into the learning. Thus ILP does not model a quantitative SAR. Instead molecules are classified as active or inactive, or a rank order of activity is learnt. The ILP models a qualitative SAR.

5 1D ILP-based SAR

The initial study involving ILP was reported in 1992 [8]. The program GOLEM was used to model the inhibition of trimethoprim analogues on *Escherichia coli* (*E. coli*) dihydrofolate reductase (DHFR) [21]. A training set of 44 compounds were used and the resultant algorithm tested on 11 further compounds [22]. The trimethoprim derivates had substituents at the 3, 4 and 5 positions. The chemical properties of these substituents were encoded manually in terms of attributes such as polarity, flexibility, size, hydrogen bond donor and hydrogen bond acceptor potential. A limitation of ILP is that it has limited capacity for quantitative reasoning. To circumvent this, the affinity of each pair of compounds were compared and expressed as compound A is more active than compound B. ILP was then used to derive rules to predict that one compound would be more active than another in terms of the chemical attributes of the substituents.

There are two major measures of the success of the SAR. First is, of course, predictive accuracy on data not used in training. We considered the prediction of ranked activity on the 11 compounds not used for learning. The result was that the rank correlation by ILP was 0.46 compared to that using Hansch regression of 0.42 – the difference is just below significance at the 5% level. The second measure is the insight into the stereochemistry. We examined the features in the rules describing active compounds and these suggested that the 3 and 5 positions should have properties of the methoxy (OCH_3) substituent whilst the 4 position should be polar. These rules agreed with the x-ray crystallographic structure of trimethoprim – *E. coli* DHFR complex.

The results of this initial study were therefore encouraging. We were only able to derive a qualitative structure-activity relationship since only the rank orders and not the numerical binding affinities were modelled. However in terms of rank, the results on test data were nearly significantly better than the widely used regression method. Importantly, the resultant rules could be interpreted steroechemically.

The general 1D ILP SAR approach was further compared to neural networks on the DHFR/trimethoprim series [23]. In addition the inhibition of DHFR by triazines was studied comparing ILP, neural networks and linear regression [24]. The conclusion from the first study was borne out by these further analyses. ILP will produce rules of comparable accuracy to neural networks and regression, ILP offered the advantage of generating rules that were easier to understand.

6 2D ILP-base SAR

The next major development in applying ILP to SAR was to use a 2D representation of the chemical connectivity of the molecules [9]. First order logic is ideally suited to describe relations between atoms and to identify important more complex chemical substructures not encoded initially. The data set was 229 aromatic and heteroaromatic nitro compounds tested for mutagenesis by the Ames test by the Hansch group [2]. The set is chemically diverse and cannot be superimposed onto a common template and therefore presents a challenge to SAR methodologies. In the

Table 1. Accuracy of different SAR approaches in predicting mutagenicity

Dataset	SAR Algorithm	Accuracy %	
		without indicator variables	with indicator variables
188 regression friendly	Regression	85.2	89.3
	Regression + sq	83.0 (*)	88.8
	Neural networks	86.2	89.4
	CART	82.5 (*)	88.3
	PROGOL I	81.4 (†)	–
	PROGOL II	87.8	–
42 regression unfriendly	Regression	66.7 (†‡)	66.7 (†‡)
	Regression + sq	71.8 (†‡)	69.0 (†‡)
	Neural networks	64.3 (†‡)	69.0 (†‡)
	CART	83.3	83.3
	PROGOL I	85.7	–
	PROGOL II	83.3	–

Accuracy is defined as (no of correct prediction)/(no of predictions made). Regession + sq is regression with squares.
(*) – Accuracy significantly worse ($P < 0.1$) than PROGOL I
(†) – Accuracy significantly worse ($P < 0.025$) than PROGOL I
(‡) – Accuracy significantly worse ($P < 0.025$) than PROGOL II

initial study be Debnath and coworkers, the data set was divided by inspection into 188 compounds considered to be amenable to regression and a further 42 that could not be readily modelled by regression. A subsequent study using neural networks [25] also used this division. Both studies used an attribute representation involving the energy of the lowest unoccupied molecular orbital (LUMO) and the molecular hydrophobicity (the octanol/water patrician coefficient, LogP). In addition two binary indicator variables were introduced (after manual expert inspection) to describe chemical features of the subsets of the compounds.

The ILP program PROGOL [20] was used in this study. The molecular representation used was obtained by inputting each molecule into a standard modelling program (QUANTA, Molecular Simulations, Burlington, MA) and obtaining a typing of the atoms and the connecting bonds. Thus:

atom(127, 127 1,C, 22, 0.191)

stated that in compound no 127, the atom no 1 was a Carbon of quanta type 22 with a partial charge of 0.191, and

bond(127, 127 1,127 6,7)

stated that in compound 127, atom no 1 and atom no 6 are connected by a bond of type 7 (aromatic). Two ILP representations were explored (I and II). In ILP representation I only the above features, which were generated automatically, were used. A more extensive representation of the molecules was also developed (II). This used the information of I with the LUMO and LogP information together with simple logical statements (PROLOG programs) that identified certain high level chemical concepts from representation I. These high level concepts included methyl groups, aromatic rings etc.

The study then compared ILP against our implementation of regression (linear and linear with squares) and neural networks to derive the SAR. In addition a decision trees was applied to the data (CART) [26]. The data set was divided into compounds that were high or low mutagencity. Each learning study was subject to leave-one-out cross validation to obtain an average accuracy. The results are given in Table 1. On the regression friendly data, PROGOL II performed comparable to the other methods when they used the manually derived indicator variables. On the 42 regression unfriendly compounds, CART and PROGOL I obtained significantly better predictions that regression or neural networks.

A series of structural alerts for mutagenicity were automatically generated by PROGOL. Importantly PROGOL was able to generate a new structural feature for mutagenicity for the 42 regression unfriendly compounds which stated that there is a double bond conjugated to a five membered aromatic ring.

7 3D ILP-based SAR

It is well recognised that the stereochemistry of molecules is often crucial in deriving a SAR. The next major step in ILP-based SAR was introducing a 3D representation [27]. A simple PROLOG program was incorporated as background knowledge that took as input the atomic coordinates of two atoms and generated their distance of separation. The system studied was the angiotensin converting enzyme (ACE) inhibitors that are a widely used form of medication for hypertension. The data set of 28 compounds was

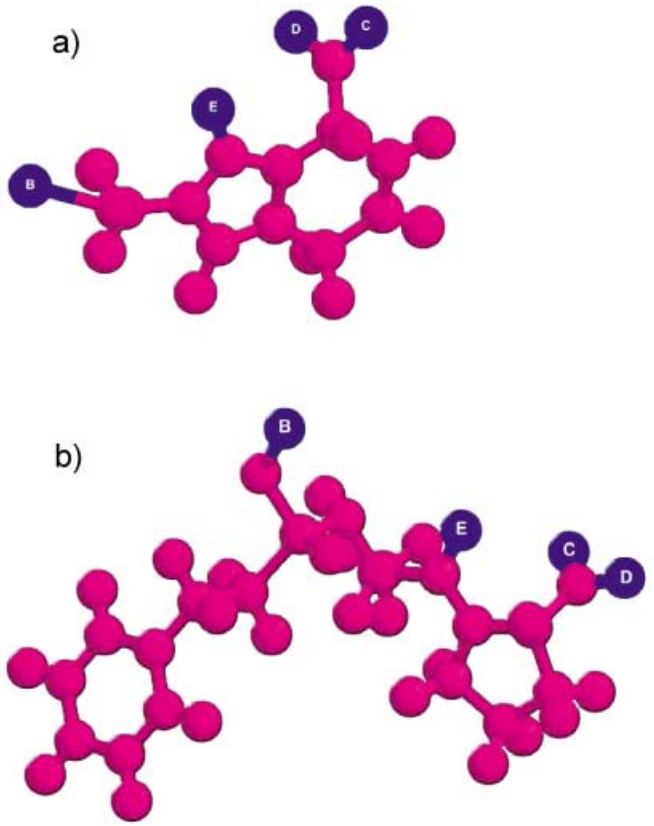

Figure 2. ACE inhibitors with the four-point pharmacophore highlighted.

previously studied by Mayer et al [28] who proposed a pharmacophore based on a set of postulated 3D structures for the molecules and a zinc binding site.

The first study using PROGOL aimed to rediscover the Mayer pharmacophore. Each molecule was represented in the conformation proposed by Mayer. Atoms and bonds were represented in a similar approach to the 2D study on nitro compounds. In addition, the potential for an atom to be a hydrogen bond donor or an acceptor was included. A specific rule was included to define a single putative zinc binding site for each of the molecules. Modifications to the learning algorithm were included to cope with positive only data and to generate rules for pharmacophore discovery including the largest number of atoms. The result was that PROGOL could rediscover a four point pharmacophore which was similar to that proposed by Mayer (Figure 2). The rule learnt was

Molecule A is an ACE inhibitor if:
it can bind to zinc at a site B, and
it contains a hydrogen acceptor C, and
the distance between B and C is 7.9 ± 1.0 Å, and
it contains a hydrogen acceptor D, and
the distance between B and D is 8.5 ± 1.0 Å, and
the distance between C and D is 2.1 ± 1.0 Å, and
it contains a hydrogen acceptor E, and
the distance between B and E is 4.9 ± 1.0 Å, and
the distance between C and E is 3.1 ± 1.0 Å, and
the distance between D and E is 3.8 ± 1.0 Å.

The second study addressed the problem that the conformer with the lowest energy might not be the active stereochemistry due to either limitations in modelling and/or that the bound conformer is not that of lowest energy. Ten low energy conformers were generated for each compound using modelling software. In addition, rules based on stereochemistry were introduced to predict potential zinc binding sites from the chemical properties of the atoms in the molecules. Thus in this study the zinc binding site (or sites) would have to be discovered by the program. The results showed that there might be two locations for zinc binding and indeed this had been suggested by others [29].

8 3D ILP Compared to CoMFA

A recent study by King and coworkers [30] has compared 3D ILP SAR to CoMFA on two systems – thermolysin zinc protease inhibitors and glycogen phosphorylase inhibitors. The key aspect of CoMFA [3] is that the molecules must first be superimposed in 3D and this can be difficult and subjective when they present a diversity of substructures. In 3D ILP that uses internal distances (as in section 7), there is no requirement for an initial superposition. In the King and co-workers study, CoMFA was performed using standard modelling software (Sybil, Tripos Associates, St Louis, MO). The results obtained were comparable in the cross validated squared correlation coefficient regression (r^2_{cv}) to those obtained by other groups studying these systems. The ILP employed the typing of atom and bonds, the inclusion of electrostatic charge, the use of chemical knowledge to define chemical entities (such as aromatic five carbon ring); hydrogen bond donors and acceptors and internal distances. The ILP program used was Aleph (developed by A Shrinivasan). One study employing the three lowest energy conformers for each molecule and for both data sets ILP performed significantly better ($P < 0.01$) than CoMFA [3] implemented by the King group. The descriptions of the pharmacophores were in agreement with those proposed by other groups.

9 Concluding Remarks

We have described a series of studies in which SARs were derived using ILP. In each of the datasets there was some commonality of chemical structure, but this could be diverse (e.g. aromatic and heteroaromatic nitro compounds). For each of these datasets chemical structures related to activity (pharmacophores or structural alerts) were derived and expressed in a form that can be interpreted by a chemist (as opposed to weights in a regression or a neural network). Major benefits of using ILP to derive SARs are that chemical substructures can be learnt without having been

previously identified and that there is no requirement for an initial 3D superposition.

Comparisons of accuracy of methods are difficult as on each test system there often are cycles of methodology refinement prior to the final study. In addition, the numbers of compounds are often too few to prove that differences of accuracy are significant. However the general conclusion from the studies described above is that ILP SARs are at least as accurate as many widely used approaches. In some studies, such as compared to an automatic implementation of CoMFA [30], ILP was shown to yield significant improvements in accuracy over widely used approaches.

One difficult with ILP is that at present it is non trivial to implement. Unlike regression, neural networks or decision trees, one cannot simply run a standard package. The present state of the software is that there is a substantial learning curve to use ILP effectively. In addition, there has been a very limited number of person years invested to date in applying ILP to SARs. The effort has been on the development of the methodology (from 1D, via 2D to 3D) rather than on developing a user-friendly program for general adoption. We consider that the approach is now ready for the development of a system suitable for widespread use by the community. Indeed one requires the use by many groups on diverse datasets to identify the areas for its improvement.

References

[1] Marshall, G. R., and Cramer, R. D., *Trends Pharm. Sci., 9*, 285 – 289 (1988).
[2] Debnath, A. K., Lopez de Compadre, R. L., Debnath, G., Shusterman, A. J., and Hansch, C., *J. Med. Chem., 34,* 786 – 797 (1991).
[3] Cramer, R. D., Patterson, D. E., and Bunce, J. D., *J. Am. Chem. Soc., 110*, 5959 – 5967 (1988).
[4] Klebe, G., Abraham, U., and Mietzner, T., *J. Med. Chem., 37*, 4130 – 4146 (1994).
[5] Klebe, G., *J. Mol. Med.*, *78*, 269 – 281 (2000).
[6] Livingstone, D. j., *J. Chem. Inf. Comput. Sci.*, *40*, 195 – 209 (2000).
[7] Winkler, D. A., *Briefings in Bioinformatics*, *3*, 73 – 86 (2002).
[8] King, R. D., Muggleton, S., Lewis, R. A., and Sternberg, M. J. E., *Proc. Nat. Acad. Sci. USA*, *89*, 11322 – 11326 (1992).
[9] King, R. D., Muggleton, S. H., Srinivasan, A., and Sternberg, M. J. E., *Proc. Nat. Acad. Sci. USA*, *93*, 438 – 442 (1996).
[10] Hansch, C., *Acc. Chem. Res.*, *2*, 232 – 239 (1969).
[11] Aoyama, T., Suzuki, Y., and Ichikawa, H., *J. Med. Chem.*, *33*, 2583 – 2590 (1990).
[12] Brint, A. T., and Willett, P., *J. Mol. Graph.*, *5*, 49 – 56 (1987).
[13] Willett, P., *J. Mol. Recognit.*, *8*, 290 – 303 (1995).
[14] Brown, N., Willett, P., Wilton, D. J., and Lewis, R. A., *J. Chem. Inf. Comput. Sci.*, *43*, 288 – 297 (2003).
[15] Barnum, D., Greene, J., Smellie, A., and Sprague, P., *J. Chem. Inf. Comput. Sci.*, *36*, 563 – 571 (1996).
[16] Muggleton, S. H., *Inductive Logic Programming*. Academic Press, **1992**.
[17] Muggleton, S. H., *Artficial Intelligence*, *114*, 283 – 296 (1999).
[18] Mitchell, T., *Machine Learning*. McGraw Hill, **1997**.
[19] Muggleton, S. H., and Feng, C., in *Inductive Logic Programming*, S. H. Muggleton, ed., Academic Press, **1992**, pp ■.
[20] Muggleton, S., *New Generation Computing Journal*, *13*, 245 – 286 (1995).
[21] Hansch, C., Li, R.-I., Blaney, J. M., and Langridge, R., *J. Med. Chem.*, *25*, 777 – 784 (1982).
[22] Roth, B., Rauckman, B. S., Ferone, R., Baccanari, D. P., Champness, J. N., and Hyde, R. M., *J. Med. Chem.*, *30*, 348 – 356 (1987).
[23] Hirst, J. D., King, R. D., and Sternberg, M. J. E., *J. Comp. Aided. Mol. Design*, *8*, 405 – 420 (1994).
[24] Hirst, J. D., King, R. D., and Sternberg, M. J. E., *J. Comp. Aided. Mol. Design*, *8*, 421 – 432 (1994).
[25] Villemin, D., Cherqaoui, D., and Cense, J. M., *J. Chim. Phys.*, *90*, 1505 – 1519 (1993).
[26] Breiman, L., Friedman, J. H., and Olshen, R. A., *Classification and Regression Trees*. Wadsworth, Belmont, **1984** .
[27] Finn, P., Muggleton, S. H., Page, D., and Srinivasan, A., *Machine Learning*, *30*, 241 – 271 (1998).
[28] Mayer, D., Naylor, C., Motoc, I., and Marshall, G., *J. Comput. Aided Mol. Design*, *1*, 3 – 16 (1987).
[29] Bohacek, R., Lombaert, S. D., McMartin, C., Priestle, J., and Grutter, M. *J. Am. Chem. Soc.*, *118*, 8231 – 8249 (1996).
[30] Marchand-Geneste, N., Watson, K. A., Alsberg, B. K., and King, R. D., *J. Med. Chem.*, *45*, 399 – 409 (2002).

Received ■; Accepted on ■