molecular pharmaceutics

articles

Discovering H-Bonding Rules in Crystals with Inductive Logic Programming

Howard Y. Ando,*,† Luc Dehaspe,‡ Walter Luyten,‡ Elke Van Craenenbroeck,‡ Henk Vandecasteele,‡ and Luc Van Meervelt§

Research Formulations, Pfizer Global Research and Development, Ann Arbor Laboratories, Ann Arbor, Michigan 48105, PharmaDM, Kapeldreef 60, B-3001 Leuven, Belgium, and Department of Chemistry, K.U.Leuven, Biomolecular Architecture, Celestijnenlaan 200F, B-3001 Leuven, Belgium

Received March 22, 2006

Abstract: In the domain of crystal engineering, various schemes have been proposed for the classification of hydrogen bonding (H-bonding) patterns observed in 3D crystal structures. In this study, the aim is to complement these schemes with rules that predict H-bonding in crystals from 2D structural information only. Modern computational power and the advances in inductive logic programming (ILP) can now provide computational chemistry with the opportunity for extracting structure-specific rules from large databases that can be incorporated into expert systems. ILP technology is here applied to H-bonding in crystals to develop a self-extracting expert system utilizing data in the Cambridge Structural Database of small molecule crystal structures. A clear increase in performance was observed when the ILP system DMAX was allowed to refer to the local structural environment of the possible H-bond donor/acceptor pairs. This ability distinguishes ILP from more traditional approaches that build rules on the basis of global molecular properties.

Keywords: Computer aided drug design; in silico modeling; crystal structure; solubility; hydrogen bonding; machine learning; inductive logic programming

Introduction

Rules and relationships are one way that knowledge is encapsulated so that it can be applied to a future situation; they are usually generalizations that were gleaned from a body of facts. As such, they often lack the specificity that may be needed for a particular situation, the statistical probability for their veracity, and a systematic means for updating them in the light of new facts. Machine learning technology, specifically inductive logic programming (ILP), also known as relational data mining,[1,2] offers one way to overcome some of these deficiencies. The technology will be applied to hydrogen bonding (H-bonding) in crystals, an area of noncovalent interactions.

H-bonding interactions are an important attribute of a drug molecule. In aqueous solutions, the hydration of pendent H-bonding donor or acceptor groups retards the permeation of a drug molecule through biological membranes.[3−6] In the crystalline state, H-bonding is one component of a crystal's cohesive energy, the net attractive noncovalent lattice

* Author to whom correspondence should be addressed. Mailing address: Research Formulations, Pfizer Global Research and Development, Ann Arbor Laboratories, Ann Arbor, MI 48105. Tel: +1 734 622-1278. Fax: +1 734 622-3609. E-mail: Howard.Ando@Pfizer.com.

† Ann Arbor Laboratories.

‡ PharmaDM.

§ K.U.Leuven.

(1) Dzeroski, S., Lavrac N., Eds. *Relational Data Mining*; Springer-Verlag: Berlin, 2001.

(2) King, R. D.; Muggleton S.; Lewis, R. A.; Sternberg M. J. E. Drug design by machine learning: The use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. *Proc. Nat. Acad. Sci. U.S.A.* **1992**, *89*, 11322−11326.

10.1021/mp060034z CCC: $33.50
Published on Web 08/16/2006

interactions that bind molecules in a crystal together and retard dissolution of lattice molecules into aqueous media. Compounds forming high cohesive energy crystals are often "discovered" by modern high throughput in vitro screening (HTS) technologies that are used in the pharmaceutical industry today, because H-bonding is one of the most important factors that determine binding of small molecules to target proteins. However, a paradox often arises. When compounds with nanomolar potency are found, they are usually very water insoluble. Such compounds have been termed high-affinity traps by Stella and Borchardt.[6]

Rules were developed by Lipinski,[7] the Rules of 5, to address the high-affinity trap paradox. These rules were gleaned from a retrospective analysis of marketed drugs, the assumption being that such drugs as an aggregate have the physicochemical properties that should serve as a baseline for molecules that are identified by HTSs. Two out of the four rules deal with H-bonding atoms: N and O. These rules place limits on these atoms and are general enough to address both permeability and the H-bonding cohesive component of high-affinity traps that produces strong lattice energy crystals. The rules themselves have changed very little over the past 15 years. This is a testament to their validity but also to the difficulty for humans to refine and adapt them to more specific situations. More specific rules regarding H-bonding have arisen out of the emerging area of crystal engineering.

A major focus in crystal engineering has been to develop a system for classifying H-bonding patterns in crystals. For inorganic crystals, Wells[8] used graph theory to describe such hydrogen-bonding patterns where atoms were represented as points and hydrogen bonds (HBs) as lines. Kuleshova and Zorky[9] extended this concept to organic molecules and classified a crystal structure database of 776 molecules. Further refinements of graph theory have been made by Etter et al.,[10,11] Davis et al.,[12] and Grell et al.[13] Motherwell[14] has automated the recognition of this type of classification in RPLUTO.

Etter[11] adapted the Cahn−Ingold−Prelog (CIP) rules for designating the absolute configuration of chiral atoms to H-bonding systems. The CIP system assigns priorities to different atoms only as a convention to assign chirality in an unambiguous way. Similarly, Etter's adaptation of the CIP system to H-bonding involves arbitrary rules that only provide a convention for consistently classifying H-bonding patterns. However, Etter has also developed some general and more functional, group-specific H-bonding rules.

The above crystal engineering approaches—while addressing different research goals—are complementary to the rule discovery techniques presented here. The present study focuses on the anticipation of crystalline H-bonding, given only the 2D structure of the molecule. In contrast, the above approaches aim at the categorization of H-bonding patterns encountered in known crystal structures (e.g., using graph sets). The results of this analysis can however be used as background knowledge by ILP (hence the complementarity). As will be explained below, some of the group-specific H-bonding rules by Etter have been used in this manner.

Although two of Lipinski's[7] four rules address H-bonding, they do not specifically address H-bonding in crystals, where close packing constraints and neighboring group effects would affect the ability for a potential intermolecular or intramolecular HB to form. Lipinski's H-bonding rules, however, do provide a convenient baseline to compare more structure and context specific rules.

One baseline for the crystalline state would be an "ideal liquid state" of the substance in which drug molecules are allowed to form cohesive HBs with all possible H-bonding atom pairs. This would include both intramolecular and intermolecular atom pairs. Atom pairs in the molecule that could participate in H-bonding would be classified as either

(3) Goodwin, J. T.; Conradi, R. A.; Ho, N. F.; Burton, P. S. Physicochemical determinants of passive membrane permeability: role of solute hydrogen-bonding potential and volume. *J. Med. Chem.* **2001**, *44*, 3721−3729. Erratum in: *J. Med. Chem.* **2002**, *45* (10), 2122.

(4) Goodwin, J. T.; Mao, B.; Vidmar, T. J.; Conradi, R. A.; Burton, P. S. Strategies toward predicting peptide cellular permeability from computed molecular descriptors. *J. Pept. Res.* **1999**, *53*, 355−369.

(5) Wang, B.; Gangwar, S.; Pauletti, G.; Siahaan, T.; Borchardt, R. T. Synthesis of an esterase-sensitive cyclic prodrug of a model hexapeptide having enhanced membrane permeability and enzymic stability using a 3-(2′-hydroxy-4′,6′-dimethylphenyl)-3,3-dimethylpropionic acid promoiety. *Methods Mol. Med.* **1999**, *23* (Peptidomimetics Protocols), 53−69.

(6) Gangwar, S.; Pauletti, G. M.; Siahaan, T. J.; Stella, V. J.; Borchardt, R. T. Synthesis of an esterase-sensitive cyclic prodrug of a model hexapeptide having enhanced membrane permeability and enzymatic stability using an acyloxyalkoxy promoiety. *Methods Mol. Med.* **1999**, *23* (Peptidomimetics Protocols), 37−51.

(7) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3−26.

(8) Wells, A. F. *Structural Inorganic Chemistry*, 3rd ed.; Clarendon Press: New York, 1962; pp 294−315.

(9) Kuleshova, L. N.; Zorky, P. M. Graphical enumeration of hydrogen-bonded structures. *Acta Crystallogr.* **1980**, *B36*, 2113−2115.

(10) Etter, M. C. Encoding and decoding hydrogen-bond patterns of organic compounds. *Acc. Chem. Res*. **1990**, *23*, 120−126.

(11) Etter, M. C.; MacDonald, J. C.; Bernstein, J. Graph-set analysis of hydrogen-bond patterns in organic crystals. *Acta Crystallogr.* **1990**, *B46*, 256−262.

(12) Davis, R. E.; Bernstein, J. Graph set analysis of hydrogen-bond patterns in molecular crystals. *Trans. Am. Cryst. Assoc*. **1999** (**1998**), *33* (ACA Transactions), 7−21.

(13) Grell, J.; Bernstein, J.; Tinhofer, G. Investigation of hydrogen bond patterns: A review of mathematical tools for the graph set approach. *Crystallogr. Rev.* **2002**, *8*, 1−56.

(14) Motherwell, W. D. S.; Shields, G. P.; Allen, F. H. Automated assignment of graph-set descriptors for crystallographically symmetric molecules. *Acta Crystallogr*. **2000**, *B56*, 466−473.

an H-donor or H-acceptor, depending on their role in the HB. For the crystalline state of matter, the Cambridge Structural Database[15−17] (CSD) of small molecules was used to obtain atom pairs in molecules that could and actually did form HBs, based on generally accepted criteria. To analyze these atom pairs, we used ILP,[1] the branch of machine learning that has the ability to deduce rules from logical principles and statistics. ILP is one of a number of artificial intelligence (AI) approaches that have been used to leverage the power of computers. The use of ILP is best illustrated by contrasting it with expert systems.

Expert systems capture human knowledge by expressing rules, that can be gleaned from human experts, in special AI languages that are designed to process facts (data), rules, and questions; Prolog is one example of such an AI language. Use of an expert system involves querying the knowledge base of rules with questions in the particular domain of interest. The limitation of expert systems is their need for rules. Humans are needed to formulate the rules from their domain of expertise. Lipinski,[7] for example, extracted from the databases of marketed drugs a set of rules that characterized their general properties. Any rule, however, is an abstraction and a generalization that may or may not apply to a specific situation and may not be in fact consistent with new data that have inevitably accumulated after its formulation. These limitations are the issues that ILP (and machine learning in general) attempts to address.

ILP uses the Prolog language to extract rules from information in databases. As such, it can use the power of computers to update its knowledge base of rules as more data are made available. In addition, the language of ILP is flexible enough to incorporate the specific characteristics of a particular domain of interest as background knowledge. And finally, ILP can use and validate human-generated rules and incorporate such knowledge as part of its background knowledge. For example, in King et al.,[2] an ILP system was used that "understands" bonds, atoms, and functional groups on organic molecules.

In the present study, ILP was used to seek rules on H-bonding from the CSD. More specifically, rules were sought that differentiate the "ideal liquid state" H-bonding from the H-bonding in crystals.

Materials and Methods

ILP Technique: Hierarchical Rule Generation with DMAX. Generation of rules was done with PharmaDM's DMAX ILP technology for the generation of hierarchical rules.[18] This technology is derived from the TILDE[19] ILP system and develops Prolog rules in a hierarchical manner following a recursive partitioning strategy, as shown in Figure 1. The most general rule (e.g., *query-1* in Figure 1) is the one with the widest range of applicability for the entire training set. Such a rule forms the basis for two alternatives or branches, a branch that satisfies this chemical rule and a branch that does not satisfy it (e.g., the *yes* and *no* databases in Figure 1). When looking for splitting rules, DMAX will select at each step the one that best discriminates between the three classes: intramolecular HB, intermolecular HB, and no H-bonding. For instance, in Figure 1, *query-1* is the rule selected by DMAX on the basis of the fact that it maximizes class-purity within the resulting *yes* and *no* databases.

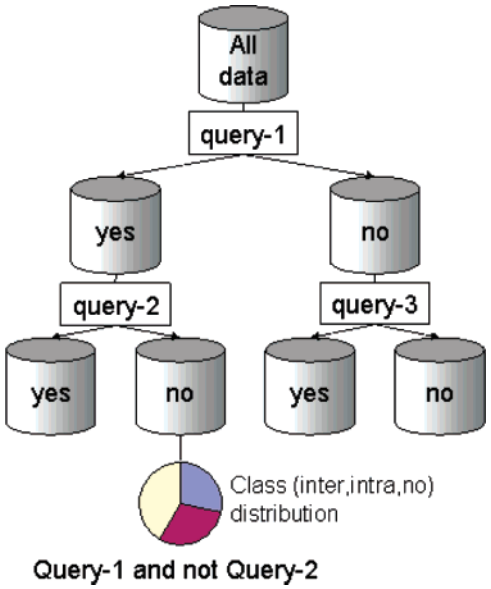

Figure 1. A decision tree resulting from hierarchical rule generation.

For each branch, more specific rules can be induced that put further constraints on the type of molecules that the rule governs (e.g., *query-1 and query-2* in Figure 1). Further rule generation for each succeeding branch produces increasingly more refined rules that only apply to smaller and more class-pure (i.e., with the majority of examples belonging to one of three classes) subsets of the training set.

Via a tree structure as shown in Figure 1, the total training data set at the top of the tree is recursively partitioned. The leaves of the tree contain data subsets that become associated with (1) a description that corresponds to the path to that leaf from the root of the tree (e.g., *query-1 and not query-2* in Figure 1); and (2) a class distribution, i.e., per class *inter*, *intra*, *no*, to represent the fractions of examples present in the leaf that form intermolecular, intramolecular, or neither intermolecular nor intramolecular HBs. To use a tree such as the one shown in Figure 1 as an expert system, new

(15) Cambridge Structural Database 2002. *ConQuest 1.5 User Guide*, database v5.24; The Cambridge Crystallographic Data Centre: Cambridge, U.K., 2002.

(16) Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr.* **2002**, *B58*, 380−388.

(17) Allen, F. H.; Taylor, R. Research applications of the Cambridge Structural Database (CSD). *Chem. Soc. Rev.* **2004**, *33*, 463−475.

(18) DMax: http://www.pharmadm.com/dmax.asp.

(19) Blockeel, H.; De Raedt, L. Top-down induction of first-order logical decision trees. *Artif. Intell.* **1998**, *101* (1−2), 285−297.

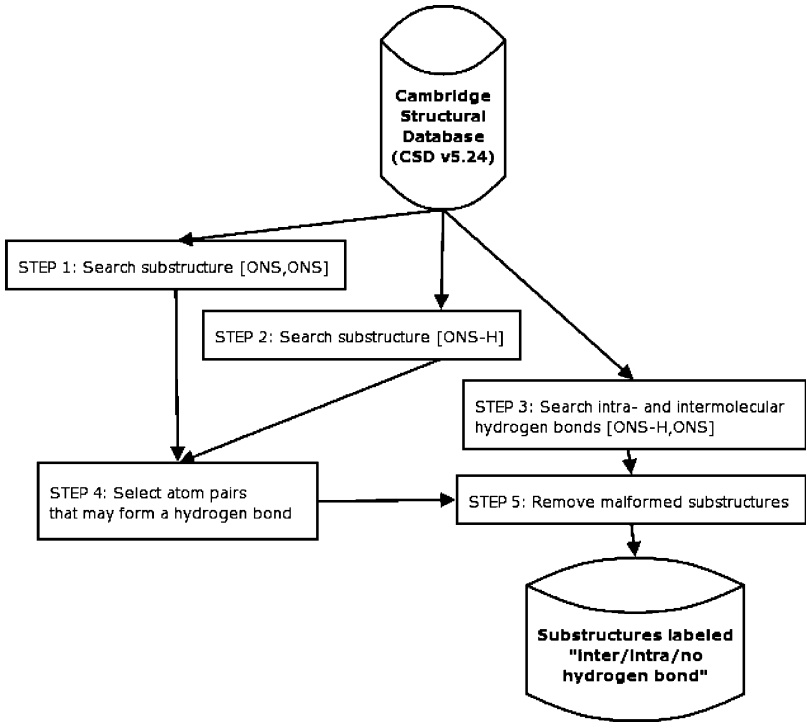

Figure 2. CSD data preparation steps 1−5.

examples are sorted down the tree, where the query in each node determines which branch the example will follow. As the example arrives in a leaf of the tree, the distribution associated with that leaf is used to assign a score between 0 and 1 per class *inter*, *intra*, *no*. The path from the root to the leaf then corresponds to an explanation of why a particular example has been associated with these three scores.

A score of 0 (i.e., 0%) for a particular class indicates that, in the training data, this class was not represented in the leaf and therefore the expert system predicts that examples sorted to this leaf do not belong to this class. At the other extreme, a score of 1 (i.e., 100%) for a particular class indicates all training data in that leaf belong to that class, which supports the prediction that new examples sorted to the same leaf also belong to that class. In practice, after pruning away those branches of the decision tree that degrade predictive accuracy, most of the class distributions found in the leaves of the decision tree are not uniform, such that most of the predicted scores for classes *inter*, *intra*, *no* are somewhere between 0 and 1. This allows us to rank the predictions per class, putting those cases where the expert system is most confident (i.e., the scores toward 1) at the top. To evaluate the quality of predictions per class, one can locate the true positive cases in the ranking. The best ranking will put all those true positives on top. More details on the evaluation procedure are given below in the Results section.

Preprocessing Data from the Cambridge Structural Database. All ILP analyses were conducted on molecules and atom pairs from the CSD version 5.24 (November 2002).[15−17] A training and validation set was selected for HB data mining in five steps as shown in Figure 2.

Step 1. Search substructure [ONS,ONS] with constraints:

all CSD filters on (CSD filters: 3D records, Rfactor < 0.05, not disordered, no errors, not polymeric, no ions, only organics); and
all atoms in the molecule are in set {C,N,O,F,P,S,Cl,Br} (this eliminates organic molecules containing Si or Mg); and
one molecule in unit cell ($Z' = 1$); and
the hydrogens are attached; and
CSD option "Normalize terminal H positions for C,N,O Defaults"
Result: 316 031 substructures [ONS,ONS] in 23 077 molecules.

(We use [ONS,ONS] to denote a substructure with two atoms that can be either O, N, or S.)

Step 2. Search substructure [ONS-H] with constraints:

all constraints of step 1; and
the distance between ONS and H is defined
Result: 17 906 substructures [ONS-H].

(Substructure [ONS-H] refers to a hydrogen bound to either O, N, or S.)

Step 3. Search intra- (respectively inter-) molecular HBs with constraints:

Step 4. Select atom pairs that may form a HB. For that purpose, the [ONS,ONS] substructures resulting from step

all constraints of step 1; and
limit to intra- (respectively inter-) molecular bonds; and
the distance between the donor and acceptor in the pair is less than or equal to the sum of the van der Waals radii of donor and acceptor plus 0.5 Å (van der Waals radii of H, O, N, and S were taken as 1.2, 1.52, 1.55, and 1.8 Å, respectively); and
the distance between the H and the acceptor is less than or equal to the van der Waals radius of the acceptor minus 0.12 Å; and
the angle donor−H−acceptor is between 100° and 180°
Result: 5044 intramolecular and 14 823 intermolecular HBs.

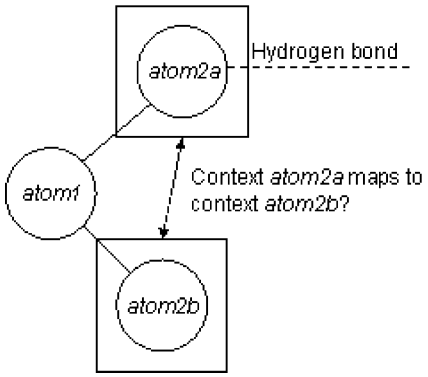

Figure 3. Detection of local symmetry during data preprocessing.

1 that share at least one atom with a [ONS-H] substructure resulting from step 2 are selected.

Result: 66 530 substructures [ONS,ONS] where at least 1 side is bound to hydrogen.

Step 5. Remove "malformed" substructures

belonging to molecules consisting of multiple parts; or
having local symmetry; or
due to errors in the structure files produced with CSD (Some substructures generated by CSD have an unexpected * or ' character after the atom identifier, or have an atom code that cannot be mapped to an atom identifier. In both cases the substructure is ignored.)

Result: 62 781 substructures labeled either "inter" (i.e., represents an intermolecular HB), "intra" (i.e., represents an intramolecular HB), or "no" (i.e., represents no HB).

In step 5, item 2, we drop cases where CSD returns only one of two possible HBs. To detect these so-called "local symmetries", we first look for atom pairs (*atom1*, *atom2a*) and (*atom1*, *atom2b*) such that in CSD *atom2a* but not *atom2b* forms a HB and *atom2a* and *atom2b* have the same atom symbol (O, N, or S); see also Figure 3. If the structural contexts of *atom2a* and *atom2b* are identical, we conclude that there is local symmetry between *atom2a* and *atom2b* and drop both atom pairs (*atom1*, *atom2a*) and (*atom1*, *atom2b*) from the dataset.

Application of ILP. The collection of labeled substructures being used as a starting point, rules were derived, tested, and deployed in four steps, as shown in Figure 4.

Step 6. Randomly split the set of labeled substructures into

training set (75%): used for derivation of rules in step 7
validation set (25%): used for evaluation of rules in step 8

Step 7. Training. Apply DMAX (see previous section) to the training set with varying sets of background knowledge activated, and add discovered rules to the expert system. Details on the background knowledge modules that were used are given below.

Step 8. Validation. Use the expert system automatically extended in the previous step to predict H-bonding for the atom pairs in the validation set. The quality of the extended knowledge was assessed via a comparison of predicted scores per class and the actual class labels.

Step 9. Deployment. Use the expert system automatically extended in step 7 to predict H-bonding labels of new substructures (i.e., substructures from molecules not in the training or validation set).

Background Knowledge Modules. Next to the data described above, background knowledge has a major impact on rules produced by DMAX. Richer and more relevant background knowledge will typically allow DMAX to construct higher quality rules. We have tested the capability of DMAX to take advantage of superior background knowledge by running experiments with increasingly sophisticated background knowledge (BK): from BK-level-0 to BK-level-2.

(A) BK-Level-0. With BK-level-0 active, DMAX has access to the presence and frequency of 95 elementary molecular substructures (rings and functional groups): ring,

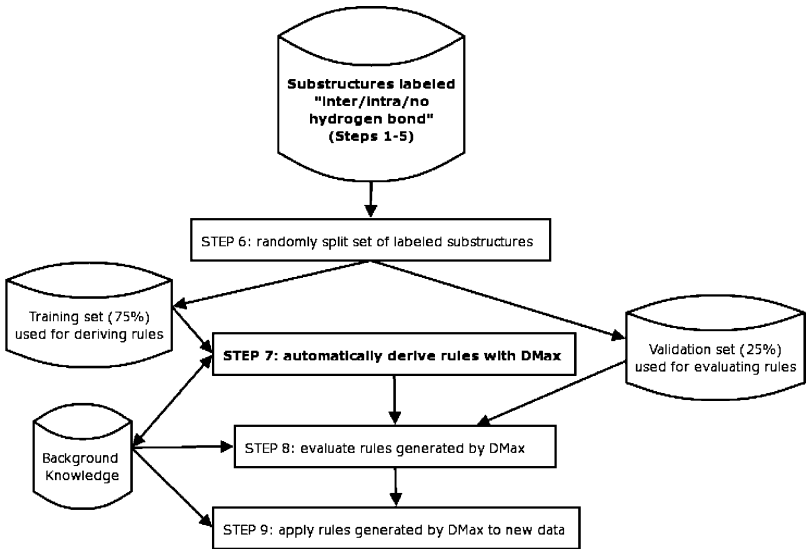

Figure 4. Application of ILP, steps 6−9.

aromatic ring, hetero-aromatic ring, non-hetero-aromatic ring, benzene ring, non-aromatic ring, hetero-non-aromatic ring, non-hetero-non-aromatic ring, pyrrole ring, furan ring, thiophene ring, pyrazole ring, imidazole ring, pyridine ring, pyridazine ring, pyrimidine ring, pyrazine ring, methyl group, phosphate group, phosphonate group, phosphinate group, miscellaneous phosphor group, acylhalide, halide, carboxylic ester, thio-*O*-carboxylic ester, methoxygroup, ether, carboxylic acid, thio-*O*-carboxylic acid, alcohol, conjugated base of a carboxylic acid, conjugated base of a thio-*O*-carboxylic acid, oxide, ketone, aldehyde, diazo group, azide, nitro group, nitrile, iminium ion, amide, thioamide, sulfonamide, sulfinamide, oxime, thioxime, imine, hydroxylamine, thiohydroxylamine, amine, *n*-hydroxyamide, *n*-sulfanylamide, hydroxyammonium, sulfanylammonium, ammonium ion, nitroso group, thio-*S*-carboxylic ester, dithiocarboxylic ester, thioether, thio-*S*-carboxylic acid, dithiocarboxylic acid, thiol group, conjugated base of a thio-*S*-carboxylic acid, conjugated base of a dithiocarboxylic acid, sulfide, *n*-hydroxythioamide, *n*-sulfanylthioamide, sulfoxide, sulfinic acid, sulfinic ester, conjugated base of a sulfinic acid, sulfonic acid, sulfonic ester, conjugated base of a sulfonic acid, sulfone, metal ion, counterion, heteroatoms, aliphatic chain, general functional group, general nonammonium acid, general ester, general non-amine base, general ether, general -ol group, general -on group, general amide, general amine, general ammonium, and general oxime.

Extra moieties that are considered to be "good" donors or acceptors according to Etter[20] were added: phenol, aniline, urea, and imide.

This BK level does not require ILP. The training data can be converted to a single table, with one row per candidate HB and a column for each of the moieties considered. The cells of that table would contain the frequency of that moiety in the molecule where the candidate HB was found. This table associated with BK-level-0 can be processed with any machine learning method. This is not the case for the BK-levels > 0.

(B) BK-Level-1. With BK-level-1 active, DMAX can detect the 95 moieties listed above (cf., BK-level-0), plus relationships between those moieties (e.g., connected, fused, or linked by aliphatic chain), plus relationships between those moieties and the candidate donor and acceptor.

This BK-level (and the next one) requires ILP. The relationships cannot be expressed with a non-ILP machine learning technique. At this level also the difference from traditional structure−activity-relationship (SAR) approaches shows. These typically rely on a set of global molecular descriptors (e.g., fingerprints or physicochemical properties). In contrast, an ILP system such as DMAX can also build rules that refer to the local context of the candidate donor and acceptor. It thus becomes possible to distinguish between donor−acceptor pairs within the same molecule that do and those that do not form an HB. Such a distinction is not possible with a method that can only refer to global molecular properties: in such a method atom pairs taken from that molecule will receive the same vector of global descriptors, even if the pairs belong to different H-bonding classes.

(C) BK-Level-2. For this highest BK level—which includes BK-level-0 and BK-level-1—some concepts were added that were estimated to be particularly relevant for H-bonding:

- conjugated paths to groups that donate or accept electrons; and
- the number of moieties (from the list of 95 shown above) within a specified distance (in terms of bonds); and
- a description of the shortest path (via covalent bonds) between donor and acceptor or from donor/acceptor to moiety (from the list of 95 shown above):
  - distance (in bonds) of path; and
  - whether the path is part of a moiety, how many atoms it shares with that moiety, and
  - number of single, double, or aromatic bonds on the path

Table 1. Data Preprocessing Statistics

HB label	no. of atom pairs	% of total
only intra	4 230	6.74
only inter	12 700	20.23
both intra and inter	601	0.96
none	45 250	72.08
total	62 781	100

Results

Data Processing. Table 1 shows that, out of the 62 781 "ideal liquid state" atom pairs, only about 28% could be classified as "intra" and/or "inter". The remaining 72% of the potential atom pairs did not take part in crystalline H-bonding. Notice that exceptionally—in 1% of the cases—a single atom pair receives both "intra" and "inter" labels. This occurs when, according to CSD, the two atoms in the pair meet the distance and angle constraints (see step 3 for details) not only when the search is limited to intramolecular bonds but also when the search is limited to intermolecular bonds. Since rules for "intra" and "inter" were learned separately, the 1% doubles was added to both sets. In total, the data set for learning "inter" and "intra" rules consisted of 8% and 21% positive examples, respectively.

The total data set was randomly split into

training set:	47 086 atom pairs (75% of total)
validation set:	15 695 atom pairs (25% of total)

Inductive Logic Programming. (A) Overall Performance of the Generated Expert System. A set of hierarchical rules was constructed from the 47 086 examples in the training set. These rules were then applied to the 15 695 cases in the validation set, such that each of these cases received a score between 0 and 1 for each the three classes *inter*, *intra*, and *no*. Per class, the examples were sorted (in descending order, see Materials and Methods section above), and for each position *n* in the resulting ranking, the number of true positives in the top *n* was counted. The corresponding cumulative response curves are shown in Figure 5.

(20) Etter, M. C. Hydrogen bonds as design elements in organic chemistry. *J. Phys. Chem.* **1991**, *95*, 4601−4610.

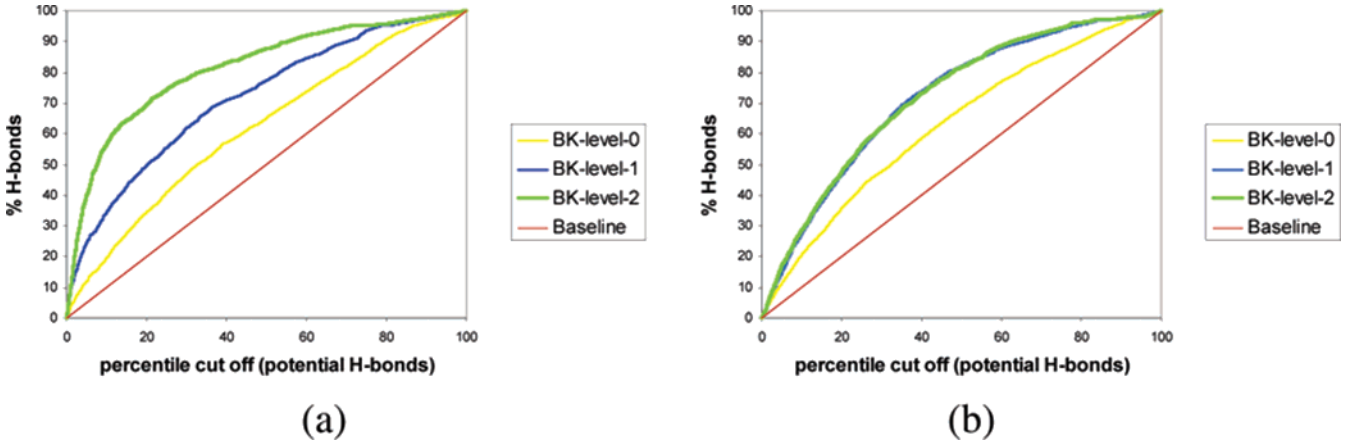

Figure 5. Cumulative response curves showing performance of expert system with varying background knowledge (BK-level-0 to -2) on two tasks: (a) intramolecular HB prediction and (b) intermolecular HB prediction. *Baseline* is a trivial expert system that models the "ideal liquid state" and classifies all examples as both intra- and intermolecular.

On the basis of the experimental data summarized in Figure 5, we can draw the following conclusions:

(1) Whatever the level of background knowledge used, the expert system generated with DMAX using 2D structural information only always significantly outperformed the "ideal liquid state" model that classifies all examples as both intra- and intermolecular: see BK-level-0 versus baseline in Figure 5a,b. Recall that in BK-level-0 only global molecular descriptors were used. Apparently these descriptors, even if they do not refer to the particular context of the atom pair, do carry information that is relevant for predicting HBs.

(2) In both the intra- and intermolecular HB cases, the quality of predictions significantly improves by using background knowledge that requires ILP: see BK-level-1 versus BK-level-0 in Figure 5a,b.

(3) HB-specific background knowledge leads to superior quality predictions of intramolecular HBs (see BK-level-2 versus BK-level-1 in Figure 5a). This is not the case for intermolecular HBs (see overlapping curves for BK-level-2 versus BK-level-1 in Figure 5b). Although the prediction of intermolecular HBs clearly benefits from the use of ILP (cf. previous item), it does not benefit from the HB specific background knowledge of BK-level-2. This observation is consistent with the intuition that intermolecular phenomena are inherently harder to predict from structural properties than intramolecular ones.

(4) The ILP system DMAX does take advantage of increasingly more sophisticated background knowledge. For instance, for intramolecular HB prediction (Figure 5a), with BK-level-2 the top 10% of the ranking contains about 3 times more true intramolecular HBs than with BK-level-0. This phenomenon is further illustrated in the next paragraph.

(B) Examples of Superior Predictions with Increasing Levels of Background Knowledge. As shown in Figure 5, DMAX is able to assign higher scores to true inter- and intramolecular HBs when it has access to higher levels of background knowledge. Below are some examples—all taken from the validation set—of such corrections obtained by moving from BK-level-0 to BK-level-1 and from BK-level-1 to BK-level-2. It will become clear from the DMAX-generated rules underlying the predictions that the extra knowledge available at BK-level-1 and BK-level-2 is indeed crucial for achieving the superior classification.

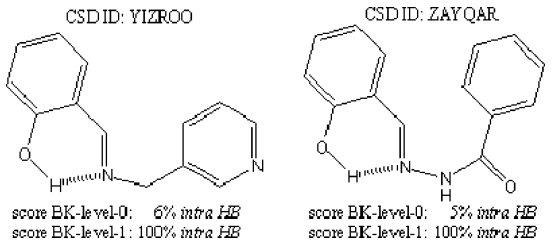

Figure 6. Two examples (with CSD identifiers YIZROO and ZAYQAR) of *intramolecular* HBs with higher scores on BK-level-1 (with ILP) than on BK-level-0 (without ILP). The HB is marked with a hatched line between the hydrogen and the nitrogen.

(1) Step 1: From BK-Level-0 to BK-Level-1. With BK-level-0 the two intramolecular HBs in Figure 6 get low scores (5% and 6%), such that they end up at the bottom of the ranking. In contrast, with BK-level-1 they get the maximum score (100%) and thus contribute to the lift of the BK-level-1 curve in Figure 5a. The BK-level-1 rule that assigns this maximum score to both examples in Figure 6 is as follows:

If	the HB donor is an oxygen that *is part of* a phenol ring;
and	an imine *containing* the HB acceptor *is connected by a single bond* to the phenol ring *in ortho position* with respect to the alcohol *containing* the HB donor;
and	the HB acceptor *is not part of* a non-aromatic ring;
and	the compound contains fewer than 7 general functional groups and at least 16 atoms;
then	there is an intramolecular HB in 100% of the cases.
(Rule coverage:	42 cases in training set, 96 cases in total dataset.)

Notice that the above rule relies heavily on the extra relational knowledge available from BK-level-1 onward. The phrases generated by DMAX to establish relationships between moieties, HB donor and HB acceptor, are shown in italics. In a non-ILP approach such phrases are absent, which explains the inferior performance of BK-level-0.

Figure 7 shows two intermolecular HBs that benefit from the additional BK-level-1 knowledge: in both cases the assigned score increases by 86% (from 14% to 100%). The DMAX-generated rule responsible for this improvement is as follows (ILP specific phrases are again shown in italics):

If	the HB acceptor *is part of* an amine that *is fused to* a non-aromatic ring;
and	on the non-aromatic ring there is a ring type *substituent in para position* with respect to the amine;
and	the HB donor is an oxygen;
and	the HB donor *is not part of* a ring with the above non-aromatic ring *fused* at *distance 3*;
and	apart from the amine, no other functional group *is fused to* the non-aromatic ring;
and	the compound contains at most one alcohol, fewer than 8 general functional groups, and at most 53 atoms;
then	there is an intermolecular HB in 100% of the cases.
(Rule coverage:	13 cases in training set, 27 cases in total dataset.)

The examples in Figure 6 and Figure 7 illustrate that intra- and intermolecular HB cases, respectively, move up in the ranking when BK-level-1 is used. A third way to improve performance is to assign lower *intra* and *inter* scores to the cases that do not represent a HB, i.e., cases belonging to class *no*. An example that gets a higher score for *no* will have lower *intra* and *inter* scores (sum of three scores is 1 for each example) and tend to move to the bottom of the "inter" and "intra" rankings. Two such examples are presented in Figure 8.

The rule covering both examples in Figure 8 can be interpreted as a constraint on HB formation:

If	the candidate HB acceptor is an oxygen that *is part of* a general ether *connected by a single bond to* an aromatic 6-ring which is not a phenol;
and	the candidate HB acceptor *is not part of* a non-aromatic ring;
and	the general ether *is not in ortho position* with a ring type substituent;
and	the general ether *is not linked to* a ketone *via* an aliphatic chain;
then	there is no HB involving the candidate HB acceptor in 90% of the cases.
(Rule coverage:	319 cases in training set, 607 cases in total dataset.)

(2) Step 2: From BK-Level-1 to BK-Level-2. The above examples mainly illustrate the advantages of the ILP setting compared to the non-ILP setting (BK-level-0). In this paragraph, we further show the benefits of additional, HB-specific background knowledge using examples that get more correct scores on BK-level-2 than on BK-level-1.

The cases shown in Figure 9 are covered by the rule below. As before, the phrases not available at the previous background knowledge level are in italics: for instance, information on the path between donor and acceptor and on conjugated systems is included in BK-level-2, but not in BK-level-1.

If	the *path between HB donor and HB acceptor shares 2 atoms with* a non-hetero-aromatic ring;
and	that *path contains at most 2 single bonds and exactly 1 double bond*;
and	the HB acceptor *is not electron donating as part of a conjugated system*;
and	there is no thioamide that *shares 2 or fewer atoms with the above path*;
and	there is no phosphorus group that *shares no atoms with the above path*;
and	the HB acceptor is not an oxygen, nor part of a general ether, nor part of an iminium ion;
then	there is an intramolecular HB in 100% of the cases.
(Rule coverage:	66 cases in training set, 214 cases in total dataset.)

Notice that, as before, the additional—in this case HB specific—background knowledge is predominant in rules that improve classification of intramolecular HBs.

The rule that corresponds to the examples in Figure 10 is as follows:

If	the HB acceptor *is electron donating as part of a conjugated system*;
and	the HB acceptor is part of a general non-amine base that is not linked via an aliphatic chain to a general -ol group;
and	there is no aliphatic chain or general ammonium that *shares no atoms with the path between the HB donor and the HB acceptor*;
and	there is no 5-ring that *shares exactly 2 atoms with the above path*;
and	there is no iminium ion that *shares fewer than 2 atoms with the above path*;
and	the *covalent bond distance between HB donor and HB acceptor* differs from 9;
and	the compound contains fewer than 9 general functional groups;
then	there is an intermolecular HB in 98% of the cases.
(Rule coverage:	51 cases in training set, 105 cases in total dataset.)

The first condition in the rule above, i.e., whether the HB acceptor is electron donating as part of a conjugated system, is the first split criterion selected by DMAX. This criterion holds for 3404 examples in the validation set, 90% of which belong to class "no". Recall that the expected fraction of "no" cases in a randomly drawn sample is 72% (cf. Table 1). A binomial test reveals that the probability of finding 90% or more "no" cases in a sample of size 3404 is extremely low ($<10^{-159}$), so from a statistical perspective the following hypothesis can be accepted: the fact that a candidate HB acceptor is electron donating as part of a conjugated system has a negative influence on its ability to participate in a HB.

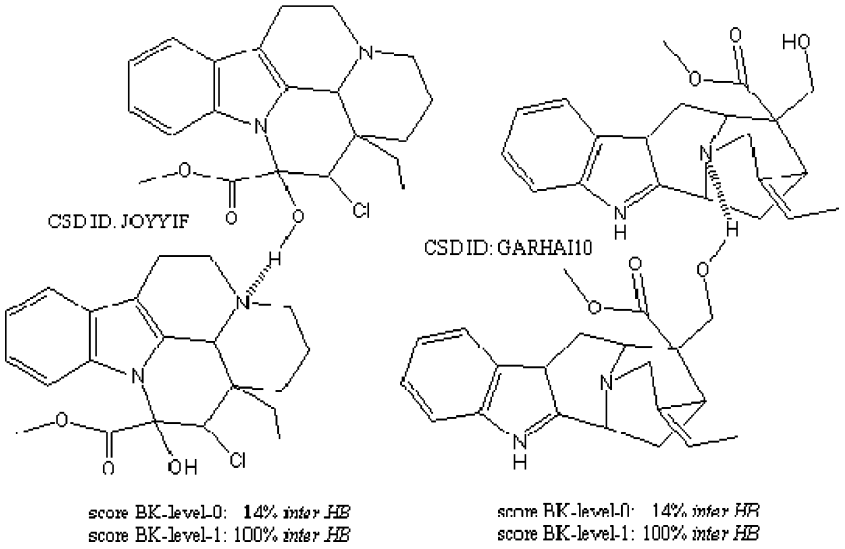

Figure 7. Two examples (with CSD identifiers JOYYIF and GARHAI10) of *intermolecular* HBs with higher scores on BK-level-1 (with ILP) than on BK-level-0 (without ILP).

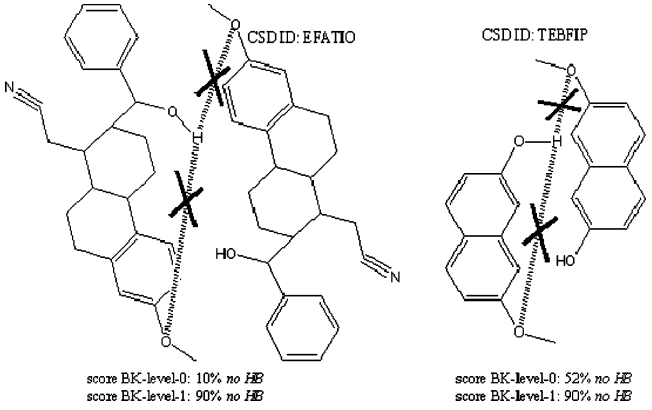

Figure 8. Two examples (with CSD identifiers EFATIO and TEBFIP) where there are no HBs (cf. the crossed out inter- and intramolecular HBs in the drawing) and that get higher scores for class *no* on BK-level-1 (with ILP) than on BK-level-0 (without ILP).

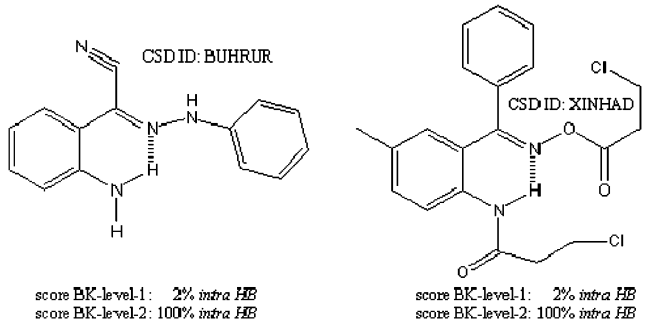

Figure 9. Two examples (with CSD identifiers BUHRUR and XINHAD) of *intramolecular* HBs with higher scores on BK-level-2 than on BK-level-1.

This general constraint also makes sense chemically since HBs form between hydrogens that have a positive partial charge, due to their polarized covalent bond to an electronegative atom, and an H-acceptor bearing a partial negative

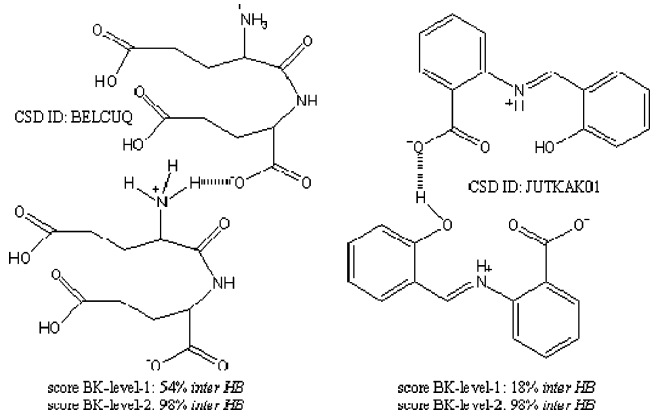

Figure 10. Two examples (with CSD identifiers BELCUQ and JUTKAK01) of intermolecular HBs with higher scores on BK-level-2 than on BK-level-1.

charge (typically due to a lone pair). If the H-acceptor donates electrons as part of a conjugated system, this will tend to shift away (lone pair) electrons and make the H-acceptor less negatively charged, thus making it less likely to form a (strong) HB (less likely, but not impossible as can be seen in Figure 10 and the corresponding rule). The "general non-amine base" mentioned as the second condition in this rule pertains mostly to carboxyl groups as acceptors. The carboxylate anion can be viewed as a resonance hybrid of the two anionic structures, or as a conjugated system of three interacting p-orbitals containing four electrons. Since COO^- is the (non-amine) base of the corresponding acid COOH, and the C═O double bond and the negative $C-O^-$ charge are delocalized, both oxygens may be able to function as acceptors (all the more so because free rotation is typically possible around the carbon bond connecting the carboxylate group to the rest of the molecule). This resonance may offset the otherwise unfavorable decrease in the negative charge on the oxygen atom, allowing an intermolecular HB to be formed.

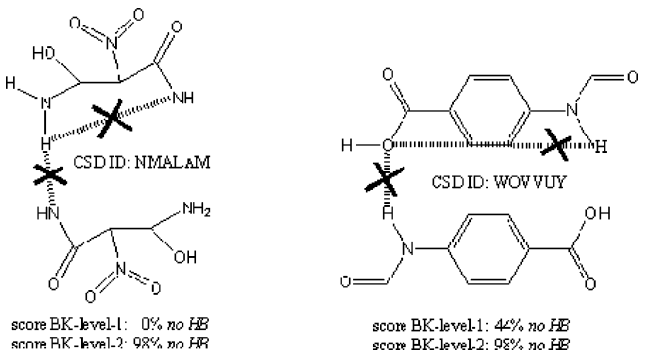

Figure 11. Two examples (with CSD identifiers NMALAM and WOVVUY) where there are no HBs (cf. the crossed out inter- and intramolecular HBs in the drawing) and that get higher scores for class *no* on BK-level-2 than on BK-level-1.

To conclude this section on illustrations of BK-level impact, we present in Figure 11 two "no" cases that are better classified due to the background knowledge about conjugated systems (cf. discussion above).

The DMAX-generated rule that covers the examples in Figure 11 is as follows:

If	both the candidate HB acceptor and donor *are electron donating as part of a conjugated system*;
and	the candidate HB acceptor is not part of a general non-amine base, general ester, ring, or oxide;
and	the c*ovalent bond distance between candidate HB donor and acceptor* differs from 3;
and	there is no aniline, aromatic ring, or hetero-non-aromatic 6-ring that *shares two atoms with the path between candidate HB donor and acceptor*;
and	there is no thio-*S*-carboxylic ester or thioether that *shares no atoms with the above path*;
and	there is no furan ring that *shares 3 atoms with above path*;
then	there is no HB involving the candidate HB acceptor and donor in 98% of the cases.
(Rule coverage:	649 cases in training set, 1298 cases in total dataset.)

Conclusion

Almost all modern computational chemistry techniques are based on the analysis of molecular properties. Such applications include multiple regression, artificial neural networks, and support vector machines. Very few address the interactions between different parts of the molecule and the interactions of molecules with one another. In this study, we have demonstrated that ILP rules that characterize H-bonding in crystals can be induced from properly preprocessed 2D structural data.

Atom pair information on H-bonding was extracted from the CSD, and the strongest rules that govern these data were induced in a hierarchy from the most general to more specific. Because these rules are induced from data with minimal human interaction, they can be readily updated as new data accumulate and are thus much more adaptable than human rules obtained by laborious reflection. Moreover, ILP has the ability to analyze massive amounts of data consistently. It also has the ability to incorporate information in a variety of formats into background knowledge. In this respect, it offers superior flexibility over other machine learning techniques such as support vector machines. Finally, expert systems that use ILP hierarchies provide a much finer degree of prediction because a given molecule can be analyzed for the specific rules that may apply to it alone. Thus ILP expert systems would be expected to outperform expert systems that rely on global molecular properties.

Recently,[21] the concept of incorporating ILP into an artificial scientist, which can analyze data, postulate hypotheses, and design experiments in a recursive fashion, has been demonstrated. Such applications of machine learning are a small step in using computational machines to expand our knowledge base and codify knowledge and information into rules that can be continuously updated.

MP060034Z

(21) King, R. D.; Whelan, K. E.; Jones, F. M.; Reiser, P. G. K.; Bryant, C. H.; Muggleton, S. H.; Kell, D. B.; Oliver, S. G. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature* **2004**, *427*, 247−252.