New Approach to Pharmacophore Mapping and QSAR Analysis Using Inductive Logic Programming. Application to Thermolysin Inhibitors and Glycogen Phosphorylase *b* Inhibitors

Nathalie Marchand-Geneste,*,† Kimberly A. Watson,‡ Bjørn K. Alsberg,† and Ross D. King†

Department of Computer Science, Computational Biology Group, The University of Wales Aberystwyth, Penglais Campus, Aberystwyth, Ceredigion SY23 3DB, Wales, England, and Laboratory of Molecular Biophysics, Department of Biochemistry, University of Oxford, South Parks Road, OX1 3QU, Oxford, England

Received April 12, 2001

A key problem in QSAR is the selection of appropriate descriptors to form accurate regression equations for the compounds under study. Inductive logic programming (ILP) algorithms are a class of machine-learning algorithms that have been successfully applied to a number of SAR problems. Unlike other QSAR methods, which use *attributes* to describe chemical structure, ILP uses *relations*. This gives ILP the advantages of not requiring explicit superimposition of individual compounds in a dataset, of dealing naturally with multiple conformations, and of using a language much closer to that used normally by chemists. We unify ILP and standard regression techniques to give a QSAR method that has the strength of ILP at describing steric structure with the familiarity and power of regression methods. Complex pharmacophores, correlating with activity, were identified and used as new indicator variables, along with the comparative molecular field analysis (CoMFA) prediction, to form predictive regression equations. We compared the formation of 3D-QSARs using standard CoMFA with the use of ILP on the well-studied thermolysin zinc protease inhibitor dataset and a glycogen phosphorylase inhibitor dataset. In each case the addition of ILP variables produced statistically better results ($P < 0.01$ for thermolysin and $P < 0.05$ for GP datasets) than the CoMFA analysis. Moreover, the new ILP variables were not found to increase the complexity of the final QSAR equations and gave possible insight into the binding mechanism of the ligand−protein complex under study.

Introduction

The problem of learning structure−property relationships is central to all applications of molecular design. In particular, in the biological sciences, quantitative structure−activity relationships (QSARs) have been studied for many years in order to either elucidate biological processes or develop new drugs. These studies are based on the concept that a biological (or pharmaceutical) effect caused by a given molecule (drug) is a function of its chemical structure.

Most existing SAR methods describe chemical structure using *attributes*, which are general properties of objects. For example, in the traditional Hansch approach to QSARs[1,2] the attributes are properties such as $\log P$ and π, which are global properties of the molecule or substituted group, whereas in the CoMFA[3] approach to QSARs, the attributes are properties of points in space that are global properties of the coordinate system used. Compounds are described as lists of attributes. This form of data representation is not well suited to describing the steric structure of molecules because it is difficult to map efficiently atoms and their bond connectivities onto a list. This is due to difficulty in defining a natural bond order and the different sizes of molecules. One intuitive approach to the problem is to use the eigenvalues of adjacency matrices.[4] Our approach is to use *relations* to describe objects and to learn QSARs using an inductive logic programming (ILP) system (specifically Aleph developed by A. Srinivasan, which supersedes P-Progol;[5] the program Aleph is available at http://www.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/aleph.pl). In a relational description, the basic elements are relations between objects (i.e., logic programs, a general form of computer program). Formally, the difference in descriptive language between attributes and relations corresponds to the difference between propositional and first-order predicate logic. Existing learning methods for QSARs are all based on propositional logic. ILP uses the more powerful representation language of predicate logic, equivalent to the ability to learn general computer programs, for prediction. To illustrate the difference between attributes and relations, consider the following hypothesis: an active compound requires a double bond conjugated to an aromatic ring. Such a hypothesis could be directly discovered and represented by a relational QSAR system using only simple atom and bond types (e.g., atom A in an aromatic ring is connected by a single bond to atom B, which is connected by a double bond to atom C). It could not be found or represented in an attribute-based language without specifically precoding the attribute "double bond conjugated with an aromatic ring". The increased generality gained using relations

* To whom correspondence should be addressed. Phone: +44 (0)1970 622357. Fax: +44 (0)1970 622455. E-mail: ntg@aber.ac.uk.
† The University of Wales Aberystwyth.
‡ University of Oxford.

10.1021/jm0155244 CCC: $22.00
Published on Web 12/19/2001

allows a more direct mapping from the chemical steric structure to its representation. In this relational representation, we can deal with multiple low-energy conformations because there is no need to explicitly prealign individual compounds to a common 3D spatial reference nor is it necessary to use a large number of points to define the fields because they could be implicitly defined using a relation (logic program). Moreover, comprehensible results are more easily produced because the use of logical relations provides a richer language to describe drug binding. Initial work was done using the program Golem to form SARs for the inhibition of dihydrofolate reductase by pyrimidines and triazines.[6−8] This work was extended by the development of the program Progol and applied to predicting the mutagenicity[9] of a series of structurally diverse nitro compounds. Finally ILP was used to predict toxicity of compounds[10] and in the first attempt of pharmacophore discovery.[11]

The aims of this study were to test whether ILP methods could be used in a pharmacophore discovery task and to form new indicator variables for regression equations in QSAR to improve the results obtained from standard SAR methods. This study is the first attempt that compares results obtained with ILP and conventional 3D-QSAR.

We used two datasets to evaluate our method: the well-studied thermolysin zinc protease inhibitor dataset and a glycogen phosphorylase *b* (GP) inhibitor dataset. Thermolysin inhibitors were chosen for two reasons: (1) the structures of native thermolysin and several complexes with bound inhibitors are available, which allows us to structurally evaluate the computational results; (2) there are a number of inhibitors of thermolysin available from the literature, which can be modeled and added to the dataset. GP inhibitors were also chosen because the X-ray crystal structures of each ligand−GP complex, and corresponding biochemical data, had been determined during a project that aimed to obtain inhibitors of GP with potential therapeutic activity as antidiabetic drugs.[12,13]

We conclude that ILP enables pharmacophore mapping and QSAR formation that avoids the aforementioned problem typically encountered in standard SAR methods. In this paper, we first introduce the dataset and the methods used. Second, the results using the three lowest conformations are presented for both datasets. Third, the results using the 10 lowest conformers for each dataset are described and the QSARs obtained with CoMFA and ILP indicator variables are discussed.

Materials and Methods

A. Data. A.1. Thermolysin Dataset. Thermolysin, isolated from *Bacillus thermoproteolyticus*, is a zinc requiring endopeptidase of M_r 34 600. Overall, the tertiary structure of thermolysin consists of two spherical domains separated by a deep cleft that constitutes the active site. Such zinc-containing proteases are widely distributed in nature and play an important role in numerous physiological processes such as digestion and blood pressure regulation. Crystal structures of 12 inhibitors (Table 1) bound to the active site of thermolysin were extracted from the Protein Data Bank.[14] The corresponding activity values[15−23] ($pK_i = -\log K_i$) ranged from 2.42 to 10.17. A Monte Carlo conformation analysis was performed on these 12 inhibitors. Then a geometry optimization and charge calculation of the 10 lowest conformers were performed with the semiempirical AM1 method.[24] An additional 19 inhibitors with K_i values ranging from 10^{-3} to 10^{-8} M were taken from the literature (Table 2) and modeled as previously described using the inhibitor-bound crystal structures as templates (the references for these structures are given in Table 2).

Table 1. Series of Inhibitors of Thermolysin Taken from the Crystallographic Structures[a]

molecule	pK_i ($-\log K_i$)	reference
bzs	2.42	15
cct	6.42	16
phosphoramidon	7.55	17
clt	7.30	18
pln	5.89	17
llnhoh	3.72	19
zfpla	10.17	20
honhbzmagna	6.37	19
zgll	8.04	20
zgpoll	5.04	21
bag	6.12	22
bppp	2.79	23

[a] Molecule names were assigned according to the reference given.

Table 2. Thermolysin Inhibitors Dataset[a]

molecule	pK_i ($-\log K_i$)	reference
zgplda	5.77	38
zgplg	6.57	38
zgplla	7.78	38
zgplnh2	6.12	38
zgplf	7.11	38
zf	3.29	23
honhbzmagnh2	6.18	39
honhbzmoet	4.70	39
paaoh	4.05	40
piaoh	6.44	40
plfoh	7.72	40
pltoh	7.82	40
pfoh	4.14	40
zggdlnhoh	3.60	39
zggllnhoh	4.41	39
zggnhoh	3.03	39
zglnhoh	4.89	39
zglnmeoh	2.65	39
zlnhoh	5.00	39

[a] These inhibitors were taken from the literature and modeled using the inhibitor-bound crystal structures as templates. Molecule names are taken from the corresponding references given.

A.2. Glycogen Phosphorylase *b* (GP) Dataset. GP inhibitors were chosen because the X-ray crystal structures of each ligand−GP complex and their corresponding activities are available.[12] We used the same numbering and structures as those used in the series of 51 glucose and thioglucose derivatives reported in Table 1 of Pastor et al.[25] The 51 glycogen phosphorylase inhibitors used have in common a glucopyranose ring, with different substitution at the C1 position in the α and/or β configurations. The inhibitors varied from small, simple monosaccharides such as α-D-glucose, to large disaccharides such as gentiobiose, and most of them were polar in nature. The activity of these molecules ranges from 10^{-2} to 10^{-6} M. To be consistent with the thermolysin dataset, we performed an AM1 geometry optimization and charge calculation of each of the lowest conformers found from a Monte Carlo conformation analysis.

B. Algorithm Used. Inductive logic programming (ILP) algorithms are a class of machine-learning algorithms that have been successfully applied to a number of SAR problems.[6−11] In ILP all the inputs and outputs are logical statements in the computer language Prolog. Such statements are readily understandable because they closely resemble natural language. The input for an ILP method is a set of positive and

negative examples defining compounds as active or inactive, respectively, and background chemical knowledge. To dichotomize the real-valued activities into two classes, we ordered the examples on the basis of their corresponding biological activity and considered the top half (high-activity compounds) to be positive examples and the bottom half (low-activity compounds) to be negative. The output of using an ILP method is a set of hypotheses, expressed as a set of rules, that predict the positive and negative examples using the background knowledge. The best hypothesis is chosen to maximize the compression function $f(c)$ defined as $f(c) = P(c) - N(c) - L(c)$, where $P(c)$ is the number of positive examples that can be proven by the clause c, taken together with the background knowledge. $N(c)$ is the number of negative examples that can be proven in the same way, and $L(c)$ is the size (number of literals) of the clause c. After generation of a single rule, the examples covered by the rule are removed from consideration and other rules are generated until all examples are removed or until no more statistically significant rules can be found.

C. Background Knowledge for ILP. The background knowledge represents the individual compounds in terms of the chemical type of the component atoms and the connecting bonds. These values were assigned using the atom types as defined in the molecular modeling program Sybyl,[26] version 6.7. The atoms of each molecule were typed automatically according to their local chemical environment and bond type. The electrostatic charges were extracted from the output of the AM1 charge calculation performed with Spartan.[27] Our ILP methodology allows a full three-dimensional representation by adding to the system background knowledge the coordinates of the atoms plus basic knowledge of 3D geometry, i.e., Pythagoras's theorem and simple trigonometry. During the learning, Aleph focuses on a single positive example and constructs a "bottom" clause containing everything, subject to language constraints, that is true of that positive example according to the background theory. The search then proceeds by beginning with the "empty" pharmacophore (0 points) and constructing progressively more complex pharmacophores in the bottom clause. The complexity of the pharmacophores comes from the number of points and the error on distances that have been set prior to the search in the background knowledge. The constructed pharmacophores are tested on the remaining molecules. Thus, arbitrary complex pharmacophores are identified using Aleph and become indicators of activity in a QSAR. When learning pharmacophores, Aleph carries out an internal search, in a computationally efficient way, to find the best alignment for prediction. The alignment carried out by the algorithm Aleph is on the internal coordinates of each molecule instead of on Cartesian coordinates as with CoMFA. The geometry of the molecule is represented by pairwise distances between points.

In standard approaches, like CoMFA, it is necessary to first explicitly superimpose individual compounds of a dataset, and only after such prealignment to a common spatial reference frame can learning take place. This has the problem that it can be very difficult to superimpose heterogeneous compounds such as those within the thermolysin inhibitor dataset. The compound-specific knowledge for each of the lowest inhibitor conformers of the dataset under study is represented by first-order atomic formulas or Prolog facts of the form

atm(m1, c1, a1, o, x1, y1, z1, 2, q1)

which asserts that the molecule m1 in the conformation c1 has an atom a1 that is an oxygen at the position (x1, y1, z1) in 3D space. This oxygen is sp^2-hybridized and bears a negative charge q1. Similarly, the relation

bond(m1, c1, b1, a1, a2, 2)

represents that in the conformation c1 of molecule m1 the bond between atoms a1 and a2 is assigned a double bond. Using this atom and bond description, we defined libraries of elementary chemical concepts (literals). For example, the following Prolog program fragment defines and can be used to detect a hydroxamic acid group (C(=O)NHOH).

Table 3. General Chemical Knowledge Defined in the ILP Method

hacc (hydrogen acceptor)	
hdonor (hydrogen donor)	
hydrophobic	
neg_charge (negative charge)	
pos_charge (positive charge)	
alcohol	hetero_non_aromatic_6c_ring
aldehyde	hydroxamic_acid
alkyl_halide	imine
amide	ketone
amine	methoxy
aromatic_5c_ring	nitro
aromatic_6c_ring	non_aromatic_5c_ring
ar_alcohol (aryl alcohol)	non_aromatic_6c_ring
ar_halide (aryl halide)	phosphorus_acid
carboxylate	phosphorus_po2
carboxylic_acid	phosphate_opo3
deoxy_amide	sulfide
ester	sulfo
ether	sulfoamide
hetero_aromatic_5c_ring	sulfone
hetero_aromatic_6c_ring	sulfonic_acid
hetero_non_aromatic_5c_ring	thiol

```
hydroxamic_acid(Molecule,Conf,Atom1,X,Y,Z):-

    atm(Molecule,Conf,Atom1,c,X,Y,Z,_,_),

    bond(Molecule,Conf,Atom1,Atom0,2),

    atm(Molecule,Conf,Atom0,o,_,_,_,o_2,_),

    bond(Molecule,Conf,Atom1,Atom2,1),

    atm(Molecule,Conf,Atom2,n,_,_,_,n_3,_),

    bond(Molecule,Conf,Atom2,Atom3,1),

    atm(Molecule,Conf,Atom3,o,_,_,_,o_3,_),

    bond(Molecule,Conf,Atom3,Atom4,1),

    atm(Molecule,Conf,Atom4,h,_,_,_,_,_).
```

The predicate hydroxamic_acid states that a Molecule in the conformation Conf represents an hydroxamic acid group centered on Atom1 located at the *X, Y, Z* coordinates. If we can find an Atom1 in the Molecule in the conformation Conf, which is a carbon "c" atom at the *X, Y, Z* coordinates, this carbon atom should be double-bonded to an sp^2 (o_2) oxygen "o" Atom0. Atom1 should also be single-bonded to an sp^3 (n_3) nitrogen "n" Atom2. Nitrogen Atom2 should form a single bond with an sp^3 oxygen, Atom3, which is itself single-bonded to a hydrogen Atom4. We defined 35 new generic structural groups compared with the five predicates of Finn and co-workers.[11] Table 3 shows the 39 generic structural groups defined in our background knowledge. The description of the pharmacophore is given by a clause of the form

```
active(M):- methyl(M, Z, A), hacc(M, Z, B),
      carboxylic_acid(M, Z, C), dist(M, A, B, 2.5, 1.0),
            dist(M, A, C, 3.0, 1.0), dist(M, B, C, 5.2, 1.0).
```

which asserts that a molecule M in the conformation Z is active if it has a methyl group A and an hacc (hydrogen acceptor) B and a carboxylic_acid C such that the distance between A and B is 2.5 ± 1.0 Å, the distance between A and C is 3.0 ± 1.0 Å, and the distance between B and C is 5.2 ± 1.0 Å. The predicate active(M) defines the head of the clause and states that a

pharmacophore should be active, while the body specifies the points of the pharmacophore (at least three points) via the generic structural predicates (in this example, methyl, hacc, carboxylic_acid) and specifies the distance between each pair of points via the predicate dist.

D. Methodology. D.1. CoMFA. All CoMFA analyses were done using Sybyl, version 6.7, running on a Silicon Graphics Octane dual R12000 operating under Iris 6.5. For each dataset, the superimposition procedure of the individual compounds was performed by rms fitting of the backbone heavy atoms of each ligand to the most active analogue compound. The steric and electrostatic interactions for the CoMFA analyses were calculated using an sp^3 carbon probe atom carrying a charge of +1.0 and a distance-dependent dielectric constant ($1/r$). The CoMFA lattice was 28 Å × 28 Å × 24 Å. The cutoff value for both the steric and electrostatic interactions was set to +30 kcal/mol. Regression analyses were done using the partial-least-squares[28] (PLS) methodology in conjunction with a full (leave-one-out) cross-validation[29] procedure.

D.2. ILP. We used the computer language Prolog to implement Aleph under the Yap Prolog compiler, version 4.1.19. The range of activity was divided into two (half−half) intervals, namely, "active" and "inactive" representing the positive and negative examples, respectively. For each of these intervals, including the background knowledge, ILP results in a set of rules predicting the pharmacophore for active and inactive classes. These rules are then used as new indicator variables, i.e., Boolean attributes that are 1 when the pharmacophore is displayed in the compound and 0 when it is absent, to form QSARs. Standard regression methods such as PLS were used to form QSARs with these new indicator variables. A linear regression method, as implemented in WEKA,[30] version 1.3.0, was also used to select the best model and to predict the values of the activity. We decided to use this method because the ILP indicator variables are Boolean attributes that can cause problems during the PLS run. The evaluation of a significant improvement, by addition of the ILP indicator variables, is calculated using a binomial test. Finally, QSARs obtained with standard a CoMFA method and with CoMFA and ILP attributes were compared.

Results

A. Results Using the Three Lowest Conformations. A.1. Thermolysin Inhibitors. We worked on two models. Model A was based on the neutral ligands, and to evaluate the QSARs with respect to changes in the electronic nature of the ligands, we developed the ionized model B. For the CoMFA superimposition, the compounds were fit to the most active analogue compound using the Zn-binding functional groups (i.e., carboxylate, hydroxamate, phosphonate, or sulfhydryl). Several studies[31,32] on thermolysin have attempted to model the data. DePriest et al.,[31] performing a PLS analysis using only the CoMFA steric and electrostatic fields as explanatory variables, reported a predicted coefficient $r^2_{CV} = 0.70$ using 61 molecules. Waller et al.,[32] using the CoMFA analysis on the dataset of DePriest et al., found $r^2_{CV} = 0.536$ and improved it to 0.596 by adding the HOMO fields. Hence, even slight variations in molecular superimposition can lead to different predicted coefficients. Our dataset of 31 molecules were entered as separate rows for each molecule in a QSAR table along with their respective pK_i values. CoMFA steric and electrostatic fields were calculated, as previously described in Materials and Methods, and entered as columns in the QSAR table. An initial PLS analysis, with a leave-one-out cross-validation, was performed to determine the optimum numbers of components using only the CoMFA steric and electrostatic fields as explanatory variables. This was followed by a non-cross-validated run using the optimum number of components to derive a predictive model. Analysis of the lowest conformation of 31 thermolysin inhibitors using only CoMFA parameters produced a correlation having a cross-validated $r^2_{CV} = 0.61$ and a conventional $r^2 = 0.99$ using eight principal components for model A and $r^2_{CV} = 0.42$ and $r^2 = 0.93$ using three principal components for model B. The correlation coefficients obtained are consistent with those obtained in previous studies.[31,32] The most relevant interactions, as proposed in the literature, between a substrate (e.g., zfpla) and the enzyme are shown in Figure 1a.

A leave-one-out cross-validation procedure, with the new indicator variables obtained from the ILP pharmacophore search on the three lowest conformers for models A and B, selected the best model and constructed regression equations relating activity to these new attributes. Table 4 presents the results of models A and B and the clausal representation of the pharmacophore constructed using the ILP system. The addition of ILP indicator variables has not made the regression equations any more complicated than those obtained with standard methods. Moreover, easily interpretable results are produced and can easily give insight into the drug-binding process.

For model A, ILP indicator variables led to a cross-validated coefficient $r^2_{CV} = 0.82$ (Table 4a) that is significantly better ($P < 0.01$) than our standard CoMFA analysis $r^2_{CV} = 0.61$ for the same dataset. The regression equation obtained is formed by the A4, A10, and A16 four-point pharmacophores represented by the boxed regions labeled 1−4 in Figure 1b. It is straightforward to interpret the coefficient of indicator variables in linear QSAR equations because the magnitude indicates the contribution of the presence of the corresponding structural feature. The interpretation of the ILP results (in Table 4) for the successful inhibition of thermolysin would be as follows. Pharmacophore A4 states that neutral compounds should have (1) an *amide* group to interact through the carboxyl oxygen to the side chain nitrogens of residue Arg 203 and the hydrogen acceptor binding site defined by Asn 112 and Ala 113, (2) a *neg_charge*, i.e., a slightly negatively charged group (to interact within the hydrophobic pockets S_1 and S_2) such as a carbobenzoxyl moiety in which delocalization of the π electrons can occur on the benzene ring, (3) an *hacc* group located on one of the phosphonamide oxygens close to Glu 143, which is strongly involved in the catalysis,[23] or on the terminal carboxylate group in contact with the S'_2 pocket and Asn 112, and (4) an *hdonor* group corresponding to the second phosphonamide oxygen, which binds the zinc ion and accepts a hydrogen bond from both Tyr 157 and His 231; this last interaction stabilizes the tetrahedral intermediate during catalysis.[33,17] From the four-point pharmacophore A10, compounds would be active if they have (1) two *hdonor* groups binding with the hydrogen bond acceptor site (amide oxygen of Ala 113, oxygen OD1 of Asn 112), as also reported by Holden et al.,[20] (2) a *carboxylic_acid* located in the S'_2 pocket, and (3) a slight *neg_charge* atom corresponding to a methyl group of a leucine side chain. Finally, the most informative pharmacophore related to activity is represented by A16 because it yields the highest regression coefficient. Pharmacophore

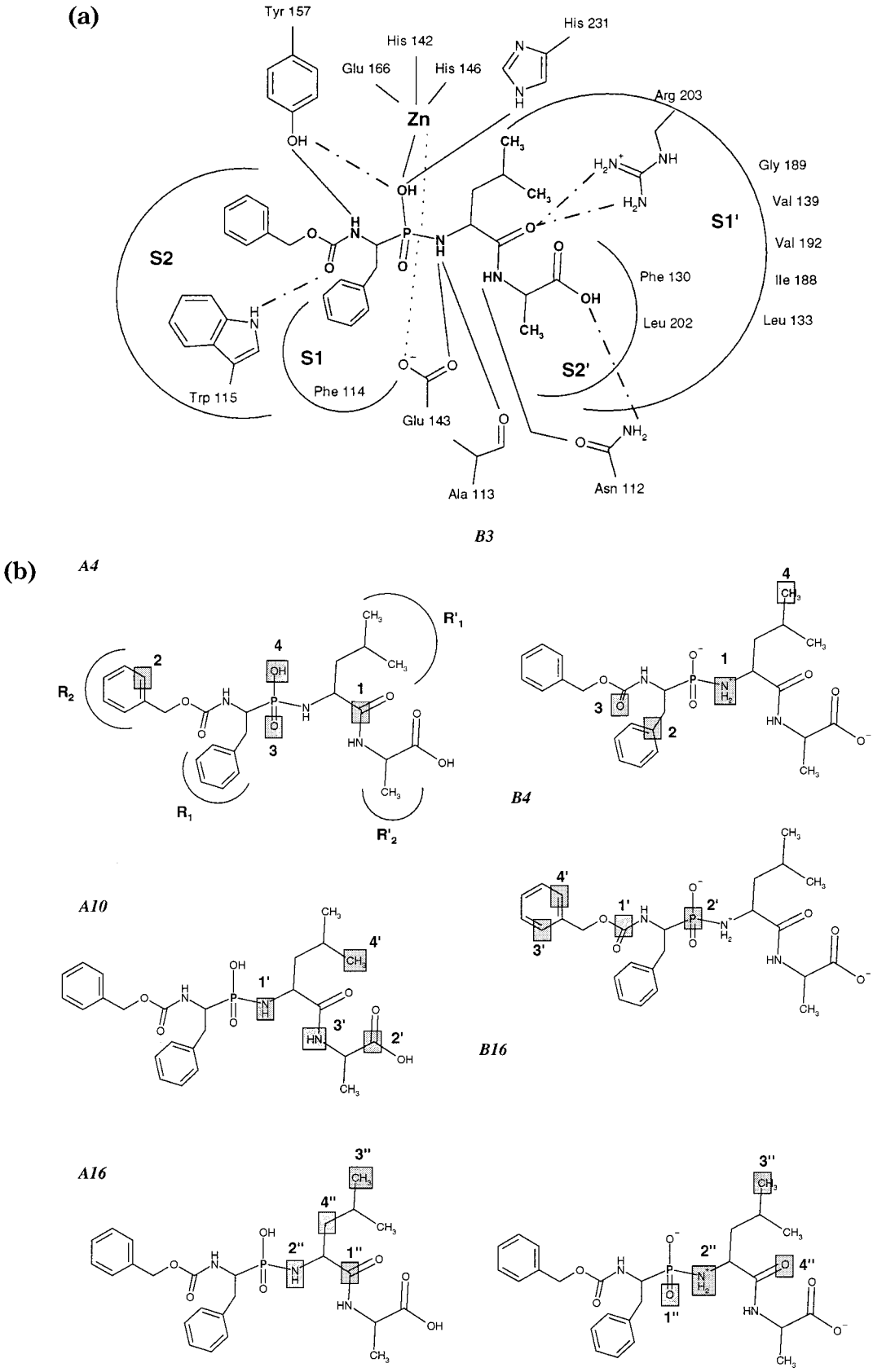

Figure 1. (a) Schematic drawing of the interactions observed in the neutral zfpla−thermolysin (4TMN) complex where S_1, S_2, S'_1, and S'_2 indicate the most significant regions of the active site of thermolysin. (b) Model A, neutral zfpla inhibitor with highlighted four-point pharmacophores A4, A10, and A16 obtained using the ILP system of the three lowest conformations for each inhibitor in the dataset. See text for detailed discussions of the boxed regions labeled 1−4. (c) Model B, ionic inhibitor zfpla with highlighted four-point pharmacophores B3, B4 and B16 obtained using the ILP system for the three lowest conformations of each inhibitor in the dataset.

Table 4. Resulting QSARs Using the Three Lowest Conformations of Thermolysin Inhibitors

(a) Model A

Regression Equation Obtained with Only ILP Attributes

r^2_{cv}ILP = 0.82, pK_i = 3.62 + 0.95(A4) + 1.27(A10) + 1.82(A16)

ILP Clausal Representation of the Pharmacophore

A4

active(A):- amide(A,B,C),
neg_charge(A,B,D), dist(A,D,C,8.0,1.0)
hacc(A,B,E), dist(A,E,D,7.2,1.0), dist(A,E,C,4.1,1.0),
hdonor (A,B,F), dist(A,F,C,4.0,1.0),
dist(A,E,F,2.4,1.0), dist(A,D,F,7.7,1.0).

A10

active(A):- hdonor(A,B,C),
carboxylic_acid(A,B,D), dist(A,D,C,5.8,1.0),
hdonor(A,B,E), dist(A,E,D,2.5,1.0), dist(A,E,C,3.6,1.0),
neg_charge(A,B,F), dist(A,F,C,4.5,1.0),
dist(A,E,F,5.0,1.0), dist(A,D,F,5.6,1.0).

A16

active(A):- amide(A,B,C),
hdonor(A,B,D), dist(A,D,C,2.4,1.0),
methyl(A,B,E), dist(A,E,D,5.5,1.0), dist(A,E,C,6.7,1.0),
neg_charge(A,B,F), dist(A,F,C,2.4,1.0),
dist(A,E,F,2.5,1.0), dist(A,D,F,2.5,1.0).

(b) Model B

Regression Equation Obtained with Only ILP Attributes

r^2_{cv}ILP = 0.80, pK_i = 3.88 + 1.91(B3) + 1.00(B4) + 2.07(B16)

ILP Clausal Representation of the Pharmacophore

B3

active(A):- amine(A,B,C),
hydrophobic(A,B,D), dist(A,D,C,6.5,1.0),
hacc(A,B,E), dist(A,E,D,4.4,1.0), dist(A,E,C,4.6,1.0),
pos_charge(A,B,F), dist(A,F,C,4.7,1.0),
dist(A,E,F,7.6,1.0), dist(A,D,F,10.6,1.0).

B4

active(A):- amide(A,B,C),
phosphorus(A,B,D), dist(A,D,C,3.8,1.0),
neg_charge(A,B,E), dist(A,E,D,4.3,1.0),
dist(A,E,C,4.6,1.0),
neg_charge(A,B,F), dist(A,F,C,5.7,1.0),
dist(A,E,F,2.4,1.0), dist(A,D,F,5.9,1.0).

B16

active(A):- neg_charge (A,B,C),
hdonor(A,B,D), dist(A,D,C,2.2,1.0),
methyl(A,B,E), dist(A,E,D,3.6,1.0), dist(A,E,C,4.1,1.0),
hacc(A,B,F), dist(A,F,C,5.5,1.0),
dist(A,E,F,4.9,1.0), dist(A,D,F,3.5,1.0).

A16 asserts that molecules should present (1) an *amide* group that can hydrogen-bond through the carboxyl group to the side chain nitrogens of Arg 203 (as found in A4), (2) an *hdonor* atom corresponding to the phosphonamide nitrogen (as in zfpla) or phosphoramide nitrogen (as in plfoh and pltoh), (3) a *methyl* group, and (4) a *neg_charge* atom belonging to the nonpolar aliphatic group R'_1 of the ligand most likely interacting with a leucine side chain in the active site.

Model B yielded a cross-validated coefficient r^2_{CV} = 0.80 using only ILP attributes (Table 4b). Figure 1c highlights the relevant four-point pharmacophores B3, B4, and B16 found in both regression equations. B3 involves (1) an *amine* group corresponding to the protonated phosphonamide or phosphonamide nitrogen that interacts with Glu 143, as reported in Kester et al.,[23] or potential interaction with the nitrogen atom of the tryptophan R'_2 side chain (found in compounds clt and pltoh), (2) a *hydrophobic* group that is a phenylalanine side chain, where the presence of this phenyl group at the R_1 position can interact more extensively with the enzyme and is reported by Morihara et al.[34] to contribute strongly to ligand binding, or alternatively it could be a leucine side chain in position R'_1 of the inhibitor (found in compounds phosphoramidon and pltoh but not shown in Figure 1c), (3) a *hacc* atom located on the carbonyl oxygen of the carbobenzoxyl moiety of the inhibitor as displayed in zfpla, which makes a hydrogen bond with the peptide of Trp 115 in the S_2 pocket, and (4) a *pos_charge* atom located on the methyl group of the leucine side chain. B4 is represented by (1) an *amide* group in the carbobenzoxyl moiety that interacts with Trp 115 and Tyr 157, (2) a *phosphorus* atom that potentially coordinates to the zinc ion, in agreement with the work of DePriest et al.,[31] and (3) two *neg_charge* atoms on the phenyl ring of carbobenzoxyl moiety. The most informative pharmacophore is B16 because it presents the highest regression coefficient and it involves (1) a *neg_charge* atom corresponding to one of the phosphonamide oxygens (P═O) arising from the delocalization of the negative charge on the second phosphonamide oxygen, (2) an *hdonor* atom (phosphonamide nitrogen), (3) a *methyl* group from the R'_1 leucine side chain of the inhibitor, and (4) an hdonor to interact with the side chain of nitrogens of Arg 203.

The analysis with CoMFA and ILP descriptors using the three lowest conformers led to correlations having cross-validated coefficient r^2_{CV} = 0.85 (model A) and r^2_{CV} = 0.79 (model B). These are significantly better ($P < 0.001$) than those found using only the CoMFA attributes. But in both models, the magnitude of the CoMFA coefficient is small enough (0.01) to justify using only our new ILP indicator variables in the analysis.

A.2. Glycogen Phosphorylase *b* Inhibitors. In the work of Pastor et al.[25] PLS models were obtained without variable selection leading to cross-validated squared correlation coefficient r^2_{CV} = 0.43 (without taking into account water molecules) and r^2_{CV} = 0.45 (with water), with GRID/SRD (smart region definition) GOLPE variable selection leading to r^2_{CV} = 0.79 (without water), and with standard GRID/GOLPE variable selection leading to r^2_{CV} = 0.76.[12] The standard GRID/GOLPE variable selection method is performed using a D-optimal design in the loading space, followed by a fractional factorial design (FFD) strategy with a fixing−excluding procedure to test variable combinations on the predictivity as described by Cruciani et al.[13]

Furthermore, the GRID/SRD GOLPE uses all these variable-selection methods in addition to a smart region definition[25] to carry out the variable selection on groups of variables chosen according to their positions in 3D space, which further reduces the number of variables and consequently yields a high cross-validated squared correlation coefficient. Thus, the most appropriate comparison is between the cross-validated squared correlation coefficient obtained using the standard GRID/CoMFA method (r^2_{CV} = 0.43) of Pastor et al. and that obtained using our ILP procedure. In the work of Venkatarangan et al.[35] using a series of modeled glycogen phosphorylase inhibitors, a 4D-QSAR and FEFF 3D-QSAR model was derived to construct ligand−receptor binding models. In the 4D QSAR study, multiple conformers, alignments, and pharmacophores in 3D-QSAR model construction were explored in the

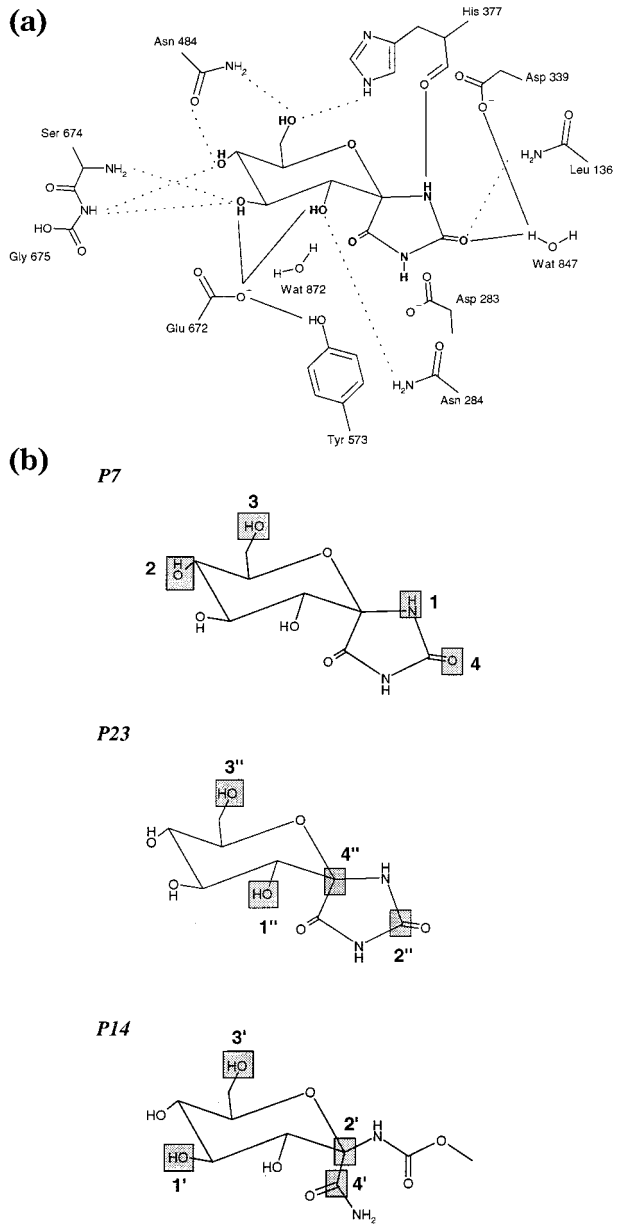

Figure 2. (a) Schematic drawing of the interactions observed in the crystallographic structure of the GP−spirohydantoin complex.[36,37] (b) Inhibitor **45** with highlighted four-point pharmacophores P7, P23 and inhibitor **35** for pharmacophore P14 obtained using the ILP system. Inhibitor numbering is adopted from Table 1 of Pastor et al. (1997). See text for detailed descriptions of the highlighted areas labeled 1−4.

search for the active conformation and binding mode for each compound. Because these studies include portions of the protein, direct comparison to our ILP model is difficult.

Watson et al.[12] reported that the substrate binds at the catalytic site of GP buried 15 Å from the surface and stabilizes the inactive T state form of the enzyme through specific interactions with a loop of residues (280s loop) that can block access to the catalytic site. Figure 2a highlights the main interactions of the most potent inhibitor[36,37] (3 μM) at the catalytic site of GP. The CoMFA analysis in this work was done following the same method as in the thermolysin dataset. Analysis of the lowest conformation of 51 GP *b* inhibitors using only CoMFA parameters produced a correlation having a cross-validated $r^2_{CV} = 0.61$ and a conventional

Table 5. Equation Obtained Using the Linear Regression Method and Clausal Representation of the Pharmacophore for the Three Lowest Conformations of Each GP Inhibitor

Regression Equation Obtained with CoMFA and ILP Attributes and with Only ILP Attributes

r^2_{cv}ILP = 0.72, pK_i = 2.25 + 1.39(P7) + 0.82(P14) + 0.84(P23)

ILP Clausal Representation of the Pharmacophore

P7

active(A):- hdonor(A,B,C),
alcohol(A,B,D), dist(A,D,C,4.8,1.0),
hdonor(A,B,E), dist(A,E,D,4.8,1.0), dist(A,E,C,2.9,1.0),
hacc(A,B,F), dist(A,F,C,2.3,1.0), dist(A,E,F,3.8,1.0),
dist(A,D,F,6.5,1.0).

P14

active(A):- alcohol(A,B,C),
hetero_non_ar_6_ring (A,B,D), dist(A,D,C,2.8,1.0),
alcohol (A,B,E), dist(A,E,D,3.8,1.0), dist(A,E,C,5.5,1.0),
amide (A,B,F), dist(A,F,C,4.9,1.0), dist(A,E,F,5.9,1.0),
dist(A,D,F,2.8,1.0).

P23

active(A):- alcohol(A,B,C),
pos_charge (A,B,D), dist(A,D,C,3.8,1.0),
alcohol(A,B,E), dist(A,E,D,4.7,1.0), dist(A,E,C,5.5,1.0),
hetero_non_ar_5_ring (A,B,F), dist(A,F,C,2.6,1.0),
dist(A,E,F,3.6,1.0), dist(A,D,F,1.8,1.0).

$r^2 = 0.98$ using eight principal components. ILP indicator variables yield a cross-validated coefficient $r^2_{CV} = 0.72$ and led to P7, P14, and P23 four-point pharmacophores (Table 5; boxed regions labeled 1−4 shown in Figure 2b). Our ILP procedure outperforms the standard GRID/CoMFA of Pastor and co-workers and CoMFA methods ($r^2_{CV} = 0.43$ and $r^2_{CV} = 0.61$, respectively) and is not doing badly against the standard GOLPE and SRD/GOLPE methods ($r^2_{CV} = 0.76$ and $r^2_{CV} = 0.79$, respectively) that use variable-selection methods that ILP does not. The highest coefficient is found for P7, which asserts that an active compound would have (1) an *hdonor* atom corresponding to the nitrogen atom of hydantocidin moiety making hydrogen bonds with the backbone oxygen of His 377 as shown in Figure 2a, (2) an *alcohol*, and (3) an *hdonor* corresponding to two hydroxyl groups located on the glucopyranose ring that can interact with the side chain atoms of Asn 484 and His 377 and each hydroxyl is involved both as hydrogen donor and as hydrogen acceptor with these residues, and (4) an *hacc* group corresponding to the amide oxygen of the hydantocidin moiety and can interact with the side chain of Leu 136 and a water molecule Wat 847; this interaction was postulated[36] to enhance the inhibitory effect 3-fold. *P23* involved (1) an *alcohol* group that can interact with the side chain atoms of Asn 284 and Glu 672, (2) a slight *pos_charge* atom corresponding to the carbonyl carbon of the β substituent of compound **45** (compound numbering is adopted from Table 1 of Pastor et al.[25]), (3) an *alcohol* group interacting with the side chain atoms of Asn 484 and His 377, and (4) a *hetero_non_ar_5_ring* group present only in the spirohydantoin derivatives (inhibitors **45**, **46**, **50**, **51** in Tables 1 and 3 of Pastor et al.[25]), the most active compounds in the series. Finally, for the successful inhibition of GP *b*, the compounds should have, according to P14, (1) an *alcohol* group that can interact with the side chain oxygens of Glu 672, and the backbone nitrogens of Ser 674 and Gly 675, (2) a *hetero_non_ar_6_ring* corresponding to the anomeric carbon (C1) of the glucopyranose ring, (3) an *alcohol* group that can interact with the side chain atoms of Asn 484 and His 377, and

Table 6. Resulting QSARs Using the 10 Lowest Conformations of Thermolysin Inhibitors

(a) Model A

Regression Equation Obtained with Only ILP Attributes

r^2_{cv}ILP = 0.80, pK_i = 3.70 + 1.57(A′5) + 1.43(A′7) + 0.91(A′14)

ILP Clausal Representation of the Pharmacophore

A′5

active(A):- hdonor(A,B,C),
pos_charge(A,B,D), dist(A,D,C,3.6,1.0),
carboxylic_acid (A,B,E), dist(A,E,D,6.2,1.0),
dist(A,E,C,5.6,1.0),
hydrophobic(A,B,F), dist(A,F,C,3.1,1.0), dist(A,E,F,6.6,1.0),
dist(A,D,F,2.2,1.0).

A′**7**

active(A):- amide(A,B,C),
hdonor(A,B,D), dist(A,D,C,4.9,1.0),
hdonor(A,B,E), dist(A,E,D,2.9,1.0), dist(A,E,C,2.4,1.0),
pos_charge (A,B,F), dist(A,F,C,1.5,1.0), dist(A,E,F,1.4,1.0),
dist(A,D,F,3.6,1.0).

A′**14**

active(A):- amide(A,B,C),
neg_charge (A,B,D), dist(A,D,C,8.5,1.0),
hacc(A,B,E), dist(A,E,D,8.3,1.0), dist(A,E,C,4.2,1.0),
hdonor (A,B,F), dist(A,F,C,3.9,1.0), dist(A,E,F,2.4,1.0),
dist(A,D,F,9.3,1.0).

(b) Model B

Regression Equation Obtained

r^2_{cv}ILP = 0.84, pK_i = 3.95 + 2.1(B′5) + 1.5(B′11) + 1.08(B′12)

ILP Clausal Representation of the Pharmacophore

B′5

active(A):- pos_charge (A,B,C),
neg_charge(A,B,D), dist(A,D,C,2.6,1.0),
carboxylate(A,B,E), dist(A,E,D,6.1,1.0), dist(A,E,C,7.8,1.0),
hydrophobic(A,B,F), dist(A,F,C,2.2,1.0), dist(A,E,F,6.3,1.0),
dist(A,D,F,3.0,1.0).

B′**11**

active(A):- amide(A,B,C),
methyl(A,B,D), dist(A,D,C,3.6,1.0),
hdonor(A,B,E), dist(A,E,D,3.7,1.0), dist(A,E,C,3.4,1.0),
neg_charge (A,B,F), dist(A,F,C,5.2,1.0), dist(A,E,F,2.4,1.0),
dist(A,D,F,6.1,1.0).

B′**12**

active(A):- amide(A,B,C),
neg_charge (A,B,D), dist(A,D,C,5.9,1.0),
methyl(A,B,E), dist(A,E,D,4.9,1.0), dist(A,E,C,5.5,1.0),
hacc (A,B,F), dist(A,F,C,6.3,1.0), dist(A,E,F,2.4,1.0),
dist(A,D,F,2.4,1.0).

(4) a C-*amide* group in the α-substituted position, as reported by Pastor et al.,[25] as the only α substituent that leads to an increase in activity.

For the GP inhibitor dataset the regression equation obtained with CoMFA and ILP descriptors is the same as that obtained using only the new ILP indicator variables. As previously stated for the thermolysin dataset, the ILP attributes are as informative as the CoMFA descriptors.

B. Results Using the 10 Lowest Conformations. B.1. Thermolysin Inhibitors. The introduction of 10 conformers is straightforward because we simply have to add the first-order atomic formulas atm and bond for these new conformers. The results of the ILP approach for the thermolysin inhibitors are summarized in Table 6. Analysis of model A, using the 10 lowest conformers from the AM1 Monte Carlo search, produced a correlation having a cross-validated $r^2_{CV} = 0.80$. The regression equation obtained using the new ILP indicator variables led to the four-point pharmacophores A′5, A′7, and A′14 (Figure 3; a prime symbol is employed to differentiate from the results of previous models A and B using the three lowest conformers). It is noteworthy that even with the use of 10 conformers, the equations formed are simple and easy to comprehend. These new pharmacophores select similar structural features found in section A using the three lowest conformations. More generally, each of these pharmacophores, using the 10 lowest conformers, requires a specific structural functionality adding to the usual hydrogen donor and acceptor interactions, i.e., all the molecules displaying A′5 have in common a leucine side chain at the R'_1 position and a terminal carboxylic acid. A nonaromatic hydrophobic side chain (Leu, Ile, Ala) is the common feature for **A′7**. And finally, all the molecules displaying A′14 have a bulky R_2 group (carbobenzoxyl moiety or rhammose ring).

Model B yields a cross-validated coefficient $r^2_{CV} = 0.84$ using only the ILP descriptors. The regression equation led to three pharmacophores: B′5, B′11, and B′12 (Figure 3b). Similar important structural features were found between the pharmacophores found with the three lowest conformations and these three. A feature that has not been found previously is the presence of a methyl group at the R'_2 position, which seems to be more likely than a bulky side chain. This new requirement has not been reported yet in the literature. Hence, ionic molecules exhibiting the B′5 pharmacophore have in common a terminal carboxylate and most have a leucine residue in the R'_1 position. Pharmacophore B′12 is represented by molecules possessing a nonbulky R'_2 group.

By taking into account the 10 lowest conformers, the regression equation in model A, obtained using the CoMFA and ILP indicator variables, led to the same four-point pharmacophores as that using the ILP descriptors only. Importantly, using the CoMFA and ILP descriptors led to lower cross-validated $r^2_{CV} = 0.79$ and $r^2_{CV} = 0.81$ for models A and B, respectively. But in both models, no CoMFA coefficient is found in the regression equation. The absence of any CoMFA coefficient emphasizes the ability of ILP variables to be more informative than the CoMFA descriptors.

B.2. Glycogen Phosphorylase *b* Inhibitors. The introduction of the 10 or 6 lowest conformers in the analysis of the GP dataset led to a poorer cross-validated squared correlation coefficient $r^2_{CV} = 0.42$ than that found with only three conformers. By adding more conformers in the analysis, we are introducing more noise than informative data. Since the 51 GP ligands have in common a glucopyranose ring, they are relatively rigid and therefore present fewer possible conformers than the thermolysin inhibitors. Hence, the "best" pharmacophores are found among the three lowest conformers.

C. Results Using the Less Active Compounds as Positive Examples. C.1. Thermolysin Inhibitors. ILP is also able to learn rules from the less active compounds by taking them as positive examples (swapping the positive and negative examples). Thus, by inclusion of the less active compounds as positive examples in the dataset, pharmacophores can be found to predict and understand the low inhibitory effects of the compounds of a series. For model A, using the three

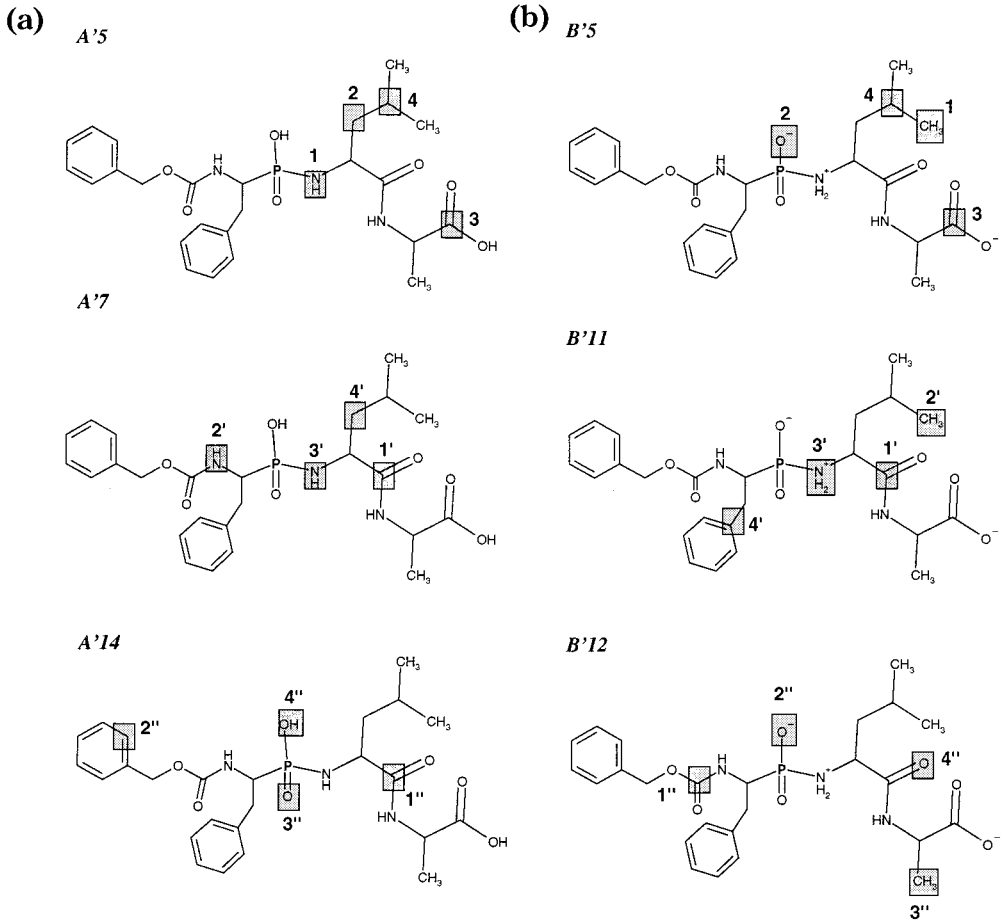

Figure 3. (a) Model A, neutral inhibitor zfpla with highlighted four-point pharmacophores A′5, A′7, and A′14 obtained using the ILP system for the 10 lowest conformers. (b) Model B, ionic inhibitor zfpla with highlighted four-point pharmacophores B′5, B′11, and B′12 obtained using the ILP system for the 10 lowest conformers.

lowest conformers, a set of "negative" rules are described by the following predicates:

```
active(A):- hydroxamic_acid(A,B,C),
        pos_charge(A,B,D), dist(A,D,C,1.5,1.0),
          hdonor(A,B,E), dist(A,E,D,1.4,1.0),
          dist(A,E,C,2.4,1.0), hdonor(A,B,F),
        dist(A,F,C,1.4,1.0), dist(A,E,F,3.0,1.0),
                                dist(A,D,F,2.5,1.0).
```

This clause means that a *hydroxamic_acid* (C(=O)-NHOH) functional group leads to a decrease in the activity. This is in agreement with the activity of the compounds (Table 2); all the molecules including a hydroxamic group represents the lowest pK_i of the dataset. A second important "negative" pharmacophore has been found and is represented by the following clause:

```
active(A):- ar_6c_ring(A,B,C), neg_charge(A,B,D),
              dist(A,D,C,1.4,1.0), hacc(A,B,E),
          dist(A,E,D,3.8,1.0), dist(A,E,C,4.5,1.0),
            hdonor(A,B,F), dist(A,F,C,5.6,1.0),
                  dist(A,E,F,4.6,1.0), dist(A,D,F,4.7,1.0).
```

It asserts that an *ar_6c_ring* at the R′$_1$ position is expected for the less active compounds (e.g., in the inhibitors bzs, βppp, pfoh, zf) instead of a leucine side chain as shown in the previous "positive" pharmacophores discovered. The ILP system for model B, using the negative examples as positive, induced the following important rule:

```
active(A):- carboxylate(A,B,C), neg_charge
  (A,B,D), dist(A,D,C,1.5,1.0), ar_6c_ring(A,B,E), dist
    (A,E,D,3.8,1.0), dist(A,E,C,4.4,1.0), neg_charge
  (A,B,F), dist(A,F,C,2.5,1.0), dist(A,E,F,2.9,1.0), dist
                                  (A,D,F,1.5,1.0).
```

As previously, this rule points out that a phenyl group at the R′$_1$ position is expected to decrease the activity. Moreover a carboxylate group, instead of a phosphonate group, as a zinc binding site decreases the activity such as in the ionic molecule bzs. Such clauses represent the key features that can provide insight into the low inhibitory effect of the less active compounds.

Using the 10 lowest conformers, we found a similar set of clauses giving rise to the conclusion that a carboxylic acid is not expected as a zinc binding functional group, that a benzene ring should be avoided as an R′$_1$ group (such as in the bzs molecule), and that a hydroxamic acid is not expected to enhance the activity. The same "negative" rule is obtained for model B as the one using only the three lowest conformers.

C.2. Glycogen Phosphorylase *b* Inhibitors. Two rules represented by the following clauses were found to be important:

```
active(A):- hydrophobic(A,B,C), pos_charge(A,B,D),
        dist(A,D,C,1.5,1.0), neg_charge(A,B,E),
        dist(A,E,D,1.4,1.0), dist(A,E,C,2.4,1.0),
           ether(A,B,F), dist(A,F,C,2.9,1.0),
                 dist(A,E,F,3.1,1.0), dist(A,D,F,3.6,1.0).

active(A):- alcohol(A,B,C), alcohol(A,B,D),
        dist(A,D,C,4.2,1.0), pos_charge(A,B,E),
        dist(A,E,D,5.6,1.0), dist(A,E,C,6.5,1.0),
        neg_charge(A,B,F), dist(A,F,C,8.3,1.0),
                 dist(A,E,F,4.4,1.0), dist(A,D,F,8.3,1.0).
```

The first rule means that an *ether* group is not expected in substituents such as molecules **4**, **8, 18**, **33**, **36**, **39** (as in Table 1 of Pastor et al.[25]). The second rule asserts that a *neg_charge* most likely corresponding to a benzene ring, a cyano, or a fluoro group is not favorable as a β substituent.

Conclusions

We have presented a new procedure for the formulation of accurate, easily interpretable QSARs. ILP can be used to improve the descriptive representation of the chemical structures of a dataset under study. First, the useful indicator variables are identified, then standard linear regression is used to form the QSAR. The relevant pharmacophores are selected from the regression equation, and their significance is given by the regression coefficient. The addition of ILP indicator variables has not made the QSAR equations any more complicated than those obtained using standard methods. The most important parameter in a CoMFA study is the relative superimposition of molecules to a common spatial reference frame whereby slight variations in the 3D prealignment will yield very different results. The key advantage of ILP here is that this method avoids the explicit prealignment procedure because it uses relations instead of attributes to represent molecules.

Models A and B derived for the thermolysin dataset led to significantly better results ($P < 0.01$) than with the CoMFA analysis provided in previous work.[31,32] The pharmacophore points found using our ILP method highlighted interactions that were consistent with the previous studies of thermolysin. The observed rules found using the ILP system for models A and B for the 3 or 10 lowest conformers led to the conclusion that the basic structural requirements for the successful inhibition of thermolysin involve (a) a Zn-binding functional group, most likely a phosphonamidate or phosphoramidate moiety, (b) an aromatic hydrophobic R_1 group such as a phenylalanine side chain, (c) a bulky side chain at the R_2 position, which can be a carbobenzoxyl moiety or a rhammose group, (d) a nonpolar aliphatic side chain at the R'_1 position found most likely to be a leucine residue, and (e) amide groups and terminal carboxylate group for hydrogen bonding to side chain residues in the active site of the enzyme. From the reverse ILP procedure, a hydroxamate functional group, or a carboxylate zinc binding functional group, and an aromatic R'_1 group produced a decrease in activity and are not favorable for producing an inhibitory effect.

The requirements obtained for glycogen phosphorylase inhibitors are in agreement with the results of the GRID/GOLPE analysis of Pastor et al.[25] They reported that hydrogen bonds and polar interactions are important for stabilizing the 280s loop and for maintaining the enzyme in its inhibited T-state form. The presence of polar interactions near His 377, Asp 283, and Asn 282 residues has been shown to enhance activity. The P23 pharmacophore has clearly identified that a nonaromatic five-member ring substituent, such as that found in the spirohydantoin derivatives, enhances the inhibitory effect. The negative rules obtained have confirmed that a large aromatic or slightly negatively charged β substituents decrease the inhibitory effect. This is consistent with the proposed carboxyanion intermediate formed during the catalysis in which electron-withdrawing groups at the C1 position would not be favorable for the stabilization of the proposed transition state.[40]

This novel procedure to map pharmacophores and form QSARs could be applied to any type of heterogeneous compounds where 3D prealignment of molecules would be difficult.

Further work will be required to extend the existing background knowledge to the active site of the protein−ligand complex to obtain more informative pharmacophores. This could be done simply by adding to the existing background knowledge new general chemical groups and interactions potentially important in the active site.

Acknowledgment. N.G. thanks the BBSRC for financial support (Grant No. 2/B11471). K.A.W. acknowledges support of the Lister Institute of Preventive Medicine. We also thank Dr. Ashwin Srinivasan of the Computing Laboratory at the University of Oxford. Finally, the authors are grateful to Dr. Brian Matthews for providing some crystallographic structures.

References

(1) Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. Correlation of biological activity of phenoxyacetic acids with Hammet substituent constants and partition coefficient. *Nature* **1962**, *194*, 178−180.
(2) Martin, Y. C. *Quantitative Drug Design: A Critical Introduction*; Marcel Dekker: New York, 1978.
(3) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(4) Muggleton, S. H. Inverse entailment and Progol. *New Gen. Comput.* **1995**, *13*, 245−286.
(5) Burden, F. R. A chemically intuitive molecular index based on the eigenvalues of a modified adjacency matrix. *Quant. Struct.−Act. Relat.* **1997**, *16*, 309−314.
(6) 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 relationship of trimephoprim analogues binding to dihydrofolate reductase. *Proc. Nat. Acad. Sci. U.S.A.* **1992**, *89*, 11322−11326.
(7) Hirst, J. D.; King, R. D.; Sternberg, M. J. E. Quantitative structure−activity relationships: neural networks and inductive logic programming compared against statistical methods. I. The inhibition of dihydrofolate reductase by pyrimidines. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 405−420.
(8) Hirst, J. D.; King, R. D.; Sternberg, M. J. E. Quantitative structure−activity relationships by neural networks and inductive logic programming. The inhibition of dihydrofolate reductase by triazines. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 421−432.
(9) King, R. D.; Muggleton, S.; Srinivasan, A.; Sternberg, M. J. E. Structure−activity relationships derived by machine learning: the use of atoms and their bond connectivities to predict mutagenicity using inductive logic programming. *Proc. Nat. Acad. Sci. U.S.A.* **1996**, *93*, 438−442.

(10) Srinivasan, A.; King, R. D.; Bristol, D. An assessment of submissions made to the predictive toxicology challenge. In *Sixteenth International Joint Conference on Artificial Intelligence*; Dean, T., Ed.; Morgan Kaufmann: San Francisco, CA, 1999; pp 270−275.
(11) Finn, P.; Muggleton, S.; Page, D.; Srinivasan, A. Pharmacophore discovery using the inductive logic programming system PROGOL. *Mach. Learning* **1998**, *30*, 241−270.
(12) Watson, K. A.; Mitchell, E. P.; Johnson, L. N.; Cruciani, G.; Son, J. C.; Bichard, C.; Fleet, G.; Oikonomakos, N.; Kontou, M.; Zographos, S. E. Glucose analogue inhibitors of glycogen phosphorylase: from crystallographic analysis to drug prediction using GRID force-field and GOLPE variable selection. *Acta Crystallogr.* **1995**, *D51*, 458−472.
(13) Cruciani, G.; Watson, K. A. Comparative molecular field analysis using GRID force-field and GOLPE variable selection methods in a study of inhibitors of glycogen phosphorylase *b*. *J. Med. Chem.* **1994**, *37*, 2589−2601.
(14) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **1977**, *112*, 535−542.
(15) (a) Bolognesi, M. C.; Matthews, B. W. Binding of the byproduct analog L-benzylsuccinic acid to thermolysin determined by X-ray crystallography. *J. Biol. Chem.* **1979**, *254*, 634−639. (b) Hausrath, A. C.; Matthews, B. W. Redetermination and refinement of the complex of benzylsuccinic acid with thermolysin and its relation to the complex with carboxypeptidase A. *J. Biol. Chem.* **1994**, *29*, 18839−18842.
(16) Holland, D. R.; Barclay, P. L.; Danilewiez, J. C.; Matthews, B. W.; James, K. Inhibition of thermolysin and neutral endopeptidase 24.11 by a novel glutaramide derivative: X-ray structure determination of the thermolysin−inhibitor complex. *Biochemistry* **1994**, *33*, 51−56.
(17) (a) Weaver, L. H.; Kester, W. R.; Matthews, B. W. A crystallographic study of the complex of phosphoramidon with thermolysin. A model for the presumed catalytic transition state and for the binding of extended substrates. *J. Mol. Biol.* **1977**, *114*, 119−132. (b) Tronrud, D. E.; Monzingo, A. F.; Matthews, B. W. Crystallographic structural analysis of phosphoramidates as inhibitors and transition-state analogs of thermolysin. *Eur. J. Biochem.* **1986**, *157*, 261−268.
(18) Monzingo, A. F.; Matthews, B. W. Binding of *N*-carboxymethyl dipeptide inhibitors to thermolysin determined by X-ray crystallography: a novel class of transition-state analogues for zinc peptidases. *Biochemistry* **1984**, *23*, 5724−5729.
(19) Holmes, M. A.; Matthews, B. W. Binding of hydroxamic acid inhibitors to crystalline thermolysin suggests a pentacoordinate zinc intermediate in catalysis. *Biochemistry* **1981**, *20*, 6912−6920.
(20) Holden, H. M.; Tronrud, D. E.; Monzingo, A. F.; Weaver, L. H.; Matthews, B. W. Slow- and fast-binding inhibitors of thermolysin display different modes of binding: crystallographic analysis of extended phosphonamidate transition-state analogues. *Biochemistry* **1987**, *26*, 8542−8553.
(21) Tronrud, D. E.; Hazel, H. M.; Matthews, B. W. Structures of two thermolysin−inhibitor complexes that differ by a single hydrogen bond. *Science* **1987**, *235*, 571−574.
(22) Monzingo, A. F.; Matthews, B. W. Structure of a mercaptan−thermolysin complex illustrates mode of inhibition of zinc proteases by substrate−analogue mercaptans. *Biochemistry* **1982**, *21*, 3390−3394.
(23) Kester, W. R.; Matthews, B. W. Crystallographic study of the binding of dipeptide inhibitors to thermolysin: implications for the mechanism of catalysis. *Biochemistry* **1977**, *16*, 2506−2516.
(24) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(25) Pastor, M.; Cruciani, G.; Watson, K. A. A strategy for the incorporation of water molecules present in a ligand binding site into a three-dimensional quantitative structure−activity relationship analysis. *J. Med. Chem.* **1997**, *40*, 4089−4102.
(26) The program Sybyl is available from Tripos Associates, Inc., 1699 S. Hanley Road, St. Louis, MO.
(27) The program Spartan is available from Wavefunction, Inc., 18401 Von Karman Avenue, Suite 370, Irvine, CA 92612.
(28) Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J. The covariance problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM *J. Sci. Stat. Comput.* **1984**, *5* (3), 735−743.
(29) Cramer, R. D.; Bunce, J. D.; Patterson, D. E. Cross-validation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR Studies. *Quant. Struct.−Act. Relat.* **1988**, *7*, 18−25.
(30) Holmes, G.; Donkin, A.; Witten, I. H. WEKA: a machine learning workbench. In *Proceedings of the Second Australia and New Zealand Conference on Intelligent Information Systems*, Brisbane, Australia, 1994; pp 357−361.
(31) DePriest, S. A.; Mayer, D.; Naylor, C. B.; Marshall, G. 3D-QSAR of angiotensin-converting enzyme and thermolysin inhibitors: a comparison of CoMFA models based on deduced and experimentally determined active site geometries. *J. Am. Chem. Soc.* **1993**, *115*, 5372−5384.
(32) Waller, C. L.; Marshall, G. R. Three-dimensional quantitative structure−activity relationship of angiotensin-converting enzyme and thermolysin inhibitors. 2. A comparison of CoMFA models based on active-analogue and complementary-receptor-field alignment rules. *J. Med. Chem.* **1993**, *36*, 2390−2403.
(33) Hangauer, D. G.; Monzingo, A. F.; Matthews, B. W. An interactive computer graphics study of thermolysin catalyzed peptide cleavage and inhibition by *N*-carboxymethyl dipeptides. *Biochemistry* **1984**, *23*, 5730−5741.
(34) Morihara, K.; Tsuzuki, H. Thermolysin: kinetic study with oligopeptides. *Eur. J. Biochem.* **1970**, *15*, 374−380.
(35) (a) Venkatarangan, P.; Hopfinger, A. J. Prediction of ligand−receptor binding free energy by 4D-QSAR analysis: application to a set of glucose analogue inhibitors of glycogen phosphorylase. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1141−1150. (b) Hopfinger A. J.; Reaka, A.; Venkatarangan, P.; Duca, J. S.; Wang, S. Construction of a virtual high throughput screen by 4D-QSAR analysis: application to a combinatorial library of glucose inhibitors of glycogen phosphorylase *b*. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1151−1160. (c) Venkatarangan, P.; Hopfinger, A. J. Prediction of ligand−receptor binding thermodynamics by free energy force field three-dimensional quantitative structure−activity relationship analysis: application to a set of glucose analogue inhibitors of glycogen phosphorylase. *J. Med. Chem.* **1999**, *42*, 2169−2179.
(36) Bichard, C. J. F.; Mitchell, E. P.; Wormald, M. R.; Watson, K. A.; Johnson, L. N.; Zographos, S. E.; Koutra, D. D.; Oikonomakos, N. G.; Fleet, G. W. J. Potent inhibition of glycogen phosphorylase by a spirohydantoin of glucopyranose: first pyranose analogues of hydantocidin. *Tetrahedron Lett.* **1995**, *36*, 2145−2148.
(37) Gregoriou, M.; Noble, M. E. M.; Watson, K. A.; Garman, E. F.; Krulle, T. M.; de la Fuente, C.; Fleet, G. W. J.; Oikonomakos, N. G.; Johnson, L. N. The structure of a glycogen phosphorylase glucopyranose spirohydantoin complex at 1.8 Å resolution and 100 K: the role of the water structure and its contribution to binding. *Protein Sci.* **1998**, *7*, 915−927.
(38) Bartlett, P. A.; Marlowe, C. K. Phosphonamidates as transition-state analogue inhibitors of thermolysin. *Biochemistry* **1983**, *22*, 4618−4624.
(39) Nishino, N.; Powers, J. C. Peptide hydroxamic acids as inhibitors of thermolysin. *Biochemistry* **1978**, *17*, 2846−2850.
(40) (a) Kam, C.-M.; Nishino, N.; Powers, J. C. Inhibition of thermolysin and carbopepdidase A by phosphoramidates. *Biochemistry* **1979**, *18*, 3032−3038. (b) Klopman, G.; Bendale, R. D. Computer automated structure evaluation (CASE): A study of inhibitors of the thermolysin enzyme. *J. Theor. Biol.* **1989**, *136*, 67−77.

JM0155244