Application of Inductive Logic Programming to Structure-Based Drug Design

David P. Enot and Ross D. King

Computational Biology Group, Department of Computer Science,
University of Wales Aberystwyth, SY23 3DB, UK
{dle,rdk}@aber.ac.uk

Abstract. Developments in physical and biological technology have resulted in a rapid rise in the amount of data available on the 3D structure of protein-ligand complexes. The extraction of knowledge from this data is central to the design of new drugs. We extended the application of Inductive Logic Programming (ILP) in drug design to deal with such structure-based drug design (SBDD) problems. We first expanded the ILP pharmacophore representation to deal with protein active sites. Applying a combination of the ILP algorithm Aleph, and linear regression, we then formed quantitative models that can be interpretated chemically. We applied this approach to two test cases: Glycogen Phosphorylase inhibitors, and HIV protease inhibitors. In both cases we observed a significant ($P < 0.05$) improvement over both standard approaches, and use of only the ligand. We demonstrate that the theories produced are consistent with the existing chemical literature.

1 Introduction

Most drugs are small molecules (ligands) that bind to proteins [19]. When knowledge of the 3D structure of the target protein is used in the drug design process, the term structure-based drug design (SBDD) is used. Knowledge of the co-crystallized protein-ligand complex structure is particularly important as it shows how a drug interacts with its target. The binding of the ligand to its target can be regarded as a key (ligand) fitting a lock (active site) (figure 1). To ensure this complementarity, a potential candidate must be the right size for the binding site, must have the correct binding groups to form a variety of weak interactions and must have these binding groups correctly positioned to maximize such interactions. These interactions are primarily hydrophobic and electrostatic (hydrogen bonds, interactions between groups of opposite charges). They are individually weak, but they lead if in sufficient number, to a strong overall interaction (*binding energy*) enabling the ligand to bind to the target site (also referred as activity). These general principles of drug interactions are now well understood, but specific relations between molecular structure and function are still too complex to be delineated from physico-chemical theory and semi-empirical approaches are necessary. From the computational side[14], SBDD involves two main sub-problems to design new active compounds: the prediction of

N. Lavrač et al. (Eds.): PKDD 2003, LNAI 2838, pp. 156–167, 2003.

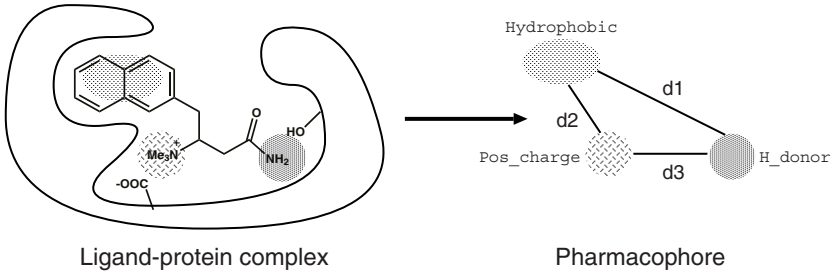

Fig. 1. Schematic representation of a ligand binding a protein illustrating the complementarity of shape and property (left). Example of a three elements pharmacophore (right) derived from the 3D structure of the ligand in the known ligand-protein complex (left).

the most likely ligand mode binding conformation (docking) and the estimation of the relative binding energy of a protein-ligand complex (scoring)[9].

The Protein Data Bank (PDB)[3] is the single worldwide repository for the processing and distribution of 3D biological macromolecular structure data and the number of co-crystallized protein-ligand complexes is rising exponentially over the years. The state-of-the-art in SBDD is to use general propositional regression functions that are designed to be applicable to any active sites (although parameterized using only a small subset of the PDB). Predictions are not generally tuned for specific active sites [20]. Here we describe an Inductive Logic Programming (ILP)/ Relational Data Mining (RDM) approach for SBDD based on generalizing over examples of ligands bound to a specific active site.

The structural nature of many chemical structure-function/property relationships has proven to be well suited to Inductive Logic Programming (ILP) [17]. We take the name of ILP to generalise all work in ILP and the related field of Relational Data Mining. In drug design, ILP has been successfully applied to model structure-activity relationships (SAR). Here the task was to obtain rules that could predict biological activity or toxicity of compounds from their chemical structure[12,16]. ILP is based on *logical relations* and differs from standard chemoinformatics approaches that use *attributes* (molecular descriptor, molecular field, etc) to encode the chemical information. For such problems, logic provides a unified way of representing the relations between objects (atoms and bonds). ILP systems have progressively been shown to be capable of handling 1[12], 2[16] and 3 dimensional[8,21] descriptions of the molecular structures, allowing the development of compact and comprehendable theories. Moreover, ILP has achieved the same predictive power or has significantly improved the traditional QSAR (Quantitative SAR) built using standard propositional learners and statistical methods[15,16].

We take the next natural step in developing the ILP approach to drug design by extending it to SBDD. The aim of this study is four-fold:

- to explore how best to represent the relationship between ligand and protein and how to adapt the ILP tools to suit our study.
- to test whether ILP can form accurate quantitative models of the binding energy of ligands.
- to compare the ILP results with conventional 3D QSAR and SBDD programs.
- to examine the insight obtained from the ILP rules.

2 Methods and Materials

This section describes the complete process we employed to address our problem. The methodology adopted for this study is organized as follows: 1) collect 3D structural data from the PDB and their corresponding biological activities in the literature; 2) transform the molecular structures into facts from a molecular modelling package and extract the features of interest to build the background knowledge; 3) form 3D structural features (pharmacophores) using ILP; 4) form regression models using the pharmacophores and assess their predictive power.

2.1 Datasets

A complete description of the protein-ligand series is reported in section 3. While the PDB gathers most of the structural data of biomolecular systems, there is no unified way to distribute biological activities and structures directly to analysis methods. A preprocessing step is necessary to clean the PDB files: isolation of the ligand, addition of missing atoms or residues, removal of useless information, etc. Despite the fact that the way ILP encodes chemical information is less sensitive to the initial preparation of the complexes than other SBDD methods (protonation state for example), extra care was required to form the proper assignment of the atom types before building the Datalog program.

2.2 Background Knowledge and Its Representation

ILP systems use background knowledge to further describe problems. The background knowledge comprises our statements about the most relevant features to explain the biological activity. This mainly involves using the most comprehensive and the most declarative representation to encode domain-dependant information. The content of the background knowledge used for this study is illustrated in figure 2.

In our representation, the three dimensional information is expressed in terms of distances between atoms or structural groups (*building blocks*) giving the final rule a *pharmacophore* like form. The concepts of (3D)pharmacophore and pharmacophore elements are very important in medicinal chemistry: a pharmacophore is an arrangement of atoms or groups of atoms which influence drastically the activity at a target receptor[19]. Pharmacophore representation expresses the potential activity in a language familiar to medicinal chemists and

```
  Predicates related to the ligand (all arity 3):
hacc,hdon,alcohol,equiv_ether,six_ring,hetero_non_ar_6_ring,amide,
carbonyl,amine_0h,methyl,lipo_seg,ar_6c_ring,halogen,five_ring.
  Predicates related to the protein (all arity 4):
prot_backc2,prot_cooh,prot_alcohol,prot_negcharge,prot_poscharge,
prot_amide,prot_guadinium,prot_lipo_seg.
  Hydrogen bonding predicate: hb/4.
  Water position predicate: water/3.
```

Fig. 2. General chemical knowledge defined in the background knowledge.

is easily convertible for searching compounds in chemical databases. A pharmacophore usually refers to the ligand only but, in the following, we apply this definition to the active site as well.

The Prolog implementation requires facts that store the location of particular groups and a predicate *dist/4* which states the Euclidean distance in 3D space between two groups. For example, the following conjunction,

```
hdonor(l10,1,A),methyl(l10,1,B),dist(A,B,6.3,1.0)
```

represents the fact that in the compound *l10* in its conformation labelled *1*, there are a methyl group A and hydrogen bond donor B separated by 6.3 ± 1.0 Angstroms.

Pharmacophore mapping with ILP avoids the need of traditional 3D QSAR and pharmacophore learning methods to prealign and superpose all the ligands to a common extrinsic coordinate system. The requirement is forced by the propositional nature of the traditional approaches[19]. ILP has the advantage that it can directly use the intrinsic coordinate system of each complex.

Some ligands may also have more than one conformation (3D structure). This is the problem which first highlighted the multiple instance problem, and most propositional machine learning algorithms require major changes to deal with it [13]. ILP has the advantage that it can naturally deal with multiple instance problems.

Only a brief summary of the predicates used for this study is presented here. A Prolog example of generating building blocks facts from molecular structure is illustrated in [21]. The pharmacophore elements, available for the present SAR analysis (figure 2) can be divided in the following two categories:

- Ligand related predicates state the position in 3D space of simple or complex chemical groups providing, for example, the definition of methyl group or aromatic rings. They can also encode some important physico-chemical properties of the atoms or the building blocks, such as their ability to form hydrogen bonds.
- The active site is described by integrating specific chemical knowledge related to a number of important amino acids and water molecules as well as representing hydrogen bonds(*hb/4*) explicitly.

2.3 Constructing Theories with Aleph

The learning algorithm used for this study is the ILP system Aleph[25]. This algorithm follows the classic ILP search engine framework[6]: given a background knowledge (i.e. relations describing the molecular structures), a set of examples (i.e. training data) and a language specification for hypotheses, an ILP system will attempt to find a set of rules that explain the examples using the background knowledge provided. We chose Aleph because it can be easily tuned to suit our learning system which proceeds by iterating through the following basic algorithm:

- The training data is formed by dichotomising the data into two sets (positive and negative) based on their biological activity. Because there is not a natural cut-off to the predictor, an example is chosen from the training data and the positive set comprises the molecules with the closest activity (1/3 of the training data are used in this study). The rest of the examples (2/3 of the molecules) are considered as negative examples.
- The most specific clause (*bottom clause*) that entails the above example is then constructed within the language restrictions provided[22]. This is known as the saturation step. The bottom clause prunes the search before it begins by identifying all the potential clauses explaining the activity of the selected molecule.
- The search is a refinement graph search: it proceeds along the space of clauses (partially ordered by Θ-subsumption) between the specific hypothesis (*bottom clause*) and the most general clause (*empty body*)[25]. We require a complete search in order to find all the possible pharmacophores consistent with the data.
- The new clause is added to the theory and the search is repeated until all the examples are saturated once. Pharmacophores are, thus, learnt for both highly and less active compounds. This contrasts with the usual ILP framework where all examples made redundant are removed (*cover removal* step[25]). Our aim is to use the rules as indicator variables to build quantitative models and the compactness will be assured by the model rather than by the ILP process.

2.4 Building QSAR Models

To combine the ILP pharmacophore into a regression model we used a vanilla in-house multiple linear regression program. The predictive power of the model was evaluated using leave-one-out cross-validation (involving the ILP and regression steps). The results are presented using the squared correlation coefficient (R^2_{cv}) between the actual and the predicted value of the activity. This is the standard measure in drug design. In the following, the activity is evaluated in logarithm units of the inhibition constant ($log(1/Ki)$).

We compare the results of the ILP models with the use of two conventional drug design approaches, CoMFA and a SBDD scoring function.

CoMFA (Comparative Molecular Field Analysis[5]) is the most commonly used 3D QSAR ligand-based approach[18]. The basic idea of CoMFA is to superimpose ligands onto a common 3D grid, and then sample their electronic structure at regular points (voxols). This has the benefit of transforming the data into a propositional form, but relies on the (often false) assumption that every molecule in the series interacts with the same target molecule and in the same way (*common receptor assumption*)[18]. It can also be difficult to know how best to superimpose molecules that do share much common structure. CoMFA also has the drawback of producing thousands of correlated attributes which requires the powerful PLS regression approach to avoid overfitting. In CoMFA, neighbouring voxol attributes are generally highly correlated, yet this information is thrown away. PLS can be used to partially regenerate this correlated structure. In the following, we present the CoMFA analysis using the observed ligand conformation in the protein-ligand complex (common receptor assumption) within an optimized molecular field (superposition/translation).

As no general scoring function has been reported to date that is able to predict binding affinities with a high degree of accuracy[10], we present results with the most accurate approach, for each series under study, among five functions available in the CScore module of Sybyl[1] to compare models including information on the active site.

3 Results and Discussion

We report results obtained from our approach on two protein targets: the glycogen phosphorylase *b* (GP) and the human immunodeficiency virus protease (HIV-PR) enzymes. Chemical structures, inhibition data and predicted biological activities can be accessed from
`http://www.aber.ac.uk/compsci/Research/bio/dss/`.

We chose to study GP and HIV-PR because: a significant amount of 3D information is available on them in the PDB, allowing an accurate validation of the method; they have already been extensively studied, giving us the opportunity to verify the meaning of the rules found by Aleph, and comparable published models; the two datasets stand at two extreme points in SBDD problems. The GP dataset is an homogeneous series of 3D structures with only slight modifications of the structure of ligands. This contrasts with the HIV-PR dataset where the structures of the inhibitor, and to a lesser extent the protein sequence, exhibit dramatic changes from one complex to the next.

3.1 Glycogen Phosphorylase *b*

The set of 51 co-crystallized inhibitors of the glycogen phosphorylase *b* has been taken from the same SBDD project[23]. In this case, the chemical structure of the GP inhibitors is homogeneous; meeting then the usual requirements of traditional 2D/3D QSAR (common receptor assumption). However, the CoMFA[5] analysis on the 51 inhibitors leads to a poor predictive power (r^2_{cv}=0.46, table 1). One

would have thought that we should have been able to derive more physical properties characterising ligand-receptor interaction but the best structure-based binding energy function accuracy is only r^2_{cv}=0.34 (FlexX[24], table 1).

Table 1. Models accuracies from the GP dataset.

Id.	Method	Accuracy (r^2_{cv})
1	CoMFA	0.46
2	FlexX	0.36
3	ILP: Ligand only	0.66
4	ILP: Ligand + *water/3*	0.74
	+ H-bonds involving ligand and water	

In the case of GP, ligands bind at the catalytic site buried deeply from the surface of the enzyme and they stabilize an inactive form of the protein mainly through specific hydrophilic interactions with the protein and some water molecules. Water molecules are well known to play a significant role in stabilizing protein-ligand complexes but they remain a challenge for many QSAR analyses as their mobility violates the common receptor assumption. Table 1 also shows a comparison between results where the background knowledge contains facts only related to the ligand and where the background knowledge also contains facts related to the water molecule position and all the possible hydrogen bonds between the ligand, the active site and the molecules of water.

The results show that our ILP approach outperforms CoMFA and FlexX ($P < 0.005$ for both cases). Addition of more informative knowledge regarding the active site improves the predictive power of the model ($P < 0.025$). The results demonstrate the need to explicitly include hydrophilic interactions in forming a good predictor. The addition of the protein and water interaction also makes the interpretation of the model easier, as they highlight the most important features involved in the binding (see below). The resulting theory and QSAR model are reported in figure 3. The first three (pharmacophores) rules P1, P2 and P3 are overlaid with a highly active ligand to illustrate the main features found by the hypothesis on the same figure. Taking into account the relative homogeneity of the inhibitors, a close inspection of the rules found by Aleph in experiment 3 found that all the key chemical groups are involved in the final model. As shown in figure 3, ILP globally simplifies the interpretation. Insight into the binding mechanism is outlined in two points:

- The amide group in the region **2** is a constant in the three rules (*amide/3*), acting, though, as the basis for the construction of the three pharmacophores. This not surprising as this group is associated with the high activity of the series. Due to the high number of possible interactions in the region **1** and **3**, the theory involves OH groups (*alcohol/3*, rules P2 and P3) rather than explicit hydrogen bonds.
- The most surprising feature denoted by our method is related to the distal part (region **4**) of the active site. Most rules involve either the position of

```
P1 : active(A) :-
hb(A,B,C,D),carbonyl(A,B,E),amide(A,B,F),dist(A,F,E,1.35,1.0),
dist(A,C,E,9.47,1.0),dist(A,D,E,10.77,1.0),dist(A,C,F,10.74,1.0),
dist(A,D,F,11.95,1.0).
P2 : active(A) :-
water(A,B,C),alcohol(A,B,D),alcohol(A,B,E),amide(A,B,F),
dist(A,C,D,14.56,1.0),dist(A,C,E,13.12,1.0),dist(A,D,E,5.98,1.0),
dist(A,F,D,4.63,1.0),dist(A,F,E,3.00,1.0),dist(A,C,F,11.12,1.0).
P3 : active(A) :-
water(A,B,C),water(A,B,D),alcohol(A,B,E),amide(A,B,F),
dist(A,C,D,4.83,1.0),dist(A,C,E,13.69,1.0),dist(A,D,E,14.38,1.0),
dist(A,F,E,4.80,1.0),dist(A,C,F,9.29,1.0),dist(A,D,F,9.75,1.0).
P4 : active(A) :-
water(A,B,C),alcohol(A,B,D),methylen(A,B,E),equiv_ether(A,B,F),
dist(A,C,D,12.50,1.0),dist(A,E,D,4.42,1.0),dist(A,C,E,8.72,1.0),
dist(A,C,F,10.03,1.0),dist(A,D,F,3.01,1.0),dist(A,E,F,1.84,1.0).

QSAR model : log(1/Ki) = 2.43 + 0.76*P1 + 0.91*P2 + 0.35*P3 - 0.49*P4
```

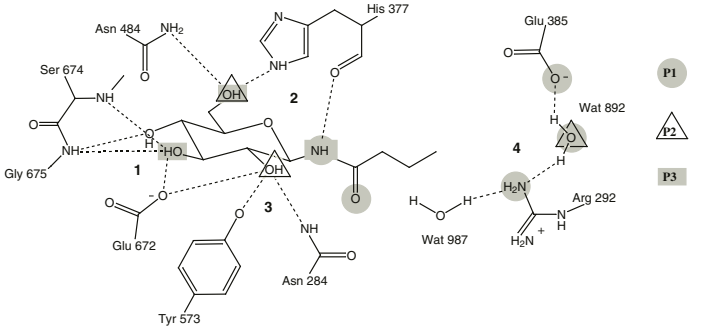

Fig. 3. Theory from experiment 4, table 1 (top). 2D representation of the interaction involved in the binding of the ligand (numbered **26** in [23]) found by our ILP approach (bottom). Shaded circles/rectangles and open triangle outline the pharmacophore elements involved in the theory. Intermolecular interactions between the inhibitor and the binding site are represented with dashed lines.

two water molecules or an explicit hydrogen bond interaction with Arg292 (*water/3* and *hb/4*). How could these interactions be involved in the binding process? We found that [4] suggested that the presence of water overlapping this region could explain a high inhibitory effect with a strong stabilization of the enzyme in the 280's loop.

3.2 Human Immunodeficiency Virus Protease

The second set concerns a series of inhibitors of the well studied human immunodeficiency virus protease. In this case, we are dealing with a series of diverse ligands, some inhibitors are present in two conformations and some residues in

Table 2. Models accuracies from the HIV-PR dataset.

Id.	Method	Accuracy (r^2_{cv})
1	CoMFA	0.58
2	ChemScore	0.35
3	ILP: Ligand only	0.62
4	ILP: Ligand + Active site + H-bonds involving the ligand + *water/3*	0.75

the protein may be mutated (i.e. the sequence of amino-acids can differ from one structure to the next). The same process as for GP is reported in table 2.

In this case, the ILP structure based model (r^2_{cv}=0.75) improves on the CoMFA (r^2_{cv}=0.58, $P < 0.05$) and the scoring function ChemScore[7] (r^2_{cv}=0.35, $P < 0.001$) prediction of the binding energy. The theory from experiment 4 (table 2) is reported in figure 4. The first three rules P1, P2 and P3 are mapped onto the highest active inhibitor (PDB code: **1hvj**).

For HIV-PR, the structural requirements for highly active ligands can seen upon two points of view:

- Polar interaction are highlighted by a specific hydrogen bond with Asp29 (region **3**) and the need of a group (*alcohol/3* in P3) able to interact with Asp25 (region **1**). This last amino acid is involved in the catalytic mechanism of HIV-PR[2]. Finally, the carbonyl group (*carbonyl/3* in P3) in region **2** interacts with the water molecule known to be crucial for the binding process.
- Hydrophobic interactions are more difficult to include in the background knowledge as they are not as local as the hydrogen bonds, for example. Nevertheless, they are implicitly involved in the theory. P1 and P2 largely encode the relative orientation/position of four aromatic rings (mapped by *lipo_seg/3* and *six_ring/3*). The hydrophobic behaviour (*prot_lipo_seg/3*) of the residues 81 and 84 (regions **4** and **5**) are revealed to be important to ensure these non polar contacts.

4 Conclusions

We have presented a new procedure for the formulation of accurate and easily interpretable QSARs to predict binding energy within a series of protein-ligand complexes. This extends the application of ILP in drug design to problems where the structure of the binding protein is known. To form the models we used a relational description of the molecular structure to find rules in the form of pharmacophores, and linear regression to combine the pharmacophores into a predictive model. We consider that the ILP approach was effective for the following reasons:

- the logical formalism is an effective representation for the diverse types of knowledge required.

```
P1 : active(A) :-
hb(A,B,C,D),lipo_seg(A,B,E),six_ring(A,B,F),dist(A,C,E,5.49,1.0),
dist(A,C,F,5.31,1.0),dist(A,D,E,7.22,1.0),dist(A,D,F,7.77,1.0).
P2 : active(A) :-
lipo_seg(A,B,C),prot_lipo_seg(A,B,84,D),six_ring(A,B,E),
dist(A,C,D,5.93,1.0),dist(A,C,E,4.88,1.0),dist(A,D,E,9.19,1.0).
P3 : active(A) :-
alcohol(A,B,C),carbonyl(A,B,D),prot_lipo_seg(A,B,81,E),
dist(A,C,D,5.30,1.0),dist(A,C,E,11.54,1.0),dist(A,D,E,8.67,1.0).
P4 : active(A) :-
carbonyl(A,B,C),pos_charge(A,B,D),prot_negcharge(A,B,29,E),
dist(A,C,D,9.23,1.0),dist(A,C,E,6.09,1.0),dist(A,D,E,9.61,1.0).

QSAR model : log(1/Ki) = 8.00 + 0.81*P1 + 0.43*P2 + 0.58*P3 - 0.90*P4
```

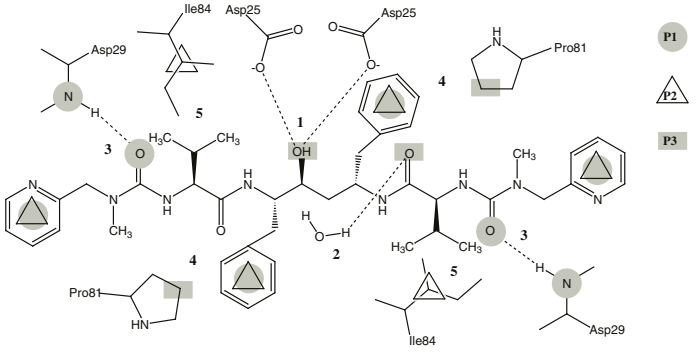

Fig. 4. Theory from experiment 4, table 2 (top). 2D representation of the interaction involved in the binding of **1hvj** found by our ILP approach. The same notation as in figure 3 is adopted.

- the coordinates of molecular structures can be used directly without the superposition or prealignment prior to some traditional approaches.
- ILP deals naturally with the multiple instances problem and can find all possible pharmacophore consistent with the background.
- the theories generated are compact and comprehensible in a language familiar to scientists.

We have tested this approach on two qualitatively different datasets. In both examples, the ILP models outperformed and yet were of equal complexity to the results of traditional SBDD approaches. The ILP models were directly interpretable by mapping the learned pharmacophore onto selected examples, and these interpretations were consistent with previous reported analysis. The derivation of so-called receptor-based pharmacophore does not only improve the predictive power of the models but allows the identification of key interaction *hotspots*. In the case of GP, ILP has brought an unexpected insight into the binding mech-

anism. Analysis of HIV-PR hypotheses shows that our approach could deal with heterogeneous series of protein-ligand. Here, we used direct information from the experimentally resolved structure of a similar protein-ligand complex to give the clues to whereabouts in the active site the ligand binds and in what conformation. Work is in progress to evaluate the applicability of our approach when such information is unavailable or insufficient. Flexible docking techniques can be used to explore the conformational space of the ligand within the active site leading to a highly diverse docking solution set: either our ILP models can be used to restrict the search space[11] or pharmacophores can be learnt from the docking set.

Acknowledgments

DPE was supported by the Biotechnology and Biological Science Research Council (grant no 2/B11471).

References

1. Sybyl 6.8. - Tripos Associates, Inc., 1699 S. Hanley Road, St. Louis, MO.
2. R.E. Babine and S.L. Bender. Molecular recognition of protein-ligand complexes: Applications to drug design. *Chemical Reviews*, 97(5):1359–1472, 1997.
3. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(Supp):235–242, 2000.
4. C.J.F. Bichard, E.P. Mitchell, M.R. Wormald, K.A. Watson, L.N. Johnson, S.E. Zographos, D.D. Koutra, N.G. Oikonomakos, and Fleet G.W.J. Potent inhibition of glycogen phosphorylase by a spirohydantoin of glucopyranose: first pyranose analogues of hydantocidin. *Tetrahedron Letters*, 36:2145–2148, 1995.
5. R.D. Cramer, D.E. Patterson, and J.D. Bunce. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society*, 110(18):5959–5967, 1988.
6. S. Dzeroski and N. Lavrac. An introduction to inductive logic programming. In Dzeroski S. and Lavrac N., editors, *Relational Data Mining*, pages 28–73. Springer-Verlag, 2001.
7. M.D. Eldridge, C.W. Murray, T.R. Auton, G.V. Paolini, and R.P. Mee. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of Computer-Aided Molecular Design*, 11(5):425–445, 1997.
8. P.W. Finn, S. Muggleton, D. Page, and A. Srinivasan. Pharmacophore discovery using the inductive logic programming system PROGOL. *Machine Learning*, 30(2-3):241–270, 1998.
9. I. Halperin, B. Ma, H. Wolfson, and R. Nussinov. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*, 47(4):409–443, 2002.
10. S. Ha, R. Andreani, A. Robbins, and I. Muegge. Evaluation of docking/scoring approaches: a comparative study based on MMP3 inhibitors. *Journal of Computer-Aided Molecular Design*, 14(5):435–448, 2000.

11. S.A. Hindle, M. Rarey, C. Buning, and T. Lengauer. Flexible docking under pharmacophore type constraints. *Journal of Computer-Aided Molecular Design*, 16(2):129–149, 2002.
12. J.D. Hirst, R.D. King, and M.J.E. Sternberg. Quantitative structure-activity relationships by neural networks and inductive logic programming. I. The inhibition of dihydrofolate reductase by pyriminides. *Journal of Computer-Aided Molecular Design*, 8(4):405–420, 1994.
13. A.N. Jain, K. Koile, and D. Chapman. Compass: predicting biological activities from molecular surface properties. performance comparisons on a steroid benchmark. *Journal of Medicinal Chemistry*, 37(15):2315–2327, 1994.
14. D. Joseph-McCarthy. Computational approaches to structure-based ligand design. *Pharmacology and Therapeuthics*, 84(2):179–191, 1999.
15. R.D. King, S. Muggleton, R. Lewis, and M.J.E Sternberg. Drug design by machine learning: The use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. *Proceedings of the National Academy of Sciences of the USA*, 89(23):11322–11326, 1992.
16. R.D. King, S. Muggleton, A. Srinivasan, and M.J.E. Sternberg. Structure-activity relationships derived by machine learning: The use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming. *Proceedings of the National Academy of Sciences*, 93:438–442, 1996.
17. R.D. King and A. Srinivasan. Relating chemical activity to structure: An examination of ILP successes. *New Generation Computing Special issue on Inductive Logic Programming*, 13(3-4):411–434, 1995.
18. H. Kubinyi. *3D QSAR in drug design. Theory methods and application.* Kluwer, Dordrecht, 1997.
19. T. Liljefors and I. Pettersson. Computer-aided development and use of three dimensional pharmacophore. In P. Krogsgaard-Larsen, U. Madsen, and T. Liljefors, editors, *A Textbook of Drug Design and Development*, pages 86–116. Taylor and Francis, London, 2002.
20. A. Logean, A. Sette, and D. Rognan. Customized versus universal scoring functions: application to class I MHC-peptide binding free energy predictions. *Bioorganic Medicinal Chemistry Letters*, 11(5):675–679, 2001.
21. N. Marchand-Geneste, K.A. Watson, B.K. Alsberg, and R.D. King. New approach to pharmacophore mapping and QSAR analysis using inductive logic programming. Application to thermolysin inhibitors and glycogen phosphorylase *b* inhibitors. *Journal of Medicinal Chemistry*, 44(18):2861–2864, 2001.
22. S. Muggleton. Inverse entailment and Progol. *New Generation Computing, Special issue on Inductive Logic Programming*, 13(3-4):245–286, 1995.
23. M. Pastor, G. Cruciani, and K.A. Watson. A strategy for the incorporation of water molecules present in a ligand-binding site into a 3D-QSAR analysis. *Journal of Medicinal Chemistry*, 40(25):4089–4102, 1997.
24. M. Rarey, B. Kramer, T. Lengauer, and G. Klebe. A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology*, 261(3):470–489, 1996.
25. A. Srinivasan. Aleph: A Learning Engine for Proposing Hypotheses http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/aleph.pl.