WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

A general approach for developing system-specific functions to score protein–ligand docked complexes using support vector inductive logic programming

Ata Amini,[1] Paul J. Shrimpton,[1] Stephen H. Muggleton,[2] and Michael J. E. Sternberg[1]*

[1] Structural Bioinformatics Group, Centre for Bioinformatics, Division of Molecular Biosciences, Imperial College London, London SW7 2AY, United Kingdom

[2] Computational Bioinformatics Laboratory, Department of Computing, Imperial College London, London SW7 2AY, United Kingdom

ABSTRACT

Despite the increased recent use of protein–ligand and protein–protein docking in the drug discovery process due to the increases in computational power, the difficulty of accurately ranking the binding affinities of a series of ligands or a series of proteins docked to a protein receptor remains largely unsolved. This problem is of major concern in lead optimization procedures and has lead to the development of scoring functions tailored to rank the binding affinities of a series of ligands to a specific system. However, such methods can take a long time to develop and their transferability to other systems remains open to question. Here we demonstrate that given a suitable amount of background information a new approach using support vector inductive logic programming (SVILP) can be used to produce system-specific scoring functions. Inductive logic programming (ILP) learns logic-based rules for a given dataset that can be used to describe properties of each member of the set in a qualitative manner. By combining ILP with support vector machine regression, a quantitative set of rules can be obtained. SVILP has previously been used in a biological context to examine datasets containing a series of singular molecular structures and properties. Here we describe the use of SVILP to produce binding affinity predictions of a series of ligands to a particular protein. We also for the first time examine the applicability of SVILP techniques to datasets consisting of protein–ligand complexes. Our results show that SVILP performs comparably with other state-of-the-art methods on five protein–ligand systems as judged by similar cross-validated squares of their correlation coefficients. A McNemar test comparing SVILP to CoMFA and CoMSIA across the five systems indicates our method to be significantly better on one occasion. The ability to graphically display and understand the SVILP-produced rules is demonstrated and this feature of ILP can be used to derive hypothesis for future ligand design in lead optimization procedures. The approach can readily be extended to evaluate the binding affinities of a series of protein–protein complexes.

Proteins 2007; 69:823–831.

INTRODUCTION

Over recent years, the use of protein–ligand docking programs has become an increasingly widely used technique in drug discovery. In parallel, protein–protein docking has become more reliable and due to the benefits that researchers in biotechnology, systems biology, and molecular biology would gain from models of protein–protein associations, it is an important and rapidly growing field. Many different protein–ligand software packages and scoring functions have been reported and compared in the literature.[1–3] In common with protein–ligand docking, various methods for protein–protein structure generation have been reported[4,5] along with a range of diverse scoring protocols to choose the best model.[6–11] It is commonly accepted that the majority of the docking programs are often successful at accurately positioning the ligand in the active site (as judged by comparison to X-ray structures).[1,3] However, ranking a series of ligands in terms of their affinity for a particular protein is more problematic and in this work, we address this problem. Because of the comparative ranking prob-

Conflict of interest: AA is now employed by Equinox Pharma Ltd; SHM and MJES are Founder Directors of Equinox Pharma Ltd, hold shares in the company, and have obtained remuneration from the company. Equinox Pharma Ltd is exploiting logic-based methods for drug discovery.
Grant sponsor: BBSRC.
*Correspondence to: Michael J. E. Sternberg, Structural Bioinformatics Group, Centre for Bioinformatics, Division of Molecular Biosciences, Imperial College London, London SW7 2AY, UK. E-mail: m.sternberg@imperial.ac.uk
Received 30 May 2007; Revised 10 July 2007; Accepted 11 July 2007
Published online 1 October 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21782

lems associated with generic scoring functions, attempts have been made to produce system-specific scoring functions to rank a series of ligands binding to a common protein.[12,13] However, such strategies normally involve many separate steps and therefore may not easily be implemented to a new system. We report here a general procedure using a machine learning approach (support vector inductive logic programming, SVILP) to use information from known X-ray structures and ligand binding affinities as background knowledge to obtain system-specific scoring functions. On the condition that sufficient background knowledge exists, the smallest test set described here contained 25 ligands; the method not only provides accurate binding affinity predictions for novel ligands but is also able to provide an insight into features of the ligands and the protein binding site, which are important for determining activity and hence may be used to guide the design of future compounds.

Several generic post-docking ligand–protein scoring functions, such as Drugscore[14,15] and Ligscore,[16] have been reported, which show better ranking than the scoring functions within the docking programs. Improved results have been demonstrated when a consensus score combining more than one scoring method for each ligand is used.[17–19] The concept of pharamacophore type restraints or bias has also been employed to assist scoring ligands to a specific target protein.[20–23] Recently, the generation of bit strings based on the interactions seen in X-ray structures of known inhibitors of a given protein has been reported. Comparison of these strings to those generated from a series of docked ligands has been shown to be an effective method of selected active ligands from a database of decoys.[24,25]

Inductive logic programming (ILP) is a qualitative method, which employs logic to learn rules that can describe certain properties of each member of a dataset. In a biological context, it has been successfully employed for automatic identification of chemical substructures that can be used to describe the toxicity or activity of a compound.[26–29] Recently the method has been made quantitative by combining ILP with support vector (SV) machine technology—support vector inductive logic programming (SVILP).[30] To date, SVILP has only been used on biological datasets containing a series of small molecules, in particular studies on the use of SVILP to identify known actives from a series of 11 activity classes using the standard benchmark dataset taken from the MDL drug data report (MDDR)[31] and to predict the toxicity of a series of compounds.[32] In this study we report an SVILP method to predict accurate binding affinities from datasets containing protein–ligand complexes.

The techniques described thus far have mainly been reported in the context of lead identification, selecting actives from a large database. We report here a general method that uses knowledge gained from ligands of known affinity to produce a system-specific scoring function that can (i) predict the binding affinity of a novel ligand and (ii) provide an insight into which interactions within the active site are important for determining the affinity and as such may be thought of as analogous to QSAR (quantitative structure-activity relationship) techniques. Rather than lead identification, the method is more suited to the lead optimization stage of a drug discovery program where the goal is to understand the effectiveness of, and improve on, features of ligands that have been previously identified as able to bind to the protein in question. An advantage over other system-specific scoring functions is that when sufficient background knowledge is available, the same procedure is shown to work on a range of protein inhibitor systems.

We have tested two variations of the method on five protein systems, HIV protease, carbonic anhydrase II (CAII), trypsin, thrombin, and factor Xa. In the first two systems, we learn rules incorporating the protein–ligand interactions from the crystal structures of known ligands bound to the proteins and predict the affinity of a computationally docked novel ligand in a leave-one-out procedure. Using inhibitors of trypsin, thrombin, and factor Xa, we also show the ability to learn rules from protein–ligand structures generated by docking when X-ray structures are not available. In this latter case, rules are learnt from a training set and predictions made for a separate testing set in order to compare with a previous report using these datasets.[33] Our predictions compare favorably with those of previous studies and we demonstrate the potential usefulness of the rules we identify in future ligand design.

METHODS

Datasets

Five protein inhibitor systems were studied, HIV protease, carbonic anhydrase II (CA II), trypsin, thrombin, and factor Xa. The inhibition constants of the inhibitors of HIV protease were taken from Vinkers et al.[12] Various sources[34–41] were used to collect the inhibition constants of the CA II dataset including http://www-mitchell.ch.cam.ac.uk/pld/energy_kinetic.php. The details of inhibition constants for ligands of trypsin, thrombin, and factor Xa are given by Bohm et al.[33] For the HIV protease and CA II datasets, X-ray structures for all inhibitor complexes were available in the protein data bank (PDB)[42] and each inhibitor was selected as the novel ligand in turn in a leave-one-out procedure. The ligands for the final three protein systems are the same in each case. However, they bind to each protein with differing relative affinities. In order to directly compare our results with those of the previous QSAR study,[33] training and testing sets were defined. As protein–ligand complexes were not available for all the inhibitors in these three datasets, rules were learnt from structures generated by

 DOI 10.1002/prot

docking. The same protein structure was used for all the docking runs (both to generate the complexes in the training set used for rule generation and the testing set).

The Tanimoto coefficient of molecule 1 in the dataset is measured against all the other molecules. The Tanimoto coefficient (T_c) was calculated based on the number of similar fragments two molecules share (see following for details of the fragmentation method):

$$T_c = \frac{\mathrm{nc}}{(\mathrm{nq} + \mathrm{nt} - \mathrm{nc})}$$

where nc equals the number of fragments to the two molecules, nq is the total number of fragments in molecule, 1 and nt the total number of fragments in molecule 2. The diversity of the ligands in each dataset was calculated using a similarity index (T) based on Tanimoto coefficient:

$$T = 100 \times \frac{N_1}{N_t}$$

When the calculated Tanimoto coefficient between molecule 1 and another molecule, for example 2, is greater than 0.8, it is suggested that molecules 1 and 2 have a large degree of similarity. N_1 equals the number of molecules in the dataset that have at least one similar molecule (with Tanimoto coefficient of greater than 0.8) in the set and N_t is total number of molecules.

A system-specific approach to predicting and ranking the inhibitors in the HIV protease dataset has previously been reported[12]; likewise the compounds known to bind to trypsin, thrombin, and factor Xa have also been studied using the CoMFA and CoMSIA QSAR methodologies.[33] In an attempt to provide a back-to-back comparison, we also ran CoMFA and CoMSIA on the HIV protease and CA II datasets (see following for details).

Support vector inductive logic programming

Inductive logic programming (ILP)[43] is a machine learning system, which learns logic rules according to observations (the known binding affinity of the ligand) and background knowledge, here distances between fragments of the ligand and protein atoms. The ligand molecules were fragmented according to a layer-based approach.[44] A central non-hydrogen atom and all of the atoms (including hydrogen atoms) directly bonded to it were defined as a fragment. The distance between each fragment's central atom and all protein atoms in residues that have at least one atom within 5 Å of a ligand atom was calculated. Each distance was stored in the format of ligand fragment—distance ±0.5 Å—protein atom in residue X. ILP calculations require that the dataset used to generate the rules is divided into positives (chemicals with higher activity) and negatives (chemicals with less activity). Here all molecules with activity equal to or above the mean value of the activity of the dataset were defined as positives and the remaining as negatives. CProgol[43,45] selects which distances have predictive power and therefore output as the rules. Support vector machine (SVM) technology is used to quantify ILP rules via the SVILP methodology.[30] A model is built from a matrix consisting of the activity of each molecule against each rule. If the rule is present in a molecule then a "1" is entered, if not a "0". A similar matrix is produced for testing molecules with unknown activities and the model constructed from training matrix is used for quantitative prediction of these molecules. For further information about SVILP, please see our recent report concerning the prediction of the toxicology of molecules.[32]

Ligand preparation and docking for HIV protease and CAII datasets

For the HIV protease and CAII datasets, where X-ray structures of all the ligands complexed to the proteins were available, the ligand from each PDB file was compared to all the other ligands using the 2D similarity function within the SuperLigands website (http://bioinf.charite.de/superligands/)[46] to find its nearest structural neighbor. The only exceptions were for ligands which were split into more than one HETATM group where a visual comparison was performed.

Ligand and protein structures were prepared using Sybyl7.2. Initially, all the PDB files were superimposed onto one member of the set using all the protein backbone heavy atoms via an in-house program. Hydrogens and Gasteiger-Marsili charges were added to all ligands assuming all carboxylic groups to be charged. To ensure that the docking calculations were not influenced by the ligand's conformation being that which is adopted upon binding, each ligand was subjected to a short simulated annealing run. The default sybyl parameters and forcefield were used and the final structure was saved in the mol2 format. Prior to the simulated annealing, each ligand was saved in its X-ray conformation to be used as input for the CoMFA and CoMSIA calculations.

Protein structures had hydrogens assigned via the biopolymer module within sybyl and were also saved as mol2 files. All protonation states were assigned assuming a pH for 7, except for the sidechain of Asp 25 in chain B of the HIV protease proteins, which was protonated in agreement with the assumption that one of the active site aspartic acids should be neutral. In the HIV protease dataset, if the protein structure of the nearest structural neighbor of a ligand included the conserved water molecule bound to Ile 50 A and Ile 50 B, then this was included in the docking run with the "toggle off" and "spin on" options in Gold3.0[47] selected. The active site metal was included in the docking in the CAII protein structures. For both proteins the search area was defined as a sphere of radius 16 Å centered in the middle of the

active site. Up to four atom–atom distance restraints were used to direct similar regions of the ligand being docked to adopt the same interactions within the binding site as those identified from the X-ray structure of the neighbor ligand. Gold3.0[47] was used to dock each ligand to the protein structure taken from the PDB file of its nearest structural neighbor. For each ligand, 10 solutions were generated and the best one as ranked by Gold Score was kept; this pose was also rescored using the online version of Drugscore (http://www.agklebe.de/drugscore) with the CSD and PAIR options selected.

Ligand preparation and docking for thrombin, trypsin, and factor Xa datasets

For the thrombin, trypsin, and factor Xa datasets, the ligands in the training set and the testing set were the same as defined by Bohm et al.[33] PDB files 1ets.pdb (thrombin), 1pph.pdb (trypsin), and 1hcg.pdb (factor Xa) were used as in the previous report.[33] The ligands were downloaded as SD files from http://www.cheminformatics.org/datasets. Babel (http://openbabel.sourceforge. net/wiki/Main_Page) was used to convert each to a mol2 file, which was visually checked in sybyl and inconsistent atom types or structural errors were corrected. Hydrogens were added consistent with the charges displayed in Tables I and V of the Bohm et al. paper[33] Each ligand was then subjected to a short minimization use the default settings and saved in the mol2 format. The three protein files were all prepared using the biopolymer module in sybyl as described. The search area for each protein was a sphere of radius 15 Å centered in the middle of the active site. Atom–atom distance restraints were used to direct the common core of all the ligands to make the same interactions as those seen in 1pph.pdb. Ten solutions were generated and the best one as ranked by Gold Score was kept and rescored using Drugscore as described.

CoMFA and CoMSIA

Comparative molecular field Analysis (CoMFA)[48] and comparative molecular similarity analysis (CoMSIA)[49] were applied to the HIV protease and CAII datasets in order to compare the results from our SVILP method to QSAR techniques in all five datasets. The QSAR calculations reported by Bohm et al.[33] were preformed using a structural alignment generated by building the ligands within the protein active site using the X-ray structure of one of the ligands as a base. Therefore, we chose to use the conformations of the ligands from their original X-ray structures (after the protein backbones had been superimposed) to form the alignment for our CoMFA and CoMSIA calculations. The CoMFA region was defined 5 Å beyond the volume of all of aligned molecules in the training set. The grid spacing was set to 2 Å in all directions and an sp_3 carbon atom with formal charge of +1 was defined as the probe atom. The maximum field values were truncated to 30 kcal/mol for the steric field energies and ±30 kcal/mol for the electrostatic field energies. The same parameters were used for CoMSIA calculations. A leave-one-out cross-validation was performed on the aligned molecules.

McNemar test

The McNemar test[50] is a measure of significance when two methods are compared. In this study, we calculated the McNemar χ^2 using the following equation:

$$\chi^2 = \frac{(B - C)^2}{B + C}$$

where B is the number of times method 1 has squared error less than that of method 2, and C is the number of times method 2 has squared error less than that of method 1. Error is the difference between the predicted and observed values. Statistical significance is then evaluated by finding the probability associated with χ^2. The improvement is significant if the two-tailed probability is <0.05.

RESULTS

Predictions

The size, diversity, and range of experimental affinities for all five datasets used in this study are summarized in Table I. The HIV protease ligands can be seen to be the most diverse as judged by their Tanomoto Similarity Index. Whilst the CAII dataset is less diverse since it is roughly half the size of the HIV protease set, it was expected to be a good test of the amount of background information needed to be able to successfully generate reasonable predictions using SVILP. The ligands for the final three protein systems are the same in each case, and these three datasets are larger and less diverse than the HIV protease and CAII ligands. However, they bind to each protein with differing relative affinities and in all three cases, the range of the binding affinities is much less than the previous two datasets, again providing a test of the ability of SVILP to be useful across a wide range of protein–ligand systems.

SVILP was used to provide binding affinity predictions for each of the five datasets using protein–ligand complexes as input to SVILP to produce a set of rules which could be used as a system-specific scoring function to use with structures obtained from computational docking experiments. Results for binding affinity predictions calculated by all the above methods are summarized in Table II. These results suggest that the use of our SVILP method to rescore inhibitor–receptor docking is a tech-

 DOI 10.1002/prot

Table I
Size and Procedure, Range of Experimental Binding Affinities, and Ligand Diversity of Each of the Five Datasets

Protein	Number of ligands and procedure	Range of pK_i	Diversity[a]
HIV protease	49, leave one out	4.30–11.30 (7.00)	24%
CAII	26, leave one out	3.90–10.00 (6.10)	50%
Trypsin	72, training; 16, testing	4.34–7.64 (3.30)	88%
Thrombin	72, training; 16, testing	4.74–8.48 (3.74)	88%
Factor Xa	72, training; 16, testing	4.28–5.51 (1.23)	88%

[a]As measured by the Tanimoto Similarity Index (see Methods).

nique that can readily be applied to different protein systems to produce results at least as good as other state-of-the-art methods for prediction and understanding of small molecule to protein binding affinities.

HIV protease and CAII datasets

Table II shows that for the diverse HIV protease dataset the SVILP predictions compare favorably with the previous system-specific HIV protease binding affinities predictions.[12] However, the docking we performed used fully flexible ligands, with only a few distance restraints to direct the program, whereas the previous study employed rigid ligand docking using the bound conformation of the ligand taken directly from the X-ray structure. In addition, this previous study was on only one system. Four compounds were excluded in our work due to incomplete protein structures or multiple nonsymmetric ligand binding modes in the available X-ray structures. The R^2_{CV} for SVILP is larger than for either CoMFA or CoMSIA and a McNemar test[50] confirms our method to be significantly more accurate than CoMFA. For the CAII dataset, SVILP and the traditional QSAR techniques (CoMFA and CoMSIA) produce comparable accurate results as judged by the R^2_{CV} figures (Table II). However, the McNemar showed SVILP to be significantly worse. Whilst not as accurate as the QSAR techniques (as judged by the McNemar test), the fact that the same SVILP procedure as that used for the HIV protease dataset produced a high R^2_{CV} using a dataset roughly half the size is a pleasing result. When either the GoldScore or the DrugScore results are used to provide binding affinity predictions for both the HIV and CAII datasets, the accuracy is much lower than SVILP, CoMFA, or CoMSIA (Table II).

Rmsd values were calculated for the docked solution of each ligand to its original X-ray structure position (after superposition of the two protein structures). For both the HIV and CAII datasets, the maximum Rmsd was 6.12 Å, with the majority of ligands being having values of less than 3 Å. Acknowledging that some restraints were used to guide similar regions of the ligands to occupy the same regions of the active site, these results demonstrate the ability of the GoldScore function to accurately position the ligands within the protein. However, the results in Table II highlight the limitation of using such generic scoring functions to rank a series of ligands. Although for these two datasets, rescoring with the generic knowledge function DrugScore shows improvement over GoldScore, our results show that using the available X-ray structural information to produce a system-specific set of rules via SVILP is a superior approach (Table II).

Trypsin, thrombin, and factor Xa datasets

The results from the three remaining datasets are also shown in Table II. The SVILP R^2_{CV} values are similar to those from both CoMFA and CoMSIA. McNemar tests across the three datasets show that there is no significant difference in all of these results. As before predictions made directly from GoldScore and DrugScore are poor (Table II). Our standard SVILP approach can produce predictions on par with state-of-the-art QSAR techniques even when the range of experimental binding affinities is smaller than in the previous two examples (Table I).

Table II
Results of the Binding Affinity Predictions, References and Methods for Each of the Five Datasets

Protein	Method	Reference	R^2_{CV}	MSE	AUC
HIV protease	SVILP	*	0.67	0.75	0.94
HIV protease	HIV protease specific	12	0.66		
HIV protease	CoMFA	*	0.26	1.75	0.85
HIV protease	CoMSIA	*	0.33	1.67	0.82
HIV protease	GoldScore	*	0.08	2.24	0.81
HIV protease	DrugScore	*	0.34	1.46	0.80
CAII	SVILP	*	0.84	0.79	0.97
CAII	CoMFA	*	0.89	0.29	0.89
CAII	CoMSIA	*	0.80	0.55	0.80
CAII	GoldScore	*	0.00	n.a.	n.a.
CAII	DrugScore	*	0.18	2.11	0.80
Trypsin	SVILP	*	0.64	0.28	0.76
Trypsin	CoMFA	33	0.65	0.30	0.79
Trypsin	CoMSIA	33	0.84	0.17	0.81
Trypsin	GoldScore	*	0.39	1.13	0.86
Trypsin	DrugScore	*	0.00	n.a.	n.a.
Thrombin	SVILP	*	0.39	0.56	0.79
Thrombin	CoMFA	33	0.47	0.45	0.72
Thrombin	CoMSIA	33	0.43	0.48	0.67
Thrombin	GoldScore	*	0.25	2.41	0.78
Thrombin	DrugScore	*	0.00	n.a.	n.a.
Factor Xa	SVILP	*	0.47	0.13	0.85
Factor Xa	CoMFA	33	0.38	0.08	0.81
Factor Xa	CoMSIA	33	0.16	0.10	0.84
Factor Xa	GoldScore	*	0.00	n.a.	n.a.
Factor Xa	DrugScore	*	0.00	n.a.	n.a.

*This work.
MSE, mean square error; AUC, area under ROC curve.

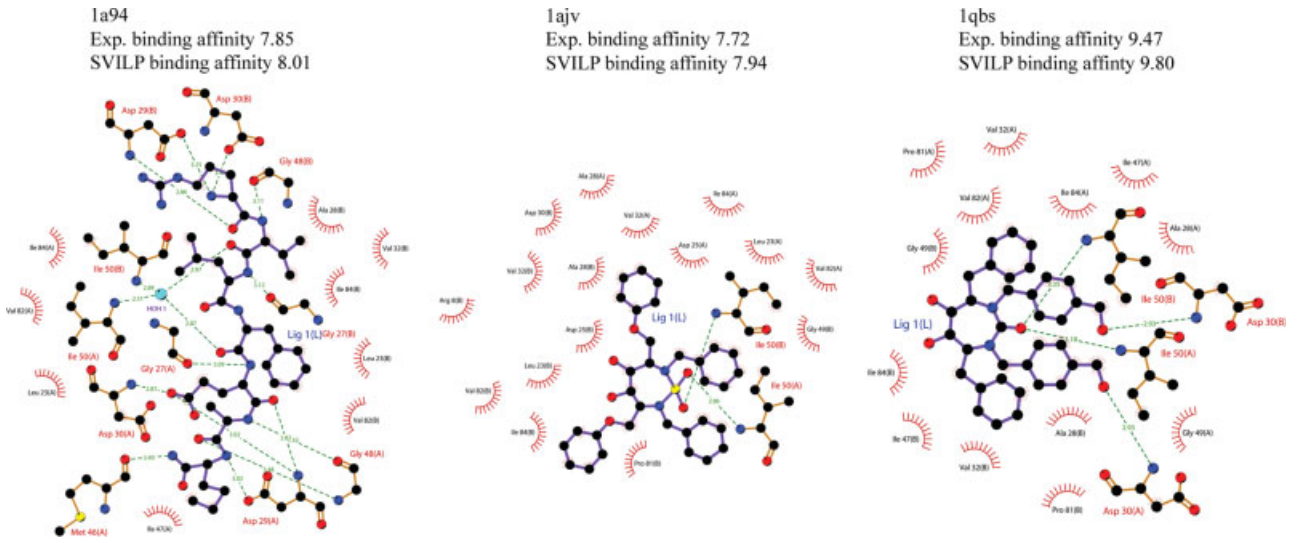

Figure 1

Ligplot[51] generated diagrams of the binding interactions of the ligands from 1a94.pdb,[52] 1ajv.pdb,[53] and 1qbs.pdb.[54] Standard Ligplot color schemes are used with hydrogen bonds indicated by dashed lines and hydrophobic interactions by red arcs. The binding affinities are quoted as pK_i values. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Insights

An important feature of our SVILP approach is that the rules used to make each prediction are output in "human readable" form. At the initial input stage, thousands of distance statements are obtained for each member of the set, which are reduced to a few hundred that have predictive power. Some of these rules will be seen as obvious but since the prediction is derived based on the presence or absence of all the learnt rules in the testing complex, SVILP will often be able to provide accurate predictions in cases where the structure and binding mode of the ligand makes expert visual analysis of the available data difficult. This point is illustrated in Figure 1, which shows LIGPLOT[51] diagrams of three ligands from the HIV protease dataset. The ligand from 1a94.pdb[52] is a peptide-based inhibitor, which mainly makes hydrogen bond interactions with the active site. In contrast, the inhibitor from 1ajv.pdb[53] is bound almost exclusively through hydrophobic contacts within the active site. However, both these ligands have been shown experimentally, and correctly predicted via SVILP, to bind with similar affinities. Moreover, the ligand from 1qbs.pdb[54] appears to be structurally similar to that from 1ajv.pdb and its mode of binding is again characterized by a majority of hydrophobic interactions, yet SVILP correctly predicts this inhibitor to bind with an affinity greater than an order of magnitude higher than 1ajv.pdb (Fig. 1).

As the rules are output ordered by their importance in a simple text file they can easily be converted to display in graphical visualization programs such as Rasmol[55] to aid understanding and future ligand design. Each line of the output file contains a ligand fragment central atom and the bonded fragment atoms, a protein atom type and associated residue number and a distance (and cutoff). For example an aromatic carbon in an aromatic ring in the ligand being 4.0 (±0.5) Å from an aromatic carbon in residue 31 would appear as:

C.ar_C.arC.arH, C.ar, 31, 4.000, 0.5

And an oxygen in a carbonyl group in the ligand being 2.8 (±0.5) Å from an amide nitrogen in residue 51 would appear as:

O.2_C.2, N.am, 51, 2.800, 0.5

Simple Perl scripts were used to select certain rules (lines) from the SVILP output and convert them to a Rasmol script file to display them. For example we collected all rules where the central ligand fragment atom was the same atom type and defined each protein atom and residue from this subset of the rules as a set which could be selected in Rasmol. When this Rasmol script is run it informs the user what ligand atom types occurred in this set of rules and the command used to highlight the protein atoms associated with each ligand atom type.

To demonstrate how displaying the rules may be useful in generating hypothesis of future ligand design the following experiment was carried out. Two ligands, 1qbs.pdb[54] and 1qbt.pdb,[56] were omitted from the HIV protease dataset. The remaining 47 ligands were used to create rules as normal. These rules were used to make a prediction of the binding affinity for the ligand from 1qbs.pdb. The pre-

 DOI 10.1002/prot

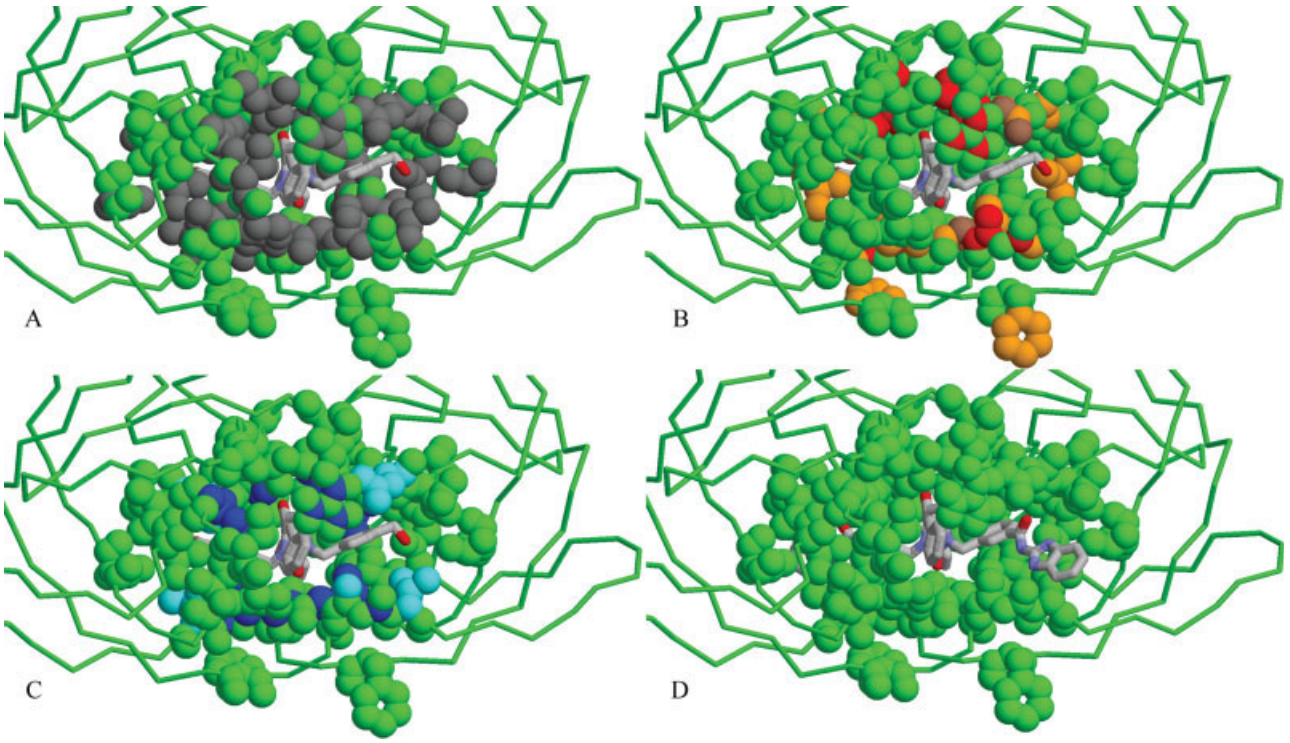

Figure 2

Rasmol depictions of rules. Proteins in green with residues involved in rules are shown as green atom spheres. Ligands are wireframe colored by atom type (carbon gray, nitrogen blue, oxygen red). ***A****: Ligand from 1qbs.pdb[54] bound in active site with protein atoms involved in rules with ligand C.ar sybyl atom types are colored gray.* ***B****: Ligand from 1qbs.pdb[54] with protein atoms involved in rules with ligand O.3 sybyl atom types are colored red; ligand O.2 atom types are colored orange and both ligand oxygen atom types are colored brown.* ***C****: Ligand from 1qbs.pdb[54] with protein atoms involved in rules with ligand N.am sybyl atom types are colored blue and ligand N.ar atom types are colored cyan.* ***D****: Ligand from 1qbt.pdb[56] bound in active site.*

dicted pK_i of 9.73 was in good agreement with observed pK_i of 9.47. The rules used to make this prediction were then displayed using Rasmol (Fig. 2). The ligand is displayed as thick wireframe bonds colored by atom type and the protein as a green backbone with the heavy atoms of all residues involved in rules as spheres. In this example all protein atoms that are involved in rules where the central ligand fragment is the same atom type are highlighted in a common color. In Figure 2(A) all the protein atoms involved in rules in which the central ligand fragment atom is an aromatic carbon have been colored gray. Aromatic carbons in the ligand can be accommodated almost everywhere in the active site. However, it can be seen that an aromatic carbon patch of the active site beyond the terminal hydroxyl group of the ligand is not being filled by this ligand. Figure 2(B) displays protein atoms involved in rules with ligand sp_3 and sp_2 hybridized oxygen atoms. The terminal hydroxyl group can be seen to be far away from any areas of the protein where it would be helping to increase the affinity of the ligand. Figure 2(C) depicts protein atoms involved in rules with amide and aromatic nitrogen atoms indicating that both these atom types are favored in the area of the active site where expansion of the ligand could occur. Hence a hypothesis could be made that affinity would be increased by producing a ligand that had an extension consisting of an aromatic ring containing nitrogen as well as carbon atoms attached to the original core via an amide linker. This is the structure of the ligand from 1qbt.pdb. Using the rules obtained from the 47 ligands (i.e. when neither 1qbs.pdb nor 1qbt.pdb was present in the training) to predict the binding affinity of the ligand leads to an affinity of 10.75, which again compares well with the observed pK_i of 10.62. The predicted binding of the ligand is shown in Figure 2(D).

CONCLUSION

We have reported the use of SVILP to provide system-specific binding affinity predictions, which when tested on five datasets produced results comparable with those of current state-of-the-art methods. This is to our knowledge the first time SVILP has been applied to biological complexes rather than ligand only datasets and demonstrates the capacity of SVILP to handle such input. The difficulties of simply using generic docking scoring functions to obtain binding affinity predictions have been

highlighted. A major benefit of our SVILP methodology is that reliable system-specific rescoring functions were obtained across a range of protein systems using the same procedure. The ability to distinguish between highly and moderately active molecules despite a high degree of structural similarity is an exciting result. The production of rules in a simple text file, which can easily be manipulated to produce graphical displays, is a feature of the SVILP method that we believe could be extremely useful for hypothesis generation in lead optimization processes within a drug discovery project. Having demonstrated the applicability of using an SVILP procedure on systems containing proteins and small molecules, we are confident that the methodology may be further extended to protein–protein systems. SVILP has shown consistently good results across five protein–ligand systems. On these same systems other methods performed less reproducibly. As there is a similar variation in both the methods employed and results obtained in predicting protein–protein associations, we are currently initiating investigations to determine the usefulness of rescoring protein–protein docking using an SVILP procedure.

REFERENCES

1. Taylor RD, Jewsbury PJ, Essex JW. A review of protein-small molecule docking methods. J Computer-Aided Mol Des 2002;16:151–166.
2. Bissantz C, Folkers G, Rognan D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. J Med Chem 2000;43:4759–4767.
3. Perola E, Walters WP, Charifson PS. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. Prot Struct Funct Bioinf 2004;56:235–249.
4. Halperin I, Ma BY, Wolfson H, Nussinov R. Principles of docking: an overview of search algorithms and a guide to scoring functions. Prot Struct Funct Genet 2002;47:409–443.
5. Mendez R, Leplae R, Lensink MF, Wodak SJ. Assessment of CAPRI predictions in rounds 3–5 shows progress in docking procedures. Prot Struct Funct Bioinf 2005;60:150–169.
6. Dominguez C, Boelens R, Bonvin A. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc 2003;125:1731–1737.
7. Kozakov D, Clodfelter KH, Vajda S, Camacho CJ. Optimal clustering for detecting near-native conformations in protein docking. Biophys J 2005;89:867–875.
8. Zhang C, Vasmatzis G, Cornette JL, DeLisi C. Determination of atomic desolvation energies from the structures of crystallized proteins. J Mol Biol 1997;267:707–726.
9. Moont G, Gabb HA, Sternberg MJE. Use of pair potentials across protein interfaces in screening predicted docked complexes. Prot Struct Funct Genet 1999;35:364–373.
10. Gabb HA, Jackson RM, Sternberg MJE. Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol 1997;272:106–120.
11. Kortemme T, Morozov AV, Baker D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. J Mol Biol 2003;326:1239–1259.
12. Vinkers HM, de Jonge MR, Daeyaert FD, Heeres J, Koymans LMH, van Lenthe JH, Lewi PJ, Timmerman H, Janssen PAJ. Inhibition and substrate recognition—a computational approach applied to HIV protease. J Computer-Aided Mol Des 2003;17:567–581.
13. Oliveira FG, Sant'Anna CMR, Caffarena ER, Dardenne LE, Barreiro EJ. Molecular docking study and development of an empirical binding free energy model for phosphodiesterase 4 inhibitors. BioorgMed Chem 2006;14:6001–6011.
14. Velec HFG, Gohlke H, Klebe G. DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. J Med Chem 2005;48:6296–6303.
15. Gohlke H, Hendlich M, Klebe G. Knowledge-based scoring function to predict protein-ligand interactions. J Mol Biol 2000;295: 337–356.
16. Krammer A, Kirchhoff PD, Jiang X, Venkatachalam CM, Waldman M. LigScore: a novel scoring function for predicting binding affinities. J Mol Graph Model 2005;23:395–407.
17. Baber JC, William AS, Gao YH, Feher M. The use of consensus scoring in ligand-based virtual screening. J Chem Inf Model 2006; 46:277–288.
18. Charifson PS, Corkery JJ, Murcko MA, Walters WP. Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. J Med Chem 1999;42:5100–5109.
19. Clark RD, Strizhev A, Leonard JM, Blake JF, Matthew JB. Consensus scoring for ligand/protein interactions. J Mol Graph Model 2002;20:281–295.
20. Fradera X, Mestres J. Guided docking approaches to structure-based design and screening. Curr Top Med Chem 2004;4:687–700.
21. Fradera X, Knegtel RMA, Mestres J. Similarity-driven flexible ligand docking. Prot Struct Funct Genet 2000;40:623–636.
22. Feher M, Deretey E, Roy S. BHB: a simple knowledge-based scoring function to improve the efficiency of database screening. J Chem Inf Comput Sci 2003;43:1316–1327.
23. Verdonk ML, Berdini V, Hartshorn MJ, Mooij WTM, Murray CW, Taylor RD, Watson P. Virtual screening using protein-ligand docking: avoiding artificial enrichment. J Chem Inf Comput Sci 2004;44: 793–806.
24. Mpamhanga CP, Chen BN, McLay IM, Willett P. Knowledge-based interaction fingerprint scoring: a simple method for improving the effectiveness of fast scoring functions. J Chem Inf Model 2006; 46:686–698.
25. Kelly MD, Mancera RL. Expanded interaction fingerprint method for analyzing ligand binding modes in docking and structure-based drug design. J Chem Inf Comput Sci 2004;44:1942–1951.
26. King RD, Muggleton SH, Srinivasan A, Sternberg MJE. Structure-activity relationships derived by machine learning: the use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming. Proc Natl Acad Sci USA 1996;93: 438–442.
27. King RD, Muggleton S, Lewis RA, Sternberg MJE. Drug design by machine learning—the use of inductive logic programming to model the structure-activity-relationships of trimethoprim analogs binding to dihydrofolate-reductase. Proc Natl Acad Sci USA 1992;89:11322–11326.
28. Finn P, Muggleton S, Page D, Srinivasan A. Pharmacophore discovery using the Inductive Logic Programming system PROGOL. Machine Learn 1998;30: 241–270.
29. Sternberg MJE, Muggleton SH. Structure activity relationships (SAR) and pharmacophore discovery using Inductive Logic Programming (ILP). Qsar Combinatorial Sci 2003;22:527–532.
30. Muggleton S, Lodhi H, Amini A, Sternberg MJE. Support vector inductive logic programming. In: Proceedings of the 8th International Conference on Discovery Science, LNAI 3735, Springer-Verlag, 2005. pp 163–175.
31. Cannon EO, Amini A, Bender A, Sternberg MJE, Muggleton SH, Glen RC, Mitchell JBO. Support vector inductive logic programming outperforms the naive Bayes classifier and inductive logic programming for the classification of bioactive chemical compounds. J Computer-Aided Mol Des 2007;21:269–280.

 DOI 10.1002/prot

32. Amini A, Muggleton SH, Lodhi H, Sternberg MJE. A novel logic-based approach for quantitative toxicology prediction. J Chem Inf Model 2007;47:998–1006.
33. Bohm M, Sturzebecher J, Klebe G. Three-dimensional quantitative structure-activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa. J Med Chem 1999;42:458–477.
34. Kim CY, Whittington DA, Chang JS, Liao J, May JA, Christianson DW. Structural aspects of isozyme selectivity in the binding of inhibitors to carbonic anhydrases II and IV. J Med Chem 2002; 45:888–893.
35. Kim CY, Chang JS, Doyon JB, Baird TT, Fierke CA, Jain A, Christianson DW. Contribution of fluorine to protein-ligand affinity in the binding of fluoroaromatic inhibitors to carbonic anhydrase II. J Am Chem Soc 2000;122:12125–12134.
36. Kim CY, Chandra PP, Jain A, Christianson DW. Fluoroaromatic-fluoroaromatic interactions between inhibitors bound in the crystal lattice of human carbonic anhydrase II. J Am Chem Soc 2001;123: 9620–9627.
37. Menchise V, De Simone G, Alterio V, Di Fiore A, Pedone C, Scozzafava A, Supuran CT. Carbonic anhydrase inhibitors: stacking with Phe131 determines active site binding region of inhibitors as exemplified by the X-ray crystal structure of a membrane-impermeant antitumor sulfonamide complexed with isozyme II. J Med Chem 2005;48:5721–5727.
38. Weber A, Casini A, Heine A, Kuhn D, Supuran CT, Scozzafava A, Klebe G. Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. J Med Chem 2004;47:550–557.
39. Scolnick LR, Clements AM, Liao J, Crenshaw L, Hellberg M, May J, Dean TR, Christianson DW. Novel binding mode of hydroxamate inhibitors to human carbonic anhydrase II. J Am Chem Soc 1997;119:850–851.
40. Grzybowski BA, Ishchenko AV, Kim CY, Topalov G, Chapman R, Christianson DW, Whitesides GM, Shakhnovich EI. Combinatorial computational method gives new picomolar ligands for a known enzyme. Proc Natl Acad Sci USA 2002;99:1270–1273.
41. Boriack-Sjodin PA, Zeitlin S, Chen HH, Crenshaw L, Gross S, Dantanarayana A, Delgado P, May JA, Dean T, Christianson DW. Structural analysis of inhibitor binding to human carbonic anhydrase II. Prot Sci 1998;7:2483–2489.
42. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucl Acids Res 2000;28:235–242.
43. Muggleton S, Deraedt L. Inductive logic programming — theory and methods. J Logic Prog 1994;20:629–679.
44. Bender A, Mussa HY, Glen RC, Reiling S. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. J Chem Inf Comput Sci 2004;44:1708–1718.
45. Muggleton S. Inverse entailment and Progol. N Generat Comput 1995;13:245–286.
46. Michalsky E, Dunkel M, Goede A, Preissner R. SuperLigands—a database of ligand structures derived from the Protein Data Bank. BMC Bioinf 2005;6:122–128.
47. Verdonk ML, Chessari G, Cole JC, Hartshorn MJ, Murray CW, Nissink JWM, Taylor RD, Taylor R. Modeling water molecules in protein-ligand docking using GOLD. J Med Chem 2005;48:6504–6515.
48. Cramer RD, Patterson DE, Bunce JD. Comparative molecular-field analysis (Comfa).1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc 1988;110:5959–5967.
49. Klebe G, Abraham U, Mietzner T. Molecular similarity indexes in a comparative-analysis (Comsia) of drug molecules to correlate and predict their biological-activity. J Med Chem 1994;37:4130–4146.
50. McNemar Q. Note on the sampling error of the difference between correlated proportions of percentages. Psychometrica 1947;12:153–157.
51. Wallace AC, Laskowski RA, Thornton JM, Ligplot—A program to generate schematic diagrams of protein ligand interactions. Prot Eng 1995;8:127–134.
52. Wu J, Adomat JM, Ridky TW, Louis JM, Leis J, Harrison RW, Weber IT. Structural basis for specificity of retroviral proteases. Biochemistry 1998;37:4518–4526.
53. Backbró K, Lowgren S, Osterlund K, Atepo J, Unge T, Hulten J, Bonham NM, Schaal W, Karlen A, Hallberg A. Unexpected binding mode of a cyclic sulfamide HIV-1 protease inhibitor. J Med Chem 1997;40:898–902.
54. Lam PYS, Ru Y, Jadhav PK, Aldrich PE, DeLucca GV, Eyermann CJ, Chang CH, Emmett G, Holler ER, Daneker WF, Li LZ, Confalone PN, McHugh RJ, Han Q, Li RH, Markwalder JA, Seitz SP, Sharpe TR, Bacheler LT, Rayner MM, Klabe RM, Shum LY, Winslow DL, Kornhauser DM, Jackson DA, EricksonViitanen S, Hodge CN. Cyclic HIV protease inhibitors: synthesis, conformational analysis, P2/P2′ structure-activity relationship, and molecular recognition of cyclic ureas. J Med Chem 1996;39:3514–3525.
55. Sayle RA, Milnerwhite EJ. Rasmol—biomolecular graphics for all. Trends Biochem Sci 1995;20:374–376.
56. Jadhav PK, Ala P, Woerner FJ, Chang CH, Garber SS, Anton ED, Bacheler LT. Cyclic urea amides: HIV-1 protease inhibitors with low nanomolar potency against both wild type and protease inhibitor resistant mutants of HIV. J Med Chem 1997;40:181–191.