J. Braz. Chem. Soc., Vol. 13, No. 6, 777-787, 2002.
Printed in Brazil - ©2002 Sociedade Brasileira de Química
0103 - 5053 $6.00+0.00

Article

Seeking Novel Leads Through Structure-Based Pharmacophore Design

Luke S. Fisher and Osman F. Güner*

Accelrys Inc., 9685 Scranton Road, San Diego, CA 92121, U.S.A

Neste artigo mostramos um procedimento desenvolvido em nossos laboratórios que identifica "novos" compostos biologicamente ativos distintos de compostos previamente conhecidos no que diz respeito à especificidade da ação e outras características. O procedimento envolve o mapeamento de uma série de compostos ativos (série de treinamento), através de interações lipofílicas e de ligações hidrogênio com o sítio de uma enzima ativa. As interações identificadas e mapeadas são, então, removidas (com exceção daqueles sítios de interação criticamente importantes) e, apenas as interações não utilizadas pelos compostos ativos são mantidas. Os modelos farmacofóricos são gerados e escalonados para os sítios previamente enumerados. Esse modelo é então usado para varreduras de bancos de dados tridimensionais e identificação de novas substâncias matrizes. Este procedimento foi aplicado para inibidores da protease do vírus HIV-1. Vários compostos com atividade moderada na ordem de "micro"(letra mi)Molar foram selecionados pelo modelo farmacofórico proposto.

We developed a procedure that identifies "novel" biologically active compounds that are expected to be distinct from known active compounds with respect to specificity and other such characteristics. The procedure involves mapping a set of known active compounds (training set) onto all possible hydrogen bonding and lipophilic interaction sites of an enzyme active site and flagging those interactions that are utilized by the training set. These flagged sites are removed (except for those that are deemed critical binding sites), leaving only potential interaction sites not utilized by the active compounds. Once unflagged sites were enumerated, pharmacophore models were then generated, scored, and prioritized where the top pharmacophore model was used to search 3D databases for identifying new leads. This procedure was applied to HIV-1 protease inhibitors. Several compounds retrieved by the top pharmacophore model were identified as moderately active (in μMolar range).

Keywords: pharmacophores, structure-based design, 3D databses

Introduction

Rational design of new chemical entities (NCE) typically involves learning from past experiences and developing a knowledge base that can be used to predict future successes. For example QSAR methodology involves development of predictive models based on a list of compounds with *known* biological activities and their structural attributes.[1] Similarly, pharmacophore methodology involves development of a model that represents the common 3D functional attributes of *known active* compounds deemed responsible for biological effect.[2] In both cases, the models are used to propose or design new compounds that are expected to be active. These approaches have been largely successful for the pharmaceutical and biotechnology industry. New active compounds have been designed using these rational approaches, and some of these designs have made it all the way through successful release of drugs in the marketplace.[3,4,5,6] However, because these new compounds were designed based on the attributes of known active compounds, they are "similar" to those known compounds. In fact, this is the basis of knowledge-driven models.

How can we discover active compounds that are sufficiently "different" than those already known? How are novel classes of compounds to be identified if we restrict our search to those characteristics that were derived from known active classes of compounds? Is there a way to expand our search into unknown territories without losing the ability to do rational design? The availability of protein structures provides us an opportunity to expand our search beyond the space of known active compounds. Such new methods will be useful to expand our horizon to search for truly novel active compounds as more and more protein structures become available as a consequence of the human genome project.

In this paper, we are proposing a new approach to pharmacophores: structure-based pharmacophore design. The

* e-mail: osman@accelrys.com

objective is to identify all possible pharmacophore configurations directly from the receptor active site, then map a set of known active compounds onto the pharmacophore space and identify the core pharmacophore features that are utilized by the active compounds… *and then delete them*! What is left then are the potential binding sites in the active site that are *not* utilized by the known active compounds (if a specific interaction is deemed absolutely essential, then that interaction may optionally be retained). One can then use these features to build pharmacophore models that can be used to search databases and retrieve compounds that can potentially bind to the receptor active site. More importantly, because we are forcing these compounds to map onto pharmacophores that are *not* utilized by the known active compounds, we are searching for compounds that are likely to be different than the known active compounds. This is the essence of our proposal: To enforce search criteria for compounds that are different than what is already known, yet still capable of binding to the active site.

The process described above has one problem: Among the potentially hundreds of pharmacophore models, which are the ones we should use to search for novel compounds? Using all of the pharmacophores is likely to be prohibitive, but not impossible. One idea is to use all of the pharmacophores to search and retrieve compounds from a database or a compound library and then concentrate on those compounds that are retrieved by multiple pharmacophores. We are considering this approach for a future report. In this paper, we have concentrated on developing some means of measurement to prioritize the pharmacophores in order to identify the leading few that would be used in database search. We offer use of a drug database to search with these pharmacophores, and score the hit lists with GH-score formula[7] (see Methods for details).

We tested the above concept using the structure of HIV-1 protease and a series of ligands bound to it. Following the enumeration of pharmacophores, they are scored via GH-score, and the top scoring pharmacophore is then used to retrieve commercially available compounds from ACD. The hit list from ACD is matched with compounds listed at the NCI antiviral database, and of the 15 compounds that were also in this database; four were listed as moderately active against HIV.

Methods

Protein selection and alignment

The first step in this work was to decide on the protein-ligand complex for study. With the development of the LigScore scoring function available in Cerius2 4.5 ccL release,[8] 82 protein-ligand complexes were used and optimized as the training set for the empirical scoring function. Additional complexes were optimized in the process of fine-tuning the training set. The selection criteria for the protein of interest was based on two points: (1) there must be multiple complexes for the same protein to simulate different binding modes based on the flexibility of the receptor, and (2) the X-ray resolution of each complex studied must be < 2.3 Angstroms. Using these criteria, we focused our efforts on two complexes of HIV-1 Protease: 1HVL[9] and 4PHV.[10] These two complexes contained different bound ligands and illustrated the flexibility of the active site because several residues were in different positions relative to each other.

1HVL and 4PHV represent the same protein structure, but some of the residues in the active site have different 3D orientations due to flexibility of the protein upon binding with a ligand. Therefore, aligning the two protein/ligand complexes was a priority. Our main objective in this alignment was to match the active site characteristics along with the bound ligands simultaneously. The methodology for aligning the active sites was an atom-match approach. Using the alignment module of Cerius2,[8] the procedure constituted manual atom matches to be assigned prior to the alignment procedure. Additionally, one of the complexes can be chosen as a target structure in the alignment procedure. No other manipulation or optimizations were done to these two complexes.

Generation of pharmacophore queries

The method of choice for producing binding feature search queries based solely on the target structure was the Cerius2 Structure Based Focusing (SBF) module.[11] In this program, the first step involves choosing a center for the active site. There are a number of choices for selecting this sphere center including utilizing the bound ligand 3D coordinates or manually placing a pseudoatom in the center of the active site. For this purpose, we used the active site finding utility within the Cerius2-LigandFit program.[8] In this step, the active site was found by looking for crevices within the protein using a flood-filling algorithm.[12] With this technology, the entire active site can be mapped and used to determine the center of an active site domain by taking the centroid location of the points in the active site map.

After the active site center was chosen, the next step was to decide on an appropriate radius for the active site sphere. This sphere was used to pick protein residues that reside within the radius cutoff distance relative to the center point. Only the selected residues were used in the query

generation procedure. We altered the radius iteratively between 5 and 10 Å using a step-interval of 0.25 Å.

After selecting the appropriate radius, the next step was to elucidate the binding features in the protein using Cerius2-SBF. First, the procedure utilized the Cerius2-LUDI program,[13] to generate a LUDI interaction map of the relevant binding features. This interaction map contained three binding features: hydrogen bond acceptors (HBA), hydrogen bond donors (HBD) and lipophilic (LIPO) regions. For Catalyst[8] queries, the name of the LIPO feature will be changed to a hydrophobic (HYD) definition. In this study, we did not augment these three features with any additional Catalyst[14] binding features that may include ring aromatic, positive ionizable, negative ionizable and charged species. Some of the features that were seen in the original interaction map were removed due to binding orientation outside the binding cavity. The active site for HIV-1 Protease is not spherical, but rather is best described as a cylindrical channel through the protein.[10] Therefore, removing some of the interactions was important since they are buried inside the protein and not accessible for ligand binding. Next, binding features in the active site were clustered to reduce the overall number of interactions into a reasonable set of binding features for use in Catalyst 3D database mining. The clustering was done using the complete linkage algorithm in Cerius2-QSAR+.[8] Additional binding features were removed due to known binding orientations. Finally, a series of structure-based pharmacophore queries were produced based on binding features. With the Cerius2-SBF interface, this procedure was automated. The user can interactively change the total number of features in each query to be generated. Additional restrictions on the combinations of features in each query can be placed on the query generation program prior to producing the hypothesis files.

The Cerius2-SBF program allows for additional restraints to the queries based on the shape of the active site domain. In our study we utilized both types of restraints based on two separate methodologies. The first method involves using volume exclusions on each heavy atom in the protein active site sphere. The second type of restraint uses the Catalyst/Shape program (catShape).[15] This shape functionality is itself a query for searching Catalyst 3D databases. However, shape-based queries can be merged with any binding-feature based query. In both methodologies, the default parameters were used for these additional shape restraints.

3D database searches

The next step was to perform the databases searches in Catalyst. In all cases, we used the fast fit method[16,17] in Catalyst for database searching. To begin to prioritize queries, we first utilized the Derwent World Drug Index version 1999 database[18] built in Catalyst 3D format (WDI99). The goal of this step was to assess the ability of a single query to retrieve known hits from this set of drug molecules, otherwise defined as the active hits (H_a). The procedure called for us to create a small subset of the WDI99 database (53964 compounds total) that contained the words "HIV" and "Protease" in the mechanism of action (MA) field. This subset database contained a total of 134 compounds (WDI-HP). Additional HIV protease inhibitors that were not present in the WDI99 index were not added to this subset of 134 molecules. Interestingly, the two bound ligands in our dataset were not part of the WDI-HP dataset since neither of the molecules contained information in their respective MA field in the catalyst-formatted WDI database. Nevertheless, to maintain consistency of the analysis, these two bound ligands were not added to the WDI-HP subset.

The next search was designed to limit the size of the entire WDI99 database hits to only compounds that contained an entry in the MA field. With this restriction, the database size was reduced to 14,912 compounds (WDI-MA) versus the 53,964 in the full WDI99 index. This restriction is needed for consistency in the study since we had observed the lack of information in the MA field for the two bound ligands. The total number of hits retrieved from these searches is denoted as H_t: the size of the hit list retrieved for a search.

Prioritization of queries

An important component of this study is to search a database of available chemicals, to retrieve compounds that could potentially bind to the receptor active site. However, the need to prioritize the queries we have generated is a crucial step toward retrieving a set of compounds with a high hit rate for the *in silico* analysis that should translate into a good hit rate for the *in vitro* assay studies. With the H_a and H_t data for each query, an analysis metric called the Goodness of Hit (GH) score was utilized (see Figure 1 and equations 1-4).[7] In this metric, the hit list for each query is analyzed for hit rate, coverage, selectivity and enrichment. Coverage is defined as the percent ratio of actives in the hit list (%A - equation 1).

$$\%A \quad \frac{H_a}{A} x100 \tag{1}$$

The selectivity is a term used to describe the yield of

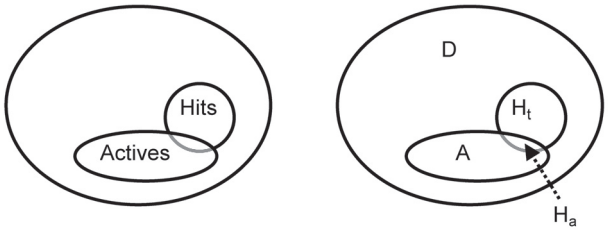

Figure 1. Diagram of terms in GH score, where *D* represents the compounds in the database, *A* represents compounds that are active and *Ht* represents hit list with *Ha* those compounds in the hit list that are active.

active compounds returned in a hit list, or the percent yield (% Y - equation 2).

$$\%Y \quad \frac{H_a}{H_t} x100 \tag{2}$$

The enrichment or enhancement is the ratio of the active compounds in the hit list to the number of compounds in the hit list over the active compounds in the database to the total number of compounds in the database (E - equation 3).

$$E \quad \frac{H_a / H_t}{A / D} \quad \frac{H_a x D}{H_t x A} \tag{3}$$

For each query, the GH score was calculated and used for ranking and prioritizing our list of Catalyst hypotheses (see equation 4).

$$GH \quad \frac{H_a(3A \quad H_t)}{4H_t A} \quad 1 \quad \frac{H_t \quad H_a}{D \quad A} \tag{4}$$

Search for new lead compounds

To continue the process of searching for new lead compounds, we used the Catalyst 3D-formatted[8] Available Chemical Directory version 1999 (ACD99):[19] These compounds are available for purchase and therefore are easily accessible for *in vitro* inhibition studies. The compounds returned from the ACD99 searches were used as input for searching to see if these structures were present within the National Cancer Institute version 2000 database[20] built in Catalyst format (NCI2000). The NCI2000 database contains information for approximately 23,000 compounds with biological activity data from HIV and cancer screens. Therefore, if compounds found from the ACD99 databases are shown to be inhibitors of the HIV virus via the NCI2000 database, we have confidence in selecting these compounds for *in vitro* studies as potential HIV-1 Protease inhibitors.

The NCI2000 database contains three data fields of interest to our work: the anti-viral screen conclusion (AVS_CONC), the anti-viral screen IC50 (AVS_IC50) and the anti-viral screen EC50 (AVS_EC50). Of particular importance for our prioritization is the AVS_CONC field. This data contains three possible entries: confirmed active (CA), confirmed moderate (CM) and confirmed inactive (CI). The AVS_CONC field was used to produce subset databases of the full NCI2000 index. The full NCI2000 index contains 238819 compounds, however the set of structures that contain information in the AVS_CONC field is only 40243. This will be denoted here as the NCI2000-AC database. Additionally, the NCI2000-AC index was further segmented into categories of compounds based on the AVS_CONC value. Therefore, we divided and denoted the following datasets: NCI2000-CA (399 compounds with CA in AVS_CONC field), NCI2000-CM (1005 compounds) and NCI-CI (38839 compounds).

Results and Discussion

Protein selection and alignment

Following the workflow diagram described in Figure 2, we aligned the two protein structures, 1HVL and 4PHV. The methodology we used required manual atom matches for the two complexes followed by target-based alignment

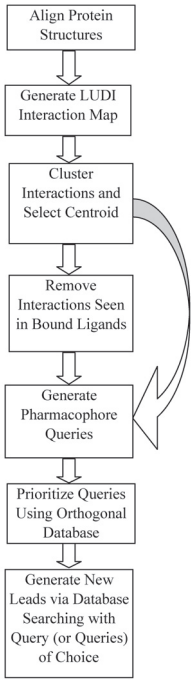

Figure 2. Workflow diagram.

of one protein to the other. We chose the 1HVL complex as the target protein in the procedure. The atom matches were comprised of equivalent heavy atoms in the following residues: ASP25, ASP29, ILE47. Since HIV-1 Protease is a dimer, there were a total of six residues that were used, or two for each residue ID number listed above. In the six residues, at total of 30 atoms were selected for the RMS alignment. After the alignment was performed, the RMS deviation for the 30 atoms was about 1.5 Å. The main deviation in the alignment occurred in the ILE47 residues while ASP25 and ASP29 were aligned without any significant difference in the 3D coordinates. This alignment met our goal of aligning the active site domains. Additionally, the bound ligands were aligned to our satisfaction in this single step for both of the protein-ligand complexes. We did not perform any additional ligand alignment. This process involved keeping the bound ligands complexed with the protein structure during the alignment and then removing the ligands following the alignment. The aligned ligands are displayed in Figure 3 in their bound conformations.

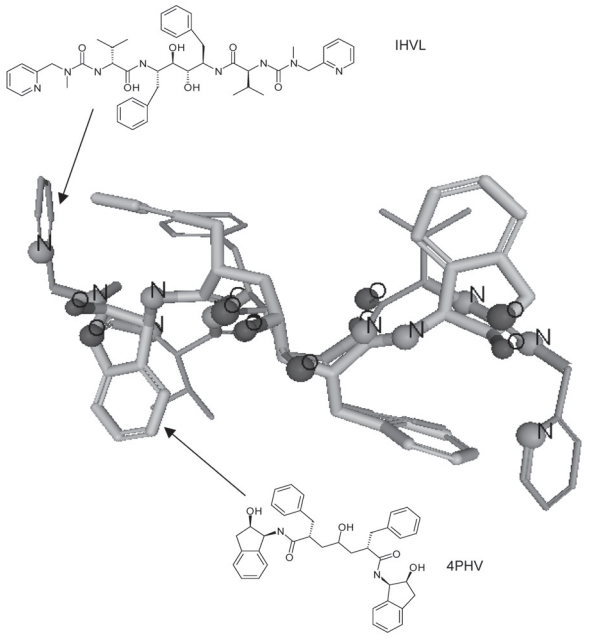

Figure 3. Alignment of the bound conformations of 1HVL and 4PHV (as listed at Brookhaven Protein Databank).

Pharmacophore query generation

The orientation and location of the active site is the initial requirement for active site-based pharmacophore generation in Cerius2-SBF. In this study, the center point was selected as described earlier using the active site map from Cerius2-LigandFit. For this work, we selected a radius of 9 Å, which met two criteria. First, all residues through the channel region of HIV-1 protease were selected, thereby giving us the ability to find new binding orientations not seen in the two bound ligands. Second, the cavity definition did not protrude into the solvent. To validate the active site domain, we overlaid the two bound ligands into their respective active sites to verify that all of the important binding features in the cavity were available for the generation of pharmacophore-based queries.

The LUDI interaction map generated as described above produced many binding features from the active site residues. Evaluation of the interaction map showed that some binding features seen were unimportant for ligand binding. Primarily, these features were buried in the protein active site cavity, as one may expect due to the active site identification method employed. A sphere as the geometric shape of this active site is not realistic. All residues either surrounding the surface of the active site or buried in the protein were selected initially. However, the interaction map was user editable. Therefore, we were able to delete the features in the interaction map that were not relevant. Additional features were removed where the vector was pointing toward solvent atoms, including the functional water molecule near the center of the active site.

To reduce the number of LUDI binding interactions, clustering the features into representative groups was the preferred option. However, the original clusters of HBD, HBA and LIPO regions needed some manipulation. Clustering can be adjusted by merging selected interactions into either a new cluster or into an existing cluster. This allows for full user-defined clusters of interactions that will be needed for the remaining steps in the procedure. After all clusters were defined appropriately, the total number for each type of interaction was as follows: HBD - 9, HBA - 7, LIPO - 9 for 1HVL and HBD - 10, HBA - 6, LIPO - 8 for 4PHV. The number of clusters is important because this is also the number of binding features that remain in the active site. To reduce the total number of binding features, the cluster centroid was selected and all other interactions were eliminated. To begin, the total number of interactions (25 for 1HVL and 24 for 4PHV) was used to generate millions of pharmacophore configurations.

We initially generated all combinations of 3, 4, 5, 6, 7 and 8 binding features for both 1HVL and 4PHV. Table 1 displays the total number of queries for each combination. As illustrated in columns B and C, there are over 3 million search queries if we consider all combinations of features. Since the time needed to perform all of these searches was considered prohibitive, we decided to place restrictions to

Table 1. Number of pharmacophore queries for each protein

A	B	C	D	E	F	G
3	2,300	2,024	23	22	165	9
4	12,650	10,626	253	231	330	36
5	53,130	42,504	1,771	1,540	462	84
6	177,100	134,596	8,855	7,315	462	126
7	480,700	346,104	33,639	26,334	330	126
8	1,081,575	735,471	100,947	74,613	165	84
Total	1,807,455	1,271,325	145,498	110,055	1,914	465

Column A is the total number of features in each query. Column B is the total number of pharmacophore configurations (queries) for 1HVL. Column C is the total number of queries of 4PHV. Column D is the total number of queries with the ASP25 HBD groups in each query for 1HVL. Column E is the total number of queries with the ASP25 HBD groups in each query for 4PHV. Column F is the total number of de novo queries for either 1HVL or 4PHV. Column G is the total number of de novo queries with the two ASP25 HBD groups in each query for either 1HVL or 4PHV.

reduce the number of queries. The Cerius2-SBF interface allows for restrictions on the type of queries generated based on many different criteria. In this work, we decided to focus our efforts on queries that contained 5 binding features (such as H-bond donor/acceptor, lipophilic interactions, etc.). For HIV-1 Protease inhibitors, 5-features queries seemed reasonable because of the high degree of flexibility in the compounds. For more rigid inhibitors, it would be logical to use a set of queries with fewer binding features to retrieve more active compounds from the database searches. This restriction reduced the number of queries to about 100,000. However, after considering the need to validate the series of queries, we decided to retain the two HBD features originating from the ASP25 residues in the site (see Figures 4 and 5). The next constraint was to force each query generated to include the two HBD features emanating from ASP25. In other words, the number of queries was restricted to combinations of 3 binding features plus the two ASP25 HBD features. These two binding features can be thought of as anchors for the generation of

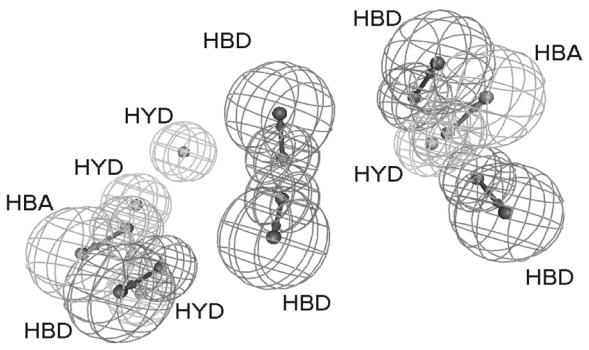

Figure 4. *De novo* binding features in the active site of 1HVL.

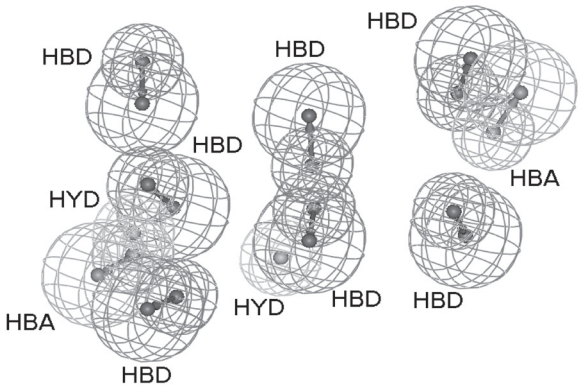

Figure 5. *De novo* binding features in the active site of 4PHV.

queries that followed. This operation substantially reduced the total number of combinations from 95,634 to 3311. Table 1 shows the breakdown of the total number of queries retained based on our selection protocol as described above.

After all of the possible binding features were identified, we focused on searching for *de novo* binding features in the HIV-1 Protease cavity. Therefore, we were interested in an interaction map that contained only binding features that were not seen in the two bound ligands of each complex. The hydrogen bonding interaction between the ASP25, ASP29, GLY27 and GLY48 residues of 1HVL and the residues ASP25, ASP29, GLY27 of 4PHV were removed, with the exception of the two ASP25 residues. Additional lipophilic sites where a hydrophobic fragment of the bound ligand was found to overlap in close proximity (less than 2 Å) to the defined centroid were also removed. This substantially reduced the set of binding features. In all, the group of ligand-based Catalyst queries was as follows: five HBD, two HBA and four HYD groups for an overall total of 11 features for 1HVL and, similarly for 4PHV, the total number of binding features was 11 with five HBD, two HBA and 2 HYD regions (see Figures 4 and 5). Requiring all combinations of these *de novo* binding features decreased the total number of queries from 3311 to 168.

Database searches and query prioritization

The full set of 3311 queries was used to search the WDI database and each query was evaluated by calculating the GH score for the hit lists.[7] Table 2 shows the top 10 ranked queries. One observation from Table 2 is that there are several queries that are redundant in the type of binding features and the 3-D coordinates. When all binding features

Table 2. The top 10 highest ranked queries by GH score (GH)

	Ha*	Ht*	% Y*	% A*	E*	GH*
WDI-MA Database	134	14912	0.9	100.0	1.0	0.000
4PHV_1132	75	633	11.8	56.0	13.2	0.220
1HVL_a7_186	75	633	11.8	56.0	13.2	0.220
1HVL_a8_205	80	828	9.7	59.7	10.8	0.210
1HVL_1131	76	737	10.3	56.7	11.5	0.209
4PHV_1125	76	737	10.3	56.7	11.5	0.209
1HVL_a8_207	68	578	11.8	50.7	13.1	0.208
1HVL_1132	78	817	9.5	58.2	10.6	0.206
4PHV_1126	78	817	9.5	58.2	10.6	0.206
1HVL_a8_209	73	712	10.3	54.5	11.4	0.204
1HVL_1140	82	1004	8.2	61.2	9.1	0.201

* see text for definition of terms.

Table 3. The top 10 highest ranked *de novo* queries by percent ratio of actives (%A)

	H_a	% A
WDI-HP Database	134	100
1HVL-84	53	39.6
1HVL-72	40	29.9
1HVL-83	39	29.1
1HVL-82	28	20.9
4PHV-83	25	18.7
1HVL-81	24	17.9
1HVL-28	21	15.7
1HVL-25	21	15.7
1HVL-64	21	15.7
1HVL-71	20	14.9

are available in each active site (1HVL and 4PHV), an overlap in binding features occurred because the active sites were identical for the specified interactions. Since this work focused on the prospect of generating novel leads through *de novo* binding features, we present the results in more detail below for this set of queries.

Each of the 168 *de novo* queries was used as the search criteria for retrieving known inhibitors of HIV-1 protease via the WDI-HP database, in order to determine the quantity H_a for each query. The measurement of H_a is a necessary value for addressing the coverage of active compounds in a particular search.[7] In this set of searches, query 1HVL-84 (see Figure 6) returned the highest hit rate for H_a: 53 out of a possible 134. The percent ratio of actives (%A) for this query was approximately 40%, i.e., 40% of the active compounds in the database were retrieved. Table 3 displays the top 10 ranked queries ranked by H_a and %A.

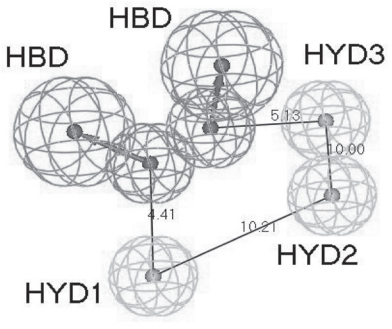

Figure 6. 1HVL-84 pharmacophore. Distance between hydrophobic groups: HYD1-HYD2 = 7.5 Å, HYD1-HYD3 = 15.0 Å, HYD2-HYD3 = 11.7 Å.

Analysis of the coverage of active compounds shows some interesting results. First, the coverage of actives drops quickly once we proceed past the highest ranked hypothesis. For 1HVL-84, we have coverage of actives of approximately 40% while the next highest ranked query, 1HVL-72, yields only about 30% coverage. 1HVL-83 gives about 30% coverage of actives, but the remaining seven searches yielded %A values less than or equal to 20%. Therefore, there is a clear delineation of priority of this set of queries when only %A values are considered.

Following the search for active compounds, we performed database searches on the WDI-MA database to retrieve the values for H_t, the number of compounds in a search hit list. The statistic of interest in this step was the percent yield of actives (%Y). The top 10 searches based on %Y ranking are in Table 4. In analyzing these results, it is clear that the 1HVL-84 query was a strong candidate because the percent yield was about 12%, i.e., 12% of the compounds in the hit list were designated as active. The percent yield may appear low, but is actually significant when compared to the 0.9% (134 of 14,912) of active compounds in the full database. Hence, we are retrieving a much higher percentage using our *in silico* search query than one may expect to retrieve from random searches. This robustness in percent yield is defined as the

Table 4. The top 10 highest ranked *de novo* queries by percent yield (%Y)

	Ha	Ht	% Y
WDI-MA Database	134	14912	0.9
1HVL-84	53	428	12.4
4PHV-83	25	300	8.3
1HVL-28	21	279	7.5
1HVL-82	28	420	6.7
1HVL-83	39	605	6.4
1HVL-63	15	258	5.8
1HVL-25	21	367	5.7
1HVL-71	20	356	5.6
1HVL-78	11	203	5.4
1HVL-23	19	358	5.3

enrichment (E) factor (see Equation 3). Table 5 ranks the top 10 ranked queries based on E. Once again, we observe that query 1HVL-84 is ranked first with an enrichment value of 13.8. This means that the probability of randomly picking up an active compound from the hit list is 13.8 times greater than that of from the full database. This result increases our confidence in using the 1HVL-84 query to generate new leads. However, we wish to rank the effectiveness of a query by measuring both the percent yield (%Y) of actives, and simultaneously, the percent ratio of actives (%A). As described above, the statistic that was utilized is known as the Goodness of Hit score (GH) (see Equation 4). The top 10 queries ranked by GH score can be seen in Table 6. In evaluating each hypothesis with this value, we again observe that the 1HVL-84 is superior (GH = 0.187) to all other *de novo* queries used to search for HIV-1 Protease inhibitors within the WDI-MA database. With all of the aforementioned evidence supporting the 1HVL-84 query, we have chosen this as the preferred query for searching for new and novel lead compounds.

One noteworthy result from this study was that 19 queries for the full 5-feature set of 3311 had a higher GH score than the preferred *de novo* query, 1HVL-84. Interestingly, each of these 19 pharmacophores contained the same binding features: 2 HBD groups and 3 HYD features. Clearly, queries containing three HYD groups produce a high rate of return on the known active compounds. It is expected to have the queries obtained from the full active site analysis to score better (*e.g.*, GH = 0.220 in Table 2) since there are no restrictions on which binding interactions that are considered. Note that even after the elimination of the known binding interactions (except those at ASP-25), the top pharmacophore, 1HVL-84, scored very well retrieving compounds with relatively high selectivity and high coverage. By using primarily the interactions that are not utilized by the known active compounds, we risked not being able to retrieve compounds that are active. But these results clearly validate the utility of the procedure proposed in this paper, as we were able to retrieve hit lists with good selectivity and coverage. The next step is to take the top queries and use them to retrieve potential leads from corporate, or commercially available chemicals databases, or even virtual libraries.

Table 5. The top 10 highest ranked *de novo* queries by enrichment (E).

	Ha	Ht	% Y	E
WDI-MA Database	134	14912	0.9	1.0
1HVL-84	53	428	12.4	13.8
4PHV-83	25	300	8.3	9.3
1HVL-28	21	279	7.5	8.4
1HVL-82	28	420	6.7	7.4
1HVL-83	39	605	6.4	7.2
1HVL-63	15	258	5.8	6.5
1HVL-25	21	367	5.7	6.4
1HVL-71	20	356	5.6	6.3
1HVL-78	11	203	5.4	6.0
1HVL-23	19	358	5.3	5.9

Table 6. The top 10 highest ranked *de novo* queries by GH score (GH).

	Ha	Ht	% Y	% A	E	GH
WDI-MA Database	134	14912	0.9	100.0	1.0	0.000
1HVL-84	53	428	12.4	39.6	13.8	0.187
1HVL-83	39	605	6.4	29.1	7.2	0.116
4PHV-83	25	300	8.3	18.7	9.3	0.107
1HVL-72	40	836	4.8	29.9	5.3	0.105
1HVL-82	28	420	6.7	20.9	7.4	0.100
1HVL-28	21	279	7.5	15.7	8.4	0.094
1HVL-25	21	367	5.7	15.7	6.4	0.080
1HVL-71	20	356	5.6	14.9	6.3	0.078
1HVL-23	19	358	5.3	14.2	5.9	0.074
1HVL-64	21	435	4.8	15.7	5.4	0.073

Search for new lead compounds

The search for novel lead compounds was performed on the ACD99 database with the highest ranked *de novo* query: 1HVL-84. To reiterate, the reasoning behind using ACD99 structures is based solely on the accessibility of these compounds for purchase and, therefore, their availability for *in vitro* studies. The ACD99 database contains a total of 231,003 compounds, with conformational models pre-built in the database. The 1HVL-84 query retrieved 1,276 compounds from ACD99. This *in silico* screen greatly reduced the set of compounds to a region in chemical space that possess the correct 3D binding orientations. One important aspect of the ACD99 is that there are many entries that contain different salt mixtures of the same primary structure. For example, one structure, chlorohexidine was returned nine times in our database search because of different salt combinations. Since our interest in this study focuses solely on the primary structure, we removed all of the multiple fragment hits from our hit list. This step reduced the size of our retrieved hits to 1,119. Each one of the 1,119 structures has the potential of being a new inhibitor of HIV-1 Protease.

We checked the presence of these structures in the NCI2000 database. This database contains information on inhibition of the HIV-1 virus. However, activity information is not target-dependent. Therefore, the reduction in HIV viral load could be attributable to many targets within the lifecycle of the HIV virus. However, if a compound is shown

to actively inhibit HIV, there is a distinct possibility that the inhibition is due to binding within the HIV-1 Protease active site that we are investigating in this study. This is the premise used to prioritize compounds for the ACD99 database searches.

Each of the 1119 structures returned from the ACD99 searches were used to scan the NCI2000-AC database, via exact structure match searches, to evaluate if any compounds existed with HIV viral screen information. Of the 1119 compounds, there were 15 structures that existed with some HIV screen information from the NCI2000-AC index. Of the 15 structures, none were confirmed active for the NCI anti-viral screen (NCI2000-CA). However, four of the structures were found to be confirmed moderate (NCI2000-CM) for HIV viral load reduction (see Table 7). Of these four structures, one compound, 60411, is too large to fit within the HIV protease active site moiety; hence its activity may be due to another mechanism. However, the other three structures (53287, 59597 and 20346; shown in Figures 7-9) are worth further investigation.

To verify that the three remaining structures could be considered new lead candidates, we used the WDI99 database, and the subset databases WDI99-HP and WDI99-MA, to see whether these structures were known HIV-1 Protease inhibitors. None of the three structures were

Table 7. HIV antiviral screen data for compounds in the ACD99 databases and the NCI2000 database of confirmed moderate activity

Compound	NSC ID [a]	IC50 [b]	EC50 [b]
53287	272671	500	352
59597	44184	193	142
20346	9222	250	250

[a] Compound ID in the NCI databases. [b] Values are in mM.

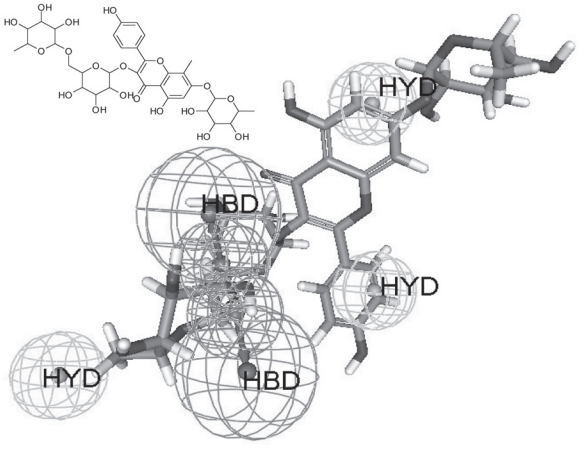

Figure 7. Compound 20346 retrieved from ACD99 database using 1HVL-84 hypothesis.

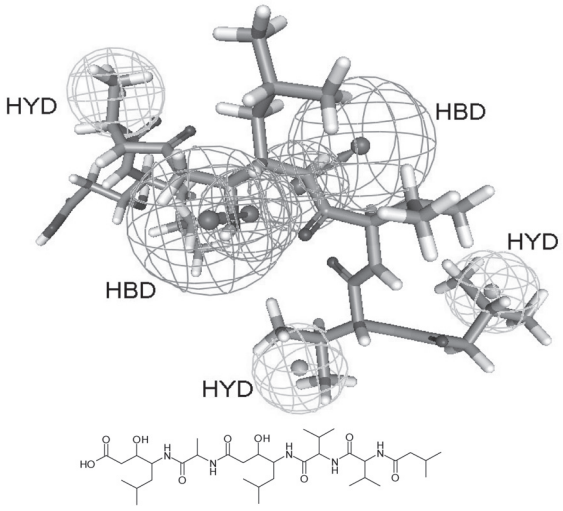

Figure 8. Compound 53287 retrieved from ACD99 database using 1HVL-84 hypothesis.

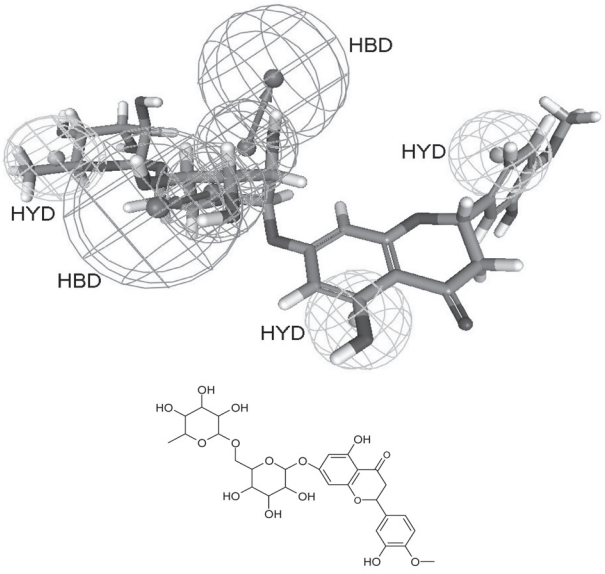

Figure 9. Compound 59597 retrieved from ACD99 database using 1HVL-84 hypothesis.

seen within the WDI-HP database. Structure 53287 (Pepstatin) was seen within the WDI99-MA database. However, the mechanism of action for this compound was identified as a peptide hydroxylase inhibitor.[18] Additionally, compound 59597 (Hesperidin) was also found in the full WDI99 database, but no MA information was provided for this compound. Structure 20346 (Robinin) was not included in the WDI99 database. Regardless of these results seen here, all three compounds are potential lead candidate for HIV-1 protease inhibition

since no evidence to the contrary was observed within our original set of structures used to determine the most important pharmacophore query.

Would the results have improved if we had chosen to enumerate the queries with six features instead of five? As previously discussed, the pharmacophore queries used in this study contained the two HBD groups derived from the two ASP25 residues in the active site (Figure 4 or 5). This resulted in a total of 252 queries (126 for each complex) containing different 6-feature combinations of HBD, HBA and HYD groups. Following the method employed for the 5-feature study against the WDI-HP and WDI-MA databases, we calculated the GH score to prioritize this set of queries (see Table 8).

Table 8. The top 10 ranked 6-feature *de novo* queries by GH score (GH).

	Ha	Ht	% Y	% A	Enrichment	GH score
WDI-MA Database	134	14912	0.9	100.0	1.0	0.000
1HVL-6f-56	4	11	36.4	3.0	40.5	0.280
4PHV-6-f-36	2	11	18.2	1.5	20.2	0.140
1HVL-6-f-111	6	37	16.2	4.5	18.0	0.133
1HVL-6-f-125	4	24	16.7	3.0	18.5	0.132
1HVL-6-f-16	1	7	14.3	0.7	15.9	0.109
1HVL-6-f-4	1	8	12.5	0.7	13.9	0.096
1HVL-6-f-55	5	45	11.1	3.7	12.4	0.092
1HVL-6-f-53	5	49	10.2	3.7	11.4	0.086
1HVL-6-f-71	1	9	11.1	0.7	12.4	0.085
1HVL-6-f-54	3	29	10.3	2.2	11.5	0.083

The results in Table 8 indicate that there is one 6-feature query that is superior to the others based on GH score: 1HVL-6f-56 (see Figure 10). Visual inspection of this query revealed that it is nearly identical to the favored query with five features (1HVL-84) with one major exception:

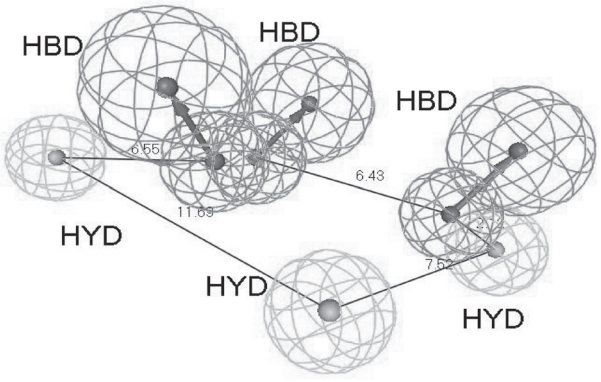

Figure 10. 1HVL-6f-56 pharmacophore. A six-feature model.

there is an additional HBA group in 1HVL-6f-56. The other five features, two HBD and the three lipophilic groups, are in similar positions as in the 1HVL-84 query.

Statistical comparisons between the 5-feature and 6-feature queries (1HVL-84 and 1HVL-6f-56, respectively) offer additional insights. First, the overall GH score is higher for the 6-feature query relative to the 5-feature query. This result is expected since the refinement of a hit list typically occurs when the number of features increase and a much smaller hit list is obtained, hence compromising the "coverage" of active space. 1HVL-84 is a hypothesis that returned nearly 40% of the active compounds, while 1HVL-6f-56 only returned about 3% of the total actives in the WDI99 database. The selectivity of the 6-feature query is far superior as seen by comparing the percent yields. 1HVL-6f-56 has a % Y of about 36% (4 of only 11 total hits) versus 12% (53 of 428) found for the 5-feature query. Hence, false positives are lower with the 6-feature query. Consequently, the enrichment value is also higher for the 6-feature query. Since GH score is a measurement of both the selectivity and the coverage, the 6-feature query appears to be the better choice based on these considerations in terms of providing a fewer but potentially higher quality leads. However, in this work, we have decided that when identifying new modes of binding in a protein target, we would prefer to have a larger set of new lead compounds versus a more selective set of structures, which may have a higher *in vitro* hit rate. In our opinion for this study, a query with higher coverage of our active compounds, in this case, is more important that one with high selectivity, hence our overall choice of the 5-feature query (1HVL-84).

The 1HVL-6f-56 query was used to search the ACD99 database. The number of hits (70 single fragments) returned was much smaller, as expected, than the 1,119 compounds from the 5-feature query. Similarly, these 70 compounds were used as 2D substructure queries for searching the NCI2000 database. Of the 70 structures, two compounds were returned with confirmed moderate HIV-1 inhibitory activity. There were no compounds returned with confirmed activity for AVS_CONC in the NCI2000 database. Interestingly, these two compounds, 59597 and 60411, were also identified by the 5-feature query: 1HVL-84. This buttresses our focus on very similar compounds since these two best-ranked queries have very similar binding features oriented in 3D space. Of the two compounds common to hit lists obtained from the 5- and 6-feature query searches of the ACD99 and NCI2000 databases, compound 59597 (Figure 9) is the new HIV-1 Protease inhibitor candidate identified in this study since we have determined that compound 60411 was too large for this active site.

Conclusions

In this work, we report a methodology for generating new lead compounds from *de novo* binding features derived from a protein active site domain. This example illustrated the utility of extracting novel structures that may alleviate a fundamental concern with the HIV virus: drug resistance. By utilizing the d*e novo* binding features not seen in known inhibitors, we have demonstrated that new classes of compounds can be found *in silico* and, hopefully, produce new alternative drug candidates via *in vitro* screening results. A key aspect of this methodology is that it can be applied to any target structure where there are at least two bound ligands in the X-ray structures. While in this proof of concept work we used only two complexes (1hvl and 4phv), utilization of additional complexes will provide a more complete coverage of the feature space of the enzyme active site. Our hope is that this procedure will gain widespread applications for other protein classes where drug resistance is a motivating factor in the generation of new and novel lead candidates.

Acknowledgments

The authors would like to express gratitude to Omoshile Clement and Marvin Waldman for their critical review of this manuscript.

References

1. Hansch, C.; Leo, A.; *Exploiting QSAR: Fundamentals and Applications in Chemistry and Biology*, ACS Professional Reference Book: USA, 1995.
2. Güner, O. F. ed.; *Pharmacophore Perception, Development, and Use in Drug Design*, International University Line: La Jolla, 2000.
3. Kawakami, Y; Inoue, A.; Kawai, T.; Wakita M.; Sugimoto, H.; Hopfinger, A. J.; *Bioorg. Med. Chem*. **1996**, *4*, 1429.
4. Weinstock, J.; Keenan, R. M.; Samanen, J.; Hempel, J.; Finkelstein, J. A.; *J. Med. Chem*. **1991**, *34*, 1514.
5. Dorsey, B. D.; Levin, R. B.; McDaniel, S. L.; Vacca, J. P.; Guare, J. P.; *J. Med. Chem*. **1994**, *37*, 3443.
6. Kaldor, S.W.; Kalish, V. J.; Davies II, J. F.; Shetty, B. V.; Fritz, J. E.; Appelt, K.; Burgess, J. A.; Campanale, K. M.; Chirgadze, N. Y.; Clawson, D. K.; Dressman, B. A.; Hatch, S. D.; Khalil, D. A.; Kosa, M. B.; Lubbehusen, P. P.; Muesing, M. A.; Patick, A. K.; Reich, S. H.; Su, K. S.; Tatlock, J. H.; *J. Med. Chem*. **1997**, *40*, 3979.
7. Güner, O. F.; Henry, D. R. In *Pharmacophore Perception, Development, and Use in Drug Design;* Güner, O. F. ed., International University Line: La Jolla, 2000, p. 191.
8. *Cerius2* and *Catalyst* are software package available from Accelrys Inc., 9685 Scranton Road, San Diego, CA, 92121, USA, http://www.accelrys.com
9. Hosur, M. V.; Bhat, T. N.; Kempf, D. J.; Baldwin, E. T.; Liu, B. S., Gulnik, S.; Wideburg, N. E.; Norbeck, D. W.; Appelt, K.; Erickson, J. W.; *J. Am. Chem. Soc*. **1994**, *116*, 847.
10. Bone, R.; Vacca, J. P.; Anderson, P. S.; Holloway, M. K.; *J. Am. Chem. Soc*. **1991**, *113*, 9382.
11. Kirchhoff, P. D.; Brown, R.; Kahn, S.; Waldman, M.; Venkatachalam, C. M.; *J. Comput. Chem*. **2001**, *22*, 993.
12. Peters, K.P.; Fauck, J.; Frommel, C.; *J. Mol. Biol.*, **1996** *256*, 201.
13. Böhm, H.-J.; *J. Comput.-Aided Mol. Design* **1992**, *6*, 243.
14. Kurogi, Y.; Güner, O. F.; *Curr. Med. Chem*. **2001**, *8*, 1035.
15. Hahn, M.; *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 80.
16. Güner, O. F.; Waldman, M.; Hoffman, R.; Kim, J.-H. In *Pharmacophore Perception, Development, and Use in Drug Design;* Güner, O. F. ed., International University Line: La Jolla, 2000, p 215.
17. Murrall, N. W.; Davies, E. K.; *J. Chem. Inf. Comput. Sci*. **1990**, *30*, 312.
18. Derwent, Alexandria, VA USA, http://www.derwent.com/worlddrugindex/. accessed in the Catalyst version of World Drug Index (WDI99).
19. MDL Information Systems, Inc. San Leandro, CA, USA, http://www.mdli.com/products/acd.html, accesed in the Catalyst version of Available Chemical Directory (ACD99).
20. National Cancer Institute, Bethesda, MD, USA, http://www.nci.nih.gov/, accesed in the Catalyst version of the National Cancer Institute database version 2000 (NCI2000).

Received: April 11, 2002
Published on the web: October 31, 2002