Analysis of HIV Wild-Type and Mutant Structures via in Silico Docking against Diverse Ligand Libraries

Max W. Chang,*,† William Lindstrom,‡ Arthur J. Olson,‡ and Richard K. Belew†,§

Bioinformatics Program, University of California−San Diego, La Jolla, California 92093, Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92037, and Department of Cognitive Science, University of California−San Diego, La Jolla, California 92093

Received February 1, 2007

The FightAIDS@Home distributed computing project uses AutoDock for an initial virtual screen of HIV protease structures against a broad range of 1771 ligands including both known protease inhibitors and a diverse library of other ligands. The volume of results allows novel large-scale analyses of binding energy "profiles" for HIV structures. Beyond identifying potential lead compounds, these characterizations provide methods for choosing representative wild-type and mutant protein structures from the larger set. From the binding energy profiles of the PDB structures, a principal component analysis based analysis identifies seven "spanning" proteases. A complementary analysis finds that the wild-type protease structure 2BPZ best captures the central tendency of the protease set. Using a comparison of known protease inhibitors against the diverse ligand set yields an AutoDock binding energy "significance" threshold of −7.0 kcal/mol between significant, strongly binding ligands and other weak/nonspecific binding energies. This threshold captures nearly 98% of known inhibitor interactions while rejecting more than 95% of suspected noninhibitor interactions. These methods should be of general use in virtual screening projects and will be used to improve further FightAIDS@Home experiments.

INTRODUCTION

The FightAIDS@Home project (FAAH; http://fightaidsathome.scripps.edu/) utilizes the World Community Grid distributed computing network to conduct virtual screens for new inhibitors against HIV protease. The project is built around AutoDock,[1] which uses a Lamarckian genetic algorithm (a hybrid of evolutionary algorithm sampling with local search methods) to search for the optimal conformation of a given ligand in relation to a target receptor structure. Currently, FAAH is installed on approximately 450 000 clients and is capable of screening almost 10 000 ligands per day. As part of a larger screening process seeking new protease inhibitors effective against both wild-type HIV and emerging drug-resistant mutants, FAAH completed an initial screen of approximately 1800 ligands against 268 HIV protease structures, totaling almost 500 000 different dockings and more than 10^{15} separate energy evaluations.

Though the scale of these in silico experiments is huge, growing databases of proteins and chemical structures have the potential to surpass these resources. As FAAH moves forward, techniques to judiciously choose informative structures and ligands remains important. The set of protease structures considered in FAAH includes a large number of modeled structures. This analysis focuses on structures taken from the Protein Data Bank (PDB),[2] consisting of 71 wild-type and mutant proteases. More specifically, these structures include 26 wild-type HIV-1, 33 mutant HIV-1, and 12 HIV-2. The ligand library used in the current FAAH experiment consists of 11 known protease inhibitors and compounds from the National Cancer Institute (NCI) Diversity Set (http://dtp.nci.nih.gov/branches/dscb/diversity_explanation.html). The NCI Diversity Set is chosen specifically to represent a broad sampling of pharmacophores, providing a characterization of protease docking modes "outside the box", beyond those represented by currently approved protease inhibitors. An overview of the current FAAH data set is shown in Table 1.

One of the long-term goals of this research is to discover compounds that can inhibit a broad range of mutant proteases, so a variety of HIV protease mutant structures are considered. A similar method has previously proven successful in developing an inhibitor effective against FIV, SIV, and HIV.[3] Due to the rapid evolution of drug resistance, such approaches are vital in the design of new inhibitors. Cross-resistance involving current Food and Drug Adminsitration (FDA)-approved drugs is also a continuing problem.[4]

Other research has addressed docking against an ensemble of protein structures.[5,6] However, these studies focus on molecular dynamics-based "snapshots" of the protein in motion. A more recent paper by Fernandes et al. addresses the problem of docking to multiple structures in an effort to develop inhibitors effective against a set of targets.[7] Their work includes the comparison of docking results from several HIV protease structures, showing that ligands adopt similar binding modes across various proteases.

Our work incorporates a much larger number of ligands and proteases. Hayashi et al. use such methods to generate ligand profiles by docking small molecules against a panel of various proteins.[8] A key feature of their work is the use of vectors of binding energies as a way of describing

* Corresponding author e-mail: mchang@ucsd.edu.
† Bioinformatics Program, University of California−San Diego.
‡ The Scripps Research Institute.
§ Cognitive Science, University of California−San Diego.

10.1021/ci700044s CCC: $37.00 © 2007 American Chemical Society
Published on Web 04/21/2007

Table 1. Overview of FAAH Ligands and Protease Structures

proteases	wild-type	mutant	HIV-2
number of structures	26	33	12
unique (by sequence)	~1	10	2
ligands	**known inhibitors**	**NCI Diversity Set**	
number of compounds	11	1760	

particular ligands. Complementary methods that focus on proteins rather than ligands are discussed in this study, with applications toward finding consensus and representative proteases. The consensus protease structure that best captures the central tendency of the larger set of structures would prove useful in more focused virtual screening experiments. Such a structure was constructed by Vinkers et al., using averaged 3D coordinates from a set of crystallographic data.[9] We describe an approach based on binding energy profiles below.

METHODS

Data Set. The ligand library used in FAAH consists of 11 known protease inhibitors and 1990 compounds from the NCI Diversity Set. Known inhibitors include eight FDA-approved compounds: amprenavir, atazanavir, indinavir, lopinavir, nelfinavir, ritonavir, saquinavir, and tipranavir. The remaining three known inhibitors are TL-3, KNI-272, and JE-2147. Structures for all of the known inhibitors can be found in the PDB (see also the Supporting Information for a list of the inhibitors and their PDB codes). Of the compounds from NCI, 153 could not be processed correctly for AutoDock, due to the presence of metal atoms or multiple fragments, and so were not included in this study. An additional 77 were removed due to extremely poor binding, leaving a total of 1760.

Characterizing "wild-type" for a quasi-species like HIV is a notoriously difficult problem. Two common characterizations of subtype B of the HIV-1 virus are the "consensus B" sequence and "HXB2".[10] Since proteases for these two differ only at positions 3 and 37, sequences matching either are considered wild-type HIV-1 structures. The protease structures analyzed include 26 wild-type HIV-1, 33 mutant HIV-1, and 12 HIV-2.

Docking Protocol. Atomic coordinates for the HIV proteases were obtained from the PDB. The ligand and crystallographic waters were removed with the exception of the water bridging the flaps. When absent from the crystal structure, a water molecule was placed with hydrogen atoms oriented to facilitate the hydrogen-bonding pattern commonly observed in HIV protease.[11] Polar hydrogens were added, and Kollman charges were assigned to all atoms. Affinity grids centered on and encompassing the active site were calculated with 0.375 Å spacing using AutoGrid 4.

The NCI Diversity Set was processed for input to AutoDock 4. Gasteiger charges were assigned to all atoms, and rotatable bonds were assigned using AutoDockTools.

AutoDock 4 was used to evaluate ligand binding energies over the conformational search space using the Lamarckian genetic algorithm. Default docking parameters were used with the following exceptions: ga_pop_size, 200; ga_num_evals, 10 000 000; and ga_run, 100. For this study, only the minimum energy found is considered.

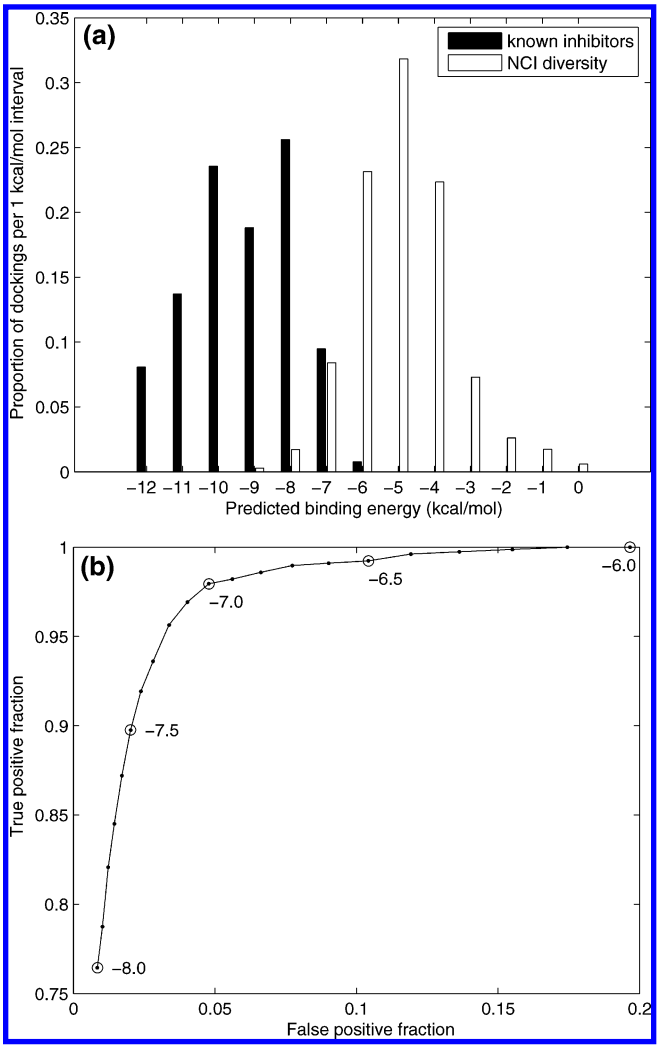

Figure 1. (a) Comparison of the distribution of binding energies for known inhibitors and NCI Diversity Set compounds. (b) ROC curve showing a sensitivity/specificity tradeoff for threshold values from −8 to −6 kcal/mol.

RESULTS

Discrimination between Specific and Nonspecific Interactions. AutoDock seeks the best interaction energy between a flexible ligand and the protein surface. Computed energies are typically favorable since the docking procedure searches widely. Since FAAH is ultimately focused on lead discovery, accurate discrimination between weak and strong binding is of vital importance. Toward this end, the differences in binding energy between NCI diversity compounds and known inhibitors can be used to determine a threshold at which interactions become significant.

Nearly all of the diversity compounds exhibit weak or moderate binding energies when compared to the known inhibitors. In order to determine the threshold at which specific binding is expected, the distribution of binding energies for the NCI Diversity Set is compared against the energies from the known inhibitors. At least with respect to HIV protease, derivation of a specific-interaction threshold represents an especially appropriate system due to the availability of a large number of positive controls (known inhibitors). The distribution of binding energies for both known inhibitors and the NCI Diversity Set is shown in Figure 1a.

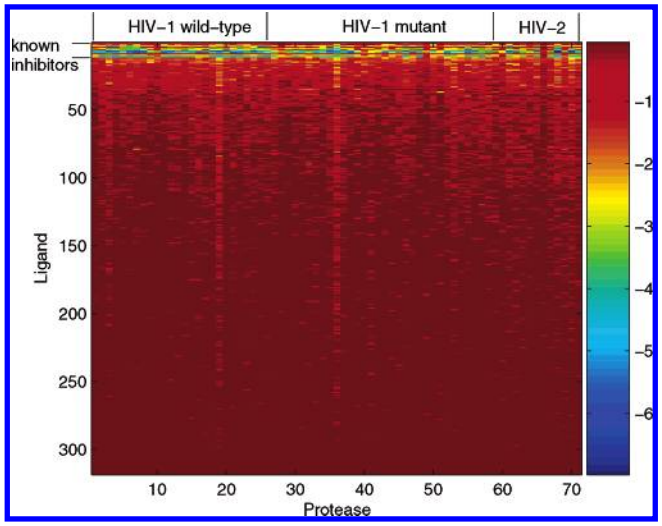

Figure 2. Specific energy interaction map. Each energy value indicates the level of binding beyond −7.0 kcal/mol. Ligands are sorted by ascending average binding energy.

The receiver operating characteristic (ROC) curve in Figure 1b demonstrates the effect of several threshold values that attempt to separate the known inhibitors and diversity compounds. For the purposes of the plot, the positive class contains only known protease inhibitors and the negative class contains all ligands from the NCI Diversity Set. A −7.0 kcal/mol threshold was selected as the significance cutoff in future experiments. At this level, only a small fraction (5.3%) of all dockings are considered "specific" interactions, which includes 97.7% of known inhibitor dockings and 4.7% of NCI Diversity Set dockings.

Figure 2 shows the degree to which predicted ligand−protease interaction energies exceed the −7.0 kcal/mol threshold. These are organized as portrayed in Table 1, with known inhibitors along the top and wild-type protease to the left. In addition, ligands have been sorted by average interaction energy, with the most favorable (i.e., most negative energies) near the top. As shown in the figure, there are wide variations in binding energy for single ligands docked against multiple proteases. Note that there are noticeable differences even among relatively homogeneous sets, such as the wild-type structures (columns 1−31). This variation underscores the importance of judicious protein structure selection in order to obtain the best binding energy estimates.

Determining a Consensus HIV Protease Structure. The large number of both ligands and protease structures tested in FAAH Stage 1A represents an opportunity to analyze the similarity of ligand/protein interactions across both dimensions of variability. In general, if a ligand binds poorly with one protease structure, it is not expected to bind strongly to others. If a single protease structure is found to capture the central tendency of the entire set, this single probe can be used as an initial probe against large libraries of ligands.

Representing protease structures as vectors of binding energies allows a direct mathematical characterization of a consensus protease structure as their *centroid*. That is, each of the proteases corresponds to a point in a high-dimensional space. The centroid is found by taking an average across all 1760-element vectors of binding energy values. (The specific interaction threshold deduced earlier could be used to eliminate less favorable docking results from this analysis. However, since the analysis gains information from the full set of both specific and nonspecific binding energies, instances of weak and nonspecific binding are retained.) When a Euclidean distance measure is used, 2BPW is the closest structure to the centroid. This remains true, whether the average is taken across only wild-type protease or across all proteases.

Table 2. Representative Protease Structures[a]

PDB ID	description
1HII*	HIV-2
1GNM	HIV-1 with V82D mutation
1BDL*	HIV-1 with heavily mutated 30-loop
2BPZ*	HIV-1 wild-type
7UPJ	HIV-1 wild-type
1AJX	HIV-1 wild-type
5UPJ	HIV-2
1HVI	HIV-1 wild-type
1HVJ	HIV-1 wild-type
1HVK	HIV-1 wild-type
1HSI*	HIV-2 apo (no ligand bound in crystal structure)
1AID*	HIV-1 with minor drug resistance mutations
3AID	HIV-1 with minor drug resistance mutations
1BDQ*	HIV-1 with heavily mutated 30-loop and drug
1MEU*	HIV-1 with major drug resistance mutations

[a] The coefficient for each structure is at least 2 standard deviations from the mean for at least one principal component. An asterisk (*) indicates proteases that are maximally loaded across at least one principal component.

Representative Protease Structures. While the centroid provides a convenient characterization of the central tendency across all proteases, identification of a larger set of "spanning" protease structures is also useful, to allow efficient screening of large libraries that capture the full breadth of observed results. To generate such a set of representative protease structures, principal component analysis (PCA) was used on the matrix of protease−ligand binding energies. By convention, columns of the matrix correspond to proteases and rows represent ligands. (As above and for the same reasons, weak and nonspecific interaction energies are included in this analysis.)

In brief, PCA identifies a small set of principal components (orthogonal basis vectors) that capture most of the variance within high-dimensional data sets. Because the principal components are linear combinations of the observed data, they cannot be interpreted directly. Since we seek a small set of spanning, nearly orthogonal protease structures, we therefore consider those proteases which *load most heavily* along each principal component.

We consider a 10-dimensional PCA. The first principal component serves as a scaling factor, accounting for approximately 90% of the variance in the data set, with all protease loading coefficients very close in value. The sum of the 2nd through 10th eigenvalues account for approximately 70% of the remaining variance. Any protease's loading coefficient on a principal component greater than two standard deviations from the mean is deemed significant, and the corresponding protease is added to the set of spanning protease. The resulting set of 16 representative structures is shown in Table 2; also shown are those structures that maximally load on a principal component.

These results align closely with expectations that major structural changes should affect binding energy. From the proteases studied, the main delineations are the presence of a heavily mutated loop region, the absence of a ligand in the crystal structure, and HIV-1 versus HIV-2. Drug-resistance mutations also seem to play an important role.

The first two principal components can be visualized in a two-dimensional space (cf. Figure 3a). The set of representative structures roughly bounds the periphery, while the consensus structure is centrally located. A multidimensional scaling plot of the same data using Sammon's nonlinear mapping (NLM)[12,13] in Figure 3b demonstrates this behavior even more clearly.

In contrast to PCA, NLM reduces dimensionality via explicit local gradient minimization of a "stress" (error) function:

$$E = \frac{\sum_{\mu=1}^{n-1} \sum_{\nu=\mu+1}^{n} \frac{[d^*(\mu,\nu) - d(\mu,\nu)]^2}{d^*(\mu,\nu)}}{\sum_{\mu=1}^{n-1} \sum_{\nu=\mu+1}^{n} d^*(\mu,\nu)} \qquad (1)$$

reflecting cumulative error $d^*(\mu,\nu) - d(\mu,\nu)$ in measuring the distance between n pairs of points μ,ν in the original d^* data space versus the reduced-dimensional space d. Given the stress function, minimization can be accomplished by a number of algorithms. Local search methods consider gradient change in stress within local neighborhoods, as potential placements in the reduced dimensional space are considered. In general, as with all local minimization procedures, there is no guarantee that the reduced dimensional solution is unique or globally optimum. In the current application, however, the Sammon mapping helps to confirm the basic pattern of the PCA solution and provides additional indications that the particular mutants identified do indeed span the larger set.

Binding Energy/Sequence Relationship. In ideal cases, relationships between protein sequence and function are obvious. For example, when dealing with protein crystal structures, factors other than sequence can have major effects. To determine the degree to which sequence and binding energy are coupled in this data set, a comparison between a sequence similarity matrix and a binding energy correlation matrix was performed. The sequence similarity matrix corresponds to the fraction of identical positions between sequence pairs. For binding energies, the matrix containing pairwise Pearson linear correlation coefficients is calculated. While the correlation between the two matrices is low, $r = 0.232$, the Mantel test demonstrates a statistically significant relationship between them. When 100 000 random permutations are used, the empirically derived p value is 0.015.

DISCUSSION

The huge computational resources provided by the FAAH project have provided a wealth of docking information. In addition to the primary purpose of identifying novel inhibitors, this data can be used to calibrate and focus future experiments. Several novel analyses were carried out using the large body of docking results. Considering protease structures in the context of in silico dockings against diverse libraries of ligands provides a perspective on how similarities and dissimilarities among them could not be anticipated, for example, on the basis of sequence identity alone.

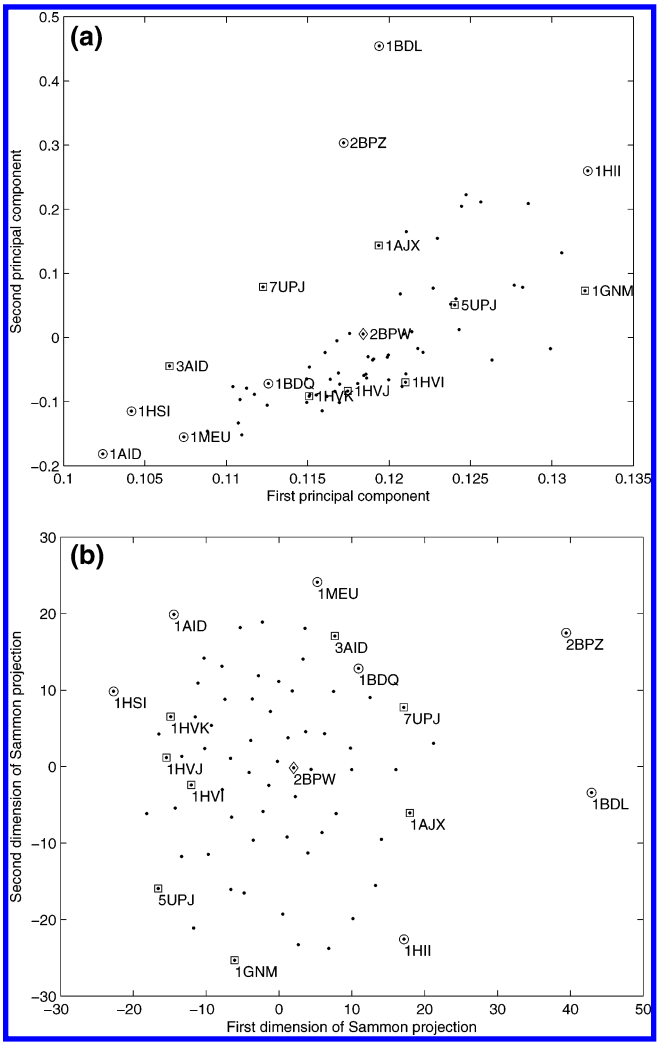

Figure 3. Representative protease structures plotted using (a) the first two principal components and (b) multidimensional scaling with Sammon mapping. Maximally loaded structures are labeled using circles; other highly loaded structures are labeled using squares. The consensus protease structure, 2BPW, is represented with a diamond.

In a comparison of binding energies between compounds specifically designed to act as protease inhibitors and approximately random compounds drawn from the NCI Diversity Set, a threshold of −7.0 kcal/mol works well to discriminate between putative specific and nonspecific binding with HIV protease. Applying this threshold to data sets may be useful in filtering out noise in weakly binding compounds. While this cutoff is specific to AutoDock and the protease system, the general approach is broadly applicable to other users of AutoDock (a widely used tool) and other docking systems.

The consensus protease structure, along with the other representatives, constitutes a limited set which captures the breadth of the entire set of protease structures. Rather than directly capturing specific structural elements, these structures are characterized by their affinity with a diverse set of ligands. The PCA-based approach for choosing representatives is able to capture protease structures that lie on the

periphery of the data set (Figure 3). Further applications of this technique may be useful in broader structural comparison and classification. In the more immediate future, the set of representatives will allow FAAH to continue screening larger libraries while maintaining breadth in its range of targets.

The enormous search capacity provided by the FAAH computing platform allows in silico experimentation on an unprecedented scale. It is not a coincidence, however, that the primary techniques we describe in this paper are all designed to *restrict* our experiments to especially informative cases; selection of the "centroid" wild-type structure and "spanning" viral structures provide two examples. Despite strong growth in computing power we can anticipate for the foreseeable future, high-throughput experimental methods and growing libraries of potential ligands generate a range of potential experiments that dwarf even these resources. Techniques supporting the judicious selection of informative structures and ligands will need to grow apace.

ACKNOWLEDGMENT

This work was supported by NIH grant PO1 GM48870. The FightAIDS@Home project is made possible by the World Community Grid, with technical and financial support by the IBM corporation. The authors would like to thank the members of the Molecular Graphics Laboratory for helpful discussions, especially David Goodsell, Garrett Morris, and Ruth Huey. The contributions from all members of WCG/FAAH are greatly appreciated.

Supporting Information Available: A table of inhibitors and their PDB codes and PDB codes for HIV-1 wild-type proteases, HIV-1 mutant proteases, and HIV-2 proteases. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem.* **1999**, *19*, 1639−1662.

(2) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(3) Lee, T.; Laco, G. S.; Torbett, B. E.; Fox, H. S.; Lerner, D. L.; Elder, J. H.; Wong, C. H. Analysis of the S3 and S3′ Subsite Specificities of Feline Immunodeficiency Virus (FIV) Protease: Development of a Broad-Based Protease Inhibitor Efficacious against FIV, SIV, and HIV in Vitro and ex Vivo. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 939−944.

(4) Kutilek, V. D.; Sheeter, D. A.; Elder, J. H.; Torbett, B. E. Is Resistance Futile? *Curr. Drug Targets: Infect. Disord.* **2003**, *3*, 295−309.

(5) Knegtel, R. M.; Kuntz, I. D.; Oshiro, C. M. Molecular Docking to Ensembles of Protein Structures. *J. Mol. Biol.* **1997**, *266*, 424−440.

(6) Carlson, H. A.; McCammon, J. A. Accommodating Protein Flexibility in Computational Drug Design. *Mol. Pharmacol.* **2000**, *57*, 213−218.

(7) Fernandes, M. X.; Kairys, V.; Gilson, M. K. Comparing Ligand Interactions with Multiple Receptors via Serial Docking. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1961−1970.

(8) Hayashi, Y.; Sakaguchi, K.; Kobayashi, M.; Kobayashi, M.; Kikuchi, Y.; Ichiishi, E. Molecular Evaluation Using in Silico Protein Interaction Profiles. *Bioinformatics* **2003**, *19*, 1514−1523.

(9) Vinkers, H. M.; de Jonge, M. R.; Daeyaert, E. D.; Heeres, J.; Koymans, L. M.; van Lenthe, J. H.; Lewi, P. J.; Timmerman, H.; Janssen, P. A. Inhibition and Substrate Recognition−A Computational Approach Applied to HIV Protease. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 567−581.

(10) Leitner, T.; Foley, B.; Hahn, B.; Marx, P.; McCutchan, F.; Mellors, J.; Wolinsky, S.; Korber, B. *HIV Sequence Compendium 2005*; Theoretical Biology and Biophysics Group, Los Alamos National Laboratory: Los Alamos, NM, 2005.

(11) Wlodawer, A. Rational Approach to Aids Drug Design through Structural Biology. *Annu. Rev. Med.* **2002**, *53*, 595−614.

(12) Sammon, J. W. A Nonlinear Mapping for Data Structure Analysis. *IEEE Trans. Comput.* **1969**, *C-18*, 401−409.

(13) Lerner, B.; Guterman, H.; Aladjem, M.; Dinstein, I.; Romem, Y. On Pattern Classification with Sammon's Nonlinear Mapping − An Experimental Study. *Pattern Recognit.* **1998**, *31*, 371−381.

(14) Sanner, M. F. Python: A Programming Language for Software Integration and Development. *J. Mol. Graphics Modell.* **1999**, *17*, 57−61.

CI700044S