Binding Response: A Descriptor for Selecting Ligand Binding Site on Protein Surfaces

Shijun Zhong and Alexander D. MacKerell, Jr.*

Computer-Aided Drug Design Center, Department of Pharmaceutical Sciences, University of Maryland, 20 Penn Street, Baltimore, Maryland 21201

Received April 26, 2007

The identification of ligand binding sites on a protein is an essential step in the selection of inhibitors of protein−ligand or protein−protein interactions via virtual database screening. To facilitate binding site identification, a novel descriptor, the binding response, is proposed in the present paper to quantitatively evaluate putative binding sites on the basis of their response to a test set of probe compounds. The binding response is determined on the basis of contributions from both the ligand−protein interaction energy and the geometry of binding poses for a database of test ligands. A favorable binding response is obtained for binding sites with favorable ligand binding energies and with ligand geometries within the putative site for the majority of compounds in the test set. The utility of this descriptor is illustrated by applying it to a number of known protein−ligand complexes, showing the approach to identify the experimental binding sites as the highest scoring site in 26 out of 29 cases; in the remaining three cases, it was among the top three scoring sites. This method is combined with sphere-based site identification and clustering methods to yield an automated approach for the identification of binding sites on proteins suitable for database screen or *de novo* drug design.

1. INTRODUCTION

Over the past decade, significant developments have been made in the field of chemical biology. Much of this effort has included the identification of inhibitors of protein−protein interactions involved in signal transduction.[1−4] Such inhibitors act as essential research tools for elucidating the role of a particular protein−protein interaction in different biological processes.[5] In addition, such compounds have the potential to be developed into novel therapeutic agents for disease states associated with perturbations of signaling pathways. Toward these goals, computational methods, in particular, virtual database screening, have made and continue to make significant contributions.[6−8] Typically, virtual screening may be used to select low-molecular-weight chemical compounds from a large database, with the selected compounds subjected to biological assays. These efforts often lead to the identification of compounds with binding affinities in the low-micromolar range, a level of affinity often sufficient for their use as research tools and as therapeutic agents.[9]

One of the challenges of applying virtual screening toward the identification of inhibitors of protein−protein interactions is the absence of a well-defined small-molecule binding site. Protein−protein interfaces are typically quite extensive, containing a number of putative binding sites.[10−13] While experimental data, such as results from mutation experiments, are useful for the identification of appropriate binding sites, such data are often not available, making theoretical approaches for binding site identification attractive. In the present work, a new method for the prediction of sites suitable for the binding of low-molecular-weight ligands is presented. The motivation for the approach is the identification of sites suitable for blocking protein−protein interactions;[14,15] however, the method may be applied to identify novel binding sites on any protein.

A prediction procedure of binding or "druggable" sites on a protein may be divided into two steps. The first step is to detect potential binding sites by accounting for properties of the protein surface. The next step is to select the ideal binding site(s) on the basis of well-defined scoring criteria. A binding site or a binding pocket on a protein usually has the concave shape of a well, cave, bowl, cleft, or valley which can be described by some geometrical quantity such as surface curvature.[16−18] Such pockets are often formed by the arrangement of the peptide backbone and may have a high level of sequence conservation. β-sheets[19] often play an important role in the formation of a binding pocket, while the positioning of side chains in the pocket will effect its shape and affect ligand binding.[20] Physicochemical properties of binding sites are important as they will impact the ability of the site to bind ligands with an acceptable affinity. However, which properties are most important for defining a binding site? The simple answer may be that no single property plays a major role in the ligand−protein interactions. A total of 408 physicochemical, structural, and geometrical attributes of cavities have been used in SCREEN (Surface Cavity REcognition and EvaluatioN)[21] to identify drug-binding sites by applying the random forest machine learning technique. Of these attributes, about 18 size- or shape-related attributes were found to be important. Thus, combinations of properties seem to be necessary for providing better empirical scores to identify binding sites.[19,22] Many of these properties are related to the concept of the protein surface as defined by Connolly's solvent-accessible surface.[23,24] Thus, while the use of physicochemical properties to characterize and define ligand binding sites appears to not have yielded

* Corresponding author e-mail: alex@outerbanks.umaryland.edu.

10.1021/ci700149k CCC: $37.00
Published on Web 09/27/2007

a unique solution, it is clear that a necessary starting point is the simple shape of the protein surface.

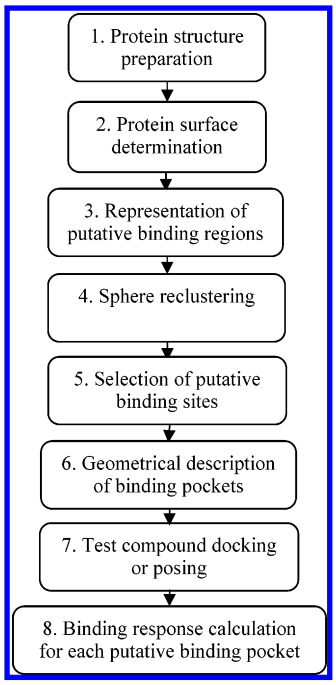
Many studies provide not only algorithms to reveal functional sites and potential binding pockets but also criteria to rank the identified sites. Some recent developments in the identification of binding sites can be found in several excellent reviews.[25−27] These methods may be divided into four categories, that is, geometry-based, energy-based, knowledge-based, and docking-based approaches. These approaches are typically able to detect functional sites including native binding sites for most protein−ligand complexes. However, separate use of each method may not be sufficient to find an ideal binding site suitable for virtual database screening.[27] Machine learning algorithms and related approaches[28] can identify residues or patches of residues related to the binding sites, but no detailed information by which to judge the suitability of the sites for virtual database screening is obtained. The geometry-based approaches,[27] including simple criteria such as pocket size, the number of constituent spheres, or the density of sphere clusters which are used to identify pockets, often provide the shape of the binding site but ignore the receptor−ligand interaction energy, information that is crucial to rank ligands in virtual database screening. The energy-based methods[29,30] and the comprehensive analysis approaches including multifactor empirical scoring function approaches[22] compute the contribution from probes or models, but not that from real ligands.

In the present paper, a novel descriptor for ranking binding sites, termed the binding response, is presented. The approach is based on the score and geometry of receptor−ligand binding poses in the candidate binding sites and, thus, requires the generation of the binding poses in a series of docking steps. For implementing the novel descriptor, a complete procedure for the identification and ranking of binding sites on a protein for use in virtual database screening is presented. The approach builds on previously published methods and includes steps for putative site detection, docking, and site ranking. An example protein−ligand complex is first analyzed to explain the usage of the novel descriptor. Then, the developed method is applied to a set of 29 protein−ligand complexes obtained from the LPDB[31] as a test of the approach to identify native binding sites.

2. METHOD

The novel descriptor defined in the present paper was designed to select binding sites appropriate for virtual database screening. One of the primary criteria in the development of the descriptor is that it must favorably score putative sites into which ligands bind in a well-defined orientation. We name the novel descriptor "binding response" and use it to describe the match between the concerned binding site and a probe compound or set of probe compounds. This site−probe match requires both favorable binding energy and geometrical overlap with the targeted binding site as expressed in the scoring function presented in eq 1 below. Before applying the novel descriptor, the docked or posed coordinates of the ligands in the binding site and the binding energy or score must be obtained. Any screening or docking method can be used to obtain the binding coordinates and the associated score, although the present work is based on the screening program DOCK 4.0[32] along with the energy-based scoring method included with the program.

Scheme 1



2.1. Overview of Binding Site Identification and Scoring. Scheme 1 gives a simple flow diagram of the complete candidate binding-site ranking protocol. The protocol involves four major processes, that is, defining the protein surface (steps 1 and 2), identification of potential binding sites using spheres complementary to the protein surface (steps 3−6), test docking against each of the putative binding sites (step 7), and ranking of these binding sites according to the calculated binding response (step 8). The following is a step-by-step description of the protocol that builds on previously published methods.[8]

1. Protein Structure Preparation. For the present study, protein−ligand complexes were taken from the LPDB,[31] which already included coordinates for the protein hydrogen atoms, and were used directly except for removal of the ligands. However, typical applications require that the 3D structure of the protein of interest be obtained from the Protein Data Bank (PDB, http://www.rcsb.org/).[33] Missing atoms including hydrogen atoms must then be added, followed by local energy minimization, performed with the program CHARMM[34] in the present study. Minimizations involved 200 steps using the steepest descent minimizer with the positions of atoms identified in the crystallographic structure fixed at their experimental values. Calculations were performed using the CHARMM all-atom protein force field[35,36] with default cutoffs for the nonbond interactions.

2. Protein Surface Determination. The protein surface is expressed as the solvent-accessible surface computed using Connolly's algorithm.[23,24] The subroutine DMS implemented in the program MIDAS,[37] which was recently updated to Chimera,[38] was used to generate the solvent-accessible surface. All hydrogen atoms on the protein were removed for generating the protein surface by DMS, which expresses the protein surface by the surface points of the probe rather than the probe center. The probe radius was set to be 1.4 Å,

and the density of surface points was set to be 0.5, as suggested for proteins. The coordinates of each surface point and its normal are recorded for generating complementary spheres in the next step of the protocol.

3. Representation of Putative Binding Regions. Putative binding regions associated with concave regions on the protein surface were defined using the sphere-based method developed in the context of the program DOCK.[39] Spheres of radii, 1.4−4 Å, complementary to the protein surface were generated by the subroutine SPHGEN implemented in the package DOCK. Each sphere touches two protein surface points and lies on the normal of one of the two points. This procedure generates a very large number of spheres, which are trimmed on the basis of the following criteria. The filtration step first selects only the largest sphere associated with each surface atom. The remaining spheres are then analyzed for overlap and grouped via a single linkage algorithm into clusters with those clusters identifying cavities on the protein surface. This filtration is applied as the resulting sphere sets are appropriate for docking;[39] this sphere-generation method has been shown to successfully identify known binding sites in which the native ligand is overlapped by the largest sphere cluster.[32] However, it should be emphasized that the sphere clusters generated by SPHGEN do not properly represent the pockets on the protein surface due to many of the clusters being either too small and only identifying a small portion of a pocket or too large, thereby extending beyond a defined binding pocket. These limitations motivated inclusion of the following sphere reclustering step

4. Sphere Reclustering. The reclustering of spheres was done using the subroutine CLUSTER[40−42] implemented in CHARMM. The clustering algorithm performs a very simple iterative process. Spheres are arbitrarily selected as the center of a cluster, and other spheres within a cutoff distance, for example, 5 Å in the present study, are added to the cluster, with the centroid sphere of the cluster re-evaluated as new spheres within the distance threshold are added. This iterative process leads to additional spheres within the distance threshold being added to a given cluster while some previously included spheres are removed. This process is repeated until all clusters do not change. It should be pointed out that it is impossible to set a distance threshold to correctly express all pockets on the protein surface due to their variable size. Accordingly, a large pocket may be divided into two or more sphere clusters, while a sphere cluster may extend beyond the confines of a small pocket though the reclustering step minimizes these limitations, as presented below.

5. Selection of Putative Binding Sites. All binding sites can be employed in the evaluation of the binding response, as used in the blind docking approach[43] which, based on the AutoDock program,[44] aims at finding ligands able to bind to any site on a protein. However, most known binding sites are located in large pockets on the protein surface,[45] and such large pockets typically contain the largest number of spheres. Therefore, putative binding sites used in the present paper for further analysis are the top 10 clusters selected on the basis of the number of the constituent spheres in each cluster. The sphere clusters defining these 10 sites are then used as the anchor space for the subsequent docking calculations as well as to define the geometric extent of the binding pocket. In all of the ligand−protein complexes studied presently, the known ligand binding sites were included in the 10 selected clusters. It is emphasized that this initial selection process may not always identify the ligand pocket; though, increasing the number of clusters selected for calculation of the binding response to values greater than 10 would diminish this possibility.

6. Geometrical Description of Binding Pockets. Determination of the binding response requires a definition of the geometric extent of each putative binding site. This will be referred to as the *effective space* and, in general terms, is defined as the region into which an inhibitor can bind and have the desired biological outcome, for example, block a protein−protein or protein−substrate interaction. While there may be many approaches by which to define the effective space, for simplicity, in the present work, it will be defined as the anchor space. The anchor space is the volume of space that the spheres defining the putative binding pocket, as discussed in the preceding section, encompass. As discussed below, the anchor space is typically defined by approximately 20 spheres and is able to envelope more than 10 ligand atoms when the radius of each sphere is 1.4 Å and each cluster is generated on the basis of a radius of 5 Å.

7. Test Compound Docking or Posing. To obtain the binding energy required for calculation of the binding response, it is necessary to dock or pose a collection of test compounds. The obtained orientation of each compound is used for determination of both the binding energy and the geometric match to the effective space of each binding site as described below. Docking is performed by using DOCK4.0[32] with the following parameters, as previously used for secondary docking calculations performed in our laboratory.[14,15,46−48]

The docking parameters and options can be clarified by describing the posing process. Each ligand is broken down into segments that overlap at each rotatable bond. The largest segment and any rigid segment containing at least five heavy atoms, typically rings, are then used as anchors in the initial step of ligand posing. The largest anchor is first placed into the pocket by overlapping it with the spheres and is then minimized. The matching proceeds until 500 orientations have been generated, with 50 optimal, diverse configurations retained. The remainder of the ligand is organized into layers according to the connection distance to the anchor and subsequently built layer by layer. In the buildup of each layer, the dihedral angle about the bond being added is rotated in 10° steps to identify the lowest energy conformation, following which, the current layer, up to and including the five innermost layers, and the anchor are reminimized. Once the entire ligand has been built, the entire ligand conformation is reminimized. The simplex minimization of the entire ligand is for up to 200 cycles or a convergence criterion of 0.5 kcal/mol. This process is repeated for all anchor fragments with the ligand orientation with the lowest interaction energy selected. The energy function for both posing and scoring is the DOCK-based sum of the electrostatic energy employing a distance-dependent dielectric constant and 6−12 Lennard-Jones energy, computed between all ligand atoms and protein atoms within 10 Å of the ligand.

The test set of compounds used in the present study was generated to maximize structural diversity and druglike properties. The creation of the 1000 compound database was from our in-house database of over 3 million compounds,[46,49] which have been energy-minimized at the empirical level

using the SYBYL[50] or MOE[51] force fields and assigned atomic partial charges using the CM2 charge model at the semiempirical AM1 level using AMSOL.[52,53] The criteria for selecting the 1000 druglike compounds were based on the rule-of-five proposed by Lipinski et al.,[54] which includes molecular weight ($\leq$500 Daltons), adequate solubility [ClogP(o/w) $\leq$ 5], and the number of hydrogen-bond acceptors ($\leq$10) and donors ($\leq$5), with the final compounds also satisfying the slightly more strict rules as defined by Oprea et al.[55] The diversity of the selected compounds was judged by the Tanimoto similarity index[56,57] based on BIT_MACCS fingerprints,[58] which is implemented in the database software MOE.[51] The database of 1000 test compounds as well as software required to facilitate the performance of the binding response calculation may be obtained from the MacKerell lab Web site.[59]

8. Binding Response Calculation for Each Putative Binding Pocket. The binding response, B_{cp}, of each test compound in each putative binding site was computed on the basis of the definition and parametrization of the binding response as described below. The average binding response of each site, B_p, is then determined over a subset of the test molecules in that site. It is the average binding response for each putative binding site that is used to evaluate the suitability of that site for virtual database searching.

It should be noted that the present method is relatively computationally intensive. The docking of 1000 compounds on each binding site takes about 6 h of CPU time on a 2 GHz, 64 bit AMD Athlon with 1 GB of memory. Thus, approximately 60 CPU hours are required when 10 putative sites on a protein are considered.

2.2. Binding Response. Both the binding energy and the geometrical measure of a binding pose are used in the present work to judge if a pocket is, or is not, suitable for binding a small molecule. The binding energy alone only defines the strength of the interaction between the receptor and the small molecule, but no information about the location of the bound molecule is included. Therefore, ligands that score well may lie outside of the target binding site. To overcome this limitation, a geometrical factor is included in the binding response. This factor is designed to determine if a small molecule is located in the putative binding site as well as to quantify the extent to which the molecule is situated in the pocket.

To facilitate an understanding of the binding response, a description of how the mathematical formula was developed follows. The basic factors used in the mathematical formula are the "sufficient binding energy" and "adequate portion" of the small molecule inside the pocket. The former can be determined on the basis of any energetic-based docking procedure. The later is based on the consideration that a portion of a ligand be located in the putative binding pocket. The model was developed to have a scoring range of 0 to 1, where a sufficent match receives a value of 1 while a poor match has a value of 0. The mathematical formula containing the two factors was then designed to quantitatively provide a continuous response between 0 and 1. The boundary conditions for the score of 1 are a sufficient binding energy and an adequate portion of the small molecule inside the binding pocket. Any binding pose with a binding energy more favorable than the sufficient binding energy still receives the value 1. A ligand with more than an adequate portion of the molecule inside the pocket also only receives the value 1. The value 0 is assigned to binding poses with zero or unfavorable binding energies or without any portion of the molecule inside the pocket. Parameters for the model were initially selected to control the rate of change of the binding response between 0 and 1 as presented below.

The binding response of pocket p to probe compound c is defined as

$$B_{cp} = B_1(n,e) - B_2(D) \tag{1}$$

where the first term B_1 contributes favorably to the binding response. It includes contributions from n, the number of ligand atoms located in the effective binding space, and the binding energy, e. The binding energy e between the compound and the protein in the present study is the DOCK interaction energy, but it can be replaced by any scoring function value used in the docking algorithm. The second term B_2 decreases the binding response if the ligand departs a distance D away from the pocket. For convenience, the values of the binding response, B_{cp}, and the two components B_1 and B_2 are designed to be in the range [0, 1] under appropriate boundary conditions. An ideal binding response has a value of 1, while a poor binding response has a value close to 0.

B_1 is given the following form

$$B_1 = \begin{cases} S\sum_{i=1}^{n}\left(\frac{1}{2}\right)^i, \text{ if } n > 0 \\ 0, \text{ if } n = 0 \end{cases} \tag{2}$$

where the summation runs over the ligand atoms in the effective space of the binding site. Its value will approach 1, starting with the values 0.50, 0.75, 0.88, 0.94, 0.97, 0.98, 0.99, and so forth for n = 1, 2, 3, 4, 5, 6, 7, and so forth, respectively. The calculation of n depends on the specification of the effective space, which is expressed as the anchor space, as described above. The determination of n is based on the number of ligand atoms within 1.4 Å of any sphere in the set used to identify the binding site.

As is obvious, the above summation is highly damped for values near 1. This summation is relatively insensitive to the value of n once n is larger than 4. Therefore, to introduce additional scaling into the binding response, a factor S is introduced. S is defined with respect to a reference interaction energy E and a reference number N.

$$S = \begin{cases} 0, \text{ if } n = 0, \text{ or } e \geq 0 \\ 1, \text{ if } n \geq N, \text{ and } |e| \geq |E| \\ \dfrac{\ln tn}{\ln tN}\dfrac{\ln u\,|e|}{\ln u\,|E|}, \text{ for others} \end{cases} \tag{3}$$

where N is a parameter defining a sufficient number of ligand atoms located in the effective space, for example, 12 in the present analysis. E is a reference value for judging favorable binding energies or scoring function values, for example, $-$30 kcal/mol in the present study for the total interaction energy used in DOCK4.0. t and u are tuning factors larger than 1 (e.g., 32 and 8) used to maintain the ratio in a targeted range, typically close to 1. The value of each logarithm will be constrained to the range [0, 1].

B_2 in eq 1 is also restricted to the range [0, 1] and has the following form:

$$B_2 = \left(\frac{D}{R}\right)^{\upsilon} \quad (4)$$

where D is a measure of the departure of the compound from the effective space and R expresses the geometrical size of the compound. There are several ways to define D and R. In the present work, R is defined as the largest distance from the center of mass of the ligand to the furthest of its constituent atoms. D is defined as the difference $(d - r)$ where d is the distance from the geometrical center of the effective space to the center of mass of the ligand and r is the radius of the effective space. D can be corrected to zero when the largest distance l between ligand atoms located in the effective space is larger than r. These geometrical quantities are illustrated in Figure 1.

The total binding response B_p for a potential binding site is given as the following average:

$$B_p = \frac{1}{M}\sum_{c=1}^{M} B_{cp} \quad (5)$$

where the sum runs over M compounds. M can be the total number of compounds in the test set or a subset of the test set selected as those compounds that have the best B_{cp} values for the targeted binding site. For example, the B_p value may be based on the average over the 500 compounds with the best B_{cp} values. Motivation for this definition is based on the reality that individual binding sites have preferences for ligands of certain structural and physiochemical properties. Thus, calculating B_p on the basis of those ligands that interact favorably with the target site may be more representative than including ligands that are intrinsically not suited for the binding pocket. A test of this approach is performed below. Finally, a ranking of different binding sites in a given protein is based on the B_p values using the same number, M, of compounds. Preferred binding sites will have B_p values approaching 1, while poorer binding sites will have smaller B_p values. The utility of this analysis is presented below via calculations on a collection of protein−ligand complexes for which the 3D coordinates of the complexes have been determined via crystallographic studies.

2.3. Determination of Parameters. In this section, the parameters will be selected via analysis of the mathematical behavior and physical meaning of these factors on the binding response. Alternatively, the parameters could be adjusted to optimize the ability of the binding response to reproduce a set of target data. However, given the success of the approach on the selected test set of complexes, such an approach was not utilized.

In eqs 1−4, the quantities (*neDR*) are directly computed from the docking result. Other quantities (*tuυEN*) are the parameters that need to be determined. On the basis of docking studies performed in our laboratory using DOCK4.0,[32] we have determined that a value of −30 kcal/mol for the reference scoring function, E, represents a significantly favorable interaction energy relative to the majority of compounds. For example, studies targeting the Y3 binding site on ASV and HIV integrases showed the most favorable DOCK score to be −25 kcal/mol when using the nonproprietary, open portion of the NCI 3-D database[61] and approximately −40 kcal/mol when using the WDI or CHEMDIV compound databases.[49] For most binding sites,

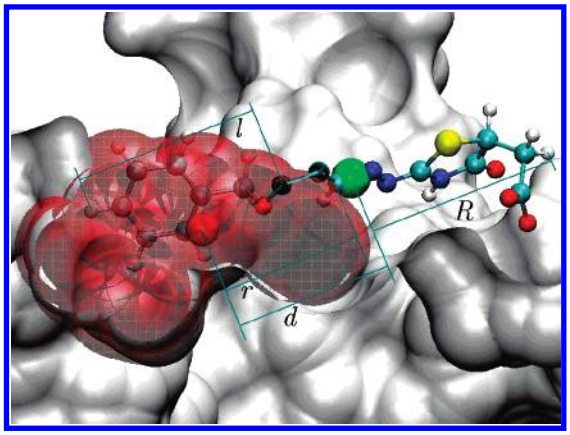


Figure 1. Definition of geometrical quantities used in eq 4. Shown is an arbitrary docked ligand in CPK representation (hydrogen, white; carbon, aqua; nitrogen, blue; oxygen, red; sulfur, yellow). The red transparent spheres represent the effective space; the solid red sphere is the geometrical center of the effective space, and the solid green sphere is the center of mass of the ligand. All figures were generated using VMD.[60]

a good geometrical match between a ligand and the binding site is achieved if there are more than 10 ligand atoms located inside the effective space of the site. Accordingly, it is reasonable to set the reference number N to be 12 or a similar number. Therefore, on the basis of these values for E and N, a binding site will be regarded as ideal for a specific ligand if the total interaction energy is lower than −30 kcal/mol and there are more than 12 ligand atoms located in the effective space. For other binding cases, that is, the interaction energy $e > -30$ kcal/mol and the number of ligand atoms in the effective space $n < 12$, the binding response values will be less than 1 and can be computed on the basis of the three parameters (*tuυ*), as discussed below.

The impact of the parameters (*tu*) on the two logarithms in eq 3 are shown in Figure 2A with $N = 12$ and in Figure 2B with $|E| = 30$, respectively. A large tuning factor (e.g., $t = 128$ or $u = 128$) will keep the logarithm values close to 1, while smaller values (e.g., $t = 2$ or $u = 2$) will spread the logarithm values over a larger range with the small value end closer to 0. When the relative impact of (tN) and ($u|E|$) is compared, the curve of $t = 32$ in Figure 2A has a similar range as that of the curve of $u = 8$ in Figure 2B. The latter curve (i.e., $u = 8$) spreads a little bit lower because ($u|E|$) = (8 × 30) is smaller than (tN) = (32 × 12). A smaller tuning factor will lead to a dominant effect by the B_1 value. This combination, that is, (*tuEN*) = (32, 8, −30, 12) means that the energy factor ($u|E|$) plays a slightly more important role than the geometrical factor (tN).

The second term, B_2, in eq 1 is introduced to negatively impact the scores of those ligands whose binding mode only overlaps a small portion of the effective space of the binding site. This typically occurs when only one end of a molecule overlaps with the effective space and is incorporated in the formalism to distinguish this type of binding from binding modes in which the central portion of a ligand is placed in the pocket. However, this negative contribution from B_2 is based on the departure of the ligand from the effective space, which is partly accounted for in the B_1 term by counting the ligand atoms in the effective space. Therefore, the B_2 term should be relatively small, which requires the parameter υ in eq 4 to be large. The impact of υ on B_2 is shown in Figure

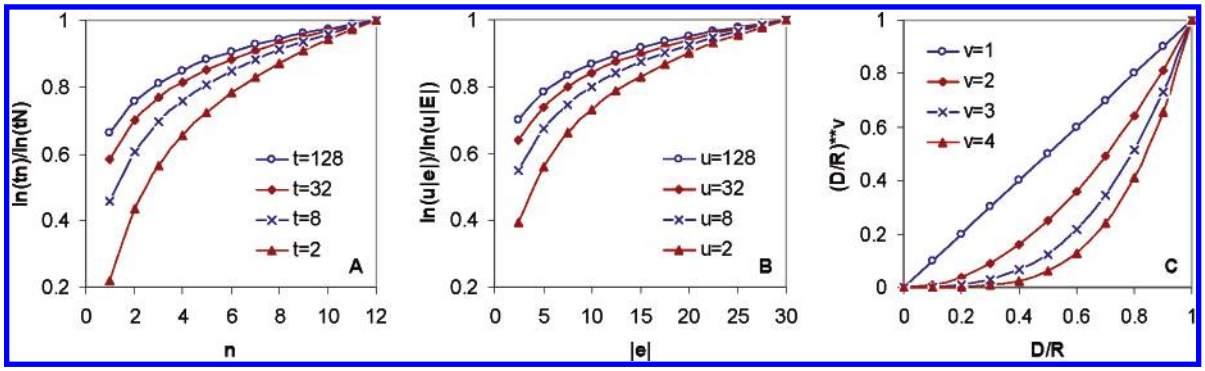


Figure 2. Impact of selected parameters in (A) $\ln(tn)/\ln(tN)$ in eq 3 as a function of atoms, n, in the effective space where $N = 12$, (B) $\ln(u|e|)/\ln(u|E|)$ in eq 3 where $|E| = 30$ as a function of e, the ligand−protein interaction energy, and (C) eq 4 where the ratio D/R is varied from 0 to 1.

2C. Those results indicate that $v = 3$ should be large enough to keep the B_2 values in a range that does not lead to this term dominating the overall binding response.

On the basis of the above analysis, the default values for the parameters are set to be $(tuvEN) = (32, 8, 3, -30, 12)$. The use of very large tuning factors (tu) is not recommended because the contribution of the scaling factor S in eq 3 will be reduced and even vanish, although the parameters can be determined over a wide range, for example, $t > 1$, $u > 1$, $v > 1$, $E < 0$, and $N > 1$ according to the above equations. In addition, it is suggested that the contributions of the energetic and geometrical factors to the binding response should be similar, that is, the ratio $(tN)/(u|E|)$ is around 1. In the default values $(tuvEN) = (32, 8, 3, -30, 12)$, the importance of the energy term is slightly higher than the geometric term due to the use of a smaller tuning factor $u = 8$ as compared to the geometrical tuning factor $t = 32$.

3. RESULTS

Presented is a method for the automated identification and selection of binding sites on protein surfaces suitable for virtual database screening. The overall approach will be explained by an example of a model system, the complex of the ligand AGB with the urokinase plasminogen activator b-chain.[62] It will then be applied to 29 ligand−protein complexes of known 3D structures to show the capability of the method to identify known ligand-binding sites.

3.1. Example of the Binding Response on Urokinase Plasminogen Activator b-Chain. To clarify the implementation and utility of the binding response evaluation, it will be applied to a known ligand−protein complex: the X-ray crystallographic structure of the urokinase plasminogen activator b-chain inhibitor complex, which contains the ligand AGB (N-(1-adamantyl)-N′-(4-guanidinobenzyl)urea, $C_{19}H_{27}N_5O$; PDB identifier 1ejn, 1.8 Å resolution).[62] When this system was used, tests were performed on the impact of the cutoff radius during reclustering (step 4) on sphere set size. This was followed by analysis of the impact of the bound orientation of different docked ligands on the value of the binding response, B_{cp}, including a series of tests of the impact of the parameters in eqs 1−4 on B_{cp}. Finally, an analysis of the use of the average binding response, B_p, for 10 binding sites on the urokinase plasminogen activator b-chain was performed.

Of the eight steps involved in the selection of binding sites for use in virtual database screening (Scheme 1), the first step that represents a departure from previously published approaches is step 4, reclustering of the sphere generated by the program SPHGEN. As stated above, this was performed using the clustering routine included in the program CHARMM, with the only parameter in the clustering algorithm being the cluster radius. Shown in Figure 3 are two clusters that are located in the experimentally determined binding site; the ligand AGB is included to indicate the extent of the binding site. For the 4 and 5 Å cutoffs, the extent of the spheres is similar, and it may be seen that both sets have significant amounts of overlap with the ligand. Upon increasing to larger cutoffs, the number of spheres increases significantly, with multiple spheres being located far from the bound ligand. As the goal of the present work is to define binding sites suitable for docking with the assumption that the screening procedure should be focused on a well-defined site, it is evident that more-focused sphere clusters, as with the 4 and 5 Å cutoffs, are desirable. Since the 5 Å cutoff was similar to the 4 Å value and to avoid limiting the cluster sizes, a value of 5 Å was selected for further studies. Similar analysis on other proteins yielded the same conclusion (not shown).

When steps 1−4, described in Scheme 1, were applied to the urokinase b-chain and a 5 Å clustering cutoff was used, a total of 633 spheres were obtained after the SPHGEN filtration process and grouped into 56 clusters. This procedure leads to the experimentally identified binding region being occupied by two sphere clusters. These two clusters, with 21 and 20 spheres, respectively, were the largest clusters identified. The effective space occupied by the first sphere cluster contains 17 ligand atoms, while the second cluster only contains six ligand atoms. Further analysis of the binding response for different ligands will be performed using the first sphere cluster.

Figure 4 shows the binding modes of selected docked ligands, labeled **A**−**I**. The quantities associated with eqs 1−4 were computed for each ligand and are listed in Table 1 with fixed parameters $(EN) = (-30, 12)$. In addition, the binding response values as a function of the parameters (tuv) are also reported. Compound A is the native ligand AGB. As shown in Figure 4A, its docking pose (ball and stick representation) is similar to that of the crystal structure (stick representation). The resulting B_{cp} value for the docked orientation of the native compound was 0.97, close to the ideal value of 1. The binding site is also suitable for other compounds, including **B**, **C**, and **D** whose binding modes

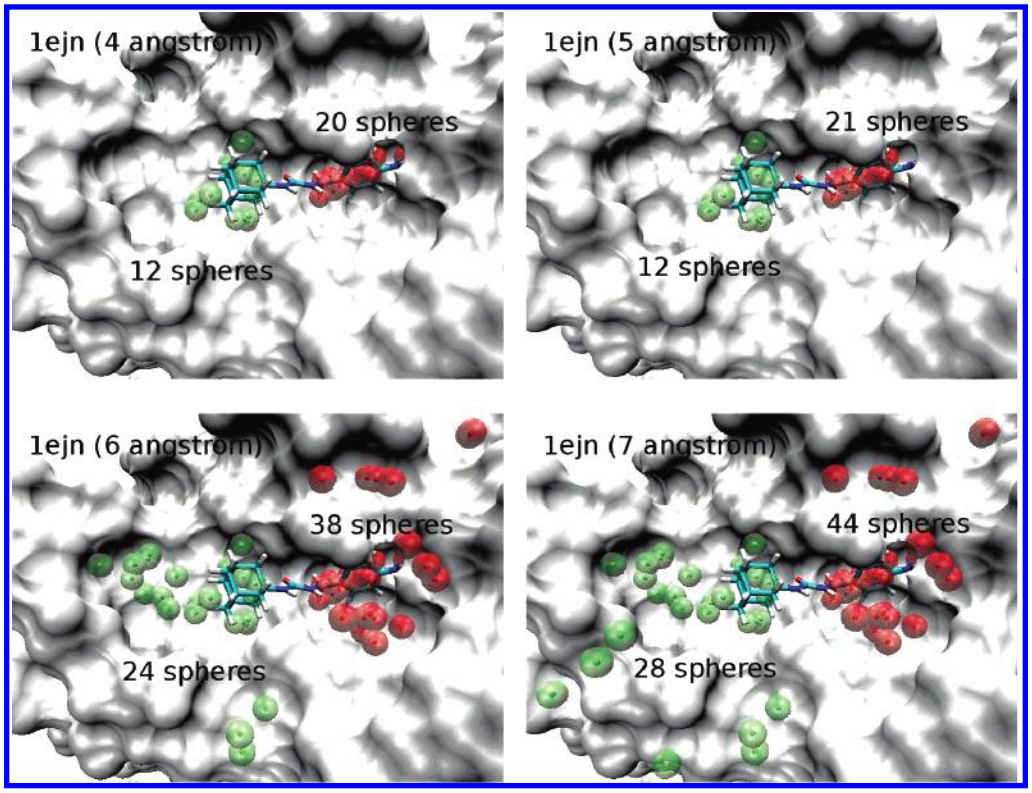


Figure 3. Sphere sets resulting from clustering with cutoffs ranging from 4 to 7 Å using two separate clusters as examples (green and red sphere sets).

are similar to that of the experimental ligand, as shown in Figure 4B, C, and D. As listed in Table 1, they have favorable interaction energies, −45.3, −37.7, and −38.3 kcal/mol, respectively, with many atoms located in the effective space. These characteristics yield high binding response values of 1.00, 0.98, and 0.97, respectively, with the parameters (tuv) set to (32, 8, 3). In contrast, **E**, **F**, and **G** are not predicted to bind well. As may be seen in Figure 4, only one end of each compound overlaps with the effective space, yielding significantly lower values of B_{cp}. Finally, the docked orientations of compounds **H** and **I** are entirely out of the effective space, leading to B_{cp} values of 0. Notably, both **F** and **H** have very favorable interaction energies, e; however, the diminished or lack of overlap with the effective space leads to B_{cp} values significantly less than 1. A detailed comparison will be given below.

A more detailed analysis may be performed with the default parameter set ($tuvEN$) = (32, 8, 3, −30, 12). Compounds **C** with **E** both have 11 atoms in the effective space, but the B_{cp} for **E** is smaller, 0.77, due to its less favorable interaction energy of −8.83 kcal/mol. This situation also occurs between **F** and **G**. Concerning the number of ligand atoms in the effective space, **D** and **F** both have similar, favorable interaction energies, though B_{cp} for **F** is smaller due to the decreased number of atoms in the effective space. Additional analysis shows the role of the departure quantity, D. Comparing **C** and **D**, which have similar n, e, and R values, shows the B_{cp} value for **D** to be slightly smaller than that for **C** because the center of mass of **D** departs 2.05 Å from the geometric center of the sphere cluster.

Additional analysis was performed by varying the parameters (tuv) with (EN) = (−30, 12). In the last four columns of Table 1, several variations of parameters (tuv) are given. The change of parameters does not significantly alter B_{cp} for both the well-bound compounds, **A**, **B**, **C**, and **D**, and poorly bound compounds, **H** and **I**. However, the parameter change does significantly affect the intermediate-bound compounds, **E**, **F**, and **G**. As analyzed above, larger tuning factors t or u will diminish the sensitivity of B_{cp}, maintaining larger values as shown for (tuv) = (128, 128, 3). Values of t and u larger that 128 do not significantly change the B_{cp} values. Smaller t or u values decrease the B_{cp} values as shown for (tuv) = (2, 2, 2). The smaller t parameter enhances the impact of the number of ligand atoms in the effective space. This leads to **E** having a larger B_{cp} than **F** and **G** as shown for (tuv) = (2, 128, 3). Alternatively, a smaller u enhances the contribution of the interaction energy e, such that **F** has a larger B_{cp} value than **E** for (tuv) = (128, 2, 3). Thus, the tuning parameters, t and u, may be systematically altered to control the sensitivity of the binding response to the interaction energy and number of ligand atoms in the effective space. For the remainder of the present work, the parameters ($tuvEN$) = (32, 8, 3, −30, 12) will be used as default parameters, although other choices can be used on the basis of specific comparisons.

Figure 5 shows the 10 largest sphere clusters generated by the reclustering with a 5 Å cutoff. Each sphere cluster is labeled by a number which is their rank according to the B_p determined from eq 5 and reported in Table 2. Cluster 1 represents the central part of the experimentally determined binding site; the ligand AGB is shown in the figure. Two additional clusters also occupied portions of the binding site. Cluster 10 is buried inside protein, and the second cluster contains only 12 spheres and thus was not a member of the 10 largest clusters. It may be argued that a single sphere cluster should fill the native ligand space. While it may be considered desirable for a single cluster to encompass each binding site, known binding sites are often part of large

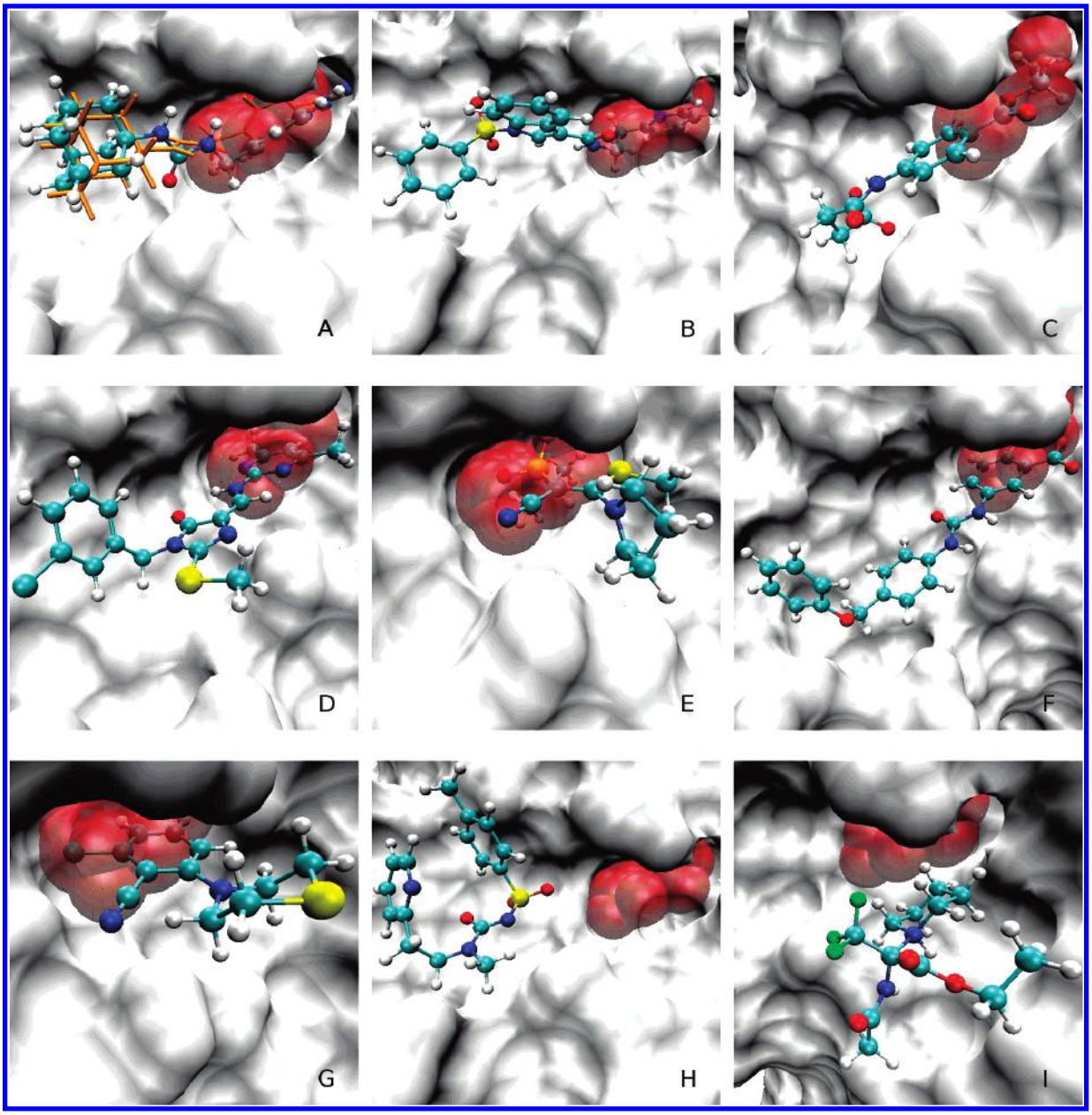


Figure 4. Selected docked ligands targeting the binding pocket indicated by the red spheres on the surface of urokinase plasminogen activator b-chain (PDB identifier: 1ejn). The binding response values are given in Table 1, and the experimental conformation of the inhibitor, N-(1-adamantyl)-N′-(4-guanidinobenzyl)urea, is shown in stick representation in panel A.

Table 1. Quantities in eqs 1−4 for the Compounds Docked against a Pocket in the Native Binding Site of 1ejn, as Shown in Figure 4[a]

					tuv = 32,8,3					128,128,3	128,2,3	2,128,3	2,2,2
#	*e*	*n*	*D*	*R*	Lne	Lnn	B_1	$-B_2$	B_{cp}	B_{cp}	B_{cp}	B_{cp}	B_{cp}
A	−45.99	10	0.00	9.38	1.00	0.97	0.97	0.00	0.97	0.97	0.97	0.94	0.94
B	−45.25	14	0.50	9.47	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00
C	−37.73	11	0.00	7.53	1.00	0.99	0.98	0.00	0.98	0.99	0.99	0.97	0.97
D	−38.27	11	2.05	8.52	1.00	0.99	0.98	−0.01	0.97	0.97	0.97	0.96	0.91
E	−8.83	11	0.00	5.29	0.78	0.99	0.77	0.00	0.77	0.84	0.69	0.83	0.68
F	−40.05	6	2.85	8.87	1.00	0.88	0.87	−0.03	0.84	0.86	0.86	0.74	0.67
G	−16.95	6	0.23	4.49	0.90	0.88	0.78	0.00	0.78	0.83	0.77	0.72	0.66
H	−42.53	0	5.03	5.46	1.00	0.00	0.00	−0.78	0.00	0.00	0.00	0.00	0.00
I	−16.95	0	4.84	5.23	0.90	0.00	0.00	−0.80	0.00	0.00	0.00	0.00	0.00

[a] The fixed parameters are $E = -30$ kcal/mol and $N = 12$. The varied parameters are (*tuv*). Lne and Lnn are the logarithm ratios of the energy *e* and the number *n* in eq 3, respectively.

cavities, and in many cases, the region occupied by the ligand is a relatively small portion of the cavity. This occurs with the binding site for AGB, as well as other binding sites (e.g., protein 1c83 as shown in Figure 6). Thus, it is not anticipated that a single cluster will map out the entire binding site in all cases. However, in the context of virtual screening, what

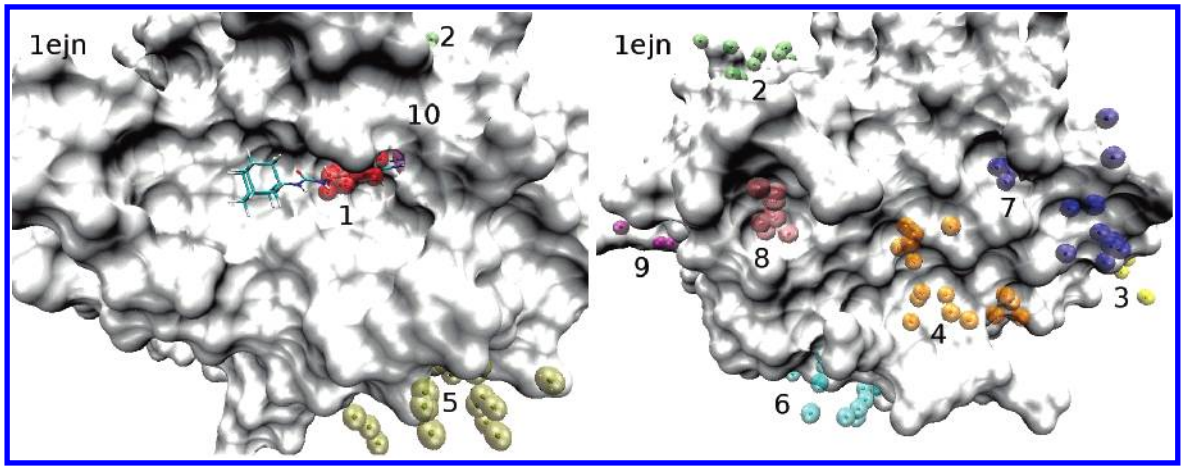


Figure 5. The 10 largest sphere clusters on the surface of urokinase plasminogen activator b-chain (PDB identifier: 1ejn). The numbers indicate their rank according to the B_p values (eq 5), as listed in Table 2. Cluster 10 overlaps a few ligand atoms and is buried inside the protein.

Table 2. The Binding Response B_p of the 10 Largest Sphere Clusters, Each Representing a Potential Binding Site on Protein 1ejn[a]

				B_p	
cluster (residue)	sphere	ligand atom	e	1000	900
1 (V213)	21	17	−30.55	0.903	0.95
2 (T188)	17	0	−21.72	0.902	0.91
3 (E84)	17	0	−21.97	0.899	0.91
4 (S48)	17	0	−23.11	0.860	0.88
5 (D97)	18	0	−20.84	0.805	0.84
6 (P236)	19	0	−21.44	0.788	0.82
7 (G113)	18	0	−20.76	0.761	0.81
8 (T208)	19	0	−22.13	0.760	0.78
9 (I163)	20	0	−20.28	0.668	0.71
10 (C42)	20	6	−12.09	0.032	0.03

[a] The approximate location of each cluster is indicated by the residue shown in parentheses. The number of its constituent spheres, the number of native ligand atoms that overlap with that sphere set, and the averaged total interaction energy e (kcal/mol) are included.

is important is that the B_p value used to rank the putative binding sites identifies appropriate sites for the binding of ligands. As is shown in Table 2 as well as for 29 additional proteins (see below), the approach does indeed identify known ligand binding pockets.

Ranking of the 10 largest clusters is shown in Table 2. Cluster 1 has the largest B_p value, indicating it to be the best binding pocket, consistent with its identification in the experimental structure. Clusters 2 and 3 have B_p values of 0.902 and 0.899, respectively, when B_p is based on all 1000 compounds. Thus, even though clusters 2 and 3 are not the experimentally observed binding sites, they are indicted to be appropriate binding sites for use in virtual screening. Notably, clusters 2 and 3 have significantly less favorable average interaction energies than cluster 1, while their B_p values are similar. This small differential in B_p is due to a number of the 1000 docked test ligands not having a significant overlap with the effective space of cluster 1, thereby reducing the first term B_1 of eq 3. Such a result is not necessarily unexpected given the diverse chemical structure of the test ligands, leading to many of those ligands not interacting well with the region of the site in the effective space of cluster 1. To account for this behavior, the B_p values of the top 10 clusters were calculated by averaging over the top 900 compounds with the best B_{cp} values for the individual sites. This leads to a larger differential between the B_p value of cluster 1 and those of clusters 2 and 3. This result suggests that using a slightly smaller subset of the docked test compounds may lead to better discrimination of the binding sites; additional analysis of this behavior is presented below.

3.2. Application of Binding Response to 29 Known Ligand−Protein Complexes. In this section, the binding response descriptor is applied to 29 experimentally determined ligand−protein complexes. These were obtained from the database LPDB,[31] which contains ligand−protein complexes collected from the PDB, corrected for the proper protonation states of ionizable groups[63] and other structural features and put into file formats suitable for CHARMM.[34] The criteria for selecting the 29 complexes were (1) proteins which contain not more than two polypeptide chains and (2) ligands located on the protein surface versus being buried inside the protein. For all 29 systems, the eight steps outlined in Scheme 1 were applied without modification to test the ability of the approach to identify known binding sites. Additional analysis was then performed on select cases to further understand the behavior of the binding response.

Table 3 presents the predicted rank of the experimentally determined binding sites according to the B_p values. When calculating B_p using all 1000 test compounds, the experimentally determined sites are ranked first with a 76% success rate. Notably, while for many of the ligand−protein complexes, including 1afl, 1c2t, 1c83, 1c84, 1c87, 1c88, and 1ctr, the known binding site is not ranked first, they are all among the top 10, with the worst ranking being six. As discussed above, calculation of B_p based on all 1000 test ligands may be misleading as ligands that are intrinsically unsuitable for a binding site are included in the average value. Accordingly, B_p values were calculated for all 29 complexes, selecting the compounds with the top 900, 800, 700, 600, and 500 B_{cp} values for all 10 sites on each respective protein. As is evident, using smaller numbers of test ligands leads to increases in the success rate as based on the highest ranking site being the experimentally determined binding site. At each decrease in the number of compounds used in the calculation of B_p, there is improvement in the predictability of the model. With 500 compounds used to score the sites, the experimentally identified site is selected 90% of the time, with that site not being the best in only three cases, where it is

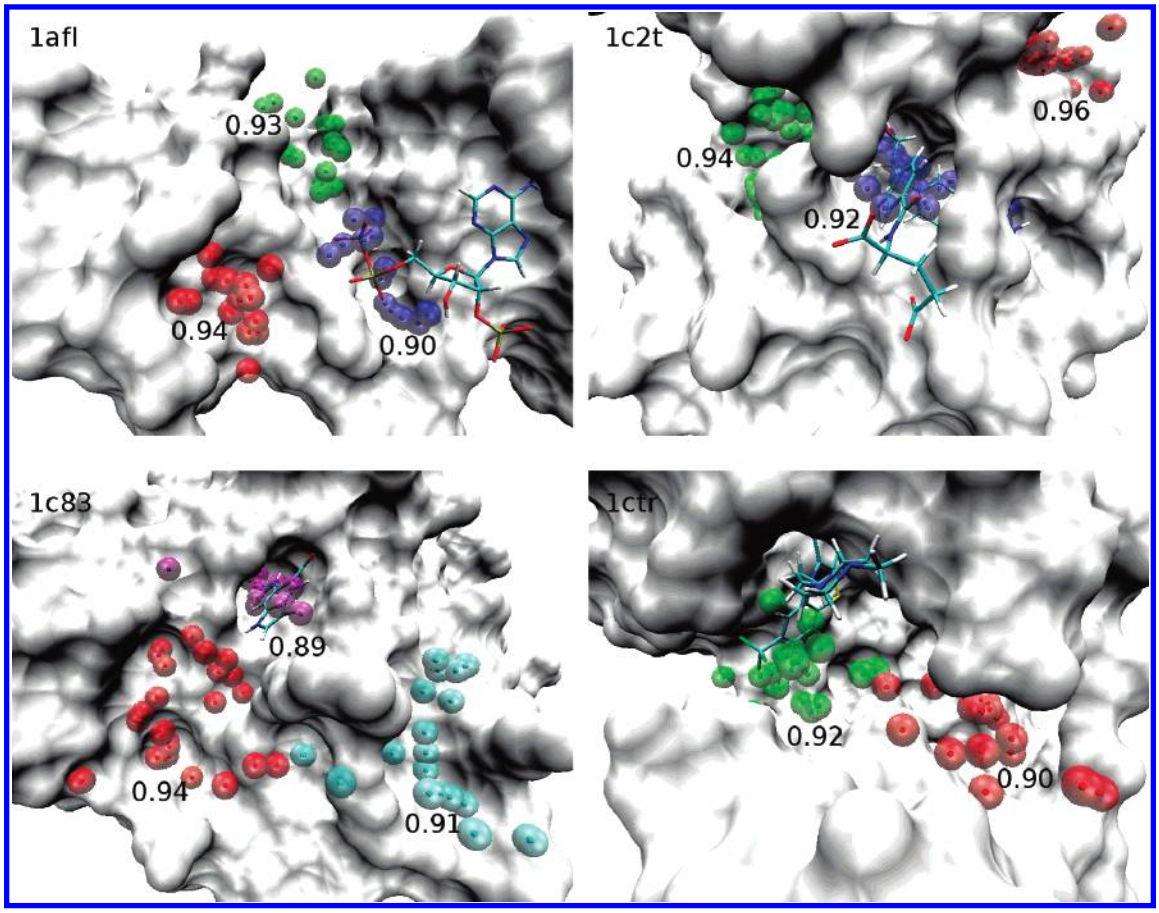


Figure 6. Four experimentally determined binding sites which were not ranked first on the basis of their B_p values. The PDB identifier is marked at the upper-left corner of each panel, and the B_p values for each binding site are marked adjacent to the respective sphere cluster. In each panel, the experimentally determined structure of the respective bound ligand is shown in stick representation. The number of compounds used in the calculation of B_p is 1000 for the four proteins.

still among the top three sites. Thus, the proposed binding site identification scheme, including the use of binding response to score the sites successfully, selects the experimentally selected site in 90% of the test cases and ranks that site among the top three in the remaining three cases.

Detailed analysis of the impact of the number of test compounds used in the calculation of B_p was performed for protein 1c2t, for which a switch in the ranking of the sites occurred (Table 3). Presented in Figure 7 are the B_p values as a function of the number of compounds included in the average for the four top-scoring binding sites. As is evident, the exclusion of compounds that do not interact favorably with the respective sites, as judged by the B_{cp} values of the individual test compounds, leads to higher B_p values in all cases. Such a result is obvious, as lower-B_{cp} compounds are being removed. However, the change in the ranking of the individual sites is important, with the change upon going from 1000 to 900 compounds in the case of site 1 being significant. In contrast, the impact is much smaller on site 2 upon going from 1000 to 900 and 800 to 700 compounds and for site 3 upon going from 900 to 800 compounds. With site 4, an almost linear increase in the B_p values occurs as the number of compounds used in the determination of its value decreases. While the importance of decreasing the number of compounds in the ranking of binding sites is evident, the differential changes in B_p for different sites suggests that this change may be indicative of the selectivity of the individual sites. For example, if decreasing the number of compounds used in the determination of B_p does not significantly impact B_p, it may be surmised that that site does not have a high degree of selectivity for ligands. Alternatively, large changes in B_p as the number of compounds in the average decreases suggest more discriminating behavior by a binding site, which may be indicative of a more selective (or specific) binding site. Further studies will be required to develop this hypothesis.

Finally, it is useful to analyze some of the proteins for which the experimental binding site did not rank first when 1000 test compounds were used in the determination of B_p. Figure 6 shows four such examples. On protein 1afl, the ligand is located on a ridge with only a small portion of the ligand in a visually well-defined pocket. However, near the experimental ligand binding site are two sphere clusters lying in well-defined pockets suitable for docking, thereby having high B_p values. With 1c2t, the experimental ligand binding site lies in a large cavity which is occupied by a second sphere cluster (green cluster in Figure 6). The ligand protrudes from this cavity through a narrow mouth; this orientation is suggested to limit the types of ligands that can bind effectively to the blue cluster. However, when 100 compounds with poor B_{cp} values are eliminated from the calculation of B_p, the blue pocket becomes the top-scoring pocket (Table 3). Protein 1c83 is representative of structures 1c84, 1c87, and 1c88, all of which had the experimental pocket not ranked first with 1000 test compounds used to calculate B_p. In this case, the experimental ligand binding

Table 3. The B_p-Based Ranking of the Sphere Clusters Representing the Native Binding Site on Each Protein Indicated by Its PDB ID

	subset					
PDB ID	1000	900	800	700	600	500
1afk	1	1	1	1	1	1
1afl	3	3	3	3	3	3
1aoe	1	1	1	1	1	1
1atl	1	1	1	1	1	1
1bxo	1	1	1	1	1	1
1c2t	3	1	1	1	1	1
1c83	6	6	4	4	3	2
1c84	3	3	3	1	1	1
1c87	2	1	1	1	1	1
1c88	5	5	2	2	3	2
1c8k	1	1	1	1	1	1
1cbs	1	1	1	1	1	1
1ctr	2	2	2	2	1	1
1ejn	1	1	1	1	1	1
1fkg	1	1	1	1	1	1
1hew	1	1	1	1	1	1
1hrn	1	1	1	1	1	1
1ppl	1	1	1	1	1	1
1ppm	1	1	1	1	1	1
2er6	1	1	1	1	1	1
2wea	1	1	1	1	1	1
2web	1	1	1	1	1	1
2wec	1	1	1	1	1	1
3er3	1	1	1	1	1	1
3er5	1	1	1	1	1	1
4er1	1	1	1	1	1	1
4er2	1	1	1	1	1	1
4er4	1	1	1	1	1	1
5er2	1	1	1	1	1	1
prediction rate (%)	76	83	83	86	90	90

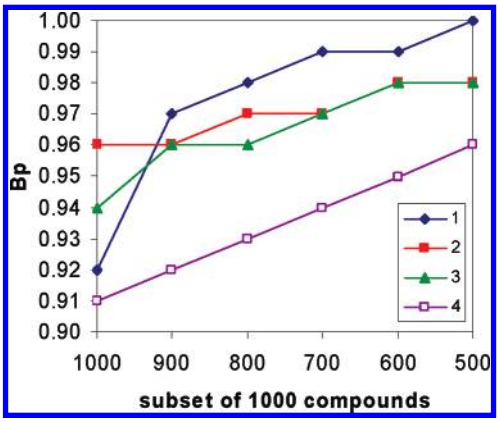


Figure 7. The variation of the B_p values of four top-binding sites on protein 1c2t when a series of subsets of a total of 1000 compounds are used in the computation of B_p values in eq 5.

pocket is a deep, narrow well that appears to limit the range of ligands that can fit into the site. This leads to a significant number of test compounds having poor B_{cp} values such that when these compounds are removed it leads to improvement in the B_p values for those sites. Finally, with protein 1ctr, the ligand is largely located on the surface of the protein, with only a portion of the ligand located in the region occupied by the green sphere cluster. While ultimately this site is ranked as the top site when 500 compounds are used in the B_p calculation, the highest ranked sphere cluster is located on the opposite side of the protein and is not shown in the figure. While the images in Figure 6 allow for a better understanding of why the experimental binding site is not always ranked as the top site, it should be emphasized that the experimental site is always among the top sites, speaking to the utility of binding response in the identification of binding sites suitable for virtual screening. It should also be emphasized that the other sites in the 29 studied proteins that have high B_p values are likely appropriate sites for ligands to bind and, therefore, suitable for database screening.

4. SUMMARY

Virtual database screening is often performed on more than 1 million compounds. While in certain cases, such as well-defined substrate binding pockets, the target binding site is obvious, in many cases, an appropriate site is not known. This is particularly true in efforts to identify small-molecular-weight inhibitors of protein−protein interactions. This requirement motivated the presented approach to systematically identify and rank putative binding sites on a protein for use in virtual screening. Central to the method is the developed binding response used to rank ligands in a site. This approach overcomes limitations in the ranking of sites based on geometric or energetic criteria alone by simply combining these aspects into one term. As the primary goal of the method is for the identification of sites for virtual screening, the use of screening itself in the procedure is considered to be central to the success of the approach. The present method can be categorized as a docking-based method, similar to the blind docking approach,[43] although the latter only uses an energy score during site selection. The advantage of the present method is the combination of energetic and geometrical factors, which provides information on both binding strength and binding location. In addition, the binding site identified by the present method can be directly used in virtual database screening, offering a significant advantage over other binding site identification methods reported in the literature. While a well-defined protocol is presented, it should be reiterated that the present approach can be implemented using a wide range of software packages. To facilitate the use of the method, a collection of scripts to perform the required calculations in the context of the programs DOCK and CHARMM is available to the academic community via the MacKerell lab Web site.[59]

ACKNOWLEDGMENT

Financial support from NIH grants CA107331 and CA120215 and the University of Maryland Computer-Aided Drug Design Center are acknowledged, and appreciation is expressed to Ganesh Kamath, Deva Priyakumar, and Chayan Acharya for helpful discussions. A collection of programs, scripts, and documentation to perform site identification and scoring is available on the MacKerell lab Web site (http://www.pharmacy.umaryland.edu/faculty/amackere/).

REFERENCES AND NOTES

(1) Jones, S.; Thornton, J. M. Principles of protein−protein interactions. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 13−20.

(2) Janin, J.; Wodak, S. J. Protein modules and protein−protein interaction. Introduction. *Adv. Protein Chem.* **2002**, *61*, 1−8.

(3) Szilagyi, A.; Grimm, V.; Arakaki, A. K.; Skolnick, J. Prediction of physical protein−protein interactions. *Phys. Biol.* **2005**, *2*, S1−S16.

(4) Fernandez-Ballester, G.; Serrano, L. Prediction of protein−protein interaction based on structure. *Methods Mol. Biol. (Totowa, NJ, U.S.)* **2006**, *340*, 207−234.
(5) Pagliaro, L.; Felding, J.; Audouze, K.; Nielsen, S. J.; Terry, R. B.; Krog-Jensen, C.; Butcher, S. Emerging classes of protein−protein interaction inhibitors and new tools for their development. *Curr. Opin. Chem. Biol.* **2004**, *8*, 442-449.
(6) Barril, X.; Hubbard, R. E.; Morley, S. D. Virtual screening in structure-based drug discovery. *Mini Rev. Med. Chem.* **2004**, *4*, 779−791.
(7) Ghosh, S.; Nie, A.; An, J.; Huang, Z. Structure-based virtual screening of chemical libraries for drug discovery. *Curr. Opin. Chem. Biol.* **2006**, *10*, 194−202.
(8) Zhong, S.; Macias, A. T.; MacKerell, A. D., Jr. Computational Identification of Inhibitors of Protein−Protein Interactions. *Curr. Top. Med. Chem.* **2007**, *7*, 63−82.
(9) Sheinerman, F. B.; Giraud, E.; Laoui, A. High affinity targets of protein kinase inhibitors have similar residues at the positions energetically important for binding. *J. Mol. Biol.* **2005**, *352*, 1134−1156.
(10) Chen, R.; Mintseris, J.; Janin, J.; Weng, Z. A protein−protein docking benchmark. *Proteins* **2003**, *52*, 88−91.
(11) Chen, Y.; Xu, D. Computational analyses of high-throughput protein−protein interaction data. *Curr. Protein Pept. Sci.* **2003**, *4*, 159−181.
(12) Zhou, H. X. Improving the understanding of human genetic diseases through predictions of protein structures and protein−protein interaction sites. *Curr. Med. Chem.* **2004**, *11*, 539−549.
(13) Shi, T. L.; Li, Y. X.; Cai, Y. D.; Chou, K. C. Computational methods for protein−protein interaction and their application. *Curr. Protein Pept. Sci.* **2005**, *6*, 443−449.
(14) Chen, F.; Hancock, C. N.; Macias, A. T.; Joh, J.; Still, K.; Zhong, S.; MacKerell, A. D., Jr.; Shapiro, P. Characterization of ATP-independent ERK inhibitors identified through in silico analysis of the active ERK2 structure. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 6281−6287.
(15) Hancock, C. N.; Macias, A.; Lee, E. K.; Yu, S. Y.; MacKerell, A. D., Jr.; Shapiro, P. Identification of novel extracellular signal-regulated kinase docking domain inhibitors. *J. Med. Chem.* **2005**, *48*, 4586−4595.
(16) Cosgrove, D. A.; Bayada, D. M.; Johnson, A. P. A novel method of aligning molecules by local surface shape similarity. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 573−591.
(17) Tsodikov, O. V.; Record, M. T., Jr.; Sergeev, Y. V. Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. *J. Comput. Chem.* **2002**, *23*, 600−609.
(18) Coleman, R. G.; Burr, M. A.; Souvaine, D. L.; Cheng, A. C. An intuitive approach to measuring protein surface curvature. *Proteins* **2005**, *61*, 1068−1074.
(19) Neuvirth, H.; Raz, R.; Schreiber, G. ProMate: a structure based prediction program to identify the location of protein−protein binding sites. *J. Mol. Biol.* **2004**, *338*, 181−199.
(20) Yang, A. Y.; Kallblad, P.; Mancera, R. L. Molecular modelling prediction of ligand binding site flexibility. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 235−250.
(21) Nayal, M.; Honig, B. On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins* **2006**, *63*, 892−906.
(22) Liang, S.; Zhang, C.; Liu, S.; Zhou, Y. Protein binding site prediction using an empirical scoring function. *Nucleic Acids Res.* **2006**, *34*, 3698−3707.
(23) Connolly, M. Analytical molecular surface calculation. *J. Appl. Crystallogr.* **1983**, *16*, 548−558.
(24) Connolly, M. L. Solvent-accessible surfaces of proteins and nucleic acids. *Science* **1983**, *221*, 709−713.
(25) Sotriffer, C.; Klebe, G. Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design. *Farmaco* **2002**, *57*, 243−251.
(26) Campbell, S. J.; Gold, N. D.; Jackson, R. M.; Westhead, D. R. Ligand binding: functional site location, similarity and docking. *Curr. Opin. Struct. Biol.* **2003**, *13*, 389−395.
(27) Laurie, A. T.; Jackson, R. M. Methods for the prediction of protein−ligand binding sites for structure-based drug design and virtual ligand screening. *Curr. Protein Pept. Sci.* **2006**, *7*, 395−406.
(28) Yang, Z. R.; Wang, L.; Young, N.; Trudgian, D.; Chou, K. C. Pattern recognition methods for protein functional site prediction. *Curr. Protein Pept. Sci.* **2005**, *6*, 479−491.
(29) Laurie, A. T.; Jackson, R. M. Q-SiteFinder: an energy-based method for the prediction of protein−ligand binding sites. *Bioinformatics* **2005**, *21*, 1908−1916.
(30) An, J.; Totrov, M.; Abagyan, R. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol. Cell Proteomics* **2005**, *4*, 752−761.
(31) Roche, O.; Kiyama, R.; Brooks, C. L., III. Ligand-protein database: linking protein−ligand complex structures to binding data. *J. Med. Chem.* **2001**, *44*, 3592−3598.
(32) Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411−428.
(33) 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.
(34) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.
(35) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.
(36) MacKerell, A. D., Jr. Empirical force fields for biological macromolecules: overview and issues. *J. Comput. Chem.* **2004**, *25*, 1584−1604.
(37) Ferrin, T. E.; Huang, C. C.; Jarvis, L. E.; Langridge, R. The MIDAS display system. *J. Mol. Graphics* **1988**, *6*, 13−27.
(38) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera−a visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.
(39) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.
(40) Carpenter, G. A.; Grossberg, S. Self-organization of stable category recognition codes for analog input patterns. *Appl. Optics* **1987**, *26*, 4919−4930.
(41) Pao, Y.-H. *Adaptive Pattern Recognition and Neural Networks*; Addison-Wesley: New York, 1989; Vol. 61.
(42) Karpen, M. E.; Tobias, D. J.; Brooks, C. L., III. Statistical clustering techniques for the analysis of long molecular dynamics trajectories: analysis of 2.2-ns trajectories of YPGDV. *Biochemistry* **1993**, *32*, 412−420.
(43) Hetenyi, C.; van der Spoel, D. Efficient docking of peptides to proteins without prior knowledge of the binding site. *Protein Sci.* **2002**, *11*, 1729−1737.
(44) 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.* **1998**, *19*, 1639−1662.
(45) Laskowski, R. A.; Luscombe, N. M.; Swindells, M. B.; Thornton, J. M. Protein clefts in molecular recognition and function. *Protein Sci.* **1996**, *5*, 2438−2452.
(46) Huang, N.; Nagarsekar, A.; Xia, G.; Hayashi, J.; MacKerell, A. D., Jr. Identification of non-phosphate-containing small molecular weight inhibitors of the tyrosine kinase p56 Lck SH2 domain via in silico screening against the pY + 3 binding site. *J. Med. Chem.* **2004**, *47*, 3502−3511.
(47) Hancock, C. N.; Macias, A.; MacKerell, J. A. D.; Shapiro, P. Mitogen Activated Protein (MAP) Kinases: Development of ATP and Non-ATP Dependent Inhibitors. *Med. Chem.* **2006**, *2*, 213−222.
(48) Markowitz, J.; Chen, I.; Gitti, R.; Baldisseri, D. M.; Pan, Y.; Udan, R.; Carrier, F.; MacKerell, A. D., Jr.; Weber, D. J. Identification and characterization of small molecule inhibitors of the calcium-dependent S100B-p53 tumor suppressor interaction. *J. Med. Chem.* **2004**, *47*, 5085−5093.
(49) Pan, Y.; Huang, N.; Cho, S.; MacKerell, A. D., Jr. Consideration of molecular weight during compound selection in virtual target-based database screening. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 267−272.
(50) *SYBYL*, version 7.3; Tripos, Inc.: St. Louis, MO, 2006.
(51) *MOE (Molecular Operating Environment*), version 2007.05; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2007.
(52) Chambers, C. C.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Model for aqueous solvation based on class IV atomic charges and first solvation shell effects. *J. Phys. Chem.* **1996**, *100*, 16385−16398.
(53) Li, J.; Zhu, T.; Cramer, C. J.; Truhlar, D. G. New class IV charges model for extracting accurate partial charges from wave functions. *J. Phys. Chem. A* **1998**, *102*, 1820−1831.
(54) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3−26.
(55) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308−1315.
(56) Tanimoto, T. *IBM Internal Report* **1957**, *Nov.*

(57) Godden, J. W.; Xue, L.; Bajorath, J. Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 163−166.
(58) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273−1280.
(59) Alex MacKerell. http://www.pharmacy.umaryland.edu/faculty/amackere/ (accessed July 24, 2007).
(60) Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graphics* **1996**, *14*, 33−38; 27−38.
(61) Chen, I. J.; Neamati, N.; Nicklaus, M. C.; Orr, A.; Anderson, L.; Barchi, J. J.; Kelley, J. A.; Pommier, Y.; MacKerell, A. D., Jr. Identification of HIV-1 integrase inhibitors via three-dimensional database searching using ASV and HIV-1 integrases as targets. *Bioorg. Med. Chem.* **2000**, *8*, 2385−2398.
(62) Sperl, S.; Jacob, U.; Arroyo de Prada, N.; Sturzebecher, J.; Wilhelm, O. G.; Bode, W.; Magdolen, V.; Huber, R.; Moroder, L. (4-aminomethyl)phenylguanidine derivatives as nonpeptidic highly selective inhibitors of human urokinase. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5113−5118.
(63) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L., III. Assessing scoring functions for protein−ligand interactions. *J. Med. Chem.* **2004**, *47*, 3032−3047.

CI700149K