Identifying and Characterizing Binding Sites and Assessing Druggability

Thomas A. Halgren*

Schrödinger, Inc., 120 West 45th Street, New York, New York 10036

Received September 8, 2008

Identification and characterization of binding sites is key in the process of structure-based drug design. In some cases there may not be any information about the binding site for a target of interest. In other cases, a putative binding site has been identified by computational or experimental means, but the druggability of the target is not known. Even when a site for a given target is known, it may be desirable to find additional sites whose targeting could produce a desired biological response. A new program, called SiteMap, is presented for identifying and analyzing binding sites and for predicting target druggability. In a large-scale validation, SiteMap correctly identifies the known binding site as the top-ranked site in 86% of the cases, with best results (>98%) coming for sites that bind ligands with subnanomolar affinity. In addition, a modified version of the score employed for binding-site identification allows SiteMap to accurately classify the druggability of proteins as measured by their ability to bind passively absorbed small molecules tightly. In characterizing binding sites, SiteMap provides quantitative and graphical information that can help guide efforts to critically assess virtual hits in a lead-discovery application or to modify ligand structure to enhance potency or improve physical properties in a lead-optimization context.

INTRODUCTION

Understanding the structure and function of protein binding sites is a cornerstone of structure-based drug design. Developing this understanding requires knowledge of both the location and physical properties of the binding site. In addition, the identification of small-molecule binding sites as modulators of protein−protein interactions is of increasing interest. Furthermore, even when a validated binding site has been identified, it is often important to find additional potential binding sites where appropriate targeting could result in different biological effects or new classes of compounds.

When the binding site is not known from a 3-D structure or from other experimental data, computational methods can be employed to suggest likely locations. Models that rely on geometric properties, such as POCKET,[1] SURFNET,[2] APROPOS,[3] LIGSITE,[4] CAST,[5] PASS,[6] and CASTp,[7] have long been used for this purpose. These methods are generally very fast and are accurate when the binding sites are well-defined pockets. In developing their refined LIGSITE^{csc} model, Huang and Schroeder have shown that the degree of conservation of key residues in the putative binding pocket can be used to improve LIGSITE's ability to detect drug-binding sites.[8] PocketFinder, a method developed by Abagyan and co-workers, expands on geometric methods by contouring a smoothed van der Waals potential for the protein to identify candidate ligand-binding sites.[9] This partially accounts for physical properties but neglects electrostatic and desolvation effects. Nayal and Honig have employed a large, comprehensive set of physiochemical, structural, and geometric descriptors in developing SCREEN,[10] which has been shown to accurately identify binding sites in a training set of 99 cocrystallized complexes. Other methods have been developed that use fragment probes to find and characterize binding sites.[11−13] These methods more accurately account for the physical properties of the putative binding sites but tend to be computationally expensive. Knowledge-based approaches have also been used with varying degrees of success.[14,15] Finally, methods that combine physics and knowledge have been developed. For example, Coleman et al. developed a method that employs a local version of the MAP$_{POD}$ approach recently proposed by Cheng et al.[16] to identify likely binding sites.[17] This method performed reasonably well for predictions on a limited set of targets. When combined with a statistical residue-coupling analysis, it also suggested a possible allosteric binding site in p38α MAP kinase that has some precedent in another kinase structure. In another example, Joughin et al. developed a method that combines the solvated electrostatic potential, surface curvature, and amino acid identity to predict phosphopeptide binding sites.[18]

When the location of the primary binding site is known, medicinal chemistry efforts to design better ligands can profit from a better understanding of the degree to which known ligands are, or fail to be, complementary to the receptor as well as from a critical assessment of the degree to which the occupancy of accessible but unexplored regions by appropriate ligand functionality can be expected to promote binding or could be used to improve the physical properties of the ligand without lessening its binding affinity. Such assessments can assist in the evaluation and optimization both of known binding molecules and of virtual screening hits. It is also important to understand the potential druggability of the site. It is estimated that 60% of small-molecule drug-discovery projects fail because the target is found to not be druggable.[19] Furthermore, it is estimated that only 10% of the proteins encoded in the human genome are druggable

* Corresponding author phone: (973)-744-0163; e-mail: halgren@schrodinger.com.

10.1021/ci800324m CCC: $40.75
Published on Web 01/20/2009

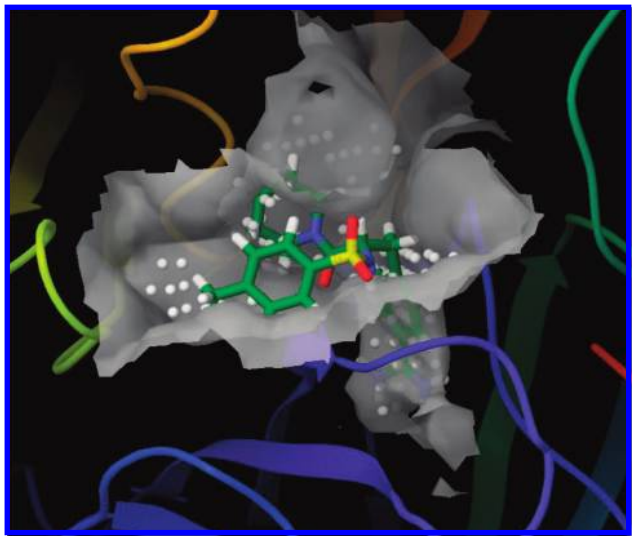

Figure 1. SiteMap surface, site points, and cocrystallized ligand for 1ett, exterior of pocket.

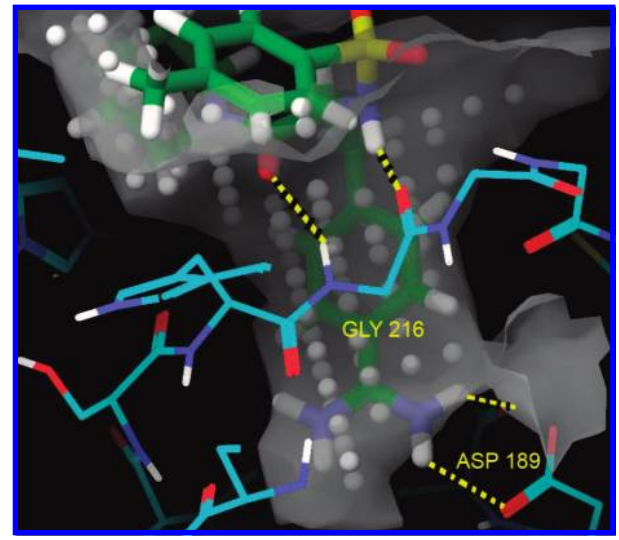

Figure 2. SiteMap surface and site points for 1ett, specificity pocket. The protein is shown in wire-frame.

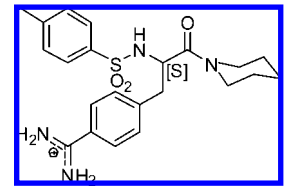

Figure 3. Thrombin 1ett inhibitor 4-TAPAP.

by oral small molecules.[20] Given that most therapeutic projects in the pharmaceutical industry continue to pursue small-molecule, orally available therapeutics, the ability to accurately predict target druggability, before much time and money is spent on discovery efforts, is invaluable. Compared to binding-site identification, less effort has been placed on computational methods for predicting druggability, but the MAP_{POD} model of Cheng et al. has been shown to have high predictive ability.[16]

In this work we describe SiteMap,[21,22] a program we have developed for identifying binding sites and for predicting their druggability. We describe in detail how SiteMap works and validate its utility across a large set of diverse proteins. We also illustrate its ability to characterize binding sites with quantitative and graphical descriptors. SiteMap is fast enough to be used routinely in drug-discovery studies. Thus, proteins with roughly 5000 atoms, including hydrogens, take about 2−3 min on a single CPU of a 2.4 GHz Intel Pentium 4 workstation, while proteins with roughly 8000 atoms typically take about 5 min and proteins with 12,000 atoms take about 9 min; for comparison, the average size of the proteins used in this work is 4250 atoms. These attributes make SiteMap a promising tool for binding-site analysis, virtual-hit assessment, and lead optimization.

RESULTS AND DISCUSSION

Site Maps. To illustrate the graphical feedback provided in a typical application, Figures 1 and 2 show the cocrystallized ligand for the thrombin 1ett receptor (Figure 3) and the "site points" for the binding site generated by SiteMap (white) in the context of the receptor structure and of the gray, translucent SiteMap surface (see the Methods section). The first focuses on relatively exposed regions of the site, while the second profiles the buried specificity pocket. Figures 4 and 5, taken from the same viewpoints, display the hydrophobic (yellow), hydrogen-bond donor (blue), and hydrogen-bond acceptor (red) maps but for clarity suppress the receptor surface. These figures show that portions of the hydrophobic groups of the ligand occupy hydrophobic regions and that the donors and acceptors of the ligand for the most part lie in or close to appropriate donor and acceptor

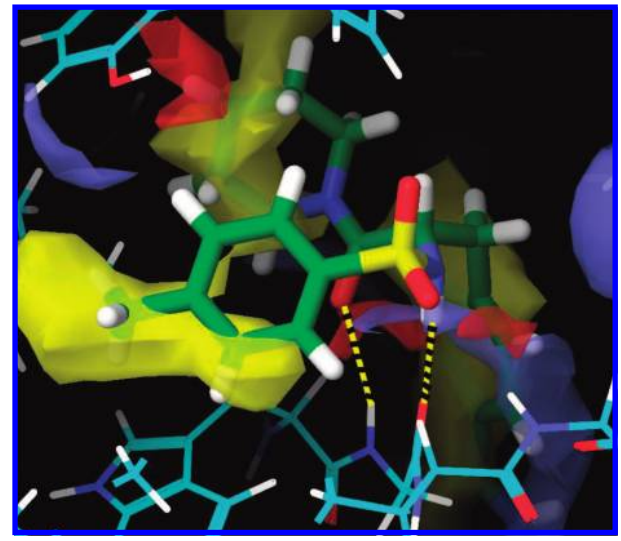

Figure 4. Hydrophobic, donor, and acceptor maps for 1ett, exterior of pocket.

regions. On close examination, it is clear that SiteMap finds that some elements of the ligand structure, such as the non-hydrogen-bonded NH_2 of the benzamidinium group, do not support the binding. In addition, the hydrogen-bonded ligand carbonyl group in Figures 4 and 5 just misses the red acceptor region. But SiteMap is correct on this score because the N···O distance, 3.33 Å, is too long for a strong hydrogen bond. Overall, however, the match is reasonably good; in many other cases, it is quite striking.

In contrast to techniques that color-code the receptor surface, the maps and the properties generated by SiteMap depend on the site as a whole, not just on the character of the closest receptor atom. Moreover, the maps explicitly show the shape and suggest the extent of the regions

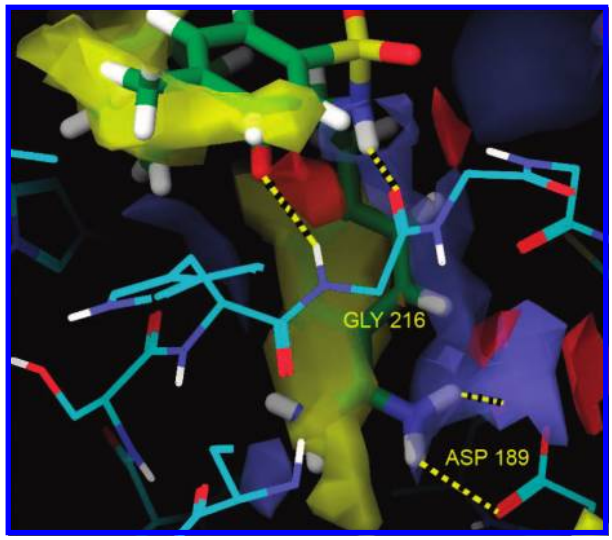

Figure 5. Hydrophobic, donor, and acceptor maps for 1ett, specificity pocket.

considered hydrophobic and hydrophilic, something a surface-based display cannot do. In lead-discovery studies, such maps can aid in the design of better ligands by highlighting regions in which the ligand and the receptor are not complementary or by revealing "targets of opportunity" that might be exploited to improve the binding—for example, hydrophobic regions that have room to accommodate a larger hydrophobic group. They can also be used to quickly evaluate virtual hits from a docking run. Regions that are neither hydrophilic nor hydrophobic are important as well because they show places at which it may be possible to modify the ligand to improve its physical properties—for example, by modifying its solubility—without compromising its binding affinity.

Identifying Binding Sites. We have applied SiteMap to an extensive set of 538 proteins taken from the PDBbind database,[23] prepared as described in the Methods section. The cocrystallized ligands for these proteins have molecular weights of at least 130 Da, and 90% lie between 170 and 600 Da. To focus specifically on drug-binding cavities, the ligands for the sites selected for this study were chosen to have few if any exceptions to druglike properties, as assessed by QikProp.[21] Of these proteins, 342 bind their ligands with affinities of 1 μM or better, and 67 are also subnanomolar binders. Of the remaining 196 ligands, 170 have binding affinities lying between 1 μM and 1 mM, and 26 are millimolar binders. This data set was used to optimize the SiteScore site-identification function by requiring that the site with the best score correspond to the cocrystallized site in as many cases as possible. SiteMap generates a site that matches the known binding site, as judged by the sitemin criterion described below, for 326 of the 342 proteins with micromolar or better ligands (95%) and for 170 of the 196 proteins that bind their ligands less tightly (87%). To facilitate comparison between sites in the same or different proteins, the 326 submicromolar binding sites found by SiteMap were used to calibrate SiteMap's contact, hydrophobic, and hydrophilic components (see the Methods section) so that the average value for each is 1.0. The most significant terms are the size of the site as measured by the number of site points found and the relative openness of the site as measured by the exposure and enclosure properties. The tightness of the site, as measured by the contact term, and the hydrophobic and hydrophilic character of the site were found to be less significant in distinguishing binding sites from other sites. Each of these terms was employed in the SiteScore function used in the original release of SiteMap,[22] which was developed using a smaller set of 230 protein complexes. However, study of the current set of 538 complexes found that nearly optimal discrimination of known binding sites could be obtained by using a version of SiteScore that employs only the size, enclosure, and hydrophilic terms. This is the version used in this work and described in the Methods section.

Table 1. Percent Success in Locating the Cocrystallized Site in 538 Proteins as a Function of the Ligand Binding Affinity[a]

comparison	all sites	<1 nM	1 μM−1 nM	1 mM−1 μM	>1 mM
best-scoring site is correct[b]	85.9	98.5	87.6	81.2	65.4
largest site is correct	78.1	89.6	80.0	72.4	65.4
binding site not found	7.8	1.5	5.5	11.8	26.8

[a] 67 sites have ligand binding affinities of <1 nM, 275 have binding affinities between 1 μM and 1 mM, 170 have binding affinities between 1 μM and 1 mM, and 26 are millimolar binders. [b] A site is considered correct if at least one atom of the cocrystallized ligand lies within 4 Å of the centroid of the site points ("sitemin" criterion).

Table 1 summarizes SiteMap's accuracy in locating the cocrystallized binding site for the 538 proteins. In this work, much as has been done for LIGSITE^{csc}, a "hit" is recorded when at least one atom of the ligand lies within 4 Å of the site centroid,[8] here represented as the centroid of the site points; we shall refer to this as the "sitemin" criterion. Our earlier work[22] employed a less stringent "refmin" criterion, which requires only that at least one atom of the ligand lie within 1 Å of a site point. The table lists the results obtained using the sitemin criterion.

As Nayal and Honig find for SCREEN and report for other methods,[10] size is a fairly good predictor of the ligand-binding site. However, SiteScore is a better predictor, correctly locating the cocrystallized site in 86% of the proteins in the full set and in 98% of the subnanomolar binders. The subnanomolar category also yields the fewest percentage of cases in which no site satisfies the sitemin criterion. Many of the 8% of cases in the full set that fail this criterion, however, satisfy the refmin criterion. In particular, just over 2% of the cases fail to locate the cocrystallized site when the latter is used, and the success percentages shown in the table typically increase by about 5%. The difference between the two criteria reflects cases in which the site points overlap the ligand but extend considerably beyond it in an asymmetric manner, thereby displacing the centroid of the site points significantly from the ligand. Large sites that have such extensions can be appropriate when the substrate is a protein or peptide that contacts the target protein over an extended region. Nevertheless, they complicate ligand-design studies because they make it difficult to locate the "hot spot" that needs to be the focus of developmental efforts.

Even when SiteMap does not find the top-scoring site to be the cocrystallized site, it may still rank that site among the top few, as Table 2 shows.

Approximately 82% of the 538 sites are enzymes, while 18% are receptors that are involved in cellular signaling

Table 2. Percent Success a Function of the Ligand Binding Affinity in Locating the Cocrystallized Site as One of the Top N Sites

sites/affinity	top 1	top 2	top 3	top 5
all sites	85.9	90.1	91.3	91.8
<1 nM	98.5	98.5	98.5	98.5
1 μM−1 nM	87.6	90.1	91.2	91.6
1 mM−1 μM	81.2	85.3	87.1	87.6
>1 mM	65.4	69.2	69.2	73.1

Table 3. Effect of Protein Preparation[a] on SiteMap Accuracy as a Function of Binding Affinity

protein preparation	all sites	<1 nM	1 μM−1 nM	1 μM−1 mM	>1 mM
remove waters, add hydrogens	81.3	98.5	84.4	73.5	57.7
apply protassign to complex	80.5	98.5	83.3	71.2	65.4
also relax complex to rmsd 0.35 Å	85.9	98.5	87.6	81.2	65.4
apply protassign to protein only	80.7	98.5	84.0	70.6	65.4
also relax protein to rmsd 0.35 Å	81.4	95.5	86.2	72.4	53.8

[a] See the Methods section.

processes or are proteins that bind ligands for some other purpose, such as for storage, transport, or immune response. The data show that SiteMap handles the enzymatic sites somewhat more accurately. In particular, 85.9% of the best-scoring sites set are correct for the full set (Table 2), as opposed to 87% of the enzymatic sites and 80% of the nonenzymatic sites. This 7% differential does not reflect significant differences in binding affinities, as the nonenzymatic sites comprise a relatively constant 15−19% of the sites in each of the activity ranges shown in Tables 1 and 2. We do not know, however, whether the differential reflects differences in the kinds of interactions used to bind ligands, as a referee has suggested (e.g., main-chain vs side-chain interactions), arises for some other physical reason or represents a statistical fluctuation associated with the relatively small size of the nonenzymatic set.

To test the sensitivity of SiteMap to the protein geometry, we examined four additional sets of protein sites, namely, the sets obtained by removing waters and small molecules and adding hydrogens, by applying the protassign procedure described in the Methods section to the resultant complexes, by applying protassign to the unligated proteins, and by relaxing the unligated proteins after applying protassign. Table 3 shows that annealing the protein−ligand complex improves the results but that other aspects of the protein preparation have little effect. The better performance found using proteins relaxed with the cocrystallized ligand present may indicate that the annealed sites are physically more realistic. For the 67 sites with binding affinities of 1 nM or better, all but one site is identified correctly for four of the five protein preparations. The last-listed preparation, which relaxes the unligated protein site, misses in three cases, but a site that satisfies the sitemin criterion ranks second in each case. The missed site for three other preparations also ranks second; that for the fourth corresponds to the binding site but narrowly fails to satisfy the sitemin criterion.

Discriminating Binding Sites from Other Sites. By setting a threshold SiteScore of 0.80 (~80% of the average found for the submicromolar sites), SiteMap can be employed to discriminate sites that bind ligands from sites not known to do so. Its performance as a classifier is summarized in Table 4, which shows that 92% of the PDBbind targets generate a site that matches the cocrystallized site when the sitemin criterion is used and that 82% both match the known binding site and have a SiteScore of 0.80 or better. We also find that 95% of the 342 submicromolecular sites reproduce the cocrystallized site, 87% with a SiteScore of at least 0.80.

Table 4. Performance in Classifying Binding Sites in Proteins

	all sites		submicromolar sites	
comparison	number	percent	number	percent
binding site has SiteScore ≥0.80	441	82.0	299	87.4
binding site has SiteScore <0.80	55	10.2	27	7.9
binding site not found	42	7.8	16	4.7
other site has SiteScore ≥0.80	463	--	285	--

These results suggest that SiteScore can be used as one criterion for deciding whether to target a given site, but they do not show that a site that scores well *is* a drug-binding site—just that it *could be*. When configured to report up to 10 sites, SiteMap finds 2286 sites for the 538 proteins or about 4 per protein; 60% have fewer than 5 sites, and only 27 have 10 or more. Of the 2286 sites, 904 have SiteScores of 0.80 or better. Table 4 shows that 441 sites match the known binding site, while 463 arise from other sites. For the 342 submicromolar targets, 1430 sites are found, 584 of which have SiteScores of at least 0.80. Of this number, 299 reproduce the cocrystallized site and 285 arise from other sites. For this data set, therefore, about half of the sites with SiteScores of 0.80 or higher are known binding sites. Some of the other sites, to be sure, may bind different ligands, but it is not possible to gauge how often this may be the case.

Classifying Druggability. As previously noted, many drug-design projects fail because the target proves not to be druggable. Accurate determination at an early stage of whether a given protein is or is not druggable therefore has the potential of saving considerable time and expense. In an important recent contribution, Cheng and co-workers[16] developed a model for predicting the maximal affinity achievable for a given target by a passively absorbed oral drug (MAP_{POD}) and used it to assess druggability. They model the maximal achievable binding energy as

$$\Delta G_{MAP_{POD}} = -300\ \text{Å}^2 f_{nonpolar}\gamma(\infty)/(1 - 1.4/r) \qquad (1)$$

where $f_{nonpolar}$ is the nonpolar fraction of the solvent-accessible surface area, $\gamma(\infty)$ is 45 cal/mol/ Å^2, and r is the radius of curvature computed for the site. In their approach, the "undruggable" and "difficult" targets are those that have the poorest predicted maximal affinity. As eq 1 shows, an increase in the fraction $f_{nonpolar}$ enhances activity, as does a decrease in the radius r. We have developed an alternative model that generates a "druggability" score for the target site. This score, Dscore, includes terms that promote ligand binding—adequate size, isolation from solvent—but offsets them with a term that penalizes increasing hydrophilicity as shown in eq 2

$$\text{Dscore} = 0.094\, n^{1/2} + 0.60\, e - 0.324\, p \qquad (2)$$

where n is the number of site points found for the site, capped at 100, e is the degree of enclosure of the site, and p is the hydrophilic score computed for the site. The equation and its derivation are discussed in the Methods section. In our model, the sites predicted to be "druggable" are those with the highest scores, while "undruggable" sites have the lowest scores and "difficult" sites have intermediate values. Though expressed in a different language, a central idea—that highly polar sites are least likely to be druggable—is much the same in our approach and in Cheng's. Nevertheless, the two models behave differently in many respects.

Cheng et al. examined 63 protein sites representing 27 proteins, 22 of which had marketed drugs or advanced-stage drug candidates as of November, 2005; despite intensive efforts by pharmaceutical companies, five did not. They regard four of the five as undruggable: PTP1B, cathepsin K, caspase 1 (ICE-1), and HIV integrase. They classify the fifth such site—MDM2, a regulator of tumor suppressor p53—as "druggable", even though there were as yet no orally active small-molecule drugs, because they regard it as an interesting example of a druggable protein−protein interaction. Among the 22 proteins with marketed drugs or advanced candidates, they assign angiotensin converting enzyme (ACE), the HIV-rt nucleotide site, inosine monophosphate dehydrogenase (IMPDH), neuraminidase, thrombin, and penicillin binding protein to the "difficult" category. The common factor is that the "difficult" targets almost universally employ prodrugs that are cleaved *in vivo* to produce ionic functionality that contributes to, and may be essential for, ligand binding. Many prodrugs, for example, are carboxylic acid esters that yield a carboxylate anion whose function is to chelate a metal ion such as Zn^{2+} or to occupy an anionic recognition site. They classify these targets as "difficult" because prodrugs sufficiently complicate the developmental process that they are avoided whenever possible. For penicillin binding protein (1qmf), many of the antibiotics are indeed administered as prodrugs, but some have unprotected carboxylates and are not. However, the latter cases appear to be actively transported.[16] Cheng et al. point out, however, that active transport is not a mechanism generally available to the drug researcher, because transporter selectivity is not well understood.[24] In all, they classify 10 sites of 4 proteins as "undruggable", 10 sites of 6 proteins as "difficult", and 43 sites of 17 proteins as "druggable".

We recognize that some of the Cheng sites might be classified differently by others, but we shall accept their classification, for the most part, because our primary objective will be to determine how well SiteMap can reproduce it. We will show that SiteMap does an excellent job but disagrees in a few cases. Our hope is that examining the basis for agreement and disagreement will shed light on the degree to which the physical properties of the site can be expected to predict druggability. We also recognize that other targets, if divided into "undruggable", "difficult", and "druggable" sites in an equally valid manner, might show different physical characteristics than those found for the Cheng set. It remains to be seen how SiteMap will fare when applied to protein targets not considered here.

Table 5 lists the ranks assigned by SiteMap and MAP_{POD} for the 10 "undruggable" and 10 "difficult" sites; a SiteMap listing for the full set of 63 targets, including values obtained for key properties, is given in Table 6. The tables show that Dscore agrees remarkably well with the Cheng classification. Thus, Dscore ranks the "undruggable" sites 1−10 and the "difficult" sites 11−15, 17, 18, 21, 25, and 45. SiteScore, however, is significantly less effective. Table 5 shows that MAP_{POD} does a bit better at differentiating "druggable" sites from other sites but that Dscore readily distinguishes "undruggable" from "difficult" targets while MAP_{POD} does not.

Table 5. Ranks of Druggable and Difficult Protein Sites Assigned by SiteMap and by Maximal Affinity Predicted for a Passively Absorbed Oral Drug (MAP_{POD})[16]

	rank				
case	category[a]	Dscore	SScore	MAP_{POD}	target
1qs4	0	1	1	1	HIV integrase
1pty	0	2	7	8	PTP1B
1onz	0	3	4	9	PTP1B
1g1f	0	4	6	10	PTP1B
1nny	0	5	5	11	PTP1B
1q1m	0	6	9	12	PTP1B
1nlj	0	7	2	17	cathepsin K
1jf7	0	8	15	13	PTP1B
1mem	0	9	3	18	cathepsin K
1bmq	0	10	8	14	caspase 1 (ICE-1)
1a4g	1	11	36	2	neuraminidase
1nnc	1	12	12	3	neuraminidase
1nf7	1	13	42	17	IMPDH
1f8b	1	14	13	4	neuraminidase
1mwe	1	15	21	5	neuraminidase
1t03	1	17	22	7	HIV-rt (nucleotide site)
2qwk	1	18	16	6	neuraminidase
1o86	1	21	49	19	ACE-1
1qmf	1	25	26	15	penicillin binding protein
1kts	1	45	37	27	thrombin

[a] "Undruggable" sites belong to category 0, "difficult" sites to category 1.

To see how sensitive the rankings are to the details of the protein preparation, we also carried out calculations for protein sites prepared: (1) by just removing water and the ligand and adding hydrogens and (2) by then applying protassign but not relaxing the resultant protein−ligand complexes. These calculations also gave good results. Specifically, the "undruggable" sites ranked 1−9 and 11 and the "difficult" sites rated 10, 12−17, 19, 37, and 43 for set (1), while the former ranked 1−5 and 7−11 and the latter sites rated 6, 12, 14−17, 20, 32, 39, and 41 for set (2). The similarities in these rankings show that the procedure used for protein preparation does not greatly affect SiteMap's ability to classify the druggability of binding sites.

The characteristics that appear to best distinguish "druggable", "difficult", and "undruggable" sites are listed in Table 7. As Table 6 shows, the "undruggable" sites typically are much smaller and somewhat less sheltered from the solvent than are either the "difficult" or "druggable" sites. Moreover, these sites also tend to have the most hydrophilic and the least hydrophobic character. All four properties combine to distinguish "undruggable" from "druggable" sites, while size and enclosure on balance distinguish "undruggable" from "difficult" sites, and hydrophilic and hydrophobic character distinguish "difficult" from "druggable" sites.

Among the "undruggable" targets, the six PTP1B sites are somewhat smaller than the average site (and than the cutoff of 100 site points used in the Dscore function). Their

Table 6. SiteMap Property Values and Dscore Ranks for the Cheng Set

case	category[a]	rank	Dscore	SScore	size	enclosure	philic	phobic
1qs4	0	1	0.179	0.413	4	0.697	1.311	0.132
1pty	0	2	0.496	0.883	57	0.792	2.115	0.188
1onz	0	3	0.529	0.840	56	0.735	1.888	0.000
1g1f	0	4	0.600	0.882	67	0.721	1.847	0.000
1nny	0	5	0.629	0.868	66	0.706	1.711	0.037
1q1m	0	6	0.721	0.925	85	0.671	1.683	0.137
1nlj	0	7	0.750	0.751	38	0.629	0.633	1.376
1jf7	0	8	0.759	1.015	100	0.720	1.881	0.000
1mem	0	9	0.782	0.804	60	0.649	1.029	0.975
1bmq	0	10	0.864	0.890	79	0.655	1.119	0.512
1a4g	1	11	0.866	1.071	100	0.804	1.706	0.070
1nnc	1	12	0.870	1.001	80	0.815	1.411	0.411
1nf7	1	13	0.876	1.091	175	0.835	1.731	0.175
1f8b	1	14	0.887	1.009	81	0.821	1.384	0.323
1mwe	1	15	0.908	1.048	92	0.815	1.479	0.370
1pwm	2	16	0.938	0.967	83	0.746	1.122	1.338
1t03	1	17	0.952	1.050	371	0.773	1.385	0.303
2qwk	1	18	0.952	1.017	85	0.809	1.227	0.643
1kzn	2	19	0.993	0.967	86	0.685	0.889	0.779
1ptw	2	20	1.001	1.099	139	0.846	1.370	0.487
1o86	1	21	1.004	1.111	283	0.864	1.393	0.477
1e1x	2	22	1.021	1.060	176	0.787	1.201	0.872
1h08	2	23	1.026	1.044	235	0.764	1.142	1.069
1zkn	2	24	1.037	1.065	145	0.795	1.166	0.725
1qmf	1	25	1.039	1.056	165	0.782	1.137	0.424
1hwl	2	26	1.041	1.047	158	0.769	1.104	0.665
1hw9	2	27	1.042	1.068	152	0.799	1.159	0.723
1m17	2	28	1.043	1.025	82	0.787	0.858	0.958
1hvr	2	29	1.044	1.054	178	0.778	1.115	0.818
1hwk	2	30	1.055	1.092	134	0.836	1.186	0.715
1hwi	2	31	1.056	1.037	117	0.747	1.019	0.557
1hwr	2	32	1.057	1.059	168	0.786	1.087	0.862
1h07	2	33	1.058	1.051	176	0.775	1.064	1.216
1qbs	2	34	1.059	1.060	151	0.787	1.084	0.873
1oyn	2	35	1.064	1.107	158	0.857	1.198	1.294
1kij	2	36	1.069	1.060	193	0.787	1.055	0.803
1ezq	2	37	1.074	1.015	101	0.645	0.777	0.815
1q9m	2	38	1.075	1.082	159	0.820	1.095	0.949
1dmp	2	39	1.080	1.058	163	0.769	0.987	0.908
1hwj	2	40	1.082	1.086	162	0.827	1.086	0.749
1hw8	2	41	1.085	1.055	107	0.748	0.931	0.711
1i2z	2	42	1.087	1.069	136	0.790	1.003	0.695
1ke6	2	43	1.094	1.067	135	0.770	0.944	1.178
1ke9	2	44	1.098	1.077	146	0.794	0.977	1.271
1kts	1	45	1.101	1.071	117	0.770	0.924	1.314
1kv1	2	46	1.102	1.103	265	0.852	1.072	1.846
1t41	2	47	1.104	1.082	123	0.799	0.969	1.676
1udt	2	48	1.119	1.116	255	0.871	1.052	1.295
1gpk	2	49	1.122	1.123	180	0.882	1.065	0.921
1ke8	2	50	1.137	1.110	129	0.829	0.922	1.577
1iep	2	51	1.140	1.115	246	0.838	0.930	1.582
1ea1	2	52	1.154	1.130	176	0.859	0.925	1.618
1c14	2	53	1.183	1.125	90	0.818	0.614	1.489
1udu	2	54	1.186	1.145	244	0.844	0.799	1.767
1rv1	2	55	1.197	1.107	229	0.706	0.511	1.623
1e66	2	56	1.201	1.161	202	0.867	0.793	1.257
1ep4	2	57	1.222	1.157	183	0.817	0.639	2.349
4cox	2	58	1.252	1.205	189	0.915	0.727	2.781
1pwl	2	59	1.264	1.195	108	0.859	0.588	3.016
1rth	2	60	1.267	1.206	166	0.888	0.630	2.744
1e9x	2	61	1.323	1.242	142	0.896	0.474	2.571
1eet	2	62	1.363	1.271	143	0.916	0.390	3.282
1fk9	2	63	1.376	1.287	121	0.943	0.400	3.656
average value:								
undruggable			0.631	0.827	61	0.698	1.522	0.336
difficult			0.871	0.995	140	0.799	1.385	0.413
druggable			1.108	1.091	156	0.807	0.926	1.374
all cases			1.011	1.048	143	0.793	1.099	1.061

[a] "Undruggable" sites belong to category 0, "difficult" sites to category 1, and "druggable" sites to category 2.

Table 7. Characteristics of Undruggable, Difficult, and Druggable Sites

category	characteristics
undruggable	(a) very strongly hydrophilic; relatively small in size; little or no hydrophobic character; or (b) requires covalent binding; or (c) very small or very shallow
difficult	sufficiently hydrophilic to require administration as a prodrug; less hydrophobic than a typical site
druggable	of reasonable size, enclosure, and hydrophobicity with unexceptional hydrophilicity

distinguishing feature, however, is that they are strongly hydrophilic and have little or no hydrophobic character. One reason for classifying these sites as "undruggable" is that they may well require charged, even multiply charged, ligand functionality that would impede passive transport[24] unless protected as prodrugs. An even more important reason, however, may be their lack of significantly sized hydrophobic regions whose occupancy by hydrophobic ligand functionality could provide a strong driving force for ligand binding.

The "undruggable" cathepsin K (1nlj, 1mem) and caspase 1 (1bmq) sites, on the other hand, are not strongly hydrophilic. However, their relatively small size and greater exposure to solvent, relative to a typical site, rank them after five of the six PTP1B sites but before the "difficult" sites. Moreover, for the cathepsin K sites and the "difficult" penicillin binding protein site (1qmf), a second factor—covalent binding—comes into play. The suitability of the site for noncovalent interactions may matter less in these cases because the driving force for ligand binding may be the energetics of the covalent bond formation itself. For cathepsin K, at least, it is reasonable to ask whether the site should be regarded as "undruggable" because the irreversible covalent binding it appears to require raises such serious issues that is avoided whenever possible, or because its site, being small, is less conducive to strong noncovalent interactions. A case can certainly be made that sites that require covalent binding may be "undruggable", or at least "difficult", because of the chemistry involved rather than because of the general physical properties of the site.

It seems appropriate to also classify sites that have very few site points as "undruggable", as occurs for HIV integrase (1qs4, 4 site points). Sites can have few site points either because the cavity is small or because the site is shallow and few candidate site points are sufficiently sheltered from solvent to be retained in the site-finding stage (see the Methods section). HIV integrase is an example of a shallow site that offers few opportunities for strong ligand binding.

Most targets that fall into the "difficult" category have good size and enclosure but are more hydrophilic and much less hydrophobic than the "druggable" sites. For these sites, having substantial hydrophilic character is relevant because it can signal the need for administration as a prodrug. Moreover, significant hydrophobic volume, which most "undruggable" and "difficult" targets lack, may be needed to achieve adequate binding affinity with a druglike compound. The previously discussed penicillin binding protein (1qmf), ranked twenty-fifth, and thrombin (1kts), ranked forty-fifth, are exceptions to this rule in that they are not strongly hydophilic. Of these, the 1qmf site scores relatively well because of its substantial size and good enclosure. Except for its low hydrophobic character, it has the earmarks of an attractive binding pocket. The reason that the MAP_{POD} scores recognize it as "difficult", while Dscore does not, may be that this site, while primarily polar, is (in SiteMap's estimation) only moderately hydrophilic. The substantial size of the site, though, is by far the most significant factor in boosting its Dscore value to within the "druggable" range. Size, however, is not overtly a factor in the Cheng approach because MAP_{POD} employs the ratio $f_{nonpolar}$ rather than the nonpolar surface area itself and thus in effect scales all sites to a common size. This difference may also be why the MAP_{POD} approach is less able to distinguish between "difficult" and "undruggable" sites. We should note, however, that the radius of curvature, r, in eq 1 may act as a surrogate for the number of site points, n, in eq 2 (or vice versa), if sites that have few site points because they are shallow also tend to have a large MAP_{POD} radius of curvature. In the absence of published values for r, however, it is difficult to make such a judgment.

The thrombin/factor Xa pairing (1kts/1ezq) presents an interesting case. One might expect that the two serine proteases would pose similar developmental challenges, but Cheng classifies thrombin as "difficult" but factor Xa as "druggable". Moreover, SiteMap actually scores the "difficult" thrombin site better, even though it finds this site to be the more hydrophilic. For thrombin, ximelagatran, the late-stage compound cited by Cheng, was ultimately withdrawn because of concerns regarding possible liver toxicity. Subsequent research produced the released compound dabigatran elixate, marketed in Europe and Canada as Pradaxa, but this compound is a prodrug that protects both carboxylate and benzamidinium functionalities. The latter moiety undoubtedly interacts with the Asp 189 carboxylate in the S1 specificity pocket in a manner similar to that shown in Figure 5. The factor Xa inhibitor rivaroxaban[25] (Xarelto) and the late-stage investigational compound apixaban,[26] in contast, are not prodrugs. These compounds lack a correspondingly basic group; while they occupy the specificity pocket and approach Asp 189 closely, they are unable to hydrogen bond to it. To avoid paying the considerable cost of desolvating a carboxylate anion, we suspect that Asp 189 is present as the neutral carboxylic acid in these complexes. We used the publicly available Karlsberg package[27] to study the effect of the environment on the pK_a and found that it gives estimated pK_a values for Asp 189 of 6.2 and 6.5 for "Vacuum" and "GBSW" calculations for 1fjs (factor Xa) and of 6.0 for the "vacuum" calculation for 1ett (thrombin). For comparison, aliphatic carboxylic acids have a pK_a of about 4.8, and the Karlsberg package gives an average pK_a of 4.2 (range, 2.8−5.4) for nine exposed Asp and Glu residues in 1fjs and 1ett that do not engage in close electrostatic interaction with the protein. These values suggest that Asp 189 could be protonated at minimal cost at pH 7. In light of reports of active thrombin inhibitors that also lack a strongly basic moiety and have good pharmacokinetic properties,[28] an orally administered thrombin therapeutic that is not a prodrug may yet be developed—in which case thrombin presumably would move in Cheng's classification from "difficult" to "druggable" (despite its demonstratively difficult development). In principle, even for a strongly hydrophilic "undruggable" target it might be possible to design an oral therapeutic that does not require prodrug administration; this might be

Table 8. Selected Dscore Values for PDBbind Cases with Subnanomolar Affinity

case	Prep 1[a]	Prep 2[b]	Prep 3[c]	Prep 4[d]	Prep 5[e]
1a42	*0.854*	*0.904*	*0.937*	*0.904*	**0.811**
1bn1	*0.850*	*0.965*	*0.918*	*0.965*	**0.819**
1bn3	*0.938*	*0.909*	*0.954*	1.005	*0.860*
1bn4	*0.897*	*0.964*	*0.890*	*0.979*	***0.828***
1bnn	*0.861*	*0.971*	*0.896*	*0.965*	*0.858*
1bnq	*0.902*	*0.926*	*0.875*	*0.926*	*0.842*
1bnt	*0.915*	*0.944*	*0.921*	*0.961*	***0.803***
1bnu	***0.757***	*0.843*	***0.778***	*0.835*	***0.681***
1bnw	*0.871*	*0.963*	*0.918*	1.011	*0.865*
1cil	*0.912*	0.991	*0.941*	1.001	*0.878*
1ezq	1.063	1.081	1.061	1.081	*0.971*[f]
1fjs	1.072	1.034	*0.925*	1.039	1.036
1if2	***0.745***	***0.824***	**0.798**	*0.837*	***0.789***
1ksn	1.055	1.028	1.075	1.031	0.983[f]
1o86	1.055	1.054	1.050[f]	1.056	1.035
1rm8	*0.950*	*0.973*	*0.936*	*0.968*	*0.908*
1swp	*0.834*	*0.966*	*0.914*	1.016	*0.976*
1we1	1.063[f]	1.134[f]	1.010	1.134[f]	1.049
2tct	1.052	*0.932*	*0.956*	1.049	1.066[f]
6cpa	***0.771***	*0.867*	*0.938*	***0.791***	*0.892*
7cpa	***0.712***	***0.808***	*0.852*	***0.756***	***0.790***
8cpa	***0.764***	*0.863*	*0.853*	***0.798***	***0.791***
Cat. 0[g]	6	2	2	3	8
Cat. 1[h]	10	14	16	9	10
Cat. 2[i]	51	51	49	55	50

[a] Remove waters, add hydrogens. [b] Apply protassign to complex. [c] Also relax complex to rmsd 0.35 Å. [d] Apply protassign to protein only. [e] Also relax protein to rmsd 0.35 Å. [f] Top-ranked SiteScore site does not meet sitemin ≤4 Å criterion. [g] Number of sites predicted to be "undruggable" (Dscore shown in bold italics). [h] Number of sites predicted to be "difficult" (Dscore shown in italics). [i] Number of sites predicted to be "druggable".

accomplished by displacing the ligand far enough from the locus of polar (and, particularly, charged) ligand—receptor interactions to minimally disturb the water atmosphere that exists in the apo state. But the question would then be whether it is possible to generate sufficient binding affinity in the absence of suitable regions for hydrophobic interaction.

To close this section, we shall examine the sensitivity of the druggability classifications to the procedure used for protein preparation. Table 8 shows Dscore values obtained using the five previously discussed procedures for 22 of the 67 subnanomolar PDBbind sites (*cf.* Table 3). In rough accord with the data on druggability classification in Table 6, we have, somewhat arbitrarily, assigned sites with Dscore values smaller than 0.83 as "undruggable" (shown in ***bold italics***), those having Dscore values between 0.83 and 0.98 as "difficult" (*italics*), and those having larger Dscore values as "druggable". The 22 sites are those that, for one or more of the preparation procedures, are predicted to be "undruggable" or "difficult" or for which the top-ranked SiteScore site does not satisfy the sitemin ≤4 Å criterion. The listing shows that the classification is independent of the method used for protein preparation in most cases. In particular, when the 45 subnanomolar sites uniformly predicted to be "druggable" are included, the probability that any two preparation methods predict the same druggability classification for a given site is found to be 89%. This probability rises to 95% when the question is simplified to whether or not the site is predicted to be "druggable". Moreover, the predictions for the 74 PTP1B, neuraminidase, ACE, thrombin, and factor Xa sites within the full set of 538 PDBbind sites are consistent with those made for the Cheng set in 80% of the cases. The largest divergence occurs for the 8 neuraminidase sites, where just over half of the assignments for the five preparation procedures rate the site as "druggable". The observed variations suggest to us that consensus assignments made when multiple structures are available for a given protein target are likely to be more reliable. For example, the 1c1v site actually predicts thrombin to be "undruggable" (Dscore 0.80), a clear outlier. This designation arises partly because the top-ranked SiteScore site is not the cocrystallized site in this instance (the second is), and partly because each of the first two sites has an atypically large hydrophilic character, as assessed by SiteMap. To be sure, the more important question is whether the druggablity classifications are accurate for systems not in the Cheng training set. This question, however, is largely beyond the scope of the present study.

METHODS

SiteMap extends a procedure for characterizing binding sites developed at the Merck Research Laboratories.[29] In this capacity it operates in a manner similar to Goodford's GRID algorithm[30] but employs a unique definition of hydrophobicity. A SiteMap calculation of the type we describe here has three stages. In the first, relevant site points are selected based on geometric and energetic properties, and the points are grouped into sets to define the sites. Next, hydrophobic, hydrophilic, and other key properties are computed at grid points and contour maps are prepared, much as was done in the original method. Finally, site properties such as those listed in Table 6 are computed. These steps are described below. First, however, we will define the procedure used to prepare the protein sets used in this work.

Protein Preparation. Approximately 100 of the 538 PDBbind structures used in the study on identifying sites have two or more subunits with equivalent (or nearly equivalent) binding sites. To keep regions between protein monomers from generating unphysical sites that score well but exist only in the crystal lattice, only protein monomers were retained, normally by keeping chain A. The proteins were then prepared using standard Schrödinger techniques, as described below. To avoid prejudicing the search, all crystallographic water was first removed, as were small molecules of crystallization. The preparation continued by assigning appropriate bond orders and formal charges to ligands and cofactors and by using the Schrödinger Maestro interface[21] to add hydrogens. The protassign procedure was then applied to the protein—ligand complexes to reorient glutamine and asparagine amide groups, where called for, and to assign the tautomeric and protonation state of histidine residues and the protonation state of Asp and Glu residues. Protassign also reorients protein and ligand hydroxyl groups to optimize hydrogen-bonding patterns. Manual modifications were then made after visual inspection in some cases, particularly for Asp and Glu protonation states in aspartyl proteases and some metalloproteases. Finally, to remove unphysical contacts, the impref procedure was employed to relax the complexes. This procedure carries out a series of short refinement stages that use increasingly smaller Cartesian restraints. When the refined structure exceeds a specified root-mean-square deviation (rmsd) to the starting coordinates

of non-hydrogen atoms, here taken to be 0.35 Å, the procedure terminates and returns the structure generated in the previous refinement stage, which by definition did satisfy the rmsd limit.

Most protein sites employed in classifying druggability also were prepared using this procedure. For 1t03, the nucleotide binding site of HIV-rt, and for 1rv1, the MDM2/p53 complex, however, we applied the protassign and impref procedures to the unliganded protein sites because neither has a typical small-molecule ligand. We also modified the procedure for the covalently bound cathepsin K (1mem and 1nlj) and penicillin binding protein (1qmf) complexes by breaking the covalent bond and excising enough of the ligand at the point of attachment to keep the remaining structure from strongly clashing with the protein in the impref step.

Finding Sites. The objective of the first stage of a SiteMap calculation is to locate, but not accurately score, the sites. A "site" is defined by a set of points on a grid that are either contiguous or are bridged by short gaps in exposed regions. The site-finding algorithm begins by placing a 1-Å grid around the entire protein or within a box centered on a reference species such as a docked or cocrystallized ligand. Each grid point is classified as being "inside" or "outside" the protein by comparing the distance to nearby protein atoms to the van der Waals (vdW) radius of those atoms. If the ratio of the squares of these distances is uniformly larger than a given threshold (default value, 2.5), the point is considered to be outside the protein.

The "outside" points are examined in the next step to determine which ones are in good vdW contact with the receptor and are reasonably well enclosed by it. Enclosure is computed by sampling 110 evenly spaced directions and determining the fraction of radial rays that strike the receptor surface within a given distance (default value, 8 Å). If the fraction is smaller than a given threshold (default value, 0.5), the point is discarded. Contact with the receptor is determined by a cutoff on the vdW interaction energy at the grid point (default value, −1.1 kcal/mol), and the point is rejected if the interaction energy is positive or is too small in magnitude. Grid points that meet both criteria are added to a list of candidate site points. In these calculations, the radius and well depth of the Lennard-Jones probe particles are taken to be 1.5 Å and 0.13 kcal/mol, respectively.

The third step combines site points into groups. For a site point to be considered for membership in a group, a minimum number of candidate site points (default value, 3) must lie within a given distance (default value, 1.76 Å). Site points that do not meet this criterion are discarded. The process starts by assigning a site point to a group and then recursively adds candidate points that are sufficiently close to an included site point (the default requirement is that the sum of the differences in grid indices be ≤ 3 and the sum of the squares of these differences be ≤ 5). When no further site points can be added, the group is considered complete and another site-point group is started. The process continues until all site points have been examined. Groups that have fewer than a required minimum number of site points (default, 3) are then discarded.

The final step merges site-point groups when the gap between them, as measured by the distance of closest approach, occurs in an exposed region and is less than or equal to a specified distance (default value, 6.5 Å). The requirement for establishing that the region is suitably exposed is that it be possible to step between the site points of closest approach by traversing a connected sequence of grid points that lie outside the protein. To limit the growth of the site-point sets, SiteMap merges groups only when at least one has no more than 100 (default value) site points.

The final groupings constitute the sites. In the present work, SiteMap was configured to return up to 10 sites (the default value is 5); when more than the requested number are generated, the largest are taken. (We returned 10 sites to better characterize cases in which a site is found that matches the known binding site but scores poorly. We found just one case of this type, one in which the cocrystallized site ranks sixth.) One site is always returned (unless no site-point groups survive), but other sites must meet a specified minimum size (default value, 15 site points).

To allow SiteMap to be adapted to the needs of a specific problem, the default quantities specified above are user settable. For example, to detect more shallow sites, such as those sometimes found in protein−protein interfaces, one might relax the −1.1 kcal/mol cutoff on the contact energy, count radial rays that intersect the protein surface at distances of greater than 8 Å, and/or accept grid points for which fewer than 50% of the rays intersect the surface.

Visualizing Sites. The second stage prepares contour maps that express the character of the sites. In this stage, van der Waals and distance-dependent electrostatic interactions of a probe placed at the grid points are employed to generate field grids. These grids are computed in rectilinear boxes (default grid-point separation, 0.7 Å) that extend beyond the site-point positions by 6 Å on a side (default value). The probe, which loosely simulates a water molecule, is represented by a vdW sphere of radius 1.6 Å, a well depth of 0.13 kcal/mol, and a point dipole moment of 2.3 Debye. To form the electric-field grid, the probe's point dipole is oriented along the field but is offset from the grid point by 0.10 Å along the field direction. The offset point lies closer to the protein at points at which a ligand might donate a hydrogen bond to the protein and further away where it might receive one. To more accurately represent the expected contact positions and interaction energies of donor and acceptor ligand atoms, the OPLS_2001 force field[21] used in this work is modified by reducing formal-charge contributions to the partial atomic charges by 50%. This reduction in formal charges, like the one employed in Glide,[31,32] helps to keep regions around charged groups from dominating the maps, which are meant to reflect interactions in solvent rather than in the gas phase. The resultant vdW and electric-field grids are then employed to generate the potentials used to produce the hydrophilic, hydrophobic, donor, acceptor, surface, and, when appropriate, metal-binding maps. When SiteMap calculations are submitted from Maestro, the site-point sets and the site maps are returned to the session for convenient visualization. The properties of the sites, described below, are also returned and are displayed in the Maestro Project Table.

1. Hydrophilic Map. SiteMap constructs a measure of hydrophilicity by adding an "electric-field reward" term to the vdW interaction energy

$$\text{Grid\_philic} = \text{vdW\_energy} + \text{oriented-dipole\_energy}$$

where the oriented-dipole energy is necessarily negative. *Hydrophilic* regions are those within which the sum of the

two terms is sufficiently negative. These regions are depicted by contouring the hydrophilic grid at a prescribed negative isosurface value, typically −8 kcal/mol.

2. Hydrophobic Map. Conversely, the quantity representing hydrophobicity is constructed by adding an oppositely signed (positive) "electric-field penalty" term to the vdW term:

$$\text{Grid\_phobic} = \text{vdW\_energy} - 0.30 * \text{oriented-dipole\_energy}$$

3. Hydrophobic Regions. Hydrophobic regions thus are regions in which "something" would like to be (as evidenced by a favorable vdW term), but water would not (as indicated by the lack of an appreciable electric field). Such regions are initially defined as the regions within which the Grid_phobic potential is more negative than a given threshold, taken by default to be −0.75 kcal/mol. Normal practice, however, is to first modify the hydrophobic grid by removing grid points that border on too many assigned hydrophilic points, have too few hydrophobic or "inside-protein" neighbors, or have too many neighboring grid points that are neither hydrophilic nor hydrophobic in character. The hydrophobic potential at the surviving grid points is then scaled down by multiplying the Grid_phobic potential by the fraction of radial rays that intersect the protein surface within 6 Å; at the same time, the threshold for defining the hydrophobic region is reduced to −0.50 kcal/mol (default value). These modifications help SiteMap focus on regions that are well sheltered from the solvent and are bordered by few polar groups, somewhat as Glide XP does, by other means, in detecting "hydrophobic enclosure".[33]

4. Donor, Acceptor, and Metal-Binding Maps. To provide better guidance, the hydrophilic map is partitioned into separate hydrogen-bond donor and acceptor maps; a metal-binding map is also formed when a multiply charged metal cation other than Ca^{2+} is present. In carrying out the partitioning, hydrophilic grid points that lie within 3 Å of such a metal center are first assigned as metal-binding grid points. The remaining hydrophilic points are then classified by displacing the sampling point in the direction of the local electric field and recomputing the field. A donor point is assigned if this displacement increases the magnitude of the field, and an acceptor point is assigned if it decreases it. These maps are also contoured, by default, at a negative isosurface value of −8 kcal/mol.

5. Surface Map. To show the size and shape of the space available to the ligand, a map representing the SiteMap surface is generated by contouring the repulsive part of the vdW grid (attractive regions are first removed) at a positive threshold value, set to +1 kcal/mol by default.

Scoring Sites. This stage uses the site-point groups (i.e., sites) produced in the first stage and the energetic properties of the grid points produced in the second stage to evaluate the sites. A number of properties are computed, some of which are used in the final site scoring. These properties are described below.

1. Number of Site Points. As a rule of thumb, 2 to 3 site points are typically found for each atom of the bound ligand, including hydrogens. As was shown in the Results section, the size of the site is often a good indicator of the preferred binding site. The site points ultimately reported are those determined in the site-finding stage plus the "extension" points identified in computing the "exposure" property (see below) that also meet either the previously cited hydrophilic or hydrophobic threshold. For the 326 submicromolar sites specified in the Results section, the average number of site points is 132. This number is inflated to some degree by contributions from a relatively small number of large sites, but the median value, 124 site points, is not much different.

2. Exposure and Enclosure. These properties provide different measures of how open the site is to solvent. To evaluate the exposure property, "extension" site points are added on the 1-Å site-point grid. These points must lie within a given distance in *x*, *y*, and *z* from an original site point (by default, 3 Å) and must make favorable vdW contact with the receptor or lie at least 4 Å from the nearest protein atom. The value of the property is the ratio of the number of extension points to the number of original plus extension points. A shallow, open site would allow more extension points to be added than would a deep or well-encapsulated site, giving a higher exposure score. The lower the score, the better; the average for the submicromolar sites is 0.52. To evaluate the enclosure property, radial rays are drawn from the site points in 110 evenly spaced directions. The enclosure score is the fraction of rays that strike the receptor surface within a distance of 10 Å, averaged over the original site points and the extension points. Here, higher scores are better; the average for the submicromolar sites is 0.76.

3. Contact. This property measures how strongly the average site point interacts with the receptor via vdW nonbonded interactions when a probe positioned at the site point is given the nominal vdW parameters used in the site-visualization stage. The score is computed by averaging the probe-receptor vdW interaction energies over the original site points and the extension points computed for the exposure calculation. To facilitate comparison between sites, the contact score is calibrated so that the average score for the submicromolar sites is 1.0.

4. Hydrophobic and Hydrophilic Character, and Balance. The first two properties measure the hydrophobic and hydrophilic character of the site as computed by averaging the Grid_phobic or Grid_philic potential over the original site points and the extension points, while the third, balance, expresses the ratio of the two. The hydrophobic and hydrophilic scores are also calibrated so that the average score for a submicromolar site is 1.0. The average balance score, on the other hand is about 1.6, not 1.0, because the ratios computed for sites that have high hydrophobic but low hydrophilic scores make large contributions to the average.

5. Donor/Acceptor Character. This property suggests the degree to which a well-structured ligand might be expected to donate, rather than accept, hydrogen bonds, as inferred from the sizes and intensities of donor and acceptor SiteMap regions. The average for the submicromolar sites is 0.76. This value indicates a moderate preference for ligand acceptor groups and is qualitatively consistent with the relative prevalence of ligand acceptors and donors in druglike molecules.[34,35]

6. SiteScore. SiteScore, the score used to identify and to compare binding sites, is based on a weighted sum of properties described above. It was determined by optimizing the number of cases among the 538 PDBbind proteins for which the site with the best SiteScore corresponds to the cocrystallized site, as assessed by the sitemin criterion (see the Results section). A more elaborate expression had been

found in earlier studies that employed a smaller number of complexes,[22] but the PDBbind set proved to be well described by an expression that uses just three terms: (1) the square-root of the number of site points, capped at 100 site points to avoid overly rewarding large sites; (2) the enclosure score; and (3) the hydrophilic score. As it happens, the latter term is not actually needed. Rather, it was included in the initial release of the suite2008 software, with the fixed coefficient shown below, in the hope that SiteScore, so configured, could serve as an optimal function for assessing druggability as well as for identifying binding sites. We ultimately decided to retain this form for binding-site identification in the current version of SiteMap[21] but, for reasons discussed below, to introduce a separate, better performing function for classifying druggability. The SiteScore function is

$$\text{SiteScore} = 0.0733\, n^{1/2} + 0.6688\, e - 0.20\, p \qquad (3)$$

where n is the number of site points (up to 100), e is the enclosure score, and p is the hydrophilic score, capped at 1.0 (the average for the submicromolar sites) to limit the impact of hydrophilicity in charged and highly polar sites. The average SiteScore for the submicromolar sites is 1.01.

7. Druggability Score. As previously specified in eq 2, Dscore uses the same properties as SiteScore but different coefficients:

$$\text{Dscore} = 0\,.\,094\, n^{1/2} + 0.60\, e - 0.324\, p$$

The number of site points is again capped at 100, but the hydrophilic score is not capped. In our view, the use of different functions for binding-site identification and for classifying druggability is justified because these are different, and sometimes conflicting, tasks. For example, ligands that bind to the PTP1B phosphate pocket with nanomolar, and even subnanomolar, affinity are known.[36] But these highly active ligands have charge structures like those of the natural phosphate substrate and are not druglike. Our view is that SiteMap should recognize that such a site can bind ligands strongly, but should not rate it as druggable. As we have seen, the larger, uncapped hydrophilic term is one of the keys for distinguishing "difficult" and "undruggable" targets from "druggable"ones. For example, this term is as large as 2.1 for the "undruggable" sites in the Cheng set. This gives a penalty of 0.7, whereas the penalty applied in the SiteScore function is at most 0.2. On the other hand, SiteScore finds 462 of the 538 cocrystallized sites in the PDBbind set as the top-scoring site, but this number decreases to 448 when Dscore is used for site indentification. Though the reduction is not large, we chose to use separate scoring functions.

Because the "druggable" sites in the Cheng set are much more hydrophobic than are either the "difficult" or "undruggable" sites (*cf.* Table 6), one might ask whether Dscore should also include a hydrophobic term. As it happens, hydrophobicity by itself correlates fairly well with the Cheng classification, giving ranks of 1−4, 6, 7, 9, 17, 38 and 48 for the "undruggable" sites and of 5, 8, 10−15, 19, and 46 for the "difficult" sites. For comparison, Dscore gives ranks of 1−10 for the "undruggable" sites and 11−15, 17, 18, 21, 25, and 45 for the "difficult" sites (Tables 5 and 6). When 30% of the hydrophobic score is added to Dscore, the "undruggable" sites rank 1−7, 12, 15, and 19 and the "difficult" sites rank 8−11, 13, 14, 16, 17, 20, and 46; together, they occupy 19 of the first 20 positions, a result equally as good as that obtained in the MAP_{POD} approach (Table 5). However, including hydrophobicity also intermixes the "difficult" and "undruggable" sites to some degree. Whether or not to include hydrophobicity, then, depends on whether separating "druggable" sites from either "difficult" or "undruggable" sites is taken to be more important than distinguishing "difficult" from "undruggable" sites. If the former is preferred, the reported Dscore value can easily be adjusted in the manner described.

To determine whether the size of the sampled region significantly affects the Dscore rankings, we carried out additional calculations that extended the box around the ligand atoms used in the site-finding stage from the default border of 6 Å′ on a side to 12 Å. In most, but not all, cases, the larger box produced few if any additional site points, with the result that the ranks of the "undruggable" sites changed relatively little—i.e., to 1−9 and 16—while those of the "difficult" sites became 10−15, 17, 25, 27, and 44. Comparison to the ranks obtained using the default border (cf. Table 6) shows that the size of the site-finding box has only a small effect on the Dscore rankings.

8. Site Volume. The volume of a protein site is well defined when the site is fully enclosed by the protein. More commonly, however, the site is open to solvent on one or more sides. To assign the volume in such a case, what needs to be decided, proceeding outward from the protein surface, is where to stop counting. Understandably, different criteria will yield different volumes. SiteMap's approach is to approximate the "shrink-wrap" volume of the site by counting enclosed regions while excluding regions that protrude into solvent. This objective is addressed by first identifying all points on the cubic mapping grid (default grid separation, 0.7 Å) that are outside the protein surface and lie within 4 Å of any site point. Uniformly spaced radial rays are then drawn in 110 directions from the candidate volume points, and those for which fewer than 60% strike the protein surface within 8 Å are discarded. The volume of the site is computed from the number of surviving volume points and the grid-box volume.

9. Reference Distance Properties. When a ligand or other reference species is identified, "refdist", "refmin", "refavg", and "sitemin" properties are also calculated. The first of these specifies the distance between the centroid of the site-point set and that of the reference species, while the second specifies the closest approach of a site point to a reference atom. The third property measures how completely the site covers the reference species by giving the average distance of closest approach of a reference atom to a site point, and the fourth, sitemin, measures the closest approach of a reference atom to the site-point centroid. The latter is the basis for the criterion used in the Results section to determine whether or not a given site corresponds to the ligand binding site.

In some cases, a small refmin value (typically, < 1 Å) is accompanied by a much larger refdist or sitemin distance. These are cases in which the site extends significantly beyond the reference species in an asymmetric manner. To be sure, these extensions are justified in some cases. In particular, in an endoprotease they may map the channels that bind the N-terminal and C-terminal strands of the peptide undergoing cleavage. Such extensions make it more difficult to determine

the "hot spot" that should be the focus of attention in a drug-design study but can indicate regions that a tight-binding ligand could usefully probe.

CONCLUSIONS

SiteMap combines a novel and highly effective algorithm for rapid binding-site identification with easy-to-use property and visualization tools. For binding-site identification, it correctly identifies the known site as the top-scoring site in 86% of a set of 538 complexes taken from the PDBbind database. Moreover, its accuracy increases to 88% when only proteins that bind their cocrystallized ligands with submicromolar affinity are considered and to 98% when the affinity is subnanomolar. SiteMap also provides useful guidance as to whether a candidate site is likely to be a ligand-binding site. In addition, SiteMap calculates a druggabilty score that accurately accounts for the division by Cheng and co-workers of 63 sites covering 27 protein targets into "druggable", "difficult", and "undruggable" sets[16] and that provides insight into the physical basis of the classifications. For binding-site analysis, SiteMap provides a wealth of information in the form of computed properties and graphical contour maps that distinguish hydrophobic, hydrogen-bond donor, hydrogen-bond acceptor, and metal-binding regions. This information can be used in a lead-discovery application to quickly evaluate docking hits or in a lead-optimization context to suggest how a ligand structure might be modified to increase its binding affinity or to improve its physical properties. SiteMap calculations typically take 2−5 min on a 2.4 GHz Intel Pentium 4 workstation for proteins with 5000−8000 atoms, including hydrogens. Such calculations can be submitted conveniently from Schrödinger's Maestro interface or from the command line.

ACKNOWLEDGMENT

I thank Dr. Matt Repasky and Mr. Vasudeva Atukuri for developing the interfaces used for submitting SiteMap jobs from Maestro and incorporating their output. I also thank Dr. Woody Sherman for helpful discussions and for drawing attention to the need to investigate SiteMap's ability to classify the druggability of target sites, Dr. Ramy Farid for first pointing out the importance of the hydrophilic term in accounting for druggability, and Dr. Kate Holloway of the Merck Research Laboratories for providing descriptors computed using a previous version of SiteMap for most proteins of the Cheng set. I also thank Dr. Walter Knapp, one of the originators of the Karlsberg protein pKa site,[27] for suggesting the use of this site to estimate the likely cost of neutralizing the specificity-pocket aspartate in serine proteases.

REFERENCES AND NOTES

(1) Levitt, D.; Banaszak, L. POCKET: A computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graphics* **1992**, *10*, 229–243.
(2) Laskowski, R. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graphics* **1995**, *13*, 323–330.
(3) Peters, K. P.; Fauck, J.; Frommel, C. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Biol.* **1996**, *256*, 201–213.
(4) Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: automatic and efficient detection of potential small-molecule binding sites in proteins. *J. Mol. Graphics Modell.* **1997**, *15*, 359–363.
(5) Liang, J.; Edelsbrunner, H.; Woodward, C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.* **1998**, *7*, 1884–1897.
(6) Brady, G. P., Jr.; Stouten, P. F. Fast prediction and visualization of protein binding pockets with PASS. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 383–401.
(7) Blinkowski, T.; Naghibzadeh, S.; Liang, J. CASTp: computed atlas of surface topography of proteins. *Nucleic Acids Res.* **2003**, *31*, 3352–3355.
(8) Huang, B.; Schroeder, M. LIGSITE^{csc}: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol [Online]* **2006**, *6*. http://www.biomedcentral.com/1472-6807/6/19 (accessed June 11, 2007).
(9) An, J.; Totrov, M.; Abagyan, R. Comprehensive identification and classification of ligand-binding envelopes. *Mol. Cell Proteomics* **2005**, *4*, 752–761.
(10) 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.
(11) Dennis, S.; Kortvelyesi, T.; Vajda, S. Computational mapping identifies the binding sites of organic solvents on proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 4290–4295.
(12) Kortvelyesi, T.; Silberstein, M.; Dennis, S.; Vajda, S. Improved mapping of protein binding sites. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 173–186.
(13) Ruppert, J.; Welch, W.; Jain, A. Automatic identification and representation of protein binding sites for molecular docking. *Protein Sci.* **1997**, *6*, 524–533.
(14) Verdonk, M. L.; Cole, J. C.; Watson, P.; Gillet, V.; Willett, P. SuperStar: improved knowledge-based interaction fields for protein binding sites. *J. Mol. Biol.* **2001**, *307*, 841–859, 2001.
(15) Bliznyuk, A.; Gready, J. Simple method for locating possible ligand binding sites on protein surfaces. *J. Comput. Chem.* **1999**, *9*, 983–988.
(16) Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C.; Huang, E. S. Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* **2007**, *25*, 71–75.
(17) Coleman, R. G.; Salzberg, A.; Cheng, A. C. Structure-based identification of small molecule binding sites using a free-energy model. *J. Chem. Inf. Model.* **2006**, *46*, 2631–2637.
(18) Joughin, B. A.; Tidor, B.; Yaffe, M. B. A computational method for the analysis and prediction of protein:phosphopeptide-binding sites. *Protein Sci.* **2005**, *14*, 131–139.
(19) Brown, D.; Superti-Furga, G. Rediscovering the sweet spot in drug discovery. *Drug Discovery Today* **2003**, *8*, 1067–1077.
(20) Hopkins, A. L.; Groom, C. R. The druggable genome. *Nat. Rev. Drug Discovery* **2002**, *1*, 727–730.
(21) *SiteMap, version 2.2 (2008 update release)*; Schrödinger, LLC: Portland, OR, 2008. *Maestro, version 8.5*; Schrödinger, LLC: Portland, OR, 2008. *QikProp*, *version 2.5*; Schrödinger, LLC: Portland, OR, 2006. *OPLS_2001*; Schrödinger, LLC: Portland, OR, 2008.
(22) Halgren, T. New method for fast and accurate binding-site identification and analysis. *Chem. Biol. Drug Des.* **2007**, *69*, 146–148.
(23) Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind database: methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111–4119. *PDBbind database, version 2004*; University of Michigan: Ann Arbor, MI, 2005.
(24) Wermuth, C. G. *The Practice of Medicinal Chemistry*, 2nd ed.; Academic Press: London, U.K. and San Diego, CA, 2003.
(25) Roehrig, S.; Straub, A.; Pohlman, J.; Lampe, T.; Pernerstorfer, J.; Schlemmer, K. H.; Reinemer, P.; Perzborn, E. Discovery of the novel antithrombotic agent 5-chloro-N-({(5S)-2-oxo-3-[4-(3-oxomorpholin-4-yl)phenyl]-1,3-oxazolidin-5-yl}methyl)thiophene-2-carboxamide (BAY 59-7939): an oral, direct factor Xa inhibitor. *J. Med. Chem.* **2005**, *48*, 5900–5908. The crystallographic structure has been solved but does not appear to have been deposited. However, a factor Xa structure with a similar group in the S1 pocket has been deposited as PDB refcode 1nfw.
(26) Pinto, D. J. P.; Orwat, M. J.; Koch, S.; Rossi, K. A.; Alexander, R. S.; Smallwood, A.; Wong, P. C.; Rendina, A. R.; Luettgen, J. M.; Knabb, R. M.; He, K.; Xin, B.; Wexler, R. R.; Lam, P. Y. Discovery of 1-(4-Methoxyphenyl)-7-oxo-6-(4-(2-oxopiperidin-1-yl)phenyl)-4,5,6,7-tetrahydro-1H-pyrazolo[3,4-c]pyridine-3-carboxamide (Apixaban, BMS-562247), a highly potent, selective, efficacious, and orally bioavailable inhibitor of blood coagulation factor Xa. *J. Med. Chem.* **2007**, *50*, 5339–5356. The crystallographic structure is deposited as PDB refcode 2p16.
(27) Karlsberg. http://agknapp.chemie.fu-berlin.de/karlsberg (accessed Aug 23, 2008). Kieseritzky, G.; Knapp, E. W. Optimizing pKa computation in proteins with pH adapted conformations. *Proteins: Struct., Funct., Bioinf.* **2008**, *71*, 1335–1348.

(28) Burgey, C. S.; Robinson, K. A.; Lyle, T. A.; Sanderson, P. E.; Lewis, S. D.; Lucas, B. J.; Krueger, J. A.; Singh, R.; Miller-Stein, C.; White, R. B.; Wong, B.; Lyle, E. A.; Williams, P. D.; Coburn, C. A.; Dorsey, B. D.; Barrow, J. C.; Stranieri, M. T.; Holahan, M. A.; Sitko, G. R.; Cook, J. J.; McMasters, D. R.; McDonough, C. M.; Sanders, W. M.; Wallace, A. A.; Clayton, F. C.; Bohn, D.; Leonard, Y. M.; Detwiler, T. J., Jr.; Lynch, J. J., Jr.; Yan, Y.; Chen, Z.; Kuo, L.; Gardell, S. J.; Shafer, J. A.; Vacca, J. P. Metabolism-directed optimization of 3-aminopyrazinone acetamide thrombin inhibitors. Development of an orally bioavailable series containing P1 and P3 pyridines. *J. Med. Chem.* **2003**, *46*, 461–473. Crystallographic structures are deposited as PDB refcodes 1mu6, 1mu8, and 1mue.

(29) Weber, A.; Halgren, T. A. Design and synthesis of P2-P1′-linked macrocyclic human renin inhibitors. *J. Med. Chem.* **1991**, *34*, 2692–2701.

(30) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.

(31) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shaw, D. E.; Shelley, M.; Perry, J. K.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.

(32) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47*, 1750–1759.

(33) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra precision Glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.* **2006**, *49*, 6177–6196.

(34) 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.* **1997**, *23*, 3–25.

(35) 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. Model.* **2001**, *41*, 1308–1315.

(36) Therien, M.; Skorey, K.; Zamboni, R.; Li, C. S.; Lau, C. K.; LeRiche, T.; Truong, V. L.; Waddleton, D.; Ramachandran, C. Synthesis of a novel peptidic photoaffinity probe for the PTP-1B enzyme. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 2319–2322.

CI800324M