Structure-Based Identification of Small Molecule Binding Sites Using a Free Energy Model

Ryan G. Coleman,[†,‡,||] Anna C. Salzberg,[§,⊥] and Alan C. Cheng*[,†,§,#]

Research Technology Center, Pfizer Global Research & Development, Cambridge, Massachusetts 02139, Tufts University, Medford, Massachusetts 02139, and Brandeis University, Waltham, Massachusetts 02454

Received June 7, 2006

We separately have shown that the maximal druglike affinity of a given binding site on a protein can be calculated on the basis of the binding-site structure alone by using a desolvation-based free energy model along with the notion that druglike ligands fall into certain physiochemical property ranges. Here, we present an approach where we reformulate the calculated druggability affinity as an additive free energy to facilitate the searching of whole protein surfaces for druglike binding sites. The highest-scoring patches in many cases represent known ligand-binding sites for druggable targets, but not for difficult targets. This approach differs from other approaches in that it does not simply identify pockets with the greatest volume but instead identifies pockets that are likely to be amenable to druglike small-molecule binding. Combining the method with a functional residue prediction method called SCA (statistical coupling analysis) results in the prediction of potentially druggable allosteric binding sites on p38α kinase.

INTRODUCTION

The computational discovery of potential ligand-binding sites on protein surfaces is useful in generating hypotheses for new druggable binding sites that can subsequently be experimentally tested through virtual and high-throughput screening or, potentially, de novo drug design. For instance, Smrcka et al. recently discovered selective small-molecule modulators of heterotrimeric G-proteins through the elucidation of a peptide-binding "hotspot" followed by computational docking of a small-molecule library.[1] In this work, we describe a structure-based computational approach for finding druglike small-molecule binding sites on protein surfaces. Conceptually, we search the protein surface for "druggable" hotspots using the maximal affinity prediction for a passively absorbed oral drug (MAP_{POD}) scoring approach that we report elsewhere[2] and describe in more detail below. We enumerate all reasonable surface patches on the protein structure and then score every patch using a "local" version of the MAP_{POD} druggability equation.

Binding-site prediction methods generally either find the largest pockets or use molecular probes to identify hotspots on the protein. Examples of the former include SURFNET, LIGSITE, CASTP, and PASS.[3−6] The recent SURFNET-ConSurf combination method uses sequence conservation among homologues to improve binding-site prediction from SURFNET alone.[7] Examples of probe-based methods include the multiple-copy simultaneous search (MCSS), computational solvent mapping, and Q-SiteFinder methods.[8−10] The original MCSS method involved the energy minimization of pools of functional group probes where probes do not interact with each other. The minimized probes are then clustered to identify probe hotspots that can potentially aid drug design.[11] The computational solvent mapping method is similar to MCSS but uses a multiple-step approach that is reported to place probes more frequently in known binding sites.[9] The recent Q-SiteFinder approach places methyl probes on a grid that envelops the protein and uses the clustering of energetically favorable probe positions to determine the potential binding sites. Q-SiteFinder differs from MCSS in its use of a discrete three-dimensional grid with 0.9 Å spacing, which leads to faster calculations on the order of 10−15 s per protein.

Currently available binding-site prediction approaches attempt to find ligand binding pockets, but not all ligand-binding pockets are capable of binding a druglike small molecule with reasonable affinity nor are they all biologically relevant. Here, we present a method that begins to address these issues directly. We address the first issue by finding pockets that are most likely to bind druglike ligands. Our approach is based on our previous work where we estimate from the binding-site structure the maximal druglike affinity, or MAP_{POD}, using a model that estimates the desolvation of hydrophobic patches taking into account the topology-dependent nature of desolvation.[2,12,13] The essential concept is that deep, hydrophobic pockets desolvate more easily, and this concept can be quantitatively expressed and used to make a computational estimation of the maximal affinity of a given pocket for a "druglike" ligand. "Druglike" small molecules have minimal formal charges, polar surface areas less than 140 $Å^2$, and molecular weights less than 500 Da.[14−16] Why does a desolvation-based approach estimate so well the variation in maximal druglike affinity? One likely contributing reason is that many of the other energetic terms are roughly constant for an ideal druglike ligand. For instance,

* Corresponding author tel.: (617) 444-5411; e-mail: alan.cheng@amgen.com.
† Pfizer Global Research & Development.
‡ Tufts University.
§ Brandeis University.
|| Currently at Genomics and Computational Biology Graduate Group, University of Pennsylvania School of Medicine, Philadelphia, PA 19104.
⊥ Currently at Pfizer Global R&D, Cambridge, MA.
Currently at Amgen, Cambridge, MA.

10.1021/ci600229z CCC: $33.50
Published on Web 11/08/2006

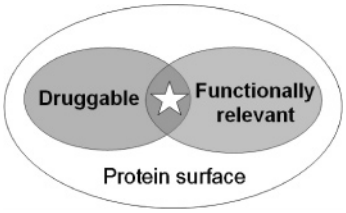

Figure 1. Overview of approach for finding druggable, functionally relevant binding sites on the protein surface. The star represents the overlap of residues identified by the druggable binding-site search and functionally relevant residue analysis methods.

the van der Waals interaction energy is roughly constant because of a maximum druglike molecular size of around 500 Da, and the rotational/translational entropy terms are roughly constant because they also scale with the molecular weight. The fact that drugs tend to have minimal formal charges also limits the electrostatic interaction contribution. These are qualitative arguments that do not cover all of the terms, and further investigation is warranted. In any case, the druglike affinity approach we present below appears to be appropriate for finding "druggable" pockets suitable for small-molecule modulation.

To address the second issue of finding pockets that are biologically relevant, we combine our MAP_{POD} search approach with a functional residue prediction method, the statistical coupling analysis (SCA), developed by Lockless and Ranganathan.[17] The method uses large sequence alignments to find evolutionarily conserved coupling of residues and is especially interesting in that it successfully finds coupled residues that are distant from each other.[18] The combined approach is conceptually depicted in Figure 1, and we demonstrate this combined approach on the p38α protein kinase.

METHODS

Surface Representation. Protein structures downloaded from the Protein Databank (PDB) are used without modification. All heteroatoms, including all waters and cofactors, are ignored. Hydrogens are generally not used in the calculation of surface areas, and we follow this precedence here.

An analytic representation of the macromolecular surface is generated as we previously reported[19] and involves first constructing the three-dimensional weighted Delaunay tessellation of the biomolecule and then subtracting the α-shape complex, as implemented by Koehl and Edelsbrunner in POCKET[20] and first described by Liang and co-workers.[5] We modified POCKET to include eight dummy atoms at extreme points around the protein in order to include atoms on the convex hull of the protein in generating the Delaunay tesselation. Atomic radii values used are the standard radii found in NACCESS. An example biomolecular surface tessellation is shown in Figure 2. Arcs representing boundaries between atoms are then calculated to allow the reconstruction of the surface. The solvent-accessible surface is a union of sphere sections, while the molecular surface additionally includes torus and sphere sections that map the re-entrant surface. The areas for both surfaces are calculated analytically as a sum of the areas of the sphere and torus sections. To calculate curvature, we use our previously reported approach of using geometric inversion to find the least-squares-fitted sphere to the molecular surface and taking the radius of the sphere as the radius of curvature.[19]

A connected representation of the surface that we call the α-skin is used to enable computational searching of the protein surface and finding of the surface patch with the highest MAP_{POD} affinity. The α-skin, depicted in Figure 2A, is related to α shapes and is composed of *nodes*, *edges*, and *triples*. The *nodes* carry information on the local MAP_{POD} affinity for a particular surface piece. Because a particular atom may be a part of more than one disconnected surface patch, we defined individual nodes in our search graph as mapping to individual connected surface patches derived from the same atom. Thus, a single atom may define more than one node. The *edges* carry information on which surface pieces are adjacent to facilitate traversal of the surface. An edge is present whenever two nodes have a surface boundary consisting of a single arc. Pairs of nodes can have multiple edges between them if they have multiple arcs defining their boundaries. The *triples* represent areas where three atomic spheres meet and are indicated in Figure 2A as blue triangles because triples have exactly three edges. Triples, like edges, carry information on which surface pieces are adjacent and facilitate traversal of the surface. All edges are involved in triples, except for a few rare edges such as those shown in Figure 2A in red. Construction of the α-skin, including the Delaunay tessellation and α-shape computation, requires 1−2 min for each protein in our data set.

Local Formulation of MAP_{POD} Equation. We elsewhere have described the derivation and application of the MAP_{POD}-computed ligand-binding affinities for defined binding sites.[2] Here, we reformulate the maximal druglike affinity formula as a piece-wise additive free energy so that the MAP_{POD} contribution for each node can be precomputed, and the graph can be searched exhaustively and rapidly. The "global" equation[2] is

$$\Delta G_{\mathrm{MAP_{POD}}} \approx -\gamma(r)\, A^{\mathrm{target}}_{\mathrm{nonpolar}} \frac{A^{\mathrm{target}}_{\mathrm{druglike}}}{A^{\mathrm{target}}_{\mathrm{total}}} + C; \quad \gamma(r) = \frac{\gamma(\infty)}{1 - \frac{1.4}{r}}$$

where $\gamma(r)$ is related to the solvent surface tension, $\gamma(\infty)$ is the desolvation of a flat surface, r is the radius of curvature for the binding site, $A^{\mathrm{target}}_{\mathrm{total}}$ is the total solvent-accessible surface area (SASA) of the protein surface involved in binding, $A^{\mathrm{target}}_{\mathrm{nonpolar}}$ is the nonpolar fraction of the SASA involved in binding, $A^{\mathrm{target}}_{\mathrm{druglike}}$ is set at 300 Å^2 and represents a binding surface corresponding to a druglike compound of approximately 500 Da in molecular weight, and C is set to zero. We use the Wesson and Eisenberg adjusted atomic solvation parameters (ASPs)[21] to define carbon atoms as nonpolar (positive ASP values) and oxygen, nitrogen, and sulfur atoms as polar and favoring solvation (negative ASP values). Assuming that the free energies we calculate are additive,[22] we can derive a "local" version of the equation:

$$\Delta G_{\mathrm{MAP_{POD}}} \approx -\frac{A^{\mathrm{target}}_{\mathrm{druglike}}}{A^{\mathrm{target}}_{\mathrm{total}}} \cdot \sum_{\mathrm{atom} \in \mathrm{HPOAtoms}} [\gamma(r_{\mathrm{atom}}) A_{\mathrm{atom}}]$$

where surface areas (A_{atom} terms) and curvatures (r_{atom} terms) are computed for each hydrophobic atom node (i.e., each atomic surface piece), and C is not shown because it is zero.

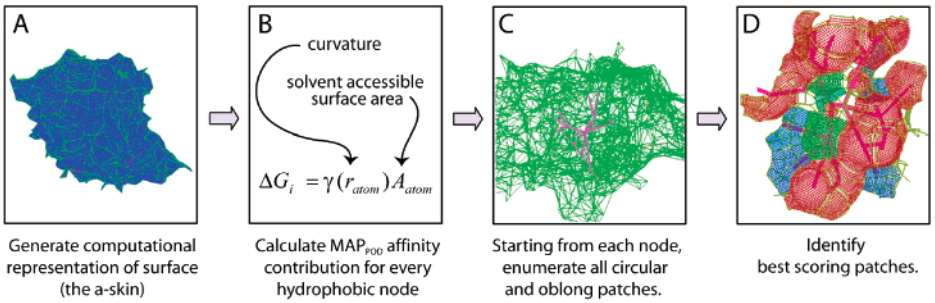

Figure 2. Structure-based binding-site search method. (A) Generate α-skin. Edges are shown in light gray and triangles shown in black. All edges are formed by three spheres with the exception of edges indicated in red, which are formed by the overlap of two spheres. (B) Calculate local affinity scores. (C) Using a shortest path tree, generate all circular, elliptical, and oblong patches starting from each surface piece (node). The bold lines in the figure indicate the traversed edges that define the set of nodes in a circular patch. (D) Record additive MAP_{POD} scores, and return the best-scoring patches. The example shows the underlying traversed edges in purple, and surfaces are colored red for convex surfaces, green for approximately planar surfaces, and blue for convex surfaces. The actual curvature is a continuous quantity calculated as previously described.[19] Only nonpolar surfaces are colored; polar surfaces are not colored because the curvature is not calculated as they do not contribute to the MAP_{POD} score.

The local surface area is computed as described above, while the curvature requires the definition of a local region. The curvature approach we use is a simplified representation of desolvation that appears to capture the role of topology.[13] Because the organization of waters is affected by the surrounding surface topology in addition to the individual surface piece, we use the *continuous* surface within 3.0 Å of the center of the surface piece in calculating the curvature. For the studies shown in this paper, we used a $\gamma(\infty) = 47.5$ cal/mol/Å^2 and required $\gamma(r)$ to be in the range of 15−140 cal/mol/Å^2 in order to limit affinity contributions from any given node. Values below the minimum or above the maximum were set to their respective endpoints. The values of 15 and 140 are arbitrary and represent values that are about a third and three times the value of $\gamma(\infty)$, respectively. Values below and above those values appear to be physically implausible relative to our $\gamma(\infty)$ value. The lower cutoff is below the lowest values we have seen reported in the literature and has a negligible effect on the results. The upper cutoff is in practice necessary to limit rare cases where we find a very large energetic contribution from a single local surface patch due to the local surface being extremely narrow with a low radius of curvature. These cases appear to be artifacts from the use of a rigid solid surface, which are amplified because of the use of the small surface patches. For proteins in our data set, precalculation of the surface areas and the $\Delta G_{MAP_{POD}}$ contribution for all nodes requires 1−2 h on a system running Linux on a 3.2 GHz Pentium 4 processor.

Searching the Surface. To generate roughly circular or elliptical patches on the biomolecular surface, we used a heuristic approach along with a well-known graph theoretic data structure, the shortest path tree. We find the approach is fast and generates reasonable surface patches in practice. To generate roughly circular patches, connected edges from the starting node of the α-skin are used to compute the shortest path tree. The length of an edge is the distance between the atom centers that generate the two nodes. Each time a node is added, the surface area is added to a running total, and patches are enumerated once the surface area reaches the minimum of 300 Å^2. Patch enumeration of the node is complete when the surface area reaches the maximum of 400 Å^2. These areas represent minimum and maximum sizes of druggable patches as determined by studying the correlation of small-molecule molecular weights with buried surface areas that we calculate. In particular, we show in Figure 3 that the druglike ligand with a maximal weight of 550 Da corresponds roughly to the 300−400 Å^2 surface area range. To generate more irregular patches, two or more neighboring nodes can be initialized to have distances of zero. The use of two starting nodes results in roughly elliptical patches, and the use of three starting nodes results in oblong or boomerang patches. In practice, the use of one, two, or three nodes generates binding-site patches within approximately 2, 10, and 120 min, respectively, on a Linux system with a 3.2 GHz Pentium 4 processor. Each patch is scored using the additive MAP_{POD} formulation, and patches with the lowest scores (i.e., maximal favorable affinity) are reported.

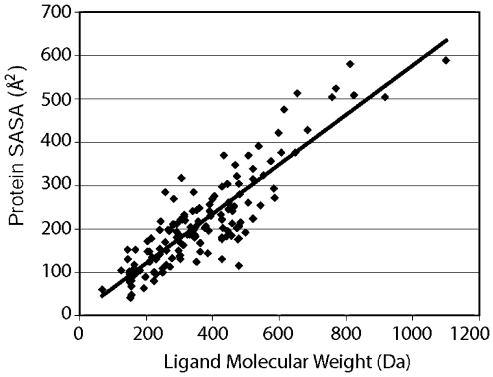

Figure 3. Ligand molecular weight correlates with protein binding pocket surface area. A 550 Da ligand corresponds to about 300−400 Å^2 on the protein surface. Data are measured from a large, diverse set of 305 protein−ligand cocrystal structures from CCDC/Astex.[35] The correlation is $r^2 = 0.77$.

Statistical Coupling Analysis. SCA[17,18] Matlab scripts were provided courtesy of R. Ranganathan (communication to A.C.S.), and all analyses are performed using the scripts provided. The S_TKc seed sequence alignment and protein sequences with the S_TKc domain are downloaded from the SMART resource.[23,24] The seed alignment is used to generate a Ser/Thr kinase domain hidden markov model (HMM) using HMMER 2.3.2.[25] The SMART Ser/Thr protein kinase sequences are then aligned using the HMM. Sequences are removed if they contain no label or the label includes the keyword "theoretical". This results in a sequence alignment

Table 1. Proteins Used in Testing Binding-Site Search Approach[a]

protein	PDB ID	one node	two node	three node
PDE 4D (apo)	1f0j	1	5	1
PDE 4D	1oyn	1	1	1
PDE 4D	1ptw	5	3	3
PDE 4D	1q9m	5	17	149
PDE 4D	1rko	1	1	1
PDE 4D	1y2e	1	1	1
PDE 5	1udt	1	1	1
PDE 5	1udu	87	150	> 250
aldose reductase	1eko	8	27	>250
streptavidin	1stp	1	1	1
streptavidin	2izg	1	1	1
Cox 2	4cox	24	22	1
Alcohol DH	1bto	1	1	1
EGFR kinase (apo)	1m14	170	>250	>250
EGFR kinase	1m17	1	1	1
HMG-CoA reductase	1hw8	>250	>250	>250
HMG-CoA reductase	1hw9	>250	>250	>250
HMG-CoA reductase	1hwi	>250	>250	>250
HMG-CoA reductase	1hwj	>250	>250	>250
HMG-CoA reductase	1hwk	>250	>250	>250
HMG-CoA reductase	1hwl	>250	>250	>250
PTP-1B (open conf)	1pty	7	1	3
PTP-1B (open conf)	1onz	52	167	>250
PTP-1B (closed conf)	1nny	19	29	>250
PTP-1B (closed conf)	1q1m	1	1	1
C-abl kinase	1fpu	1	1	1
C-abl kinase	1iep	3	7	1

[a] Enzyme names are given along with the PDB ID for the structure used. Apo structures are indicated; all other proteins are cocomplex structures with either a small-molecule or natural substrate. Searches were performed using one, two, and three node searches, and the ranking of the known inhibitory binding site is given, and values are colored in gray if the known binding site is not one of the top 10 patches. Further details are found in the main text.

of 3170 identified Ser/Thr kinase domains, which then forms the basis for the prediction of functional residues by the SCA method. The full alignment is available in the Supporting Information.

RESULTS AND DISCUSSION

We have developed a method of searching for druglike binding sites on the basis of a maximal druglike affinity approach previously shown to be useful in estimating druggability for a single binding site.[2] To do this, proteins are represented using our previously reported approach[19] along with a new computational geometric construct we call the α-skin. The α-skin allows for the analytic representation of the molecular surface as atomic sections that are aware of their neighboring atomic sections and, thus, facilitates rapid graph-based searching for a patch that optimizes the MAP_{POD} druggability score. To enable the searching of protein surfaces within a reasonable time frame, the MAP_{POD} equation was reformulated as a localized equation. Further details are provided in the Methods section.

To assess the correlation of the local MAP_{POD} formulation with the original global formulation, we computed scores for 65 000 patches on 27 proteins (see Table 1 and Figure 4). Scores from the two formulations have a correlation of $r^2 = 0.73$ and a Spearman rank correlation of $\rho^2 = 0.70$. Manual inspection of a subset of patches suggests that the differences are largely due to (1) the global formulation being less sensitive to local topological variations in curvature because it uses a single curvature and (2) the presence of convex surfaces in local patches, which is again averaged out in the global formulation. We did not expect the quantitative correlation to be perfect, and in practice, we find the local formulation correlates less well with known druggabilities (data not shown). Nevertheless, the local formulation allows for rapid searching and has decent correlation with our validated global formulation and, thus, should be useful in identifying potential druglike binding sites.

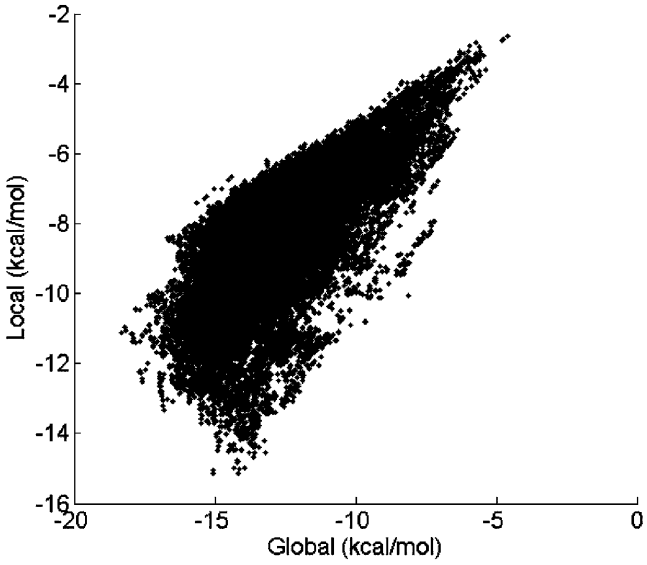

Figure 4. Scatter plot of values from local additive MAP_{POD} and global MAP_{POD} methods for the 25 proteins listed in Table 1. Plot generated using Matlab 7.1 (Mathworks, Natick, MA).

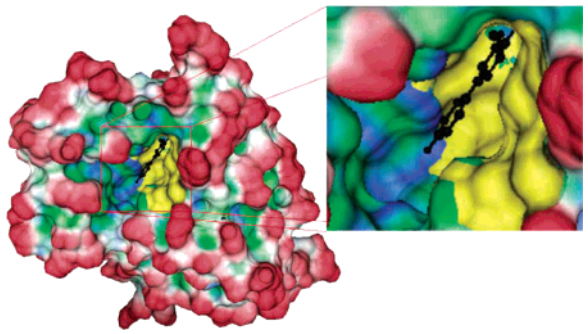

Figure 5. Highest affinity patch found on the surface of PDE-4D (PDB ID: 1y2e) using circular patches generated from single nodes. The patch is highlighted in yellow.

Binding-Site Searching. We applied the method to search a protein, PDE4D (PDB ID: 1Y2E), using circular patches. To our delight, the top-scoring patch found, shown in Figure 4, is not only inside the known active site of PDE-4D but also represents the common inhibitor binding region of compounds such as rolipram, roflumilast, and piclamilast.[26] In particular, the patch does not include the metal binding sites on the lower-left-hand corner of the binding site in Figure 5 because it is relatively polar and is avoided by most drugs in clinical trials. The MAP_{POD} approach attempts to find maximum druglike affinity sites where no strong electrostatics, such as metal binding, are involved. Drugs that chelate metals to achieve potency can be more difficult to design selectivity for, in part, because metals are physiologically ubiquitous, and the drug may have strong interactions with metals in other enzyme-binding sites. None of the

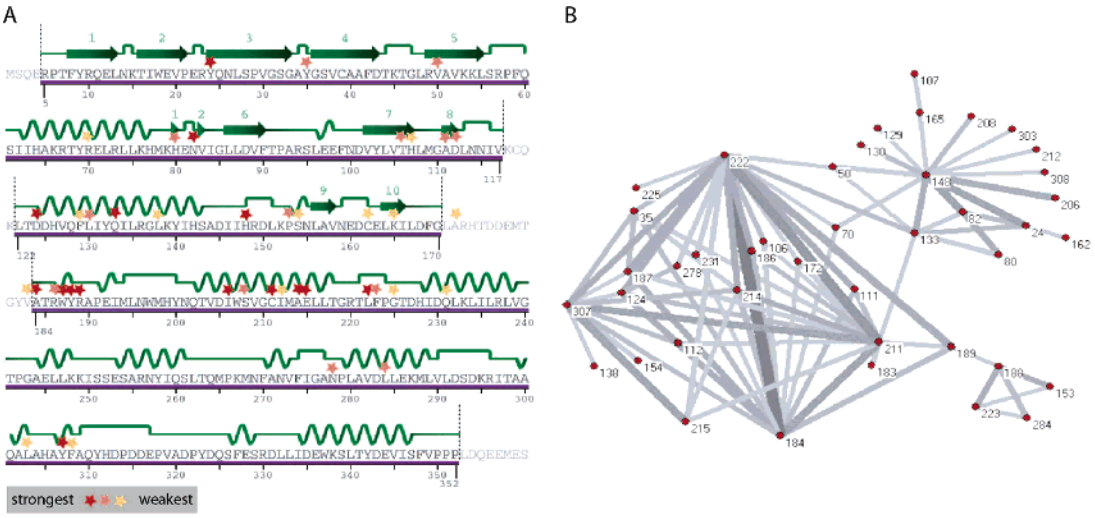

Figure 6. SCA analysis of Ser/Thr protein kinases. The residue numbering is based on p38α (PDB ID: 1kv1). (A) Annotation of identified coupled residues superposed onto a depiction of the sequence and secondary structure for p38α. Stars indicate coupled residues and are colored red, orange, or yellow to indicate the strength of coupling. The figure is taken from the PDB 1kv1 entry of the RCSB resource,[36] and the secondary structure is automatically assigned using Stride.[37] (B) Network of statistically coupled residues, where thicker and darker lines indicate stronger statistical coupling. Residue numbering is for p38α. The figure is produced using SCA scripts (see Methods) and the Pajek software.[38]

PDE-4 and PDE-5 inhibitors on the market or having entered phase II or phase III clinical trials directly interact with the binding-site metals.[26−28] The patchiness of the predicted site (see Figure 5 insert) is the result of the method selecting predominantly hydrophobic surfaces and avoiding hydrophilic surfaces. The top-scoring patches also include the lip of the binding site, which is hydrophobic but is difficult to design for. The top five scoring patches also fall within the same region. Because surface patches are generated starting from every node, patches inevitably are found multiple times, and there is significant redundancy in the identified patches. We find it effective to identify the top-scoring surface patches (we use the top 250) and aggregate patches that have overlapping residues.

We next applied the method to the larger set of proteins shown in Table 1. The use of only circular patches (one node) finds the known drug binding site as the top-scoring site for three of the five PDE-4D structures and finds the known drug-binding site within the top five scoring sites for all cases. We define successful identification of a binding site by the location of greater than 50% of the residues within 5 Å of the bound ligand. For the apo structure (PDB ID: 1F0J), the known ligands define the known binding site. The method scores poorly for HMG-CoA reductase because the binding site itself scores as marginal for druglike ligands,[2] and known drugs have a high polar surface area (>150 Å^2) and molecular weight. If we remove HMG-CoA reductase structures and look at the 20 remaining structures for nine different proteins, we find that the use of circular patches finds the binding site as one of the top five scoring patches in 13 of the cases. The use of oval patches (two nodes) finds the binding site in 12 of the cases, and the use of oblong patches (three nodes) finds the binding site in 14 of the cases. In the case of the two apo proteins listed in Table 1, the approach successfully finds the binding site for PDE-4D, while it does not for EGFR. The failure in the case of EGFR appears to be due to (1) changes in conformation between the apo and bound forms (PDB IDs are 1M14 and 1M17, respectively), which results in a closing down of the pocket, leading to a smaller curvature value in the bound form, and (2) the movement of the sulfur of Cys751 from being part of the apo pocket to being away from the pocket in the bound structure. The cysteine sulfur here is considered a polar residue on the basis of work by Wesson and Eisenberg[21] on ASPs. However, the ASPs are continuous values, and the use of a binary binning results in sulfurs being binned with the somewhat more polar oxygen and nitrogen atoms. The incorporation of continuous values may help. The EGFR example highlights limitations not only due to conformational change but also due to the discrete parametrization of polar and nonpolar atoms. While the method cannot find all binding sites, the results suggest it is nevertheless useful for hypothesis generation.

Identifying Functionally Relevant Binding Sites. While the method presented here locates potential drug-binding sites, binding is necessary but not sufficient for therapeutic modulation; the binding site must additionally be functionally relevant. To identify binding sites that are both capable of binding a druglike compound and functionally relevant, we combine our binding-site search method with the statistical coupling analysis method for identifying allosteric coupled residues.[17,18] Residues identified by both methods are potential druggable allosteric sites. We use a p38α kinase structure[29] to demonstrate this approach. SCA using a sequence alignment of 3170 identified kinases from the SMART online resource[23] finds 42 residues that are significantly coupled to at least one other residue, where significance is defined by the coupling strength being greater than three standard deviations from the normal distribution mean.[17] For the sequence alignment we used, the cutoff criteria is $\Delta\Delta G \geq 1.14kT$*. In Figure 6A, these coupled residues are indicated superposed onto the sequence of p38α. In Figure 6B, the network of connected residues is depicted with stronger couplings indicated by thicker, darker lines.

The MAP$_{POD}$ binding-site search method identifies four patches on the p38α structure. The patch at the top right of p38α as depicted in Figure 7 is defined by residues 59−63 and 331−337 (numbering based on the 1KV1 PDB struc-

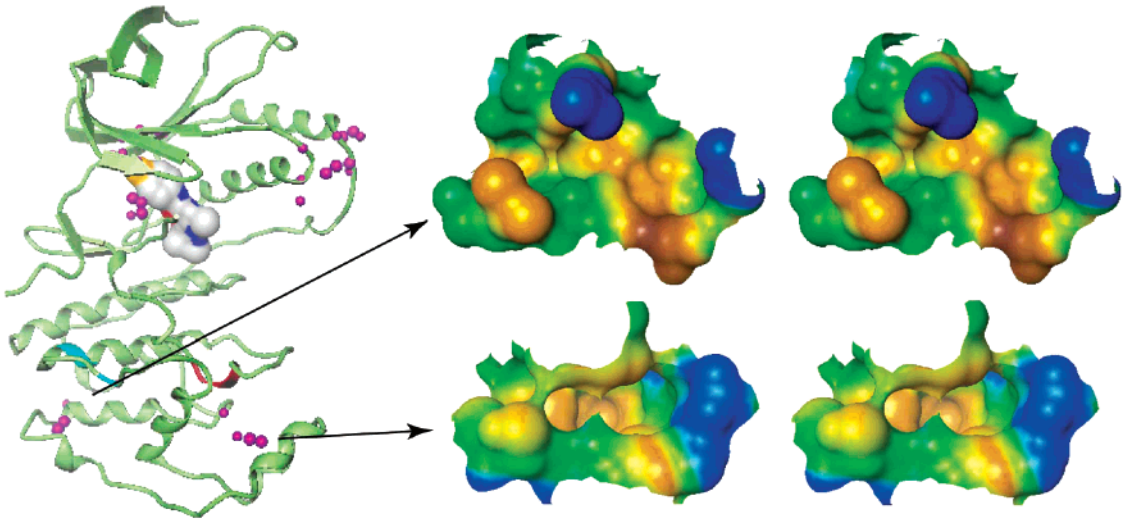

Figure 7. Predicted druggable binding sites. Purple spheres on the protein ribbon diagram approximately indicate identified binding sites on the basis of searching the protein surface, where the sphere positions are the average coordinates of residues involved in the binding site. The two predicted druggable, allosteric ligand binding sites in p38α protein kinase are shown on the right side as stereoview surface depictions. The top site involves statistically coupled p38α residues Ala214, Glu215, Leu222, and Phe223 (ribbon colored blue). The bottom site involves the highly conserved APE motif (ribbon colored red). Binding-site surfaces are colored by hydrophobicity, where brown and yellow indicate hydrophobic regions, green indicates polar regions, and blue indicates charged regions.

ture). This pocket appears to be hydrophobic but lacks a distinct pocket. While parts of the surface are concave, the whole surface itself is convex, suggesting that the additive curvature approach is not effective at finding patches with a convex "global" curvature. In practice, these convex patches can be triaged manually. The cluster in the middle is the ATP binding site. The bottom right patch is defined by residues 191−192, 195, 197, 200−201, 232, 236, 242, 245−246, 249−250, 259, and 291−294 and is adjacent to the conserved APE motif. The bottom left patch is defined by residues 217−218, 222, 238, and 272−278 and is adjacent to coupled residues 214−215 and 222−223. This patch is rather shallow, as can be seen from the stereoview depiction provided in Figure 7.

On the other hand, the surface patch adjacent to the APE motif is deep and enclosed and appears to be sufficiently hydrophobic and enclosed to bind a druglike small molecule. The patch, however, does not in fact touch any of the residues identified by SCA, although the residues lining the pocket are adjacent in sequence to coupled residues 185−189. A ligand in the binding site can conceivably affect these residues. Interestingly, however, the pocket does contact the APE motif. Residues such as those in the APE motif are so highly conserved that SCA analysis is not possible, because it is not possible to find alignment subsets that satisfy the SCA alignment acceptance criteria.[17] In addition to the APE motif, highly conserved sequence motifs include the K53 and E71 residues, catalytic RD motif (residues 149−150), and the DFG motif (residues 168−170). Although these residue positions are not detectable by the computational SCA approach, they are almost always important functional residues because substantial evolutionary pressure holds them fixed. Thus, inhibitors using this pocket are potentially allosteric. It is unclear why p38α has a pocket at this position, although it is interesting that an imidazolyl-pyridine p38α inhibitor was found to bind in this pocket in another p38α crystal structure.[30] The APE motif itself is thought to play primarily a structural role.[31]

The bottom left patch is the same region involved in myristic acid binding in C-Abl, where a myristolated peptide corresponding to the C-Abl N-terminal region is known to stabilize the inactive conformation of C-Abl.[32] Furthermore, Adrian et al.[33] recently reported a small-molecule inhibitor that appears to utilize this site to allosterically inhibit Bcr-abl-dependent proliferation in cultured cells at an IC_{50} of 138 nM. The equivalent binding site in Abl structures, however, is significantly deeper and more hydrophobic than that seen in p38α structures. Nevertheless, others have recently reported that similar binding sites can be identified on many kinase structures and that these patches are especially hydrophobic and have the potential of binding small molecules.[34] Thus, it is an intriguing possibility that this site is capable of small-molecule allosteric modulation.

The MAP_{POD} search approach does have drawbacks. The method identifies a surface patch composed of residues 59−64 and 332−335 (cluster at the top of protein structure in Figure 7), which does not contain a cleft to facilitate specific binding of small molecules. This particular patch involves a crystal contact, which may explain its relatively high exposure of nonpolar surface area. An additional problem with many methods that rely on static crystal structure snapshots is that proteins undergo significant dynamics that can reveal additional binding sites. Our search method is rapid enough to be applied to a few hundred snapshots from molecular dynamics simulations, and this combination may be useful in generating hypotheses for allosteric binding sites. Applying the method to multiple crystal structures instead of a single structure can also increase the completeness of this approach.

CONCLUSIONS

We have described an approach to finding small molecule binding sites on biomolecules using an additive version of our previously reported free energy equation[2] that essentially quantifies the idea that deep, hydrophobic pockets are more

likely to bind druglike small molecules. We demonstrate the approach on known drug targets and find that the method is largely successful in finding known druggable binding sites but, at least from our limited study, is only partially successful in finding binding sites using apo proteins. Nevertheless, the method can be used to generate hypotheses from the known structural conformations. We therefore applied this method to the prediction of allosteric pockets on p38α kinase. Because binding is necessary but not sufficient for allosteric small-molecule modulation of protein function, we combined our MAP_{POD} search approach with results from the SCA approach, which has been shown to predict allosteric residues. The combined approach results in the prediction of two allosteric binding sites on p38α. De novo design of small-molecule modulators to these sites is difficult and will likely proceed only after significant experimental validation studies. We note, however, that one of the sites is similar to an allosteric small-molecule drug site recently reported for Bcr-abl.

ACKNOWLEDGMENT

We thank Beth Lunney, Xin Huang, Diane Souvaine, and Pfizer Research Technology Center colleagues for helpful discussions, and R. Ranganathan for providing SCA scripts.

Supporting Information Available: Kinase sequence alignment used for SCA analysis available in PFAM sequence alignment format (txt file). This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Bonacci, T. M.; Mathews, J. L.; Yuan, C.; Lehmann, D. M.; Malik, S.; Wu, D.; Font, J. L.; Bidlack, J. M.; Smrcka, A. V. Differential Targeting of Gbetagamma-Subunit Signaling with Small Molecules. *Science* **2006**, *312*, 443−446.

(2) Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D.; Salzberg, A.; Huang, E. S. Structure-Based Maximal Affinity Model Predicts Small-Molecule Druggability. Submitted for publication.

(3) Laskowski, R. A. SURFNET: A Program for Visualizing Molecular Surfaces, Cavities, and Intermolecular Interactions. *J. Mol. Graphics* **1995**, *13*, 323−330.

(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.; Fu, P.; Sudhakar, P. V.; Subramaniam, S. Analytical Shape Computation of Macromolecules: I. Molecular Area and Volume through Alpha Shape. *Proteins* **1998**, *33*, 1−17.

(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) Glaser, F.; Morris, R. J.; Najmanovich, R. J.; Laskowski, R. A.; Thornton, J. M. A Method for Localizing Ligand Binding Pockets in Protein Structures. *Proteins* **2006**, *62*, 479−488.

(8) Miranker, A.; Karplus, M. Functionality Maps of Binding Sites: A Multiple Copy Simultaneous Search Method. *Proteins* **1991**, *11*, 29−34.

(9) Silberstein, M.; Dennis, S.; Brown, L.; Kortvelyesi, T.; Clodfelter, K.; Vajda, S. Identification of Substrate Binding Sites in Enzymes by Computational Solvent Mapping. *J. Mol. Biol.* **2003**, *332*, 1095−1113.

(10) Laurie, A. T.; Jackson, R. M. Q-SiteFinder: An Energy-Based Method for the Prediction of Protein−Ligand Binding Sites. *Bioinformatics* **2005**, *21*, 1908−16.

(11) Caflisch, A.; Miranker, A.; Karplus, M. Multiple Copy Simultaneous Search and Construction of Ligands in Binding Sites: Application to Inhibitors of HIV-1 Aspartic Proteinase. *J. Med. Chem.* **1993**, *36*, 2142−2167.

(12) Sharp, K. A.; Nicholls, A.; Fine, R. F.; Honig, B. Reconciling the Magnitude of the Microscopic and Macroscopic Hydrophobic Effects. *Science* **1991**, *252*, 106−109.

(13) Southall, N. T.; Dill, K. A. The Mechanism of Hydrophobic Solvation Depends on Solute Radius. *J. Phys. Chem. B* **2000**, *104*, 1326−1331.

(14) 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.

(15) Palm, K.; Stenberg, P.; Luthman, K.; Artursson, P. Polar Molecular Surface Properties Predict the Intestinal Absorption of Drugs in Humans. *Pharm. Res.* **1997**, *14*, 568−571.

(16) Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties that Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45*, 2615−2623.

(17) Lockless, S. W.; Ranganathan, R. Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. *Science* **1999**, *286*, 295−299.

(18) Suel, G. M.; Lockless, S. W.; Wall, M. A.; Ranganathan, R. Evolutionarily Conserved Networks of Residues Mediate Allosteric Communication in Proteins. *Nat. Struct. Biol.* **2003**, *10*, 59−69.

(19) 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.

(20) Edelsbrunner, H.; Koehl, P. The Weighted-Volume Derivative of a Space-Filling Diagram. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 2203−2208.

(21) Wesson, L.; Eisenberg, D. Atomic Solvation Parameters Applied to Molecular Dynamics of Proteins in Solution. *Protein Sci.* **1992**, *1*, 227−235.

(22) Dill, K. A. Additivity Principles in Biochemistry. *J. Biol. Chem.* **1997**, *272*, 701−704.

(23) Schultz, J.; Milpetz, F.; Bork, P.; Ponting, C. P. SMART, a Simple Molecular Architecture Research Tool: Identification of Signaling Domains. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 5857−5864.

(24) Letunic, I.; Copley, R. R.; Schmidt, S.; Ciccarelli, F. D.; Doerks, T.; Schultz, J.; Ponting, C. P.; Bork, P. SMART 4.0: Towards Genomic Data Integration. *Nucleic Acids Res.* **2004**, *32*, D142−4.

(25) Eddy, S. R. Profile Hidden Markov Models. *Bioinformatics* **1998**, *14*, 755−763.

(26) Huai, Q.; Wang, H.; Sun, Y.; Kim, H. Y.; Liu, Y.; Ke, H. Three-Dimensional Structures of PDE4D in Complex with Roliprams and Implication on Inhibitor Selectivity. *Structure* **2003**, *11*, 865−873.

(27) Sung, B. J.; Hwang, K. Y.; Jeon, Y. H.; Lee, J. I.; Heo, Y. S.; Kim, J. H.; Moon, J.; Yoon J. M.; Hyun, Y. L.; Kim, E.; Eum, S. J.; Park, S. Y.; Lee, J. O.; Lee, T. G.; Ro, S.; Cho, J. M. Structure of the Catalytic Domain of Human Phosphodiesterase 5 with Bound Drug Molecules. *Nature* **2003**, *425*, 98−102.

(28) Menniti, F. S.; Faraci, W. S.; Schmidt, C. J. Phosphodiesterases in the CNS: Targets for Drug Development. *Nat. Rev. Drug Discovery* **2006**, *5*, 660−670.

(29) Pargellis, C.; Tong, L.; Churchill, L.; Cirillo, P. F.; Gilmore, T.; Graham, A. G.; Grob, P. M.; Hickey, E. R.; Moss, N.; Pav, S.; Regan, J. Inhibition of p38 MAP Kinase by Utilizing a Novel Allosteric Binding Site. *Nat. Struct. Biol.* **2002**, *9*, 268−272.

(30) Tong, L.; Pav, S.; White, D. M.; Rogers, S.; Crane, K. M.; Cywin, C. L.; Brown, M. L.; Pargellis, C. A. A Highly Specific Inhibitor of Human p38 MAP Kinase Binds in the ATP Pocket. *Nat. Struct. Biol.* **1997**, *4*, 311−316.

(31) Kannan, N.; Neuwald, A. F. Evolutionary Constraints Associated with Functional Specificity of the CMGC Protein Kinases MAPK, CDK, GSK, SRPK, DYRK, and CK2. *Protein Sci.* **2004**, *13*, 2059−2077.

(32) Nagar, B.; Hantschel, O.; Young, M. A.; Scheffzek, K.; Veach, D.; Bornmann, W.; Clarkson, B.; Superti-Furga, G.; Kuriyan, J. Structural Basis for the Autoinhibition of c-Abl Tyrosine Kinase. *Cell* **2003**, *112*, 859−71.

(33) Adrian, F. J.; Ding, Q.; Sim, T.; Velentza, A.; Sloan, C.; Liu, Y.; Zhang, G.; Hur, W.; Ding, S.; Manley, P.; Mestan, J.; Fabbro, D.; Gray, N. S. Allosteric Inhibitors of Bcr-abl-Dependent Cell Proliferation. *Nat. Chem. Biol.* **2006** *2*, 95−102.

(34) Milo, M.; Goldblum, A. A. Myristoyl Binding Site in Protein Kinases. 3D-SIG Laptop Session Abstract. *Conference on Intelligent Systems for Molecular Biology*, Fortaleza, Brazil, August 6−10, 2006.

(35) Nissink, J. W.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A new test set for validating predictions of protein−ligand interaction. *Proteins* **2002**, *49*, 457−471.

(36) 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.

(37) Frishman, D.; Argos, P. Knowledge-based Protein Secondary Structure Assignment. *Proteins* **1995**, *23*, 566−579.

(38) Batagelj, V.; Mrvar, A. Pajek − Analysis and Visualization of Large Networks. In *Graph Drawing Software*; Jünger, M., Mutzel, P., Eds.; Springer, Berlin, 2003; pp 77−103.

CI600229Z