High-Throughput Virtual Screening of Proteins Using GRID Molecular Interaction Fields

Simone Sciabola,*[,†] Robert V. Stanton,[†] James E. Mills,[‡] Maria M. Flocco,[‡] Massimo Baroni,[§] Gabriele Cruciani,[||] Francesca Perruccio,[⊥] and Jonathan S. Mason[#]

Pfizer Research Technology Center, Cambridge, Massachusetts 02139, Pfizer Global Research and Development, Ramsgate Road, Kent CT13 9NJ, Sandwich, United Kingdom, Molecular Discovery Limited, 215 Marsh Road, HA5 5NE, Pinner, Middlesex, United Kingdom, Laboratory of Chemometrics, University of Perugia, Via Elce di Sotto, 10 I-60123, Perugia, Italy, Syngenta, Schaffhauserstrasse, 4332 Stein AG, Switzerland, Lundbeck A/S, Ottiliavej 9, DK-25000, Copenhagen, Denmark

Received September 1, 2009

A new computational algorithm for protein binding sites characterization and comparison has been developed, which uses a common reference framework of the projected ligand-space four-point pharmacophore fingerprints, includes cavity shape, and can be used with diverse proteins as no structural alignment is required. Protein binding sites are first described using GRID molecular interaction fields (GRID-MIFs), and the FLAP (fingerprints for ligands and proteins) method is then used to encode and compare this information. The discriminating power of the algorithm and its applicability for large-scale protein analysis was validated by analyzing various scenarios: clustering of kinase protein families in a relevant manner, predicting ligand activity across related targets, and protein−protein virtual screening. In all cases the results showed the effectiveness of the GRID-FLAP method and its potential use in applications such as identifying selectivity targets and tools/hits for new targets via the identification of other proteins with pharmacophorically similar binding sites.

INTRODUCTION

There have been a number of revolutions in molecular biology in the past few decades, one of the most recent being completion of immense effort to sequence the human genome.[1,2] However, the genome itself only represents the first layer of complexity. Biological function is not carried out by the static genome but mainly by the dynamic population of proteins determined by the interplay of gene and protein regulation with extracellular influences. Even the composition of mature proteins cannot be predicted from the genome sequence alone. There are numerous instances of differential splicing and post-translational protein modifications (e.g., phosphorylation, glycosylation, ubiquitination, and methylation) that can govern the behavior of proteins. For these reasons, there is increasing interest in the field of proteomics,[3] or the large-scale study of proteins as a complement to genomics and functional genomics. Indeed, by means of proteomics and powerful bioinformatics tools, there is hope that gene variants which contribute to multifactorial diseases or genes in nonhuman infectious agents can be identified.

As a consequence of these advances, a large number of targets suitable for drug intervention are being revealed. X-ray crystallography and NMR spectroscopy can be used on amenable targets to obtain a comprehensive three-dimensional (3D) structural view of the protein structures. Increasingly, this allows access to a large number of experimentally resolved protein structures, even before their functions are known, together with homology-based models of further protein targets, recent examples being human G-protein coupled receptor (GPCR) structures.

The availability of the 3D structures of a very large number of protein targets leads to a critical and ever increasing need for methods to analyze and compare proteins based on descriptors relevant to their function and interactions with drugs and other ligands, such as the projected pharmacophoric ligand space of binding sites. Nevertheless, inferring the biological function of proteins remains a challenging problem, given that it strongly depends on the in vivo biological context (e.g., localization, post-translational modifications) in which the protein is found. An ability to identify similar proteins in terms of binding site similarity is a first step.

Protein structural information can traditionally be divided into three levels of knowledge: amino acids sequence, backbone structure, and local arrangement of atoms. The earliest comparative algorithms are based on sequence information (FASTA,[4,5] BLAST[6]) and routines implemented into publicly available databases (SWISS-PROT,[7] PROSITE,[8] or OWL[9]). Although these tools provide efficient ways to extract similar sequences from databases containing millions of entries and to correlate them with biological functions using sequential patterns, they reach their limits when comparing sequences with a low degree of homology. Indeed, while high sequence similarity usually correlates with

* Corresponding author phone: 617-551-3327; fax: 617-551-3117; e-mail: simone.sciabola@pfizer.com.

[†] Pfizer Research Technology Center.

[‡] Pfizer Global Research and Development.

[§] Molecular Discovery Limited.

[||] University of Perugia.

[⊥] Syngenta.

[#] Lundbeck A/S.

10.1021/ci9003317 © 2010 American Chemical Society
Published on Web 11/17/2009

high structure similarity and similar biological function, the opposite that low sequence similarity corresponds to structural dissimilarity is not necessarily given.[10]

The need for better algorithms has been partially addressed by moving from sequence similarity searches to comparison-based methods which infer protein similarity in terms of the overall three-dimensional (3D) fold. Structural comparison is superior to sequence analysis, since it takes into account the spatial geometric structure of the molecules involved in the recognition event and not solely the amino acid composition and order in the primary sequence.

Computationally these algorithms can be divided into three essential parts: (1) an adequate molecular representation of the proteins being compared, (2) a rigid transformation (rotation and translation) for spatial superposition, and (3) an efficient comparison algorithm. To make the similarity analysis fast, often an approximate representation of protein structure is used, based, for example, on C^{α}-atom coordinates.

A wide variety of algorithms have been developed, from distance matrix methods[11,12] and geometric hashing techniques[13,14] to genetic algorithms.[15] Although remarkable improvements over sequence-only techniques have been achieved, these methods usually do not provide a biologically relevant alignment when there is no fold similarity between the aligned structures. Moreover, it has been shown that proteins with the same fold, like triosephosphateisomerase (TIM) barrels, can have multiple functions.[16]

Thus, a third level of inferring protein similarities has been devised, allowing comparison to be focused on smaller subregions. The key aspect of such an approach is that proteins are assumed to perform similar functions if they share similar binding patterns and recognize similar binding partners. The programs TESS[17] (and its updated version JESS[18]) and ASSAM[12] use geometric hashing and clique detection, respectively, to retrieve templates of predefined 3D amino acid patterns. Park et al.[19] presented a sequence alignment like binding site comparison using a clique detection algorithm to identify sequence substitution patterns of important residues in evolutionary-related binding sites. Jones et al.[20] reviewed a wide range of computational methods for recognition of functional sites in proteins.

Important contributions to this area were also published by Schmitt et al. (Cavbase)[21] and Shulman-Peleg et al. (SiteEngine).[22,23] They defined generic pseudo-centers that efficiently encode the pharmacophoric/physicochemical properties important for molecular interactions such as hydrogen-bond donor, hydrogen-bond acceptor, mixed donor/acceptor, hydrophobic aliphatic, and aromatic (π) contacts. Each amino acid residue of a protein is represented as a set of such centers. While in Cavbase a clique detection algorithm is used to retrieve cavities that are similar to a specified query, SiteEngine uses a heuristic algorithm based on efficient hashing and matching of triangle centers of physicochemical properties. In a related study,[24] the pseudo-centers concept was applied to analyze the properties of protein−protein interfaces at the level of physicochemical interactions to reveal functionally important patterns of interactions conserved throughout evolution.

Jambon et al.[25] described SuMo, an innovative bioinformatics method to detect similar 3D sites in proteins. They applied two levels of structure representation: the first based on local atomic density computed for each atom and used as discriminative estimation of the burial of a given atom and the second based on the concept of chemical groups defined for each amino acid and consisting of a predefined set of atoms. The chemical groups are then used to build triangles. The final representation of the protein is obtained by connecting adjacent triangles to make a graph in which each triangle forms a vertex. Finally, the comparison of two proteins is obtained by using a heuristic algorithm which first generates a graph representation of pairs of similar triangles coming from each of the two molecules and then identifies and scores common patches made by independent subsets of pairs of similar triangles that are geometrically consistent. Kupas et al.[26] successfully described a new method for large-scale analysis of protein binding sites WaveGeoMap which uses both shape and physicochemical descriptors of local patches of the solvent-accessible surface to characterize substructures in protein binding sites. Overall, the algorithm was capable of detecting the differences between the active sites correctly in all the three data sets where it was applied.

In the present paper, we describe an application of FLAP (fingerprints for ligands and proteins)[27,28] for comparing and clustering proteins. In FLAP a protein is represented and compared with other proteins by its ligand "image", which contains the "hotspots" for binding of key ligand functional groups, such as a hydrogen-bond acceptor (carbonyl group), hydrogen-bond donor (amide N−H), lipophilic/hydrophobic, and charged chemical probes (nitrogen in basic amines, oxygen in carboxylate anions). Differing from previously reported methods for protein similarity and clustering,[21,22,25] this method is distinctive in that it only uses this "ligand-space" image of the protein derived from GRID molecular interaction fields (MIFs),[29,30] including, but not focused on, the shape of the binding site, not the sequence or specific amino acids, and very importantly does not use or require a structural alignment of the proteins (which can introduce significant bias and restrictions).

The 4-point pharmacophores that are generated from the MIF "hotspots" in this ligand image of the protein binding site give a common frame of reference for comparing different proteins and, as described earlier,[28] for comparing ligands to protein structures, with applications such as de novo design, virtual screening/docking, and structure-based combinatorial library design. The approach used here thus is of high interest (particularly when considering issues of selectivity/cross-reactivity/off-target binding, function, etc.) in that it characterizes proteins by how a ligand would "see" them.

A key differentiating factor of the FLAP approach is that it brings together into a common frame of reference ligands and proteins using a descriptor (GRID MIFs) that has already been found to be an effective method to analyze proteins and drive ligand−protein interactions (docking etc.) and ligand similarity studies.[27,28] The GRID energetic survey of a binding site is not a protein-focused theoretical approach but is based on a force field that has been derived principally from an analysis of protein−ligand X-ray complexes. The utility of the approach to protein−protein comparison and virtual screening is shown in this paper, with its ability to group proteins into binding site relevant clusters and identify proteins with similar function and ligands, including those with very different sequences. An additional goal of this

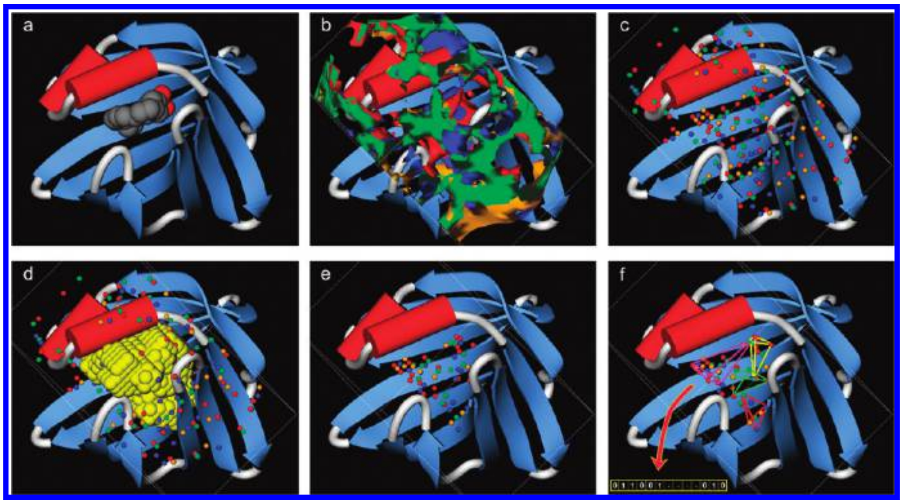

Figure 1. Protein binding site characterization workflow. (a) Protein−ligand X-ray starting structure. (b) Molecular interaction fields (MIFs) obtained by running GRID with selected probes. (c) Information contained in the GRID MIFs is analyzed and condensed into few energy minima points called MINI. (d and e) Cavity constraint derived from CAVGEN is applied to filter out all the nonrelevant MINI points. (f) MINI which survived to the filtering procedure are then exhaustively combined into a pharmacophore key representative of the given protein binding site.

study was to further validate the GRID-FLAP method as previously described for docking and molecular similarity (e.g., for scaffold hopping, etc.) by showing that it is a descriptor relevant for describing and comparing protein binding sites.

A representative set of case studies is presented in the final section to demonstrate the effectiveness of the algorithm, in particular with respect to its applicability for data mining and large-scale analysis of protein binding sites.

THEORY

The procedure developed here is divided into the following steps: (1) automatic detection and extraction of potential binding sites from proteins, (2) characterization of these sites into pharmacophoric/physicochemical properties by use of GRID MIFs, (3) generation of 3D pharmacophore fingerprints of the binding sites, and (4) data analysis.

1. Cavity Definition. Since protein recognition and function usually occur in well-characterized clefts or cavities of protein surfaces (Figure 1a), the analysis can be focused by identifying and comparing only this portion of protein space. In the past decade, several computational biology tools have been developed to locate depressions on protein surfaces as potential binding sites using various strategies.

SiteID[31] identifies protein pockets by solvating the structure to locate regions where solvent spheres tend to cluster. SuperStar[32] outputs cavities as part of the hydrogen-bond map generation, using a flood-fill technique[33] starting from a point defined by the user. SURFNET[34] also uses a flood-fill technique to generate molecular surfaces and gaps between surfaces, considering all relevant pairs of atoms in turn, and placing a sphere midway between each pair, shrinking its size until no clashes will occur with any neighboring atoms and keeping it only if the final sphere radius will be above some minimum threshold. Other cavity-detection algorithms, so-called grid-based approaches, POCKET,[35] LIGSITE,[36] and LIGSITE^{csc}[37] localize pocket regions in the protein structure by placing a trial sphere at points on a three-dimensional grid.

Fully geometric approaches have also been used to automatically locate protein pockets, such as APROPOS[38] and CASTp,[39] based mainly on the α-shape algorithm implemented by Edelsbrunner et al.[40] The α-shape algorithm describes protein surfaces as lists of adjacent triangles. Depending on the α parameter, a more or less detailed description of molecule shape can be achieved. For $\alpha \rightarrow \infty$ (crude shape) the convex hall representation is obtained, while for $\alpha = 0$ one obtains the set of points itself; α values between these two extremes give a more or less accurate description of the form of the set of points (fine shape). Comparison of these geometrical representations allows location of cavities on the solvent-accessible surface, taking into account only those cavities where a significant deviation between crude and fine shape is present.

Another pocket detection method is PocketPicker.[41] This algorithm operates on a regular rectangular grid and employs a sophisticated scanning process to locate protein surface depressions. The scanning procedure comprises the calculation of “buriedness” of probe points installed in the grid to determine their atom environment. The buriedness of grid points is interpreted as a pocket accessibility index. When the pocket detection routine implemented in PocketPicker was compared to results achieved with the existing methods CAST, LIGSITE, LIGSITE^{csc}, and SURFNET, the success rates of PocketPicker were comparable to those of LIGSITE^{csc} and outperformed the other tools.

The Pfizer in-house approach implemented in our procedure, CAVGEN,[42] is based on the SURFNET algorithm, but modifications have been made to improve the speed and performance of the program. Each cavity is described by a set of spheres. For each possible pair of atoms in the protein a sphere is generated by placing its center at the midpoint of the two atoms, then shrinking it until it no longer clashes with the van der Waals surface generated by any atoms in the protein. The new sphere is included in the set of spheres defining the active site only if its radius remains above 1 Å. The intention of using midpoints of atom pairs as the sphere centers is that this prevents the cavity going beyond the surface of the protein. Simply picking random points on a grid containing the protein would give rise to one continuous surface all round the protein.

CAVGEN addresses a number of issues not covered by SURFNET: (1) Using every possible pair of atoms to generate spheres is computationally intensive and also provides redundancy because many of the spheres have significant overlap. CAVGEN only generates a fraction of the maximum possible number of spheres, obtaining the same coverage of space with lower density and therefore much more quickly. (2) Spheres are split up into discrete clusters, each representing a different cavity. (3) Surface grooves are not identified by SURFNET in known examples of surface binders because they are too shallow for the algorithm to identify. If required, CAVGEN accentuates these clefts by using midpoints of surface points rather than of atoms as the sphere centers, effectively raising the "sea level". (4) The last step consists of computing the volume and accessibility of each potential active site, allowing the user to select either the preferable ligand binding cavity (cavities can also be selected according to their proximity to a ligand or protein residues) or all the potential cavities from a given protein and store them in the in-house database of protein binding sites.

2. Cavity Mapping. To compute potential ligand interaction similarities for cavities across a large sample set, GRID[29,42] was used to generate isopotential energy surfaces for all the binding sites under investigation. An energy-based clustering procedure was then applied to condense the GRID MIF information into a reduced number of energy minima points representing locations at which the selected probe might be able to make strong (energetically favorable) nonbonded interactions.

GRID is a computational procedure for detecting energetically favorable binding regions in proteins and small molecule drugs of known 3D structure. The energies are calculated using the electrostatic, hydrogen-bond, Lennard−Jones, and entropic interactions of chemically selective probes with the chosen biological target. The method of images[29] is used by GRID in order to account for the dielectric influence of solvent water, since incorrectly modulated electrostatic interactions would give misleading results in the absence of explicit water molecules. The program works by defining a three-dimensional grid of points that contains the chosen substrate binding site (Figure 1a and 1b). At each node of the grid, the energy between the probe and the target is calculated as indicated in eq 1

$$E_{xyz} = \sum E_{\mathrm{EL}} + \sum E_{\mathrm{HB}} + \sum E_{\mathrm{LJ}} + S \qquad (1)$$

where E_{EL} is the appropriately modulated electrostatic energy, E_{HB} is the hydrogen-bonding energy, E_{LJ} is the Lennard−Jones potential energy, and S is the entropic contribution. Each individual term in the summations relates to one pairwise interaction between the probe at position xyz and a single atom of the protein, and the summations extend over all protein atoms. The same calculation is repeated for each node in the three-dimensional grid and for each probe being considered. The results of these calculations are a collection of three-dimensional matrices, one for each probe−target interaction. A detailed description of the GRID program, the force field parameters, and details of calculations can be found elsewhere.[29,30,43]

In order to use GRID sequentially over a large set of protein cavities, the implementation of an automated MIF generation procedure was essential, the problem being the localization and definition of the GRID three-dimensional cage for multiple nonaligned binding sites. This was accomplished by using the precomputed cavity information calculated by CAVGEN. For each protein in the database, the corresponding information about its binding site coordinates and size is extracted and used to define the GRID cage where the calculation is run. After defining the cage, several GRID runs using different probes (N1, hydrogen-bond donors; O, hydrogen-bond acceptors; N^+/O^-, positive and negative charge centers; DRY, hydrophobic centers) are carried out. The information present in the MIF is automatically analyzed and condensed to a few energy minima points called MINI, showing the best interaction energy between the probe and cavity-flanking residues of the receptor. A MINI point, for a given energetic threshold, is defined as a point inside the three-dimensional cage, which is completely surrounded by points at which the GRID energy is greater (more positive). Each MINI will be located at the center of a block of 3 × 3 × 3 grid points, making 27 points in the whole block. The neighboring 26 points will have greater energies than the central MINI point. In this way, the likelihood of such a point lying on the outer surface of the grid is reduced because an outside point will not have 26 neighbors.

However, due to the way in which the GRID MIFs are generated, not all points located inside this three-dimensional grid are relevant for further calculations. Part of them in fact might lie in regions of the cage where substantial interactions cannot be made with any amino acids in the active site (Figure 1c). Therefore, in order to focus only on those interactions clearly present within the binding site, the cavity constraint (Figure 1d) information from the CAVGEN database of protein binding sites was again used in order to filter out all the MINI points outside the binding site (Figure 1e).

3. Cavity Fingerprint. A 3D pharmacophore is defined by the critical arrangement of molecular features forming a necessary but not sufficient condition for biological activity.[44] In theory a three-dimensional pharmacophore fingerprint could be derived either for a small molecule or for a protein binding site.[45–47] In a small molecule, a pharmacophore can be defined as the atoms or groups which may have critical interactions with a receptor, whereas in a macromolecule a pharmacophore can be defined as the combined set of all the complementary "hotspots" (for ligand-relevant features)[48] located within the macromolecule active site (calculated taking into account the environment of all the amino acid residues in the site rather than being assigned on a residue by residue basis).

FLAP[28] was used to generate the binary pharmacophore fingerprints for the studied binding sites. FLAP is a computational procedure to explore the pharmacophore space of ligands and proteins. The potential 4-point 3D pharmacophores expressed by ligands and/or receptors are calculated taking conformational flexibility and molecular or receptor shape into account. With 4-point pharmacophores, chirality is calculated, with a significant increase in the amount of

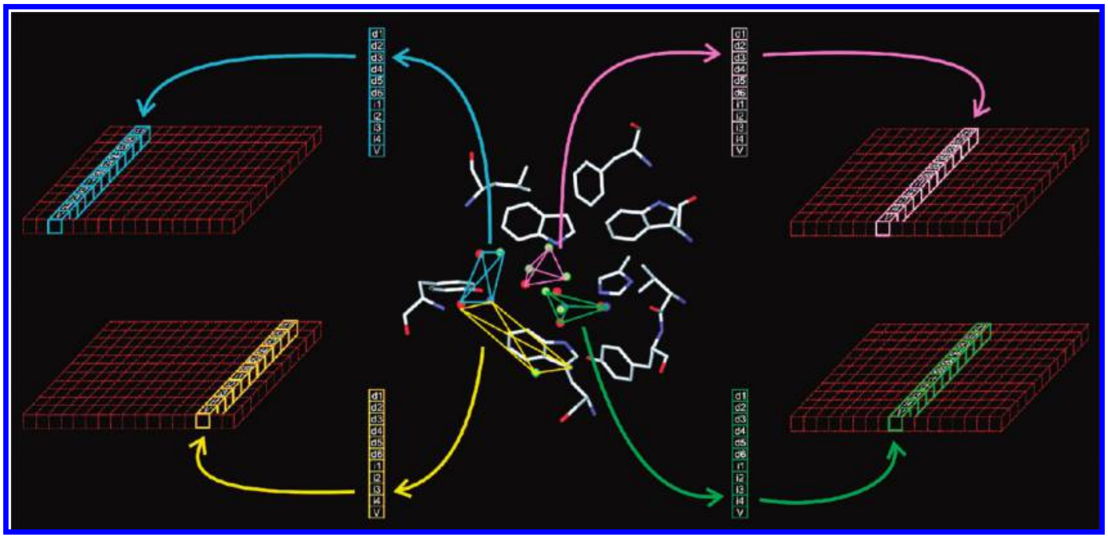

Figure 2. Data structure used by FLAP to encode the 4-point pharmacophores combinations into the receptor bitstring. Every tetrahedral configuration in FLAP is described as an array of consecutive packets of 11 integer values containing the six distances (*d*1, *d*2, *d*3, *d*4, *d*5, *d*6), the four coordinates of the corresponding vertices (*i*1, *i*2, *i*3, *i*4) and a value (*V*) proportional to the sum of the energy of the 4 points.

information, as a fundamental requirement for ligand−receptor recognition. The method FLAP uses to encode the binary bitstring represents a new technology where the receptor bitstring exists only “conceptually”.[28] Unlike previously reported approaches[49,50] storing receptor information in a 4-point pharmacophore fingerprint, FLAP uses a diversified data structure for the chirality and physicochemical/pharmacophoric properties of the points.

An array of variable length is therefore generated for each of these possible combinations of chemical features describing the cavity-flanking residues. This array is composed of consecutive packets of 11 integer values (limited to the range 0−255 as only one byte is associated to each value), with each integer value corresponding to both a determined quadruplet of points of a certain type and the vertices of a solid possessing a determined chiral configuration.

The 11 fields are assigned so they contain the six values of the distances (*d*1, ..., *d*6), the indices of the 4 site-points (*id*1, ..., *id*4, through which all the required information is accessed by referring to another data structure), and a value by default proportional to the sum of the energy of the 4 points. There is no need to introduce information concerning type and chirality because the array is specific to a certain type of data sequence of atoms/site points. Therefore, the vector is associated with the 4 given points of a certain type and a tetrahedron of a determined volume (see Figure 2). In the end, millions of potential 4-points pharmacophores are thus considered for protein binding sites together with up to some hundred thousand for each ligand.

Using FLAP, the pharmacophore key for protein binding sites can be generated by exhaustively combining GRID MINI site points. The use of MINI points as features describing protein binding sites should be compared to other well-known approaches such as Cavbase,[21] implemented in Relibase[51−54] and SiteEngine.[22] In their work,[21,22] excellent results have been achieved condensing the amino acid physicochemical/pharmacophoric properties into a restricted sets of generic pseudo-centers corresponding to five essential properties for molecular recognition (DO, hydrogen-bond donor; AC, acceptor; DA, donor−acceptor; AL, hydrophobic aliphatic; PI, aromatic contact). The location of these pseudocenters comes directly from the atomic coordinates of the amino acid atoms in the binding site, derived according to empirical rules described in detail in their papers. However, in our approach, the location of the MINI site points used to describe the cavity are not derived from the positions of the amino acid atoms, but they explore the “mirror” (complementary) space of the cavity-flanking residues, the so-called “ligand space” (Figure 3). Therefore, GRID-derived MINI site points represent protein hot spots for ligand atoms on the corresponding MIF maps. Their positions define locations at which the selected ligand atoms might be able to make strong nonbonded interactions with the amino acids atoms in the cavity (Figure 3a), thus representing pharmacophoric features in proteins with a common frame of reference with pharmacophoric feature points in the ligands (Figure 3b).

4. Data Analysis. Two different approaches were exploited to analyze the pharmacophore fingerprints for the studied binding sites. The first approach consists of comparing the fingerprint for every binding site to all the other fingerprints in the protein data set, the same as in a clustering analysis, in order to discover possible structural patterns, commonalities, and specificity trends (Figure 4a). A second approach is also used consisting of starting from the fingerprint representation of a given binding site (query) and then using this information to rank a corporate database of protein binding sites, accordingly to their degree of similarity with the starting query (Figure 4b).

RESULTS

In this section we analyze the experimental results obtained by applying our procedure in several different case studies using different protein structure data sets downloaded from the Protein Data Bank (PDB) Web site.[55,56]

Chorismate Case Study: Different Sequence Does Not Mean Different Fold. The example discussed first has been studied by Rosen et al.[10] using sparse critical points[57] derived from the Connolly surface of manually obtained binding pockets and by Schmitt et al.[21] using Cavbase. They

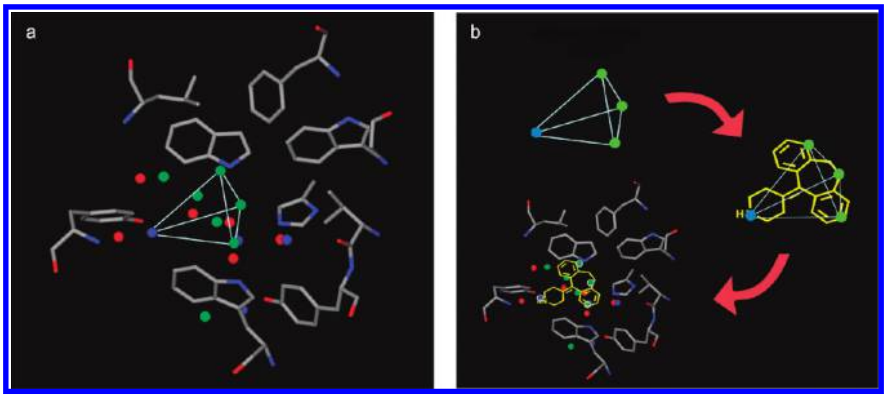

Figure 3. Ligand space concept in FLAP. MINI points in the binding site define configurations (a) at which corresponding arrangements of atoms in a ligand (b) can make strong and favorable interactions with the amino acids in the protein active site.

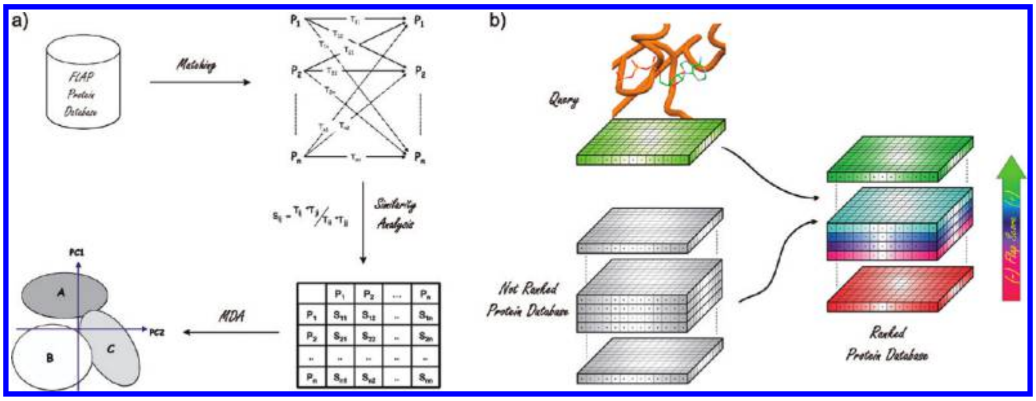

Figure 4. FLAP binding site analysis. (a) "ALL against ALL" approach where every protein active site is compared to itself and to all the others in the data set, giving rise to the corresponding similarity matrix table which can then be analyzed by means of multivariate statistical techniques. (b) Protein virtual screening approach where FLAP is used to rank every protein in the database accordingly to its degree of similarity with the query active site.

studied the binding pockets extracted from two Chorismate mutases originating from *Saccharomyces cerevisiae* (PDB entry 4csm) and *Escherichia coli* (PDB entry 1ecm).[58,59] These two proteins show sequence identity below 20%; however, they both adopt a similar fold. The bound ligand, a bicyclic transition-state analogue inhibitor, is recognized mostly through side-chain interactions. Furthermore, in the *S. cerevisiae* enzyme the binding pocket is composed of residues emerging from one peptide chain, whereas in the *E. coli* protein two chains contribute to interact with the co-crystallized inhibitor. As we will show later in the manuscript, such situations would represent a limitation for the applicability of methods using sequence alignments in order to detect protein binding sites similarities while in theory being very well handled by an approach like the one implemented in our procedure.

A database consisting of 990 randomly selected protein binding sites[60] containing the chorismate mutase from *E. coli* was created while excluding *S. cerevisiae*, which was used as a query. The previously described protocol was used to compare the query pharmacophore fingerprint to those of all the database entries. The top scored solution was found for the cavity from *E. coli*. This first case study demonstrates the importance and success of a procedure based on alignment independency and physicochemical/pharmacophoric 3D descriptors to capture relevant hydrogen-bond patterns, hydrophobic contacts, and charge interactions which are effective for discriminating similar from randomly selected binding sites.

Clustering Kinase Subfamilies. Kinases represent a large class of potential drug targets which have been heavily pursued by the pharmaceutical industry in recent years. Their involvement in the regulation of all aspects of cellular function makes kinase inhibition a key strategy for the treatment of both oncology and nononcology indications. However, because of the high degree of structural homology across different protein kinases, predicting selectivity based on the binding site shape, pharmacophores, or sequence remains a challenging task.

As a preliminary study, we tested FLAP's ability to cluster protein active sites belonging to specific kinases in distinct subfamilies. For this study, 23 X-ray structures from six specific kinase subfamilies were selected and reported in Table 1. Two well-known Serine-Threonine subfamilies, cyclin-dependent kinase 2/cyclin A (CDK2-CyclinA) and glycogen synthase kinase 3 beta (GSK3β), were included in the data set as previous studies[61,62] have shown that CDK2 inhibitors tend to be also quite potent on GSK3β, and a major problem in discovering CDK2 kinase inhibitors is to achieve selectivity with GSK3β.[61] Mitogen-activated protein kinase (P38α MAPK) was added to the data set for its importance as a therapeutic target in the treatment of a number of inflammatory and autoimmune diseases.[63] Two other Serine-Threonine protein structures were included: the proto-oncogene kinase Pim1, involved in several biological functions including cell survival, proliferation, and differentiation and the 3-phosphoinositide-dependent protein kinase-1 (PDK1), a key protein kinase which regulates the activity of a group

Table 1. Structural Details About the 23 Protein Kinases Data Set Used in the Clustering Analysis with FLAP

PDB entry	compound	subfamily	resolution (å)
1h1s	CDK2	Ser/Thr	2.0
1oi9	CDK2	Ser/Thr	2.1
1oiu	CDK2	Ser/Thr	2.0
1oiy	CDK2	Ser/Thr	2.4
1q3d	GSK3β	Ser/Thr	2.2
1q3w	GSK3β	Ser/Thr	2.3
1q41	GSK3β	Ser/Thr	2.1
1q5k	GSK3β	Ser/Thr	1.9
1ouk	P38	Ser/Thr	2.5
1ouy	P38	Ser/Thr	2.5
1r3c	P38	Ser/Thr	2.0
1wbo	P38	Ser/Thr	2.2
1uu3	PDK1	Ser/Thr	1.7
1uu7	PDK1	Ser/Thr	1.9
1uu8	PDK1	Ser/Thr	2.5
1uvr	PDK1	Ser/Thr	2.8
1xws	PIM1	Ser/Thr	1.8
1yi3	PIM1	Ser/Thr	2.5
1yi4	PIM1	Ser/Thr	2.4
1qpc	LCK	Tyr	1.6
1qpd	LCK	Tyr	2.0
1qpe	LCK	Tyr	2.0
1qpj	LCK	Tyr	2.2

of related protein kinases through phosphorylation and important as a target for the treatment of diabetes and cancer.[64,65] The last kinase included in this study, the lymphoid cell kinase (LCK), belongs to the Tyrosine protein kinase subfamily (TK), which plays a key role in the regulation of cell proliferation, malignancy, and signal transduction. The importance of LCK lies in the fact that it regulates T cell maturation and activation, and it is perhaps the best-studied and best-understood member of cytoplasmatic, nonreceptor TK of the Src family.[66]

The protein kinase selection was performed trying to maximize the following criteria: the structure should be of pharmaceutical interest, it should derive from a human source, its X-ray protein−ligand complex structure should be available with a resolution of less than 2.5 Å, and the corresponding B factor was required to be less than 40. Moreover, a sequence alignment analysis was carried out to check for any mutations or gaps present in three of the most important regions within the ATP binding site, that is, the activation loop, the glycine-rich loop, and the DFG-in/out conformation.

Molecular interaction fields were generated for each target using the GRID force field with five chemical (functional group) probes: H-bond donor ("N1", NH of amide), H-bond acceptor ("O", C═O of carbonyl), negative charged ("O═", resonating oxygen, i.e., phosphates), hydrophobic probe ("DRY"), shape probe ("H"). The 4-point pharmacophore fingerprints were subsequently generated starting from the obtained MIFs. By way of example, Figure 5 illustrates the contours and pharmacophore features for the 1H1S kinase active site. The ensemble of favorable MIF locations, called "hotspots", was treated as a hypothetical molecule interacting at all favorable positions in the binding site, and its pharmacophore fingerprint was calculated and analyzed from these hotspots in the same way as for any ordinary ligand.[28]

FLAP was then applied to perform target-pair comparisons for all protein kinase binding sites under investigation, including comparisons of different structures for the same protein site. This operation produces a numerical description for each binding site containing the number of tetrahedral configurations in common between one protein site and another for all protein sites under investigation. For example, the PIM1 kinases 1XWS and 1YI3 showed 59 (14 DRY, 21 N1, 10 O, 14 O═) and 51 (9 DRY, 18 N1, 11 O, 13 O═) hotspots that, respectively, lead to 430 037 and 224 817 4-point pharmacophores, with 9870 in common, decreasing to only 193 when molecular shape filtering was turned on. Similarly the LCK kinase 1QPC showed 57 hotspots (8 DRY, 21 N1, 13 O, 15 O═), leading to 365 920 4-point pharmacophores, with 7650 in common with 1XWS and 4601 in common with 1YI3, reduced to 28 and 7, respectively, with shape filtering on.

Comparison of one protein site with another may give rise to a different result if the comparison is inverted (comparison of protein *I* with protein *J* may be different from the comparison of protein *J* with protein *I*) due to the asymmetry of the steric effect. For this reason, an ad hoc function for computing binding sites similarities was implemented and is defined as follows:

$$S_{IJ} = S_{JI} = \frac{T_{IJ}T_{JI}}{T_{II}T_{JJ}}$$

Here, $S_{IJ(JI)}$ is the absolute similarity score obtained for the comparison of protein *I* and *J*, T_{IJ} and T_{JI} are, respectively, the 4-point pharmacophore combinations in common between protein *I* and protein *J*, and vice versa. T_{II} and T_{JJ} are the number of common pharmacophores for the comparison of protein *I* and *J* with themselves, and they were used to normalize the corresponding similarity values from 0 to 1.

The final similarity matrix is composed of numbers that change according to the data set under investigation, 0 and 1 being the minimum and maximum similarity, respectively. The data in this matrix can be analyzed using multivariate statistical analysis in an attempt to discover possible patterns or trends across different families or subfamilies of protein binding sites. The PCA scores plot in Figure 6 (left side) represents the relative position of the objects (kinase active sites) in the space of the principal components (PCs). In this new coordinate space set by PC1 and PC2, the 23 protein kinases have been subdivided into three main clusters, the first containing Pim1, the second P38α and LCK, and the third including GSK3β, CDK2, and PDK1. As the principal components are extracted in decreasing order of importance, the first PC is derived in order to explain the maximum amount of variation and, therefore, when there are clusters of objects to distinguish among them. In our case, the first PC separates proteins of the Pim1 subfamily from all the other in the data set ($\mathrm{PC}_1^{\mathrm{ExplainedVariance}} = 35.2\%$). Previous studies for Pim1[67] revealed a unique hinge region lacking a hydrogen-bond donor (common active site motif among protein kinase), suggesting potential for the development of specific Pim kinase inhibitors and explaining the outlier behavior highlighted by the PCA analysis. The second and third PCs (Figure 6, right side) were able to differentiate subfamilies of kinases within the second and third cluster ($\mathrm{PC}_2^{\mathrm{ExplainedVariance}} = 12.7\%$, $\mathrm{PC}_3^{\mathrm{ExplainedVariance}} = 8.3\%$). Taken together, these results shows how the new pharmacophore-based descriptors encoded in FLAP combined with an effective multivariate statistical algorithm can be successfully applied to classify protein kinase subfamilies, making this

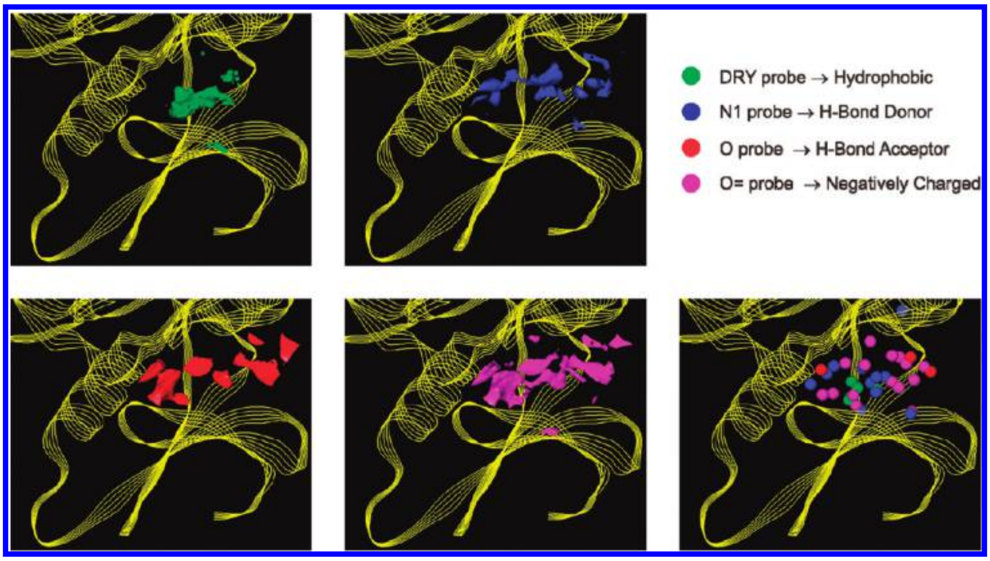
● DRY probe → Hydrophobic
● N1 probe → H-Bond Donor
● O probe → H-Bond Acceptor
● O= probe → Negatively Charged

Figure 5. GRID-MIF for 1H1S kinase and the target-site points (lower right) used for pharmacophore fingerprint calculations.

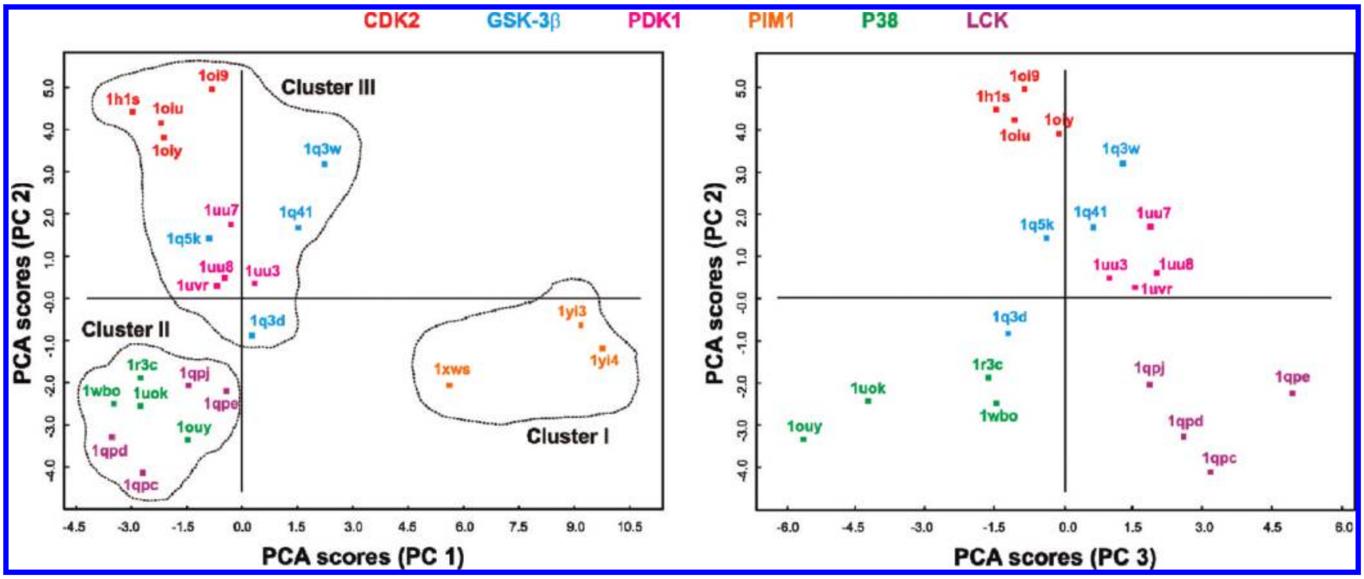

Figure 6. "ALL against ALL" results obtained by applying principal component analysis (PCA) to the data set of 23 protein kinases (PKs). The first principal component (PC) explains the structural differences in the PIM1 subfamily when compared to the other kinase families (left). Overall, the first two PCs grouped the data into three main clusters and the third PC is necessary in order to further differentiate the remaining subfamilies of PKs within the second and third cluster (right).

procedure a valuable tool for large-scale analysis of protein binding sites.

Kinase Specificity for Staurosporine Inhibition. There are currently over 30 known kinase inhibitors in clinical trials or approved for use in humans, and the particular set of kinases inhibited by a compound may drastically affect its therapeutic usefulness. Although based only on relatively small data sets of kinases, previous studies have shown how molecular specificity varies widely among known inhibitors,[68] and this variation is not dictated by the general chemical scaffold of an inhibitor (e.g., EGFR inhibitors, belonging to the quinazoline/quinoline class, range from highly specific to quite promiscuous) or by the primary, intended kinase target toward which the particular inhibitor was initially optimized (e.g., compounds considered TK inhibitors also bind to Ser-Thr kinases and vice versa).

To further investigate kinase specificity, a well-characterized kinase inhibitor, staurosporine, was chosen to probe its differential propensity to inhibit a specific set of protein kinases. Staurosporine is highly promiscuous, and it binds to at least 100 diverse kinase proteins with affinities evenly distributed from picomolar to low micromolar.

The same procedure, described in the previous section on clustering, was applied to test the ability of FLAP to identify binding site features correlating with the reported staurosporine inhibition profile across several kinase structures.[69] For this purpose, 14 protein kinases were selected based on the following criteria: proteins were of potential interest as pharmaceutical targets, available as X-ray structure in the PDB, and with the corresponding cavity information already deposited in the in-house database of protein binding sites. Additional filters were related to the resolution at which the structure was solved and the integrity of the kinase domain in both the ATP binding site and the activation/glycine-rich loop region. Details about the selected protein kinases are given in Table 2.

To best describe the structural features of the 14 selected protein active sites, an initial cross-validation analysis was

Table 2. PDB Entries Belonging to Different Kinases Used for Studying Their Different Propensity to Staurosporine Inhibition

PDB code	protein name	EC number	resolution (Å)	IC50 (nM)	P(IC50) (nM)
1aq1	CDK2	2.7.1.37	2.00	5.8	−0.76
1byg	CSK	2.7.1.112	2.40	22	−1.34
1nxk	MAPKAPK-2	2.7.1.−	2.70	140	−2.15
1nvr	CHK1	2.7.1.−	1.80	1.6	−0.20
1ol5	AURORA-A	2.7.1.37	2.50	5.8	−0.76
1jqh	IGF-1R	2.7.1.112	2.10	400	−2.60
1q3d	GSK-3β	2.7.1.37	2.20	5.7	−0.76
1stc	PKA	2.7.1.37	2.30	3.3	−0.52
1u59	ZAP-70	2.7.1.112	2.30	3.3	−0.52
1xbc	SYK	2.7.1.112	2.00	1.2	−0.08
1xkk	EGFR	2.7.1.112	2.40	100	−2.00
1xjd	PKC-θ	2.7.1.−	2.00	0.52	+0.28
1yvj	JAK3	2.7.1.112	2.55	0.34	+0.47
1m52	ABL1	2.7.1.112	2.60	71	−1.85

carried out in an attempt to find the most relevant combination of GRID probes to use in the cavity mapping generation step (see section 2 of Theory). From the original pool of 12 specific GRID probes,[70] 5 of them were each time extracted exhaustively for all the possible 5-probe combinations, used to generate MIFs for the proteins in the data set, and the corresponding isopotential maps analyzed by the FLAP algorithm (see section 3 in Theory). For each tested combination of GRID probes, a similarity matrix was obtained and exploited as block of descriptors to train the corresponding partial least squares (PLS) model. The negative logarithm of staurosporine inhibition was used as dependent variable in the regression model (pIC_{50}).

The quality of the model was assessed through leave one out validation (LOO),[71] and the best combination of probes that came out of the analysis, based on Q^2_{LOO}, was chosen to energetically describe the 14 kinase active sites.[72] FLAP was then applied to compute all the common pharmacophores among the proteins in the data set, and the resulting similarity matrix was subjected to a variable selection routine (based on a fractional factorial design) to remove irrelevant variables from the descriptors space, which finally amounted to 803 4-point pharmacophore configurations. To test the predictive power of the final PLS model, LOO cross-validation was used and an optimal model with two latent variables was derived (Q^2 = 0.63, standard deviation of the error in prediction 'SDEP' = 0.55). As shown in Figure 7, all the protein kinases were satisfactorily assigned in terms of staurosporine sensitive(blue)/not-sensitive(red) classes, indicative of the FLAP's ability to efficiently cluster protein binding sites according to the inhibition value induced by a certain compound (staurosporine in this case).

A future extension of such a study would be identification of the structural features in the binding sites responsible for the inhibition trend under investigation. This analysis will support the rational design of specific inhibitors with appropriate specificity that target single or multiple kinases involved in the disease process while avoiding kinases implicated in different biological pathways.

High-Throughput Protein Virtual Screening. Virtual high-throughput screening or virtual screening is an in silico technique used in drug design research for the rapid assessment of large collections (libraries) of chemical structures in order to guide the selection of compounds likely to have activity on a specific target. Ligand-based (once one or more active inhibitor(s) has been identified) and/or structure-based (when a target protein structure or model is available) virtual screening can be used to search for molecules that closely resemble the compounds in terms of structure and/or properties (e.g., 1−3D descriptors, pharmacophore-based descriptors, shape matching) and/or are complementary to the protein binding site. Importantly, the search is not restricted to compounds that have already been synthesized but may include synthetically accessible compounds that exist only in virtual form and in the case of structure-based approaches can involve de novo design approaches, including the use of fragments.

The idea behind this study was to extend the pharmacophore-driven virtual screening protocol, as used frequently in pharmaceutical research, from small drug-like molecules to the analysis of protein binding sites. This is particularly relevant as many more protein structures are becoming available through the recent efforts made by structural genomics projects. Various initiatives are focused on the 3D structure determination of proteins of interest to human disease and/or all proteins of a given organism using experimental methods such as X-ray crystallography, NMR spectroscopy, or computational homology modeling. Such

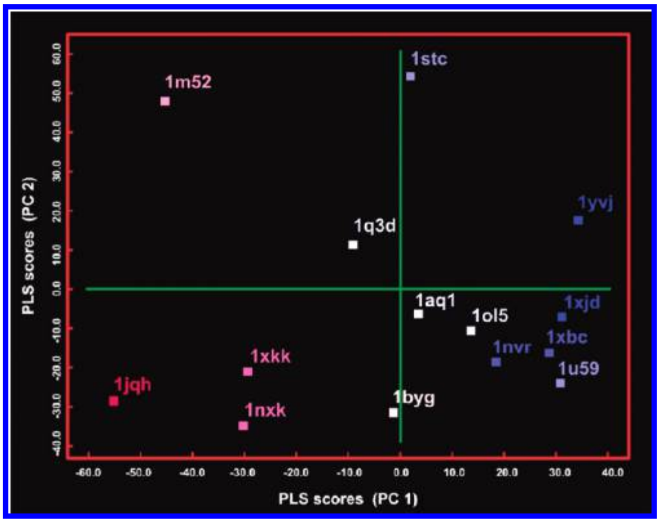

Figure 7. PLS scores plot representing 14 different protein kinases in the latent variables space, colored by their activity (red = low inhibition, blue = high inhibition) against staurosporine inhibition. Kinases with different staurosporine inhibition levels are correctly ranked from the left (lower inhibition) to the right (higher inhibition) side of the PLS scores plot.

Table 3. Description of the Protein Classes Employed, Showing the Corresponding PDB Entries for Both the Query and the Database Binding Sites

protein family	PDB query	PDB database hits	decoys
HIV protease	1hvs	1hbv, 1hpx, 1hvc, 1hvi, 1hvj, 1hvk, 1hvl, 1aaq, 1sbg, 2upj	990 randomly selected binding sites
L-arabinose	1abf	1abe, 1apb, 1bap, 5abp, 6abp, 7abp, 8abp, 9abp, 1dbp, 2dri	
ALBP	1lid	1al8, 1ab0, 1acd, 1adl, 1alb, 1lib, 1lic, 1lie, 1lif, 2ans	
ERα	3ert[b]	(1g50, 1qkt, 1qku, 3erd)[a], (1sj0, 1xp1, 1xp6, 1xp9, 1xpc, 1yim)[b]	

[a] Estrogen receptor α (cocrystallized with an agonist). [b] Estrogen receptor α (cocrystallized with an antagonist).

Table 4. Description of the Database Composition for the Four Structural Classes of Ligands Employed[a]

ligand name	PDB query	PDB database hits	decoys
ATP	1e2q	1b8a, 1gn8, 1hck, 1j09, 1kay, 1kj8, 1kj9, 1n75, 1yag, 2bek	990 randomly selected binding sites
citrate	1az2	1emd, 1gcz, 1gd0, 1hqs, 1l8d, 1reo, 2acs, 2acu, 2cts, 6csc	
maltose	1mpd	1a7l, 1cdg, 1cgw, 1cgx, 1cxf, 1kcl, 1n3w, 1nl5, 1qho, 1qhp	
retinoic acid	1gx9	1cbr, 1cbs, 1epb, 1fby, 1fem, 1fm6, 1fm9, 1g5y, 1k74, 1n4h	

[a] PDB entry names for both the query and the hit proteins in the database are also reported.

protein structural initiatives add to the need for higher throughput as well as ligand-relevant approaches.

Unlike traditional structural biology, the determination of a protein structure through a structural genomics effort often comes before anything is known regarding the protein function; therefore, there is a need for fast bioinformatics tools able to infer protein functions directly from their 3D structures. The need to identify proteins with similar binding sites regardless of sequence identity (e.g., for selectivity and/or to identify potential hits) has led us to also develop a tailored computational procedure, based on the FLAP engine, where the similarity to a given protein active site (query) can be rapidly assessed against publicly available databases of solved protein structures.

The employed work flow derives from applying FLAP for describing protein binding sites on the basis of their shape and 4-feature pharmacophores properties, as seen in the clustering analysis. However, this time the comparison is not clustering based but more oriented toward a virtual screening study. Every protein binding site stored in the database is compared, in a pairwise mode, to the selected query, and a FLAP score is given to each of them, representing their similarity in terms of common 4-point pharmacophore configurations.

To test the robustness of such a procedure, eight different databases were artificially created for this purpose (see Tables 3 and 4); the approach consists of incorporating a set of protein active sites (for which their function is known) in a larger database of randomly selected protein cavities, also called decoys (assumed to have unrelated functions). After that, the ability of the model to query the database for retrieving similar binding sites is evaluated based on its performance compared to a random search (enrichment factor).

Two different studies were carried out in which FLAP was tested for both (i) its performance on retrieving protein binding sites belonging to the same family as the query binding site and (ii) in the case of "off-target" ligands, where proteins with no evident sequence homology bind the same ligand.

Four different databases were made for testing the first hypothesis, one for each protein family under investigation: (a) HIV-protease, (b) L-arabinose, (c) adipocyte lipid binding protein (ALBP), (d) estrogen receptor alpha (ER-alpha). These proteins were selected given their importance as pharmaceutical targets and the high number of crystal structures deposited in the PDB. For each target, 10 different protein entries belonging to the same family were used to enrich a database of 990 randomly selected protein structures.[60] In addition, one more PDB entry was selected and used as "query" for each protein family (Table 3). Figure 8 reports the enrichment curves as obtained after using FLAP on these 4 protein data sets.

For both ALBP and L-arabinose families, satisfactory results were obtained in terms of enrichment, with all 10 protein hits being retrieved within the first 4% of the database screened. The screening performed on HIV-protease gave 7 out of 10 on-target proteins retrieved within the first 5% of the database screened. Looking at the false-positive ranked highest, the 12th position was assigned to protein "human rhinovirus" (PDB entry 1hrv) that, in terms of sequence homology, is completely unrelated to the query (PDB entry 1hvs).

However, the binding site of protein 1hrv shows surprising similarities to that of protein 1hvs in terms of molecular interaction fields (Figure 9). Both cavities are surrounded by two large hydrophobic regions with a hydrogen-bond donor area located between them. This is in agreement with the position of the corresponding cocrystal ligand structure features inside the cavity, where a common pharmacophore through which both ligands could interact with the binding site is highlighted (Figure 9). Surprisingly, human rhinovirus and HIV-protease proteins share multiple nodes within the gene ontology (GO) tree browser (GO ID 16020, 16032, 16740, 16779, 16787, 19012, 19028, 3723, 3824, 5198, 5737, 6508, 8233), indicating similarity also in terms of cellular location, biological processes, and molecular functions.[73] This finding was completely unexpected and should be taken as a warning whenever the performance of a virtual screening protocol is evaluated on the basis of artificial data sets where no functional annotation is given to their objects.

Valuable results were also obtained in the virtual screening analysis performed on ER-alpha receptor. Starting from PDB entry 3ert as query, the method was able to retrieve 6 out of 10 hits within the first 7 positions in the scoring list (obtaining nearly an ideal enrichment) while showing

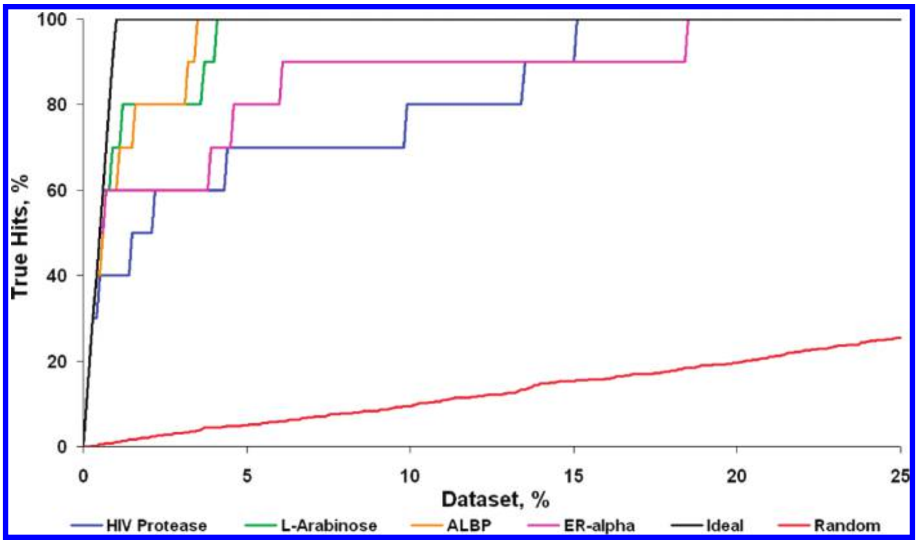

Figure 8. Enrichment curves showing the FLAP performance on retrieving proteins belonging to the same family as the query.

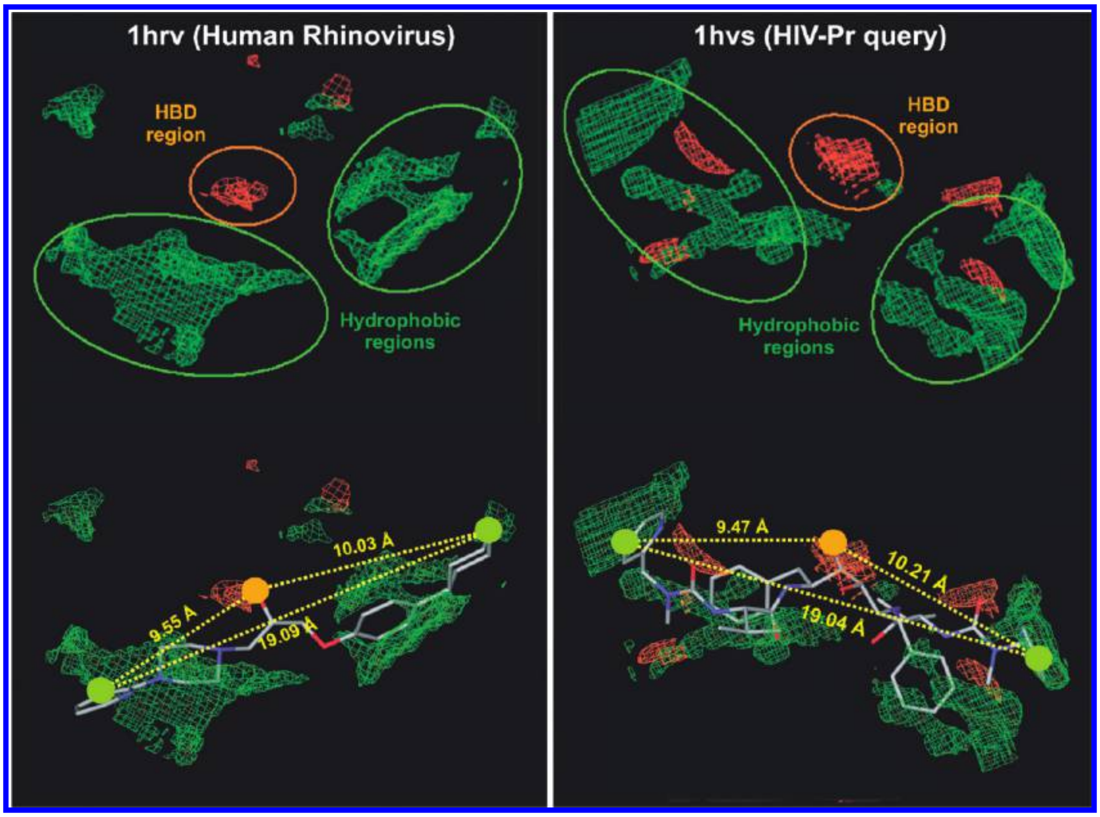

Figure 9. GRID molecular interaction fields showing hydrogen-bond donor (HBD) and hydrophobic (DRY) regions for 1hrv and 1hvs protein binding sites (top). X-ray intrafeatures ligand distances are compared with the position of high relevant regions of interaction in both active sites. A possible common pharmacophore is highlighted (bottom).

problems retrieving the remaining 4. A possible explanation can be found in the structural basis of ER activation, as these 6 proteins were cocrystallized in the presence of an ERα antagonist as was the query, while the remaining 4 crystal structures of proteins scored lower by FLAP were obtained in the presence of an ERα agonist. Indeed, the ER ligand binding domain is capable of existing in multiple response states depending on the nature of the bound ligand and its active site geometry undergoes significant structural modifications upon the agonist/antagonist binding. The resultant conformation reflects the size and shape of the ligand and more importantly determines what type of coregulator is recruited, CoA coactivator in the case of agonists binding and CoR corepressor after antagonists binding.[74]

The unexpected findings obtained by testing the first hypothesis (i.e., HIV-protease) suggested to us the investigation of another possible application of FLAP for retrieving protein active sites showing very low sequence homology and binding the same ligand. As stated before, there are proteins with neither sequence nor fold homology that perform similar biological functions since they share similar binding patterns and recognize similar binding partners.

Four additional protein databases were made for this analysis, one for each ligand class under investigation (ATP,

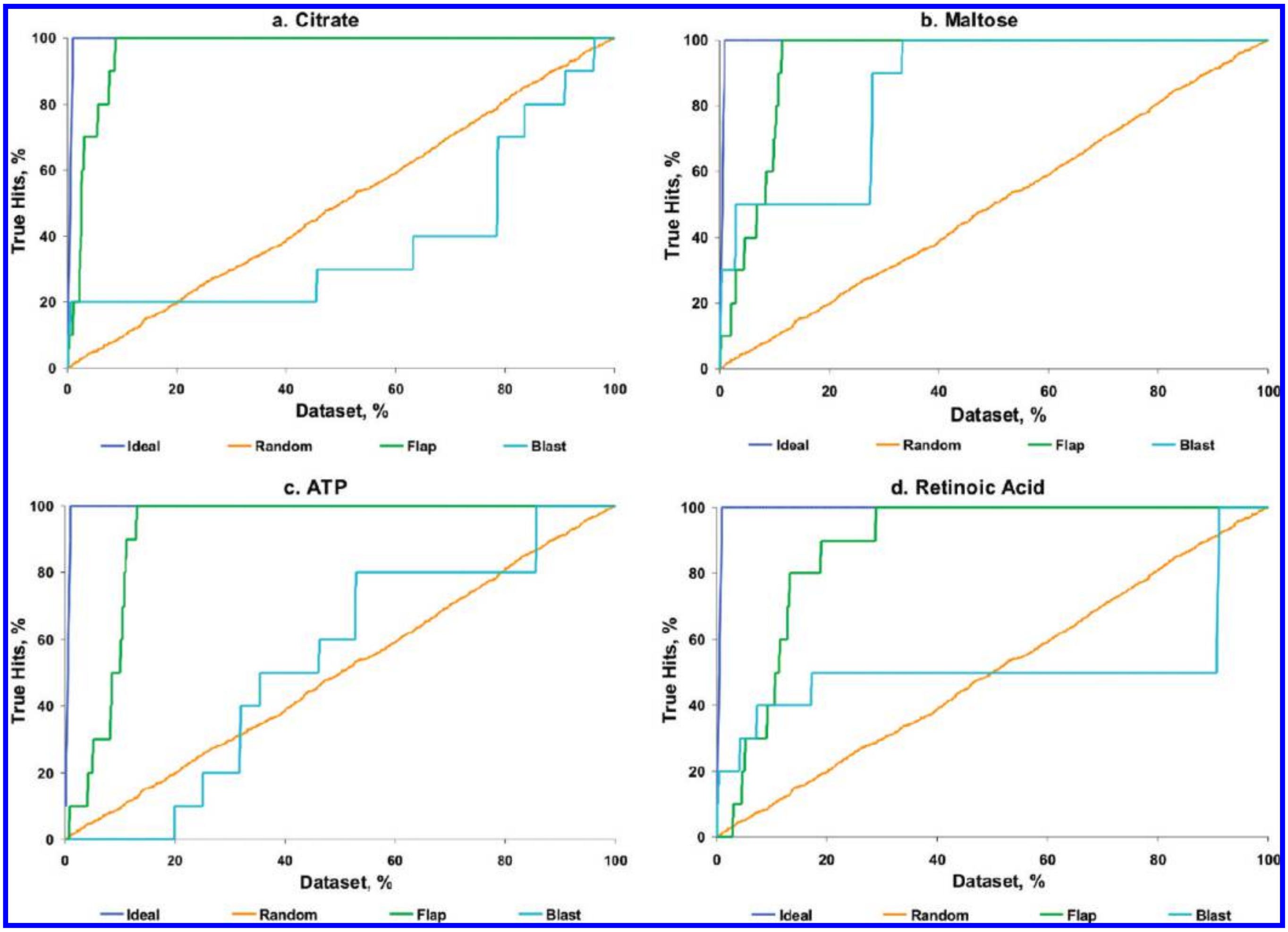

Figure 10. Enrichment curves showing the FLAP performance on retrieving unrelated proteins binding the same functional ligand, and comparison with a well-known sequence similarity method BLAST.

citrate, maltose, retinoic acid). These ligands were selected given their importance as biological molecules and the high number of cocrystal structures deposited in the Protein Data Bank. For each of them, 10 different PDB entries bound to the same ligand were used to enrich the same database of 990 randomly selected protein structures.[60] In addition, one more PDB entry was selected and used as "query" for each ligand class (Table 4).

Results were compared with those obtained by performing a sequence similarity search using BLAST.[6] For this purpose, the same proteins used in FLAP were also used to build a FASTA[4,5] database. BLAST[6] was then used to compare a query sequence to the other sequences collected in the FASTA database in a pairwise manner. Each sequence was given a score reflecting its degree of similarity to the query, and those scores, together with the FLAP scores, were used to build the corresponding enrichment curves. The results for this comparison are reported in Figure 10.

The screening analysis performed on the citrate, maltose, and ATP data sets gave satisfactory results regarding the FLAP performance, with all 10 proteins retrieved within the first 10% of the screened database. Moreover, the results obtained with BLAST confirm the sequence similarity weakness in recognizing any such kind of "off-target" cross-reactivity exploited by certain classes of ligands, since the signal is quite often near to random.

The data set where FLAP performed slightly worse was the one used for retrieving protein binding to retinoic acid. In this study, FLAP failed to find any of the 10 different protein−(retinoic acid) complexes within the first 30 positions in the scoring list. However, when these first 30 positions in the scoring list are examined, it is possible to observe a high degree of structural similarity between the retinoic acid and the ligands found in some of the cocrystal structures (Figure 11), explaining the slightly lower performance for this data set.

CONCLUSIONS AND OUTLOOK

The use of protein 3D pharmacophore fingerprints (4-feature), derived by the GRID/FLAP procedure using "hotspots" complementary to the protein sites, has been shown to be a powerful new approach for protein−protein comparisons (clustering, virtual screening) in addition to the previously described use for ligand−protein (and ligand−ligand) comparisons. The use of "ligand-relevant" descriptors, pharmacophoric features, enables relationships to be identified that are not found by amino acid sequence-based approaches. The computational algorithm used, FLAP, can be applied to large numbers of ligands and proteins, enabling new approaches for similarity (virtual screening, clustering, etc.) for all combinations: ligand−ligand, ligand−protein, protein−protein. The work described here further validates the use of GRID-FLAP to compare ligands (including scaffold hopping) and for ligand−receptor (e.g., docking and virtual screening) studies as well as providing a ligand- (and

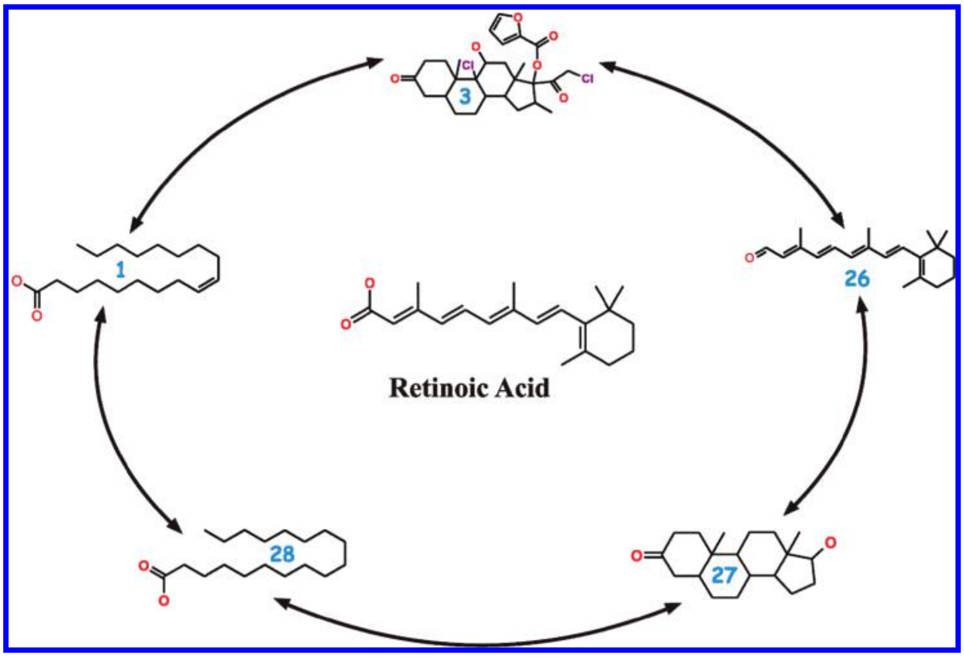

Figure 11. 2D depiction of the ligand structures found in some of the protein active sites with high FLAP score, compared with retinoic acid structure. The number corresponds to the ranking position in the output scoring list.

potentially function-) relevant approach to compare proteins. In terms of protein−protein comparisons (virtual screening) the GRID-FLAP approach can be used to take a (potential) binding site of a protein (including those for which no ligand information is known) and search a database of other proteins to find ones that have a similar binding site (in terms of protein-desired ligand properties). This information can be used to find potential ligands for the query protein, such as for tool compounds to aid in target validation studies or for hit/lead identification, by screening (in silico and in vitro) ligands previously found to be active on these similar proteins.

A further application of the method is to use the results of the protein−protein virtual screening to identify potential selectivity targets for the query protein, including those that would not be found by a sequence similarity approach. The GRID-FLAP approach has a significant advantage for this task in that it uses an energetically derived ligand "image" of the binding site for its comparison and can search completely diverse proteins with no structural alignment needed, enabling the structures of new protein classes such as the G-protein-coupled receptors (GPCRs) to be rapidly leveraged.

Representatives of the majority of the drug target classes are now available with high-resolution 3D X-ray structures, with recently the β1 and β2 adrenergic[75–77] and A_{2A} adenosine[78] GPCRs having been solved. Overall, the number of available protein and protein−ligand structures is increasing rapidly from both "structural genomics" initiatives and the publication of results from the enormous structural biology efforts in industry and academia. Methods that use a ligand-space-relevant characterization of the protein binding site, such as the one described here, FLAP, based on GRID descriptors, will be increasingly useful to determine inter-protein relationships based on binding site properties as well as providing useful structure-based virtual screening/docking approaches. The identification of tool and potential lead compounds will continue to be very important, and both the direct (virtual screening) and indirect (from similar proteins) approaches described are enabling for this need. The pharmacophoric description of the binding site can also be exploited for other key drug design applications such as de novo design, where fragments (or "components" of ligands) can be identified that match pharmacophoric patterns in a putative binding site, with a pose where pharmacophoric interactions are being made (with steric clashes avoided and steric fit optionally optimized). Thus, the availability of approaches that bring a relevant common 3D property frame of reference to the worlds of the ever increasing numbers of proteins and compounds (potential ligands) enables enhanced analysis possibilities and use of these fundamental components for rational drug discovery.

Abbreviations: 3D, 3-dimensional; BLAST, basic local alignment search tool; FLAP, fingerprints for ligands and proteins; GO, gene ontology; GPCR, G protein-coupled receptor; LOO, leave one out; MIF, molecular interaction fields; NMR, nuclear magnetic resonance; PC, principal component; PCA, principal component analysis; PLS, partial least squares; SDEP, standard deviation of the error in prediction.

ACKNOWLEDGMENT

Pfizer Sandwich U.K. is acknowledged for its support of the FLAP project. The authors gratefully thank Molecular Discovery for providing the FLAP program and the laboratory for Chemometrics and Cheminformatics at the University of Perugia for help on testing and improving the FLAP code. Personal gratitude is expressed by S.S. and G.C. to Dr. Jonathan Mason for his inspiration and constructive discussions during the development of the FLAP program and the preparation and revision of this manuscript.

Supporting Information Available: Physicochemical properties of the 990 proteins used as decoy for the protein virtual screening analysis and listing of all the PDB IDs for the 990 proteins together with the protein family name, ligand name, and experimental resolution. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Lander, E. S.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860–921.

(2) Venter, J. C.; Adams, M. D.; Myers, E. W.; Li, P. W.; Mural, R. J. The sequence of the human genome. *Science* **2001**, *291*, 1304–1351.

(3) Wilkins, M. R.; Pasquali, C.; Appel, R. D.; Ou, K.; Golaz, O. From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Biotechnology* **1996**, *14*, 61–65.

(4) Pearson, W. R.; Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 2444–2448.

(5) Pearson, W. R. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* **1990**, *183*, 63–98.

(6) Altschul, S. F.; Gish, W.; Miller, W.; Meyers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215*, 403–410.

(7) Bairoch, A.; Boeckmann, B. The SWISS-PROT protein sequence data bank: current status. *Nucleic Acids Res.* **1994**, *22*, 3578–3580.

(8) Bairoch, A. PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res.* **1991**, *19*, 2241–2245.

(9) Bleasby, A. J.; Akrigg, D.; Attwood, T. K. OWL a non-redundant composite protein sequence database. *Nucleic Acids Res.* **1994**, *22*, 3574–3577.

(10) Rosen, M.; Lin, S. L.; Wolfson, H.; Nussinov, R. Molecular shape comparisons in searches for active sites and functional similarity. *Protein Eng.* **1998**, *11*, 263–277.

(11) Russel, R. B. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J. Mol. Biol.* **1998**, *279*, 1211–1227.

(12) Artymiuk, P. J.; Poirrette, A. R.; Grindley, H. M.; Rice, D. W.; Willett, P. A Graph-theoretic Approach to the Identification of Three-dimensional Patterns of Amino Acid Side-chains in Protein Structures. *J. Mol. Biol.* **1994**, *243*, 327–344.

(13) Nussinov, R.; Wolfson, H. J. Efficient Detection of Three-Dimensional Structural Motifs in Biological Macromolecules by Computer Vision Techniques. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 10495–10499.

(14) Wolfson, H. J.; Rigoutsos, I. Geometric hashing: an overview. *IEEE Comput. Sci. Eng.* **1997**, *11*, 263–278.

(15) Lehtonen, J. V.; Denessiouk, K.; May, A. C. W.; Johnson, M. S. Finding local structural similarities among families of unrelated protein structures: A generic non-linear alignment algorithm. *Proteins: Struct. Funct. Genet.* **1999**, *34*, 341–355.

(16) Nagano, N.; Orengo, C. A.; Thornton, J. M. One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol.* **2002**, *321*, 741–765.

(17) Wallace, A. C.; Borkakoti, N.; Thornton, J. M. TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.* **1997**, *6*, 2308–2323.

(18) Barker, J. A.; Thornton, J. M. An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics* **2003**, *19*, 1644–1649.

(19) Park, K.; Kim, D. A Method to Detect Important Residues Using Protein Binding Site Comparison. *Genome Inform.* **2006**, *17*, 216–225.

(20) Jones, S.; Thornton, J. M. Searching for functional sites in protein structures. *Curr. Opin. Chem. Biol.* **2004**, *8*, 3–7.

(21) Schmitt, S.; Kuhn, D.; Klebe, G. A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology. *J. Mol. Biol.* **2002**, *323*, 387–406.

(22) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Recognition of Functional Sites in Protein Structures. *J. Mol. Biol.* **2004**, *339*, 607–633.

(23) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. SiteEngines: recognition and comparison of binding sites and protein-protein interfaces. *Nucleic Acids Res.* **2005**, *33*, W337–341.

(24) Shulman-Peleg, A.; Shatsky, M.; Nussinov, R.; Wolfson, H. Spatial chemical conservation of hot spot interactions in protein-protein complexes. *BMC Biol.* **2007**, *5:43*, 1–11.

(25) Jambon, M.; Imberty, A.; Deléage, G.; Geourjon, C. A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins: Struct. Funct. Genet.* **2003**, *52*, 137–145.

(26) Kupas, K.; Ultsch, A.; Klebe, G. Large scale analysis of protein-binding cavities using self-organizing maps and wavelet-based surface patches to describe functional properties, selectivity discrimination, and putative cross-reactivity. *Proteins: Struct., Funct., Bioinf.* **2008**, *71*, 1288−1306.

(27) Perruccio, F.; Mason, J. S.; Sciabola, S.; Baroni, M. FLAP: 4-Point Pharmacophore Fingerprints from GRID. In *Molecular Interaction Fields*; Cruciani, G., Ed.; Wiley-VCH: New York, 2006; Vol. 27, pp 83−102.

(28) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A Common Reference Framework for Analyzing/Comparing Proteins and Ligands. Fingerprints for Ligands And Proteins (FLAP): Theory and Application. *J. Chem. Inf. Model.* **2007**, *47*, 279–294.

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

(30) Carosati, E.; Sciabola, S.; Cruciani, G. Hydrogen Bonding Interactions of Covalently Bonded Fluorine Atoms: From Crystallographic Data to a New Angular Function in the GRID Force Field. *J. Med. Chem.* **2004**, *47*, 5114–5125.

(31) *SYBYL*, version 7.3; Tripos Inc.: St. Louis, MO.

(32) 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*, 841–59.

(33) Ho, C. M. W.; Marshall, G. R. Cavity search: An algorithm for the isolation and display of cavity-like binding regions. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 337–354.

(34) Laskowski, R. A. SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.* **1995**, *13*, 323–330.

(35) Levitt, D. G.; Banaszak, L. J. POCKET: A computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graph.* **1992**, *10*, 229–234.

(36) Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: Automatic and efficient detection of potential small moleculebinding sites in proteins. *J. Mol. Graph. Model.* **1997**, *15*, 359–363.

(37) Huang, B.; Schroeder, M. LIGSITE^{csc}: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.* **2006**, *6:19*, 1−11.

(38) Peters, K. P.; Fauck, J.; Frömmel, 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.

(39) 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.

(40) Edelsbrunner, H.; Mucke, E. P. Three-Dimensional Alpha Shape. *ACM Trans. Graph.* **1994**, *13*, 43–72.

(41) Weisel, M.; Proschak, E.; Schneider, G. PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chem. Cent. J.* **2007**, *1:7*, 1–17.

(42) *CAVGEN*, version 1.0; Pfizer: Kent, Sandwich, U.K.

(43) *GRID*, version 22; Molecular Discovery Ltd.: Pinner, Middlesex, U.K.

(44) Boobbyer, D. N. A.; Goodford, P. J.; McWhinnie, P. M.; Wade, R. C. New hydrogen-bond potentials for use in determining energetically favorable binding sites on molecules of known structure. *J. Med. Chem.* **1989**, *32*, 1083–1094.

(45) Mason, J. S.; Good, A. C.; Martin, E. J. 3D-Pharmacophores in Drug Discovery. *Curr. Pharm. Des.* **2001**, *7*, 567–597.

(46) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C. R.; Labaudiniere, R. F. New 4-Point Pharmacophore Method for Molecular Similarity and Diversity Applications: Overview of the Method and Applications, Including a Novel Approach to the Design of Combinatorial Libraries Containing Privileged Substructures. *J. Med. Chem.* **1999**, *42*, 3251–3264.

(47) Mason, J. S.; Cheney, D. L. Ligand-receptor 3-D similarity studies using multiple 4-point pharmacophores Pac. *Symp. Biocomput.* **1999**, *4*, 456–467.

(48) Eksterowicz, J. E.; Evensen, E.; Lemmen, C.; Patrick Brady, G.; Kevin Lanctot, J.; Bradley, E. K.; Saiah, E.; Robinson, L. A.; Grootenhuis, P. D. J.; Blaney, J. M. Coupling structure-based design with combinatorial chemistry: application of active site derived pharmacophores with informative library design. *J. Mol. Graph. Model.* **2002**, *20*, 469–477.

(49) Gohlke, H.; Hendlich, M.; Klebe, G. Predicting binding modes, binding affinities and 'hot spots' for protein-ligand complexes using a knowledge-based scoring function. *Perspect. Drug Discovery Des.* **2000**, *20*, 115–144.

(50) Mason, J. S.; Pickett, S. D. Combinatorial Library Design, Molecular Similarity and Diversity Applications. In *Burger's Medicinal Chemistry and Drug Discovery*, 6th ed.; Abraham, D. J., Ed.; John Wiley & Sons: New York, 2003; Vol. 1, pp 187−242.

(51) Good, A. C.; Mason, J. S.; Pickett, S. D. Pharmacophore Pattern Application in Virtual Screening, Library Design and QSAR. In *Virtual Screening for Bioactive Molecules*; Bohm, H. J., Schneider, G., Eds.; Wiley-VCH: New York, 2000.

(52) Hemm, K.; Aberer, K.; Hendlich, M. Constituting a Receptor-Ligand Information Base from Quality-Enriched Data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1995**, *3*, 170–178.

(53) Hendlich, M. Databases for protein-ligand complexes. *Acta Crystallogr.* **1998**, *D54*, 1178–1182.

(54) Hendlich, M.; Bergner, A.; Gunther, J.; Klebe, G. Relibase: Design and Development of a Database for Comprehensive Analysis of Protein-Ligand Interactions. *J. Mol. Biol.* **2003**, *326*, 607–620.

(55) Gunther, J.; Bergner, A.; Hendlich, M.; Klebe, G. Utilising Structural Knowledge in Drug Design Strategies: Applications Using Relibase. *J. Mol. Biol.* **2003**, *326*, 621–636.

(56) RCSB Protein Data Bank. http://www.pdb.org, 02/142007.

(57) 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.

(58) Lin, S. L.; Nussinov, R.; Fischer, D.; Wolfson, H. J. Molecular surface representations by sparse critical points. *Proteins: Struct. Funct. Genet.* **1994**, *18*, 94–101.

(59) Strater, N.; Schnappauf, G.; Braus, G.; Lipscomb, W. N. Mechanisms of catalysis and allosteric regulation of yeast chorismate mutase from crystal structures. *Structure* **1997**, *5*, 1437–1452.

(60) Lee, A. Y.; Karplus, P. A.; Ganem, B.; Clardy, J. Atomic structure of the buried catalytic pocket of Escherichia coli chorismate mutase. *J. Am. Chem. Soc.* **1995**, *117*, 3327–3362.

(61) Decoy protein data set selection workflow. All structures come from a Pfizer in-house version of the PDB database. We started with the set of proteins for which a co-crystallized ligand was present in the active site. This gave us 25 809 protein−ligand structures. We did not want to bias our selection by any previous knowledge about protein families, and we decided to use the physicochemical information of the bound ligands to define the protein decoy dataset. For this purpose the following filters were applied to the set of 25 809 ligand structures: (i) MW $\leq$ 800, (ii) HB-acceptor $\leq$ 15, (iii) HB-donor $\leq$ 10, (iv)−7 $\leq A \log P \leq 7$, and (v) PSA $\leq$ 350 Å^2. This left us with 10 275 PDB entries after duplicates and bad valence states were removed. In the last step, 990 binding sites were randomly selected out of the 10 275 left (see Supporting Information for more details about protein information and ligand physicochemical properties).

(62) Vulpetti, A.; Crivori, P.; Cameron, A. J.; Bertrand, J.; Brasca, M. G.; D'Alessio, R.; Pevarello, P. Structure-Based Approaches to Improve Selectivity: CDK2-GSK3β Binding Site Analysis. *J. Chem. Inf. Model.* **2005**, *45*, 1282–1290.

(63) Leost, M.; Schultz, C.; Link, A.; Wu, Y.-Z.; Biernat, J.; Mandelkow, E.-M.; Bibb, J. A.; Snyder, G. L.; Greengard, P.; Zaharevitz, D. W.; Gussio, R.; Senderowicz, A. M.; Sausville, E. A.; Kunick, C.; Meijer, L. Paullones are potent inhibitors of glycogen synthase kinase-3beta and cyclin-dependent kinase 5/p25. *Eur. J. Biochem.* **2000**, *267*, 5983–5994.

(64) Noble, M. E. M.; Endicott, J. A.; Johnson, L. N. Protein Kinase Inhibitors: Insights into Drug Design from Structure. *Science* **2004**, *303*, 1800–1805.

(65) Biondi, R. M.; Komander, D.; Thomas, C. C.; Lizcano, J. M.; Deak, M.; Alessi, D. R.; vanAalten, D. M. F. High resolution crystal structure of the human PDK1 catalytic domain defines the regulatory phosphopeptide docking site. *The EMBO J.* **2002**, *21*, 4219–4228.

(66) Kumar, A.; Mandiyan, V.; Suzuki, Y.; Zhang, C.; Rice, J.; Tsai, J.; Artis, D. R.; Ibrahim, P.; Bremer, R. Crystal Structures of Protooncogene Kinase Pim1: A Target of Aberrant Somatic Hypermutations in Diffuse Large Cell Lymphoma. *J. Mol. Biol.* **2005**, *348*, 183–193.

(67) Isakov, N.; Biesinger, B. Lck protein tyrosine kinase is a key regulator of T-cell activation and a target for signal intervention by Herpesvirus saimiri and other viral gene products. *Eur. J. Biochem.* **2000**, *267*, 3413–3421.

(68) Bullock, A. N.; Debreczeni, J.; Amos, A. L.; Knapp, S.; Turk, B. E. Structure and Substrate Specificity of the Pim-1 Kinase. *J. Biol. Chem.* **2005**, *280*, 41675–41682.

(69) Fabian, M. A.; Biggs, W. H.; Treiber, D. K.; Atteridge, C. E.; Azimioara, M. D.; Benedetti, M. G.; Carter, T. A.; Ciceri, P.; Edeen, P. T.; Floyd, M.; Ford, J. M.; Galvin, M.; Gerlach, J. L.; Grotzfeld, R. M.; Herrgard, S.; Insko, D. E.; Insko, M. A.; Lai, A. G.; Le'lias, J.-M.; Mehta, S. A.; Milanov, Z. V.; Velasco, A. M.; Wodicka, L. M.; Patel, H. K.; Zarrinkar, P. P.; Lockhart, D. J. A small molecule-kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol.* **2005**, *23*, 329–336.

(70) *Kinase Profiling Book*; CarnaBio USA, Inc.: Natick, MA, Feb 7, 2006; www.carnabio.com.

(71) GRID probes evaluation. DRY (hydrophobic probe), C3 (methyl CH_3 group), C1= (sp^2 CH aromatic or vinyl), N1 (neutral flat NH, e.g., amide), N2 (neutral flat NH_2, e.g., amide), N1$^+$ (sp^3 amine NH cation), N2$^+$ (sp^3 amine NH_2 cation), N$^+$ (sp^3 cationic nitrogen), O (sp^2 carbonyl oxygen), O= (O of phosphate), O−(sp^2 phenolate oxygen), O1 (alkyl hydroxy OH group).

(72) Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating optimal linear PLS estimations (GOLPE): an advanced chemometric tool for handling 3D-QSAR problems. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9–20.

(73) GRID probes best combination. C3 (methyl CH_3 group), C1=(sp^2 CH aromatic or vinyl), N1 (neutral flat NH, e.g., amide), N$^+$ (sp^3 cationic nitrogen), O1 (alkyl hydroxy OH group).

(74) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25–29.

(75) Pike, A. C. W. Lessons learnt from structural studies of the oestrogen receptor. *Best Pract. Res. Clin. Endocrinol. Metab.* **2006**, *20*, 1–14.

(76) Warne, T.; Serrano-Vega, M. J.; Baker, J. G.; Moukhametzianov, R.; Edwards, P. C.; Henderson, R.; Leslie, A. G. W.; Tate, C. G.; Schertler, G. F. X. Structure of a β_1-adrenergic G-protein-coupled receptor. *Nature* **2008**, *454*, 486–491.

(77) Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G. F.; Thian, F. S.; Kobilka, T. S.; Choi, H.-J.; Kuhn, P.; Weis, W. I.; Kobilka, B. K.; Stevens, R. C. High-Resolution Crystal Structure of an Engineered Human β_2-Adrenergic G Protein Coupled Receptor. *Science* **2007**, *318*, 1258–1265.

(78) Rasmussen, S. G. F.; Choi, H.-J.; Rosenbaum, D. M.; Kobilka, T. S.; Thian, F. S.; Edwards, P. C.; Burghammer, M.; Ratnala, V. R. P.; Sanishvili, R.; Fischetti, R. F.; Schertler, G. F. X.; Weis, W. I.; Kobilka, B. K. Crystal structure of the human β_2 adrenergic G-protein-coupled receptor. *Nature* **2007**, *450*, 383–387.

(79) Jaakola, V.-P.; Griffith, M. T.; Hanson, M. A.; Cherezov, V.; Chien, E. Y. T.; Lane, J. R.; Ijzerman, A. P.; Stevens, R. C. The 2.6 Angstrom Crystal Structure of a Human A_{2A} Adenosine Receptor Bound to an Antagonist. *Science* **2008**, *322*, 1211–1217.

CI9003317