A Common Reference Framework for Analyzing/Comparing Proteins and Ligands. Fingerprints for Ligands And Proteins (FLAP): Theory and Application

Massimo Baroni,† Gabriele Cruciani,*,‡ Simone Sciabola,‡ Francesca Perruccio,§ and Jonathan S. Mason||

Molecular Discovery Limited, 215 Marsh Road, Pinner, Middlesex, London HA5 5NE, United Kingdom, Laboratory for Chemometrics and Cheminformatics, Chemistry Department, University of Perugia, Via Elce di sotto 10, I-06123 Perugia, Italy, Pfizer Global Research and Development, Sandwich Laboratories, Sandwich, Kent CT13 9NJ, United Kingdom, and Lundbeck A/S, Ottiliavej 9, DK-2500 Copenhagen, Denmark

Received June 20, 2006

A fast new algorithm (Fingerprints for Ligands And Proteins or FLAP) able to describe small molecules and protein structures using a common reference framework of four-point pharmacophore fingerprints and a molecular-cavity shape is described in detail. The procedure starts by using the GRID force field to calculate molecular interaction fields, which are then used to identify particular target locations where an energetic interaction with small molecular features would be very favorable. The target points thus calculated are then used by FLAP to build all possible four-point pharmacophores present in the given target site. A related approach can be applied to small molecules, using directly the GRID atom types to identify pharmacophoric features, and this complementary description of the target and ligand then leads to several novel applications. FLAP can be used for selectivity studies or similarity analyses in order to compare macromolecules without superposing them. Protein families can be compared and clustered into target classes, without bias from previous knowledge and without requiring protein superposition, alignment, or knowledge-based comparison. FLAP can be used effectively for ligand-based virtual screening and structure-based virtual screening, with the pharmacophore molecular recognition. Finally, the new method can calculate descriptors for chemometric analysis and can initiate a docking procedure. This paper presents the background to the new procedure and includes case studies illustrating several relevant applications of the new approach.

INTRODUCTION

How many bond interactions must be present between a ligand and a protein for these two entities to be considered as bound to each other? Obviously, there must be at least one ligand atom able to make a sufficiently stable interaction with a protein group in the active site, but on average, the number of energetically favorable interaction points will be equal to or greater than 4.[1]

This finding has led to the development of a key concept in drug design called the pharmacophore, which is commonly defined as a three-dimensional (3D) arrangement of molecular features or fragments forming a necessary, but not necessarily sufficient, condition required for binding.[2,3] 3D pharmacophore descriptors have been successfully used for many years to represent key interactions between a ligand and a protein-binding site and enable many key drug discovery needs. These include both ligand-based and structure-based methods for virtual screening to identify new leads and the design of compounds and libraries, addressing potency; selectivity; and some adsorption, distribution, metabolism, excretion, and toxicity properties.

The use of pharmacophores derived from ligands is well-established, and many methods are available for their perception. Some approaches have recently been reviewed by van Drie,[4] and applications have been described by Martin and others.[5−10]

Pharmacophores have been widely used as inputs for 3D database searching[11,12] in order to generate new leads, for automated 3D design, and for quantitative structure−activity relationships. Their application has been further expanded by the concept of pharmacophore “fingerprints”, which represent a systematic view of the potential pharmacophores that a molecule can exhibit. 3D pharmacophores have been used as diversity and similarity tools for the design of combinatorial libraries[13−16] as well as for virtual screening (using both single defined pharmacophores and fingerprints of potential pharmacophores).[17,18]

In considering the primary molecular recognition process for a ligand to its binding site, the electrostatic properties will be a key driving force, and a 3D pharmacophoric descriptor provides a simplified representation of that, through the different features that are distinguished and quantified in 3D space. Commonly used features are hydrogen-bond acceptors and donors, positively (basic) and negatively (acidic) charged groups, and hydrophobic and aromatic areas; 3D “shapes” of these features thus are powerful representations of key properties for molecular recognition and binding. The “shape” element is evident for

[*] Corresponding author phone: +39-0755855550; fax: +39-07545646; e-mail: gabri@chemiome.chm.unipg.it.
[†] Molecular Discovery Limited.
[‡] University of Perugia.
[§] Pfizer Global Research and Development.
[||] Lundbeck A/S.

10.1021/ci600253e CCC: $37.00 © 2007 American Chemical Society
Published on Web 01/09/2007

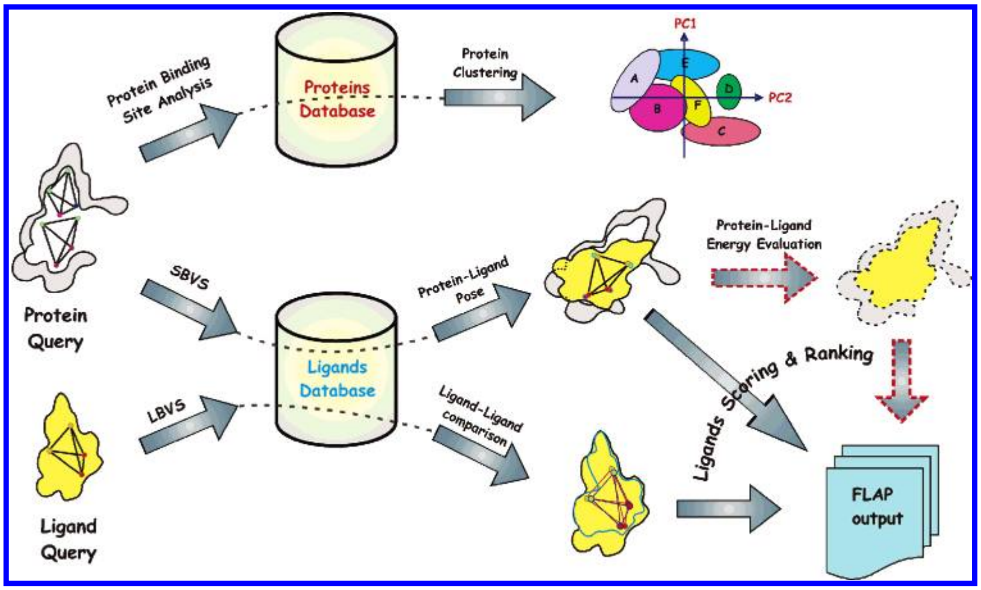


Figure 1. Several applications presented by FLAP. It can be used to investigate protein similarity, for structure-based virtual screening (SBVS), ligand-based virtual screening (LBVS), and to eventually optimize the pose computation in docking studies.

pharmacophores with four or more points (four = a 3D shape that can distinguish chirality, a tetrahedron; three = planar, a triangle; two = vector, a line).

While the use of pharmacophore descriptors from just ligands is very powerful, with many reported impactful uses in the drug discovery process as discussed above, the use of pharmacophore descriptors derived from protein-binding sites enables a common frame of reference to be used for both ligands and receptors. Thus, by generating pharmacophore feature points that are complementary to a binding site, an image that represents the "perfect" ligand is produced; fingerprints or models produced from such complementary points can thus be directly used together with pharmacophoric information from ligands.

Mason and Cheney were the first to generate 3D pharmacophore fingerprints complementary to a protein binding site, to provide this common frame of reference framework for the analysis of both ligands and their binding sites. They used them to drive docking and to compare serine protease binding sites and ligands of different selectivities. A GRID[19−21] force field analysis was used with only a semiautomated generation of the complementary site points, an automated generation of the pharmacophore fingerprint from these site points, but without further using the shape of the sites during the comparisons (the four-point pharmacophores used encode some shape information). When using four-point (but not three-point) pharmacophores, they were able to differentiate ligand binding and match ligands in the correct similarity order to their different serine protease targets. However, 3D pharmacophore matching alone is unlikely to be sufficient for binding, as shape is known to be a fundamental characteristic for addressing target-ligand selectivity, and clearly, there can be many invalid pharmacophore matches, in that a steric clash of the ligand to the protein will occur when the common features are aligned.

There is thus a need for automated methods of generating and using complementary pharmacophores of protein binding sites together with their shape- and ligand-based pharmacophores. This requirement has now led to the development of a new advanced procedure (Fingerprints for Ligands And Proteins or FLAP), which encompasses several additional capabilities, such as protein selectivity and protein similarity studies, and fast generation of lattice-independent molecular descriptors for 3D quantitative structure−property relationships. The FLAP approach enables the automated identification of the potential complementary ligand pharmacophoric features for a protein binding site, with additive effects from multiple residues in the context of the 3D structure of the site automatically being taken into account, and flexibility of the side chains in response to the probe group are also able to be taken into account to include cooperative effects; this site context enables an enhanced approach over, for example, approaches that use some type of lookup table for interactions from residues in the site.

The FLAP method described in this paper is a major advance as the shape of the site is taken into account in comparing both ligand and protein target fingerprints, and the generation of the fingerprints can be automated. The successful application to virtual screening, docking, and protein clustering is described here and in a related paper.

MATERIALS AND METHODS

FLAP has been developed to exploit the relevant information from crystallographic structures, from virtual screening molecules and/or from their complexes. The method comprises the quantification of the (macro)molecular fingerprints which allow the application of rational strategies to generate de novo virtual structures or to compare and cluster protein families without any bias from previous knowledge. Many different applications are possible using this new method, and a flowchart summarizing the main case studies is reported in Figure 1.

A. Target-Based Pharmacophores Obtained from Molecular Interaction Fields. Many computational techniques have been developed to exploit the relevant information in X-ray crystallographic structures. A widely used method is the GRID program in which a more realistically shaped and charged probe, with a predefined hydrogen-bonding pattern, is used instead of a more traditional neutral sphere. These

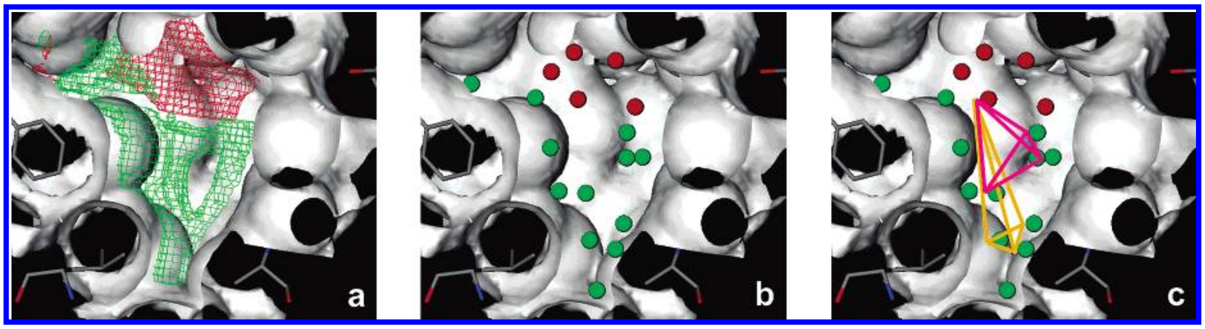


Figure 2. (a) Molecular interaction fields (MIF) calculated in the active site of a protein structure. (b) MIF is condensed in few target-based pharmacophoric points. (c) All possible arrangements of four pharmacophoric points are generated.

new probes were carefully parametrized using high-resolution X-ray data and their energetic interaction with the protein structure called molecular interaction fields (MIFs). MIFs can be drawn at various energy levels by using negative-energy-level contours representing the attractive regions of the protein surface. Other programs may be used to compute MIFs, but the GRID program is state-of-the-art and one of the most used programs in the field of structure-based ligand design.[22]

GRID maps were developed with the aim of predicting where ligands would bind to biological macromolecules and, so, improve the user's understanding of the factors involved in binding. A consequence of this improvement in understanding should also be the design of improved ligands. MIFs describe the spatial variation of the interaction energy between a (macro)molecular target and a chosen probe. The target may be a macromolecule, a low-molecular-weight compound, or a molecular complex. The probe may be a molecule, a fragment of a molecule, or a single atom.

In a small molecule, a pharmacophore can be defined by the atoms which may have critical interactions with a target receptor together with their relative spatial arrangement. Such atoms, when isolated from the molecular context and positioned in an active macromolecule site, may interact attractively or repulsively with the target, and their precise behavior is simulated by GRID probes and recorded in the corresponding molecular interaction field. Therefore, a ligand-based pharmacophore can be viewed as being a set of GRID probes together with their relative spatial orientation.

A pharmacophore can be obtained by using an alternative method that examines the binding site of the macromolecular target itself. By studying the MIFs produced by different GRID probes in the target binding site, a negative pharmacophoric image can be generated. For example, if it is assumed that the binding site contains positively charged, hydrogen-bond donor and hydrophobic amino acids, these groups would be capable of forming ionic bonds, hydrogen bonds, and attractive hydrophobic interactions respectively, coded in their MIFs. Therefore, the drugs that would interact with this binding site would contain all or some negative-charged group(s), hydrogen-bonding acceptor group(s), and hydrophobic group(s).

The required pharmacophore for such drugs could be defined by identifying the regions in the target space that are potentially able to accommodate the chemical probes contained in the drug molecules. Therefore, by using GRID probes on a macromolecular target, target-based pharmacophores are automatically produced within a reference framework of probe-based pharmacophores in ligand molecules. This approach can be very useful because now all (macro)molecules can be compared with all the others, as reported below, using descriptors relevant to ligand binding. Moreover, promising applications can be foreseen, such as the analysis of novel targets being discovered by genomic projects. Here, the need is to identify a lead compound as quickly as possible, the targets are totally novel, and the natural ligand or chemical messenger is not known.

B. The New Procedure. FLAP is a new computational procedure able to explore the 3D-pharmacophoric space of ligands and proteins and to provide quantitative information for the complementarity of their interactions, using common reference frameworks to allow ligand−ligand, ligand−protein, or protein−protein comparison.

The FLAP procedure analyzes a protein cavity by using GRID molecular interaction fields (GRID-MIFs) obtained by running a limited series of chemical probes over a part or all of the proteins (see Figure 2a). As far as possible, the probes should be chosen so that they represent all possible interactions between functional groups of potential ligands and the amino acids that occur in the receptor cavity. The entire process is based on the assumption that the recognition/affinity of one molecule for the binding site is an additive function of the recognition/affinity of the individual atoms that contribute to the interaction process. The probes enable the identification of energetically both favorable and unfavorable interactions (e.g., for selectivity etc.).

The information contained in the GRID-MIFs is then condensed into fewer target-based pharmacophoric points by using a weighted energy-based and space-coverage function. Figure 2b illustrates this step of the procedure. As illustrated in Figure 2c, FLAP now generates all possible energetically favorable arrangements of four pharmacophoric points in the regions chosen to map the receptor.

In terms of evaluating the affinity between the molecule and the target, and given a certain conformation of the potential ligand, a favorable event is formed by the simultaneous positioning by FLAP of four atoms in the energetically favorable areas (3D-MIF), combined with an (optional) absence of repulsion between all of the other atoms and the target. Other properties can be associated with this favorable event such as the value of the sum of the energy associated with four favorable ligand atom−target interactions. The larger the number of favorable events, the greater the probability of ligands binding with targets and the greater the likely affinity between the two.

Because the molecular interaction fields in the target produced by the GRID force field are used to identify the

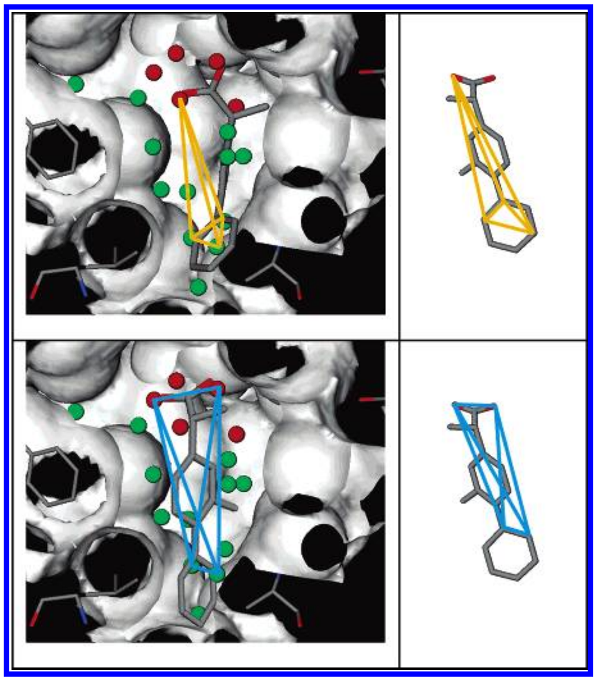


Figure 3. Fit of two conformers of a ligand over their corresponding target-based pharmacophoric points.

most favorable positions for each atom of the potential ligand, it is obvious that the method chosen to classify atoms of the potential ligand is simply that based on GRID atomic types. Because as many probes are available as ligand atom types, ideally, there are no limitations on the selection of probes. However, in reality, this route cannot be taken at present, because several reasons combine to prevent more than a total of six probes being used. It is therefore important to choose these six probes as appropriately as possible, from all of the possible probes in the program GRID, so that a good description of the main types of interaction is obtained.

Once the probes have been selected, each atom of the ligand needs to be associated with the probes that best approximate its behavior inside the receptor. FLAP normally does this automatically, but the procedure can be personalized by the user if requested. A potentially energetically favorable ligand−receptor interaction is likely when a molecule in a defined conformation possesses four atoms disposed in space so that they are superimposable on many pharmacophoric points of the same type (see Figure 3).

This leads to the impression that the FLAP software is able to directly effect a type of docking, but the reality is different. In fact, while the energy between a ligand and a protein in its complete form is estimated in a docking program, FLAP only uses a few contributions at a time, considering four anchorage points to be a necessary and sufficient condition as well as the optional absence of a number of atom/macromolecule repulsions above a certain threshold. It thus produces “poses”, which can be further optimized to “docked” fits to a target binding site. The other fundamental difference is that FLAP operates in an absolutely discrete space, extracting only the essential points to describe the GRID-MIFs by approximating them sufficiently closely. These complementary pharmacophore features do enable a powerful way to match ligands to proteins, by looking for similar patterns, providing a way to fit (ready for docking optimization) compounds using a primary recognition driving force that is more electrostatic (binding features) in nature than the shape-based approaches to docking. Complementarity of the shape is evaluated in the next step, with conformational flexibility of the ligand in the site being possible. If molecular recognition is at least partially driven by these 3D electrostatic/binding forces, then this approach should provide the basis of an effective method of docking, with the potential to outperform methods that work in reverse; that is, docking is driven primarily by shape matching, with the complementarity of electrostatics/pharmacophoric features as the second step. Encouraging results are described below.

C. The FLAP Fingerprints. Only the ligand−target case will be referred to from here onward, but it will be shown how the ligand−ligand and target−target methods are natural extensions of the approach now described.

For each of the fields produced by the GRID probes in the selected space containing the active site, the group of most important points is selected from both the energetic and the spatial point of view. The selection can be made automatically or manually, and this is a critically important stage in the overall job because, de facto, it is when the potential target-based pharmacophoric points within the binding site are chosen. The importance and delicacy of this selection is accentuated because there must not be too many target points (usually no more than 25−30 selected points per probe) so that combinatorial problems are to be avoided. Mapping of the receptors using chosen pharmacophoric centers is therefore essential and must be precise, while at the same time, it must also be sufficiently informative.

By the end of this procedure, a total of perhaps 100 or 200 points may have been selected, and each one of these is characterized by several different properties: (i) the type of energetic interaction (e.g., hydrophobic, hydrogen-bond donors, etc.), (ii) its entity (e.g., interaction energy value in kilocalories per mole), and (iii) three Cartesian coordinates that identify its position in the receptor.

FLAP starts with a combination of these points and produces a mathematical model composed by the group of all possible 3D pharmacophoric configurations obtained by combining four points. When a ligand molecule in a low-energy conformation possessed four atoms disposed in space so that they are superimposable on as many pharmacophoric points of the same type on a receptor, a potentially energetically favorable ligand−receptor interaction has been detected.

Obviously, a perfect superimposability of the ligand atoms and the corresponding pharmacophoric points on the receptor is hardly ever found in practice, so a certain degree of approximation in the Cartesian space needs to be accepted when the positions of the points are evaluated. For example, if the two tetrahedrons formed by combining four atoms and four points are superimposed, the result is considered satisfactory if each atom is less than 1 Å away from the corresponding point. If a more precise superimposition is required, the threshold distance can be reduced to 0.5 or 0.25 Å, and two entities would be considered spatially coincident if enclosable in a 0.25 Å sphere.

Accordingly, FLAP models the receptor as a group of quadruplets (N.B., in mathematics, a tuple is a finite group of objects and a quadruplet is written as 4-tuple) of objects each belonging to one or more types (the number depends on the number of fields used to map the site) which in space

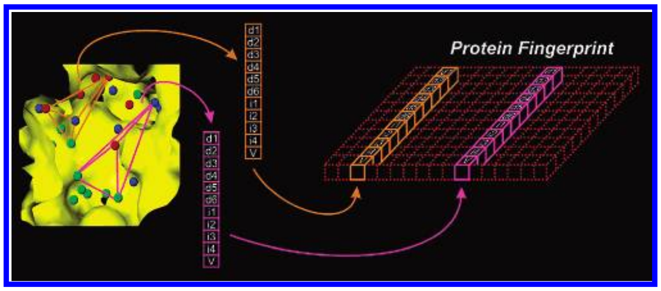


Figure 4. All the possible target-based four-point pharmacophores built within the active site of a protein recorded in a protein fingerprint. The process produces a large matrix, called a protein fingerprint.

assume relative chiral positions and are further represented by a vector of six integers, corresponding to as many multiple lengths of a value Δ (to represent distance).

In the past, this receptor perception was stored in a four-point pharmacophore fingerprint-type approach,[17,18] associating a bit in a bitstring of prefixed length to each spatial configuration, chirality, and combination of atom types/features (typology) of the points. An analogous bitstring would then be extracted for a ligand, in a defined conformation, and the degree of ligand−target affinity would be determined by a parameter extracted from the number of common bits in the two bitstrings.

This method was quite fast but had serious limitations. First of all, to limit the number of bits in the string, the number of distance bins used had to be limited (e.g., to 7 or 10). Furthermore, the description does not consider the "physicality" of the bit, which is seen as a spatial configuration of points without considering either from where they are effectively placed or how many effectively correspond to that bit. As the fingerprints were very sparse (hundreds to thousands of bits set from possible millions), just the set bits were usually saved (coded by the four feature types and the six distance bins). FLAP overcomes these limitations by using a new technology to encode the bitstring. The receptor bitstring in FLAP only exists "conceptually". De facto, it uses diversified data structures for the chirality (positive and negative volumes) and for the typology combination of atoms types of the points. An array of variable lengths is therefore generated for each of these possible combinations. This array is composed of consecutive packets of 11 integer values (limited to the range 0−255 as only one byte is associated with 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 respectively assigned so that they contain the six values of the distances, the indices of the four site points (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 four points. There is no need to introduce information concerning the type and chirality because the array is specific to a certain type of data sequence of atoms/site points. Therefore, the following vector is associated with the four given points A, B, C, and D of a certain type and a tetrahedron of a determined volume (see Figure 4).

During the process of pharmacophoric modeling of the receptor, all the site points could be combined in all possible ways by considering all permutations of the 4-tuples. The sign of the tetrahedron volume is always calculated for each one of these (using the exact coordinates of the four vertices), and this together with the typology of the four points will determine the array to which the vector of the 11 described integer values will be added.

As previously mentioned, the length of the edges on each tetrahedron occupying the first six places are the rounded off values of the chosen approximation Δ. Because only one byte is reserved for each length, the maximum representable quantity is equal to $255 \times \Delta$. Moreover, this is a value sufficient for very small values of Δ (e.g., 0.2 Å). The procedure for constructing vectors terminates with an internal reordering of the 11 byte packets based on the first six values only, which is essential for fast searches to be effected.

In practice, given two tetrahedrons, all the paired distances would be compared, prioritizing the six values in decreasing order. If $d1_{AB} < d2_{AB}$, then the first tetrahedron is placed before the second. If, on the contrary, $d1_{AB} > d2_{AB}$, then the order is reversed. However, if the two values are equal ($d1_{AB} = d2_{AB}$), then $d1_{AC}$ and $d2_{AC}$ are compared, and should they be equal, the search moves on to $d1_{AD} < d2_{AD}$, and so on. This ordering allows searches to be performed using very efficient and fast search algorithms. Finally, once the model has been constructed, the degree of pharmacophoric similarity of the receptor can be evaluated using any molecule. Once the atoms of the potential ligand have been classified by associating the probes used to map the site to them, they combine to form all the quadruplets possible but without effecting any permutations; that is, the four different atoms bring about only one tetrahedron and not 12 as is the case with the receptor.

Each combination of atoms will be characterized by both a determined "typology" and a determined chirality. These two elements are immediately used to identify which 11-tuplet ordered vector will be used to search for any correspondence between the six interatomic distances, rounded off to an integer, and the first six elements of each 11-byte packet. When the search is successful, it does not stop, however, because more equivalent tetrahedrons may be present but generated by other points. More information on this topic is provided in the manual distributed with the FLAP[23,24] software and in the technical details section.

D. The Elimination Stage: Using Shape. 3D pharmacophore matching is not necessarily sufficient to ensure binding because shape is also a fundamental characteristic influencing target−ligand selectivity, and there can be no valid ligand−protein pharmacophoric complementarity when steric clashes are present.

An important key feature in FLAP is therefore its ability to bias that pharmacophoric complementarity selection by using both the shape of the ligand and that of the receptor. After determining all possible quadruplet interactions, FLAP identifies all the ways in which up to four atoms of a ligand could make attractive interactions with the target. The overall number of saved positions is usually several thousand, but many of these can be eliminated very quickly if shape constraints are taken into consideration. Hydrogen GRID maps are used for this job, leaving only a few possible modes of interaction for further detailed consideration.

FLAP finally evaluates the magnitude of any remaining steric clashes and allows a certain amount of protein or ligand

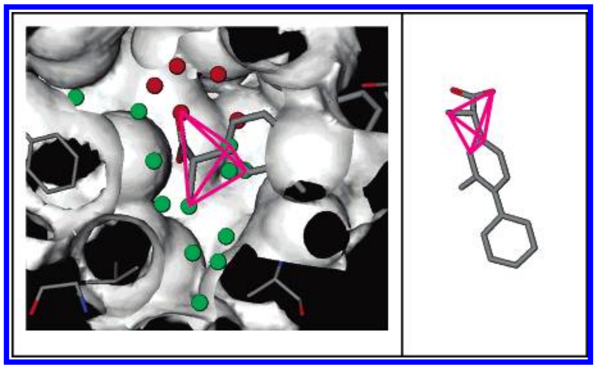


Figure 5. Steric hindrance acting as a filter for the many solutions found for each ligand when "docked" in the protein active site.

flexibility before docking. At the end of the process, if unacceptable clashes are still present, like those reported in Figure 5, the pose is eliminated.

E. Protein Flexibility. Flexible proteins may require a conformational sampling of side-chain amino acids, and the FLAP method uses "on-the-fly" generation of side-chain conformations while the protein site points are being generated.

When flexibility occurs, hydrophobic amino acids will tend to move in toward the hydrophobic group represented by the hydrophobic probe, while polar amino acids will tend to move in the opposite direction toward the aqueous environment. Using this model, what actually happens depends on the overall balance between these two effects.

When the flexible option is turned on, favorable interaction regions of the GRID maps become larger, while regions of steric hindrance in the maps become significantly smaller. Thus, the maps represent dynamic behavior of the protein in response to the probe movements. It is important to note that, because the maps are different, the 3D locations of the energy minima points are also different, and so the pharmacophoric points of the quadruplets are changed.

The conformation initially assigned to the flexible side chains of the target by the user is no longer critically important when flexibility is permitted in a GRID computation, because the program will not use the initial torsion angles when dealing with a flexible chain. However, care must be taken not to place flexible atoms unacceptably close to each other (within van der Waals touching distance!), and it is therefore important to choose a "sensible" starting conformation for the target.

RESULTS AND DISCUSSION

Structure-Based Virtual Screening (SBVS). The FLAP program starts by identifying the pharmacophores that are common between a ligand and a putative active site. First, the protein pharmacophores are generated from MIFs and recorded together with the shape of the cavity. The probes used for the calculations of the molecular interaction fields within the protein active site are normally selected by default, as reported in Table 1. All this is usually done "on the fly", although the user can customize the selection of the probes and may especially want to do this if an interaction between the target and some bound ligand is of particular interest.

Conformational sampling methods (random or systematic) are used to generate the ligand pharmacophores, because flexible ligands need conformational sampling. On-the-fly generation of the conformers is carried out automatically at search time, and a quick evaluation of each conformation is performed on the basis of an internal steric contact check to reject poor or invalid structures. The method automatically selects rotamers so that their modifications will produce the maximum variation of molecular atomic positions. Once these initial rotamers have been selected, a random perturbation generates a population of possible rotamer solutions, or as an alternative to this random generation, the user can select a systematic search method. In the latter, customizable angular steps and steric bump factors can be selected to tune the number of solutions. Moreover, the systematic search solutions can be selected in order to reduce the final number of rotamers. Then, in contrast to many other pharmacophoric methods which append the fingerprint for each of the conformers in a unique resulting fingerprint, FLAP produces a single fingerprint for each of the molecule conformations. Finally, for the pharmacophores of each conformation of each ligand under investigation, FLAP computes protein−ligand matches with all the possible pharmacophores of the putative active site of the protein. Keywords such as regions can be used to define a sphere within which each pharmacophore needs to have at least one point, and the selection of a particular probe can also be made.

FLAP has a unique integrated feature that can be used to "bias" (filter) the generation of ligand conformers. They are generated not only to populate the conformational space but also to match the ligand pharmacophore with the shape and the chemical features of the protein cavity. FLAP docks the ligand pharmacophoric features into the protein pharmacophoric features, and the resulting matches are accepted only when they show shape complementarities and feature complementarities. The resulting matches are thus strongly biased by protein−ligand shape similarity, and the new docked ligand coordinates are written to a file, together with the number of matches and other similarity indicators (see Figure 6).

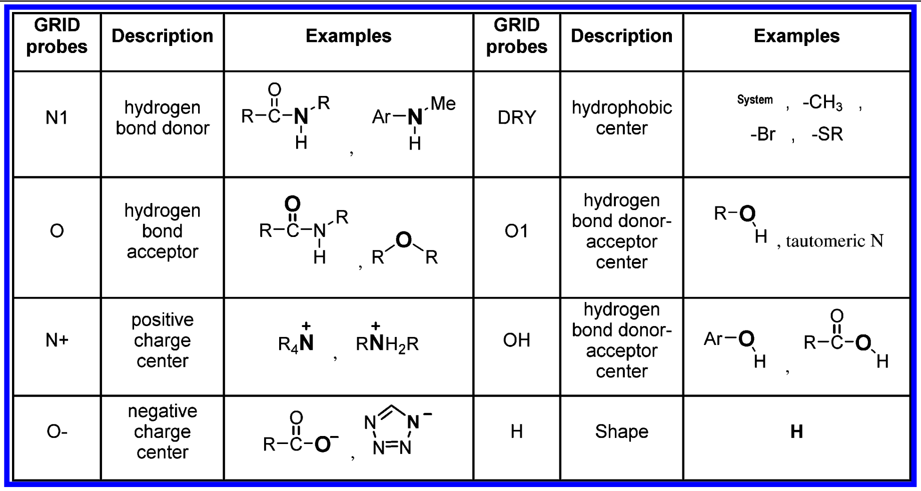
The first example of structure-based virtual screening using FLAP was performed on Factor Xa ligands. The crystal structure of protein entry 1nfu was downloaded from the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB)[25] and used as a target, with 22 known active compounds as ligands (including the 10 public compounds reported in Table 2). The data set was completed by 1138 compounds from the MDL Drug Data Report (MDDR) database from which compounds were randomly selected with the constraint that they had to have a similar molecular weight to that of the known active compounds.

Although FLAP is not as such a docking method, here it was compared with three well-known docking methods, that is, GOLD,[26] GLIDE,[27] and DOCK[28,29] (which in their "fast" modes for SBVS also approximate the docking process).

Figure 7 reports the enrichment factor obtained after applying the methods to the same data set. It is interesting to note that the results produced by FLAP were obtained in 1.5 h, compared to the 19 h for GLIDE, 24 h for GOLD, and 12 h for DOCK.

Thymidine kinase (TK) was the second protein studied, and this is a challenging test case because of its flexible active site and the presence of several water molecules that participate in ligand binding.[30] Starting from the coordinates of the protein complexed with deoxythymidine (PDB code:

Table 1. Some of the Standard Probes Used to Compute MIF within the Protein Active Site



GRID probes	Description	Examples	GRID probes	Description	Examples
N1	hydrogen bond donor	R–C(=O)–N(H)–R , Ar–N(H)–Me	DRY	hydrophobic center	System , $-CH_3$, -Br , -SR
O	hydrogen bond acceptor	R–C(=O)–N(H)–R , R–O–R	O1	hydrogen bond donor-acceptor center	R–O–H , tautomeric N
N+	positive charge center	$R_4\overset{+}{N}$, $R\overset{+}{N}H_2R$	OH	hydrogen bond donor-acceptor center	Ar–O–H , R–C(=O)–O–H
O-	negative charge center	$R–C(=O)–O^-$, tetrazolate	H	Shape	H

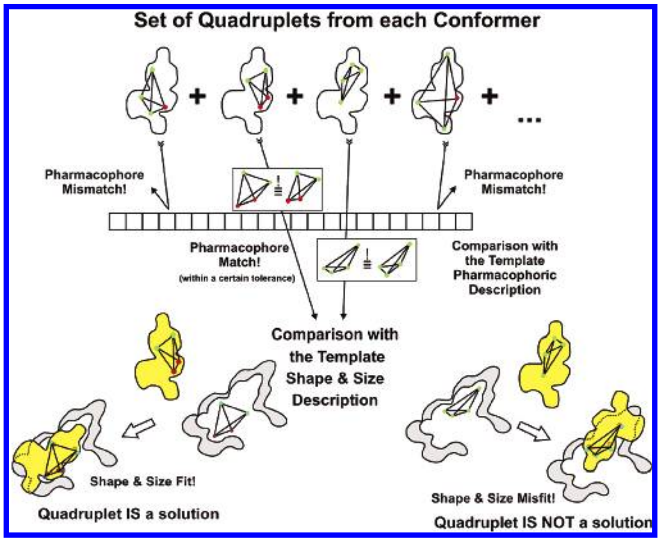

Set of Quadruplets from each Conformer
+
+
+
+
...
Pharmacophore Mismatch!
Pharmacophore Mismatch!
Pharmacophore Match! (within a certain tolerance)
Comparison with the Template Pharmacophoric Description
Comparison with the Template Shape & Size Description
Shape & Size Fit!
Quadruplet IS a solution
Shape & Size Misfit!
Quadruplet IS NOT a solution

Figure 6. Flowchart of the SBVS procedure using FLAP.

1kim), a 1000-compound database containing 10 known actives (see Table 2) was generated. The 990 other compounds were randomly extracted from the MDDR database with the same constraint that they had to have a similar molecular weight to that of the 10 true active compounds.

As before, the virtual screening performances were tested and compared using the DOCK, GOLD, and GLIDE docking methods. As shown in Figure 8, very satisfactory results were obtained using FLAP in only a fraction of the time required by the other methods (0.5 h for FLAP, against the 6 h for GLIDE and GOLD and 10 h for DOCK).

It is interesting to note that TK ligands demonstrated the most pharmacophoric and shape overlap, with more than 80 common solutions using four-point pharmacophores, while nonligands clearly lacked complementarity, reporting an average of less than 20 solutions.

Ligand-Based Virtual Screening (LBVS). In ligand-based virtual screening, ligands are compared with each other using a method similar to that used for comparing ligands and a protein structure.

FLAP computes the ligand pharmacophores and is able to identify the pharmacophores that are common between a ligand template and the other ligands under investigation. Ligand−ligand complementarity may be generated using conformational sampling biased by using both shape complementarity and feature complementarities with one or more template molecules. The shape can be defined around a unique template molecule, or around a combination of template molecules. The resultant matches are then written to a file, as described above.

If the 3D structure of the target under investigation is known, another possible approach is to compare ligands by using the shape of the protein as a shape constraint and features in the protein cavity as additional constraints. As in the case of structure-based virtual screening, keywords such as regions are used to define a sphere within which each

Table 2. Public Ligand Structures Used for Virtual Screening Analysis[a]

target	PDB entry	resolution (Å)	ligand name
Factor Xa	1ezq	2.20	RPR[b]
	1f0r	2.10	815[b]
	1f0s	2.10	PR2[b]
	1ksn	2.10	FXV[b]
	1lpk	2.20	CBB[b]
	1lqd	2.70	CMI[b]
	1nfu	2.05	RRP[b]
	1nfw	2.10	RRR[b]
	1nfy	2.10	RTR[b]
			Tapap
Thymidine Kinase	1e2k	1.70	TMC[b]
	1e2m	2.20	HPT[b]
	1e2n	2.20	RCA[b]
	1e2p	2.50	CCV[b]
	1ki2	2.20	GA2[b]
	1ki3	2.37	PE2[b]
	1ki6	2.37	AHU[b]
	1ki7	2.20	ID2[b]
	1kim	2.14	THM[b]
	2ki5	1.90	AC2[b]
Estrogen Receptor α	1sj0	1.90	E4D[b]
	1uom	2.28	PTI[b]
	1xp1	1.80	AIH[b]
	1xqc	2.05	AEJ[b]
	3ert	1.90	OHT[b]
			EM-343
			LY-357489
			nafoxidene
			sumitomo-biphenol
			ZK-11901

[a] 3D-SDF ligand coordinates are listed in the Supporting Information. [b] Ligand PDB code.

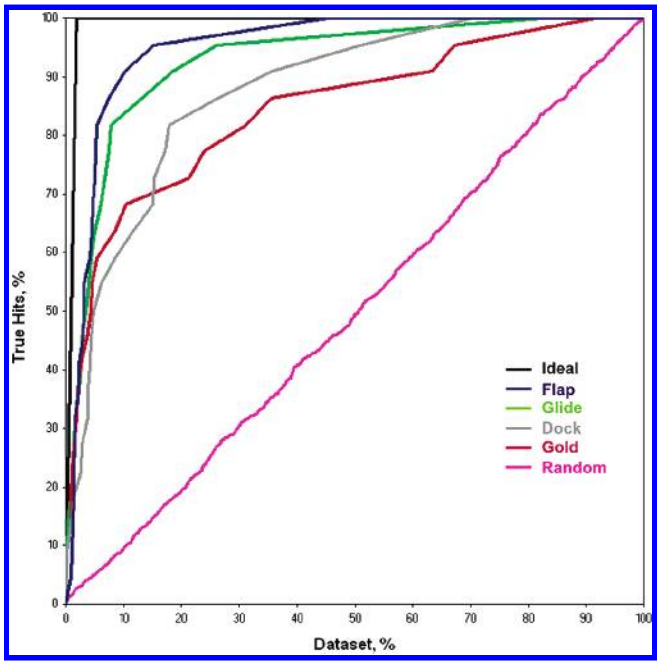


Figure 7. Enrichment factors (Factor Xa) obtained using different docking methods compared with FLAP.

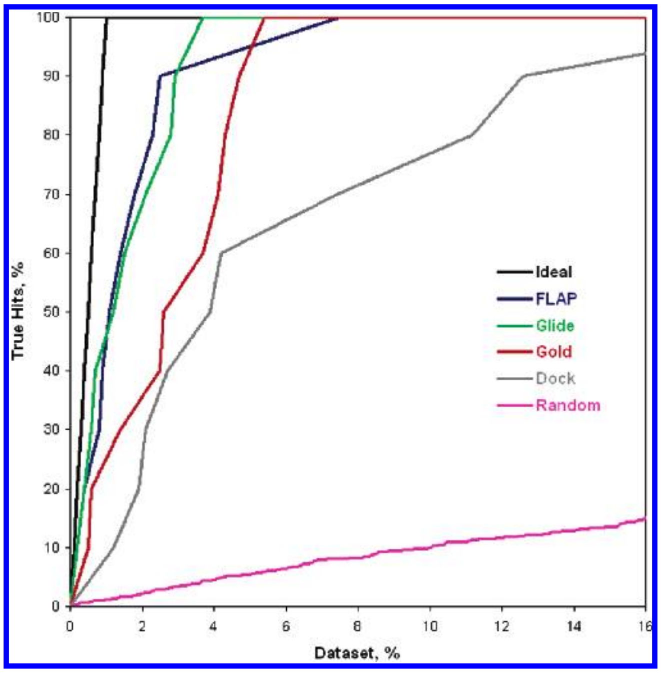


Figure 8. Enrichment factors (thymidine kinase) obtained using different docking methods compared with FLAP.

pharmacophore needs to have at least one point, and the selection of a particular probe can also be made.

The first example of ligand-based virtual screening performed using FLAP was an in-house project at Pfizer, Sandwich Laboratories. For reasons of confidentiality, details of the structures for the project cannot be disclosed. Both active ligands (338 active) and inactive structures (833 inactive) were available.

Seven different pharmacophores were built in FLAP representing different series of active ligand molecules compared to the same target (see the Computational Methods section). The virtual screening ranked the seven pharmacophores one at a time so that a given library was ranked. The results from the seven runs of ligand-based virtual screening were then merged summing up the FLAP score from each library. The corresponding procedure was applied in Catalyst[31] in order to compare the two methods (see Figure 9 and the Computational Methods section).

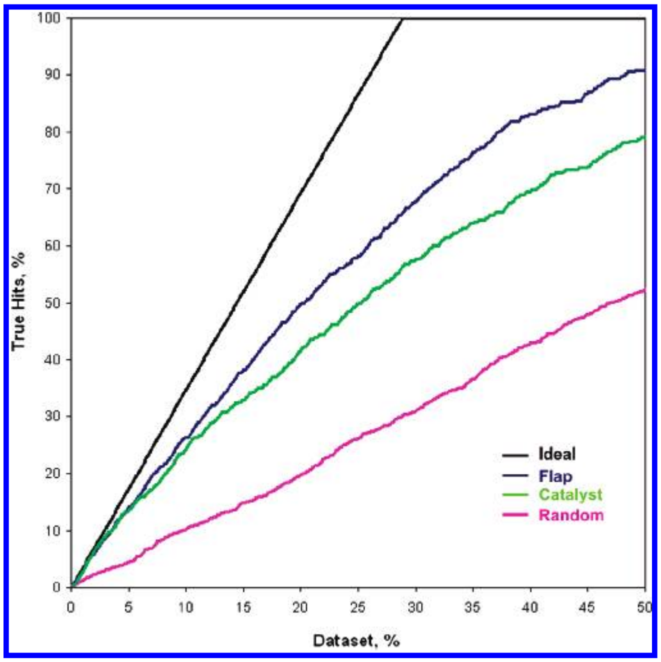


Figure 9. Enrichment plot comparison between Catalyst and FLAP.

The second example was conducted using the ERα receptor, which has already been successfully used as virtual screening target.[30,32] Starting from the coordinates of the ERα receptor complexed with 4-hydroxy-tamoxifen (PDB: 3ert), a 1000-compound database containing 10 known ERα antagonists was generated (see Table 2). The 990 compounds were randomly extracted by the MDDR database with the constraint of having to have a similar molecular weight to that of the 10 known active compounds.

The virtual screening performance was tested and compared using Catalyst, FLAP, ChemX,[33] and Barnard[34] fingerprints. Computational details and protocols are reported in the Computational Methods section below, and very satisfying results were obtained using FLAP (Figure 10), where all the actives were found after screening only 2% of the database.

Pairwise Protein Similarity. With the recent advances in genomics and protein structure determination methods, a wealth of new sequence and structural data is now available. Mining this information for familial resemblance and other protein characteristics offers several novel opportunities for drug design. For example, having reference sets of related sequences and/or structures can be extremely helpful in homology modeling. Furthermore, pharmacophoric markers can be developed for protein families, and these markers can then be used to virtually screen sets' compounds for inhibitory potential against any novel target in a family.

An earlier study of serine proteases by Mason and coworkers[13,14] used site-derived fingerprints to quantify the range of different pharmacophoric complementarities of protein binding sites and illustrated the large differences in 3D pharmacophoric fingerprints between related targets. These can be exploited for selectivity, whereas common pharmacophores would represent common binding motifs.

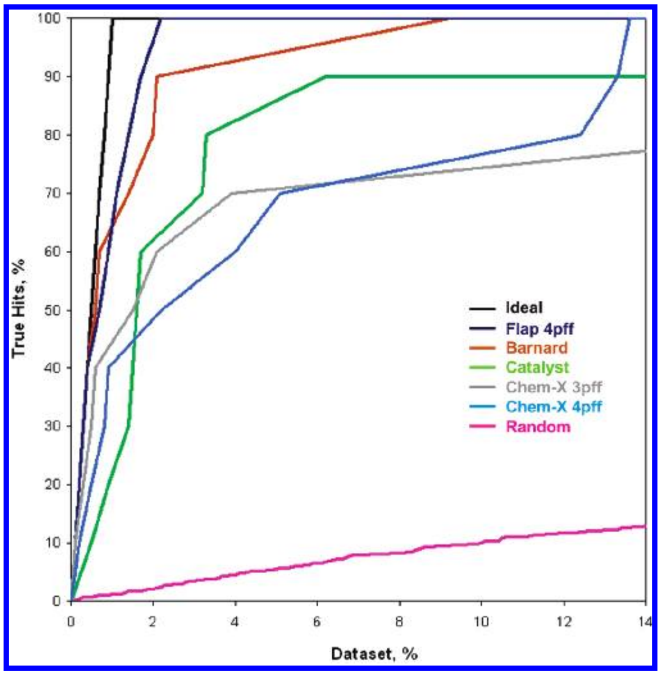


Figure 10. Enrichment plot (ERα receptor) using FLAP compared with Catalyst, Barnard, and ChemX.

FLAP is able to compare and cluster protein families into target classes, without any bias from previous knowledge. It is important to note that FLAP utilizes only the 3D structure of the proteins described by MIFs and does not require protein superposition, alignment, or knowledge-based comparison.

The Kinase Case Study. Kinases are involved in the regulation of all aspects of cellular functions, and kinase inhibition is a widely applied strategy for the treatment of various diseases. However, with over 500 kinases in the human genome, the interpretation of selectivity is a daunting task. In this field, the pharmacophoric and shape similarity of the proteins can play a central role, driving a better understanding of the structural features which govern binding and inhibition, and thus directly help in the search for novel selective inhibitors. A total of 23 target kinases belonging to four distinct kinase families (CDK2, GSK3*β*, P38α, and LCK) were selected for this study and are reported in Table 3.

It was decided to include two well-known serine−threonine subfamilies: cyclin-dependent kinase 2/cyclin A (CDK2−CyclinA) and glycogen synthase kinase 3 *β* (GSK3*β*). Previous studies[35] regarding these two kinases have shown that CDK2 inhibitors are also quite potent on GSK3*β*, and a major problem in discovering selective CDK2 kinase inhibitors is achieving selectivity by avoiding interaction with GSK3*β*.

P38α mitogen-activated protein kinase (MAPK) was included in the training set because selective compounds can lead to increased activity of proinflammatory cytokines such as tumor necrosis factor and interleukin 1*β*. Selective inhibition of this kinase subfamily could be a therapeutically useful route in treating a number of inflammatory and autoimmune diseases.[36]

Finally, given the important role of tyrosine protein kinases (TKs) in the regulation of cell proliferation, malignancy, and signal transduction, it was decided to add the lymphoid cell kinase (LCK) subfamily of TKs. The importance of LCK lies in the fact that it regulates T-cell maturation and activation and is perhaps the best studied and best understood member of the cytoplasmatic, nonreceptor tyrosine protein kinases of the Src family.

Table 3. The 23 Protein Kinases Studied

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
1ol1	CDK2	Ser/Thr	2.9
1okv	CDK2	Ser/Thr	2.4
1oku	CDK2	Ser/Thr	2.9
1ouy	P38	Ser/Thr	2.5
1r3c	P38	Ser/Thr	2.0
1w7h	P38	Ser/Thr	2.2
1wbo	P38	Ser/Thr	2.2
1w84	P38	Ser/Thr	2.2
1q3d	GSK3*β*	Ser/Thr	2.2
1q3w	GSK3*β*	Ser/Thr	2.3
1q5k	GSK3*β*	Ser/Thr	1.9
1pyx	GSK3*β*	Ser/Thr	2.4
1h8f	GSK3*β*	Ser/Thr	2.8
1i09	GSK3*β*	Ser/Thr	2.7
1qpc	LCK	Tyr	1.6
1qpd	LCK	Tyr	2.0
1qpe	LCK	Tyr	2.0
1qpj	LCK	Tyr	2.2
3lck	LCK	Tyr	1.7

The intrafamily target selection was based on the following criteria: the availability of a human source of pharmaceutical interest, availability of an X-ray diffraction technique, a resolution of less than 2.5 Å (only four PDB entries reported resolution larger than 2.5 Å), and when possible, B factors 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.

All of the kinase targets were pretreated by the GRID force field using five chemical probes: H-bond donor (NH of amide), H-bond acceptor (C═O), negative charged (O═), hydrophobic probe (Dry), and shape probe and the four-point pharmacophore fingerprints were subsequently generated from the MIFs obtained. By way of an example, Figure 11 illustrates the contours and the pharmacophore features for the 1H1S kinase active site.

The ensemble of favorable MIF locations, called “hotspots”, was treated as a hypothetical molecule that interacts at all favorable positions in the binding site, and it’s pharmacophore fingerprint was calculated and analyzed from these hotspots in the same way as for any ordinary ligand. For example, the CDK2 kinases 1OI9 and 1OIU showed 72 and 71 hotspots that respectively lead to 1 094 498 and 974 882 four-point pharmacophores, with 63 084 in common, decreasing to only 18 432 when shape filtering was turned on.

Similarly, the P38 kinase 1W84 showed 90 hotspots (10 DRY, 28 N1, 22 O, 30 O═), leading to 2 660 998 four-point pharmacophores, with 55 209 in common with 1OI9 and 48 694 in common with 1OIU, reduced to 1071 and 1011, respectively, with shape filtering on.

Common pharmacophores such as these provide a set of useful common binding motifs that can be used to drive

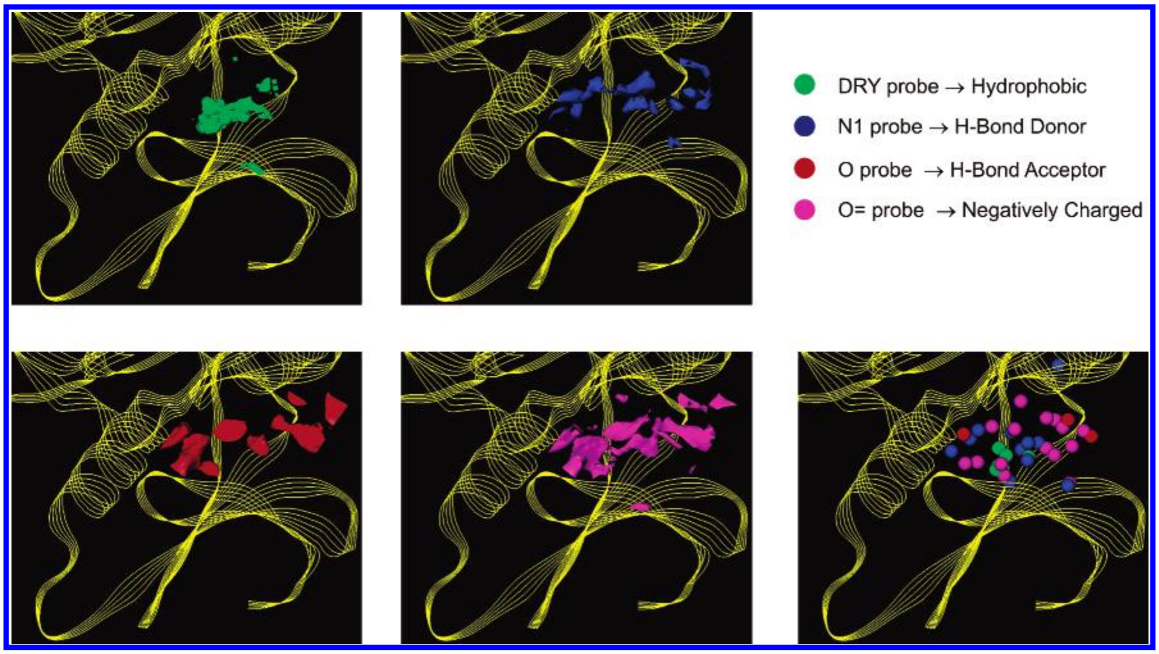


Figure 11. GRID-MIFs for 1H1S kinase and the target-site points (lower right) used for pharmacophore fingerprint calculations.

docking studies and to either differentiate the sites in order to gain selectivity or identify common binding features. Using four-point fingerprints to compare these protein-derived pharmacophore fingerprints with known ligands shows that they can be used to search for novel ligands within a database and that they are sufficiently specific to capture ligand selectivity between similar proteins. The comparisons possible in FLAP (see the SBVS section) enable binding site characteristics such as shape to be retained when comparing proteins and proteins to ligands, greatly enhancing the signal and providing the ability required to deal with protein selectivity.

FLAP is able to compare multiple protein targets. Although such a comparison could be performed over entire proteins, or over protein domains, reference is only made here to the comparison of protein active sites. This comparison is made as before by using a combination of four points of minima, which have been previously calculated by sampling each protein site using GRID probes.

The larger the number of combinations of GRID minima in common between two sample proteins, the greater the similarity between those two proteins is. The results are obtained from the calculations of all the tetrahedral configurations obtained by each of the probes interacting with the protein cavity (see Figure 12). FLAP performs target-pair comparisons as many times as there are protein sites under investigation (including a comparison between the same protein site, i.e., a self-comparison).

This operation produces a file for each binding site that contains the number of tetrahedral configurations in common between one protein site and another for all protein sites under investigation. The comparison of one protein site with another protein site may give rise to a different result if the comparison is inverted (the comparison of ProtA to ProtB may be different from the comparison of ProtB to ProtA). This is due to the asymmetry of the steric effect.

Finally, the total number of shape coincidences $T(ij)$ (the combinations of four points between two protein sites) is given for each protein (i). Having N protein sites, it can therefore be assumed that $1 \leq j \leq N$. This result is transformed in the final table of similarity. The similarity is given by the following expression (for the i-protein site and the j-protein site):

$$S(ij) = S(ji) = [T(ij) \times T(ji)]/[T(ii) \times T(jj)]$$

The 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 similarity matrix can be analyzed using multivariate statistical analysis in an attempt to rationalize the data and to find possible patterns or trends across different families or subfamilies of protein binding sites (see Figure 13).

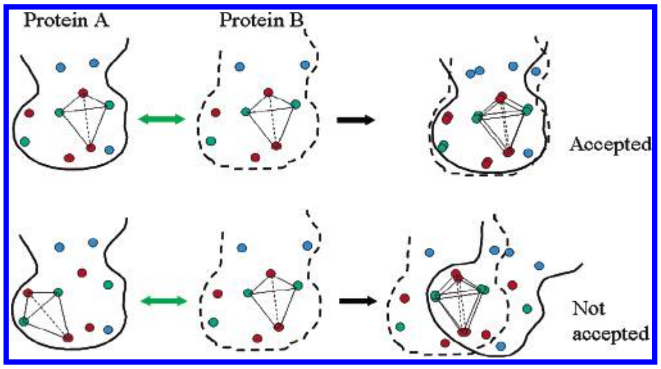


Figure 12. Two similar protein active sites are compared. All four-point pharmacophores are computed for proteins A and B. When a four-point pharmacophore in A is equivalent to a four-point pharmacophore in B, and the protein cavity shapes are similar (with some tolerances defined by the user), the pharmacophore represents a common solution for similarity ranking. Conversely, when the protein cavity shapes are different, the pharmacophore does not represent an acceptable solution.

Docking of Ligands into X-ray Structures. The number of algorithms available to assess and rationalize molecular docking studies is large and ever increasing. Many algorithms share common methodologies with novel extensions, and the diversity in both their complexity and computational speed provides a plethora of techniques to deal with modern structure-based drug design problems.[37]

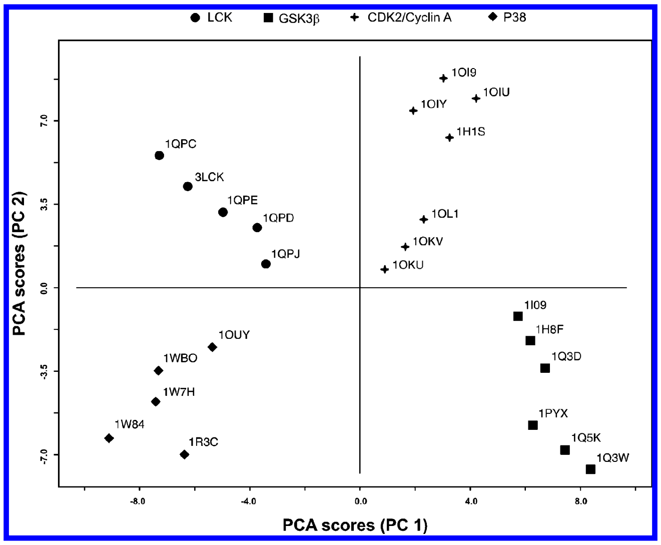


Figure 13. Principal component analysis performed on the similarity matrix as obtained from FLAP. The first principal component separates the LCK and P38 families from the CDK2 and GSK3*β* families. The second component separates LCK from P38 and CDK2 from GSK3*β*. It is important to point out that this result was obtained without any protein superposition, alignment, or knowledge-based comparison.

The FLAP program fits ligand molecules into a set of MIF-GRID maps of a target protein structure. So, although FLAP is not directly docking software, its pose-algorithm program could be used as part of a docking procedure.

As explained above, the four-point pharmacophoric features for a (macro)molecule are automatically identified. Once this has been done, all the accessible geometries for all the combinations of four features are calculated and stored in a fingerprint of the binding site. Afterward, an iterative procedure identifies all the ways in which four atoms of the ligand could bind to the target by pairing every atom to the nearest MIF used. Hydrophobic and polar atoms of the ligand for which several conformers are quickly produced are fitted over their corresponding attractive energy location, sometimes giving rise to millions of ligand arrangements, which are temporarily stored in memory.

Then, a large number of arrangements are quickly eliminated because of redundancy and steric hindrance constraints. Redundancy occurs whenever two or more arrangements are close enough to each other, that is, the root mean standard deviation (RMSD) calculated over their 3D structures is lower than 2.0 Å: they are therefore grouped by a clusterization process, and only one arrangement will be the candidate representing the entire group. Conversely, steric hindrance occurs whenever part of the ligand clashes into the binding site: if possible, the clashing part is accommodated along the site, otherwise the arrangement is excluded. Indeed, this refinement means only reliable arrangements are processed in the following step.

FLAP was tested as the pose algorithm for the GLUE docking procedure,[38,39] and the results were evaluated by comparing them with the results obtained using three of the most recognized docking programs, DOCK, GLIDE, and GOLD (refer to the Computational Methods for more details about the data set and settings used for this case study).

Defining as "Best Pose" a well-docked solution with an RMSD to the X-ray structure less than 2.00 Å, GLUE was able to dock 90 out of 100 complexes within this RMSD threshold. By comparison DOCK, GLIDE, and GOLD obtained 46, 78, and 81 results, respectively (Figure 14a). However, when the ranking is looked at, the best pose obtained by GLUE was among the first three solutions for 63 out of 100 complexes studied, while DOCK, GLIDE, and GOLD obtained 20, 38, and 31, respectively, in this test case (Figure 14b).

TECHNICAL DETAILS OF THE FLAP METHOD

From a computational point of view, molecular atoms and protein pharmacophoric points are separate independent spatial entities, and investigation into all possible coincidences between four atoms of a molecule and the same number of pharmacophoric points is a particularly onerous procedure.

Specifically, when considering the first (A1) of the four molecular atoms, FLAP would place it on each of the corresponding pharmacophoric points of the target protein, by translating the entire molecular structure. While keeping A1 fixed at protein position T1, FLAP would then search for a second corresponding pharmacophoric point (T2) on the target, at which one of the other atoms (A2) of the ligand could be placed, slightly rotating and translating the structure. Analogously, it could then proceed to a third atom and therefore to a fourth, but this is entirely inefficient, at least in terms of the calculation times required.

In fact, for the third atom, there must be two atom−atom distances agreeing with the protein point−point distances (dist. 1−3 and dist. 2−3), and in successful cases, a search is required for the fourth point which is as distant from the other three already identified as the fourth atom is distant from the first three (dist. 1−4, 2−4, 3−4). Furthermore, unlike the search for the three previous ones, an additional complication arises in assigning the fourth point to the fourth atom. In fact, it is not enough for them to be at the right distance from the three points already identified, but they also need to be found together with the fourth atom in the same part of the plane defined by the other three points/atoms.

This is a consequence of the problem of chirality which arises for tetrahedrons lacking an element of symmetry. So, if the edges are also coincident, two tetrahedrons could be the mirror images of each other.

For an average sized system, the number of distances and rototranslations to be calculated would be enormous, even in a restricted case of at most 20 pharmacophoric points for each of the six probes (or fewer). Obviously, there is a notable increase in the efficiency of the coincidence search algorithm when the total of possible distances between the points that map the receptors are calculated once and for all.

The point coordinates could be input into a bidimensional matrix from which the distance between all coupled pharmacophoric points can be easily extracted and used to check superimposability on the atoms. By combining all the atoms into groups of four and calculating the six interatomic distances, any combinations of points of the same type which present six coincident values can therefore be identified in the matrix, at least within the limits of the approximation required. However, in successful cases, the coincidence of the two chiralities needs to be checked.

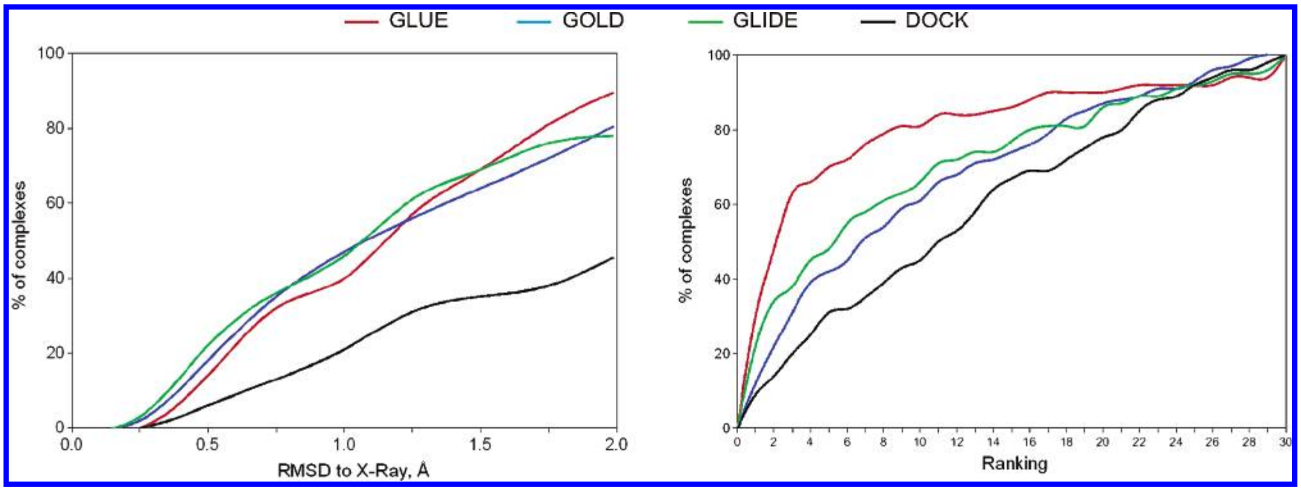


Figure 14. FLAP RMSD and ranking compared with the most used docking programs.

However, given the very large number of comparisons required, this search method is still not fast enough. For example, if a molecule is composed of 20 heavy atoms and a receptor mapped using 100 pharmacophoric points, the number of possible comparisons between tetrahedrons constructed on the molecule and those present in the receptor is greater that 1013!

In reality, this number is significantly reduced when only the tetrahedrons composed of atoms and related target pharmacophoric points are compared, and the large number of comparisons stop on reaching the fourth, third, or even the second atom. However, the problem still remains computationally onerous, and FLAP therefore calculates and memorizes all possible combinations of four pharmacophoric points in a simple but efficient format so that searches for tetrahedrons possessing determined size characteristics work much more quickly.

If it were not for the asymmetries that can often make non-superimposable solids equivalent, one being the mirror image of the other, this type of three-dimensional geometry could be unambiguously identified from the length of the six edges. Nevertheless, the problem can be resolved by assigning a flag to the chiral configurations so that they are clearly distinct, and the sign (positive/negative) of the tetrahedron volume can be used for this purpose.

At this point, it should be clear that the problem of the permutations is an inconvenience for any search algorithm based on comparing vector signs. Obviously, representing a spatial configuration of four points in 12 different ways could be a little inefficient.

In effect, a solution could be found in an algorithm capable of imposing an unambiguous order on the four vertices so that a unique vector, and only one chirality, is produced. This method always selects only one permutation from the 12 permutations, and by applying this method both to the tetrahedron constructed on the molecule and to that defined in a receptor, the two vectors and the sign of the chirality can be compared immediately.

Unfortunately, it has not been possible to find an effective algorithm that imposes an unambiguous order on the four vertices on the basis of relative distances only. Moreover, even it were found, it would create the problem of how to manage any symmetries that, even if there are some in tetrahedrons, are not for example found in molecules.

For example, suppose four equidistant atoms are connected, at least within the limits of the approximation of their distances adopted, and are equivalent from a chemical point of view (for example, four hydrophobic atoms).

If four pharmacophoric points are present in the site and together these form a regular tetrahedron with edges of the same sizes, after applying the ordering algorithm, one unique superimposition between the two tetrahedrons would be obtained. Because all the vertices are perfectly equivalent in this particular case, it would produce any one of the 12 methods of listing the vertices.

The point is that, if it is true that the pharmacophoric equivalence is not considered, the same cannot be said for what physically occurs when the molecule is rototranslated to superimpose the four atoms on the four points. In fact, by changing the way in which atoms and points are coupled, 12 different methods of arranging a hypothetical ligand in the receptor are produced, and these positions are anything but equivalent.

The same problem also occurs in the case in which the symmetries are less pronounced, for example, when they have edges of equal lengths. Here too, in both the molecule and receptor, the tetrahedral equivalence of several permutations never corresponds to an equivalence of the inside of the entire system. For this reason, and others of a computational nature, the receptor model is essentially composed of the group of all possible combinations of four potentially pharmacophoric points in all possible permutations.

On the other hand, the tetrahedrons constructed on the molecule are not permuted and are generated by adopting the order in which the atoms in the molecule are listed. These atomic configurations can therefore correspond to various combinations of pharmacophoric points in the site, and this diversity is due both to connected points and sometimes simply to the sequence in which they are combined.

It has already been stated that a certain degree of approximation Δ of all the lengths, which is typically a value of about 1 Å, will be applied to the entire system (molecule + target) for which two edges of a tetrahedron that differ by less than this value, optional but which once chosen, remains a constant within the model, are considered equal. Rounding off the lengths like this brings about a variation in the measurement of the Euclidean space for which, rather than using coordinates expressed in real numbers, simple integers can be used to indicate the values, which are simply multiples of the quantity Δ adopted.

However, in FLAP, the "objects" placed in the Euclidean space keep their real coordinates, and the approximation of the lengths is only applied to the edges of the tetrahedrons for both molecules and receptors. Obviously, if the value of Δ is very small, it will be very difficult to find spatial configurations of four coincident atoms with as many pharmacophoric points. A very high value of Δ could lead to a lack of ability to discriminate between molecules that might even present very different affinities compared to the receptors.

In FLAP, the receptor model is physically composed of the group of all possible combinations of four potential pharmacophoric points identified in the active site.

From a mathematical point of view, each quadruplet is identified by the typology of the four points, the six approximated distances between the four points, and the relative spatial arrangement that determines the type of chirality. Moreover, all the characteristic information is expressed by $6 + 4 + 1 = 11$, that is, distances, the typology of points, and the sign of chirality.

By using these 11 values, the mathematical receptor model can be defined as a data array which not only memorizes and identifies each 4-tuple unequivocally but also carries out extremely quick searches using algorithms based on appropriate ordering techniques. The procedure now outlined is used to evaluate the degree of pharmacophoric similarity between the receptor and any molecule.

First, the atoms of the potential ligand are classified so that they can be associated with the probes used to map the site, and then they are combined to form all possible quadruplets, each of which is searched for by the receptor model. These might be more than one match because several equivalent 4-tuples may be located at different positions in the site.

Whenever the form filter is activated, when FLAP finds a solution (match between the six distances that define the form of the two tetrahedra respectively constructed by the points and the atoms + the same chirality), the molecule is rototranslated to superpose the atoms as far as possible on the corresponding pharmacophoric points. The match between the six distances within the limits of the required tolerance is not always accompanied by an equally satisfactory superposition of the four atoms on the four pharmacophoric points, and in this case, it may be necessary to reject the solution. If the superposition of the four points is good, the position assumed in space by each atom is analyzed, including atoms such as the hydrogens that do not normally form part of the quadruplet.

Whenever the atom is in a sterically impeded area, a variable (named P) is increased by two units. On the other hand, if the atom is situated in the void but outside the grid, the variable P increases by one; otherwise, it remains unchanged. Consequently, at the end of the process, a value of $P \geq 0$ is obtained. If the number of atoms effectively analyzed is indicated by N, then $R = P/N$ is the value to be compared with a FLAP parameter (L). The solution is accepted if $R < L$ but is rejected should the contrary occur. Obviously, the greater L is, the more "tolerant" is the filtering of solutions on the basis of shape, and nothing else will be filtered for $L \geq 2$.

COMPUTATIONAL METHODS

Structure-Based Virtual Screening (SBVS). 1. DOCK4.0.1. Active site spheres were generated using SPHGEN[28] using default parameters. Sphere clusters were examined visually, and the cluster(s) that best filled the binding site were retained. To compute interaction energies, a 3D grid of 0.3 Å resolution was centered in both TK and Factor Xa active sites. Energy scoring grids were obtained using an all-atom model and a distance-dependent dielectric function ($\epsilon = 4r$) with a 10 Å cutoff. Amber7 FF99 atomic charges were assigned to all protein atoms as implemented in the biopolymer module in Sybyl7.0.[40] Flexible ligand docking (peripheral search and torsion drive) followed by energy minimization was carried out for all data set molecules, and the top solution corresponding to the best DOCK energy score for each ligand was then stored in a single multi mol2 file.

2. GOLD3.0. Active site detection was determined using the flood−fill procedure by defining the sequential number of an atom within the active site (558 NE2 GLU125 for TK and 1421 OD1 ASP189 for FXA). For each ligand, a number of 10 independent genetic algorithm (GA) runs were used and a maximum number of 1000 GA operations were performed on a single population of 50 individuals. Operator weights for crossover, mutation, and migration were set to 100, 100, and 0, respectively.

The annealing parameters, van der Waals and hydrogen bonding, allow poor hydrogen bonds to occur at the beginning of a genetic algorithm run, in the expectation that they will evolve to better solutions. The maximum distance between hydrogen donors and fitting points was set to 5 Å, and nonbonded van der Waals energies were cut off at a value equal to 10 kij, where kij is the depth of the van der Waals energy well between atoms i and j. To further speed up the calculation, GOLD was instructed to stop docking a ligand if it reaches a state in which the best three solutions found so far are all within 1.5 Å RMSD of each other. Only the top-scored position was kept for each ligand.

3. GLIDE2.5. Starting from Protein Data Bank files for both TK and Factor Xa, the protein preparation step was performed in order to check chemical correctness, assign protein atom charges, add hydrogens, and neutralize side chains that are not close to the binding cavity and do not participate in salt bridges.

To energetically describe protein active sites, a 3D grid of 1 Å resolution was located at the center of the bound ligand, and its dimension was automatically deduced from the ligand size and was able to fit the entire active site. For each ligand, the GlideScore function was used to keep the top-scored solution.

4. FLAP. Starting from Protein Data Bank files for both TK and Factor Xa, the protein preparation step was performed in order to check chemical correctness, assign protein atom charges, add hydrogens, and neutralize the charges of side chains that are not close to the binding cavity using GRID software. A 3D 1 Å resolution grid was then placed around the center of a bound ligand, and its overall dimensions were automatically deduced from the ligand size and the overall size of the entire active site. Target site points were automatically generated using C1=, N2, O, and O=S GRID probes (for Factor Xa) and C1=, N1, N:=, O, and O1 (for TK) plus shape. The tolerance for the distance comparison was set to 1.0 Å, with the unselective level at 0.2 (this option allows solutions showing few atoms clashing against the positive surface). The protein cavity was expanded by 1 Å to reduce steric clashes, and active solutions containing equivalent points were excluded to avoid redundant quadruplets of pharmacophoric points. Conformers were generated using random conformations by rotating between three and eight rotatable bonds. For TK, the maximum number of accepted conformers was 50.

Ligand-Based Virtual Screening. 1. Catalyst. The algorithm termed HipHop was used within Catalyst to generate seven different hypotheses, starting from the same chemical series used for FLAP. HipHop pharmacophore models are derived by comparing a set of conformational models and a number of 3D configurations of chemical features shared among the training set molecules. Compounds of the training set may or may not fit all features of each resulting hypothesis, depending on the settings for the parameters "Maximum Omitted Features", "Misses", and "Complete Misses".

The program Catalyst allows the generation of pharmacophore models, also termed hypotheses. A Catalyst pharmacophore model consists of a 3D arrangement of a collection of features necessary for the biological activity of the ligands. Whereas some features [like hydrogen-bond acceptor (HBA) or hydrogen-bond donor (HBD)] are defined as vectors, others [for example, hydrophobic (H) features] are located at centroids of the corresponding (e.g., hydrophobic) ligand atoms. The features in Catalyst are associated with location constraints, displayed as colored spheres, which allow a certain spherical tolerance surrounding the ideal position of a particular feature in 3D space. Catalyst models may be used as queries to search 3D coordinate databases of organic molecules for structurally new, potentially bioactive ligands. To be retrieved as a hit, a molecule must possess appropriate functional groups that match the features of the pharmacophore model. An automatic pharmacophore generation process in Catalyst requires the input of several active ligands that share the same binding mode. Depending on the properties and information content of these training set molecules, qualitative or quantitative hypotheses may be generated. Alternatively, the features of the pharmacophore may be placed manually, guided by the X-ray structure of a receptor ligand complex. In the present study, the latter, so-called structure-based approach, was used.

The starting point for the hypothesis generation process was the PDB entry 1ERR detected at a resolution of 2.6 Å, which contains human estrogen receptor in complex with raloxifene. The basic pharmacophore consists of four chemical features (two hydrophobic, one positive ionizable, and one HBA/HBD). The molecular shape constraint has been added to the pharmacophoric query. These features were placed manually on the corresponding groups of the ligand in its bioactive conformation.

A database made of the same ligands used with the other methods has been built in the Catalyst format and used to test the generated hypothesis in 3D database screening.

2. ChemX 3D Pharmacophore Fingerprints. The 3D structures of the query and actives + decoys are subjected to a conformation analysis after which ChemX fingerprints are generated (three and four points, seven distance ranges from 2.5 to 28 Å, with six pharmacophoric features for each point: basic, acidic, aromatic, hydrophobc/lipophilic, H-bond donor, H-bond acceptor; multiple features can be assigned to an atom to deal with H-bond donor + acceptor etc.). Query and active/decoy are then compared by using the Tanimoto equation as modified by Mason et al.:[13]

$$\frac{N_{\text{common}}}{(N_{\text{mol\_only}}0.25W_{\text{dynamic}}) + (N_{\text{refmol\_only}}0.05) + N_{\text{common}}}$$

where

$$W_{\text{dynamic}} = 1 - \frac{N_{\text{common}}}{N_{\text{refmol}}}$$

This equation is used to calculate the dynamically weighted Tanimoto-style coefficient similarity values and the resulting values using dynamic and nondynamic ($W_{\text{dynamic}} = 1$) weighting. N_{common} = pharmacophores common to "mol" and "refmol"; $N_{\text{mol\_only}}$ = pharmacophores exhibited by the analyzed molecule "mol" not common with the "reference" molecule; $N_{\text{refmol\_only}}$ = pharmacophores exhibited by the "reference" molecule not common with "mol".

Either molecule in a pair of molecules could be the "refmol" and the greatest of similarity retained; when searching a database using pharmacophoric data from an active molecule(s), that data is normally "N_{refmol}".

3. Barnard 2D Fingerprint. BCI structural fingerprints are 4096-long bit strings based on the presence or absence of 2D structural features of a molecule, listed in a predefined fragment dictionary that contains six different families of fragments: augmented atoms, atom/bond sequences, atom pairs, ring composition fragments, ring fusion fragments, and ring ortho fragments. When the query and decoy structures are used as starting points, the 2D fingerprints are generated and the Tanimoto coefficient is used to compare similarities.

4. FLAP. Starting from the Protein Data Bank file of the estrogen receptor α in complex with 4-hydroxy-tamoxifen (PDB entry: 3ert), ligand coordinates have been extracted and processed with the program FLAP, producing two output files. The first called the "pseudoreceptor" file represents the FLAP format of the ligand crystal structure. Atoms within the 4-hydroxy-tamoxifen file were classified in DRY, ACPT, and DONN according to GRID force field parameters (HB capability plus atomic charge) and stored together with their corresponding coordinates in the pseudoreceptor file. These site points are then used to generate all the possible combinations of four-point pharmacophores showed by the probe. The second file generated during this preprocessing step is called "mini" and represents the molecular shape of the reference ligand. This shape was then used as a geometric constraint during the ligand−ligand comparison analysis.

The tolerance for the distances comparison was set to 1 Å, with an unselective level set to 0.2 (the option allowing solutions showing few atoms clashing against the positive surface) and protein cavity expansion to 1 Å for reducing steric clashes.

To avoid quadruplets of pharmacophoric points made of the same kind of interactions (i.e., DRY−DRY−DRY−DRY or N1−N1−N1−N1), active solutions containing equivalent points were filtered out. Tetrahedral solutions were forced to contain at least one ACPT atomic site. Three regions were used as geometrical constraints, on the basis of the visual inspection of the complex between the ligand and the protein 1ERR. For each ligand, a maximum number of 50 conformers were generated using random conformations by rotating five rotatable bonds.

Pairwise Protein Similarity. The kinase PDB structures were downloaded directly from the RCSB Protein Data Bank. For each PDB entry, a protein preparation step was performed in order to check chemical correctness, assign protein atom charges, add hydrogens, and neutralize side chains that were not close to the binding cavity using GRID software. For this study, crystallographic water molecules and metal ions were removed from the protein coordinates. Protein mapping and protein fingerprinting were carried out as previously described in the Structure-Based Virtual Screening section, with the exception that the tolerance for the distance comparison was set to 1 Å, the unselective level being set to 0.3 (the option allows atoms partially clashing against the positive surface), and the protein cavity was expanded by 2 Å to reduce steric clashes.

Docking. The crystal structures of 100 protein−ligand complexes[41] from the RCSB PDB were used to generate a separate set of coordinates for the whole protein and its ligand. For DOCK, GLIDE, and GOLD, both the structures preparation and the docking stage were carried out as described by Rognan et al. in previous works.[41,42] The input conformations of the ligands were obtained starting from Smiles notation and transforming them into 3D geometry using CORINA.[43] The same 3D coordinates for the proteins and the ligands were used within the software GLUE, and both the structure preparation and the docking studies were run according to the previously reported work.[38]

CONCLUSIONS

The FLAP program represents a promising approach to obtain information from the molecular interaction field calculated by the GRID software for a selected region of a protein structure and from GRID probes representing a ligand molecule. The key feature of this software is the production of a common frame of reference for analyzing and comparing proteins and ligands together with an exhaustive compilation of all four-point pharmacophores in the protein region under investigation, followed by the codification of all this information into a compact fingerprint. Moreover, the simultaneous use of the (macro)molecular shape together with information about protein and ligand flexibility makes the FLAP procedure a very powerful tool for comparing protein and ligand pharmacophore fingerprints, pairwise ligand fingerprints, and pairwise protein pharmacophore fingerprints. This approach can be exploited very straightforwardly in structure-based drug design and in predocking calculations, ligand-based virtual screening, and protein similarity studies. The examples shown in this paper offer a validation of the method and illustrate the potential effectiveness of the approach.

A key feature is FLAP's ability to take into consideration the flexibility and shape of the ligand and the active site of the protein. The user can set constraints and keywords to describe particular features of the protein active site or of the ligand molecules. Moreover the calculation of the pharmacophore fingerprints is very fast, so that a reasonably large number of molecules can be handled within a few seconds.

ACKNOWLEDGMENT

Molecular Discovery thanks Pfizer Sandwich U. K. for their support in the FLAP project and the laboratory for Chemometrics and Cheminformatics at the University of Perugia for the help in testing the FLAP code. Particular thanks are due to Dr. Jonathan Mason for his inspiration and constructive discussions.

Supporting Information Available: A 3D SDF file of all the hits molecules used to enrich the databases studied in ligand-based and structure-based virtual screening. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) The study, performed on 200 protein−ligand complexes at high resolution, when recorded at −3.0 kcal/mol for hydrogen-bond donors/acceptors and −0.5 kcal/mol for hydrophobic interaction energies, showed that 85% of ligands interact with four or even more atomic groups with the target macromolecules.

(2) Gund, P. Three-Dimensional Pharmacophoric Pattern Searching. *Prog. Mol. Subcell. Biol.* **1977**, *5*, 117−143.

(3) Marshall, G. R. Binding-Site Modeling of Unknown Receptors. In *3D QSAR in Drug Design: Theory, Methods and Applications*; Kubinyi, H., Ed.; Escom: Leiden, The Netherlands, 1993; pp 80−116.

(4) Van Drie, J. H. *Pharmacophore Discovery: A Critical Review*; Marcel Dekker: New York, 2004; pp 437−460.

(5) Ghose, A. K. W. J. J. Pharmacophore Modelling: Methods, Experimental Verification and Applications. *Perspect. Drug Discovery Des.* **1998**, *9*−*11*, 253−271.

(6) Milne, G. W. A.; Nicklaus, M. C.; Wang, S. Pharmacophores in Drug Design and Discovery. *SAR QSAR Environ. Res.* **1998**, *9*, 23−38.

(7) Van Drie, J. H.; Nugent, R. A. Addressing the Challenges Posed by Combinatorial Chemistry: 3D Databases, Pharmacophore Recognition, and Beyond. *SAR QSAR Environ. Res.* **1998**, *9*, 1−21.

(8) Martin, Y. C. Pharmacophore Mapping. In *Des. Bioact. Mol.*; Martin, Y. C., Willett, P., Ed.; American Chemical Society: Washington, DC, 1998; pp 121−148.

(9) Bures, M. G. Recent Techniques and Applications in Pharmacophore Mapping. In *Practical Application of Computer-Aided Drug Design*; Charifson, P. S., Ed.; Marcel Dekker: New York, 1997; pp 39−72.

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

(11) Good, A. C.; Mason, J. S. Three-Dimensional Structure Database Searches. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Ed.; VCH Publishers: New York, 1995; Vol. 7, pp 67−117.

(12) Warr, W. A.; Willett, P. In *Des. Bioact. Mol.*; American Chemical Society: Washington, DC, 1998; pp 73−95.

(13) 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.

(14) Mason, J. S.; Cheney, D. L. Ligand−Receptor 3-D Similarity Studies Using Multiple 4-Point Pharmacophores. *Proc. Pac. Symp. Biocomput.* **1999**, *4*, 456−467.

(15) Mason, J. S.; Cheney, D. L. Library Design and Virtual Screening Using Multiple Point Pharmacophore Fingerprints. *Proc. Pac. Symp. Biocomput.* **2000**, *5*, 576−587.

(16) Mason, J. S.; Beno, B. R. Library Design Using BCUT Chemistry-Space Descriptors and Multiple Four-Point Pharmacophore Fingerprints: Simultaneous Optimization and Structure-Based Diversity. *J. Mol. Graphics Modell.* **2000**, *18*, 438−451.

(17) 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.

(18) Good, A. C.; Mason, J. S.; Pickett, S. D. Pharmacophore Pattern Application in Virtual Screening, Library Design and QSAR. In

Methods and Principles in Medicinal Chemistry; Bohm, H. J., Schneider, G., Eds.; Wiley-VCH: New York, 2000; Vol. 10, pp 131−159.

(19) Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28* (7), 849−857.

(20) Goodford, P. J. Multivariate Characterization of Molecules for QSAR Analysis. *J. Chemom.* **1996**, *28*, 107−117.

(21) 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* (21), 5114−5125.

(22) GRID is currently licensed to more than 1000 research centers worldwide and is available at www.moldiscovery.com.

(23) 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.

(24) Molecular Discovery Limited, 215 Marsh Road, Pinner, Middlesex, London HA5 5NE, U. K. www.moldiscovery.com (accessed Sept 2006).

(25) Berman, H. M.; Henrick, K.; Nakamura, H. Announcing the Worldwide Protein Data Bank. *Nat. Struct. Biol.* **2003**, *10* (12), 980.

(26) Jones, G.; Willett, P.; Glen, C. R.; Leach, R. A.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267*, 727−748.

(27) Friesner, A. R.; Banks, L. J.; Murphy, B. R.; Halgren, A. T.; Klicic, J. J.; Mainz, T. D.; Repasky, P. M.; Knoll, H. E.; Shelley, M.; Perry, K. J.; Shaw, E. D.; Francis, P.; Shenkin, S. P. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.

(28) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule−Ligand Interactions. *J. Mol. Biol.* **1982**, *161* (2), 269−288.

(29) Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: Search Strategies for Automated Molecular Docking of Flexible Molecule Databases. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411−428.

(30) Bissantz, C.; Folkers, G.; Rognan, D. Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.

(31) *Catalyst*, version 4.9.1; MSI: San Diego, CA.

(32) Baxter, C.; Murray, C. W.; Waszkowycz, B.; Li, J.; Sykes, R. A.; Bone, R. G. A.; Perkins, T. D. J.; Wylie, W. New Approach to Molecular Docking and Its Application to Virtual Screening of Chemical Databases. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 254−262.

(33) *Chem-X*; Oxford Molecular: Medawar Centre, Oxford Science Park, Oxford OX4 4GA, England.

(34) Barnard, J. M.; Downs, G. M. Chemical Fragment Generation and Clustering Software. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 141−142.

(35) Vulpetti, A.; Crivori, P.; Cameron, A.; 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.

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

(37) Abagyan, R.; Totrov, M. High-Throughput Docking for Lead Generation. *Curr. Opin. Chem. Biol.* **2001**, *5*, 375−382.

(38) Sciabola, S.; Baroni, M.; Carosati, E.; Cruciani, G. In *Recent Improvements in the GRID Force Field. 1. The docking procedure GLUE*. 15th European Symposium on QSAR & Molecular Modelling, Istanbul, Turkey, September 5−10, 2004; Aki-Sener, E., Yalcin, I., Ed.; CADD&D Society: Istanbul, Turkey, 2004; pp 47−49.

(39) Glue is part of the GRID software from version 22 and is available at www.moldiscovery.com (accessed Sept 2006).

(40) *SYBYL*, v. 7.1; Tripos Inc.: St. Louis, MO.

(41) Paul, N.; Rognan, D. ConsDock: A New Program for the Consensus Analysis of Protein−Ligand Interactions. *Proteins* **2002**, *47*, 521−533.

(42) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative Evaluation of Eight Docking Tools for Docking and Virtual Screening Accuracy. *Proteins* **2004**, *57*, 225−242.

(43) *Corina, Generation of 3D Coordinates*; Molecular Networks GmbH: Erlangen, Germany. http://www.mol-net.com/software/corina/index.html (accessed Sept 2006).

CI600253E