DOI: 10.1002/cmdc.200800355

SHOP: A Method For Structure-Based Fragment and Scaffold Hopping

Fabien Fontaine,[b] Simon Cross,[c] Guillem Plasencia,[b] Manuel Pastor,[a] and Ismael Zamora*[a]

A new method for fragment and scaffold replacement is presented that generates new families of compounds with biological activity, using GRID molecular interaction fields (MIFs) and the crystal structure of the targets. In contrast to virtual screening strategies, this methodology aims only to replace a fragment of the original molecule, maintaining the other structural elements that are known or suspected to have a critical role in ligand binding. First, we report a validation of the method, recovering up to 95% of the original fragments searched among the top-five proposed solutions, using 164 fragment queries from 11 diverse targets. Second, six key customizable parameters are investigated, concluding that filtering the receptor MIF using the co-crystallized ligand atom type has the greatest impact on the ranking of the proposed solutions. Finally, 11 examples using more realistic scenarios have been performed; diverse chemotypes are returned, including some that are similar to compounds that are known to bind to similar targets.

Introduction

Medicinal chemistry is one of the key disciplines in the small-molecule drug discovery process, and has suffered several paradigm changes over the years.[1] Recently, most drug discovery projects start with a high-throughput screening (HTS) campaign followed by a lead-generation step and then a lead-optimization process before one or more promising compounds with the desired pharmacodynamic, pharmacokinetic, and toxicological profiles (and free from external intellectual property restrictions) are selected for the first application in humans.[2] Most of the chemical resources are devoted to the optimization phase, synthesizing a large number of compounds in sufficient quantity and purity to be tested in various in vitro assays or in the lead-generation phase to evaluate (through SAR analysis, re-synthesis, scaffold modification, etc.) the potential hits obtained from the HTS. During the lead-generation phase a compound family is selected to be optimized, and therefore the selection of an appropriate scaffold during this step is a crucial point to continue the discovery process.

To aid lead generation, an increasing number of virtual screening (VS) experiments are performed on in-house, commercially available, or virtual libraries of compounds.[3] Compounds passing primary filters (such as Lipinski's rule of five) can be virtually screened, decreasing the number of compounds to test experimentally and providing novel chemotypes, with less cost than a full HTS campaign. Although in principle VS procedures can decrease the number of compounds to test, HTS campaigns are still performed on in-house compound collections to ensure that no active compounds are overlooked. VS experiments performed on commercial datasets provide compounds that are generally available; hence a further derivatization is needed to obtain intellectual property and to improve the overall properties of the potential new candidate drugs. Finally, using VS on virtual libraries of compounds has the advantage that a greater number of compounds are available, however there may be problems with synthetic accessibility.

There is currently a wide variety of VS methods available, using 2D[4–6] or 3D[7–10] representations of the compounds.[11] Ligand-based approaches use one or more structures of known active compounds as templates;[12] structure-based approaches use the structure of the protein[13] along with various scoring algorithms to select and rank the compounds. In each case, the basic procedure is a three-step process: 1) *compound preparation*: preparation of the virtual compound library and their descriptors, which can be from different sources (in-house, commercial, or virtual[14]); 2) *query preparation*: prepared from one structure, an ensemble of structures, or a protein; and 3) *scoring*: using similarity analysis[15] or scoring functions,[16,17] which are then used to rank the compounds in the database. Each variety of VS procedure yields a set of molecules to test, purchase, or synthesize, thus providing various starting points to evaluate. This has a limited impact in the lead-optimization phase, in which the compound family has already been selected and characterized.

[a] *Prof. M. Pastor, Prof. I. Zamora*
Universidad Pompeu Fabra-IMIM
Dr. Aiguader 88, 08003 Barcelona, Barcelona (Spain)
Fax: (+34) 935843347
ismael.zamora@telefonica.net

[b] *F. Fontaine, G. Plasencia*
Lead Molecular Design, S.L.
Av. Cerdanyola 92-94, 08173 Sant Cugat del Vallés, Barcelona (Spain)

[c] *S. Cross*
Molecular Discovery, Ltd.
215 Marsh Road, 1st Floor, HA5 5NE Pinner, Middlesex (UK)

In addition to these procedures that "jump" from one or more template molecules to new structures, there are other methods that aim to substitute one fragment of the molecule, leaving the rest of the molecule as it is in the original template.[18–22] These methods follow the same three steps described above for virtual screening: 1) compound or fragment preparation, 2) query preparation, and 3) scoring and ranking the possible solutions. This methodology may be called scaffold hopping, fragment hopping, or in the case of using chemical reactants in the initial database, reagent selection. When it is called scaffold hopping it might be confused with the VS procedures.

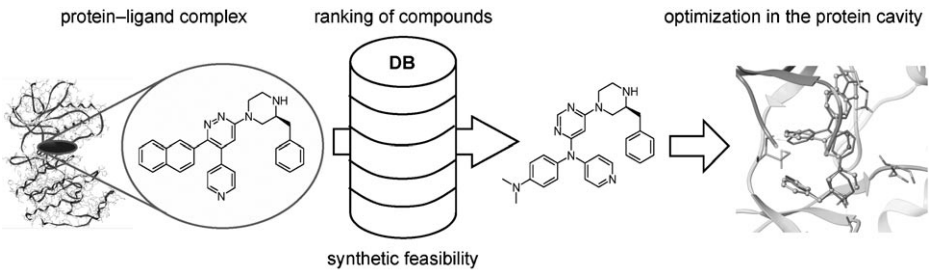


Figure 1. Search strategy: the query fragment to be substituted is indicated.

In the case of a fragment substitution using the bioisosterism paradigm, one of the key factors to consider is the synthetic feasibility of the suggested fragment replacement. Therefore, any such method needs an accurate definition of the query fragment and the solutions must include information about how they can be joined together. Although with this method reagents can come from in-house or chemical providers, it is intrinsically virtual, as the final compound, built from the suggested fragments and the other parts of the original molecule, usually has no guarantee of synthetic success.

As mentioned above, VS methods (either whole-molecule or fragment-based) can be ligand based (starting from the structure of a template ligand) or structure based (starting from a target binding cavity).[23] This work is focused on the latter alternative, introducing a fragment-hopping method that starts from the structure of the target binding site (protein-based SHOP). In addition, this method can integrate synthetic feasibility considerations and can also be linked to ADME prediction tools. Herein we start with a detailed description of the method and then present the results of some validation procedures, the first of which was performed on 11 diverse targets by analyzing the recovery of the known crystallographic fragments from an ad hoc generated database. Moreover, some practical cases are analyzed using an alternative database that includes our approach to address synthetic feasibility.

Materials and Methods

All of the computations were performed with a dual Xeon 3.0 GHz machine running a Linux operating system. All scripts were written in Perl programming language and can be accessed at http://www.leadmolecular.com.

The methodology to perform a SHOP scaffold-hopping search is straightforward (Figure 1). First, a database with the potential solution fragments is built to use in subsequent searches. Second, a query is prepared to search the database. In the case of a structure-based query, a ligand–receptor complex is required. Once the user has selected the database and the query fragment to be replaced in the ligand from the receptor–ligand complex, a 3D similarity analysis comparing the query and the fragments in the database is performed. In this way all of the fragments in the database are scored and sorted from the most to least similar. These sorted solution fragments are aligned with the original and placed in the protein cavity to check for steric clashes If there are no collisions with the protein atoms (with a certain customizable tolerance), the fragment is presented for user analysis. When the user selects fragments of interest, the entire ligand is rebuilt and optimized in the protein cavity. The result from this optimization process is a comparative analysis of the energy of interaction as computed by the GRID[24] program for each amino acid present in the cavity.[25]

SHOP database

The database contains all of the solution fragments with specified attachment atoms (anchor points) and must be built before starting the procedure. The fragments can be obtained from different sources; crystal structures of ligands, ad hoc designed fragments selected taking into account their synthetic feasibility, datasets of fragments already published, or even sets of heterocycles collected from described organic synthesis. In the first case the results from the search might be less interesting from the point of view of intellectual property, and therefore the crystal-structure-derived database is used for the purposes of validation. The program SHOP can import the fragments in standard 2D formats and carry out the 2D-to-3D conversion and conformational analysis.[13] The program can also determine the ionization state of the fragments during this step.[26] The structures are encoded into a vector of suitable descriptors containing the following four blocks:[18]

SHOP-geom

Two sets of descriptors are computed based on the distance between the anchor points in one case and the dihedral angle between the vector formed by the anchor point and the atom bonded to it in the fragment in the second.

SHOP-GRID

GRID Molecular Interaction Fields (MIFs) are computed around the fragment under analysis and are used to derive distance-based descriptors.[24] The GRID fields are computed using five

different probes that simulate the behavior of small chemical groups around the fragment under analysis. The probes used are: DRY, which describes the hydrophobic interaction; N1 (amide nitrogen probe with one hydrogen atom attached to it, the H-bond donor), which describes the hydrogen bond acceptor capabilities of the fragment; O (carbonyl oxygen probe with two lone pairs, H-bond acceptor), which describes the hydrogen bond donor capabilities of the fragment; N3+ (charged amine nitrogen probe, positive and H-bond donor), which describes the hydrogen bond acceptor capabilities and the electrostatic interaction with a negative charge of the fragment; and O− (charged oxygen probe, negative and H-bond acceptor), which describes the hydrogen bond donor capabilities and the electrostatic interaction with a positive charge of the fragment. The GRID computation is performed using all of the default parameters, but the number of planes (NPLA) per Ångstrom is set to 2 (grid spacing of 0.5 Å). The molecular interaction points which typically represent more than 10 000 grid points are then simplified by selecting the most relevant points. The anchor point is converted into a "hydrogen" atom prior to the MIF computation, and because the filter option has been set to "on" the MIF points coming from these "hydrogen" atoms are removed from the analysis. Finally, a vector for each anchor point and the selected points for each MIF is obtained by using the energy computed with the probe under analysis at the grid point and the distance between the anchor and the grid point position.[27]

SHOP-shape

The positive energy MIF points computed with the N1 probe are used to describe the shape of the fragment under analysis. A vector is defined considering the frequency of the distance between anchor point and the selected positive MIF points.[18]

SHOP-finger

For this descriptor set the atoms in the fragment are classified as hydrophobic, hydrogen bond donor, hydrogen bond acceptor, formally positively charged, and formally negatively charged following the same scheme reported in the MetaSite procedure.[28] Next, a vector for each anchor point and atom type classification is obtained by using the distance between the anchor atom and the different atom positions. Each of the distances is represented by using a Gaussian around the distance, with an alpha factor of 0.5.

In this way a fragment with one anchor point is represented in the SHOP database by 11 vectors: one for the shape analysis, five (one for each probe) for the GRID-derived description, and another five (one for each atom type analyzed) for the atom-type-derived description. A fragment with two anchor points is represented in the SHOP database by 24 vectors: one for the distance between the anchor points, one for the dihedral angle, two for the shape analysis (one per anchor point), ten for the GRID-derived description (five for each anchor point), and ten (five for each anchor point) for the atom-type-derived description. Finally, a fragment with three anchor points is represented in the database by 35 vectors: one for the distance between the anchor points, one for the dihedral angle, three for the shape analysis (one per anchor point), 15 for the GRID-derived description (five for each anchor point), and 15 (five for each anchor point) for the atom-type-derived description.

Query description

Although SHOP can be used to perform ligand-based searching, only structure-based searching is described herein. In this case the query is defined from a protein–ligand complex, in which the fragment of the ligand to be substituted can be selected by manually selecting the anchor points. Once the fragment has been selected in the ligand, it is removed from the ligand–receptor complex, and a GRID computation in the protein cavity is performed by using the same five probes described in the database building process (DRY, N1, O, N3+, and O−) (Figure 2). The most relevant grid points are selected using predefined field threshold values for each probe, and their positions are encoded in a vector describing the distance between every MIF point and the anchor atoms. At the end, five vectors are obtained (one for every probe). Such points represent the position where a ligand atom of the same kind of the probe used would produce a favorable interaction. Therefore these vectors match precisely the fourth block of fragment descriptors (SHOP-finger) described above: a frag-

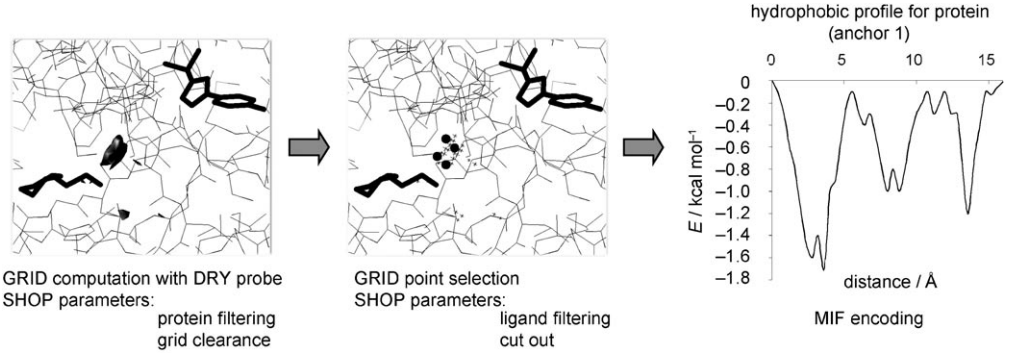


Figure 2. Encoding of the molecular interaction fields (MIFs).

ment with a perfect match is likely to place every atom in a favorable position. There are four parameters that can be modified during this process:

Protein filtering (PF)

This option controls which protein atoms are to be used in the GRID computation. Depending on the filtering level selected, no atom is deleted (none) from the protein prior to the MIF computation, or the atoms in the protein can be filtered to remove HETATM records (e.g. water molecules and counterions) (auto). Each of these two levels has been used in this analysis.

Grid clearance

This parameter controls the definition of the grid box used in the MIF computation. The box is defined as the maximum and minimum coordinates of the fragment to be replaced; the box is then enlarged in each dimension by the defined parameter. In this study this parameter has been set to 2 Å in all searches done.

Cut out

This parameter controls the grid points that are to be submitted to the MIF point selection algorithm. By definition the grid box is a regular parallelepiped, but typically the corners of this box are not interesting for the analysis, as the ligand does not interact with those regions of the receptor. This parameter removes those MIF points that are distant from the fragment under consideration. In this study this parameter has been set to 1.5 Å in all the searches done (i.e. any MIF point that is $>$1.5 Å from the fragment is eliminated from the analysis).

Ligand filtering (LF)

When applied, the MIF points are compared with the closest ligand atom. If the MIF point matches the ligand atom type in terms of hydrophobicity and H-bond donor and acceptor character, they are kept, otherwise they are discarded. In this analysis the searches were carried out both with and without ligand filtering. The effects of protein filtering (PF) and ligand filtering (LF) are analyzed exhaustively in this study.

Database search

Once the database is generated and the query has been defined, the search procedure can start. This process has three steps (Figure 3):

1. All of the fragments in the database with the same number of anchor points are sorted by a similarity analysis against the query description. A Pearson similarity index is used to compare the vectors describing the query with those describing the fragments in the database. The SHOP-geom descriptors of the query are compared with the same type of descriptors in the database. The MIF-derived descriptors for the query are compared with the SHOP-finger descriptors in the database. In this comparison the descriptors for the hydrophobic interaction field for the query are compared with the hydrophobic atoms of the database fragments; the O, N1, N3+, and O− probe-derived descriptors in the query are compared with the hydrogen bond acceptor, donor, positive, and negative atoms in the database fragments, respectively. Also, the atom-type fingerprint descriptors for the query are compared with the SHOP-GRID descriptors of the fragments in the database. As in the case of the MIF-derived descriptors, the complementarities on the interaction partners are considered. It is possible to control the influence of the different vector types in the ranking of the fragments. In this validation study, the influence of the charge descriptors and the atom-based fingerprint description of the query are evaluated.
2. The fragments have been scored and sorted; they are aligned to the original query in the protein cavity in order to evaluate their shape complementary. In this process

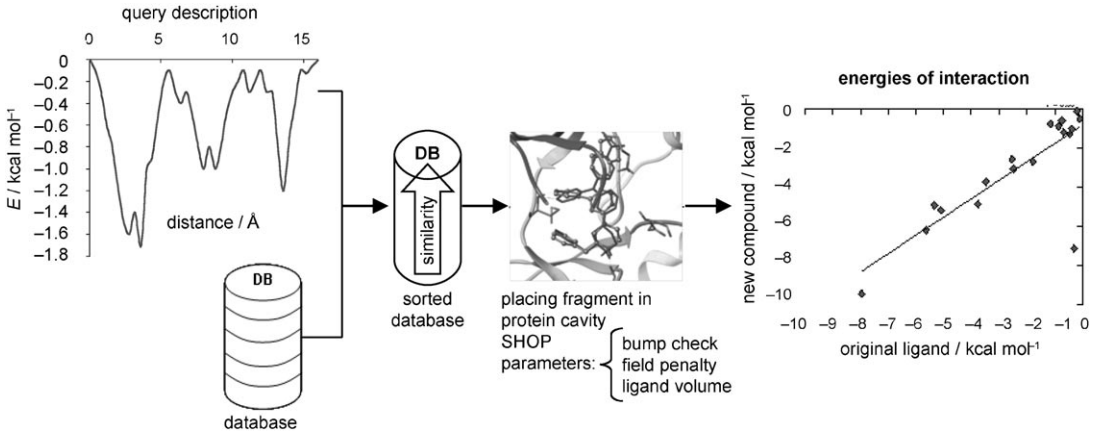


Figure 3. Search procedure: 1) ranking of solutions from the SHOP database, 2) placing the resulting scaffolds in the binding site, and 3) interaction energy computation.

there are three possible parameters that control the alignment procedure:

Tolerance: When the compound is placed in the cavity, a certain tolerance for the distance between the atoms of the fragment and the protein is allowed. The default value used in this study is 0.8 times the sum of the van der Waals radius of the atoms under consideration.

Field penalty: When placing the fragment in the cavity it is possible to analyze the complementarities of the atom types of the fragment with the GRID-computed field of the protein. If a field is unoccupied or a ligand atom type does not match a corresponding field, then the penalty is applied to the resulting atom-MIF complementarity score. In this analysis, this option was either not used, or set to 0.1. The greater the field penalty value, the more it influences the atom-MIF complementarity score and consequently the alignment.

Ligand volume: This parameter has two states: in the 'off' mode it is not used for the alignment of the compound in the protein cavity, while in the 'on' mode the fragment must not only fit in the protein cavity, but also within a volume envelope of the query fragment with a certain defined tolerance. The lower the value, the more restrictive the search. In this analysis we have used both the 'off' state and the 'on' with a value of 2.5 Å.

3. Optimization of the compound in the protein cavity. The best fragments that can fit in the protein cavity are submitted to a further structural refinement. In this procedure the fragments of the original ligand that were not defined as part of the query are joined to the selected solution fragment, building a complete molecule which is then minimized within the protein cavity using our own implementation of the MM3 force field. In the minimization process any amino acid that has any atom closer than 8 Å to any atom of the original ligand is considered. Moreover, the amino acids are kept fixed during the optimization process. The minimization process is used to compute the energy of interaction of the new suggested compound with the amino acids on the cavity, and it is not used to re-rank or re-prioritize the selected fragments.

Validation, parameter analysis, and fragment hopping

To optimize each of the searches, an analysis of the various factors affecting each of the studies was performed, with three important questions in mind. First, given a number of protein–ligand complexes, can the method return known ligand fragments from a pool of decoys? Second, which of the parameters influences the method most? Finally, and crucially, can the method find relevant fragments that are interesting from a medicinal chemistry perspective?

To answer these questions, 164 fragment queries from 11 targets were evaluated after automatic definition using an ad hoc developed script. The fragments were elected to represent part of the ligands that undergo key interactions with the protein as well as those parts that may not be so relevant for the ligand–protein interaction. These fragments were already present in the SHOP database generated from the PDBbind validation database described below. SHOP searches were performed using these queries, which consisted of one, two, or three anchor point searches, and considering six factors at two levels each, yielding 64 searches for each fragment query (Table 1). A total of 10496 searches were performed; the quality of each search was measured using the position of the fragment query in the 100 top-ranking solutions from the database (i.e. the query finding itself in the solutions). Notably, in this method the similarity of the fragments to the interaction in the protein cavity is used to rank the compounds in the database. Therefore, the validation of the method is not by an enrichment of the database but the capture of the fragment that is known (from crystal structure information) that can interact well with the cavity under analysis. Finally, a subset of these fragment queries were used to search different databases to provide an idea of the solutions returned in a more realistic scenario.

Table 1. Factors considered in the structure-based SHOP study.

Factor	Letter code	Low level	High level
Protein filtering	PF	None	SemiAuto
Ligand filtering	LF	Off	On
Protein atom descriptors	PAD	Off	On
Charge descriptor	CD	Off	On
Field penalty	FP	Off	0.1
Ligand volume	V	Off	2.5

Datasets

Database fragments

Validation database: The SHOP database used in the evaluation analysis was generated from the PDBbind dataset.[29,30] The ligands from the entire set in mol2 format were treated to create the SHOP fragment database. Because the compounds were already in 3D conformation, no additional optimization was performed on these compounds. Several steps were followed to prepare fragments from the compounds:

- A script was developed to cut each of the acyclic single bonds in each molecule, yielding two different fragments from each bond.
- Duplicate topological fragments were removed.
- Fragments with more than 50 and fewer than four heavy atoms were removed from this analysis.
- This process was performed in three cycles. In the first cycle, fragments with one attachment point were obtained (7455 fragments with one anchor point). In the second cycle the fragments obtained in step 1 were subsequently cut to generate fragments with two anchor points (19286 fragments with two anchor points); these, in turn, were followed to generate the three-anchor-point fragments (32442 fragments with three anchor points).

- The fragments were submitted to a SHOP database computation as described above.

Combinatorial chemistry database: One way to introduce synthetic feasibility into a SHOP database is to incorporate fragments that have been used in combinatorial chemistry. This was done by extracting 2D structures from combinatorial chemistry reviews.[31] These structures were converted to 3D and a conformational analysis was performed. The resulting fragments were then submitted to a SHOP database computation.

Heterocycle database: This database was built to introduce different heterocycles that can mimic the interaction of several chemical features in the protein binding site. In this case, no synthetic feasibility is considered; only the bioisosteric properties of the fragments are considered. The heterocycles with different substitution patterns were imported in 2D, then converted to 3D with conformational analysis. The fragments were then submitted to a SHOP calculation. The database is available from http://www.leadmolecular.com

Commercial database: The database was built considering commercial building blocks (http://www.specs.com) and using methodology already published by the SHOP virtual reaction utility.[18]

Query fragments

To analyze the quality of the strategy proposed in this study with different parameters, 164 queries were carried out, starting from the structure of 11 different ligand–receptor complexes extracted from PDBbind. The proteins were chosen among the structures with better resolution, trying to cover a wide number of protein/enzyme classes (Table 2).[31–41] The proteins were selected from the PDBbind dataset to cover various protein functions with a diverse set of ligands that have a wide range in affinity for, or inhibition of the target. From this collection of 164 queries, 91 fragment queries were used with one anchor point, 55 used two anchor points, and 18 used three anchor points.

During the validation process it was ensured that the query fragments were present in the database. For this study we carried out queries for the 164 fragments from the 11 different targets with one, two, or three anchor points and considering six factors at two levels each, yielding 64 searches for each fragment query (Table 2). Therefore a total of 10496 searches were performed. The quality of every search was quantified by noting the position of the fragment query in the first 100 first results.

Results and Discussion

The bioisosteric fragment replacement method SHOP, already described for the ligand-based approach,[18] has been extended to operate starting from the structure of a ligand–receptor complex. Unlike the original SHOP method, the search is based on the complementarity between the fragment and the MIF computed in the binding site. The fragments are first ranked by using a similarity index, then they are aligned inside the protein cavity to check for shape complementarity before rebuilding the original ligand with the query replaced by the new fragment and optimizing the whole ligand in the target environment. In this way the method is able to generate new potentially active compounds and can incorporate synthetic feasibility. The methodology was validated with a set of 164 fragments obtained from 11 diverse targets and a database built from a collection of ligands with known ligand–receptor complexes (PDBbind). A study was performed in order to evaluate the impact of the various parameters in the outcome of the search using the same database. Finally, one fragment for each target was used as a query in a more real scenario using databases that consider the potential synthetic feasibility of the compounds (see Materials and Methods above for a detailed description).

Methodology validation

Prediction quality assessment

To validate the structure-based fragment hopping, the recovery of the query fragments that were included in the SHOP database was analyzed by using the initial search results prior to any cavity optimization. The results for the best-ranking position for each query fragment across all search conditions is shown in Figure 4. More than 40% of the single-anchor-point fragments are recovered in the top five rank positions, whereas for two and three anchor points the results are around 90%. The effect of the number of anchor points is easily explained by the fact that there are more descriptor vectors for the three-anchor search than for the two-anchor search, and the fewest of all for the one-anchor search. Nevertheless, the one-anchor-point

Table 2. Query proteins used in the SHOP validation analysis.

Target name	PDB code	K_i	Resolution	# Fragments		
				1	2	3
Thrombin[33]	1SL3	1.4 pM	1.81	11	20	0
AmpC β-lactamase[34]	1XGJ	1.0 μM	1.97	6	0	0
p38α[34]	2BAJ	4.0 nM[a]	2.25	11	3	0
cAMP-Dependent protein kinase[36]	1RE8	0.3 nM	2.1	13	5	0
β-Trypsin[36]	1O2Q	21 nM	1.5	5	9	5
HIV-1 protease[37]	2I0D	0.8 pM	1.95	6	7	8
Estrogen receptor[38]	2AYR	0.51 nM	1.9	11	0	0
Tyrosine protein kinase (SRC)[39]	1IS0	100 nM[a]	1.9	6	6	5
Factor Xa[40]	1MQ5	1 nM	2.1	13	0	0
Acetylcholinesterase[41]	1E66	0.13 nM	2.1	3	0	0
Purine nucleoside phosphorylase[42]	1B8O	23 pM	1.5	6	5	0

[a] K_d values; the 2D structures of the fragments to be substituted are given in the Supporting Information.

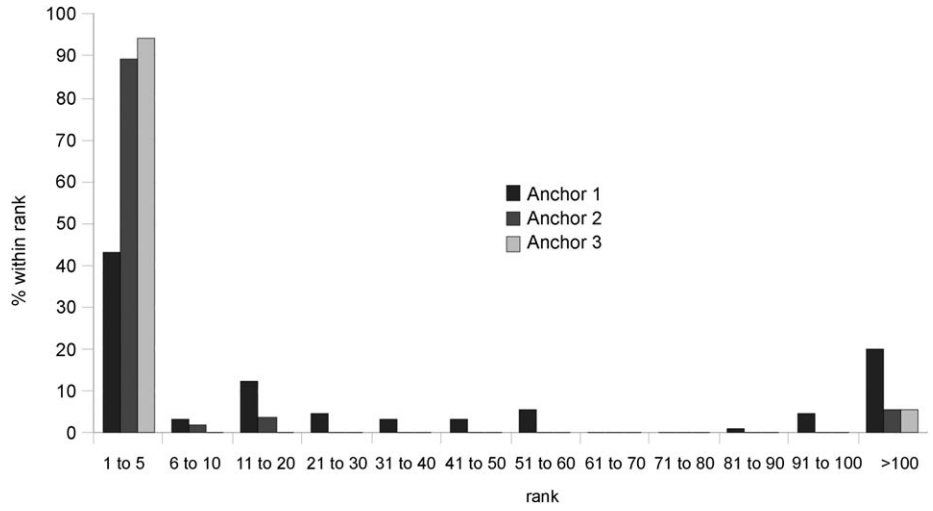


Figure 4. Histogram of best-ranking position for the query fragments found from all of the search criteria conditions, grouped by anchor number.

fragments are recovered up to 80% when the first 100 solutions are considered.

Analysis of the various factors

To analyze the effect of the various search parameters in the ranking position of the query fragments present in the SHOP database, a factorial design analysis was performed. In this investigation the main effects and the combination effects are studied (Figure 5). Ligand filtering (LF) is the factor that has the greatest contribution in the average effect analysis of all query fragments as well as for the analysis per query or even per target. Including the charge descriptors (CD) in the similarity analysis has a negative contribution (positive average effect) to the ranking of the query fragment. This is probably due to the fact that the information for these descriptor sets is already included in the hydrogen bond donor and acceptor ones. Therefore the inclusion of these descriptors increases the number of solutions with formal charges, which are not found in the fragment used as the query. Including the protein atom descriptors (PAD) has a positive contribution (negative average effect) to the ranking of the query fragment. Therefore, these vectors enrich the description of the fragment. It was also observed that the contribution for the volume (V) is greater with one anchor point than with two anchor points, and it has a negative contribution with three anchor points. In the case

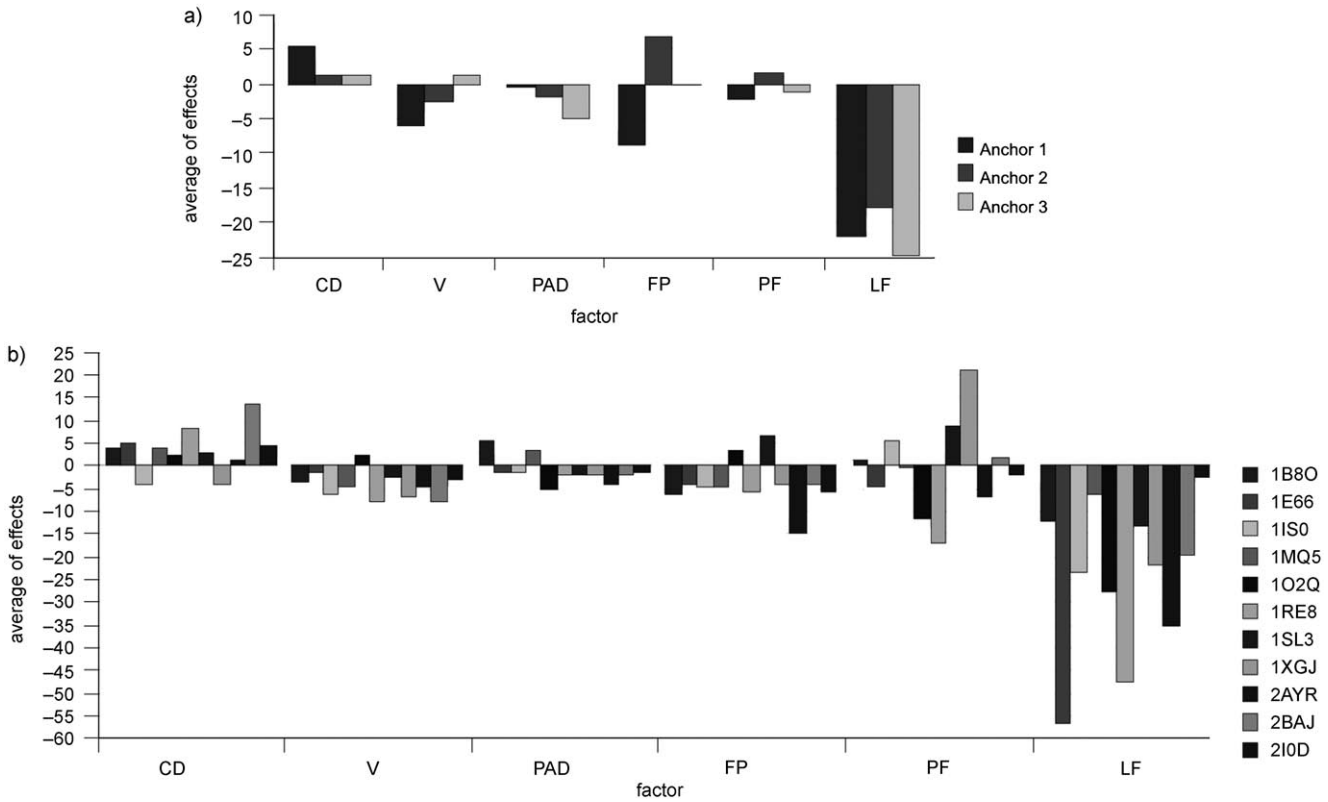


Figure 5. Factorial design analysis for the different effects: a) average of effects of main factors for all queries grouped by anchor number; b) average of effects of main factors for all queries per target analysis.

that there is only one anchor point the alignment is only defined by the anchor point itself, and so the inclusion of external volume restriction may be beneficial for the ranking. For the three-anchor-point cases the alignment is defined by the position of the three points, and consequently an external volume may not add anything to the alignment process. The field penalty (FP) and protein filtering (PF) have a different effect depending on the protein–ligand complex under consideration. In the case of the field penalty this may be related to a potential relationship between this factor and the volume and ligand-filtering factors. The protein filtering effect most likely depends on whether or not the water molecules included in the analysis make hydrogen bonds with the query fragment. Most of the combination effects have limited impact on the average effect analysis (data not shown).

To better demonstrate the effect of two of the factors that affect the query fragment description, the ligand and protein filtering are analyzed with an example:

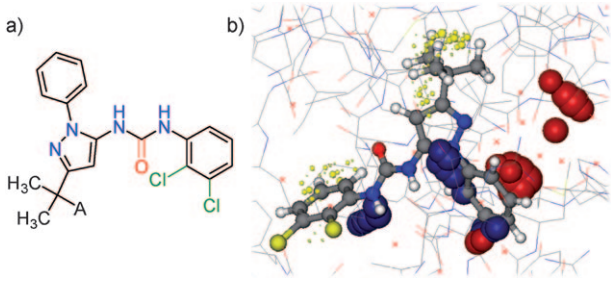


Figure 6. Ligand filtering effect on the descriptor and similarity computation: a) an example of a query fragment (A = anchor point); b) MIF computed by GRID using the DRY probe (yellow), N1 (blue), and O (red) inside the 2BAJ binding site; the ligand was extracted to perform the computation.

Ligand filtering (LF) effect

This parameter controls the grid points that are submitted to the descriptor calculation for the MIF description of the query. It has two states: in the "off" state, all of the grid points after the cut out process are submitted to the selection algorithm; in the "on" state only those grid points of the same nature as the ligand atom types are selected for the next step. This has a great impact in the case of some fragments under consideration. For example, with 2BAJ, there are hydrophilic grid points that are coincident with atoms that have hydrophobic character (phenyl ring) (Figure 6). In this case if the field is not filtered by the atom type, the query considers that an optimal fragment to substitute the phenyl ring is a polar-interacting group, and consequently the query fragment will be down-ranked. Therefore, it was observed that for the query fragment shown from target 2BAJ, the average ranking position for all search conditions with ligand filtering "on" is 12.3, while with ligand filtering "off" the average ranking position is >40.

Protein filtering (PF)

This filter removes atoms marked as HETATM in the PDB file, thus discarding water molecules and counterions from the computation. The effect of this parameter depends on the interaction of the fragment with these molecules and therefore has to be treated on a case-by-case basis without a general rule. As an example, a query fragment from target 1O2Q is shown in Figure 7a–c. In the case of using no filtering the water molecules are included, and the query fragment is not found among the first 100 solutions, because the water molecules undergo hydrophilic interactions with an aromatic ring. With filtering, and therefore with the water molecules absent, the query fragment appears in the ranking at position 3. An example for a positive contribution of the water molecules on the ranking of the query fragments is shown in Figure 7d–f for

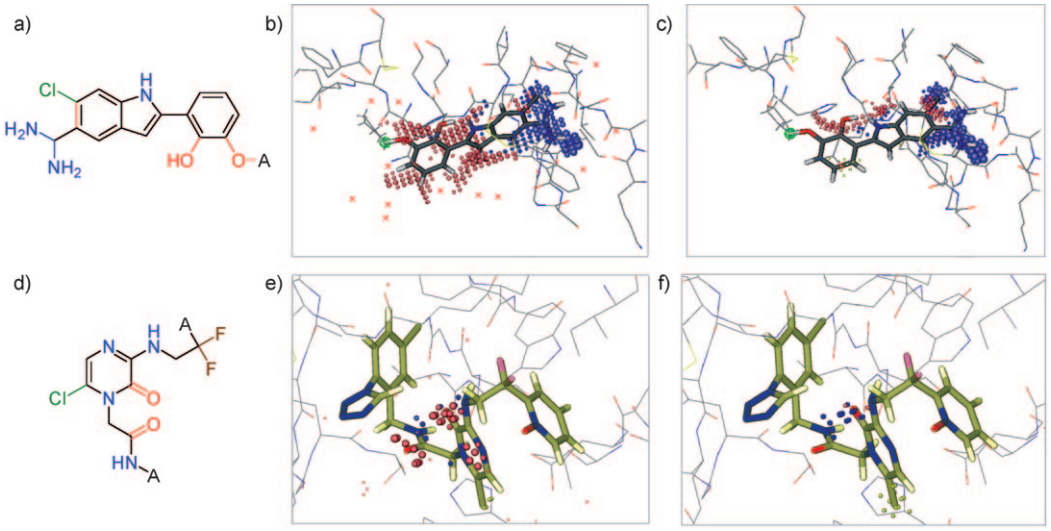


Figure 7. a) 1O2Q query fragment: interaction in the query fragment regions b) including the water molecule interaction and c) without the water molecule interaction. d) 1SL3 query fragment: interaction with the query fragment regions e) including the water molecule interaction and f) without the water molecule interaction. Red: hydrogen bond acceptor, blue: hydrogen bond donor, yellow: hydrophobic.

the target 1SL3. In this case, the carbonyl group in the query fragment makes a hydrogen bond with a water molecule that is part of a water network that interacts with the protein; inclusion of the water molecules in this case improves the ranking of the query fragment under analysis.

Real scenarios

In addition to the validation process based on the recovery of the query fragment from the database, the nature of the other solutions was analyzed. As an example the query fragment for p38α (2BAJ) was used with the following parameters set: the ligand filtering in the "on" state, the water molecules were included, the charge descriptors were not used, the protein atom descriptors were applied, the volume restriction was set to "on", and the field penalty was set to 0.1. Therefore, no hydrophobic grid points coincident with the phenyl group in the query were selected (Figure 8). The best solution in this search was a fragment from 1KV1, which is also a p38α inhibitor that has a methyl group instead of the phenyl in this part of the fragment; the rest of the molecule is identical to the query fragment. The query fragment is obtained in the second position of the search analysis, because there is no hydrophobic MIF region from the protein's perspective coincident with the phenyl ring. This simple example demonstrates that SHOP is able to return alternative solutions that are experimentally valid.

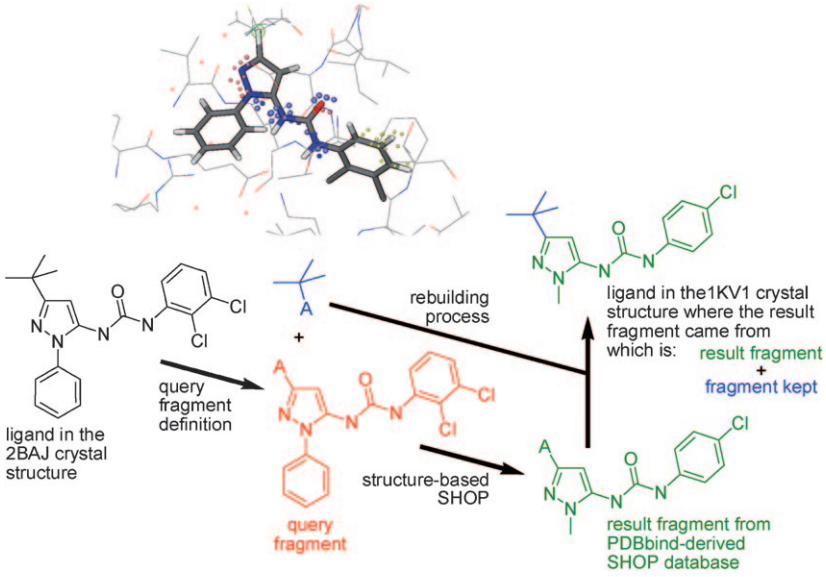


Figure 8. The ligand from the p38α structure 2BAJ is used to define a query fragment (red) and the rest of the molecule (blue). After a structure-based SHOP search, the top fragment solution from a PDBbind SHOP database search came from another p38α ligand 1KV1 (green), which, after attaching the fragment that was kept in the query definition, yields a final molecule that is identical to the ligand in 1KV1.

Another interesting example is the result from the search performed using a fragment from an HIV-1 protease inhibitor shown in Figure 9. In this case the search was performed using the same parameters as in the previous example. In the top 10 solutions, four were fragments from an HIV protease (1BV7, 1HXB, 1BWB, and 1ZP8), another two were proteases (2ER9 and 1MQ6), and the other four were from different enzyme classes: glycoxalase I (1QIN), protein tyrosine phosphatase (1ECV), acetyltransferase (2I79), and collegenase 3 (2DIN). The two ligands from the crystal structures 2I0D (query fragment) and 1HXB (one of the top 10 solutions) are shown in Figure 9. In Figure 10 the structure of the new compound formed by substitution of the query fragment by this solution optimized in the protein cavity is shown. This example demonstrates that the method was able to substitute a more complex fragment than that of Figure 8 by a different chemotype, yet retaining the same interaction pattern and mimicking the structures obtained from crystallography.

SHOP used with different databases

To demonstrate the potential use of the methodology in a prospective manner, one example of fragment hopping is shown in Table 3 for each protein target used in the validation study. The fragments found in the different databases range from very similar solutions to the query fragment (for example the ligands from 1SL3 and 1XGJ) to very diverse ones. In the first case (1SL3) the solution is almost identical, different by one carbonyl group, although the interaction pattern for both compounds is very similar, which suggests that this carbonyl does not contribute to the ligand–protein interaction. In the second case (1XGJ) the solution has one hydroxy group fewer so that it makes a weaker interaction with a tyrosine residue and with a water molecule in the protein cavity. However, the suggested compound has a stronger interaction with serine 64.

The fragments for the other nine cases are quite different in structure relative to the query, although they typically have similar interaction patterns. In some cases these solutions can be compared with ligands from other crystal structures of the same target. For example in the case of cAMP-dependent protein kinase (1RE8), one high-scoring solution is similar to the corresponding fragment from the ligand in the 1ATP crystal structure for the same target.[43] Another example is β-trypsin (1O2Q), where a high-scoring solution is similar to the corresponding fragment found in the ligand from the 1TX7[44] crystal structure. Finally, there are other solutions for which we could not find correlation with

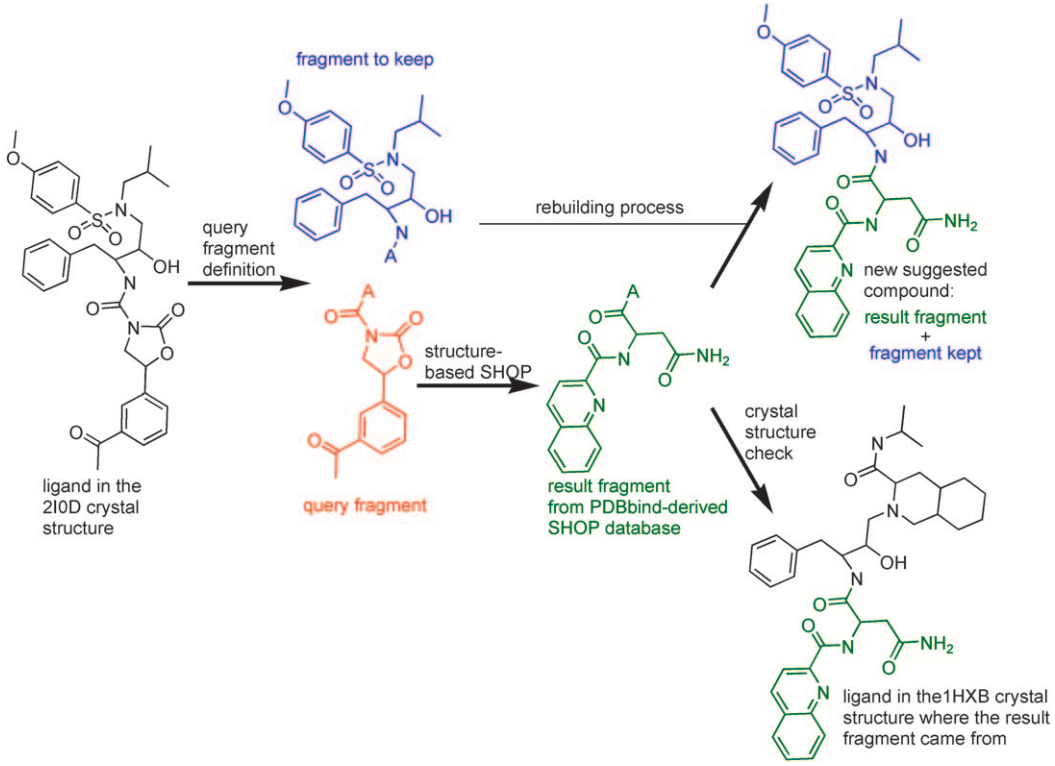


Figure 9. Fragment query used in the analysis and the result obtained from the SHOP analysis.

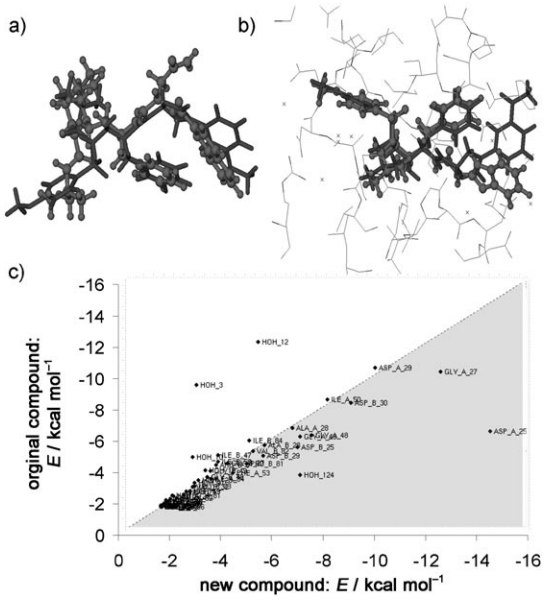


Figure 10. a) The 2I0D (stick) and 1HXB (ball-and-stick) ligand aligned by the backbone of the corresponding protein structure. b) New compound formed (ball-and-stick) optimized in the 2I0D protein cavity and the original ligand (stick). c) Energy profile for all of the amino acids in the protein cavity; the gray area represents those residues that interact more favorably with the new compound relative to the original.

known crystal structures, therefore the proposed structures were only corroborated by the optimization of the compound in the protein cavity.

Conclusions

Many computational methods used in drug discovery can generate new families of compounds with potential biological activity. All of these strategies are based on similarity or scoring methods, comparing a large database of in-house, commercial, or virtual compounds with some template molecule, pharmacophore, or protein structure. Nevertheless, many of the virtual screening approaches totally change the compound structure without keeping any part of the original lead compound. Therefore, these techniques are usually restricted to lead generation or even to a previous stage by selecting compounds from chemical providers to enhance an in-house compound library. We have presented a method that replaces one fragment of a molecule by another, keeping some structural elements of the original molecule, using the structure of the target of interest.

A validation study was performed for 11 relevant targets covering proteases, kinases, and an estrogen receptor; 164 fragments from these targets were used as queries in recovery experiments, with an ad hoc developed database from known crystal structure ligands derived from the PDBbind dataset with one, two, or three anchor points. More than 80% of the fragments with one anchor point were recovered among the first 100 solutions found by this procedure, and this recovery percentage reached 95% when two or three anchor points were used.

A detailed study of six factors that affect the search which are customizable by the user has been performed regarding the recovery of these query fragments. The most significant

Table 3. Search results for each target.

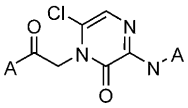
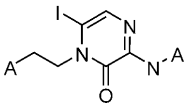
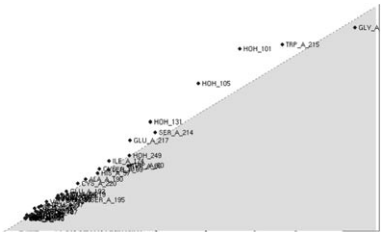
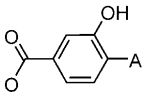
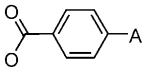
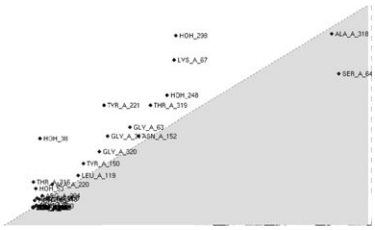
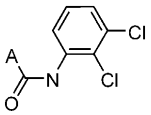
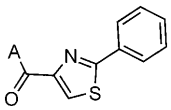
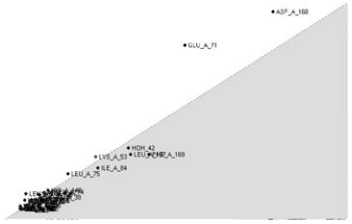
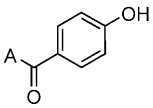
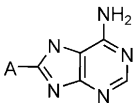
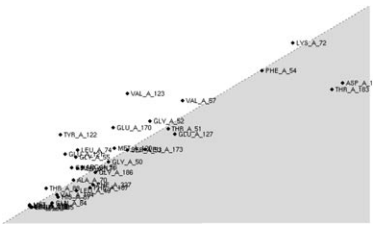
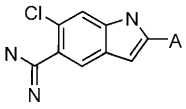
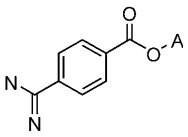
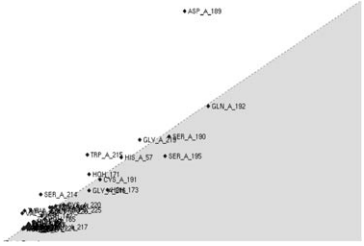
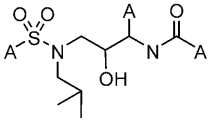
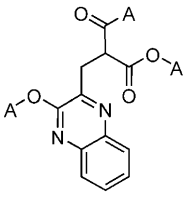
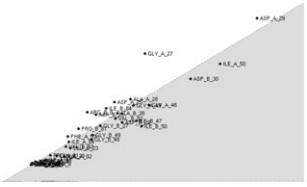
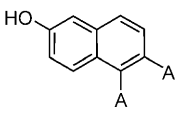
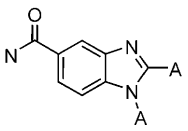
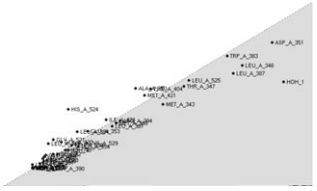
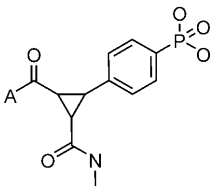
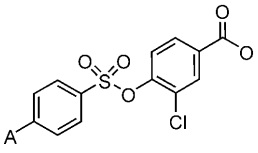
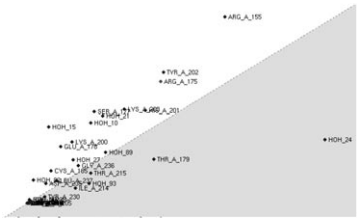
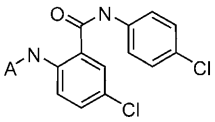
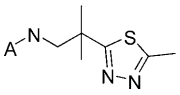
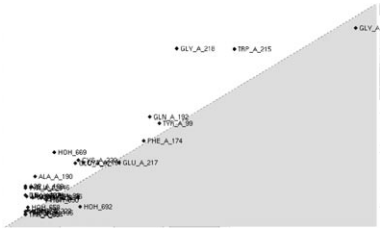
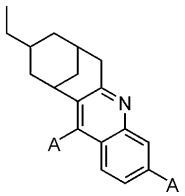
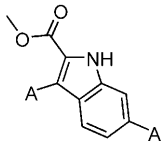
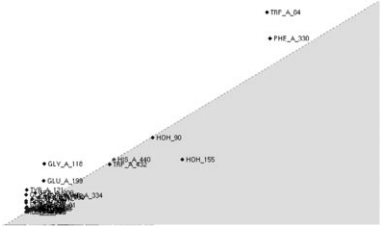
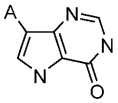
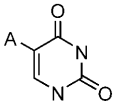
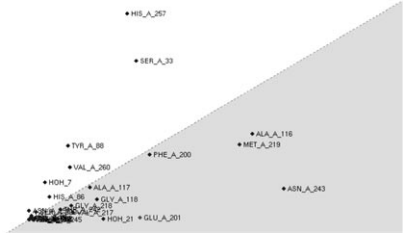
Target	Query	DB[a]	Result	Energy[b]
1SL3		A		
1XGJ		B		
2BAJ		A		
1RE8		C		
1O2Q		B		
2I0D		A		
2AYR		B		

Table 3. (Continued)

Target	Query	DB[a]	Result	Energy[b]
1IS0		A		
1MQ5		A		
1E66		B		
1B8O		C		

[a] DB = SHOP database: A, fragments from building block DB; B, fragments from the compendium of scaffolds used in combinatorial chemistry; C, fragments from heterocycles. [b] Energy: interaction energy between the compounds and the amino acids in the binding site; *x* axis: energy of interaction with the result fragment; *y* axis: energy of interaction for the original compound.

factor was the ligand filtering, in which the molecular interaction fields that describe the protein are filtered unless they are coincident with corresponding ligand atom types (polar or hydrophobic). Intuitively, filtering the solutions by the volume of the query fragment is also relevant. The field penalty and the protein filtering options have mixed effects depending on the ligand under consideration. In the case of the field penalty this is probably related to a potential relationship between this factor and both the volume and ligand filtering factors. The protein filtering effect most likely depends on whether there are any water molecules included in the analysis that make hydrogen bonds with the query fragment.

To test the technique in a more realistic scenario, one fragment for each of the targets was submitted to a structure-based SHOP analysis using three different databases: a bioisostere database developed from chemically accessible heterocycles, a database developed from known scaffolds from combinatorial chemistry reviews, and a database derived from commercial building blocks and common chemical reactions. The solutions from these searches ranged from those that were chemically very similar (and less interesting) to the query to those that were chemically very different and more interesting from a medicinal chemistry perspective. In the cases analyzed, the solutions mostly do not represent bioisosteric replacements of the original fragment, as the solutions do not present identical interactions to the queries. This is because the structure-based method is not based on similarity to the query fragment but to the interaction with the protein in the region defined by the query. Conversely, features in the query fragment may contribute to its SHOP description; if these features do *not* make interactions with the receptor, the fragment will not be scored as highly, which also explains why some of the queries are not found in the top solutions in the validation study.

In summary, we have demonstrated that it is possible to jump from one chemical fragment to other relevant chemotypes, keeping structural elements of the molecule fixed. The

procedure can be used for reactant selection for focused library design, lead generation, and also lead optimization of compounds during the drug discovery process.

Acknowledgements

This project received partial funding from the Spanish Ministerio de Educación y Ciencia (project SAF2005-08025-C03), the Instituto de Salud Carlos III (Red HERACLES RD06/0009), and the CIDEM (project RDITCRD07-2-0005-00).

Keywords: compound libraries · drug discovery · scoring functions · similarity analysis · virtual screening

[1] J. G. Lombardino, *Nat. Rev. Drug Discovery* **2004**, *3*, 853–862.
[2] H. Zhao, *Drug Discovery Today* **2007**, *12*, 149–155.
[3] J. J. Irwin, *J. Comput.-Aided Mol. Des.* **2008**, *22*, 193–199.
[4] L. Franke, O. Schwarz, L. Müller-Kuhrt, C. Hoernig, L. Fischer, S. George, Y. Tanrikulu, P. Schneider, O. Werz, D. Steinhilber, G. Schneider, *J. Med. Chem.* **2007**, *50*, 2640–2646.
[5] E. Gregori-Puigjané, J. Mestres, *J. Chem. Inf. Model.* **2006**, *46*, 1615–1622.
[6] N. Stiefl, I. A. Watson, K. Baumann, A. Zaliani, *J. Chem. Inf. Model.* **2006**, *46*, 208–220.
[7] K. Tsuchida, H. Chaki, T. Takakura, H. Kotsubo, T. Tanaka, Y. Aikawa, S. Shiozawa, S. Hirono, *J. Med. Chem.* **2006**, *49*, 80–91.
[8] J. L. Jenkins, M. Glick, J. W. Davies, *J. Med. Chem.* **2004**, *47*, 6144–6159.
[9] M. M. Ahlström, M. Ridderström, K. Luthman, I. Zamora, *J. Chem. Inf. Model.* **2005**, *45*, 1313–1323.
[10] E. Carosati, R. Mannhold, P. Wahl, J. B. Hansen, T. Fremming, I. Zamora, G. Cianchetta, M. Baroni, *J. Med. Chem.* **2007**, *50*, 2117–2126.
[11] J. Venhorst, S. Nuñez, J. W. Terpstra, C. G. Kruse, *J. Med. Chem.* **2008**, *51*, 3222–3229.
[12] Y. Koide, K. Uemoto, T. Hasegawa, T. Sada, A. Murakami, H. Takasugi, A. Sakurai, N. Mochizuki, A. Takahashi, A. Nishida, *J. Med. Chem.* **2007**, *50*, 442–454.
[13] M. Baroni, G. Cruciani, S. Sciabola, F. Perruccio, J. S. Mason, *J. Chem. Inf. Model.* **2007**, *47*, 279–294.
[14] M. Boehm, T.-Y. Wu, H. Claussen, C. Lemmen, *J. Med. Chem.* **2008**, *51*, 2468–2480.
[15] N. Wale, I. A. Watson, G. Karypis, *J. Chem. Inf. Model.* **2008**, *48*, 730–741.
[16] Q. Zhang, I. Muegge, *J. Med. Chem.* **2006**, *49*, 1536–1548.
[17] K. Tsunoyama, A. Amini, M. J. E. Sternberg, S. H. Muggleton, *J. Chem. Inf. Model.* **2008**, *48*, 949–957.
[18] R. Bergmann, A. Linusson, I. Zamora, *J. Med. Chem.* **2007**, *50*, 2708–2717.
[19] M. Bohl, B. Loeprecht, B. Wendt, T. Heritage, N. J. Richmond, P. Willett, *J. Chem. Inf. Model.* **2006**, *46*, 1882–1890.
[20] P. R. N. Wolohan, L. B. Akella, R. J. Dorfman, P. G. Nell, S. M. Mundt, R. D. Clark, *J. Chem. Inf. Model.* **2006**, *46*, 1188–1193.
[21] F. Dey, A. Caflisch, *J. Chem. Inf. Model.* **2008**, *48*, 679–690.
[22] P. Ertl, S. Jelfs, J. Mühlbacher, A. Schuffenhauer, P. Selzer, *J. Med. Chem.* **2006**, *49*, 4568–4573.
[23] B. Rikke, T. Liljefors, M. D. Sørensen, I. Zamora, unpublished results.
[24] P. J. Goodford, *J. Med. Chem.* **1985**, *28*, 849–857.
[25] B. Kjellander, C. M. Masimirembwa, I. Zamora, *J. Chem. Inf. Model.* **2007**, *47*, 1234–1247.
[26] F. Milletti, L. Storchi, G. Sforna, G. Cruciani, *J. Chem. Inf. Model.* **2007**, *47*, 2172—2181.
[27] F. Fontaine, M. Pastor, I. Zamora, F. Sanz, *J. Med. Chem.* **2005**, *48*, 2687–2694.
[28] I. Zamora, L. Afzelius, G. Cruciani, *J. Med. Chem.* **2003**, *46*, 2313–2324.
[29] R. Wang, X. Fang, Y. Lu, S. Wang, *J. Med. Chem.* **2004**, *47*, 2977–2980.
[30] R. Wang, X. Fang, Y. Lu, C.-Y. Yang, S. Wang, *J. Med. Chem.* **2005**, *48*, 4111–4119; http://sw16.im.med.umich.edu/databases/pdbbind/index.jsp (accessed December 19, 2008).
[31] a) R. E. Dolle, K. H. Nelson, *J. Comb. Chem.* **1999**, *1*, 235–282; b) R. E. Dolle *J. Comb. Chem.* **2000**, *2*, 383–433; c) R. E. Dolle, *J. Comb. Chem.* **2001**, *3*, 477–517; d) R. E. Dolle, *J. Comb. Chem.* **2002**, *4*, 369–418; e) R. E. Dolle, *J. Comb. Chem.* **2003**, *5*, 693–753; f) R. E. Dolle, *J. Comb. Chem.* **2004**, *6*, 623–679; g) R. E. Dolle, *J. Comb. Chem.* **2005**, *7*, 739–798; h) R. E. Dolle, B. Le Bourdonnec, G. A. Morales, K. J. Moriarty, J. M. Salvino, *J. Comb. Chem.* **2006**, *8*, 597–635.
[32] M. B. Young, J. C. Barrow, K. L. Glass, G. F. Lundell, C. L. Newton, J. M. Pellicore, K. E. Rittle, H. G. Selnick, K. J. Stauffer, J. P. Vacca, P. D. Williams, D. Bohn, F. C. Clayton, J. J. Cook, J. A. Krueger, L. C. Kuo, S. D. Lewis, B. J. Lucas, D. R. McMasters, C. Miller-Stein, B. L. Pietrak, *J. Med. Chem.* **2004**, *47*, 2995–3008.
[33] D. Tondi, F. Morandi, R. Bonnet, M. P. Costi, B. K. Shoichet, *J. Am. Chem. Soc.* **2005**, *127*, 4632–4639.
[34] J. E. Sullivan, G. A. Holdgate, D. Campbell, D. Timms, S. Gerhardt, J. Breed, A. L. Breeze, A. Bermingham, R. A. Pauptit, R. A. Norman, K. J. Embrey, J. Read, W. S. Vanscyoc, W. H. Ward, *Biochemistry* **2005**, *44*, 16475–16490.
[35] P. Akamine, Madhusudan, L. L. Brunton, H. D. Ou, J. M. Canaves, N. H. Xuong, S. S. Taylor, *Biochemistry* **2004**, *43*, 85–96.
[36] B. A. Katz, K. Elrod, E. Verner, R. L. Mackman, C. Luong, W. D. Shrader, M. Sendzik, J. R. Spencer, P. A. Sprengeler, A. Kolesnikov, V. W. Tai, H. C. Hui, J. G. Breitenbucher, D. Allen, J. W. Janc, *J. Mol. Biol.* **2003**, *329*, 93–120.
[37] A. Ali, G. S. Reddy, H. Cao, S. G. Anjum, M. N. Nalam, C. A. Schiffer, T. M. Rana, *J. Med. Chem.* **2006**, *49*, 7342–7356.
[38] C. W. Hummel, A. G. Geiser, H. U. Bryant, I. R. Cohen, R. D. Dally, K. C. Fong, S. A. Frank, R. Hinklin, S. A. Jones, G. Lewis, D. J. McCann, D. G. Rudman, T. A. Shepherd, H. Tian, O. B. Wallace, Y. Wang, J. A. Dodge, *J. Med. Chem.* **2005**, *48*, 6772–6775.
[39] J. P. Davidson, O. Lubman, T. Rose, G. Waksman, S. F. Martin, *J. Am. Chem. Soc.* **2002**, *124*, 205–215.
[40] M. Adler, M. J. Kochanny, Y. Bin, G. Rumennik, D. L. Light, S. Biancalana, M. Whitlow, *Biochemistry* **2002**, *41*, 15514–15523.
[41] H. Dvir, D. M. Wong, M. Harel, X. Barril, M. Orozco, F. J. Luque, D. Munoz-Torrero, P. Camps, T. L. Rosenberry, I. Silman, J. L. Sussman, *Biochemistry* **2002**, *41*, 2970–2981.
[42] A. Fedorov, W. Shi, G. Kicska, E. Fedorov, P. C. Tyler, R. H. Furneaux, J. C. Hanson, G. J. Gainsford, J. Z. Larese, V. L. Schramm, S. C. Almo, *Biochemistry* **2001**, *40*, 853–860.
[43] J. Heng, E. A. Trafny, D. R. Knighton, N. H. Xuong, S. S. Taylor, L. F. Ten Eyck, J. M. Sowadski, *Acta Crystallogr. Sect. D* **1993**, *49*, 362–365.
[44] J. Cui, F. Marankan, W. Fu, D. Crich, A. Mesecar, M. E. Johnson, *Bioorg. Med. Chem.* **2002**, *10*, 41–46.

Received: October 23, 2008
Revised: November 19, 2008
Published online on January 16, 2009

