Computational Fragment-Based Approach at PDB Scale by Protein Local Similarity

Fabrice Moriaud,*,† Olivia Doppelt-Azeroual,† Laetitia Martin,† Ksenia Oguievetskaia,† Kerstin Koch,‡ Artem Vorotyntsev,† Stewart A. Adcock,† and François Delfaud†

MEDIT SA, 2 rue du Belvédère, 91120 Palaiseau, France, and IBBMC, Université Paris Sud CNRS UMR-8619, Orsay 91405, France

Received August 31, 2008

The large volume of protein−ligand structures now available enables innovative and efficient protocols in computational FBDD (Fragment-Based Drug Design) to be proposed based on experimental data. In this work, we build a database of MED-Portions, where a MED-Portion is a new structural object encoding protein-fragment binding sites. MED-Portions are derived from mining all available protein−ligand structures with any library of small molecules. Combined with the MED-SuMo software to superpose similar protein interaction surfaces, pools of matching MED-Portions can be retrieved from any binding surface query. The rapidity of this technology allows its application to a diverse set of 107 protein binding sites. The selectivity of the protocol is shown by a qualitative correlation between the average hydrophobicity of the pools of MED-Portions and those of the binding sites. To generate hitlike molecules, MED-Portions are combined in 3D with the MED-Hybridise toolkit. Our MED-Portion/MED-SuMo/MED-Hybridise protocol is applied to two targets that represent important protein superfamilies in drug design: a protein kinase and a G-Protein Coupled Receptor (GPCR). We retrieved actives molecules of PubChem bioassays for the two targets. The results show the potential for finding relevant leads from any protein 3D structure since the occurrence of interfamily MED-Portions is 25% for protein kinase and almost 100% for the GPCR.

INTRODUCTION

Our knowledge based on known 3D structures of protein targets plays a major role in designing and optimizing compounds that bind to specific targets. The number of macromolecule structures publicly available from the Protein Data Bank (PDB) is raising exponentially,[1] with more than 600 new entries released every month. The total number of deposits was over 50,000 macromolecule structures on August 12th, 2008 of which 48,000 are protein structures, among them many popular pharmaceutical targets. Interestingly, the targets of some proteins in the PDB are cocrystallized with diverse ligands allowing to decipher induced fit and identify key residues for affinity and selectivity. Besides ligands, small organic molecules from buffer or solvent may be a rich source of information for pocket or subpocket detection. Allen et al. have introduced an experimental approach to map the surfaces of crystalline protein with solvent molecules.[2]

Fragment-based drug discovery is an established paradigm (for reviews see refs 3−5 and references therein) in contrast to conventional high throughput screening (HTS) where fully built, "drug-sized" chemical compounds are screened for activity. Small chemical structures or fragments (100−250 Da) that intrinsically have weaker binding affinity (100 μM to 10 nM range) are screened to probe the complete binding site and then to identify larger molecules based on one or multiple binding fragments. Exciting advances in recent years have come from the use of high-throughput X-ray analysis and NMR in structural screening of fragments.[6] NMR spectroscopy was the first structural technique exploited for use in fragment screening using a method termed 'SAR by NMR'.[7] Ciulli et al. have combined NMR and others biophysical approaches to probe hot spots at protein−ligand binding sites.[8] An early use of X-ray crystallography as a tool to identify fragment 'hits' was described by Verlinde et al.[9] There is an increasing number of case studies, and a more recent example among many others is the iterative structure-based design from an initial fragment hit to a nanomolar protein kinase inhibitor.[10] The impact on human diseases has been compiled recently: 47 potent inhibitors (IC_{50} < 100 nM) have been discovered with three of them having entered clinical trials.[11]

Obtaining experimental structural information on fragments or ligands complexed to a target protein is a key element and also a major limitation to the number and types of target that are amenable to fragment-based drug discovery. Consequently, computational methods play a key role in deriving structural information for designing compounds that fit a particular site on a given protein. If the three-dimensional structure of the protein is known, this information can be directly exploited for the retrieval and design of new ligands. These structure-based drug design methods fall into three categories, each distinguished by the level of detail and complexity of the problem.[12] The first category includes docking/scoring protocols. These tend to predict complementary of shape and interactions[13] and focus on existing molecules of potentially limited novelty. The second includes *de novo* design protocols. These involve incremental and interactive combination of fragments having best predicted interactions into the binding pocket to design new ligands.[14,15]

* Corresponding author phone and fax: +33-160148743; e-mail: fmoriaud@medit.fr.
† MEDIT SA.
‡ IBBMC, Université Paris Sud CNRS UMR-8619.

10.1021/ci8003094 CCC: $40.75
Published on Web 01/28/2009

The third category includes the map-based predictive methods. These can provide new perspectives, such as submitting a protein full surface. Goodford mapped a receptor active site in the drug design program GRID, followed by a fragment assembly process in which some of the favorable positions found for individual molecular fragments were connected into a single viable molecule.[16] This active site mapping and fragment assembly strategy is implemented in a number of drug design programs. Another interesting approach to mapping is the MCSS method,[17,18] which optimizes the free energy of numerous ligand copies simultaneously. Novel molecules can be constructed from fragments at the MCSS minima. Increasing computer performance now allows application of MCSS to bigger molecules than solventlike fragments and ligands.

In contrast to the previously described simulation protocols, some PDB-based methods are becoming available. These are based on the extraction of similar structural information observed in known structures of protein−ligand complexes to generate new compounds. Taking advantage of the structural experimental data of protein−ligand interactions is of high value for drug design. Fast methods based on structural interaction fingerprints[19] can be used to classify and mine protein−ligand complexes from the PDB and in a more challenging way from high-throughput docking results, but they are not intended for *de novo* design. Taking advantage of the entire PDB as a starting point of drug design is a relatively new idea because the PDB has only recently become rich in terms of diversity of proteins and of ligands. In an effort to increase the success rate of ligand generation, the nontraditional *de novo* design technique BREED[20] utilizes a completely different idea: a set of ligand-bound target structures from the same protein superfamily is superposed, and these ligands in their active three-dimensional conformations are exhaustively recombined. The recombination is performed by swapping the fragments of different ligands where their bonds are seen to overlap in space. This procedure is performed recursively, so that the candidate compounds that emerge from recombination are added to the pool of known ligands and participate in subsequent cycles of recombination. As the ligands are structurally superposed in their active conformation, the generated structures inherit this active conformation and are effectively predocked to the receptor. Relevant results were obtained with protein kinases and aspartic proteases which are easy to align and well represented in the PDB, but the method is severely limited for targets with only a few known ligand-bound structures like G-Protein Coupled Receptors (GPCRs). As suggested by the authors in reference to Schmitt et al.,[21] this sort of inhibitor design by ligand combination might be applied to ligands of unrelated targets with topologically similar binding pockets. Several approaches are now published on the detection of pocket similarities and their alignment[21−28] that can assist the prealignment of ligand-bound target structures. As a consequence, the process would become fully automated and could be applied to any protein−ligand structure of the PDB independently of the sequence identity. As a consequence, the structurally superimposable protein−ligand material is substantially richer. Moreover approaches that can include subpockets in their database would be potentially able to detect and exploit local similarities. To our knowledge only the work of Jambon et al. (SuMo)[24] and Ramensky et al.[29] can achieve this task.

The work of Ramensky et al. is the first to emphasize protein local similarities at PDB scale to improve drug design results:[29] the comparison of a query protein binding site against the 3D structure of another protein in complex with a ligand enables fragments from a ligand to be transferred to the query protein. They applied the concept to populate a binding site with a cloud of atoms which is used as either a scoring function to favor docked ligands that overlap with this cloud or a source of substituents for lead optimization because they extend the sets of atoms to fragments with an additional step of computation. One limitation is presumably that the transferred moieties are atom sets and not fragments which are more useful chemical moieties as they usually occupy a subpocket and two or three fragments are sufficient for combining into a new ligand with (presumably) a high probability of synthetic accessibility. A second limitation is that their description of the protein environment as graphs of atoms causes presumably poor performances compared to graphs of chemical features or graphs of triplets of chemical features as used in MED-SuMo.

Here we present a new computational drug design protocol combining local similarity of protein surfaces and a fragment-based approach. It brings together their respective advantages in an attractive way. The protocol is intended for fragment library design, lead discovery, and lead optimization. We will show in this work that the protocol based on MED-SuMo is very efficient, allowing application of the entire PDB and therefore maximizing the probability of finding relevant local similarities. The MED-SuMo database can contain thousands of pockets and hundreds of thousands of local protein descriptions and queries apply against the whole protein surface. Such full surface queries can help detecting where ligands would bind, like the map-based method MCSS and qualitatively address target druggability.

A first advantage of our fragment-based approach is an innovative fragmentation protocol that defines multiple protein-fragment patterns from any protein−ligand structure. These 3D patterns, called MED-Portions, include chemical moieties which are matching molecules from a chemical library and substructures of protein-bound ligand. MED-Portions, which are the MED-SuMo representation of protein-fragment patterns defined by several criteria: (1) a chemical moiety where atoms are topologically matching with a molecule from molecular libraries, e.g. synthetically accessible molecules or building blocks, (2) open valences filled by 'dummy atoms' that indicate where it was connected in the original ligand, and (3) the protein interaction surface surrounding that chemical moiety described by the MED-SuMo Surface Chemical Features (SCFs). MED-Portions chemical moieties are real molecules, like with the RECAP,[30] but is therefore unlike methods using rings, linkers, and substituents. The main difference with RECAP is that our chemical moieties have a significant overlap. Thus, for example, a ring can be in more than one fragment, and we found this useful for the hybridization algorithm.

A second advantage of this protocol is that the local similarity of proteins is addressed by a database of MED-Portions built from the protein−ligand complexes of the PDB and browsed using the MED-SuMo software:[31] protein-fragment interaction surfaces which are found locally similar

to a protein surface of interest (query) are superposed, and the MED-Portion chemical moieties are collected in the reference frame of the query and form a pool which is used to generate new 3D hybrid compounds by hybridization. The protocol populates any protein surface with MED-Portions.

We applied the protocol and collected a pool of MED-Portions for 107 protein binding sites and report in the Results section the performances in terms of the number of fragments and their physicochemical properties. We found, as expected for a computational fragment-based protocol, a qualitative correlation between the average hydrophobicity of the pools of MED-Portions and those of the binding sites, illustrating the selectivity of the protocol. To generate hitlike molecules, we applied an additional step of hybridization of MED-Portions with MED-Hybridise to two targets that represent important protein superfamilies in drug design: a protein kinase VEGFR-2 (vascular endothelial growth factor receptor 2) in a DFG-out conformation and a β2-adrenergic GPCR (G-Protein Coupled Receptor). On one hand, only a few known bound ligands are known in the PDB for GPCR, and therefore the generated hybrids rely mainly on interfamily hits. On the other hand, the protein kinase example emphasizes that new and relevant scaffolds can be generated from several well-known PDB ligands and that interfamily hits are significant as well. These two target proteins are complementary cases where the retrieved local similarities originate respectively mainly from intra- and interfamily hits.

METHODOLOGY 1. MINING THE PDB TO EXTRACT MED-PORTIONS

The results presented in this paper were generated using software developed by MEDIT SA, except where stated otherwise.

MED-SuMo Technology. MED-SuMo is derived from the SuMo software.[32] MED-SuMo is able to locate similar regions on macromolecular surfaces associated with a defined chemical function. It applies a heuristic based on a 3D representation of macromolecular surfaces using Surface Chemical Features (SCF) like, for example, H-bond donor, H-bond acceptor, formal positive and negative charges, hydrophobic, aromatic, or more specific features like amide and guanidinium. Each feature describes a putative physical interaction including its precise geometrical characteristics. The MED-SuMo comparison methodology can be divided into two major steps: (1) *The Graph Formation:* SCFs are positioned on the macromolecular structure through a lexicographic analysis of the atoms in the PDB files, *i.e.*, a residue is represented by a set of representative SCFs. Their positions and orientations are filtered, discarding any SCFs likely to be involved in intramolecular interactions or too buried to interact with a potential ligand. Remaining SCFs are assembled into triplets (triangles) with specific geometric characteristics including edge length, perimeter length, and angles. The resulting triplet network is stored as a graph data structure in the MED-SuMo database where the triplets are represented as vertices and where two adjacent triplets are connected if they share an edge. As the graph is a surface mesh of chemical features on a protein surface, it can be viewed as a surface description, and we will use the term interaction surface in the following. (2) *The Graph Comparison*: MED-SuMo searches for triplets that are composed of the same SCFs and geometrically compatible. These common triplets are termed "seeds". When a seed is detected, MED-SuMo iteratively extends the comparison to the adjacent triplets until no more similarities are found. The results of these comparisons are similar surface patches or "hits" ranked by a MED-SuMo score which takes into account both the number of common features and the local protein shape similarity. A filtering rule is applied to keep only the hits with a shape similarity above a given threshold (*shape_threshold*).

These interaction surface comparisons are usually performed between a query and a database of precomputed graphs from all the structures available from the PDB. Two distinct types of MED-SuMo databases are commonly used: the first one consists of the protein surface regions surrounding any cocrystallized ligand or small peptide (*the binding site database*), and the second contains the whole protein surface (*the full surface database*). In this paper we describe a new kind of database (*the protein-fragment database*) consisting of protein surface regions surrounding the presence of chemical moieties matching molecules from chemical libraries.

Specific MED-SuMo parameters are optimized for the protein-fragment database: (1) The volume of the fragment environment is defined by the *ligand_radius*, for the current database, and its value is set to 4.5 Å around the fragment atoms in order to discard the weak protein-fragment interactions. (2) The triplet graph construction is constrained with a maximum length of a triplet edge (*edgemax*) of 14 Å; this value allows a rather exhaustive number of SCF triangles, and therefore a representative graph, without considering all of them. (3) Three additional hydrophobic SCFs were added into the dictionary compared to the original publication,[33] two on the aryl chain of the methionine and one on the aryl chain of the lysine to take into account the hydrophobic contribution of these side chains. (4) To favor more interfamily hits by being less restrictive on the shape similarity when two surface regions are compared, the parameter *shape_threshold* is lowered from 0.65 to 0.45 meaning that a hit volume needs to share only 45% of its volume with the query volume.

Automatic Ligand Detection on Input Protein−Ligand Structures. Protein structures are downloaded from the Brookhaven Data Bank, using the remediated PDB archive FTP server[34] (accessed July 16, 2008). Ligand detection is used to build the binding site database and the protein-fragment database and to assist in building automatically the query. A ligand is defined by (i) residues or chains with more than one heavy atom and fewer than 100 (HETATM); (ii) chains with fewer than ten residues (peptide or heteropeptide); and (iii) covalent ligands (HETATM). Water molecules are consequently not treated as ligands in this work. For each PDB file, ligand information is transmitted to MED-SuMo-Fragmentor.

MED-SuMo-Fragmentor (Detection of MED-Portions). The MED-SuMo-Fragmentor's role is to extract all MED-Portions from a set of protein−ligand structures. MED-Portions are the MED-SuMo representation of protein-fragment patterns and are defined by (1) a chemical moiety where atoms are topologically matching with a molecule from molecular libraries, e.g. synthetically accessible molecules or building blocks, (2) open valences filled by 'dummy atoms' (Du) that indicate where it was connected in the original ligand, and (3) the protein interaction surface

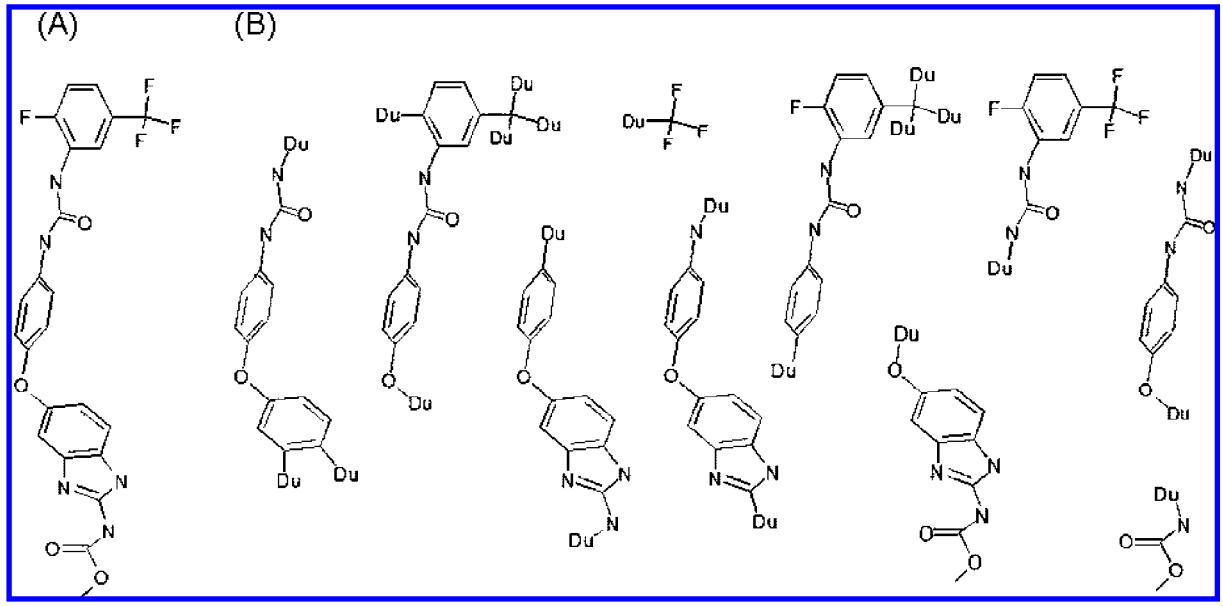

Figure 1. The VEGFR-2 ligand GIG (PDB code 2oh4) is described by 10 MED-Portions: **A.** structure of GIG ligand from 2oh4 PDB file and **B.** structures of the MED-Portions (for clarity, the surrounding SCFs and the protein environment are not shown).

surrounding that chemical moiety described by the MED-SuMo Surface Chemical Features (SCFs).

Bond orders are assigned to PDB ligands from their geometries using OpenBabel 2.2.0,[35,36] and PDB-formatted files are converted to the SDFiles format.[37] No attempt to correct bond order typing errors (e.g., due to erroneous ligand geometry) was undertaken at this stage as they are not significant in the context of intermediate results for this protocol. As a result, manual validation of bond orders may be required before presenting results at the end of the overall protocol. For the presented results, however, no manual corrections were required.

A PDB ligand is considered by MED-SuMo-Fragmentor only if it has six or more atoms. Thus, glycerol is fragmented, whereas small molecules like phosphate or imidazole are already considered as potential MED-Portion chemical moieties. Fragmentation is performed through a substructure matching step between any molecule from a small molecule database and each PDB ligand.

The database of small molecules used in this work was built using PubChem Compounds.[38] For those compounds with more than one molecule, only the largest molecule is retained. Filtering rules were applied to discard undesirable molecules: (1) molecules containing atoms other than the subset [H, B, C, N, O, F, P, S, Cl, Se, Br, I, As, Te, Si]; (2) molecules with MW > 250 Da; and (3) molecules with MW < 70 Da except those containing at least one ring. 2,112,444 molecules were selected out of 19,202,121. After removing structural duplicates, 1,577,071 molecules are stored.

MED-Portions generation is performed by a substructure matching with a depth-first maximum common substructure algorithm. This algorithm disregards bond orders but only matches those atoms with identical implied hybridization states. Known chirality and any hydrogens are also disregarded.

Matching chemical moieties are sorted by decreasing the number of heavy atoms. Very similar matches are removed to reduce redundancy according to the following protocol: the first match is stored; the next match (having n heavy atoms) is stored if it is not a substructure of a previously stored match; if it is a substructure of a previously stored match (having m heavy atoms), then it is stored only if $n \leq x.m$ with x depending on n: $x = 1.5$ if $n \geq 10$, $x = 2$ if $n \leq 5$, x is a linear function of n between 5 and 10. All stored matching chemical moieties include the relevant 3D Cartesian coordinates from their source PDB file.

Extracting a substructure of a PDB ligand implies cutting several bonds and generating a chemical moiety with open valences which is not a physical molecule. When a bond is cut, a dummy atom is added to the chemical moiety, replacing the atom which was removed, so that a virtual bond is maintained. The dummy atom has the Cartesian coordinates and the atomic number of the removed atom. Each resultant match is stored as a MED-Portion in the MED-SuMo protein-fragment database, a Sqlite[39] relational database, which includes the chemical moiety, the dummy atoms, and the protein interaction surface surrounding that chemical moiety described by the MED-SuMo Surface Chemical Features (SCFs).

This new protocol allows most of the PDB protein−ligand complexes to be converted into MED-Portions with a computationally reasonable number of MED-Portions: 81% the 11,011 unique PDB ligands (according to their three letter PDB code) have all their atoms described with an average number of 13.8 MED-Portions (18% of them are described with 1 MED-Portion, 46% with 1 to 5, and 65% with 1 to 10). Two examples are the PDB ligand GIG of VEGFR-2 (PDB code 2oh4[40]) which is shown in Figure 1 and the PDB ligand MMI of beta-secretase (PDB code 1xs7[41]) which is fully described with 47 MED-Portion chemical moieties, a comparatively large number resulting from its 52 heavy atoms and its 16 atom macrocycle. The degree of overlap of the MED-Portions chemical moieties is measured, for each ligand, by the number of atoms in more than one moiety divided by the total number of described atoms: overlap is less than 20% for 20% of the PDB ligands and more than 75% for 50%. As expected, if ignoring peptide ligands, the small molecules are described with significantly fewer MED-Portion chemical moieties: 87% of 7313 ligands are fully described with an average number of 7.1 MED-Portion chemical moieties and with a similar degree of overlap.

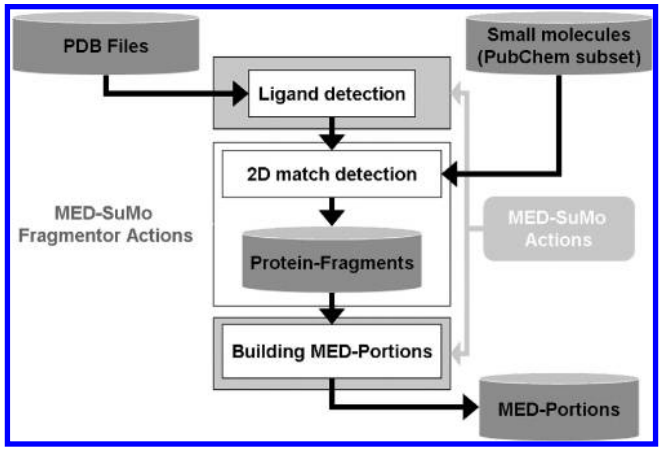

Figure 2. Software architecture to generate MED-Portions (protein-bound fragment patterns).

MED-Portions Database (Protein-Fragment Interactions Surfaces). A MED-SuMo database is built for all detected MED-Portions, *i.e*. the MED-SuMo representation of protein-fragment interactions which are defined for each entry by a chemical moiety, dummy atoms (open valences after bond cuts), the 3D protein structures, and the SCFs to describe the local environment. This MED-SuMo protein-fragment database contains 395,360 MED-Portions. Each PDB entry is associated with a list of MED-Portions (Figure 2).

METHODOLOGY 2. POPULATION OF PROTEIN SURFACES WITH MED-PORTIONS

Application of MED-SuMo for Fragment Mapping: Populating a Potential Protein Interaction Surface with a Pool of MED-Portions. The MED-Portions Protein-Fragment interaction database is queried with the same interfaces (MED-SuMo GUI, MED-SuMo CLI, Scitegic Pipeline Pilot 7.0,[42] MOE 2003[43]) as the other MED-SuMo databases. A protein surface is submitted as a query, and the results are collected as a pool of MED-Portions in the reference frame of the query surface. The query may be any user-defined protein surface region. For example, a query could be the environment of a cocrystallized ligand, a user-defined set of residues, the whole surface, or any logical combination of those. We chose an unusually large distance of 10 Å for the *ligand_radius* to define a query sufficiently large to explore the whole pocket of any protein−ligand complex with a default value.

A resultant MED-Portion has SCFs in common with the query protein: a SCF binary signature is built where 1 is associated with a matched SCF and 0 with an unmatched SCF. Matched SCF are in the 4.5 Å neighborhood and, therefore, have a significant contribution to the fragment-protein interaction. The interaction is qualitatively estimated through the MED-SuMo score which is proportional to the number of matched SCFs. A quantitative estimate of the fragment-protein interaction energy is not determined in this work. The list of the top 50,000 collected MED-Portions (hits), sorted by decreasing MED-SuMo score, is output for further analysis. The output is either an XML file or an annotated SD file. A run takes typically 20 min for a binding site and 1 h for the whole surface of a 100 kDa protein (2 Xeon Quad Core 5335, 16 GB RAM).

In order to yield a nonredundant pool of MED-Portions, 3D duplicates were removed. MED-Portions were deemed duplicates when chemical moieties were identical in 2D topology and had atoms with similar 3D coordinates. The topological similarity is quickly assessed simply by comparing a canonical representation of the molecule using SMILES.[44] Our canonical representation differs from that in the original publication,[45] CANGEN, although the basic algorithm remains the same. The atomic properties that we consider invariant for determining the ordering of atoms in the canonical representation are, in decreasing order of significance, as follows: number of attached heavy atoms, atomic number, sign of formal charge, absolute value of formal charge, number of hydrogen atoms, and hybridization state. We find this selection of invariant properties better matches our internal representation of the structures (both chemically valid and hypothetical).

MED-Portions Selection and Analysis. MED-Portions can be filtered, selected, and analyzed in different ways considering their (1) MED-SuMo hit descriptors, e.g. MED-SuMo score and rmsd of the matched SCFs; (2) bioinformatic descriptors, e.g. PFAM domain[46] annotation of the chain(s) to which the PDB ligand is bound; (3) atoms (excluding dummy atoms) to compute physicochemical properties, cheminformatic descriptors, steric clashes (bumps) with the query protein, and affinity scores (not considered in this work); and (4) SCFs environment (not considered in this work).

The analysis of physicochemical properties and bumps is performed with Discovery Studio 2.0:[47] the number of pairs of contacts with the query is computed after the removal of water molecules, hydrogen atoms, and all heteroatoms (except HEM).

A MED-SuMo score cutoff at 3.1 is used to remove most of the irrelevant hits. An additional score, MED-SuMo score efficiency, is used to take into account the size of the ligand by dividing by the heavy atom count (HAC), by analogy to Hopkins et al.:[48] MED-SuMo score efficiency = MED-SuMo score * HAC_{150}/HAC where HAC_{150} is the average number of heavy atoms found in MED-Portion chemical moieties with a MW of 150 Da (HAC_{150} = 11, determined empirically). This means that hybrids with more than 11 heavy atoms would need a higher MED-SuMo score than 3.1 to be not filtered out. A few MED-Portion chemical moieties shown in Figure 1 have 16 heavy atoms, which corresponds to a cutoff MED-SuMo score of 4.5 which is much more significant than 3.1 for a MED-SuMo hit and needed for such large chemical moieties.

For all targets considered in the Results section, we selected the MED-Portions according to (1) MED-SuMo score ≥3.1; (2) MED-SuMo score efficiency ≥3.1; and (3) bumps <10.

METHODOLOGY 3. HYBRIDIZATION OF MED-PORTIONS

MED-Hybridise (Combining MED-Portion Chemical Moieties). MED-Hybridise is a tool for recombining chemical moieties from MED-Portions into putative ligand molecules that are meaningful in the context of drug discovery and development. It also provides general-purpose functionality for filtering and analyzing both the input MED-Portions sets and the generated hybrid molecules. Only the aspects that were applied to create the results of this paper are described. Several algorithms are available for combining substructures into molecules. For this work, we applied an

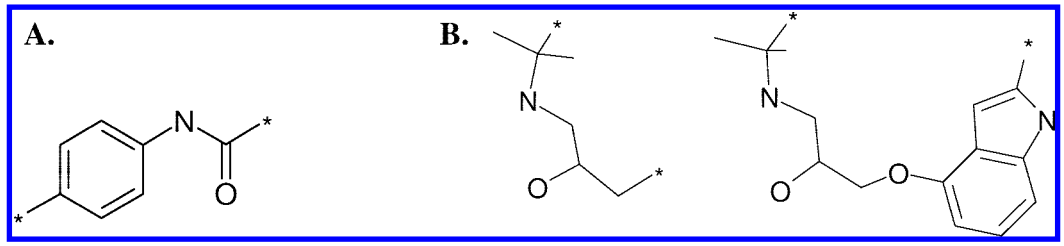

Figure 3. Set of potentially interesting moieties used in the Results section. **A.** Moiety used for VEGFR-2: phenylamide from GIG ligand in 2oh4 PDB file. **B.** 2 of the 156 moieties for GPCR considered since they have a nitrogen atom close to N18 of CAU ligand in 2rh1 PDB file.

Table 1. Results of the Diverse Set of Protein Pockets

pocket characterization[a]							pool of MED-Portions characterization[d]									
rank	protein	PDB	LIG	SCF	surface[b]	%hydr.[c]	Nb	MW	HBD	HBA	ALogP	ClogSw	PSA	RT	RO3	%RO3
1	HIV-1 protease	1dif	A85	156	2496	0.4	10705	155	1.9	2.9	0.7	−1.6	49	3.7	1993	19
2	intestinal fatty acid binding protein	1icn	OLA	152	2432	0.6	3508	142	1.2	2.2	1.4	−1.6	37	2.2	1592	45
3	beta-secretase 1	1xs7	MMI	181	2896	0.3	6287	144	1.8	2.6	0.7	−1.4	45	3.6	1274	20
4	vascular endothelial growth factor receptor 2	2oh4	GIG	158	2528	0.4	1474	149	1.3	2.4	1.8	−1.8	39	1.2	943	64
5	epididymal retinoic acid-binding protein	1epb	REA	129	2064	0.6	2018	137	0.7	1.4	2.2	−2.0	23	2.3	918	45
6	alpha-thrombin	1dwc	MIT	139	2224	0.3	1958	154	1.4	1.5	0.9	−1.6	46	1.7	835	43
7	major urinary protein i	1i05	LTL	98	1568	0.7	1769	133	0.8	1.6	1.7	−1.7	28	2.2	819	46
8	trypsin	1mtw	DX9	93	1488	0.3	1734	152	1.6	2.7	0.9	−1.6	47	2.3	704	41
9	vitamin d nuclear receptor	1db1	VDX	197	3152	0.5	1374	131	0.9	1.6	1.7	−1.8	28	2.2	618	45
10	Lck kinase	1qpe	PP2	122	1952	0.4	1029	162	1.6	3.1	1.3	−1.8	48	1.2	563	55
11	orphan nuclear receptor pxr	1ilh	SRL	210	3360	0.5	1117	136	1.0	1.9	1.5	−1.6	32	2.2	545	49
12	constitutive androstane receptor	1xnx	ATE	133	2128	0.6	964	135	0.9	1.8	1.5	−1.6	32	1.8	537	56
13	vitamin d3 receptor	1ie9	VDX	182	2912	0.6	1242	132	0.8	1.6	1.7	−1.8	28	2.2	535	43
14	bile acid receptor	1osv	CHC	201	3216	0.4	1001	140	0.9	1.8	1.6	−1.7	33	2.0	505	50
15	prostaglandin h2 synthase-2	1cvu	ACD	167	2672	0.5	882	137	1.0	1.8	1.7	−1.7	30	2.0	452	51
16	Kes1 protein	1zhy	CLR	157	2512	0.5	980	138	0.9	1.6	1.8	−1.9	28	2.2	444	45
17	Fk506 binding protein	1fkf	FK5	122	1952	0.4	822	134	1.1	2.2	0.9	−1.3	38	1.8	435	53
18	acetylcholinesterase	1acj	THA	126	2016	0.5	753	128	1.4	2.1	0.9	−1.2	38	1.9	386	51
19	leukotriene a-4 hydrolase	1hs6	BES	114	1824	0.3	905	134	1.8	2.9	0.2	−0.7	53	2.0	383	42
20	sex hormone-binding globulin	1lhu	EST	124	1984	0.5	908	135	0.7	1.5	2.0	−2.0	24	2.4	371	41
21	estrogen receptor beta	1hj1	AOE	150	2400	0.6	732	153	0.7	1.4	2.4	−2.4	24	1.8	368	50
22	retinol binding protein	1fen	AZE	144	2304	0.5	651	135	1.1	1.9	1.4	−1.5	33	1.8	366	56
23	thymidine kinase	1e2k	TMC	135	2160	0.4	884	140	1.7	3.0	0.5	−0.9	53	1.8	362	41
24	prostaglandin h2 synthase-1	1ht8	34C	112	1792	0.5	612	140	0.9	1.9	1.9	−1.7	32	2.0	318	52
25	uteroglobin	1utr	PCB	157	2512	0.5	618	132	0.7	1.5	1.9	−1.9	25	1.9	315	51
26	retinoid x receptor, beta	1h9u	LG2	141	2256	0.5	651	141	0.8	1.6	2.0	−2.0	27	2.4	286	44
27	glycolate oxidase	1al7	HST	170	2720	0.3	641	135	1.5	2.6	0.6	−1.0	45	2.1	275	43
28	phospholipase a2	5p2p	DHG	146	2336	0.3	309	156	1.5	3.0	0.8	−1.4	50	3.6	237	77
29	calmodulin	1ctr	TFP	128	2048	0.4	373	129	0.9	1.6	1.9	−1.7	27	1.6	232	62
30	6-phosphofructo-2-kinase/ fructose-2,6- bisphosphatase	1bif	ATG	199	3184	0.3	1189	146	2.7	4.3	−0.8	−0.8	88	2.2	230	19
31	cytochrome p450 3a4	2v0m	KLN	183	2928	0.4	417	131	0.9	1.7	1.5	−1.6	30	2.1	214	51
32	Igg2a-kappa 26−10 fab	1igj	DGX	111	1776	0.3	576	154	1.5	2.4	1.1	−1.7	41	2.3	202	35
33	sialidase	2sim	DAN	104	1664	0.2	444	142	1.7	3.2	0.1	−0.6	55	2.0	191	43
34	protein (dihydrofolate reductase)	1boz	PRD	140	2240	0.5	433	147	1.5	2.5	1.4	−1.5	44	2.1	172	40
35	beta-2-adrenergic receptor/ t4-lysozyme chimera	2rh1	CAU	106	1696	0.4	400	137	1.7	2.8	0.3	−0.9	52	2.2	171	43
	vascular endothelial growth factor receptor 2	2oh4	full[e]	885	14160	0.3	1677	148	1.4	2.5	1.4	−1.6	42	1.4	965	20

[a] "Pocket characterization" columns correspond to the MED-SuMo query, *i.e.* protein environment around the ligand described by surface chemical features (SCFs). The "Pool of MED-Portions characterization" columns correspond to the output of MED-SuMo resulting from the query on the protein-fragment database. [b] Discontinuous surfaces areas 10 Å around the ligand are estimated in Å^2 by extrapolation from the number of SCFs multiplied by a empirically derived factor. This factor has been derived from a regression obtained for a set of protein full surfaces, where the surface can be more straightforwardly estimated. This factor has thus been derived to be 16. [c] %Hydr: hydrophobicity of the pocket estimated as the number of hydrophobic SCFs divided by the sum of hydrophobic, H-bond acceptor, H-bond donor, and formal charges SCFs. [d] Nb: number of unique MED-Portions in the pool. MW: molecular weight. HBD: H-bond donor count. HBA: H-bond acceptor count. PSA: polar surface area. RT: rotatable bonds count. AlogP, ClogSw, PSA, and RT are the average values for the pool. RO3 is the count of rule of 3 compliant MED-Portions, and %RO3 = RO3 * 100/Nb. Only the top 35 best targets in terms of RO3 are reported here (4th is 2oh4 and 35th is 2rh1) from a set of 107 diverse surfaces (see the Supporting Information). [e] The query is not a pocket but the full surface of the protein.

algorithm that we term Chain Combine. This algorithm resembles a component of the classic BREED method[20] which was originally designed for the hybridization of ligand molecules, i.e., PDB ligands prealigned in 3D by structural alignment (fold or C-alpha atoms) of the corresponding ligand-bound proteins. MEDIT has developed a new approach which applies to MED-Portions, a local description of protein-bound ligand interaction. This method has the significant advantage, in this context, of favoring a local alignment instead of a global protein fold alignment. The basic principle of Chain Combine is to look for overlapping bonds between a pair of aligned MED-Portions. Chemical

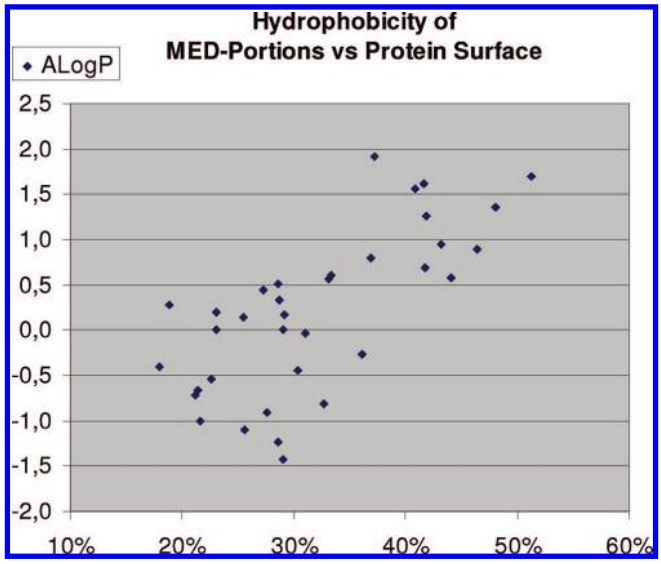

Figure 4. The hydrophobicity of the pools of MED-Portions (average ALogP of MED-Portions chemical moieties) is plotted versus the hydrophobicity of the corresponding query protein surface.

moieties are cut at this bond and hybrids are constructed from the alternate pairs. This matching of bonds was inspired by a component of the BREED method where a bond is considered matching if both bonds have the same bond order, both pairs of corresponding atoms are within a predetermined distance, and the directional vector describing the two bonds are within a predetermined angle. A distance of 1.0 Å and an angle of 15 degrees are used by default. The coordinates of the hybrid atoms involved in the matching bond are set to the mean of the original atoms' coordinates.

The hybridized molecules generated by our protocol have advantages compared to the classic BREED: as we use a query protein and we exploit local similarities with MED-Portions, the hybrids are likely to fit without bumps within a query; interfamily hits can occur and hybrids can be generated in principle for targets without known ligands in the PDB (*cf.* Results section); our protocol is fully automated from the protein-binding site to the hybrids generation.

The hybridization favors (presumably) synthetically accessible molecules. On one hand, MED-Portions contain a chemical moiety where atoms are topologically equivalent to a molecule from PubChem Compounds which are very likely to be synthetically accessible. Therefore, an arbitrary hybridization is more likely to be synthetically accessible than a combination of arbitrary substructures or sets of atoms. On the other hand, dummy atoms allow generation of bonds between chemical moieties where substitutions were actually occurring, and these bonds are therefore more likely to occur in a synthetically accessible molecule.

Combining MED-Portions into Potential Ligands. We considered two targets to apply a generation of hybrids from the pool of collected MED-Portions with MED-SuMo: a protein kinase receptor VEGFR-2 (PDB code 2oh4[40]) and a β2-adrenergic GPCR (2rh1[49]). For each target, the first hybridization step is between a set of potentially interesting starting chemical moieties and the list of selected MED-Portions. One or more starting chemical moieties are selected for each case study (2D depiction is presented in Figure 3). Each step of hybridization starts with all input MED-Portions and the hybrids of the precedent step. Between each hybridization step we applied different filters. First, we filtered the hybrids to keep only those which possessed a particular substructure in 3D coordinates: a phenylamide for VEGFR-2 and an N atom near the N18 of carazolol (3 Å) for the GPCR. The hybridization is therefore biased toward the desired hybrids. Additional filters are (1) hybrids containing a ribose (very frequently found and not relevant for GPCR), (2) 3D duplicates, (3) hybrids with unusual valences, and (4) hybrids with internal steric clashes. In the case of GPCR, we calculated the intermolecular bumps between the hybrids and the query protein. This step was unnecessary in the case of the kinase since most MED-Portion chemical moieties derive from intrafamily hits. We then filter all the hybrids having more than five intermolecular bumps after the second hybridization step. We report in the Results section the union of all steps, though the last step contains almost all of the unique hybrids.

Characterizing the Hybrids. The generated hybrids were subjected to comparisons in terms of scaffold and framework to the PDB ligands, to PubChem Compounds and to known actives according to PubChem Compound bioassay results specific for each superfamily.

Scaffolds (SC) and framework (FW) are computed according to Bemis and Murcko.[50] In the case of the protein kinase case study, we modified the scaffold definition by an additional constraint: if a phenylamide moiety is found in the molecule, then that is kept in the scaffold. This means that the results are more accurate in the protein kinase study where a phenylamide moiety is decorated: the scaffolds of the hybrids and those found in PubChem and the PDB do contain the phenylamide moiety. Frameworks (FW) are two-dimensional molecular structures computed from the scaffolds irrespective of atom type, hybridization, and bond order. The hybrids are also characterized by their physicochemical properties after the terminal dummy atoms are removed, and nonterminal dummy atoms are replaced with carbon atoms.

Additionally, in the case of the GPCRs the entire set of hybrids was compared by 2D similarity to a list of specific β2-adrenergic receptor ligands. This comparison was performed, since, in contrary to the kinases, there are only three structures with GPCR ligands in the PDB. The 2D similarity between two molecules was assessed using a topological fingerprint that resembles Daylight fingerprints[51] but with some variations in implementation. This particular fingerprint is designed for unbiased molecular similarity comparisons, considering just the 2D topology of molecules. Each set of atoms that comprises a linear path of one to seven atoms, inclusive, through the bonds of a molecule are detected. Hydrogen atoms are disregarded. These paths are each described with an integer hash value based on the identity of each atom's element and the order of each connecting bond. The integer values are mapped onto a bitstring of 1024 bits, and this is termed the MEDIT Topological Path Fingerprint. The fingerprints of two molecules, and therefore the topological structures of those molecules, can be assessed using any standard discrete similarity metric; in this work the Tanimoto metric[52] is used. This yields a value between 0.0 and 1.0, inclusive, where 1.0 corresponds to identical molecules and 0.0 to molecules with no common topological paths.

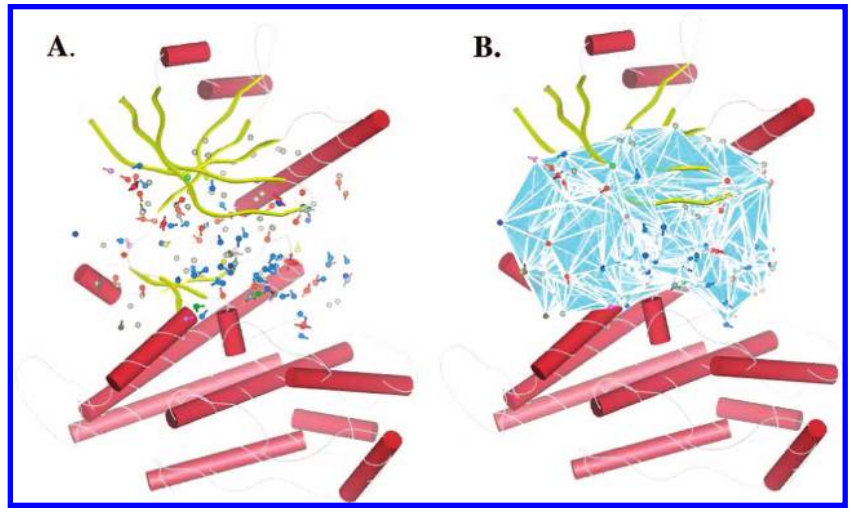

Figure 5. The query definition for the MED-SuMo search. The binding site of the crystallized structure of the VEGFR-2 protein kinase. PDB code: 2oh4[40]. **A.** The different SCFs that were considered for the query (colored balls or ball and stick). **B.** The query is a graph of triangles which are shown here. They form a surface mesh of chemical features on the protein surface. Top: N lobe; bottom: C lobe.

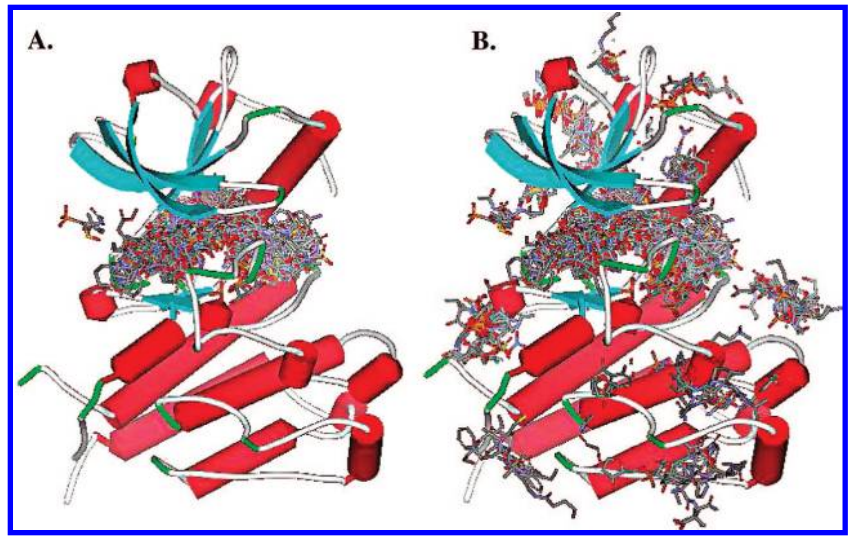

Figure 6. The protein kinase MED-Portions chemical moieties. A. 1474 are collected in the GIG ligand pocket. B. 1677 are collected on the full surface. Only 200 more are collected by using a full surface query instead of a binding site query: MED-Portions are collected mostly on interaction sites.

RESULTS

A. Characterization of the Pool of MED-Portions for a Diverse Set of Protein Pockets. We selected (rather arbitrarily) a tentatively unbiased set ($n = 107$) of ligand-bound protein from the PDB by picking from two sets designed for two different purposes: (1) We selected targets in most of the therapeutic areas from the database of therapeutically relevant targets PDTD (Potential Drug Target database).[53] (2) We added a few pockets from the pocket detection work Q-SiteFinder from Laurie et al.[54] which are presumably unbiased in terms of pocket size and hydrophobicity. (3) We also added the recently deposited GPCR structure 2rh1. We describe each pool of MED-Portions in Table 1 by the number of unique MED-Portions, their average physicochemical properties, and the hydrophobicity of the pocket. We found that the hydrophobicity of the protein surface is qualitatively correlated to the average hydrophobicity of the MED-Portions (Figure 4). This result is highly anticipated for a computational fragment-based protocol and emphasizes the relevance of the MED-Portion chemical moieties and the selectivity of the protocol. The number of collected MED-Portions (Nb) and the number and percentage (%) of them which are rule of 3 (RO3) compliant[55] are dependent on many factors: occurrence of the query protein with various ligands from the PDB, physicochemical properties of these ligands, and surface and hydrophobicity of the query among others. From these results a correlation could be established that would allow predicting the pockets which are amenable to this fragment based drug design protocol (not analyzed here).

Hybridization of the MED-Portions into Potential Ligands. In this work, we report the hybridization results of MED-Portions which are a local description of protein−ligand structures. To demonstrate the advantages of using MED-Portions instead of whole ligands, we compared these two approaches with MED-SuMo in similar conditions for MED-Portions (*protein-fragment database*) and whole ligands (*binding site database*): query = 10 Å around ligand, MED-SuMo score ≥3.1, and a filter of 10 bumps with the query. Since the collected hits from each database are both originating from the PDB ligands, either whole ligand or portion of, the comparison is done on the number of unique

Table 2. Results of the Hybridization Protocol[a]

characterization of the hybrids	protein kinase (2oh4)	GPCR (2rh1)
hybrid count	221 281	40 860
hybrid count (unique 2D)	218 085	3818
MW	400 ± 79	244 ± 77
rotatable bond count	6.1 ± 1.9	4.8 ± 2.6
ring count	3.5 ± 0.9	2.0 ± 1.1
framework (FW)	1257 (100%)	121 (100%)
FW found in PubChem	832 (66%)	84 (70%)
FW found in corresponding PubChem bioassay	188 (15%)	50 (41%)
FW found in the PDB	56 (4%)	20 (17%)
FW found in the PDB ligands of the same Pfam domain	14	3 (2%)
scaffold (SC)	9936 (100%)	729 (100%)
SC found in PubChem	549 (3%)	110 (15%)
SC found in PubChem bioassay	27 (3%)	43 (6%)
SC found in the PDB	175 (2%)	67 (9%)
SC found in the PDB ligands of the same Pfam	83 (1%)	3 (0.4%)

[a] Hybrids were generated as described. Hybrids contain dummy atoms and are filtered for 3D duplicates. The terminal dummy atoms in the 'Unique 2D' hybrids were deleted, and the nonterminal dummy atoms were replaced with a carbon atom prior to 2D filtering. '±' indicates an average value on its left and a standard deviation on its right.

pairs of PDB-LIG in the hit list (for instance 1s9i-5EA, 1s9i-ATP, 1iep-STI, and 1xbb-STI are 4 unique pairs). This number indicates the volume of data which is hit from the PDB and which we assumed to be relevant for drug design. Although, this comparison is performed with specific software, it has presumably a more general value because the mining is tractable at PDB scale and both intrafamily and interfamily hits are collected. For the protein kinase (PDB code 2oh4), 147 ligands are collected (765 ligands prior to bump filter). Ultimately 144 pairs are unique. Whereas we have collected 2407 MED-Portions (1474 with the ligand efficiency filter) originating from 968 unique pairs which is a 7-fold improvement compared to the whole ligand approach (968 vs 144). This high improvement is mainly due to the fact that the query is a DFG-out protein kinase and that most of the DFG-in ligands which are much more frequent in the PDB (40 times) cannot fit as a whole in a DFG-out pocket, but substructures of them can. Similar numbers are obtained with GPCR: 60 ligands are collected (1387 ligands prior to bump filter). Ultimately 57 are unique. Whereas we have collected 817 MED-Portions (400 with the ligand efficiency filter) originating from 449 unique PDB ligands which is a 8-fold improvement compared to the whole ligand approach (449 vs 57). In conclusion of the cases studied: intrafamily ligands are likely to fit in proteins sites of other proteins of the same superfamily since the binding sites are similar on the whole, while with interfamily hits the binding sites are more likely to be globally different but locally similar. In the following, we used an efficiency filter for the MED-Portions in order to focus on the most promising MED-Portions which ideally have a high MED-SuMo score and a small number of heavy atoms.

A. Protein Kinase As a Validation for Intrafamily Drug Design. The kinase superfamily[56,57] is one of the largest in the genome. Their common key function in signal transduction for all organisms makes it a very attractive target class for therapeutic interventions in many disease states such as cancer, diabetes, and inflammation such as arthritis.[58,59] The ATP binding site is located at the interface between two lobes forming the specific kinase fold. In 2001, the first orally administered kinase inhibitor imatinib was approved as a potent agent against chronic myelogenous leukemia (CML), proving that the structurally conserved ATP pocket can be selectively addressed by small molecules. The crystal structure was solved[60] and revealed that imatinib captures a specific inactive conformation (i.e., DFG-out) by occupying the ATP pocket and created a new pocket by a large displacement of F of the DFG motif. The two pockets are connected by the region of the gatekeeper residue playing a crucial role in the binding of DFG-out ligand.

In this work, we selected a DFG-out structure of one of the most pursued protein kinase VEGFR-2. The query protein with the SCFs and triplets from MED-SuMo is shown in Figure 5.

The described protocol focuses on generation of potential DFG-out inhibitors. We generate hybrids in 3D in a DFG-out pocket, and we strictly filter hybrids as containing the phenylamide substructure of the GIG ligand in the gatekeeper region of the 2oh4 structure. This is a critical choice for our DFG-out ligand generation as the phenylamide close to the gatekeeper is observed only in DFG-out protein-bound ligands and therefore favorably biasing our hybrids. It is worth noting that the protocol can generate DFG-in inhibitors as a DFG-out inhibitor can bind with a different conformation to a DFG-in structure like imatinib to the c-Abl kinase DFG-in[60] and to the syk kinase DFG-out.[61] The protein-fragment database contains both DFG-in and DFG-out structures which makes sense as the important hinge region is not modified, and therefore DFG-in MED-Portions contribute to DFG-out hybrids. After the MED-SuMo run and the selection of the MED-SuMo Portions as described above, we obtained 1474 MED-Portions (Figure 6). Among them, 25% came from proteins from a different PFAM Clan than protein kinase and are therefore interfamily hits. This percentage is significant, and interfamily hits probably play a role in the hybridization results, though we did not investigate their exact impact and relevance on the generated hybrids' activities.

The example of a full surface query shows that even by submitting the full surface of a protein kinase, the results on the pool of MED-Portion chemical moieties are very similar (see Figure 6 and Table 1) to those with only the binding pocket around the GIG ligand: 88% of the MED-Portions are in the pocket. In that specific case the protocol is able to detect the binding pocket. In future work we hope to evaluate on a diverse set of full surface whether this fragment-based drug design protocol is able to detect cavities and assess druggability.

To validate our study we searched chemical libraries to identify potential DFG-out ligands and to check whether they are annotated as active on protein kinase. Success rate with a library of known actives cannot be quantitatively estimated because DFG-in ligands are unlikely to be generated by our protocol and would, therefore, be a source of misinterpreted false negatives. The results of the hybrization protocol are given in Table 2. We performed a scaffold analysis of the hybrids, generating more than 9000 unique scaffolds. 175 are matching PDB ligands (19 are unique), demonstrating that the protocol is able to retrieve PDB ligands and, more

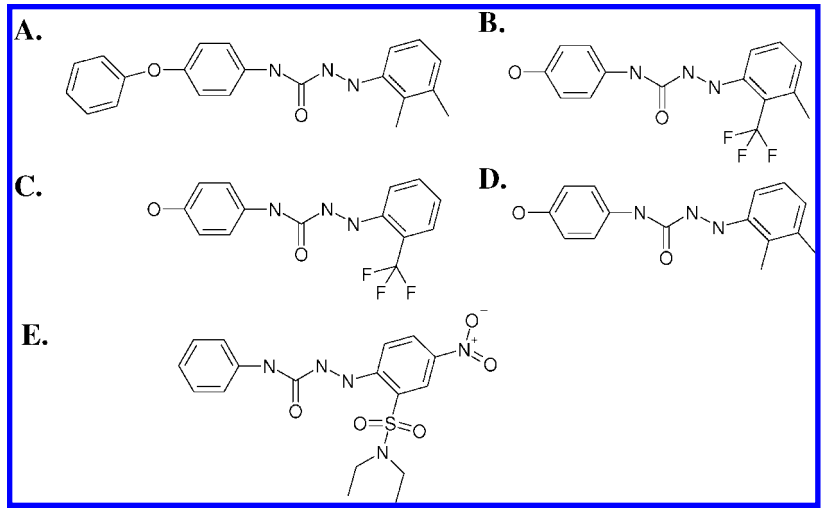

Figure 7. A-D: 2D structures of the four hybrids containing the scaffold of the PubChem compound as a substructure (CID 3527591) represented in **E.**

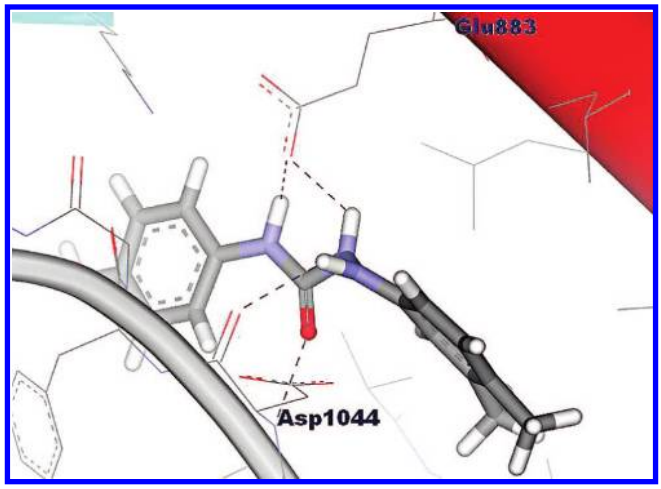

Figure 8. Molecule D from Figure 7, represented with its 3D coordinates in the query (VEGFR-2 structure 2oh4). This molecule is a hybrid of 4 MED-Portions from the PDB files: 2hz0, 2oh4, 1wbn, 3ctq (all four are protein kinase DFG-out structures) and the starting phenylamide moiety. The geometry of the hybrid had been energy minimized to relax the benzohydrazide group (the phenylamide was kept fixed). An H-bond is found for the DFG-out ligand between the carbonyl of D1044 (DFG motif) and the benzohydrazide group.

interestingly, to generate more than 8000 new scaffolds not seen in the PDB. 83/175 matches are protein kinase PDB scaffolds (the scaffolds are substructure of PDB ligands: GIG, L09, BMU, L10, 1PP, G2G, BAX, 2RL) which is an indication that the hybrids are focused on protein kinase. The comparison to a large database of molecules (PubChem Compounds) shows that there are 549 matches (294 unique scaffolds) which are potentially leads for DFG-out ligands. In addition to this study of known actives, we searched the database of VEGFR-2 decoys from DUD (dud.docking.org).[62] We found only two scaffolds (scaffolds of molecules ZINC00341936 and ZINC00570337), giving indication on the presumably small rate of generated false positives. Interestingly, 27 matches (13 unique) are annotated as active on protein kinase, corresponding to 735 ligands in PubChem protein kinase bioassays. Potentially, these could be tested on different protein kinase targets known to bind DFG-out ligands.

A few interesting scaffolds for a retrospective validation are found in actives of PubChem protein kinase bioassays but not in the PDB ligands. Most of them have minor differences when compared to PDB ligands, e.g., N to O on a linker or C to N on an aromatic ring. One particular ligand is interesting and corresponds to the PubChem Compound with CID = 3527591. This is annotated as a FAK kinase inhibitor and contains an original benzohydrazide moiety compared to the PDB. Our results suggest that CID 3527591 is a potential DFG-out ligand of protein-kinase FAK (Figures 7 and 8). This of course would need to be validated experimentally. Interestingly, a recent structure deposited in the PDB (3c1x[63]) is a protein kinase complex with a DFG-out ligand containing a malonamide. That ligand shares some similarity with our hybrid as it is different from the usual amide or urea moiety found at this position.

B. GPCR As a Validation for Interfamily Drug Design. G protein-coupled receptors (GPCRs), or seven transmembrane domain receptors, constitute one of the largest superfamilies in the human genome.[64] They are found only in eukaryotes, where they are active in nearly every organ. The GPCRs mediate responses to visual, olfactory, hormonal, neurotransmitter, and other signals outside the cell and activate inside signal transduction pathways through stimulation of members of the ubiquitously expressed family of heterotrimeric G proteins or β-arrestins, thereby regulating the intracellular levels of various second messengers and thus cellular responses.[65,66]

Among more than 791 GPCR genes encoded by the human genome, about 375 are of pharmaceutical interest (the nonolfactive ones).[67] Currently, the drugs available on the market address only 30 GPCRs. Thus the large majority of the GPCRs still remain promising drug targets in areas including cancer, cardiac dysfunction, diabetes, central nervous system disorders, obesity, inflammation, and pain relief so that some of the pharmaceutical companies are focused exclusively on this superfamily.[68] Nonetheless, for a long time the crystal structure of only a single GPCR has been solved: the visual sensory protein rhodopsin in its ground and photoactivated state.[69,70] Now the way at last appears cleared for more rapid progress with the appearance of several papers presenting structures of the β2-adrenergic

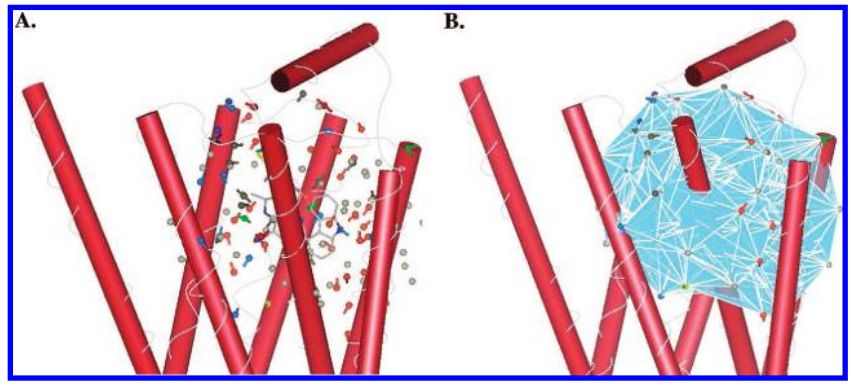

Figure 9. The query definition for the MED-SuMo search. The binding site of the crystallized structure of the β2-adrenergic receptor with carazolol. PDB code: 2rh1.[49] **A.** The different SCFs considered for the query (colored balls and sticks). **B.** The triangles built from the SCF are shown. Top: extracellular; bottom: intracellular.

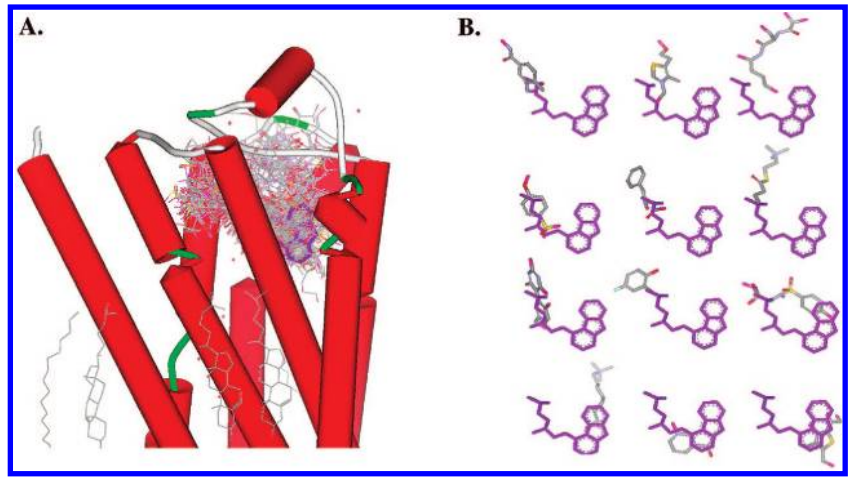

Figure 10. The hit selection. The hits were selected with a MED-SuMo score >3.1 and a maximum of 10 bumps with the query. Thus we obtained 400 MED-Portions chemical moieties that are depicted in **A.** The initial ligand (carazolol) is drawn in purple. Some of these chemical moieties are depicted in **B.** Top: extracellular; bottom: intracellular.

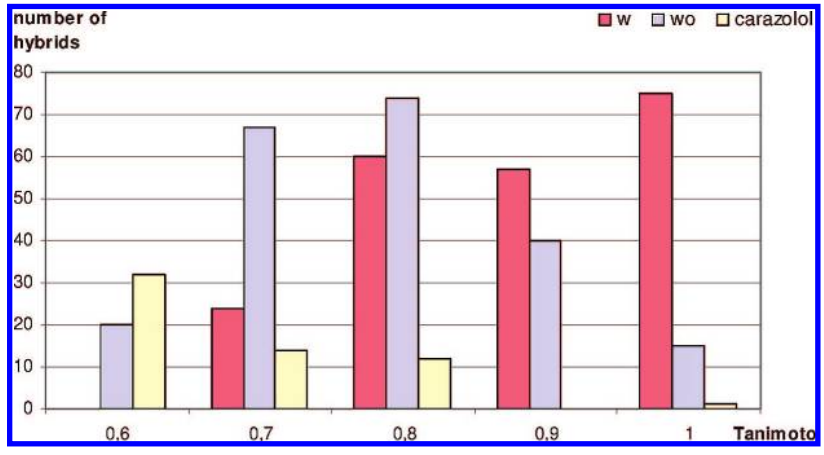

Figure 11. Repartition of the hybrids similar to known β2-adrenergic receptor ligands obtained by using all the MED-Portions retrieved with MED-SuMo (w, pink) or only from the interfamily MED-Portions (wo, blue). A simple 2D ligand similarity search gives the result depicted in yellow (carazolol). In the case of only interfamily MED-Portions, we still obtain more then 50 hybrids that are similar to known ligands with a Tanimoto similarity >0.8.

receptor (PDB code 2rh1,[49,71] 3d4s[72]) and β1-adrenergic receptor (PDB code 2vt4[73]).

For our study we submitted the surface from the β2-adrenergic receptor structure around the ligand, carazolol (PDB code 2rh1), to MED-SuMo on our database of MED-Portions (Figure 9). After the selection of the MED-Portions as described above we obtained a subset of 400 MED-Portions. Interestingly, the chosen MED-Portions populated the binding site in a similar way to that obtained in a High Throughput Docking study by Topiol et al.[74] (Figure 10.A). Some of the obtained MED-Portions chemical moieties are depicted in Figure 10.B, compared to the position of the initial ligand carazolol.

The selected MED-Portions were exported from MED-SuMo and subjected to several hybridization steps as described above. Prior to analysis, we selected the most promising compounds based on physicochemical properties,

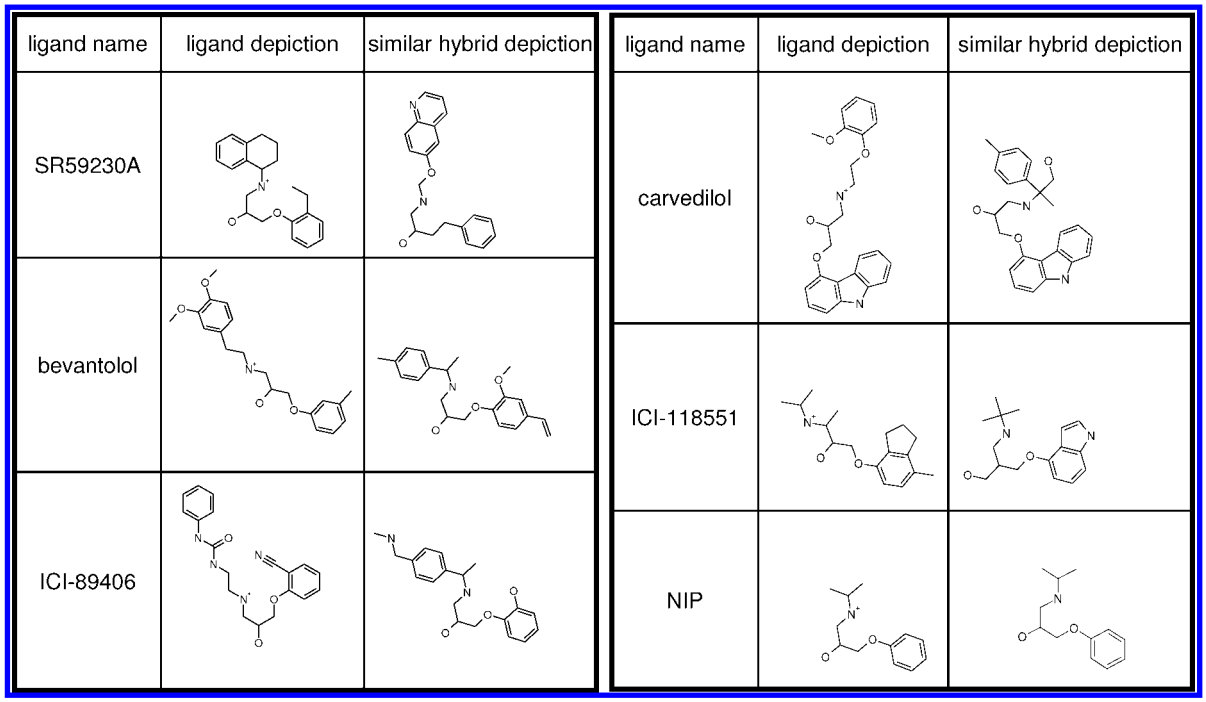

Figure 12. Some of the hybrids similar to known β2-adrenergic receptor ligands. For this comparison we used MEDIT topological path fingerprints followed by Tanimoto metric analysis.

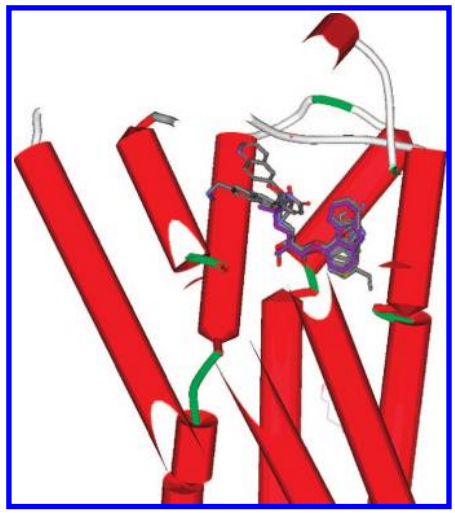

Figure 13. 3D depiction of the superimposed hybrids inside the binding pocket compared to the carazolol position from the crystal structure. Top: extracellular; bottom: intracellular.

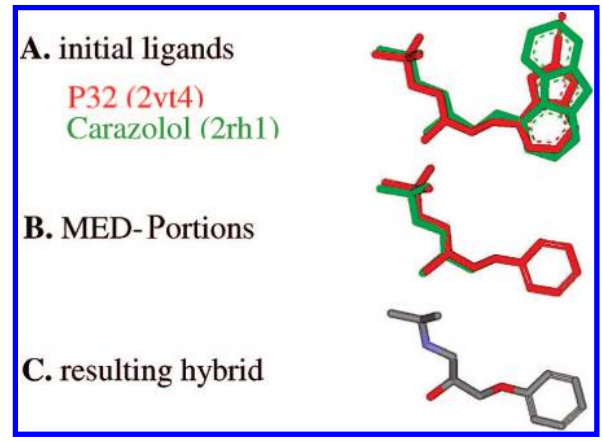

Figure 14. Generation of the NIP-hybrid. **A.** The initial ligands are depicted (the red P32 comes from the crystal structure of the turkey β1-adrenergic receptor (PDB code 2vt4); the green from the human β2-adrenergic receptor (PDB code 2rh1). **B.** The MED-Portion chemical moieties that are generated in our MED-SuMo database from the initial ligands that were used for the hybridization (for clarity, the SCFs and the dummy atoms are not represented here). **C.** The hybrid generated from these two MED-Portion hits corresponds to the existing NIP ligand.

and, after removing 2D duplicates, 3818 unique molecules were obtained.

As for the kinase study, a scaffold analysis of the hybrids was performed. In the case of the GPCRs, 729 scaffolds were generated which is fewer than for the kinases (Table 2). This is not surprising since the PDB material is less substantial, with only three GPCR structures available. Nevertheless we successfully retrieved 6% (43 out of 729) of the scaffolds present in the PubChem bioassays on the GPCRs. The comparison in terms of frameworks gave us 41% (50 out of 121) of the frameworks that matched those from PubChem bioassay on the GPCRs. This difference between the comparison in terms of scaffolds and frameworks implies that we are able to generate diverse hybrids in terms of scaffolds, which is partly due to the use of dummy atoms in the MED-Portions we are using for hybridization. Thus, we are able to generate novel and diverse putative ligands even on a difficult target like a GPCR in the sense that the starting PDB material is poor.

We compared the similarity of the obtained hybrids to a list of 54 reference ligands of β2-adrenergic receptor kindly provided by Stefano Costanzi and Santiago Vilar Varela (Laboratory of Biological Modeling, NIDDK, NIH). For the comparison the terminal dummy atoms were removed in MED-Hybridise, and the remaining nonterminal dummy atoms were replaced with carbon atoms. The hybrids were then filtered to remove 2D (topological) duplicates, finally yielding 3818 unique compounds.

In the case of the GPCR, we performed an additional hybridization protocol with only the MED-Portions from interfamily hits. The obtained hybrids as well as those

obtained by using all the MED-Portions retrieved with MED-SuMo were compared by topological fingerprint similarity to a list of known β2-adrenergic receptor ligands (Figure 11). By using all the MED-Portions we obtained 132 molecules with a Tanimoto similarity >0.8. By using only the interfamily MED-Portions, we still obtained 55 molecules. As compared to a simple ligand similarity search to the initial cocrystallized carazolol, the only similar molecule we obtained having a Tanimoto similarity >0.8 was the carazolol itself.

These hybrids contained compounds similar to the initial 3 ligands (PDB codes 2rh1, 3d4s, 2vt4). We also obtained 11 other ligands from the list of 54 (CGP12177, ICI-118551, SR59230A, alprenolol, carvedilol, pindolol, NIP, bevantolol-S, nebivolol, timolol, bucindolol; Tanimoto similarity 0.9−1.0 for 7 ligands, 0.8−0.9 for 5 ligands). Some of these hybrids are depicted in 2D next to the similar ligand (Figure 12). This shows that the protocol can generate molecules similar to known active ligands. This is a significant retrospective validation as these GPCR ligands are not present in the PDB. The 3D superposition inside the binding site of these hybrids in comparison to the carazolol crystal position is shown in Figure 13.

Thus, the presented technology can obtain molecules more diverse than those present in the PDB, since only three GPCR ligands are present in the PDB. An example of hybridization is shown in Figure 14. The NIP-hybrid was generated by a hybridization of two MED-Portions originally from the 2rh1 (CAU) and 2vt4 (P32). This example shows the benefit to match ligands from the PDB on a specified chemical library to bring many more diversities into MED-Portions chemical moieties, including in this case some mono ring system from an initial multiple ring system.

The application of our complete protocol to the β2-adrenergic receptor 2rh1 structure, cocrystallized with an antagonist, carazolol, should more likely give us potential antagonists. To estimate the chances for finding agonists or antagonists with our protocol, we looked for the presence of the catechol moiety that is thought to interact with the serine residues in the fifth trans-membrane helix.[75] Any hybrids containing this moiety would more likely be agonists.

Only 48 hybrids, out of 5791, contained the catechol substructure, and none of the hybrids has the two hydroxyl groups available. We searched for hybrids that made H-bonds by means of a hydroxyl group with one of the serine residues, finding 42 hybrids out of 5791. Thus, using our protocol on an inactive GPCR structure gives hybrids that are more likely antagonists.

CONCLUSION

The exponential growth of the PDB (over 50,000 entries) is a clear incentive to explore innovative and efficient computational protocols in fragment-based drug design for use with experimental data. We have developed a complete automated protocol based on MED-Portion/MED-SuMo/MED-Hybridise technologies starting from a protein binding site query to superpose similar protein-fragments from the PDB and to combine those chemical moieties into new hybrid compounds. We have described here an original scheme to convert any protein−ligand structure to a set of protein-fragment patterns (MED-Portions) by matching substructure of small molecules from molecular libraries. Any source of protein−ligand complexes PDB files can be used as input, and therefore all efforts to curate the PDB would benefit the quality of the generated hybrids. Any source of small molecules can be used, and we chose PubChem as the source of molecules very likely to be synthetically accessible. We have defined a new structural object, MED-Portion, which consists of a chemical moiety, dummy atoms where the matching molecule was connected to the rest of the original ligand (topological connexions to the original ligand), and its associated local protein interaction surface. Combined with the MED-SuMo software to superimpose similar interaction surfaces, this allows various protein surfaces, either binding sites or full surfaces, to be populated with a pool of MED-Portions by target hopping at PDB scale. We characterized the results on 107 diverse protein surfaces and found that the protocol provides a pool of MED-Portions for diverse proteins and binding pockets, and the hydrophobicity of the surface is qualitatively correlated to the average hydrophobicity of the MED-Portions. This result is highly anticipated for a computational fragment-based protocol and emphasizes the relevance of the MED-Portion chemical moieties and the selectivity of the protocol. We were able to retrieve scaffolds of known actives for two targets: a GPCR and a protein kinase which are the two extremes in terms of propensity in the PDB. The cases studied are focused on growing a seed of the size of a fragment to druglike molecules that can be searched for in a compound database. The final results are real molecules that can be potentially purchased or synthesized.

ACKNOWLEDGMENT

We thank Stefano Costanzi and Santiago Vilar Varela for providing us useful comments and suggestions for our GPCR application as well as a list of β2-adrenergic receptor ligands (Laboratory of Biological Modeling, NIDDK, NIH). Kerstin Koch is a postdoctoral fellow from the Université Paris Sud supported by the POPS collaborative project (http://www.pops-systematic.org/) and funded by the French office "Direction Générale des Entreprises". She was involved in the application of the methods (Results section). All the software developments were implemented by and at MEDIT SA. This work was supported by the Carriocas collaborative project (http://www.carriocas.org/) and funded by the French office "Direction Générale des Entreprises".

Supporting Information Available: Statistics for the diverse set of 107 protein surfaces (Supplement Table 1: 107 pocket entries instead of 35). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–42.

(2) Allen, K. N.; Bellamacina, C. R.; Xiaochung, D.; Jeffery, C. J.; Mattos, C.; Petsko, G. A.; Ringe, D. An experimental approach to mapping the binding surfaces of crystalline proteins. *J. Phys. Chem.* **1996**, *100*, 2605–2611.

(3) Erlanson, D. A. Fragment-based ligand discovery meets phage display. *ACS Chem. Biol.* **2007**, *2*, 779–82.

(4) Jahnke, W. Perspectives of biomolecular NMR in drug discovery: the blessing and curse of versatility. *J. Biomol. NMR* **2007**, *39*, 87–90.

(5) Leach, A. R.; Hann, M. M.; Burrows, J. N.; Griffen, E. J. Fragment screening: an introduction. *Mol. Biosyst.* **2006**, *2*, 430–46.
(6) Congreve, M.; Murray, C. W.; Blundell, T. L. Structural biology and drug discovery. *Drug Discovery Today* **2005**, *10*, 895–907.
(7) Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. Discovering high-affinity ligands for proteins: SAR by NMR. *Science* **1996**, *274*, 1531–4.
(8) Ciulli, A.; Williams, G.; Smith, A. G.; Blundell, T. L.; Abell, C. Probing hot spots at protein-ligand binding sites: a fragment-based approach using biophysical methods. *J. Med. Chem.* **2006**, *49*, 4992–5000.
(9) Verlinde, C. L. M. J., Kim, H., Bernstein, B. E., Mande, S. C.; Hol, W. G. J. Antitrypanosomiasis drug development based on structures of glycolytic enzymes. In *Structure-Based Drug Design*; Veerapandian, P., Ed.; Marcel Dekker: New York, 1997; pp 365−394.
(10) Saxty, G.; Woodhead, S. J.; Berdini, V.; Davies, T. G.; Verdonk, M. L.; Wyatt, P. G.; Boyle, R. G.; Barford, D.; Downham, R.; Garrett, M. D.; Carr, R. A. Identification of inhibitors of protein kinase B using fragment-based lead discovery. *J. Med. Chem.* **2007**, *50*, 2293–6.
(11) Hajduk, P. J.; Greer, J. A decade of fragment-based drug design: strategic advances and lessons learned. *Nat. Rev. Drug Discovery* **2007**, *6*, 211–9.
(12) Kortvelyesi, T.; Dennis, S.; Silberstein, M.; Brown, L.; Vajda, S. Algorithms for computational solvent mapping of proteins. *Proteins* **2003**, *51*, 340–51.
(13) Cummings, M. D.; DesJarlais, R. L.; Gibbs, A. C.; Mohan, V.; Jaeger, E. P. Comparison of automated docking programs as virtual screening tools. *J. Med. Chem.* **2005**, *48*, 962–76.
(14) Bohm, H. J. A novel computational tool for automated structure-based drug design. *J. Mol. Recognit.* **1993**, *6*, 131–7.
(15) Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discovery* **2005**, *4*, 649–63.
(16) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849–57.
(17) Miranker, A.; Karplus, M. Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins* **1991**, *11*, 29–34.
(18) Evensen, E.; Joseph-McCarthy, D.; Weiss, G. A.; Schreiber, S. L.; Karplus, M. Ligand design by a combinatorial approach based on modeling and experiment: application to HLA-DR4. *J. Comput.-Aided Mol. Des* **2007**, *21*, 395–418.
(19) Deng, Z.; Chuaqui, C.; Singh, J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.* **2004**, *47*, 337–44.
(20) Pierce, A. C.; Rao, G.; Bemis, G. W. BREED: Generating novel inhibitors through hybridization of known ligands. Application to CDK2, p38, and HIV protease. *J. Med. Chem.* **2004**, *47*, 2768–75.
(21) Schmitt, S.; Kuhn, D.; Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **2002**, *323*, 387–406.
(22) Artymiuk, P. J.; Poirrette, A. R.; Grindley, H. M.; Rice, D. W.; Willett, P. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J. Mol. Biol.* **1994**, *243*, 327–44.
(23) Wallace, A. C.; Borkakoti, N.; Thornton, J. M. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.* **1997**, *6*, 2308–23.
(24) Jambon, M.; Imberty, A.; Deleage, G.; Geourjon, C. A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins* **2003**, *52*, 137–45.
(25) Barker, J. A.; Thornton, J. M. An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics* **2003**, *19*, 1644–9.
(26) Stark, A.; Sunyaev, S.; Russell, R. B. A model for statistical significance of local similarities in structure. *J. Mol. Biol.* **2003**, *326*, 1307–16.
(27) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. SiteEngines: recognition and comparison of binding sites and protein-protein interfaces. *Nucleic Acids Res.* **2005**, *33*, W337–41.
(28) Gold, N. D.; Jackson, R. M. Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. *J. Mol. Biol.* **2006**, *355*, 1112–24.
(29) Ramensky, V.; Sobol, A.; Zaitseva, N.; Rubinov, A.; Zosimov, V. A novel approach to local similarity of protein binding sites substantially improves computational drug design results. *Proteins* **2007**, *69*, 349–57.
(30) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP--retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–22.
(31) Jambon, M.; Andrieu, O.; Combet, C.; Deleage, G.; Delfaud, F.; Geourjon, C. The SuMo server: 3D search for protein functional sites. *Bioinformatics* **2005**, *21*, 3929–30.
(32) Doppelt, O.; Moriaud, F.; Bornot, A.; de Brevern, A. G. Functional annotation strategy for protein structures. *Bioinformation* **2007**, *1*, 357–9.
(33) Jambon, M. *A bioinformatic system for searching functional similarities in 3D structures of proteins*;Université Claude Bernard Lyon 1, 2003.
(34) Release of remediated PDB archive. ftp://ftp.wwpdb.org/ (accessed July 16, 2008).
(35) Guha, R.; Howard, M. T.; Hutchison, G. R.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J.; Willighagen, E. L. The Blue Obelisk-interoperability in chemical informatics. *J. Chem. Inf. Model* **2006**, *46*, 991–8.
(36) The Open Babel Package. http://openbabel.sourceforge.net/ (accessed July 5, 2008).
(37) CTFile Formats. www.mdl.com/downloads/public/ctfile/ctfile.pdf (accessed November 1, 2007).
(38) The PubChem Project. http://pubchem.ncbi.nlm.nih.gov (accessed June 9, 2008).
(39) SQLite. http://oandrieu.nerim.net/ocaml/mlsqlite/ (accessed October 26, 2008).
(40) Hasegawa, M.; Nishigaki, N.; Washio, Y.; Kano, K.; Harris, P. A.; Sato, H.; Mori, I.; West, R. I.; Shibahara, M.; Toyoda, H.; Wang, L.; Nolte, R. T.; Veal, J. M.; Cheung, M. Discovery of novel benzimidazoles as potent inhibitors of TIE-2 and VEGFR-2 tyrosine kinase receptors. *J. Med. Chem.* **2007**, *50*, 4453–70.
(41) Ghosh, A. K.; Devasamudram, T.; Hong, L.; DeZutter, C.; Xu, X.; Weerasena, V.; Koelsch, G.; Bilcer, G.; Tang, J. Structure-based design of cycloamide-urethane-derived novel inhibitors of human brain memapsin 2 (beta-secretase). *Bioorg. Med. Chem. Lett.* **2005**, *15*, 15–20.
(42) *Scitegic Pipeline Pilot, 7.0*; Accelrys Software Inc.: 10188 Telesis Court, S.S.D., CA 92121, U.S.A., 2004−2008.
(43) *MOE*; Chemical Computing Group Inc.: 1010 Sherbrooke St. W, S.M., Quebec, Canada H3A 2R7, 2003−2004.
(44) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
(45) Weininger, D. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.
(46) Finn, R. D.; Tate, J.; Mistry, J.; Coggill, P. C.; Sammut, S. J.; Hotz, H. R.; Ceric, G.; Forslund, K.; Eddy, S. R.; Sonnhammer, E. L.; Bateman, A. The Pfam protein families database. *Nucleic Acids Res.* **2008**, *36*, D281–8.
(47) *Discovery Studio, 2.0*; Accelrys Software Inc.: 10188 Telesis Court, S.S.D., CA 92121, U.S.A., 2005−2007.
(48) Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today* **2004**, *9*, 430–1.
(49) Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G.; Thian, F. S.; Kobilka, T. S.; Choi, H. J.; Kuhn, P.; Weis, W. I.; Kobilka, B. K.; Stevens, R. C. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science* **2007**, *318*, 1258–65.
(50) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–93.
(51) Daylight fingerprints. http://www.daylight.com/dayhtml/doc/theory/theory.finger.html (accessed November 1, 2007).
(52) Tanimoto, T. T. *IBM Internal Report 17th Nov.*; IBM: 1957.
(53) Gao, Z.; Li, H.; Zhang, H.; Liu, X.; Kang, L.; Luo, X.; Zhu, W.; Chen, K.; Wang, X.; Jiang, H. PDTD: a web-accessible protein database for drug target identification. *BMC Bioinf.* **2008**, *9*, 104.
(54) Laurie, A. T.; Jackson, R.M. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* **2005**, *21*, 1908–16.
(55) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A 'rule of three' for fragment-based lead discovery. *Drug Discovery Today* **2003**, *8*, 876–7.
(56) Taylor, S. S.; Radzio-Andzelm, E. Protein kinase inhibition: natural and synthetic variations on a theme. *Curr. Opin. Chem. Biol.* **1997**, *1*, 219–26.
(57) Bossemeyer, D. Protein kinases--structure and function. *FEBS Lett.* **1995**, *369*, 57–61.
(58) Levitzki, A. Protein kinase inhibitors as a therapeutic modality. *Acc. Chem. Res.* **2003**, *36*, 462–9.
(59) Fabbro, D.; Garcia-Echeverria, C. Targeting protein kinases in cancer therapy. *Curr. Opin. Drug Discovery Dev.* **2002**, *5*, 701–12.
(60) Nagar, B.; Bornmann, W. G.; Pellicena, P.; Schindler, T.; Veach, D. R.; Miller, W. T.; Clarkson, B.; Kuriyan, J. Crystal structures of the kinase domain of c-Abl in complex with the small molecule inhibitors PD173955 and imatinib (STI-571). *Cancer Res.* **2002**, *62*, 4236–43.
(61) Atwell, S.; Adams, J. M.; Badger, J.; Buchanan, M. D.; Feil, I. K.; Froning, K. J.; Gao, X.; Hendle, J.; Keegan, K.; Leon, B. C.; Muller-Dieckmann, H. J.; Nienaber, V. L.; Noland, B. W.; Post, K.;

Rajashankar, K. R.; Ramos, A.; Russell, M.; Burley, S. K.; Buchanan, S. G. A novel mode of Gleevec binding is revealed by the structure of spleen tyrosine kinase. *J. Biol. Chem.* **2004**, *279*, 55827–32.
(62) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789–801.
(63) Schroeder, G. M.; Chen, X. T.; Williams, D. K.; Nirschl, D. S.; Cai, Z. W.; Wei, D.; Tokarski, J. S.; An, Y.; Sack, J.; Chen, Z.; Huynh, T.; Vaccaro, W.; Poss, M.; Wautlet, B.; Gullo-Brown, J.; Kellar, K.; Manne, V.; Hunt, J. T.; Wong, T. W.; Lombardo, L. J.; Fargnoli, J.; Borzilleri, R. M. Identification of pyrrolo[2,1-f][1,2,4]triazine-based inhibitors of Met kinase. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 1945–51.
(64) Lagerstrom, M. C.; Schioth, H. B. Structural diversity of G protein-coupled receptors and significance for drug discovery. *Nat. Rev. Drug Discovery* **2008**, *7*, 339–57.
(65) Pierce, K. L.; Premont, R. T.; Lefkowitz, R. J. Seven-transmembrane receptors. *Nat. Rev. Mol. Cell Biol.* **2002**, *3*, 639–50.
(66) Lefkowitz, R. J.; Rajagopal, K.; Whalen, E. J. New roles for beta-arrestins in cell signaling: not just for seven-transmembrane receptors. *Mol. Cell* **2006**, *24*, 643–52.
(67) Bjarnadottir, T. K.; Gloriam, D. E.; Hellstrand, S. H.; Kristiansson, H.; Fredriksson, R.; Schioth, H. B. Comprehensive repertoire and phylogenetic analysis of the G protein-coupled receptors in human and mouse. *Genomics* **2006**, *88*, 263–73.
(68) Filmore, D. It's a GPCR world. *ACS Modern Drug Discovery* **2004**, *7*, 24–28.
(69) Palczewski, K.; Kumasaka, T.; Hori, T.; Behnke, C. A.; Motoshima, H.; Fox, B. A.; Le Trong, I.; Teller, D. C.; Okada, T.; Stenkamp, R. E.; Yamamoto, M.; Miyano, M. Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* **2000**, *289*, 739–45.
(70) Salom, D.; Lodowski, D. T.; Stenkamp, R. E.; Le Trong, I.; Golczak, M.; Jastrzebska, B.; Harris, T.; Ballesteros, J. A.; Palczewski, K. Crystal structure of a photoactivated deprotonated intermediate of rhodopsin. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 16123–8.
(71) Rasmussen, S. G.; Choi, H. J.; Rosenbaum, D. M.; Kobilka, T. S.; Thian, F. S.; Edwards, P. C.; Burghammer, M.; Ratnala, V. R.; Sanishvili, R.; Fischetti, R. F.; Schertler, G. F.; Weis, W. I.; Kobilka, B. K. Crystal structure of the human beta2 adrenergic G-protein-coupled receptor. *Nature* **2007**, *450*, 383–7.
(72) Hanson, M. A.; Cherezov, V.; Griffith, M. T.; Roth, C. B.; Jaakola, V. P.; Chien, E. Y.; Velasquez, J.; Kuhn, P.; Stevens, R. C. A specific cholesterol binding site is established by the 2.8 A structure of the human beta2-adrenergic receptor. *Structure* **2008**, *16*, 897–905.
(73) Warne, T.; Serrano-Vega, M. J.; Baker, J. G.; Moukhametzianov, R.; Edwards, P. C.; Henderson, R.; Leslie, A. G.; Tate, C. G.; Schertler, G. F. Structure of a beta1-adrenergic G-protein-coupled receptor. *Nature* **2008**, *454*, 486–91.
(74) Topiol, S.; Sabio, M. Use of the X-ray structure of the Beta2-adrenergic receptor for drug discovery. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 1598–602.
(75) Rosenbaum, D. M.; Cherezov, V.; Hanson, M. A.; Rasmussen, S. G.; Thian, F. S.; Kobilka, T. S.; Choi, H. J.; Yao, X. J.; Weis, W. I.; Stevens, R. C.; Kobilka, B. K. GPCR engineering yields high-resolution structural insights into beta2-adrenergic receptor function. *Science* **2007**, *318*, 1266–73.

CI8003094