MED-SuMoLig: A New Ligand-Based Screening Tool for Efficient Scaffold Hopping

Olivier Sperandio,[†,‡] Olivier Andrieu,[‡] Maria A. Miteva,[†] Minh-Quang Vo,[‡] Marc Souaille,[‡] François Delfaud,[‡] and Bruno O. Villoutreix*[,†]

INSERM U648, University Paris V, 45 rue des Sts peres, 75006 Paris, France, and MEDIT SA, 2 rue du Belvedere, 91120 Palaiseau, France

Received January 26, 2007

The identification of small molecules with selective bioactivity, whether intended as potential therapeutics or as tools for experimental research, is central to progress in medicine and in the life sciences. To facilitate such study, we have developed a ligand-based program well-suited for effective screening of large compound collections. This package, MED-SuMoLig, combines a SMARTS-driven substructure search aiming at 3D pharmacophore profiling and computation of the local atomic density of the compared molecules. The screening utility was then investigated using 52 diverse active molecules (against CDK2, Factor Xa, HIV-1 protease, neuraminidase, ribonuclease A, and thymidine kinase) merged to a library of about 40 000 putative inactive (druglike) compounds. In all cases, the program recovered more than half of the actives in the top 3% of the screened library. We also compared the performance of MED-SuMoLig with that of ChemMine or of ROCS and found that MED-SuMoLig outperformed both methods for CDK2 and Factor Xa in terms of enrichment rates or performed equally well for the other targets.

INTRODUCTION

Virtual screening (VS) methods are becoming an important aspect of drug discovery, and several success stories have already been reported.[1−3] VS and experimental high-throughput screening (HTS) are commonly used in modern drug discovery campaigns and are indeed complementary techniques, although several examples can be given where VS outperformed HTS.[4] Two main computer-based approaches can be distinguished: structure-based (SBVS) and ligand-based (LBVS) virtual screening methods. The former relies on docking small molecules in the 3D structure of a biological receptor and quantifying the resulting interactions. The latter is based on the assumption that structurally similar compounds are likely to exhibit similar biological activities. Both in silico screening approaches aim to enrich a list of molecules with putative potent compounds. However, rather than opposing these two major computer techniques, recent studies have emphasized the benefit of combining them.[5,6] The numerous methods used in LBVS focus either on 2D/3D structure−activity relationships[7−10] or on molecular similarity searches.[11] Molecular similarity methods encompass 2D-similarity-search, shape-based, and pharmacophore-based engines. The latter has been extensively explored, and various studies can be cited to comfort the pertinence of such a technique.[12−14] One important challenge for ligand-based methods is to find the appropriate balance between search specificity and search flexibility, the former being linked to the receptor selectivity and the latter to the legitimate wish for "*scaffold hopping*" (i.e., alleviating any ligand scaffold dependency). Both shape-based and pharmacophore-based methods use 3D structures as input. Even though the algorithms behind such tools can be different in essence, more and more programs tend to combine the two concepts. Some methods such as rapid overlay of chemical structures (ROCS)[15] use primarily the molecular shape as a criterion for molecular similarity but perform better when the molecular overlay is biased by the chemical nature of the molecules in play.[16] Other methods such as Catalyst[17] which are driven by pharmacophore matching can use exclusion volumes to emulate the bulky presence of the receptor and therefore use the notion of shape stringency. Finally, new methods are in essence founded on both shape and chemistry. They are based on the so-called *shape signature* that projects both the topological profile (shape-based) of the query ligand and the chemical nature of its substructures into low-dimensional descriptors, thereby allowing for very fast and accurate pairwise comparisons.[18]

The program MED-SuMo[19,20] was originally developed to detect structural similarities between proteins, and more specifically protein active sites. This package can be used for different purposes: characterizing an unknown protein binding site or investigating active-site selectivity within a protein family. The approach of MED-SuMo relies primarily on a physicochemical description of molecules such as hydrogen-bond donor, hydrogen-bond acceptor, aromatic group, hydrophobic group, and so forth but also considers the notion of molecular shape through an evaluation of local atomic densities of the compared proteins. On the basis of what we observed with MED-SuMo on protein active-site comparisons, we decided to enhance MED-SuMo, such as using its powerful algorithm to detect similarities between ligands and to superimpose them. Several other reasons legitimated the release of this program. First, MED-SuMo is very fast. This becomes a great advantage when screening several thousands of molecules. Second, because its algorithm relies on the detection of physicochemical properties with a local shape-matching evaluation, it is a prirori well-

* Corresponding author e-mail: bruno.villoutreix@univ-paris5.fr.
[†] University Paris
[‡] MEDIT SA.

10.1021/ci700031v CCC: $37.00
Published on Web 05/04/2007

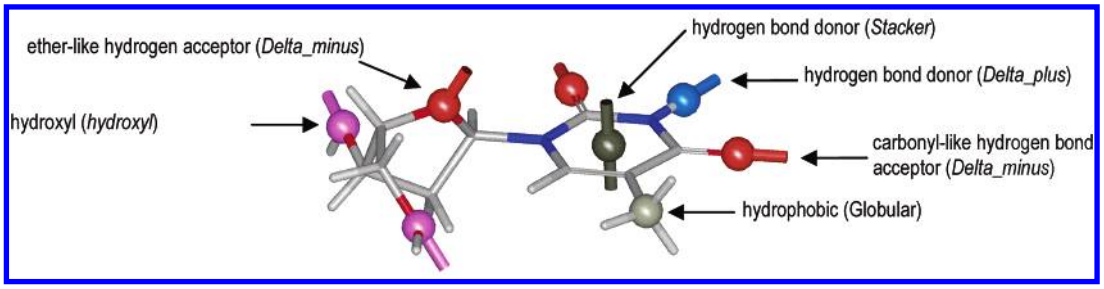

Figure 1. Example of MED-SuMo object mapping for small molecules. The molecule is represented in terms of physicochemical properties (aromatic, hydrogen bond donor, etc.). Those properties are transposed through the constructors into MED-SuMo objects that correspond to their position, orientation, and symmetry. These data for a small molecule were indeed generated with MED-SuMoLig since the original MED-SuMo could not treat effectively such compounds.

suited for scaffold hopping and thus for enriching a list of compounds with diversified potent ligands. Third, it only detects shared pharmacophore features without penalizing the discrepancies. This favors the identification of specific chemical functions with no attempt to force global alignments and therefore allows chemical diversity.

Our new program, MED-SuMoLig, targets the specific situation where one or several ligands are known as actives on a given receptor (with the availability of cocrystals) and are meant to be individually used as pharmacophoric hypotheses in a virtual screening campaign. Indeed, in some cases, it might be more interesting to combine the enriched lists resulting from molecular similarity runs on each of the actives[6] rather than building a consensual hypothesis. This becomes particularly true when the available actives turn out to have different binding modes or when they cannot be rationally aligned.

Any pharmacophoric-based method first needs to identify the physicochemical properties of the treated molecules. The previous version of MED-SuMo was based on a lexicographic analysis of the structure file, which listed possible atom names (i.e., essentially amino acid atom names). This solution was well-suited for protein coordinate files but is not applicable for small molecules databases. To circumvent the chemical diversity issue and in order to make our tool totally flexible, we have implemented the SMARTS[21] language (SMiles ARbitrary Target Specification). This pattern-matching method is very efficient and allows one to handle chemical wild cards instead of explicit atom names and types, thereby facilitating fast identification of chemical functions and substructures. The transposition of the physicochemical properties into MED-SuMo objects is now represented as SMARTS expressions. Beside the implementation of the SMARTS language itself, several subroutines had to be written, for example, methods to correctly type the atoms, to transpose the molecule connectivity, or to detect aromatic rings.

In order to evaluate the performance of MED-SuMoLig and test if it would be suited for large virtual screening experiments, we created a validation set composed of 52 molecules active against six different proteins: cyclin-dependent kinase 2 (CDK2), coagulation factor Xa (FXa), HIV-1 protease (HIV-1p), neuraminidase (NA), ribonuclease A (RNase), and thymidine kinase (TK). We merged these active compounds with the diversity set of the ChemBridge database [37 907 molecules after absorption, distribution, metabolism, excretion, and toxicity (ADMET) filtering] to monitor both the quality of the superimposition between coactives and the enrichment profiles on a given protein. Because MED-SuMoLig only superimposes rigid molecules, a multiconformational version of the database was constructed with Omega 2.0.[22] We show that in all cases our program recovered more than half of the actives in the top 3% of the screened library. We also compared, on the same validation set, the performance of our tool to those of ChemMine[23] and ROCS,[15] which are respectively 2D similarity-based and shape-based programs. MED-SuMoLig either outperformed both ChemMine and ROCS in terms of enrichment rates or performed similarly.

METHODS

MED-SuMo Algorithm. In the present study, we significantly modified the detection routine of MED-SuMo objects in order to handle small molecules (see next section), whereas the core of the comparison algorithm remains the same (see ref 24 for a detailed description of the method). Thus, prior to reporting the introduced changes, we will briefly recall the key concepts of the MED-SuMo algorithm and the changes implemented to make the package able to handle small organic molecules.

The MED-SuMo approach was originally designed to deal with amino acid residues. Each residue of the protein is mapped to discrete MED-SuMo objects (hydrogen-bond donor, aromatic, acyl group, etc.) with geometrical constructors (globular symmetry, simple polarization, symmetric polarization, etc.). These constructors are intimately associated with both the nature and the symmetry of the substructure to which they correspond (Figure 1). Among the various MED-SuMo objects available for proteins, we focused on those having a major signification when dealing with the notion of pharmacophores. Five objects and thus five constructors were kept, but modified from the original MED-SuMo, to treat small organic molecules. The constructor that generates an aromatic MED-SuMo object is defined as a symmetric polarization called a *stacker*. This *stacker* characterizes the anisotropic cones pointing out on each side of an aromatic ring. The constructor *Point* (globular-isotropic object) is used to define the hydrophobic MED-SuMo object due to the spherical symmetry of this physicochemical property. Finally, three different constructors (simple polarization vector) generate, respectively, hydrogen-bond donors (*Delta_plus*), hydrogen-bond acceptors (*Delta_minus*), and hydroxyl groups (*hydroxyl*). These constructors reflect the specific symmetry of such chemical functions. *Delta_minus* has its direction and origin defined by the nature of its chemistry: ether-like acceptor or carbonyl-like acceptor.

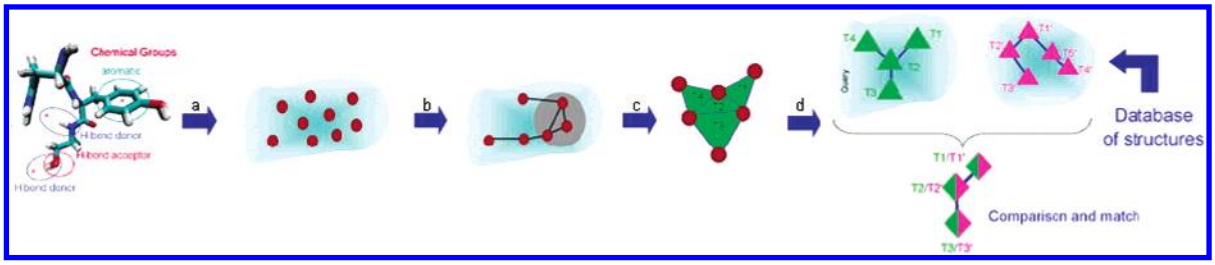

Figure 2. MED-SuMo comparison procedure. (a) The molecular structure is scanned and mapped to a dictionary of physicochemical/geometric constructors (i.e., Stacker, Point, Delta_minus, etc.) that detect the MED-SuMo objects (red circles). (b) The geometric rules are applied to build a triangle network. (c) This network is transposed into a graph data structure (green triangle). (d) The resulting graph is compared to a similarly constructed graph from a precompiled database of structures.

Table 1. Med-SuMo Objects, Substructures, and SMARTS Codes

MED-SuMo object	substructure	SMARTS code
H-bond donor (Delta_plus)	nitrogen proton donor	[#7][$([#1]−*)]
H-bond acceptor (Delta_minus)	ether-like hydrogen-bond acceptor	[$([$([#7D2;!H]),$([#8D2;!H]),$([#16D2;!H])])](∼[*])
H-bond acceptor (Delta_minus)	carbonyl-like hydrogen-bond acceptor	[#6,#7,#16]∼[OD1]
aromatic (Stacker)	six-ring aromatic	[a]1[a][a][a][a][a]1
hydrophobic (Point)	cyclohexyl	[Cv4][Cv4;H1;R]1[Cv4;H2;R][Cv4;H2;R][Cv4;H2;R][Cv4;H2;R][Cv4;H2;R]1
hydrophobic (Point)	*sec*-butyl	[Cv4;H1;!R](−[CH2;!R][CH3;!R])[CH3;!R]

A set of parameters was associated with every MED-SuMo object. These parameters were divided into two categories: those defining the construction of the object itself (position and orientation) and those defining the extent to which a superimposition of this object will be validated (e.g., angle tolerance upon object superimposition).

The MED-SuMo architecture therefore finds its foundations in a physicochemical description of the protein structure that is independent from the notion of protein sequence. The MED-SuMo objects (Figure 2) are assembled into triangles (one object per triangle summit) through a set of specific rules: geometric rules and density rules. These rules address the combinatorial explosion issue that stands behind the generation of such a triangle network. The geometric rules specify, for example, acceptable ranges between triangle summits and acceptable angles between a triangle's edges. The density rules prevent overlapping triangles by precluding autoencapsulation of several triangles.

This triangle network is stored as a graph data structure, with the triangles as vertices and with edges connecting adjacent triangles. The core of the comparison algorithm operates on this graph data structure. To compare two graphs, MED-SuMo first looks for similar (compatible) triangles then assembles nearby triangles in a patch. The query graph is matched with each graph of the database using a pairwise comparison. A comparison graph is built on the basis of the compatibility (isomorphism) between the query graph and the database graphs, that is, in fine, compatible triangles of concordant MED-SuMo objects. The connected components of the comparison graph determine which parts of the molecules are to be matched. The local density of the atoms neighboring the triangles is taken into account along the process of pairwise triangle comparison. This permits obtaining local matches with the query rather than averaged alignments that are more approximate. The final score is simply given by the size of the final patch of triangles found in common (*main scoring function*).

Comparisons between proteins are made by comparing graphs of MED-SuMo object triangles through a heuristic algorithm which detects patches of local common subgraphs. Consequentially, this makes every triangle of the query molecule a "sub-query" that will seek for its own local match. MED-SuMo runs typically consist of submitting a protein query to a database of precompiled MED-SuMo graphs and obtaining a list of matching protein structures superimposed onto the query protein.

The MED-SuMoLig Approach. MED-SuMoLig was engineered such that not only proteins but also small molecules could be compared. Several C++ subroutines were added to the core of MED-SuMo written in Caml in order to modify the procedures that scan molecules and to define and detect the MED-SuMoLig objects. MED-SuMoLig reads SD files as input and parses atomic coordinates and connectivity. The developed subroutines then use this information to map a list of SMARTS expressions which corresponds to each MED-SuMoLig object to be created. The list of MED-SuMoLig objects was tuned toward a more pharmacophoric profiling of the molecules compared to the initial version. In order to accommodate the changes in the MED-SuMo objects, certain parameters were modified as well. A *space occupancy threshold* of 0.3, a *shape threshold* of 0.75, and a *relative deformation* of 0.15 (see ref 23 for details) were used during the validation of the program. These three parameters are among the ones that have the largest influence on the comparison procedure because they respectively drive the way the triangles themselves are built (*space occupancy threshold*) and compared (*shape threshold* and *relative deformation*). The afore-mentioned values showed the best results for both the enrichment assays and the quality of the superimposition. The results obtained from MED-SuMoLig consist of a list of hit molecules that can be easily superimposed onto the query molecule and sorted through a proprietary graphical interface.

MED-SuMoLig Object Mapping. A major change has been made concerning the routine that detects the MED-SuMo objects. In the original version, this detection was driven by a dictionary (lexicographic analysis), mapping atom names to the above-mentioned object constructors. This

supposed a foreknowledge of every single atom name in the molecule, which is the case for protein (e.g., CA, NE1, ND2, etc.). This could not be anticipated for ligands because of the size of the chemical space ($>10^{60}$ molecules).[25] A more flexible routine was therefore implemented in order to address this issue. The SMARTS[21] language was fully incorporated into the MED-SuMo package. SMARTS is derived from SMILES (Simplified Molecular Input Line Entry System) and allows one to manipulate chemical substructures with the use of wild cards (e.g., aromatic atoms or any heavy atom bound to a hydrogen) instead of an explicit atom type. A list of SMARTS expressions was designed such that each MED-SuMoLig object was associated with one or several SMARTS codes. Some examples of these SMARTS expressions are shown in Table 1. Such an implementation required the development of several subroutines besides the SMARTS language itself. First, MED-SuMo had to be modified such that an MDL SDF rather than a PDB file could be manipulated. This format is obviously more suited for small molecules when considering their physicochemical properties, for example, hydrogen-bond donors/acceptors or aromaticity, which depend on the protonation states and the bond orders.

Moreover, we have developed our own subroutine to detect ring aromaticity. The subroutine first flags the rings that are only potentially aromatic, called *Huckelable* rings, that is, precluding rings having, for instance, sp^3 carbons, or sulfur having exocyclic bonds. Then, an iterative procedure detects clusters of *Huckelable* rings and applies the Huckel rule ($4n + 2\ \pi$ electrons) on each of them. If the whole cluster of rings matches the Huckel rule, it then is declared as aromatic and all the rings in this cluster are flagged as such. If the number of counted electrons does not match the $4n + 2$ rule, then all possible subclusters within this cluster are tested recursively until determining the ring aromaticity character of every ring in the cluster. The Huckel electron count used is very similar to the one included in the Marvin suite:[26] an sp^2 carbon with endocyclic double bonds counts for one π electron; oxygen and divalent sulfur within the ring count for two π electrons; divalent nitrogen counts for one π electron; trivalent nitrogen counts for two π electrons, and an sp^2 carbon having an exocyclic double bond with either a terminal oxygen or a terminal sulfur counts for zero but does not preclude the global electron count on the ring and/or the cluster.

The list of SMARTS expressions has been generated in order to cover as much as possible the "druglike" chemical space. One SMARTS chain was created to define the hydrogen-bond donor as any nitrogen carrying a hydrogen atom. Four codes define hydrogen-bond acceptors (ether-like, carbonyl-like, nitrile, and thioketone), one code to detect hydroxyl functions, and two codes for aromatic rings with five or six atoms. Finally, about 25 codes define hydrophobic functions such as propyl, isopropyl, pentyl, cyclohexyl, etc.

XML Descriptor File. The list of MED-SuMoLig object definitions was incorporated into an XML (*extensible markup language*) descriptor file. The file was primarily organized by physicochemical property (*point_phar*) and, as a second classification level, by constructor type (*constructor*), which correlates to the geometry of the property. A field *smarts_chain* was used to define the SMARTS code and the number of atom IDs expected as a result of the SMARTS query. The following fields use the resulting list of atom IDs as arguments. Depending on the geometry of the constructor, the field *geometry* is tuned toward a point (globular symmetry), a vector (unidirectional symmetry), or a double vector (bidirectional symmetry). When it is required by MED-SuMoLig, a field *option* is also present to set the construction and comparison parameters. The XML parameter file also serves to describe substructures that can have multiple MED-SuMoLig objects such as hydrophobic substructures (Figure 3).

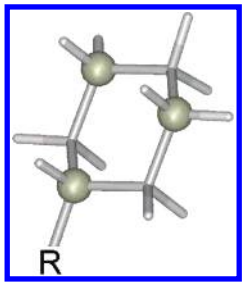

Figure 3. Example of multiple objects matching. The cyclohexyl group is characterized by three hydrophobic objects (gray spheres) placed on the ring that were derived from one unique SMARTS expression. [Cv4][Cv4;H1;R]1[Cv4;H2;R][Cv4;H2;R][Cv4;H2;R]-[Cv4;H2;R][Cv4;H2;R]1.

The XML format allows a total parametrization of the MED-SuMoLig objects and therefore makes the program easily customizable regarding either the very existence of an object or the parameters used for its detection and comparison. Figure 4 presents the outline of two MED-SuMoLig objects from the XML parameter file. The mapping of the object is applied to the whole surface of the ligands present in the database and can be visualized in a proprietary interface. This interface allows checking of the validity of the query and launching of the MED-SuMoLig runs. The interface is also used to visualize the results in a clickable chart displaying the hits ranked by the SuMo *main scoring function* and more importantly to visualize the superimposition of the hits onto the query ligand.

The global procedure, from detecting the MED-SuMoLig objects to the comparison routine, is shown in Figure 5. This illustrates the changes undertaken with respect to the previous version (for proteins).

Database Generation. As for the previous version of MED-SuMo, MED-SuMoLig needs to precompile a database which stores all the graphs of triangles and information about the atomic density around those triangles for every molecule to be screened. Even though the disk space required can be important (30 GB for 1 150 000 molecules) and the generation of the database takes 2−3 days, it has to be done only once. The screening itself is very rapid (50 min for 1 150 000 molecules on a standard Linux workstation).

This whole study has been carried out using the diversity set (50 000 compounds) from the ChemBridge database (http://chembridge.com/chembridge) filtered for ADMET properties.[27] The remaining collection contained 37 907 different molecules with molecular weights between 200 and 900 Da, computed logP values between −5.0 and 6.0, polar surface area between 0.0 and 160, a maximum number of rotatable bonds of 20, a maximum number of hydrogen-bond donors of 8 and a minimum of 0, a maximum number of hydrogen-bond acceptors of 12 and a minimum of 0, and at least two heteroatoms per compound. The validation set

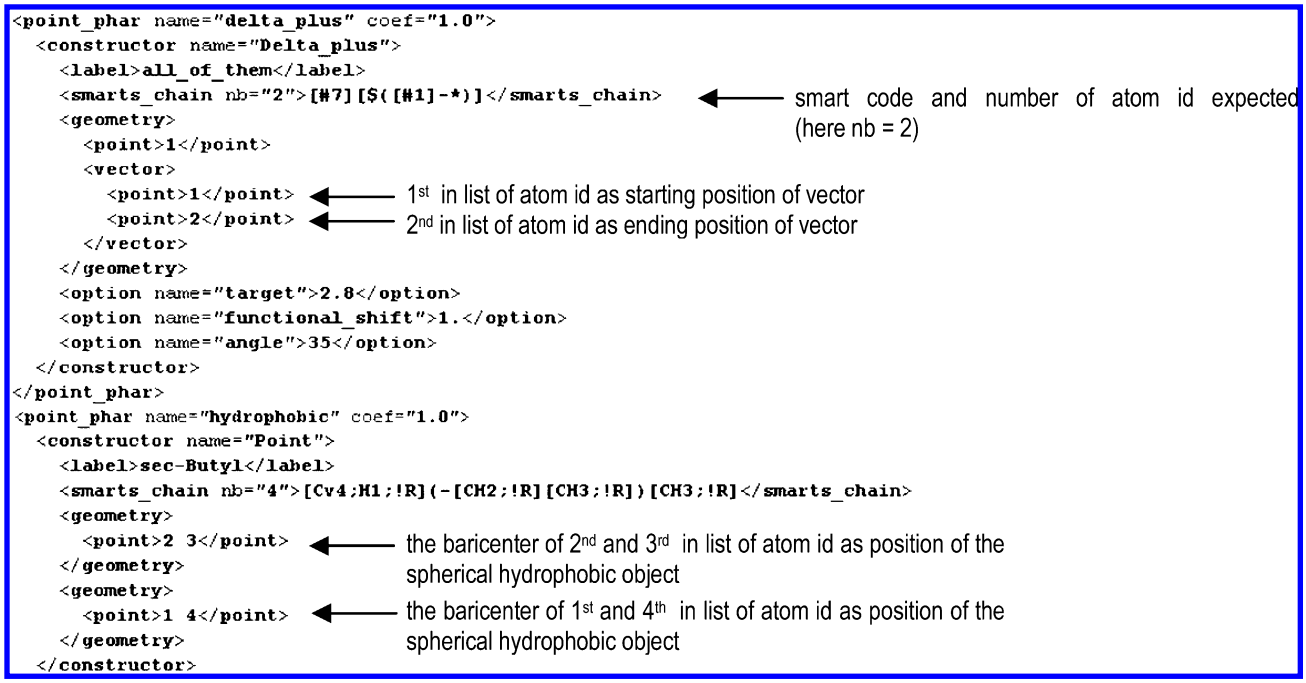

Figure 4. XML parameter file of MED-SuMo.

(Supporting Information) used for enrichment assays was composed of inhibitors against six protein families with a total of 52 inhibitors: 10 for CDK2, nine for FXa, five for HIV-1p, 10 for NA, eight for RNase, and 10 for TK.

The multiconformational database was generated from the SMILES strings[28] with Omega 2.0[22] with an energy window of 25 kcal mol^{-1}, a diversity threshold (root-mean-square deviation; RMSD) of 1 Å, and a maximum of 100 conformers per molecule. These parameters represented an appropriate balance between speed and accuracy and agreed well with the recent study reported by Kirchmair et al.[29] The resulting database contained 1 150 000 structures (an average of 30 conformers per molecule).

MED-SuMoLig Runs. The MED-SuMoLig runs were performed on each of the 52 active compounds of the validation set using their bioactive conformation (whole ligand) as queries. The validation procedure based on these runs first consisted of monitoring the pertinence of the comparison engine of MED-SuMoLig on small molecules and its ability to correctly detect and align hit molecules on the query molecule. We defined as *experimental alignment* the experimentally derived superimposition of the cocrystalized ligands by superimposing the protein structures from which they were extracted.

The second major validation assessment performed from the runs was correlated to the notion of enrichment in coactive compounds and therefore to the notion of scoring. A modified version of the ranking criterion was used for the enrichment rate calculations. Besides the MED-SuMo main scoring function that simply accounts for the size of the patch of triangles found in common with the query molecule, three other functions are also computed during the screening: RMSD between the objects found in common, *deformation*, and *penalty*. The parameter *deformation* is a global evaluation of the deformation of distances between elementary objects and the changes in their orientations by taking into account their possible symmetry. The parameter *penalty* monitors the difference of accessibility of the objects present in a triangle. All three parameters are meant to be as low as possible in order to obtain a good match with the query. Therefore, the following combination of these scores has been implemented. The main scoring function (see previous paragraph) was used as a first scoring criterion. The ranking was then hashed into scoring bins. The size of each bin was 0.2 of the main score. Within these bins, a reranking was made on the basis of the sum of the three above-mentioned functions (*RMSD*, *deformation*, and *penalty*) with lower values at the top of each scoring bin. This ensured that highly populated bins were ranked according to an accurate criterion besides the size of the patch of triangles.

ChemMine Study. The 2D similarity search was performed with the program ChemMine. This program implements an improved two-dimensional fragment-based similarity search based on two user-defined structural descriptors (atom pairs and atom sequences) and Tanimoto coefficients[30] as a similarity measure. The combination of atom sequence and Tanimoto was used for enrichment rate calculations.

ROCS Study. The 3D shape-based search has been performed with the program ROCS 2.2. This program evaluates the superimposability of two molecules by quantifying the maximum overlap of their shape. The volume of the molecules is approximated by Gaussian-based functions which are rapid to calculate and compare.

The runs have been launched on the same multiconformational database (1 150 000 structures). As for the MED-SuMoLig runs, the query inhibitor was always the active conformation taken from the crystal structure with no minimization.

The ROCS chemical force-field option (*optchem*) that takes into account the chemical nature of the molecule to bias the molecular overlay has been used because it did provide significant improvements for enrichment factor (EF) evaluations in the present study. It was combined with the *combo* ranking rather than the usual *shape tanimoto* alone. The *combo* ranking is just the sum of the *shape tanimoto* and the *scaled color* score. The *scaled color* score evaluates

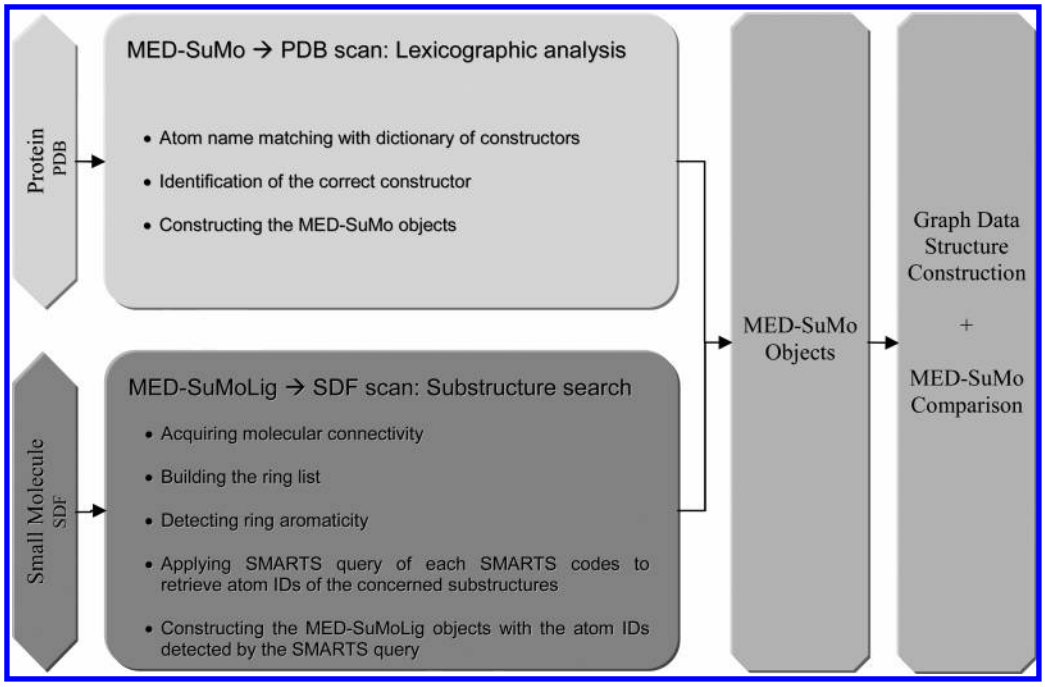

Figure 5. Diagram of MED-SuMo object construction for MED-SuMo and MED-SuMoLig. The following steps (graph data structure construction and graph comparison) remain identical except for the comparison parameters that have been tuned differently.

the chemical match between the two molecules and is between 0 and 1 as for the *shape tanimoto*.

Enrichment Tests. The enrichment study has been carried out using the following definition for the enrichment factor: where $Hits_{sel}$ is the number of target-specific molecules (true

$$\text{Enrichment Factor (EF)} = \frac{Hits_{sel}}{Hits_{tot}} \times \frac{NC_{tot}}{NC}$$

positives) selected by the screening algorithm at a specific percentage level of subsetting, $Hits_{tot}$ is the total number of target-specific molecules for the target in question in the database, NC_{tot} is the total number of molecules screened in the database, and NC is the total number of compounds within the chosen percentage level of subsetting.

The database was composed of a total of 37 959 compounds (NC_{tot} = 37 959). The enrichment factors were evaluated at 10, 5, 3, and 1% levels of subsetting which makes NC equal to 3795, 1898, 1139, and 380, respectively. The maximal EF value for each subsetting level is 10 for 10%, 20 for 5%, 33 for 3%, and 100 for 1%.

RESULTS AND DISCUSSION

Screening experimentally all the molecules present in a large compound collection is no longer viable. As such, many research groups attempt to develop or optimize virtual screening technologies to facilitate the drug discovery process. It is generally accepted that similar compounds can have similar activities and similar binding modes. However, similarity can be measured in a variety of ways.[4] We developed a new ligand-based VS tool called MED-SuMoLig and evaluated its performance on several bioactive ligand queries. The query ligands were systematically taken from their crystallographic structure in their bioactive conformation as done by Chen et al.[31] We first addressed the notion of ligand alignment to check how well a ligand detected as a hit was superimposed onto the query ligand. This aspect is essential in a drug design project particularly when working on lead optimization. Second, we tested the ability of MED-SuMoLig to detect active ligands among a collection of compounds seeded with diverse bioactive molecules. The goal here was to test whether or not our program would be suited for a screening campaign. Finally, we compared the enrichment capabilities of MED-SuMoLig to two different programs: ChemMine and ROCS 2.2. The choice of these two programs resides in the following conceptual observations: ROCS is primarily based on a shape-based comparison of the evaluated molecules but can take into account their chemistry as well to bias both the superimposition (*optchem*) and the ranking (*combo*), while ChemMine is a 2D-similarity search method. By incorporating these two programs in the MED-SuMoLig evaluation, we aimed at verifying whether its algorithm (based on both chemistry and shape) was capable of competing with a tool purely based on chemistry (ChemMine) and a tool predominantly based on shape (ROCS).

Superimposability. We ran the program on each of the 52 actives of the validation set and compared the MED-SuMoLig alignment with the *experimental alignments* as defined in the Methods section. The cocrystalized ligand conformations for a given protein target were therefore superimposed and used as a positive control. It is important, when considering ligand alignment with 3D ligand-based tools, to rationalize the orientation of the ligand with respect to its protein target. The vicinity of the protein residues determines in part the ligand binding mode, so any comparison with *experimental alignment* has to be undertaken with an appropriate consideration for the specific interactions between the protein residues and the different chemical groups of the ligand. Therefore, in the following paragraphs, several references will be made to the nature of the protein interactions with the corresponding active ligands.

In Figure 6 are represented for the six protein targets the *experimental alignments* (panels A1, B1, C1, D1, E1, and

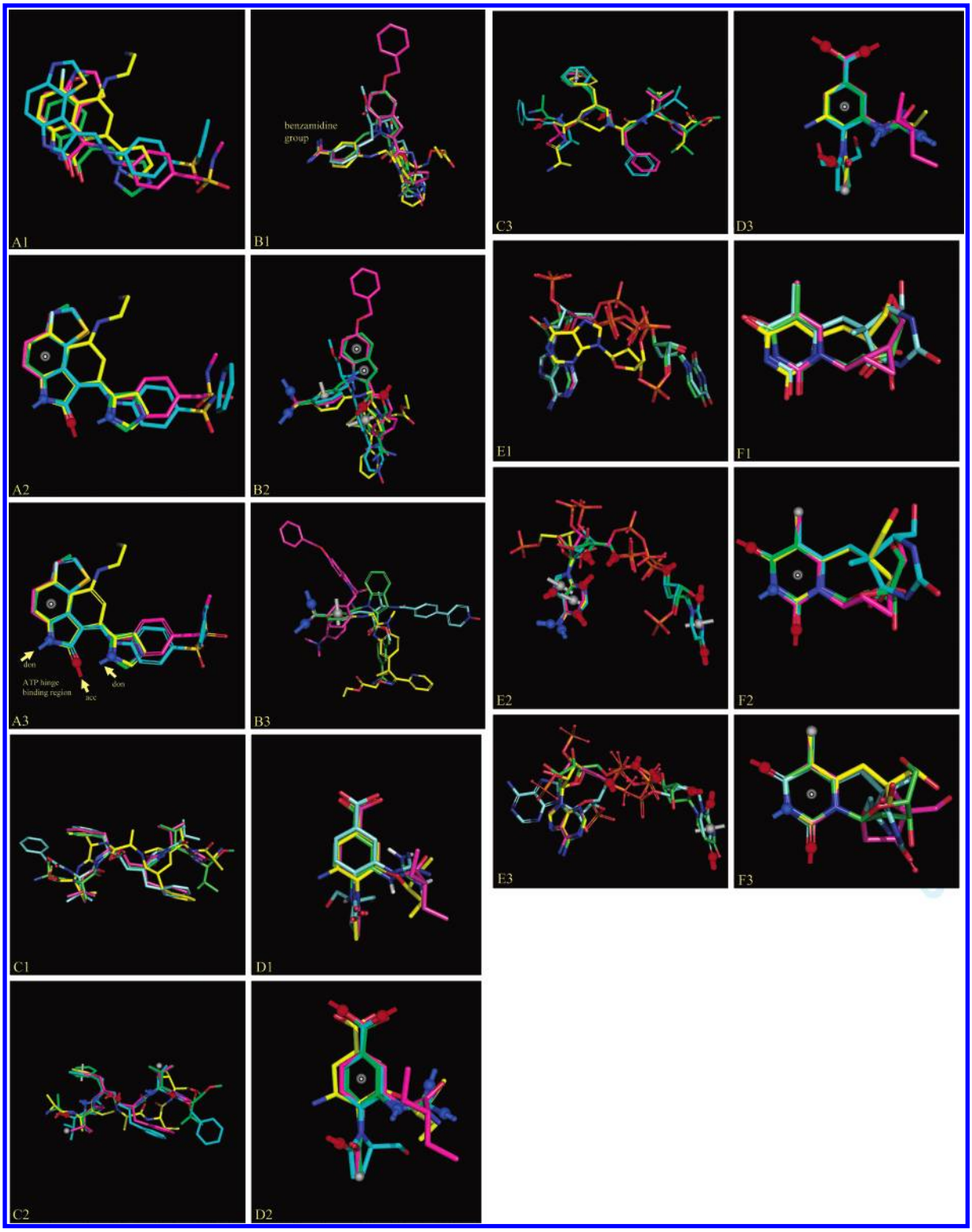

Figure 6. Evaluation of MED-SuMoLig superimposition. Panels A, B, C, D, E, amd F refer to the six proteins of the validation set. Within each panel, 1 shows the experimental alignment (by the matching of the proteins), 2 shows the MED-SuMoLig alignment using the crystallographic conformations of the ligands, and 3 shows the MED-SuMoLig alignment using the multiconformer library (de novo conformations). Each of these panels shows the superimposition of the three first MED-SuMoLig hits on the query ligand (green): first hit (cyan carbon), second hit (pink carbon), and third hit (yellow carbon). Thus, only the green ligand has exactly the same conformation on the three panels for each protein because it is the X-ray ligand conformation. Panel A's (CDK2): 1PF8 (query) on which are superimposed 1FVV (first), 1KE6 (second), and 1P2A (third). The hinge binding region of the CDK2 actives is composed of two hydrogen-bond donors (don) and one hydrogen-bond acceptor (acc). Panel B's (FXa): 1LPK (query) on which are superimposed 1KSN (first), 1LPG (second), and 1G2L (third). Panel C's (HIV-1p): 1AAQ (query) on which are superimposed 1HPS (first), 1SBG (second), and 1HBV (third). Panel D's (NA): 1INF (query) on which are superimposed 1B9T (first), 1B9S (second), and 2QWK (third). Panel E's (RNase): 1JN4 (query) on which are superimposed 1QHC (first), 1AFK (second), and 1O0O (third). Panel F's (TK): 1KIM (query) on which are superimposed 1E2N (first), 1E2K (second), and 1E2P (third). Also represented are the MED-SuMoLig objects detected in common with the query for only one of the three first hits (for clarity reasons): blue objects are hydrogen-bond donors, red objects are hydrogen-bond acceptors, gray spheres are hydrophobic objects, and gray Stackers are the aromatic objects.

F1), the MED-SuMoLig alignments using the crystallographic conformations of the ligands (panels A2, B2, C2, D2, E2, and F2), and finally the MED-SuMoLig alignments using the multiconformer database (panels A3, B3, C3, D3, E3, and F3). The query ligand chosen for the figure was the one having the best enrichment results. For each protein, the first three hits (cyan carbon, pink carbon, and yellow carbon) are superimposed onto the query ligand (green carbon). The MED-SuMoLig objects detected in common with the ligand query are also represented only for the first hit (cyan carbon) with the exception of FXa (second hit) for which the second hit shows a better contrast between the crystallographic MED-SuMoLig alignment (panel B2) and the de novo MED-SuMoLig alignment (panel B3). In the following, ligands are identified by the PDB code from which they were extracted. To differentiate them from the protein structure they will be written in *italics*.

The MED-SuMoLig alignments were consistent with the *experimental alignment* 70% of the time. When the alignment was only partial, like for some of the FXa actives, the key substructures were identified and used to drive the superimposition. Nevertheless, we observed two kinds of situations for which our program might provide nonoptimal alignments. First, we noticed that in the presence of inaccurate conformers (i.e., the structure was too different from the bioactive conformation) MED-SuMoLig might not always be able to fill the conformational space gap despite its built-in tolerance routine. If the conformational inaccuracy triggers the loss of one or several objects in the global alignment, this can obviously impede the correct alignment of the hit molecule onto the query. This can be observed in Figure 6, especially in the case of FXa and RNase ligands (see the differences between panels B2 and B3 on the one hand and panels E2 and E3 on the other hand). In the case of FXa, the lack of conformational accuracy triggers a misalignment of the indole group which is correctly aligned when using the crystallographic conformations. In the case of RNase, the de novo conformation did not allow recovery of the adenine-based part of the ligand query. Second, in the case of a molecule having one or several axes of symmetry or when symmetry can be found on the patch of the objects detected by MED-SuMoLig, the atomic density evaluated during the comparison procedure can turn out to be insufficient to discriminate the wrong alignment from the correct one.

On panels A (CDK2), MED-SuMoLig clearly superimposed the three first active hits correctly on the query ligand 1PF8. A well-known feature about CDK2 active sites as for several other kinases is the nature of the interaction between known inhibitors and the so-called hinge region of the ATP binding site. This region involves, on the ligand side, two hydrogen-bond donors and a hydrogen-bond acceptor that interact with the backbone of Leu83 (N and O) and Glu81 (O),[32] respectively. MED-SuMoLig managed to highlight these preponderant pharmacophoric features by correctly aligning the hydrogen-bond donors and acceptors of the hit molecules on the corresponding objects in the query molecule. This allowed the program to correctly superimpose ligands that have a quite different scaffold, illustrating a good example of *scaffold hoping*.

For FXa (panels B), the MED-SuMoLig superimpositions are quite different from the *experimental alignment*. The experimental binding modes of the FXa actives are characterized by the quasi-systematic presence of a benzamidine group [$Ar-C(NH_2)NH$)] in the S1 pocket of FXa.[33] Despite a lack of optimal superimposition of the FXa actives, the benzamidine function was systematically detected (when present) by MED-SuMoLig. The program therefore provided a superimposition of the ligands on the basis of this sole chemical function. This explains the alignment observed in Figure 6. The remaining part of the ligands is characterized by an absence of consensual chemistry or even physicochemical properties, therefore impeding the performance of pharmacophore-based ligand screening method such as MED-SuMoLig (see below). Again, when superimposing all FXa actives, the only recurrent substructure to emerge is indeed the benzamidine function.

The results obtained with HIV-1p ligands are quite satisfactory as seen on panels C. For most of the runs, the ligands were correctly aligned. This result allows us to make interesting observations for some ligands. This concerns the nature of the superimposition of 1HBV (third hit) on 1AAQ (query). The MED-SuMoLig alignment seems to be flipped over and oriented in the wrong direction compared to the *experimental alignment*. But when considering the pure chemistry of both ligands, the MED-SuMoLig alignment seems to be correct. We checked the electron density map of 1HBV using the UPPSALA server[34] and found out as expected that the ligand in the crystal structure was in the correct orientation. So, it is interesting to see that the sole presence of the azepin in 1HBV is responsible for the flipping over of the ligand with respect to the other HIV-1p actives. We think that neither a pharcomophore-based method nor a shape-based ligand screening method could anticipate such a reorientation taking place without considering the presence of the receptor. Moreover, studies have reported that one ligand can have several modes of binding and be captured into the so-called reversed-binding mode.[35−38]

The results obtained for NA (panels D) are in good agreement with the experiments. An example of alignment obtained with MED-SuMoLig for NA actives is shown in Figure 6. This figure shows 1INF (green carbon), which is characterized by two major interactions with NA. The first involves the guanidinium group, which forms an electrostatic interaction with Glu275, and the second concerns the carboxylate function that makes salt bridges with three arginines (Arg116/292/374) in the active site.[39] These two groups were used by MED-SuMoLig to correctly orient the other actives with respect to 1INF. These results are quite satisfactory because NA is well-known to be a very difficult target for structure-based methods because of the relative symmetry displayed by several charged residues within the active site and since the active site is open and relatively flat. Such a binding pocket illustrates well cases where SBVS methods tend to fail (i.e., the docking can be correct but the ranking false), suggesting that the 3D ligand-based screening method could be used in combination with SBVS approaches on these difficult cases assuming the availability of an experimental cocrystal structure.

The alignment provided for RNase actives is illustrated on panels E. Overall, the alignments obtained were of good quality even though 1QHC (cyan carbon) was only partially superimposed on the query (1JN4, green carbon) as shown in Figure 6. We observed an interesting alignment pattern for 1O0O. The adenosine-based part of the ligand (yellow

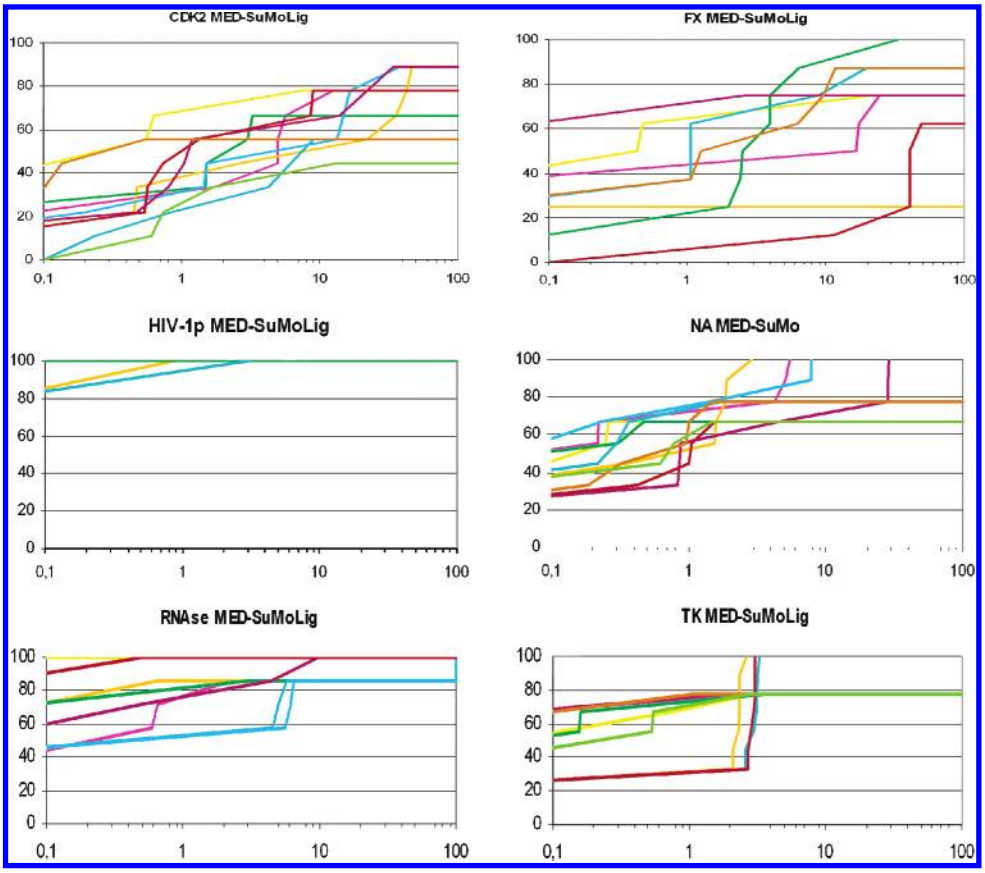

Figure 7. MED-SuMoLig enrichment curves. The *y* axis is the percentage level of recovered actives on a given protein, and the *x* axis is the logarithmic expression of the percentage of the database screened. Each color represents a different query ligand of the validation set for each protein.

carbon) was superimposed by MED-SuMoLig onto the adenosine part of 1JN4. This pattern is not observed in the experimentally derived alignment, but interestingly the adenine part of the ligands in question interacts in both complexes with His119 of RNase through a π/π stacking. In fact, His119 undergoes a complete flip over upon binding depending on which ligand is present in the active site. So in a way, the MED-SuMoLig alignment is a reasonable approximation of the experimental alignment considering the observed topological change in the binding cavity. This was observed in a recent study with the program Surflex-Sim[40] in the case of TK ligands, regarding the superimposition of the heterocycle of the guanine-based ligands onto the heterocycle of the thymidine-based ligands. In that case, Gln125 was the residue to be flipped over depending on the ligand in place. This type of behavior represents one of the limits of the ligand-based screening methods. Interestingly, we also made this observation with MED-SuMoLig on some of the TK actives.

Concerning TK and aside from the observations mentioned above, the alignment of the actives was very accurate as shown in panels F. The uridine-based motif of 1KIM (query) was correctly detected and superimposed on 1E2N (first, cyan carbon), 1E2K (second, pink carbon), and 1E2P (third, yellow carbon). The remaining part of the ligand was also aligned correctly using the presence of hydroxyl groups and/or hydrophobic functions depending on the query. In the case of small ligands like those of TK, a RMSD threshold of 1 Å used in the conformer generation can preclude certain conformations to arise. The conformations available in the database were not as close to the bioactive conformation as we expected; thus even though the alignments were satisfactory, those obtained using the X-ray conformation for the TK actives were better.

Enrichments. *Enrichment Curves.* A recent study[31] reported a screening campaign with various tools (structure- and ligand-based) on several protein targets. The database used contained about 20 000 ligands, and they calculated enrichment factors at 10% of the subsetting for the ligand-based approaches. Here, we report the enrichment factors calculated on a database that contains over 39 000 druglike molecules at four percentage levels of subsetting (10%, 5%, 3%, and 1%).

The output of the 52 runs was processed using the rescoring procedure described in the Methods section. This provided the enrichment curves (ECs) depicted in Figure 7. If N actives were present for protein P, the goal was to recover the $N - 1$ other actives on P. Because the database was multiconformational, the best conformation of each of the $N - 1$ hits was taken for the enrichment factor calculation.

The first observation is that the ECs are dependent on the query ligand. The runs were launched with the whole surfaces of the ligands as queries (i.e., even the nonconsensual parts). This was therefore not a pharmacophoric hypothesis representing the essential chemical functions of the ligands. As mentioned above, the MED-SuMo algorithm is built such that each part of the query molecule represents in itself a subquery. Each triangle from the triangle network represents a subquery and therefore seeks for its own local match. It is

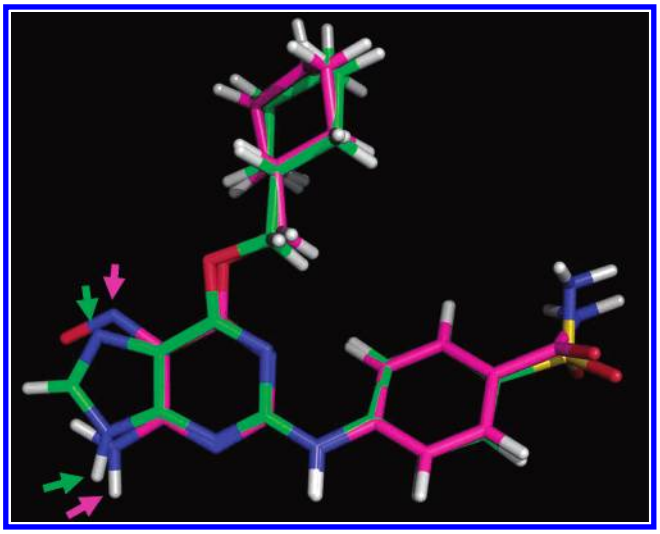

Figure 8. Example of scaffold hopping for CDK2 ligands. MED-SuMoLig superimposition of 1H1S (green carbon) on 1OGU (pink carbon). The presence of the amino and the nitro groups at the 4 and 5 position of the pyrimidine allowed MED-SuMoLig to superimpose a pyrimidine-based scaffold on a guanine-based scaffold. The pink and green arrows represent the hydrogen-bond acceptors and the hydrogen-bond donors (in each molecule) used for the superimposition.

only when the fusion of several triangles is possible (when respecting the geometric rules mentioned in the Methods section) that the union of these triangles represents a bigger patch and therefore a more relevant hit.

The dependence on the query is particularly true with FXa (panel B) for which the EC shows a poor recovering rate for three of the nine actives (red, 1F0S, and light orange, 1MQ5, and in a lesser way purple, 1FOR). Interestingly, these three ligands do not possess a benzamidine function, which confirmed the observation made in the previous section concerning the importance of the benzamidine group detection to recover a partial alignment. 1FOR has a relatively poor recovering rate (recovery of 75% of the actives around 25%) thanks to the amine function present on the isoquinoline group which represents two hydrogen-bond donors and an aromatic group that managed together to partially match the benzamidine function of the other ligands. One explanation for these results with FXa is that this subset of inhibitors is quite diverse chemically and, therefore, represents a real challenge for the ligand-based method (see the next section). For the six other actives on FXa, the recovering rate is satisfactory, with 75% percent of the actives recovered before 10% of the library was screened.

The results for CDK2 (panel A) are more homogeneous and display globally good recovering rates considering the chemical diversity of the ligands and their binding modes (see the Superimposability section). The worst recovering rates correspond to 1PXL (olive-green) which is specifically, with 1OGU (deep purple), the only CDK2 of the validation set ligand that has a pyrimidine-based scaffold. 1OGU did not display a bad recovering rate like 1PXL did because of the presence of an amino group (H-bond donor) and a nitro group (H-bond acceptor) at the 4 and 5 positions, respectively, of the pyrimidine. The presence of these groups has permitted the superimposition of the pyrimidine-based scaffold onto the guanine-based scaffold (Figure 8) of 1H1S and 1G5S as observed in the experimental alignment (data not shown). This again illustrates the ability of the method to undertake to some extent a "scaffold hoping" between structures with different chemical patterns.

For HIV-1p, MED-SuMoLig was particularly efficient and recovered the $N - 1$ actives in all cases. This might be due to the chemical nature of the active compounds (on HIV-1p), which is highly similar to oligopeptides, and therefore facilitates certainly the enrichment compared to more drug-like compounds (present in our collection). The presence of successive triangle patches associated with the amide groups (one hydrogen-bond acceptor + one hydrogen-bond donor) represent a graph motif that is easily tractable by MED-SuMoLig. This gives therefore a good foundation to extend the triangle patch found in common with the query and prevent the hit molecule from being trapped in the noise.

The NA runs gave good enrichment rates with more than 60% of the actives recovered right after the 1% level of subsetting. Three of the runs recovered 100% of the actives before the 5% level of subsetting.

For RNase, the results are also satisfactory with more than 80% of the actives recovered in the top 5% of the library screened. It is interesting to note that the two runs having the "worst" recovering rate are 1O0M (cyan) and 1O0N (blue), the only two RNase ligands of the validation set with no adenosine-based scaffold. Interestingly, MED-SuMoLig considered them as "outliers" but still did recover them before the 5% level of subsetting.

The results on TK show a good enrichment rate as well, with at least 78% of the actives recovered in the top 3% of the library. The results also show the specificity of three actives (2KI5, light orange; 1KI2, red; and 1KI3, cyan) with respect to the seven other compounds. Those are actually the guanine-based compounds of the TK validation set. As mentioned in the previous section, the superimposition of the guanine-based compounds on the thymine-based compounds is not exactly the one observed experimentally. Nevertheless, in the case of guanine-based ligands, the enrichment shows that the program managed to recover the $N - 1$ actives despite a lack of scaffold similarity, showing one more time an example of scaffold hopping.

The enrichment curves show an overall good recovering rate throughout the six sets of ligands tested. This represents an encouraging result considering that the rest of the molecules screened were also druglike molecules. The curves also emphasized the sensitivity of the approach at recovering ligands with a relatively distant scaffold but at the same time the discriminatory power to distinguish a molecule belonging to a different subclass. Therefore, we anticipate that we could use MED-SuMoLig on subsets of active molecules as a distance measurement in a hierarchical classification procedure (clustering) in order to segregate different potential actives into subclasses. This could facilitate the selection of cluster representatives that will in turn become a more legitimate MED-SuMoLig query toward large chemical libraries.

Enrichment Factors and Comparison to Tiers Programs. In order to further evaluate MED-SuMoLig, we performed screening experiments on a bank of 37 959 druglike molecules (see the Methods section) using 52 active ligands (i.e., the active molecules against our six protein targets) as query input (the entire bioactive X-ray structure for 3D search and

Table 2. Enrichment Factors[a]

protein	MED-SuMoLig				2D similarity				ROCS				ROCS-cff			
% database	10%	5%	3%	1%	10%	5%	3%	1%	10%	5%	3%	1%	10%	5%	3%	1%
CDK2	57.8	51.1	46.7	36.7	40.0	30.0	27.8	18.9	17.8	15.6	14.4	8.9	53.3	46.7	43.3	36.7
(10)	(13.9)	(11.3)	(13.0)	(14.9)	(12.4)	(13.2)	(15.1)	(10.0)	(12.4)	(8.9)	(8.7)	(8.3)	(18.5)	(17.2)	(14.4)	(15.9)
FX	54.7	48.4	45.3	39.1	37.5	26.4	19.4	12.5	38.9	29.2	26.4	16.7	45.8	37.5	23.6	16.7
(9)	(28.6)	(24.6)	(22.5)	(22.9)	(18.6)	(15.0)	(13.3)	(13.2)	(15.0)	(13.2)	(10.9)	(10.2)	(20.4)	(12.9)	(19.3)	(13.3)
HIV	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	80.0	70.0	65.0	60.0	95.0	90.0	85.0	75.0
(5)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(18.7)	(24.5)	(30.0)	(25.5)	(10.0)	(12.2)	(20.0)	(22.4)
NA	78.9	75.6	72.2	60.0	66.7	53.3	47.8	45.6	88.9	76.7	67.8	43.3	100.0	97.8	97.8	96.7
(10)	(14.4)	(10.9)	(11.4)	(8.9)	(30.1)	(27.3)	(24.8)	(23.7)	(30.0)	(28.5)	(25.2)	(17.3)	(0.0)	(6.7)	(6.7)	(10.0)
RNAse	91.1	82.1	76.8	73.2	100.0	100.0	100.0	100.0	37.5	26.8	25.0	17.9	70.71	57.1	55.4	44.2
(8)	(6.9)	(17.1)	(21.4)	(20.7)	(0.0)	(0.0)	(0.0)	(0.0)	(15.9)	(16.7)	(13.8)	(13.8)	(17.1)	(16.0)	(18.1)	(12.2)
TK	84.4	84.4	76.7	54.4	92.2	78.9	72.2	57.8	82.2	67.8	54.4	33.3	87.8	86.5	82.2	73.3
(10)	(10.2)	(10.2)	(10.5)	(21.3)	(8.7)	(23.0)	(21.8)	(21.5)	(18.1)	(16.1)	(17.5)	(16.5)	(6.7)	(10.0)	(14.1)	(18.6)
weighted average	75.6	71.2	66.9	56.7	69.7	60.8	56.8	50.7	56.5	46.7	40.1	27.85	74.3	68.4	63.7	56.7

[a] The enrichment factor values have been calculated at 10, 5, 3 and 1% and are given in this table as percentages. As an example, for CDK2, the value 57.8% obtained at 10% subsetting has been calculated by averaging the EF values of the 10 CDK2 actives at 10% and represents the percentage of actives recovered at 10%. All enrichment factors are represented with the standard deviation observed on the given subset of actives. The weighted average stands for an average of all EF percentage values over all protein targets weighted by the number of actives present for each protein target: 75.6% for MED-SuMoLig at 10% subsetting.

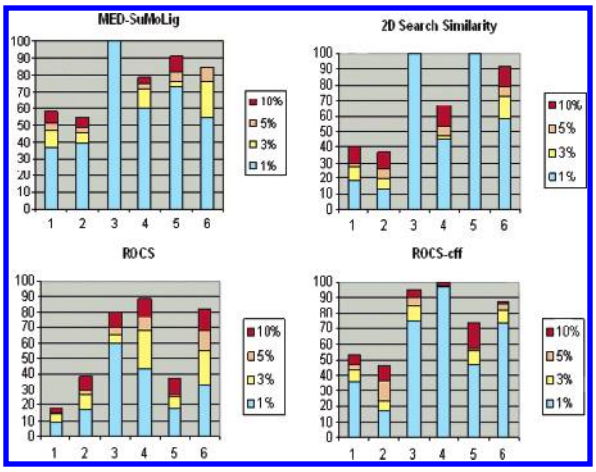

Figure 9. Cumulated enrichment rates. The histograms show the cumulated percentage of recovered actives for the six protein targets (1-CDK2, 2-FXa, 3-HIV-1p, 4-NA, 5-RNase, and 6-TK). This gives both the total averaged percentage of actives recovered and the detailed averaged enrichment rates at the four different percentage levels of subsetting. As an example for TK and with MED-SuMoLig, more than 80% of the actives were recovered when 10% of the whole database was screened, and more than 50% were recovered in the first percentage level of subsetting.

2D search). The overall strategy for evaluation of 3D/2D ligand-based screening computations follows the one reported by Chen et al.[31] The screening experiments were also carried out with ChemMine[23] and ROCS.[15] ChemMine is known to be a very efficient 2D search method,[23] and ROCS is a well-known 3D shape-matching package that has been, for instance, used successfully in the identification of a set of novel inhibitors of the ZipA-FtsZ protein−protein interaction.[41] We are aware that comparing in silico screening methods without inadvertently (dis)favoring the performance of one package over another is difficult, but we have strived to run the different programs carefully and checked the different control parameters. For both tiers programs tested, we ensured that, to the best of our knowledge, the best-performing parameters were used.

For ChemMine, recent studies[42] have strongly suggested that the atom sequence combined with the Tanimoto coefficient tends to provide better results at least in the scope of the descriptors tested. The 2D similarity search has therefore been performed using these parameters. The enrichment factors have been calculated using the Tanimoto coefficient as the "scoring" function.

For ROCS, as for the MED-SuMo Lig runs, the query inhibitor was always the bioactive conformation taken from the crystal structure with no minimization. This protocol has indeed been shown to give better performance with ROCS.[31] The score used for the enrichment calculation was either the *shape tanimoto* (ROCS-shape-only) or *combo* score (ROCS-cff) because a real discrepancy can be observed between the two, the latter providing clearly better results. Given the fact that the *optchem* option and the *combo* ranking give significantly better results, we will mainly discuss the EF values calculated for ROCS-cff, but we found that it was interesting for the sake of the discussion to present the EF values for ROCS-shape-only and also because we think these data highlight the importance of taking into account both the chemistry and the shape of the molecules.

Enrichment factors were calculated for each active molecule and for the three above-mentioned ligand-based screening tools as described in the Methods section. The EF values reported in Table 2 are an average per protein (that is, from all actives as a query on a given protein) and not one particular active. A global weighted average has also been calculated to determine the global behavior of the methods at diverse subsetting levels (%LS) across the six subsets of actives. EF values have been established at subsetting levels 10, 5, 3, and 1% of the input database, and as such, our investigation is much more aggressive than the one used by Chen et al.,[31] who only considered a generous 10% LS when evaluating the performance of 2D/3D ligand-based screening experiments. Figure 9 represents the cumulated percentage of actives recovered at 10, 5, 3, and 1% LS for each of the six protein targets.

Given the present validation set, the first observation resulting from our computations is that MED-SuMoLig is the only method that does not display bad EF values for a given protein target as opposed to ChemMine (e.g., CDK2,-

and FXa) or ROCS-cff (e.g., FXa). The second observation is that all methods perform well on HIV-1p and TK inhibitors, even though all fail to correctly align (this point is irrelevant for ChemMine) the guanine-based compounds on the thymine-based compounds for TK (data not shown). For HIV-1p, the methods obtain a perfect or nearly perfect enrichment even at 1% LS. Interestingly, MED-SuMoLig is the only tool tested that recovered more than half of the actives in the top 5% of the database for every single target tested, suggesting that it is robust and efficient regardless of the nature of the ligands.

More specifically, for CDK2 at a generous 10% LS, MED-SuMoLig displays the best averaged EF with a value of 57.8% while ChemMine and ROCS-cff obtain 40.0% and 53.3%, respectively. The tendency remains the same for the other subsetting levels except at 1% where ROCS-cff performs equally well. Those values which are not among the best EFs (see the following paragraphs) reflect the diversity of the CDK2 validation subset in terms of chemistry and size. This impedes the performance of the three methods. This also emphasizes that this diversity was differently handled depending on the screening tool used. It seems that a purely chemical-based method such as ChemMine encountered more difficulties than MED-SuMoLig or ROCS-cff to retrieve the correct active molecules. It is interesting to observe that, on the other hand, ROCS-shape-only has the worst EF value even at 10% and 17.8%, while we note 53.3% for ROCS-cff. This point underlines the benefit of using the chemistry as well as the shape for ligand-based VS experiments.

For FXa at 10%LS, MED-SuMoLig (54.7%) performs better than ROCS-cff (45.8%) and significantly better than ChemMine (37.5%) while at 1%LS the gap between MED-SuMoLig (39.1%) and the two other programs is quite important (12.5% for ChemMine and 16.67% for ROCS-cff). Even though the ligands of FXa have been obviously a real challenge for the three methods, MED-SuMoLig managed to recover, on average, more than one-third of the actives at 1%LS. Again, we suggest that the chemical diversity of the different sub-structure elements decreases the efficacy of the three programs even for methods purely based on shape such as ROCS-shape-only.

As previously stated, the results for HIV-1p were excellent for all three methods and more specifically for MED-SuMoLig and ChemMine, which display a perfect score at the four %LS values. It seems that the peptido-mimetic nature of the HIV-1p ligands favored the identification of active compounds over the others (for reasons exposed in the previous section), more druglike molecules, present in the database. It has to be noted that the large flexibility of the HIV-1p actives (up to 21 rotamers) impeded neither the performances of MED-SuMoLig nor those of ROCS-cff.

For NA, ROCS-cff outperforms both MED-SuMoLig and ChemMine, with the excellent score of 100% at 10% LS and a nearly perfect score at other %LS values. However, MED-SuMoLig performs well even at 1% LS (60.0%). The results for ROCS-shape-only at lower %LS values were not as good as those of MED-SuMoLig; therefore, the chemistry of such ligands helped to correctly rank the actives versus the decoys by favoring the match of persistent and critical chemical functions such as the acyl groups (data not shown).

In the case of RNase, ChemMine displays perfect scores at all %LS values. Nevertheless, MED-SuMoLig also performs well on these ligands with a very good EF of 73.2% at 1% LS, whereas ROCS-cff only displays an EF value of 44.2% even though it is drastically better than ROCS-shape-only (17.9%). The score discrepancy between ChemMine and ROCS-shape-only seems to show that a chemical criterion is more suited to retrieve coactive molecule for this specific family of ligands. Another explanation might be that the RNase actives have different sizes which impede more certainly an algorithm that focuses on shape matching[16] (ROCS) rather than on substructure detection.

For TK ligands, the three programs perform well at 10% and 5% LS. At 3% LS, ChemMine (72.2%) does not display as good results as MED-SuMoLig (76.7%) and ROCS-cff (82.2%). Finally, at 1% LS, only ROCS-cff maintains a relatively high EF (73.3%), whereas MED-SuMoLig displays 54.4%, which is still reasonable at this %LS.

Among the three programs tested, MED-SuMoLig is certainly the most robust across the various ligand families tested. It displays the weighted average EF at all %LS values (Table 2). When ligands display a more pronounced diversity in terms of chemistry and size (e.g., for CDK2 and FXa), MED-SuMoLig is the only method that maintains a recovering rate near 50% even at 3% LS, while the two other programs perform poorly. As mentioned in the previous sections, the MED-SuMo comparison algorithm is built such that every substructure in the query molecule becomes an independent subquery. Only for substructures with compatible geometry and density, the method merges more substructures in a wider patch of common subgraphs. It seems that the good enrichment rates obtained with MED-SuMoLig are mainly due to this characteristic, which correlates to favoring molecular similarities without penalizing molecular discrepancies.

The above observations made on TK and RNase actives concerning the induced fit of Gln125 and His119 address a reasonable question with regard to the validity of ligand-based approaches. On the one hand, it is legitimate to seek for optimal superimposition with respect to what would be present in hypothetical cocrystals, because this represents almost unquestionably the effective binding mode of the ligands. But on the other hand, the plasticity of certain ligand-based methods such as MED-SuMoLig or Surflex-Sim might bring new rationales concerning our vision of what is to be called canonical superimposition (and not binding mode). Such "irregular" superimpositions when founded on legitimate assumptions, that is, physically and/or chemically relevant, might provide useful information. Moreover, it cannot be ignored that ligands can have several binding modes with close binding energies.

A recent study[16] compared the performance of ROCS and ROCS-cff to structure-based screening packages. The authors also suggest that ROCS-cff performs significantly better than the original ROCS at least on the ligand tested, and that ROCS-cff tends to perform better than the structure-based packages except on one target, Factor VIIa. Surprisingly, in our tests, the worst performances of ROCS-shape-only and ROCS-cff were on the serine protease Factor Xa, which is highly homologous to Factor VIIa. By performing at least equally as well as ROCS-cff, we expect our program to compete with structure-based tools. In that recent ROCS

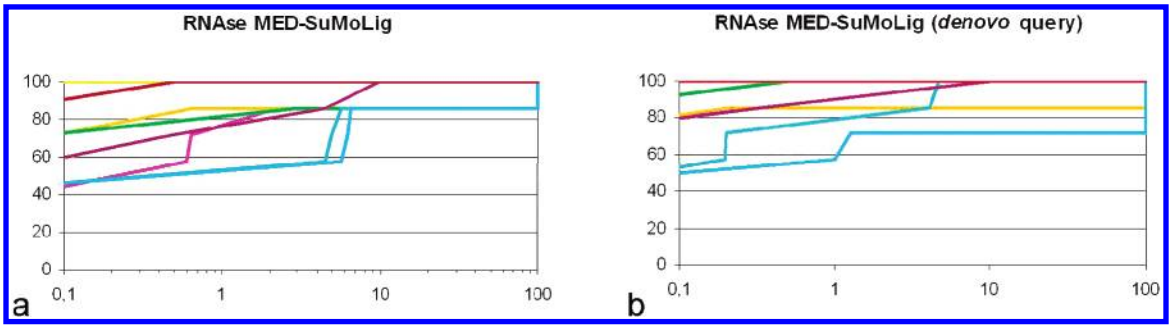

Figure 10. MED-SuMoLig enrichment curves for RNase. The *y* axis is the percentage level of recovered actives on a given protein, and the *x* axis is the logarithmic expression of the percentage of the database screened. Each color represents a different query ligand of the validation set. Panel a: All query molecules in the crystallographic conformation. Panel b: All query 3D molecules generated de novo with Omega.

study, the authors also mentioned that "replacing the crystallographic conformation of a query with a low-energy conformer obtained with Omega has essentially no impact on ROCS's performance" which is in contradiction with a previous study.[31] To investigate this issue, we also performed all the RNase runs using a low-energy conformer from Omega as the query molecule rather than with the crystallographic conformation (Figure 10). The choice of RNase actives was motivated by the fact that they are relatively flexible (≈seven rotamers per molecule on average) as opposed to those of TK or NA. Therefore, these ligands constitute a fair test to assess the capacity of MED-SuMoLig to retrieve from the diversity set RNase coactives by using exclusively low-energy conformation query molecules. It is clear from Figure 10 that the enrichments obtained with the de novo 3D conformations are comparable to those obtained with the bioactive conformations. It appears that, at least in this case, programs like MED-SuMoLig manage to identify molecule matches regardless of the conformations used as queries. We can consider the bioactive conformations of a given set of coactives as a meeting point in the vast conformational space, but in some cases other meeting points can be found as long as the molecules in play share a certain level of compatibility in terms of chemistry and degrees of freedom (rotamers). If this hypothesis is true, then it is not surprising to obtain enrichments similar to those obtained with the bioactive conformations. However, many additional studies are required to confirm and validate this observation. Nevertheless, the perspective of using de novo conformations as queries opens a much wider range of applications for 3D ligand-based computer tools, because it alleviates the burden of having an experimental structure of the query ligand cocrystallized in the desired target. Therefore, when no other options are available, it might be reasonable to use the de novo conformation of one or several ligands as input to programs such as MED-SuMoLig.

CONCLUSION

We have developed a new ligand-based method called MED-SuMoLig and evaluated its performance on six protein targets, a total of 52 diverse active molecules, and a compound collection containing 37 959 druglike molecules. Our program performs a 3D pharmacophore search and considers the local atomic density of the compared molecules. MED-SuMoLig performs best when the 3D structure of a bioactive molecule is known, although predicted 3D structures can also be used as query input. The method tends to perform better than 2D search approaches such as ChemMine and 3D shape-matching tools like ROCS on our validation set. The program is fast as large compound collections can be screened in a day on a single workstation. In fact, our ligand-based approach is very effective as compared to other ligand-based methods like quantitative structure−activity relationships and comparative molecular field analysis as these latter ones usually require extensive computer resources (and availability of a significant amount of experimental data to define appropriate descriptors).

We suggest that combining ligand-based methods with structure-based approaches will become a standard in tomorrow's medicinal chemistry research. The approaches are indeed highly complementary and definitively not mutually exclusive. MED-SuMoLig can easily be incorporated within a multistep computational scheme to screen large libraries of compounds in search of new hits or for lead optimization. We anticipate that MED-SuMoLig can become a tool of choice in many situations and indeed be used instead of structure-based approaches for difficult proteins like NA.

Software Licensing. Commercial Information about MED-SuMoLig is available at www.medit.fr. Questions about MED-SuMoLig licensing should be addressed to info@medit.fr. Researchers from the Inserm Institute U648 have no financial interests in MEDIT and collaborated with this company only for the present project. Therefore, MEDIT SA has the exclusivity for MED-SuMoLig sales.

ACKNOWLEDGMENT

We would like to thank the French Ministry of Research for co-funding this work and the INSERM institute for support. We thank OpenEye Scientific Software for providing Omega and ROCS.

Supporting Information Available: A table showing the validation set used for enrichment assays. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Kubinyi, H. Success Stories of Computer-Aided Design. In *Computer Applications in Pharmaceutical Research and Development*; Wang, B., Ed.; Wiley-Interscience: New York, 2006; Wiley Series in Drug Discovery and Development; pp 377−424.
(2) Shoichet, B. K.; McGovern, S. L.; Wei, B.; Irwin, J. J. Lead Discovery Using Molecular Docking. *Curr. Opin. Chem. Biol.* **2002**, *6* (4), 439−46.
(3) Shoichet, B. K. Virtual Screening of Chemical Libraries. *Nature* **2004**, *432* (7019), 862−5.

(4) Jalaie, M.; Shanmugasundaram, V. Virtual Screening: Are We There Yet? *Mini Rev. Med. Chem.* **2006**, *6* (10), 1159−67.
(5) Moro, S.; Braiuca, P.; Deflorian, F.; Ferrari, C.; Pastorin, G.; Cacciari, B.; Baraldi, P. G.; Varani, K.; Borea, P. A.; Spalluto, G. Combined Target-Based and Ligand-Based Drug Design Approach As a Tool To Define a Novel 3D-Pharmacophore Model of Human A3 Adenosine Receptor Antagonists: Pyrazolo[4,3-e]1,2,4-triazolo[1,5-c]pyrimidine Derivatives as a Key Study. *J. Med. Chem.* **2005**, *48* (1), 152−62.
(6) Davies, J. W.; Glick, M.; Jenkins, J. L. Streamlining Lead Discovery by Aligning in Silico and High-Throughput Screening. *Curr. Opin. Chem. Biol.* **2006**, *10* (4), 343−51.
(7) Honorio, K. M.; Garratt, R. C.; Polikarpov, I.; Andricopulo, A. D. 3D QSAR Comparative Molecular Field Analysis on Nonsteroidal Farnesoid X Receptor Activators. *J. Mol. Graphics Modell.* **2007**, *25* (6), 921−7.
(8) Kadam, R. U.; Roy, N. Cluster Analysis and Two-Dimensional Quantitative Structure−Activity Relationship (2D-QSAR) of Pseudomonas Aeruginosa Deacetylase LpxC Inhibitors. *Bioorg. Med. Chem. Lett.* **2006**, *16* (19), 5136−43.
(9) Sivaprakasam, P.; Xie, A.; Doerksen, R. J. Probing the Physicochemical and Structural Requirements for Glycogen Synthase Kinase-3alpha Inhibition: 2D-QSAR for 3-Anilino-4-phenylmaleimides. *Bioorg. Med. Chem.* **2006**, *14* (24), 8210−8.
(10) Zhao, W. G.; Wang, J. G.; Li, Z. M.; Yang, Z. Synthesis and Antiviral Activity against Tobacco Mosaic Virus and 3D-QSAR of Alpha-substituted-1,2,3-thiadiazoleacetamides. *Bioorg. Med. Chem. Lett.* **2006**, *16* (23), 6107−11.
(11) Eckert, H.; Bajorath, J. Molecular Similarity Analysis in Virtual Screening: Foundations, Limitations and Novel Approaches. *Drug Discovery Today* **2007**, *12* (5−6), 225−33.
(12) Doddareddy, M. R.; Choo, H.; Cho, Y. S.; Rhim, H.; Koh, H. Y.; Lee, J. H.; Jeong, S. W.; Pae, A. N. 3D Pharmacophore Based Virtual Screening of T-type Calcium Channel Blockers. *Bioorg. Med. Chem.* **2007**, *15* (2), 1091−105.
(13) Low, C. M.; Buck, I. M.; Cooke, T.; Cushnir, J. R.; Kalindjian, S. B.; Kotecha, A.; Pether, M. J.; Shankley, N. P.; Vinter, J. G.; Wright, L. Scaffold Hopping with Molecular Field Points: Identification of a Cholecystokinin-2 (CCK2) Receptor Pharmacophore and Its Use in the Design of a Prototypical Series of Pyrrole- and Imidazole-based CCK2 Antagonists. *J. Med. Chem.* **2005**, *48* (22), 6790−802.
(14) Schuster, D.; Maurer, E. M.; Laggner, C.; Nashev, L. G.; Wilckens, T.; Langer, T.; Odermatt, A. The Discovery of New 11beta-Hydroxysteroid Dehydrogenase Type 1 Inhibitors by Common Feature Pharmacophore Modeling and Virtual Screening. *J. Med. Chem.* **2006**, *49* (12), 3454−66.
(15) *ROCS*, version 2.2; Openeye Scientific Software LLC: Santa Fe, NM, 2006.
(16) Hawkins, P. C.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50* (1), 74−82.
(17) *Catalyst*; Accelrys Software Inc.: San Diego, CA.
(18) Meek, P. J.; Liu, Z.; Tian, L.; Wang, C. Y.; Welsh, W. J.; Zauhar, R. J. Shape Signatures: Speeding Up Computer Aided Drug Discovery. *Drug Discovery Today* **2006**, *11* (19−20), 895−904.
(19) Jambon, M.; Andrieu, O.; Combet, C.; Deleage, G.; Delfaud, F.; Geourjon, C. The SuMo Server: 3D Search for Protein Functional Sites. *Bioinformatics* **2005**, *21* (20), 3929−30.
(20) Jambon, M.; Imberty, A.; Deleage, G.; Geourjon, C. A New Bioinformatic Approach to Detect Common 3D Sites in Protein Structures. *Proteins* **2003**, *52* (2), 137−45.
(21) *SMARTS*; Daylight Chemical Information Systems, Inc.: Aliso Viejo, CA, 2005.
(22) *Omega 2.0*; Openeye Scientific Software LLC: Santa Fe, NM, 2006.
(23) Girke, T.; Cheng, L. C.; Raikhel, N.; ChemMine. A compound Mining Database for Chemical Genomics. *Plant Physiol.* **2005**, *138* (2), 573−7.
(24) Jambon, M., *IBCP*; CNRS: Lyon, France, 2003. http://martin.jambon.free.fr/phd/sumo-letter-oneside.pdf (accessed Apr 2007).
(25) Bohacek, R. S.; McMartin, C.; Guida, W. C. The Art and Practice of Structure-Based Drug Design: A Molecular Modeling Perspective. *Med. Res. Rev.* **1996**, *16* (1), 3−50.
(26) *Marvin*; Chemaxon Kft.: Budapest, Hungary, 2006.
(27) Miteva, M. A.; Violas, S.; Montes, M.; Gomez, D.; Tuffery, P.; Villoutreix, B. O. FAF-Drugs: Free ADME/tox Filtering of Compound Collections. *Nucleic Acids Res.* **2006**, 34 (Web Server issue), W738−44.
(28) *SMILES*; Daylight Chemical Information Systems, Inc.: Aliso Viego, CA, 2005.
(29) Kirchmair, J.; Wolber, G.; Laggner, C.; Langer, T. Comparative Performance Assessment of the Conformational Model Generators Omega and Catalyst: A Large-Scale Survey on the Retrieval of Protein-Bound Ligand Conformations. *J. Chem. Inf. Model.* **2006**, *46* (4), 1848−61.
(30) Willett, P. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(31) Chen, H.; Lyne, P. D.; Giordanetto, F.; Lovell, T.; Li, J. On Evaluating Molecular-Docking Methods for Pose Prediction and Enrichment Factors. *J. Chem. Inf. Model.* **2006**, *46* (1), 401−15.
(32) Liu, J. J.; Dermatakis, A.; Lukacs, C.; Konzelmann, F.; Chen, Y.; Kammlott, U.; Depinto, W.; Yang, H.; Yin, X.; Chen, Y.; Schutt, A.; Simcox, M. E.; Luk, K. C. 3,5,6-Trisubstituted Naphthostyrils as CDK2 Inhibitors. *Bioorg. Med. Chem. Lett* **2003**, *13* (15), 2465−8.
(33) Matter, H.; Defossa, E.; Heinelt, U.; Blohm, P. M.; Schneider, D.; Muller, A.; Herok, S.; Schreuder, H.; Liesum, A.; Brachvogel, V.; Lonze, P.; Walser, A.; Al-Obeidi, F.; Wildgoose, P. Design and Quantitative Structure−Activity Relationship of 3-Amidinobenzyl-1H-indole-2-carboxamides as Potent, Nonchiral, and Selective Inhibitors of Blood Coagulation Factor Xa. *J. Med. Chem.* **2002**, *45* (13), 2749−69.
(34) Kleywegt, G. J.; Harris, M. R.; Zou, J. Y.; Taylor, T. C.; Wahlby, A.; Jones, T. A. The Uppsala Electron-Density Server. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60* (12, 1), 2240−2249.
(35) Maignan, S.; Guilloteau, J. P.; Choi-Sledeski, Y. M.; Becker, M. R.; Ewing, W. R.; Pauls, H. W.; Spada, A. P.; Mikol, V. Molecular Structures of Human Factor Xa Complexed with Ketopiperazine Inhibitors: Preference for a Neutral Group in the S1 Pocket. *J. Med. Chem.* **2003**, *46* (5), 685−90.
(36) Reich, S. H.; Melnick, M.; Davies, J. F., II; Appelt, K.; Lewis, K. K.; Fuhry, M. A.; Pino, M.; Trippe, A. J.; Nguyen, D.; Dawson, H. Protein Structure-Based Design of Potent Orally Bioavailable, Nonpeptide Inhibitors of Human Immunodeficiency Virus Protease. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92* (8), 3298−302.
(37) Stoll, V.; Stewart, K. D.; Maring, C. J.; Muchmore, S.; Giranda, V.; Gu, Y. G.; Wang, G.; Chen, Y.; Sun, M.; Zhao, C.; Kennedy, A. L.; Madigan, D. L.; Xu, Y.; Saldivar, A.; Kati, W.; Laver, G.; Sowin, T.; Sham, H. L.; Greer, J.; Kempf, D. Influenza Neuraminidase Inhibitors: Structure-Based Design of a Novel Inhibitor Series. *Biochemistry* **2003**, *42* (3), 718−27.
(38) Stubbs, M. T.; Reyda, S.; Dullweber, F.; Moller, M.; Klebe, G.; Dorsch, D.; Mederski, W. W.; Wurziger, H. pH-Dependent Binding Modes Observed in Trypsin Crystals: Lessons for Structure-Based Drug Design. *ChemBioChem* **2002**, *3* (2−3), 246−9.
(39) Sudbeck, E. A.; Jedrzejas, M. J.; Singh, S.; Brouillette, W. J.; Air, G. M.; Laver, W. G.; Babu, Y. S.; Bantia, S.; Chand, P.; Chu, N.; Montgomery, J. A.; Walsh, D. A.; Luo, M. Guanidinobenzoic Acid Inhibitors of Influenza Virus Neuraminidase. *J. Mol. Biol.* **1997**, *267* (3), 584−94.
(40) Jain, A. N. Ligand-Based Structural Hypotheses for Virtual Screening. *J. Med. Chem.* **2004**, *47* (4), 947−61.
(41) Rush, T. S., III; Grant, J. A.; Mosyak, L.; Nicholls, A. A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein−Protein Interaction. *J. Med. Chem.* **2005**, *48* (5), 1489−95.
(42) Chen, X.; Reynolds, C. H. Performance of Similarity Measures in 2D Fragment-Based Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1407−14.

CI700031V