JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Subpocket Analysis Method for Fragment-Based Drug Discovery

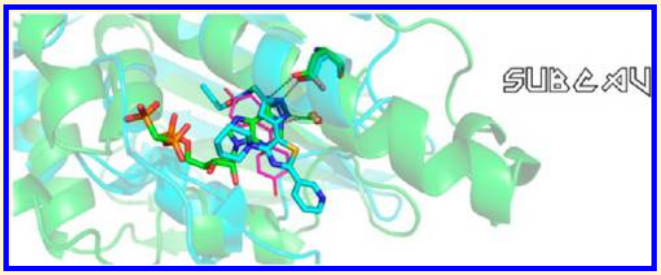
Tuomo Kalliokoski,[†] Tjelvar S. G. Olsson,[‡] and Anna Vulpetti*[,†]

[†]Novartis Institutes for Biomedical Research, Postfach, CH-4002 Basel, Switzerland

[‡]Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, United Kingdom

Supporting Information

ABSTRACT: Although two binding sites might be dissimilar overall, they might still bind the same fragments if they share suitable subpockets. Information about shared subpockets can be therefore used in fragment-based drug design to suggest new fragments or to replace existing fragments within an already known compound. A novel computational method called SubCav is described which allows the similarity searching and alignment of subpockets from a PDB-wide database against a user-defined query. The method is based on pharmacophoric fingerprints combined with a subpocket alignment algorithm. SubCav was shown to be effective in producing reasonable alignments for subpockets with low sequence similarity and be able to retrieve relevant subpockets from a large database of structures including those with different folds. It can also be used to analyze subpockets inside a protein family to facilitate drug design and to rationalize compound selectivity.



1. INTRODUCTION

Deriving knowledge from protein structural data to inform molecular design decisions is a key tenet of structure based drug design. Traditionally this has been achieved by solving structures of the target protein bound to a compound of interest on an ad hoc basis. The solved structure can then be used to generate new design ideas and hypotheses that could be tested, for example, by solving the structure of the new compound bound to the target. However, as the number of structures in the public domain, and in in-house pharmaceutical company repositories, increase there is the potential to make more use of these historic structures to help informed design decisions in a prospective manner.

Fragment-based drug discovery (FBDD) is now widely used both in industry and academia.[1] The underlying assertion in FBDD is that 'chemical space' can be more effectively sampled with small molecules (fragments) than with libraries of bigger compounds. Fragments that bind with low-affinity are then grown into larger molecules with higher affinity and specificity. Most methods rely on the availability of high-resolution structural information from X-ray crystallography. FBDD is also possible without such data, as recently shown by Konrat et al. who demonstrated that the meta-structure approach provides an alternative route to rational lead/fragment identification in cases where no 3D structure information about the biological target is available.[2]

Chemical space sampled by fragments is considerable, even when just by looking at publicly available fragments. Zuegg and Cooper analyzed 8 million compounds from a large number of chemical vendors.[3] Approximately 400,000 of these passed the commonly used fragment-like filter 'the rule of three'[4], and thus experimentally considering all possible fragments in drug design is not feasible. In order to facilitate the rational fragment library design process, there have been several studies on the binding preferences of proteins for fragments. These studies have previously been done by analyzing the distributions of residues that bind fragments in the Protein Databank (PDB).[5] Chan and co-workers looked at the chemical fragments that form hydrogen bonds to the most common residues found in binding sites (aspartic acid, glutamic acid, arginine, and histidine).[6] They showed that some fragments do indeed interact more frequently with certain amino acids. Wang et al. mapped the residue preferences of computationally fragmented ligands in the PDB.[7] This mapping could be used to identify whether a fragment was located in a favorable environment or if it could be modified to fit the site better.

Fragment binding environments can also be analyzed using binding site similarity methods. It is important to realize that a pair of proteins could differ significantly when comparing their whole active sites but could share similarity at the subpocket level. This can be the reason why two cavities with overall low sequence similarity can bind to identical chemical fragments. Examples of such are the similar sulfonamide- and trifluoromethyl-binding subpockets between cyclooxygenase-2 and the isoenzymes of the carbonic anhydrase family which bind the same functional groups in comparable orientations.[8]

There is a plethora of literature describing methods for analyzing binding sites.[9−11] Very different approaches have been applied, with methods ranging from BLOSUM62-based sequence comparisons to more complex approaches such as self-organizing fuzzy graphs that fill the pocket volume.[12,13] However, most of the existing work focuses on the similarity of whole binding sites, and there are far fewer studies focusing on

Received: October 31, 2012
Published: January 17, 2013

dx.doi.org/10.1021/ci300523r | J. Chem. Inf. Model. 2013, 53, 131−141

subpockets.[14] Med-SuMo[15] is commercially available software for locating similar regions on protein surfaces which are linked to certain chemical function. Wallach and Lilien developed a method for predicting subcavity binding preferences using pharmacophoric features.[16] CrystalDock is an approach for finding suitable fragments that match the pocket-lining residues.[17] Weisel et al. described Protein Ligand Interaction Explorer (PROLIX) which can be used to mine crystal structure databases for certain pocket environments that take part in the specific protein−ligand interactions.[18] Vulpetti et al. developed a 3D pharmacophoric fingerprint descriptor (named F-SPE-FP-PH3) to investigate the nature of the protein environment around each fluorine atoms in fluorinated ligands deposited in PDB.[19] During the preparation of this manuscript, Wood et al. published a method named Key Representation of Interaction in POckets (KRIPO).[20] KRIPO is also based on the idea of pharmacophoric fingerprints and was shown to be able to generate reasonable suggestions for bioisosteric replacements.

In this study, the F-SPE-FP-PH3 descriptor developed to analyze atom protein environments (i.e., around fluorine atoms) was extended and optimized for fragments' protein environments (subpockets). The resulted novel method is called SubCav, which is a tool for comparing and aligning subpockets. The performance of SubCav was evaluated using a data set of nonredundant PDB complexes hosting the same fragments. In addition, a prospective screen was done using the adenine portion of phosphomethylphosphonic acid adenylate ester (ACP) bound to Heat Shock Protein 90 (HSP90) to illustrate the potential application of the developed tool for drug design purposes. Finally, the cofactor binding sites of histone methyl-transferases were analyzed as another application of SubCav.

2. MATERIAL AND METHODS

This section is divided into the description of SubCav (2.1), the validation studies (2.2), and the analysis of the histone-methyl-transferases (2.3).

2.1. Description of the SubCav-Method. SubCav works in two steps. In the first step, the pregenerated fingerprints are compared, and then, if possible, the subpockets are aligned.

2.1.1. Subpocket Extraction. As the definition of a subpocket is not trivial and is a separate problem from similarity comparisons, a commonly used approach of defining pockets is to focus on protein atoms/residues at a certain distance from any atom of the bound ligand.[21] In this study all protein atoms at 4.5 Å distance from any of the fragment atoms were defined to be a part of a subpocket. All water and other nonprotein atoms were excluded from the subcavity. The PDB atom types were converted into eight pharmacophoric features (Table 1). These features were derived from the original nine used in the previously developed F-SPE-FP-PH3 descriptor[19] by only removing the fluorine atom type.[22] Only subpockets that had in total more than six protein atoms at 4.5 Å distance from any of the fragment atoms were considered as subpockets.

2.1.2. Subpocket Description and Similarity Calculation. The fingerprint generation resembles the previously described the F-SPE-FP-PH3 approach.[19] Triangles were enumerated between the features of the subpocket, and the distances were then binned (Figure 1A).

The four distance bins and eight features generate a 7680 element long fingerprint. The 2D similarity between the two subpocket fingerprints was calculated using two metrics. The

Table 1. Conversion Table Used To Convert PDB Atom Types to SubCav Pharmacophoric Features

SubCav atom type	description (feature type)	PDB atom types
0	donor (D)	LYS.NZ
1	α-carbon (CA)	ALA.CA ARG.CA ASN.CA ASP.CA CYS.CA GLN.CA GLU.CA GLY.CA HIS.CA ILE.CA LEU.CA LYS.CA MET.CA PHE.CA PRO.CA SER.CA THR.CA TRP.CA TYR.CA VAL.CA
2	carbon of the carbonyl (C)	ALA.C ARG.C ARG.CZ ASN.C ASN.CG ASP.C ASP.CG CYS.C GLN.C GLN.CD GLU.C GLU.CD GLY.C HIS.C ILE.C LEU.C LYS.C MET.C PHE.C PRO.C SER.C THR.C TRP.C TYR.C VAL.C
3	neutral donor and acceptor (P)	HIS.ND1 HIS.NE2 SER.OG THR.OG1 TYR.OH
4	hydrophobe (H)	ALA.CB ARG.CB ARG.CD ARG.CG ASN.CB ASP.CB CYS.CB CYS.SG GLN.CB GLN.CG GLU.CB GLU.CG HIS.CB HIS.CG ILE.CB ILE.CD1 ILE.CG1 ILE.CG2 LEU.CB LEU.CD1 LEU.CD2 LEU.CG LYS.CB LYS.CD LYS.CE LYS.CG MET.CB MET.CE MET.CG MET.SD PHE.CB PRO.CB PRO.CD PRO.CG SER.CB THR.CB THR.CG2 TRP.CB TYR.CB VAL.CB VAL.CG1 VAL.CG2
5	donors with sp2 character (π-donor) (D=)	ALA.N ARG.N ARG.NE ARG.NH1 ARG.NH2 ASN.N ASN.ND2 ASP.N CYS.N GLN.N GLN.NE2 GLU.N GLY.N HIS.N ILE.N LEU.N LYS.N MET.N PHE.N SER.N THR.N TRP.N TRP.NE1 TYR.N VAL.N
6	acceptors with sp2 character (π-acceptor) (A=)	ALA.O ARG.O ASN.O ASN.OD1 ASP.O ASP.OD1 ASP.OD2 CYS.O GLN.O GLN.OE1 GLU.O GLU.OE1 GLU.OE2 GLY.O HIS.O ILE.O LEU.O LYS.O MET.O PHE.O PRO.O SER.O THR.O TRP.O TYR.O VAL.O
7	π-hydrophobe (H=)	HIS.CD2 HIS.CE1 PHE.CD1 PHE.CD2 PHE.CE1 PHE.CE2 PHE.CG PHE.CZ TRP.CD1 TRP.CD2 TRP.CE2 TRP.CE3 TRP.CG TRP.CH2 TRP.CZ2 TRP.CZ3 TYR.CD1 TYR.CD2 TYR.CE1 TYR.CE2 TYR.CG TYR.CZ
8	ignored	PRO.N and all HETATM

dx.doi.org/10.1021/ci300523r | J. Chem. Inf. Model. 2013, 53, 131−141

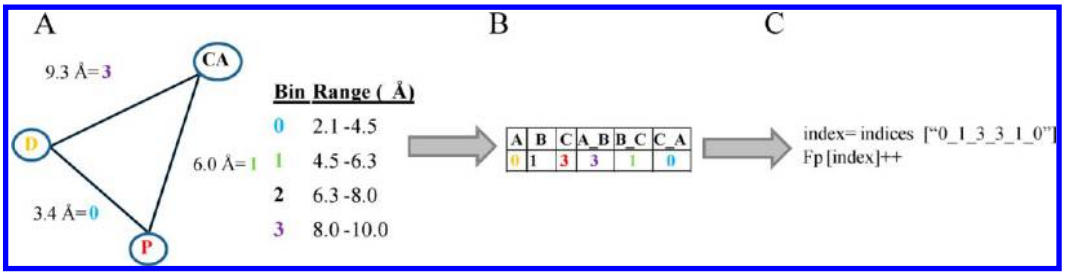


Figure 1. Encoding and distance binning of the subpocket features into a fingerprint. A: All atoms in the subpocket are encoded into atom types described in Table 1, and the distances between these points are binned. B: The triangles are encoded into canonical representation. C: Finally, the value in the correct fingerprint element is increased.

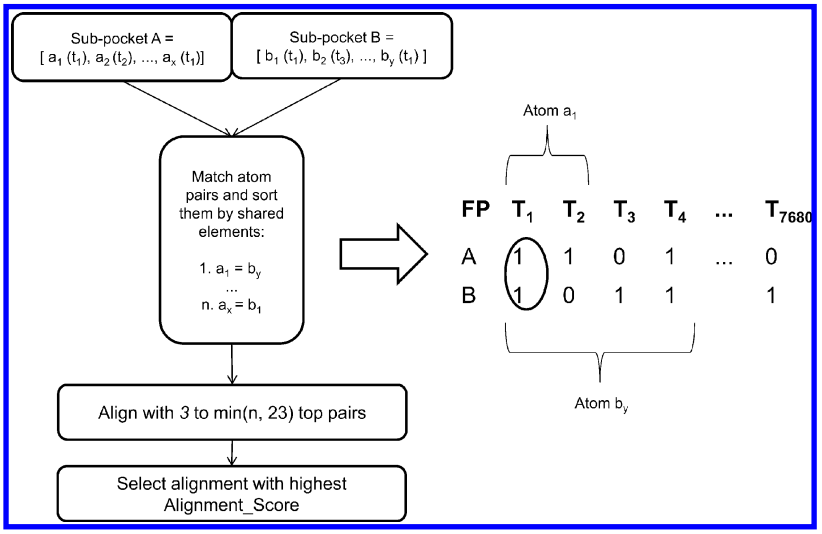


Figure 2. The alignment algorithm. The atoms in the two subpockets are paired by the atom types (t_1, t_2, and t_3). For each pair, the shared elements (i.e., T_1 in the pair $a_1 = b_y$ taken as an example on the right) are determined between the fingerprint elements containing the atoms of the pair (T_1 and T_2 for atom a_1 in the fingerprint generated for the subpocket A, and T_1, T_3 and T_4 for atom b_y in the subpockect B fingerprint). The whole fingerprint (FP) contains 7680 elements (i.e., the enumerated triangles T_1-T_{7680}). The list of paired atoms is sorted in order of decreasing number of shared elements of the atoms and the best alignment is selected by Alignment_Score.

normalized Tanimoto similarity measures the similarity between two fingerprints A and B by comparing the number of common elements in the fingerprints (#AB) to the minimum number of elements between the two fingerprints

$$\text{Tanimoto} = \#AB/\min(\#A, \#B)$$

where #A is the number of elements in fingerprint A, and #B is the number of elements in fingerprint B.

Inspired by the work of Ito et al., Cosine similarity was also implemented.[23] The similarity between the two fingerprints A and B (encoded as a high-dimensional feature vector that describes the frequency of the triangles) is calculated as the cosine of the two vectors

$$\text{Cos}(A, B) = A \cdot B / \| A \| \| B \|$$

where ||A|| and ||B|| are the norms of the vectors A and B. The norm of a vector Q is calculated as such

$$\| Q \| = \sqrt{\sum_{i=1}^{7680} q_i^2}$$

where q is an element of vector Q.

2.1.3. Subpocket Alignment. The alignment of two subpockets was implemented by using the Procrustes-like alignment method of Kabsch[24] implemented in BioPython,[25] which requires assignment of matching atoms between the two subpockets. Because the generated fingerprints have a large number of elements, a superpositioning method based on matching triangles would result in a combinatorial explosion. The selection of atoms of the same atom type to be used in the superpositioning was therefore prioritized based on atoms sharing the most fingerprint elements (Figure 2). The algorithm first calculates which fingerprint elements belong to each atom. The number of shared elements is then calculated for each pair of atoms having matching atom types between the two subpockets. The atom pairs with the highest shared triangles are listed. In order to make the algorithm deterministic this list is ranked in order of decreasing number of triangles they share and by their atom numbers.

The algorithm selects top n pairs from the list of pairs of size N (n = 3,...,min(23,N)) for the alignment. The maximum number of selected pairs was limited to 23 for computational efficiency. For each superposition the root-mean-square deviation (RMSD) of n selected protein atoms pairs is calculated (the lower limit for RMSD is set to 1×10^{-14} to avoid division by zero), and the best alignment is selected as the one with the maximum Alignment_Score defined as

$$\text{Alignment\_Score} = n/\text{RMSD}$$

The 3D similarity between the two aligned subpockets is quantified by calculating how many of the protein atoms with

dx.doi.org/10.1021/ci300523r | J. Chem. Inf. Model. 2013, 53, 131−141

the same pharmacophoric feature type match. The overlap of SubCav features (O_{sc}) is calculated as follows

$$O_{sc} = \text{atoms\_matched}/\min(\#a, \#b)$$

where #a is the number of atoms in subpocket A, #b is the number of atoms in subpocket B, and atoms_matched is the count of matched atoms. Two atoms are considered matched if they have the same SubCav atom type and the distance between them is less than 1.0 Å. Since the algorithm can produce slightly different alignments for two subpockets depending on which one is selected as query, SubCav generates both alignments and outputs the one with higher overlap value. After the SubCav alignment the RMSD between the two identical fragments bound to the subpockets was also calculated ($RMSD_{sc}$).

2.1.4. Additional Protein Annotations. Sequence similarity and protein annotations were used together with SubCav similarity to facilitate the analysis of the results in order to highlight nontrivial matches. The sequence similarity of the proteins was calculated using the Needleman-Wunsch method[26] implemented in EMBOSS.[27] The default values for opening penalty of 10.0 and extension penalty 0.5 were used.

Uniprot annotations for the protein structures were extracted from Uniprot Knowledgebase.[28] Two commonly structural classifications for proteins were also used (SCOP[29,30] family annotation and PFAM[31]).

2.1.5. Software Implementation and Hardware Used. SubCav was written in Python and makes use of RDKit[32] for ligand operations as well as BioPython[25] and OpenBabel[33,34] for protein structure manipulation. Calculations were performed on a Linux-cluster with x86_64 architecture. The software is available from the authors upon request.

2.2. Validation of SubCav. Unlike when validating a whole pocket similarity method, the validation of a novel subpocket similarity algorithm is far from trivial. There is no readily available data set which one could use or even a commonly accepted validation criteria. Here, subpockets were defined similar if they share similar 3D distribution of pharmacophoric features and, in addition, bind the same fragment in similar pose. Pairs of subpockets satisfying both criteria were identified by superposing all pairs of subpockets using the coordinate transformations given by the fragment atoms alignment (fragment-based pairwise alignment). The RMSD between the fragments ($RMSD_{frag}$) and the overlap of the subpocket atoms after the fragment-based alignment were calculated (O_{frag}). These criteria are quantifiable, and the similarity between two subpockets can also be visually verified from the alignment of the two subpockets. It must be pointed out that subpocket similarity, much like chemical similarity between ligands, is a subjective concept, and this naturally complicates the validation of subpocket similarity methods in general.

We defined "true" similar subpockets as those with $RMSD_{frag}$ less than 1.5 Å and O_{frag} greater than a defined threshold ($O_{threshold}$ ranging from 0.5 to 1). Most studies done on protein−ligand docking methods consider the docking having been successful if the docked ligand has a RMSD less than 2.0 Å to the crystal binding mode. However, Verdonk and coworkers have shown that this cutoff is often too generous for fragments, and therefore the stricter limit was used.[35]

The defined "true" similar subpocket data set was used for validating the new SubCav method by analyzing the ability of the method in correctly aligning these subpockets based on the protein subpocket information only (i.e., not using any ligand information).

In sections 2.2.1 and 2.2.2 details on fragment generation and crystal structures selection process are reported.

2.2.1. Fragment Generation. As there are not enough crystal structures of pure fragments in the PDB, the bound ligands were fragmented to produce a database of protein environments around those fragments. There are many fragmentation schemes described in the literature.[36] A well-known example of a fragmenting scheme is the REtrosynthetic Combinatorial Analysis Procedure (RECAP),[37] which is a knowledge-based method. A more advanced version of the original RECAP is the Breaking of Retrosynthetically Interesting Chemical Substructures (BRICS),[38] which was selected as the fragmentation scheme in this study. The RDKit implementation of the algorithm was used. The definition of a fragment was as follows: molecular weight <300 Da, calculated LogP < 3 and either number of heavy atoms >5 or number of rings >0. LogP was calculated by the method by Wildman and Crippen[39] implemented in RDKit. These definitions are permissive in order to maximize the coverage of the crystal structures in term of fragments. Fragments were grouped together by Topological Fingerprints in RDKit with a Tanimoto similarity >0.99. Topological Fingerprints enable the differentiation between identical fragments with different substitution patterns. Rare fragments that were found in less than three different protein structures were excluded from the study.

2.2.2. Data Sets. Two data sets of fragment protein environments were generated for SubCav. For validation purposes a nonredundant subset of PDB structures was selected. A larger database, for prospective screening, based on the majority of PDB structures was also generated.

The crystal structures for the initial validation study were selected from the Protein-Small Molecule Database (PSMDB).[40] This database provides a nonredundant subset of the PDB, considering both the ligand and protein. The construction of PSMDB has been described in detail in the original publication and thus is only briefly explained here. All structures of reasonable quality are extracted from the latest PDB release. The similarity between the proteins is compared using BLAST. The ligand similarity is calculated using Daylight fingerprints and a Tanimoto coefficient. All PDB structures are compared in a pairwise manner, and the resulting similarity matrix is used for the selection of a diverse subset. There are several different subsets available at PSMDB with various similarity thresholds. A set with 90% nonredundant protein and 85% nonredundant ligands with a minimum of 13 heavy atoms were selected for this study because it is the most diverse and has ligands of reasonable size for the fragmentation (see 2.2.1). From the 6,097 protein structures listed in PSMDB, 3,886 were used in the final data set after applying the fragment filters described in section 2.2.1. 332 different fragments types and 17,044 subpockets were extracted.

A database containing all fragments in the PDB was also constructed to facilitate PDB-wide searches (FRAGPDB). The SMILES strings of all PDB ligands were downloaded from LigandExpo[41] and fragmented (see 2.2.1). The PDB structures that had a fragment, compliant with fragment definition, were then checked against structure quality criteria. Crystal structures with a resolution of at least 2.8 Å and with at least one chain with over 50 amino acids were considered. PDB files that had multiple models of the same structure in them were

discarded. [PDB codes are as follows: 1m0k, 1m0m, 1o0a, 1ohh, 1p8h, 1p8u, 1t18, 1t19, 1t1a, 1t1b, 1t1c, 1t3n, 1ts0, 1ts6, 1ts8, 1x0i, 1yk0, 1yrq, 2icy, 1ts7, 2q4t, 2q4y, 2q3u, 2q3w, 2q46, 2q4g, 2q4h, 2q4v, 2q4x, 2q3m, 2q3o, 2q3r, 2q3t, 2q4b, 2ull.] The fragments and subpockets were then extracted and saved into a database, and the fingerprints were calculated. The 194,356 subpocket-database was based on 31,077 PDBs with 9,304 different ligands.

2.3. Analysis of Histone Methyl Transferase Binding Sites. Here a new class of drug targets called histone methyltransferases were analyzed using the SubCav method. A similar data set to that used by Campagna-Slater et al.[42] consisting of arginine methyl transferases (RMT) and lysine methyl transferases (KMT) was created. The PDB codes used for the analysis are shown in Table 2. The cofactor pocket of this target family was analyzed by fragmenting the bound S-adenosylmethionine (SAM) or S-adenosyl-L-homocysteine (SAH) ligand in three: adenine, ribose, and tail fragments (Figure 3).

Table 2. Data Set Used for the Histone Methyl Transferase Subpocket Analysis

protein name	PDB code	type
PRMT1	1or8	RMT
PRMT3	2fyt	RMT
CARM1	3b3f	RMT
DOT1L	1nw3	KMT
GLP	2rfi	KMT
G9a	3k5k	KMT
MLL1	2w5z	KMT
SETD2	3h6l	KMT
SET7	1o9s	KMT
SET8	1zkk	KMT
SETMAR	3bo5	KMT
SMYD3	3mek	KMT
SUV39H2	2r3a	KMT

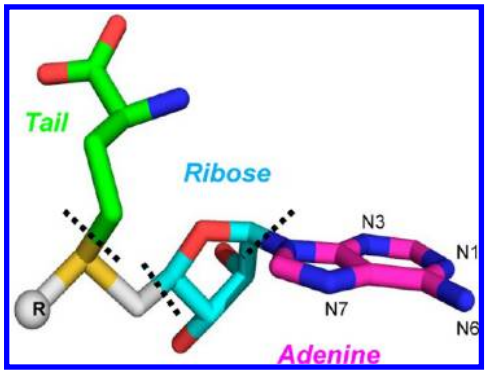


Figure 3. Fragmentation of SAM (R = methyl) or SAH (R = nothing) into adenine, ribose, and tail fragments.

The corresponding dissimilarity matrix for the subpockets tailored around each fragment was generated by calculating the dissimilarity ($1\text{-}O_{sc}$) for each pair of proteins. The matrix was then clustered using hierarchical cluster analysis with average agglomeration in the statistical software package R.[43]

3. RESULTS AND DISCUSSION

3.1. Validation Using PSMDB. To validate the proposed SubCav method 3,394,572 pairs of subpockets that had identical fragments bound to them were extracted from the fragments derived by fragmenting the ligands in the nonredundant data set of PDB (PSMDB). Only those pairs of subpockets which could bind identical fragments with a $RMSD_{frag} \leq 1.5$ Å after the fragment-based pairwise alignment were compared with fragment- and SubCav-based alignments (in total 3,268,620 pairs). Table 3 reports in the second column the number of subpocket pairs with a calculated O_{frag} value $\geq O_{threshold}$ and $RMSD_{frag}$ between the fragments $\leq$1.5 Å after the fragment-based pairwise alignment. This event is rare as only 2.73% of all the pairs considering $O_{threshold} = 0.5$ and 0.05% with $O_{threshold} = 1.0$ matched these criteria. SubCav was fairly successful in correctly producing the alignments of these pairs in a similar fashion as produced by fragment-based alignment in 72.8% to 85.2% of the cases, depending on the $O_{threshold}$ value used for the definition for a matched pair (i.e., $O_{sc} > O_{threshold}$ with $O_{threshold} = 0.5$ and $O_{threshold} = 1.0$, respectively).

Table 3. Numbers of Different Pairs of Alignments with Several Overlap Thresholds $O_{threshold}$ for Fragment- and SubCav-Based Alignments[a]

$O_{threshold}$	fragment-based OK	SubCav-based OK	both OK	not matched
0.50	89130	104756	64859	3118540
0.60	50967	66217	41228	3185231
0.70	30448	38782	25822	3222237
0.80	16772	20229	14646	3245273
0.90	6975	8020	6216	3259590
1.00	1521	1641	1296	3266643

[a](Total number of subpocket pairs after removal of the pairs with $RMSD_{frag} \geq 1.5$ Å is 3,268,620). Fragment-based OK corresponds to the number of pairs with $O_{frag} \geq O_{threshold}$ and $RMSD_{frag} \leq 1.5$ Å. SubCav-based OK corresponds to the number of pairs with $O_{sc} \geq O_{threshold}$ and $RMSD_{sc} \leq 1.5$ Å. Not matched corresponds to the number of pairs with either ($O_{frag} < O_{threshold}$ and $RMSD_{frag} \leq 1.5$ Å) or ($O_{sc} < O_{threshold}$).

However, more subpockets could be correctly aligned using the subpocket pharmacophoric features than when using the fragment heavy atoms to perform the superposition (i.e., compare columns two and three of Table 3). Few randomly selected instances where the subpocket alignment succeeded and the fragment based superposition failed (at $O_{threshold} = 1.0$) were inspected. The small differences between the conformations of the two fragments can generate invalid subpocket alignments (see the Supporting Information for an example). These cases (with $O_{sc} \geq O_{threshold}$ and $RMSD_{sc} \leq 1.5$ Å), not found by fragment-based alignment, have also to be considered to be correct matches.

The retrieval performance of the 2D similarity metrics (Tanimoto and Cosine, see 2.1.2) for matching subpockets was also investigated similarly to ref 44. Fingerprint comparison is faster than the alignment of subpockets and could have been used as a filter in the screening process. However, both of the 2D similarity metrics yielded unsatisfactory recall and precision values between 0.44 and 0.53 (see the Supporting Information), and thus the 2D similarity scores were not used further. It is possible that the fingerprint could be optimized for better performance. However, as the screening of the whole FRAGPDB (ca. 195,000 subpockets) requires only about an

dx.doi.org/10.1021/ci300523r | J. Chem. Inf. Model. 2013, 53, 131−141

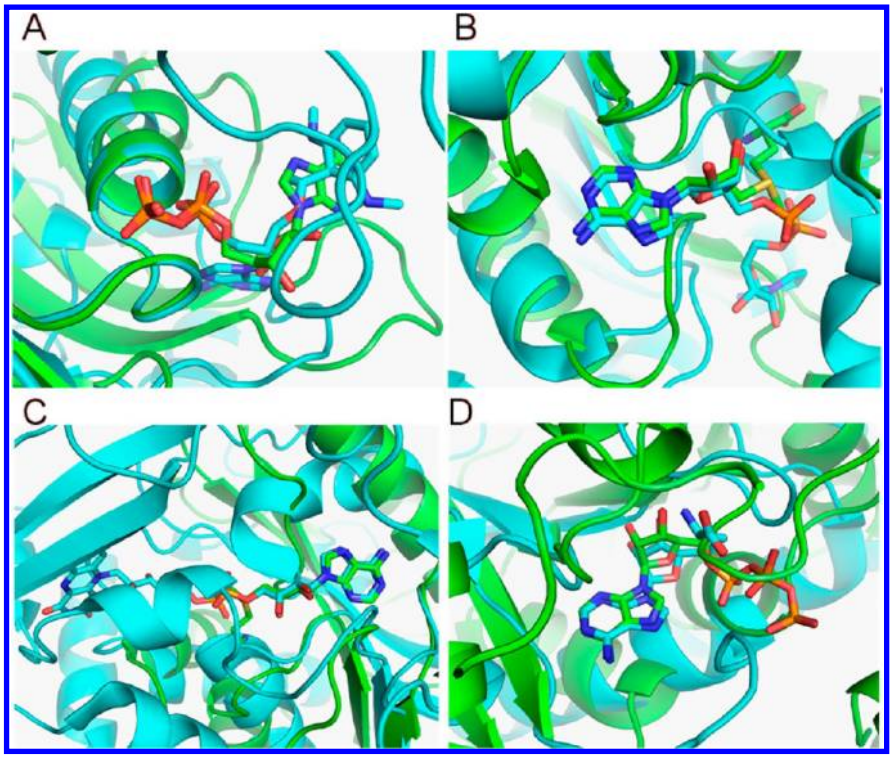


Figure 4. Four examples of subpocket pairs aligned by SubCav from proteins sharing low sequence similarity. A: 2ff7[47] (green) aligned to 1lvk[48] (cyan) using the phosphate subpocket. B: 2avd (green) aligned to 1p1r[49] (cyan) using the adenine subpocket. C: 2plw (green) aligned to 1ryi[50] (cyan) using the adenine subpocket. D: 3hgm[51] (green) aligned to 1mxi[52] (cyan) using the adenine subpocket. PyMol sessions are provided as Supporting Information.

Table 4. Proteins Found from HSP90 Screening with Sequence Similarity <30% and O_{sc} > 0.50[a]

entry	UniProt ID(s)	PDB	Lig[b]	Tan[c]	Cos[d]	O_{sc}[e]	Seq[f]
1	ODP2_HUMAN(P10515); PDK2_RAT(Q64536)	3crl	ANP	0.29	0.57	0.81	17.6
2	PDK2_HUMAN(Q15119)	2bu8	ADP	0.29	0.54	0.84	20.4
3	ODP2_HUMAN(P10515); PDK3_HUMAN(Q15120)	1y8p	ATP	0.37	0.6	0.86	17.9
4	PDK4_HUMAN(Q16654)	2zdx	P4A	0.25	0.56	0.64	16.8
5	BCKD_RAT(Q00972)	1gjv	SAP	0.3	0.61	0.82	14.1
6	GYRB_ECOLI(P0AES6)	3g7e	B46	0.2	0.28	0.81	29.7
7	TOP2_YEAST(P06786)	1qzr	ANP	0.2	0.31	0.64	21.8
8	TOP2A_HUMAN(P11388)	1zxm	ANP	0.17	0.39	0.64	18.9
9	TOP6B_SULSH(O05207)	1z59	ADP	0.43	0.55	0.82	16.1
10	CHEA_THEMA(Q56310)	1i5a	ACP	0.24	0.46	0.77	24
11	DESK_BACSU(O34757)	3ehg	ATP	0.22	0.3	0.59	17.5
12	MUTL_ECOLI(P23367)	1b63	ANP	0.53	0.69	0.86	23.9
13	PMS1_YEAST(P14242)	3h4l	ANP	0.4	0.59	0.77	25.6
14	PMS2_HUMAN(P54278)	1ea6	ADP	0.4	0.59	0.86	24.7
15	MLH1_HUMAN(P40692)	3na3	ATP	0.34	0.65	0.77	23.2
16	ENPL_CANFA(P41148)	2o1v	ADP	0.29	0.58	0.82	23.6
17	PARE_ECOLI(P20083)	1s16	ANP	0.3	0.46	0.82	21.3
18	SP2AA_GEOSE(O32726); SP2AB_GEOSE(O32727)	1til	ATP	0.24	0.41	0.77	4.1

[a]For clarity only the PDB with highest overlap for each protein is shown and proteins with unknown UniProt IDs are excluded. [b]Ligand bound to the PDB. [c]Tanimoto similarity. [d]Cosine similarity. [e]Overlap. [f]Sequence similarity of the protein chains.

hour of calculation time on 100 CPUs, the method presented in this paper is sufficiently fast for drug design. Visual inspection and user judgment of the alignment for the top-ranked solutions is, as in many other computational workflows (e.g., docking pose analysis), still an unavoidable step.

Some examples of meaningful and not-trivial alignments of subpockets (with O_{sc} > 0.5) from proteins sharing low sequence similarity (≤10%) and different bound cofactors are illustrated in Figure 4: ADP-MNT in Figure 4A, SAM-NAI in Figure 4B, SAM-FAD in Figure 4C, and ATP-SAH in Figure 4D. Other groups have also found out that targets with no sequence homology can still have similar binding sites and bind the same cofactor or a portion of it.[45,46]

3.2. Prospective Screen with HSP90 and FRAGPDB. To illustrate the potential of using subpocket searches to identify fragment-like bioisosteric replacements, a prospective screen on

dx.doi.org/10.1021/ci300523r | J. Chem. Inf. Model. 2013, 53, 131−141

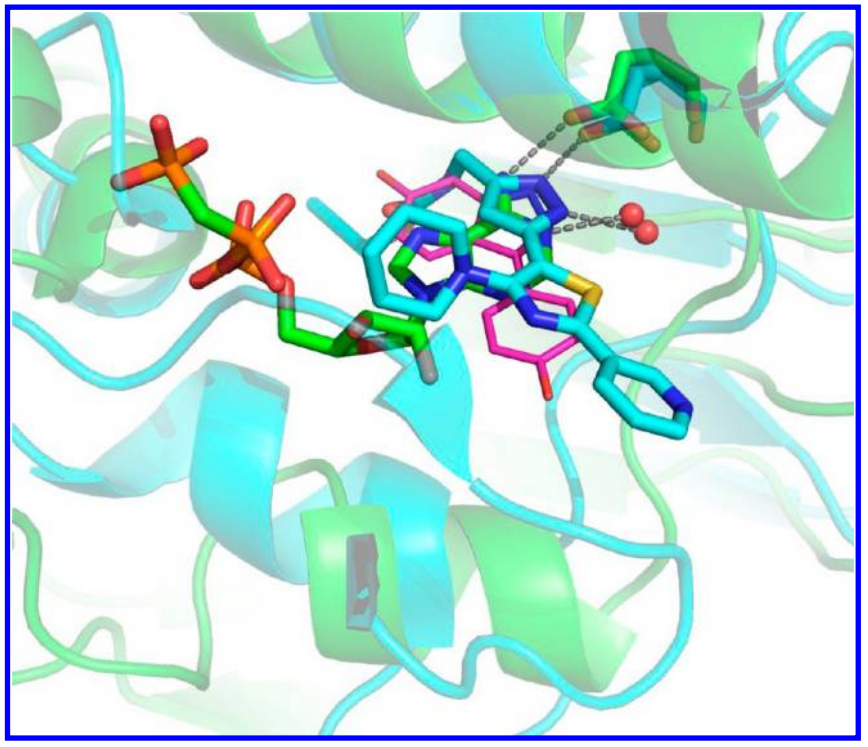
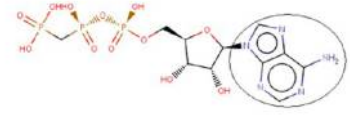
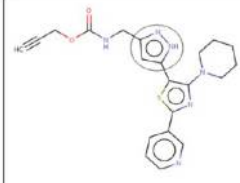
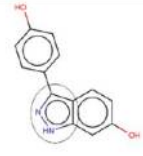


Figure 5. The query 3t10 (green) aligned with SubCav to 3g7e (cyan). HSP90 with indazole-containing ligand from 4eeh is shown in purple (protein structure omitted for clarity). The aspartate interacting residues from HSP90_HUMAN (Asp93) and GYRB_ECOLI (Asp73) are shown as sticks. There is a conserved water (red sphere) also forming interaction with the fragments.

Table 5. Three Protein Structures Shown in Figure 5[a]

Protein (UNIPROT)	PDB ID	Ligand ID	2D Structure of the Ligand
HSP90_HUMAN	3t10	ACP	
GYRB_ECOLI	3g7e	B46	
HSP90_HUMAN	4eeh	HH6	

[a]The matching fragments are highlighted on ligand structures.

FRAGPDB was carried out using the adenine portion of ACP bound to HSP90 (PDB code 3t10). The PDB structure 3t10 was recently published and did not form part of the LigandExpo database used to build FRAGPDB. A subpocket query was manually defined with PyMol's graphical user interface[53] to include all protein atoms at a 5 Å distance from any atoms of the adenine fragment. The longer query distance was selected as it is one of the presets in PyMol.

An O_{sc} value of greater than 0.50 was used to select the most promising cases for further inspection. By using the UNIPROT annotation, all the subpockets corresponding to other HSP90 PDB structures present in FRAGPDB were removed, which were correctly top ranked based on O_{sc}. The subpockets corresponding to proteins with the same SCOP family name (i.e., d.122.1.1) and with a sequence similarity higher than 30% were discarded, and the remaining 1416 subpocket pairs were then manually checked. Some of the retrieved PDBs are highlighted in Table 4. Despite the low sequence similarity other proteins belonging to the ATPase/kinase superfamily were identified, such as the pyruvate and alpha-ketoacid

dx.doi.org/10.1021/ci300523r | J. Chem. Inf. Model. 2013, 53, 131−141

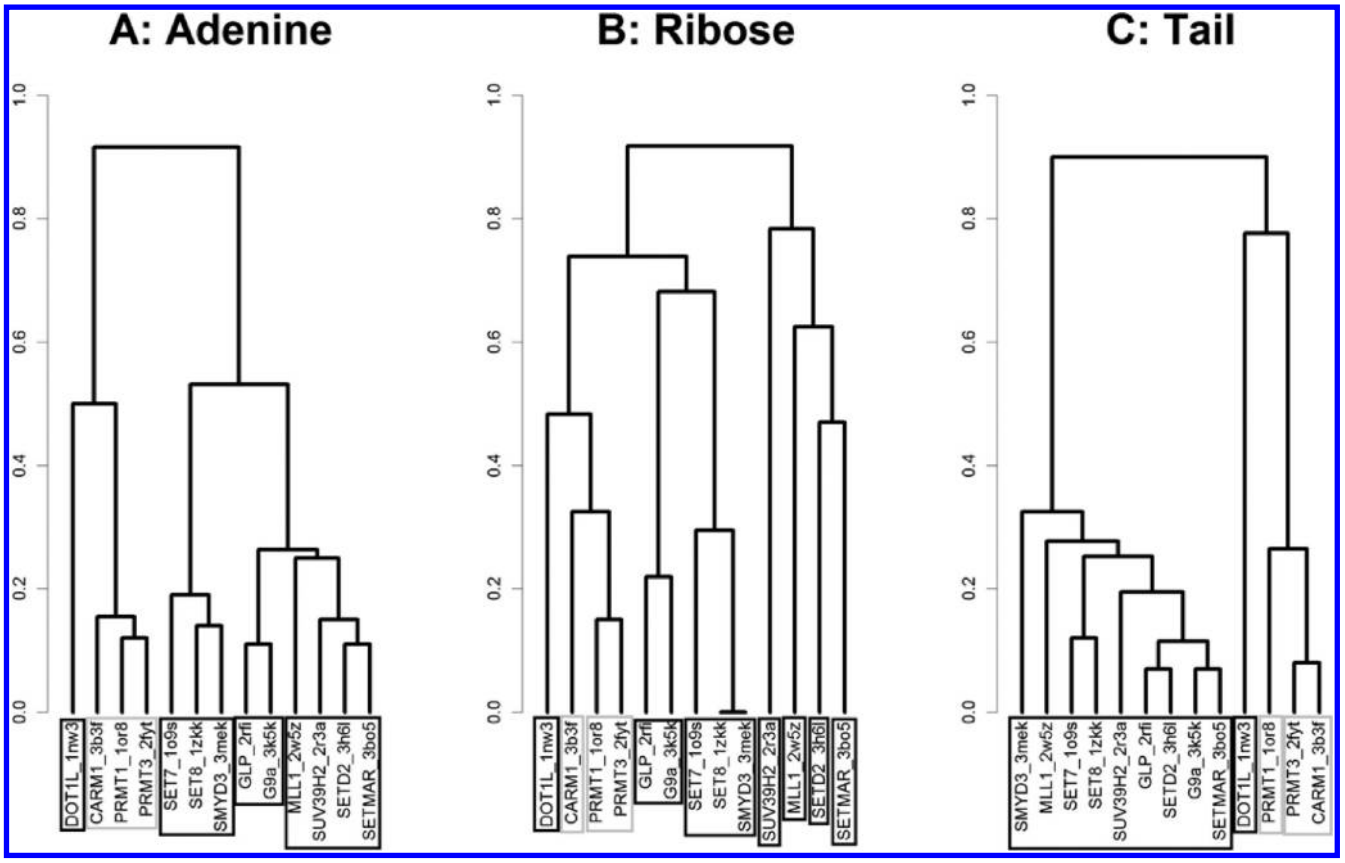


Figure 6. A: Dendrogram obtained by clustering the subpockets proteins reported in Table 2 around the adenine fragment (shown in purple in Figure 3); B: Dendrogram obtained by clustering the subpockets around the ribose fragment (shown in cyan in Figure 3); C: Dendrogram obtained by clustering the subpockets around the tail fragment (shown in green in Figure 3). The black boxes correspond to the clustering of the lysine methyl-transferases, and the gray boxes correspond to the clustering of the arginine methyl-transferases.

dehydrogenase kinase (entries 1−5 of Table 4) and DNA topoisomerase gyrase B (entry 6) as well as various topoisomerases (entries 7−9), histidine kinases (entries 10−11), and the DNA repair enzyme MutL-like proteins (entries 9−12). It is worth pointing out that the interactions with the ribose and the phosphates are not conserved among protein histidine kinases and the other ATPases.[54] The cytosolic paralog of HSP90 (entry 16) was also found. Good overlap was also observed with the nucleotide-binding site of the anti-σ and serine kinase spoIIAB bound to ADP and to the anti-σ spoIIAA protein (entry 18), with an extremely low sequence similarity of 4.1%.

The 3D overlay of the HSP90 query subpocket with that of entry 6 of Table 4 is taken as an example to explain the potential of how the method could be used to suggest new fragments for the query target of interest. Figure 5 reports the overlay produced by SubCav: the 3t10 query protein is shown in green and the matched 3g7e protein in cyan. The whole protein is shown for clarity. The adenine of 3t10 and the pyrazole of 3g7e have similar interactions in the two cavities. The hydrogen bonds with an aspartate residue (Asp93 in 3t10, and Asp73 in 3g7e) are shown with dashed gray lines. The similar interaction of the N6 nitrogen of adenine and pyrazole fragment with a water molecule (in ball in Figure 5) is also shown (note that water molecules were not considered in the description of the subpockets). This finding suggests the possibility of using a pyrazole as bioisosteric replacement for the adenine. As a consequence, various fragments containing this structural motif could be screened or molecules could be designed starting from the pyrazole fragment bound to a protein with a similar subpocket. To show that this SubCav finding was reasonable, ligands bound to HSP90 in the PDB containing the pyrazole substructure were examined. Recently, a crystal structure (PDB ID: 4eeh[55]) of HSP90 cocrystallized with an indazole ligand (Ligand ID: HH6) was released (Table 5). The 4eeh protein was aligned based on the sequence using the align function of PyMol to 3t10. The ligand HH6 (see Table 5) is colored in purple in Figure 5. The perfect match of the two pyrazole scaffolds highlights the value of subpocket similarity based suggestions for design/screening purposes.

3.3. Analysis of Histone Methyl-Transferase Binding Sites. Subpockets can also be used to probe enzyme specificity by analyzing the diverse intrafamily proteins using in silico methods. In this respect, protein kinases, phosphatases, and phosphodiesterases represent a well-known example of an enzyme superfamily which has been heavily investigated with such approaches.[56−59] The recent crystal structure determination of GPCR proteins has also opened binding sites comparison to this family. Here we will use subpocket alignments to analyze a set of methyl transferases, which constitute an epigenetic target class.

Epigenetics is an emerging field that can provide promising targets for drug discovery. Several epigenetic medicinal chemistry projects are currently ongoing both in academia and in pharmaceutical companies. Methyl transferases represent one of the epigenetic target classes of great interest with potential in various disease areas. This family binds the SAM cofactor which methylates the arginine or lysine residues of the

histone in various positions. In this section, the use of SubCav is described for more systematically analyzing the local structural similarity of the cofactor site of this family. Three dendrograms from the hierarchical clustering of the dissimilarity matrix generated by calculating the pairwise O_{sc} values of all the selected proteins (see Table 2) are reported in Figure 6A–C. By taking the value 0.3 for dissimilarity (see 2.3) based on the visual inspection as the level for clustering, a total of 5, 9, and 4 clusters were obtained for the adenine, ribose, and tail subpockets, respectively.

The clustering of the cofactor binding site by subpockets around each specific fragment revealed different levels of local similarity within the selected proteins set. The four adenine subpocket typologies of binding for the lysine methyl transferases (KMT) are shown in Figure 7. The KMT

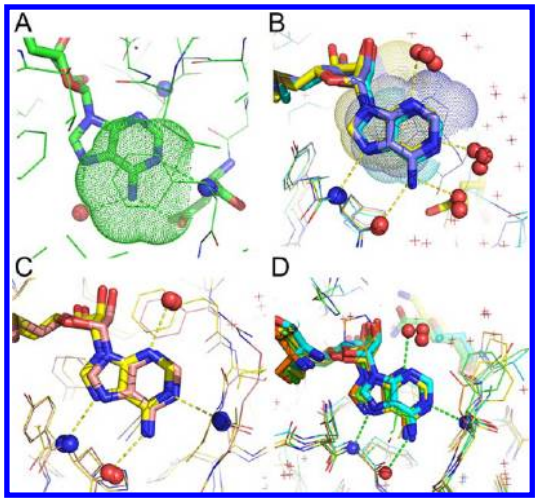


Figure 7. Close-up of adenine subpockets in the crystal structures of lysine methyl transferases of Table 3 corresponding to the four clusters of Figure 6A. A: DOT1L (PDB 1nw3,[60] green); B: SET7 (PDB 1o9s,[61] yellow), SET8 (PDB 1zkk,[62] violet), SMYD3 (PDB 3mek, cyan); C: GLP (PDB 2rfi,[63] pink), G9a (PDB 3k5k,[64] cyan); D: SETD2 (PDB 3h6l, green), SUV39H2 (PDB 2r3a,[65] cyan), SETMAR (PDB 3bo5, yellow), MLL1 (PDB 2w5z,[66] orange).

DOTL1 SAM pocket is very distinct from other KMT and much more similar to RMT. This finding is consistent with the classification using global pocket clustering.[42] The N6 nitrogen (see Figure 3 for nomenclature) of adenine of the SAM molecule bound to DOT1L makes two hydrogen bonds: with Asp222 residue and with a water molecule, which is also close to N7. The N1 nitrogen makes one hydrogen bond with the NH of the backbone of Phe223.The side chain of this residue lies above the aromatic core of adenine (highlighted with dots in Figure 7A). Similar hydrogen bond interactions involving N6 are observed in RMT (with Glu129 in PRMT1, with Glu331 in PRMT3, and with Glu224 in CARM1, data not shown). The adjacent residues involved in making hydrogen bond with N1 are instead aliphatic (Val128 in PRMT1, Ile330 in PRMT3 and Val243 in CARM1). The $\pi-\pi$ stacking between Phe223 side-chain and the adenine core is thus not present in the analyzed RMT. The N3 nitrogen of adenine bound to DOT1L and all the RMT is close to a NH of the backbone (see the blue ball in Figure 7A corresponding to Lys187 in DOT1L).

The SET7, SET8, and SMYD3 cluster shares the two hydrogen bonds made by the N6 and N7 nitrogen atoms of the adenine core with the backbone. In addition SET7 makes a hydrogen bond to Glu356 (in stick in Figure 7B). This Glu356-N6 interaction is replaced by water molecules in SET8 and SMYD3 complexes. This difference results in being SET8 and SMYD3 more similar to each other than to SET7 (see Figure 6A). The water molecules are reported in the picture but are not considered in the description of the subpockets (see Material and Methods section). The residue below the adenine plane (highlighted as cloud of dots) is aromatic in all three complexes and is highly involved in $\pi-\pi$ stacking (Trp352, Trp349, and Phe259 in SET7, SET8, and SMYD3, respectively). The N1 nitrogen of adenine is involved in water hydrogen bond in all three complexes. This water is replaced by a NH of the backbone in all the other KMT (see Figure 7A, C, D).

The other two adenine subpocket based clusters differ from each other by the type of residue below and above the aromatic adenine core. GLP and G9a still have an aromatic residue underneath the adenine core (Trp1216 and Trp1158, respectively, not shown in Figure 7C) and locate a methionine residue above the adenine ring (Met1105 and Met1048, respectively). The bigger cluster, made by MLL1, SUV39H2, SETD2, and SETMAR (see Figure 6A and Figure 7D), shares similar hydrogen bonding interactions but have an aliphatic residue underneath the adenine core (Leu3968, Leu298, Leu1689, and Leu284, respectively) and a positive charged residue for the three most similar one (Lys1560 in SETD2, Arg150 in SUV39H2, and Lys135 in SETMAR). This residue corresponds to Ile3838 in MLL1.

The tail subcavities are more similar among each other in the KMT as they are grouped in one single cluster, see Figure 6C. The ribose subcavities are instead more diverse within the KMT set. The results of this analysis could be exploited to design subfamily adenine, ribose, and tail fragment mimetics.

3.4. Current Limitations and Future Directions. Only structures with bound ligands in PDB were used in this study, which means that a major part of the PDB was excluded from the study. Defining subpockets from apo-structures would be a natural extension to this work. This would require software for subpocket detection and generation like the DoGSite-method developed by Volkamer and co-workers.[67]

4. CONCLUSIONS

The same fragments can bind to very different subpockets. However, the identification of the pairs of subpockets which host identical fragments provided a large data set against which a method aiming at finding local subpocket similarity could be validated. The SubCav method developed here was successful in aligning most of these pairs, highlighting its potential for fragment-based drug design. The potential use of SubCav in FBDD was shown using HSP90 as a case study, where the method appears to be efficient in retrieving suitable fragments for screening out of PDB-wide subpocket database searches. Moreover, SubCav has also been proven as an informative tool for detailed analysis of intrafamily subpockets with the help of clustering algorithms.

■ ASSOCIATED CONTENT

Supporting Information

Figure S1, an example of failed fragment-based subpocket alignment; **Figure S2**, the retrieval performance of 2D similarity metrics; and PyMol sessions of Figure 4. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Phone: +41 61 3241016. E-mail: anna.vulpetti@novartis.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

T.K. thanks the Novartis Institutes for BioMedical Research for Presidential Postdoctoral Fellowship. Dr. Guido Kirsten (Chemical Computing Group) is acknowledged for his suggestions regarding the 3D alignment protocol.

REFERENCES

(1) Erlanson, D. A. Introduction to fragment-based drug discovery. *Top. Curr. Chem.* **2012**, *317*, 1−32.

(2) Henen, M. A.; Coudevylle, N.; Geist, L.; Konrat, R. Towards rational fragment-based lead design without 3D structures. *J. Med. Chem.* **2012**, *55*, 7909−7919.

(3) Zuegg, J.; Cooper, M. A. Drug-likeness and increased hydrophobicity of commercially available compound libraries for drug screening. *Curr. Top. Med. Chem.* **2012**, *12*, 1500−1513.

(4) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A 'rule of three' for fragment-based lead discovery? *Drug Discovery Today* **2003**, *8*, 876−877.

(5) 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.* **2010**, *28*, 235−242.

(6) Chan, A. W.; Laskowski, R. A.; Selwood, D. L. Chemical fragments that hydrogen bond to Asp, Glu, Arg, and His side chains in protein binding sites. *J. Med. Chem.* **2010**, *53*, 3086−3094.

(7) Wang, L.; Xie, Z.; Wipf, P.; Xie, X. Q. Residue preference mapping of ligand fragments in the Protein Data Bank. *J. Chem. Inf. Model.* **2011**, *51*, 807−815.

(8) Weber, A.; Casini, A.; Heine, A.; Kuhn, D.; Supuran, C. T.; Scozzafava, A.; Klebe, G. Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. *J. Med. Chem.* **2004**, *47*, 550−557.

(9) Kellenberger, E.; Schalon, C.; Rognan, D. How to measure the similarity between protein ligand-binding sites? *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 209−220.

(10) Henrich, S.; Salo-Ahen, O. M. H.; Huang, B.; Rippmann, F.; Cruciani, G.; Wade, R. C. Computational approaches to identifying and characterizing protein binding sites for ligand design. *J. Mol. Recognit.* **2010**, *23*, 209−219.

(11) Nisius, B.; Sha, F.; Gohlke, H. Structure-based computational analysis of protein binding sites for function and druggability prediction. *J. Biotechnol.* **2012**, *159*, 123−134.

(12) Madala, P. K.; Fairlie, D. P.; Bodén, M. Matching cavities in G protein-coupled receptors to infer ligand-binding sites. *J. Chem. Inf. Model.* **2012**, *52*, 1401−1410.

(13) Reisen, F.; Weisel, M.; Kriegl, J. M.; Schneider, G. Self-organizing fuzzy graphs for structure-based comparison of protein pockets. *J. Proteome Res.* **2010**, *9*, 6498−6510.

(14) Vulpetti, A.; Kalliokoski, T.; Milletti, F. Chemogenomics in drug discovery: computational methods based on the comparison of binding sites. *Future Med. Chem.* **2012**, *4*, 1971−1979.

(15) Moriaud, F.; Doppelt-Azeroual, O.; Martin, L.; Oguievetskaia, K.; Koch, K.; Vorotyntsev, A.; Adcock, S. A.; Delfaud, F. Computational fragment-based approach at PDB scale by protein local similarity. *J. Chem. Inf. Model.* **2009**, *49*, 280−294.

(16) Wallach, I.; Lilien, R. H. Prediction of sub-cavity binding preferences using an adaptive physicochemical structure representation. *Bioinformatics* **2009**, *25*, i296−304.

(17) Durrant, J. D.; Friedman, A. J.; McCammon, J. A. CrystalDock: a novel approach to fragment-based drug design. *J. Chem. Inf. Model.* **2011**, *51*, 2573−2580.

(18) Weisel, M.; Bitter, H. M.; Diederich, F.; So, W. V.; Kondru, R. PROLIX: rapid mining of protein-ligand interactions in large crystal structure databases. *J. Chem. Inf. Model.* **2012**, *52*, 1450−1461.

(19) Vulpetti, A.; Schiering, N.; Dalvit, C. Combined use of computational chemistry, NMR screening, and X-ray crystallography for identification and characterization of fluorophilic protein environments. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 3281−3291.

(20) Wood, D. J.; Vlieg, J. D.; Wagener, M.; Ritschel, T. Pharmacophore fingerprint-based approach to binding site subpocket similarity and its application to bioisostere replacement. *J. Chem. Inf. Model.* **2012**, *52*, 2031−2043.

(21) Feldman, H. J.; Labute, P. Pocket similarity: are alpha carbons enough? *J. Chem. Inf. Model.* **2010**, *50*, 1466−1475.

(22) Müller, K.; Faeh, C.; Diederich, F. Fluorine in pharmaceuticals: looking beyond intuition. *Science* **2007**, *317*, 1881−1886.

(23) Ito, J.; Tabei, Y.; Shimizu, K.; Tomii, K.; Tsuda, K. PDB-scale analysis of known and putative ligand-binding sites with structural sketches. *Proteins: Struct., Funct., Bioinf.* **2012**, *80*, 747−763.

(24) Kabsch, W. A solution of the best rotation to relate two sets of vectors. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1976**, *32*, 922−923.

(25) Cock, P. J.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422−1443.

(26) Needleman, S. B.; Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **1970**, *48*, 443−453.

(27) Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16*, 276−277.

(28) UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acid Res.* **2012**, *40*, D71−D75.

(29) Murzin, A. G.; Brenner, S. E.; Hubbard, T. J. P.; Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **1995**, *247*, 536−540.

(30) Andreeva, A.; Howorth, D.; Chandonia, J. M.; Brenner, S. E.; Hubbard, T. J. P.; Chothia, C.; Murzin, A. G. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.* **2008**, *36*, D419−D425.

(31) Punta, M.; Coggill, P. C.; Eberhardt, R. Y.; Mistry, J.; Tate, J.; Boursnell, C.; Pang, N.; Forslund, K.; Ceric, G.; Clements, J.; Heger, A.; Holm, L.; Sonnhammer, E. L.; Eddy, S. R.; Bateman, A.; Finn, R. D. The Pfam protein families database. *Nucleic Acids Res.* **2012**, *40*, D290−D301.

(32) RDKit: Open-source cheminformatics. http://www.rdkit.org.

(33) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.

(34) The Open Babel Package, Version 2.3.1. http://openbabel.org.

(35) Verdonk, M. L.; Giangreco, I.; Hall, R. J.; Korb, O.; Mortenson, P. N.; Murray, C. W. Docking performance of fragments and druglike compounds. *J. Med. Chem.* **2011**, *54*, 5422−5431.

(36) Lounkine, E.; Batista, J.; Bajorath, J. Random molecular fragment methods in computational medicinal chemistry. *Curr. Med. Chem.* **2008**, *15*, 2108−2121.

(37) 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*, 511−522.

(38) Degen, J.; Wegscheid-Gerlach, C.; Zalian, A.; Rarey, M. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem* **2008**, *3*, 1503−1507.

(39) Wildman, S. A.; Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.
(40) Wallach, I.; Lilien, R. The protein-small-molecule database, a non-redundant structural resource for the analysis of protein-ligand binding. *Bioinformatics* **2009**, *25*, 615−620.
(41) Feng, Z.; Chen, L.; Maddula, H.; Akcan, O.; Oughtred, R.; Berman, H. M.; Westbrook, J. Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics* **2004**, *20*, 2153−2155.
(42) Campagna-Slater, V.; Mok, M. W.; Nguyen, K. T.; Feher, M.; Najmanovich, R.; Schapira, M. Structural chemistry of the histone methyltransferases cofactor binding site. *J. Chem. Inf. Model.* **2011**, *51*, 612−623.
(43) R Core Team. R: A Language and Environment for Statistical Computing. 2012. http://www.R-project.org.
(44) Weill, N.; Rognan, D. Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. *J. Chem. Inf. Model.* **2010**, *50*, 123−35.
(45) Denessiouk, K. A.; Rantanen, V. V.; Johnson, M. S. Adenine recognition: a motif present in ATP-, CoA-, NAD-, NADP-, and FAD-dependent proteins. *Proteins: Struct., Funct., Bioinf.* **2001**, *44*, 282−291.
(46) Stegemann, B.; Klebe, G. Cofactor-binding sites in proteins of deviating sequence: Comparative analysis and clustering in torsion angle, cavity, and fold space. *Proteins: Struct., Funct., Bioinf.* **2012**, *80*, 626−648.
(47) Zaitseva, J.; Oswald, C.; Jumpertz, T.; Jenewein, S.; Wiedenmann, A.; Holland, I. B.; Schmitt, L. A structural analysis of asymmetry required for catalytic activity of an ABC-ATPase domain dimer. *EMBO J.* **2006**, *25*, 3432−3443.
(48) Bauer, C. B.; Kuhlman, P. A.; Bagshaw, C. R.; Rayment, I. X-ray crystal structure and solution fluorescence characterization of Mg.2′(3′)-O-(N-methylanthraniloyl) nucleotides bound to the Dictyostelium discoideum myosin motor domain. *J. Mol. Biol.* **1997**, *247*, 394−407.
(49) Venkataramaiah, T. H.; Plapp, B. V. Formamides mimic aldehydes and inhibit liver alcohol dehydrogenases and ethanol metabolism. *J. Biol. Chem.* **2003**, *278*, 36699−36706.
(50) Mörtl, M.; Diederichs, K.; Welte, W.; Molla, G.; Motteran, L.; Andriolo, G.; Pilone, M. S.; Pollegioni, L. Structure-function correlation in glycine oxidase from Bacillus subtilis. *J. Biol. Chem.* **2004**, *279*, 29718−29727.
(51) Schweikhard, E. S.; Kuhlmann, S. I.; Kunte, H. J.; Grammann, K.; Ziegler, C. M. Structure and function of the universal stress protein TeaD and its role in regulating the ectoine transporter TeaABC of Halomonas elongata DSM 2581(T). *Biochemistry* **2010**, *49*, 2194−2204.
(52) Lim, K.; Zhang, H.; Tempczyk, A.; Krajewski, W.; Bonander, N.; Toedt, J.; Howard, A.; Eisenstein, E.; Herzberg, O. Structure of the YibK methyltransferase from Haemophilus influenzae (HI0766): a cofactor bound at a site formed by a knot. *Proteins: Struct., Funct., Bioinf.* **2003**, *51*, 56−67.
(53) The PyMOL Molecular Graphics System, Version 1.2r3pre; Schrödinger, LLC.
(54) Bilwes, A. M.; Quezada, C. M.; Croal, L. R.; Crane, B. R.; Simon, M. I. Nucleotide binding by the histidine kinase CheA. *Nat. Struct. Biol.* **2001**, *8*, 353−360.
(55) Buchstaller, H. P.; Eggenweiler, H. M.; Sirrenberg, C.; Grädler, U.; Musil, D.; Hoppe, E.; Zimmermann, A.; Schwartz, H.; März, J.; Bomke, J.; Wegener, A.; Wolf, M. Fragment-based discovery of hydroxy-indazole-carboxamides as novel small molecule inhibitors of Hsp90. *Bioorg. Med. Chem. Lett.* **2012**, *22*, 4396−4403.
(56) Milletti, F.; Vulpetti, A. Predicting polypharmacology by binding site similarity: from kinases to the protein universe. *J. Chem. Inf. Model.* **2010**, *50*, 1418−1431.
(57) Kalliokoski, T.; Vulpetti, A. Large-scale evaluation of CavBase for analyzing the polypharmacology of kinase inhibitors. *Mol. Inf.* **2011**, *30*, 923−925.
(58) Kuhn, D.; Weskamp, N.; Hullermeier, E.; Klebe, G. Functional classification of protein kinase binding sites using CavBase. *ChemMedChem* **2007**, *2*, 1432−1447.
(59) Kinnings, S. L.; Jackson, R. M. Binding site similarity analysis for the functional classification of the protein kinase family. *J. Chem. Inf. Model.* **2009**, *49*, 318−329.
(60) Min, J.; Feng, Q.; Li, Z.; Zhang, Y.; Xu, R. M. Structure of the catalytic domain of human DOT1L, a non-SET domain nucleosomal histone methyltransferase. *Cell* **2003**, *112*, 711−723.
(61) Xiao, B.; Jing, C.; Wilson, J. R.; Walker, P. A.; Vasisht, N.; Kelly, G.; Howell, S.; Taylor, I. A.; Blackburn, G. M.; Gamblin, S. J. Structure and catalytic mechanism of the human histone methyltransferase SET7/9. *Nature* **2003**, *421*, 652−656.
(62) Couture, J. F.; Collazo, E.; Brunzelle, J. S.; Trievel, R. C. Structural and functional analysis of SET8, a histone H4 Lys-20 methyltransferase. *Genes Dev.* **2005**, *19*, 1455−65.
(63) Wu, H.; Min, J.; Lunin, V. V.; Antoshenko, T.; Dombrovski, L.; Zeng, H.; Allali-Hassani, A.; Campagna-Slater, V.; Vedadi, M.; Arrowsmith, C. H.; Plotnikov, A. N.; Schapira, M. Structural biology of human H3K9 methyltransferases. *PLoS One* **2010**, *5*, e8570.
(64) Liu, F.; Chen, X.; Allali-Hassani, A.; Quinn, A. M.; Wasney, G. A.; Dong, A.; Barsyte, D.; Kozieradzki, I.; Senisterra, G.; Chau, I.; Siarheyeva, A.; Kireev, D. B.; Jadhav, A.; Herold, J. M.; Frye, S. V.; Arrowsmith, C. H.; Brown, P. J.; Simeonov, A.; Vedadi, M.; Jin, J. Discovery of a 2,4-diamino-7-aminoalkoxyquinazoline as a potent and selective inhibitor of histone lysine methyltransferase G9a. *J. Med. Chem.* **2009**, *52*, 7950−7953.
(65) Wu, H.; Min, J.; Lunin, V. V.; Antoshenko, T.; Dombrovski, L.; Zeng, H.; Allali-Hassani, A.; Campagna-Slater, V.; Vedadi, M.; Arrowsmith, C. H.; Plotnikov, A. N.; Schapira, M. Structural biology of human H3K9 methyltransferases. *PLoS One* **2010**, *5*, e8570.
(66) Southall, S. M.; Wong, P. S.; Odho, Z.; Roe, S. M.; Wilson, J. R. Structural basis for the requirement of additional factors for MLL1 SET domain activity and recognition of epigenetic marks. *Mol. Cell* **2009**, *33*, 181−191.
(67) Volkamer, A.; Griewel, A.; Grombacher, T.; Rarey, M. Analyzing the topology of active sites: on the prediction of pockets and subpockets. *J. Chem. Inf. Model.* **2010**, *50*, 2041−2052.