



# Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery

Stéphanie Pérot<sup>1</sup>, Olivier Sperandio<sup>1,2</sup>, Maria A. Miteva<sup>1</sup>, Anne-Claude Camproux<sup>1</sup> and Bruno O. Villoutreix<sup>1,2</sup>

<sup>1</sup>Inserm UMR-S 973, Molécules Thérapeutiques *in silico*, Université Paris Diderot, 35 rue Hélène Brion, 75013 Paris, France

<sup>2</sup>CDithem Platform, IGM, 27 rue Juliette Dodu, 75010 Paris, France

Detection, comparison and analyses of binding pockets are pivotal to structure-based drug design endeavors, from hit identification, screening of exosites and de-orphanization of protein functions to the anticipation of specific and non-specific binding to off- and anti-targets. Here, we analyze protein–ligand complexes and discuss methods that assist binding site identification, prediction of druggability and binding site comparison. The full potential of pockets is yet to be harnessed, and we envision that better understanding of the pocket space will have far-reaching implications in the field of drug discovery, such as the design of pocket-specific compound libraries and scoring functions.

The growing track record of small molecule leads derived from structure-based drug design (SBDD), together with the increasing number of experimental macromolecular structures [1–3] and high-quality homology models reported at unprecedented speed [4], prompt drug designers to develop and refine concepts about ligand-binding sites through different levels of abstraction. In the 1990s, binding pockets were essentially seen as concavities with some shape and chemical complementarity with the ligands, whereas today, new concepts about ligand-binding sites are emerging, from the notions of ‘druggable’ pockets [5–8] and binding site centric chemical space [9] to the observation of conserved binding sites across different target classes [10,11].

Binding pockets are today one of the cornerstones of modern drug discovery projects and at the crossroad of several research fields, from structural biology to mathematical modeling [12–17]. This increased awareness stresses that pockets are the actual drug targets and as such, complement the notion of protein families and superfamilies [18,19]. Medium-to-large-scale surveys of high-quality protein–ligand complexes through the ‘lens of the pocket concept’ could have far-reaching implications for drug discovery scientists because they support the idea that it is possible to develop pocket-dependent chemical libraries and scoring functions. Antici-

pating similarities between binding pockets should also facilitate the understanding of compound binding to off- and anti-targets and help compound profiling and/or de-orphanization of protein cavities and/or functions. Enhanced understanding of binding pockets at the protein–protein or protein–membrane interfaces will also be of major importance to address these difficult target classes [13,20–24]. For instance, Fuller *et al.* [13] have reported recently that pockets at the protein–protein interface are statistically significant different from those present at protein–ligand interfaces. The importance of studying binding pockets is further highlighted in a recent research article entitled ‘Form follows function’ [25]. Although it is still not possible to answer this long-lasting question (i.e. which is first, form or function?), that study will certainly boost research in the field. Here, we review several key studies dealing with binding pockets, pocket detection methods and concepts about druggable pockets. We also discuss pocket similarity search engines and other recent trends in the field. We end this review with a case study, in which we analyze high-resolution pocket–ligand complexes extracted from the Astex data set to illustrate some of the points discussed in the other paragraphs [26].

## Anatomy of protein pockets and associated ligands

A major issue that hampers the analysis of binding pockets is the lack of a standard definition of what constitutes a pocket. Geometric descriptions of the depth or size and, therefore, volume,

Corresponding author: Villoutreix, B.O. (bruno.villoutreix@inserm.fr), (bruno.villoutreix@univ-paris-diderot.fr)

amino acid composition, and so on of binding pockets are method-dependent and, to some extent, subjective [13]. Ligand-binding sites vary widely in size and shape; they can be nearly spherical or form a curved groove composed of several interconnected subpockets [27]. Catalytic sites usually occur in large and deep clefts on the protein surface, and drugs have been observed to essentially bind into the largest surface cavities [27–29]. Overall, it has been found that larger cavities tend to be less spherical [30]. Pockets' volume is known to be of major importance for predicting binding sites [31]. The average volume of a drug-binding cavity was found to be around 930 Å<sup>3</sup> using a geometry-based method [28], whereas in another report using an energy-based approach, the average envelope volume enclosing pockets was found to be approximately 610 Å<sup>3</sup> [32]. These values cannot be compared directly because volumes are computed differently and because there are still no gold standards in the field. From the ligand side, the average volume of drug-like compounds was found to be 439 Å<sup>3</sup> and most often above 200 Å<sup>3</sup> [32]. There is, in general, a good correlation between the ligand volume and the binding site volume when the pocket volume is less than 700 Å<sup>3</sup> [27], and this is even more pronounced when the pockets tend to be closed.

Shape and chemical complementarity are the underlying bases of molecular recognition. Comparisons between high- and low-affinity enzymatic and non-enzymatic complexes have also been performed, for instance using the Binding Mother of All Databases [33]. The results showed that enzymes' high-affinity ligands tend to be much larger than those with low affinity, indicating that the addition of complementary functional groups is likely to improve the affinity of an enzyme inhibitor; however, this process might not be as fruitful for nonenzymes' ligands. High- and low-affinity ligands of nonenzymes are nearly the same size, so modest modifications and isosteric replacement might be most productive. Nonenzymes were found to have greater ligand efficiencies than enzymes, and these differences seem to come not from the ligands but from the pockets [34]. Interestingly, protein–protein interaction inhibitors in pockets located at the interface tend to have ligand efficiencies similar to those computed for protease inhibitors, suggesting that at least some members of this target class can be drugged [21]. Geometrical complementarity in general is not sufficient to fully drive molecular recognition. Nevertheless, it has been shown – when considering only shape and size – that a significant proportion of the recognition 'power' of a binding pocket for its ligand resides in its shape. In addition, a 'buffer zone', or a region of free space between the ligand and protein, is often observed, which results in binding pockets being, on average, three times larger than the ligand they bind [16]. This is somewhat expected because in many cases, small drug ligands bind to catalytic sites, regions that usually accommodate much larger peptide or protein substrates. With regard to SBDD methods, it has been noted that depending on the shape and nature of the binding pockets (and, in some cases, on the nature of the ligands), some docking and scoring approaches seem to be more efficient than others [35,36].

### Protein pocket and druggability predictions

There are issues associated with finding pockets because they can change shape considerably upon (or before) ligand binding.

Assuming a somewhat preformed cavity, pocket detection methods can be applied using as input the 3D structure of the receptor. These methods can be divided into two major categories: geometric algorithms (e.g. SURFNET, LIGSITE, PocketDepth and PocketPicker [37–40]) and probe and/or energy-based methods (e.g. GRID, CS-Map, QSiteFinder, AutoLigand [41–43] and ICM-PocketFinder [32]), although some approaches can use several principles. Furthermore, two other methods can be considered: evolutionary methods (structure and/or sequence alignment) can be applied, but they tend to be used for identifying protein–protein interaction sites and docking methods that can be used to predict drug-like molecule binding sites (Table 1).

Geometric pocket detection algorithms cover a variety of techniques, from the fitting of virtual spheres into the solvent-accessible space between protein atoms to the use of Delaunay triangulation or of the alpha-shapes approach to delineate cavities (for a recent review, see Ref. [14]). Several techniques have been developed for estimating the interaction energy between a probe molecule (e.g. methyl, hydroxyl or amine groups) at a given point and a protein (probe-mapping algorithms) [44]. These tools incorporate some levels of protein physics into the pocket identification process. For example, AutoLigand [43] uses a grid-based representation of the binding affinity potential to define envelopes of maximal affinity. Affinity potentials are generated for six atom types (aliphatic carbon, aromatic carbon, hydrogen, oxygen, nitrogen and sulfur), and the best envelope within the energy grid is calculated using a three-step process of flood fill, local migration and ray-casting neighborhood search. The energy-based method, called QSiteFinder [42], describes zones where methyl probes can interact favorably at the protein surface. The computational solvent mapping (CS-Map) algorithm [45] can be used for the identification of hot spots (i.e. regions of protein binding sites that are major contributors to the binding energy and, hence, are prime targets in drug design) [46]. The ICM-PocketFinder method, an energy-based method, performs a Gaussian convolution of the Lennard–Jones potential around a protein [32]. The resulting field calculated as a 3D grid map is contoured to produce envelopes, the location, shape and volume of which are indicative of the ligand-binding pockets. Like pure geometric approaches, the package is fast and capable of identifying clefts and cavities regardless of the nature of the substrate [15].

Finding pocket cavities is a first step in predicting protein druggability, and the next step attempts to define whether the predicted pocket can bind a drug-like molecule. Research groups have thus been developing tools that compute a 'druggability index' [31]. Although some of the tools mentioned above score pockets and in some ways can be considered to predict druggability, we report below some recent methods dedicated to druggability prediction. For instance, Soga *et al.* [47] have developed a method on the basis of the specific amino acid composition observed at the ligand-binding sites of ligand–protein complexes determined by X-ray analysis (the Alpha Site Finder implemented in the software system MOE was used to detect concavities on the surface of the proteins). A profile representing the preference of each of the 20 standard amino acids at the binding sites of drug-like molecules was obtained, and an index termed 'propensity for ligand binding' was created from these profiles. The authors found particularly interesting that the binding sites could be predicted

TABLE 1

**Main algorithms to search for binding pockets.**

<i>Method type</i>	<i>Name</i>	<i>Refs.</i>	<i>Search method</i>	<i>URL</i>
<b>Geometric and genomic</b>	LigSite <sup>csc</sup> (csc = Conolly Surface and Conservation)	[87]	Optimized version of LigSite: re-ranking of predicted pockets by the degree of conservation of the closest surface residues	<a href="http://projects.biotech.tu-dresden.de/cgi-bin/index.php">http://projects.biotech.tu-dresden.de/cgi-bin/index.php</a>
<b>Geometric and genomic</b>	SURFNET-ConSurf	[88]	Optimized version of SURFNET: re-ranking to further increase prediction accuracy with conservation scores	<a href="http://www.biochem.ucl.ac.uk/~roman/">http://www.biochem.ucl.ac.uk/~roman/</a>
<b>Geometric</b>	APROPOS (Automatic PROtein Pocket Search)	[89]	Identification of pockets using the alpha-shape principles	<a href="http://www.csb.yale.edu/userguides/datamanip/apropos/manual.html">http://www.csb.yale.edu/userguides/datamanip/apropos/manual.html</a>
<b>Geometric</b>	Binding-response	[90]	Putative binding regions are defined with the sphere-based method of DOCK; re-clustering of spheres is done with a clustering algorithm implemented in CHARMM. Development of a novel descriptor, the binding response, to quantitatively evaluate putative binding sites	<a href="http://mackerell.umaryland.edu/CADD/CADD_bindingresponse.html">http://mackerell.umaryland.edu/CADD/CADD_bindingresponse.html</a>
<b>Geometric</b>	CAST (Computed Atlas of Surface Topography)	[27]	Identification and measurement of surface accessible pockets using the weighted-Delaunay triangulation and the alpha-shape principles	
<b>Geometric</b>	CASTp	[91]	Server for CAST incorporating some new utilities as compared to the original version	<a href="http://sts.bioengr.uic.edu/castp/">http://sts.bioengr.uic.edu/castp/</a>
<b>Geometric</b>	CAVER	[92]	A 3D-grid that is constructed over a molecule and stripped to its convex hull. Nodes are evaluated using a cost function and the algorithm finds the lowest cost pathway between a point and the surface of the molecule	<a href="http://loschmidt.chemi.muni.cz/caver/">http://loschmidt.chemi.muni.cz/caver/</a>
<b>Geometric</b>	Fpocket	[93]	Clustering of alpha-shape spheres and scoring of each pocket with atoms properties	<a href="http://sourceforge.net/projects/fpocket/">http://sourceforge.net/projects/fpocket/</a>
<b>Geometric</b>	GHECOM (Probe-based HECOMi finder)	[94]	Placing small and large probe spheres on the protein VdW surface: pocket regions are defined as a space into which a small probe can enter, but a large probe cannot	<a href="http://biunit.naist.jp/ghecom/">http://biunit.naist.jp/ghecom/</a>
<b>Geometric</b>	LigSite	[38]	Scanning along 14 search vectors (x-, y- and z- axes plus the cubic diagonals) to delineate pockets	
<b>Geometric</b>	McVol	[95]	Identification of pockets by solving a Monte Carlo algorithm originally used for the computation of the van der Waals volume and the molecular volume of proteins	<a href="http://www.bisb.uni-bayreuth.de/People/ullmann/mcvol/mcvol.html">http://www.bisb.uni-bayreuth.de/People/ullmann/mcvol/mcvol.html</a>
<b>Geometric</b>	PASS (Putative Active Sites with Spheres)	[96]	Cavities in a protein are filled with a set of spheres	<a href="http://www.ccl.net/cca/software/UNIX/pass/index.shtml">http://www.ccl.net/cca/software/UNIX/pass/index.shtml</a>
<b>Geometric</b>	POCKET	[97]	3D grid and spherical probes to map protein surface and pockets	
<b>Geometric</b>	PocketDepth	[39]	Division of a given space into multiple subspaces using a grid and computation of their depths, which are used to retain and cluster only the high-depth subspaces, corresponding to pockets	<a href="http://proline.physics.iisc.ernet.in/pocketdepth/">http://proline.physics.iisc.ernet.in/pocketdepth/</a>
<b>Geometric</b>	PocketPicker	[40]	Identification of clusters of grid points with a buriedness index	<a href="http://gecco.org.chemie.uni-frankfurt.de/pocketpicker/">http://gecco.org.chemie.uni-frankfurt.de/pocketpicker/</a>
<b>Geometric</b>	Screen (Surface Cavity REcognition and EvaluationN)	[28]	Geometrical definition of surface cavity in terms of the empty space between the protein's molecular surface and an envelope surface constructed by rolling a probe and characterization with molecular descriptors	<a href="http://interface.bioc.columbia.edu/screen/">http://interface.bioc.columbia.edu/screen/</a>
<b>Geometric</b>	SplitPocket	[98,99]	Triangulation of a protein with weighted-Delaunay method and computation of a discrete flow algorithm with customized probes to obtain the pockets. The concept of a split pocket (comparison between the pocket found with or without ligand) is used to identify the functional surface of the protein	<a href="http://pocket.uchicago.edu/">http://pocket.uchicago.edu/</a>
<b>Geometric</b>	SURFNET	[37]	Fitting a virtual spheres into the solvent-accessible space between protein atoms	<a href="http://www.biochem.ucl.ac.uk/~roman/surfnet/surfnet.html">http://www.biochem.ucl.ac.uk/~roman/surfnet/surfnet.html</a>

TABLE 1 (Continued)

Method type	Name	Refs.	Search method	URL
Geometric	TravelDepth	[100]	Grid-based approach consisting of coating the protein with a 3D-grid and searching for grid points not in the protein satisfying some conditions	
Geometric	VICE (Vectorial Identification of Cavity Extends)	[101]	Determination of grid points, which are scored according to a metric roughly similar to degree of burial.	
Geometric	VOIDOO	[102]	VOIDOO and Flood are tools to delineate cavities	<a href="http://xray.bmc.uu.se/usf/voidoo.html">http://xray.bmc.uu.se/usf/voidoo.html</a>
Geometric	Xie and Bourne	[103]	Application of a geometric potential depending on both the global shape of the protein structure as well as the surrounding environment of each residue for binding-site prediction. Note that the protein structure is represented by C $\alpha$ atoms only	
Geometric and energy-based	SiteMap	[48]	Relevant points are selected on a 3D-grid based on geometric and energetic properties and the points are grouped into sets to define the sites	<a href="http://www.schrodinger.com/">http://www.schrodinger.com/</a>
Energy-based	ICM-PocketFinder	[32]	Calculation of a Lennard-Jones potential over a grid and prediction of the envelopes representing the shape and size of potential binding sites	<a href="http://www.molsoft.com/">http://www.molsoft.com/</a>
Energy-based	Q-SiteFinder	[42]	Positioning of a methyl probe at grid points and calculation of an interaction energy with the protein	<a href="http://www.modelling.leeds.ac.uk/qsitefinder/">http://www.modelling.leeds.ac.uk/qsitefinder/</a>
Energy-based	SITEHOUND	[104]	Identification of the regions characterized by favorable van der Waals interactions and computation of a cluster algorithm.	<a href="http://bsbbsinai.org/SHserver/SiteHound/download.html">http://bsbbsinai.org/SHserver/SiteHound/download.html</a>
Probe-mapping/energy-based	AutoLigand	[43]	The method searches the space surrounding the protein and finds the contiguous envelope with the specified volume of atoms which has the largest possible interaction energy with the protein	<a href="http://mgltools.scripps.edu/downloads">http://mgltools.scripps.edu/downloads</a>
Probe-mapping/energy-based	GRID	[41]	GRID is a computational procedure for determining energetically favorable binding sites on molecules of known structure.	<a href="http://www.moldiscovery.com/soft_grid.php">http://www.moldiscovery.com/soft_grid.php</a>
Probe-mapping/energy-based	Surflex-Protomol	[44]	Coating of the protein surface with 3 types of probes (hydrophobic, donor and acceptor probes)	<a href="http://www.biopharmics.com">www.biopharmics.com</a>
Docking	MEDock (Maximum-Entropy based Docking)	[105]	Binding sites found via docking	<a href="http://medock.csie.ntu.edu.tw/">http://medock.csie.ntu.edu.tw/</a>

accurately by the specific amino acid composition surrounding the concavities on the surface of proteins. Another type of druggability score has been developed and implemented in SiteMap [48]. The druggability of targets has also been addressed through simplified descriptions of the binding site (e.g. the hydrophobic nature and an additional term taking into account the concavity of the pocket, among others) [49] or from first-principle molecular simulations to quantify the maximal binding affinity without the need of any training set [50]. A direct application of these findings could be that if a pocket is ranked as difficult, it could be beneficial to develop a prodrug rather than trying to design a regular compound [48,49].

### Similarity between binding sites

It is known that one compound can bind to different targets. One example involves celecoxib (Celebrex), a nonsteroidal anti-inflammatory drug that binds to the cyclooxygenase COX-2 (PDB code 6cox, all alpha protein) and the totally unrelated carbonic anhydrase (PDB code 1oq5, all beta protein) [51]. Interestingly, both proteins share structural and/or physico-chemical similarities in the binding pocket area. In this particular case, resemblance between the binding pockets could be inferred through the use

of CavBase [52], a database of pockets associated with a tool to compare binding sites. This illustrates previous observations that targets, unrelated in sequence or structure, can bind the same ligands [12]. Thus, binding site prediction and comparison can have a crucial role for the annotation of protein function. Furthermore, toxicity can be due to a drug binding to anti-targets. Several methodologies and databases have been developed to annotate or compare binding sites (Table 2). In general, structure-based methods for local comparison of unrelated proteins use a simplified representation of the cavity residues [10]. These patterns are then structurally aligned and a scoring function is applied to quantify the similarity of the aligned features. The methods usually differ depending on the way the cavity or the entire surface is represented (all surface or pseudo-atoms), the way the molecular information of the binding pocket is encoded and how the best structural alignment is identified.

### Ligand-binding pockets, chemical space, drug discovery and chemical biology

Structural analysis and computational modeling of ligand-binding sites bring key information to designing drugs and annotating proteins [16,17,46,53–56]. Among the different challenges facing

TABLE 2

**Methods for evaluating binding site similarities and related databases.**

<i>Name</i>	<i>Refs.</i>	<i>Method</i>	<i>URL</i>
<b>CavBase</b>	[52]	Condense the physico-chemical properties of the cavity-flanking residues into a restricted set of generic pseudocenters (interpreted as nodes of a graph) corresponding to properties essential for molecular recognition. Then find similarities using a clique-detection algorithm. The solutions are ranked according to the similarity of property-based surface patches	<a href="http://relibase.ccdc.cam.ac.uk">http://relibase.ccdc.cam.ac.uk</a> <a href="http://relibase.rutgers.edu">http://relibase.rutgers.edu</a>
<b>CPASS (Comparison of Protein Active Site Structures)</b>	[106]	Determines the optimal sequence and structural alignment between two binding sites without maintaining sequence connectivity. An iterative search for best rotation/translation is then performed	<a href="http://bionmr-c1.unl.edu/CPASS_OV/CPASS.htm">http://bionmr-c1.unl.edu/CPASS_OV/CPASS.htm</a>
<b>CSC (Common Structural Cliques)</b>	[107]	A method for locating functionally relevant atoms in protein structures. The search method is based on comparison of local structure features of proteins that share a common biochemical function and does not depend on overall similarity of structures and sequences	
<b>eF-seek</b>	[108]	Performs clique detection on the vertices of the triangulated solvent-accessible surface. eF-seek and a database of binding sites (eF-site) are available	<a href="http://ef-site.hgc.jp/eF-seek/top.do">http://ef-site.hgc.jp/eF-seek/top.do</a>
<b>FINDSITE</b>	[109]	A method for ligand-binding site prediction and functional annotation based on binding site similarity across groups of weakly homologous template structures identified from threading	<a href="http://cssb.biology.gatech.edu/skolnick/files/FINDSITE/">http://cssb.biology.gatech.edu/skolnick/files/FINDSITE/</a>
<b>IsoCleft</b>	[110]	Graph matching-based method for the detection of pairwise local 3D atomic similarities without utilizing any sequence alignment information	Can be obtained from the authors
<b>MultiBind</b>	[111]	The method aims at finding binding patterns common to a set of protein structures. It performs a multiple alignment between protein binding sites in the absence of overall sequence, fold or binding partner similarity. MultiBind recognizes common spatial arrangements of physico-chemical properties in the binding sites. It applies an efficient geometric hashing technique to detect a potential set of multiple alignments of the given binding sites. To overcome the exponential number of possible multiple combinations, a very efficient filtering procedure is applied	<a href="http://bioinfo3d.cs.tau.ac.il/MultiBind/">http://bioinfo3d.cs.tau.ac.il/MultiBind/</a>
<b>Park and Kim</b>	[112]	Binding sites are transformed into graphs, which consist of nodes. Nodes are defined as C $\alpha$ carbon of each binding site residue. Comparison is then made by a maximum clique-detection algorithm	
<b>PROSURFER (PROteinSURFaceExploreR)</b>	[113]	A method for the structural alignment of atoms in the solvent-accessible surface of proteins that uses similarities in the local atomic environment	<a href="http://www.tsurumi.yokohama-cu.ac.jp/fold/database.html">http://www.tsurumi.yokohama-cu.ac.jp/fold/database.html</a>
<b>Query3d</b>	[114]	A method that integrates many existing databases and programs for 3D functional annotation together with a fast structural comparison algorithm	<a href="http://pdbfun.uniroma2.it/">http://pdbfun.uniroma2.it/</a>
<b>Ramensky et al.</b>	[115]	Comparison of a query protein binding site (target) against the 3D structure of another protein (analog) in complex with a ligand enables ligand fragments from the analog complex to be transferred to positions in the target site, so that the complete protein environments of the fragment and its image are similar. The revealed environments are similarity regions and the fragments transferred to the target site are considered as binding patterns	
<b>SiteAlign</b>	[116]	The method measures distances between protein cavities. Starting from user-defined ligand-binding sites, eight topological and physico-chemical properties are projected from cavity-lining protein residues to an 80 triangle-discretized sphere placed at the centre of the binding site, thus defining a cavity fingerprint	<a href="http://bioinfo-pharma.u-strasbg.fr/template/jd/pages/download/download.php">http://bioinfo-pharma.u-strasbg.fr/template/jd/pages/download/download.php</a>



TABLE 2 (Continued)

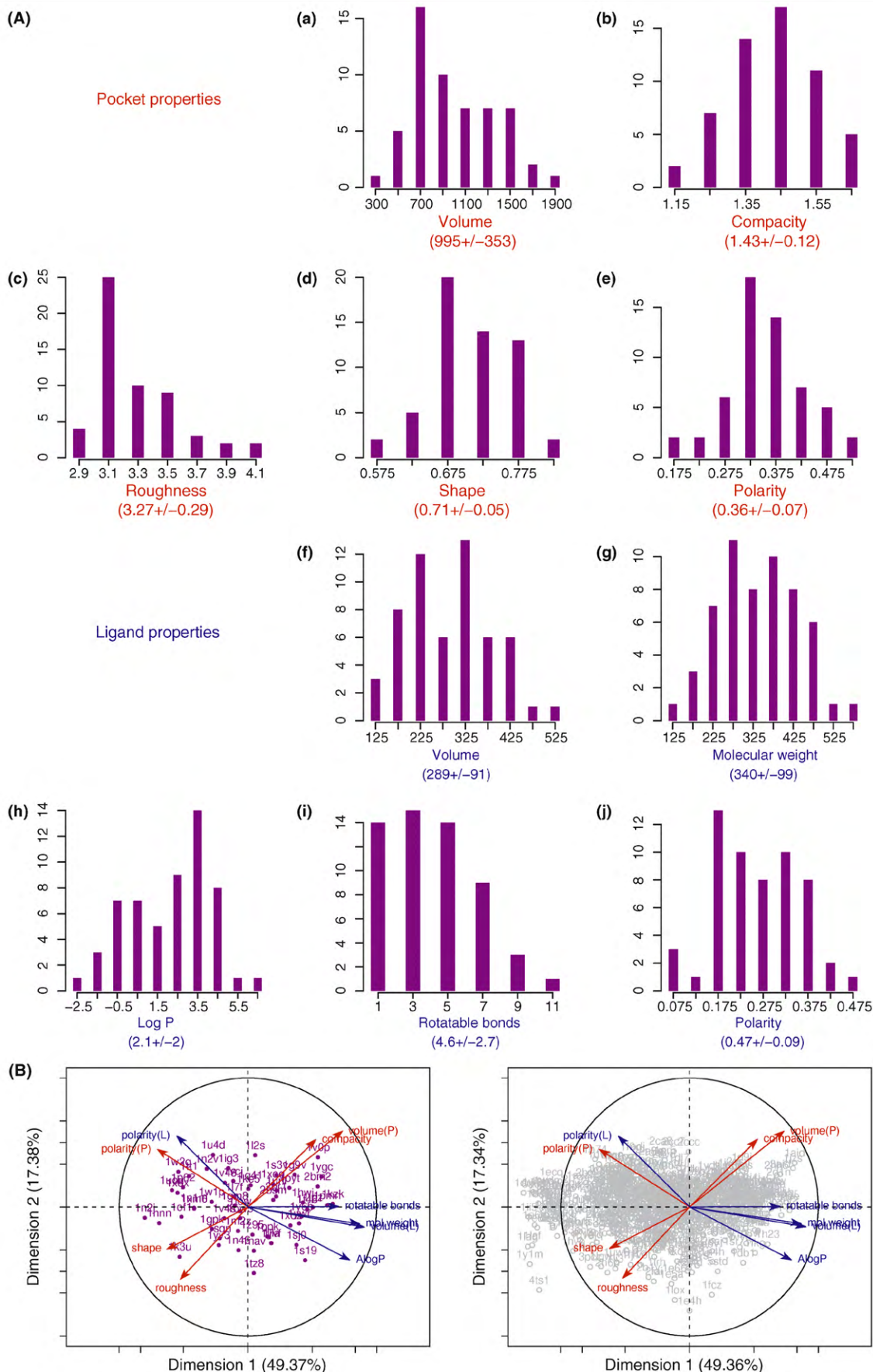
Name	Refs.	Method	URL
SiteBase	[117]	Comparison of binding sites using geometric matching to detect similar three-dimensional structure using a maximum clique-detection method	<a href="http://www.modelling.leeds.ac.uk/sb">http://www.modelling.leeds.ac.uk/sb</a>
SiteEngine	[118]	Efficient hashing and matching of triangles of centers of physico-chemical properties (hydrogen-bond donor, acceptor, mixed donor and acceptor, hydrophobic aliphatic and aromatic) and fast hierarchical scoring of all solutions	<a href="http://bioinfo3d.cs.tau.ac.il/SiteEngine/">http://bioinfo3d.cs.tau.ac.il/SiteEngine/</a>
SMID-BLAST (Small Molecule Interaction Database BLAST)	[119]	Identification of putative small molecule binding sites in proteins for which a crystal-structure has not yet been determined by finding structural domains from the CDD (Conserved Domain Database) that are in complex with small compounds	
SuMo	[120]	Chemical groups of atoms are used to build triangles of chemical groups, the comparison of two molecules starts from the graphs of triangles representing the input molecules, pairs of similar triangles that are geometrically consistent are made and form patches that are refined by using a selection procedure	<a href="http://sumo-pbil.ibcp.fr">http://sumo-pbil.ibcp.fr</a>
VA (Vincinity Analysis)	[121]	Identifies similarities between protein binding sites based on their three-dimensional structure and the chemical similarity of matching residues	
Weskamp <i>et al.</i>	[122]	A two-step method that allows detection of common substructures in proteins. The tool combines the advantages from both clique-detection and geometric hashing approaches	Can be obtained from the authors
@TOME-2	[123]	Pipeline for comparative modeling of protein–ligand complexes	<a href="http://abcis.cbs.cnrs.fr/AT2/">http://abcis.cbs.cnrs.fr/AT2/</a>

binding site prediction and analysis, we comment below on three different topics: recent advances in handling pocket flexibility, the possibility of tuning a scoring function for a given cavity type instead of a target or family type and the design of pocket-focused compound collections.

### Pocket flexibility

As mentioned above, receptor and pocket flexibility can make binding site predictions and comparisons difficult because it can affect the properties of the pockets. Although some protein–ligand complexes apparently still fit the ‘lock and key’ hypothesis, two additional mechanisms have been proposed and probably represent the vast majority of the protein–ligand partnerships: induced fit and conformational selection theory [57]. In the latter theory, the unbound protein explores the energy landscape, spending most of the time in the lowest energy conformations, but also occupies higher-energy states, some of which, potentially, are structurally similar to the bound conformation. In the course of binding, because of favorable interactions with a ligand, these protein conformers get preferentially selected and the population of protein microstates shifts in the direction of the bound conformation. As such, in the case of induced fit, optimal binding is achieved by specific structural changes, whereas in the conformational selection theory, binding takes place through selection from the already present unbound ensemble. Another observation about binding sites and more specifically catalytic sites was reported some years ago. Using structure-based thermodynamic stability analysis of non-structurally homologous proteins for which high-resolution structures of their complexes with specific ligands were available, it was shown that for all 16 proteins

considered, the binding sites had a dual character with very low structural stability and high-stability regions [58]. Small molecular motions can be handled by pocket comparison methods such as SiteAlign because of the fuzzy representation of the protein sites. Pockets have still to be formed and visible in the receptor structures, however, and they have to be somewhat similar to a bound conformation. Obviously, if several experimental structures are available, they could be used for pocket analysis and prediction, assuming they are structurally related to the bound conformation. In many cases, these data are missing and simulation tools such as molecular dynamics, Monte Carlo methods, normal mode analysis, Gaussian network models or essential dynamics can be applied [59,60]. It seems that the unliganded protein pockets open frequently, as shown by standard molecular dynamics simulation runs [61]. Moreover, either for docking purposes or to explore alternative pocket conformations, some specific methods have been developed recently, including SCARE (e.g. flexible side chains in the binding pocket should be temporarily deleted rather than wrongly positioned), active site fumigation, or active site resin pressurization [15,62,63]. One key problem remains, however: the selection of the relevant protein conformations among ‘wrong’ structures (i.e. which pocket structures could represent a bound conformation or could be acceptable for virtual screening or binding site prediction and comparison). If inappropriate, this multi-receptor state can turn out to be counterproductive compared with just using a single rigid receptor structure. Some protocols attempting to address this issue have been proposed recently: they usually make use of molecular dynamics runs and involve the pruning of receptor ensembles to generate a small number of possibly relevant conformations [64–67]. However,



important work is still required in this field to understand, predict, characterize and fully exploit biophoric patterns in drug discovery and chemical biology projects.

## Scoring

Scoring functions for the purpose of ranking molecules during the course of structure-based virtual screening projects have several weaknesses. In fact, most packages incorporate a generalized scoring function that has been derived through the use of several different receptor and ligand structures. Some binding pockets are polar, however, whereas others are essentially hydrophobic. Such scoring functions still succeed in prioritizing molecules in some cases but generate many false positive and negative hits. Consensus scoring can be used and so can target-specific scoring functions [68–70]. Although consensus scoring generally improves enrichment by compensating for the deficiencies of each scoring function, the strategy of how individual scoring functions are selected remains a challenging task when only a few known active compounds are available. With regard to target-specific optimization of scoring functions, the limitations include the fact that the function has to be tuned for a single protein or for a family (assuming that sufficient experimental data are available). The process is also time consuming and cannot be applied to a new unknown target. It is now known that tuned scoring functions toward a protein family can effectively increase virtual screening enrichment rates [71] such that it would seem possible to develop and tune master scoring functions dedicated to a pocket class and, thus, be able to simultaneously hit multiple targets not belonging to the same family with the same scoring engine.

## Compound collections

Target-specific compound libraries (e.g. serine proteases or kinases) are known to efficiently improve hit rates while reducing the overall cost of experimental screening [72]. Yet by analyzing protein structures, it has been suggested that only about 1000 different folds exist [73]; thus, the number of binding pockets might also be limited. In fact, the size, shape and distribution of functional groups within the pockets could dictate the design of a limited number of compound collections instead of designing a collection for each target or each family. A generic compound collection-filtering scheme functioning along this line could initially follow the guidelines proposed by Djuric's group [74]. The working hypothesis in their study was that the shapes of the ligands that could bind a regular target cavity (e.g. a catalytic site) were likely to be similar to the shape of known bioactive ligands. Yet the shape of bioactive ligands is, in part, dependent on the shape and nature of the binding pockets. Using a shape descriptor

[75], the authors noted that most drugs congregate in a region of the chemical space found to be between the 'rod' and 'pancake' but not spherical shapes, suggesting that most active-site proteins can only bind drugs that do not adopt a spherical shape. The analysis revealed that approximately 5–10% of the Abbott compound collection could not energetically achieve the rod or pancake shape. Detailed investigations of this kind of behavior could enable a first level of filtering and reduce the size of the initial collection. Furthermore, because it is known that proteins with different folds and low sequence identity can bind similar ligands, it is likely that the pocket space could be used to design pocket-focused compound collections and avoid the generation of a collection for each target family. A recent study managed to classify 623 binding sites into only 23 independent clusters with specific topological properties [18]. Works around the pocketome [32,76,77] have been essentially devoted to investigation of cross-reactivities (same ligand binding to different proteins that have a similar pocket), hit design, optimization and de-orphanization but not fully to the design of compound collections, whereas it has been shown recently that the cavity space has higher relevance than the sequence space [76]. The concept of 'binding site centric chemical space' has been coined recently [9,78] and supports the above discussion. In these two studies, the authors pursued the idea that binding sites with similar geometrical and/or chemical properties would recognize similar functional groups. Insights into the features of selective binding mode of carboxylic, sulfonic, phosphonic, amine and amidine moieties have been gained and suggest, for instance, that highly polar and large binding sites endowed with broad conformational flexibility have the propensity to bind guanidine and primary amine groups. The authors noted that the guanidine and primary amine binding sites have respectively triple and double the number of hydrogen bond acceptors on median values than the secondary, tertiary and quaternary amine binding sites. Such studies clearly support the notion of charting the chemical space based not only on known activities and/or target family but also on ligand-binding cavities.

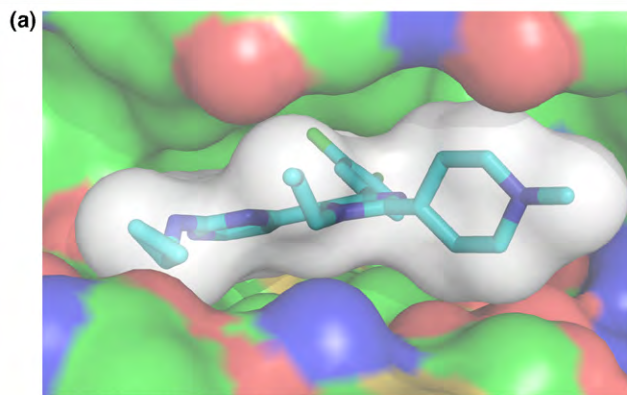
## Case study: revisiting pocket–ligand complexes through computational means

We decided to revisit pocket–ligand pairs with a recent and high-resolution set of protein pockets in complexes with drug-like ligands (typically, compounds with a molecular mass between 150 and 600 Da), keeping in mind some of the concepts reviewed above. We analyzed 56 high-resolution structures (kinases, nuclear receptors, serine proteases, members of the phosphodiesterase family, and so on) extracted from the Astex set [26]. This dataset was selected because it has been carefully generated and manually

### FIGURE 1

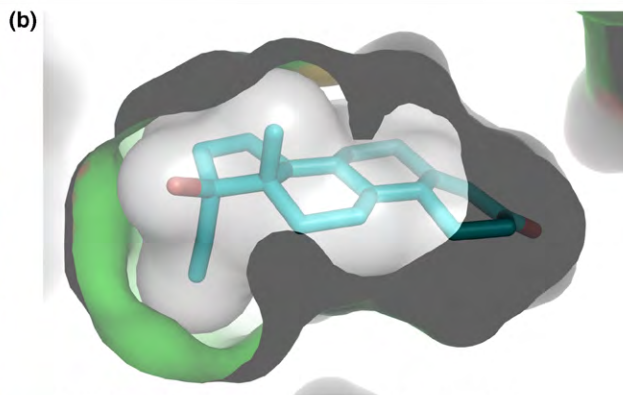
Analysis of pocket and ligand properties. **(A)** To define pockets, we used the protocol utility of Surflex [83]; pockets atoms were defined as accessible atoms at a radius of 4 Å from any protomol probe. The histograms of pocket properties [(a)–(e)] and ligand properties [(f)–(j)] were computed on the Astex set (magenta). Means and standard deviations are given in parentheses. The considered descriptors are volume (MSMS package implemented in Chimera [84]), compacity [31], roughness [81], shape [30] and polarity ratio [61]. Ligand properties were volume, polarity ratio [61], molecular weight, A log *P* and rotatable bonds as implemented in FAF-Drugs2 [85]. **(B)** The principal component analysis (PCA) of pocket and ligand properties was computed on the two sets. PCA performs a linear projection of the data points from the high-dimensional space to a low dimensional space while accounting for a maximum of total variability. This illustrates that the Astex subset (magenta, left) samples well the pocket and ligand space represented by the larger PDBbind subset (gray, right). As indicated by the position of the variables (red for pocket and blue for ligand) close to the correlation circle, all descriptors contribute to the variability of the data capturing more than 65% on the two first axes. It can be observed that pocket roughness and shape are opposed to volume and compacity, whereas pocket and ligand polarity are correlated, as well as pocket and ligand volume. Rotatable bonds, molecular weight and log *P* also tend to be correlated.



**1PMN (protein kinase)**

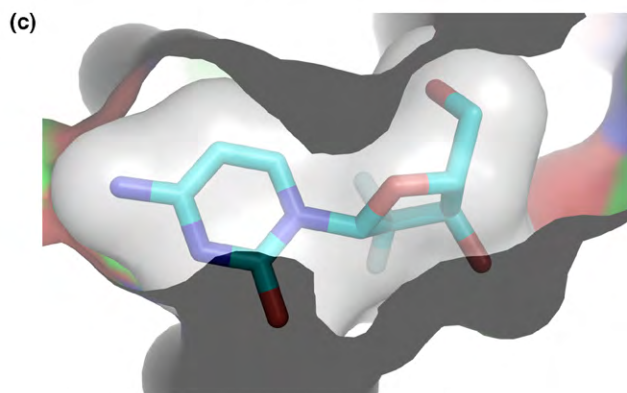
Crystal structure of JNK3 in complex with an imidazole-pyrimidine inhibitor. The ligand has very tight contacts with the pocket that forms a narrow and elongated cavity.

Pocket: vol=1458A3 ; pol=0.3 ; hba=8 ; hbd=12  
Ligand: vol=440A3 ; pol=0.19 ; hba=3 ; hbd=1  
Interaction (IC50=7nM): hbb=2 ; hbs=0 ; cp=1 ; ar=0 ; sb=0

**1SQN (nuclear receptor ligand-binding domain)**

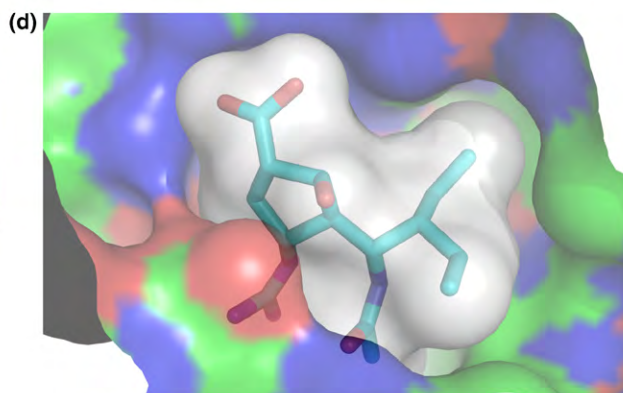
Progesterone receptor ligand binding domain with bound norethindrone. The ligand is fully buried inside the pocket, a very tight shape complementarity is observed.

Pocket: vol=590A3 ; pol=0.41 ; hba=0 ; hbd=5  
Ligand: vol=274A3 ; pol=0.09 ; hba=3 ; hbd=1  
Interaction (Kd=0.4 nM): hbb=0 ; hbs=2 ; cp=0 ; ar=0 ; sb=0

**1P62 (nucleotide and nucleoside kinase)**

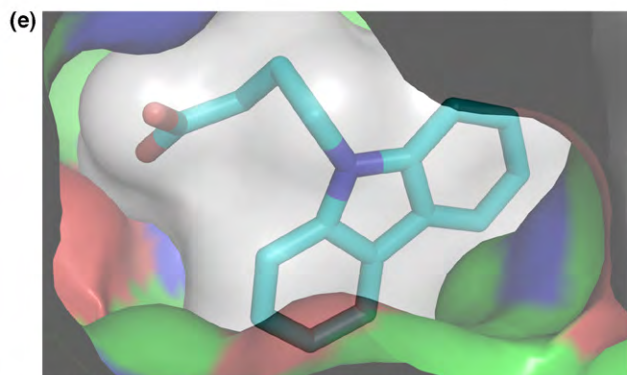
Structure of human DCK complexed with gemcitabine and ADP-MG. The ligand is essentially buried in the pocket with a very tight shape complementarity.

Pocket: vol=575A3 ; pol=0.45 ; hba=2 ; hbd=6  
Ligand: vol=189A3 ; pol=0.44 ; hba=5 ; hbd=3  
Interaction (Km=22000nM): hbb=0 ; hbs=5 ; cp=1 ; ar=1 ; sb=1

**1L7F (neuraminidase)**

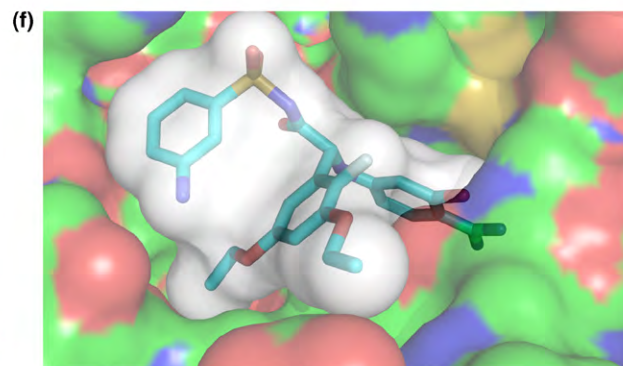
Crystal structure of influenza virus neuraminidase in complex with BCX-1812. The ligand is in part exposed but for the regions that are in contact with the protein, the shape complementarity is high.

Pocket: vol=815A3 ; pol=0.36 ; hba=7 ; hbd=10  
Ligand: vol=306A3 ; pol=0.35 ; hba=5 ; hbd=6  
Interaction (IC50=0.8 nM): hbb=2 ; hbs=12 ; cp=0 ; ar=0 ; sb=5

**1TOW (fatty acid binding protein-like)**

Crystal structure of human adipocyte fatty acid binding protein in complex with a carboxylic acid ligand. The ligand is essentially buried in the cavity but the shape complementarity is not very high, some free space is visible around the molecular surface of the ligand.

Pocket: vol=915A3 ; pol=0.38 ; hba=7 ; hbd=7  
Ligand: vol=330A3 ; pol=0.16 ; hba=2 ; hbd=1  
Interaction (IC50=570nM): hbb=0 ; hbs=3 ; cp=1 ; ar=1 ; sb=1

**1YGC (eukaryotic protease)**

Factor VIIa-inhibitor complex. The ligand is exposed but for one region that inserts into a deep pocket. The shape complementarity is tight for some regions but in some other parts, there is room between the molecular surface of the protein and the one of the ligand.

Pocket: vol=1403A3 ; pol=0.6 ; hba=9 ; hbd=9  
Ligand: vol=459A3 ; pol=0.32 ; hba=7 ; hbd=6  
Interaction (Ki=0.35 nM): hbb=2 ; hbs=6 ; cp=0 ; ar=1 ; sb=1

*Drug Discovery Today*

curated, through – for instance – the generation of the electronic density maps around the ligands. The test set originally contained 85 complexes, but we removed 29 proteins with ion in the binding pocket or having a cofactor next to the ligand. To ensure that our analysis could be extended to other proteins, we also investigated 564 protein–ligand complexes extracted from PDBbind [79] (see [Supplementary Data online](#)). Hydrogen bonds, cation–pi, and aromatic stacking between the ligands and the proteins were analyzed (Figs 1 and 2; see figure legends for details about computations). Seventy-nine percent of these selected pockets have no cation–pi interactions, 68% have no aromatic stacking and 29% have no salt bridge. Interestingly, in more than 44% of these pockets, we found no hydrogen bonds with the protein backbone. By contrast, in 91% of the pockets, we noted hydrogen bonds with the protein side chains. Although incomplete, this analysis sheds light on pocket properties, ligand properties and the correlations between the two spaces. We observed that pockets in our dataset tend to have a low or medium polarity ratio ( $0.36 \pm 0.07$ ) and are thus essentially hydrophobic or of mixed nature (polar and lipophilic), in agreement with previously reported data [31,49,80]. Through the use of principal component analysis, we saw that our selected Astex subset was representative of the larger PDBbind subset (i.e. pockets and ligands essentially belong to the same chemical space as the larger set). Smallest pockets, which are also the least compact ones, tend to be rougher (as shown previously) [81]; more spherical, in agreement with a recent study [30]; and more polar. These particular pockets seem to bind smaller ligands, with a lower log *P* value and fewer rotatable bonds. These trends are confirmed by the use of classification methods (data not shown). This kind of analysis also suggests that scoring functions could be developed for a pocket type, instead of – or complementing – generic scoring and/or target-family-tuned scoring functions: we note, for example, that three proteins from three different families (kinase 2BR1, phosphatase 1XOQ and serine protease 1OYT) bind ligands with related volume and polarity ratio ( $327 \text{ \AA}^3$  and 0.24,  $309 \text{ \AA}^3$  and 0.23, and  $345 \text{ \AA}^3$  and 0.21, respectively). In fact, the binding seems to be guided by the pocket itself because these three proteins share similar pocket properties in terms of volume ( $1222 \text{ \AA}^3$ ,  $1480 \text{ \AA}^3$  and  $1333 \text{ \AA}^3$ ), roughness (3.30, 3.18 and 2.89) and shape (0.70, 0.67 and 0.69). We also observe that two proteins from the same family (nuclease 1Z95 and 1S19) bind two different types of ligand in terms of volume ( $315 \text{ \AA}^3$  and  $417 \text{ \AA}^3$ , respectively) and polarity ratio (0.28 *versus* 0.10). The

related pockets are also different in volume ( $699 \text{ \AA}^3$  and  $1034 \text{ \AA}^3$ ) and in polarity ratio (0.32 *versus* 0.23). A recent study along this line of reasoning further highlights that even when sequence identity is high, low binding site similarities can be found [82]. Although we have not yet investigated whether the concept of designing a scoring function for a binding pocket type rather than for a single protein or a protein family for the targets mentioned above (2BR1, 1XOQ, 1OYT) could provide interesting results, we [36] and others [35] have noticed that some docking and/or scoring engines perform better on some specific binding pocket classes (e.g. polar or lipophilic).

## Concluding remarks

Ligand-binding site prediction is an active field of research, and several different approaches are being used, from geometric to energy-based methods. No method is 100% successful, and each has its own strengths and weaknesses, but functional site location is extremely important for predicting function, for drug discovery and for chemical biology projects. Prediction of druggability and binding site comparisons can also be performed. Statistical analysis of binding pockets can provide important information to a design strategy, from evaluating the likelihood of a docking pose to the right selection of a medicinal chemistry protocol. Enhanced understanding of the cavity space will obviously help to understand binding to anti-targets and off-targets and, therefore, should contribute notably to drug discovery. A difficulty in this field comes from flexibility of the pockets and receptors, a well-known limitation of present SBDD tools. Generation or selection of the relevant target conformers among a large set of 3D structures will require extensive research efforts. Finally, it seems that charting the chemical space of the pocketome could be valuable to prepare a compound collection and to design tuned scoring functions dedicated to pocket classes, in contrast to receptor classes.

## Acknowledgements

We would like to thank the Inserm Institute, the University of Paris Diderot, the CODDIM 'region Ile-de-France', and the academic French IbiSa funds for financial support.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.drudis.2010.05.015](https://doi.org/10.1016/j.drudis.2010.05.015).

## References

- Weigelt, J. *et al.* (2008) Structural genomics and drug discovery: all in the family. *Curr. Opin. Chem. Biol.* 12, 32–39
- Fedorov, O. *et al.* (2007) Insights for the development of specific kinase inhibitors by targeted structural genomics. *Drug Discov. Today* 12, 365–372
- Kolb, P. *et al.* (2009) Docking and chemoinformatic screens for new ligands and targets. *Curr. Opin. Biotechnol.* 20, 429–436
- Cavasotto, C.N. and Phatak, S.S. (2009) Homology modeling in drug discovery: current trends and applications. *Drug Discov. Today* 14, 676–683

## FIGURE 2

Ligand and pocket properties of some selected targets (see also the legend of Fig. 1): (a) 1PMN, (b) 1SQN, (c) 1P62, (d) 1L7F, (e) 1TOW and (f) 1YGC. *Vol* is the abbreviation of volume, *pol* of polarity, *hba* of hydrogen-bond acceptors and *hbd* of hydrogen-bond donors. With regard to interactions, *hbb* and *hbs* correspond to the number of hydrogen bonds between the ligand and either the protein backbone (*hbb*) or the amino acid side chains (*hbs*) lining the ligand-binding pocket. The three other interactions evaluated correspond to the presence (1) or absence (0) of at least one cation–pi interaction (*cp*), aromatic stacking (*ar*) and salt bridge (*sb*) between the protein and the ligand. The protein families in brackets come from the SCOP database [86]. For the ligands, oxygen atoms are in red, N in blue and S in yellow. For the protein, carbon atoms are in green, and they are cyan for ligands. Green atoms in the ligand are Cl, and F atoms are shown in light blue. All pictures were generated with PyMOL. Gray surface corresponds to the volume of ligands generated with the surface tool of PyMOL.



- 5 Vajda, S. and Guarnieri, F. (2006) Characterization of protein-ligand interaction sites using experimental and computational methods. *Curr. Opin. Drug Discov. Devel.* 9, 354–362
- 6 An, J. *et al.* (2004) Comprehensive identification of 'druggable' protein ligand binding sites. *Genome Inform.* 15, 31–41
- 7 Hunter, W.N. (2009) Structure-based ligand design and the promise held for antiprotozoan drug discovery. *J. Biol. Chem.* 284, 11749–11753
- 8 Keller, T.H. *et al.* (2006) A practical view of 'druggability'. *Curr. Opin. Chem. Biol.* 10, 357–361
- 9 Macchiarulo, A. and Pellicciari, R. (2007) Exploring the other side of biologically relevant chemical space: insights into carboxylic, sulfonic and phosphonic acid bioisosteric relationships. *J. Mol. Graph. Model.* 26, 728–739
- 10 Kellenberger, E. *et al.* (2008) How to measure the similarity between protein ligand-binding sites? *Curr. Comput. Aided Drug Design* 4, 209–220
- 11 Andersson, C.D. *et al.* (2009) Mapping of ligand-binding cavities in proteins. *Proteins* 78, 1408–1422
- 12 Chalk, A.J. *et al.* (2004) PDBLIG: classification of small molecular protein binding in the Protein Data Bank. *J. Med. Chem.* 47, 3807–3816
- 13 Fuller, J.C. *et al.* (2009) Predicting druggable binding sites at the protein-protein interface. *Drug Discov. Today* 14, 155–161
- 14 Laurie, A.T.R. and Jackson, R.M. (2006) Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening. *Curr. Protein Pept. Sci.* 7, 395–406
- 15 Abagyan, R. and Kufareva, I. (2009) The flexible pocketome engine for structural chemogenomics. *Methods Mol. Biol.* 575, 249–279
- 16 Kahraman, A. *et al.* (2007) Shape variation in protein binding pockets and their ligands. *J. Mol. Biol.* 368, 283–301
- 17 Gunasekaran, K. and Nussinov, R. (2007) How different are structurally flexible and rigid binding sites? Sequence and structural features discriminating proteins that do and do not undergo conformational change upon ligand binding. *J. Mol. Biol.* 365, 257–273
- 18 Weisel, M. *et al.* (2010) Architectural repertoire of ligand-binding pockets on protein surfaces. *ChemBioChem* 11, 556–563
- 19 Campagna-Slater, V. *et al.* (2010) Pharmacophore screening of the Protein Data Bank for specific binding site chemistry. *J. Chem. Inf. Model* 50, 358–367
- 20 Segers, K. *et al.* (2007) Design of protein membrane interaction inhibitors by virtual ligand screening, proof of concept with the C2 domain of factor V. *Proc. Natl. Acad. Sci. U.S.A.* 104, 12697–12702
- 21 Wells, J.A. and McClendon, C.L. (2007) Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature* 450, 1001–1009
- 22 Villoutreix, B.O. *et al.* (2008) In silico-in vitro screening of protein-protein interactions: towards the next generation of therapeutics. *Curr. Pharm. Biotechnol.* 9, 103–122
- 23 Higueruelo, A.P. *et al.* (2009) Atomic interactions and profile of small molecules disrupting protein-protein interfaces: the TIMBAL database. *Chem. Biol. Drug Des.* 74, 457–467
- 24 Dahlback, B. and Villoutreix, B.O. (2003) Molecular recognition in the protein C anticoagulant pathway. *J. Thromb. Haemost.* 1, 1525–1534
- 25 Weisel, M. *et al.* (2009) Form follows function: shape analysis of protein cavities for receptor-based drug design. *Proteomics* 9, 451–459
- 26 Hartshorn, M.J. *et al.* (2007) Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* 50, 726–741
- 27 Liang, J. *et al.* (1998) Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.* 7, 1884–1897
- 28 Nayal, M. and Honig, B. (2006) On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins* 63, 892–906
- 29 Laskowski, R.A. *et al.* (1996) Protein clefts in molecular recognition and function. *Protein Sci.* 5, 2438–2452
- 30 Sonavane, S. and Chakrabarti, P. (2008) Cavities and atomic packing in protein structures and interfaces. *PLOS Comput. Biol.* 4, e1000188
- 31 Hajduk, P.J. *et al.* (2005) Druggability indices for protein targets derived from NMR-based screening data. *J. Med. Chem.* 48, 2518–2525
- 32 An, J. *et al.* (2005) Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol. Cell. Proteomics* 4, 752–761
- 33 Hu, L. *et al.* (2005) Binding MOAD (Mother Of All Databases). *Proteins* 60, 333–340
- 34 Carlson, H.A. *et al.* (2008) Differences between high- and low-affinity complexes of enzymes and nonenzymes. *J. Med. Chem.* 51, 6432–6441
- 35 Schulz-Gasch, T. and Stahl, M. (2003) Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. *J. Mol. Model.* 9, 47–57
- 36 Miteva, M.A. *et al.* (2005) Fast structure-based virtual ligand screening combining FRED, DOCK, and Surflex. *J. Med. Chem.* 48, 6012–6022
- 37 Laskowski, R.A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.* 13, 323–330
- 38 Hendlich, M. *et al.* (1997) LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model.* 15, 359–363
- 39 Kalidas, Y. and Chandra, N. (2008) PocketDepth: a new depth based algorithm for identification of ligand binding sites in proteins. *J. Struct. Biol.* 161, 31–42
- 40 Weisel, M. *et al.* (2007) PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chem. Cent. J.* 1, 7
- 41 Goodford, P.J. (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* 28, 849–857
- 42 Laurie, A.T.R. and Jackson, R.M. (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* 21, 1908–1916
- 43 Harris, R. *et al.* (2008) Automated prediction of ligand-binding sites in proteins. *Proteins* 70, 1506–1517
- 44 Ruppert, J. *et al.* (1997) Automatic identification and representation of protein binding sites for molecular docking. *Protein Sci.* 6, 524–533
- 45 Kortvelyesi, T. *et al.* (2003) Algorithms for computational solvent mapping of proteins. *Proteins* 51, 340–351
- 46 Landon, M.R. *et al.* (2007) Identification of hot spots within druggable binding regions by computational solvent mapping of proteins. *J. Med. Chem.* 50, 1231–1240
- 47 Soga, S. *et al.* (2007) Use of amino acid composition to predict ligand-binding sites. *J. Chem. Inf. Model.* 47, 400–406
- 48 Halgren, T. (2007) New method for fast and accurate binding-site identification and analysis. *Chem. Biol. Drug Des.* 69, 146–148
- 49 Cheng, A.C. *et al.* (2007) Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* 25, 71–75
- 50 Seco, J. *et al.* (2009) Binding site detection and druggability index from first principles. *J. Med. Chem.* 52, 2363–2371
- 51 Weber, A. *et al.* (2004) Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. *J. Med. Chem.* 47, 550–557
- 52 Schmitt, S. *et al.* (2002) A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* 323, 387–406
- 53 Keiser, M.J. *et al.* (2009) Predicting new molecular targets for known drugs. *Nature* 462, 175–181
- 54 Renner, S. *et al.* (2009) Bioactivity-guided mapping and navigation of chemical space. *Nat. Chem. Biol.* 5, 585–592
- 55 Campillos, M. *et al.* (2008) Drug target identification using side-effect similarity. *Science* 321, 263–266
- 56 Gupta, A. *et al.* (2009) Structural models in the assessment of protein druggability based on HTS data. *J. Comput. Aided Mol. Des.* 23, 583–592
- 57 Wlodarski, T. and Zagrovic, B. (2009) Conformational selection and induced fit mechanism underlie specificity in noncovalent interactions with ubiquitin. *Proc. Natl. Acad. Sci. U.S.A.* 106, 19346–19351
- 58 Luque, I. and Freire, E. (2000) Structural stability of binding sites: consequences for binding affinity and allosteric effects. *Proteins (Suppl. 4)*, 63–71
- 59 Cozzini, P. *et al.* (2008) Target flexibility: an emerging consideration in drug discovery and design. *J. Med. Chem.* 51, 6237–6255
- 60 McCammon, J.A. (2005) Target flexibility in molecular recognition. *Biochim. Biophys. Acta* 1754, 221–224
- 61 Eyrich, S. and Helms, V. (2007) Transient pockets on protein surfaces involved in protein-protein interaction. *J. Med. Chem.* 50, 3457–3464
- 62 Withers, I.M. *et al.* (2008) Active site pressurization: a new tool for structure-guided drug design and other studies of protein flexibility. *J. Chem. Inf. Model.* 48, 1448–1454
- 63 Bottegoni, G. *et al.* (2008) A new method for ligand docking to flexible receptors by dual alanine scanning and refinement (SCARE). *J. Comput. Aided Mol. Des.* 22, 311–325
- 64 Ekonomiuk, D. *et al.* (2009) Flaviviral protease inhibitors identified by fragment-based library docking into a structure generated by molecular dynamics. *J. Med. Chem.* 52, 4860–4868
- 65 Bolstad, E.S. and Anderson, A.C. (2008) In pursuit of virtual lead optimization: the role of the receptor structure and ensembles in accurate docking. *Proteins* 73, 566–580
- 66 Rueda, M. *et al.* (2010) Recipes for the selection of experimental protein conformations for virtual screening. *J. Chem. Inf. Model.* 50, 186–193
- 67 Sperandio, O. *et al.* How to choose relevant multiple receptor conformations for virtual screening: a test case of Cdk2 and normal mode analysis. *Eur. Biophys. J.* (in press).
- 68 Teramoto, R. and Fukunishi, H. (2008) Consensus scoring with feature selection for structure-based virtual screening. *J. Chem. Inf. Model.* 48, 288–295
- 69 Feher, M. (2006) Consensus scoring for protein-ligand interactions. *Drug Discov. Today* 11, 421–428

- 70 Seifert, M.H. *et al.* (2007) Virtual high-throughput screening of molecular databases. *Curr. Opin. Drug Discov. Devel.* 10, 298–307
- 71 Seifert, M.H. (2009) Targeted scoring functions for virtual screening. *Drug Discov. Today* 14, 562–569
- 72 Orry, A.J. *et al.* (2006) Structure-based development of target-specific compound libraries. *Drug Discov. Today* 11, 261–266
- 73 Cuff, A.L. *et al.* (2009) The CATH classification revisited—architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res.* 37 (Database issue), D310–D314
- 74 Akritopoulou-Zanze, I. *et al.* (2007) Topography-biased compound library design: the shape of things to come? *Drug Discov. Today* 12, 948–952
- 75 Sauer, W.H. and Schwarz, M.K. (2003) Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity. *J. Chem. Inf. Comput. Sci.* 43, 987–1003
- 76 Weskamp, N. *et al.* (2009) Merging chemical and biological space: structural mapping of enzyme binding pocket space. *Proteins* 76, 317–330
- 77 Gold, N.D. and Jackson, R.M. (2006) Fold independent structural comparisons of protein–ligand binding sites for exploring functional relationships. *J. Mol. Biol.* 355, 1112–1124
- 78 Macchiarulo, A. *et al.* (2009) Charting the chemical space of target sites: insights into the binding modes of amine and amidine groups. *J. Chem. Inf. Model.* 49, 900–912
- 79 Wang, R. *et al.* (2004) The PDBbind database: collection of binding affinities for protein–ligand complexes with known three-dimensional structures. *J. Med. Chem.* 47, 2977–2980
- 80 Halgren, T.A. (2009) Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inf. Model.* 49, 377–389
- 81 Pettit, F.K. and Bowie, J.U. (1999) Protein surface roughness and small molecular binding sites. *J. Mol. Biol.* 285, 1377–1382
- 82 Kinnings, S.L. and Jackson, R.M. (2009) Binding site similarity analysis for the functional classification of the protein kinase family. *J. Chem. Inf. Model.* 42, 318–329
- 83 Jain, A.N. (2003) Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* 46, 499–511
- 84 Pettersen, E.F. *et al.* (2004) UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612
- 85 Lagorce, D. *et al.* (2008) FAF-Drugs2: free ADME/tox filtering tool to assist drug discovery and chemical biology projects. *BMC Bioinformatics* 9, 396
- 86 Murzin, A.G. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536–540
- 87 Huang, B. and Schroeder, M. (2006) LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.* 6, 19
- 88 Glaser, F. *et al.* (2006) A method for localizing ligand binding pockets in protein structures. *Proteins* 62, 479–488
- 89 Peters, K.P. *et al.* (1996) The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Biol.* 256, 201–213
- 90 Zhong, S. and MacKerell, A.D. (2007) Binding response: a descriptor for selecting ligand binding site on protein surfaces. *J. Chem. Inf. Model.* 47, 2303–2315
- 91 Binkowski, T.A. *et al.* (2003) CASTp: computed atlas of surface topography of proteins. *Nucleic Acids Res.* 31, 3352–3355
- 92 Petrek, M. *et al.* (2006) CAVER: a new tool to explore routes from protein clefts, pockets and cavities. *BMC Bioinformatics* 7, 316
- 93 Le Guilloux, V. *et al.* (2009) Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics* 10, 168
- 94 Kawabata, T. and Go, N. (2007) Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites. *Proteins* 68, 516–529
- 95 Till, M.S. and Ullmann, G. (2009) McVol—a program for calculating protein volumes and identifying cavities by a Monte Carlo algorithm. *J. Mol. Model.* 16, 419–429
- 96 Brady, G.P. and Stouten, P.F. (2000) Fast prediction and visualization of protein binding pockets with PASS. *J. Comput. Aided Mol. Des.* 14, 383–401
- 97 Levitt, D.G. and Banaszak, L.J. (1992) POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graph.* 10, 229–234
- 98 Tseng, Y.Y. *et al.* (2009) SplitPocket: identification of protein functional surfaces and characterization of their spatial patterns. *Nucleic Acids Res.* 37 (Web Server issue), W384–389
- 99 Tseng, Y.Y. and Li, W.-H. (2009) Identification of protein functional surfaces by the concept of a split pocket. *Proteins* 76, 959–976
- 100 Coleman, R.G. and Sharp, K.A. (2006) Travel depth, a new shape descriptor for macromolecules: application to ligand binding. *J. Mol. Biol.* 362, 441–458
- 101 Tripathi, A. and Kellogg, G.E. (2010) A novel and efficient tool for locating and characterizing protein cavities and binding sites. *Proteins* 78, 825–842
- 102 Kleywegt, G.J. and Jones, T.A. (1994) Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallogr. D: Biol. Cryst.* 50, 178–185
- 103 Xie, L. and Bourne, P.E. (2007) A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. *BMC Bioinformatics* 8 (Suppl. 4), S9
- 104 Ghersi, D. and Sanchez, R. (2009) Improving accuracy and efficiency of blind protein–ligand docking by focusing on predicted binding sites. *Proteins* 74, 417–424
- 105 Chang, D.T.-H. *et al.* (2005) MEDock: a web server for efficient prediction of ligand binding sites based on a novel optimization algorithm. *Nucleic Acids Res.* 33 (Web Server issue), W233–238
- 106 Powers, R. *et al.* (2006) Comparison of protein active site structures for functional annotation of proteins and drug design. *Proteins* 65, 124–135
- 107 Milik, M. *et al.* (2003) Common structural cliques: a tool for protein structure and function analysis. *Protein Eng.* 16, 543–552
- 108 Kinoshita, K. *et al.* (2002) Identification of protein functions from a molecular surface database, eF-site. *J. Struct. Funct. Genom.* 2, 9–22
- 109 Brylinski, M. and Skolnick, J. (2008) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc. Natl. Acad. Sci. U.S.A.* 105, 129–134
- 110 Najmanovich, R. *et al.* (2008) Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites. *Bioinformatics* 24, i105–i111
- 111 Shulman-Peleg, A. *et al.* (2008) MultiBind and MAPPIS: web servers for multiple alignment of protein 3D-binding sites and their interactions. *Nucleic Acids Res.* 36 (Web Server issue), W260–264
- 112 Park, K. and Kim, D. (2008) Binding similarity network of ligand. *Proteins* 71, 960–971
- 113 Minaai, R. *et al.* (2008) Method for comparing the structures of protein ligand-binding sites and application for predicting protein–drug interactions. *Proteins* 72, 367–381
- 114 Ausiello, G. *et al.* (2005) Query3d: a new method for high-throughput analysis of functional residues in protein structures. *BMC Bioinformatics* 6 (Suppl. 4), S5
- 115 Ramensky, V. *et al.* (2007) A novel approach to local similarity of protein binding sites substantially improves computational drug design results. *Proteins* 69, 349–357
- 116 Schalon, C. *et al.* (2008) A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins* 71, 1755–1778
- 117 Brakoulis, A. and Jackson, R.M. (2004) Towards a structural classification of phosphate binding sites in protein–nucleotide complexes: an automated all-against-all structural comparison using geometric matching. *Proteins* 56, 250–260
- 118 Shulman-Peleg, A. *et al.* (2004) Recognition of functional sites in protein structures. *J. Mol. Biol.* 339, 607–633
- 119 Snyder, K.A. *et al.* (2006) Domain-based small molecule binding site annotation. *BMC Bioinformatics* 7, 152
- 120 Jambon, M. *et al.* (2005) The SuMo server: 3D search for protein functional sites. *Bioinformatics* 21, 3929–3930
- 121 McGready, A. *et al.* (2009) Vicinity analysis: a methodology for the identification of similar protein active sites. *J. Mol. Model.* 15, 489–498
- 122 Weskamp, N. *et al.* (2007) Multiple graph alignment for the structural analysis of protein active sites. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 4, 310–320
- 123 Pons, J.L. and Labesse, G. (2009) @TOME-2: a new pipeline for comparative modeling of protein–ligand complexes. *Nucleic Acids Res.* 37 (Web Server issue), W485–491