JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

TRAPP: A Tool for Analysis of *Tran*sient Binding *P*ockets in *P*roteins

Daria B. Kokh,*[,†] Stefan Richter,[†] Stefan Henrich,[†] Paul Czodrowski,[‡] Friedrich Rippmann,[‡] and Rebecca C. Wade*[,†,§]

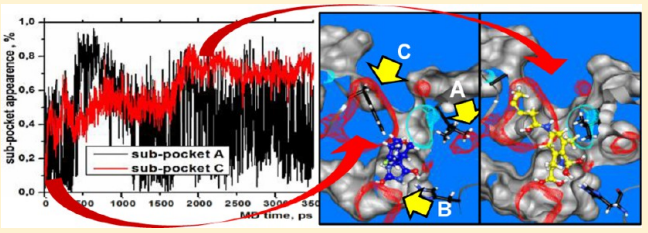
[†]Molecular and Cellular Modeling Group, Heidelberg Institute for Theoretical Studies (HITS), Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany

[‡]Global Computational Chemistry, Merck Serono, Merck KGaA, Frankfurter Strasse 250, 64293 Darmstadt, Germany

[§]Zentrum für Molekulare Biologie (ZMBH), Heidelberg University, 69120 Heidelberg, Germany

Supporting Information

ABSTRACT: We present TRAPP (TRAnsient Pockets in Proteins), a new automated software platform for tracking, analysis, and visualization of binding pocket variations along a protein motion trajectory or within an ensemble of protein structures that may encompass conformational changes ranging from local side chain fluctuations to global backbone motions. TRAPP performs accurate grid-based calculations of the shape and physicochemical characteristics of a binding pocket for each structure and detects the conserved and transient regions of the pocket in an ensemble of protein conformations. It also provides tools for tracing the opening of a particular subpocket and residues that contribute to the binding site. TRAPP thus enables an assessment of the druggability of a disease-related target protein taking its flexibility into account.



INTRODUCTION

Most drugs that target disease-related proteins bind at an active site used by a natural ligand. The binding site is usually located inside a concave pocket which allows the ligand to make many favorable contacts with the target protein. Shape complementarity of a compound and a binding pocket is therefore often the first requirement for selecting a drug candidate. During the last two decades, a variety of computational methods has been developed for predicting the position, shape, and interaction properties of ligand binding pockets from protein sequences and structures (see reviews[1,2]). The majority of these methods explore the shape of the protein surface to detect cavities and then score them according to their potential ability to bind small molecules. It is common to use a single static structure of a target protein for detection of a binding pocket, which may be available from crystallographic or NMR experiments or derived by comparative or ab initio modeling procedures. This is, however, often a rather inadequate representation for modeling ligand–protein recognition for the following reasons. First, even high-resolution crystal structures may harbor some uncertainty because of deviations of side chain orientations and solvent-exposed loops in the crystal state compared to solution.[3] Second, the geometry of the binding site in holo-structures may be affected by the ligand itself, resulting in a bias of compounds selected to those having steric properties similar to the cocrystallized ligand.[4] Finally, transitions between the free and bound forms of a target (or between various bound forms) may be driven by an essential backbone rearrangement, including distortion or repacking of the secondary structure, loop fluctuations, or even entire domain motions, which makes design of a potential ligand practically impossible without taking into account the dynamics of the binding site since even small side chain motions may alter the spatial and physical properties of a binding site and, therefore, the druggability of the pocket considered (see, for example, refs 5 and 6 for reviews).

It is widely recognized nowadays that neglect of protein flexibility adversely affects cross-docking results and is a source of biasing in virtual screening.[7−9] However, complete sampling of all protein degrees of freedom is currently not feasible, and the treatment of receptor flexibility is commonly restricted to side chain rotation or small backbone rearrangement at most. To take global backbone motion into account, protocols for sequential docking or screening against an ensemble of static structures (for example, cocrystallized with different ligands) are most commonly used.[10,5] Although such procedures enable the diversity of detected compounds to be enlarged, they are still strongly biased toward the steric and physical properties of binders that are already known. Knowledge about protein motions might help to overcome these limitations since the majority of bound conformations are also found to be accessible in free proteins.[11]

Indeed, methods to simulate protein dynamics and molecular docking procedures have been successfully combined in a number of studies. For several pharmaceutically relevant targets, binding subpockets that were absent in the unbound structures were found to open frequently during molecular dynamics (MD) simulations.[12−17] It has also been demonstrated that MD

Received: January 14, 2013

   dx.doi.org/10.1021/ci4000294 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

snapshots can improve blind virtual screening predictions.[18] To explore global structural rearrangements inaccessible by standard MD, more approximate methods, such as tCONCOORD[19] or normal-mode analysis (NMA), have been successfully applied for docking of known ligands to ligand-free structures that demonstrate large (with RMSD up to 10 Å) conformational changes relative to structures cocrystallized with ligands.[20−23]

To date, four computational tools for analysis of binding pocket variations in an ensemble of protein structures have been reported: EPOS^{BP} [13] designed for tracking some pocket properties along the MD trajectory (as volume, depth, and polarity); MDpocket[17] for identification of binding site and migration channels opening in MD trajectory or observed in an ensemble of crystallography structures, which also enables simulation of quite large range of pocket descriptors (for example, channel radius, hydrophobicity, polar, and apolar surface areas in addition to those included in the EPOS^{BP} method); PocketAnalyzer^{PCA} [16] that employs pocket clustering and PCA analysis for automated selection of diverse pocket conformers from MD trajectories; and Provar[24] that provides the probability that each atom or residue of a protein borders a predicted pocket in a set of protein structures. These tools employ different types of geometry-based approaches for protein cavity identification. Particularly, in PocketAnalyzer^{PCA},[16] a grid-based LIGSITE[25] algorithm is used, where the cavity position and the shape are directly bound to protein atomic coordinates mapped onto a rectangular grid. In contrast, the PASS algorithm,[26] employed in EPOS^{BP},[13] is pocket-based and virtually independent of protein orientation, where protein pockets are described by spheres that fill protein cavities. Provar employs either the PASS or LIGSITE programs to identify atoms that compose pocket boundary. In MDpocket,[17] which is based on FPocket,[27] a pocket is defined by a set of virtual atoms placed at the centers of alpha-spheres whose radii directly relate to the local curvature defined by the nearest four protein atoms. Then positions of alpha-spheres are mapped onto a grid, hereby converting identified cavity shape from the pocket-based to a grid-based representation.

None of these pocket detection methods (and others mentioned in ref 1) is completely suitable for our purposes for the following reasons. First, the pocket-centered approaches (employed in PASS or FPocket) seem to be quite unstable with respect to small alterations of atomic positions within a binding site. For example, the shift of the alpha-sphere center may be notably larger than the movement of the protein atom causing this shift (see the Supporting Information). This drawback becomes more pronounced if binding pockets are represented by spheres (pseudoatoms), since whether a pseudoatom is assigned to a particular pocket point or not depends on the chosen parameter threshold (for example, the smallest and largest radii of the alpha-spheres that are employed for pocket definition in FPocket/MDpocket). As a consequence, (i) details of the pocket structure within the pocket−protein boundary are lacking, even if the position and the approximate volume of the pocket are relatively well-identified; (ii) the number and/or position of pseudoatoms may change notably from one snapshot to another without significant alteration of the atomic positions, and only average statistical characteristics of the pocket are reliable. Another drawback of many pocket identification approaches is that they employ a set of parameters that must be adjusted to the target to be studied and can have a major influence on the results.[17] For example, the threshold values for the largest and smallest alpha-sphere radii to be considered in FPocket/MDpocket, the distance between pairs of vertices that must be taken into account in MDPocket, the threshold of the burial count that represents the extent to which the probe is excluded from solvent in PASS, and the minimal number of neighbors and minimal cluster size in PocketAnalyzer^{PCA},[16] all affect the shape and the size of the pocket detected and must be optimized with respect to the specific protein system under study. These parameters enable the method application to systems with substantially different cavity shapes (for example, for detection either of small closed or of shallow open cavities), but they introduce uncertainty into the results if the pocket characteristics change considerably along a trajectory. A further problem concerns identification of a particular binding pocket shape along a motion trajectory. Most approaches dealing with a static or almost static structure use the surrounding atoms as the reference for pocket identification, as for example in the pocket superimposing procedure proposed in ref 28. A similar approach is also applied in the EPOS^{BP} method, where pocket identification along a trajectory is based on the surrounding atoms. However, cavity dynamics may lead to changes in the set of residues surrounding a cavity, making pocket identification uncertain along a protein motion trajectory, especially when large conformational changes of a binding site take place.

Here, we present a new method for exploring binding pocket dynamics and identifying transient pockets and subpockets from a large collection of protein structures. The method is aimed at providing useful hints for the design of new ligands having distinct structural and interaction properties from ligands observed in cocrystallized protein structures. To this end, we have developed a new grid-based pocket detection program that performs calculations of the shape and physical properties of a binding site. Here, we use the term "binding site" to define a region of a protein within which a known ligand (native or non-native) binds and in whose vicinity we wish to identify transient pockets or subpockets that can be exploited in ligand design or optimization. The method has been implemented as part of a general platform, TRAPP (TRAnsient Pockets in Proteins), designed for tracking and analysis of pocket dynamics and identification of conserved and transient pockets or subpockets in protein motion trajectories or ensembles of structures. Two alternative algorithms have been implemented in TRAPP for the detection of transient pocket regions: (1) using principal component analysis (PCA) technique for analysis of the correlated pocket variations and (2) using computation of an averaged deviation of the pocket shape in a trajectory/ensemble of structures from that of a reference (crystal) structure. The latter algorithm will be referred to hereafter as the averaged relative deviation from a reference structure (ARDR) approach. Following identification, the transient pocket regions are split into subpockets using a clustering procedure, which enables structures where a particular transient subpocket of interest is open to be identified or the opening of a subpocket along a protein motion trajectory to be traced. TRAPP also allows the residues that contribute to a binding pocket to be traced. Additionally, we have defined pocket−pocket similarity and ligand−pocket complementarity measures that allow us to evaluate pocket simulation accuracy using experimental data on the pocket conformational variations. Finally, to facilitate analysis of the simulation results, TRAPP generates scripts for the PyMol[29] and Jmol,[30] molecular graphics software, which enable visualization of the shape and properties of the identified binding pockets as well as the position of conserved and transient pocket regions.

dx.doi.org/10.1021/ci4000294 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

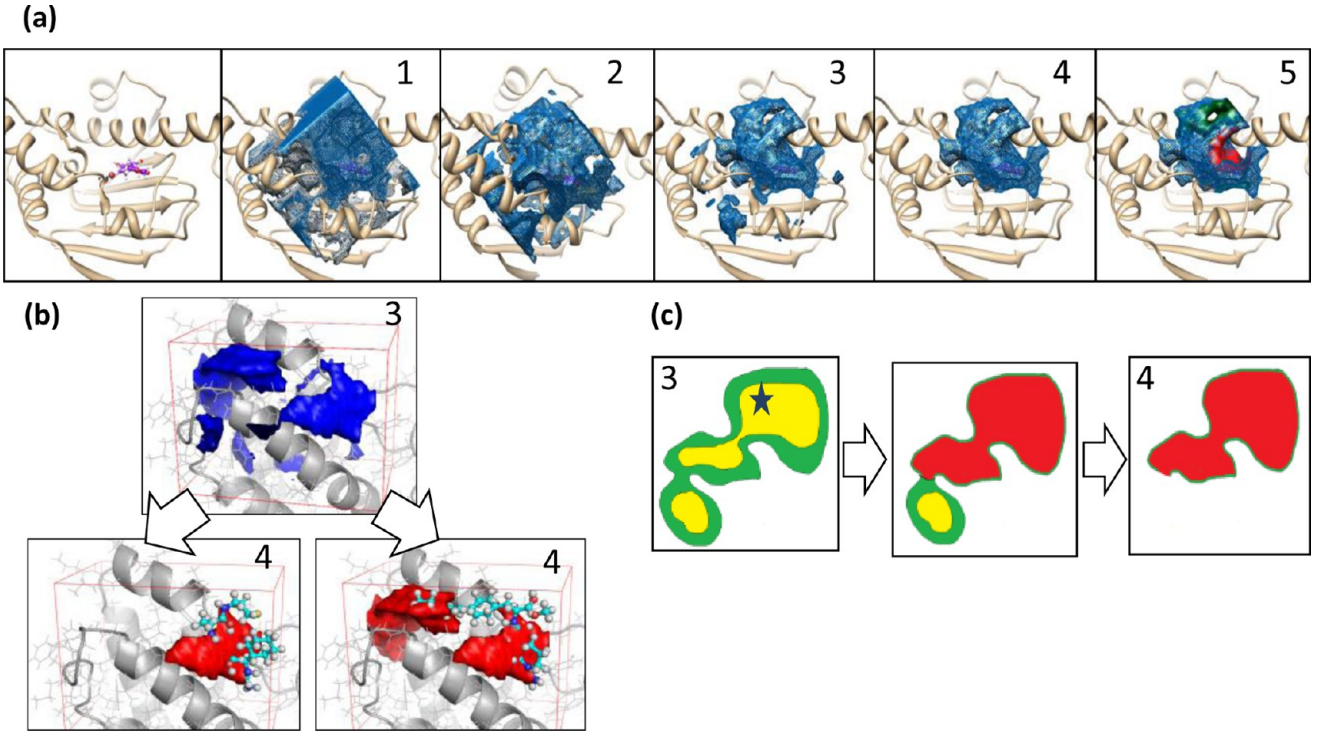


Figure 1. Illustration of the computational methods. (a) The procedure for identification of the protein cavity is illustrated for HSP90; the protein pocket distribution $G(r)$ is shown by a blue mesh at each computation step: (1) cavity regions detected around a binding site; (2) external protein regions have been eliminated; (3) the cavity boundary has been smoothed and small pockets have been removed; (4) the central binding pocket has been detected, and all small pockets that are not connected with the central one have been eliminated; (5) physicochemical properties (contacts with negatively and positively charged residues) of the pocket have been computed and are shown by green and red surfaces. (b) The binding site of the IL-2 protein (see Figure 2d) is represented by two separate pockets. In the lower plot, both subpockets are identified if they overlap with the reference ligand (right-hand plot); otherwise, only one central pocket is detected (left-hand plot) (reference ligands are shown in ball-and-stick representation). (c) Schematic illustration of the procedure for selecting a central pocket part: the star indicates a binding site center. A cavity is assigned to the central binding pocket (shown in red) if it contacts directly with the pocket region where $G(r) = 1$ (yellow area; the pocket shell where $0.6 < G(r) < 1$ is shown in green), the rest of the cavity is removed from the pocket distribution. The first and the last snapshots correspond to steps 3 and 4, respectively of the workflow shown in part a.

We have tested TRAPP on four well-characterized drug targets that demonstrate a wide range of conformational differences in their binding sites in structures cocrystallized with different ligands and determined at high resolution: the heat-shock protein HSP90 (HSP90), MAP kinase p38 (P38), interleukin-2 (IL-2), and aldose reductase (AR). For each target, we have chosen a set of holo-structures that encompass the major variations of the binding site.

To validate the pocket shape identification procedure, we estimated the steric compatibility of each cocrystallized ligand with the binding pocket obtained from its native structure and other crystal structures. Next, we evaluated the ability of TRAPP to identify conserved and transient regions from a set of holo-structures and from protein motion trajectories. For this purpose, we employed trajectories available in the MoDEL[31] database generated by standard MD simulation for 10 ns in explicit solvent for P38 and HSP90, and simulated 6 ns MD trajectories of the AR and IL-2 proteins. Additionally, we employed the tCONCOORD[19] approach to generate ensembles of structures with substantial backbone rearrangements in HSP90 and P38, which are inaccessible in short-time MD simulations. The MD trajectories, tCONCOORD ensembles, and sets of holo-structures were then used for TRAPP analysis of the pocket dynamics in the test targets and comparison of the transient pockets identified. These examples show which characteristics of the pocket dynamics can be derived from TRAPP analysis.

MATERIALS AND METHODS

1. TRAPP Workflow. TRAPP incorporates a workflow (a diagram is shown in Figure 1S of the Supporting Information) consisting of the following steps: first, a set of binding site residues is defined for a reference protein structure, which provides a reference for the binding site position in a protein and is used for sequential alignment and superposition of protein structures from an ensemble or a trajectory. Then the shape and physicochemical properties of the binding site region are computed for each structure and stored on a grid. A set of pocket shapes is used for analysis of the pocket dynamics, computation of pocket similarity and pocket−ligand complementarity, as well as for data visualization. The software for the TRAPP workflow consists of several modules, which have been implemented in Python using the NumPy[32] and BioPython[33] libraries and are described in the next section.

Definition of the Set of Binding Site Residues and Structure Superposition. For the application of TRAPP, it is assumed that at least one experimentally determined target structure is available at high resolution (this will be referred to hereafter as the reference structure) and that the position of the binding site is known. A binding site of a reference structure is determined by a set of residues that can either be defined by the user, automatically selected using the position of a ligand relative to the reference structure, or derived from coordinates of the binding site center provided by the user. As a reference ligand, an

dx.doi.org/10.1021/ci4000294 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

experimentally determined ligand from a holo-structure may be used, or any set of atoms placed inside the binding site to be analyzed. Residues that have at least one non-hydrogen atom within a user-defined distance around the reference ligand or binding site center are assigned to the binding site.

A set of binding site residues of a reference structure is used to identify a particular binding site for each ensemble/trajectory. Here, we assume that all structures in an ensemble/trajectory have an identical primary structure, i.e. have identical sequence, while the sequences in different ensembles/trajectories to be analyzed need not be identical or complete. We use sequential alignment of the binding site residues in a reference structure to corresponding residues in a representative one of an ensemble (the alignment procedure is described in the Supporting Information) to find binding site residues in each ensemble/trajectory. Then all members of the ensemble/trajectory are overlaid by superimposing the backbone atoms of the binding site residues with those of the reference structure. Next, a suitably sized grid for subsequent pocket computation is constructed such that all the binding site residues are positioned inside the grid in each structure in the ensemble/trajectory.

Pocket Detection. The grid-based pocket detection approach has been designed to meet the following requirements: (1) enable processing of multiple structures within a reasonable computational time; (2) provide simulation accuracy high enough to detect even small alterations of side chain positions, i.e. small variations in a protein structure must give rise to comparably small variations in the pocket shape and properties; (3) ensure that a wide spectrum of cavity shapes and sizes can be identified without any adjustment of input parameters. In particular, all cavities that are associated with a binding site in a trajectory or in an ensemble must be captured.

The principles of our binding pocket detection algorithm are 3-fold: first, a pocket should not overlap with protein atoms considering their atomic radii; second, the pocket curvature must permit at least one atom to fit inside, but not be infinitely small (i.e., not a flat surface); third, the binding pocket is characterized by a boundary region (shell) where a ligand, when positioned inside, can form noncovalent bonds to protein atoms.

A pocket is described by a distribution function ($G(\boldsymbol{r}, p)$, where $\boldsymbol{r} = (x, y, z)$, and p is a particular protein structure in an ensemble of snapshots) mapped onto the previously defined grid. A system of coordinates and an origin of the vector $\boldsymbol{r}$ are defined by atomic coordinates in a PDB file of a reference structure and remain fixed in all simulations. The grid spacing must be small enough to cover the van der Waals interaction distance with several grid points (we used 0.75 Å in the present calculations). The procedure for computing $G(\boldsymbol{r}, p)$ is illustrated in Figure 1a. First, all atoms of a binding site are scanned and the cavity distribution function is computed as

$$G(\boldsymbol{r}, p) = \begin{cases} 0 & \text{if } l < 0.5\sigma_{\text{Prot}} \text{ for at least one protein atom} \\ 1 & \text{if } l > 2.0\sigma_{\text{Prot}} \text{ for all protein atoms} \\ \min(1 - e^{-[(l-0.5\sigma_{\text{Prot}})/0.5\sigma_{\text{Prot}}]}) & \text{taken over all protein atoms otherwise} \end{cases}$$

where l is the distance between the grid point $\boldsymbol{r}$ and a protein atom, σ_{Prot} is an atomic interaction radius (that is analogous to the corresponding parameter in the Lennard-Jones potential; we have used σ_{Prot} = 1.2, 1.5, and 1.7 Å for H, O/N, and all other atom types, respectively; if the structures do not contain hydrogen atoms, σ_{Prot} = 1.8 Å is used for all atom types, see the Supporting Information for details). The iso-surface at the value of $G_{\text{vdW}} \sim 0.6$ (which corresponds to a surface at $\sim\sigma_{\text{Prot}}$) defines the van der Waals surface of the protein around a pocket. Pocket points described by $G(\boldsymbol{r}, p) = 1$ and $G_{\text{vdW}} < G(\boldsymbol{r}, p) < 1$ correspond to the internal and boundary (shell) regions, respectively. The latter indicate pocket regions where ligand atoms can form noncovalent bonds with protein atoms.

Next, the exterior and interior regions of cavities are separated by using a buriedness index,[25] $\alpha(\boldsymbol{r})$, which is computed for all pocket points where $G(\boldsymbol{r}, p) > 0$ and is the fraction of the equally distributed points on a sphere of a radius $|\boldsymbol{u}| \equiv |\boldsymbol{r}' - \boldsymbol{r}|$, where $G(\boldsymbol{r}', p) < 1$. i.e. that are inside a protein van der Waals surface or in a pocket shell. A reasonable choice of the sphere radius $|\boldsymbol{u}|$, is the cutoff used for computing van der Waals interactions (~7 Å). Thus, a value of α_0 in the range of 0.5 to 0.6 corresponds to points in a shallow cavity on a protein surface with a radius smaller than $\sim|\boldsymbol{u}|$. The larger cavities will be assigned to the protein exterior. On the other hand, small cavities on the protein surface may be missed in this procedure. To improve the accuracy of the protein surface detection, we used a two-step calculation of the buriedness: first with half of the vector length, $|\boldsymbol{u}|/2$, and then, if $\alpha < \alpha_0$, repeating the procedure using the vector length $|\boldsymbol{u}|$. The total number of vectors used is 64 for the first step and 144 for the second one. Further increasing the number of vectors does not improve accuracy, while a smaller number reduces computational time at the expense of accuracy. Grid points with $\alpha < \alpha_0$ are identified as being in the protein exterior and are labeled with $G(\boldsymbol{r}, p) = -1$. The region defined by $G(\boldsymbol{r}, p) \geq 0$ will hereafter be referred to as the protein envelope.

After pocket detection, small buried protein cavities (i.e., cavities whose volume is too small to place at least one non-hydrogen atom of radius of 1.5 Å inside) are eliminated by spatial averaging of the pocket distribution $G(\boldsymbol{r}, p)$ over a sphere of radius 3 Å. Practically, the grid points where $\langle G(\boldsymbol{r}, p)\rangle < G_{\text{vdW}}$ are set to 0. This procedure also allows the pocket boundaries to be smoothed out (see Figure 1a, step 3).

Finally, the central pocket is selected by using a procedure that mimics filling a cavity. Namely, a starting grid node is defined as the center of the binding pocket (the geometric center of the reference ligand or, if this center is outside the cavity, the nearest pocket point). All grid nodes around the starting node are assigned to the subgrid and, if their grid density value G_{vdW}, also to the central pocket. Then, the procedure is repeated for all direct neighbor grid nodes of the subgrid. Thus, the subgrid increases up to the size of the whole grid. At each step, a particular node is assigned to the central pocket if the grid density value $G(\boldsymbol{r}, p) > G_{\text{vdW}}$ and if it has at least one direct neighbor grid node that belongs to the central pocket (see Figure 1c). Thus, all cavities that are not directly connected to the central pocket are excluded (as illustrated in Figure 1a, step 4), while only the pocket region located around the starting point is selected. In some cases, however, several separate pockets can be associated with a binding site (such an example is shown in Figure 1b). To detect several separate pockets, grid regions occupied by a reference ligand are scanned. If some grid nodes of this region belong to a protein pocket ($G(\boldsymbol{r}, p) > G_{\text{vdW}}$) but are not assigned to the central pocket, the procedure is repeated from the new starting point. This procedure enables protein pockets separated into several subpockets to be detected if their position is defined by a reference ligand (see Figure 1b). The user may, however,

dx.doi.org/10.1021/ci4000294 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

choose to skip the central pocket selection procedure completely and thus to include all protein cavities in the subsequent analysis of binding pocket dynamics. This is recommended if transient pocket regions may appear as separate small pockets in some structures and no information about possible pocket dynamics is available in advance.

The time required for the complete pocket detection procedure using the default parameters strongly depends on the grid size. For 20 × 20 × 20 grid points (a pocket of ∼7 Å radius with a 0.75 Å grid spacing), the pocket detection procedure takes less than 1 s on a desktop workstation. However, simultaneous computation of the physicochemical properties of the pocket (see below) may increase the computational time several fold. The parameters used in the present work and also as default in TRAPP are summarized in the Supporting Information.

Computation of the Physicochemical Properties of a Cavity. Simultaneously with the pocket shape identification, some physicochemical properties of the binding site can be computed as distribution functions and stored in separate grids. Currently, we have implemented the three distribution functions that show pocket contact with

(i) residues of specific types

$$G^{\mathrm{resi}}(\boldsymbol{r}, p) = \sum_{\mathrm{resi}} 1/n(\mathrm{resi}) \sum_{i}^{n(\mathrm{resi})} \mathrm{e}^{-[(l_i - R_{\mathrm{resi}})R_{\mathrm{resi}}^{-1}]^2}$$

(ii) charged atoms

$$G^{\mathrm{ch}}(\boldsymbol{r}, p) = \sum_{\mathrm{resi}} \sum_{i}^{n(\mathrm{resi})} q_i e^{-[(l_i - R_{\mathrm{ch}})R_{\mathrm{ch}}^{-1}]^2}$$

(iii) specific atoms

$$G^{\mathrm{at}}(\boldsymbol{r}, p) = \sum_{i} \mathrm{e}^{-[(l_i - R_{\mathrm{at}})R_{\mathrm{at}}^{-1}]^2}$$

where summation is over all non-hydrogen atoms (of a total number n) in a side chain of the user-defined residue type resi (i) and/or over all user-defined atoms i (ii); l_i is the distance from atom i to a particular grid point $\boldsymbol{r}$; (R_{resi}, R_{ch}, R_{at}) are characteristic user-defined interaction distances; and $G^{\mathrm{resi}}(\boldsymbol{r}, p) = G^{\mathrm{ch}}(\boldsymbol{r}, p) = G^{\mathrm{at}}(\boldsymbol{r}, p) = 0$ where $G(\boldsymbol{r}, p) \leq 0$.

An overlap between a pocket distribution $G(\boldsymbol{r}, p)$ and a distribution function of particular pocket properties provides a descriptor that enables tracking the presence of specific residue and/or atom types in the binding site along a trajectory (this is demonstrated below for the case of the MD trajectory of AR protein).

Analysis of the Cavity Dynamics along a Protein Motion Trajectory. Three ways to characterize the pocket dynamics are implemented in TRAPP:

(i) An averaged pocket distribution over a set of structures $\{p\}$, which shows a fraction of structures where a pocket appears in a particular point. It is computed as $G^{\mathrm{A}}(\boldsymbol{r}) = \langle G^{\mathrm{A}}(\boldsymbol{r}, p)\rangle_p$, where $G^{\mathrm{A}}(\boldsymbol{r}, p)$ is

$$G^{\mathrm{A}}(\boldsymbol{r}, p) = \begin{cases} 1, & \text{if } G(\boldsymbol{r}, p) > G_{\mathrm{vdW}} \\ G(\boldsymbol{r}, p)/G_{\mathrm{vdW}} & \text{if } 0 < G(\boldsymbol{r}, p) \leq G_{\mathrm{vdW}} \end{cases}$$

The pocket regions that appear in every structure of an ensemble can be visualized with an iso-surface at $G^{\mathrm{A}}(\boldsymbol{r}) \sim 1$ and will be referred to hereafter as the conserved pocket regions.

(ii) An average deviation of pocket distributions, $G(\boldsymbol{r}, p)$, in an ensemble of structures, $\{p\}$, relative to that of a reference structure, $G(\boldsymbol{r}, p = 0)$ (average relative deviation from a reference (ARDR)). The ARDR transient pocket distribution function is computed as $G^{\mathrm{t}}(\boldsymbol{r}) = \langle G^{\mathrm{t}}(\boldsymbol{r}, p)\rangle_p$, where $G^{\mathrm{t}}(\boldsymbol{r}, p) = \pm 1$ if the pocket variation relative to the reference structure $|G(\boldsymbol{r}, p) - G(\boldsymbol{r}, p = 0)| > G_{\mathrm{vdW}}$ and $G(\boldsymbol{r}, p = 0) \geq 0$ (i.e., for a reference structure $\boldsymbol{r}$ is inside a protein). Positive/negative values of $G^{\mathrm{t}}(\boldsymbol{r})$ indicate appearing/disappearing pocket regions, respectively. Practically, the $G^{\mathrm{t}}(\boldsymbol{r})$ value roughly defines the fraction of structures in which a particular pocket region has been found but does not exist in the reference structure or, vice versa, does exist in the reference structure but disappears in some snapshots of a trajectory/ensemble. For visualization, we have used iso-surfaces at $G^{\mathrm{t}} = \pm 0.5$, which corresponds to the appearance of a particular transient pocket region in about half of all structures.

Transient regions often consist of subpockets that may come from roughly independent motion of different elements of a binding site. It is, therefore, useful to analyze them separately. For this, we implemented a two-step clustering procedure. First, a hierarchical clustering method is applied to divide the ARDR $G^{\mathrm{t}}(\boldsymbol{r})$ regions that do not contact each other at a particular iso-value (i.e., occurrence) into separate subpockets, $G_i^{\mathrm{t}}(\boldsymbol{r})$. Additionally large, complex ARDR transient pockets are further divided into smaller, compact ones, using a k-means clustering procedure. Finally, the projection of each subpocket distribution function, $G_i^{\mathrm{t}}(\boldsymbol{r})$, onto the pocket distribution, $G(\boldsymbol{r}, p)$, for each structure p in an ensemble or a snapshot of a trajectory, is computed. The resultant overlap function may vary from zero (a subpocket is completely closed) to 1 (subpocket is completely open), and thus provides a simple way to identify structures or snapshots where a particular subpocket is open or to trace the appearance of a subpocket along a protein motion trajectory.

(iii) Principal component analysis, PCA, of the pocket distribution in a trajectory/ensemble describes pocket variations by a set of vectors. PCA is applied to the same set of $G(\boldsymbol{r}, p)$ distributions. The first few PCA vectors show pocket regions with maximum correlated variation from the pocket distribution averaged over all structures in a trajectory/ensemble. Elements of the pocket covariance matrix for a set of N pocket distributions are defined as

$$G(\boldsymbol{r}_i, \boldsymbol{r}_j) = \frac{1}{N-1} \sum_{p=1}^{N} (G'(\boldsymbol{r}_i, p) - G'(\boldsymbol{r})) (G'(\boldsymbol{r}_j, p) - G'(\boldsymbol{r}))$$

where

$$G'(\boldsymbol{r}, p) = \begin{cases} 1, & \text{if } G(\boldsymbol{r}, p) > G_{\mathrm{vdW}} \text{ and } G(\boldsymbol{r}, p) \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

and

$$G'(\boldsymbol{r}) = \langle G'(\boldsymbol{r}, p)\rangle_p$$

Elements $G'(\boldsymbol{r})$ that are zero for all p are eliminated from the correlation matrix to reduce the matrix size. Negative and positive PCA eigenvectors show regions of correlated appearance and disappearance of the pocket, respectively. The motion of the whole pocket can thus be reduced to a small set of PCA vectors if the sum of their eigenvalues is close to 1. In this case, the method enables the most significant changes in the pocket shape to be described by a few PCA vectors.

dx.doi.org/10.1021/ci4000294 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

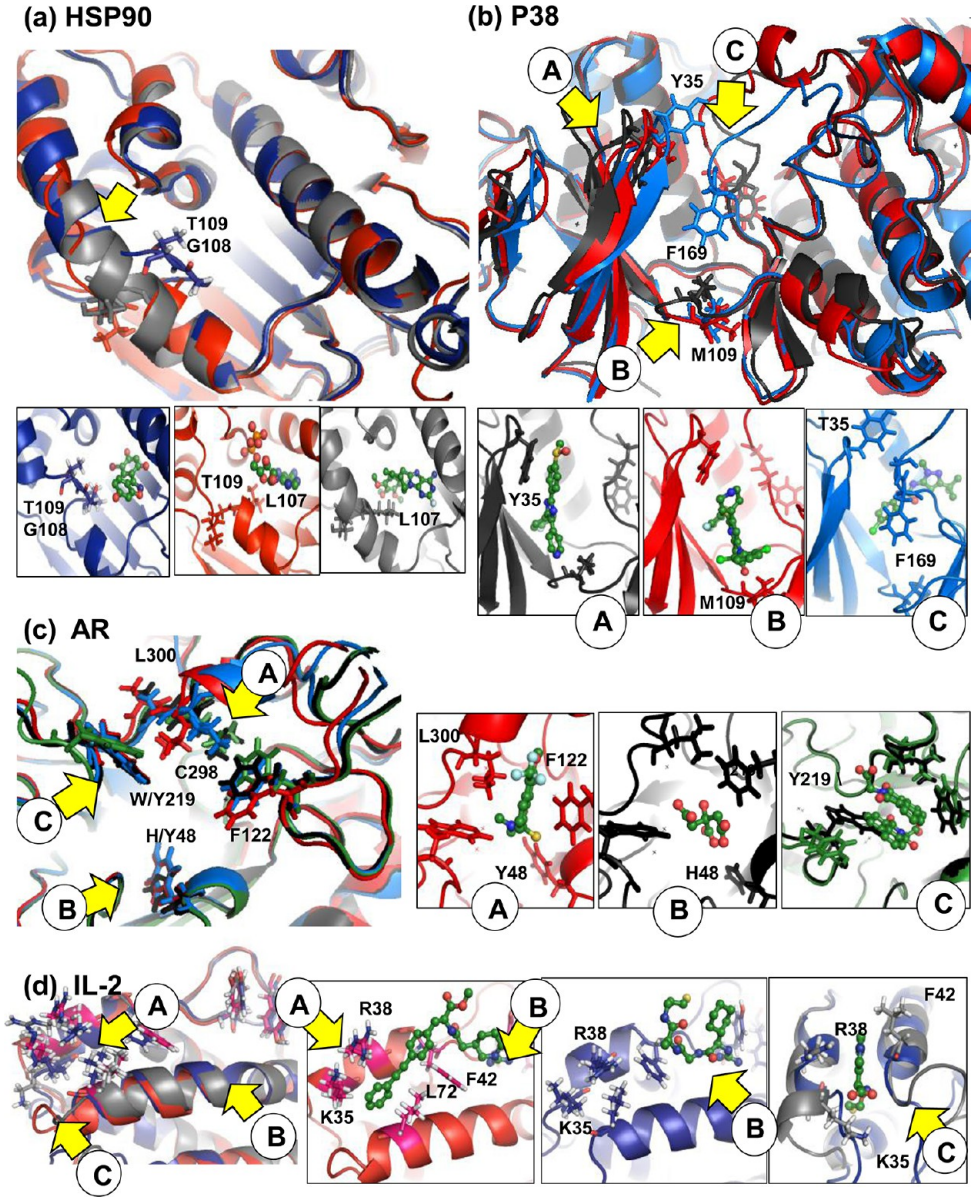


Figure 2. Conformational variations of active sites observed in crystallographically determined holo-structures of the four test proteins used for the method evaluation. Transient pocket regions are marked by yellow arrows and shown in the insets with representative cocrystallized ligands in ball-and-stick representation. Flexible side chains whose position affects the shape of the pocket are shown in stick representation. (a) HSP90. The transient binding pocket appears due to the partial unwinding of the α-helix: structures 3bm9, 1byq, and 1uyd are shown in blue, red, and gray, respectively and in the insets along with the corresponding ligands BXZ (3bm9), ADP (1byq), and PU8 (1uyd); the residue numbering is as in PDB file 1byq. (b) P38. Three transient subpockets denoted by A, B, and C arise from the motions of the beta-sheet and two loops: structures 1kv1, 1ouy, and 1a9u are shown in blue, red, and black with cocrystallized ligands BMU (1kv1), 094 (1ouy), and SB2 (1a9u) in the insets; the residue numbering is as in PDB file 1a9u. (c) AR. Side chain and backbone fluctuations give rise to the transient regions, A, B, and C: structures 1us0, 2acu, 1ah3, and 2az1 are shown in blue, black, red, and green; the structures with ligands in the insets are TOL (1ah3), CIT (2acu), and ALR (2az1); the residue numbering is as in PDB file 1ah0. (d) IL-2. Surface transient pocket A opens due to side chain flexibility; subpocket B is conserved, and deeply buried subpocket C appears due to helix distortion; structures 1m48, 1m4b, and 1m4a are shown in red, blue, and gray, and in the insets with their corresponding ligands FRG (1m48), NMP (1m4b), and MPE (1m4a). The residue numbering is as in PDB file 1m48.

Measure of the Ligand–Pocket Steric Complementarity. We express ligand–pocket complementarity by the overlap between the region occupied by a ligand ($L(\boldsymbol{r}) > 0$) and a pocket ($G(\boldsymbol{r}, p) > 0$). Accordingly, the ligand–pocket overlap (LPO) index is calculated as

$$S^{\mathrm{LPO}} = \frac{\sum_{\boldsymbol{r}} Q(\boldsymbol{r})}{\sum_{\mathbf{r}} q(\boldsymbol{r})}$$

where the summation is over all grid points $\boldsymbol{r}$ and

$$Q(\mathbf{r}) = \begin{cases} 1, & \text{if } G(\boldsymbol{r}, p) > 0 \text{ and } L(\boldsymbol{r}) > 0 \\ 0, & \text{if } G(\boldsymbol{r}, p) < 0 \text{ or } L(\boldsymbol{r}) = 0 \\ -1, & \text{if } G(\boldsymbol{r}, p) = 0 \text{ and } L(\boldsymbol{r}) > 0 \end{cases}$$

The index is normalized to the volume occupied by the ligand part positioned inside the protein envelope:

dx.doi.org/10.1021/ci4000294 | J. Chem. Inf. Model. XXXX, XXX, XXX–XXX

$$q_i = \begin{cases} 1, & \text{if } G(\boldsymbol{r}, p) \geq 0 \text{ and } G^{\mathrm{L}}(\boldsymbol{r}) > 0 \\ 0, & \text{otherwise} \end{cases}$$

where the ligand distribution function $G^{\mathrm{L}}(\boldsymbol{r})$ is defined as 1 within a distance of $0.5\sigma_{\mathrm{Lig}}$ of each ligand atom center and 0 otherwise. An interaction radius of σ_{Lig} = 1.5 Å was used for all types of ligand atoms.

Thus, according to definition of the protein and ligand distribution functions, $G(\boldsymbol{r}, p)$ and $G^{\mathrm{L}}(\boldsymbol{r})$, a contribution of a particular grid point, $Q(\boldsymbol{r})$, to the LPO index is negative if the distance between a ligand and a protein atom centers at the point $\boldsymbol{r}$ is less than $0.5(\sigma_{\mathrm{Lig}} + \sigma_{\mathrm{Prot}})$ and positive if the grid point belongs to the protein pocket and the ligand simultaneously. Any parts of a ligand that are positioned outside the protein ($G(\boldsymbol{r}, p) < 0$) are not taken into account. S^{LPO} reaches a maximum value of 1 if the ligand does not clash with any protein atoms and −1 if the ligand overlaps completely with the protein. One should note that the LPO index is primarily designed for validation of pocket simulation procedure and does not provide a measure of the protein−ligand binding energy (particularly, the LPO value equals 1 also when a small ligand is placed in the center of a large pocket and does not have direct contacts with protein atoms).

Visualization of Pocket Dynamics and Physical Characteristics. The pocket shape and properties for each snapshot/structure, as well as conserved/transient pocket regions and subpockets, are stored in dx grid format and can be visualized using a molecular visualization program, such as PyMol,[29] VMD,[34] Jmol,[30] or Chimera.[35] One should note, however, that visualization of pocket dynamics using a grid representation requires either uploading many grids simultaneously (which is memory consuming and not practically feasible for a trajectory of more than several tens of snapshots) or sequentially (which can be quite time-consuming). Besides, many molecular graphics programs do not provide the possibility to upload different grids into frames of a trajectory. We therefore also implemented a cavity representation using pseudoatoms for visualization. For each structure, the pocket volume defined by an iso-value of G_{vdW} is filled with spheres of radius R_{LJ} (the radius value can also be defined by the user, the default value is 1.5 Å) and saved as a PDB file with atom names describing the physical properties of the corresponding grid region. The generated PDB files are combined into a pocket motion trajectory that can then be visualized.

2. Test Target Proteins. The HSP90 heat-shock protein is a potential target for cancer treatment by inhibition of its ATPase activity. The ADP binding pocket contacts an unstable α-helix that undergoes distortion in the middle, converting into two short helices connected by a loop as shown in Figure 2a. The complete α-helix is associated with the most open binding pocket and is observed in the bound form cocrystallized with the purine-based inhibitors, PU8 [9-butyl-8-(2-chloro-3,4,5-trimethoxy-benzyl)-9*H*-purin-6-ylamine] and PU1 [8-(2-chloro-3,4,5-trimethoxy-benzyl)-2-fluoro-9-pent-4-ylnyl-9*H*-purin-6-ylamine] (structures 1uyd and 1uyf, respectively; throughout this paper, we refer to ligands by their residue names in the corresponding PDB files of the cocrystallized structures). In the apo-structure, 1yes, and in the majority of the holo-structures, the α-helix is distorted and the side chains of its loop part move into the binding pocket, closing it to differing extents. Specifically, L107 partially occupies the pocket space in the holo-structures 1byq/1ior/3bm9 cocrystallized with adenosine-diphosphate, ADP, and in the apo-structure, 1yer. In the holo-structures, 2vci/2vcj/3bm9, the T109 side chain is also oriented into the binding pocket resulting in the most closed pocket conformations.

The p38 mitogen-activated protein (MAP) kinase (P38) is an intracellular signaling protein involved in cytokine synthesis and is an important target for the treatment of osteoarthritis and inflammation. P38 has a deep pocket that contacts a flexible β-sheet (ATP-binding site labeled by A in Figure 2b) and two flexible loops (subpockets labeled by B and C in Figure 2b). In crystal structures 1kv1 and 1kv2 (not shown in Figure 2b), a transient subpocket C (binding site for a diaryl urea class of inhibitors, indicated by C in Figure 2b) appears due to reorganization of the activation loop containing the highly conserved Asp−Phe−Gly motif. Furthermore, the flip of the M109 residue accompanied by a small backbone shift also opens a part of the pocket, B (see Figure 2b), that is occupied by a dihydropyrido-pyrimidine inhibitor (094) in the 1ouy holo-structure. It is noteworthy that most holo-structures exhibit only one open transient subpocket, suggesting correlated motion of the binding site.

Aldose reductase, AR, is an NADPH-dependent enzyme, whose inhibitors are of interest as potential therapeutic agents for preventing long-term diabetic complications. The AR binding site adopts slightly different shapes in experimental structures due to side chain rotation and backbone fluctuations. Specifically, a transient part observed in structures 1us0, 1ah3, and 1iei (complexed with the inhibitors IDD594 [LDT], tolrestat [TOL], and zenarestat [ALR], respectively; structures 1us0 and 1ah3 are illustrated in Figure 2c) arises from motion of L300 (labeled A in Figure 2c); while inhibitor IDD384 (I87) from 1ek0 requires a small rotation of T48 (mutated to H48 in the PDB structure 2acu; see Figure 2c) for binding (transient region B). Motion of the residue W219 in C298A/W219Y mutant AR (1az1) opens space for the two-molecule alrestatin inhibitor, ALR, as shown in the insets in Figure 2c (transient region C). The subpocket C is closed when mutant AR is cocrystallized with CIT (1az2, not shown in figures). Recently Craig et al.[16] have analyzed conformational variation of the AR binding pocket along a 10 ns MD trajectory and found that fluctuations of the C298 side chain lead to opening of a novel subpocket that has not yet been observed experimentally and might be used for the design of new binders.

Interleukin 2, IL-2, is a cytokine. It is an example of a target with a solvent-exposed surface binding site used for inhibiting protein−protein binding.[13,36] An inhibitor that can bind to the interfacial binding site occupying a groove that is not present in the apo structure of the protein was found empirically.[37] The IL-2 binding site has two subpockets, labeled A and B in Figure 2d. The shape of pocket A is highly adaptive[36] due to the flexibility of several side chains around R38/L72 and was used as a test example in several developments of pocket detection procedures.[15,20,38,13] Additionally, some global protein flexibility is observed outside the main binding site, which includes partial unwinding of a helix and loop motion (see structure 1m4a shown in gray inset in Figure 2d). Loop motion together with flipping of the side chain of K35 leads to closure of the deeply buried subpocket C as observed in the 1m4a holo-structure.

Structure Preparation. All water molecules and ligands were removed from the protein structures used in the simulations (the cofactor $NADP^+$ in AR was kept as in the crystal structure). Hydrogen atoms were added to all protein structures assuming standard protonation states at pH 7 and missing side chains and loops (that did not belong to a binding site) were modeled into the 1kv1 and 1kv2 structures of P38 using the Prime module of

dx.doi.org/10.1021/ci4000294 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

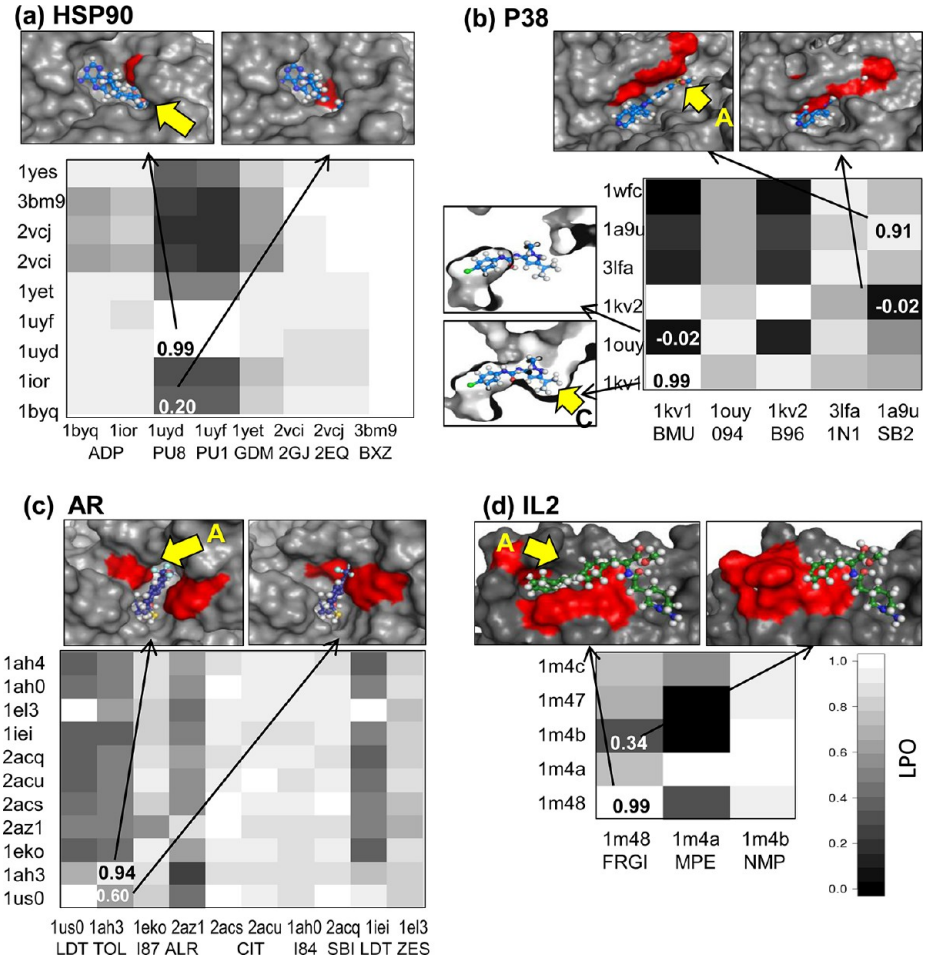


Figure 3. Ligand−pocket overlap LPO matrices for the four test targets: horizontal and vertical axes correspond to ligand and protein (employed for pocket generation) structures, respectively. The negative LPO values are set to 0. The ligand position in the binding pockets is visualized in the insets for several representative cases (the pocket notation corresponds to that used in Figure 1), and the magnitude of the ligand−pocket overlap is given for the corresponding matrix element. The residues contributing most to ligand−protein clashes are highlighted in red.

the Maestro Schrödinger software.[39] In the ligand−pocket overlap calculations, all holo-structures with ligands were superimposed (see the TRAPP workflow section), and the protein and ligand structures were stored in separate PDB files. Next, a set of residues within a distance of 3.0 Å from a reference ligand were assigned to a binding site. As reference ligand structures, we have used the ADP inhibitor from 1byq (HSP90) and several ligands combined in one PDB file for other targets: FRG (1m48) and MPE (1m4a) for IL-2, B96 (1kv2) and SB2 (1a9u) for P38, and LDT (1us0), TOL (1ah3), and ALR (2za1) for AR. This resulted in the following sets of binding site residues: (HSP90) N51, S52, A55, K57, D93, I96, M98, N106−I120, G135-F138, T184; (P38) V30, Y35, V38, A51, K53, R67, R70, E71, L74, L75, M78, V83−G86, Y104−G110, I141, I146, H148, A157, A167−A173; (AR) W20, V47, H48, K77, W79, H110−T113, F115, F122, R217−W219, C298−C303; and (IL-2) N33−L36, R38, M39, T41−F44, P65, V69, L72, A73 (residue numbering as in PDB files 1byq, 1a9u, 1ah0, and 1m48 for HSP90, P38, AR, and IL-2, respectively).

MD Simulations. Standard explicit solvent MD simulations of the AR and IL-2 proteins (with starting PDB structures, 2acu and 1m4b, respectively) were carried out using the GROMACS 4.5.3 software.[40] After deleting the ligand from the structure and adding all hydrogen atoms, the protein was immersed in a box of TIP3P water molecules extending at least 1 nm beyond the protein surface. The OPLS-AA force field[41] for IL-2 and the GROMOS[42] force field for AR (which includes parameters for the cofactor $NADP^+$) were used together with the particle mesh Ewald method for computing electrostatic interactions with a Coulomb cutoff distance of 12 Å. The protein structure was energy minimized with 200 steepest descent (SD) steps (for water only), and 500 SD steps and 500 conjugate gradient steps (whole system). Each system was gradually heated from 10 to 300 K in 1 ns using a time step of 2 fs and then equilibrated at 300 K for 2 ns in the *NVT* ensemble (the number of particles N, the volume V, and the temperature T are fixed in simulations). All bond lengths were constrained using the Linear Constraint Solver (LINCS) algorithm.[43] Temperature was maintained with a Nose-Hoover thermostat (τ = 0.1 K). Finally, 6 ns MD production simulations were carried out. Since side chain and loop motions in the binding site are readily observed over the picosecond time-scale, we extracted snapshots at 2 ps intervals from the MD trajectory (3000 snapshots for each target) for the TRAPP analysis.

Trajectories of HSP90 (1yer) and P38 (2acu) downloaded from the MoDEL server[31] were obtained from 10 ns explicit solvent MD simulations that used AMBER8.0 (with the PARM99 force field and the TIP3P water model). Since the downloaded trajectories do not contain hydrogen atoms, we added them to the protein structures using the GROMACS tool

dx.doi.org/10.1021/ci4000294 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

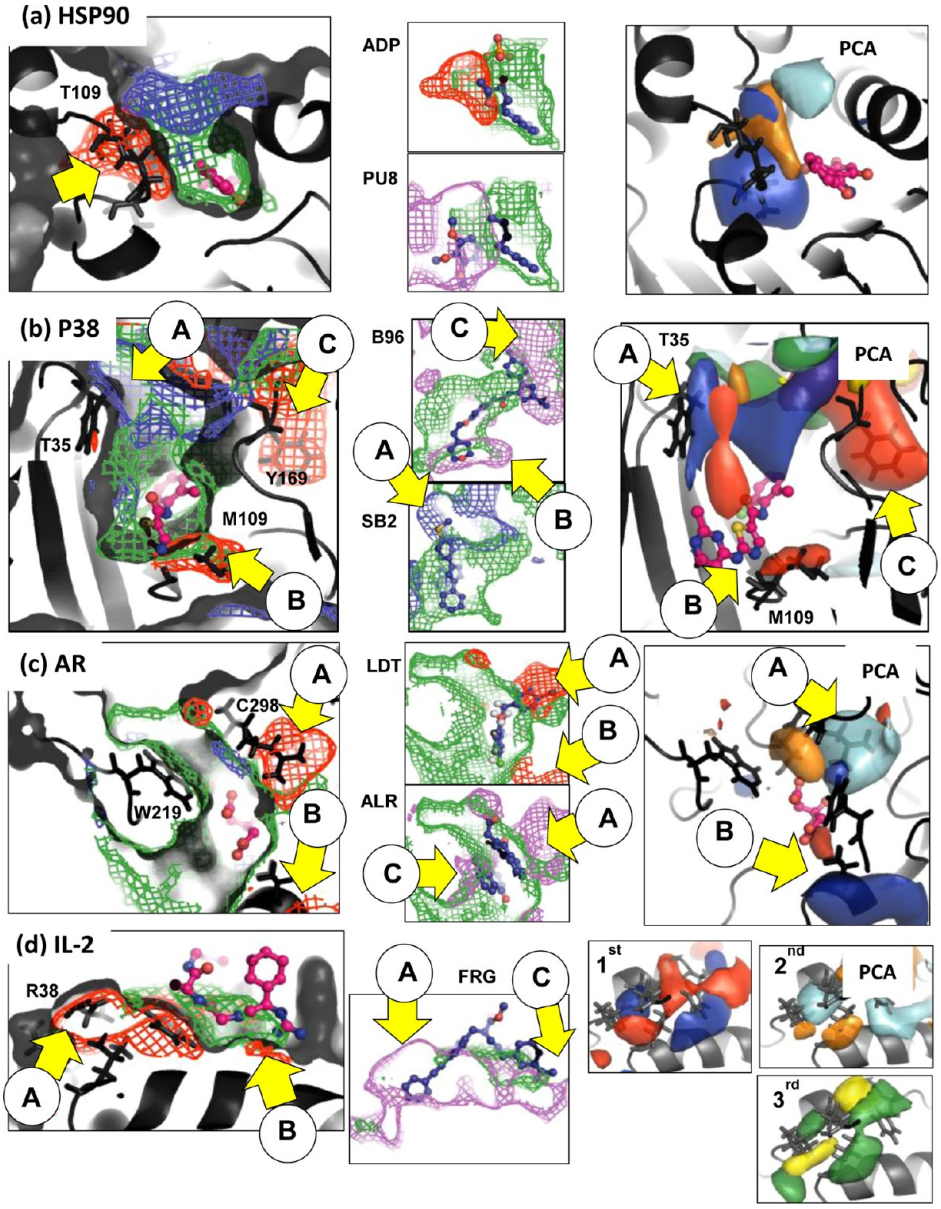


Figure 4. Analysis of the pocket variations in a set of holo-structures for four test targets. Reference PDB structures 3bm9 (HSP90), 1a9u (P38), 2acs (AR), and 1m48 (IL-2) are shown in black cartoon representation and by surface contours in the left-hand plots; flexible residues are represented by sticks. The green mesh contoured at $G^A = 1$ defines the most conserved pocket regions that are observed in all structures. Ligands that fit inside these regions are shown with carbon atoms in red: BXZ (3bm9 of HSP90), 1N1 (3lfa of P38), CIT (2acu of AR), NMP (1m4b of IL-2). Transient regions simulated by the ARDR approach contoured at $G^t = 0.5$, 0.25, and −0.5 are shown in red, magenta, and blue meshes, respectively. The positions of cocrystallized ligands that occupy transient pocket regions are illustrated in insets with carbon atoms in blue. The right-hand plots represent results of the PCA. Positive/negative parts of the 1st, 2nd, and 3rd PCA vectors are shown in red/blue, orange/cyan, and yellow/green at the vector value equal to 1/2 of its maximum vector magnitude.

pdb2gmx.[40] Unlike AR, in HSP90 and P38, conformational variations are caused mainly by long time-scale backbone motion and, to a much lesser extent, by short time-scale side chain fluctuations. Therefore, for the TRAPP analysis, we used only 1000 snapshots extracted from the 10 ns MD trajectories with an interval of 10 ps between snapshots.

Simulations of Backbone Motion in HSP90 and P38 with tCONCOORD. tCONCOORD[19] is a method that generates an ensemble of protein conformations based on a set of geometric constraints applied to the protein degrees of freedom that represent covalent bonds, hydrogen-bonds, and Coulombic and hydrophobic interactions in a protein. Since tCONCOORD is able to simulate protein flexibility from side chain rotation to global rearrangement in secondary and tertiary structures (see refs 20 and 21), we used it for generating conformations that were not observed in the MD trajectories of the P38 and HSP90 proteins but that appear in crystal structures. In particular, conformational variations of the loose part of the distorted α-helix in the transient region of HSP90 and a loop affecting transient subpocket C in P38 (see Figure 2a and b). For HSP90, the conformation with the most closed pocket (2vci) was used as the starting structure for tCONCOORD calculations, and for P38, the same reference structure 1a9u was used as in the MD simulations.

In both proteins, the flexibility of the loop to be generated was restrained by several polar contacts with nearby side chains of the

dx.doi.org/10.1021/ci4000294 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

protein body. These nonbonded interactions were preserved in the tCONCOORD procedure, which does not allow large changes in the loop conformations to be induced. To initiate the rearrangement of the backbone observed in experimental structures, we employed tCONCOORD keeping only the primary constraints and removing those of the secondary structure on the loops whose motion was generated. This procedure allowed breaking of hydrogen bonds of a particular protein fragment while preserving the structure of the rest of the protein. Specifically, residues 165−176 in P38 and 105−119 in HSP90 were described as unconstrained, and residues 96−104 in HSP90 were described as rigid in the tCONCOORD input. Then, 500 structures generated by tCONCOORD were used for the TRAPP analysis.

RESULTS

1. Validation of the TRAPP Approach Using Crystal Structures of Ligand−Receptor Complexes. *Ligand−Pocket Steric Compatibility Quantification by the Ligand−Pocket Overlap, LPO, Index.* To evaluate the accuracy of the pocket computation procedure, we first computed the LPO indexes between the binding pockets of the test protein structures and the ligands extracted from them, which provides a measure of the pocket−ligand steric complementarity. The overlap matrices for the four test targets are shown in Figure 3.

As expected, the diagonal elements of the LPO matrices corresponding to the overlap between the pocket of a holo-structure and its cocrystallized ligand are close to 1. A small deviation from 1.0 (less than 10%) may be caused by uncertainty in the position of hydrogen atoms and inaccuracy in the estimation of the van der Waals surface of either the protein or the ligand.

Cross-terms of the matrix represent the overlap between the binding pockets and non-native ligands. They approach the unit value only if either the conformations of the binding sites are similar or the ligand is positioned inside a pocket region that is conserved in the corresponding structures. Indeed, there are three groups of similar structures (RMSD is 1.5−2.5 Å within each group) for HSP90: (1uyd/1uyf), (1byq/1ior), and (2vci/2vcj/3bm9). These groups can also be clearly identified in the LPO matrix as those having similar overlap indices (see Figure 3a). Similarly, for P38 ligands BMU (1kv1) and B96 (1kv2) demonstrate similar LPO indices since both occupy the conserved pocket part and the transient subpocket C that is closed in all structures, except 1kv1 and 1kv2. Figure 3a−d also shows that the small ligands (such as 2GJ, 2EQ, and BXZ for HSP90; 1N1 for P38; NMP for IL-2; or CIT, I84, and SBI for AR) that fit well inside a conserved part of the pocket (see for example, the benzisoxazole inhibitor BXZ or citric acid CIT in Figure 2a and c, respectively) have the largest overlap indices for all structures.

The LPO values that are notably lower than 1 indicate clashes of the corresponding protein structure and ligands, which are illustrated in insets in Figure 3 (protein−ligand overlap regions are shown in red). For example, in HSP90, the PU1 and PU8 inhibitors occupy a deeply buried part of the binding pocket that appears only in the structures 1uyd/1uyf where the α-helix is preserved, whereas in all non-native structures, these ligands clash with the side chain of L107 or/and T109 depicted in Figure 2a. The magnitude of the LPO index for these ligands decreases from the partially closed (1yet, 1yes, 1byq, 1ior) to the fully closed pocket structures (2vcj, 2vci, and 3bm9). In P38, the ligands BMU (1kv1) and B96 (1kv2) occupy the rather large transient subpocket C (see Figure 3b) and have a small or even negative LPO index with the pockets of all non-native structures in which the C subpocket is closed (as illustrated in insets in Figure 3a for BMU). On the other hand, transient subpocket A, occupied by SB2 (1a9u) is closed by a β-sheet in 1kv1 and 1kv2, and the LPO index of the pocket in 1kv2 and the ligand SB2 (1a9u) is negative (see Figure 3b). A transient region labeled by letter B in P38 arises from the motion of only one residue, M106 (see Figure 2b), and its closure in the 1wfc, 1a9u, and 3lfa structures leads to only a small decrease of the LPO index (down to ∼0.6) for the dihydropyrido-pyrimidine inhibitor 094 (1ouy) because the rest of the ligand is positioned in the conserved pocket region. Similarly, the motion of several residues within the binding site of the AR protein caused only small changes in the pocket shape leading to mostly moderate reductions of the LPO for the pockets of non-native structures (see Figure 3c). In IL-2, the MPE (1m4a) compound occupies a deeply buried part of subpocket A that is completely open only in the native structure, which leads to a corresponding LPO index below 0.5 for all but the cocrystallized structure.

Identification of Conserved and Transient Regions in the Test Holo-Structures by ARDR and PCA Methods. Next, we examined whether the conserved and transient pocket regions identified by the ARDR procedure and PCA analysis agree with the observed variations in the pocket shape and the binding positions of ligands extracted from cocrystallized structures. The results are illustrated in Figure 4.

The aforementioned compounds that have large values of the LPO index for all holo-structures (BXZ, 2GJ, 2GQ in HSP90; 1N1 in P38; CIT, I87, and SBI in AR; NMP in IL-2) are found to be positioned within the pocket regions identified as most conserved in all the holo-structures. This is illustrated in the left-hand plots in Figure 4 where conserved pocket regions are represented by a green mesh and one representative ligand is shown for each target.

Transient regions identified by the ARDR procedure as appearing/disappearing (shown in Figure 4 at iso-values of $G^t(\boldsymbol{r}) = \pm 0.5$ and $G^t(\boldsymbol{r}) = 0.25$ by red/blue and magenta meshes, respectively) are adjacent to the most flexible elements of the binding sites. For example, distortion and conformational variations in the middle part of the α-helix observed in HSP90 holo-structures (see Figure 2a) result in opening of the transient region that appears in 50% of structures, shown in Figure 4a by the red mesh, that defines a partially open pocket occupied by ADP (1byq and 1ior) as well as GDM (1yet) and PU1/PU8 (1uyd/1uyf). A completely open pocket, occupied by only the PU1 and PU8 inhibitors, appears in two out of the eight structures, i.e. 25% (shown in magenta in Figure 4a, right-hand panel). In P38, all three flexible regions (A, B, and C) are identified as transient. Regions B and C are detected as positive transient subpockets because they are closed in the structure used as a reference (1a9u), whereas subpocket A, observed only in 1a9u, was detected as disappearing (Figure 4b). Similarly, in AR, the motion of residue L300 gives rise to transient subpocket A which consists of appearing and disappearing regions relative to the reference structure, 2acu (Figure 4c). In the case of IL-2, FRG (1m48) and MPE (1m4a) have LPO indices below 0.6 and 0.4, respectively, for all holo-structures, except the native one. Indeed, the inset in Figure 4d shows that ligand FRG fits inside the detected transient region only in less than 25% of the structures (transient region shown in magenta in the right-hand panel). Similar, the subpocket C, needed for binding MPE, opens

dx.doi.org/10.1021/ci4000294 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

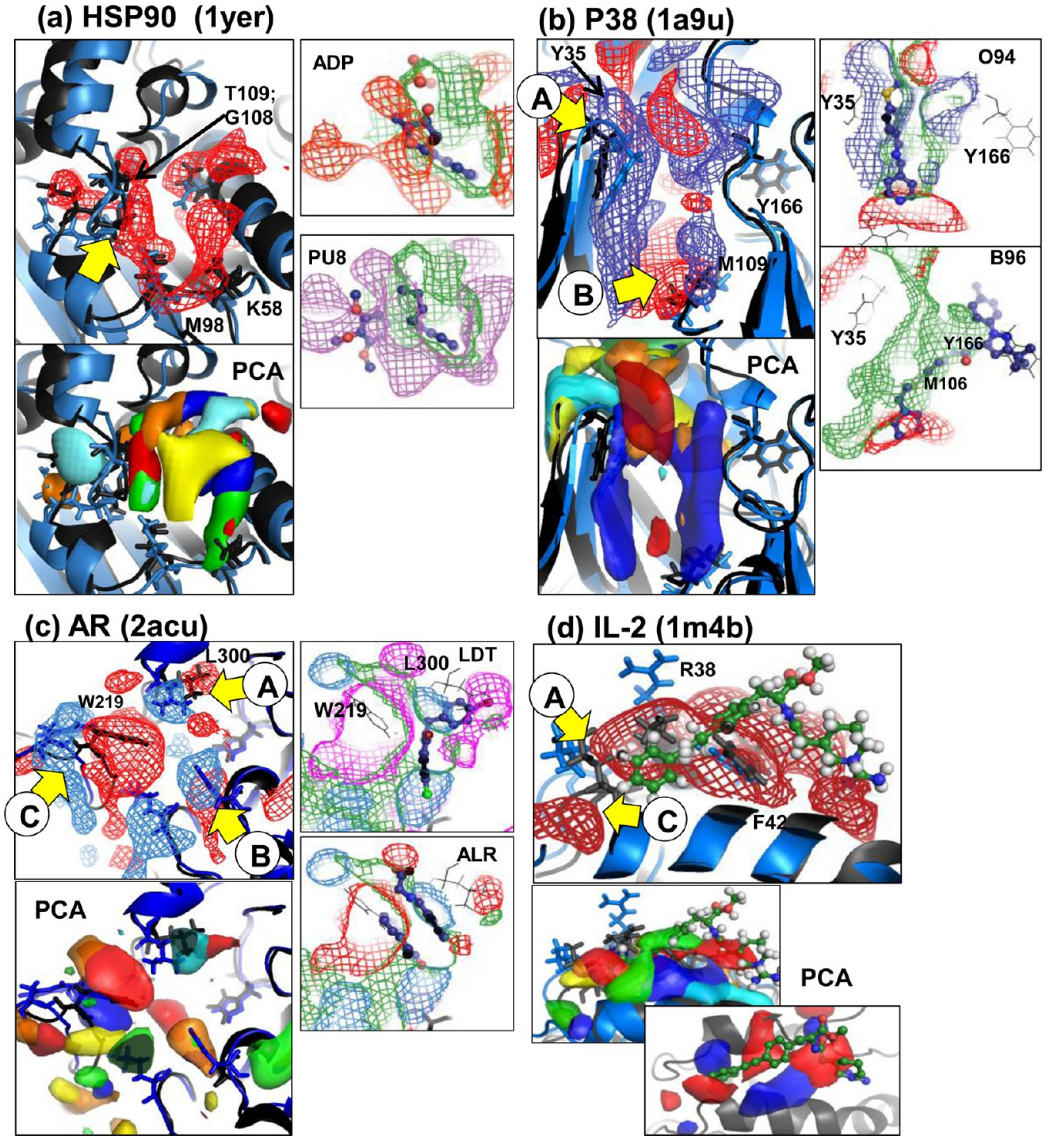


Figure 5. Analysis of the pocket dynamics in MD trajectories. The reference structure (PDB code is given in parentheses along with a protein name) and a representative MD snapshot are shown in black and blue, respectively, and key flexible residues are represented by sticks. Transient/conserved regions simulated by the ARDR method as well as PCA vectors are shown using the same color scheme and iso-values as in Figure 4. The positions of ligands extracted from holo-structures are illustrated in the insets: HSP90–PU8/ADP (1byq), P38–094/B96 (1ouy/1kv2), AR-LDT/ALR (1us0/2az1).

completely only in less than 20% of structures (iso-surface not shown in the figure).

Transient regions identified by the first two PC vectors are represented in Figure 4 by surfaces contoured at half of the maximum vector magnitude. The agreement between the transient regions found by ARDR and PCA approaches is quite good. One should note that the first three PCA vectors cover about 60–80% of pocket variations (eigenvalues are 0.32/0.25/0.13, 0.34/0.21/0.17, 0.34/0.20/0.14, and 0.38/0.26/0.19 in HSP90, P38, AR, and IL-2, respectively). A relatively large first PCA eigenvalue may result from a small number of structures being included in the analysis. The value of the PCA procedure for the detection of correlations in the dynamics of the transient pocket regions is, however, apparently case-dependent. For example, in P38, the first PC vector alone defines all three transient pockets observed in the holo-structures and detects their correlations. Specifically, pockets B and C are identified as opening in-phase, and A and B/C in opposite phase. Indeed, subpockets B and C are occupied simultaneously by ligand B96 in the structure 1kv2, while subpocket A is partially closed when subpocket C opens (in particular, the overlap between structure 1kv2 and ligand SB2, which occupies subpocket A, is negative; see Figure 3). In IL-2, however, correlations between the transient regions are not so obvious because, in particular, all three PC vectors contribute to the subpocket A and B, as demonstrated in Figure 4d.

2. Identification of Transient Pockets in MD Trajectories. For assessment of the transient pocket detection procedure, we have employed MD trajectories that start from conformations with closed or partially closed transient pockets. Specifically, in the HSP90 structure 1yer (which is almost identical to 3bm9 shown in blue in Figure 2a), the transient subpocket is completely closed; in P38 (1a9u), transient subpockets B and C (occupied by ligands BMU/B96 and 094,) are closed, whereas subpocket A is open; in AR (2acu), the subpockets A and C are closed (occupied by ligands IDD594 and tolrestat in 1us0, 1ah3, 1iei, and two drug molecules of alrestatin, ALR, in 1az1, respectively); in the 1m4b structure of IL-2, the helix of the

dx.doi.org/10.1021/ci4000294 | J. Chem. Inf. Model. XXXX, XXX, XXX–XXX

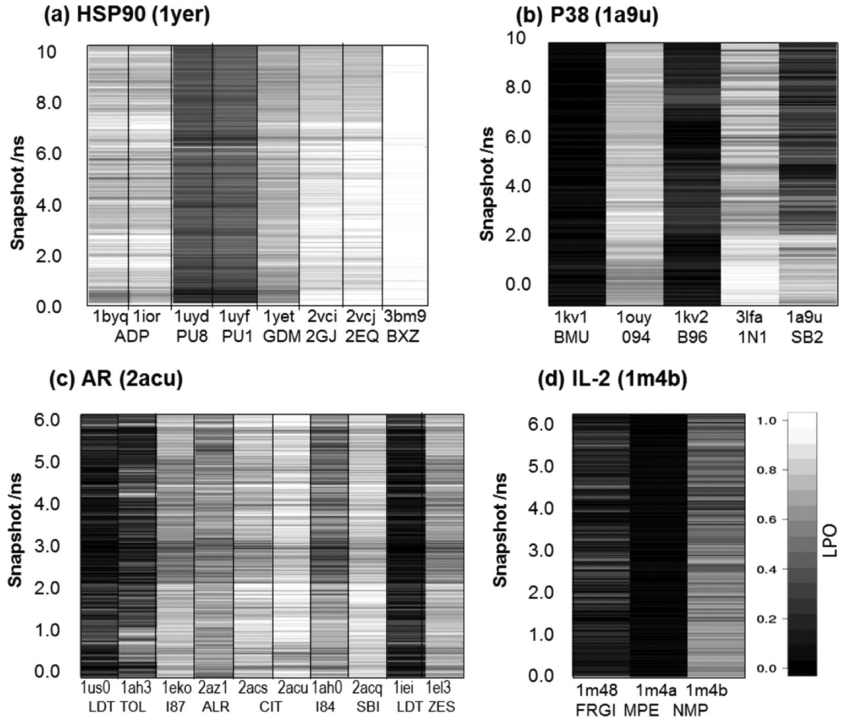


Figure 6. Variations of the LPO index along MD trajectories of the four test proteins (PDB files used for MD simulations are indicated in parentheses); the ligand structures are extracted from holo-structures (horizontal axis). The negative LPO index values are set to 0.

binding site is partially unwound and the transient binding regions A and C are occupied by side chains R38/L72 and K35, respectively.

A period of 6−10 ns of MD simulation time has been shown to be adequate for simulation of pocket opening caused by side chain and loop motion.[15] Indeed, as will be seen from the following pocket dynamics analysis, such relatively short simulations readily generate fluctuations of side chains and mobile backbone regions of the binding sites in the AR protein and a protein binding site in IL-2 (subpocket A), as well as the β-sheet motion in P38 which closes subpocket A (typical structural variations are illustrated in Figures 5a−d for each target). This simulation time was, however, found to be too short to reveal the full spectrum of binding site mobility observed in the P38, HSP90, and IL-2 holo-structures. In particular, conformational changes of the loop in P38 that regulates opening of the binding site for the BMU/B96 (1kv1/1kv2) inhibitors (transient subpocket C) have been observed only in 390 ns explicit solvent MD simulations.[44] Moreover, the binding kinetics of these inhibitors was found to be quite slow,[45] suggesting a high energy barrier between the open- and closed-pocket forms, which may be caused by noncovalent bonds that restrain the binding loop mobility. Similarly, the distorted part of the α-helix in HSP90 is stabilized by multiple contacts with the rest of the α-helix and reveals rather limited backbone motion in MD simulations, which leads to only partial opening of the binding pocket without global change in the loop shape or α-helix rebuilding. As in HSP90, the helix in the IL-2 binding site remains disturbed in the MD simulations (starting from the 1mb4 structure with a partially unwound helix) and its loop undergoes only subtle conformational changes, keeping the subpocket C closed.

All the tendencies mentioned above in the dynamics of the binding pocket shape are clearly indicated by the LPO index computed along the MD trajectories (see Figure 6a−d, the same ligands extracted from holo-structures as in the previous analysis were used). It can be seen from these plots that ligands that fit well in the diverse holo-structures (such as BXZ, 2GJ, and 2EQ in HSP90; CIT and SBI in AR, and 1N1 in P38; see Figure 3) also show large overlap with the binding pocket along MD trajectories. This indicates that the most conserved pocket regions observed in the experimental structures are in general preserved in the MD simulations as well.

As for the analysis of holo-structures, transient regions are detected around the most flexible residues (shown in stick representation in Figure 5) for all the test targets. Specifically, in HSP90, the transient pocket region is occupied by residues L107−T109 in the starting structure (shown in black), and opens periodically along the MD trajectory due to side chain reorientation. In at least 50% of the snapshots, this transient region can accommodate ADP (see inset in Figure 5a), but not the PU8 and PU1 inhibitors (1uyd, 1uyf) which are positioned in a more buried part of the transient subpocket. Accordingly, their LPO index remains below 0.5 along the MD trajectory (Figure 6a). In P38, the global motion of the β-sheet (transient region A) plays a major role in the pocket shape variations. It tends to move toward the pocket center, giving rise to a large negative (disappearing) transient region shown by the blue mesh in Figure 5b. As a result, the subpocket A occupied by the ligand SB2 in the reference structure (1a9u) closes within the first 2 ns causing a general decrease of the LPO index of the corresponding ligand (see Figure 6b). As noted above, the subpocket C in P38 occupied by ligands BMU/B96 (1kv1/1kv2), remains closed (see inset in Figure 5b) in the MD simulations: this is indicated by a small overlap index value for the corresponding ligands over the whole MD trajectory. Finally, the relatively small transient region B appears due to the fluctuations of residue M109, which is also in accordance with some increase of the LPO value of the dihydropyrido-pyrimidine inhibitor (094) in the first 2 ns.

The flexibility of the three main elements of the binding site in AR and of the R38/L72 residues in IL-2 (transient subpocket A)

dx.doi.org/10.1021/ci4000294 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

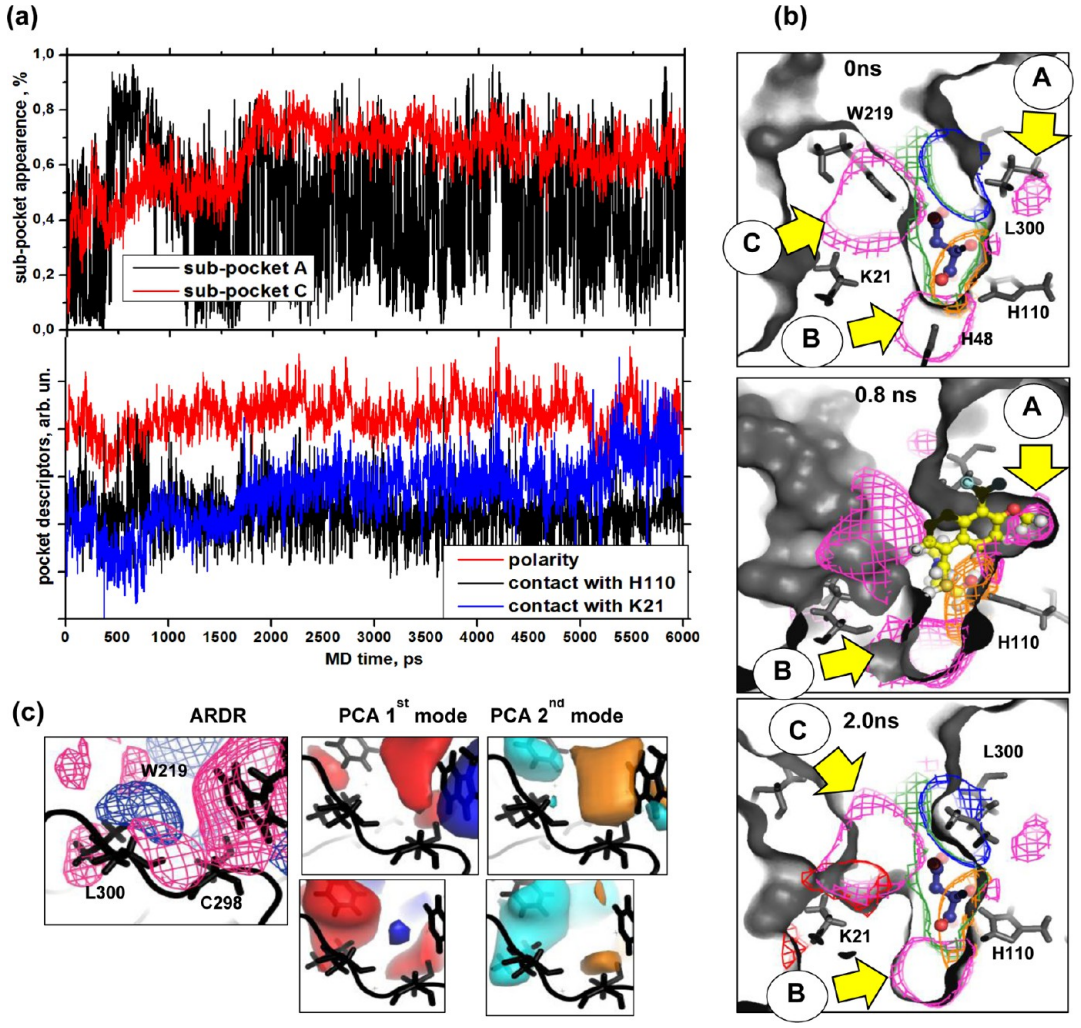


Figure 7. (a) Periodic opening of subpockets A and C in AR illustrated by variation of the overlap of a corresponding subpocket with a binding pocket along the MD trajectory (upper plot); variations of two pocket descriptors (lower plot) showing pocket contact with a catalytic site (H110) and charged residue (K21), as well as with charged residues (computed using $G^{at}(r, p)$ at R_{at} = 2 Å and $G^{ch}(r, p)$ at R_{ch} = 2.5 Å, respectively). (b) Protein conformation observed in the starting MD snapshot and after 0.8 and 2.0 ns of MD simulations are shown by black surfaces contours. The conserved pocket is shown in green and transient pockets computed by ARDR method in magenta and blue meshes (contoured at G^A = 0.5, G^t = 0.5, and G^t = −0.5, respectively); the orange and red meshes indicate the pocket regions that contacts the catalytic site of AR (H110) and K21, respectively. The ligands CIT (2acu) and TOL (1ah3) are represented with carbons in blue and yellow, respectively. TOL occupies transient subpockets A and B as well as the conserved pocket region. (c) Illustration of how the 1st and 2nd PCA vectors for the AR trajectory depend on the pocket region included in the PCA procedure. In the upper plot, a region of subpocket C is included; whereas in the lower plot, it is excluded from the PCA.

is readily observed in 10 ns MD simulations. The corresponding LPO index fluctuates from negative values up to 1 within a picosecond interval (see for example, LPO values of LDT and TOL ligands in last snapshots MD trajectory of AR in Figure 6c) showing that the shapes of the pockets are strongly influenced by side-chain rotation. Noteworthy is that subpocket A in AR, occupied by the LDT and TOL ligands (1us0, 1ah3, and 1iei) is only rarely open completely, which is indicated by the neighboring negative and positive transient regions corresponding to different positions of the L300 side chain (see inset in Figure 5c). Furthermore, unlike TOL, LDT additionally uses subpocket B (which depends on the orientation of Y48) for binding. As a result, the number of snapshots where the LPO index is close to 1 is much smaller for LDT than for the TOL inhibitor as can be seen from Figure 6c.

Interestingly, the subpocket around residues P218−Y219 in the C298A/W219Y mutant AR, which is occupied by ALR (2az1) and denoted as subpocket C, also appears in the present MD simulations. This is because the P218−W219 fragment of the AR loop tends to move, opening subpocket C within the first 2 ns of simulation, which is illustrated in Figure 7a and b. According to the plot in Figure 7a, the loop motion (subpocket C) is characterized by nanosecond time-scale fluctuations whereas the side chain rotation (subpocket A) occurs within picoseconds and is responsible, in particular, for frequent opening of the transient subpocket A. The magnitude of the overlap between a binding pocket and particular subpocket, as shown in Figure 7a, can also be used to identify snapshots where a particular subpocket is open. The changes in the pocket with time are illustrated in Figure 7b where the protein structure and the corresponding conserved and transient pocket regions are shown at the beginning of the MD simulation, after 0.8 ns and after 2.0 ns. It can be seen in this figure that the new pocket shape provides additional space for binding a compound. Furthermore, due to flipping of K21 side chain, a transient subpocket C is characterized by a positively charged region (shown by red mesh

dx.doi.org/10.1021/ci4000294 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

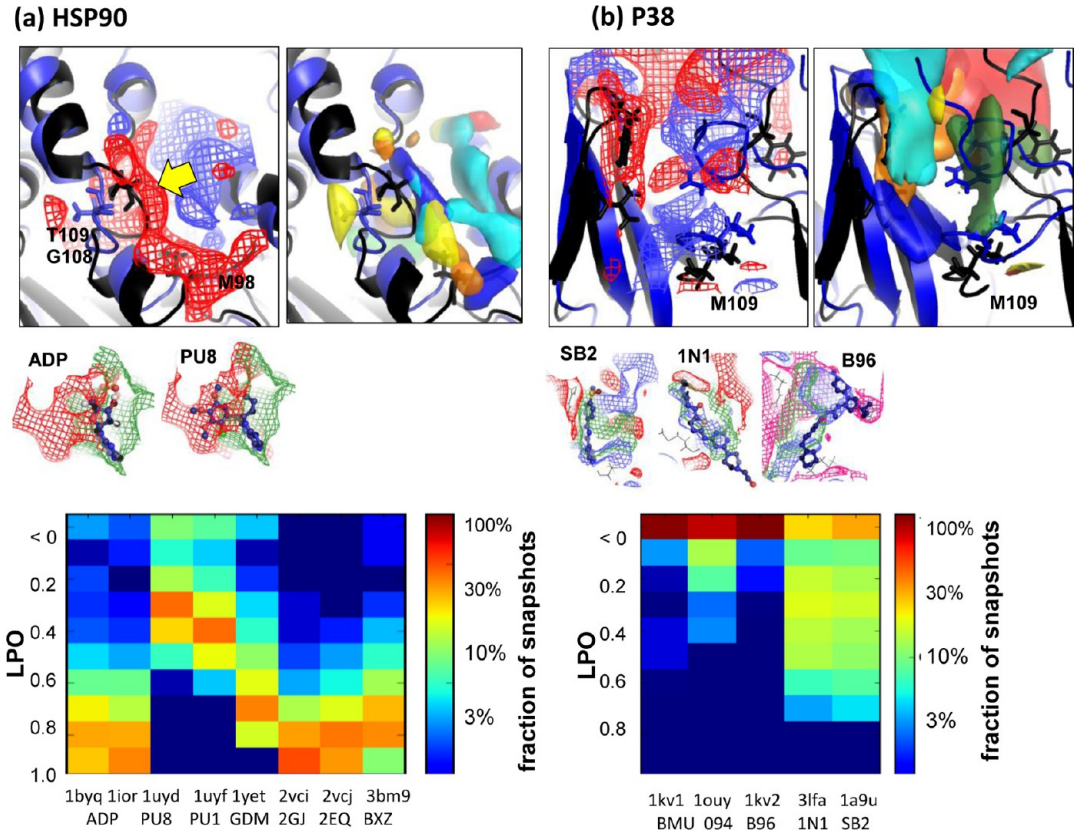


Figure 8. Illustration of the pocket dynamics characteristics identified from structures generated by tCONCOORD for HSP90 and P38. The reference (HSP90 (2vci), P38 (1a9u)) and a representative tCONCOORD structure are shown in black and in blue, respectively. Transient/conserved regions and PCA vectors are shown using the same color scheme and iso-values as in Figure 4. The positions of cocrystallized ligands that occupy transient pocket regions are illustrated in the insets. The lower plots show the statistics of the LPO index value over a tCONCOORD ensemble for each ligand.

in insert to Figure7b). To illustrate TRAPP capability to track change of physicochemical properties of the binding site, we have plotted in Figure 7a (lower panel) variation of descriptors that show contact of K21 and catalytic residue H110 with the binding pocket along MD trajectory. This figure clearly indicates that contribution of K21 to the binding site tends to increase, while contact with the catalytic residue H110 (shown by the orange mesh in Figure 7b) is preserved during protein motion. Although a single short MD trajectory cannot be considered as a reliable basis for making conclusions about the dynamic behavior of the AR binding pocket, it illustrates TRAPP application for identification of new transient subpockets and their physicochemical characteristics.

Noteworthy is that the PCA eigenvalues reveal modest correlations in the pocket motion during the MD trajectories. For all four targets, the cumulative eigenvalues of 100−200 PCA eigenvectors are required to describe 80% of the pocket motion. The eigenvalues of the first two PCA eigenvectors are 0.16/0.12, 0.25/0.08, 0.13/0.05, and 0.18/0.1 for HSP90, P38, AR, and IL-2, respectively, suggesting that the PCA is most informative for P38. The first three eigenvectors, visualized in Figure 5 for each target identify most but not all regions of conformational changes. As expected, they mostly show the global correlated pocket changes (such as the closing of subpocket A in P38 caused by the β-sheet motion or the appearance of subpocket C in AR due to a loop shift), while other smaller transient regions are not revealed by the first few vectors. For example, motion of M109 and opening of the related subpocket B in P38, fluctuation of L300/C298 in AR (transient subpocket A) and M98 in HSP90, do not appear in the first PCA vector. This shows that to analyze small subpockets, one should carry out the PCA procedure, limiting the pocket space to the region of interest. For example, limiting the simulation box so that residue W219 in AR is excluded from the PCA analysis readily shifts the position of the first PCA vector from the region of the global motion of W219 backbone to local fluctuations of the L300/C298 residues as illustrated in Figure 7c.

For IL-2, however, the PCA assists understanding pocket dynamics within a local subpocket region. As one can see from the inset in Figure 5d for IL-2, the subpocket regions around K35-R38 and L72 residues are identified by different lobes (positive/negative) of vectors, which indicate predominantly off-phase motions of these residues determining opening and closing of subpocket A. Thus, PCA provides a possible explanation for the rare and mostly incomplete opening of subpocket A and the relatively low overlap index for the inhibitor, FRG. It should be noted that, in this case, the ARDR method does not distinguish these transient regions and identifies only an averaged opening pocket.

3. Identification of Transient Pockets in tCONCOORD Ensembles. The ensemble of HSP90 conformations generated by tCONCOORD from the 2vci structure reveals substantial mobility of the loop part of the distorted α-helix. The two representative tCONCOORD structures shown in Figure 8a demonstrate the range of conformational changes which are much larger than those observed in the MD simulation (compare Figure 5a). Accordingly, the transient region arising from the loop motion is larger and deeper, providing, in particular, the

dx.doi.org/10.1021/ci4000294 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

space necessary for accommodating ADP (1byq) and GDM (1yet) and the major part of the purine-based inhibitor, PU8, as shown in the insets in Figure 8a. tCONCOORD structures with high LPO values (above 0.8) were found for all ligands except for PU8 (1uyd) and PU1 (1uyf) for which the maximum index value reaches 0.6 (see the corresponding statistics of the LPO index in Figure 8a). This is not surprising since the α-helix is undisturbed in the 1uyd/1uyf structures, a conformation that cannot be generated by tCONCOORD starting from a disturbed structure.

As for HSP90, the changes in protein conformation generated by the tCONCOORD approach in P38 are much larger than those obtained in MD simulations. Particularly, the β-sheet movement, observed in MD but with a notably smaller amplitude, tends to further increase the size of transient subpocket A already opened in the starting structure 1a9u, which gives rise to a large transient region around Y35 (see Figure 8b). In addition, the buried loop of the catalytic site becomes flexible and opens a transient subpocket C in 25% of the structures (iso-mesh shown in magenta in the insert in Figure 8b). However, in many tCONCOORD structures, this loop moves from subpocket C into the central part of the pocket, which appears as a negative transient region (shown in blue in Figure 8b). Moreover, subpocket B remains closed in all structures since the M109 side chain has a close-contact of ~2 Å with K165 (not shown in the figure), which is defined as a constraint and preserved in the tCONCOORD procedure. Instead, the backbone around M109 moves, partially closing the binding pocket, which give rise to the disappearing transient region around M109. This observation indicates an obvious shortcoming of the tCONCOORD method resulting from constraint definitions that might in some cases be remedied by using several tCONCOORD ensembles based on different reference structures. Thus, in spite of the notable structural changes generated, only a few structures have binding pockets similar to those observed in the holo-structures of P38. Accordingly, the magnitude of the LPO (Figure 8b) is mostly lower than in MD simulations for all ligands, except for BMU and B96. These ligands occupy subpocket C which remains closed in MD simulations.

Finally, it should be noted that the motions of the different elements of the binding pocket generated by the tCONCOORD method are highly uncorrelated, leading to very low PCA eigenvalues. The eigenvalues of the first and second eigenvectors are only 0.16/0.14 and 0.08/0.07 for HSP90 and P38, respectively, and thus lower than the values for the MD simulations. Accordingly, variations in the pocket distribution within most of the transient regions are described by several eigenvectors.

DISCUSSION

Choice of the Binding Site Residues. In TRAPP, a binding site is defined as a set of residues surrounding (within a user-defined interaction distance) a reference ligand and is used in two ways: for structure alignment and superposition and for definition of the pocket grid size. Thus, the shape and the position of the user-defined reference ligand affect the protein region to be analyzed. To ensure that no transient subpockets are missed in the analysis, it is reasonable to construct a reference ligand as a combination of ligands extracted from different superimposed holo-structures and/or increase the distance around the ligand to be used for binding site identification. Alternatively, a set of binding site residues may be defined by the user, making it possible to capture pocket parts that are not occupied or closed in available holo-structures. On the other hand, the size of the grid analyzed can be minimized by omitting some pocket parts, thus significantly reducing the computational time. It should also be noted that the set of binding site residues may have considerable influence on the pocket analysis when conformational differences are large. Checking the definition of the binding site residues by varying interaction distance, ligand shape, or directly the residues of the binding site is advisable in such cases.

Assignment of TRAPP Parameters. The pocket detection algorithm has been designed to identify different types of cavities without changing or fitting parameters used for the analysis. Consequently, the values of the parameters defined in the Materials and Methods section and validated for the test targets are not intended to be varied for other systems. This is an obvious advantage of the TRAPP approach with respect to other pocket detection methods; in particular, methods that have been applied recently for the analysis of transient pockets in conformational ensembles of proteins (MDpocket[17] and PocketAnalyzer[PCA 16]). In particular, the pocket computation procedure used in TRAPP enables shallow solvent-exposed pockets on a protein surface as well as closed buried pockets to be found simultaneously (comparison of existing methods for tracing and analysis of pocket dynamics with a TRAPP approach is given in section 5 of the Supporting Information).

Nevertheless, most of the TRAPP parameters, as well as the usage protocols, may be altered by the user for adjustment to particular needs (but not for a particular target). For example, the computational time and memory needed for the computation of the pocket shape can be notably reduced by changing several parameters, albeit at the expense of accuracy. Specifically, the number of vectors used for the buriedness estimation may be reduced to 64 and, instead of the two-step procedure, one may choose to use only one step. Furthermore, the solvent-exposed region of the cavity can be decreased by increasing the buriedness index threshold, α_0. Practically, the value of $\alpha_0 = 0.4-0.5$ enables detection of the both closed and very shallow pockets on the protein surface, while $\alpha_0 \sim 0.7$ helps to identify only completely buried parts of a cavity ($\alpha_0 = 0.6$ is the default value, used for this work).

Applicability of the Transient Pocket Identification Tools. We have implemented three mutually complementary approaches for analysis of pocket dynamics.

(i) The averaged pocket distribution $G^A(r)$ defines the fraction of structures in which a particular pocket region is present, and it can be mainly applied for identifying conserved parts of the binding site. This characteristic is analogous to the pocket frequency map used in MDpocket[17] for visualization of pocket opening frequencies (the results of pocket dynamic analysis using MDpocket and the TRAPP method are compared in the Supporting Information).

(ii) Next, the ARDR distribution, $G^t(r)$, gives the fraction of structures where a particular pocket region appears or disappears relative to a reference experimental structure and, thus, shows transient pocket regions. Averaging of transient regions over all structures enables smoothing of the noise of individual structures, which is important for TRAPP functionality, since the method is intended to be applied not only for analysis of experimental protein structures or MD simulation snapshots, but also to extract information on pocket flexibility from structures obtained

dx.doi.org/10.1021/ci4000294 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

from less accurate approaches, such as NMA, tCONCOORD,[19] or RIP.[46] On the other hand, it leads to loss of information on pocket conformers. This problem is solved by using a clustering of the ARDR distribution implemented in TRAPP, which enables splitting of all transient regions at different iso-values into a set of transient subpockets and identification of particular structures where a transient subpocket or several subpockets of interest are open. Note that this is different from the clustering of protein structures according to their scores along PCA as implemented in the PocketAnalyzer$^{PCA\ 16}$ method, which provides a set of structures in which the largest correlated conformational changes of the pocket observed.

(iii) Finally, the PCA method aims to reduce the dimensionality of a space of the pocket variations to a small number of vectors that provide information on the correlation in opening and closing of different transient pocket regions. Importantly, the value of the PCA method depends on the degree of correlation in motion of the different elements of a binding site, which is highly case-dependent. Indeed, the present analysis shows that PCA analysis is most easily interpretable if the number of structures analyzed is small (in the case of holo-structure analysis), or the binding site flexibility is mainly represented by global motion of some binding site elements. For most of the targets considered in the present analysis, however, the largest PC eigenvalues do not exceed 0.2 and many of the lower modes have comparable eigenvalues, suggesting that the first few PCA vectors cannot reduce the pocket motion space effectively. To represent the most tendencies of pocket motion space, the number of vectors should be much larger (usually more than 100), which makes the visual analysis of transient regions infeasible and interpretation of simulation results ambiguous. Another limitation comes from the fact that the distribution of the PC vectors depends on the pocket region included in the analysis: the first few PCA modes encompass the largest correlated regions, while motion of small subpockets is described by higher components and can easily be overlooked in the visual analysis (as shown for the AR target). Taking also into account the cumbersome and extremely computationally consuming procedure of the correlation matrix diagonalization, the practical application of PCA procedure seems to be limited to only small pockets arising from fluctuation of several side chains.

Recently a PCA-based protocol for selection of diverse pocket shapes from MD trajectories or ensembles of target structures, PocketAnalyzer^{PCA}, has been published by Craig et al.[16] The protocol is based on the LIGSITE[25] pocket detection algorithm and combines PCA and clustering techniques. The method was applied to an MD trajectory of AR, and thus, the results can be qualitatively compared with those described here. In both studies, quite small values of the first two PCA eigenvalues (only 21% of the total variance in ref 16 and about 18% in the present work) were obtained, though the authors of ref 16 used only 40 conformations (from a 20 ns MD trajectory, selected at 500 ps intervals) in contrast to the 3000 snapshots used in the present analysis. The first two PCA vectors in ref 16 were found to describe pocket variations around residues L300−C298. In the present analysis, some fluctuations of the C298 side chain were observed (see Figure 7c), but these give a rather small contribution to the total pocket variation, while the first PC arises predominantly from motion of the P218−W219 loop. Only excluding the region occupied by the P218−W219 residues makes the transient region around L300−C298 clearly pronounced in the first dominant PCA modes (see Figure 7c). Since the analysis in ref 16 is focused on the L300−C298 subpocket, it is unclear whether motion of P218−W219 was observed in this study.

Finally, one should note that the detected pocket dynamics is subject to the limitations of the trajectory or ensemble of protein conformations analyzed. Pocket analysis based on a short MD trajectory or on a limited number of MD snapshots selected at large time intervals can result in important pocket flexibility being overlooked. Using several conformational ensembles obtained from different methods and/or distinct starting conformations for structure generation may enlarge the space of conformational changes. This may also help to smooth inaccuracy that might arise from each individual approach. Currently, we are working on an automated protocol for the preparation of conformational ensembles using different approaches for the generation of protein flexibility, which will be implemented in the TRAPP platform.

CONCLUSIONS

We have presented a novel computational tool, TRAPP, for automated detection of transient regions of binding pockets in ensembles of protein conformations, which may be obtained from experimental structures or from simulations. TRAPP enables tracing of the residues that compose a binding site, the binding pocket shape and the physicochemical properties along a protein motion trajectory/ensemble of structures that may consist of several thousands of structures/snapshots. The ARDR method can be used to compute the fraction of structures/snapshots in which a particular pocket region is open and to detect structures where a transient subpocket of interest appears. The PCA technique implemented allows correlated motion of (sub)pocket regions to be identified. Additionally, TRAPP enables the similarity between pockets from different structures, and the spatial complementarity between a pocket and a particular ligand to be measured.

TRAPP is able to accurately identify the pocket shape for a very large range of protein conformational changes, from side chain rotation to domain motion, without any adaptation or fitting of calculation parameters to the particular system under study. The pocket detection method implemented in TRAPP has been evaluated against the holo-structures of four protein targets: HSP90, P38, AR, and IL-2. For this purpose, we have developed a ligand−pocket overlap (LPO) index that provides a measure of the ligand−pocket compatibility. As expected, the index reaches its maximum value of 1 for the overlap computed between a ligand and the pocket of its own holo-structure. In other cases, its magnitude reveals the extent of ligand−protein clashes if the binding pocket is partially closed. The LPO index serves as a measure of whether a ligand can be accommodated inside a particular binding pocket or not. A procedure of sampling ligand positions in a binding pocket for each protein and of screening ligand conformers to discard compounds that are not suitable, would extend the proposed approach for first-stage screening of molecular libraries taking protein flexibility fully into account.

The ability of TRAPP to detect transient regions has been validated using 6−10 ns MD trajectories for each test target and ensembles of structures generated using tCONCOORD (for HSP90 and P38). For each test case, we have employed a starting

dx.doi.org/10.1021/ci4000294 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

structure with closed or partially closed binding pockets. Transient pocket regions observed in the holo-structures are found in structures generated by MD and tCONCOORD methods. There are, however, transient parts for each target that are detected either in the MD trajectory or in the tCONCOORD ensemble but not in both. Specifically, MD simulations are able to identify pockets formed by breaking the hydrogen-bond network, whereas tCONCOORD explicitly preserves most hydrogen bonds. On the other hand, tCONCOORD can generate substantial structural variations of large portions of the protein structure, which is not easily achieved in 10 ns standard MD simulations. These distinctions in the accessible conformational variations are shown to be readily transformed into differences in the transient regions detected. The present analysis suggests that the combination of conformational ensembles obtained from distinct starting (reference) conformations and the use of different sampling methods should provide the best basis for finding transient pockets in proteins.

The TRAPP procedure for tracing the physicochemical properties of a binding pocket provides a basis for computation of variation of druggability along a protein motion trajectory or in an ensemble of structures. Our approach opens up promising perspectives for the design of novel ligands for known protein structures. Whereas protein crystallography can only provide a limited number of protein conformations, and consequently also a limited variation in the size and shape of potential binding pockets, TRAPP will find hitherto unknown additional pockets or subpockets, if the inherent protein flexibility allows for them. This can be of great value for the design of unique inhibitors that will address a so far unknown transient binding site.

The TRAPP webserver is available at http://www.mcm.h-its.org/trapp/.

■ ASSOCIATED CONTENT

Supporting Information

Illustration of the TRAPP workflow; description of the sequence alignment and superimposition procedures; robustness of the pocket shape identification procedure (effects of the superposition procedure, grid position/orientation, and hydrogen atoms on the pocket shape); details of the clustering algorythm; comparison of TRAPP with other available programs for analysis of transient/conserved binding pockets. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: daria.kokh@h-its.org (D.B.K.) and rebecca.wade@h-its.org (R.C.W.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We gratefully acknowledge the support of the German Federal Ministry for Education and Research (BMBF) Biotech Cluster Rhein-Neckar (BioRN) (Project INE-TP03) and the Klaus Tschira Foundation. We thank Musa Özboyaci for assistance in carrying out MD simulations and testing of the web server.

■ REFERENCES

(1) Henrich, S.; Salo-Ahen, O. M.; Huang, B.; Rippmann, F. 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.

(2) Laurie, T. R. A.; Jackson, R. M. Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual screening. *Curr. Protein Peptide Sci.* **2006**, *7*, 395−406.

(3) Fraser, J. S.; van der Bedem, H.; Samelson, A. J.; Lang, P. T.; Holton, J. M.; Echols, N.; Alber, T. Accessing protein conformational ensemble using room-temperature X-ray crystallography. *Proc. Nat. Acad. Sci. U.S.A.* **2011**, *108*, 16247−16252.

(4) Cozzini, P.; Kellogg, G. E.; Spyrakis, F.; Abraham, D. J.; Constantino, G.; Emerson, A.; Fanelli, F.; Gohlke, H.; Kuhn, L. A.; Morris, G. M.; Orozco, M.; Pertinhez, T. A.; Rizzi, M.; Sotriffer, C. A. Target flexibility: an emerging consideration in drug discovery and design. *J. Med. Chem.* **2008**, *51*, 6237−6255.

(5) Kokh, D. B.; Wade, R. C.; Wenzel, W. Receptor flexibility in small-molecule docking calculations. *WIREs Comput. Mol. Sci.* **2011**, *1*, 298−314.

(6) Waszkowycz, B.; Clark, D. E.; Gancia, E. Outstanding challenges in protein-ligand docking and structure-based virtual screening. *WIREs Comp. Mol. Sci.* **2011**, *1*, 229−259.

(7) Brown, S. P.; Hajduk, P. J. Effects of conformational dynamics on predicted protein druggability. *ChemMedChem* **2006**, *1*, 70−72.

(8) Gunasekaran, K.; Nussinov, R. How different are structurally flexible and rigid regions? Sequence and structural features discriminating proteins that do and do not undergo conformational changes upon ligand binding. *J. Mol. Biol.* **2007**, *365*, 257−273.

(9) Fauman, E. B.; Rai, B. K.; Huang, E. S. Structure-based druggability assessment — identifying suitable targets for small molecule therapeutics. *Curr. Opin. Chem. Biol.* **2011**, *15*, 463−468.

(10) Totrov, M.; Abagyan, R. Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr. Opin. Struct. Biol.* **2008**, *18*, 178−184.

(11) Stein, A.; Rueda, M.; Panjkovich, A.; Orozco, M.; Aloy, P. A systematic study of the energetics involved in structural changes upon association and connectivity in protein interaction networks. *Structure* **2011**, *19*, 881−889.

(12) Mangoni, M.; Roccatano, D.; Di Nola, A. Docking of flexible ligands to flexible receptors in solution by molecular dynamics simulation. *Proteins* **1999**, *35*, 153−162.

(13) Eyrisch, S.; Helms, V. Transient pockets on protein surfaces involved in protein-protein interactions. *J. Med. Chem.* **2007**, *50*, 3457−3464.

(14) Schames, J. R.; Henchman, R. H.; Siegel, J. S.; Sotriffer, C. A.; Ni, H.; McCammon, J. A. Discovery of novel binding trench in HIV Integrase. *J. Med. Chem.* **2004**, *47*, 1879−1881.

(15) Eyrisch, S.; Helms, V. How transient pockets open on the surface of the MDM2 protein. *Chem. Central J.* **2008**, *2* (Suppl I), P34.

(16) Craig, I. R.; Pfleger, C.; Gohlke, H.; Essex, V.; Spiegel, K. Pocket-space map to identify novel binding-site conformations in proteins. *J. Chem. Inf. Model.* **2011**, 2666−2679.

(17) Schmidtke, P.; Bidon-Chanal, A.; Luque, F. J.; Barril, X. MDpocket: open source cavity detection and characterization on molecular dynamics trajectories. *Bioinformatics* **2011**, *27*, 3276−3285.

(18) Nichols, S. E.; Baron, R.; Ivetac, A.; MaCammon, J. A. Predictive power of molecular dynamics receptor structures in virtual screening. *J. Chem. Inf. Model.* **2011**, *51*, 1439−1446.

(19) Seeliger, D.; Haas, J.; de Groot, B. Geometry-based Sampling of conformational transitions in protiens. *Structure* **2007**, *15*, 1482−1492.

(20) Eyrisch, S.; Helms, V. What induces pocket openings on protein surface patches involved in protein-protein interactions? *J Comput.-Aided Mol. Des.* **2009**, *23*, 73−86.

(21) Seeliger, D.; de Groot, B. Conformational transitions upon ligand binding: holo-structure prediction from apo-conformations. *PLoS Comput. Biol.* **2010**, *6*, e1000634−e1000641.

(22) Cavasotto, C. N.; Kovacs, J. A.; Abagyan, R. A. Representing receptor flexibility in ligand docking through relevant normal modes. *J. Am. Chem. Soc.* **2005**, *127*, 9632−9640.

(23) Dietzen, M.; Zotenko, E.; Hildebrandt, A.; Lengauer, T. On the Applicability of Elastic Network Normal Modes in Small-Molecule Docking. *J. Chem. Inf. Mod.* **2012**, *52*, 844−856.

(24) Ashford, P.; Moss, D. S.; Alex, A.; Yeap, S. K.; Povia, A.; Nobeli, I.; Williams, M. A. Visualisation of variable binding pockets on protein surfaces by probabilistic analysis of related structure sets. *BMC Bioinf.* **2012**, *13*, 1471.
(25) Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: Automatic and efficient method detection of potential small molecule-binding sites in proteins. *J. Mol. Graphics. Modell.* **1997**, *15*, 359−363.
(26) Brady, G. P.; Stouten, P. Fast prediction and visualization of protein binding pockets with PASS. *J Comput.-Aided Drug. Des.* **2000**, *14*, 383−401.
(27) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinf.* **2009**, *10*, 168−179.
(28) Hoffmann, B.; Zaslavskiy, M.; Vert, J.-P.; Stoven, V. A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction. *BMC Bioinf.* **2010**, *11*, 99−115.
(29) *The PyMOL Molecular Graphics System*, version 1.4.1; Schrödinger, LLC, New York; http://www.pymol.org (accessed January 2012).
(30) Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/ (accessed Jan 2012).
(31) Meyer, T.; D'Abramo, M.; Hospita, l. A.; Rueda, M.l.; Ferrer-Costa, C.s; Pérez, A.; Carrillo, O.; Camps, J.; Fenollosa, C.; Repchevsky, D.; Gelpí, V.; Orozco, M. MoDEL (Molecular Dynamics Extended Library): A Database of Atomistic Molecular Dynamics Trajectories. *Structure* **2010**, *18*, 1399−1409 ; http://mmb.pcb.ub.es/MoDEL/ (accessed Oct 1, 2012)..
(32) Jones, E.; Oliphant, T.; Peterson, P. SciPy: Open Source Scientific Tools for Python. http://www.scipy.org (accessed Apr.1, 2012)
(33) 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*, 1422−1423.
(34) Humphrey, W.; Dalke, A.; Schulten, K. VMD - Visual Molecular Dynamics. *J. Molec. Graphics* **1996**, *14*, 33−38.
(35) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera−a visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.
(36) Arkin, M. R.; Randal, M.; DeLano, W. L.; Hyde, J.; Luong, T. N.; Oslob, J. D.; Raphael, D. R.; Taylor, L.; Wang, J.; McDowell, R. S.; Wells, J. A.; Braisted, A. C. Binding of small molecules to an adaptive protein-protein interface. *Proc. Nat. Acad. Sci. U.S.A.* **2003**, *100*, 1603−1608.
(37) Tilley, J. W.; Chen, L.; Fry, D. C.; Emerson, S. D.; Powers, G. D.; Biondi, D.; Varnell, T.; Trilles, R.; Guthrie, R.; Mennona, F.; Kaplan, G.; Lemahieu, R. A.; Carson, M.; Han, R. J.; Liu, C. M.; Palermo, R.; Ju, G. Identification of small molecule inhibitor of the IL-2/IL-2Rα receptor interaction which binds to IL-2. *J. Am. Chem. Soc.* **1997**, *119*, 7589−7590.
(38) Metz, A.; Pfleger, C.; Kopitz, H.; Pfeiffer-Marek, S.; Baringhaus, K. H.; Gohlke, H. Hot spots and transient pockets: Predicting the determinants of small-molecule binding to a protein-protein interface. *J. Chem. Inf. Mod.* **2012**, *52*, 120−133.
(39) *Maestro*, version 9.2; Schrödinger, LLC, New York, 2011; http://www.schrodinger.com/.
(40) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.
(41) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.
(42) Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25*, 1656−1676.
(43) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997**, *18*, 1463−1472.
(44) Frembgen-Kesner, T.; Elcock, A. H. Computational Sampling of a Cryptic Drug Binding Site in a Protein Receptor: Explicit Solvent Molecular Dynamics and Inhibitor Docking to p38 MAP Kinase. *J. Mol. Biol.* **2006**, *359*, 202−214.
(45) Pargellis, C.; Tong, L.; Churchill, L.; Cirillo, P. F.; Gilmore, T.; Graham, A. G.; Grob, P. M.; Hickey, E. R.; Moss, N.; Pav, S.; Regan, J. Inhibition of p38 MAP kinase by utilizing a novel allosteric binding site. *Nat. Struct. Biol.* **2002**, *9*, 268−272.
(46) Ho, B. K.; Agard, D. A. Probing the flexibility of large conformational changes in protein structures through local perturbations. *PLoS Comput. Biol.* **2009**, *5*, e1000343−e1000313.