BIOINFORMATICS ORIGINAL PAPER

Vol. 25 no. 5 2009, pages 621–627
doi:10.1093/bioinformatics/btp036

Structural bioinformatics

Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques

Ryan Brenke[1,†], Dima Kozakov[2,†], Gwo-Yu Chuang[2,†], Dmitri Beglov[2,†], David Hall[2], Melissa R. Landon[1], Carla Mattos[3] and Sandor Vajda[2,*]

[1]Program in Bioinformatics, [2]Structural Bioinformatics Laboratory, Department of Biomedical Engineering, Boston University, Boston, MA, USA and [3]Department of Molecular and Structural Biochemistry, North Carolina State University, Raleigh, NC, USA

Received on October 9, 2008; revised on January 8, 2009; accepted on January 13, 2009

Advance Access publication January 28, 2009

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: The binding sites of proteins generally contain smaller regions that provide major contributions to the binding free energy and hence are the prime targets in drug design. Screening libraries of fragment-sized compounds by NMR or X-ray crystallography demonstrates that such 'hot spot' regions bind a large variety of small organic molecules, and that a relatively high 'hit rate' is predictive of target sites that are likely to bind drug-like ligands with high affinity. Our goal is to determine the 'hot spots' computationally rather than experimentally.

Results: We have developed the FTMAP algorithm that performs global search of the entire protein surface for regions that bind a number of small organic probe molecules. The search is based on the extremely efficient fast Fourier transform (FFT) correlation approach which can sample billions of probe positions on dense translational and rotational grids, but can use only sums of correlation functions for scoring and hence is generally restricted to very simple energy expressions. The novelty of FTMAP is that we were able to incorporate and represent on grids a detailed energy expression, resulting in a very accurate identification of low-energy probe clusters. Overlapping clusters of different probes are defined as consensus sites (CSs). We show that the largest CS is generally located at the most important subsite of the protein binding site, and the nearby smaller CSs identify other important subsites. Mapping results are presented for elastase whose structure has been solved in aqueous solutions of eight organic solvents, and we show that FTMAP provides very similar information. The second application is to renin, a long-standing pharmaceutical target for the treatment of hypertension, and we show that the major CSs trace out the shape of the first approved renin inhibitor, aliskiren.

Availability: FTMAP is available as a server at http://ftmap.bu.edu/.

Contact: vajda@bu.edu

Supplementary information: Supplementary Material is available at *Bioinformatics* online.

1 INTRODUCTION

It has been recognized for many protein–ligand complexes that certain regions of the binding surface, often called 'hot spots', contribute a disproportionate amount to the binding free energy (Hajduk *et al.*, 2005). Such regions are more likely to bind small drug-like compounds with high affinity than the rest of the binding site, and hence their identification is important for drug design (DeLano, 2002; Vajda and Guarnieri, 2006). Both NMR (nuclear magnetic resonance) and X-ray crystallography techniques have been used to find 'hot spots' of proteins. Using NMR the ^{15}N-labeled protein is screened against a library of small probe compounds (Hajduk *et al.*, 2005; Vajda and Guarnieri, 2006). Applications to a variety of proteins demonstrate that the 'hot spots' bind a variety of small molecules, and a high 'hit rate' is a good predictor of druggability. Indeed, a high correlation was observed between the number of different probes binding to a site, and the ability to identify high-affinity druglike ligands that bind there (Hajduk *et al.*, 2005). The X-ray technique, known as Multiple Solvent Crystal Structures (MSCS) method, is based on solving the structure of the protein in aqueous solutions of various probe compounds, primarily organic solvents (Mattos and Ringe, 1996). Each structure shows a few organic molecules associated with the protein surface in the first shell of water molecules. The power of the method arises from superimposing a number of structures solved in different solvents. Most probes generally cluster in the binding site, and the overlapping probe clusters form 'consensus' sites (CSs) that delineate the functionally most important subsites. As demonstrated by applications to porcine elastase (Allen *et al.*, 1996; Mattos and Ringe, 1996; Mattos *et al.*, 2006) and thermolysin (English *et al.*, 1999, 2001), some probes may also bind at crystal contacts or in small buried pockets, but the large CSs occur in the 'hot spots' of the binding site.

Since the identification of 'hot spots' by NMR or X-ray crystallography is very expensive, it is important to explore whether similar information can be obtained by computations. Computational mapping methods place molecular probes on the protein surface in order to explore its binding properties. A number of methods identify potential binding sites (An *et al.*, 2005; Glaser *et al.*, 2006; Laurie and Jackson, 2005). Some early methods such as GRID (Goodford, 1985) and Multiple Copy Simultaneous Search

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first four authors should be regraded as joint First Authors.

(MCSS) (Caflisch *et al.*, 1993; Miranker and Karplus, 1991) have been developed to find favorable binding positions for specific molecules or functional groups rather than to identify 'hot spots'. Both methods result in many energy minima, and it is difficult to determine which of the minima are actually relevant (Mattos and Ringe, 1996).

In this article, we describe the FTMAP algorithm specifically developed to reproduce NMR and X-ray mapping results using a number of organic probe molecules. For each probe the algorithm generates 2000 bound positions using rigid body docking, refines the positions by energy minimization, clusters the resulting conformations and ranks the clusters on the basis of this average free energy. The docking step is based on the fast Fourier transform (FFT) correlation approach which helps to efficiently sample billions of probe positions on dense translational and rotational grids, but can use only sums of correlation functions for scoring and hence, is generally restricted to very simple energy expressions. The novelty of FTMAP is that we were able to incorporate and represent on grids a detailed energy expression which includes attractive and repulsive van der Waals terms, electrostatic interaction energy based on Poisson–Boltzmann calculations, a cavity term to represent the effect of non-polar enclosures and a structure-based pairwise interaction potential. The energy expression is key to the accuracy of FTMAP, and the algorithm matches or exceeds the accuracy of our earlier mapping method (Dennis *et al.*, 2002; Silberstein *et al.*, 2003), although FTMAP requires only about one-sixth of the CPU time. Here, we show that FTMAP reproduces the X-ray mapping results for elastase very well (Allen *et al.*, 1996; Mattos *et al.*, 2006). The second application presented is to renin, a long-standing pharmaceutical target for the treatment of hypertension, with the first renin inhibitor approved in 2007 (Rahuel *et al.*, 2000; Wood *et al.*, 2003). For renin we do not have experimental mapping results, and our goal here is to show that mapping can reliably identify the 'hot spots' that substantially contribute to the free energy of ligand binding and hence should be the primary targets of drug design efforts. Two more applications are described in Supplementary Material. In addition, the FTMAP server page provides mapping results for both unbound and ligand-bound structures of 10 drug target proteins (http://ftmap.bu.edu/).

2 METHODS

2.1 Outline of the mapping algorithm

The FTMAP algorithm consists of five steps as follows.

Step 1: rigid body docking of probe molecules. Protein structures are downloaded from the Protein Data Bank (PDB) (Berman *et al.*, 2000). All bound ligands and water molecules are removed. For each structure, we use 16 small molecules as probes (ethanol, isopropanol, isobutanol, acetone, acetaldehyde, dimethyl ether, cyclohexane, ethane, acetonitrile, urea, methylamine, phenol, benzaldehyde, benzene, acetamide and *N*,*N*-dimethylformamide). The selection of this probe library is further discussed in the Supplemantary Material. For each probe, billions of docked conformations are sampled by the rigid body docking step that will be discussed in more details. Note that mapping requires only the atomic coordinates of the two molecules, i.e. no a priori information on the binding site is used. The 2000 best poses for each probe are retained for further processing.

Step 2: minimization and rescoring. The free energy of each of the 2000 complexes, generated in Step 1, is minimized using the CHARMM potential with the Analytic Continuum Electrostatic (ACE) model representing the electrostatics and solvation terms as implemented in version 27 of CHARMM (Brooks *et al.*, 1983) using the parameter set from version 19 of the program. During the minimization by an adopted basis Newton–Raphson method the protein atoms are held fixed, while the atoms of the probe molecules are free to move.

Step 3: clustering and ranking. The minimized probe conformations from Step 2 are grouped into clusters using a simple greedy algorithm. The lowest energy structure is selected and the structures within 3 Å RMSD are joined in the first cluster. The members of this cluster are removed, and we select the next lowest energy structure to start the second cluster. This step is repeated until the entire set is exhausted. Clusters with less than 10 members are excluded from consideration, thereby avoiding narrow energy minima with low entropy (Ruvinsky and Kozintsev, 2006). The clusters are ranked on the basis of their Boltzman averaged energies, and for each probe six clusters with the lowest average free energies are retained. Further details on the selection of the clustering parameters are given in the Supplementary Material.

Step 4: determination of CSs. To determine the 'hot spots' defined by overlapping probe clusters, the clusters of different probes are clustered using the distance between the centers of mass of the cluster centers as the distance measure. Again we use a simple greedy algorithm. The cluster with the maximum number of neighbors (defined as cluster centers within 4 Å) is selected as the center of CS1, and the clusters within 4 Å as the members of CS1. The clusters in CS1 are removed from consideration, and the procedure is repeated until all clusters are assigned to a CS. Having formed the CSs, their centers are fixed, but occasionally clusters have to be moved to assure that each is closest to the center of its own CS. Finally the CSs are ranked based on the number of their clusters. Duplicate clusters of the same type are considered in the count.

Step 5: characterization of the binding site. First, we select the largest CS1 that generally identifies the most important subsite (or 'hot spot'). CS1 forms the kernel of the binding site. Since additional clustering of probes close to the main CS is likely to indicate other subsites of the binding site, we expand the binding site by adding any CS (irrespective of its size) within 7 Å from any CS already in the binding site, and continue this procedure until no further expansion is possible. The probes in the resulting set of CSs are used to describe the binding site. In particular, we count the non-bonded interactions and hydrogen bonds between all atoms of these probes and the individual protein residues using the HBPLUS program (McDonald and Thornton, 1994).

2.2 The FFT correlation approach to mapping

In Step 1, we perform exhaustive evaluation of an energy function in the discretized 6D space of mutual orientations of the protein (receptor) and a small molecular probe (ligand). The center of mass of the receptor is fixed at the origin of the coordinate system. The translational space is represented as a grid of 0.8 Å displacements of the ligand center of mass, and the rotational space is sampled using 500 rotations based on a deterministic layered Sukharev grid sequence which quasi-uniformly covers the space (Lindemann *et al.*, 2004).

The energy function describing the receptor–ligand interactions is defined on the grid and is expressed as the sum of P correlation functions for all possible translations α, β, γ of the ligand at a given rotation:

$$E(\alpha,\beta,\gamma)=\sum_{p}\sum_{i,j,k}R_p(i,j,k)L_p(i+\alpha,j+\beta,k+\gamma)$$

where $R_p(i,j,k)$ and $L_p(i,j,k)$ are the components of the correlation function defined on the receptor and the ligand, respectively. This expression can be efficiently calculated using P forward and one inverse FFTs, denoted by FT and IFT, respectively:

$$E(\alpha,\beta,\gamma)=\mathrm{IFT}\Big(\sum_{p}^{P}\{\mathrm{FT}^{*}\{R_p\}\mathrm{FT}\{L_p\}\}\Big)(\alpha,\beta,\gamma)$$

$$\mathrm{FT}\{F\}(l,m,n)=\sum_{i,j,k}F(i,j,k)\exp^{-2\pi\mathbf{i}(li/N_1+mj/N_2+nk/N_3)}$$

$$\mathrm{IFT}\{f\}(i,j,k)=C\sum_{l,m,n}f(l,m,n)\exp^{2\pi\mathbf{i}(li/N_1+mj/N_2+nk/N_3)}$$

where $\mathbf{i}=\sqrt{-1}$; N_1, N_2 and N_3 are the dimensions of the grid along the three coordinates; and $C=1/(N_1N_2N_3)$. If $N_1=N_2=N_3=N$, the efficiency of this approach is $O(N^3\log(N^3))$ as compared with $O(N^6)$ when all evaluations are performed directly.

For each rotation of the ligand we generate the $\mathrm{FT}(L_p)$ function on the grid and then calculate the sum of the correlation functions using the above formula, resulting in scoring function values for all possible translations. Since the function may have multiple minima, we retain the four lowest energy regions of the translational space for each rotation. To derive the first region we select the lowest energy solution, remove the surrounding volume of the 27 Å^3, and repeat this step three more times. Finally, results from different rotations are collected and sorted.

2.3 Energy function in the rigid body docking step

The energy expression in Step 1 includes the simplified van der Waals energy E_{vdw} with attractive (E_{attr}) and repulsive (E_{rep}) contributions, the electrostatic interaction energy E_{elec}, a cavity term E_{cavity} describing the contributions from hydrophobic enclosures and the statistical knowledge-based pairwise potential E_{pair} representing other solvation effects:

$$\begin{aligned}
E &= E_{\mathrm{vdw}}+w_2E_{\mathrm{elec}}+w_3E_{\mathrm{cavity}}+w_4E_{\mathrm{pair}}\\
E_{\mathrm{vdw}} &= E_{\mathrm{attr}}+w_1E_{\mathrm{rep}}\\
E_{\mathrm{elec}} &= \sum_{i=1}^{N_l}q_i\phi_{rPB}\\
E_{\mathrm{pair}} &= \sum_{i=1}^{N_R}\sum_{j=1}^{N_L}\varepsilon_{ij}
\end{aligned}$$

where N_R and N_L denote the numbers of atoms in the receptor and the ligand, respectively. The coefficients $w_1=11.1$, $w_2=44.4$, $w_3=0.88$ and $w_4=3.33$ weight the different contributions to the scoring function. The weight w_1 of the repulsive term is selected to avoid substantial steric clashes, but to allow for some atomic overlaps accounting for the (limited) flexibility of the receptor. The coefficients w_2 and w_3 are selected according to calorimetric considerations and are scaled to the van der Waals term. The goal of adding the attractive cavity term E_{cavity} is to bias the sampling toward the cavities of the protein. In the remainder of this section we describe the biophysical origin of these energy terms, as well as their implementation on a grid.

2.3.1 van der Waals energy We use stepwise functions to represent the attractive and repulsive steric terms as suggested by Vakser (Vakser and Aflalo, 1994). The repulsive interactions are cutoff at the van der Waals radius ($r_{\mathrm{vdw})}$ plus 1.8 Å because we want the penalty function to be tolerant enough and to allow for differences between bound and unbound structures. The attractive part is truncated at 6 Å. On the grid, the functions describing the receptor and the ligand can be represented as follows.

$$\begin{aligned}
R_p(l,m,n) &= -c_{l,m,n}+w_1r_{l,m,n}\\
L_p(l,m,n) &= \begin{cases}1 \text{ if } & (l,m,n)\ni(a_j\in J)\\ 0 \text{ otherwise}\end{cases}
\end{aligned}$$

where $c_{l,m,n}$ is the number of atoms that are at the distance $d<r<D$ from the grid point (l,m,n), $r_{l,m,n}$ is the number of atoms that are at the distance $r<d$ from the same grid point and $(l,m,n)\ni(a_j\in J)$ means that the grid point (l,m,n) overlaps with atom a_j of atom type J. As mentioned, $D=6$ Å and $d=r_{\mathrm{vdw}}+1.8$ Å. Thus, the correlation of R_p and L_p provides a shape complementarity term representing both repulsive and attractive interactions, the former for the distances $r<d$, and the latter in the range $d<r<D$.

2.3.2 Electrostatic interactions We approximate the electrostatic energy as the interaction energy between the electrostatic potential ϕ_{rPB} of the solvated protein and the atomic charges q_i of the probe. Thus, the influence of the probe on the electrostatic potential of the protein-solvent system is neglected, assuming that the probe is small and not strongly charged. Using a dielectric continuum model with low ion concentration (corresponding to 0.1 mol salt concentration), the electrostatic potential of the solvated protein is calculated by solving the linearized Poisson–Boltzmann equation

$$\nabla\epsilon(\overrightarrow{r})\nabla\phi_{rPB}(\overrightarrow{r})-\kappa^2(\overrightarrow{r})\phi_{rPB}(\overrightarrow{r})=-4\pi\rho(\overrightarrow{r}),$$

where $\epsilon(\overrightarrow{r})$, $\kappa(\overrightarrow{r})$ and $\rho(\overrightarrow{r})$ are the dielectric constant, the modified Debye–Hückel screening factor, and the fixed charge density of the protein, respectively. The dielectric boundary between the low dielectric protein region and the external bulk solvent is placed to account for the reduced water mobility and hence reduced polarization in binding site cavities. This is achieved by dividing atoms of the protein into 'cavity' and 'non-cavity' sets. Atoms are considered 'cavity' if they are not accessible to a large spherical probe of 5.75 Å radius. The size of the probe is selected to represent the typical dimensions of protein active sites. Each cavity atom is assigned a dielectric radius equal to its van der Waals radius ($r_{\mathrm{vdw})}$ plus 1.4 Å. In contrast, each non-cavity atom has a small fixed dielectric radius of 0.1 Å. These radii define a continuous surface that separates the low dielectric protein and its extension into the cavities ($\epsilon=4.0$) from the bulk solvent ($\epsilon=80.0$). We use the Poisson–Boltzmann module PBEQ (D.Beglov and B.Roux, unpublished data) of CHARMM (Brooks *et al.*, 1983) to calculate the potential ϕ_{rPB}. The electrostatic energy is then expressed as the vector product of the functions

$$\begin{aligned}
R_p(l,m,n) &= \phi_{rPB}(l,m,n)\\
L_p(l,m,n) &= \begin{cases}q_j \text{ if } & (l,m,n)\ni(a_j\in J)\\ 0 \text{ otherwise}\end{cases}
\end{aligned}$$

defined on the receptor and on the ligand grids, respectively. The potential is truncated at 15.0 kcal/mol.

2.3.3 Pairwise statistical potential The general form of a pairwise contact potential is

$$E_{\mathrm{pair}}=\sum_{i=1}^{N_R}\sum_{j=1}^{N_L}\varepsilon_{ij}.$$

For a pair of atoms a_i and a_j of types I and J, respectively, $\varepsilon_{ij}=\varepsilon_{IJ}$, where ε_{IJ} is the contact energy between atoms of types I and J, if $d<r_{ij}<D$; otherwise $\varepsilon_{ij}=0$. We use the DARS (decoys as the reference state) potential that originally has been developed for protein–protein docking (Chuang *et al.*, 2008; Kozakov *et al.*, 2006), and has been extended to describe the interactions between proteins and the molecular probes considered here. The DARS parameters used in this work are listed in the Supplementary Material. In order to evaluate the energy function using FFT it must be written as a sum of correlation functions. Based on the eigenvalue–eigenvector decomposition of the matrix of pairwise interaction coefficients ε_{IJ}, these coefficients can be written as

$$\varepsilon_{IJ}=\sum_{p=1}^{K}\lambda_pu_{pI}u_{pJ}$$

where λ_p is the p-th eigenvalue of the interaction matrix, and u_{pI} is the I-th component of the p-th eigenvector. Each term in the eigenvalue–eigenvector decomposition represents an energy contribution proportional to the absolute value of the eigenvalue λ_p, and such contributions are independent due to the orthogonality of the eigenvectors. We have shown that restricting consideration to the first four terms yields around 10% error in the energy values, comparable with the error of representing the energies on a grid (Kozakov *et al.*, 2006). The energy term with the p-th eigenvalue of the pairwise potential is defined by the correlation of the functions

$$R_p(l,m,n) = \sum_{i=1}^{N_r} u_{pI}\delta_i$$

$$L_p(l,m,n) = \begin{cases} u_{pJ} & \text{if} \quad (l,m,n) \ni (a_j \in J) \\ 0 & \text{otherwise} \end{cases}$$

where δ_i is 1 if atom i of the receptor is closer than 6 Å to the grid point (l,m,n).

2.3.4 Cavity term Non-polar cavities disrupt water structure and create favorable environment for ligand binding (Young *et al.*, 2007). To represent this effect we place a Gaussian ball, with $\sigma = 10$ Å, at each grid point, and calculate its correlation with the C_α atoms of non-polar residues. For each point in space, this function measures the fraction of the ball occupied by the non-polar regions of the protein:

$$R_p(l,m,n) = \sum_{i=1}^{N_R} \frac{1}{\sigma\sqrt{2\pi}} \exp(\frac{-r^2_{i,(lmn)}}{\sigma^2})$$

$$L_p(l,m,n) = \begin{cases} 1 \text{ if} \quad (l,m,n) \ni A_L \\ 0 \text{ otherwise} \end{cases}$$

where N_R is the number of atoms in the receptor, $r_{i,(lmn)}$ is the distance from atom i to grid point (l,m,n), and A_L is an atom of the ligand probe. The correlation of R_p and L_p is large in non-polar cavities, and it is small on protrusions and flat surfaces. This way the function, added to the energy expression with a negative sign, improves the sampling of the cavity regions.

3 RESULTS AND DISCUSSION

3.1 The mapping of elastase

Porcine pancreatic elastase is a serine endopeptidase with two β-barrel domains. The catalytic residues Ser203, His60 and Asp108 lie in a cleft between the domains and the peptide substrate binding site lies roughly perpendicular to the inter-domain cleft, with the $S_4 - S_1$ subsites in one domain and the $S'_1 - S'_3$ in the other. Mattos and co-workers (Allen *et al.*, 1996; Mattos *et al.*, 2006) solved the crystal structure of elastase in the presence of 95% acetone, 55% dimethylformamide, 80% 5-hexene-1,2-diol, 80% isopropanol, 80% ethanol and 40% trifluoroethanol. In addition, a crystal structure was solved in a mixture of 40% benzene, 50% isopropanol and 10% water, and one in 40% cyclohexane, 50% isopropanol and 10% water. Isopropanol was bound to elastase in these last two structures, but there were no bound benzene or cyclohexane molecules observed in the electron density maps. Most other organic solvents cluster in the active site, delineating five subsites (Fig. 1A and B). The S_1 pocket binds acetone, dimethylformamide, 5-hexene-1,2-diol, ethanol, two molecules of trifluoroethanol and three molecules of isopropanol (Fig. 1A). Further probe binding in the active site occurs at S_4, the oxyanion hole, and two sites on the leaving group side of the catalytic triad close to the S'_3 and S'_1 pockets (Fig. 1B). S'_1 binds only trifluoroethanol, but all the other subsites bind at least three different types of solvent molecules.

We have mapped an elastase structure co-crystalyzed with the inhibitor trifluoroacetyl-L-lysyl-L-prolyl-*p*-isopropylanilid (PDB code 1ela) (Mattos *et al.*, 1994). The structure also includes an acetic acid molecule, a sulfate ion, a calcium ion and 166 water molecules. All ligands, ions and water molecules were removed before the mapping. The largest CS (CS1) is a supercluster of 20 probe clusters in the S_1 pocket (Fig. 1C). All the 16 probes are represented here (including three clusters of benzaldehyde), even benzene and cyclohexane that were not observed to bind in the MSCS experiments (Mattos *et al.*, 2006). Figure 1D shows CS1 (cyan) and the adjacent CSs. CS4 (salmon) is located in the S'_3 pocket and includes nine probe clusters. CS5 (sky blue) is in the S'_1 pocket, and includes eight probe clusters. CS7 (orange) has seven probe clusters in the oxyanion hole. Finally, CS8 (pale green) is in the S_4 subsite and has four probe clusters. A comparison of Figure 1B and D shows that using FTMAP we identify all the five subsites that bind any organic molecule in the experiments.

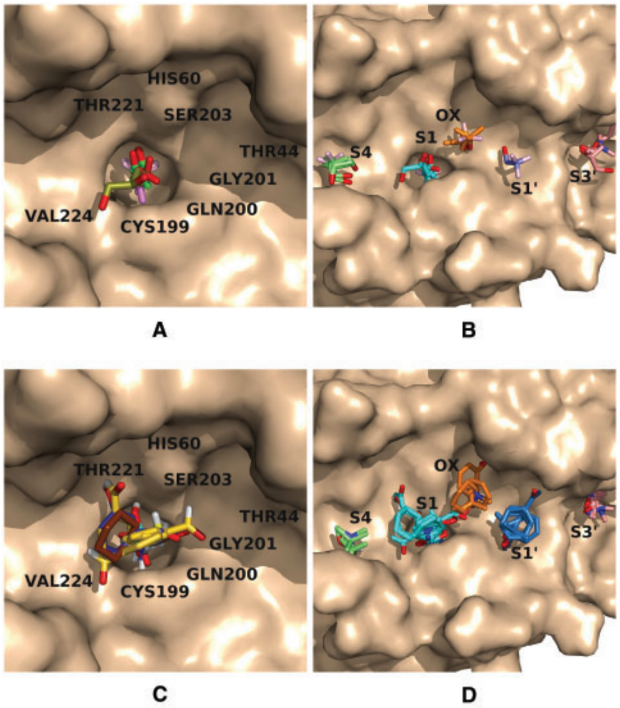

Fig. 1. Binding of organic solvents to elastase, determined by X-ray crystallography (Allen *et al.*, 1996; Mattos *et al.*, 2006) and computational mapping. (**A**) Probe molecules in the S_1 pocket of elastase, based on superimposing elastase structures solved in acetone, dimethylformamide, 5-hexene-1,2-diol, isopropanol, ethanol and trifluoroethanol. Probes are color-coded to distinguish between different molecules. (**B**) Probe molecules in the active site of elastase, based on superimposing the structures listed in (A). The probes are color-coded to distinguish between the different subsites. (**C**) Centers of probe clusters in the largest CS of elastase, located in the S_1 pocket, from mapping the protein using the 16 probes. Probes are color-coded to distinguish between different molecules. (**D**) Centers of probe clusters in the five CSs located in the active site of elastase as determined by the mapping. The probes are color-coded to distinguish between the different CSs: CS1, cyan; CS4, salmon; CS5, sky blue; CS7, orange; and CS8, pale green.

Figure 2A and B compare the protein–probe interactions and hydrogen bonds from the experimental and computational mapping results shown in Figure 1B and D, respectively. The most important residues in the S_1 subsite are Gln200, Cys199, Val224 and Thr221 (Fig. 2A). The interactions are primarily provided by the aliphatic portions of Gln200 and Thr221, and none of these residues participate in hydrogen bonds (Fig. 2B), in good agreement with the preference for substrates with apolar residues in the S_1 pocket

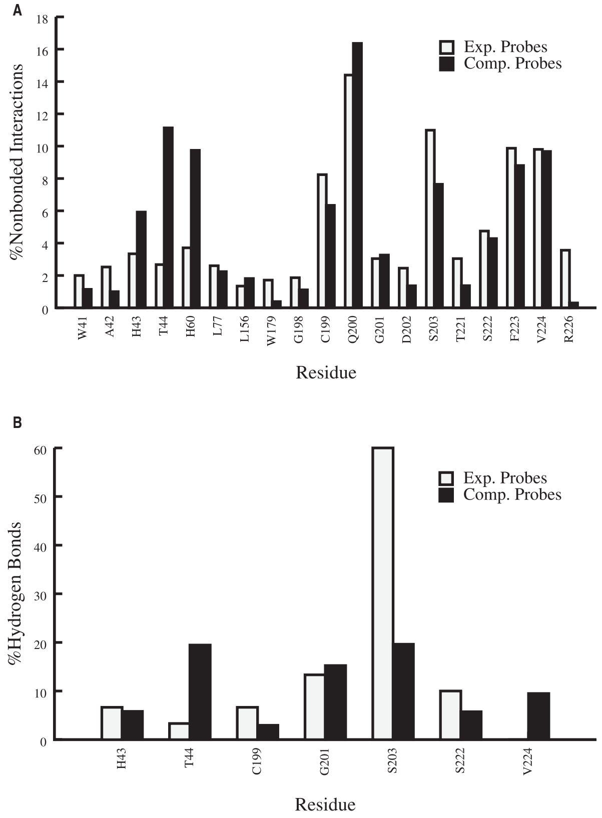

Fig. 2. (**A**) Intermolecular non-bonded interactions between probes and elastase residues, determined by X-ray crystallography (Mattos *et al.*, 2006) and computational mapping. The experimental and computational results are based on the interactions found between various elastase residues and the probes in the clusters shown in Figure 1B and D, respectively. (**B**) The same as panel (A), but for hydrogen bonds rather than non-bonded interactions.

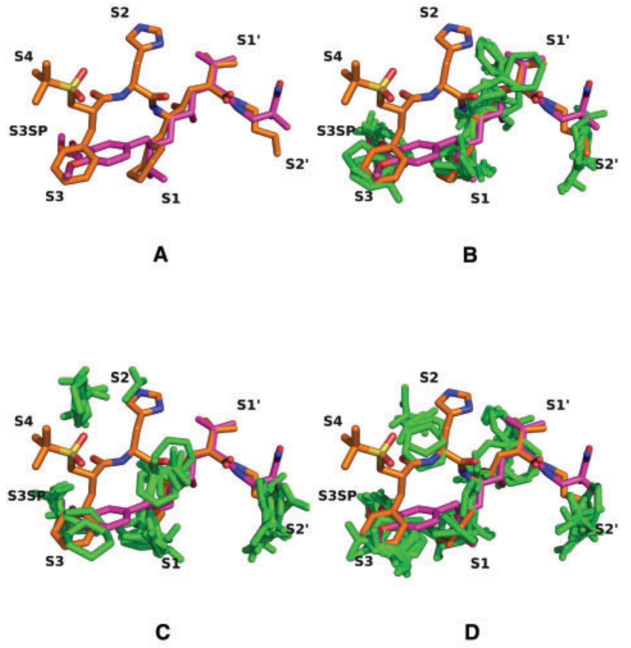

Fig. 3. Mapping results for renin. (**A**) Superposition of aliskiren (magenta) and a peptidomimetic inhibitor (orange) in the binding site of renin. The important subsites are labeled. (**B**) Mapping of the aliskiren-bound structure (PDB code 2v0z). Shown are the clusters centers in the CSs located in the renin-active site, superimposed on the inhibitor structures shown in panel (A). (**C**) Mapping of the ligand-free renin structure (PDB code 2ren). Cluster centers as in panel (B). (**D**) Mapping of renin co-crystallized with a peptidomimetic inhibitor (PDB code 1rne). Cluster centers as in panel (B).

(Mattos *et al.*, 1994). The next set of important residues, Ser203, Thr44, His60 and Gly201, are in the largely polar oxyanion hole. Several probes also form hydrogen bonds with Ser203, Thr44 and the backbone of Gly201 (Fig. 2B). The S_4 subsite is lined by Phe223 and the aliphatic portion of Arg226. The only other important residues are in S'_3 (His43 and Leu77) and in S'_1 (Leu156, which also interacts with S'_3). The binding of probes coincides with the way these sites are occupied by elastase inhibitors. For example, in the Supplementary Material we compare the residue–probe contacts from our mapping results with the contacts between the elastase residues and a peptidomimetic aminimide inhibitor (Peisach *et al.*, 1995). The agreement is excellent for all important residues.

3.2 Renin

Renin, an aspartic endopeptidase, catalyzes the rate-limiting step in the renin-angiotensin system (RAS), the cleavage of the Leu10-Val11 peptide bond of angiotensinogen to form angiotensin I. Angiotensin I is converted by the angiotensin converting enzyme (ACE) into angiotensin II, which increases blood pressure. Renin is a long-standing pharmaceutical target for the treatment of hypertension, and a variety of stable peptide-like analogs of angiotensinogen have been developed over the past 20 years that were shown to inhibit renin and lower blood pressure (Fisher and Hollenberg, 2001). The peptidomimetic inhibitors, however, suffer from limited bioavailability due to poor lipophilicity and a relatively large size.

Aliskiren, the first orally effective renin inhibitor with better pharmaceutical properties, was approved for the treatment of hypertension in 2007 (Rahuel *et al.*, 2000; Wood *et al.*, 2003). Figure 3A shows the renin-bound aliskiren structure (PDB code 2v0z, shown in magenta), superimposed with a typical peptidomimetic inhibitor (PDB code 1rne, shown in orange). The peptide-like inhibitors bind to renin subsites S_4-S_1, S'_1 and S'_2, while aliskiren is lacking the P_4 and P_2 analogs, and hence it does not interact with the S_2 or S_4 subsites. The P_3 to P_1 pharmacophore of aliskiren is accommodated by the complementary pockets S_3 and S_1 (Wood *et al.*, 2003). Aliskiren has an additional group binding to the S_3^{sp} subsite, a deep, narrow and fairly polar pocket behind S_3, which is considered essential for high binding affinity (Rahuel *et al.*, 2000).

We present mapping results for three renin structures: (i) co-crystallized with aliskiren (PDB code 2v0z), (ii) ligand-free (PDB code 2ren) and (iii) co-crystallized with a peptidomimetic inhibitor (PDB code 1rne). Table 1 lists the largest 10 CSs located in the active site for each structure. The three largest CSs for the aliskiren-bound structure are the subsites S_1 (extending toward the S'_1 pocket),

Table 1. Summary of renin mapping results

No.	2V0Z		2REN		1RNE	
	Aliskiren		Apo-Enzyme		Peptidomimetic	
	Size	Pocket	Size	Pocket	Size	Pocket
1	18	$S_1-S'_1$	22	S'_2	14	S'_2
2	13	S_3^{SP}	16	$S_1-S'_1$	14	S_3^{SP}
3	12	S'_2	12	S_3^{SP}	12	$S_1-S'_1$
4	11	–	10	S_1	10	S_3
5	10	–	9	S_2-S_4	8	S_1
6	8	S'_1	6	–	7	–
7	7	S'_1	5	–	6	–
8	7	–	4	S_3	6	–
9	4	S_3	3	–	5	S'_1
10	2	–	3	–	4	S_2-S_4

S_3^{sp} and S'_2, with 18, 13 and 12 probe clusters, respectively (Table 1 and Fig. 3B). There are two much smaller superclusters in S'_1, and another in S_4. Since this is the aliskiren-bound structure, one could assume that the large S_3^{sp} subsite is induced by ligand binding.

However, this is definitely not the case. As shown in Table 1 and Figure 3C, the mapping of the renin apostructure (2ren) places the third largest CS (CS3), with 12 probe clusters, in the S_3^{sp} subsite, thus the pocket exists before any ligand binding. The S_3^{sp} subsite is also found, with 14 probe clusters, in the structure co-crystallized with the peptidomimetic inhibitor (PDB code 1rne). More generally, in all three structures the three most important subsites are the same, i.e. S'_2, S_3^{sp} and S_1, the latter extending toward S'_1, and only the relative sizes (i.e. the numbers of probe clusters) vary. The most important difference between the three structures is a CS between the S_2 and S_4 subsites, present in both the apostructure (Fig. 3C) and the peptidomimetic-bound protein (Fig. 3D), but absent in the aliskiren-bound structure (Fig. 3B). However, this site is relatively small, containing substantially fewer probe clusters than S_3^{sp}.

Figure 4A shows the distribution of non-bonded interactions that the bound peptidomimetic, aliskiren and the probes establish with individual renin residues. The probe interactions are based on mapping the structure co-crystallized with the peptidomimetic inhibitor rather than aliskiren. Nevertheless, the residues in the S_3^{sp} subsite (Thr12, Gln13 and Tyr14) interact with many probes. Figure 4A also shows that these residues, particularly Tyr14 deep in the S_3^{sp} pocket, interact with aliskiren, but slightly or not at all with the peptidomimetic.

Tyr14 also forms hydrogen bonds with both aliskiren and the probes, but not with the peptidomimetic (Fig. 4B). In contrast, residues Tyr220, Ser222, His287 and Met289 in the S_2 subsite that interact with the peptidomimetic but not with aliskiren attract very few probes, although Figure 4A is based on the mapping of the peptidomimetic-bound structure. Other residues interacting with the peptidomimetic backbone are Ser76 and Thr77, both located between the S_1 and S_2 pockets. These residues also bind very few probes, attesting that the probe clusters trace the shape of aliskiren rather than that of the peptide. We note that Thr77 forms hydrogen bonds with the peptide, but with neither aliskiren nor the probes (Fig. 4B).

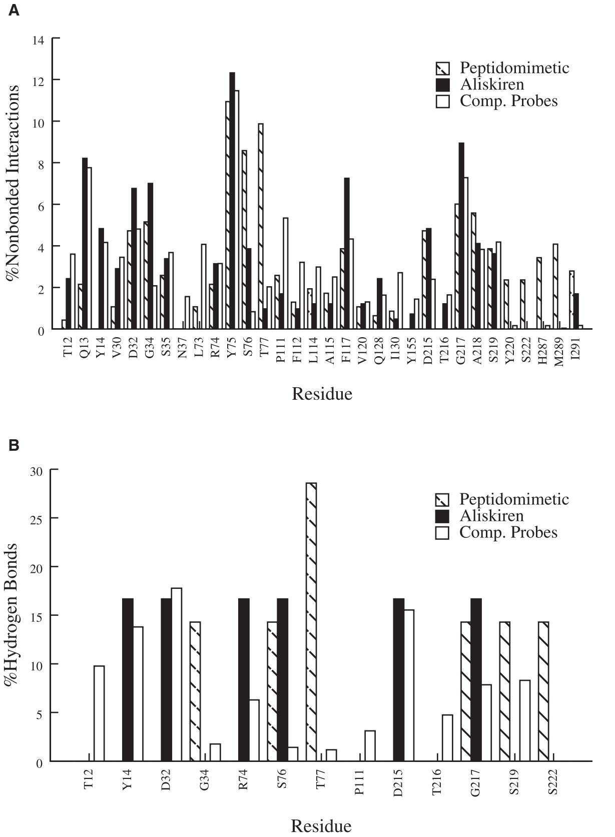

Fig. 4. (**A**) Intermolecular non-bonded interactions between inhibitors, probes and renin residues. The interactions are shown for the peptidomimetic inhibitor, aliskiren and the probes. The probe interactions are based on mapping the structure co-crystallized with the peptidomimetic inhibitor (PDB code 1rne). Two atoms are considered to interact if located within 5 Å from each other. (**B**) Same as in panel (A), but hydrogen bonds rather than non-bonded interactions.

4 CONCLUSION

Finding the most druggable pockets of target proteins, and identifying molecular fragments or functional groups that tend to bind there, are important steps in drug design. Such information can be obtained, at considerable costs, by determining the structure of a protein in a number of organic solvents by X-ray crystallography, or by NMR-based screening programs that detect the binding of fragment-sized compounds. In both methods, it has been observed that the small organic compounds cluster in the important pockets of the binding site, and this provides information on their druggability. Here, we describe the FTMAP algorithm which can be used to perform a close computational analog of such experimental druggability analyses. The main novelty of FTMAP is that it combines the highly efficient FFT correlation method of sampling protein–probe complexes with the use of a detailed and highly accurate energy function for scoring the complex conformations.

The two applications presented here demonstrate the type of information that can be obtained by FTMAP. For elastase the algorithm correctly identifies all important subsites, in good agreement with the results obtained by X-ray crystallography. For renin, a long-standing pharmaceutical target, the few largest CSs trace out the shape of the first approved renin inhibitor, aliskiren (Rahuel *et al.*, 2000; Wood *et al.*, 2003), rather than that of peptidomimetic inhibitors that have been studied for several decades but did not provide any successful drug candidate. It is important that the mapping reveals the better fit of aliskiren into the 'hot spots' even when applied to a renin structure without any bound ligand, or to structures co-crystallized with peptidomimetic inhibitors. Two more applications are described in the Supplementary Material.

FTMAP is substantially more efficient than our earlier mapping method CS-Map (Dennis *et al.*, 2002; Silberstein *et al.*, 2003). The FFT-based search is twice as fast as the non-linear minimization used in CS-Map, and takes about 30 min of CPU time on a single 1 GHz PIII processor for a small protein. The primary speed-up is achieved in the local minimization. Due to the improved sampling by FTMAP, for each probe it is sufficient to refine the 2000 lowest energy poses rather than over 6000 as we do in CS-Map (Dennis *et al.*, 2002; Silberstein *et al.*, 2003). Since the refinement of a single protein–probe complex (with the protein fixed) takes about 30 s, for each probe the minimization on a single CPU is reduced from 54 h to 18 h. Although the minimizations can be easily distributed among multiple processors, this is still a substantial gain, improving the applicability of the method. FTMAP is available as a server at http://ftmap.bu.edu/. The server requires only the PDB file or PDB code of the protein to be mapped. The output data include (i) a PDB file with the original protein and the six lowest energy cluster representatives for each probe type, grouped into CSs; (ii) the number of non-bonded interactions between the probes and each residue; and (ii) the number of hydrogen bonds between the probes and each residue. The FTMAP program is also available to academic users free of charge.

Funding: National Institutes of Health (GM064700 to S.V.); National Institute of Environmental Health Sciences (P42ES07381 to S.V.); National Science Foundation (MCB-0237297 to C.M.).

Conflict of Interest: none declared.

REFERENCES

Allen,K. *et al.* (1996) An experimental approach to mapping the binding surfaces of crystalline proteins. *J. Phys. Chem.*, **100**, 2605–2611.

An,J.H. *et al.* (2005) Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol. Cell. Proteomics*, **4**, 752–761.

Berman,H. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Brooks,B.R. *et al.* (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.*, **4**, 187–217.

Caflisch,A. *et al.* (1993) Multiple copy simultaneous search and construction of ligands in binding sites: application to inhibitors of HIV-1 aspartic proteinase. *J. Med. Chem.*, **36**, 2142–2167.

Chuang,G. *et al.* (2008) DARS (decoys as the reference state) potentials for protein-protein docking. *Biophys. J.*, **95**, 4217–4127.

DeLano,W.L. (2002) Unraveling hot spots in binding interfaces: progress and challenges. *Curr. Opin. Struct. Biol.*, **12**, 14–20.

Dennis,S. *et al.* (2002) Computational mapping identifies the binding sites of organic solvents on proteins. *Proc. Natl Acad. Sci. USA*, **99**, 4290–4295.

English,A.C. *et al.* (1999) Locating interaction sites on proteins: the crystal structure of thermolysin soaked in 2% to 100% isopropanol. *Proteins*, **37**, 628–640.

English,A.C. *et al.* (2001) Experimental and computational mapping of the binding surface of a crystalline protein. *Protein Eng.*, **14**, 47–59.

Fisher,N. and Hollenberg,N. (2001) Is there a future for renin inhibitors? *Expert Opin. Investig. Drugs*, **10**, 417–426.

Glaser,F. *et al.* (2006) A method for localizing ligand binding pockets in protein structures. *Proteins*, **62**, 479–488.

Goodford,P.J. (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.*, **28**, 849–875.

Hajduk,P.J. *et al.* (2005) Druggability indices for protein targets derived from NMR-based screening data. *J. Med. Chem.*, **48**, 2518–2525.

Kozakov,D. *et al.* (2006) PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins*, **65**, 392–406.

Laurie,A.T.R. and Jackson,R.M. (2005) Q-sitefinder: an energy-based method for the prediction of protein–ligand binding sites. *Bioinformatics*, **21**, 1908–1916.

Lindemann,S. *et al.* (2004) Incremental grid sampling strategies in robotics. In *Proceedings of the Sixth International Workshop on the Algorithmic Foundations of Robotics*. Springer, Berlin/Heidelberg, Zeist, Netherlands, pp. 313–328.

Mattos,C. and Ringe,D. (1996) Locating and characterizing binding sites on proteins. *Nat. Biotechnol.*, **14**, 595–599.

Mattos,C. *et al.* (1994) Analogous inhibitors of elastase do not always bind analogously. *Nat. Struct. Biol.*, **1**, 55–58.

Mattos,C. *et al.* (2006) Multiple solvent crystal structures: probing binding sites, plasticity and hydration. *J. Mol. Biol.*, **357**, 1471–1482.

McDonald,I. and Thornton,J. (1994) Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, **238**, 777–793.

Miranker,A. and Karplus,M. (1991) Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins*, **11**, 29–34.

Peisach,E. *et al.* (1995) Interaction of a peptidomimetic aminimide inhibitor with elastase. *Science*, **269**, 66–69.

Rahuel,J. *et al.* (2000) Structure-based drug design: the discovery of novel nonpeptide orally active inhibitors of human renin. *Chem. Biol.*, **7**, 493–504.

Ruvinsky,A.M. and Kozintsev,A.V. (2006) Novel statistical-thermodynamic methods to predict protein–ligand binding positions using probability distribution functions. *Proteins*, **62**, 202–208.

Silberstein,M. *et al.* (2003) Identification of substrate binding sites in enzymes by computational solvent mapping. *J. Mol. Biol.*, **332**, 1095–1113.

Vajda,S. and Guarnieri,F. (2006) Characterization of protein-ligand interaction sites using experimental and computational methods. *Curr. Opin. Drug. Discov. Dev.*, **9**, 354–362.

Vakser,I. and Aflalo,C. (1994) Hydrophobic docking: a proposed enhancement to molecular recognition techniques. *Proteins*, **20**, 320–329.

Wood,J.M. *et al.* (2003) Structure-based design of aliskiren, a novel orally effective renin inhibitor. *Biochem. Biophys. Res. Commun.*, **308**, 698–705.

Young,T. *et al.* (2007) Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding. *Proc. Natl Acad. Sci. USA*, **104**, 808–813.