Journal of Computer-Aided Molecular Design **17:** 173–186, 2003.

Improved mapping of protein binding sites

Tamas Kortvelyesi[1,2], Michael Silberstein[3], Sheldon Dennis[1,4] & Sandor Vajda[1,*]
[1]*Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02215, USA;* [2]*Department of Physical Chemistry, University of Szeged, H-6720 Szeged, Rerrich B. Sq 1., Hungary;* [3]*Program in Bioinformatics, Boston University, Boston, Massachusetts 02215, USA;* [4]*Current address: Biotechnology Research Institute, Montreal, Canada*

MS received 29 November 2002; accepted for publication 3 January 2003

Key words: ligand binding, drug design, ligand motion, site mapping, enzyme active site, consensus site, reaction path, molecular dynamics.

Summary

Computational mapping methods place molecular probes – small molecules or functional groups – on a protein surface in order to identify the most favorable binding positions by calculating an interaction potential. Mapping is an important step in a number of flexible docking and drug design algorithms. We have developed improved algorithms for mapping protein surfaces using small organic molecules as molecular probes. The calculations reproduce the binding of eight organic solvents to lysozyme as observed by NMR, as well as the binding of four solvents to thermolysin, in good agreement with x-ray data. Application to protein tyrosine phosphatase 1B shows that the information provided by the mapping can be very useful for drug design. We also studied why the organic solvents bind in the active site of proteins, in spite of the availability of alternative pockets that can very tightly accommodate some of the probes. A possible explanation is that the binding in the relatively large active site retains a number of rotational states, and hence leads to smaller entropy loss than the binding elsewhere else. Indeed, the mapping reveals that the clusters of the ligand molecules in the protein's active site contain different rotational-translational conformers, which represent different local minima of the free energy surface. In order to study the transitions between different conformers, reaction path and molecular dynamics calculations were performed. Results show that most of the rotational states are separated by low free energy barriers at the experimental temperature, and hence the entropy of binding in the active site is expected to be high.

Introduction

The well-established approach to drug design, exemplified by DOCK [1], is based on screening a database of small compounds for binding to a target protein. The method works only if the database includes a molecule with significant affinity for the target. There has been substantial interest in fragment-based approaches that identify favorable positions for weakly binding fragments using molecular probes, i.e., functional groups or small organic molecules, and then combine them to form a compound with significant affinity. The idea has been introduced by Goodford [2] in the GRID program, which places small probes (water, methyl group, amine nitrogen, carboxyl oxygen, and hydroxyl) at regularly spaced grid points within the active site, determining the most favorable scores in a grid-based energy evaluation. The resulting energy contour surfaces delineate regions of attraction between probe and protein, and provide information on favorable positions for a number of functional groups. With some local minimization, a small number of positions can be identified for each group, at which it is both sterically acceptable and is likely to interact favorably with the surrounding side chains of the receptor.

Other drug design programs also employ site mapping as their first step, and then connect the individual molecular fragments into a single, viable molecule.

*Corresponding author. E-mail: vajda@bu.edu

CLIX [3] screens a small molecular database and, for each molecule, attempts is make to make a pair of the substituent's chemical groups spatially coincide with a pair of favorable interaction sites proposed by GRID in the binding site of the protein. LUDI [4] positions molecular fragments into the binding site of an enzyme in such a way that hydrogen bonds with the enzyme are formed, and hydrophobic pockets are filled. These fragments are then linked together by suitable spacers. HOOK [5] employs Multiple Copy Simultaneous Search (MCSS) in the mapping stage, and the connection of the minima found is performed by searching a database of molecular scaffolds for possible connectors. MCSS, an interesting approach to mapping, uses numerous ligand copies, each transparent to the others but subject to the full force of the receptor [6–8]. Caflisch and co-workers [6] used MCSS to map a binding site, and constructed possible ligands by building bonds to connect the various minima found.

In spite of their extensive use in drug design, the mapping methods exemplified by GRID and MCSS have a major shortcoming, which is the large number of resulting energy minima [9]. In contrast, current experimental methods reveal that the binding of small, non-specific ligands occur only at a limited number of sites. In particular, Mattos and Ringe solved the X-ray structure of elastase in a variety of organic solvents [9–11]. English et al. [12, 13] applied the same method to thermolysin (TLN), and determined high resolution crystal structures in aqueous solutions of isopropanol, acetone, acetonitrile, and phenol. NMR also provides a method for detecting the binding of small molecules to a protein in solution [14–17]. These results show that the protein structure remains virtually identical to the native structure in the selected organic solvents, and that only a limited number of organic molecules associate with the protein surface. The analysis becomes particularly interesting when five or six structures of a protein, solved in different solvents, are superimposed [9, 11]. In the case of an enzyme, the different solvent molecules tend to cluster in the active site, forming a "consensus" site that delineates the binding pockets. All other binding sites are either in crystal contact, occur only at high ligand concentrations, or are in small, buried pockets that bind only a subset of the solvent molecules rather than all of them.

The limited number of bound organic solvents seen in the x-ray and NMR experiments clearly contradicts to the many minima obtained by the current computational mapping methods [9, 13]. In addition, MCSS generally assigns different preferred sites to different groups [13], in contrast to the observation that such regions overlap [9, 13]. These problems with computational mapping may be responsible for the fact that, in spite of their great promise, drug design methods based on the site mapping and fragment assembly approach seem to be less successful than the ones that use direct docking of small compounds [1]. In an attempt to overcome this difficulty, a number of experimental methods that identify small molecules binding to proteins in a weakly specific manner have been recently developed, including the already mentioned Multiple Solvent Crystal Structures approach [9–13], SAR by NMR [15, 16], and a site-directed ligand discovery technique based on tethering [18]. The fact that these methods are increasingly used as tools for systematic screening in drug discovery [15, 18, 19] demonstrates that the interest in mapping is not waning.

In this paper we describe the CS-Map algorithm, an improved procedure for the mapping of protein binding sites [20]. The algorithm differs from traditional mapping methods in three major respects: (1) there is better sampling of protein surface regions with favorable electrostatics and desolvation, (2) the scoring potential accounts for desolvation, and (3) the docked ligand positions are clustered, and the clusters are ranked on the basis of their average free energies. The combination of these elements yields an algorithm that shows remarkable robustness against variations in the protein structure and changes in energy parameters. We have shown that all three factors (i.e., appropriate sampling, desolvation term in the scoring potential, and ranking on the basis of cluster free energies) are needed to avoid spurious local energy minima [21]. Similarly to experimental solvent mapping, computational mapping places clusters of different probes in the active site, and thus provides a powerful tool for the identification and characterization of protein active sites. We present applications to hen egg white lysozyme (HEWL) and thermolysin (TLN), because these proteins have been experimentally mapped using a number of organic solvents. In addition, we map the binding site of protein tyrosine phosphatase 1B (PTP-1B), an extensively studied drug target, in order to demonstrate that the mapping provides information that can be useful for drug design.

In the second part of the paper we try to better understand why different organic molecules cluster in the active site, regardless of their size and polarity. Both solvent mapping experiments and our calcula-

tions confirm that such clustering exists, although there are many other pockets on a protein surface that could bind the probes [9, 13]. Our hypothesis is that the ligand bound in the active site retains more conformational freedom and hence loses less entropy than ligands bound anywhere else. Indeed, the mapping results reveal that the consensus site accommodates each ligand in a number of conformational states. The large thermal factors of the bound organic solvents in the x-ray structures also confirm the existence of multiple states. Thus, after soaking the protein in the water/organic solution, either the organic solvent molecules bind in different conformations with different primary interactions with the residues in the binding pocket, or the bound small molecules can move (i.e. rotate/translate) among a number of states, thereby alternating their interactions with the residues in different positions. It means that the bound state retains substantial conformational entropy, and these entropic contributions to the binding free energy may render the usually large active site more favorable than the generally much smaller other crevices on the protein surface. To classify the type of the motion and its role in the binding of small organic solvent molecules in the binding site(s), we describe preliminary results on the possible transitions between local free energy minima, obtained by either reaction paths calculations or molecular dynamics simulations.

Materials and methods

The CS-Map algorithm

The five computational steps of the algorithm are as follows [20, 21]:

Step 1: Rigid body search for regions with favorable electrostatics and desolvation. We use a multi-start simplex method [22] to move the probes around the protein, starting from a number of evenly distributed points over the entire protein surface. Thus, no *a priori* assumption is made about the location of the binding site. The scoring function in the search is given by

$$\Delta G_s = \Delta E_{elec} + \Delta G_{des} + V_{exc} \quad (1)$$

where ΔE_{elec} denotes the direct (coulombic) part of the electrostatic energy, ΔG_{des} is the desolvation free energy, and V_{exc} is an excluded volume penalty term such that $V_{exc} = 0$ if the ligand does not overlap with the protein. Note that ΔG_s does not include a van der Waals term. Indeed, we assume that the solute-solute and solute-solvent interfaces are equally well packed, and hence the intermolecular van der Waals interactions in the bound state are balanced by solute-solvent interactions in the free state [23–26].

As described in [20], in this preliminary search step we use a quasi-Coulombic model for the electrostatic interactions, i.e., determine the electrostatic field Φ of the solvated protein by a finite difference Poisson–Boltzmann (FDPB) method [27-29], and then use $\Delta E_{elec} = \Sigma_i \Phi_i q_i$, where q_i is the charge of the ith probe atom. The desolvation term, ΔG_{des}, is obtained by the Atomic Contact Potential (ACP) model [30], an atomic level extension of the Miyazawa-Jernigan potential [31]. The Atomic Contact Potential describes local interactions by the sum $\Sigma_i \Sigma_j e_{ij}$, where e_{ij} denotes the atomic contact energy of interacting atoms i and j, and the sum is taken over all atom pairs that are less than 6 Å apart [30]. According to the quasi-chemical approximation [31], e_{ij} is the effective free energy change when a solute-solute bond between two atoms of type i and j is replaced by a solute-solvent bond.

Step 2: Minimization and re-scoring. Step 1 produces a large number of protein-ligand complexes at various local minima of ΔG_s. The free energy of each complex is refined, allowing the ligand to be fully flexible, by the local minimization of the free energy potential

$$\Delta G = \Delta E_{elec} + \Delta E_{vdw} + \Delta G^*_{des} \quad (2)$$

where ΔE_{vdw} denotes the receptor-ligand van der Waals energy, and the superscript in ΔG^*_{des} emphasizes that the desolvation term includes the change in the solute-solvent van der Waals interaction energy. The sum $\Delta E_{elec} + \Delta G^*_{des}$ is obtained by the Analytic Continuum Electrostatic (ACE) model [32], as implemented in version 27 of Charmm [33], using the parameter set from version 19 of the program. The model includes a surface area dependent term to account for the solute-solvent van der Waals interactions. The minimization is performed using an adopted basis Newton-Raphson method as implemented in Charmm [33].

Step 3: Clustering and ranking. Since we seek broad free energy minima rather than narrow, isolated ones, the minimized probe conformations from Step 2 are grouped into clusters based on Cartesian coordinate information [20]. Very small clusters are excluded

from consideration. The threshold is defined by the average clusters size $t = m/n$ if $t < 20$, where m is the total number of probes and n is the number of clusters. Otherwise $t = 20$, i.e., clusters with more than 20 elements are always retained. For each retained cluster, we calculate the probability $p_i = Q_i/Q$, where the partition function Q is the sum of the Boltzmann factors over all conformations, $Q = \Sigma_j \exp(-\Delta G_j/RT)$, and Q_i is obtained by summing the Boltzmann factors over the conformations in the ith cluster only. The clusters are ranked on the basis of their average free energies $<\Delta G>_i = \Sigma_j p_{ij} \Delta G_j$, where $p_{ij} = \exp(-\Delta G_j/RT)/Q_i$, and the sum is taken over the members of the ith cluster.

Step 4: Subcluster analysis. For each ligand, the cluster with the minimum average free energy, $<\Delta G>$, is further divided into subclusters based on probe orientations and free energies. The latter are included, because similar conformations with very different free energies usually have different mechanisms of binding (e.g., different hydrogen bonding interactions), and hence it is preferable to group them into different subclusters [20]. The subclusters of the ith cluster are ranked on the basis of the probabilities $p_{ij} = Q_{ij}/Q_i$, where Q_i is the sum of the Boltzmann factors over all conformations of the ith cluster, and Q_{ij} is obtained by summing the Boltzmann factors over the conformations in the jth subcluster only.

Step 5: Determination of consensus sites. The CS-Map algorithm is primarily used to find 'consensus' sites at which many different probe molecules cluster [18]. The advantage of focusing on the consensus site is that some false positives can be tolerated. For example, if the probability of obtaining a false positive in a particular binding pocket is as high as 20%, but the false positives for the different probes are independently distributed over the protein surface, then the probability of obtaining a false consensus site using six probes is less than 0.01%. In reality the situation is less favorable since the false positives tend to be in relatively large pockets and hence are not independent. Nevertheless, as we will show, mapping with six probes usually gives very good results.

In order to find the consensus sites we select the minimum free energy conformation in each of the five lowest average free energy clusters for each solvent. The structures are superimposed, and the position at which most probes of different types overlap is defined as the main consensus site. An additional clustering of probes close to the main consensus site is likely to indicate another subsite of the active site, and hence the probes in the second cluster are added to those already in the consensus site.

Reaction path calculations

The reaction paths between two states (subclusters) of the small organic solvent molecules were calculated using the free energy potential defined by Eq. 2, i.e., based on the Analytic Continuum Electrostatic (ACE) model [30]. The Trajectory Refinement Algorithm with the conjugate Peak Refinement (CPR) method of Fischer and Karplus [34] was used to determine the saddle points along the reactions path. We have specified 50 as the maximum number of path points, and refined the path 6 times by CPR in 1000 overall refinement cycles. The gradient criteria for accepting the saddle point was 0.05 kcal/mol/Å. The free energy change and the barrier heights for the forward and the reverse processes were calculated. If several barriers were found along a path, only the highest was considered.

Molecular dynamics simulation in explicit water

Molecular dynamics simulations were performed by GROMACS [35, 36] using a modified GROMOS87 force field. The protein with an organic solvent in its active site was placed in a periodic box filled with SPC/E water, using 6592–6595 and 11259–11262 water molecules for HEWL and TLN, respectively. The small molecules were placed at the position they occupy in the first subcluster of the first cluster, as determined by the mapping. After 100 steps of steepest descent minimization, water molecules were substituted by the necessary number of counterions, 8 Cl^- for HEWL and 12 Na^+ for TLN, to neutralize the system in the box. The most positive and most negative potential energy positions, respectively, were used to place the Cl^- and Na^+ ions. After repeating the 100 step steepest descent minimization with the protein position restrained, the system consisting of the protein and the small organic molecule was solvated by a 20 ps position restrained simulation using the NVT ensemble. We have used the PME (particle Mesh Ewald) method with a 9 Å cutoff for the long range electrostatics, and also 9 Å cutoff for the van der Waals interactions. The weak coupling method was used for the pressure and temperature with 0.5 ps and 0.1 ps relaxation times, respectively. The step size

was 2 fs. To relax the protein, LINCS [37] was applied with a 10^{-4} kJ mol^{-1} convergence criterion. The productive dynamics run was 1–2 ns with the NPT ensemble. Based on the trajectories we have determined the formation and breaking of hydrogen bonds between the ligand and either protein polar groups or water molecules. The movement of the probe in the pocket was assessed by calculating its RMSD from the initial state, and by monitoring the distances between probe atoms and certain atoms on the protein.

Results

Solvent mapping of proteins binding sites

Hen egg-white lysozyme

Liepinsh and Otting [14] examined the binding for methanol, methylene chloride, acetonitrile, acetone, isopropanol, t-butanol, urea, and dimethyl-sulfoxide on hen egg-white lysozyme (HEWL) in aqueous solution by NMR techniques. It was found that all the eight molecules bind in site C of the protein active site, although methylene chloride is not as deep inside the pocket as the other ligands. On the basis of the intermolecular NOE's, the small organic molecules interact with the amide proton of N59, and protons on the side chains of W63, I98, A107, and W108. Methanol and methylene chloride also show NOEs with protons located at a second site in the interior of HEWL. Some weak NOEs were detected for acetone and isopropanol with additional atoms close to site C (residues W62, V109, and A110). No other interactions are seen between any of the probes and the protein.

We have used the same eight solvents as probes to map the entire surface of HEWL. Table 1 shows the clusters with the lowest Boltzmann average free energies, including size, probability, average free energy, average electrostatic energy, van der Waals and desolvation free energy contributions, the number of the subclusters in each cluster, the distance between the hub of the cluster and any residue in site C (N59, W62, W63 and A107), and the residue closest to the cluster (not necessarily in site C). The lowest free energy clusters were found to be in site C of HEWL for all solvents, although DMSO and methylene chloride are not as deeply inside the pocket as the other six ligands. As discussed in [20], there is fairly good agreement between the experimentally observed NOEs and the calculated intermolecular proton-proton distances. The mapping finds site C as the consensus site, binding all the eight solvents. As mentioned, the NOEs indicate other locations that bind only certain ligands, e.g., an internal site that binds methanol and methylene chloride, but these positions are not found by the mapping.

The minimum average free energy cluster for each ligand is further divided into subclusters as described in the Methods. Figures 1A through 1H show the highest probability subclusters for each solvent. According to these results, site C accommodates most molecules in a number of rotational states. The hydrophobic part of the ligand is in a pocket formed by I98 and the non-polar portions of the W62, W63, and W108 side chains. This pocket is surrounded by six polar groups: N59 NH, Q57 O, A107 O, W62 $N^{\varepsilon 1}$, W63 $N^{\varepsilon 1}$, and W108 $N^{\varepsilon 1}$. All ligands (with the exception of methylene chloride) bind in a number of conformations, in many cases forming a hydrogen bond with one of these groups, and the polar parts of the ligands point toward various polar patches on the protein even in cases where no explicit hydrogen bonds are formed. As we will further discuss, the existence of multiple bound states, resulting in higher entropy, may be important for explaining why all ligands bind in site C of the active site.

Thermolysin

The X-ray structures of thermolysin [12, 13], determined in isopropanol (IPA), acetone (ACN), acetonitrile (CCN), and phenol (IPH), show that the only location that binds all fours organic molecules is the main substrate specificity pocket S_1', and the binding is detected even at low solvent concentrations. As the solvent concentration is increased, isopropanol, acetone, and phenol start to bind at additional sites, but many of these are at crystal contacts or interacting with another bound ligand, and can never accommodate all the four solvents. The S_1' pocket is formed by a number of hydrophobic side chains (F130, L133, V139, F114 and L202), and is surrounded by polar groups, primarily the side chains of E143, R203, and H231. The binding of substrates, products, and inhibitors to thermolysin always involves this pocket, with longer ligands extending toward the subsites S_1 and S_2. In each complex, at least four hydrogen bonds are formed, mainly with the side chains of N112, E143, Y157 and R203, and with the polar backbone atoms of N111, A113 and W115. The S_1 site is very close to S_1', but some dominant interactions are different because of the different side chains. Hydrogen bonds are formed mainly with Y157 and E166. Additional polar

Table 1. The lowest average free energy clusters of the eight organic ligands bound to Hen egg-white lysozyme (HEWL)[a]

Ligand	T/K	Size	p	$<\Delta G>$	$<\Delta E_{elec}>$	$<\Delta E_{vdw}>$	$<\Delta G^*_d>$	Sub-cluster	D/Å
MET	309.16	97	0.74	−7.18	−0.87	−8.26	1.95	5	2.3
IPA	288.16	51	0.92	−11.60	−1.04	−12.82	2.27	8	2.8
BUT	309.16	76	0.48	−17.53	0.45	−16.85	−1.13	7	2.0
DMS	293.16	100	0.51	−14.05	−0.72	−12.82	−0.51	7	2.7
ACN	309.16	139	1.00	−13.14	−1.00	−13.56	1.42	6	2.2
CCN	309.16	26	0.28	−10.18	−0.82	−8.19	−1.17	4	2.4
URE	309.16	40	0.91	−12.10	−0.79	−15.38	4.07	7	2.2
MCL	288.16	63	0.28	−3.70	−3.01	−4.27	3.57	3	3.5

[a]The table contains the ligand name, the cluster size, the probability of the cluster, the Boltzmann average binding free energy, electrostatic, van der Waals and desolvation contributions to the average binding free energy, the number of the subclusters, and the distance of the cluster center from the closest residue in site C. MET – methanol, IPA – isopropanol, BUT – t-butanol, DMS - dimethyl-sulfoxide, CAN – acetone, CCN – acetonitrile, URE – urea, MCL – methylene chloride.

side chains that extend toward the S_1 and S_2 pockets are F114, H142, E143, H146 and H231.

In contrast to HEWL, whose surface is dominated by a single large pocket, thermolysin exhibits a substantial number of crevices, so its mapping is more difficult. We have performed mapping calculations using isopropanol, acetone, acetonitrile, and phenol as probes, and compared the results to the X-ray structures determined by English et al. [12, 13] in the same organic solvents. Table 2 show the mapping results for isopropanol and acetone, including the cluster size, probability, average free energy, and nearest protein atom. The three lowest free energy clusters for isopropanol are in the subsites S_1, S'_1, and S_2, respectively, of the active site. However, the lowest Boltzmann average free energy cluster is in the subsite S_1, around isopropanol IPA8 in the x-ray structure, in spite of the fact that all bound solvents first appear in the S'_1 pocket close to IPA1, and the x-ray structures show binding in the S_1 site only at much higher isopropanol concentrations [12]. CS-Map places the lowest average free energy cluster of acetone in a relatively large pocket close to R260, 12.8 Å from the closest acetone position in the experimental structure. In the x-ray structure this pocket contains three crystallographic waters. While there is a possibility that the extra electron density in the site is actually due to a very mobile acetone, the finding emphasizes that the CS-Map algorithm does not necessarily removes all false positives. Indeed, as we will discuss, the method does not account for the potential contribution of configurational entropy that may render the active site more favorable than other positions. However, the only concensus site, i.e., the location where all four solvent molecules bind, is found in the S'_1 pocket of the active site, in good agreement with the x-ray data [12, 13].

Protein tyrosine phosphatase-1B (PTP1B)

Protein tyrosine phosphatases, working in concert with protein tyrosine kinases, regulate a vast array of cellular events by catalyzing the removal of the phosphoryl group from the phosphotyrosine (pTyr) residues in protein substrates [38]. PTP1B hydrolyzes phosphotyrosines on the insulin receptor, deactivating it. Overproduction of this enzyme has been implicated in the onset of type II diabetes, and it is therefore an intensely studied target for drug discovery [39, 40]. While low nanomolar binders, found by high throughput virtual screening, have been reported in the literature, the race to find a suitable drug candidate is still open.

Our goal here is to analyze the information, provided by the mapping, on the binding site of PPT1B, and hence the search was restricted to the corresponding region of the surface. The calculations were performed on a PTP1B structure that has been co-crystallized with two phosphotyrosine molecules (PDB code 1pty). One of the pTyr residues occupies a deep pocket, with its phosphate group deeply buried in the protein, the other, lower affinity site is much closer to the surface. While the pTyr residues have been removed before the mapping, they will be shown in our figures in order to indicate the binding site.

Table 2. The lowest average free energy clusters for isopropanol (IPA) and acetone (ACN) bound to thermolysin (TLN)[a]

Ligand	Cluster[b]	Size	p	$<\Delta G>$	$<\Delta E_{elec}>$	$<\Delta E_{vdw}>$	$<\Delta G^*_d>$	Sub-cluster	D/Å
IPA	1	129	0.40	−10.71	−2.74	−12.44	4.48	12	0.6/8
	2	30	0.06	−10.49	−2.32	−12.81	4.64	11	0.7/1
	3	132	0.26	−10.31	−2.03	−12.23	3.96	14	1.0/5
ACN	1	69	0.77	−13.10	−1.37	−14.15	2.41	3	12.8/2
	2	104	0.15	−11.89	−5.59	−13.23	6.92	6	3.2/1
	3	251	0.06	−10.96	−1.92	−12.26	3.22	9	5.4/1
	4	30	0.01	−10.75	−0.16	−13.26	2.67	4	0.6/1

[a]The table contains the ligand name, the cluster size, the probability of the cluster, the Boltzmann average binding free energy, electrostatic, van der Waals and desolvation contributions to the average binding free energy, the number of the subclusters, and the distance of the cluster center from the center of the ligand molecule in the x-ray structure. IPA – isopropanol, ACN – acetone. S'_1, S_1, and S_2 are subsites of the thermolysin active site.

[b]For IPA, clusters 1, 2, and 3 are in pockets is in S_1, S'_1, and S_2, respectively. For ACN, cluster 1 is in a pocket far from the binding site, whereas clusters 2, 3, and 4 are in the pockets S_1, S_2, and S'_1, respectively.

Table 3. The lowest average free energy clusters of the eight organic ligands bound to protein tyrosine phosphatase 1B (PTP1B)[a]

Ligand	Size	p	$<\Delta G>$	$<\Delta E_{elec}>$	$<\Delta E_{vdw}>$	$<\Delta G^*_d>$	Sub	D/Å
ACN	25	1.00	−15.50	−3.95	−16.57	5.02	3	1.4
IPA	30	1.00	−14.98	−2.78	−15.94	3.74	4	1.2
CCN	29	0.95	−12.08	−4.23	−8.97	1.12	4	1.4
URE	29	0.99	−14.20	−3.33	−16.96	6.09	9	1.0
MET	31	0.75	−8.16	−3.05	−8.95	3.85	5	0.9
PHN	26	0.69	−20.55	−1.32	−21.80	2.57	6	0.9

[a]The table contains the ligand name, the cluster size, the probability of the cluster, the Boltzmann average binding free energy, electrostatic, van der Waals and desolvation contributions to the average binding free energy, the number of the subclusters, and the distance of the cluster center from the P atom of the phosphotyrosine molecule in the higher affinity (deeper) binding site of the x-ray structure 1pty. ACN – acetone, IPA – isopropanol, CCN – acetonitrile, URE – urea, MET – methanol, PHN – phenol.

We have used isopropanol (IPA), acetone (ACN), acetonitrile (CCN), urea (URE), methanol (MET) and phenol (IPH) as probes. Figure 3A shows, for each of the six probes, the lowest free energy conformation in the lowest free energy cluster in the PTP1B binding site, superimposed onto the two phosphotyrosine molecules seen in the x-ray structure. As shown, the probes cluster in the main substrate binding pocket S_1, essentially at the location of the phosphate group of the higher affinity phosphotyrosine. The binding site is surrounded by hydrophilic residues R47, S216, R221, and the (partially) hydrophobic residues Y46, P180, A217, and C215. Each of the backbone amino groups in residues 216–221 can form hydrogen bonds with the probes. In fact, the phosphate group of pTyr in the high affinity pocket has hydrogen bonds with the backbone of the residues S216, A217, G218, and G220. Since the most favorable position occurs at this location for each of the probes, any drug candidate should have a strong hydrogen bond acceptor group in this region.

For five of the six solvents (IPA, ACN, CCN, URE, and MET), the binding site shown in Figure 3A is so preferable that all clusters outside this site have substantially higher free energy. In contrast, phenol has a number of clusters with free energy comparable to that of the cluster shown in Figure 3A, and three of these low energy clusters are also in the binding site (Figure 3B). Notice that two clusters (cluster 1 and 3) overlap each other and only the OH directions are different, which means that the two clusters have the same interactions between the aromatic plane but two different hydrogen bonds of the OH group with Q262

and D181 side-chain O atoms, respectively. The other cluster representants form hydrogen bonds with R221 NH, F182 NH, Q265 NH_2, C214 S, A217 NH, and S216 OH. The relative positions of Y46 and the phenol in its several low free energy clusters indicate $\pi-\pi$ interactions. The isopropanol lowest free energy cluster forms three hydrogen bonds with C214 S and the side chain N atoms of R221. Methanol forms hydrogen bonds both as donor and acceptor. The main interacting residues are similar to that of phenol, i.e., C214, S216, and R221. The lowest free energy clusters of acetone, urea and acetonitril were found in interactions with the previously mentioned side chains. Urea has more interactions because of the four donor and three acceptor possibilities.

It is interesting to compare the mapping results with those from high throughput docking [40]. The latter reveals that the active compounds found have strong negative charges, most frequently sulphate or carboxylic groups, which interact with the backbone and side-chains of G220, S216, A217, G218, and R221. According to the mapping, in their lowest free energy clusters all probes interact with the same residues. $\pi-\pi$ interactions are also important, because the main interactions of the aromatic rings are with F182 and Y46. As a matter of fact, many of the best binders are multi-ring structures, and some of the rings overlap with the low free energy phenol clusters shown in Figure 3b. Thus, the mapping predicts reasonable binding positions for the selected probes in the pocket, suggesting that the interactions of functional groups can be considered additive. While results of this type is expected to help the find higher affinity ligands, it is clear that the set of probes needs to be extended to include the major functional groups that occur in many drug molecules. This work is underway in our laboratory.

Studying the multiplicity of bound states

Hen egg-white lysozyme

We recall that the hubs of the subclusters, found by the mapping, are local minima of the free energy potential given by Eq. (2). Reaction path calculations were carried out using the same potential in order to study the transitions among these minima. As already mentioned, in good agreement with experimental data, the first clusters for methylene chloride and acetonitrile are not so deep in the binding site as those for the other molecules. In addition, the lowest free energy cluster is dominated by a single subcluster both for acetonitrile (Figure 1F) and methylene chloride (Figure 1H), and hence in our reaction path calculation we focus on the other six probes.

Table 4. Calculated free energy barriers between the subclusters of the lowest free energy methanol clusters in HEWL

Transition	ΔG_1 (kcal/mol)[a]	ΔG_2 (kcal/mol)[b]	$\Delta\Delta G$ (kcal/mol)[c]
$1 \rightarrow 2$	–	–	0.38
$1 \rightarrow 3$	0.78	0.54	0.24
$1 \rightarrow 4$	1.26	0.16	1.10
$1 \rightarrow 5$	0.79	0.14	0.65
$2 \rightarrow 3$	0.40	0.54	−0.14
$2 \rightarrow 4$	0.89	0.16	0.73
$2 \rightarrow 5$	0.89	0.62	0.27
$3 \rightarrow 4$	–	–	0.86
$3 \rightarrow 5$	0.55	0.15	0.40
$4 \rightarrow 5$	–	–	−0.46

[a]Forward transition.
[b]Backward transition.
[c]Free energy difference.

The calculations confirmed that all subclusters are local minima and hence can be considered as individual states. Some of these states are separated by high free energy barriers (10–20 kcal/mol or more), and hence no transitions are possible. However, most free energy barriers are much lower, around 1 to 2 kcal/mol. Table 4 shows the free energy barriers of the forward and the backward transitions between the subclusters of the lowest free energy methanol cluster in HEWL. The free energy difference between the two end states along each path is also indicated. As an example, Figure 4A shows the free energy change along the reaction path from subcluster 1 to subcluster 3. More generally, the highest barrier for any transition among methanol subcluster was found to be only 1.3 kcal/mol. Although the typical free energy barriers are somewhat higher for the other five molecules, and not all minima are connected with low free energy reaction paths, the bound ligands clearly retain substantial degrees of rotational/translational freedom in the active site. However, we emphasize that the reaction path calculations neglect the effect of individual water molecules in the pocket, and hence the low free energy barriers do not necessarily imply that actual transitions occur on a fast time scale.

In order to study the kinetics of transitions, we have performed 2 ns molecular dynamics simulations of methanol bound to HEWL using a periodic box

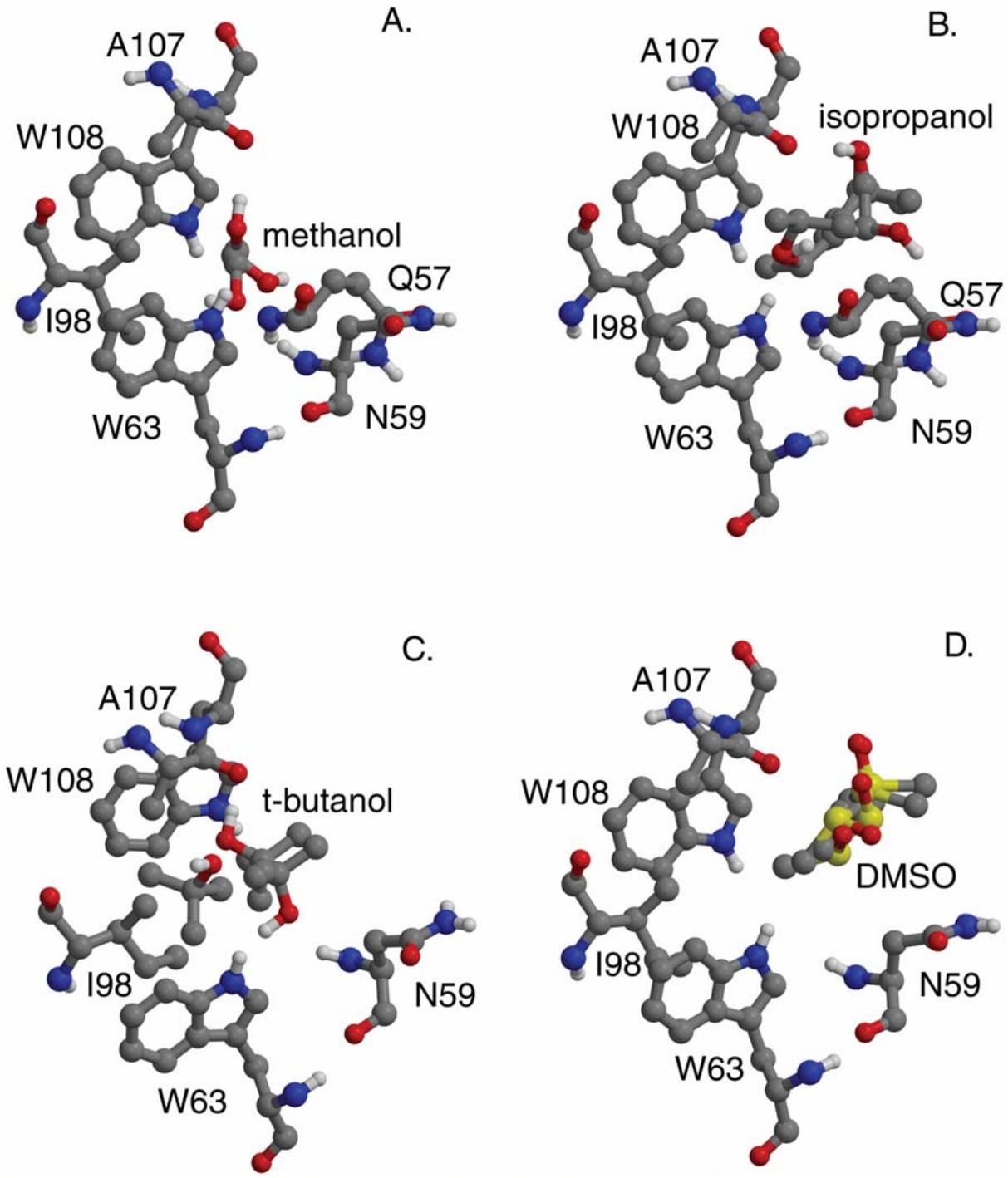

Figure 1. Conformations representing the high probability subclusters in the lowest free energy cluster of eight organic solvents, obtained by the mapping of hen egg-white lysozyme (HEWL). A. methanol, B. Isopopanol, C. t-butanol, D. DMSO, E. acetone, F. acetonitril, G. urea, H. methylene chloride.

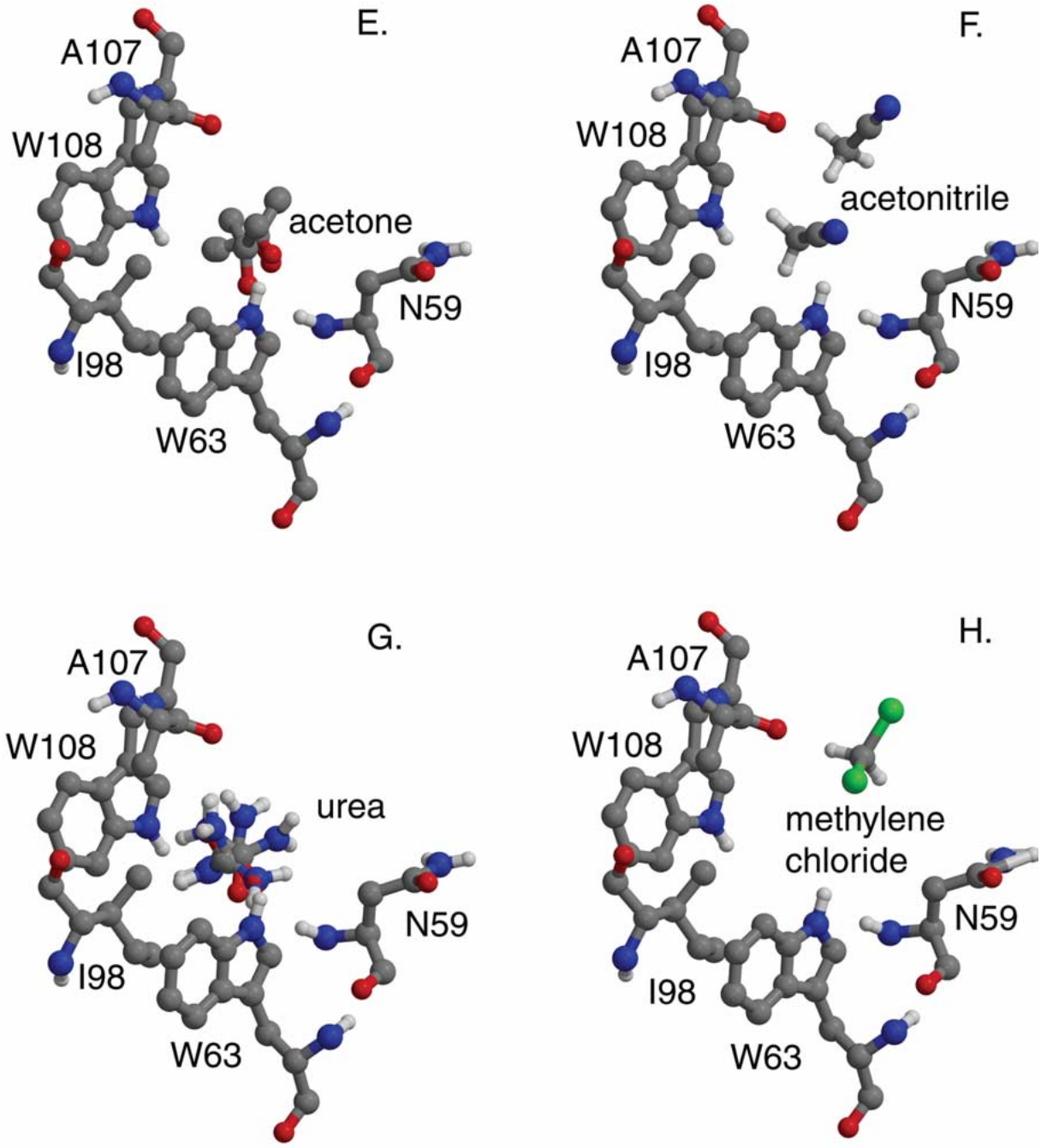

Figure 1. (Continued)

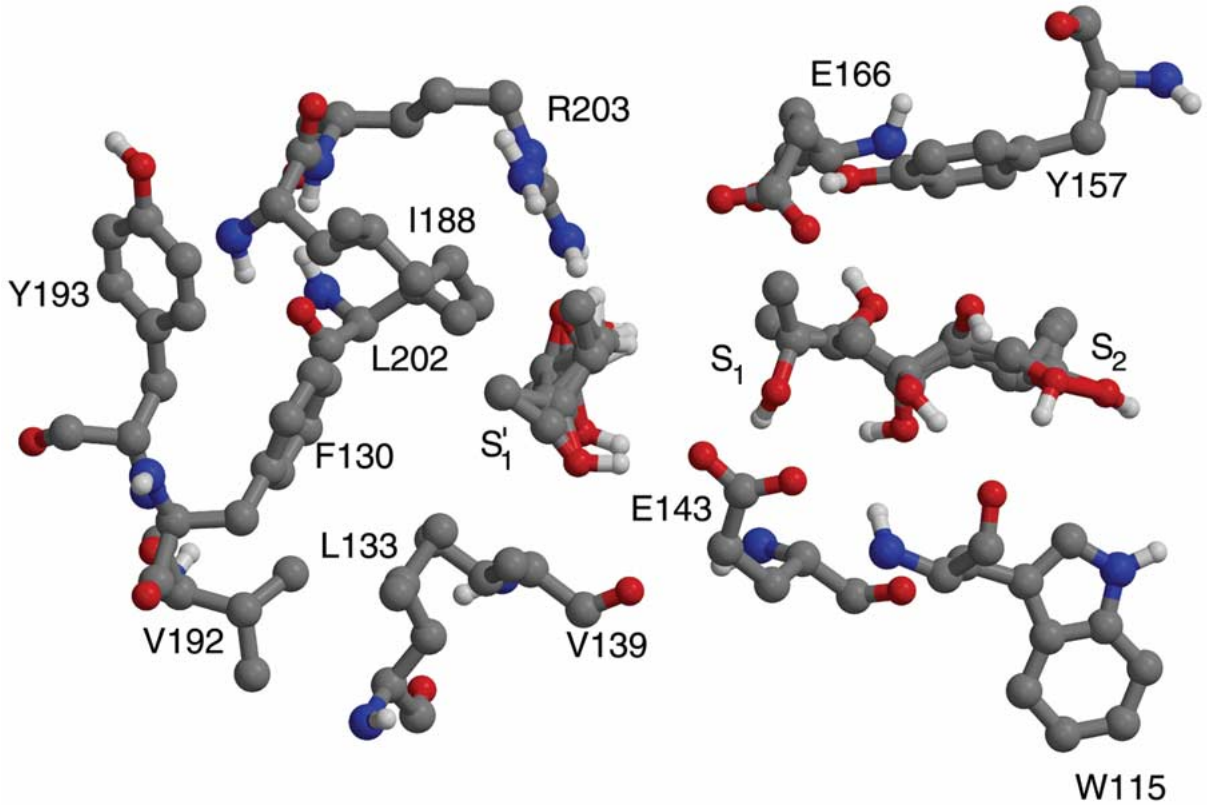

Figure 2. Conformations representing the high probability subclusters in the three lowest free energy clusters of isopropanol, obtained by the mapping of thermolysin. S_1', S_1, and S_2 indicate the major subsites of the thermolysine active site.

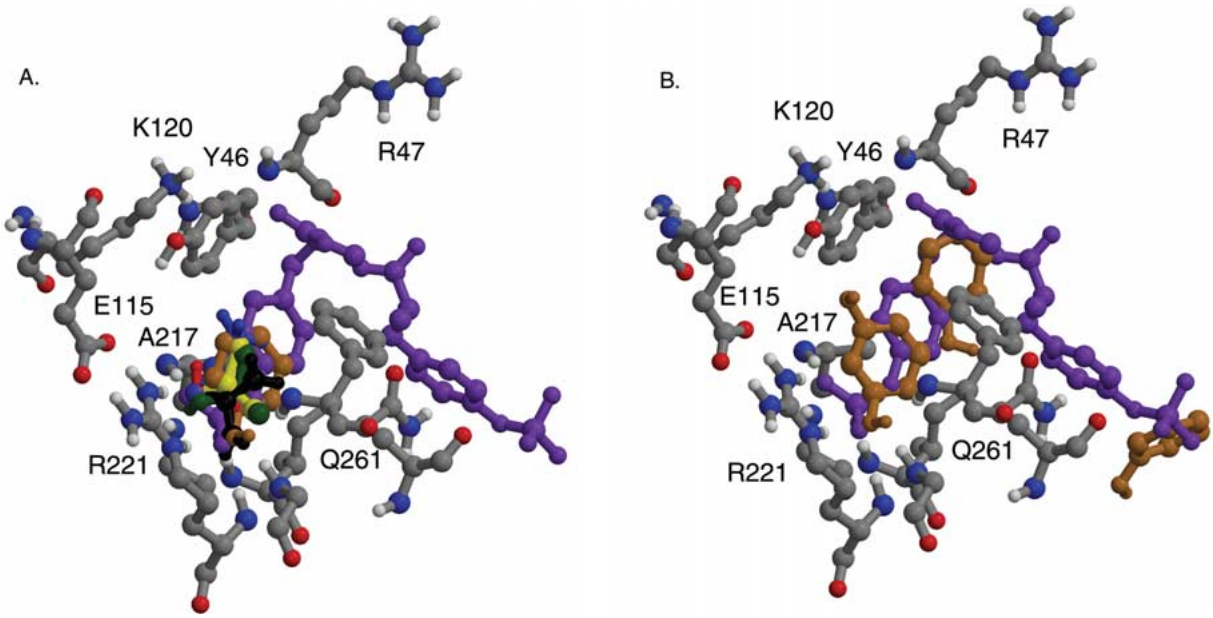

Figure 3. Conformations representing the lowest free energy clusters of six organic solvents, obtained by mapping the binding site of Protein Tyrosine Phosphatase 1B (PTP1B). A. Lowest free energy clusters of methanol (red), isopopanol (dark green), acetone (yellow), acetonitrile (blue), phenol (ochre), and urea (black). The two phosphotyrosine molecules in the x-ray structure are also shown (purple). B. The four lowest free energy clusters of phenol.

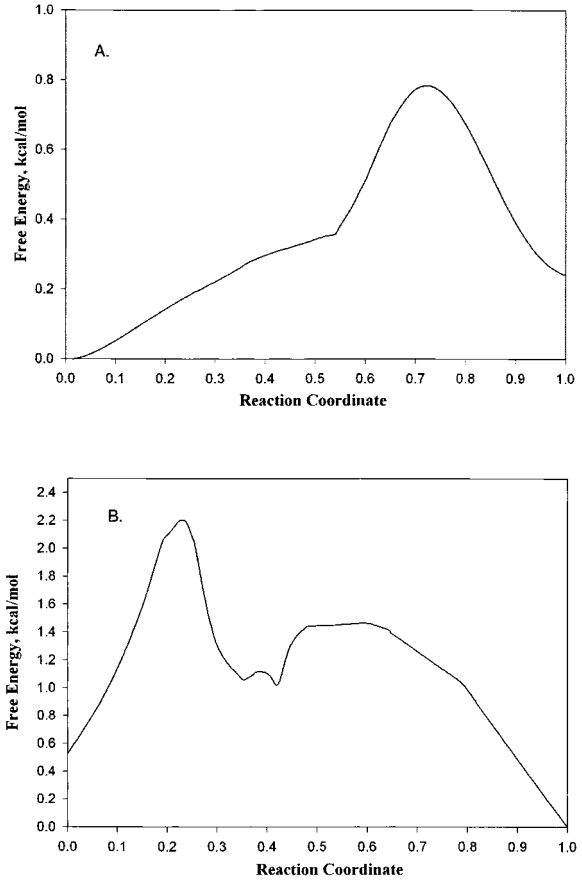

Figure 4. Reaction path calculation. A. Free energy change along the transition path between first and the third lowest free energy subclusters of methanol bound to HEWL. The free energies are shown using the first subcluster as the reference state. B. A transition path between the minimum free energy isopropanol conformations in the subsites S_1 and S_2 of the thermolysine binding site (see Figure 2). The free energies are shown using the minimum in S_2 as the reference state.

filled with 6593 SPC/E water molecules. During the simulation at 309.16 K the methanol remained in the pocket. On the average, there were 4.9 water molecules present in a 5 Å shell around the methanol. H-bonds were observed between the methanol OH as donor, and the atoms A107 O, 52D O^{δ} and N57 O as acceptor, as well as the same OH as acceptor, and N59 NH, W62 $N^{\epsilon}H$, and W63 $N^{\epsilon}H$. These interactions changed during the simulation as methanol rotated and translated in the pocket. We have also found a number of indirect hydrogen bonded water bridges between methanol OH and protein polar atoms, most frequently W62 $N^{\epsilon}H$, N57 O, and D52 O^{δ}. Based on the snapshots we have extracted, these three water bridges are present in 22%, 14% and 20%, respectively, of the total simulation time.

Dimethyl sulfoxide (DMSO) has only one oxygen atom for the formation of hydrogen bonds with the polar groups in the binding pocket, and such bonds are formed with N59 NH and W63 NH. There are almost the same number of water molecule in a 5 Å shell around DMSO as around methanol (4.6 vs. 4.9), but much fewer bridges involving a water. The motion of DMSO is generally slower than that of the methanol, resulting in less frequent transitions. The free energy barriers separating some minima are very high, i.e., the corresponding transitions do not occur. Urea can participate in hydrogen bonds both as an acceptor (CO) and a donor (NH_2 groups). During the 1.5 ns simulation the C–O vector swept only some 60 degree angle around the C atom with a flip-flop of the NH_2–C–NH_2 group. The C atom has been translated only very slightly in the site, virtually retaining its position. Generally, a number of hydrogen bonds were formed between the ligand and the protein, most frequently one to three, and up to five in some instances. The groups N59 NH and W63 $N^{\epsilon}H$ served most frequently as donors, whereas L56 O, N57 O, A107 O, W108 N, D52 O2, I58 N and W63 N^{ϵ} were the most frequent acceptors. Since urea could serve both as an acceptor and a donor, a lot of water insertions were observed, where mainly N57 and W63 played an important role, as residues in the pocket.

Thermolysin

Reaction path calculations for isopropanol were performed between the subclusters of each of the first three clusters in pockets S_1', S_1 and S_2, respectively, and between the subclusters of the clusters in S_1' and S_1, as well as in S_1 and S_2. Results confirm that transitions between the subclusters within a single pocket occur with high probability, since many of the free energy barriers are as low as 1 kcal/mol. Between the states in S_1 and those in S_2 we also see some low barriers, but the majority of activation energies are in the range of 2–3 kcal/mol. For example, the free energy profile along the reaction path from subcluster 3 in the S_2 pocket and subcluster 1 in S_1 reveals a barrier of 2.2 kcal/mol, although for the inverse transition it is only 1.8 kcal/mol (Figure 4). The lowest free energy barrier between the states in S_1 and those in S_1' was less than 2 kcal/mol, but there were also some much higher barriers. The relatively low barriers between the major subsites of the thermolysin active site imply that the ligands can migrate among them. However,

these transitions are likely to be affected by the water molecules, and hence they are probably not as fast as the low barriers would imply. Molecular dynamics simulations of small ligands in the active site of TLN show that water insertion plays important role in the movement of the ligands. For example, isopropanol moved from the initial S_1' position to S_1 position after about 1 ns (Kortvelyesi et al., in preparation).

Conclusions

Computational mapping methods place molecular probes – small molecules or functional groups – on a protein surface in order to identify the most favorable binding positions by calculating an interaction potential. Mapping is an important step in a number of flexible docking and drug design algorithms. In spite of substantial efforts, some of the frequently used mapping techniques such as GRID and MCSS are not selective enough to correctly reproduce the available NMR and x-ray data on the binding of organic solvents to proteins [9, 13]. While the calculations generally result in hundreds of energy minima, the experiments show only a few binding sites.

In this paper we describe an improved mapping algorithm, and show that it reproduces the binding of eight organic solvents to lysozyme as observed by NMR, and the crystal structures of thermolysin, determined in four different solvents. Both experimental and computational mapping confirm that the small ligands cluster in the functional sites of a protein, forming a consensus site, and hence mapping can be used for the identification and analysis of such sites. The origin of clustering is not fully understood. Since functional sites bind each small polar molecule in a number of rotational states, it is possible that the relatively high entropy makes binding at this position more favorable that binding at alternative sites. The distribution of ligands among their orientational states was studied by reaction pathway calculations, and by nanosecond-scale molecular dynamics simulations. Result imply that, as expected from the existence of several subclusters found by the mapping, the bound ligands reserve substantial degrees of rotational/translational freedom.

Acknowledgements

This research has been supported by grants DBI-0213832 from the National Science Foundation, GM064700 from the National Institute of Health, and P42 ES07381 from the National Institute of Environmental Health Sciences.

References

1. Ewing, T.J.A and Kuntz, I.D., J. Comp. Chem., 18 (1997) 1175–1189.
2. Goodford P.J., J. Med. Chem., 28 (1985) 849–875.
3. Lawrence M.C. and Davis P.C., Proteins 12 (1992) 31–41
4. Bohm H.J., J. Comp.-Aid. Mol. Des., 6 (1992) 131–147.
5. Eisen, M.B., Wiley, D.C., Karplus, M. and Hubbard, R.E., Proteins, 19 (1994) 199–221.
6. Miranker, A. and Karplus, M., Proteins: Struc. Func. Gen., 11 (1991) 29–34.
7. Caflisch, A., Miranker, A. and Karplus, M., J. Med. Chem., 36 (1993) 2142–2167.
8. Evensen, E., Joseph-McCarthy, D. and Karplus, M. MCSS version 2.1, Harvard University, Cambridge, MA, USA, 1997.
9. Mattos, C. and Ringe, D. Nat. Biotech., 14 (1996) 595–599.
10. Allen, K.N., Bellamacina, C.R., Ding, X., Jeffery, C.J., Mattos, C., Petsko, G.A. and Ringe, D., J. Phys. Chem., 100 (1996) 2605–2611.
11. Mattos, C. and Ringe, D., Curr. Opin. Struct. Biol., 11 (2001) 761–764.
12. English, A.C., Done, S.H., Caves, L.S., Groom C.R. and Hubbard, R.E., Proteins, 37 (1999) 628–640.
13. English, A.C., Groom C.R. and Hubbard, R.E., Protein Eng., 14 (2001) 47–59.
14. Liepinsh, E. and Otting, G., Nat. Biotech., 15 (1997) 264–268.
15. Shuker, S.B., Hajduk, P.J., Meadows, R.P. and Fesik, S.W., Science, 274 (1996) 1531–1534.
16. Pellechchia, M., Sem, D.S. and Wutrich, K., Nature Reviews, 1 (2002) 211–219.
17. Knubovets, T., Osterhout, J.J. and Klibanov, A.M., Biotechnol. Bioeng., 63 (1999) 242–248.
18. Erlanson, D.A., Braisted, A.C., Raphael, D.L., Randal, M., Stroud, R.M., Gordon, E.M. and Wells, J.A., Proc. Natl. Acad. Sci. USA, 97 (2000) 9367–9372.
19. Moore, J.M., Curr. Opin. Biotech., 10 (1999) 54–58.
20. Dennis, S., Kortvelyesi, T. and Vajda, S., Proc. Natl. Acad. Sci. USA, 99 (2002) 4290–4295.
21. Kortvelyesi, T., Dennis, S., Silberstein, M., Brown, L. III and Vajda, S., Proteins, 51 (2003) 340–351.
22. Dennis, S., Camacho, C.J. and Vajda, S., In: Optimization in Computational Chemistry and Molecular Biology, Floudas, C. A., Pardalos, C., Eds. Kluwer Academic, Norwell, MA, USA; 2000, 243 p.
23. Vajda, S., Weng, Z., Rosenfeld, R. and DeLisi, C., Biochemistry, 33 (1994) 13977–13987.
24. Weng, Z., Vajda, S. and DeLisi, C. Protein Sci., 5 (1996) 614–626.
25. Jackson, R.M. and Sternberg, M.J.E., J. Mol. Biol., 250 (1995) 258–275.
26. Krystek, S., Stouch, T. and Novotny, J., J. Mol. Biol., 234 (1993) 661–179.
27. Gilson, M.K. and Honig, B., Proteins, 4 (1988) 7–18.
28. Honig, B. and Nicholls, A., Science, 268 (1995) 1144–1149.
29. Bruccoleri, R.E., J. Comp. Chem., 14 (1993) 1417–1422.
30. Zhang, C., Vasmatzis, G., Cornette, J.L. and DeLisi, C., J. Mol. Biol., 267 (1996) 707–726.

31. Miyazawa, S. and Jernigan, R., Macromolecules, 18 (1985) 534–552.
32. Schaefer, M. and Karplus, M., J. Phys. Chem., 100 (1996) 1578–1599.
33. Brooks, B.R., Bruccoleri, R.E., Olafson, B., States, D.J., Swaminathan, S. and Karplus, M., J. Comp. Chem., 4 (1983) 197–214.
34. Fischer, S. and Karplus, M., Chem. Phys. Letters, 194 (1992) 252–262.
35. Berendsen, H.J.C., van der Spoel, D. and van Drunen, R., Comp. Phys. Comm., 91 (1995) 43–56.
36. Lindahl, E., Hess, B. and van der Spoel, D., J. Mol. Mod., 7 (2001) 306–317.
37. Hess, B., Bekker, H., Berendsen, H.J.C. and Fraaije, J.D.E.M., J. Comp. Chem., 18 (1997) 1463–1472.
38. Burke, T.R. and Zhang, Jr. Z.-Y., Biopolymers, 47 (1998) 225–241.
39. Sarmiento, M., Wu, L., Keng, Y.F., Song, L., Luo, Z., Huang, Z., Wu, G.Z., Yuan, A.K. and Zhang, Z.Y., J. Med. Chem., 43 (2000) 146–155.
40. Doman, T.N., McGovern, S.L., Witherbee, B.J., Kasten, T.P., Kurumbail, R., Stallings, W.C., Connolly, D.T. and Shoichet, B.K., J. Med. Chem., 45 (2002) 2213–2221.
41. Brady Jr., G.P. and Stouten, P.F.W., J. Comp. Aided Mol. Des., 14 (2000) 383–401