PROTEINS: Structure, Function, and Genetics 51:340–351 (2003)

Algorithms for Computational Solvent Mapping of Proteins

Tamas Kortvelyesi,[1,2] Sheldon Dennis,[1] Michael Silberstein,[3] Lawrence Brown III,[1] and Sandor Vajda[1]*

[1]*Department of Biomedical Engineering, Boston University, Boston, Massachusetts*
[2]*Department of Physical Chemistry, University of Szeged, Szeged, Hungary*
[3]*Program in Bioinformatics, Boston University, Boston, Massachusetts*

ABSTRACT **Computational mapping methods place molecular probes (small molecules or functional groups) on a protein surface to identify the most favorable binding positions by calculating an interaction potential. We have developed a novel computational mapping program called CS-Map (computational solvent mapping of proteins), which differs from earlier mapping methods in three respects: (i) it initially moves the ligands on the protein surface toward regions with favorable electrostatics and desolvation, (ii) the final scoring potential accounts for desolvation, and (iii) the docked ligand positions are clustered, and the clusters are ranked on the basis of their average free energies. To understand the relative importance of these factors, we developed alternative algorithms that use the DOCK and GRAMM programs for the initial search. Because of the availability of experimental solvent mapping data, lysozyme and thermolysin are considered as test proteins. Both DOCK and GRAMM speed up the initial search, and the combined algorithms yield acceptable mapping results. However, the DOCK-based approaches place the consensus site farther from its experimentally determined position than CS-Map, primarily because of the lack of a solvation term in the initial search. The GRAMM-based program also finds the correct consensus site for thermolysin. We conclude that good sampling is the most important requirement for successful mapping, but accounting for desolvation and clustering of ligand positions also help to reduce the number of false positives. Proteins 2003;51:340–351.**

Key words: ligand binding; docking; scoring; enzyme active site; clustering; solvation

INTRODUCTION

Solvent mapping is an important experimental technique for locating and characterizing binding sites on proteins. The multiple solvent crystal structures (MSCS) method, developed by Ringe and coworkers, involves soaking protein crystals in organic solvents, and then determining the structure by X-ray crystallography.[1–5] The resulting structures show that a limited number of solvent molecules replace crystallographic waters and cluster in the active site, delineating the binding pockets.[3] All other bound solvent molecules are either in crystal contact, occur only at high-ligand concentration, or are in small, buried pockets, where only a few types of solvent molecules cluster compared to the active site. NMR also provides a method for detecting the binding of small molecules to a protein in solution.[6,7]

A number of methods have been developed to perform mapping computationally rather then experimentally. Computational mapping methods place molecular probes (small molecules or functional groups) on a protein surface to identify the most favorable binding positions by calculating an interaction potential. Such calculations actually predate the first solvent mapping experiments. Goodford mapped the receptor active site in the drug design program GRID, followed by a fragment assembly stage in which some of the favorable positions found for individual molecular fragments were connected into a single viable molecule.[8] This site mapping and fragment assembly strategy has been implemented in a number of drug design programs.[8–13] An interesting approach to mapping is the MCSS method, which optimizes the free energy of numerous ligand copies simultaneously, each transparent to the others but subject to the full force of the receptor.[14–17]

Although frequently used in drug design algorithms, traditional mapping techniques generally fail to reproduce the available NMR and X-ray data on the binding of organic solvents to proteins. The major problem with approaches exemplified by GRID and MCSS is that they result in too many energy minima on the surface of the protein, and it is difficult to determine which of these minima is actually relevant.[1] In addition, MCSS generally assigns different preferred sites to different groups,[14–17] in contrast to the observation that such regions overlap.[1–6] These problems were recently demonstrated by English et al.,[5] who used both GRID and MCSS to map thermolysin for the binding sites of isopropanol, acetone, acetonitrile, and phenol and compared the results with those of mapping experiments. Although they found local minima close to the experimentally observed binding positions, the closest minima were generally not among those with the

Grant sponsor: National Science Foundation: Grant number: DBI-0213832; Grant sponsor: National Institutes of Health; Grant number: GM61867; Grant sponsor: National Institute of Environmental Health Sciences; Grant number: P42 ES07381; Grant sponsor: National Institute of Health; Grant number: GM64700.

L. Brown's present address is Fish and Neave, New York, NY 10020.

*Correspondence to: Sandor Vajda, Department of Biomedical Engineering, Boston University, Boston, MA 02215. E-mail: vajda@bu.edu

Received 19 June 2002; Accepted 5 August 2002

lowest free energies, resulting in false positives (i.e., configurations with favorable energy, which are not located near any experimentally observed binding site). The authors identify their failure to account for solvation as the main cause of the problem,[6] and we mention in addition that neither GRID nor MCSS uses search strategies beyond local minimization.

In an attempt to overcome the shortcomings of traditional mapping methods, we recently developed a three-step mapping algorithm that will be referred to as CS-Map (computational solvent mapping of proteins).[18] Step 1 of CS-Map is a rigid body search that moves the probes in the desolvation and electrostatic fields of the protein. Atomic overlaps are penalized, but attractive van der Waals interactions are not taken into account, thereby enabling the ligand to travel extensively along the protein surface toward regions with favorable electrostatics and desolvation. In step 2, we switch to a free energy potential that accounts for the van der Waals interactions and uses a more accurate continuum model to calculate the electrostatic and solvation contributions. The free energy of receptor-ligand complexes is minimized, starting from the conformations found in step 1 and allowing for ligand flexibility. In step 3, the docked ligand positions are clustered, and the clusters are ranked on the basis of their average free energies.

The CS-Map algorithm has been applied to two proteins, hen egg-white lysozyme[6] (HEWL) and thermolysin[4,5] (TLN), which have been experimentally mapped by using a number of organic solvents. The computational mapping finds a consensus site for each enzyme (i.e., a position at which the different solvents all bind), and the findings are in good agreement with the results of the mapping experiments. These results and preliminary data for additional enzymes show that the mapping consistently finds a major subsite of the substrate-binding site, and the residues that frequently interact with the probes also contact many of the enzymes' ligands (substrate analogs, products, and inhibitors). Thus, computational solvent mapping can provide reliable and detailed information on the substrate-binding sites of enzymes. Structural genomics approaches are likely to produce structures for an increasing number of poorly characterized proteins, and at this point there are relatively few computational methods for identifying functional sites on the basis of protein structure,[19–22] suggesting the need for high-throughput computational mapping.

In this article, we explore the CS-Map algorithm and a number of potential alternatives for three reasons. First, although GRID and MCSS were shown to have limited use for mapping, it is not clear how docking programs such as DOCK[23] and GRAMM[24] might handle the initial sampling problem. Neither DOCK nor GRAMM was developed with mapping in mind, but both use much more sophisticated search strategies than GRID or MCSS and so should not be dismissed without a thorough analysis. Second, testing the alternative approaches will enable us to determine whether all three steps of the CS-Map algorithm (i.e., initial search for regions with favorable electrostatics and desolvation, flexible refinement, and clustering) are necessary for avoiding spurious local minima. Accordingly, we will use alternative search strategies in step 1, test different scoring functions in step 2, and rank the conformations without clustering and averaging in step 3. Third, both DOCK and GRAMM are orders of magnitude faster than the multistart simplex search we use in the CS-Map program, and hence are potentially useful for the development of a high-throughput mapping method. The tests will involve HEWL and TLN, because of the availability of experimental mapping results for these proteins.

MATERIALS AND METHODS

The CS-Map Algorithm

The three computational steps of the algorithm are as follows[18].

Step 1: Rigid body search for regions with favorable electrostatics and desolvation

A multistart simplex method[25,26] is used to move the probes around the protein, starting from a number of evenly distributed points over the entire protein surface (i.e., 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.[27–30] This assumed van der Waals cancellation causes the free energy function ΔG_s to be relatively smooth, that is, there is only moderate sensitivity to small structural perturbations.

The (direct) electrostatic energy is determined by the expression $\Delta E_{elec} = \Sigma_i \Phi_i q_i$, where q_i is the charge of the ith probe atom, and Φ_i is the electrostatic field of the solvated protein at that point. The electric field Φ is determined by a finite difference Poisson–Boltzmann (FDPB) method[31,32] using the CONGEN program.[33] Dielectric constants $\epsilon = 4$ and $\epsilon = 78$ are used for the protein and the solvent, respectively. We use the template partial charges provided by the Quanta program[34] (Molecular Simulations, Inc) for the probe molecules. The desolvation term, ΔG_{des}, is obtained by the Atomic Contact Potential (ACP) model,[35] an atomic level extension of the Miyazava–Jernigan potential.[36] 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 <6 Å apart.[35] According to the quasichemical approximation,[36] 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 rescoring.

Step 1 produces a large number of protein-ligand complexes at various local minima of ΔG_s. The free energy of each complex is minimized by using the more accurate 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,[37] as implemented in version 27 of Charmm[38] 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 by using an adopted basis Newton–Raphson method as implemented in Charmm.[38] During the minimization, the protein atoms are held fixed, whereas the atoms of the probe molecules are free to move. At most, 1000 minimization steps are allowed, although most complexes require far fewer steps to achieve convergence.

Step 3: Clustering and ranking

The minimized probe conformations from step 2 are grouped into clusters based on Cartesian coordinate information. The method creates an appropriate number of clusters so that the maximum distance between a cluster's hub and any of its members (the cluster radius) is smaller than half of the average distance between all the existing hubs. We have slightly modified this algorithm by introducing an explicit upper bound U on the cluster radius to account for the physical dimensions of the different probe molecules. U is set equal to 2.0 Å for methanol, whereas a value of 4.0 Å is used for the other ligands. 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 >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.

Mapping Algorithms Using DOCK

Although DOCK was not designed for docking small organic solvent molecules, it works very well when docking slightly larger ligands, generating thousands of docked conformations within minutes. Thus, it may be potentially useful in our mapping protocol as a replacement for the rigid body search in step 1 of the CS-Map algorithm, which requires on the order of hours. Because DOCK is being applied to an unconventional problem, it is necessary to experiment with the various procedures and their parameter sets within the DOCK suite of programs. The first of these programs is *sphgen,* which generates a set of spheres around a number of site points that fill all pockets on the protein surface, resulting in a negative image of the receptor. The set is filtered by removing the overlapping spheres. The remaining spheres are clustered, and the next step in usual DOCK applications would be the removal of clusters that are far from the binding site. Because the search is restricted to the smallest rectangular box that contains the retained sphere clusters, reducing the number of clusters considerably speeds up the calculation. In the current article, we attempt to avoid using a priori information on the location of the binding site. Nevertheless, the number of clusters can be reduced on the basis of a geometric study of protein-ligand complexes, given that the binding site is in the largest pocket in 83% of proteins and is always among the five largest pockets.[39] Therefore, we rank the sphere clusters according to their size and retain the five largest clusters of spheres. For some of the solvents, both proteins were also mapped by using all clusters, and these calculations show that restricting consideration to the five largest clusters has little effect on the results.

Ligands can be placed by a random search within the search box or by using the (automated or manual) matching process to fit the ligand atoms to the selected site points. If random search is used, the only role of these site points is to define the box. Multiple orientations are generated, with each receiving a score quantifying its intermolecular interactions. This score is based on the electrostatic and van der Waals terms of an AMBER force field[40] that are precalculated on a grid within a selected box, saving considerable computational time. Each orientation can be adjusted by performing a short simplex search. The program can be used to dock both rigid and flexible ligands, in the latter case generating a number of ligand configurations (orientations and conformations) at each docked position.

In the *sphgen* procedure, we used a sphere radius of 1.4 Å, and 5 dots/Å^2 for the density of points on the surface. The docking parameters were 0.3 Å for the grid step, 15 Å for the cutoff distance of nonbonded interactions, 0.75 for the bump condition, and $\epsilon = 4r$ for the (distance-dependent) dielectric. No bumps were allowed. In various applications, we generated between 5000 and 50,000 ligand orientations and retained 2000–6660 of these for further analysis. In addition to rigid body docking, we also tested the effect of accounting for rotational flexibility along the C-O axis in isopropanol, phenol, and t-butanol. The conformations were minimized by using 250–1000 iterations of the simplex method in the DOCK program. The parameters selected were 0.01 kcal/mol/Å for the energy convergence condition, either 0.1 or 0.5 for the cycle convergence (the RMSD over a minimization cycle) and 0.0–1.0 kcal/mol for the energy termination. The minimization was limited to 10–20 cycles.

The primary application of DOCK in this article is to generate a number of docked probe orientations for refine-

ment and rescoring in step 2 of the CS-Map algorithm. However, we have also tested the performance of DOCK without rescoring (i.e., by clustering the 2000–6660 conformations with the lowest DOCK energies and ranking the clusters on the basis of their average DOCK energies). In a third procedure, the top five clusters (lowest average DOCK energies) were retained, and the ligand positions in these clusters were used as initial points to start a complete CS-Map run, including the multistart simplex search.

Mapping Algorithm Using GRAMM

We also explored the applicability of the docking program GRAMM[24] for replacing the multistart simplex search in step 1 of the CS-Map algorithm. GRAMM (Global RAnge Molecular Matching) was developed primarily for protein–protein docking, and it requires only the atomic coordinates of the two molecules (i.e., no a priori information on the binding site is used). The program places each molecule on a grid and performs an exhaustive six-dimensional search through the relative intermolecular translations and rotations using a very efficient Fast Fourier Transform (FFT) correlation technique and a simple scoring function that measures shape complementarity and penalizes overlaps. GRAMM may be used for docking both high-resolution molecules and low-resolution structures. We used 1.5 Å grid step for translations and 15° increments for rotations. Between 1000 and 3000 docked conformations were retained for refinement in step 2 of the CS-Map algorithm.

Determination of Consensus Sites

The primary goal of mapping is to find consensus sites at which different probe molecules cluster. For enzymes, such consensus sites occur in major subsites of the active site, thus providing detailed information on the residues that interact with the specific ligands (substrates, transition state analogs, products, and inhibitors).[18] 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 the highest number of probes 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.

RESULTS

Mapping Hen Egg-White Lysozyme

Liepinsh and Otting[6] examined the binding of methanol, methylene chloride, acetonitrile, acetone, isopropanol, t-butanol, urea, and DMSO to hen egg-white lysozyme (HEWL) in aqueous solution by using NMR techniques. They found that all eight molecules cluster in site C of the protein active site, although methylene chloride is not as deep inside the pocket as the other ligands. Based on the observed intermolecular NOEs, the solvents interact with the amide proton of N59 and protons on the side-chains of W63, I98, A107, and W108. Methanol and methylene chloride also showed NOEs with protons located at a second site in the interior of HEWL, and some weak NOEs were detected for acetone and isopropanol with additional atoms close to site C (residues W62, V109, and A110). No significant binding was observed elsewhere, and thus the only consensus site that binds all solvents is site C.

The computational mapping of HEWL was performed by using the above eight solvents as probes. Protein coordinates were downloaded from the PDB file 2lym and subjected to 100 steps of Charmm energy minimization with harmonic constraints on the positions of heavy atoms to remove steric hindrances and strains. We use the results obtained by the CS-Map algorithm[18,19] as a reference when comparing the performance of the different methods. Table I shows a number of low-average free energy clusters for the different methods. For each cluster, we list its size, probability, average free energy, the distance between the hub of the cluster, any residue in site C (N59, W62, W63, and A107), and the residue closest to the cluster (not necessarily in site C). The CS-Map algorithm places the lowest-average free energy cluster of each solvent at site C. Methylene chloride and DMSO are sligthly farther away from the bottom of the pocket than the other probes, in good agreement with the NMR data.[6]

In the DOCK calculations, the search was restricted to the first five sphere clusters generated by the *sphgen* program. These clusters contain only 170 of the 2067 nonfiltered and nonclustered site points but extend over a large fraction of the protein surface and cover the very large binding site of HEWL with its subsites A–F [Fig. 1(a)]. The rectangular box that contains the selected site points includes most of the protein. Considering all clusters [Fig. 1(b)] had little effect on the results, and we present only the results obtained by using the first five sphere clusters. Fifty thousand probes were randomly oriented in the search box and minimized. For comparison with the CS-Map results, the top 6660 probes were retained (Table I). The lowest-average energy clusters obtained by DOCK without rescoring are smaller than those obtained by CS-Map, and we find a false-positive cluster for acetonitrile. The rescoring in step 2 increased the size and the probability of the lowest free energy clusters and also removed the false positive. Using flexible docking for isopropanol and t-butanol yielded slightly larger clusters in site C, but the change in the orientations was very small.

Although Charmm rescoring yields larger clusters and increases the calculated probability of binding at site C, the results are very good even when using the average DOCK potential to rank the clusters. Thus, we conclude that site C, which is by far the largest pocket of HEWL, can be easily located by DOCK when restricting consideration to five sphere clusters. In fact, under these conditions, the conformation with the lowest DOCK energy is itself in site C for every solvent except acetonitrile. As mentioned, these results remain essentially unaffected if all sphere clusters are considered (thus slightly increasing the size of the search box). The outcome changes dramatically if we

TABLE I. Lowest-Average Free Energy Clusters of the Eight Organic Ligands Bound to HEWL[a]

	CS-Map					DOCK without rescoring					DOCK with rescoring				
Ligand	Size	*p*	⟨ΔG⟩	d/Å	*R*	Size	*p*	⟨ΔG⟩	d/Å	*R*	Size	*p*	⟨ΔG⟩	d/Å	*R*
Methanol	97	0.74	−7.18	2.3	N59	19	0.08	−6.18	2.5	A107	53	0.51	−6.87	3.6	W108
	98	0.14	−6.60	3.7	V109	49	0.13	−6.03	6.6	C94	51	0.12	−6.23	3.6	E35
	51	0.00	−5.68	7.6	K97	30	0.05	−5.56	3.7	E35	30	0.02	−5.54	8.5	G67
						26	0.02	−5.25	8.4	T69	29	0.01	−5.21	11.2	N27
						42	0.03	−5.24	2.9	N106	64	0.04	−5.32	6.4	K97
Methylene chloride	63	0.28	−3.70	3.5	Q57	53	0.53	−9.28	2.9	W108	91	0.18	−3.60	3.5	Q57
	373	0.40	−2.91	6.8	N44	27	0.01	−7.77	9.0	T69	128	0.04	−2.95	7.7	N77
						77	0.04	−7.61	3.9	Q57	155	0.04	−2.91	13.8	W123
						52	0.01	−7.31	2.6	W63	149	0.03	−2.70	2.4	N59
						151	0.04	−7.08	3.9	D52	284	0.05	−2.60	2.6	E35
Acetone	139	1.00	−13.14	2.2	N59	43	0.37	−9.71	2.4	N59	154	0.98	−12.85	2.1	N59
	63	0.00	−9.45	11.0	G117	98	0.19	−8.94	7.0	K97	23	0.00	−9.63	9.1	T69
						42	0.05	−8.67	8.6	T69	173	0.00	−9.52	11.1	G117
						121	0.06	−7.84	15.7	T118	82	0.00	−8.98	9.6	S81
						65	0.03	−7.79	3.1	V109	89	0.00	−8.95	2.2	W63
Isopropanol	51	0.92	−11.60	2.8	Q57	53	0.14	−7.56	3.7	E35	69	0.53	−14.04	3.0	W108
	50	0.01	−8.94	2.3	W63	48	0.07	−7.68	2.2	N59	85	0.43	−13.49	2.1	N59
	37	0.00	−8.81	16.9	D119	148	0.03	−6.86	15.0	W123	40	0.00	−10.98	8.8	T69
						114	0.07	−6.58	2.3	N59	69	0.00	−10.96	11.2	W108
						133	0.06	−6.78	5.7	D52	329	0.00	−9.23	7.0	V99
Tert-butanol	76	0.48	−17.53	2.0	N59	50	0.16	−8.22	2.9	G117	69	0.96	−17.05	2.1	A107
	72	0.47	−16.90	2.1	N59	158	0.08	−7.04	7.2	V99	202	0.02	−13.72	2.3	W63
	37	0.00	−14.32	6.9	E35	122	0.06	−7.04	3.4	C127	179	0.01	−12.96	15.9	S86
						79	0.04	−6.90	2.4	W62	270	0.00	−12.75	5.0	V99
						117	0.05	−6.83	2.9	D101	177	0.00	−11.7	2.7	S50
						94	0.03	−6.68	10.0	G117					
DMSO	100	0.51	−14.05	2.7	A107	109	0.08	−8.42	2.6	Q57	80	0.75	−13.49	2.8	W109
	44	0.46	−13.76	2.9	Q57	138	0.02	−7.80	7.1	A90	72	0.28	−12.63	3.0	Q57
	263	0.01	−10.69	6.8	N44	127	0.02	−7.74	15.6	W123	151	0.00	−10.03	15.7	R14
						93	0.01	−7.47	15.9	R14	174	0.00	−9.96	2.4	N59
						73	0.02	−7.82	13.3	S86	32	0.00	−9.97	8.5	T69
Urea	40	0.91	−12.10	2.2	W108	21	0.31	−11.11	2.5	A107	28	0.39	−13.58	2.0	N59
	94	0.06	−10.54	2.6	E35	28	0.13	−10.43	3.5	35E	57	0.42	−13.48	7.9	T69
	21	0.00	−9.65	8.1	T69	43	0.09	−10.40	7.7	69T	38	0.04	−12.25	2.7	E35
						89	0.14	−10.32	6.7	78I	60	0.02	−11.43	10.7	W118
						93	0.07	−9.84	5.7	35E	95	0.02	−11.29	2.0	S50
Acetonitrile	26	0.28	−10.18	2.4	W63	47	0.14	−8.15	8.0	T69	108	0.41	−9.74	2.5	W63
	154	0.39	−9.41	3.2	V109	47	0.25	−7.81	2.5	N59	52	0.06	−9.15	8.4	T69
	40	0.04	−8.81	7.7	K97	104	0.06	−6.56	7.0	K97	168	0.08	−8.49	6.8	K97
						49	0.02	−6.47	2.5	N59	212	0.05	−8.27	17.4	R14
						98	0.04	−6.34	15.4	W40	146	0.05	−8.26	6.8	N44

[a]Size, number of conformations in cluster; ⟨ΔG⟩, Boltzmann average free energy in kcal/mol; d, distance from the nearest atom of a site C residue in the protein in Ångstrom: R, the nearest residue of the protein to the small molecule.

reduce the number of conformations generated. For example, when only 6660 orientations are generated for methanol rather than 50,000, the mapping yields a large number of very small clusters, and the five lowest free energy clusters with >20 conformations are all far from the binding site. Rescoring fails to substantially rerank the clusters. Thus, generating a large number of conformations and then selecting a subset of these is absolutely necessary, which also shows that ranking on the basis of the DOCK energy provides a useful filter.

The main goal of the mapping is to locate the consensus site that binds the highest number of different ligands. Table II shows the ranking of clusters that define such consensus sites, with lower numbers indicating lower free energies. The distance of the cluster from the closest atom in site C is also shown. According to this table, CS-Map [Table II(a)] and DOCK with rescoring [Table II(b)] identify the same consensus site (site 1 in both cases), comprised of all the solvents except methylene chloride. As we mentioned, this finding is in good agreement with the results of the solvent mapping experiments.[6] DOCK without rescoring [Table II(c)] also yields a single consensus site that binds all solvents except for DMSO [site 1 in Table II(c)].

Mapping Thermolysin

The X-ray structure of thermolysin has been determined in aqueous solutions of isopropanol (IPA), acetone (ACN),

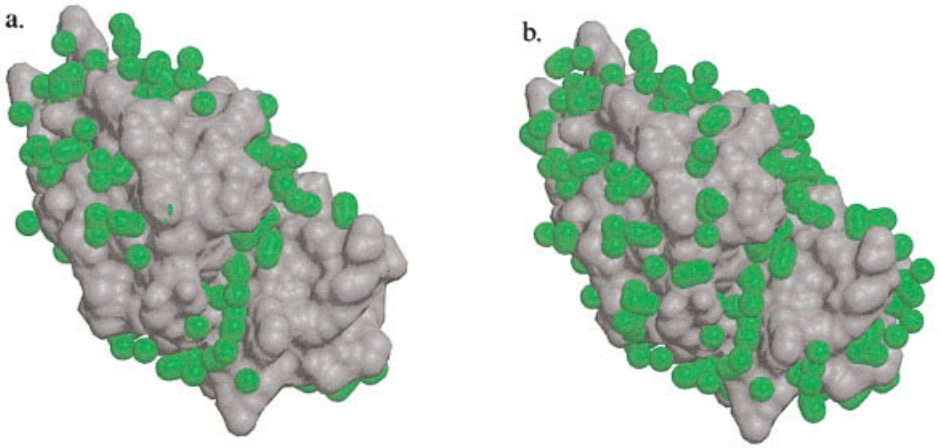

Fig. 1. Filtered and clustered site points described by 1.4 Å radius spheres on HEWL: (**a**) the first five sphere clusters; (**b**) all sphere clusters.

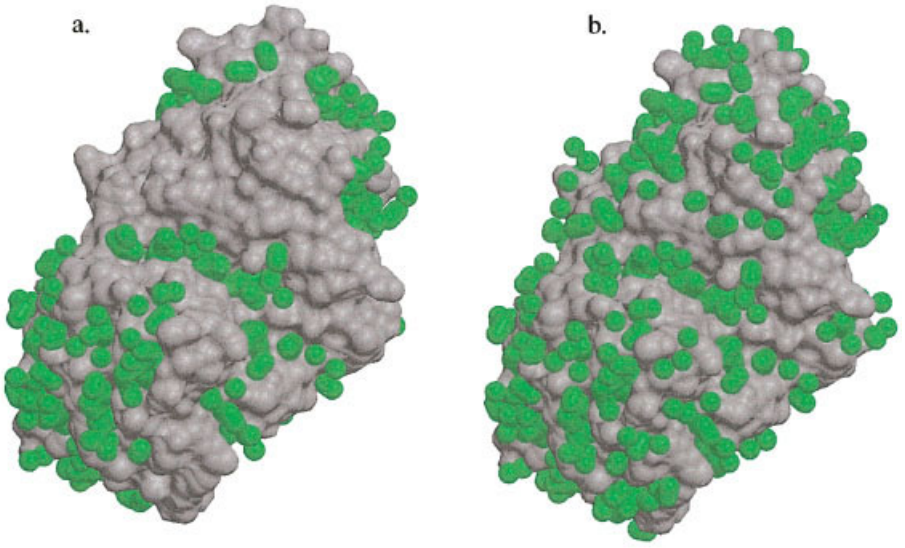

Fig. 2. Filtered and clustered site points described by 1.4 Å radius spheres on TLN: (**a**) the first five sphere clusters; (**b**) all sphere clusters.

acetonitrile (CCN), and phenol (IPH).[4,5] All four solvents bind to the main substrate specificity pocket S_1', which contains ligand molecules identified as IPA1, ACN1, CCN1, and IPH1 by the crystallographer. There are additional binding sites for IPA, ACN, and IPH; however, these appear only under certain conditions: at high organic solvent concentrations, at crystal contacts, or interacting with another bound ligand. They generally bind only a subset of the solvent molecules rather than all of them. In particular, a buried, hydrophobic pocket binds IPA2, ACN2, and IPH2, but not acetonitrile. The consensus site in the S_1' pocket is formed by a number of hydrophobic side-chains (L202, F130, L133, V139, and F114) and is surrounded by polar groups, primarily E143, R203, and H231. The binding of substrates, products, and inhibitors to thermolysin always involves the S_1' subsite, with longer ligands extending toward subsites S_1 and S_2. At least four hydrogen bonds are formed in each complex, most frequently with the side-chains of R203, E143, Y157, and N112, and with the polar backbone atoms of W115, A113, and N111.

To map thermolysin, we used the PDB coordinate file 2tlx and isopropanol, acetone, acetonitrile, and phenol as probes. Before mapping, we performed 100 steps of Charmm energy minimization with harmonic constraints on the positions of heavy atoms and removed all ligands, namely, a VK dipeptide, a DMSO molecule, and the Zn^{2+} ion bound to the active site, as well as the crystallographic water molecules. Tables III, IV, and V show the mapping results. The cluster size, probability, average free energy, and nearest protein atom are shown. The tables also give the distance between the particular cluster and the position of the closest experimentally determined ligand binding site (e.g., 0.6/8 for isopropanol in Table III indicates that the lowest free energy cluster is at 0.6 Å from IPA8.

In contrast to lysozyme, whose surface is dominated by a single large pocket, thermolysin exhibits a substantial number of crevices, and its mapping is more difficult. The

TABLE IIA. Consensus Sites of the Eight Solvent Molecules on HEWL Calculated by CS-Map

Cons. site	Methanol	Methylene chloride	Acetone	Isopropanol	Tert-Butanol	DMSO	Urea	Acetonitrile
1	1 (2.3)	—	1 (2.2)	1 (2.3)	1 (2.0)	2 (2.9)	1 (2.2)	1 (2.3)
2	2 (3.7)	1 (3.5)	—	—	2 (2.1)	1 (2.7)	—	2 (3.2)
3	5 (11.1)	5 (12.0)	2 (11.0)	4 (11.1)	—	5 (10.3)	4 (11.1)	—
4	3 (7.6)	—	—	5 (7.7)	5 (7.1)	—	4 (6.2)	3 (7.7)

TABLE IIB. Consensus Sites of the Eight Solvent Molecules on HEWL Calculated by DOCK With Rescoring

Cons. site	Methanol	Methylene chloride	Acetone	Isopropanol	Tert-Butanol	DMSO	Urea	Acetonitrile
1	1 (3.6)	—	1 (2.1)	1 (3.0) 2 (2.1)	2 (2.3)	1 (2.8)	1 (2.0)	1 (2.5)
2	2 (3.6)	1 (3.5)	—	—	1 (2.1)	2 (3.0)	3 (2.7)	—
3	—	—	—	5 (8.0)	—	—	—	—
4	4 (11.2)	2 (11.8)	3 (11.1)	4 (11.2)	—	—	4 (10.7)	—
5	3 (8.5)	—	2 (9.1)	3 (8.9)	—	—	2 (7.9)	2 (8.4)
6	—	—		—	5 (2.7)	5 (2.4)	5 (2.0)	—

TABLE IIC. Consensus Sites of the Eight Solvent Molecules on HEWL Calculated by DOCK Without Rescoring

Cons. site	Methanol	Methylene chloride	Acetone	Isopropanol	Tert-Butanol	DMSO	Urea	Acetonitrile
1	1 (2.5)	4 (2.6) 1 (2.9)	1 (2.4)	2 (2.2)	5 (2.4)	—	1 (2.5)	2 (2.5) 4 (2.5)
2	5 (2.9)	5 (3.9)	5 (3.1)	1 (3.7)	1 (2.9)	1 (2.6)	2 (3.5)	—
3	4 (8.4)	2 (9.0)	3 (8.6)	—	—	—	3 (7.7)	1 (8.0)
4	2 (6.6)	—	2 (7.0)	—	—	3 (7.1)	4 (6.7)	3 (7.0)

CS-Map results are good, but not perfect.[18] For acetonitrile, the cluster with the lowest free energy is close to CCN1, and for phenol the two lowest free energy clusters are in the S_1' pocket, the second only 0.4 Å away from IPH1. For isopropanol, the three lowest free energy clusters are in the active site, very close to IPA8, IPA1, and IPA5, respectively. However, the lowest-average free energy cluster is around IPA8 rather than IPA1, even though the latter site binds isopropanol at a much lower ligand concentration in experiments.[4] For acetone, CS-Map places the lowest-average free energy cluster at 12.8 Å from the closest acetone position in the X-ray structure, close to R260. It is not clear whether this site is a false positive, or a bound acetone is missed in the crystal structure.[18] In fact, crystallographic methods can only identify organic solvent molecules that bind in a relatively ordered manner; otherwise, the extra electron density is usually interpreted as a bound water molecule, and this particular site contains three crystallographic waters. As we discuss, the problems with isopropanol and acetone are not very important, however, because CS-Map shows only the S_1' pocket binding all four solvents, in good agreement with the experimental mapping data. Because the search in the CS-Map algorithm focuses on relatively open surface regions with favorable electrostatics and desolvation, it does not find the buried pocket containing IPA2, IPH2, and ACN2.[4,5]

Table III shows the DOCK results obtained by restricting the search to the top five sphere clusters from the *sphgen* procedure [Fig. 2(a)]. The box containing these clusters covers almost the entire protein surface. Fifty thousand orientations were generated in the box after bump filtering, and 5000 of these structures were used in the rescoring and clustering procedure. For acetone and phenol, the results are very similar to those given by CS-Map, but the lowest free energy clusters are substantially smaller. For isopropanol, the lowest free energy cluster is a false positive, found in a large pocket close to R260. However, both CS-Map and DOCK place acetone in the same pocket. As we argued, it is possible that some solvents actually bind at this location. Clusters 2 and 3 are close to the experimentally observed isopropanol positions IPA8 and IPA1, respectively. For acetonitrile, the lowest free energy cluster is again a false positive, >20 Å from the observed location of CCN1, although the highest probability cluster (cluster 2 in Table III) is in the active site. Overall, the DOCK results are somewhat weaker than the ones provided by CS-Map and, as we show, these differences interfere with identifying the correct consensus site.

The mapping was repeated by using all sphere clusters, which cover almost the whole protein [Fig. 2(b)], and gave very similar results to those obtained by using only five sphere clusters. The difference is expected to be small, because in random search the orientations are generated randomly within the smallest box containing the spheres (i.e., independently of the site points) and proceeding from five clusters to all clusters has little effect on the size of the box. We also tested the use of random search with match-

TABLE III. Lowest Average Free Energy Clusters of the Four Organic Ligands Bound to Thermolysin[a]

	CS-Map					DOCK without rescoring					DOCK with rescoring				
Ligand	Size	*p*	⟨ΔG⟩	d/Å	*R*	Size	*p*	⟨ΔG⟩	d/Å	*R*	Size	*p*	⟨ΔG⟩	d/Å	*R*
Acetone	69	0.77	−13.10	12.8/2	R260	68	0.48	−9.83	4.7/1	W115	46	0.82	−13.58	13.5/2	R260
	104	0.15	−11.89	3.2/1	Y157	53	0.07	−9.40	14.1/2	R260	76	0.13	−12.23	4.3/1	Y157
	251	0.06	−10.96	5.4/1	W115	69	0.04	−9.01	13.7/6	N233	21	0.00	−11.14	12.0/2	R35
	30	0.01	−10.75	0.6/1	R203	20	0.01	−8.88	12.0/2	R35	146	0.01	−10.80	13.8/6	N233
	98	0.01	−10.19	17.6/1	S161	73	0.07	−8.78	8.4/1	W115	155	0.01	−10.49	8.1/1	W115
Phenol	134	1.00	−28.43	2.6/1	E166	70	0.83	−16.26	10.8/2	V79	73	1.00	−29.19	3.7/1	E166
	30	0.00	−25.58	0.4/1	E143	29	0.05	−15.32	3.5/1	E143	36	0.00	−24.19	11.7/1	D185
	62	0.00	−23.33	13.1/1	D185	134	0.04	−14.32	8.7/1	D150	122	0.00	−23.42	10.6/1	D200
						53	0.04	−14.84	13.5/2	R260	68	0.00	−22.43	13.2/2	L175
						94	0.01	−14.71	8.5/1	D150	75	0.00	−20.92	26.5/2	D59
Isopropanol	129	0.40	−10.71	0.6/8	W115	63	0.44	−9.63	0.6/8	W115	44	0.79	−13.17	14.1/2	R260
	30	0.06	−10.49	0.7/1	R203	64	0.09	−9.81	2.6/8	N165	76	0.06	−10.80	1.5/8	Y157
	132	0.26	−10.31	1.0/5	D150	82	0.08	−8.06	1.3/8	Y157	93	0.03	−10.14	1.6/5	D150
	90	0.06	−10.01	13.6/8	S161	65	0.02	−7.52	13.4/3	N233	98	0.02	−10.10	14.0/8	S161
						96	0.04	−7.50	21.0/6	D43	129	0.02	−9.95	12.8/2	H88
Acetonitrile	241	0.63	−10.42	2.0/1	Y157	102	0.35	−8.28	2.9/1	W115	67	0.04	−10.80	19.8/1	N21
	75	0.06	−9.90	18.4/1	R260	145	0.12	−7.62	22.4/1	A56	177	0.57	−10.46	2.5/1	Y157
						72	0.07	−7.49	18.7/1	R260R	81	0.12	−10.13	18.3/1	R260
						76	0.07	−7.49	6.9/1	N165	20	0.01	−9.36	0.8/1	R203
						61	0.03	−7.24	20.7/1	R35	130	0.04	−9.33	27.0/1	F63

[a]Size, number of conformations in cluster; *p*, cluster probability; ⟨ΔG⟩, Boltzmann average free energy in kcal/mol; d, distance from the experimentally determined small organic solvent molecule bound to the protein in Ångstrom; R, nearest residue of the protein to the small molecule.
[b]See text for details.

TABLE IV. Lowest Average Free Energy Clusters of the Four Organic Ligands Bound to Thermolysin Calculated by DOCK Combined With CS-Map

Ligand	Algorithm 1					Algorithm 2					Algorithm 3				
	Size	*p*	⟨ΔG⟩	d/Å	*R*	Size	*p*	⟨ΔG⟩	d/Å	*R*	Size	*p*	⟨ΔG⟩	d/Å	*R*
Acetone	19	0.65	−13.20	12.8/2	R260	80	0.28	−12.23	4.3/1	Y157	21	0.80	−13.61	13.5/2	R260
	93	0.26	−11.80	3.2/1	Y157	106	0.02	−10.89	13.9/6	N233	49	0.16	−12.23	4.2/1	Y157
	101	0.07	−10.96	7.0/1	W115	115	0.02	−10.49	8.1/1	W115	26	0.00	−10.96	13.7/6	N233
	16	0.01	−10.71	11.3/2	R35	86	0.00	−7.83	15.9/2	N21	46	0.01	−10.57	3.4/1	A56
	106	0.01	−10.03	12.7/6	N233						28	0.01	−10.51	13.2/2	H88
Phenol	43	0.28	−28.49	2.6/1	E166	47	0.77	−29.22	3.7/1	E166	50	1.00	−29.19	3.2/1	E166
	67	0.72	−28.35	3.4/1	E166	44	0.23	−29.08	3.9/1	E166	46	0.00	−22.44	13.6/2	L175
	53	0.00	−22.96	11.5/2	Y83	100	0.00	−23.50	10.6/1	D200	39	0.00	−19.45	5.0/1	W115
						61	0.00	−22.45	13.2/2	L175	47	0.00	−18.42	11.7/1	D200
						39	0.00	−18.40	11.4/2	N21	39	0.00	−18.40	11.4/2	N21
Isopropanol	53	0.32	−10.81	0.7/8	W115	65	0.36	−10.76	1.5/8	Y157	31	0.15	−10.95	1.5/8	Y157
	19	0.10	−10.75	1.7/8	E143	93	0.26	−10.16	1.6/5	D150	11	0.05	−10.76	0.7/1	143E
	13	0.07	−10.73	0.7/1	R203	78	0.09	−9.66	13.2/3	N233	47	0.07	−10.05	1.6/5	150D
	85	0.40	−10.48	1.1/5	D150						11	0.01	−10.06	12.8/2	H88
	78	0.08	−9.77	12.7/3	N233						29	0.04	−10.06	20.6/6	A56
Acetonitrile	189	0.70	−10.27	2.0/1	Y157	161	0.79	−10.47	2.5/1	Y157	121	0.70	−10.46	2.5/1	Y157
	61	0.18	−10.04	18.4/1	R260	60	0.13	−10.09	18.3/1	R260	45	0.12	−10.14	18.3/1	R260
	92	0.08	−9.39	23.49	A56	66	0.03	−9.11	9.4/1	N165	30	0.04	−9.54	27.0/1	F63
	38	0.04	−9.36	7.4/1	H146	93	0.03	−8.99	22.7/1	A56	58	0.04	−9.18	22.7/1	A56
	51	0.01	−8.24	14.0/1	Y193	53	0.01	−8.80	13.2/1	Y193	22	0.01	−9.07	9.6/1	N165

ing, which places the ligands in random orientations, but at the site points. For each solvent, the lowest free energy cluster was found in the buried pocket which contains IPA2, ACN2, and IPH2. Because this pocket is not the observed consensus site, matching does not seem to provide any advantage. In addition, the search with matching requires a number of parameters that depend on the geometry of the binding site, which is generally known in docking but not necessarily known in mapping applications. Thus, in the remainder of the article, we restrict consideration to random search without matching.

Because the DOCK scoring potential does not account for solvation, we expected that DOCK without rescoring would give worse results than with rescoring, but this does

TABLE V. Lowest Average Free Energy Clusters of the Four Organic Ligands Bound to Thermolysin Calculated by GRAMM With Rescoring (2000 Probes Each)

Ligand	CS-Map					GRAMM with rescoring				
	Size	p	$\langle\Delta\rangle$	d/Å	R	Size	p	$\langle\Delta\rangle$	d/Å	R
Acetone	69	0.77	−13.10	12.8/2	R260	79	0.98	−10.88	0.6/2	Y81
	104	0.15	−11.89	3.2/1	Y157	36	0.01	−8.52	12.0/2	N21
	251	0.06	−10.96	5.4/1	W115	58	0.00	−7.53	14.0/6	N233
	30	0.01	−10.75	0.6/1	R203	71	0.00	−6.66	13.2/2	H88
	98	0.01	−10.19	17.6/1	S161	38	0.00	−6.75	0.4/1	N112
Phenol	134	1.00	−28.43	2.6/1	E166	25	1.00	−32.13	3.6/1	E166
	30	0.00	−25.58	0.4/1	E143	76	0.00	−24.14	0.6/1	E143
	62	0.00	−23.33	13.1/1	D185	28	0.00	−22.73	11.4/1	D185
						25	0.00	−21.47	12.5/2	Y83
						154	0.00	−20.14	0.4/2	L144
Isopropanol	129	0.40	−10.71	0.6/8	W115	65	0.99	−10.59	0.4/2	I100
	30	0.06	−10.49	0.7/1	R203	57	0.01	−7.60	7.5/6	N21
	132	0.26	−10.31	1.0/5	D150	48	0.00	−7.33	12.8/3	G228
	90	0.06	−10.01	13.6/8	S161	26	0.00	−6.82	0.7/1	E143
						50	0.00	−6.11	12.6/2	K265
Acetonitrile	241	0.63	−10.42	2.0/1	Y157	39	0.20	−7.27	19.9/1	N21
	75	0.06	−9.90	18.4/1	R260	157	0.76	−7.04	14.7/1	Y93
						49	0.01	−5.24	1.2/1	E143
						62	0.02	−5.33	0.7/1	R203
						74	0.01	−4.86	17.4/1	N233

TABLE VI. Ranking of Probe Clusters Within the Consensus Sites[a] on TLN

Algorithm	Con. site	Probe			
		Acetone	Phenol	Isopropanol	Acetonitrile
CS-Map	**1 (S_1')**	**4 (0.8)**	**2 (0.7)**	**2 (0.7)**	**3 (0.3)**
	2 (S_1)	2 (3.7)	1 (3.5)	—	1 (3.9)
	3	1 (18.1)	4 (17.3)	—	2 (17.4)
DOCK without rescoring	**1 (S_1')**	—	—	—	—
	2 (S_1)	**1 (3.7)**	**2 (2.7)**	**1 (4.1)**	**1 (3.3)**
	3	5 (7.6)	3 (8.7)	2 (9.7)	4 (7.3)
	4	5 (7.6)	3 (8.7)	2 (9.7)	4 (7.3)
			5 (8.7)	3 (8.5)	
	5	4 (21.0)	1 (19.0)	—	5 (21.4)
DOCK with rescoring	1 (S_1')	—	—	—	4 (0.6)
	2 (S_1)	**2 (3.6)**	**1 (3.6)**	**2 (3.6)**	**2 (2.9)**
	3	**1 (18.8)**	**4 (17.0)**	**1 (18.5)**	**3 (18.1)**
DOCK combined with CS-Map	1 (S_1')	—	—	3 (0.7)	—
	2 (S_1)	**2 (3.7)**	**1 (3.5)**	**2 (3.7)**	**1 (3.9)**
			2 (3.9)		
	3	3 (7.0)	—	4 (9.2)	4 (9.2)
GRAMM with rescoring	**1 (S_1')**	**5 (0.2)**	**2 (0.3)**	**4 (0.7)**	**4 (0.1)**
	2	1 (14.5)	—	1 (14.4)	2 (14.2)
	3	2 (20.1)	—	2 (20.2)	1 (19.8)
	4	3 (17.5)	—	3 (16.9)	5 (16.7)

[a]Consensus sites shown in bold, distance from IPA1 in brackets.

not seem to be the case. The highest probability clusters are in the active site for acetone, isopropanol, and acetonitrile, but the rescoring improves the results for phenol. Because the DOCK search is extremely fast, it makes sense to somehow use its ability to find relatively good conformations. In Table IV, we show the results for three algorithms that use an initial DOCK search combined with the CS-Map method in different ways. In algorithm 1, we cluster the DOCK results, rank the clusters based on the average DOCK energy, and retain only the top five clusters for rescoring and clustering in steps 2 and 3 of CS-Map. Thus, in contrast to the combined search in Table III, we use DOCK for selecting a small number of orientations before rescoring. In particular, for thermolysin, the top five DOCK clusters contain between 254 and 434 orientations.

In algorithm 2, we select and rescore the same number of top conformations from DOCK, but we use the individual energies rather than cluster averages. Algorithm 3 initially proceeds as algorithm 1, but the conformations in the five clusters with the lowest-average DOCK energies

are now used as starting points in a CS-Map search, in which a single simplex minimization is performed. The major potential advantage of all three strategies is to reduce the number of structures that need to be minimized and rescored in step 2 of the mapping.

As mentioned, the only difference between DOCK with rescoring (Table III) and algorithm 2 (Table IV) is that in the latter we retain far fewer orientations from DOCK for minimization and rescoring. The two methods yield very similar results for acetone and phenol. Although the low free energy clusters are slightly larger in Table III than in Table IV, the change in cluster size is small relative to the >10-fold difference in the number of structures that were minimized. It is interesting that retaining fewer probes actually improves the results for isopropanol and acetonitrile. Although DOCK with rescoring yields false positives for these two probes (Table III), the clusters at the same locations become very small when algorithm 2 is used. We show these clusters in Table IV to support this statement, but clusters this small are usually removed (see Materials and Methods). Comparing the results of algorithm 2 with those of algorithm 1 shows that retaining the top five DOCK clusters yields slightly larger low free energy clusters than retaining the same number of conformations with low DOCK energies. Finally, the results of algorithm 3 show that the search in step 1 of the CS-Map procedure generally moves the low free energy clusters toward the experimental ligand positions and increases their sizes.

Table V compares the results of GRAMM-based docking with those of the CS-Map algorithm. The five lowest free energy clusters produced by GRAMM always include one that is very close to the experimental ligand-binding position, but this particular cluster generally does not have the lowest free energy, and the method can yield as many as four false positives. Similar to the DOCK search with random matching, GRAMM also finds conformations close to ACN2, PHN2, and IPH2, which are in a buried pocket. However, as we see, GRAMM also places an acetonitrile molecule in this pocket, although the X-ray structure shows acetonitrile binding only in the active site. The GRAMM results are meaningless if we use its own score function with clustering.

Table VI shows the ranking of clusters that define the consensus site for thermolysin obtained by using the different methods, with lower numbers indicating lower free energies. The table also shows the distance of these clusters from ligand IPA1. IPA1, ACN1, CCN1, and IPH1 overlap in the S_1' pocket. The CS-Map algorithm yields a single consensus site in the S_1' pocket, in good agreement with experimental mapping, although the clusters that define the site do not have the lowest free energy for any of the probes. A second site, which is close to the S_1 pocket, binds three of the probes. All DOCK-based methods (without rescoring, with rescoring, and combined with CS-Map) find only this second location as the consensus site. In fact, these methods do not show any binding in the S_1' pocket. The only exception is the combined algorithm, which places only an isopropanol cluster in that region.

The GRAMM-based mapping finds the consensus site in the S_1' pocket (Site 1 for GRAMM in Table VI). The method places the lowest free energy clusters of acetone and isopropanol, and the second lowest free energy clusters of acetonitrile in an almost completely buried pocket (Site 2 for GRAMM in Table VI). It is interesting that this pocket has been shown experimentally to bind three of the four solvents, but the exception being acetonitrile rather than phenol.

DISCUSSION

Computational solvent mapping of proteins is an unusual and challenging problem for two reasons. First, we attempt to avoid any a priori assumption on the binding site and, thus, search the entire protein surface. Accurate calculation of a potential around the protein on a regular grid would require a very large number of probes, and hence, clever strategies are needed to achieve better sampling of the regions of interest. Second, because the probes are small and not specific to the target protein, one has to compare very weak interactions to rank the binding positions. This is very different from the problem of screening a database of small compounds for molecules that bind to a known site with high affinity. The success of docking programs, such as DOCK,[23] shows that the latter problem can be solved by using simple but sensitive target functions (e.g., the van der Waals energy or a measure of shape complementarity). In contrast, successful mapping is likely to require the use of a potential that can consistently represent the various contributions to the binding free energy, including desolvation, and possibly even entropic effects. In view of these difficulties, it is not particularly surprising that current mapping methods result in hundreds of energy minima, although experimental solvent mapping by X-ray crystallography or NMR shows that the organic solvents bind only to a limited number of sites on a protein.[1–6]

We attempted to overcome the above difficulties and developed the CS-Map mapping algorithm, which differs from traditional mapping methods in three major respects: (i) there is better sampling of protein surface regions with favorable electrostatics and desolvation; (ii) the scoring potential accounts for desolvation; and (iii) the docked ligand positions are clustered, and the clusters are ranked on the basis of their average free energies. To better understand both the mapping problem and the CS-Map algorithm, we compare CS-Map with the powerful docking programs DOCK and GRAMM, by adding a rescoring step to the latter two. We emphasize that the initial search step in the CS-Map algorithm differs substantially from DOCK and GRAMM. This difference is illustrated in Figure 3. The target function in GRAMM rewards shape complementarity and penalizes overlaps by using negative and large positive terms, respectively. The target function of DOCK includes van der Waals and electrostatic terms. Both functions result in a rugged energy surface with large numbers of local minima, separated by high-energy barriers. Thus, the probes cannot move far away from their initial locations, and the regions of interest must be

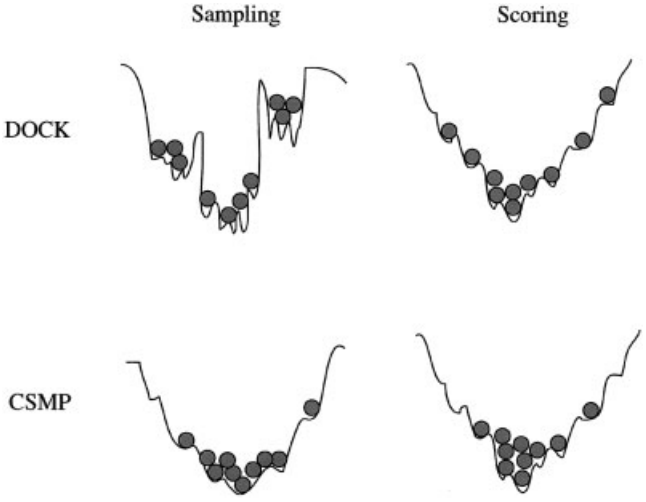

Fig. 3. Sampling and scoring differences between CS-Map and DOCK calculations.

thoroughly sampled. Rescoring with CHARMM and accounting for desolvation removes some of these energy barriers and gives the free energy surface a somewhat broader minimum around the native complex. In contrast, in step 1 of the CS-Map algorithm, we assume van der Waals cancellation, and the potential includes only electrostatic and desolvation terms (Eq. 1), because $V_{exc} = 0$ if the ligand does not overlap with the protein. The use of this potential yields a relatively smooth energy surface, and the probes move an average distance of 3.5 Å toward regions of the protein with favorable electrostatics and desolvation. Relatively good sampling of these regions can be achieved by moving the probes around the protein by the simplex method. Because assuming van der Waals cancellation is only a first-order approximation, the resulting conformations need to be rescored by using the complete potential (Eq. 2), which does account for the van der Waals energy.

The CS-Map method also differs from GRAMM and DOCK in accounting for the accessibility of the binding site. In GRAMM, the ligands are placed on a grid, whereas DOCK uses either random search or matching of the site points. However, both programs can find almost completely buried pockets if they are large enough to accommodate the ligand. In contrast, the simplex search in CS-Map moves the ligand along the protein surface, to a certain degree emulating the physical binding process. Indeed, step 1 finds a recognition site where electrostatics and desolvation broadly define a region favorable for binding, whereas step 2 accounts for the increasing role of van der Waals and other short-range interactions and ligand flexibility. This type of search cannot cross the energy barriers because of steric overlaps and, hence, does not find any buried pockets. At the same time, the method used to initially place the probe molecules on the surface can have a significant effect on the results.

Despite their differences, CS-Map and DOCK with rescoring yield similar results for lysozyme. The results are not substantially different when using DOCK without rescoring, even though in this case the target function does not account for the desolvation. Thus, according to these results, site C of lysozyme can be found by any reasonable method. Therefore, we did not test other strategies on lysozyme and proceeded to the mapping of thermolysin, a more difficult problem.

Table VI, the consensus sites for thermolysin, shows that the three methods yield very different results. The only consensus site found by CS-Map is in the S_1' pocket, in good agreement with the experimental data. In contrast, all DOCK-based methods place the consensus site in the S_1 pocket. Because one of the intended uses of the mapping is the identification of enzyme active sites and because both S_1' and S_1 are major subsites of the thermolysin active site, finding S_1 rather than S_1' is not a terrible mistake, but it clearly disagrees with the experimental mapping results. An inspection of the thermolysin structure reveals that S_1' is a very well defined, largely hydrophobic pocket, whereas the adjacent S_1 and S_2 subsites form a relatively narrow and more polar crevice. Because the DOCK scoring function accounts for direct electrostatics but not for desolvation, the lack of a hydrophobic effect biases the search toward sampling of the S_1 site.

As shown in Table VI, once the sampling has been done by DOCK, neither rescoring nor an additional CS-Map search can move the consensus site from S_1 to S_1'; in fact, the CS-Map method finds only one solvent, isopropanol, in its correct position in the S_1' site. The recently released version 5 of the DOCK program[41] includes the possibility of using a GB/SA solvation model,[42] and this may be able to solve some of the problems we have encountered in the present work.

GRAMM finds the correct consensus site in the S_1 subsite. It also places three different probes in a buried pocket which actually binds three of the ligands, and is not found by the other two methods. Thus, we conclude that both CS-Map and GRAMM provide a reliable tool for the identification of enzyme active sites by computational mapping. Nevertheless, the performances of the two algorithms show important differences, CS-Map generally finding relatively open, large binding sites, whereas GRAMM also entering into almost completely buried pockets that are not accessible to CS-Map. This property of GRAMM is an advantage in some cases. For example, in order to find the active site of haloalkane dehalogenase, located in the middle of a long and narrow channel, one would require a GRAMM-based search and small probes that fit into the channel. Indeed, CS-Map places clusters only at the two ends of the channel. Thus, we can increase the reliability of mapping by overlapping the positions given by CS-Map and GRAMM, an approach we use in our most current work.

As mentioned, the CS-Map algorithm differs from other mapping methods in three major respects (i.e., sampling, scoring function, and clustering). Table VI shows that the DOCK sampling is not good enough to find the consensus site. According to Table II, the rescoring can be crucial, and it generally increases the size of the lowest free energy

clusters. Table IV shows that clustering at the DOCK level increases the sizes of low free energy clusters and that running a CS-Map search after DOCK slightly improves the results but is unable to change them substantially. Without clustering, there would be no posibility of finding the groups of molecules in different binding sites. Thus, we conclude that all three factors play important roles in avoiding spurious local minima, and the CS-Map initial search can probably be singled out as the most important contribution.

REFERENCES

1. Mattos C, Ringe D. Locating and characterizing binding sites on proteins. Nat Biotech1996;14:595–599.
2. Allen KN, Bellamacina CR, Ding X, Jeffery CJ, Mattos C, Petsko GA, Ringe D. An experimental approach to mapping the binding surfaces of crystalline proteins. J Phys Chem 1996;100:2605–2611.
3. Mattos C, Ringe D. Protein in organic solvents. Curr Opin Struct Biol 2001;11:761–764.
4. English AC, Done SH, Caves LS, Groom CR, Hubbard RE. Locating interaction sites on proteins: the crystal structure of thermolysin soaked in 2% to 100% isopropanol. Proteins 1999;37: 628–640.
5. English AC, Groom CR, Hubbard RE. Experimental and computational mapping of the binding surface of a crystalline protein. Protein Eng 2001;14:47–59.
6. Liepinsh E, Otting G. Organic solvents identify specific ligand binding sites on protein surfaces. Nat Biotech 1997;15:264–268.
7. Shuker SB, Hajduk PJ, Meadows RP, Fesik SW. Discovering high affinity ligands for proteins: SAR by NMR. Science 1996;274:1531–1534.
8. Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem 1985;28:849–875.
9. Bohm HJ. The computer program LUDI: a new method for the de novo design of enzyme inhibitors. J Comput Aided Mol Des 1992;1:61–78.
10. Lawrence MC, Davis PC. CLIX: a search algorithm for finding novel ligands capable of binding proteins of known three-dimensional structure. Proteins 1992;12:31–41.
11. Verlinde CLMJ, Rudenko G, Hol WGJ. In search of new lead compounds for trypanosomiasis drug design: a protein structure-based linked-fragment approach. J Comput Aided Mol Des 1992;6: 131–147.
12. Rotstein SH, Murcko MA. GroupBuild: a fragment-based method for de novo drug design. J Med Chem 1993;36:1700–1710.
13. King BL, Vajda S, DeLisi C. Empirical free energy as a target function in docking and design: application to HIV-1 protease inhibitors. FEBS Lett 1996;384:87–91.
14. Miranker A, Karplus M. Functionality maps of binding sites: a multiple copy simultaneous search method. Proteins 1991;11:29–34.
15. Caflisch A, Miranker A, Karplus M. Multiple copy simultaneous search and construction of ligands in binding sites: application to inhibitors of HIV-1 aspartic proteinase. J Med Chem 1993;36:2142–2167.
16. Evensen E, Joseph-McCarthy D, Karplus M. MCSS version 2.1. Cambridge, MA: Harvard University; 1997.
17. Eisen MB, Wiley DC, Karplus M, Hubbard RE. HOOK: a program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecular binding site. Proteins 1994;19:199–221.
18. Dennis S, Kortvelyesi T, Vajda S. Computational mapping identifies the binding sites of organic solvents on proteins. Proc Natl Acad Sci USA 2002;99:4290–4295.
19. Liang J, Edelsbrunner J, Woodward C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci 1998;7:1884–1897.
20. Skolnick J, Fetrow JS, Kolinski A. Structural genomics and its importance for gene function analysis. Nat Biotech 2000;18:283–287.
21. Ondrechen MJ, Clifton JG, Ringe D. THEMATICS: a simple computational predictor of enzyme function from structure. Proc Natl Acad Sci USA 2001;98:12473–12478.
22. Brady GP Jr, Stouten PFW. Fast prediction and visualization of protein binding pockets with PASS. J Comput Aided Mol Des 2000;14:383–401.
23. Ewing TJA, Kuntz ID. Critical evaluation of search algorithms used in automated molecular docking. J Comp Chem 1997;18:1175–1189.
24. Vakser IA, Matar OG, Lam CF. A systematic study of low-resolution recognition in protein-protein complexes. Proc Natl Acad Sci USA 1999;96:8477–8482.
25. Nelder JA, Mead R. A simplex method for function minimization. Computer J 1964;7:308–313.
26. Dennis S, Camacho CJ, Vajda S. Exploring potential solvation sites of proteins by multistart local minimization. In: Floudas CA, Pardalos C, editors. Optimization in computational chemistry and molecular biology. Norwell, MA: Kluwer Academic; 2000. 243 p.
27. Vajda S, Weng Z, Rosenfeld R, DeLisi C. The effect of conformational flexibility and solvation on receptor-ligand binding free energies. Biochemistry 1994;33:13977–13987.
28. Weng Z, Vajda S, DeLisi C. The prediction of protein complexes using empirical free energy functions. Protein Sci 1996;5:614–626.
29. Jackson RM, Sternberg MJE. A continuum model for protein–protein interactions: application to the docking problem. J Mol Biol 1995;250:258–275.
30. Krystek S, Stouch T, Novotny J. Affinity and specificity of serine endopeptidase-protein inhibitor interactions. J Mol Biol 1993;234: 661–679.
31. Gilson MK, Honig B. Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies, and conformational analysis. Proteins 1988;4:7–18.
32. Honig B, Nicholls A. Classical electrostatics in biology and chemistry. Science 1995;268:1144–1149.
33. Bruccoleri RE. Grid positioning independence and the reduction of self-energy in the solution of the Poisson-Boltzmann equation. J Comp Chem 1993;14:1417–1422.
34. QUANTA/CHARMM Program, Molecular Simulations Inc., Waltham, MA, USA; 1994.
35. Zhang C, Vasmatzis G, Cornette JL, DeLisi C. Determination of atomic desolvation energies from the structures of crystallized proteins. J Mol Biol 1996;267:707–726.
36. Miyazava S, Jernigan R. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–552.
37. Schaefer M, Karplus M. A comprehensive analytical treatment of continuum electrostatics. J Phys Chem 1996;100:1578–1599.
38. Brooks BR, Bruccoleri RE, Olafson B, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comp Chem 1983;4: 197–214.
39. Laskowski RA, Luscombe NM, Swindells MH, Thornton JM. Protein clefts in molecular recognition and function. Protein Sci 1996;5:2438–2452.
40. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona P, Profeta S Jr, Weiner P. A new force field for molecular mechanical simulation of nucleic acids and proteins. J Am Chem Soc 1984;106: 765–784.
41. DOCK 5.0.0. San Francisco, CA: University of California; 2002.
42. Qiu D, Shenkin PS, Hollinger FP, Still WC. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. J Phys Chem A 1997:101:3005–3014.