Automated generation of MCSS-derived pharmacophoric DOCK site points for searching multiconformation databases

Automated Generation of MCSS-Derived Pharmacophoric DOCK Site Points for Searching Multiconformation Databases

Diane Joseph-McCarthy* **and Juan C. Alvarez**
Wyeth Research, Biological Chemistry Department, Cambridge, MA

ABSTRACT **All docking methods employ some sort of heuristic to orient the ligand molecules into the binding site of the target structure. An automated method, MCSS2SPTS, for generating chemically labeled site points for docking is presented. MCSS2SPTS employs the program Multiple Copy Simultaneous Search (MCSS) to determine target-based theoretical pharmacophores. More specifically, chemically labeled site points are automatically extracted from selected low-energy functional-group minima and clustered together. These pharmacophoric site points can then be directly matched to the pharmacophoric features of database molecules with the use of either DOCK or PhDOCK to place the small molecules into the binding site. Several examples of the ability of MCSS2SPTS to reproduce the three-dimensional pharmacophoric features of ligands from known ligand–protein complex structures are discussed. In addition, a site-point set calculated for one human immunodeficiency virus 1 (HIV1) protease structure is used with PhDOCK to dock a set of HIV1 protease ligands; the docked poses are compared to the corresponding complex structures of the ligands. Finally, the use of an MCSS2SPTS-derived site-point set for acyl carrier protein synthase is compared to the use of atomic positions from a bound ligand as site points for a large-scale DOCK search. In general, MCSS2SPTS-generated site points focus the search on the more relevant areas and thereby allow for more effective sampling of the target site. Proteins 2003;51:189–202.** © **2003 Wiley-Liss, Inc.**

Key words: **multiple copy simultaneous search; fragment positioning methods; theoretical pharmacophores; virtual screening; pharmacophore-based molecular docking; MCSS2SPTS; PhDOCK**

INTRODUCTION

Although high-throughput robotic methods[1]have accelerated the process of screening large corporate databases and combinatorial libraries, it is still not possible to screen all available compounds experimentally. In addition, recent advances in combinatorial chemistry have dramatically increased the number of new compounds that can be synthesized.[2,3] As a result, there is a critical need for fast, reliable computational methods for virtually screening large, three-dimensional (3D) libraries and molecular databases. This includes methods for structure-based screening of diverse libraries for potential binding to a specific target.

Due to genome sequencing projects,[4] as well as structural genomics efforts,[5–8] there is an increasingly large number of homology models and experimentally determined structures for therapeutically relevant targets. Structure-based virtual screening methods, which utilize the information contained in the 3D structure of a macromolecular target, will therefore play an increasingly important role in the identification of new lead compounds for drug discovery.[9–11] Hits from the virtual screen can be used to identify novel scaffolds for binding to a particular target, as well as to guide the design and synthesis of focused combinatorial libraries.

Docking methods, in general, employ some sort of heuristic to orient the ligand molecules into the binding site of the target structure. In the program DOCK,[12–14] ligand atoms are matched to predefined site points in the binding region of the target structure, whereas in the related PhDOCK,[15] pharmacophore points are matched to site points to orient pharmacophoric ensembles of conformers. PhDOCK is implemented in DOCK4.0.[16] For both methods, site points can be chemically labeled to indicate the type of atom that they are allowed to match, and it can be required that at least one site point from a subset, or a critical cluster, be matched. The way to generate default site points is to use the SPHGEN utility that accompanies DOCK, which creates a set of overlapping spheres that fill the binding site and are tangent to the molecular surface at only two points. The sphere centers are taken as unlabeled site points. Crystallographic water molecules or experimental positions of known ligand atoms can also be taken as site points. Chemically labeling the site points can significantly reduce the search time by restricting the search space to areas relevant to the target, thereby reducing the combinatorial problem. Labeling a set of site points manually, however, is a time-consuming and often ambiguous task.

*Correspondence to: Diane Joseph-McCarthy, Wyeth Research, Biological Chemistry Department, 87 Cambridge Park Dr., Cambridge, MA 02140. E-mail: djoseph@wyeth.com

Received 12 December 2001; Accepted 12 September 2002

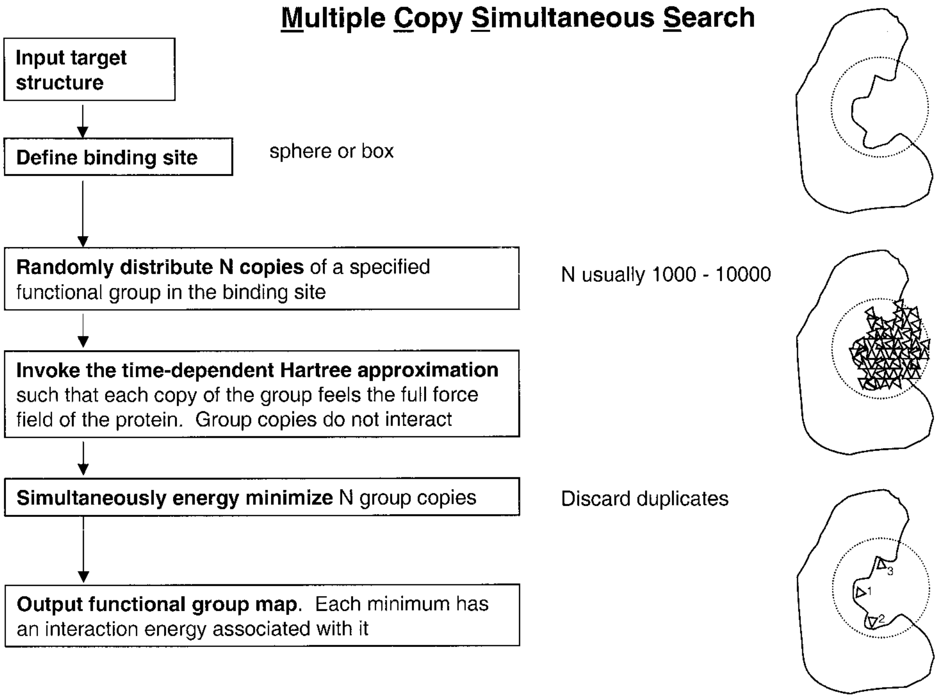

Multiple Copy Simultaneous Search
Input target structure
Define binding site
sphere or box
Randomly distribute N copies of a specified functional group in the binding site
N usually 1000 - 10000
Invoke the time-dependent Hartree approximation such that each copy of the group feels the full force field of the protein. Group copies do not interact
Simultaneously energy minimize N group copies
Discard duplicates
Output functional group map. Each minimum has an interaction energy associated with it
3
1
2

Fig. 1. A flowchart of the MCSS method.

In this article, we describe an automated method, MCSS2SPTS, for generating chemically labeled DOCK site points that are derived from theoretical pharmacophores calculated for a given target using the Multiple Copy Simultaneous Search (MCSS) approach.[17–19] Examples that validate MCSS2SPTS by reproducing the pharmacophoric points present in ligand–protein complex structures are presented. MCSS2SPTS was written as an auxilliary to PhDOCK, so that pharmacophore points, including ring centroids, could be efficiently matched to pharmacophoric site points. It is, however, a general-purpose method for generating chemically meaningful, target-derived site points. The use of these pharmacophoric site points with multi-conformation DOCK, as well as PhDOCK, is discussed.

METHODS

MCSS to Generate Target-Derived Pharmacophores

Fragment positioning methods determine energetically favorable binding-site positions for various functional group types or chemical fragments; three well-known programs are GRID,[20] MCSS,[17,18] and LUDI.[21,22] With the MCSS program (Fig. 1), probe groups are fully flexible, and individual atoms are represented with the use of the CHARMM[23] potential energy function. Very briefly, several thousand copies of a given functional group are randomly distributed in the binding site, which can either be a sphere or a box. The functional group copies are then simultaneously minimized in the time-dependent Hartree approximation such that each copy of the group feels the full force field of the protein, but the group copies do not interact with each other. For the purposes of generating DOCK site points with MCSS2SPTS, for example, a calculation is run with 1000 copies of benzene. At the end of the calculation, a functional group map obtained for benzene identifies the preferred binding positions (potential energy minima) of benzene in the target structure. The resulting MCSS maps are somewhat analogous to experimental mapping of a protein surface by determining its 3D structure in various organic solvents.[24–26] Fragment positioning methods, such as MCSS, can thereby be used to determine or combinatorially generate possible structure-based pharmacophores. Traditionally, a pharmacophore is the set of features common to a series of active molecules. A 3D pharmacophore specifies the spatial relationship between the groups or features, often defining distances or distance ranges between groups, angles between groups or planes, and exclusion spheres.

MCSS2SPTS to Generate DOCK Site Points

Chemically labeled site points can be generated in an automated fashion with the use of the script MCSS2SPTS (Fig. 2), which takes as input the coordinates for the macromolecule [in (CRD) format] (Molecular Simulations, Inc., 2001, San Diego, CA) and for the box surrounding the binding site [in Protein Data Bank (PDB) format]. The box can be obtained with the use of the DOCK utility SHOWBOX, and the same box can be used for subsequent DOCK calculations. Protein coordinates can be prepared in Quanta or InsightII (Molecular Simulations, Inc., 2001, San Diego, CA) by adding polar or all hydrogens. MCSS2SPTS runs a series of MCSS calculations on the macromolecular struc-

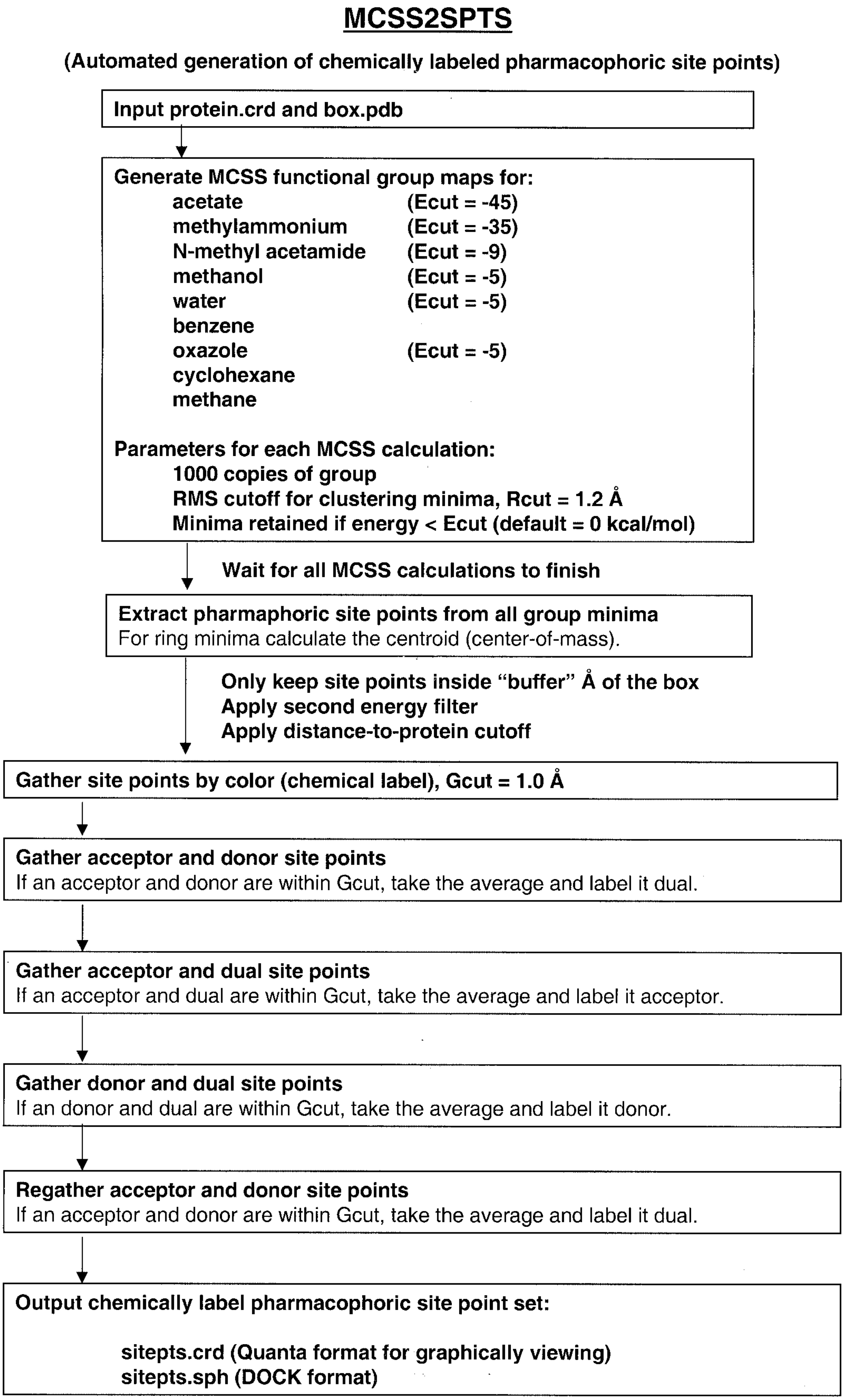
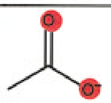
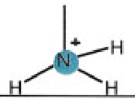
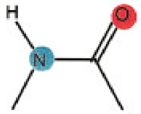
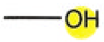
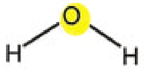
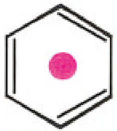
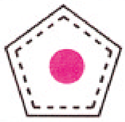
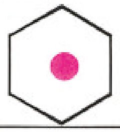

MCSS2SPTS
(Automated generation of chemically labeled pharmacophoric site points)
Input protein.crd and box.pdb
Generate MCSS functional group maps for:
acetate (Ecut = -45)
methylammonium (Ecut = -35)
N-methyl acetamide (Ecut = -9)
methanol (Ecut = -5)
water (Ecut = -5)
benzene
oxazole (Ecut = -5)
cyclohexane
methane
Parameters for each MCSS calculation:
1000 copies of group
RMS cutoff for clustering minima, Rcut = 1.2 Å
Minima retained if energy < Ecut (default = 0 kcal/mol)
Wait for all MCSS calculations to finish
Extract pharmaphoric site points from all group minima
For ring minima calculate the centroid (center-of-mass).
Only keep site points inside "buffer" Å of the box
Apply second energy filter
Apply distance-to-protein cutoff
Gather site points by color (chemical label), Gcut = 1.0 Å
Gather acceptor and donor site points
If an acceptor and donor are within Gcut, take the average and label it dual.
Gather acceptor and dual site points
If an acceptor and dual are within Gcut, take the average and label it acceptor.
Gather donor and dual site points
If an donor and dual are within Gcut, take the average and label it donor.
Regather acceptor and donor site points
If an acceptor and donor are within Gcut, take the average and label it dual.
Output chemically label pharmacophoric site point set:
sitepts.crd (Quanta format for graphically viewing)
sitepts.sph (DOCK format)

Fig. 2. Overview of the MCSS2SPTS method.

TABLE I. Automated Generation of Chemically Labeled Pharmacophoric Site Points

Group Name	Pharmacophoric site point(s)	Chemical label	Database ligand atom matches
Acetate		Acceptor	Acceptor, Dual
Methylammonium		Donor	Donor, Dual
N-Methylacetamide		Acceptor, Donor	Acceptor, Donor, Dual
Methanol		Dual	Acceptor, Donor, Dual
Water		Dual	Acceptor, Donor, Dual
Benzene		Hydrophobic	Hydrophobic
5-membered Aromatic Ring		Hydrophobic	Hydrophobic
Cyclohexane		Hydrophobic	Hydrophobic
Methane		Neutral	Any Atom

ture and automatically derives the chemically labeled site points from the resulting functional group maps.

Maps are calculated for acetate, methyl ammonium, *N*-methyl acetamide, methanol, water, cyclohexane, a five-membered aromatic ring, benzene, and methane (see Table I). Either the polar-hydrogen representation or the hybrid-hydrogen representation (all hydrogens on the protein and polar hydrogens on the functional group copies) can be specified. The script allows the MCSS calculations to be run simultaneously on multiple processors or serially, if only one processor is available. Site points are extracted from the lower energy minima (Table I) and further clustered based on type, position, and energy to determine a set of chemically labeled site points (Fig. 2). Site points are chemically labeled to indicate the type of atom that they are allowed to match. The site-point types can be hydrogen-bond acceptors, hydrogen-bond donors, and duals that can act as acceptors or donors, ring centroids, and neutrals.

As an example, from acetate minima, acceptor site points are located in the binding site. Ring centroids are calculated as the center of mass of ring minima (for 5- and 6-membered aromatic and 6-membered saturated rings). After the site points are extracted, only those well within the box boundaries (box coordinates minus a buffer distance of 2 Å) are saved. Although all the functional group copies are originally placed within the box, they may drift outside the box during minimization. Then, a second energy filter is applied. More specifically, for each site-point type, the new energy cutoff is set to a scale factor times the lowest energy site point of that type still inside the box. Scale factors of 0.6 for neutrals and centroids, 0.4 for donors and duals, and 0.35 for acceptor site points were chosen, based on their general ability to reproduce the 3D

TABLE II. Site Points for HIV1 Protease 1DIF Used to Dock Multiple Known Ligands

pdb	Pharmacore points[a]	Points within 1 Å of a matching site point[b]	Docked[c]	Docked in the binding site[c]	Energy score[d]	RMSD with X-ray[d]
1dif	16	8	Yes	Yes	−63.1	0.24
1hvk	16	9	Yes	Yes	−58.2	0.35
1ody	17	4	Yes	Yes	−56.0	0.58
9hvp	15	7	Yes	Yes	−47.0	0.80
1qbu_1	11	2	Yes	Yes	−44.1	0.53
1tcx	11	3	Yes	Yes	−38.9	0.71
1hvl	16	12	Yes	Yes	−36.3	0.31
1upj	6	5	Yes	Yes	−32.0	0.40
1qbu_2	11	4	Yes	Yes	−30.7	0.84
2aid_1	8	1	Yes	Yes	−29.5	0.53
2aid_2	8	3	Yes	Yes	−25.3	1.70
8hvp	21	5	Yes	Yes	−25.1	1.51
1hsg	14	5	Yes	Yes	−24.9	10.99
1hpv	11	7	Yes	Yes	−22.9	9.23
2bpv	13	5	Yes	Yes	−21.2	0.69
1a30	11	4	Yes	Yes	−20.7	0.62
1hxw	16	6	Yes	Yes	−18.6	1.23
4phv	13	6	Yes	Yes	−18.5	0.54
1htg_2	19	4	Yes	Yes	−13.1	0.16
1a8g	18	6	Yes	Yes	−11.8	0.29
1hvr	9	4	Yes	—	−10.1	13.12
1ohr	12	5	Yes	Yes	−7.7	9.26
1hxb	16	4	—	—		
1odw	11	1	—	—		
4hvp	19	5	—	—		
1htf_1	13	4	—	—		
1htf_2	13	5	—	—		
1htg_1	19	9	—	—		

[a] The number of pharmacophore features in the ligand.

[b] The number of ligand pharmacophore points that are within 1 Å of an appropriately labeled site point, where ligand acceptors match to acceptor or dual site points, donors match to donor or dual site points, duals match to acceptor, donor, or dual site points, centroids match to centroids, and all pharmacophore points can match to neutral site points.

[c] A dash indicates that the ligand was not docked with a negative score.

[d] Energies are in kcal/mol and root-mean-square deviations are in Å.

pharmacophoric features of ligands in X-ray protein complex structures. For site points derived from *N*-methyl acetamide minima, an additional restriction is applied, because two site points are obtained from each minimum. Only site points within 3.6 Å of a protein atom within the box are retained. This ensures that each site point obtained from a given minimum is in fact interacting with the protein target.

At this point, like site points are clustered together with the use of a distance cutoff, so that acceptors are clustered with acceptors, donors with donors, duals with duals, centroids with centroids, and neutrals with neutrals. For all types except centroids, the position of the site point extracted from the lower energy minima is retained. For centroids, the average position of the pair is retained. Next, acceptors are clustered with donors such that if an acceptor site point overlaps with a donor site point, the average position is retained and relabeled as a dual site point. If an original dual site point (from a water or methanol minimum) overlaps with an acceptor (or donor), the average position is retained and labeled as an acceptor (or donor). The latter clustering occurs because an original dual site point may in effect be acting only as an acceptor or a donor site at that position. The script outputs a site-point file in SPH format readable by PhDOCK or DOCK, and in CRD format file for graphically viewing in Quanta (Molecular Simulations, Inc., 2001, San Diego, CA). The CRD format site-point file can be further pared down manually and then reconverted to the reduced SPH format file for DOCK input with the use of an accompanying awk script.

DOCK Methodology

DOCK is one of the more widely used computational docking programs.[2–14] It systematically attempts to fit each compound from a database into the binding site of the target structure such that three or more of the atoms in the database molecule overlap with a set of predefined site points (or a clique) in the target binding site. Each acceptable orientation of a ligand in the binding site is scored on a grid throughout the macromolecular target with the use of precalculated values for the protein part of the interaction energy. The site points used for matching can be chemically labeled; if, for example, it is known from existing experimental information on bound ligands, or by

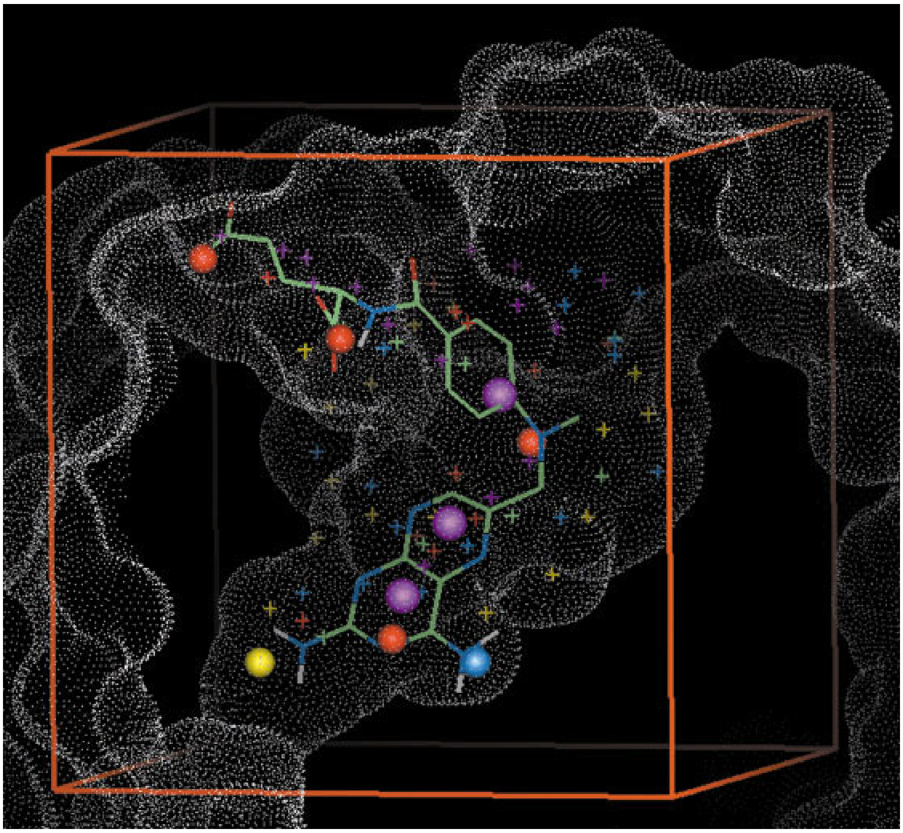


Fig. 3. A set of 75 site points generated with MCSS2SPTS for the DHFR structure. Acceptor sites are shown in red; donor sites, blue; dual sites, which can match both acceptors and donors, yellow; ring centroids, purple; and neutrals, green. Spheres indicate site points that overlay with similar features in MTX bound to DHFR. MTX is shown colored by element: A molecular surface for DHFR is in white, and the box used for the MCSS and subsequent DOCK calculations is in orange. A total of 15 acceptor, 16 donor, 17 dual, 19 centroid, and 8 neutral sites was determined.

examination of the target structure, that a particular site point should only be matched to a ligand donor atom that can be specified. Alternatively, chemically labeled site points can be derived from theoretical pharmacophores calculated for the target structure with the above described method.

PhDOCK Methodology

MCSS-derived site points based on theoretical pharmacophores are expected to yield better results with PhDOCK.[15] In a PhDOCK database, conformers of the same and different molecules are overlaid based on their largest 3D pharmacophore. Pharmacophore points include hydrogen-bond acceptors, donors, duals capable of simultaneously donating and accepting a hydrogen bond, and ring centroids. Each ensemble of conformers is simultaneously docked into the binding site, matching only the pharmacophore points to the site points. The acceptable orientations of a pharmacophore in the binding site of the target (based on a partial match) are used to dock all conformers associated with that pharmacophore. The conformers are scored, and only the best scoring conformer of each molecule is saved in the final hit list. Chemical labeling and critical clustering of site points are fully functional.

Use of MCSS-Derived Site Points

For PhDOCK calculations, acceptor site points match to ligand ensemble pharmacophore points labeled as acceptor or dual. Likewise, site points labeled as donors match to donor or dual pharmacophore points, dual site points can match to dual, acceptor, or donor pharmacophore points, and centroids to centroid (Table I). Neutral site points represent positions in the binding site where a methyl probe has a minimum; if included, these site points can match to any pharmacophore point. MCSS2SPTS-generated site points can be used with DOCK or Ensemble DOCK[27] as well. Because, with standard DOCK, site points match directly to ligand atoms, the ring centroid site points are discarded, and the neutral site points included. In this case, the neutral site points can either be labeled so that they match only carbon atoms, or they can be left unlabeled to match any atom. Furthermore, if a calculated site-point set is to be used to screen a database that has defined aliphatic hydrophobic features, these hydrophobes could be exclusively matched to the "neutral" site points.

Preparation of Biologic Test Systems

We used high-resolution ligand–protein complex X-ray structures as test systems to validate the methodology.

TABLE III. Comparison of Ligand Pharmacophore Features with Site-Point Sets

Ligand–protein complex	No. of ligand pharmacophore points	No. of pharmacophore points overlapping with a matching site point[a]			% within 1.25 Å of a matching site point
		0.75 Å	1 Å	1.25 Å	
MTX-DHFR	15	5	8	10	67%
ENO-MIF	5	2	3	5	100%
CoA-ACPS	26	4	8	12	38%
A79285-HIV1 protease	16	5	8	16	100%

[a]The number of ligand pharmacophore points that are within the stated cutoff of an appropriately labeled site point, where ligand acceptors match to acceptor or dual site points, donors match to donor or dual site points, duals match to acceptor, donor, or dual site points, centroids match to centroids, and all pharmacophore points can match to neutral site points.

The following structures were used: Dihydrofolate reductase (DHFR) complexed with reduced nicotinamide adenine dinucleotide (NADPH) and methotrexate (MTX) solved to 1.7 Å resolution (PDB3DFR)[28]; macrophage migration inhibitory factor (MIF) complexed with hydroxyphenylpyruvate (ENO) at 2.5 Å resolution (PDB1CA7)[29]; acyl carrier protein synthase (ACPS) complexed with coenzyme A (CoA) at 1.5 Å resolution[30]; and human immunodeficiency virus 1 (HIV1) protease complexed with a symmetric difluoroketone-containing inhibitor A79285 at 1.7 Å resolution (PDB1DIF).[31]

In general, all hydrogens were added to Protein Data Bank (PDB) files and, after removing any ligands, the hydrogens were minimized with CHARMm (Molecular Simulations, Inc., 2001, San Diego, CA).[23] For the NADPH cofactor bound to DHFR, partial charges for the MCSS and DOCK calculations were taken from the CHARMm all-atom force field[32] and inferred from similar compounds that were parameterized. For ACPS, after adding polar hydrogens, the entire ACPS–CoA structure was subjected to 100 steps of Powell minimization, with harmonic constraints on the heavy atoms decreasing from 20 to 12 kcal/mol · Å to relieve close contacts between the ligand and the protein. The parameters for CoA were derived from those for DNA. The CoA and Ca^{2+} ions were then removed from the structure.

PhDOCK Searches

A PhDOCK database consisting of the X-ray conformers of the ligands from 1DIF and 23 other HIV1 protease complex structures (listed in Table II) was prepared. Each ligand was translated outside of the 1DIF active site (in cases where there was already an overlap with that structure), hydrogens were added, and atom types and Gasteiger charges were assigned using Sybyl (Tripos Associates, 2000, St. Louis, MO). Pharmacophores were determined for each ligand, in its X-ray conformation, with the use of a Sybyl spl script, as described,[15] and a multi-mol2 format database was generated. The database consists of 28 pharmacophores, each with one associated conformer (4 of the 24 complex structures give two alternative conformers for the bound ligand).

This HIV ligand database was searched, specifying maximum orientations 5000, distance tolerance 0.75 Å, a bump maximum of 3, manual matching with minimum nodes 4 and maximum nodes 15, and chemical matching. These parameters require that all 15-point partial pharmacophores are matched first (if the database pharmacophore has 15 or more features), followed by 14-point ones, and so on, until the maximum number of orientations has been attempted. Energy scoring with 25 steps of rigid-body simplex minimization for each acceptable orientation was done. Scoring grids for the 1DIF structure were calculated using a 0.25 Å grid spacing, and the all-atom Amber force field,[33] with a dielectric constant of 4 and a 12-Å cutoff on the nonbonded terms. After the search, each protein structure was superimposed onto the 1DIF structure, matching only the protein backbone atoms, and the root-mean-squares difference (RMSD) of the resulting ligand position with the top ranked docked position in the 1DIF protein was calculated (Table II).

DOCK Searches

We employed a modified version of DOCK3.5 that allows multiple conformations of each molecule to be docked and retains only the best scoring conformer for each molecule. For the matching step, a distance tolerance of 1.0 Å, nodes minimum and maximum of 4, ligand and receptor binsize of 0.5 Å, ligand and receptor overlap of 0.2 Å, and a bump maximum of 1 was specified. The large distance tolerance (1.0 Å compared to 0.75 Å used for the PhDOCK searches) allows for greater sampling, which may be necessary given that the ACPS site is more open. The bump maximum of 1 (vs. 3) is more restrictive. The binsize and overlap parameters are specific to DOCK3.5 and are not used with PhDOCK or DOCK4.0. We used chemical matching and labeled ligand atoms that could match to acceptor sites (Sybyl atom types N.1, N.2, N.ar, O.3, S.3, O.2, O.co2), donor sites (C.cat, N.am, N.3, N.ar, N.4, O.3, S.3), and dual sites (N.1, N.2, N.am, N.3, N.ar, O.3, S.3, O.2). All other atom types were left unlabeled and therefore could match to any site point. Energy grids were calculated using a 0.3-Å grid spacing, a dielectric constant of 4, and a 12-Å cutoff on the nonbonded terms. For each acceptable orientation, 20 iterations of rigid-body simplex minimization of the energy function were performed. A DOCK3.5 version of the Available Chemicals Directory (ACD) database with, on average, about 20 Catalyst (30 kcal/mol energy window)-generated conformers per molecule and a solvation correc-

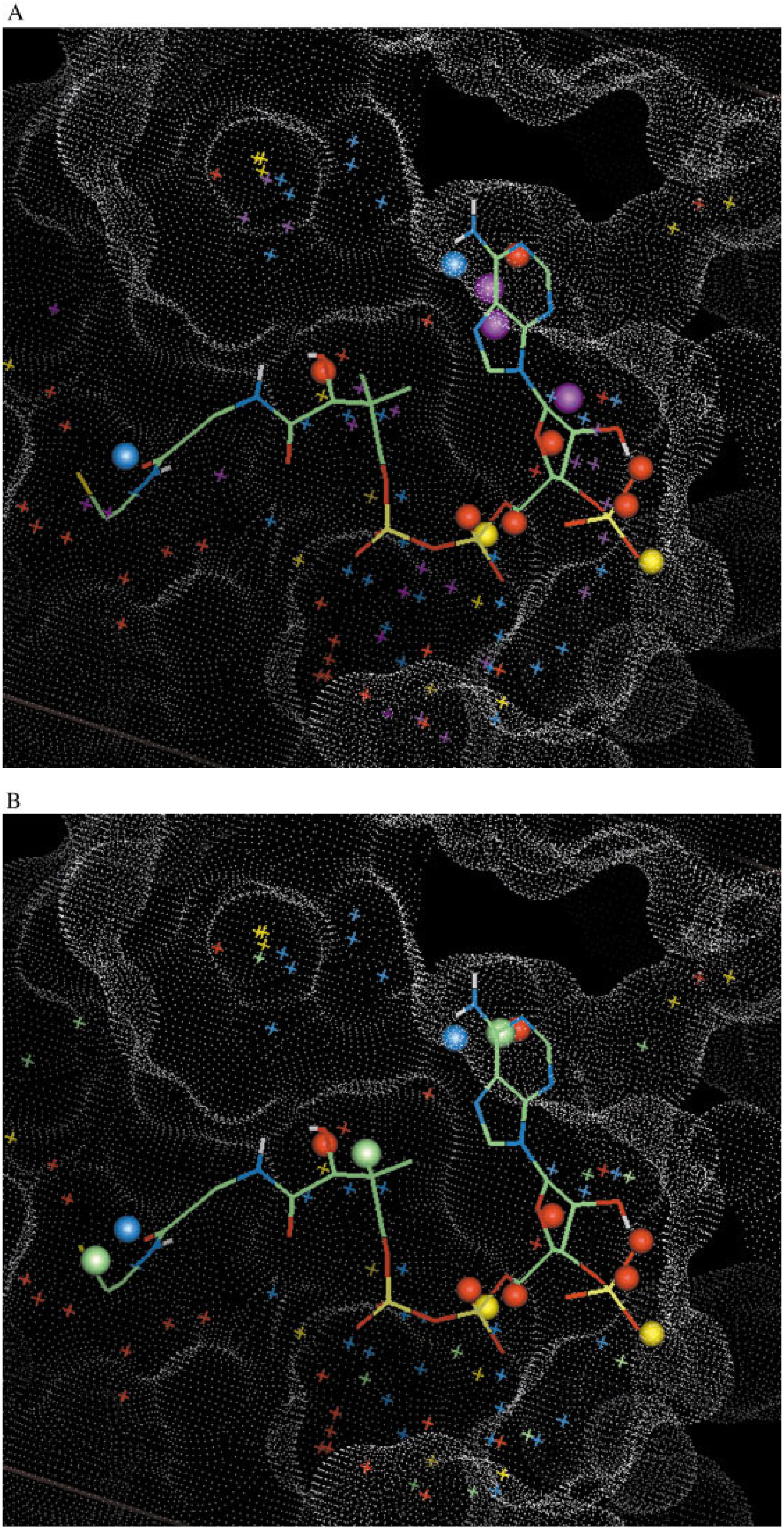


Figure 4.

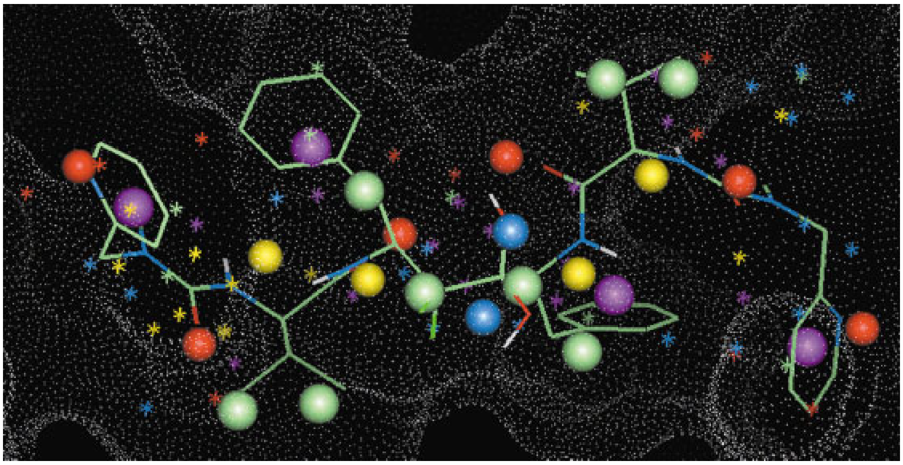


Fig. 5. A set of 90 site points generated with MCSS2SPTS for the HIV1 protease structure. Acceptor sites are shown in red; donor sites, blue; dual sites, yellow; ring centroids, purple; and neutrals, green. Spheres indicate site points that overlay with similar features in A79285 bound to HIV1 protease. A79285 is shown colored by element: A molecular surface for HIV1 protease is in white. A total of 17 acceptor, 19 donor, 17 dual, 21 centroid, and 16 neutral sites was determined.

tion factor for each molecule[34] was screened. All timings are for an SGI R10000, 250 MHz processor.

RESULTS AND DISCUSSION

MCSS2SPTS runs a series of MCSS calculations to determine theoretical pharmacophores for a given target and then automatically generates a chemically labeled site-point set. To test the method, site-point sets were obtained for various targets and compared to bound ligand structures to determine whether the pharmacophoric features of the known ligands were reproduced. If this were the case, PhDOCK (or DOCK) should be able to orient correctly the ligand in the binding site of the target using the site-point set, assuming sufficient conformational sampling in the database and orientational sampling during the docking search. An accurate scoring function would then identify this as the top ranked orientation.

Site-point sets were calculated for DHFR, MIF, ACPS, and HIV1 protease. The DHFR set contains 75 site points and reproduces the features of the bound ligand MTX well (Fig. 3). Of the 15 pharmacophoric centers in MTX, 8 are within 1 Å of an appropriately labeled site point, and 10 are within 1.25 Å (see Table III for additional details). If the neutral site points are not included, 5 (instead of 8) are within 1 Å of an appropriately labeled site point. We previously used this site-point set without the neutrals to orient MTX correctly in the binding site using only a 0.5 Å distance tolerance (during the docking, distances between pairs of appropriately labeled site points are matched).[15] The MIF site-point set is much smaller (35 site points) and reproduces some of the features of the bound ENO ligand. This structure is not nearly as high a resolution as the DHFR structure (2.5 vs. 1.7 Å); therefore, the match is not expected to be as good. The site-point set is, however, sufficient to dock the X-ray conformation of ENO correctly into the MIF structure with the use of a 1 Å distance tolerance.[15] This example illustrates that the method is not only valid for the highest resolution structures. For ACPS, the bound ligand is CoA, which is a very large and flexible ligand (with 17 rotatable bonds). The resulting site-point set reproduces a number of the features of CoA (Fig. 4) and should be sufficient to dock the bound conformation of CoA correctly. The lower percentage of site points matched by CoA may explain its low affinity ($K_m \sim 50\ \mu M$). Many of the functional groups in CoA, while occupying allowed positions, may not be occupying optimal positions. The MCSS2SPTS results for ACPS for use with PhDOCK and standard DOCK are shown in Figure 4(A and B), respectively. The final example site-point set is that for HIV1 protease (Fig. 5). This set also reproduces very well the features of the known ligand A79285. See Table III for a summary the overlap of the ligand pharmacophore features with the site-point set calculated for each of these protein structures.

Because, as a first step, MCSS2SPTS automatically calculates a series of nine MCSS functional group maps for the target, these intermediate results can also be examined for *de novo* design efforts. Benzene minima (Fig. 6) indicate how many of the ring centroid site points for HIV1 protease originate and clearly show that they are pharmacophoric in nature. Figure 7 shows various functional group minima from which the site points in the final set were derived. Typically, MCSS2SPTS generates a manage-

Fig. 4. A set of site points generated with MCSS2SPTS for the ACPS structure. In (**A**), the set of 104 site points with ring centroids included for use with PhDOCK is shown, whereas in (**B**), the set of 90 site points with neutrals sites included instead for use with standard DOCK is shown. Acceptor sites are shown in red; donor sites, blue; dual sites, yellow; ring centroids, purple; and neutrals, green. Spheres indicate site points that overlay with similar features in CoA bound to ACPS. CoA is shown colored by element: A molecular surface for ACPS is in white, and the box used for the MCSS and subsequent DOCK calculations is in orange. A total of 31 acceptor, 33 donor, 13 dual, 27 centroid, and 13 neutral sites was determined.

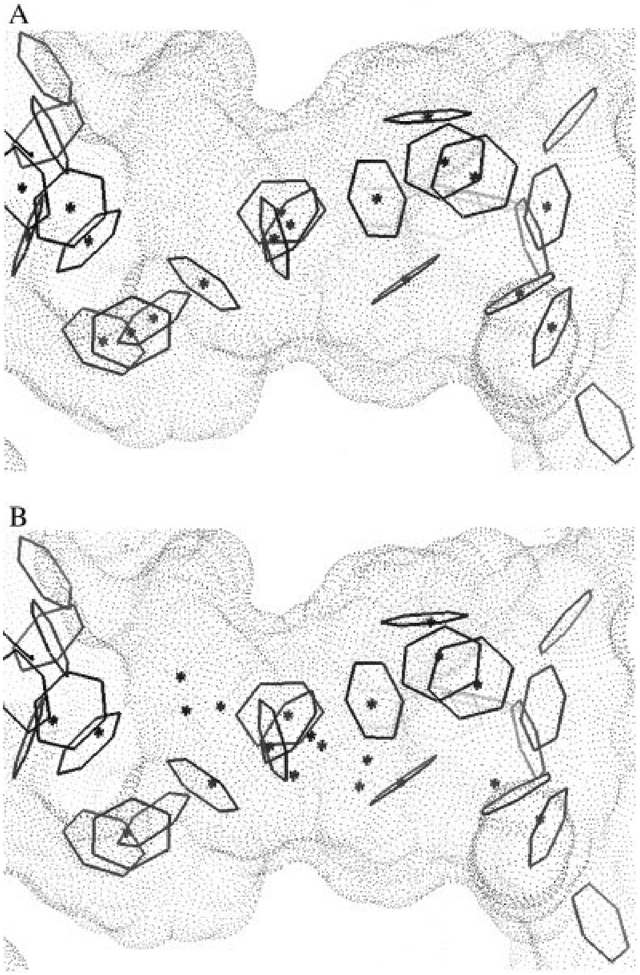


Fig. 6. Benzene minima calculated for the HIV1 protease structure with the use of MCSS are shown in (**A**) superimposed with centroids for those minima well inside the box, and in (**B**) with the ring centroid site points in the final MCSS2SPTS generated set. The ring centroid site points in (**B**) are derived from all ring minima (benzene, oxazole, and cyclohexane). These centroid site points have been clustered, and for each overlapping pair, the centroid originating from the lower energy ring minima is retained.

able number of chemically labeled site points ready for input into DOCK or PhDOCK. By comparison, SPHGEN results often have to be edited manually to select an appropriate number of site points and to include labeling.

The HIV 1 protease site points calculated for the 1DIF structure were compared to the 24 ligands listed in Table II. After we superimposed the other HIV1 protease structures with the 1DIF structure as described in the Methods and Results section, 22 out of 24 ligands had at least 4 site points within 1 Å of a site point. In the actual PhDOCK run, 20 out of 24 ligands were docked. Of these 20, 19 docked into the binding site; the other one had parts of the ligand protruding into what would be solvent. Two of the 19 ligands were docked in the binding site, overlapping well with the position of the 1DIF ligand and making reasonable interactions with the 1DIF protein, but their orientations were flipped relative to their X-ray positions (i.e., 1 hsg and 1 hpv). Given that these are highly symmetric ligands binding to a symmetric site, this is not that surprising. In fact, in four of the X-ray structures alternate, flipped orientations for the ligand in the pocket are observed experimentally (see Table II). Figure 8 shows an example of one of these ligands for which both X-ray conformers are correctly docked. Using the 1DIF site-point set, PhDOCK was able to dock correctly 16 of the 24 ligands listed in Table II (see Fig. 9).

Of the 4 ligands that were not docked into the site (with a negative score), the 1HTF ligand in either X-ray conformation would clash with the side chain of Arg 8 in the 1DIF structure, as would the 4HVP ligand. Similarly, the 1HXB ligand would clash with Ile 84 in the 1DIF protein, and one of the two X-ray conformations of 1HTG would make a close contact with Asp 30. (Because the other conformation of 1HTG is correctly docked, that ligand is not counted among the 4 ligands that are incorrectly docked.) Although the 1ODW ligand does not make any bad contacts with the 1DIF protein, it does not make as many favorable van der Waals contacts with the 1DIF protein as with its own protein structure; the side chains of Ile 47 and 84 are shifted away from the ligand relative to their positions in the 1ODW protein structure. The fact that 1ODW did not dock, however, is due to a lack of sampling; when the maximum number of orientations is increased from 5000 to 10,000, the ligand does dock correctly. For the ligand that only partially docked into the binding site, 1HVR, in its X-ray position, close contacts with Lys 45 and Asp 30, respectively, would result with the 1DIF protein structure. If this database of HIV1 protease ligands was suitably conformationally expanded, for the 6 ligands that did not correctly dock, it is possible that a conformation close to the X-ray one might; that is, in some cases, allowing conformational flexibility in the ligand can compensate for not allowing it in the protein structure.

We carried out a multiconformation DOCK (version 3.5) search for ACPS using the set of 90 MCSS2SPTS-generated site points [Fig. 4(B)] with the neutrals unlabeled. We also conducted the same search using 11 CoA ligand atom positions, 4 acceptors, 1 donor, and 6 unlabeled neutral site points. A comparison of the energies of the top 50 hits (Fig. 10) shows that the distribution of scores is nearly identical using either site-point set. Therefore, the effective sampling is similar overall. Our search with the MCSS-derived site points took 62.9 h of central-processing-unit (CPU) time compared to 44.3 h for the search using the 11 ligand atom positions. Although the latter was faster, a bound ligand structure, providing a known solution, does not always exist, and an optimal set of ligand atom positions may not be chosen for the site-point set. When all 48 of the CoA heavy atoms were used as site points (a set consisting of 18 acceptors, 3 donors, 2 duals, and 25 unlabeled neutrals), the search was prohibitively long (729.2 CPU h). This example clearly shows that MCSS-derived site points (with neutral points substituted for centroids) can be used with standard DOCK to obtain favorable results. A number of examples illustrating the use of MCSS-derived site-point sets with PhDOCK are described.[15] Furthermore, a number of large-scale screens

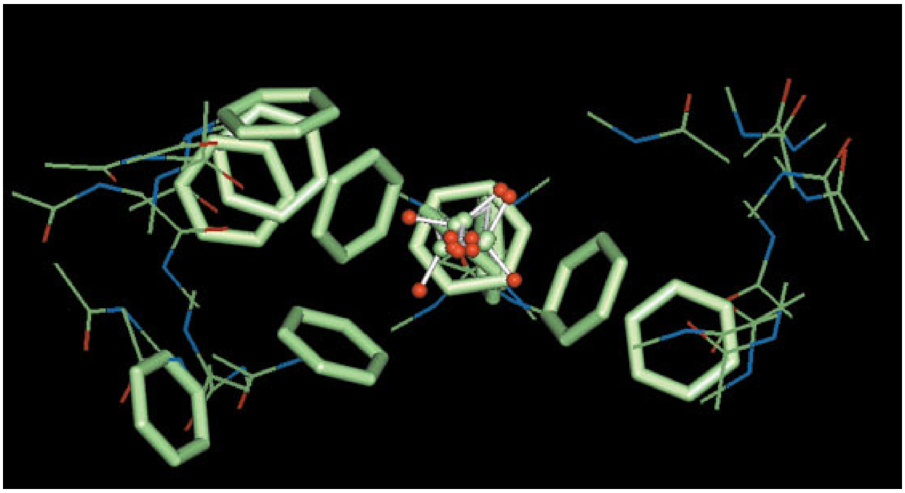


Fig. 7. The minima for acetate, *N*-methylacetamide, and benzene selected by MCSS2SPTS for HIV1 protease. Site points were subsequently extracted from these minima and clustered together with those from the other six groups to determine the final site-point set for HIV1 protease. Atoms are colored by element type. Benzene minima are shown in a liquorice model, acetate in ball-and-stick representation, and *N*-methylacetamide in thin lines.

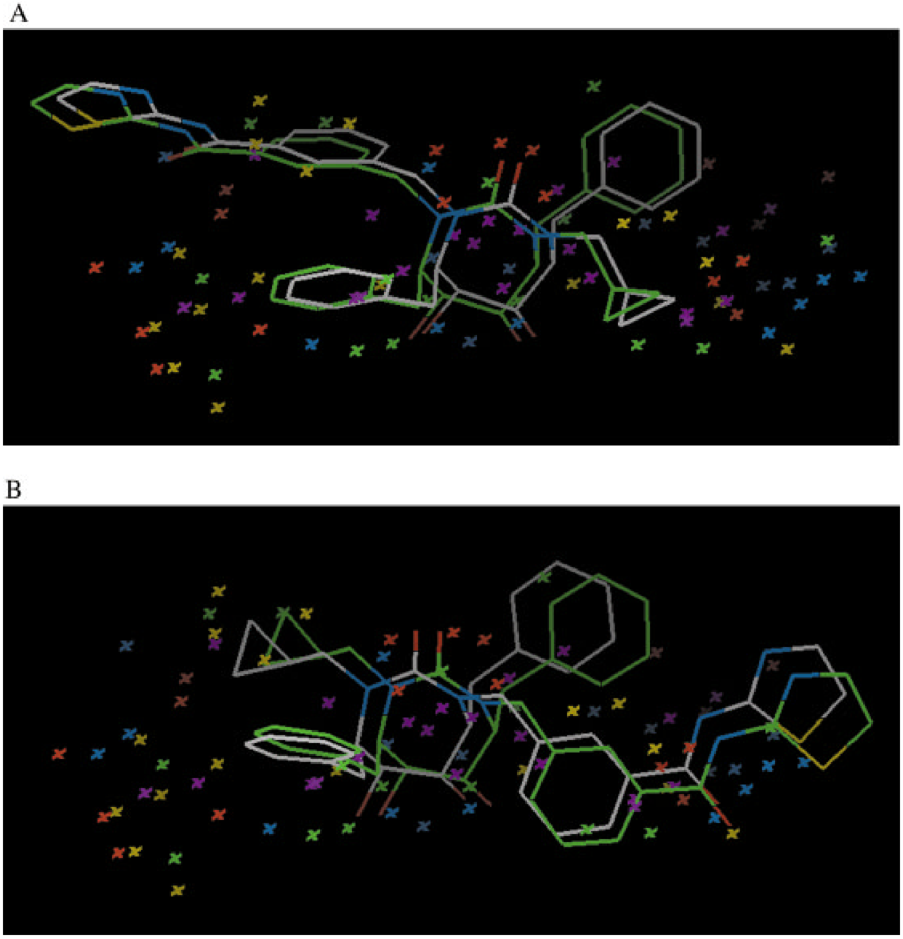


Fig. 8. Superposition of the X-ray position in the 1QBU structure and the docked position in the 1DIF structure for conformer 1 (**A**) and conformer 2 (**B**) of the ligand. The X-ray conformers are shown colored by element, with carbons colored green, whereas the docked conformers are shown with carbons colored grey.

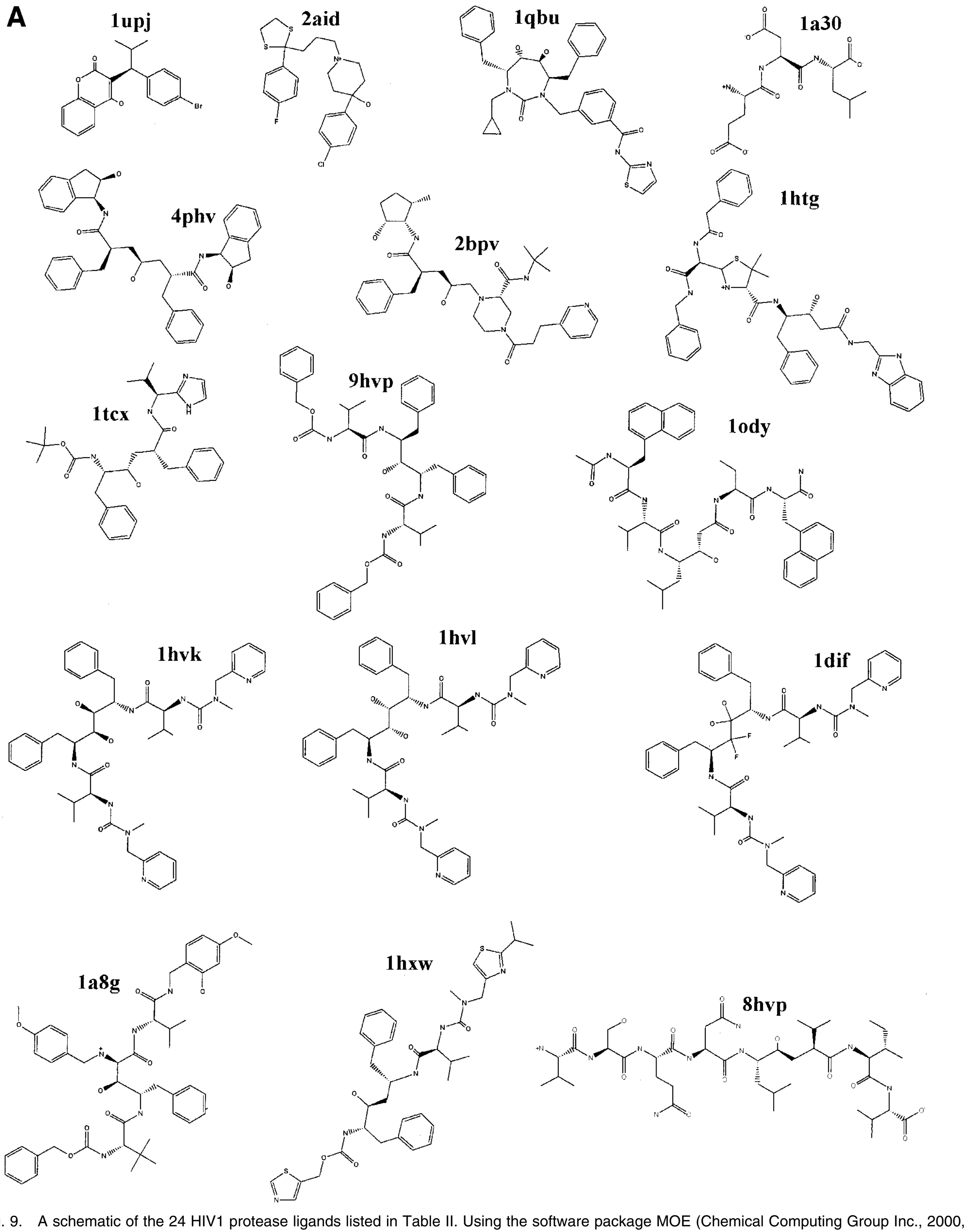


Fig. 9. A schematic of the 24 HIV1 protease ligands listed in Table II. Using the software package MOE (Chemical Computing Group Inc., 2000, Montreal, Canada), we employed the MACCS structural key fingerprint for each ligand to cluster the ligands by similarity; a Tanimoto coefficient of 0.70 for similarity and overlap resulted in 17 clusters, including 11 singletons, indicating that the ligands are fairly diverse. The 16 (out of the 24) ligands that docked correctly (RMSD <1.75 Å) to the 1DIF structure are shown in bold in (**A**), whereas those that did not dock correctly are shown in (**B**).

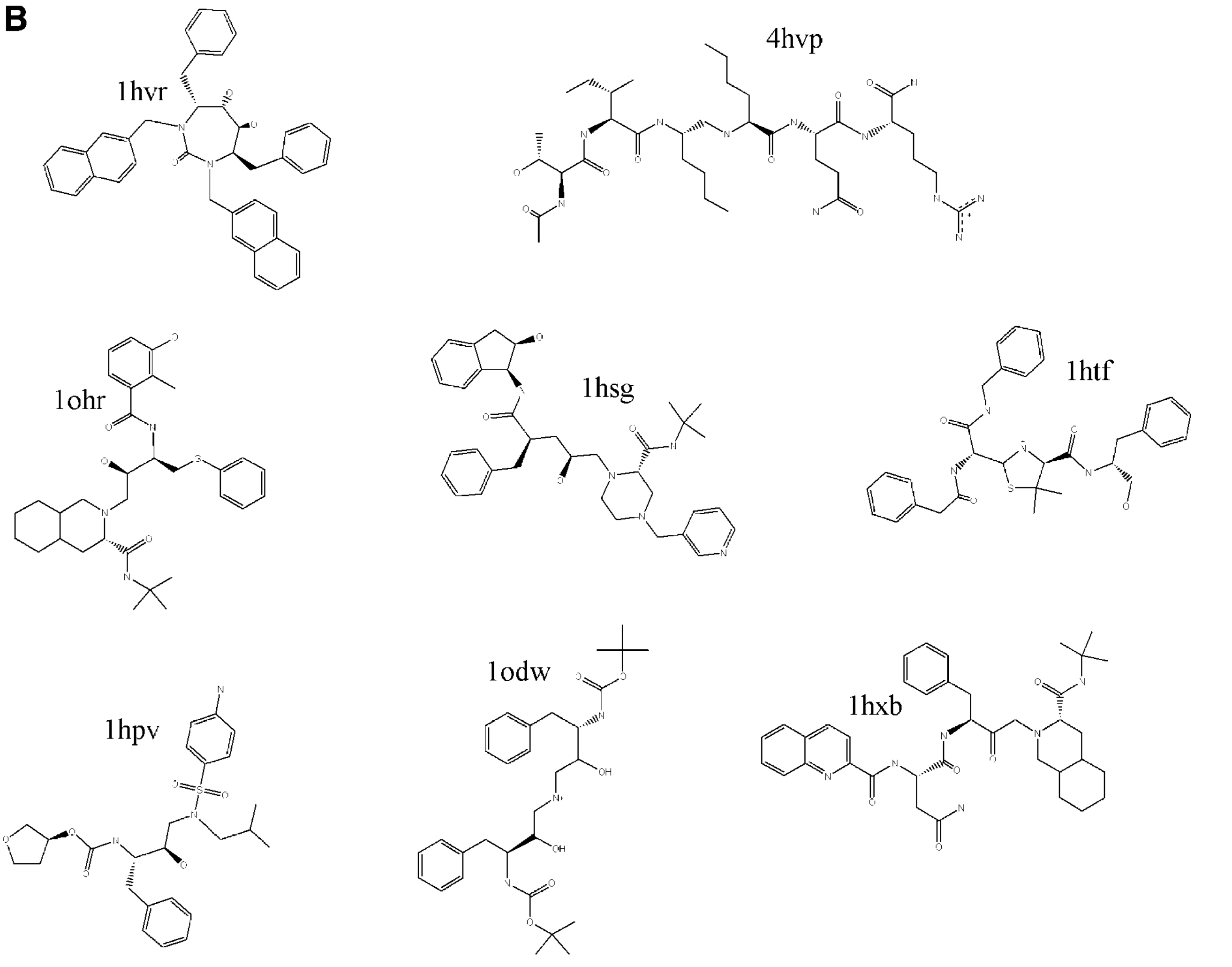


Figure 9. (Continued.)

of the entire PhDOCK ACD have been carried out for proprietary targets with the use of MCSS2SPTS site points; those screens for which top ranking compounds have been experimentally tested have produced micromolar hits (unpublished data).

In the future, as even more structural information becomes readily available, the use of the target structure to screen large databases of compounds and virtual libraries will become increasingly important in the drug discovery process. Large virtual libraries will be constructed based on available chemistry or a set of existing combinatorial scaffolds, as well as known drug properties. Improved scoring functions, faster computers, and better database storage methods will facilitate progress. The MCSS2SPTS approach is a step toward efficiently addressing the sampling issue. Effectively, the space to be sampled is reduced or focused during the setup of the search (prior to running) by calculation of the pharmacophoric site points, or "hot spots," and then during the run by only matching ligand pharmacophore features to them.

Clearly, once the issue of sampling has been adequately addressed, the success of docking molecules into a target site is dependent on the accuracy of the scoring function that ranks the compounds; that is, it is ultimately dependent on how well the corresponding relative binding affinities can be predicted. A next step with MCSS2SPTS and PhDOCK will be to include solvation corrections in the scoring and possibly to incorporate some target flexibility. Furthermore, it should be possible to extend this method to examine families of proteins. Work is in progress to use the pharmacophoric site points that are common to a family or subfamily of protein structures to screen large databases for compounds likely to bind to that family. At this time, MCSS2SPTS and PhDOCK together provide an efficient and chemically sensible approach for screening large 3D molecular databases.

ACKNOWLEDGMENTS

Our thanks to Bert Thomas IV, Thomas Rush, Eric Manas, and Ken Foreman for helpful comments and initial testing of the method, and to Iain McFadyen for help processing the HIV1 protease ligands.

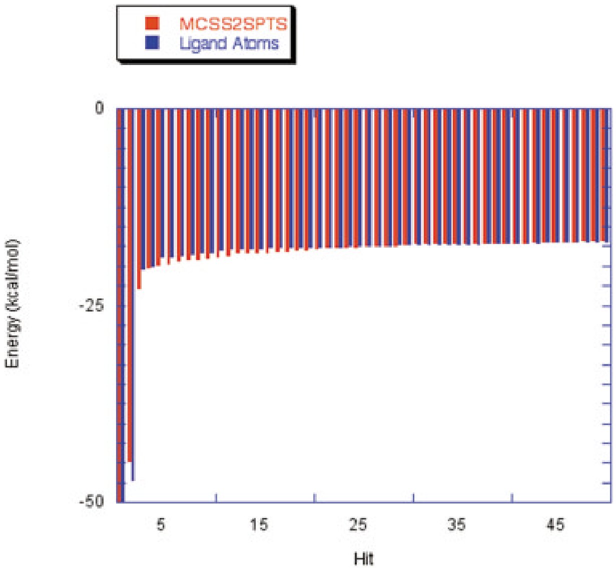


Fig. 10. A histogram comparing the energies [kcal/mol] of the top 50 hits from the multi-conformation DOCK search with the use of bound ligand atom positions as site points (red) with those using the MCSS2SPTS set (blue).

REFERENCES

1. Houston JG, Banks M. The chemical–biological interface: Developments in automated and miniaturised screening technology. Curr Opin Biotech 1997;8:734–740.
2. Gordon K, Balasubramanian, S. Recent advances in solid-phase chemical methodologies. Curr Opin Drug Discovery Dev 1999;2: 342–349.
3. Suto MJ. Developments in solution-phase combinational chemistry. Curr Opin Drug Discovery Dev 1999;2:377–384.
4. Andrade MA, Sander C. Bioinformatics: From genome data to biological knowledge. Curr Opin Biotech 1997;8:675–683.
5. Westhead DR, Thornton JM. Protein structure prediction. Curr Opin Biotech 1998;9:383–389.
6. Onuchic JN, Luthey-Schulten Z, Wolynes PG. Theory of protein folding: The energy landscape perspective. Ann Rev Phys Chem 1997;48:545–600.
7. Dunbrack RL, Gerloff DL, Bower M, Chen XW, Lichtarge O, Cohen FE. Meeting review: The Second Meeting on the Critical Assessment of Techniques for Protein Structure Prediction (CASP2), Asilomar, CA, December 13–16, 1996. Fold Des 1997;2: R27–R42.
8. Rost B. Marrying structure and genomics. Structure 1998;6:259–263.
9. Kuntz I. Structure-based strategies for drug design and discovery. Science 1992;257:1078–1082.
10. Walter WP, Stahl MT, Murcko MA. Virtual screening—an overview. Drug Discov Today 1998;3:160–178.
11. Joseph-McCarthy D. Computational approaches to structure-based ligand design. Pharmacol Therapeut 1999;84:179–191.
12. Kuntz ID, Blaney JM, Oarley SJ, Langridge R, Ferrin TE. A geometric approach to macromolecule–ligand interactions. J Mol Biol 1982;161:269–288.
13. Meng EC, Shoichet BK, Kuntz ID. Automated docking with grid-based energy evaluation. J Comput Chem 1992;13:505–524.
14. Shoichet BK, Stroud RM, Santi DV, Kuntz ID, Perry KM. Structure-based discovery of inhibitors of thymidylate synthase. Science 1993;259:1445–1450.
15. Joseph-McCarthy D, Thomas BE IV, Belmarsh M, Moustakas D, Alvarez JC. Pharmacophore-based molecular docking to account for ligand flexibility. Proteins 2003;51:172–188.
16. Ewing T, Makino S, Skillman A, Kuntz I. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. J Comput Aided Mol Des 2001;15:411–428.
17. Miranker A, Karplus M. Functionality maps of binding sites: A multiple copy simultaneous search method. Proteins 1991;11:29–34.
18. Evensen E, Joseph-McCarthy D, Karplus M. MCSSv2. 2.1 ed. Cambridge, MA: Harvard University; 1997.
19. Joseph-McCarthy D. Structure-based combinatorial library design and screening: Application of the Multiple Copy Simultaneous Search method. In: Ghose AK, Viswanadhan VN, editors. Combinatorial library design and evaluation: principles, software tools, and applications in drug discovery. New York: Marcel Dekker; 2001. p 503–529.
20. Goodford P. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem 1985;28:849–857.
21. Bohm HJ. Ludi-rule-based automatic design of new substituents for enzyme-inhibitor leads. J Comput Aided Mol Des 1992;6:593–606.
22. Bohm HJ. On the use of Ludi to search the Fine Chemicals Directory for ligands of proteins of known 3-dimensional structure. J Comput Aided Mol Des 1994;8:623–632.
23. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 1983;4: 187–217.
24. Shuker SB, Hajduk PJ, Meadows RP, Fesik SW. Discovering high-affinity ligands for proteins: SAR by NMR. Science 1996;274: 1531–1534.
25. Joseph-McCarthy D, Fedorov AA, Almo SC. Comparison of experimental and computational functional group mapping of an RNase A structure: Implications for computer-aided drug design. Protein Eng 1996;9:773–780.
26. Allen KN, Bellamacina CR, Ding XC, 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.
27. Lorber DM, Shoichet BK. Flexible ligand docking using conformational ensembles. Protein Sci 1998;7:938–950.
28. Bolin JT, Filman DJ, Matthews DA, Hamlin RC, Kraut J. Crystal structures of Escherichia coli and Lactobacillus casei dihydrofolate reductase refined at 1.7 Å resolution: I. General features and binding of methotrexate. J Biol Chem 1982;257:13650–13662.
29. Lubetsky JB, Swope M, Dealwis C, Blake P, Lolis E. Pro-1 of macrophage migration inhibitory factor functions as a catalytic base in the phenylpyruvate tautomerase activity. Biochemistry 1999;38:7346–7354.
30. Parris KD, Lin L, Tam A, Mathew R, Hixon J, Stahl M, Fritz CC, Seehra J, Somers WS. Crystal structures of substrate binding to *Bacillus subtilis* holo-(acyl carrier protein) synthase reveal a novel trimeric arrangement of molecules resulting in three active sites. Struct Fold Des 2000;8:883–895.
31. Silva AM, Cachau RE, Sham HL, Erickson JW. Inhibition and catalytic mechanism of HIV-1 aspartic protease. J Mol Biol 1996;255:321–340.
32. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 1998;102:3586–3616.
33. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A 2nd generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules. J Am Chem Soc 1995;117:5179–5197.
34. Shoichet BK, Leach AR, Kuntz ID. Ligand solvation in molecular docking. Proteins 1999;34:4–16.