A Comprehensive Docking Study on the Selectivity of Binding of Aromatic Compounds to Proteins

Csaba Hetényi,*,†,‡ Uko Maran,† and Mati Karelson†

Department of Chemistry, Tartu University, 2 Jakobi Street, 51014 Tartu, Estonia, and Department of Medical Chemistry, University of Szeged, Dóm tér 8, 6720 Szeged, Hungary

Received March 19, 2003

Generally, computer-aided drug design is focused on screening of ligand molecules for a single protein target. The screening of several proteins for a ligand is a relatively new application of molecular docking. In the present study, complexes from the Brookhaven Protein Databank were used to investigate a docking approach of protein screening. Automated molecular docking calculations were applied to reproduce 44 protein−aromatic ligand complexes (31 different proteins and 39 different ligand molecules) of the databank. All ligands were docked to all different protein targets in altogether 12 090 docking runs. Based on the results of the extensive docking simulations, two relative measures, the molecular interaction fingerprint (MIF) and the molecular affinity fingerprint (MAF), were introduced to describe the selectivity of aromatic ligands to different proteins. MIF and MAF patterns are in agreement with fragment and similarity considerations. Limitations and future extension of our approach are discussed.

INTRODUCTION

X-ray crystallography and nuclear magnetic resonance spectroscopy (NMR) measurements are undoubtedly the most reliable sources of high-resolution structures of protein−ligand complexes. Despite the increase of companies and research laboratories that carry out experimental elucidation of structures of macromolecular complexes of biological interest (the number of determined protein molecules is rapidly increasing[1]), in silico drug screening and design techniques remain indispensable tools of in vitro or in vivo high throughput screening (HTS) and design methods. Using experimentally determined ligand−protein complexes as references, computational docking is one of the most important in silico HTS methods[2] and can easily be combined with combinatorial chemistry.[3] Moreover, even if knowledge of the binding site of the ligand molecule is missing, recently introduced computational approaches can be applied for scanning entire macromolecular targets with small[4] or larger, flexible ligands[5−7] (*blind docking*). Further information on molecular docking and its applications can be found in refs 8−10.

The most important components of each docking algorithm are the search method and the scoring function. Scoring functions provide fast binding energy calculation during the minimum search of docking simulations. Scoring formulas can be applied independently to estimate binding free energies, if there is not appropriate computational capacity for free energy calculation from e.g. MD trajectories[11] or there are too many systems to be calculated. In the present study, the scoring of AutoDock 3.0[12] was applied. This scoring function is based on the Lennard-Jones and screened Coulombic terms of the 2.4 version. The original terms were scaled, and additional solvation and torsional considerations were introduced to obtain a better fit to binding free energy values. However, it should be remarked, that the absolute values of the estimated free energies are obviously not error-free, e.g. because the experimental binding constants involve some uncertainty and the approximations of solvation effects, etc. has limited power, as well. Thus, in the present study, the use of relative (subtracted) binding free energy values was preferred.

Rapid calculation of correct binding geometries and estimation of the conjugated binding free energies are essential not only solely in “traditional” HTS applications, i.e., if thousands of drug candidates are scanned for the same protein target, but also when several proteins or (more correctly) the different binding pockets are screened for the same ligand molecule.

The latter application of docking was used for prediction of drug side effect and toxicity by Chen et al.[13] In their study, docking simulations of drug molecules were used to select the proteins, which might play important role in the biochemical pathways of side effects. A large set of protein binding pockets was used as a basis of the selection. A successful application of AutoDock was reported[14] for virtual protein screening and drug side effect prediction, as well. However, screening of several proteins (instead of ligands) is a relatively new direction in docking studies. Further applications are required to test the approach and find the solution for the problems outlined in the aforementioned papers.

In the present work, a series of protein−aromatic ligand complexes was selected to investigate the efficiency of docking and scoring for selection of appropriate protein(s) for aromatic ligands. The reliability of the scoring (free energy) function and search method of AutoDock 3.0 was verified by structural match of the lowest energy conformers

* Corresponding author phone: +372-7-375254; fax: +372-7-375264; e-mail: csabahete@yahoo.com. Corresponding author address: Department of Chemistry, Tartu University, 2 Jakobi Street, 51014 Tartu, Estonia.

† Tartu University.

‡ University of Szeged.

10.1021/ci034052u CCC: $25.00 © 2003 American Chemical Society
Published on Web 07/11/2003

of the docking experiments (jobs) to native crystallographic structures of ligand molecules of all complexes.

Most of the ligand molecules are structurally similar benzene derivatives (some naphthalene and indole compounds were also involved in the study) and, therefore, fragment considerations and structural similarities allow qualitative verification of the results. Molecular interaction and molecular affinity fingerprints (MIF and MAF, respectively) were introduced to get a comprehensive picture on the selectivity of binding of the aromatic molecules on proteins.

METHODS

Docking. Forty-four complexes of 39 different, small, and middle-size aromatic ligand molecules and 31 different protein molecules (Table 1) were selected from the Brookhaven Protein Databank[15] (PDB), partially with the aid of PDBsum[16] server. All selected complexes are free of close contacts, and the ligand is not buried inside the protein by the side-chains, i.e., only systems with no or moderate induced fits were considered. The latter criterion is important, as the applied docking method handles only rigid target protein molecules and therefore no induced effects can be modeled. Moreover, systems with binding sites containing several water molecules around the ligand were omitted, as those molecules can be involved in specific binding of some ligands.[5,17,18] Thus, if crystallographic water molecules were constrained at the site and involved in docking calculations, the binding site of a certain protein would be appropriate only for docking of its original ligand molecule or very similar molecules.

Ligand and water molecules and all ions were removed from the original PDB file. If the PDB file contained several identical chains, the one bound to the ligand molecule of the lowest B-factors was selected. Essential hydrogen atoms, charges of the Kollman united atom type, and solvation parameters were added to the residues of the protein chain. Generally, the native ligand molecules were used, as their bond lengths and angles were found adequate for this purpose. In the case of systems **4d**, **7a**, **15**, and **32**, the ligand molecules were model-built in MOLDEN[19] and optimized with the aid of TINKER[20] using a modified MM3 force field. Babel[21] and VEGA[22] programs were used for file manipulations. Ligand molecules were equipped with all hydrogen atoms and Gasteiger−Marsili charges.[23] Autotors (an AutoDock[12] tool) was applied for creation of united atom representation and definition of the torsions of the ligands. A uniform procedure was applied for all the ($31 \times 39 =$) 1209 docking jobs. Affinity (grid) maps of 60 grid points in each Cartesian directions and 0.375 Å spacing were generated with the aid of Autogrid.[12] Maps were centered on the original ligand molecules. AutoDock parameter set and distance dependent dielectric function were used in the calculation of the van der Waals and the electrostatic terms, respectively. Docking simulations were performed using the Lamarckian genetic algorithm and the Solis & Wets local search method of Autodock. All torsion angles of the ligand molecules (except of amide bonds and some conjugated or rigid bonds) were released during docking. The numbers of released torsions (RT) of the ligands are listed in Table 1. Initial position, orientation, and torsions of the ligands were set randomly. During the search, translational step of 0.2 Å, quaternion, and torsion steps of 5° were applied. A population of 50 members and a maximum number of 2.5 million energy evaluations were used. Ten docking runs were performed for each job.

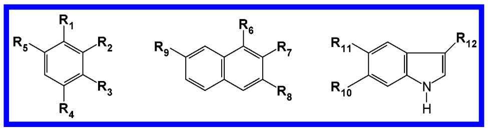
Evaluation of the Results. A C program was used for the evaluation and RMSD (Root Mean Square Deviation) calculation of all data. Binding free energies of each job were collected and the minima were selected. RMSD was calculated for the resulted 10 structures using ligand structures of minimum energies (in Table 2: crystallographic structures) at each job as references. A 2.5 Å tolerance was used to form clusters of the closest structures. The atoms of groups of C_{2v} and C_{3v} symmetries were considered identical during RMSD calculations. Average energies of the clusters of each job were calculated and collected in the $\mathbf{E_1}$ (39×31) data matrix. Standard deviations and number of cluster members were calculated and collected too. An additional $\mathbf{E_2}$ matrix was produced using a distance criterion, as follows. The distances between the centers of all resulting structures and that of the crystallographic (reference) ligand were calculated for each run. If the distance between the center of the structure of minimum energy of a job and that of the reference was smaller than the length of the native ligand (limit) of the protein, then the current member of the matrix $\{e_2\}$ was set equal to $\{e_1\}$. Otherwise, the structure with the smallest distance was selected as a reference and a new $\{e_2\}$ was calculated. If all distances of the 10 structures were beyond the aforementioned limit, then the energy value was declared undefined for the actual $\{e_2\}$. (All matrices of this study are available upon request.)

VMD[24] and Raster-3D[25] programs were used for visualization and presentation of the results.

RESULTS AND DISCUSSION

Docking. The results of docking calculations of 44 protein−aromatic ligand complexes (Table 1) is presented in Table 2. Matches of some docked ligands to the crystallographic structure with various RMSD values are depicted on Figure 1. For 40 systems, the energy minimum structure obtained as a result of 10 runs (one docking job) was the closest to the crystallographic position of the ligand. In four cases, the second best ranked structure was the closest to the crystallographic structure. These molecules are marked with letter b in the RMSD(m) column. The RMSD of the energy minimum conformations from the original crystal position was less than 1 Å in 55%, less than 2 Å in 82%, and less than 3 Å in 100% of cases. Summarily, in 82% of the cases good fit was obtained, in the remaining cases the result was acceptable. The distribution of the results is similar (67%; 91%; and 100%) considering the average RMSD values calculated for the groups of structures having RMSD less than 2.5 Å ($\mathrm{RMSD}_{2.5}$ rank). The fraction of dockings of good match increased to 91% after ranking. In 86% of all systems, the number of docked ligand conformations in the $\mathrm{RMSD}_{2.5}$ rank was between 5 and 10, i.e., more than 50% of the runs of a job matched the crystal structure (including the minima of all but four jobs). This result indicates that jobs of 10 runs are adequate to get correct docking results for the systems studied by using the procedure described in the Methods section. The calculated AutoDock minimum

Table 1. Investigated Protein−Ligand Systems Ordered and Numbered According to the Increasing Molecular Weight (MW) of the Ligands[a]



	protein				ligand		
ID	name	PDB code	binding site residues	res. (Å)	-R groups	RT	MW
1	insulin	1mpj	ACHIL	2.30	$\mathbf{R_1}$: $-OH$	1	94.1
2	insulin	1ev3	ACHIL	1.78	$\mathbf{R_1}$: $-OH$ $\mathbf{R_3}$: $-CH_3$	1	108.1
3	insulin	1qiz	ACHIL	2.00	$\mathbf{R_1, R_3}$: $-OH$	2	110.1
4a	transcription factor malt domain III	1hz4	HML	1.45	$\mathbf{R_1}$: $-COO^{(-)}$	0	121.1
4b	chloroperoxidase T	1a8u	FHLMSW	1.60	$\mathbf{R_1}$: $-COO^{(-)}$	0	121.1
4c	human peroxiredoxin	1hd2	CFILPRT	1.50	$\mathbf{R_1}$: $-COO^{(-)}$	0	121.1
4d	bacterial cocaine esterase	1ju4	FWY	1.63	$\mathbf{R_1}$: $-COO^{(-)}$	0	121.1
5a	*β*-trypsin	3ptb	C*DQSVY	1.70	$\mathbf{R_1}$: $-C(NH_2)_2^{(+)}$	0	121.2
5b	urokinase-type plasminogen activator	1f5k	C*DSV	1.80	$\mathbf{R_1}$: $-C(NH_2)_2^{(+)}$	0	121.2
6	bovine trypsin	1tnj	C*DV	1.80	$\mathbf{R_1}$: $-(CH_2)_2-NH_3^{(+)}$	3	122.2
7a	beta-acrosin from RAM spermioza	1fiw	C*DQST	2.10	$\mathbf{R_1}$: $-C(NH_2)_2^{(+)}$ $\mathbf{R_4}$: $-NH_2$	0	136.2
7b	human coagulation factor IXA	1rfn	C*DS	2.80	$\mathbf{R_1}$: $-C(NH_2)_2^{(+)}$ $\mathbf{R_4}$: $-NH_2$	0	136.2
8	prostaglandin H2 synthase-1	1pth	AILRVY	3.40	$\mathbf{R_1}$: $-COO^{(-)}$ $\mathbf{R_2}$: $-OH$	1	137.1
9	esterolytic and amidolytic 43C9 antibody (immunoglobulin)	43ca	FHQY	2.30	$\mathbf{R_1}$: $-OH$ $\mathbf{R_4}$: $-NO_2$	1	139.1
10	insulin	1tym	AC*IL	1.90	$\mathbf{R_1}$: $-OH$ $\mathbf{R_4}$: $-NH-CO-CH_3$	3	151.2
11	poly (ADP-ribose) polymerase	3pax	HSY	2.40	$\mathbf{R_1}$: $-CO-NH_2$ $\mathbf{R_3}$: $-O-CH_3$	3	151.2
12	phenylalanyl-tRNA synthetase	1b70	AEFHQRSVW	2.70	$\mathbf{R_1}$: $-CH_2-CH[COO^{(-)}]-NH_3^{(+)}$	4	165.2
13	protocatechuate 3,4-dioxygenase	3pcn	HIPRTWY	2.40	$\mathbf{R_1}$: $-OH$ $\mathbf{R_3}$: $-CH_2-COO^{(-)}$ $\mathbf{R_5}$: $-OH$	4	167.1
14	human serum albumin	1e7a	C*FILNV	2.20	$\mathbf{R_1}$: $-OH$ $\mathbf{R_2, R_5}$: $-C_3H_7$	3	178.3
15	macrophage migration inhibitory factor (MIF)	1ca7	FIKMNPSVWY	2.50	$\mathbf{R_1}$: $-OH$ $\mathbf{R_4}$: $-CH_2-CO-COO^{(-)}$	3	179.1
16	des-(Ile318-Arg417)-tyrosyl-tRNA synthetase	4ts1	DLQTY	2.50	$\mathbf{R_1}$: $-OH$ $\mathbf{R_4}$: $-CH_2-CH[COO^{(-)}]-NH_3^{(+)}$	5	181.2
17	aromatic amino acid transferase	2ay5	DFILNRSTWY	2.40	$\mathbf{R_{12}}$: $-(CH_2)_2-COO^{(-)}$	3	188.2
18	N1G9 FAB fragment	1ngp	HKRSWY	2.40	$\mathbf{R_1}$: $-NO_2$ $\mathbf{R_3}$: $-CH_2-COO^{(-)}$ $\mathbf{R_5}$: $-OH$	4	196.1
19	prostaglandin H2 synthase-1	1eqg	AILRVYW	2.61	$\mathbf{R_1}$: $-CH(CH_3)COO^{(-)}$ $\mathbf{R_4}$: $-C(CH_3)_3$	3	205.3
20	protein tyrosine phosphatase 1B	1c85	CDFIKQSVY	2.72	$\mathbf{R_1}$: $-COO^{(-)}$ $\mathbf{R_2}$: $-NH-CO-COO^{(-)}$	2	207.1
21	carboxypeptidase A	1hdu	AEHIR	1.75	$\mathbf{R_1}$: $-CH_2-CH[COO^{(-)}]-NH-CO-NH_2$	5	207.2
22	tyrosine phosphatase	1d1q	CDFHLRW	1.70	$\mathbf{R_1}$: $-NO_2$ $\mathbf{R_4}$: $-OPOO_2^{(2-)}$	2	217.1
23	carboxypeptidase A	3cpa	ADEINRTY	2.00	$\mathbf{R_1}$: $-CH_2-CH[COO^{(-)}]-NH-CO-CH_2-NH_3^{(+)}$ $\mathbf{R_4}$: $-OH$	7	238.2
24	influenza virus B/LEE/40 neuraminidase (sialidase)	1ivb	DERY	2.40	$\mathbf{R_2}$: $-OH$ $\mathbf{R_3}$: $-NH-CO-CH_3$ $\mathbf{R_4}$: $-NO_2$ $\mathbf{R_5}$: $-COO^{(-)}$	4	239.2
25	protein tyrosine phosphatase 1B	1c83	ACDFIKQVY	1.80	$\mathbf{R_{10}}$: $-NH-CO-COO^{(-)}$ $\mathbf{R_{11}}$: $-COO^{(-)}$	3	246.2
26	protein tyrosine phosphatase 1B	1c84	ACDFIKQVY	2.35	$\mathbf{R_7}$: $-NH-CO-COO^{(-)}$ $\mathbf{R_8}$: $-COO^{(-)}$	3	257.2
27	cellobiohydrolase I	1dy4	DEHQRSTWY	1.90	$\mathbf{R_6}$: $-O-CH_2-(S)CH(OH)-CH_2-NH-CH(CH_3)-CH_3$	7	259.3
28	streptavidin	1sri	ALSVWY	1.65	$\mathbf{R_1}$: $-COO^{(-)}$ $\mathbf{R_2}$: $-N{=}N-Ph(4\text{-}OH; 3,5\text{-}diMe)$	4	271.3
29	indole-3-glycerolphosphate synthase	1a53	EFKLNRSW	2.00	$\mathbf{R_{12}}$: $-CH(OH)-CH(OH)-CH_2-O-POO_2^{(2-)}$	7	285.2
30	protein-tyrosine phosphatase 1B	1bzj	ACFIRVY	2.25	$\mathbf{R_8}$: $-COO^{(-)}$ $\mathbf{R_9}$: $-CF_2-POO_2^{(2-)}$	2	299.1
31	48G7 hybridoma line FAB	1gaf	HLRWY	1.95	$\mathbf{R_1}$: $-NO_2$ $\mathbf{R_4}$: $-OPOO^{(-)}-(CH_2)_4-COO^{(-)}$	7	301.2
32	calcium binding domain VI of porcine calpain	1alw	FHIKLQRW	2.03	$\mathbf{R_1}$: $-I$ $\mathbf{R_4}$: $-CH{=}C(SH)(Z)\text{-}COO^{(-)}$	4	305.1
33	protein tyrosine phosphatase 1B	1ecv	DFIKQRVY	1.95	$\mathbf{R_1}$: $-NH-CO-COO^{(-)}$ $\mathbf{R_2}$: $-COO^{(-)}$ $\mathbf{R_4}$: $-I$	3	333.0
34	29G11 FAB	1a0q	FHKLRWY	2.30	$\mathbf{R_1}$: $-OPOO^{(-)}-CH(C_4H_9)-NH-CO-(CH_2)_2-COO^{(-)}$	10	341.3
35	catalytic antibody 28B4 fragment	1kel	FKNRWY	1.90	$\mathbf{R_1}$: $-NO_2$ $\mathbf{R_4}$: $-CH_2-N[CH_2-POO_2^{(2-)}]-(CH_2)_4-COO^{(-)}$	10	343.2
36	bovine trypsin	1az8	C*DQTV	1.80	$\mathbf{R_1}$: $-C(NH_2)_2^{(+)}$ $\mathbf{R_3}$: $-CH(CH_2-COOCH_3)-(CH_2)_2-Ph[4\text{-}C(NH_2)_2^{(+)}]$	7	354.5
37	human estrogen receptor (α)	3ert	ADEILMTW	1.90	$\mathbf{R_1}$: $-C(C_2H_5){=}CPh(4\text{-}OH)-Ph[4\text{-}O-(CH_2)_2-N(CH_3)_2]$	9	387.5
38	glutathione S-transferase	1guh	DFLQRTVY	2.60	$\mathbf{R_1}$: $-CH_2-S-CH_2-CH[CO-NH-CH_2-COO^{(-)}]-NH-CO-(CH_2)_2-CH[COO^{(-)}]-NH_3^{(+)}$	13	396.4
39	dihydrofolate reductase	4dfr	DFILKRT	1.70	$\mathbf{R_1}$: $-CO-NH-CH[COO^{(-)}]-(CH_2)_2-COO^{(-)}$ $\mathbf{R_4}$: $-N(CH_3)-CH_2-(4\text{-aminopteridin-6-yl})$	7	457.5

[a] The default −R groups are hydrogen atoms. Res: resolution of the protein structure. RT: the number of released torsions. C*: Cys in disulfide bridge.

docking energy and the free energy of binding was always lower than the energies corresponding to the original crystal structures. The decrease (or similarity) of the latter energy values indicates that the docked conformation is in an energy minimum, likewise to the crystal conformation. The free energies of binding of the minimum structures and the average free energy of binding of the members of the RMSD$_{2.5}$ rank were very close to each other, in some cases

Table 2. Verification of the Docking Method for the Investigated Protein−Ligand Complexes of the Present Study[a]

ID	E_d(c)	E_d(m)	ΔG(c)	ΔG(m)	RMSD(m)	N	ΔG(a)	SD	RMSD(a)	SD
1	−2.48	−3.61	−2.17	−3.29	0.64	10	−3.29	0.00	0.65	0.01
2	−3.91	−4.39	−3.60	−4.08	2.44	10	−4.08	0.00	2.44	0.00
3	−4.58	−5.31	−3.97	−4.67	2.53	0[c]				
4a	−3.68	−4.31	−3.68	−4.31	0.87	10	−4.31	0.00	0.87	0.00
4b	−5.80	−6.58	−5.80	−6.58	0.31	8	−6.58	0.01	0.29	0.01
4c	−4.66	−5.36	−4.66	−5.36	0.64	10	−5.36	0.00	0.64	0.00
4d	−4.82	−5.13	−4.78	−5.13	0.63	4	−5.12	0.01	0.64	0.01
5a	−7.89	−8.18	−7.89	−8.18	0.32	10	−8.18	0.00	0.32	0.00
5b	−7.28	−7.93	−7.28	−7.93	1.99[b]	9	−7.63	0.01	1.00	0.75
6	−7.15	−8.63	−6.35	−7.72	2.13	10	−7.70	0.01	2.13	0.01
7a		−9.29		−9.29	0.83	8	−9.29	0.00	0.83	0.00
7b	−8.33	−8.79	−8.33	−8.79	0.61	9	−8.79	0.00	0.61	0.00
8	−3.82	−5.61	−3.50	−4.98	2.70[b]	4	−4.46	0.00	2.36	0.01
9	−5.33	−5.64	−5.01	−5.31	0.66	10	−5.31	0.01	0.67	0.00
10	−1.83	−4.91	−1.39	−3.90	1.74	10	−3.88	0.01	1.72	0.02
11	−5.39	−7.20	−5.73	−6.20	1.30	9	−6.09	0.17	1.57	0.50
12	−7.93	−10.36	−6.83	−8.89	0.84	10	−8.88	0.01	0.91	0.05
13	−4.73	−6.81	−3.48	−5.42	2.64	2	−5.23	0.01	1.07	0.02
14	−5.04	−6.59	−5.28	−6.20	1.05	10	−6.20	0.00	1.05	0.01
15		−8.93		−8.08	0.96	10	−8.01	0.05	0.90	0.07
16	−6.78	−8.72	−5.35	−7.14	0.41	7	−7.05	0.09	0.75	0.77
17	−9.36	−10.77	−8.97	−9.92	2.10	9	−9.51	0.22	1.07	0.58
18	−7.22	−9.32	−6.51	−8.17	0.77[b]	6	−7.68	0.11	0.74	0.03
19	−8.35	−8.81	−7.39	−7.73	0.81	9	−7.72	0.00	0.79	0.02
20	−9.15	−11.31	−8.99	−10.31	1.03	9	−10.30	0.02	0.98	0.06
21	−9.27	−10.49	−8.06	−8.84	0.59	10	−8.80	0.03	0.58	0.02
22	−11.15	−11.79	−10.54	−11.18	0.80	10	−11.17	0.01	0.97	0.39
23	−4.65	−8.45	−6.77	−8.91	1.08	10	−8.70	0.21	1.12	0.15
24	−6.50	−7.60	−6.38	−6.67	0.34	8	−6.65	0.02	0.37	0.06
25	−10.85	−13.09	−10.78	−11.87	0.65	9	−11.81	0.05	0.70	0.04
26	−10.32	−11.92	−10.21	−10.85	1.70	9	−10.79	0.06	0.89	0.33
27	−10.61	−11.26	−8.68	−9.12	0.41[b]	6	−8.70	0.17	0.87	0.49
28	−0.14	−8.17	−2.10	−9.47	1.05	7	−9.33	0.06	0.99	0.05
29	−10.56	−11.77	−9.79	−10.50	1.31	4	−10.21	0.42	1.19	0.23
30	−13.78	−14.25	−12.60	−13.07	0.38	10	−13.06	0.01	0.33	0.03
31	−10.98	−12.80	−8.57	−10.37	0.58[d]	8	−10.10	0.44	0.72	0.32
32		−6.94		−5.66	1.51[e]	9	−5.62	0.03	1.83	0.31
33	−10.72	−12.85	−10.56	−11.59	0.76	8	−11.57	0.02	0.76	0.01
34	−10.72	−13.36	−7.47	−9.64	2.61	3	−9.25	0.25	2.15	0.29
35	−12.73	−15.85	−10.71	−12.59	2.01	5	−12.47	0.15	1.57	0.35
36	−12.75	−14.06	−11.87	−12.30	0.51	10	−12.22	0.06	0.59	0.27
37	−1.43	−7.23	−8.30	−10.21	1.81	7	−9.66	0.36	1.42	0.18
38	−14.88	−16.38	−10.82	−12.14	0.77	5	−11.94	0.18	0.86	0.14
39	−11.71	−14.47	−11.17	−13.77	1.22	5	−13.70	0.13	1.23	0.03

[a] E_d(c): docked energy of the crystallographic ligand (kcal/mol); E_d(m): docked energy of the docked conformation of the lowest free energy of binding of the ligand; ΔG(c): calculated free energy of binding of the crystallographic ligand (kcal/mol); ΔG(m): free energy of binding of the lowest free energy of the ligand; RMSD(m): root-mean-square deviation of the docked conformation of the lowest free energy of binding of the ligand (Å); N: number of conformations in the $RMSD_{2.5}$ rank (the crystallographic structure was used for reference, see text); ΔG(a): average free energy of binding of the $RMSD_{2.5}$ rank; RMSD(a): average RMSD of the $RMSD_{2.5}$ rank; SD: standard deviation. [b] The RMSD value corresponds to the minima of the second best group of docked ligands. [c] The RMSD of docked conformations were only slightly above 2.5 Å (acceptable fit) but $RMSD_{2.5}$ rank was not defined in this case. [d] Parts of the ligand of high B-factors were omitted during RMSD calculation. [e] The crystallographic ligand had erroneous structure. RMSD was calculated omitting the erroneous part.

they were equal. This similarity between the minimum and average energies corresponds to the structural similarity of the members of $RMSD_{2.5}$ rank.

In summary, the docking results of this section provide verification for the further investigations on the 44 systems using our standard docking protocol. It should be noted, that at systems **8** and **19**, an octylglucoside molecule covers the binding site. The proteins of these systems were not involved in further studies of the 31 different proteins, as the site seems to be specifically arranged for the original ligands. See also Methods for details on selection of the systems for the study.

As a further test of reproducibility of the docking results on the investigated systems, 31 jobs of different complexes were repeated using new seed parameters for the random number generator at each job. Despite the total randomization of starting positions of the ligand, excellent reproducibility was obtained (Supporting Information, Figure A). This finding strengthens reliability of data in the **E** matrices.

Molecular Interaction Fingerprints (MIFs). The following subtraction was applied to the rows of the **E** matrices defined in the Methods section, to obtain the $\mathbf{MIF_Y}$ matrices

$$\mathbf{MIF_Y} = \mathbf{E_Y} - (\mathbf{REF_{MIF}})^{\mathbf{T}} \qquad (1)$$

where Y = 1 or 2, denoting the two kinds of methods of evaluation, and $\mathbf{REF_{MIF}}$ is the vector of reference energies of each ligand. The reference energies correspond to the binding of the 39 different ligand molecules in complex with their original proteins. Thus, $\mathbf{MIF_Y}$ matrices contain rows, i.e., the molecular interaction fingerprints with values of

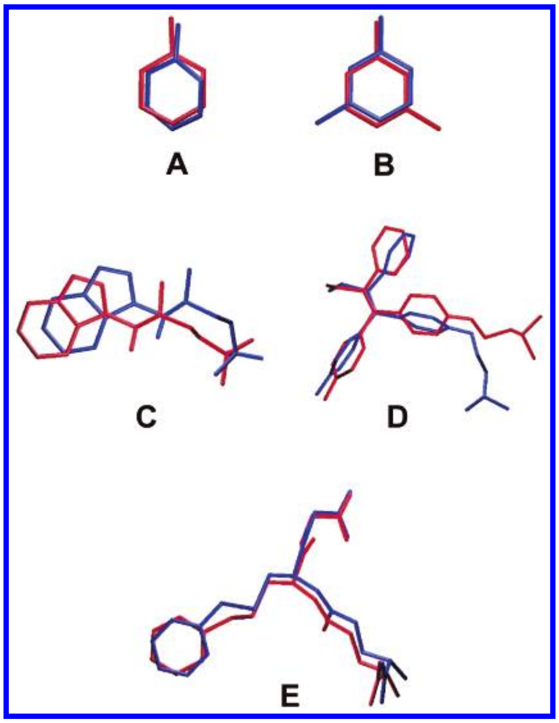


Figure 1. Comparison of docked (energy minimum; blue), and the crystallographic conformations (red) of different RMSD values. A: ligand **1**; B: ligand **2**; C: ligand **29**; D: ligand **37**; E: ligand **38**. Note that in case of small ligands, like **2**, the alteration of positon of only one methyl group causes significantly larger RMSD, due to the small number of atoms, even if the positions of other atoms match with those of the reference ligand.

relative binding free energy of each aromatic compound and the 31 different proteins. The familiar term "fingerprint" was used here to indicate that the interaction profile of each row is specific to only one compound.

A central problem of screening of proteins is the choice of binding energy threshold for selection of the appropriate targets. For this purpose, Chen et al.[13] used an energy value calculated from the correlation with the number of atoms of the ligand as a threshold. However, only an $R^2 \approx 0.5$ can be obtained when correlating binding free energies with the number of heavy atoms of ligand molecules (Supporting Information, Figure B). While the correlation with the number of atoms alone seems to be insufficient for estimation of binding free energies even for a group of similar compounds, the use of a threshold was avoided in our study.

Using the MIF approach, experimental data (binding free energies calculated with the reference structures) form a basis of the relative comparison of the binding affinities of the 31 proteins. For ligands having several crystal complexes, the minima of the individual binding energies could be used as a more precise reference.

The resulting $\mathbf{MIF_1}$ matrix is presented in Figure 2a. According to eq 1, negative or zero values indicate the possibility of interaction between the ligand and the proteins.

In protein screening studies, a further critical point is the verification of the results. It may be rather difficult, considering the expenditure required for determination of the X-ray structures of all possible (in this study: 1209) combinations of ligands and proteins or measurement of experimental binding free energies (or inhibition constants). In some cases, indirect biochemical data are available that relate the activity of selected proteins to drug side effects or toxicities of the compounds.[13,14] However, validation of only the positive results is possible, i.e., when drugs are involved in considerable interaction with the proteins.[14] For biologically inactive compounds (negative result) no or very few data are available.

In the present study, structural considerations were used to perform a check on the aforementioned positive selection. The principle of fragment docking[26−29] and fragment-based drug design[30] asserts that fragments of certain molecules bind to the same sites as the whole molecules. Here, this simple principle was applied to select out the proteins, which are a priori targets of a given compound, having the same (boxes of solid outline on Figure 2a) or similar (boxes of dashed outline) fragments as the native ligand. A definition by Wang et al.[27] was used to classify the "same" fragments of the investigated ligands. For example, salicylic acid (**8**) and 2-(oxalylamino)benzoic acid (**20**) have the same fragment (benzoic acid), the 4-(acetylamino)-3-hydroxy-5-nitrobenzoic acid (**24**) was considered only "similar" to salicylic acid in this study, having more than one group linked to the benzene ring.

Comparison of the location of boxes of Figure 2a, and the corresponding energy values showed good agreement: the proteins (sites) selected by fragment considerations (boxes) have, with few exceptions, zero or negative energy value (Figure 2a). The fragment-based comparison was made for ligands **1**−**14**. As larger compounds have site-specific or bulky groups attached to the aromatic rings, the simple fragment considerations are not as straightforward as they were at the small molecules. For example, ligand **28** and protein sites **4a**−**b** hardly match even if benzoic acid is a common "fragment" of ligands **28** and **4**, because the attached group makes molecule **28** considerably larger than benzoic acid. However, in some cases striking agreement with the rational fragment or similarity considerations can still be found, for larger compounds, as well. One example is the match of tyrosine (**16**) to the site of native ligand phenylalanine (**12;** Figure 2a, box with red outline): the two compounds differ only by a hydroxyl group. A further nice match is that of ibuprofen (**19**) and the site of ligand **17**. Both ligands have carboxyl groups and large hydrophobic parts (isobutylphenyl and indole groups, respectively) arranged at similar distance. There are some other systems, which have negative value in $\mathbf{MIF_1}$ and high degree of similarity between the ligands. One example is ligand **13**, which fits to the binding site of **17**: both ligands have carboxyl groups linked to the benzene ring (the linker differs only in one $-CH_2-$ unit), which play a pivotal role in forming the complexes. Furthermore, the existence of identical binding site residues I, R, T, W and Y (Table 1) at both proteins indicate the similarity of binding pockets (like at the other pairs of systems discussed above). In conclusion, the rows of $\mathbf{MIF_1}$ matrix contain reasonable "interaction fingerprints" of the ligand molecules, which are in good agreement with fragment and similarity considerations.

Interestingly, there is a clear trend in $\mathbf{MIF_1}$ toward the higher selectivity for larger ligands. The only exception is ligand **32**, iodobenzene. This trend seems reasonable as ligands of smaller molecular weight are fragments of the larger molecules in many cases and therefore can fit to the sites of the proteins corresponding to larger ligands. At the same time, larger molecules have more specific groups to

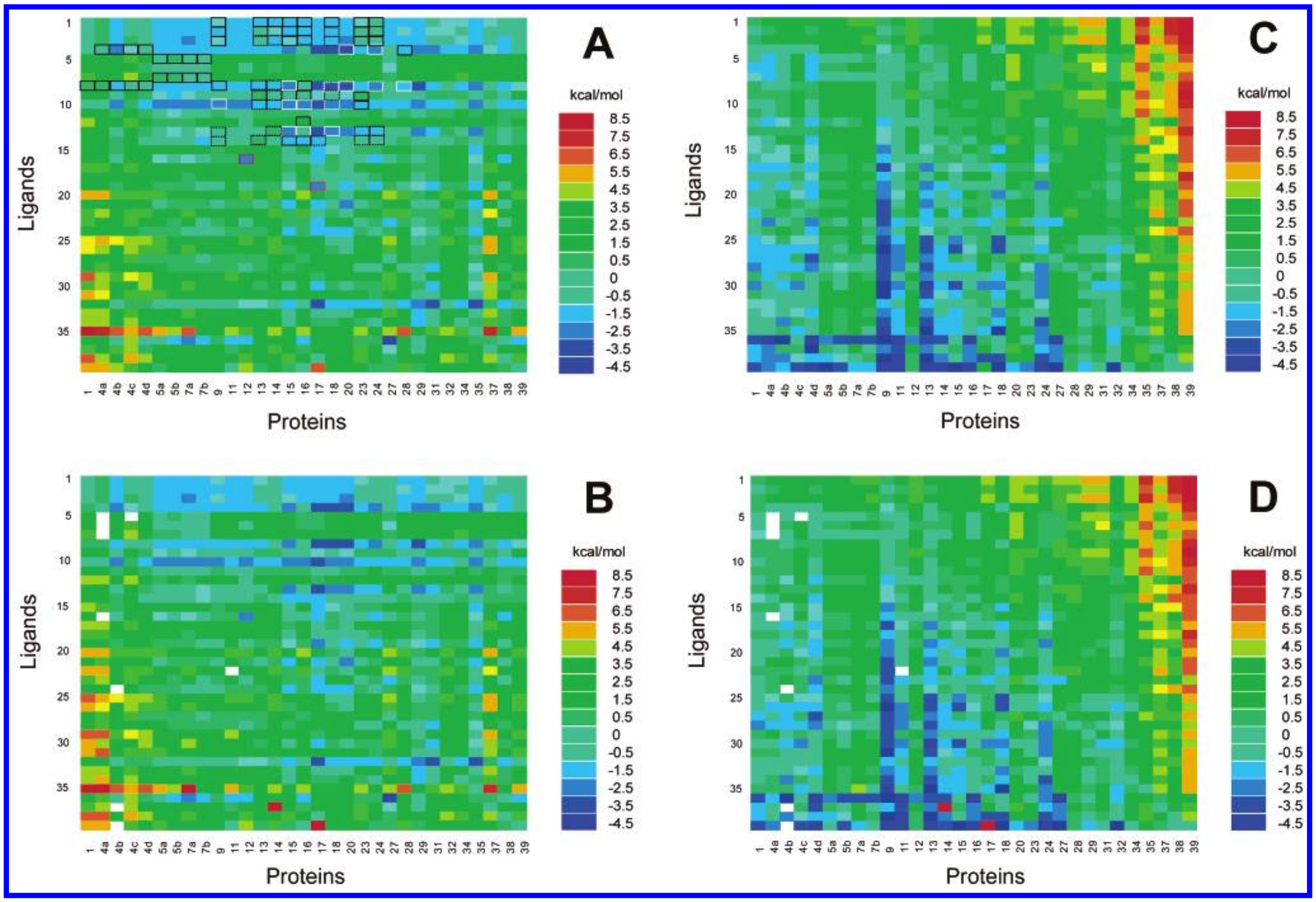


Figure 2. $\mathbf{MIF_1}$ (A), $\mathbf{MIF_2}$ (B), $\mathbf{MAF_1}$ (C), and $\mathbf{MAF_2}$ (D) matrices. The rows are ordered according to the serial number (molecular weight) of the ligand molecules. See text for meaning of other notations. White boxes mean undefined values. Solid (and dashed) outlines mark boxes with same (or similar) fragments of the ligand of the actual protein and the ligand of the actual row. Red outlines correspond to similarities, as well. (See text for details on the latter ones.) The reference crystal structures correspond to boxes of zero energy values and positioned roughly on the diagonal of the matrix. The rows of each matrix are named "fingerprints", including specific data for the ligands.

interact with their specific target and/or there is no space to fit them in the sites of smaller ligands. (In case of system **32**, the heavy iodine atom does not contribute to the specificity relatively as much as it increases the molecular weight. Hence, this compound is an exception to the trend.)

In matrix $\mathbf{MIF_2}$, a similar trend can be observed (Figure 2b). In this case, the docked structures situated far from the center of the subsite of the original crystallographic ligand were filtered out (see Methods for details). In 131 of 1209 cases (11%) the values of $\mathbf{E_1}$ were changed during the generation of $\mathbf{E_2}$, due to the large distance from the reference subsite. In nine cases white boxes depict in the figure, that in all cases the resulting structures were far from the original subsite. In some simple cases a rational explanation can be given for the appearance of such deviations. For example, in case of the benzamidine derivatives and 2-phenylethylamine (**5**, **6**, and **7**), no match was found with the subsite of **4a**, a benzoic acid target (Figure 2b). This subsite, attracting benzoic acid, is repulsive for the positively charged compounds. This repulsion causes a shift of e.g. ligand **5** far away from the original **4a** site (Figure 3) and the appearance of the white box in the $\mathbf{MIF_2}$ matrix. In cases when only the subsite of a protein pocket is of interest, the filtering used to obtain $\mathbf{MIF_2}$ could be essential. For instance, this situation occurs e.g. when the other parts of the pocket are not involved in the key interactions or interact with water molecules. Hence, this filtering helps to avoid overestimating to actual role of the positive interactions with the inactive part of the pocket and thus of small importance. The rows (MIFs) of the $\mathbf{MIF_2}$ matrix reflect relative structural and energy information at the same time and contain therefore more information than $\mathbf{MIF_1}$ (or methods used in refs 13 and 14), which have not applied e.g. a distance criterion to filter ligands bound to nonactive subsites.

Molecular Affinity Fingerprints (MAFs). While the rows of the $\mathbf{E_Y}$ matrices form the basis of the calculation of MIFs, the columns can be considered as "slices" of conventional in silico HTS, which is performed routinely in drug design for a protein site. A subtraction, similar to (1)

$$\mathbf{MAF_Y} = \mathbf{E_Y} - \mathbf{REF_{MAF}} \qquad (2)$$

was performed for the columns to obtain matrices of molecular affinity fingerprints ($\mathbf{MAF_Y}$). ($\mathbf{REF_{MAF}}$ is not completely identical with $\mathbf{REF_{MIF}}$, as the former one has 39 elements instead of the 31.) In this case, the columns of $\mathbf{MAF_Y}$ contain relative energies of ligands for each protein site. However, also the rows of the matrix have valuable meaning, describing the competitive affinity of a compound for the different protein sites relative to the original crystallographic ligand of the site. Therefore, the rows of $\mathbf{MAF_Y}$ were called the molecular affinity fingerprints (MAFs) of the compounds. The diagonal of $\mathbf{MAF_1}$ (Figure 2c) divides the MAF values roughly into two groups: the negative values

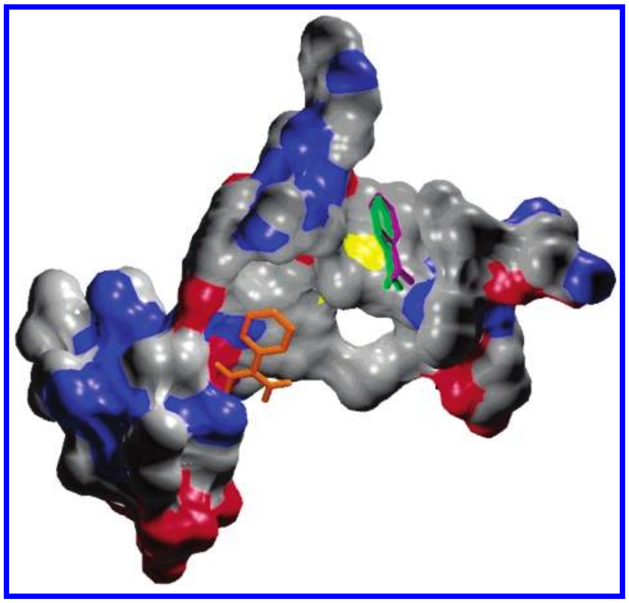


Figure 3. Solvent accessible surface representation of the binding pocket of Transcription factor Malt domain III (protein **4a**). The crystallographic positions of the original ligand benzoic acid (**4**), its docked conformation, and the docked benzamidine (**5**) are marked with green, purple, and orange colors, respectively. The match of the docked and the crystallographic benzoic acid is excellent. The benzamidine molecule is at a distance of ca. 10 Å from the subsite of benzoic acid. Benzamidine is attracted by the negatively charged (red) subsite and repulsed by the H-donor (blue) subsite of benzoic acid. The free energy of binding of benzamidine is smaller (−5.41 kcal/mol) than that of benzoic acid (−4.31 kcal/mol). Thus, the selection of the appropriate binding pocket may be erroneous, using only the free energies as filters. The distance criterion at **MIF$_2$** (see text) decreases the number of the wrongly selected pockets or proteins.

are gathered at the left bottom corner and the positive values are at the right top corner of the matrix. This finding indicates that smaller compounds are weaker competitors at foreign sites, while the relative competitive affinity of the larger compounds is better at foreign sites, as well. However, beyond system **28**, the competitive efficiency seems to decrease for any compound. This limit indicates that if the binding site is large and has a native ligand of more than ca. 300 Da, then the competitive strength of other ligands at the site is smaller. The quantitative results presented in this section are in agreement with the rational considerations: small compounds with fewer interacting points are plausibly weak competitors at sites of larger ligands. The crude trend of decreasing binding free energies with an increasing number of atoms of ligands (see Section "Molecular Interaction Fingerprints" for details) indicates a similar trend, as well. Summarily, the MAFs reflect a realistic picture of the competition.

The **MAF$_2$** matrix (Figure 2d) contains changes at similar positions as **MIF$_2$**. In some cases, a large shift to the positive values can be observed (e.g. at ligand **37** − protein **14** or ligand **39** − protein **17**) after the second filtering.

CONCLUSIONS

In the present study, a docking approach of protein screening was investigated. Based on a validated set of reproduced crystal complexes, relative measures of molecular interaction and affinity fingerprints were introduced and used for the comparison of ligand-protein selectivity. MIF and MAF patterns were found to be in good agreement with rational considerations. As the generation of the corresponding matrices is relatively fast (one docking job took an average of a half an hour of CPU time on a PIV 1.7 GHz processor), the data set can be expanded with further proteins and ligands in an effective way. Furthermore, using the same scoring function, the results on 1209 complexes remain comparable. This statement is highly important, as the number of the available experimental inhibition constants or binding free energy values is very limited, and sometimes the measured values are not comparable with each other due to the different experimental setups.[10] Thus, protein screening would be impossible if to use only the available experimental binding affinity values. An additional advantage of the computational docking approach is that the docked complexes can be precisely compared to the original crystal complex of the site and in the case of the other ligands, structural information of the place of subsite can be involved in the evaluation (filtering) together with the corresponding energy values. Moreover, if no data exist for the reference crystal complex, the binding site can be found by methods based on pocket search[31] or by blind docking.[5]

However, general problems of docking applications, protein flexibility[32] and the role of structural waters,[5,18] should be considered in future protein screening studies. The error of docking caused by structural waters can be reduced with the above-mentioned filtering using subsite information. The elimination of the induced effects is more problematic.[33]

Practically, MIFs can be used for comparison (selection) of protein targets involved in drug side effects and toxicity, while MAFs contain information on the competitive affinity of a compound at a site and therefore may be applied for relative estimation of the possibility of interference of a drug with others at their corresponding binding sites (proteins).

Generally, MIFs and MAFs calculated on the basis of a larger free energy database may aid the exploration of the appropriate biochemical interaction route of any compounds, e.g. by automated comparison of their fingerprints with those of ligands of elucidated biochemistry.

ACKNOWLEDGMENT

This work was supported by a grant from the IMAGETOX EU Research and Training Network: Contract No. HPRN-CT-1999-00015. Support from Estonian Science Foundation is also acknowledged (Grant No. 4571).

Supporting Information Available: Job-by-job reproducibility of 31 docking jobs performed on different protein-(-ligand) systems (Figure A) and the trend between the calculated binding free energies and the number of heavy atoms of the ligands (Figure B). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Blundell, T. L.; Jhoti, H.; Abell, C. High-throughput crystallography for lead discovery in drug design. *Nature Rev.* **2002**, *1*, 45−54.

(2) Shoichet, B. K.; McGovern, S. L.; Wei, B.; Irwin, J. J. Lead discovery using molecular docking. *Curr. Opin. Chem. Biol.* **2002**, *6*, 439−446.

(3) Böhm, H. J.; Stahl, M. Structure-based library design: molecular modelling merges with combinatorial chemistry. *Curr. Opin. Chem. Biol.* **2000**, *4*, 283−286.

(4) Dennis S.; Körtvélyesi, T.; Vajda, S. Computational mapping identifies the binding sites of organic solvents on proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 4290−4295.

(5) Hetényi, C.; van der Spoel, D. Efficient docking of peptides to proteins without prior knowledge of the binding site. *Protein Sci.* **2002**, *11(7)*, 1729−1737.

(6) Hetényi, C.; Szabó, Z.; Klement, É.; Datki, Z.; Körtvélyesi, T.; Zarándi, M.; Penke, B. Pentapeptide amides interfere with the aggregation of beta amyloid peptide of Alzheimer's disease. *Biochem. Biophys. Res. Comm.* **2002**, *292*, 931−936.

(7) Hetényi, C.; Körtvélyesi, T.; Penke, B. Mapping of the possible binding sequences of two beta-sheet breaker peptides on beta amyloid peptide of Alzheimer's disease. *Bioorg. Med. Chem.* **2002**, *10(5)*, 1587−1593.

(8) Halperlin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **2002**, *47*, 409−443.

(9) Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. A review of protein-small molecule docking methods. *J. Comput.-Aided Mol. Design* **2002**, *16*, 151−166.

(10) Brooijmans, N.; Kuntz, I. D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335−373.

(11) Marelius, J.; Hansson, T.; Åqvist, J. Calculation of ligand binding free energies from molecular dynamics simulations. *Int. J. Quantum Chem.* **1998**, *69*, 77−88.

(12) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19(14)*, 1639−1662.

(13) Chen, Y. Z.; Zhi, D. G. Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. *Proteins* **2001**, *43*, 217−226.

(14) Rockey, W. M.; Elcock, A. H. Progress toward virtual screening for drug side effects. *Proteins* **2002**, *48*, 664−671.

(15) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Databank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(16) Laskowski, R. A. PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res.* **2001**, *29*, 221−222.

(17) Minke, W. E.; Diller, D. J.; Hol, W. G. J.; Verlinde, C. L. M. J. The role of waters in docking strategies with incremental flexibility for carbohydrate derivatives: heat-labile enterotoxin, a multivalent test-case. *J. Med. Chem.* **1999**, *42*, 1778−1788.

(18) Pang, Y.-P.; Perola, E.; Xu, K.; Prendergast, F. G. EUDOCK: A computer program for identification of drug interaction sites in macromolecules and drug leads from chemical databases. *J. Comput. Chem.* **2001**, *22(15)*, 1750−1771.

(19) Schaftenaar, G.; Noordik, J. H. Molden: a pre- and postprocessing program for molecular and electronic structures *J. Comput.-Aided Mol. Design* **2000**, *14*, 123−134.

(20) Pappu, R. V.; Hart, R. K.; Ponder, J. W. Analysis and application of potential energy smoothing and search methods for global optimization. *J. Phys. Chem. B* **1998**, *102*, 9725−9742.

(21) Walters, P.; Dolata, M. S. Babel − A Molecular Structure Information Interchange Hub. Department of Chemistry, University of Arizona, Tucson, AZ 85721. (http://smog.com/chem/babel/).

(22) Pedretti, A.; Villa, L.; Vistoli, G. Vega: a versatile program to convert, handle and visualize molecular structure on windows-based PCs *J. Mol. Graph.* **2002**, *21*, 47−49.

(23) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity − A rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3288.

(24) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **1996**, *14*, 33−38.

(25) Merritt, E. A.; Bacon, D. J. Raster3D: Photorealistic molecular graphics. *Methods Enzymol.* **1997**, *277*, 505−524.

(26) Friedman, A. R.; Roberts, V. A.; Tainer, J. A. Predicting molecular interactions and inducible complementarity: fragment docking of Fab-peptide complexes. *Proteins* **1994**, *20*, 15−24.

(27) Wang, J.; Kollman, P. A.; Kuntz, I. D. Flexible ligand docking: A multistep strategy approach. *Proteins* **1999**, *36*, 1−19.

(28) Jackson, R. M. Q-fit: A probabilistic method for docking molecular fragments by sampling low energy conformational space. *J. Comput.-Aided Mol. Design* **2002**, *16*, 43−57.

(29) Majeux, N.; Scarsi, M.; Apostolakis, J.; Ehrhardt, C.; Caflisch, A. Exhaustive docking of molecular fragments with electrostatic solvation. *Proteins* **1999**, *37*, 88−105.

(30) Lesuisse, D.; Lange, G.; Deprez, P.; Benard, D.; Schoot, B.; Delettre, G.; Marquette, J. P.; Broto, P.; Jean-Baptiste, V.; Bichet, P.; Sarubbi, E.; Mandine, E. SAR and X-ray. A new approach combining fragment-based screening and rational drug design: Application to the discovery of nanomolar inhibitors of Src SH2. *J. Med. Chem.* **2002**, *45*, 2379−2387.

(31) Brady, G. P.; Stouten, P. F. W. Fast prediction and visualization of protein binding pockets with PASS. *J. Comput.-Aided Mol. Design* **2000**, *14(4)*, 383−401.

(32) Fradera, X.; de la Cruz, X.; Silva, C. H. T. P.; Gelpi, J. L.; Luque, J. F.; Orozco, M. Ligand-induced changes in the binding sites of proteins. *Bioinformatics* **2002**, *18(7)*, 939−948.

(33) Carlson, H. A. Protein flexibility and drug design: how to hit a moving target. *Curr. Opin. Chem. Biol.* **2002**, *6*, 447−452.

CI034052U