Docking of Small Ligands to Low-Resolution and Theoretically Predicted Receptor Structures

MAREK WOJCIECHOWSKI, JEFFREY SKOLNICK
Laboratory of Computational Genomics, Donald Danforth Plant Science Center,
893 North Warson Rd., St. Louis, Missouri 63141

Received 25 April 2001; Accepted 5 June 2001

Abstract: We have developed a simple docking procedure that is able to utilize low-resolution models of proteins created by structure prediction algorithms such as threading or *ab initio* folding to predict the conformation of receptor–small ligand complexes. In our approach, using only approximate, discretized models of both molecules, we search for the steric and quasi-chemical complementarity between a ligand and the receptor molecules. This averaging procedure allows for the compensation of numerous structural inaccuracies resulting from the theoretical predictions of the receptor structure. The best relative orientation of these two models is obtained by an exhaustive scan over the rigid body's six-dimensional translational and rotational degrees of freedom. The search method is based on a real space grid-searching algorithm, unlike docking methods based on the fast Fourier Transform algorithm. We have applied this algorithm to rebuild structures of several complexes available in the Protein Data Bank. The structures of the receptors are produced by means of our threading algorithm PROSPECTOR, subsequently refined, and then utilized in the docking experiment. In many cases, not only is the localization of the binding site on the receptor surface correctly identified, but the proper orientation of the bounded ligand is also reasonably well reproduced within the level of accuracy of the modeled receptor itself.

© 2002 John Wiley & Sons, Inc.   J Comput Chem 23: 189–197, 2002

Key words: ligand docking; protein structure prediction; low-resolution protein models; cocrystallized complexes; induced fit

Introduction

One of the important, if not the most important, elements of protein function is the protein's ability to interact with and bind various ligands. This ability is closely related to the three-dimensional structure of the protein. Although the number of known primary sequences of proteins grows rapidly, their quaternary structures usually remain unknown due to the relatively difficult and time-consuming procedure of experimental structure determination. Recently, the quality of theoretical structure prediction methods has been greatly improved, and sometimes results in structures whose quality is similar to low-resolution experimental structures.[1, 2] Thus, there is a clear need for a docking procedure that will be able to utilize these theoretical models of proteins for the prediction of conformations of receptor–small ligand complexes. Such an approach might also be helpful in rebuilding the quaternary structures of multimeric proteins when the structures of particular subunits of the protein are also theoretically predicted.

There are many approaches to the docking problem, and many algorithms developed by various groups have been devoted to this problem.[3−7] Many approaches consider both the ligand and the receptor to be rigid,[8−10] while still others try to deal with the ligands and, to some extent, the receptor's flexibility.[5, 11, 12] There are also various methodologies used for scoring the quality of the resulting complexes. Some implementations use simple geometric criteria such as surface and shape complementarity to define the binding site,[13, 14] while others use some type of potential energy function to distinguish between good and bad solutions.[15−17]

Shape complementarity plays an important role in protein–protein interactions,[18, 19] and various techniques have proven to be efficient tools for generating near-native conformations of complexes, even from unbound components.[20] Recently, methods based on correlation functions have become very popular.[7, 8, 21, 22] In these algorithms, the structures of the molecules to be docked are first discretized by projecting them onto a three-dimensional grid and then the value of a correlation function that accounts for the shape complementarity of these two discrete representations is calculated in a search over the six-dimensional rigid body degrees of

Correspondence to: J. Skolnick; e-mail: skolnick@danforthcenter.org

Contract/grant sponsor: NIH; contract/grant number: RR12255

freedom. This search can be very efficiently performed by applying a Fourier transformation. An additional advantage of algorithms that utilize a correlation function is their ability to accept, to some extent, inaccuracies in the models by computationally changing the grid size. This type of protocol should make docking calculations for low-resolution structures possible.[22–25]

Although algorithms for the successful docking of low-resolution structures of pairs of proteins have recently improved,[16, 26] the problem of docking small ligands to such receptors is unexplored. In this article, we approach the problem of docking small ligands to inaccurate receptor structures by searching for both the steric and chemical complementarity between the ligand and the receptor molecule. Because our main focus is on docking to low-resolution structures that are in most cases the results of theoretical predictions, we use only approximate, discretized models of both the ligand and its protein receptor. Previously, it was shown that by averaging the structural details and by smoothing the potential energy surface, it is possible to drive the ligand towards the real binding site; thus avoiding, in many cases, the local minima problem.[21, 27] In our case it also turns out that this averaging procedure allows for the compensation of numerous structural errors resulting from theoretical predictions of the receptor's tertiary structure.

We have applied our new algorithm to rebuild structures of several complexes available in the Protein Data Bank. The structures of these receptors were first predicted from our threading algorithm,[28] refined using our generalized comparative model protocol,[29] and then utilized in the docking experiments. In many cases, not only has the localization of the binding site on the receptor surface been correctly identified, but the proper orientation of the bound ligand was reasonably restored, well within the level of accuracy of the modeled receptor.

Methods

Our docking procedure is a grid-based, complete search over the six-dimensional space defined by the rigid body translation of the ligand in three dimensions and its rotation over three Euler angles. No additional information regarding the binding site is required. Before performing the actual docking procedure, we assign "properties" to every atom of the ligand molecule. This assignment just defines to which of the 19 predefined chemical groups (see Table 1) the particular atom of the ligand belongs.

In the first step of the algorithm, both the receptor and the ligand are discretized by projecting them onto a uniform cubic lattice of grid size 2 Å. The projection of the ligand onto the grid is performed such that if the distance of the centroid of any cell to an atom of the ligand is smaller than the size of the cell, then this cell is marked by the property of the group of which this atom is a member.

The structure of the receptor is projected in a slightly different manner. This process is divided into two stages (see Fig. 1). In the first stage, all lattice cells that lie within the distance of double the cell size from any atom of an amino acid side chain or its alpha carbon are marked as having the property of that particular amino acid. Backbone atoms, excluding the alpha carbons, are projected as an additional, virtual, amino acid type. At the same time, all the lattice cells projected by side chain atoms are marked as the receptor's "SHELL," while lattice cells projected by backbone atoms are marked as receptor "COAT" cells. In the second stage, the projection of the receptor is repeated. This time the particular cell is marked as belonging to the attractive COAT of the receptor if its centroid is located within the distance of one cell size from the side chain atom of any amino acid. A cell is marked as belonging to the repulsive CORE of the receptor molecule if it is located within the distance of one cell size from any protein backbone atom, excluding the alpha carbons. In the case of the receptor molecule, not only are the exact positions of the side chain atoms used for the projection, but the positions of these atoms resulting from all the rotameric states of every residue are also used.

Table 1. Fragments and Functional Groups Used in the Definition of the Ligand–Amino Acid Statistical Potential.

Number	Symbol	Group Description
1	—COOH	carboxylic acid
2	—CONH—	amide
3	—NH_2	amine
4	—NH_2	amine by multiple bond
5	—OH	hydroxyl
6	—SH	thiol
7	—Ph	phenyl
8	—C—C—	chain of aliphatic carbons
9	—C=C— or —C≡C—	fragment of chain with multiple CC bonds
10	—NHC(NH_2)NH	guanidinium
11		heterocyclic ring
12	—C—S—C—	thioether
13	—C—O—C—	ether
14	>C=O	carbonyl
15	—SO_2—	sulfone
16	—SO_3H	sulfonic acid
17	—PO_4—	phosphate
18	—NO_2	nitro group
19	—X	halogene (F, Cl, Br)

After both molecules are discretized, their best relative orientation or, to be more precise, the best relative orientation of their discretized images, is obtained by an exhaustive search over the entire grid space, which is conducted by moving the set of cells representing the ligand molecule. This movement is performed using a one grid cell step. When the scanning of the grid by the ligand cells' translation is complete, the ligand molecule is rotated by one of the Euler angles. Then, its new orientation is again projected onto the lattice, and the whole search process starts over again from the beginning. The projection, grid scanning, and ligand rotation steps are repeated until the entire six-dimensional relative translation-orientation space is exhaustively searched.

During the search, each position of the discretized ligand molecule is scored according to its steric complementarity with the particular area of the receptor grid calculated for the value of the correlation function. Additionally, the energy of the interaction is calculated according to the scoring by the knowledge-based pairwise potential.

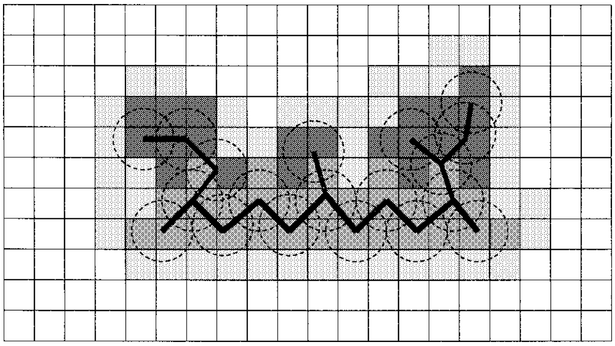

Figure 1. The projection of the receptor residues to the cubic lattice. Light gray represents the SHELL cubes, medium gray represents the CORE cubes, and dark gray represents the COAT cubes.

The value of the steric complementarity is evaluated by means of a simple correlation function:

$$S_{\alpha,\beta,\gamma} = \sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} a_{l,m,n} * b_{l+\alpha,\ m+\beta,\ n+\gamma} \tag{1}$$

where:

$$a_{l,m,n} = \begin{cases} -5 & \text{CORE} \\ 0.5 & \text{COAT} \\ 0.5 * n & \text{SHELL} \\ 0 & \text{environament} \end{cases} \qquad b = \begin{cases} 1 \\ 0 \end{cases}$$

In this representation, the only repulsive part of the interaction comes from the ligand cubes overlapping with CORE receptor cubes, i.e., those cubes, which are the discretized representation of the receptor-backbone atoms. The cubes that represent the receptor side chains contribute an attractive part in the interaction score. In particular, cubes in the SHELL of the molecule are attractive with the strength of this interaction proportional to n, the number of the receptor amino acids projected into this cube.

Because docking calculations based only on steric complementarity (especially when dealing with small ligand molecules) usually lead to incorrect results due to a large number of false positives, we additionally use a pairwise statistical potential to score the resulting complexes according to their quasichemical complementarity. The specific part of the interaction in our scoring function is based on a pairwise statistical potential built on the basis of over 300 known structures of various complexes available in the Protein Data Bank (PDB).[30] The structures used to derive the statistical potential were selected from the whole PDB database according to the following rule: only those structures with ligands containing at least above 5 heavy atoms were chosen. Those structures with any other ligands (listed as HET records in the PDB file) within the range of 8 Å from the chosen one were rejected. All sequences with a sequence identity above 50% to any other sequence in the set were also removed from the database. Two structures with an identity above 50% were accepted into the training set of protein structures only if their ligands had different PDB three-letter code names and, more importantly, if their sizes differed by at least five heavy atoms. Extrinsic to the training set, a testing set of 20 complexes was selected according to similar criteria, except that none of the structures included in the testing set was allowed to have a sequence identity higher than 20% to any of the structures already included in the above-mentioned training set.

To build the potential, we defined 19 functional groups used to decompose the structure of the ligand into quasichemical building blocks. The groups used in our analysis are listed in Table 1. The parameters of our potential were obtained by a statistical analysis of the training set of complexes described above, and then applying eq. (2).

$$E_{i,j} = -\ln \frac{n_{i,j}}{N \cdot x_i \cdot x_j} \tag{2}$$

where $n_{i,j}$ is the number of observed contacts of the functional group i with amino acid of type j, x_i, and x_j are the mol fractions of groups i and amino acids j, respectively, and N is the total number of contacts in the database. In our approach, a contact between i and j occurs when the ligand cell marked with property i overlaps with the receptor cell marked with the property of the amino acid j. Obviously, this potential depends on the lattice cell size. In all of our calculations, we used a fixed grid size of 2 Å. This was arrived at on the basis of several computational experiments, which showed that the best results are obtained with this grid size.

The specific interaction score was calculated according to eqs. (3) and (4), respectively, as

$$p = \sum_{i=1}^{21}\sum_{j=1}^{19} E_{i,j} * n_i \tag{3}$$

$$P_{\alpha,\beta,\gamma} = \sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} p_{l+\alpha,\ m+\beta,\ n+\gamma} \tag{4}$$

The final docking score of the particular complex was calculated by means of the linear combination of both the steric and the potential terms according to eq. (5).

$$\mathrm{DS} = \mathrm{corr} * S + P \tag{5}$$

where corr is a correction term that depends on the size of the ligand molecule, S is the value of the steric match term calculated by eq. (1), and P is the value of the potential specific interaction term calculated by eq. (4). The correction term is used to ensure that both P and S have a similar influence on the value of the docking score DS, and it is calculated by eq. (6), in which n is the number of atoms in the ligand molecule. The coefficients of this equation are estimated as the least-square approximations of the values of S and P on the basis of calculations for the set of native complexes.

$$\mathrm{corr} = \frac{4.23 * n + 40.08}{1.32 * n + 22.20} \tag{6}$$

After the entire six-dimensional space is searched, all of the solutions are scored according to eq. (5) and then sorted.

Results and Discussion

We used a program written on the basis of the above algorithm to dock some small ligands to the structures of the receptors available from the PDB, as both are complexes with these ligands as well as with a number of theoretically predicted models of the receptors. The complete list of ligands and receptors used can be found on our Web page http://bioinformatics.danforthcenter.org/. To be able to easily verify the docking results in the test cases, when the correct geometry of the complex was known, we need a measure that scores the quality of the predicted complexes relative to the native ones. We decided to use (as a quality indicator) the percentage of the native contacts (NC) from the original native complex that were preserved in the predicted complex as well. We count a contact as being native if ligand atom number i is in contact with receptor residue number j in both the predicted and the native complexes. Additionally, we also used the percentage of nonspecific contacts (nsNC) as a measure of the success of the localization of the binding site residues. The percentage of nonspecific contacts is defined as the fraction of residues that are in contact with the ligand in the predicted complex relative to the fraction of the residues that are in contact with the ligand in the native structure.

First, we tested our algorithm on the database of 318 complexes that we used to generate our statistical potential (see Table 2). We repeated these calculations with only the steric complementarity term used for scoring, then with only the statistical potential term, and finally with the complete term used for scoring the complexes. In all runs, a grid size of 2 Å was used, and the ligand rotation step was set to 20°. All calculations were performed on a 733-MHz Pentium III cluster. The average job took about 5–10 min on 20 processors.

The results of these calculations are presented in Figure 2. As one can easily see, the scoring function that combines both the steric complementarity term and the qasichemical potential performs better than either alone. This combined function allows us to predict the most complexes with a particular number of preserved native contacts. However, a closer analysis of the curves showing the results for the steric only and potential only docking reveals an interesting relationship. Scoring by steric complementarity only gives a few complexes with very good quality (the percentage of preserved native contacts is above 80–90%), but it is able to predict about 200 structures out of 318 with at least one native contact. Although the best quality predictions made by means of the potential-only scoring have fewer, mainly only 50–60%, of their native contacts preserved, the percentage of the predicted contacts drops more slowly than in the previous case to give 230 complexes overall, with at least one native contact. This result clearly indicates that the binding site and its surroundings maintain specificity towards the ligand even in this low-resolution representation. On the other hand, the number of correctly predicted complexes obtained by scoring only with the steric complementarity term is also surprisingly high, especially when one keeps in mind that a fuzzy representation of the receptor side-chain positions is used in the calculations, and the ligands are relatively small.

The results using the combined set of terms, rank ordered by decreasing quality of the fraction of native ligand–receptor contacts that are correctly predicted, is summarized in Table 2. The fraction of correctly predicted contacts ranges from 87 to 0%.

Obviously these calculations do not say much about the real predictive power of our algorithm. To get a more reliable verification of its performance, we repeated the docking calculations for the smaller database of 20 test complexes, but this time with the additional restriction that not only were none of these structures present in the training set of complexes, but also that none of these structures had a sequence identity higher than 20% to any of the structures from this set. Table 3a shows the set of testing proteins along with their corresponding ligand. As shown in Table 3b, in this case, 11 predicted complexes had 20% or more of the specific native contacts preserved, and 13 of them had at least 10% of the specific contacts preserved. This looks very promising in terms of the predictive power of the algorithm.

To test our algorithm under conditions closer to real-life problems when the geometry of the binding site differs sometimes significantly from the geometry in the cocrystallized complex, we performed additional docking calculations on a few examples of the ligands and receptors that were crystallized separately and whose structures are available in the PDB in both forms. Because, in most cases, the conformation of the receptor side chains readjust to the ligand only upon binding (the alpha carbon root mean square derivative, RMSD, for the receptors crystallized in the free form and the ones that cocrystallized with the ligands in most cases are below 1 Å) and because, in our model, the crystallographic side chain positions are not used explicitly for docking, not surprisingly we did not notice any significant differences in complexes obtained by redocking ligands to their cocrystallized receptors and the same complexes obtained after docking these ligands to the apropriate receptors crystallized in the free form.

Although in most cases the ligand-binding process involves only small side-chain rearrangements among the binding pocket residues, it may sometimes induce a wide range of the structural changes in a large part of the protein, including the hinge movements of the entire receptor subdomains. In our selected subset of the PDB, we found a few structures of cocrystallized complexes with the receptors also crystallized separately where the RMSD differences of the alpha-carbon positions between both the free and ligand-bound forms of the protein were even above 7 Å. We applied our algorithm to dock the ligands found in the cocrystallized forms to the free forms of these receptors. An example of a protein that shows significant ligand-induced domain movements is the maltodextrin binding protein.[31] The structure of this protein is available in the PDB in both the free form (1omp) as well as in the form of complexes with various ligands (1anf, 3mbp). In the free form, the binding site is open and accessible to the water molecules (Fig. 3a).

Upon ligand binding, this protein undergoes a hinge-bending and a twisting kind of motion between its two domains, so that, once bound, the ligand is closed inside the binding pocket (Fig. 3b). The RMSD of the alpha-carbon positions between these two forms of the receptor is about 3.7 Å; however, most of the differences are concentrated in the area of the binding site. The subdomains themselves behave, during ligand binding, almost as rigid bodies, and their internal geometries do not change much. Despite these structural differences, the docking of maltose (the ligand molecule from the 1anf structure) as well as maltotriose (the ligand molecule

Table 2. The List of All Training PDB Structures and the Percentage of Preserved Native Contacts.

No.	PDB ID	NC[a]	No.	PDB ID	NC[a]	No.	PDB ID	NC[a]
1	1rom	87	58	1bj9	33	115	1rms	17
2	1ars	86	59	7gch	32	116	2cbs	16
3	1c24	85	60	1fkh	32	117	1qcp	16
4	1dt1	78	61	1c1h	32	118	1ck6	16
5	1spa	75	62	1db1	32	119	1ayw	16
6	1icm	75	63	2hmb	32	120	1bgq	16
7	1ddt	65	64	1bep	31	121	1qti	16
8	1icn	63	65	1pax	30	122	1au3	16
9	1map	62	66	1pbk	29	123	1hvq	16
10	1co6	62	67	3pyp	29	124	2nlr	16
11	451c	61	68	1aba	29	125	1ci3	16
12	1c22	60	69	3pax	29	126	1d4o	16
13	2cmd	58	70	1cxc	27	127	1hlb	16
14	1maq	57	71	1bek	27	128	2ypn	16
15	1iol	57	72	1drh	27	129	1bb6	15
16	1ivr	57	73	1cyo	26	130	1gsq	15
17	5yas	57	74	1c1l	26	131	1lce	15
18	1akc	57	75	2hbg	26	132	1oyc	15
19	1aod	55	76	5tln	26	133	2ack	15
20	1bxm	54	77	5cyt	25	134	1akb	15
21	1hcz	54	78	1myt	25	135	4pax	15
22	1zsb	52	79	1b8o	25	136	1cr1	15
23	3c2c	52	80	4lbd	25	137	1eno	15
24	2dri	51	81	1oce	25	138	1cbs	15
25	1dtp	51	82	2mm1	25	139	1fkl	15
26	1oxp	50	83	1bvd	24	140	1ayv	15
27	5rhn	50	84	3dhe	24	141	1ojt	14
28	1cot	48	85	1a53	24	142	1fem	14
29	1ylv	48	86	1bb7	24	143	1fen	14
30	1zid	47	87	3lbd	23	144	1ra9	13
31	3rhn	45	88	1yet	23	145	3cbs	13
32	5bu4	44	89	2fcr	23	146	1ndh	13
33	1ctj	44	90	1flp	23	147	1arc	13
34	3ert	43	91	1bgo	22	148	1mnp	13
35	1rpj	43	92	1d7r	22	149	7taa	12
36	2dap	41	93	5fit	22	150	1shv	12
37	1lih	41	94	1htp	22	151	4mbp	12
38	1d7v	40	95	1drv	22	152	1cef	12
39	1c75	40	96	5eat	21	153	2fam	11
40	1drm	39	97	1b56	21	154	1rg7	11
41	6qch	38	98	6nul	21	155	1cr2	11
42	1a3k	38	99	1c9e	20	156	1bp4	11
43	1kpf	38	100	1a4h	20	157	1qs2	11
44	21hb	38	101	1fkd	20	158	1jdd	11
45	1ceq	37	102	3a3h	20	159	2fke	11
46	1ptg	37	103	1bso	20	160	1d06	11
47	1mrk	37	104	1nje	20	161	1blh	11
48	7odc	36	105	7ccp	20	162	5fx2	10
49	1fkg	36	106	1dmb	20	163	1vzc	10
50	1cpq	36	107	1cxy	19	164	1aec	10
51	1fsz	35	108	1au2	19	165	1b02	10
52	1mpd	35	109	1gne	19	166	1eco	10
53	1pmt	35	110	2sim	18	167	1b8n	10
54	1llo	34	111	1fhe	17	168	1aim	9
55	1b9i	34	112	1tyn	17	169	1bo8	9
56	1ngh	34	113	1qsr	17	170	2lh5	9
57	1qkq	33	114	1cgo	17	171	1b9h	9

Table 2. (Continued)

No.	PDB ID	NC[a]	No.	PDB ID	NC[a]	No.	PDB ID	NC[a]
172	1tsl	9	221	1b9t	0	270	1obt	0
173	2pax	9	222	1qpk	0	271	1rbn	0
174	1b0f	8	223	1drw	0	272	1upj	0
175	2cbr	8	224	1a39	0	273	2tdm	0
176	1ayu	7	225	1axb	0	274	1dgy	0
177	1hbp	7	226	1cg6	0	275	1b39	0
178	1bdu	7	227	1inv	0	276	1bsj	0
179	1lif	7	228	1bzc	0	277	1pjc	0
180	1rx7	7	229	1bzj	0	278	3cox	0
181	1au0	6	230	1cy6	0	279	1b8v	0
182	1hnl	6	231	1b0e	0	280	3eng	0
183	1hna	6	232	1jdx	0	281	1qan	0
184	4tmk	6	233	1bdb	0	282	3jdw	0
185	1dad	6	234	1b1c	0	283	1qg2	0
186	1a26	6	235	1by2	0	284	1qgf	0
187	2dhn	6	236	1a27	0	285	1du7	0
188	1erb	6	237	1may	0	286	3rab	0
189	4fiv	6	238	1dru	0	287	1qra	0
190	1fel	5	239	1zfj	0	288	1enu	0
191	1hmr	5	240	1mrj	0	289	4a3h	0
192	1vot	5	241	1dud	0	290	2csn	0
193	4rsk	5	242	1a5w	0	291	1aj6	0
194	1br6	5	243	1cgk	0	292	1eus	0
195	1adg	5	244	2aim	0	293	1ama	0
196	1lid	4	245	1aqm	0	294	1lsp	0
197	5tmp	4	246	2cah	0	295	1amq	0
198	1rob	4	247	1cip	0	296	2enb	0
199	1adf	4	248	1cjc	0	297	5a3h	0
200	1trb	3	249	1cet	0	298	1btn	0
201	1ofv	3	250	1cg4	0	299	1rsm	0
202	1gr2	3	251	2a3h	0	300	1rvd	0
203	4cd2	3	252	1b0u	0	301	1bvq	0
204	1fdr	3	253	1cpt	0	302	1bws	0
205	1bib	3	254	2ang	0	303	1frq	0
206	1c9w	2	255	1ctq	0	304	1skj	0
207	2cnd	2	256	1aq7	0	305	1fxs	0
208	1rpf	2	257	1iam	0	306	1sth	0
209	1tcs	2	258	1cw7	0	307	6cts	0
210	1lmc	2	259	1mbt	0	308	6fiv	0
211	2dpg	2	260	1cy4	0	309	1gym	0
212	1rx5	2	261	1fmb	0	310	2q21	0
213	1byg	2	262	2gnk	0	311	1uib	0
214	2dpm	2	263	1bx6	0	312	1hdr	0
215	1ifu	1	264	1mtw	0	313	7jdw	0
216	1enz	1	265	1d01	0	314	1vpt	0
217	1b0o	1	266	1d6h	0	315	1diw	0
218	1a8p	1	267	1d6f	0	316	8est	0
219	1cgz	1	268	2ncd	0	317	9est	0
220	1dih	1	269	1nox	0	318	9icd	0

[a] NC is the fraction of preserved native contacts.

from the 3mbp structure) to the free receptor (1omp) resulted in the structure of a complex with the ligands positioned in the correct area of the binding site (Fig. 3).

Another group of proteins that are known to undergo significant conformational changes upon binding are kinases.[32] Figure 4 shows adenylate kinase complexed with its inhibitor (Fig. 4a with the complex's PDB code 1ake) and the same receptor crystallized in the free form (PDB code 4ake), but complexed with the inhibitor by means of our docking program (Fig. 4b). In this case, the RMSD between the receptor in its free (open) and the ligand-

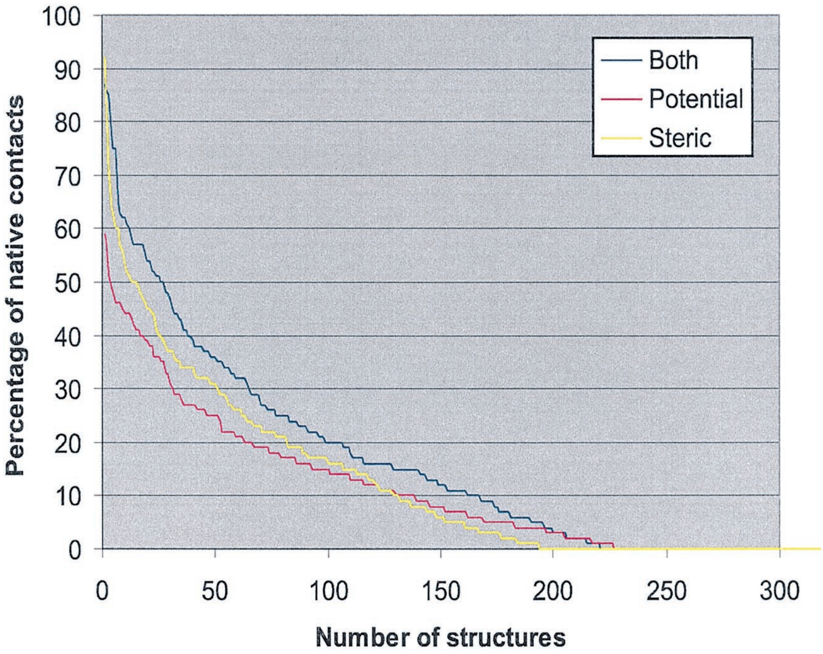

Figure 2. The number of complexes predicted by means of various scoring functions with the particular number of specific native contacts preserved.

bound (closed) form is 7.1 Å. Although the geometry of the binding site as well as the geometry of a large part of the protein is completely different in the free and in the complexed form of the kinase, even in this case our program was able to correctly identify the binding-site residues. When comparing residues in contact with the ligand molecule in the cocrystallized native complex with residues in contact in the complex obtained by means of our docking algorithm, in the predicted complex the docked inhibitor is in contact with a subset of residues that are also in contact with this ligand in the cocrystallized complex. Obviously, it is only a subset of these residues due to the significantly different geometry of the binding site in the open form of the receptor (Fig. 4b). However, the orientation of the bound inhibitor in the docked complex is similar to the one in the native cocrystallized complex.

Our main goal here was to develop an algorithm that would be able to dock smaller ligands to low-resolution, theoretically

Table 3a. The List of Ligands Used in the Nonhomologous Testing Set of PDB Complexes.

PDB ID	The Name of a Ligand as it Appears in the HETNAM (or HETSYN When Available) Records
1af7	S-Adenosyl-L-homocysteine
1b3n	Cerulenin
1b59	Ovalicin
1bj4	Pyridoxal-5′-phosphate
1bym	Glucose
1cen	Glucose
1dmw	7,8-Dihydrobiopterin
1dve	Protoporphyrin IX containing FE
1dvp	Citric acid
1mai	D-Myo-inositol-1,4,5-triphosphate
1mdr	(S)-Atrolactic acid
1npx	Flavin-adenine dinucleotide
1nst	Adenosine-3′5′-diphosphate
1oth	N-(Phosphonoacetyl)-L-ornithine
1qkp	Retinal
1rne	[[[3-(2-Methyl-propane-2-sulfonyl)-1-benzenyl]-2-propyl]-carbonyl-histidyl]-amino-[Cyclohexylmethyl]-[2-hydroxy-4-isopropyl]-pentan-5-oic acid butylamide
1ukd	P1-(adenosine-5′-P5-(uridine-5′)pentaphosphate
2hmy	S-Adenosylmethionine
2izj	Biotin
5pnt	2-(N-Morpholino)-ethanesulfonic acid

Table 3b. Percentage of Preserved Specific and Nonspecific Native Contacts in the Docked Complexes for the Nonhomologous Testing Set.

PDB ID	nNC[a] (%)	NC (%)[b]	PDB ID	nNC[a] (%)	NC (%)[b]
1oth	89	87	2izj	35	23
1cen	77	61	1npx	55	18
1dvp	80	47	1nst	54	11
1bj4	75	45	1ukd	70	7
1b59	79	41	1af7	40	5
1mai	48	39	1byc	47	3
1qkp	60	30	2hmy	0	0
1dmw	31	27	1rne	3	0
1b3n	53	27	1mdr	0	0
5pnt	48	25	1dve	21	0

[a] NC is the fraction of preserved native contacts.

[b] nNC is the percentage of nonspecific contacts, as defined in the Results and Discussion section.

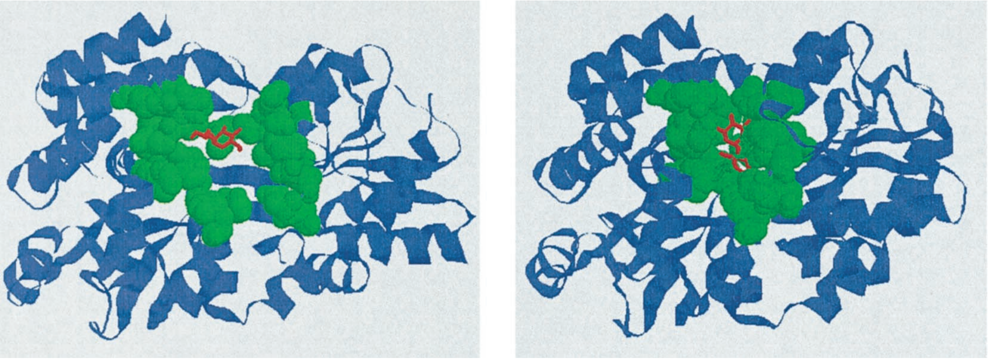

Figure 3. Maltodextrin binding protein. (a) The free (open) form with the ligand, maltose, docked. (b) The complexed (closed) form with the same ligand cocrystallized. Residues in contact with the ligand in the cocrystallized form are green.

predicted structures of receptors. Docking small molecules to the structures of receptors with significant differences from native is the real challenge. To test the efficiency of our protein structure prediction algorithms, we used a set of standard benchmarks, including the Fischer Database.[33] Taking advantage of the fact that some of the structures present in the Fischer Database[34] are also present in the PDB in the form of complexes with small ligands, we tested our docking procedure by trying to rebuild these complexes using our homology modeled structures of these receptors instead of the X-ray ones. The structures that were successfully modeled with reasonable accuracy and at the same time were available in the form of complexes with some small ligands have PDB ID codes as follows: 1bbh, 1c2r, 1mdc, 2cmd, 2sar. The quality of these models is in the range of 3 to 6 Å RMSD when compared to the appropriate PDB structures.

The results of the docking calculations are shown in Table 4. Even for the structures as far as almost 6 Å from native, up to 47% of the specific native contacts are preserved. Only for the predicted 1mdc structure did docking fail to recognize the binding site. This structure is the fatty acid binding protein of a sulfate ion bound together with the ligand palmitic acid. The presence of this ion in the binding site was not taken into account in the docking experiment, but in this case it could be crucial for the correct ligand binding.[35] All other complexes from this set were successfully rebuilt within the accuracy of the modeled receptor.

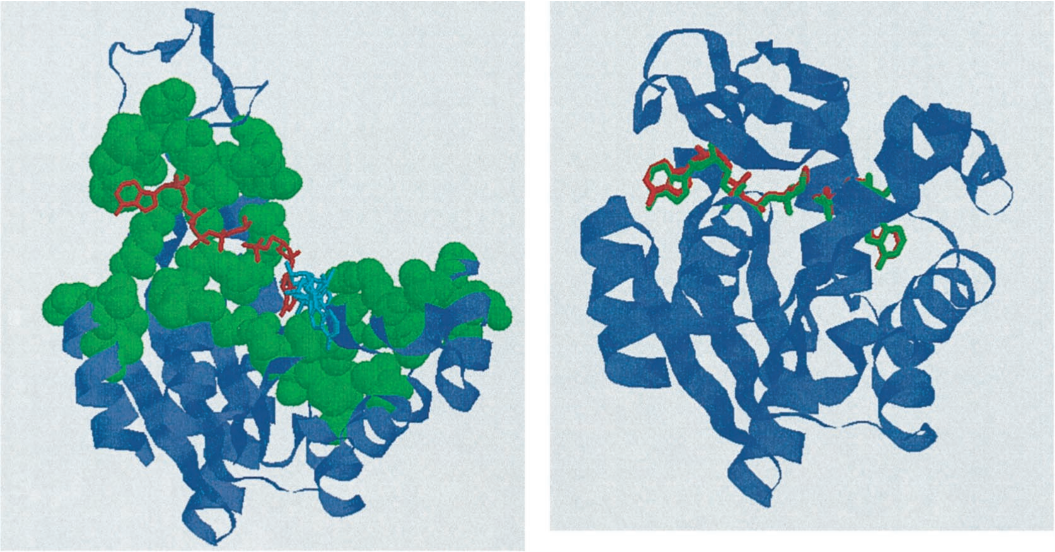

Figure 4. Adenylate kinase. (a) The free (open) form with the inhibitor docked by means of our algorithm. Two of the best scored complexes are shown. (b) The cocrystallized (closed) form. The original and redocked positions of the inhibitor are shown. Residues in contact with the ligand in the cocrystallized form are green.

Table 4. Results of Docking to the Theoretically Predicted Receptor.[a]

PDB ID	RMSD (Å)	nsNC (%)	NC (%)
1bbhA	3.16	75	47
1c2rA	4.94	71	30
1mdc_	4.92	0	0
2cmd_	5.57	74	47
2sarA	5.99	42	59

[a] RMSD of the modeled receptor from the experimental one. nsNC and NC are the percentage of nonspecific native contacts and percentage of specific native contacts, respectively.

Discussion

These results indicate that our docking routine is able to utilize structural information still present even in the low-resolution structures of receptors and to use this information to place small ligands in the binding site in the correct orientation. Local steric and physicochemical properties of the receptor binding sites are definitely responsible for the final locking of the ligand molecule in the correct position and orientation; however, these local preferences extend beyond the immediate neighborhood of the binding site itself. The general physicochemical properties of the receptor, mimicked here by our potential term, drive ligands toward the correct location on the receptor surface. It was shown previously[22] that global structural features are an important factor in the first stages of protein–protein recognition. On the basis of our calculations, we conclude that, similarly, the global topology of the receptor is significant for binding small ligands as well, especially in the early stages. In the small ligand binding, we seem to be mimicking the first stage of the binding process, when the ligand is probing the surface of the receptor trying to find the areas with favorable interactions. At this point, the detailed structural features and high-resolution interactions like hydrogen bonding do not yet play an important role. It was shown that including such features, even in the docking of unbound molecules, does not influence the result.[19] However, hydrogen bonds are obviously crucial in the next stage of binding, and including them in the scoring function significantly improves the results when restoring cocrystallized complexes.[36]

Although our algorithm is not perfect and, in some cases, fails to recognize the binding site for the particular receptor altogether, our results clearly indicate that in many cases it is possible to utilize even low-quality structures in successful docking experiments with small ligands. Obviously this procedure does not lead to a unique atomic-level solution. The resulting complexes must be further refined either by simple energy minimization or molecular dynamics calculations; however, when used in combination with other tools, our approach may prove to be very valuable for the genome-scale products of the ligand binding site.

References

1. Pillardy, J.; Czaplewski, C.; Liwo, A.; Lee, J.; Ripoll, D. R.; Kazmierkiewicz, R.; et al. Proc Natl Acad Sci USA 2001, 98, 2329.
2. Marchler-Bauer, A.; Bryant, S. H. Proteins 1999, 37, 218.
3. Gschwend, D. A.; Good, A. C.; Kuntz, I. D. J Mol Recognit 1996, 9, 175.
4. Read, R. J.; Hart, T. N.; Cummings, M. D. Ness, S. R. Supramol Chem 1995, 6, 135.
5. Leach, A. R. J Mol Biol 1994, 235, 345.
6. Blaney, J. M.; Dixon, J. S. Perspect Drug Dis Design 1993, 1, 301.
7. Katchalski-Katzir, E.; Shariv, I.; Eisenstein, M.; Friesem, A. A.; Aflalo, C.; Vakser, I. A. Proc Natl Acad Sci USA 1992, 89, 2195.
8. Gabb, H. A.; Jackson, R. M.; Sternberg, M. J. E. J Mol Biol 1997, 272, 106.
9. Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. A. J Mol Biol 1982, 161, 269.
10. Jiang, F.; Kim, S. H. J Mol Biol 1991, 219, 79.
11. Desmet, J.; Wilson, I. A.; Joniau, M.; DeMaeyer, M.; Lasters, I. FASEB J 1997, 11, 164.
12. Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. J Comput Aid-Mol Design 1996, 10, 293.
13. Peters, K. P.; Fauck, J.; Frommel, C. J Mol Biol 1996, 256, 201.
14. Laskowski, R. A.; Luscombe, N. M.; Swindells, M. B.; Thornton, J. M. Protein Sci 1996, 5, 2438.
15. Weng, Z. P.; Vajda, S.; Delisi, C. Protein Sci 1996, 5, 614.
16. Robert, C. H.; Janin, J. J Mol Biol 1998, 283, 1037.
17. Muegge, I.; Martin, Y. C. J Med Chem 1999, 42, 791.
18. Norel, R.; Lin, S. L.; Wolfson, H. J.; Nussinov, R. J Mol Biol 1995, 252, 263.
19. Norel, R.; Petrey, D.; Wolfson, H. J.; Nussinov, R. Protein Struct Funct Genet 1999, 36, 307.
20. Palma, P. N.; Krippahl, L.; Wampler, J. E.; Moura, J. J. G. Protein Struct Funct Genet 2000, 39, 372.
21. Vakser, I. A. Protein Eng 1996, 9, 37.
22. Vakser, I. A. Biopolymers 1996, 39, 455.
23. Vakser, I. A.; Matar, O. G.; Lam, C. F. Proc Natl Acad Sci USA 1999, 96, 8477.
24. Vakser, I. A. Protein Struct Funct Genet 1997, 1, 226.
25. Vakser, I. A. Protein Eng 1995, 8, 371.
26. Ritchie, D. W.; Kemp, G. J. L. Protein Struct Funct Genet 2000, 39, 178.
27. Trosset, J. Y.; Scheraga, H. A. Proc Natl Acad Sci USA 1998, 95, 8011.
28. Skolnick, J.; Kihara, D. Proteins 2001, 42, 319.
29. Kolinski, A.; Rotkiewicz, P.; Ilkowski, B.; Skolnick, J. Proteins 1999, 37, 592.
30. Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; et al. J Mol Biol 1977, 112, 535.
31. Sharff, A. J.; Rodseth, L. E.; Spurlino, J. C.; Quiocho, F. A. Biochemistry 1992, 31, 10657.
32. Schulz, G. E. Faraday Discuss 1992, 93, 85.
33. Fischer, D.; Elofsson, A.; Rice, D.; Eisenberg, D. Pac Symp Biocomput 1996, 300.
34. http://www.doe-mbi.ucla.edu/people/fischer/BENCH/table1.html. UCLA. 1996, Abstract.
35. Benning, M. M.; Smee, A. F.; Wells, M. A.; Holden, H. M. J Mol Biol 1992, 228, 208.
36. Meyer, M.; Wilson, P.; Schomburg, D. J Mol Biol 1996, 264, 199.