Prediction of ligand-binding sites of proteins by molecular docking calculation for a random ligand library

Yoshifumi Fukunishi[1,2]* and Haruki Nakamura[1,3]

[1]Protein Structural Information Analysis Team, Biological Information Research Center (BIRC), National Institute of Advanced Industrial Science and Technology (AIST), Koto-ku, Tokyo 135-0064, Japan

[2]Pharmaceutical Innovation Value Chain, BioGrid Center Kansai, 1-4-2 Shinsenri-Higashimachi, Toyonaka, Osaka 560-0082, Japan

[3]Laboratory of Protein Informatics, Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan

Received 27 July 2010; Revised 10 September 2010; Accepted 23 October 2010
DOI: 10.1002/pro.540
Published online 9 November 2010 proteinscience.org

Abstract: A new approach to predicting the ligand-binding sites of proteins was developed, using protein-ligand docking computation. In this method, many compounds in a random library are docked onto the whole protein surface. We assumed that the true ligand-binding site would exhibit stronger affinity to the compounds in the random library than the other sites, even if the random library did not include the ligand corresponding to the true binding site. We also assumed that the affinity of the true ligand-binding site would be correlated to the docking scores of the compounds in the random library, if the ligand-binding site was correctly predicted. We call this method the molecular-docking binding-site finding (MolSite) method. The MolSite method was applied to 89 known protein-ligand complex structures extracted from the Protein Data Bank, and it predicted the correct binding sites with about 80–99% accuracy, when only the single top-ranked site was adopted. In addition, the average docking score was weakly correlated to the experimental protein-ligand binding free energy, with a correlation coefficient of 0.44.

Keywords: protein pocket prediction; ligand pocket; ligand binding site prediction; protein-compound docking; protein-ligand binding free energy

Introduction

Finding functional sites on protein molecular surfaces is crucial for revealing the mechanisms of molecular signaling involving target proteins. Ligand-binding sites are among the most promising targets for drug candidates, whose actions depend upon the inhibition or regulation of the target protein functions. However, in some cases, such ligand-binding sites must be predicted, because little or no experimental information about the protein's functional sites exists. For example, such information is lacking for many of the proteins with tertiary structures that have been determined by structural genomics projects. In addition, non-native ligand-binding sites need to be searched because of putative protein structural changes, due to allosteric effects.

Many researchers have published studies on the prediction of ligand-binding sites.[1–30] Some methods use the mathematical shapes of target proteins, and identify the corresponding concavities as potential ligand-binding sites.[1–23] In these approaches, the volumes of the binding pocket can be determined, but the boundaries between the pocket and non-pocket regions are often unclear. Other methods use a spherical probe with a suitable radius, and calculate the pseudo energy between the protein and the probe.[24–27] A site that is predicted to bind the probe strongly is selected as a ligand-binding site.

Grant sponsors: New Energy and Industrial Technology Development Organization of Japan (NEDO); Ministry of Economy, Trade, and Industry (METI) of Japan.

*Correspondence to: Yoshifumi Fukunishi Biological Information Research Center (BIRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-3-26, Aomi, Koto-ku, Tokyo 135-0064, Japan. E-mail: y-fukunishi@aist.go.jp

Evolutionary information is also employed, under the assumption that the amino-acid sequence of the ligand-binding site must be conserved in the evolution process. Thus, conserved sequences are potential candidates for ligand-binding sites. The propensity of each amino-acid residue to exist at a ligand-binding site has also been calculated from a database of protein-ligand complex structures.[2,28–30] Trp, Phe, Tyr, and Arg tend to form ligand-binding sites, while Gly, Ala and Pro do not. Methods combining these bodies of knowledge have also been examined,[30] and suggested the presence of multiple potential ligand-binding sites, even if only one true binding site exists.

Besides ligand-binding site prediction, the functional analysis and classification of the pockets have been reported.[31–33] As with the folds of the proteins, the structural variety of the ligand-binding sites should be finite, and common structural motifs of ligand-binding sites are frequently observed, even if the global folds or amino-acid sequences are quite different.[31] These findings suggest that the structural variety of ligands, as well as the variety of ligand-binding pockets, may be limited, and thus a finite number of compounds could form a probe set for detecting a ligand-binding pocket by a protein-compound docking study.

Currently, the accuracy of the above predictions is roughly 50–80%, when the top-ranked predicted pocket is adopted as the candidate binding site.[1] When the top three to five predicted sites are adopted, the probability of finding the true site among them increases to 70–90%.[1,28] However, it is still necessary to improve the accuracy and to estimate the affinity of the pocket. Information about the pocket location and size is not very useful if the pocket lacks strong affinity to any compound. Smith *et al.*[34] showed that the average IC_{50} of known drugs is 30 n*M*, and that 70% of known drugs have an $IC_{50} < 50$ n*M*.

In the current study, we proposed a new method for ligand-binding site prediction based on protein-compound docking simulation, and we tried to estimate the protein compound binding free energy without known ligands. In our method, actual compounds were randomly selected from a large compound library, and their three-dimensional (3D) structures were used as the probes, instead of a spherical probe. Each compound was then docked onto the protein surface by the ligand-flexible docking of *in silico* drug screening. We assumed that the true ligand-binding site would show stronger affinity than the other sites to the compounds in the random library, even if the random library did not include the true binders to the binding site. This assumption was based on a previous experiment, in which we applied an *in silico* structure-based drug screening method to more than ten target proteins, and found that the docking poses of almost all compounds were localized around their true ligand-binding sites. We also assumed that the affinity of the true ligand-binding site would be correlated to the docking scores of the compounds in the random library if the ligand-binding site was correctly predicted, because the affinity of a drug depends on the shape of the ligand-binding pocket. In general, a deep binding pocket shows stronger affinity than a shallow pocket. We call this method the molecular-docking binding-site finding (MolSite) method. The MolSite method was applied to 89 proteins, and it predicted the ligand-binding sites of these proteins fairly correctly. The MolSite method was also used to examine the protein-compound affinity for 50 proteins, but only weak correlations were found between the predicted and experimental binding free energies when the volumes of the pockets were small.

Results

Ligand-binding site prediction

The Cα carbons were selected as the centers of the scoring grids for the 89 target proteins, as mentioned in the previous section. On average, 29.3 Cα carbons were selected for each target protein, and a 3D mesh was generated around the Cα carbons. The minimum number of selected Cα carbons was 10, for 2pk4, and the maximum number was 53, for 1dbj. The 10,000 compounds of the C1, C2 and C3 probe sets were docked to the whole surfaces of the target proteins, using the above mesh potential (see Table I and Materials section). The protein-compound docking was flexible, allowing up to 100 conformers for each compound, and was performed by the Sievgene/myPresto program, with an average docking time of about 2 seconds per compound. The S_{avg}, S_1 and S_{top} scores were calculated for each scoring grid from these docking scores for the C1, C2 and C3 sets.

Table II shows the probabilities of finding the true scoring grid among the selected scoring grids by the three different scores, when the C1 set was used as the probe set (see Fig. 1 and Method section). The definition of "true scoring grid" is the condition in which the selected scoring grid contains the average atomic coordinates of the experimentally determined bound ligand. If multiple grids contain the average atomic coordinates of the experimentally determined bound ligand, then the true scoring grid was selected among these grids as follows: the distance between the center Cα carbon of the true scoring grid and the average coordinates of the bound ligand must be the minimum value among the values obtained from these grids. When the S_{avg} score was used for the selection, all true scoring grids were selected. If the scoring grid had been randomly selected, then the probability of finding the true scoring grid would have been 3.4%, since the

Table I. *Summary of the Number of Heavy Atoms in Each Compound Set*

Compound sets	Min.	Average	Max.
C1	7	20.37	26
C2	9	20.36	26
C3	8	20.36	26
Ligand[a]	5	21.25	46

Min., Max., and Average represent the minimum, maximum and average numbers of heavy atoms of the compound sets, respectively.
[a] Ligands of the protein-compound complex structures summarized in Appendix A.

average number of scoring grids was 29.3 (3.4% = 100%/29.3). The S_1 score gave the second best prediction probability. The S_{top} score was inferior to both the S_{avg} and S_1 scores. When the C2 and C3 sets were used as the probe set, the results were almost the same as those obtained with the C1 set, as shown in Table II. The probability of finding the true scoring grid did not depend on the choice of the probe set. The difference in the probability was less than 3%. When the S_{avg} score was used for the selection, the true scoring grids were selected at 100% probability.

Table III shows the prediction results for the centers of the ligand-binding sites, when the C1 set was used as the probe set. The scoring grids were selected using the S_{avg} score. The distances between the center of the predicted ligand-binding site and the center of the bound ligand (Dc) were calculated for the 89 proteins. This distance was adopted as the measure of the prediction accuracy by Brylinski and Skolnick.[28] The minimum distances between the center of each predicted binding site and any atom of the bound ligand (Dmin) were also calculated, as shown in Table III. This minimum distance was adopted as the measure of the prediction accuracy by Huang.[1] The Dmin values are smaller than the Dc values in many cases, but there are some exceptions. The bound ligand of 1a6w adopts an "L" shape, and that of 1epo has a "C" shape. In these cases, the average coordinates of the ligand's atoms are outside of the molecule, and thus the Dc values are smaller than the Dmin values.

Table II. *Probability that the Selected Scoring Grids Include the True Ligand-Binding Sites, Depending on the Scores Used for Selection*

Library	S_{avg}	S_{top}	S_1
C1[a]	98.89	68.18	89.77
C1[b]	100.00	68.63	90.20
C2[b]	100.00	70.59	88.24
C3[b]	100.00	80.39	88.24

Values in the table are in %. C1, C2, and C3 are the three different compound libraries.
[a] All 88 proteins were used (Appendix A).
[b] The 50 proteins with ΔG values were used (Appendix B).

Table III. *Summary of the Dc and Dmin Values for All 89 Target Proteins*

		Distance	
Number	Target protein	Dc (Å)	Dmin (Å)
1	1a6w	0.50	0.82
2	1acj	1.87	0.69
3	1bid	3.79	1.72
4	1blh	2.19	2.57
5	1byb	2.12	1.78
6	1cdo	1.65	1.01
7	1fbp	1.10	0.78
8	1gca	0.41	1.00
9	1hew	5.28	1.02
10	1hfc	2.03	0.95
11	1hyt	5.74	1.17
12	1ida	1.45	0.88
13	1igj	5.11	1.77
14	1imb	1.41	1.48
15	1inc	4.89	1.04
16	1ivd	1.50	0.88
17	1mrg	3.08	1.56
18	1mtw	4.08	0.73
19	1okm	1.88	1.06
20	1pdz	1.98	0.53
21	1phd	2.62	1.30
22	1pso	2.02	0.52
23	1qpe	0.64	0.42
24	1rbp	2.12	1.42
25	1rne	1.58	0.68
26	1rob	0.88	0.59
27	1snc	0.87	0.48
28	1srf	1.81	0.87
29	2ctc	1.69	1.30
30	2h4n	3.22	0.69
31	2pk4	3.22	0.69
32	2sim	1.10	1.10
33	2tmn	0.82	0.95
34	3gch	2.87	0.90
35	3mth	2.39	1.02
36	4phv	39.98	27.29
37	5cna	1.82	1.05
38	5p2p	2.24	1.77
39	6rsa	2.54	0.88
40	1dwd	0.79	0.98
41	1stp	1.17	1.18
42	1ulb	3.37	1.18
43	2ifb	0.41	2.03
44	3ptb	3.60	1.37
45	2ypi	1.56	1.17
46	4dfr	3.69	0.53
47	7cpa	3.67	0.73
48	1apu	3.73	1.78
49	1abe1	7.15	9.64
50	1abe2	7.15	9.64
51	1abf1	5.78	7.39
52	1abf2	4.72	7.40
53	1cbx	1.66	3.22
54	1dbb	0.94	2.59
55	1dbj	1.29	1.57
56	1dog	0.96	2.70
57	1ebg	0.59	1.51
58	1epo	1.03	3.05
59	1etr	1.12	2.51
60	1ets	1.07	2.54
61	1ett	1.58	3.06
62	1hpv	1.47	0.44
63	1hsl	0.76	1.12

Table III. *(Continued)*

Number	Distance		
	Target protein	Dc (Å)	Dmin (Å)
64	1htf1	1.68	3.08
65	1htf2	1.40	4.14
66	1hvr	0.62	1.33
67	1mbi	4.05	4.92
68	1mnc	1.26	5.23
69	1nsd	1.23	2.40
70	1pgp	4.01	5.43
71	1phf	1.11	1.75
72	1phg	1.04	0.85
73	1ppc	1.80	2.05
74	1pph	1.97	2.34
75	1rbp	0.57	2.48
76	1tmn	0.84	3.30
77	1tng	2.21	4.19
78	1tnh	0.81	3.78
79	2cgr	0.43	3.05
80	2cpp	0.68	0.81
81	2gbp	1.96	2.95
82	2phh	2.53	3.51
83	2r04	2.45	8.62
84	2tsc	0.66	4.12
85	5abp1	5.18	7.48
86	5abp2	5.13	7.57
87	5cpp	0.91	0.62
88	5tln	0.96	2.96
89	6cpa	0.69	3.60
	Average	3.34	1.82

This prediction is based on the S_{avg} score. D_c: the distance between the center of the predicted ligand-binding site and the center of the bound ligand. D_{min}: the minimum distance between the center of the predicted binding site and any atom of the bound ligand.

In the MolSite method, the center of the predicted docking pocket is given by the average coordinates of the docking poses of all compounds in the candidate scoring grid. In the current study, the center of the predicted docking pocket was the average coordinates of the 10,000 docking poses. The distribution (root-mean square deviation; RMSD) of the average coordinates of the 10,000 docked poses was calculated for each target protein. The average RMSD value for the 88 targets was 1.61 Å. This average RMSD value means that most of the docked poses were overlapped around the center of the predicted docking pocket for each target protein, when the compounds were selected randomly. With a greater number of compounds, we can expect the average coordinates of the docking pose to converge to a point.

For several targets (4phv, 1abe, 2r04, 1abf, etc.), the MolSite method failed in the prediction of the binding sites. Therefore, the probe-size dependence of the prediction was examined. Table IV shows the results obtained by the "Small" compound set, consisting of the ligands with HA < 13 atoms in the protein-ligand complex structures listed in

Table IV. *Summary of the Dc and Dmin Values for all 89 Target Proteins Obtained by the Small-Compound Set*

Number	Distance		
	Target protein	Dc (Å)	Dmin (Å)
1	1a6w	1.49	0.64
2	1acj	2.18	1.00
3	1bid	4.04	1.88
4	1blh	2.50	1.84
5	1byb	3.29	0.89
6	1cdo	1.60	1.52
7	1fbp	1.83	1.96
8	1gca	0.73	0.80
9	1hew	6.31	0.72
10	1hfc	4.04	0.52
11	1hyt	5.60	1.36
12	1ida	1.53	0.46
13	1igj	7.05	0.56
14	1imb	3.00	0.73
15	1inc	4.51	0.54
16	1ivd	0.77	0.64
17	1mrg	1.19	0.79
18	1mtw	5.28	1.21
19	1okm	3.10	0.65
20	1pdz	1.12	0.65
21	1phd	1.59	1.05
22	1pso	1.81	0.86
23	1qpe	3.68	0.83
24	1rbp	2.45	1.48
25	1rne	1.48	0.80
26	1rob	1.33	1.44
27	1snc	2.97	1.92
28	1srf	2.94	0.65
29	2ctc	0.80	1.48
30	2h4n	1.75	0.59
31	2pk4	1.48	0.79
32	2sim	1.71	1.05
33	2tmn	1.16	1.21
34	3gch	4.12	1.30
35	3mth	3.86	2.61
36	4phv	4.65	0.59
37	5cna	0.73	1.24
38	5p2p	4.04	1.10
39	6rsa	2.28	0.77
40	1dwd	5.46	2.95
41	1stp	1.61	1.32
42	1ulb	3.86	2.08
43	2ifb	4.11	1.65
44	3ptb	1.53	1.48
45	2ypi	4.07	1.50
46	4dfr	4.28	1.05
47	7cpa	3.67	2.62
48	1apu	3.10	1.40
49	1abe1	1.76	0.80
50	1abe2	1.76	0.80
51	1abf1	1.42	0.51
52	1abf2	1.49	0.67
53	1cbx	2.42	1.34
54	1dbb	3.53	0.75
55	1dbj	3.14	0.67
56	1dog	1.06	1.05
57	1ebg	1.16	0.32
58	1epo	2.30	0.74
59	1etr	3.73	1.91
60	1ets	9.06	3.42
61	1ett	7.01	3.84
62	1hpv	0.79	0.85

Table IV. *(Continued)*

Number	Distance		
	Target protein	Dc (Å)	Dmin (Å)
63	1hsl	1.18	1.00
64	1htf1	3.19	1.15
65	1htf2	2.80	1.76
66	1hvr	3.74	1.82
67	1mbi	7.32	6.36
68	1mnc	8.16	4.13
69	1nsd	1.81	1.29
70	1pgp	3.97	2.43
71	1phf	0.78	0.52
72	1phg	0.44	1.53
73	1ppc	4.60	1.47
74	1pph	3.42	1.54
75	1rbp	4.28	1.55
76	1tmn	4.39	1.12
77	1tng	1.63	1.39
78	1tnh	1.67	1.94
79	2cgr	5.64	0.55
80	2cpp	0.51	0.86
81	2gbp	0.97	1.03
82	2phh	1.48	1.04
83	2r04	7.80	3.08
84	2tsc	5.87	1.98
85	5abp1	1.10	0.38
86	5abp2	1.29	0.51
87	5cpp	0.49	1.36
88	5tln	3.16	0.65
89	6cpa	3.86	0.48
	Average	3.12	1.53

This prediction is based on the S_{avg} score. D_c: the distance between the center of the predicted ligand-binding site and the center of the bound ligand. D_{min}: the minimum distance between the center of the predicted binding site and any atom of the bound ligand.

Appendix C. The number of molecules was 21. After the scoring grid was determined using the C1 set, these 21 ligands were docked to the target proteins and the average coordinates of the docking poses were calculated. For 4phv, 1abe1, 1abe2, 1abf1, 1abf2, 5abp1 and 5abp2, the Dc and Dmin values were drastically decreased using the Small set, as compared to the results summarized in Table III. In actual ligand-binding site prediction, we usually do not know the ligand *a priori*. Thus, we did not use the results from the Small set for comparison with those obtained by other docking programs.

The prediction accuracies of our MolSite method, obtained with the C1 library, are shown in Tables V and VI with the accuracies of other methods,[1,28] where the 48 target proteins are the same ones used for the validation of LIGSITE, PASS, Q-site finder and SURFNET.[1] The target protein set was different from the set used for the validation of FINDSITE.[28] Since the results obtained from the C2 and C3 libraries were exactly the same as those obtained from the C1 set when the 50 protein set (protein numbers 40-89 in Tables III and IV) was employed, only the C1 library was used.

When we used the 48 proteins to evaluate the prediction accuracy of the current method, the probabilities of Dc $<$ 4 Å and Dmin $<$ 4 Å were 87.50% and 97.92%, respectively. When we used all 89 proteins (=50+48–9) to evaluate the current method, the probabilities of Dc $<$ 4 Å and Dmin $<$ 4 Å were 88.76% and 85.40%, respectively. Therefore, the prediction performance of the MolSite method was better than that of the other methods. The Dc values determined by the MolSite method in this table were obtained from the top-ranked prediction site. On the contrary, the Dc values for the top 5 predicted sites were the lowest for the FINDSITE and LIGSITE methods.[28] The Dmin values were obtained from the top-ranked prediction site.[28]

We examined the breakdown of the docking scores and the differences between the docking scores at the true binding site and the other sites.

Table V. *Summary of the Pocket-Prediction Accuracies of Various Methods. Prediction Accuracy of MolSite for the 48 Target Proteins, Which were Used for the Validation of LIGSITE, PASS, Q-site Finder, and SURFNET, as Shown in Appendix A*

Method No of Pockets[a]	MolSite Top 1	MolSite Top 1	FIND SITE Top 5	LIG SITE Top 5	Meta Pocket Top 1	LIG SITE Top 1	PASS Top 1	Q-site Finder Top 1	SURFNET Top 1
Distance (D)	Dc	Dmin	Dc	Dc	Dmin	Dmin	Dmin	Dmin	Dmin
D $<$ 8 Å[b]	97.92	97.92	—	—	—	—	—s	—	—
D $<$ 6 Å[b]	97.92	97.92	—	—	—	—	—	—	—
D $<$ 5 Å[b]	91.67	97.92	—	—	—	—	—	—	—
D $<$ 4 Å[b]	87.50	97.92	70.9[c]	51.3[c]	83.0[d]	81.0[d]	58.0[d]	75.0[d]	42.0[d]
D $<$ 3 Å[b]	68.75	97.92	—	—	—	—	—	—	—

The values in this table are in %. This prediction is based on the S_{avg} score. D_c: the distance between the center of the predicted ligand-binding site and the center of the bound ligand. D_{min}: the minimum distance between the center of the predicted binding site and any atom of the bound ligand.

[a] Number of pockets indicates the number of predicted pockets used for the analysis. If one of the predicted pockets is correct, then the prediction is counted as a successful prediction.

[b] The 48 target proteins were used.

[c] Reference 28.

[d] Reference 1.

Table VI. *Summary of the Pocket-Prediction Accuracies of Various Methods. Prediction accuracy of MolSite for the Total of 89 Target Proteins, as Shown in Appendixes A and B*

Method No. of pockets[a] Distance	MolSite Top 1 Dc	MolSite Top 1 Dmin
$D < 8$ Å[b]	98.88	95.51
$D < 6$ Å[b]	98.88	91.01
$D < 5$ Å[b]	92.13	89.89
$D < 4$ Å[b]	88.76	85.40
$D < 3$ Å[b]	79.78	76.40

[a] Number of pockets indicates the number of predicted pockets used for the analysis. If one of the predicted pockets is correct, then the prediction is counted as a successful prediction.
[b] All 89 target proteins were used.

The Sievgene docking score consists of an accessible surface term (mainly hydrophobic interaction), an electrostatic term, a hydrogen bonding term and a van der Waals term. The total score, and the accessible surface, electrostatic, hydrogen bonding and van der Waals terms at the true binding site were 1.29, 1.28, 0.78, 1.32, and 1.26 times larger than those values at the other sites, respectively. The contributions of the accessible surface, electrostatic, hydrogen bonding and van der Waals terms were 89.9%, 0.88%, 5.41% and 3.78%, respectively. On average, 65% of the total accessible surface of the compound was buried in the protein. At the true binding site, 78% of the total accessible surface of the compound was buried in the protein. These results suggest that the surface complementarity and the hydrophobic interactions between the protein and the compounds are important in distinguishing the true binding site from the other sites.

The MolSite method was also applied to the unbound (apo) structures of 20 target proteins, which were the (5*n + 1)th and (5*n + 3)th (where n=0,.., 9) proteins of Table IV of reference 2. The prediction results are summarized in Table VII. To calculate the Dc and Dmin values, the Cα carbons of the unbound protein were superimposed on those of the bound (holo) protein, and the Dc and Dmin values were calculated based on the ligand coordinates of the bound protein. The apo and holo proteins are summarized in Table VII. The C1, C2 and C3 sets were used as the probe set. The scoring grids were selected using the S_{avg} score. The MolSite method worked well for both the unbound and bound structures. The prediction results did not depend on the choice of the probe set. The probabilities of Dc < 4 Å and Dmin < 4 Å were 80% and 100%, respectively. This prediction accuracy for the apo protein was almost the same as that for the holo protein.

Figure 2(A,B) show the ligand-binding site prediction results obtained by the current MolSite method, summarizing the data in Tables III, IV, and VII. It is clear that the prediction results were only slightly dependent on the selected library, and that the prediction performance for apo proteins was similar to that for holo proteins. As compared with the performances of other methods (i.e., Fig. 2 of

Table VII. *Summary of the Dc and Dmin Values for 20 apo Proteins*

Library Distance Apo	Holo	C1 Dc RMSD (Å)	C1 Dmin RMSD (Å)	C2 Dc RMSD (Å)	C2 Dmin RMSD (Å)	C3 Dc RMSD (Å)	C3 Dmin RMSD (Å)
1hel	1hew	4.85	1.04	4.86	1.03	4.86	1.03
1hsi	1ida	3.02	2.03	3.03	2.05	2.99	2.03
1krn	2pk4	2.48	1.29	2.48	1.28	2.47	1.28
1pdy	1pdz	1.66	0.90	1.67	0.91	1.67	0.91
1stn	1snc	4.30	0.50	4.31	0.49	4.33	0.49
1swb	1stp	7.41	1.23	7.41	1.22	7.40	1.20
3app	1apu	3.08	1.11	3.08	1.11	3.05	1.13
3p2p	5p2p	0.50	0.98	0.52	0.97	0.52	0.97
3tms	1bid	0.60	0.99	0.62	0.98	0.57	1.00
5dfr	4dfr	2.00	1.72	1.97	1.72	2.03	1.70
1mrg	1ahc	2.59	1.37	2.58	1.37	2.58	1.36
1blh	1djb	3.26	1.90	3.24	1.90	3.24	1.91
1inc	1esa	3.51	0.88	3.47	0.86	3.51	0.88
1dwd	1hxf	4.15	0.66	4.17	0.66	4.16	0.68
2ifb	1ifb	1.15	1.90	1.16	1.91	1.15	1.92
2tmn	1l3f	2.36	0.95	2.38	0.96	2.38	0.97
1pso	1psn	3.17	0.55	3.19	0.56	3.18	0.57
5cna	2ctv	1.56	0.26	1.58	0.25	1.60	0.23
1stp	2rta	0.90	1.28	0.90	1.28	0.88	1.27
6rsa	7rat	2.45	0.90	2.46	0.89	2.45	0.89
	Average	2.75	1.12	2.75	1.12	2.75	1.12

This prediction is based on the S_{avg} score.

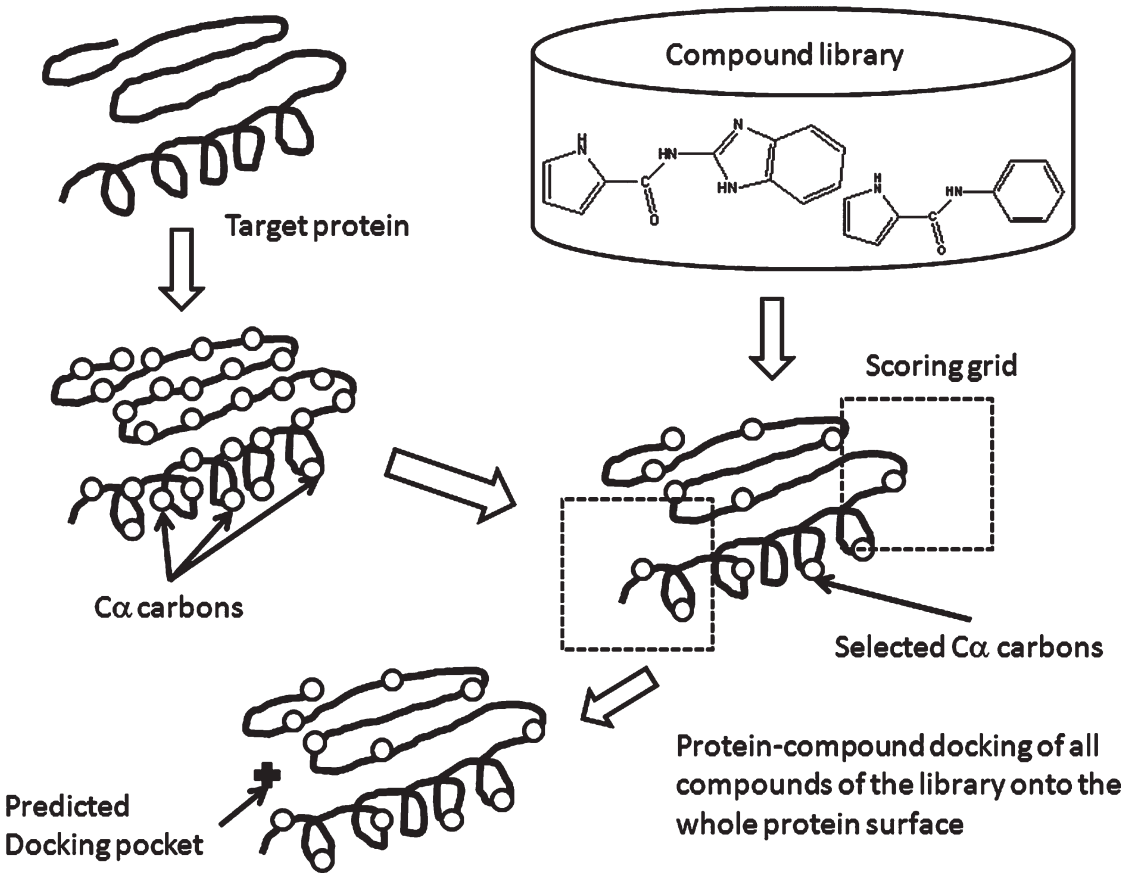

Figure 1. Schematic representation of the molecular-docking binding-site finding (MolSite) method.

reference 28), the performance of MolSite is equivalent or superior to those of other methods, although the data sets were different.

We also examined the dependence of the number of compounds on the prediction accuracy. For the 20 apo structures, 10, 100, and 1,000 randomly selected compounds were used, instead of the 10,000 compounds. The prediction accuracy was improved by increasing the number of compounds. Namely, the probabilities of Dc $<$ 4 Å were 60%, 60%, 70% and 70% for 10, 100, 1,000 and 10,000 compounds, respectively. The probabilities of Dc $<$ 5 Å were 70%, 80%, 80% and 90% for 10, 100, 1,000 and 10,000 compounds, respectively. The average Dc values were 3.34 Å, 3.52 Å, 3.42 Å and 2.99 Å for 10, 100, 1,000 and 10,000 compounds, respectively. The MolSite method could work with even 10 compounds, and the accuracy was not drastically improved with more compounds. Therefore, 10^4 compounds should be sufficient to use the MolSite method effectively.

When only one ligand of the target protein was used, the MolSite method still worked, but failed in some cases. The ligands were prepared for 16 apo structures (see Appendix D). The probabilities of Dc $<$ 4 Å were 75% and Dmin $<$ 4 Å were 81.25%. However, for three target proteins (1djb, 1esa and 1l3f), the MolSite method failed in prediction (Dc $>$ 20 Å and Dmin $>$ 18 Å).

Prediction of ligand-binding affinity

Figure 3 shows a comparison between the experimental ΔG values and the S_{avg} values. The PDB identifiers of the 50 proteins used are summarized in Appendix B (protein numbers 40–89 in Tables III and IV). When all 50 results were used, there was almost no correlation between the experimental ΔG value and the S_{avg} or S_1 value. The correlation coefficients of S_{avg} and S_1 to the experimental ΔG values were 0.195 and 0.186, respectively. As shown in Tables III and IV, the prediction failed in several cases, but all of the results were used to calculate the correlation. The docking scores were distributed around 3–4, suggesting that the difference between the strongest and weakest affinities was about 30%. On the contrary, the experimental ΔG values ranged from -2 kcal/mol to -18 kcal/mol. Thus, the ΔG value cannot be predicted well by any score obtained by the MolSite method. The correlation coefficients of S_{avg} and S_1 were higher than those of the S_{top} score.

The volumes of the ligands in this dataset were widely distributed. As shown in Table I, the smallest

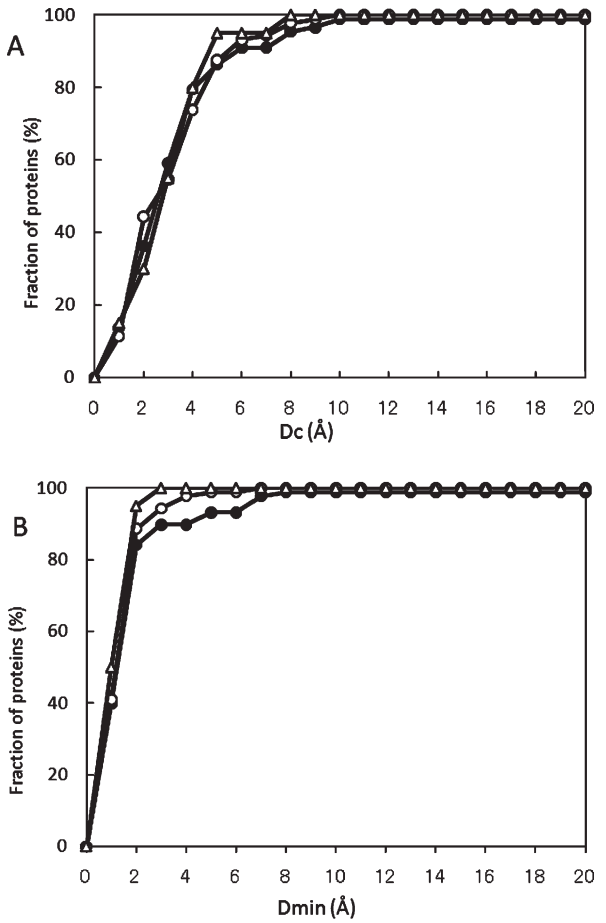

Figure 2. Performance of the MolSite method. The results are presented as the fraction of proteins with Dc and Dmin. Filled circles, open circles and open triangles represent the results of all 89 holo target proteins with the C1 library, those with the small compound set, and those of the 20 apo target proteins with the C1 library, respectively. (A) Fraction of proteins vs. Dc value, (B) fraction of proteins vs. Dmin value.

number of heavy atoms in a ligand was only 5, and the largest number was 46. In contrast, the smallest number of heavy atoms in the compounds was 7-9, and the largest number of heavy atoms was 26. If the pocket is too large, as compared with the compounds of the decoy set, then those probes cannot estimate the protein-compound interaction of the whole pocket. Moreover, if the MolSite method fails in pocket prediction, then the docking score is not meaningful for estimating the protein-compound interaction of the true pocket. In this analysis, ligands with HA $\leq$ 26 were selected, and the prediction cases with Dc $<$ 4 Å were chosen. Then, 22 target proteins (complex structures) were selected for the ΔG prediction.

Figure 4 shows a comparison between the experimental ΔG values and the S_{avg} values for the 22 selected target proteins. A weak correlation appears between the experimental ΔG value and the S_{avg} or S_1 value. Namely, the correlation coefficient between ΔG

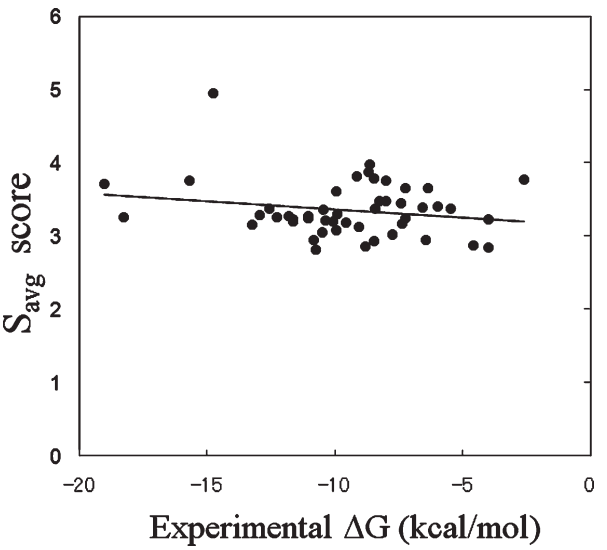

Figure 3. Correlation between the experimental ΔG value and the S_{avg} value obtained by the Sievgene protein-compound docking program for all 50 target proteins. The solid line represents the linear regression result by the least-squares fit.

and S_{avg} was 0.441 (linear regression yielded $\Delta G = -3.429\ S_{avg} + 3.559$), and that between ΔG and S_1 was 0.425 (linear regression yielded $\Delta G = -3.306\ S_1 + 3.201$). The S_{avg} score was slightly better than the S_1 score for the affinity prediction, even though the true binder was unknown. When the volume of the binding pocket was limited, the affinity of the binding pocket could be predicted, but the accuracy was not very high.

Discussion

In Table III, the MolSite method failed to predict the binding sites of 4phv, 1abe, 2r04, 1abf and 5abp.

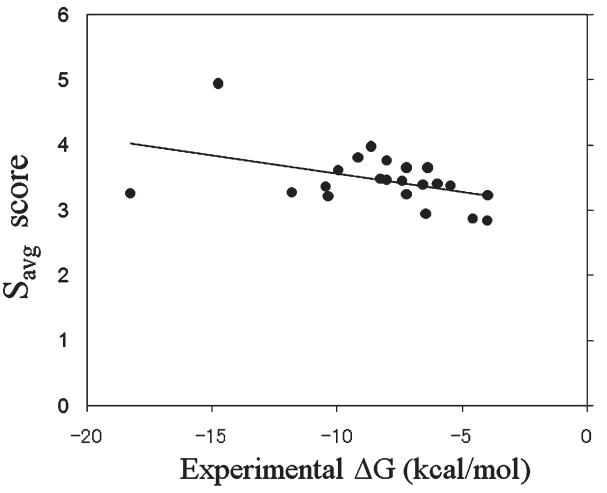

Figure 4. Correlation between the experimental ΔG value and the S_{avg} value obtained by the Sievgene protein-compound docking program for the selected target proteins. The solid line represents the linear regression result by the least-squares fit.

Three of these—1abe, 1abf and 5abp—are sugar-binding proteins, with small binding pockets buried in the proteins. The ligands of 1abe, 1abf and 5abp have only 5–8 heavy atoms. The compounds of sets C1–C3 are much larger than sugars, and no members of our probe set could bind to the pocket correctly. All of the compounds were docked onto the protein surface, and no docking poses were generated for the actual buried binding pocket. We hypothesized that if the probe set had consisted of small compounds with sizes similar to those of sugars, then the MolSite method could have predicted the true binding pocket.

Therefore, we used the set of small compounds as the probe set, and the results are shown in Table IV. The small binding pockets in 1abe, 1abf, and 5abp were well predicted. In addition, the prediction of the buried pocket in 4phv was greatly improved. Therefore, when the ligand for a target protein is known to be small, we can use the Small set for better prediction accuracy.

For *in silico* screening, the conditions of Dc <4 Å and Dmin <4 Å proved too severe. As shown in Table II, 98.89% of the ligand-binding sites were included in the selected scoring grids by the S_{avg} score. This prediction of the scoring grid should be sufficiently high for *in silico* screening. Such high accuracy should be required for docking pose analysis.

A comparison of Figure 3 with 4 reveals that the ΔG values show better correlation with the S_{avg} or S_1 values for smaller ligands with HA $\leq$ 26, rather than larger ligands with HA $>$ 26. The molecular size of the C1–C3 sets was generally smaller than that of the ligands of the 50 protein-ligand complexes. Namely, the average number of heavy atoms of the C1–C3 sets was 20.4, and the range of the number of heavy atoms (HA) was 7–9 $\leq$ HA $\leq$ 26. If the probe set consisted of large compounds, then the S_{avg} or S_1 value would show better correlation to the experimental ΔG than that in the current study. The precise reproduction of these ΔG values would be difficult, since the ΔG value depends on the ligand that binds the same pocket. The ligand efficiency (LE) has been proposed as a measure of druggability, with LE = -ΔG /HA.[35,36] The LE values of known drugs range from 0.1 kcal/mol/atom to 0.7 kcal/mol/atom, and the average LE value is 0.4 kcal/mol/atom.[35,36] In the current study, the LE values ranged from 0.104 kcal/mol/atom to 1.640 kcal/mol/atom. Three LE values of the current data were extremely high (LE $>$ 0.9 kcal/mol/atom). The ΔG value of a compound with an extremely low LE value could be improved by some chemical modifications, but we could not determine the maximum affinity. The maximum affinity of the pocket is not defined well enough to allow the prediction of the affinity without a known active compound.

Our previous work showed that the ΔG value predicted by our own docking program, Sievgene, is not highly accurate, with a correlation coefficient between the experimental and predicted ΔG values of about 0.7, which is the same as those of other docking programs.[37] The previous work was based on exactly the same 50 protein-compound complex structures used in the current study. This accuracy is the upper limit by a naïve docking study. In the current study, the correlation coefficient between the experimental and predicted ΔG values was about 0.4, without using the true ligands (binders). A correlation coefficient of 0.4 is not very high, but it may not be so bad, considering the upper limit, 0.7.

One of the drawbacks of the MolSite method is the computational time. The MolSite method is a sort of ensemble docking study. When the protein surface is divided into 50 scoring grids and the library consists of 10,000 compounds, the total number of dockings is 50 $\times$ 10,000 dockings. Thus, the total CPU time is 1,000,000 seconds (=278 hours) for one target protein by one processor. However, completely parallel computation can easily be performed, by distributing the individual docking procedures to different CPUs. Therefore, despite this drawback, the MolSite method is useful for reliably predicting the possible ligand-binding sites for a target protein.

The MolSite method can easily be improved by increasing the compound database, and its prediction accuracy is high. Once the ligand-binding pocket is predicted, the subsequent *in silico* drug screening docks millions of compounds. In recent ensemble docking studies, multiple target-protein structures have been used, and the CPU time required for *in silico* screening is much longer than that needed for the MolSite method.

In the current study, Sievgene was adopted as the protein-compound docking program. Many protein-compound docking programs have been reported,[38–40] and there is no clearly superior method.[41] In the current study, we used the ordinary Sievgene score, rather than the special function. Thus, other protein-compound docking programs besides Sievgene could be used for the MolSite method.

Materials and Methods

Protein-compound docking on the whole protein surface was performed using a random compound library, and the ligand-binding sites were predicted based on the docking scores of these compounds. We assumed that the ligand-binding site would show stronger affinity to a compound than the other regions do, even if the compound was not the true binder of the site. Figure 1 provides a schematic representation of this MolSite method.

Since it is too time-consuming for a docking program to dock a compound on the whole protein surface, the entire protein surface was first divided into many small grid boxes for the docking procedure with the random compound library. The centers of these boxes were set to the positions of the Cα carbons. The grid boxes are called the "scoring grids" hereafter. The Cα carbons were selected to reduce the computational time. The minimum distance between two Cα carbons was set at 8 Å. The Cα carbon of the first residue was adopted first. If two Cα carbons were closer than 8 Å, then the Cα carbon belonging to the latter residue was neglected. As a result, 10–53 Cα carbons were selected for the target proteins.

The scoring grid was a cubic region with a cell size of 30 × 30 × 30 Å^3, composed of a 3D mesh with grid points separated by 1 Å, each with the potential energy for ligand-binding to the target protein. Neighboring grids overlapped. All compounds of the random library were docked in these scoring grids.

To evaluate the "binding site likeness" of the grid, we prepared the following three scores.

- Type1: S_{avg}, The average value of all docking scores of compounds of the library.
- Type 2: S_{top}, The best docking score among all docking scores of compounds of the library.
- Type 3: S_1, where $S_1 = S_{avg} + \sigma$, where σ is the deviation of the docking scores, and

$$\sigma = \sqrt{\sum_{i=1}^{N} \frac{(\text{score}_i - S_{avg})^2}{N-1}},$$

  score$_i$ and N are the docking score of the i-th compound and the number of compounds in the library, respectively.

Usually, only one compound among the 10,000 randomly selected compounds is a truly active compound that experimentally shows strong affinity to the target protein. Thus, the protein-compound affinity is given by this truly active compound, among the many compounds. The highest docking score should be closer to the true docking score than the other scores, and docking scores lower than the highest score should be meaningless. Based on this logic, the S_{top} score should be the most reliable indicator.

However, the S_{top} score strongly depends on the choice of the random library. If the library contains a strong binder, then the S_{top} value should be very high. On the contrary, the S_{top} value may not be so high if the library lacks a strong binder. To reduce this library dependency, the average score value and deviation were introduced. Using the S_1 or S_{avg} values, the library dependency in the prediction of the ligand-binding sites could be reduced.

Each of these three scores (S_1, S_{avg}, and S_{top}) was calculated for each scoring grid, and the grid that yielded the best score was selected as the candidate scoring grid that contains the ligand-binding pocket. The center of the predicted docking pocket was then determined by the average coordinates of the docking poses of all compounds in the candidate scoring grid. The predicted affinity was given by each of the three scores (S_1, S_{avg}, and S_{top}) obtained in the candidate scoring grid.

To examine the MolSite method, we performed a protein-ligand docking simulation based on the known complex structures registered in the Protein Data Bank. The protein set consisted of two sets originating from the references. Here, 50 complexes with their experimental binding free energy values were selected from the database that was used for the determination of the ΔG scores of the PRO_LEADS (protein numbers 40–89 in Tables III and IV).[42] Forty-eight complexes were selected from the database that was used for the ligand-binding site prediction of the LIGSITE CSC (protein numbers 1–48 in Tables III and IV).[2] Ten proteins were redundant (protein numbers 40–48 in Tables III and IV). Thus, a total of 89 (=50+48–9) proteins were used in the current study. The PDB identifiers are summarized in Appendix A. All water molecules were removed from the proteins, and all missing hydrogen atoms were added to form all-atom models of the proteins. For ligand-flexible docking, the Sievgene/myPresto program (protein-compound docking program) was used to generate up to 100 conformers for each compound.[43] The Sievgene/myPresto program is available for free, from the web sites http://presto.protein.osaka-u.ac.jp/myPresto4/ and http://medals.jp/myPresto/index.html. Sievgene reconstructed 27.7%, 56.9%, and 66.2% of the total of 180 complexes that were adopted in the previous study[43] with RMSDs < 1 Å, 2 Å, and 3 Å, respectively, and the average computational time was 2 CPU seconds.

Three compound libraries (C1, C2 and C3) were prepared. Each library consisted of 10,000 randomly selected compounds from the LigandBox compound database.[44] The atomic charges of each compound were determined by the Mulliken charge, using MOPAC AM1 quantum chemical calculations (Quantum Chemistry Program Exchange, (QCPE), Indiana University, Bloomington, IN). The molecular weight (MW) of each compound was restricted to 150 Da < MW < 340 Da. The minimum, maximum and average numbers of heavy atoms (HA) are summarized in Table I.

The atomic charges of the proteins were the same as those in AMBER parm99.[45] The minimum, maximum and average numbers of heavy atoms

(HA) of the ligands of those protein-ligand complex structures are also summarized in Table I.

Conclusion

We developed the MolSite method for the prediction of the ligand-binding sites of a target protein. In this method, many compounds in a random library are docked over the whole surface of the target protein, and the ligand-binding site is predicted based on the resulting docking scores. We assumed that the actual ligand-binding sites would show statistically better docking scores and higher affinities to compounds in the random library than other sites do, even if the compounds are not the true ligands.

We applied the MolSite method to 89 known protein-ligand complex structures extracted from the PDB. The ligand-binding sites were predicted for the bound states of these target proteins. The center of the ligand-binding site was defined as the average coordinates of the bound ligand of the original complex structure. The center of the predicted ligand-binding site was defined as the average coordinates of all of the docked compounds of the probe set. The prediction accuracy was measured by the distance between the predicted center of the pocket and the actual center of the original complex structure.

The prediction accuracy of the MolSite method was higher than those of the other methods. Namely, the ligand-binding sites were predicted with 87.5% and 97.9% accuracies for the Dc value < 4 Å and the Dmin value < 4 Å for the bound structures, respectively, when only the single top-ranked site was adopted. The MolSite method worked well for both the unbound and bound structures. We also examined the prediction of the affinity of the ligand-binding site. When the pocket was small, the average docking score showed weak correlation to the experimental binding free energy. The results generated by the MolSite method did not depend on the choice of the compound data set.

Appendix A: Forty-Eight Proteins for Ligand-Binding Site Prediction and Binding Free Energy Estimation

The following PDB identifier complexes were used: 1bid, 1cdo, 1fbp, 1gca, 1hew, 1hyt, 1inc, 1rbp, 5cna, 1a6w, 1acj, 1blh, 1ivd, 1mtw, 1okm, 1phd, 1qpe, 1srf, 2h4n, 2sim, 3gch, 3mth, 5p2p, 1imb, 6rsa, 1rob, 4phv, 1byb, 1hfc, 1ida, 1igj, 1mrg, 1pdz, 1pso, 1rne, 1snc, 2ctc, 2pk4, 1apu, 1dwd, 1stp, 1ulb, 2ifb, 3ptb, 2ypi, 4dfr, 2tmn, and 7cpa.

Appendix B: Fifty Proteins for Binding Free Energy Estimation

The following PDB identifier complexes were used: 1abe, 1abf, 1apu, 1cbx, 1dbb, 1dbj, 1dog, 1dwd, 1ebg, 1epo, 1etr, 1ets, 1ett, 1hpv, 1hsl, 1htf, 1hvr, 1mnc, 1nsd, 1pgp, 1phf, 1phg, 1ppc, 1pph, 1rbp, 1stp, 1tmn, 1tng, 1tnh, 1ulb, 2cgr, 2cpp, 2gbp, 2ifb, 2phh, 2r04, 2tmn, 2tsc, 2ypi, 3ptb, 4dfr, 5abp, 5cpp, 5tln, 6cpa, and 7cpa. For 1abe, 1abf, 5abp, and 1htf, two receptor pockets were prepared, since each of these proteins binds two ligands.

Appendix C: Small Set

The small set consisted of the ligands of the following protein-ligand complex structures: 1mbi, 1tng, 1dwb, 1ebg, 1tnh, 2ypi, 3ptb, 1abe, 2phh, 1abf, 1dog, 1hsl, 1phf, 1ulb, 2cpp, 5cpp, 2gbp and 5abp. For 1abe, 1abf, 5abp, and 1htf, two receptor pockets were prepared, since each of these proteins binds the ligand with two different ligand-binding poses.

Appendix D: Sixteen Proteins for Ligand-Binding Site Prediction Using Only One Ligand Included in the Bound Complex Crystal

The following PDB identifier complexes were used: 1ahc, 1bid, 1djb, 1esa, 1hew, 1hxf, 1ida, 1ifb, 1l3f, 1pdz, 1snc, 2ctv, 2pk4, 2rta, 4dfr, and 5p2p. The used ligands for prediction of these proteins are the ligands of the holo structures summarized in Table VII.

References

1. Huang B (2009) MetaPocket: a meta approach to improve protein ligand binding site prediction. OMICS 13:325–330.
2. Huang B, Schroeder M (2006) LIGSITE^{csc}: predicting ligand binding sites using the Connoly surface and degree of conservation. BMC Struct Biol 6:19–30.
3. Laskowski RA (1995) SURFNET: a program for visualizing molecular surfaces, cavities and intermolecular interactions. J Mol Graph 13:323–330.
4. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM (1996) Protein clefts in molecular recognition and function. Protein Sci 5:2438–2452.
5. Brady GP, Jr, Stouten PFW (2000) Fast prediction and visualization of protein binding pockets with PASS. J Comput Aided Mol Des 14:383–401.
6. Zhong S, MacKerell AD, Jr (2007) Binding response: a descriptor for selecting ligand binding site on protein surfaces. J Chem Inf Model 47:2303–2315.
7. Harris R, Olson AJ, Goodsell DS (2008) Automated prediction of ligand-binding sites in proteins. Proteins Struct Funct Bioinf 70:1506–1517.
8. Cheng AC, Coleman RG, Smyth KT, Cao Q, Soulard P, Caffrey DR, Salzberg AC, Huang ES (2007) Structure-based maximal affinity model predicts small-molecule druggability. Nat Biotechnol 25:71–75.
9. Weisel M, Proschak E, Schneider G (2007) PocketPicker: analysis of ligand binding-sites with shape descriptors. Chem Central J 1:7.
10. Liang J, Edelsbrunner H, Woodward C (1998) Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci 7:1884–1897.
11. Peters KP, Fauck J, Frommel C (1996) The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. J Mol Biol 256:201–213.

12. Xie L, Bourne PE (2007) A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. BMC Bioinformatics 8:S9.
13. Hendlich M, Rippman F, Barnickel G (1997) LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model 15:359–363.
14. Kalidas Y, Chandra N (2008) PocketDepth: a new depth based algorithm for identification of ligand binding sites in proteins. J Struct Biol 161:31–42.
15. Coleman RG, Sharp KA (2010) Protein pockets: inventory, shape, and comparison. J Chem Inf Model 50:689–603.
16. Kim D, Cho CH, Cho Y, Ryu J, Bhak J, Kim DS (2008) Pocket extraction on proteins via the Voronoi diagram of spheres. J Mol Graphics Model 26:1104–1112.
17. Kahraman A, Morris RJ, Laskowski RA, Thornton, JM (2007) Shape variation in protein binding pockets and their ligands. J Mol Biol 368:283–301.
18. Lichtarge O, Sowa ME (2002) Evolutionary predictions of binding surfaces and interactions. Curr Opin Struct Biol 12:21–27.
19. Glaser F, Morris RJ, Najmanovich RJ, Laskowski RA, Thornton JM (2006) A method for localizing ligand binding pockets in protein structures. Struct Funct Bioinf 62:479–288.
20. Chen BY, Bryant DH, Fofanov VY, Kristensen DM, Cruess AE, Kimmel M, Lichtarge O, Kavraki LE (2007) Cavity scaling: automated refinement of cavity-aware motifs in protein function prediction. J Bioinf Comp Biol 5:353–382.
21. Joughin BA, Tidor B, Yaffe MB (2005) A computational method for the analysis and prediction of protein: phosphopeptide-binding sites. Protein Sci 14:131–139.
22. Pettit FK, Bare E, Tsai A, Bowie JU (2007) HotPatch: a statistical approach to finding biologically relevant features on protein surfaces. J Mol Biol 369:863–879.
23. Kawabata T (2010) Detection of multiscale pockets on protein surfaces using mathematical morphology. Proteins Struct Funct Bioinf 78:1195–1211.
24. Kawabata T, Go N (2007) Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites. Proteins Struct Funct Bioinf 68:516–529.
25. Laurie ATR, Jackson RM (2005) Q-siteFinder: an energy-based method for prediction of protein-ligand binding sites. Bioinformatics 21:1908–1916.
26. Ming D, Cohn JD, Wall ME (2008) Fast dynamics perturbation analysis for prediction of protein functional sites. BMC Struct Biol 8:5.
27. Coleman RG, Salzberg AC, Cheng AC (2006) Structure-based identification of small molecule binding sites using a free energy model. J Chem Inf Model 46:2631–2637.
28. Brylinski M, Skolnick J (2008) A threading-based method FINDSITE for ligand-binding site prediction and functional annotation. Proc Natl Acad Sci USA 105:129–134.
29. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput Biol 5:e1000585.
30. Soga S, Shirai H, Kobori M, Hirayama N (2007) Use of amino acid composition to predict ligand-binding sites. J Chem Inf Model 47:400–406.
31. Kinjo AR, Nakamura H (2009) Comprehensive structural classification of ligand binding motifs in proteins. Structure 17:234–246.
32. Hoffmann B, Zaslavskiy M, Vert JP, Stoven V (2010) A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction. BMC Bioinformatics 11:99.
33. Abagyan R, Kufareva I (2009) The flexible pocketome engine for structural chemogenomics. Methods Mol Biol 575:249–279.
34. Smith AJT, Zhang X, Leach AG, Houk KN (2009) Beyond picomolar affinities: quantitative aspects of noncovalent and covalent binding of drugs to proteins. J Med Chem 52:225–233
35. Orita M, Ohno K, Niimi T (2009) Two "Golden ratio" indices in fragment-based drug discovery. Drug Discov Today 14:321–328.
36. Abad-Zapatero C, Metz JT (2005) Ligand efficiency indices as guideposts for drug discovery. Drug Discov Today 10:464–469.
37. Fukunishi Y, Mikami Y, Kubota S, Nakamura H (2005) Multiple target screening method for robust and accurate in silico ligand screening. J Mol Graph Model 25: 61–70.
38. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE (1982) A Geometric approach to macromolecule-ligand interactions. J Mol Biol 161:269–288.
39. Rarey M, Kramer B, Lengauer T, Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. J Mol Biol 261:470–489.
40. Jones G, Willet P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267:727–748.
41. Warren GL, Webster Andrews C, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS (2006) A critical assessment of docking programs and scoring functions. J Med Chem 49:5912–5931.
42. Baxter CA, Murray CW, Clark DE, Westhead DR, Eldridge MD (1998) Flexible docking using tabu search and an empirical estimate of binding affinity. Proteins 33:367–382.
43. Fukunishi Y, Mikami Y, Nakamura H (2005) Similarities among receptor pockets and among compounds: analysis and application to in silico ligand screening. J Mol Graph Model 24:34–45.
44. Fukunishi Y, Sugihara Y, Mikami Y, Sakai K, Kusudo H, Nakamura H (2009) Advanced in-silico drug screening to achieve high hit ratio-development of 3D-compound database. Synthesiology 2:64–72.
45. Case DA, Darden TA, Cheatham TE, III, Simmerling CL, Wang J, Duke RE, Luo R, Merz KM, Wang B, Pearlman DA, Crowley M, Brozell S, Tsui V, Gohlke H, Mongan J, Hornak V, Cui G, Beroza P, Schafmeister C, Caldwell JW, Ross WS, Kollman PA (2004) AMBER 8, UCSF.