Improved conformational search for protein-ligand docking based on optimal arrangement of multiple small search grids

Tomohiro Ban[1,2], **Takashi Ishida**[1], **and Yutaka Akiyama**[1,2]
[1]Department of Computer Science, Graduate School of Information Science,
Tokyo Institute of Technology, W8-76, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8550, JAPAN
[2]Education Academy of Computational Life Sciences,
Tokyo Institute of Technology, W8-93, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8550, JAPAN

Abstract— *The Glide protein-ligand docking algorithm often fails to find the correct binding mode. This is because the search process can easily fall into local minima when the search target area is widely distributed across the protein's surface and the search grid is relatively large. In this research, we propose a novel method that improves the search efficiency in such cases by dividing a single, large search grid into multiple small search grids. In addition, we propose a method to minimize the number of small grids by converting the problem into a set cover problem. We present experimental results to compare the performance of the proposed approach with that of the standard protocol under two different settings.*

Keywords: protein-ligand docking, Glide, set cover problem, conformation search

1. Introduction

The technique of protein-ligand docking aims to predict the binding mode of a protein and a small chemical compound (ligand) from their three-dimensional structures. This approach is now used in many fields, such as drug discovery and molecular biology [1] [2]. To date, various research groups from both commercial and academic organizations have developed protein-ligand docking software, such as AutoDock [3], GOLD [4], FlexX [5], and Glide[6]. In particular, Glide has demonstrated good accuracy using various benchmarks, and is recognized as one of the best docking software applications [7] [8] [9]. However, even Glide does not always return the correct binding mode. Therefore, an improvement in the accuracy of protein-ligand docking is highly desirable and would have a significant positive impact in various fields. The low prediction accuracies given by protein-ligand docking software are often caused by two substantial problems. One is the estimation of the binding free energy, and the other is the problem of searching the whole conformational space. The former problem is caused by the coarse model resolution and simplified potential energy function, which are intended to reduce the computational cost. The latter problem is a result of the huge number of conformations to be searched. In particular, this problem becomes more serious if the binding site of a target protein is unknown. This is because only a narrow region need be searched if the binding site is well known; if this is not the case, the entire protein surface must be searched. Thus, the conformational space search requires more computational resources, and this can become a serious problem.

To tackle this, several software packages, such as PocketFinder [10] and SiteMap [11], have been developed to predict the ligand binding sites. In the standard Glide docking protocol, multiple binding sites are predicted from the tertiary structure of a protein using SiteMap, and then a search grid is set to cover these predicted binding sites. Finally, only the region within the grid is searched in the Glide docking simulation process. However, even using this protocol, Glide sometimes fails to find the correct binding mode. The search easily falls into local minima if the predicted binding sites are widely distributed across the protein's surface and a large search grid is used. The search algorithm of Glide tends to intensively search narrow regions near positions that score highly in the initial search stage, and overlook good conformations far from such regions. Therefore, the search accuracy of Glide often becomes lower if a large number of binding sites are predicted over a widespread area.

In this study, we propose a method to improve the search efficiency of protein-ligand docking when many binding sites are predicted and the search grid is large. To avoid the problem of local minima, we use multiple small search grids instead of one large grid. Additionally, to minimize the number of small search grids, we translate this arrangement into a set cover problem, and successfully reduce the number of grids.

2. The Glide conformation search algorithm

The Glide search algorithm [6] is a four-part process that determines the conformation with the lowest binding free energy. In the first stage, the algorithm uses simple criteria to determine candidate positions on the protein that are likely to bind with a ligand. In the second stage, the algorithm arranges ligands at these points, and calculates their binding score using a rough score function. In the third

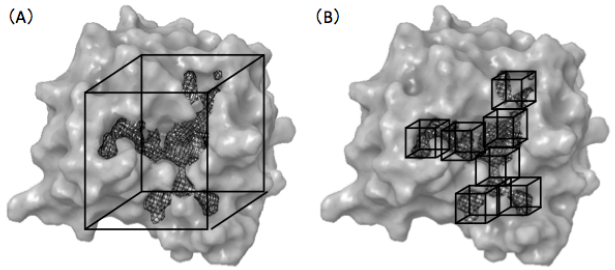

Figure 1: (A) A grid in the Glide standard protocol (B) Grids generated by the proposed method

stage, to minimize the binding free energy, the algorithm optimizes the structure of the ligand by dihedral angle rotation and rigid body transformation. In the final stage, the algorithm selects the best score conformation using a precise score function named GlideScore [6]. In particular, the second stage consists of two different processes. The first is the calculation of a GreedyScore, and the other is a refinement process. In the GreedyScore calculation process, ligands are arranged at the positions selected in the first stage, and the top 5000 conformations are selected according to their ChemScore [12]. In the refinement process, these 5000 conformations are refined by moving the center of the compound within $\pm$ 1Å and the top 400 conformations are finally selected.

The number of selected conformations is a fixed parameter, regardless of the size of a search grid. As a result, the algorithm often fails to find the correct conformation when the search target area is widely distributed over the protein surface and a large search grid is used. Of course, the parameter can be changed manually. However, the range is limited by the interface, and it is difficult to determine an appropriate value empirically.

3. The proposed method

Using the default Glide protein-ligand docking, the conformational search sometimes fails because of insufficient sampling. To solve this problem, we propose a method to improve the search efficiency by dividing a large search grid into multiple small search grids (Figure 1). For a search grid of optimal size, the Glide conformation search algorithm works well, even with the default settings, and we can generally obtain accurate conformations. Thus, in our proposed method, a large search grid is divided into multiple small grids, and then a conformational search is performed for each small grid. Finally, the output of all conformational searches is collated, and the final prediction results are selected according to the GlideScore.

Our proposed method has the clear disadvantage that the computational cost increases in proportion to the number of search grids, meaning that the cost of our method is larger

```
The algorithm of proposed method
 1 :  Let S be the site-points obtained by SiteMap;
 2 :  Let A and B and C be the empty sets;
 3 :  While (Until S is empty) do
 4 :    for (Until select the all elements of S) do
 5 :      Let gc be the selected element of S;
 6 :      for (Until select the all elements of S) do
 7 :        Let s be the selected element of S;
 8 :        if (s is included in the grid whose gc is the center) then
 9 :          Add s into the set A;
10 :        end if
11 :      end for
12 :      if (The number of A is greater than the number of B) then
13 :        Let gc be a candidate of the center of a grid;
14 :        Overwrite A to B;
15 :      end if
16 :      Empty A;
17 :    end for
18 :    Add gc into the set C;
19 :    Remove the all elements of B from S;
20 :  end while
21 :  Return C as a set of the centers of grid.
```

Figure 2: Algorithm of the proposed method

than that of a standard protocol. To reduce this harmful influence, we also propose a grid arrangement method to minimize the number of search grids. We convert this grid arrangement problem into a set cover problem [13]. In the set cover problem, given a table set U made of n elements, a subset group of U expressed as S={S_1,S_2, ... ,S_l}, and a cost function $c : S \rightarrow Q_+$ (Q_+ is a set of positive rational numbers), we must identify the subset of S covering all elements of U with the lowest cost. In our optimal grid arrangement problem, we use the site-points obtained by SiteMap as the table set, and the site-points included on a grid whose center is one of the elements of the table set is the subset group. The cost is the number of elements of the table set included in each grid. In this way, we can convert the grid arrangement problem into a set cover problem. We use an approximate algorithm to solve this, because the set cover problem is known to be NP-hard [14]. The algorithm consists of seven steps: (i) Input the site-points obtained by SiteMap and (ii) prepare the empty set C. Next, (iii) select the highest-cost grid G and (iv) add the center of grid G to C. After that, (v) remove all of the site-points included in grid G, and (vi) repeat (iii)–(v) until S is empty. Finally, (vii) use the site-points in C as the centers of grids in the dispersion setting. Figure 2 shows the pseudo-code of this algorithm. The computational complexity is O(n^3), where n is the number of elements in the table set.

Figure 3 shows the behavior of the algorithm on a two-dimensional space. Both white and black dots represent site-points, and are elements of the table set. The black dots are selected as the center of a search grid by the proposed method, and squares represent each search grid. All of the dots are included in the union of these grids. In particular, the algorithm minimizes the number of grids. In this case, the algorithm successfully covers 30 dots with only 11 grids.

We implemented the proposed method by altering the XGlide.py python script in the Glide cross docking [15].

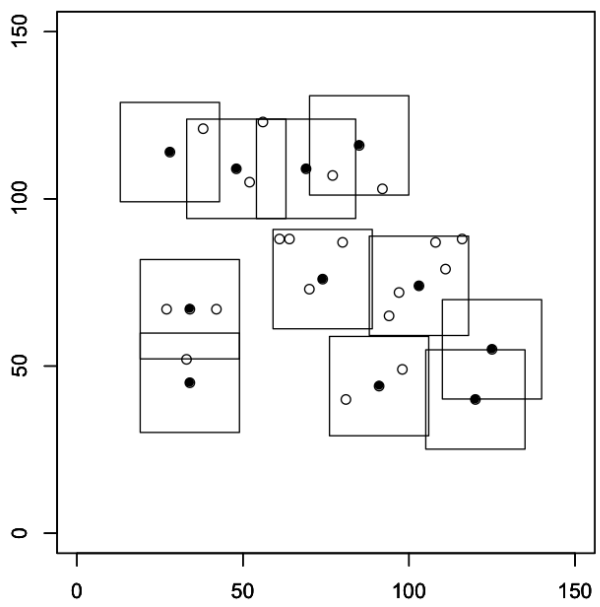

Figure 3: Example 2D grid arrangement given by the proposed method

4. Evaluation experiment

In this experiment, we confirm that the proposed method has better search efficiency than the Glide standard protocol under its default settings. We use the docking score and computation time to evaluate the search efficiency. We also compare the efficiency of the proposed method to that of the Glide standard protocol under the "heavier" setting, which makes the conformation search more onerous but more accurate. This is because a direct comparison of the proposed method and standard protocol with default settings is difficult, as the proposed method has a greater inherent computational complexity.

4.1 Dataset

We used a protein-ligand complex dataset called CCDC/Astex [16]. Because of limitations in computational power and the number of Glide software licenses, we randomly selected the following 20 proteins that did not cause errors in the docking process: 1A4G, 1AJ7, 1B9V, 1DBB, 1EJN, 1FAX, 1FKG, 1HDC, 1IBG, 1MMQ, 1QBR, 1RNE, 1TPH, 1XKB, 2DBL, 2H4N, 2TMN, 2TPI, 3ERD, 7CPA (complex structures 1GPY, 1RT2, and 4CTS were selected at first, but these were replaced by 1EJN, 1FAX, and 2TMN because of such errors). Before applying the docking calculation, the protein-ligand complexes were divided into a protein and a ligand using the Maestro software (Schrodinger, Inc.). The protein structure was optimized by the "Protein Preparation Wizard" within Maestro. This process includes five functions: "Remove cofactors", "Preprocess", "Optimize", "Remove waters", and "Minimize". The potential ligand conformations were generated by the "LigPrep" and "Epik" functions of Maestro.

4.2 Protocol to generate conformation search grids

The conformation search area for the docking simulations is determined based on the results of SiteMap. The SiteMap software predicts potential binding sites based on the protein's structural characteristics. In this experiment, we used SiteMap's default parameters and settings, except for the number of max reports, which was changed from 5 to 10 because the default value is too small for larger proteins.

Search grids were generated by the "Glide Grid Generation" function of Maestro. In the standard protocol, a search grid is located on the centroid of the site-points obtained by SiteMap. The edge size of the INNERBOX (the center of a ligand is restricted to this box through the docking process) is given by the ligand diameter, and the edge size of the OUTERBOX (all atoms of a ligand are restricted to this box) is set to the INNERBOX edge size + 16Å. In the proposed method, search grids are arranged at each of the selected site-points by our grid arrangement algorithm. The edge size of the INNERBOX and OUTERBOX are fixed to 10Å and 26Å, respectively. Therefore, the search grids generated by the proposed method are different from those in the Glide standard protocol. However, both methods satisfy the condition that all site-points given by SiteMap are included in any grid.

4.3 Protein-ligand docking using Glide

The docking results are highly dependent on the initial pose of the ligand. Thus, before the docking simulation, a sufficient number of initial ligand conformations were generated using "LigPrep" with its default settings. The protein-ligand dockings were performed using the "Ligand Docking" Glide function with default settings. Glide has two prediction modes, standard precision (SP) and extended precision (XP). Compared with SP mode, XP is slower but more accurate. In consideration of the computational cost, we used SP to predict the binding mode in this experiment.

As mentioned above, a direct comparison of the efficiency of the proposed method with that of the standard protocol under the default settings is difficult. Thus, we used the "heavier" setting in the standard protocol to enable a reasonable comparison. It is possible to improve the conformation search by increasing the number of searches, although this entails a heavier calculation. Under the heavier setting, the standard protocol forms one grid, as for the default setting. Therefore, we implemented the standard protocol with this heavier setting, and increased the number of conformation searches to that of the proposed method.

Table 1: Performance comparison of three methods

	Standard (default)			Proposed			Standard (heavier)		
PDB	Score [kcal/mol]	RMSD [Å]	time [sec]	Score [kcal/mol]	RMSD [Å]	time [sec]	Score [kcal/mol]	RMSD [Å]	time [sec]
1A4G	-7.34	23.9	5783	-8.08	23.5	9527	-7.34	23.6	25426
1AJ7	-7.33	1.8	1413	-8.02	2.3	2856	-7.71	2.2	1734
1B9V	-6.16	22.9	721	-7.15	23.6	1626	-5.43	23.7	997
1DBB	-8.74	0.5	1244	-9.13	0.5	2671	-8.73	0.5	1418
1EJN	-6.77	12.3	502	-8.84	1.0	712	-7.60	1.1	632
1FAX	-8.47	8.7	727	-8.78	11.1	1117	-9.16	4.4	1521
1FKG	-7.76	1.6	128	-6.81	5.1	120	-7.76	1.6	130
1HDC	-8.07	6.1	956	-7.96	6.1	2443	-8.05	6.1	1550
1IBG	-8.66	2.3	6705	-8.84	1.2	15601	-8.66	2.3	27230
1MMQ	-8.15	9.5	185	-7.58	9.7	310	-8.24	1.5	233
1QBR	-8.39	10.5	1108	-8.39	11.6	1427	-11.23	1.8	1278
1RNE	-13.64	1.5	43850	-15.57	0.6	75126	-13.64	1.5	68986
1TPH	-6.48	1.1	315	-6.26	1.2	460	-6.48	1.1	363
1XKB	-7.78	9.0	923	-11.52	2.1	1621	-11.51	2.0	1528
2DBL	-8.67	1.1	2082	-9.02	1.1	4682	-8.67	1.1	5865
2H4N	-5.02	6.3	526	-5.32	15.7	728	-5.03	6.3	809
2TMN	-5.23	2.6	469	-5.66	3.1	636	-5.97	4.2	664
2YPI	-8.16	0.8	513	-7.90	3.7	1083	-7.99	1.0	632
3ERD	-9.87	0.5	541	-9.95	0.6	804	-9.87	0.5	630
7CPA	-8.21	4.5	953	-9.21	4.5	1698	-8.71	4.2	1691
Average	-7.95	6.4	3482	-8.50	6.4	6262	-8.39	4.5	7166

4.4 Results of the evaluation experiment

Table 4.4 shows the docking scores, root mean square deviation (RMSD), and execution time for the proposed method and standard protocol with the default and heavier settings. The docking score is essentially the same as GlideScore, but is compensated by Epik state penalties [19]. Conformation searches are performed using GlideScore in the docking process, but the final output of Glide is a docking score. Therefore, we used the docking score as an evaluation metric in this experiment. This score represents the binding energy between a protein and a ligand, and so smaller values are better. The "Score" column shows the value of the lowest docking score. We also show the RMSD of all atoms superposed by a protein between the conformation of the crystal structure and the conformation of the complex with the best docking score. RMSD is often used to evaluate the accuracy of dockings. However, we did not use RMSD to measure the conformational search performance, because in many cases a better docking score has a larger RMSD. This is because the RMSD is highly dependent on the scoring function as well as the search performance. Therefore, we only used the docking score to evaluate the conformation search performance in this work.

From the results in Table 4.4, we can see that the proposed method exhibits the best search performance of the three methods considered. In addition, the docking score of the proposed method is better than that of the standard protocol with default settings for 15/20 complexes, and outperforms the standard protocol with the heavier setting in 13/20 cases.

The execution time of each method is shown in the "Time" column. This includes the time required by the proposed method to determine the optimal grid arrangement, as this is trivial compared with the overall execution time. From Table 4.4, we can see that the execution time of the proposed method is approximately twice that of the standard protocol with default settings. However, the proposed method is approximately 15% faster than the standard protocol with the heavier setting.

5. Discussion

5.1 Statistical significance of the improvement

The proposed method gives the best average docking score, and beats the docking score of the standard protocol with default settings in 75% of cases. Thus, we believe that the search performance of the proposed method is considerably better than that of standard protocols. To confirm this, we conducted a statistical test to check whether the difference is significant. Assuming non-parametric distributions, we applied a two-sample paired Wilcoxon signed rank test

[20] to the docking scores. This is a non-parametric statistical hypothesis test to assess the significance of differences between two related samples. We used the "wilcox.test" function of R 3.0.0 with the "pair" option.

First, we compared the standard protocol with default settings with the proposed method. The p-value of the test was 0.02, and the difference in performance was found to be statistically significant at the 0.05 level. Thus, the proposed method has significantly better conformational search performance, although its computational cost is greater.

We also compared the results from the standard protocol with the heavier setting with those given by the proposed method. The p-value of this test was 0.41, indicating that there is no significant difference in performance at the 0.05 level. Thus, from this experiment, it is impossible to conclude that the search performance of the proposed method is better than that of the standard protocol with the heavier setting. However, the proposed method is faster, and thus more efficient, than the standard protocol with the heavier setting.

5.2 The effect of optimal grid arrangement

Our grid arrangement algorithm is designed to minimize the number of search grids. In this experiment, arranging a grid for every site-point obtained by SiteMap would require an average of 4893.6 grids. However, using our optimal grid arrangement algorithm, this number decreased to only 14.2. Because the computational cost increases in proportion to the number of search grids, our grid arrangement method reduces the cost by a factor of approximately 350.

Of course, it is possible to employ other grid arrangement methods. To show the advantage of our method, we implemented another simple grid arrangement method that divides the large grid into small uniform grids at even intervals. Figure 4 shows an example arrangement given by this division method. In the figure, the white dots are site-points obtained by SiteMap, and the small crosses denote the centers of each search grid. We removed all grids that did not include any site-points. The union of the grids generated by the algorithm can also include all dots. We applied this arrangement method to the dataset used in the experiment. This algorithm generated an average of 22.7 search grids, which is more than in the proposed method. These results indicate that our proposed method can effectively decrease the number of search grids, and thus the computational cost.

6. Conclusion

In this study, we aimed to improve the conformation search of protein-ligand docking by avoiding local minima in large search areas. Thus, we proposed a method to improve the search efficiency by dividing one large search gird into multiple small search grids. In addition, we developed a technique that minimizes the number of such grids by converting the problem into a set cover problem. The results

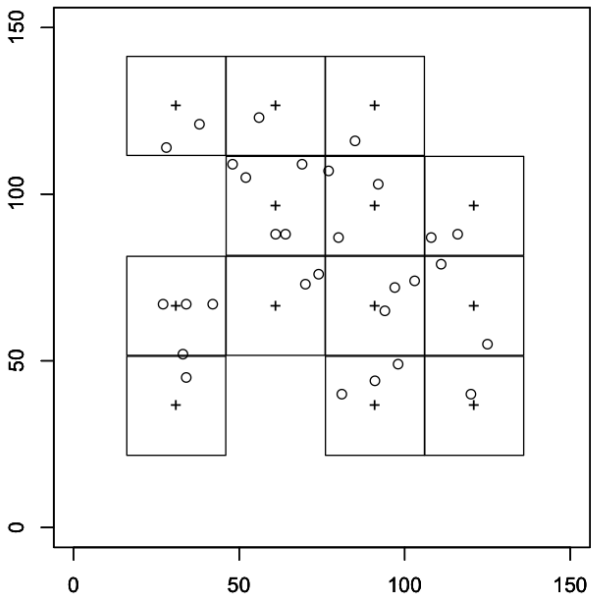

Figure 4: Example 2D grid arrangement given by the simple division method

of an evaluation experiment show that the proposed method improves the docking score relative to the standard protocol. Unfortunately, however, statistical tests did not show a clear improvement over the standard protocol with the heavier setting. The computational cost of the proposed method was lower than that of the standard protocol with the heavier setting, which indicates that our method has better search efficiency than the standard protocol. In this research, the standard protocol with the heavier setting predicts the binding mode of a crystal structure better than the proposed method. We think this is due to the inaccuracy of the docking score. Thus, in future work, we will investigate the relationship between the docking score and the RMSD, and refine the score function to improve conformational searches.

7. ACKNOWLEDGMENTS

This work was supported in part by the Education Academy of Computational Life Sciences (ACLS) at Tokyo Institute of Technology. The authors would like to thank the TSUBAME supercomputer system at the Global Scientific Information and Computing Center, Tokyo Institute of Technology.

References

[1] Xie, Li, et al. "Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors." PLoS Computational Biology 5.5 (2009): e1000387.

[2] Ramirez, Ursula D., et al. "Docking to large allosteric binding sites on protein surfaces." Advances in Computational Biology. Springer New York, 2010. 481-488.

[3] Goodsell, David S., Garrett M. Morris, and Arthur J. Olson. "Automated docking of flexible ligands: applications of AutoDock." Journal of Molecular Recognition 9.1 (1996): 1-5.
[4] Jones, Gareth, et al. "Development and validation of a genetic algorithm for flexible docking." Journal of Molecular Biology 267.3 (1997): 727-748.
[5] Rarey, Matthias, et al. "A fast flexible docking method using an incremental construction algorithm." Journal of Molecular Biology 261.3 (1996): 470-489.
[6] Friesner, Richard A., et al. "Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy." Journal of Medicinal Chemistry 47.7 (2004): 1739-1749.
[7] von Korff, Modest, Joel Freyss, and Thomas Sander. "Comparison of ligand-and structure-based virtual screening on the DUD data set." Journal of Chemical Information and Modeling 49.2 (2009): 209-231.
[8] Kellenberger, Esther, et al. "Comparative evaluation of eight docking tools for docking and virtual screening accuracy." Proteins: Structure, Function, and Bioinformatics 57.2 (2004): 225-242.
[9] Marcou, Gilles, and Didier Rognan. "Optimizing fragment and scaffold docking by use of molecular interaction fingerprints." Journal of Chemical Information and Modeling 47.1 (2007): 195-207.
[10] Laurie, Alasdair TR, and Richard M. Jackson. "Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites." Bioinformatics 21.9 (2005): 1908-1916.
[11] Halgren, T. A. : "New Method for Fast and Accurate Binding-site Identification and Analysis", Chem. Biol. Drug Des., 2007, 69, 146.
[12] Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J. Comput.-Aided Mol. Des. 1997, 11, 425-445.
[13] Vijay V. Vazirani., "Approximation algorithms" Springer p15-p26
[14] M.R.Garey and D.S. Johnson, "Computers and Intractability: A Guide to the Theory of NP-Completeness", W.H. Freeman and Company (1979)
[15] http://www.schrodinger.com/scriptcenter/#Docking
[16] Nissink, J. Willem M., et al. "A new test set for validating predictions of protein-ligand interaction." Proteins: Structure, Function, and Bioinformatics 49.4 (2002): 457-471.
[17] http://helixweb.nih.gov/schrodinger-2013.3-docs/glide/glide_user_manual.pdf, p58
[18] http://www.schrodinger.com/kb/348
[19] http://www.schrodinger.com/newsletter/12/68/
[20] Wilcoxon, Frank. "Individual comparisons by ranking methods." Biometrics bulletin 1.6 (1945): 80-83.