Docking into Knowledge-Based Potential Fields: A Comparative Evaluation of DrugScore

Christoph A. Sotriffer,* Holger Gohlke, and Gerhard Klebe

Department of Pharmaceutical Chemistry, Philipps-University Marburg, Marbacher Weg 6, D - 35032 Marburg, Germany

Received January 18, 2002

Abstract: A new application of DrugScore is reported in which the knowledge-based pair potentials serve as objective function in docking optimizations. The Lamarckian genetic algorithm of AutoDock is used to search for favorable ligand binding modes guided by DrugScore grids as representations of the protein binding site. The approach is found to be successful in many cases where DrugScore-based re-ranking of already docked ligand conformations does not yield satisfactory results. Compared to the AutoDock scoring function, DrugScore yields slightly superior results in flexible docking.

Introduction. The accurate prediction of protein−ligand interaction geometries is essential for the success of virtual screening approaches in structure-based drug design. It requires docking tools that are able to generate suitable configurations and conformations of a ligand within a protein binding site and scoring functions that appropriately translate interaction geometries into an energetic measure describing the quality of the interaction.[1−5] While the first aspect has been considered an almost solved problem based on the results of a public docking competition,[6] much more attention is currently devoted to the second aspect of scoring and affinity prediction.[7−9] As an approach to the latter, the knowledge-based scoring function DrugScore has recently been developed.[10,11] It has been derived by converting data from crystal structures of 1376 protein−ligand complexes of the PDB[12] (as stored in RELIBASE[13,14]) into distance-dependent pair preferences. Using a test set of 158 PDB complexes, a first evaluation of DrugScore has been carried out by re-ranking ("post-scoring") docking results generated by the program FlexX.[15] Although this improved the ranking of the results remarkably,[10] it was also observed that in more than 25% of the cases the top-ranked result showed an rmsd (root-mean-square deviation) of more than 2 Å from the experimental position while at the same time the crystallographically observed binding mode received a better score than any of the docked configurations. This clearly indicates that appropriate placements of the ligand had frequently not been generated by the docking program—in contrast to the assumption that the generation of suitable conformations is only a minor problem. However, it also suggests that using the scoring function not only for post-scoring but already as objective function during the docking process itself could alleviate the problem and possibly improve the overall success rate.

* Corresponding author. Phone: +49/6421/2825822. Fax: +49/6421/2828994. E-mail: sotriffer@mailer.uni-marburg.de.

Guided by these observations, DrugScore has now been tested as objective function for docking. For this purpose an "energy-driven" docking method was required that is solely based on the direct optimization of an energy criterion or scoring function[16] and does not rely on other geometric or combinatorial rules for generating docked ligand orientations and conformations. Accordingly, AutoDock was selected to test the performance of DrugScore in guiding docking searches. Traditionally, Monte Carlo simulated annealing has been used as optimization algorithm within the AutoDock program.[17,18] In the most recent version available, a so-called Lamarckian genetic algorithm has been implemented as more efficient alternative.[19] To speed-up the energy evaluation during the docking process, a grid-based representation of the binding site is used where every grid point corresponds to an affinity value calculated for the interaction of a probe atom with the protein. In AutoDock, the incorporated empirical free energy function is normally used to calculate the grids. Here, grids were computed on the basis of the DrugScore pair potentials, as outlined in the Methods section.

The new approach was tested on a set of 41 PDB complexes derived from the 200 test cases that had served to evaluate FlexX.[20] This smaller set consists of all those noncovalent complexes already used in the original DrugScore validation,[10] for which DrugScore re-ranking of FlexX-generated positions had given a result with an rmsd of >2 Å at rank 1, yet a better score for the experimental binding mode. To put the new DrugScore-based docking into proper perspective, all complexes were also subjected to docking runs using the regression-based energy function of AutoDock, i.e., the "pure" AutoDock approach. As a side effect, this also represents one of the largest evaluations of the AutoDock method hitherto made public.

Results and Discussion. First indications that DrugScore pair potentials might be suitable to perform grid-based, energy-driven docking came from a hot spot analysis.[11] For the same test set of 158 protein−ligand complexes mentioned above, the spatial coincidence of hot spots (obtained from a grid-based evaluation of DrugScore potentials within the binding site) with experimentally observed ligand atoms of corresponding type was analyzed. Depending on the atom-type classification, overall prediction rates between 74 and 85% were obtained. A visual analysis of contoured hot-spot areas superimposed with the experimental binding mode of the ligand supports this finding.

As an initial test for the general applicability of the approach, rigid docking runs were carried out for the entire test set, using the experimentally determined conformation of the complexed ligand. The purpose of this procedure was to avoid complications arising in flexible docking, such as issues of convergence or the coupling with the intramolecular force field, and to analyze whether DrugScore grids fulfill certain minimum requirements for successful docking. In 90% (37/41) of the test cases, correct predictions were obtained, with top-ranked docking results deviating by less than 1 Å from the experimental reference. In most of

10.1021/jm025507u CCC: $22.00
Published on Web 04/20/2002

Table 1. Classification of Docking Results for the Single Test Cases Given by PDB Code[a]

scoring function	rank 1 result with rmsd < 2 Å	degenerate results with rmsd < 2 Å	top-ranked results with rmsd > 2 Å
DrugScore	**1aha**, 1apt, **1bbp**, 1eap, **1ela**, **1ele**, **1frp**, **1hvr**, **1ida**, **1mmq**, **1nco**, **1phg**, 1poc, **1ppi**, **2cht**, **3hvt**, 4phv, 4tmn	**1did**, 1etr, 1hdc, 1hfc, 1hgj, 1srj, **6rnt**, 8gch	**1bma**, **1eed**, **1elc**, 1eld, **1glq**, **1hef**, 1igj, 1ldm, **1pph**, **1ppl**, **1ppm**, **1rne**, **4hvp**, **7cpa**, **9hvp**
AutoDock	**1aha**, **1bbp**, **1ela**, **1ele**, **1frp**, 1hdc, 1hgj, **1hvr**, **1ida**, 1igj, 1ldm, **1mmq**, **1nco**, **1phg**, **1ppi**, **2cht**, **3hvt**, 8gch	**1did**, 1eld, **6rnt**	1apt, **1bma**, 1eap, **1eed**, **1elc**, 1etr, **1glq**, **1hef**, 1hfc, 1poc, **1pph**, **1ppl**, **1ppm**, **1rne**, 1srj, **4hvp**, 4phv, 4tmn, **7cpa**, **9hvp**

[a] Docking runs were carried out using the standard search protocol and grids of 1 Å spacing. Results were ranked based on the total docking energy. Complexes classified identically with DrugScore and AutoDock are highlighted in bold.

Table 2. Success Rates Obtained for the Test Set of 41 Complexes by Docking with the Standard Search Protocol, Using Grids of 1 Å Spacing[a]

	DrugScore		AutoDock	
	no. of cases	%	no. of cases	%
rigid, rank 1	37	90	38	93
rigid, rank 1 + deg ranks	38	93	41	100
flexible, rank 1 (E_d)	18	44	18	44
flexible, rank 1 (E_i)	20	49	18	44
flexible, rank 1 + deg ranks (E_d)	26	63	21	51
flexible, rank 1 + deg ranks (E_i)	26	63	21	51
test cases with <15 torsions only: flexible, rank 1 + deg ranks (E_d)	23 (of 30)	77	19 (of 28)	68

[a] The table shows the number and percentage of complexes for which results deviating less than 2 Å from the reference were observed either at rank 1 or a degenerate (deg) rank. The ranking is based on the total docking energy (E_d) or the intermolecular energy (E_i); for rigid docking $E_d = E_i$, since intramolecular terms need not be evaluated.

these cases, all 10 independently performed docking runs for a single ligand converged to the same result. In only four cases, the goal of correctly reproducing the X-ray structure at rank 1 was not achieved. For one of these test cases, however, a slightly lower ranked, essentially degenerate result (cf. below) was obtained that differs by less than 1.5 Å from the reference and brings the overall success rate to 93%.

Not unexpectedly, the success rate was lower when docking was performed with ligand flexibility. Using the standard search protocol, flexible docking correctly predicted the binding mode of 18 complexes (44%) (cf. Tables 1 and 2). This refers to the most restricted definition of "correct prediction", where exclusively the result at rank 1 is considered, which is required to exhibit an rmsd of less than 2 Å. Although it would be highly desirable for any docking method to *always* yield the result with lowest rmsd at rank 1, there are both practical and physical reasons which suggest this to be an unrealistic expectation (e.g., approximations inherent in all currently existing scoring and docking approaches, ruggedness of the energy landscape, uncertainties in experimental structures, alternative or multiple binding modes, residual mobility of a ligand in the binding site, "single-structure approximation" of configurational and conformational ensembles). As a consequence, it normally makes sense to additionally consider docking results with a score very similar to the score of the top rank, i.e., "degenerate" results in terms of the docking energy. Taking into account also the results that match the score at rank 1 within 0.5 kcal/mol (such degenerate results were observed for 21 of the 41 test cases), the number of complexes with an rmsd of <2 Å result rises to 26, equal to a success rate of 63%. The tolerance of 0.5 kcal/mol should be adequate to highlight essentially degenerate results, given the grid approximation and uncertainties in the experimental structures which may lead to differences in the score of more than 0.5 kcal/mol for structures deviating less than 1 Å from each other. This has, for example, been observed for the docking results of 1eap and 1elc, as well as for the scores of the NAPAP ligand bound to thrombin in two independently solved crystal structures (1dwd, 1ets).

The ligand configurations generated upon docking are normally ranked according to the total docking energy, which is the sum of the intermolecular score and the intramolecular energy. In AutoDock, the latter is calculated based on nonbonded interaction terms and serves mainly the purpose of avoiding strained conformations. Obviously, this term must be in appropriate balance with the intermolecular term. Since in a standard AutoDock run the same Lennard-Jones parameters are used for both, a sufficient balance can normally be assumed. However, if the grids are calculated with another function, i.e., DrugScore, this is not necessarily the case, although the primary DrugScore values were scaled to bring them into the range of the AutoDock grid point energies (cf. Methods section). Accordingly, it was checked whether different overall results would be obtained by a ranking based solely on the intermolecular score, i.e., the DrugScore value. Using this ranking method, a top ranked result within 2 Å rmsd was obtained for 20 complexes (49%), 17 of which, however, had been correctly ranked already by the standard method. Counting also degenerate results, the same success rate (63%) was obtained as with the standard ranking. Taken together, this suggests that the balance between inter- and intramolecular term seems not to be a major problem.

A closer look at the 15 test cases for which no successful prediction was obtained (cf. Table 1) reveals a high proportion of very flexible molecules: eight ligands have 15 or more torsional degrees of freedom, (up to 29 for 4hvp). This is not only beyond the level of complexity AutoDock and its standard optimization protocol was designed for, it also exceeds the number of rotatable bonds that molecules useful in the context of drug design, especially in the lead-finding process, normally have. Of the 41 ligands tested, 30 have less than 15 torsions. For these, the standard search protocol applied here is sufficiently exhaustive to lead to near-convergence, and accordingly in 23 cases correct predictions were obtained at rank 1 or a degenerate rank, corresponding to a success rate of 77%. In contrast, for only three (1poc, 1apt, 1ida) of the 11 molecules with ≥15 torsions a correct prediction was obtained. For

seven of the eight remaining cases, the standard search parameters were apparently not sufficient to obtain a converged result, as assessed by comparison with the intermolecular energies obtained in the corresponding rigid docking runs. Performing runs with a 5-fold enhanced search protocol gave two additional successful predictions (for 1ppl and 1rne), while for the other five cases convergence was still not accomplished. In the case of complex 9hvp (20 torsions), however, the result seems acceptable even though the optimal energy was not reached and a top-ranked result with an rmsd of 2.12 Å was obtained. Here the deviation results to a large extent from two terminal phenyl rings that are rather exposed in the native structure and do not show significant contacts with the protein. If these terminal rings are not considered, the rmsd falls well below 2 Å.

Issues of convergence also play a role in some of the less flexible test cases that have failed. For 1bma and 1glq (12 torsions each), correct predictions at rank 1 (1bma) or a degenerate rank (1glq) could be obtained with an enhanced sampling protocol. Crystal packing contacts influencing the experimental binding mode could play a role in some other cases (e.g., 1igj, 1hdc),[21] since the crystallographic packing environment is of course not taken into account by docking to a single protein binding site.[22] Complex structures exhibiting crystal packing effects should actually not be part of test sets used for evaluating docking methods and scoring functions.

The grid spacing is a further issue with possible influence on the results. The standard grid spacing of 1 Å used here is larger than the default used in AutoDock (0.375 Å). Grids with larger spacing are computationally more efficient, but some information gets necessarily lost. The docking runs for all 23 cases, for which no top-ranked result within 2 Å rmsd had been obtained (cf. Table 2), were therefore repeated using grids of 0.375 Å spacing. For most of the cases, this did not lead to significant improvements. For the five complexes 1bma, 1did, 1etr, 1hgj, and 1srj, however, a top ranked result with an rmsd of ≤1.5 Å could now be observed. Since, in four of these cases, "degenerate" results close to the reference structure had already been observed with the coarse grids, it appears that more detailed DrugScore grids help to discriminate more precisely between native and non-native results when different positions with very similar energy are possible. In summary, coarse grids appear acceptable, but smaller grid sizes are clearly preferable.

To put the use of DrugScore as objective function for docking into perspective, the entire test set was also docked using the AutoDock empirical free energy function. As can be seen in Tables 1 and 2, the overall success rates are comparable. Rigid docking with the AutoDock function provides the correct result at rank 1 in 93% of the test cases, with a near-degenerate result reproducing the experimental binding mode in the remaining three cases. Flexible docking yields correct predictions at rank 1 for 18 cases (44%). Considering also near-degenerate results, this figure rises to 51%, which is somewhat lower than the 61% achieved with DrugScore. Ranking by intermolecular energy instead of the total docked energy does not change the overall success rates. Again, major problems arise with the extremely flexible ligands. Leaving these cases apart, docking with the AutoDock function succeeds in 68% of the cases.

It is worth noting that there is significant overlap between DrugScore and AutoDock with respect to the cases that are correctly predicted and those which remain problematic or fail (cf. Table 1). Although there are certainly test cases that are generally easier to handle than others, this is an interesting observation, given that the two scoring functions have completely different origins and characteristics (e.g., DrugScore: 17 Tripos atom types;[23] no charges; no hydrogen atoms. AutoDock: atom types limited to different elements, except for aromatic/aliphatic carbon; polar hydrogens and charges required). It should be mentioned, though, that correct predictions obtained with DrugScore are usually closer to the experimental reference than the corresponding results obtained with the AutoDock function: for the 13 test cases correctly predicted at rank 1 by both methods (cf. Table 1), the rmsd of the DrugScore results is on average 0.22 Å lower compared to the AutoDock results.

Although the overall success rates in flexible docking with DrugScore and AutoDock may appear low at first sight, they are considerable given that the test set consisted entirely of complexes not correctly predicted by an earlier approach. A certain bias toward "difficult" test cases seems therefore to be given. The success rates reported here are nevertheless perfectly comparable with those reported for other well-established programs (e.g., ICM: 51.0%, based on 51 complexes;[16] FlexX: 46.5%, based on 200 complexes[20]). However, as long as no standard evaluation suite exists against which all docking methods and scoring functions are routinely benchmarked, objective comparisons are difficult to be made. Still, with respect to DrugScore itself, it is remarkable that purely knowledge-based atomic distance preferences combined with a simple repulsion term can be used successfully for guiding docking searches, at least as well as empirical free energy functions based on force field terms.

Conclusion. DrugScore pair potentials have been tested as objective function for docking searches using the AutoDock program. The study has shown that knowledge-based pair potentials can be successfully applied for docking optimizations. With respect to the prediction of experimental binding modes by flexible docking, the PDB-derived DrugScore potentials and the empirical AutoDock scoring function are found to be of comparable quality, with slight advantages for DrugScore in the overall statistics. Docking of highly flexible ligands (≥15 torsional degrees of freedom) was found to remain problematic regardless of the scoring function being used; although surely of limited relevance for docking in the context of lead discovery, the efficient generation of appropriate conformations for large flexible ligands still awaits a more satisfying solution.

Methods. Protein Setup. Protein structures were taken from the PDB.[12] Ligands and solvent molecules were removed, retaining, however, cofactors (1ldm, 1phg) and metal ions (1did, 1frp, 1poc, 1mmq, 4tmn, 7cpa) near the binding site. For use with the AutoDock free energy function, polar hydrogens were added with the PROTONATE utility from AMBER,[24] AMBER united

atom force field charges were assigned,[25] and solvation parameters were added using the ADDSOL utility of AutoDock3.0.

DrugScore grids were calculated by evaluating the pair preferences for a given ligand atom type at every grid point, summing over all protein atoms within the 6 Å definition range of the DrugScore potentials. The pair potentials were combined with a Gaussian-type repulsive term at short interatomic distances. The primary grid values were scaled for proper combination with the intramolecular force field terms used by AutoDock. A scaling factor of 2.5×10^{-5} was found to be appropriate to yield grid point minimum values in the same order of magnitude as the AutoDock function. (Note: although in the Results section the scoring energies are reported in "kcal/mol", the scale is actually irrelevant for the DrugScore values; in the context of this work, these values should only be regarded as relative scores for binding geometries.) The grids were centered on the ligand in its crystallographic binding mode. A grid spacing of 1 Å was used, and the grids were dimensioned sufficiently large to extend at least 6 Å beyond any ligand atom in its crystallographic binding mode. DrugScore grids and AutoDock grids had identical dimensions.

Ligand Setup. Ligand structures were obtained in mol2-format from the FlexX-200 test set[20] and modified where necessary. For docking runs with the AutoDock free energy function, hydrogens were added to the ligands and Gasteiger partial atomic charges were assigned.[26] This step was not required in the context of DrugScore. Flexible torsions were defined with the help of AutoTors. In general, these were all acyclic, nonterminal single bonds (excluding amide bonds) in a given ligand molecule. Ligands with -OH/-NH_2 groups had additional rotatable bonds assigned in docking runs with the AutoDock function (these -OH/-NH_2 rotors were kept rotatable also in the "rigid" docking runs).

Docking. All docking runs were performed with version 3.0 of the program AutoDock, using the Lamarckian genetic algorithm. The standard docking protocol for rigid and flexible ligand docking consisted of 10 independent runs per ligand, using an initial population of 50 randomly placed individuals, a maximum number of 1.5×10^6 energy evaluations, a mutation rate of 0.02, a crossover rate of 0.80, and an elitism value of 1. The probability of performing a local search on an individual in the population was 0.06, using a maximum of 300 iterations per local search. Results differing by less than 1 Å rmsd from each other were clustered together and represented by the result with the best docking energy (corresponding to the sum of inter- and intramolecular score). Depending on the ligand size and the number of flexible torsions, a single docking run using these search parameters required CPU times ranging from 55 s (6 atoms, 1 torsion) to 35 min (54 atoms, 29 torsions) on a PIII 800 running Linux.

Acknowledgment. Financial support from the German Federal Ministry of Education and Research (ReLiMo project, Grant 0311619) is gratefully acknowledged.

References

(1) Abagyan, R.; Totrov, M. High-throughput docking for lead generation. *Curr. Opin. Chem. Biol.* **2001**, *5*, 375−382.
(2) Kuntz, I. D.; Meng, E. C.; Shiochet, B. K. Structure-based molecular design. *Acc. Chem. Res.* **1994**, *27*, 117−123.
(3) Lengauer, T.; Rarey, M. Computational methods for biomolecular docking. *Curr. Opin. Struct. Biol.* **1996**, *6*, 402−406.
(4) Dixon, J. S.; Blaney, J. M. Docking: Predicting the structure and binding affinity of ligand−receptor complexes. In *Designing bioactive molecules: three-dimensional techniques and applications*; Martin, Y. C., Willett, P., Eds.; American Chemical Society: Washington, DC, 1997; pp 175−198.
(5) Muegge, I.; Rarey, M. Small molecule docking and scoring. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, 2001; Vol. 17; pp 1−60.
(6) Dixon, J. S. Evaluation of the CASP2 docking section. *Proteins* **1997**, *Suppl*, 198−204.
(7) Gohlke, H.; Klebe, G. Statistical potentials and scoring functions applied to protein−ligand binding. *Curr. Opin. Struct. Biol.* **2001**, *11*, 231−235.
(8) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.
(9) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.
(10) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.
(11) Gohlke, H.; Hendlich, M.; Klebe, G. Predicting binding modes, binding affinities and 'hot spots' for protein−ligand complexes using a knowledge-based scoring function. *Perspect. Drug Discovery Des.* **2000**, *20*, 115−144.
(12) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(13) Hendlich, M. Databases for protein−ligand complexes. *Acta Crystallogr. D Biol. Crystallogr.* **1998**, *54*, 1178−1182.
(14) Bergner, A.; Guenther, J.; Hendlich, M.; Klebe, G.; Verdonk, M. Use of Relibase for retrieving complex three-dimensional interaction patterns including crystallographic packing effects. *Biopolymers (Nucleic Acids Sciences)* In press.
(15) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.
(16) Totrov, M.; Abagyan, R. Protein−ligand docking as an energy optimization problem. In *Drug-receptor thermodynamics: introduction and applications*; Raffa, R. B., Ed.; Wiley: Chichester, 2001; pp 603−624.
(17) Goodsell, D. S.; Olson, A. J. Automated docking of substrates to proteins by simulated annealing. *Proteins* **1990**, *8*, 195−202.
(18) Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 293−304.
(19) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.
(20) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FLEXX incremental construction algorithm for protein- ligand docking. *Proteins* **1999**, *37*, 228−241.
(21) Cole, J. C.; Bergner, A. Personal communication, 2001.
(22) Sotriffer, C. A.; Ni, H. H.; McCammon, J. A. HIV-1 integrase inhibitor interactions at the active site: prediction of binding modes unaffected by crystal packing. *J. Am. Chem. Soc.* **2000**, *122*, 6136−6137.
(23) SYBYL Molecular Modeling Software, version 6.7; Tripos Inc.: St. Louis, MO.
(24) Case, D. A.; Pearlman, D. A.; Caldwell, J. W.; Cheatham, T. E., III; Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R. V.; Cheng, A. L.; Vincent, J. J.; Crowley, M.; Tsui, V.; Radmer, R. J.; Duan, Y.; Pitera, J.; Massova, I.; Seibel, G. L.; Singh, U. C.; Weiner, P. K.; Kollman, P. A. AMBER 6; University of California, San Francisco: CA, 1999.
(25) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **1984**, *106*, 765−784.
(26) Gasteiger, J.; Marsili, M. Iterative partial equilization of orbital electronegativity − a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3228.

JM025507U