DOI: 10.1002/cmdc.200600073

Addressing Protein Flexibility and Ligand Selectivity by "in situ Cross-Docking"

Matthias Zentgraf, Jasmine Fokkens, and Christoph A. Sotriffer*[a]

To overcome the "single-structure paradigm" in current methods for computational protein–ligand docking, we have recently introduced the "in situ cross-docking" (ISCD) approach to simultaneously address multiple targets,[1] using the grid-based AutoDock program as search engine.[2] Whereas the feasibility of ISCD for dealing with dissimilar binding sites and non-cross-reactive, tight-binding ligands had been demonstrated, it remained to be shown whether it could also be applied to different conformations of the same target (to address protein flexibility) or to closely related targets binding the same ligand with varying affinity (to address selectivity). Here, we investigate the first issue using aldose reductase as a test case, and the second using a recently introduced series of thrombin and trypsin inhibitors.

Aldose reductase (AR), a target against late-onset diabetic complications, catalyzes the reduction of aldoses and other aldehydes to the corresponding alcohols.[3] Its substrate promiscuity is in part due to an "induced-fit"-like mechanism of ligand binding, whereby a specificity pocket can be closed or opened in different conformations, depending on the ligand being bound.[4] Crystal structures of AR–inhibitor complexes have revealed three major binding-pocket conformations, best represented by the complexes with sorbinil (PDB 1AH0), tolrestat (1AH3), and IDD594 (1US0) as shown in Figure 1.

For docking and structure-based ligand design AR poses the obvious problem that a single conformation of the protein is not sufficiently representative as a target structure; instead, at least three major conformations need to be addressed. In standard docking, this would be done sequentially, using each protein conformer for separate docking simulations, thus requiring for each ligand as many separate simulations as there are protein conformers to investigate. With ISCD, instead, the conformers can be combined to a single search space such that only one simulation must be run per ligand.

Using 1AH0, 1AH3, and 1US0 as structures for the three AR binding-site conformers, separate AutoDock grids were first calculated for each of them (further details about the methods are provided as Supporting Information A) and AutoDock runs were carried out on the separate single grids, proving that the experimental binding mode can indeed be reproduced by standard flexible docking (sorbinil to 1AH0: docking result on rank 1 shows a root-mean-square deviation (RMSD) of 0.26 Å with respect to the crystal structure; tolrestat to 1AH3: 0.94 Å; IDD594 to 1US0: 1.39 Å for rank 1, 0.84 Å for rank 2. A detailed tabular report of these results is provided as Supporting Information B). This is obviously a prerequisite for testing the ISCD approach, which can only be successful if single-grid docking with the applied scoring function is able to provide the correct binding mode. Clearly, an appropriate scoring function is essential for any (cross-) docking procedure.

To setup ISCD, the single grids were then combined to a joined grid with repulsive layers between them.[1] The joined grid representing all three binding-site conformers was used to test whether ISCD is able to identify the native binding pocket of a given ligand (sorbinil, tolrestat, and IDD594) in a single docking calculation. For this purpose the same standard docking parameters were used as applied before in the single-binding-site docking. Table 1 illustrates that with ISCD the native binding pocket and binding mode is indeed found in all three cases: for each ligand, the top-ranked cluster corresponds to the correct binding mode in the correct protein conformer, and the cluster size (corresponding to the occurrence frequency of the result) is excellent for the most rigid ligand (sorbinil) and fully acceptable for the most flexible compound (IDD594).

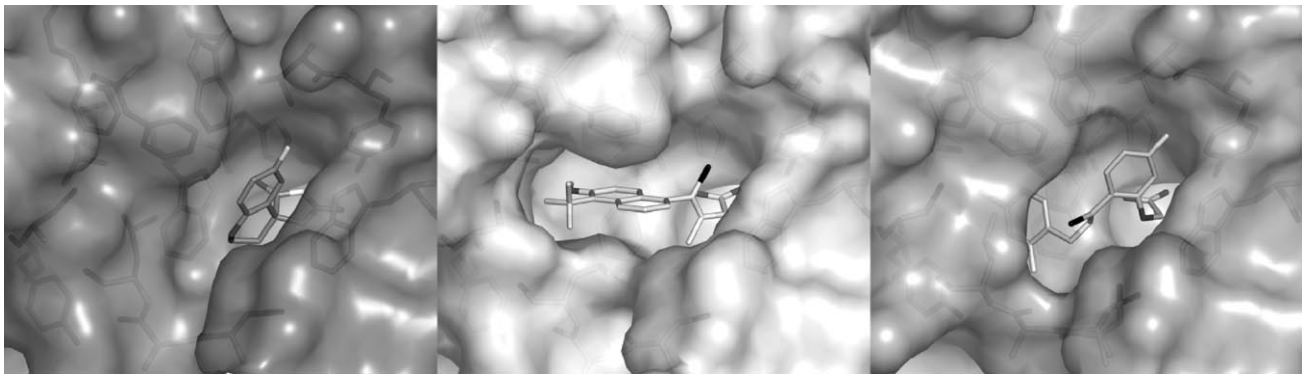

Figure 1. Comparison of the three different AR binding-site conformers, as exemplified by the complexes with sorbinil (left), tolrestat (middle), and IDD594 (right), seen from an identical perspective.

[a] *M. Zentgraf, Dr. J. Fokkens, Dr. C. A. Sotriffer*
Department of Pharmaceutical Chemistry
Philipps-Universität Marburg
Marbacher Weg 6, 35032 Marburg (Germany)
Fax: (+49) 6421/2828994
E-mail: sotriffer@staff.uni-marburg.de

Supporting information for this article is available on the WWW under http://www.chemmedchem.org or from the author.

Table 1. Results of 100 ISCD runs for six AR ligands, using the joined 1AH0–1AH3–1US0 grid.[a]

Ligand	E_d [kcal mol^{-1}]	RMSD [Å]	Cluster Size	$N \leq 2$ Å RMSD
sorbinil	−9.38	0.26	96	96
tolrestat	−11.77	1.00	35	51
IDD594	−12.16	0.84	5	6
fidarestat	−9.58	0.51	52	99
zenarestat	−13.15	1.37	8	15
pyridazinone	−10.98	0.49	30	34

[a] For each ligand, the top-ranked docking result is reported, with the docking score ("docked energy" E_d), the RMSD to the experimental binding mode, the size of the top-ranked result cluster, and the total number of results showing an RMSD of less than 2 Å.

As the inhibitors tested so far were the exact ligands of the protein conformers used to construct the joined grid, a further test for running ISCD with the joined 1AH0–1AH3–1US0 grid was carried out using the compounds fidarestat and zenarestat (cf. Figure 2). In the crystal structure, fidarestat (PDB 1PWM) shows a sorbinil-like binding mode, whereas zenarestat (PDB 1IEI) binds to a similar AR conformer as IDD594 (to date, no ligand other than tolrestat is known to induce a 1AH3-like conformation). For zenarestat, ISCD correctly suggests preferential binding to the 1US0 conformation (cf. Table 1): the top-ranked result shows a clear score difference of 1.14 kcal mol^{-1} with respect to binding modes in other pockets, and all top 15 results yield an RMSD < 1.4 Å with respect to the crystal structure (after best-fit alignment of the 1IEI protein structure with 1US0). For fidarestat, the correct binding mode is obtained in almost all of the 100 runs, though in two different pockets: 47 times in the 1AH0 pocket (with an RMSD of 0.69 Å after best-fit alignment of the 1PWM protein structure with 1AH0) and 52 times in the 1US0 pocket (RMSD 0.51 Å after best-fit alignment of 1PWM with 1US0). Both results show a virtually identical score (−9.46 kcal mol^{-1} versus −9.58 kcal mol^{-1}, respectively), indicative of the fact that although 1US0 corresponds to the open conformation and fidarestat is expected to bind to the closed conformation (1AH0), the 1US0 conformation is also compatible with the native fidarestat binding mode. This is in contrast to sorbinil, where a clash with Cys 298 in 1US0 precludes docking in the native mode. Fidarestat shows a larger distance to Cys 298, thus allowing a near-native binding mode to be observed also in 1US0.

In a further test, predictive docking was carried out for a new and very potent sulfonyl pyridazinone inhibitor for which no crystal structure was available at that time (compound 24 described by Mylari et al.;[5] cf. Figure 2). ISCD suggested a binding mode similar to IDD594, with the best score achieved in the 1US0 pocket and 34 out of 100 runs ending up in this pocket. The correctness of the prediction was later confirmed by the crystal structure of the pyridazinone–AR complex,[6] which yields an RMSD of 0.49 Å for the docking result after superposition of the protein in the new crystal structure with 1US0.

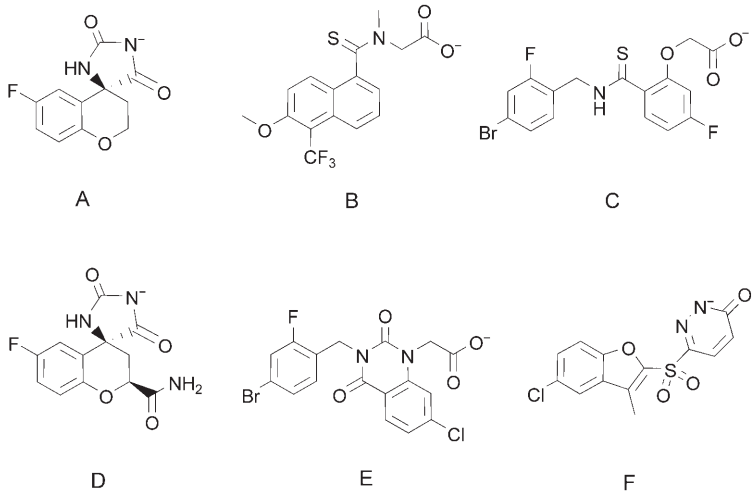

Figure 2. Chemical formulae of the AR inhibitors investigated by ISCD: sorbinil (A), tolrestat (B), IDD594 (C), fidarestat (D), zenarestat (E), and the pyridazinone inhibitor (F).

Taken together, the results indicate that with ISCD alternative protein conformations can be successfully addressed simultaneously. The major advantage of the method, however, is that differences among the target structures need not to be confined to small side-chain reorientations. Accordingly, issues of selectivity with respect to different targets can also be addressed. Whereas in the original proof-of-concept very different proteins (and ligands with no expected cross-reactivity) were used,[1] here the selectivity with respect to two closely related enzymes is investigated. The serine proteases thrombin and trypsin and a set of seven new inhibitors with varying selectivity for thrombin plus two trypsin-selective inhibitors were used for this purpose. The inhibitors **1**–**7** (numbered as in Fokkens et al.[7]) consist of a common rigid nonpeptidic core, substituted in position 1 with different aliphatic and aromatic groups, as shown in Figure 3 (ligands **3** and **6** differ only with respect to the core ring system). The selectivity ratio of the inhibition constants K_i(trypsin)/K_i(thrombin) ranges between 2.6 for ligand **4** and 760 for ligand **1** (measured for the racemic mixture).[7] For all ligands except **4**, crystal structures in complex with trypsin could be obtained;[7,8] for **1**, **2**, **3**, and **6** crystal structures in complex with thrombin are available as well.[7–10] The structures show that all ligands share a common binding mode and that the (+)-enantiomer is exclusively bound to the enzyme; this is also the enantiomer drawn in Figure 3 and used in all docking calculations. Ligands **8** and **9** are two closely related benzothiophene-carboxamidine inhibitors (**8** = APC-6860, **9** = APC-7377 in Katz et al.[11]). They show more favorable binding to trypsin, with a selectivity ratio K_i(trypsin)/K_i(thrombin) of 0.022 (**8**) and 0.017 (**9**), respectively. Crystal structures in complex with trypsin are available for both ligands (PDB 1C5Q, 1C5S), a thrombin complex is available for **8** (PDB 1C5N).

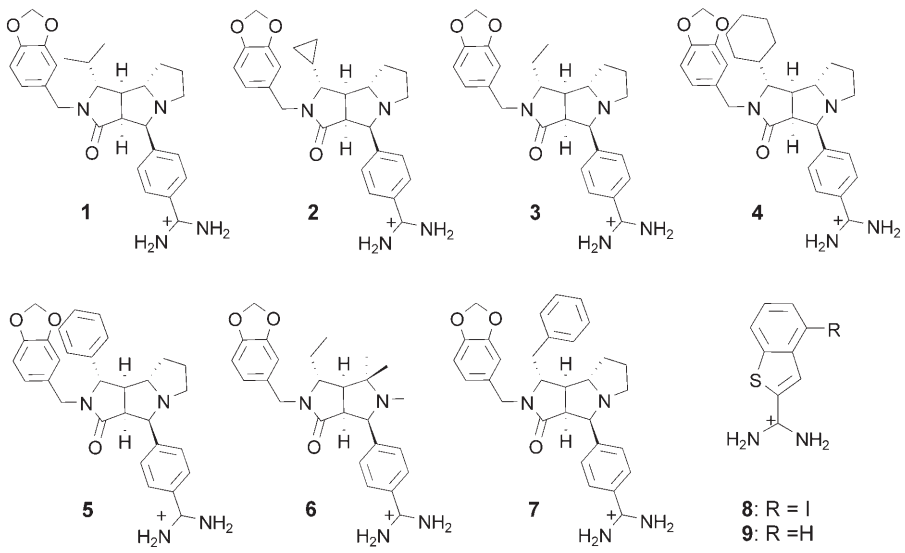

Figure 3. Chemical formulae of the trypsin and thrombin ligands investigated by ISCD.

Using the crystal structures of trypsin and thrombin in complex with ligand **1**, grids were calculated for the two binding sites and combined to one joined grid. As shown in Table 2, performing ISCD with 50 independent runs per ligand revealed energetically preferred binding to thrombin for ligands **1** to **7**: the top-ranked result for these ligands was found in the thrombin grid, whereas the best-ranked docking result for trypsin occurred between rank 2 and 7. All best-ranked trypsin results are structurally correct, as indicated by the low RMSD values. For thrombin, this is the case for five of the seven ligands, since the top-ranked result for ligand **5** and **7** does not show the presumably correct binding mode (for both ligands no experimental complex structure with thrombin is available). For the two trypsin-selective ligands (**8**, **9**) preferred binding to trypsin is indeed observed: for **8**, the top-ranked result corresponds precisely to the experimental binding mode in trypsin, whereas the result on rank 2 reproduces the thrombin binding mode. For **9**, all 50 docking results are virtually identical and correspond to the crystallographic binding mode in trypsin.

Thus, looking at the best-ranked result in each grid, ISCD reveals the correct binding mode of all ligands for trypsin and of six ligands for thrombin. In addition, the binding preference is correctly predicted, with the better score observed for the preferred target known from experiments. Despite the energetic preference, however, the large majority of the results is found in the trypsin grid, also in the case of the thrombin-selective ligands (cf. Table 2 and Figure 4). For the trypsin-selective ligand **9** none of the results is found in thrombin with the applied search protocol. Apparently, the open trypsin binding site represents an easily accessible minimum, whereas the access to the optimum given by the thrombin binding site is restricted by the "60-loop", which makes thrombin binding modes more difficult to reach. (Also in single-grid docking, convergence for trypsin is generally much better than for thrombin; cf. Supporting Information C, which contains further details about the results, as well as the results of separate docking runs to trypsin and thrombin).

These observations illustrate the main problem associated with ISCD when applied to cross-reactive targets (where a given ligand binds to different targets) in contrast to "orthogonal" targets (where the ligand binds to only one of the targets). In the latter case only the global optimum is of interest, whereas in the former, the lower-ranked results must also be obtained if the task is to reveal a binding mode for each cross-reactive target. If so, full convergence to the global optimum would not be desired in ISCD, because otherwise binding to the protein showing lower affinity would not be found at all. Instead, ISCD should reveal all experimentally observable binding modes, and the best-ranked result in each grid should pro-

Table 2. Results of 50 ISCD runs for the joined trypsin and thrombin grid.[a]

	Trypsin					Thrombin				
Ligand	Cluster Rank	E_d [kcal mol^{-1}]	RMSD [Å]	Cluster Size	$N \leq 2$ Å RMSD	Cluster Rank	E_d [kcal mol^{-1}]	RMSD [Å]	Cluster Size	$N \leq 2$ Å RMSD
1	7	−11.43	0.54	42	42	1	−13.72	0.90	2	8
2	5	−11.66	0.52	40	40	1	−13.11	0.89	3	9
3	3	−11.51	0.85	42	42	1	−13.65	0.85	3	6
4	2	−12.21	0.86	7	43	1	−13.28	1.06	1	1
5	5	−10.84	1.10	28	29	1	−11.70	8.60	5	2
6	4	−11.20	0.97	1	39	1	−13.59	1.03	5	9
7	2	−12.59	0.55	45	45	1	−12.76	8.65	1	0
8	1	−8.85	0.49	43	43	2	−7.86	0.61	7	7
9	1	−8.00	0.47	50	50	–	–	–	–	–

[a] For each ligand, the top-ranked docking result is reported, with the docking score ("docked energy" E_d), the RMSD to the experimental binding mode, the size of the top-ranked result cluster, and the total number of results showing an RMSD less than 2 Å. The RMSD for ligands **1**–**7** was calculated for the common ligand substructure consisting of the core ring system and the benzamidine anchor, which is the part of the ligands that is well defined in all available crystal structures.

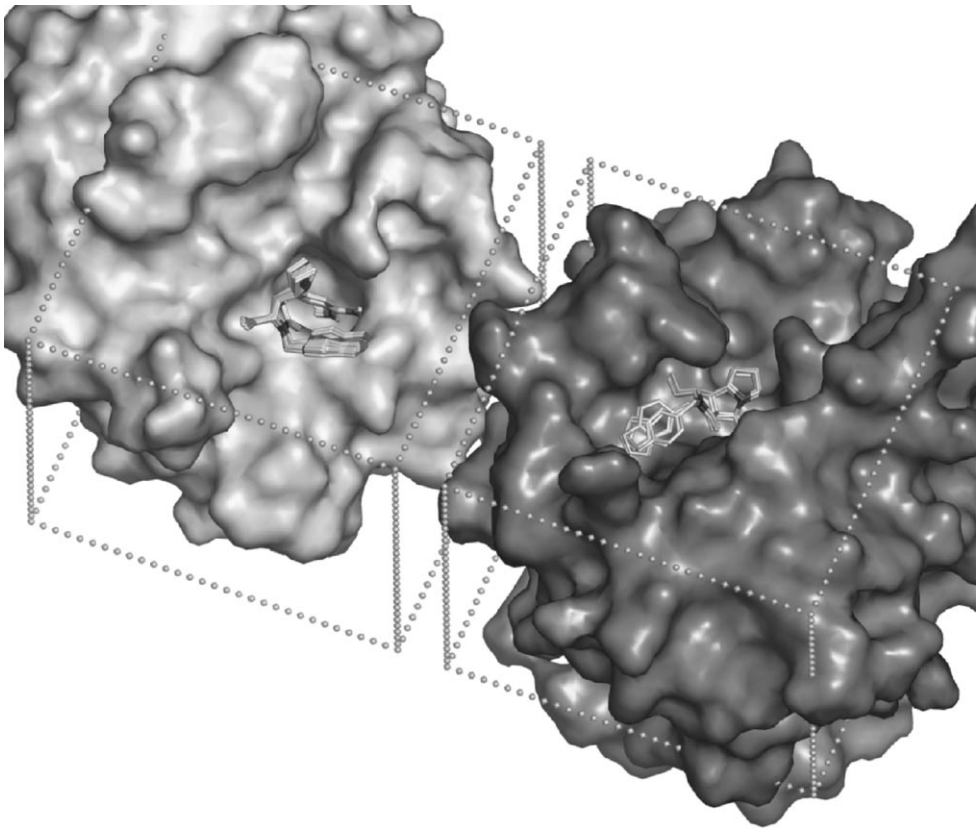

Figure 4. ISCD docking results for ligand **3**, illustrating the members of the three best-ranked clusters, which together cover 96% of the results (48 of 50 docking runs). Clusters 1 and 2 are found in thrombin (right) and contain three results each; cluster 3 is found in trypsin (left) and contains 42 closely overlapping results. The borders of the two grids joined to a single ISCD grid are shown; the gap between them corresponds to the repulsive layer required for avoiding artifacts.

vide the correct binding mode for the corresponding protein, as observed here for most of the ligands. Clearly, this represents a dilemma for a docking method based on an optimization algorithm, which is generally supposed to converge to the global optimum. The challenge is to use a search protocol which finds the global optimum, but still produces lower-ranked binding modes. In fact, if the search is carried out with a very large number of steps and converges well, lower-ranked results may not be observed at all (as apparent with **9**, where no thrombin result is found with the applied search protocol; however, if the search length is halved, results in thrombin start appearing also for **9**). On the other hand, using a search protocol that is too short risks missing the true global minimum. Obviously, knowing in advance which protocol may be suitable for a given system is rather difficult and may require extensive parameter studies. This is a limitation of ISCD in comparison to standard cross-docking to separate grids, where a binding mode is guaranteed for every target. Application of ISCD is more advantageous if the task is only to reveal the preferred target.

In summary, this study has shown that addressing protein flexibility and ligand selectivity with ISCD is feasible. With a single docking calculation, the appropriate binding-site conformation can be selected from three different alternatives in case of the flexible AR enzyme. Likewise, a single docking calculation is sufficient to reveal the preferred binding modes of a given ligand for two closely related proteins, trypsin and thrombin, if an appropriate search protocol is used. Alternative approaches for dealing with protein flexibility in ligand docking are available, such as FlexE[12] (which had also been tested on AR[12]) and grid-based averaging[13,14] (which we have applied to AR for comparative purposes; cf. Supporting Information D), but these are all restricted to relatively small structural differences among the investigated systems. ISCD does not have this limitation and can be applied to protein structures of any desired difference. The challenge, however, will be to develop efficient protocols for simultaneously considering more than two or three structures to extend the scope of ISCD.

Keywords: aldose reductase · drug design · molecular modeling · thrombin · trypsin

[1] C. A. Sotriffer, I. Dramburg, *J. Med. Chem.* **2005**, *48*, 3122.
[2] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, A. J. Olson, *J. Comput. Chem.* **1998**, *19*, 1639.
[3] C. Yabe-Nishimura, *Pharmacol. Rev.* **1998**, *50*, 21.
[4] C. A. Sotriffer, O. Krämer, G. Klebe, *Proteins Struct. Funct. Genet.* **2004**, *56*, 52.
[5] B. L. Mylari, S. J. Armento, D. A. Beebe, E. L. Conn, J. B. Coutcher, M. S. Dina, M. T. O'Gorman, M. C. Linhares, W. H. Martin, P. J. Oates, D. A. Tess, G. J. Withbroe, W. J. Zembrowski, *J. Med. Chem.* **2003**, *46*, 2283.

[6] H. Steuber, M. Zentgraf, A. Podjarny, A. Heine, G. Klebe, *J. Mol. Biol.* **2006**, *356*, 45.
[7] J. Fokkens, G. Klebe, *Angew. Chem.* **2006**, *118*, 1000; *Angew. Chem. Int. Ed.* **2006**, *45*, 985.
[8] J. Fokkens, Dissertation, University of Marburg (Marburg), **2005**.
[9] U. Obst, Dissertation, ETH Zürich (Zürich), **1997**.
[10] U. Obst, D. W. Banner, L. Weber, F. Diederich, *Chem. Biol.* **1997**, *4*, 287.
[11] B. A. Katz, R. Mackman, C. Luong, K. Radika, A. Martelli, P. A. Spengler, J. Wang, H. Chan, L. Wong, *Chem. Biol.* **2000**, *7*, 299.
[12] H. Claussen, C. Buning, M. Rarey, T. Lengauer, *J. Mol. Biol.* **2001**, *308*, 377.
[13] R. M. Knegtel, I. D. Kuntz, C. M. Oshiro, *J. Mol. Biol.* **1997**, *266*, 424.
[14] F. Osterberg, G. M. Morris, M. F. Sanner, A. J. Olson, D. S. Goodsell, *Proteins Struct. Funct. Genet.* **2002**, *46*, 34.

Received: March 21, 2006
Revised: August 8, 2006
Published online on October 6, 2006

