Performance of 3D-Database Molecular Docking Studies into Homology Models

Connie Oshiro,* Erin K. Bradley, John Eksterowicz, Erik Evensen, Michelle L. Lamb, J. Kevin Lanctot, Santosh Putta, Robert Stanton, and Peter D. J. Grootenhuis*

Deltagen Research Laboratories, 740 Bay Road, Redwood City, California 94063, and 4570 Executive Drive, Suite 400, San Diego, California 92121

Received February 18, 2003

The performance of docking studies into protein active sites constructed by homology model building was investigated using CDK2 and factor VIIa screening data sets. When the sequence identity between model and template near the binding site area is greater than approximately 50%, roughly 5 times more active compounds are identified than would be found randomly. This performance is comparable to docking to crystal structures.

Introduction

Despite the progress in X-ray crystallography and high-field NMR structure elucidation methods, it is still fairly common not to have the structures of the target receptors available, particularly in the early phases of a drug discovery project. Often one will try to construct 3D-models based on the structures of homologous proteins. The performance of homology model building approaches is well understood and documented. In general, such models correctly predict the overall fold, but in regions of low sequence homology (often the ligand binding site) the model may be less accurate. The technology for generating homology models has been optimized to the point where completely automated generation of homology models for entire genomes is now feasible.[1,2]

One of the possible applications of homology-built models is in the docking and scoring of 3D-databases of compounds. The 3D-database docking into experimentally determined protein structures is known to be a very effective strategy to identify novel binders and chemotypes.[3,4] For example, Doman et al. recently showed that docking into the structure of protein tyrosine phosphatase-1B led to a very high enrichment with actives relative to high-throughput screening of a corporate collection against this target.[5] Various authors have successfully used homology models for their docking studies. However, to the best of our knowledge, no one has done a systematic analysis of the performance of molecular docking methods applied to homology models in comparison to crystal structures. In this paper, we describe the results of a retrospective analysis of homology docking using datasets from factor VIIa inhibitor and CDK2-antagonist projects. The datasets were derived from both screening libraries and directed combinatorial synthesis. We will show that when the sequence homology in the binding site region is greater than 50%, homology models can be used very effectively for docking, while the results obtained with less homologous models vary but are never worse than random selection of compounds.

Methods

Homology Model Building. Reference crystal structures for factor VIIa had PDB reference codes[6] 1dan and 1cvw. Eight homology models were constructed (method described below) for factor VIIa. The models were all based on serine proteases with sequence identity near the binding site vs factor VIIa varying from 37% to 77%. The template structures upon which the homology models were based were trypsin (2tbs, 1tnk), factor Xa (1f0r, 1fjs), heparin binding protein (1a7s, 1fy3), and collagenase (1azz, 2hlc). These eight serine protease structures themselves were also used in these docking studies to identify factor VIIa ligands and are referred to later in this paper as “similar structures” or “template structures.”

Reference crystal structures for CDK2 were 1hck and 1ckp. Four homology models were constructed on the basis of other kinase structures with sequence identity vs CDK2 varying from 43% to 60% near the binding siete. These template kinase structures were MAP kinase (3erk, 1pme) and Erk2 kinase (1lp4, 1daw). Again, these kinase structures themselves (“similar” structures) were also used for docking to identify CDK2 ligands.

All homology models were generated using MOE.[7] Traditionally, the sequence from the template structure is first aligned to or threaded along the target structure. In our case, rather than doing a sequence-to-sequence alignment, the template crystal structure was structurally superposed onto one of the reference target crystal structures (1dan for factor VIIa; 1hck for CDK2). This ensured that appropriate residues would be used in the model structure. By use of this alignment, homology models were generated in MOE using default parameters.

Screening Data. For factor VIIa, approximately 21 000 small molecules were docked. Approximately 13 000 of these were chemically diverse compounds from a general screening library; 8000 were compounds synthesized by project chemists. All of these compounds had been screened for activity. Those molecules with a K_i less than 10 μM or inhibition greater than 50% at 30 μM were considered active. There were 352 active compounds in this set, with 18 active scaffolds. Three scaffolds had only one representative structure in the database.

The CDK2 screening set consisted of approximately 17 000 compounds. Approximately 13 000 of these were chemically diverse compounds taken from a general screening library; 3000 were combinatorial synthetic compounds designed and synthesized by project chemists and 1000 were compounds selected from the ACD[8] based on chemical similarity to known literature actives. All compounds were tested, and those with an IC_{50} less than 25 μM or an inhibition greater than 50% at 10 μM were considered active. There were 367 active compounds in the dataset on 15 different scaffolds.

Molecular Docking. The program UCSF DOCK (version 4.0)[9−11] was used to dock and score different conformations and configurations of molecules in the databases. Rigid body

* To whom correspondence should be addressed. For C.O.: (address) Roche Pharmaceuticals, 3431 Hillview Avenue, Palo Alto, CA 94304; (phone) 650-855-5646; (fax) 650-852-1875; (e-mail) connie.oshiro@roche.com. For P.D.J.G.: (address) Vertex Pharmaceuticals, 11010 Torreyana Road, San Diego, CA 92121; (phone) 858-404-6676; (fax) 858-404-6713; (e-mail) peter_grootenhuis@sd.vrtx.com.

10.1021/jm0300781 CCC: $27.50 © 2004 American Chemical Society
Published on Web 01/03/2004

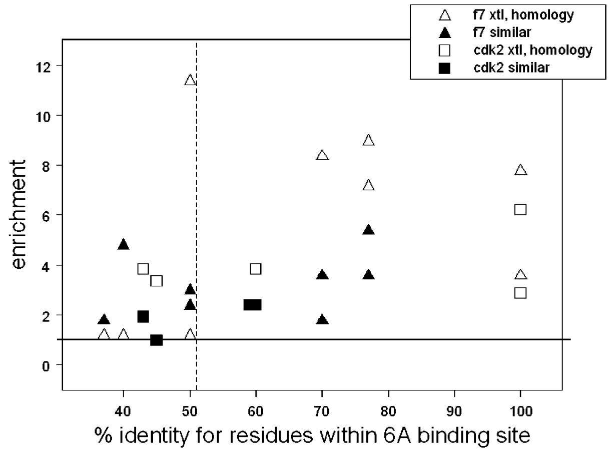

Figure 1. Plot of enrichment of active compounds selected by docking vs sequence identity of residues near the binding site. Values at 100% identity are the actual crystal structures (i.e., 1dan and 1cvw for factor VIIa; 1hck and 1ckp for CDK2). Random (enrichment equal to 1) is given by horizontal line. Note overplot of kinase homology enrichmet (value = 2.3) by "similar" structure enrichment.

Table 1. Average Enrichment of Actives for Homology Models[a]

target	crystal structure	sequence id >50%	sequence id <50%
factor VIIa	5.7	6.6	1.2[a]
CDK2	4.5	3.1	3.5
both	5.1 ± 2.3	5.4 ± 3.2	2.1[a] ± 1.3

[a] Outlier removed from average calculations.

docking was done for the target crystal structures and homology models. For "similar" structures, docking was skipped and the docked configurations of the molecules using the crystal structure were rescored, following commonly used practices in consensus scoring. The DOCK energy scoring was used to rank all the molecules in all cases. For each molecule in the dataset, 10 different conformations per stereoisomer were generated by an in-house conformational analysis program (CONAN)[12] and were rigidly docked. The lowest DOCK energy score of the 10 conformations was used as the score of the molecule for ranking. The top 100 ranking molecules were used in the analysis. For factor VIIa, 100 molecules represent 0.5% of the dataset. For CDK2, they represent 0.6%. A small number of molecules was selected because we were considering virtual screening as a mechanism to limit the actual number of compounds purchased or synthesized.

Energy scoring was done using Kollman united atom parameters for the protein,[13] while Kollman all-atom van der Waal parameters[14] and Gasteiger charges[15,16] were used for the molecules.

Results

We define "enrichment" of actives and scaffolds as the ratio of the number of active compounds (scaffolds) identified in the top 100 ranked compounds to the number of active compounds (scaffolds) identified in a randomly picked set of 100 compounds. The random number of actives was calculated analytically (% database search × total number of actives); the number of random scaffolds was determined computationally by sampling 100 compounds randomly for 5000 trials. Plots of the compound and scaffold enrichment as a function of the sequence identity of the target structure vs the template structure are given in Figures 1 and 2, respectively. Values at 100% identity are the actual crystal structures (i.e., 1dan and 1cvw for factor VIIa; 1hck and 1ckp for CDK2). Only residues within 6 Å of the bound crystal ligand were considered for these plots.

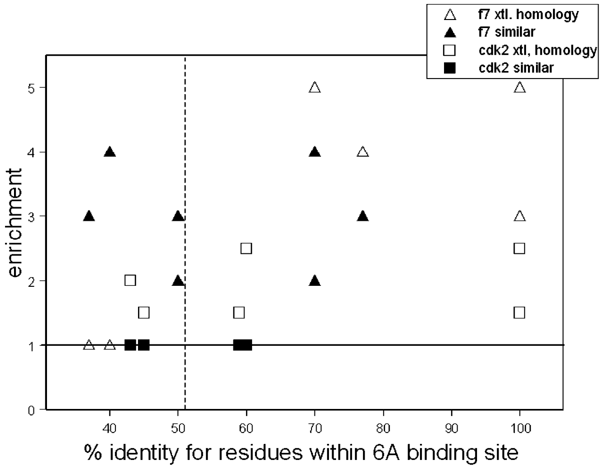

Figure 2. Plot of enrichment of scaffolds selected by docking vs sequence identity of residues near the binding site. Values at 100% identity are the actual crystal structures (i.e., 1dan and 1cvw for factor VIIa; 1hck and 1ckp for CDK2). Random (enrichment equal to 1) is given by horizontal line.

Table 2. Average Scaffold Enrichment Using Homology Models

target	crystal structure	sequence id >50%	sequence id <50%
factor VIIa	4.0	3.8	1.7
CDK2	2.0	2.0	1.8
both	3.0 ± 1.5	3.1 ± 1.4	1.6 ± 0.5

As can be seen in Figure 1, enrichment is always better than random. In addition, the data in Figure 1 can be divided roughly into two groups: those with enrichment values from homology with sequence identity greater than ~60% and those with sequence identity less than or equal to 50%. Using the two-sample *t*-test to test whether the two groups could come from the same distribution and thus have the same mean (null hypothesis) yields $P = 0.059$ ($t = 2.16$ for df = 9), with a 95% confidence interval of the true difference of the mean values −0.16 to 6.73. Although the *P* value is somewhat high, in conjunction with additional statistical analysis dividing the data into such groups (described below) we feel it is sufficient to consider these groups separately. For simplicity, we refer to the data into those with sequence identity greater than 50% and those with less than 50% identity.

When the sequence identity is above ~50%, the enrichment is consistently far better than random. This is true for both homology models and template similar structures. The enrichment for the homology models with sequence identity above 50% is, on average, 5 times better than random, roughly the same as for the actual target crystal structures. For the "similar" structures, the enrichment is about 3 times better than random. With lower sequence identity, enrichment is roughly 2 times better than random, taking both homology models and "similar" structures together. (Note that in Figure 1, the symbol for the enrichment value (2.3) of a kinase homology model is hidden underneath the enrichment value for a "similar" structure.)

In Figure 1, one of the homology models with sequence identity of approximately 50% performs extremely well. This outlier behavior disappears when larger numbers of top-ranking molecules are considered; in particular, when the 1000 top-ranking molecules are

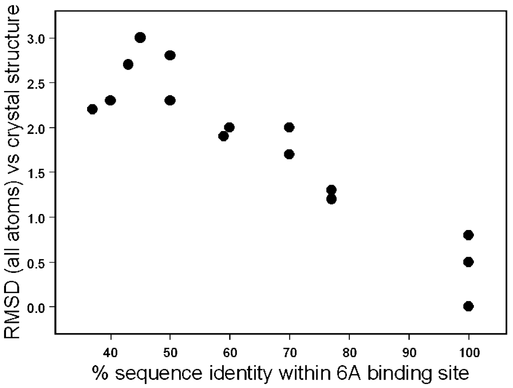

Figure 3. Plot of rmsd (of homology model relative to crystal structure) vs sequence identity of residues near the binding site. Three groups are evident: (a) sequence identity 100%, rmsd < 1; (b) sequence identity of >50%, rmsd < 2; (c) sequence identity of <50%, rmsd > 2.

used in the analysis, the enrichment for this model is reduced to the level of the other like homology models.

Average enrichment values for the reference target crystal structures and homology models only are summarized in Table 1. The mean and standard deviations for the homology models taken together is tabulated there. We note that for CDK2, the performance of the homology models is essentially flat; that is, we consider an enrichment of 3.1 roughly equivalent to 3.5.

Scaffold enrichment is given in Figure 2. Again, values at 100% identity are the actual crystal structures (i.e., 1dan and 1cvw for factor VIIa; 1hck and 1ckp for CDK2). As can be seen there, when the sequence identity is above 50%, the scaffold enrichment is more likely to be better than random. For homology models, the enrichment is roughly 3 times better than random, again comparable to enrichments for the target crystal structure. For "similar" protein structures, the scaffold enrichment is 2 times better than random. (Note that some data points are coincident in Figure 2.) With lower sequence identity, scaffold enrichment is roughly 2 times random, taking both homology models and "similar" structures together. Average scaffold enrichment values for the target crystal structures and homology models are summarized in Table 2. The mean and standard deviations for the homology models taken together are also tabulated in Table 2.

Considering both the factor VIIa and CDK2 datasets together, enrichment of active compounds and scaffolds for homology models generated from templates with sequence identity greater than 50% is similar to enrichments from the target crystal structure. An explanation for this 50% cutoff may be found in the quality of these homology models: the better they resemble the crystal structure, the more the performance becomes comparable. This can be seen in Figure 3, which contains a plot of the root-mean-square deviation (rmsd) between the binding site residues of the homology models and the target crystal structures (one "reference" crystal structure was used, 1dan for factor VIIa and 1hck for CDK2) as a function of sequence identity. All backbone atoms of all residues near the binding site were used in these calculations.

Roughly three clusters of data appear in Figure 3: (i) the reference crystal structures (100% homology), which typically have rmsd values less than 1 Å; (ii) the homology models with sequence identities greater than 50%, which have rmsd values less than 2 Å; (iii) the homology models with sequence identities less than 50%, which display rmsd values greater than 2 Å. Using the two-sample *t*-test reveals that groups ii and iii come from different population distributions ($P = 0.0018$; $t = -4.35$; df = 9). Thus, when the homology model is within 2 Å of the true crystal structure, virtual screening of a database of compounds is likely to identify active compounds and novel scaffolds. Since there is no a priori way to know the quality of the homology model in the absence of the corresponding crystal structure, the 50% sequence identity rule-of-thumb seems a pragmatic way to estimate the performance of database docking studies.

The performance of homology models and the "template" structure upon which the homology model was based was also compared. We examined the different number of active molecules identified as well as the different number of scaffolds identified. We found that when the sequence identity to the target structure was greater than 50%, the homology models on average identified three more active molecules and two more active scaffolds. At lower sequence identity, the number of active molecules and scaffolds identified was roughly the same. This suggests that for proteins with sequence identity greater than 50%, the time and effort to generate a homology model would be worthwhile, since a greater number of actives and, in particular, more scaffolds would be found. That is *not* to say that crystal structures of similar targets, with high sequence identity to the target structure, would not identify active compounds; they do. Rather, the homology models generated from such structures would identify even more scaffolds. On the other hand, at lower sequence identity, using a homology model or a "similar" structure would identify a comparable number of actives; consequently, generating such a model in this case may not be necessary.

Discussion

It has become routine in drug discovery projects to perform virtual screening using computer programs such as DOCK[4,17−19] in order to prioritize compounds for biological testing. In nearly all published studies, one or more crystal structures of the target protein have been used for the docking. An exception is the work of Ring et al.[20] who identified inhibitors of serine and cysteine proteases on the basis of homology models. However, to the best of our knowledge, a systematic study of the performance of homology models in the context of identifying novel leads has not been done before. Schafferhans and Klebe[21] used homology models to correctly identify crystal binding modes of bound ligands. They found, similarly to us, that a 40% sequence identity was needed in order to obtain reliable results.

It remains possible that a sequence identity lower than 50% would be sufficient to generate a homology model that would *consistently* identify active compounds. The 20−30% homology band has been called, by some authors, the "twilight zone" because the quality of homology models may vary widely.[22] Generally, when the sequence identity exceeds 30%, reliable homology models can be constructed.[23] It should be noted that our 50% cutoff was derived using homology models generated with only a single template structure. Other homology modeling programs use multiple structures rather than single structures.[24] Such models could have an rmsd vs the true target crystal structure of less than 2 Å and, as our data suggest, identify novel leads and have enrichments comparable to that of the true crystal structure. In this work, we were not trying to evaluate the quality of homology models but rather to provide some guidelines for when homology models could be useful in docking.

It is also possible that alternative methods of scoring could improve our results. We simply used the DOCK energy scoring. A variety of different scoring functions exist, and their ability to identify leads has been shown to vary.[25] In addition, consensus scoring has been used to improve the hit rates in docking.[19] Conceivably, different homology models based on the same template structure could be used and a consensus score could be derived in this manner.

The enrichment values for the model-built structures are very good when compared to the crystal structure enrichment. The crystal structure enrichments for retrieving active compounds (5.7 for factor VIIa; 4.5 for CDK2) are in fact rather low in light of the maximal possible enrichment (60 for factor VIIa; 46 for CDK2). There could be many reasons for this. Although the crystal structures used here are all of reasonably high resolution (for 1dan, 2.0 Å; for 1cvw, 2.28 Å; for 1hck, 1.90 Å; for 1ckp, 2.05 Å), representing an estimated standard deviation in atomic coordinates of less than 0.4 Å,[26] they each represent only a single low-energy conformation of the protein. For factor VIIa there can be considerable side chain movement and some inhibitors cannot dock with the protein adopting these rotamer conformations. For CDK2, all compounds of particular scaffolds could be docked into the ATP binding site, indicating again that these particular crystal conformations could not accommodate these scaffolds.

Finally, we note that although we have used for simplicity 50% as a cutoff, our data cover only sequence identity greater than ∼60% and sequence identity less than or equal to 50%. To be exact, two values are needed to characterize our data set. At greater than ∼60% identity, the enrichment is ∼5 times better than random, but with sequence identity less than ∼50%, the enrichment is ∼2 times better than random.

Conclusions

When a homology model is generated using a template structure with sequence identity greater than 50% to the target receptor, the number of active molecules identified was on average 5 times the number that would be found randomly while the number of active scaffolds was on average 3 times better than random. "Similar" structures (the underlying template structure used to generate the homology model) identified fewer active molecules and active scaffolds at this level of sequence identity. We suggest that the 50% cutoff is due in part to the quality of the homology model; above this value, the rmsd vs the crystal structure is in general less than 2 Å.

References

(1) Baker, D.; Sali, A. Protein structure prediction and structural genomics. *Science* **2001**, *294*, 93−96.
(2) Sanchez, R.; Pieper, U.; Melo, F.; Eswar, N.; Marti-Renom, M. A.; et al. A Protein structure modeling for structural genomics. *Nat. Struct. Biol.* **2000**, November (Suppl.), 986−990.
(3) Schneider, G.; Bohm, H.-J. Virtual screening and fast automated docking methods. *Drug Discovery Today* **2002**, *7*, 64−70.
(4) Shoichet, B. K.; McGovern, S. L.; Wei, B.; Irwin, J. J. Lead discovery using molecular docking. *Curr. Opin. Chem. Biol.* **2002**, *6*, 439−446.
(5) Doman, T. N.; McGovern, S. L.; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R.; et al. Molecular docking and high throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J. Med. Chem.* **2002**, *43*, 2213−2221.
(6) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. E.; Brice, M. D.; et al. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **1977**, *112*, 535−542.
(7) Chemical Computing Group, Montreal, Canada.
(8) Available Chemicals Directory; Molecular Design, Ltd., San Leandro, CA.
(9) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. A geometric approach to macromolecular-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.
(10) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated docking with grid-based energy evaluation. *J. Comput. Chem.* **1992**, *13*, 505−524.
(11) Ewing, T.; Kuntz, I. D. Critical Evaluation of Search Algorithms for Automated Molecular Docking and Database Screening. *J. Comput. Chem.* **1997**, *18*, 1175−1189.
(12) Smellie, A.; Henne, R.; Stanton, R.; Teig, S. Conformational Analysis by Intersection: CONAN. *J. Comput. Chem.* **2003**, *24*, 10−20.
(13) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; et al. A new force field for molecular mechanics simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **1984**, *106*, 765−784.
(14) Weiner, S. J.; Kollman, P. A.; Hguyen, D. T.; Case, D. A. An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem.* **1986**, *7*, 230−252.
(15) Marsili, M.; Gasteiger, J. π charge distribution from molecular topology and p orbital electronegativity. *Croat. Chem. Acta* **1980**, *52*, 601−614.
(16) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity: a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3210−3228.
(17) Somoza, J. R.; Skillman, A. G.; Munagala, N. R.; Oshiro, C. M.; Knegtel, R. M. A.; et al. Rational design of novel antimicrobials: blocking purine salvage in a parasitic protozoan. *Biochemistry* **1998**, *37*, 5344−5348.
(18) Kick, E. K.; Roe, D. C.; Skillman, A. G.; Liu, G.; Ewing, T. J.; Sun, Y.; Juntz, I. D.; Ellman, J. A. Structure-based design and combinatorial chemistry yield low nanomolar inhibitors of cathepsin D. *Chem. Biol.* **1997**, *4*, 297−307.
(19) Charifson, P.; Corkery, J.; Murcko, M.; Walters, W. P. Consensus Scoring: A Method for Obtaining Improved Hit Rates for Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem* **1999**, *42*, 5100−5109.
(20) Ring, C. S.; Sun, E.; McKerrow, J. H.; Lee, G. K.; Rosenthal, P. J.; et al. Structure-based inhibitor design by using protein models for the development of antiparasitic agents. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 3583−3587.
(21) Schafferhans, A.; Klebe, G. Docking ligands onto binding site representations derived from proteins. *J. Mol. Biol.* **2001**, *307*, 407−427.
(22) Sternberg, M. J. E. Protein structure prediction: principles and approaches. *Protein Structure Prediction*; Oxford University Press: Oxford, 1996; pp 1−30.
(23) Sanchez, R.; Sali, A. Advances in comparative protein-structure modelling. *Curr. Opin. Struct. Biol.* **1997**, *7*, 206−214.
(24) Insight II, 2000, Accelrys, San Diego, CA.
(25) Bissantz, C.; Folkers, G.; Rogman, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.
(26) Klebe, G.; Holger, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem., Int. Ed.* **2002**, *41*, 2644−2676.

JM0300781