JCTC Journal of Chemical Theory and Computation

Ensemble Docking from Homology Models

Eva Maria Novoa,[†] Lluis Ribas de Pouplana,[‡,§] Xavier Barril,[§,||] and Modesto Orozco*[,†,⊥]

Joint IRB-BSC Research Program in Computational Biology, Institute for Research in Biomedicine, Josep Samitier 1−5, Barcelona 08028, Spain, Cell and Developmental Biology, Institute for Research in Biomedicine, Josep Samitier 1−5, Barcelona 08028, Institució Catalana per la Recerca i Estudis Avançats, Passeig Lluis Companys 23, Barcelona 08010, Spain, Departament de Fisicoquímica, Facultat de Farmàcia, Avgda Diagonal sn, Barcelona 08028, Spain, and Structural Bioinformatics Node Instituto Nacional de Bioinformática, Institute of Research in Biomedicine, Josep Samitier 1−5, Barcelona 08028, Spain

Received May 10, 2010

Abstract: We present here a systematic exploration of the quality of protein structures derived from homology modeling when used as templates for high-throughput docking. It is found that structures derived from homology modeling are often similar in quality for docking purposes than real crystal structures, even in cases where the template used to create the structural model shows only a moderate sequence identity with the protein of interest. We designed an "ensemble docking" approach based on the use of multiple homology models. The method provides results which are usually of better quality than those expected from single experimental X-ray structures. The use of this approach allows us to increase around five times the universe of use of high-throughput docking approaches for human proteins, by covering over 75% of known human therapeutic targets.

Introduction

New algorithms and computers are making possible the use of atomistic docking approaches in a high-throughput (HTD) regime, being possible to screen in silico libraries containing 10^5-10^6 compounds against a limited number of protein targets.[1−3] However, we cannot ignore that the requirement of computational efficiency implies the introduction of severe simplifications in both the description of molecular interactions and the coverage of the conformational space of ligands and proteins.[4−6] As a result, docking methods have problems in representing ligand-induced conformational changes in the protein, and in general the quality of docking algorithms decreases as the docked drug differs from that bound in the crystal structure.[7,8] However, despite all these limitations, the power of current docking algorithms is beyond all doubt, and many authors have demonstrated that their use largely enriches the possibility to find a good binder from a large library of decoys and that the proposed optimal poses are good starting points for lead-optimization processes.[9−12] It is not surprising, then, that virtual screening based on docking algorithms is a routine task in medicinal chemistry laboratories.[11,13,14]

The inputs of docking algorithms are ligand and protein structures, and the outputs are a series of "poses", i.e., possible configurations of the protein−ligand complex, which are then scored using an empirically refined function yielding to a small subset of preferred binding modes with the associated binding affinity.[15−17] Given the number of approximations done in a docking algorithm, the practical purpose of HTD is not the accurate ranking of potential

* Corresponding author phone: 0034-93-4037155; e-mail: modesto@mmb.pcb.ub.es.

[†] Joint IRB-BSC Research Program in Computational Biology, Institute for Research in Biomedicine.

[‡] Cell and Developmental Biology, Institute for Research in Biomedicine.

[§] Institució Catalana per la Recerca i Estudis Avançats.

[||] Facultat de Farmàcia.

[⊥] Structural Bioinformatics Node Instituto Nacional de Bioinformática, Institute of Research in Biomedicine.

10.1021/ct100246y
Published on Web 07/16/2010

binders, but the enrichment of true binders among the top-ranked compounds and the recovery of good leads for refinement.

The need to have a three-dimensional structure of the target protein strongly limits the use of docking algorithms, and despite the impressive advance of structural genomics, the number of proteins for which experimental structure is known represents only a small fraction of the total proteome. Thus, the 2010 version of the Protein Data Bank (PDB) contains around 60000 entries, but only 42.5% (25560) of them correspond to unique proteins from which only 15% (3935) are human.[18,19] In comparison, sequence analysis suggests that the total number of human proteins ranges between 20332—Swissprot[20]—and 93110—RefSeq[21]—probably twice or more if spliced forms are considered,[22] which means that the PDB covers only between 2 and 19% of the human proteins. The gap between structure and sequence becomes even larger if we consider proteins from virus, bacteria, or other pathogens for which little structural information exists.

Protein structure can be predicted by a variety of computational methods,[23] homology modeling (also named comparative modeling) being the most accurate one in cases where there is a clear sequence identity between the target protein and at least one template with known three-dimensional structure.[24,25] The quality of the structure derived from homology modeling roughly correlates with the sequence identity between the target protein and template proteins.[26] Thus, it is accepted that for sequence identities below 30% less than half of the residues have their C_α correctly placed.[27,28] The percentage of correctly placed residues increases to 85% for identities ranging from 30 to 50%, and most of the C_αs are well-positioned for sequence identities above 50%. Inside the high-quality range no direct correlation exists between the accuracy of the model and the sequence identity with the template, and evaluation of the expected quality of a model is still an unsolved problem.[29] In fact, the concept of "goodness" is not unique, since it depends on its planned use.[30] For example a model with an accuracy around 3.5 Å in backbone positioning may be good enough for understanding protein function or designing mutations but is expected to be of small utility for prediction of ligand binding.[26−31]

Different authors have tried to evaluate the quality of homology models for docking experiments. Thus, McGovern and Shoichet performed high-throughput docking on 10 target enzymes for which apo, holo, and homology model structures were available, finding that they were useful for enriching the screening, but not as powerful as the holo-crystal structure.[32] Diller and Li reported good enrichments (in some cases similar to those obtained with the crystal structure) when model structures of six kinases obtained for identities in the range of 30−50% were used to screen a large library.[33] Similar results were obtained by Oshiro et al.[34] in the study of two targets (CDK2 and factor VIIa), by Gilson's group with a set of five targets,[35] and by Ferrara and Jacoby in the analysis of insulin growth factor I receptor.[36] In a very recent paper Fan et al.[37] found good results when ensembles of homology models of several proteins were used to screen for ligands in the DUD database[38] using the DOCK computer program.[39] All these studies illustrate the power of homology models to guide docking experiments but also underline their limitations related to the lack of "a priori" evaluations on the quality of the model for docking purposes and on the problems of selecting a priori a structural model from the battery of solutions given by homology modeling routines (for discussion see ref 36).

The introduction of protein flexibility is the next step in docking, and there is a significant amount of work focused in this direction.[40,41] Among the different approaches suggested, "ensemble docking" (also known as multiple docking) is one of the most popular ones. It assumes that the effect of target flexibility in docking can be represented by using a Boltzmann ensemble of conformations for the protein instead of just a single rigid structure. Different methods for generating ensembles have been proposed, including molecular dynamics[42,43] (from a known experimental structure of the target), crystallographic (X-ray),[44−47] and spectroscopic (NMR).[48,49] All these approaches require experimental knowledge of protein structure and are then able to cover just a small fraction of proteome. In this contribution, following the pioneering work by Fan et al.,[37] we explore the possibility of using ensembles derived from homology (comparative) modeling. This approach is simple and fast and, if successful, would allow us to dramatically expand the range of applicability of ensemble docking approaches. We explored, with a wide range of metrics and for a large number of proteins, not only the ability of the approach to enrich in active ligands drug libraries but also the structural quality of the docking predictions, a crucial element in lead optimization procedures. We designed and tested a procedure to perform ensemble docking based on the combination of Modeller[50] and Glide,[51] finding that the results are in general of better quality than those expected when a single-crystal structure is used as a template in docking experiments.

Methods

Protein Data Sets. We defined two sets of proteins of our study: one for training and another for testing. The training set was defined considering proteins for which at least 30 crystal ligand-bound structures were available in PDB (with the same sequence or at most one single mutation). PDBs with point mutations were only used to build the set of active ligands but were not included in the set of docked proteins. This set of proteins includes thrombin (2cn0, 1ay6, 1bmm, 1tom, 1xm1), renin (2g24, 1bil, 1hrn, 1rne, 2g1r), cyclin-dependent kinase 2—CDK2—(1aq1, 1e1v, 1gz8, 1jsv, 3ddq), and protein tyrosine phosphatase 1B—PTP-1B—(2f71, 1c83, 1g7g, 1ony, 2h4g). The test set was created using less restrictive conditions in terms of the number of crystallized structures available—at least eight—and contained α-momorcharin (1mrg, 1aha, 1mom, 1f8q, 1ahb), trypsin (1tng, 1tnl, 1f0t, 1lqe, 2by5), p38 kinase (3hp2, 1w7h, 2baj, 3c5u, 3cg2), HIV retrotranscriptase (3jyt, 1dtq, 1rt1, 1s6p, 3dol), factor Xa (2vvc, 1ezq, 1fax, 1lpk, 1nfu), and heat shock protein 90—HSP90—(1yet, 1osf, 1uy6, 1yc4, 2ccs).

Homology Modeling. The derivation of model structures was performed using scripts designed for HTD production

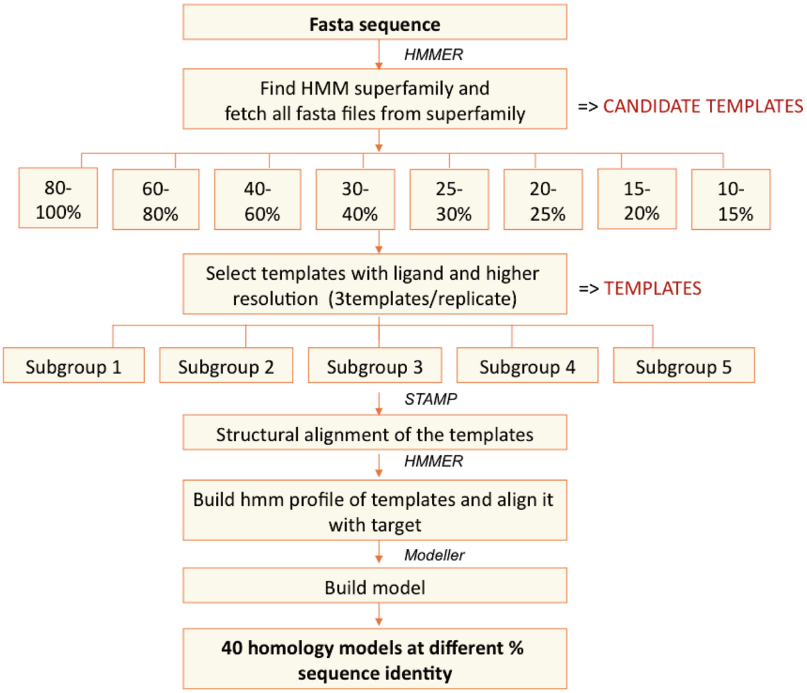
Fasta sequence
HMMER
Find HMM superfamily and fetch all fasta files from superfamily
=> CANDIDATE TEMPLATES
80-100%
60-80%
40-60%
30-40%
25-30%
20-25%
15-20%
10-15%
Select templates with ligand and higher resolution (3templates/replicate)
=> TEMPLATES
Subgroup 1
Subgroup 2
Subgroup 3
Subgroup 4
Subgroup 5
STAMP
Structural alignment of the templates
HMMER
Build hmm profile of templates and align it with target
Modeller
Build model
40 homology models at different % sequence identity

Figure 1. Comparative model building work flow. The process is automated such that a FASTA sequence is given as input, and a total of 40 homology models that range from 10 to 100% sequence identity are obtained. The software used at each step is detailed in Methods.

trying to mimic the standard expert procedure for homology modeling (see Figure 1). We are aware that by using automatic protocols homology modeling might be prone to errors, related mostly to misalignment problems, which can be easily corrected by manual refinement. However, to evaluate a pure HTD scenario, no human refinement was done here, which means that results presented here can be considered a lower limit of accuracy for the technique. Accordingly, the sequence of each target protein was extracted from the PDB, transformed to FASTA-format, and launched against the Pfam-A database[52] using HMMER[53] to assign the sequence to a superfamily. All the FASTA files for which there is a PDB corresponding to this same superfamily—all the candidate templates—were retrieved and aligned to the target sequence using ClustalW.[54] After this procedure each template was placed into different categories depending on its percentage sequence identity: 80−100, 60−80, 40−60, 30−40, 25−30, 20−25, 15−20, and 10−15%. For each sequence-identity category we selected 15 templates considering only proteins bound to ligand and solved at the highest resolution possible. Each set of 15 templates was divided into five subsets in order to build five different models per sequence-identity category. It is important to remark that the five models per sequence-identity category were built on the basis of different templates. Such templates were structurally aligned by STAMP,[55] creating then a profile using HMMER, which was introduced as a meta-template for alignment of the target sequence (see Figure 1). Finally, the 9v5 version of MODELLER[50] was used to create structural models using default options.

Ligand Selection. The active ligands to dock were downloaded from the PDB database (www.rcsb.org), by selecting all available X-ray ligands from PDB complexes for each of the proteins of the study. All the available ligands were subjected to similarity analysis using MOE[56] implementation of MACCS structural fingerprints[57] and distributed in 80% identity clusters. Only one compound per cluster was selected, which guarantees the diversity of the ligands, avoiding bias derived from the overrepresentation of the same scaffold. The set of known ligands was mixed with 1000 diverse “decoys” (molecules not described as binders for these proteins) which were selected from the most populated clusters obtained using Reynolds’ algorithm at a similarity cutoff level of 60%[58] on a local database containing 1.7 million commercially available compounds—already filtered by drug-likeness criteria: Lipinski rules, Veber rules, and lack of reactive groups.[59−61] The percentage of active ligands ranged from 0.5 to 10%, depending on the protein.

Docking Procedure. Ligand screening and docking was performed using the Glide 5.0 program.[51] The extraprecision Glide docking (Glide XP) protocol was used for the training set, while the standard-precision (Glide SP) protocol was used for the test set, trying then to mimic a normal HTD procedure (in practice, we found very small differences between both scoring functions). Starting from the PDB structures, ligands were prepared using the LigPrep[62] facility in Schrödinger utility MAESTRO,[63] by generating low-energy ionization and tautomeric states within the range of pH 7.0 ± 2.0. All ligands were energy-minimized using the OPLS_2005 force field implemented in MAESTRO.[63] The setup of proteins was done with the Protein Preparation Wizard facility, which included hydrogen optimization, protonation, and geometry optimization using again the OPLS_2005 force field. The receptor grid defining the docking universe was built centered on the crystallographic ligand, which was then removed as any other nonprotein molecule.

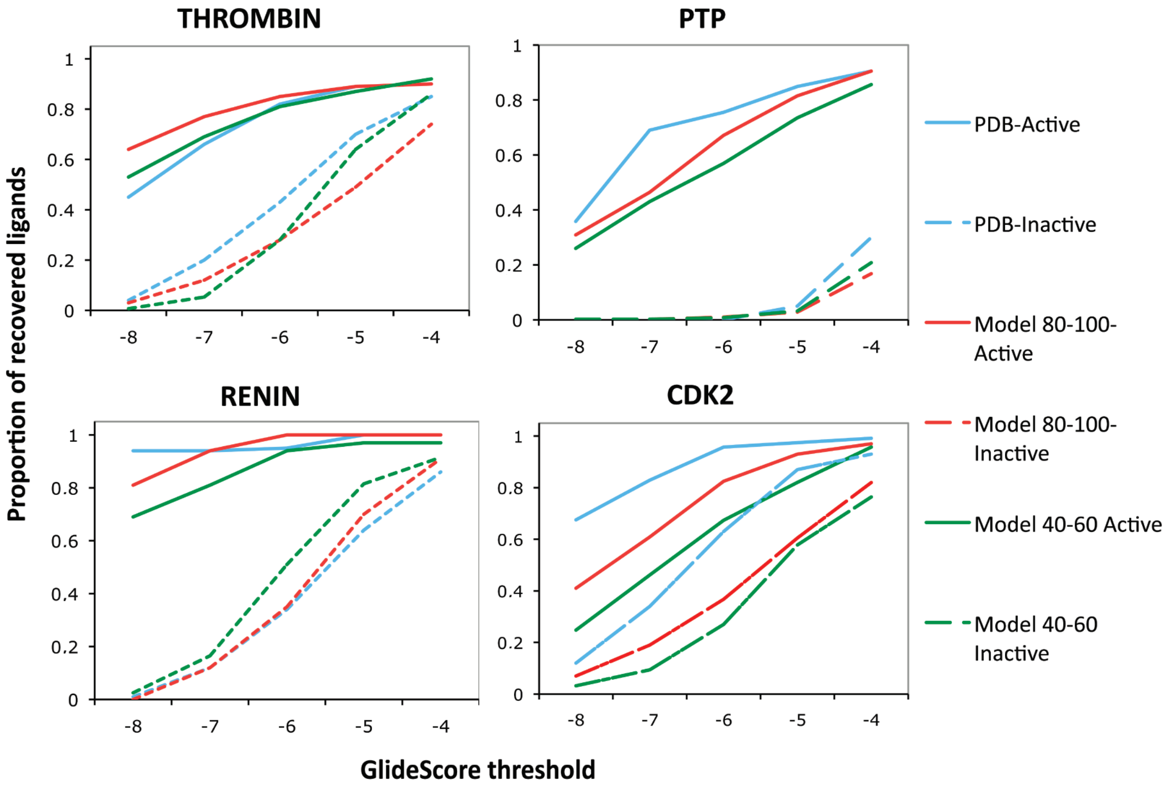

Figure 2. Recovery of active (nondashed lines) and inactive ligands (dashed lines) for each of the proteins of the training set. As can be seen in the plots, both active and inactive ligand recovery increases as the GS threshold decreases. The different colors correspond to different sequence identity ranges: blue (PDB), red (model 80−100), and green (model 40−60). The selected PDBs are all high-resolution holo conformations: 2cn0, 1aq1, 2g24, and 2f71, which correspond to thrombin, cdk2, renin, and PTP-1B, respectively.

Metrics for Preevaluation of Model Quality. The structural quality of the model was evaluated using both global and local parameters. The global quality indexes included global root-mean-square deviation (rmsd; model-reference PDB), global sequence identity, number of gaps in the alignment, and sequence coverage of the model. The local parameters were always referred to the binding site (defined as the set of residues with at least one atom at less than 5 Å from the crystal ligand) and included binding site rmsd, binding site sequence identity, and atom conservation in binding site structure. All rmsd measures were computed using the MMTSB tool set.[64]

Metrics for Evaluation of Success in Docking. The success of docking was measured by analyzing the following: (1) the ability of the models to predict the structure of the ligand−protein complex and (2) the applicability of the models for virtual screening purposes. The ability of models to predict the structure of the complex was assessed by (i) measuring the proportion of docked poses with rmsd below 2 Å from crystal structure using an SVL script in MOE, (ii) measuring the rmsd obtained when comparing the best-docked pose (rmsd-based selection) and the best-ranked pose (GlideScore-based selection) with the crystallographic ligand, and (iii) measuring the similarity between ligand−protein contact maps in models and crystal structures, which are determined by comparing the number of atoms that are conserved from those found at less than 5 Å from the docked ligand−compared to the original PDB where the docked ligand is found. Thus, for each docked ligand, a different ligand−protein map is built and compared to its corresponding PDB.

The utility of the models for virtual screening purposes was evaluated by assessing the performance of the homology models to discriminate between active compounds and decoys (inactive). A virtual screening run selects a list of molecules (n) from a given database of N entries, which includes both actives (true positive compounds, TP) and decoys (false positive compounds, FP). Actives (A) that have not been found by the screening method are false negatives (FN), and decoys that have not been selected are true negatives (TN). The optimum screening is that able to recover all true positives, without recovering any false positive.

Many different enrichment descriptors described in the literature have been considered in this work.[65,66] First we computed the *sensitivity* (true positive rate; TPR; see eq 1) and the *specificity* (true negative rate; TNR see eq 2) indexes. The first indicates the ability of the method to recover the real ligands, while the second informs on its ability to avoid decoys.

$$\text{sensitivity} = \text{TPR} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{1}$$

$$\text{specificity} = \text{TNR} = \frac{\text{TN}}{(\text{FP} + \text{TN})} = 1 - \text{FPR} \tag{2}$$

where FPR stands for false positive rate.

The *accuracy* (Acc; eq 3) index was used to describe the percentage of molecules which have been correctly classified by the screening protocol, while *precision* (positive predictive value; PPV) was used to describe the proportion of true positives among the list of selected compounds given by the docking (eq 4).

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{N} = \frac{A}{N}\mathrm{TPR} + \left(1 - \frac{A}{N}\right)\mathrm{TNR} \qquad (3)$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})} \qquad (4)$$

To assess the ability of the models to obtain true actives among the first ranked compounds (an extra requirement in HTD studies[67]), the *enrichment factor* (EF, eq 5) was used.

$$\mathrm{EF} = \frac{\mathrm{TP}/n}{A/N} \qquad (5)$$

ROC (receiver operating characteristic; true positive versus false positive rates) curves and the associated AUC curves (area under the ROC curve) have also been used to determine the discriminatory power of the virtual screening procedure. These metrics are especially powerful since they are not dependent on the ratio of actives to decoys of the database.[68−70]

Results and Discussion

Structural Quality of the Models. Modeller[50] provides good global models when using structural templates with sequence identities above 25% (Supporting Information Figures S1 and S2). The use of templates with sequence identities below such a threshold can yield wrong structures due mainly to alignment errors or to the presence of large unfolded regions. The atom conservation—i.e., the similarity between ligand−protein contact maps—at binding sites grows faster than global sequence identity, and for identities as small as 25−30% around 60−70% of the atoms at the experimental binding site are conserved in the model (Supporting Information Figure S3). The heavy-atoms rmsd between model and real binding sites are typically below 2 Å for sequence identities above 25% (Supporting Information Figure S4). Clearly, then, structural models created using homology modeling not only reproduce well global protein structure but also provide quite important details of the binding site. Whether or not the quality of these details is enough for drug docking studies will be the main subject of discussion in the remaining of our communication.

Docking Enrichment Using Single-Structure Homology Models. The second point to analyze was the quality of single homology models when used to recover specifically active ligands from a mixture of ligands and decoys. Within the Glide framework the number of hits recovered in a docking depends on the scoring (GS) threshold. For very restrictive GS values very few decoys (false positives) are recovered, but many real ligands might be lost. On the contrary, when very permissive GS values are used, all real ligands are recovered, but at the expense of increasing dramatically the number of incorrectly selected decoys. Results shown in Figure 2 demonstrate that using a single PDB structure as

Table 1. Training Set Enrichment Descriptors[a]

		PDB		model 80−100		model 60−80		model 40−60		model 30−40		model 25−30		model 20−25		model 15−20		model 10−15	
	GS	SG	ENS	SG	ENS	SG	ENS	SG	ENS	SG	ENS	SG	ENS	SG	ENS	SG	ENS	SG	ENS
sensitivity	−8	0.56	0.81	0.49	0.77	0.43	0.74	0.38	0.67	0.32	0.60	0.26	0.55	0.22	0.37	0.14	0.26	0.11	0.20
	−7	0.76	0.92	0.66	0.86	0.62	0.88	0.57	0.83	0.47	0.75	0.45	0.79	0.39	0.66	0.27	0.50	0.20	0.38
	−6	0.87	0.97	0.82	0.94	0.78	0.92	0.74	0.91	0.63	0.88	0.62	0.91	0.54	0.80	0.41	0.70	0.30	0.51
specificity	−8	0.95	0.85	0.94	0.88	0.96	0.88	0.96	0.88	0.97	0.91	0.98	0.93	0.98	0.96	0.99	0.97	0.99	0.97
	−7	0.88	0.71	0.88	0.76	0.90	0.76	0.90	0.75	0.94	0.80	0.94	0.77	0.94	0.81	0.91	0.86	0.94	0.90
	−6	0.69	0.44	0.73	0.51	0.74	0.48	0.71	0.45	0.80	0.51	0.81	0.49	0.82	0.56	0.80	0.66	0.85	0.75
EF (1%)	–	24.40	26.49	21.72	26.09	22.3	24.58	19.01	21.49	17.95	23.08	15.19	19.66	17.67	21.29	11.35	13.62	6.84	9.74
accuracy	−8	0.93	0.85	0.91	0.87	0.92	0.87	0.92	0.87	0.93	0.90	0.93	0.91	0.93	0.92	0.93	0.92	0.93	0.92
	−7	0.88	0.72	0.87	0.77	0.88	0.77	0.88	0.76	0.91	0.80	0.91	0.78	0.90	0.80	0.87	0.83	0.88	0.86
	−6	0.70	0.48	0.74	0.53	0.74	0.51	0.71	0.48	0.80	0.54	0.80	0.52	0.80	0.57	0.77	0.66	0.81	0.73
PPV	−8	0.52	0.37	0.33	0.29	0.40	0.31	0.39	0.32	0.40	0.34	0.47	0.50	0.49	0.45	0.48	0.41	0.48	0.41
	−7	0.42	0.31	0.36	0.27	0.39	0.30	0.43	0.33	0.53	0.35	0.46	0.39	0.49	0.39	0.42	0.33	0.41	0.33
	−6	0.26	0.18	0.26	0.17	0.25	0.16	0.25	0.16	0.26	0.16	0.28	0.22	0.28	0.20	0.23	0.18	0.23	0.18
AUC	–	0.85	0.95	0.86	0.92	0.84	0.90	0.83	0.89	0.82	0.88	0.78	0.88	0.73	0.83	0.70	0.78	0.56	0.59

[a] Six different enrichment descriptors (sensitivity, specificity, EF for the top 1% ranked compounds, accuracy, PPV and AUC) have been computed for models and PDBs using both single docking (SG) and ensemble docking (ENS) approaches. For each of the cases, enrichment descriptors have been quantified taking different scoring (GS) thresholds: −8, −7, and −6, allowing us to see the difference between models and PDBs not only depending on the docking approach used—single or ensemble—but also depending on the chosen GS threshold. In the case of the EF (1%) and AUC enrichment descr iptors, no GS threshold has been used, given that these descriptors are GS-threshold independent.

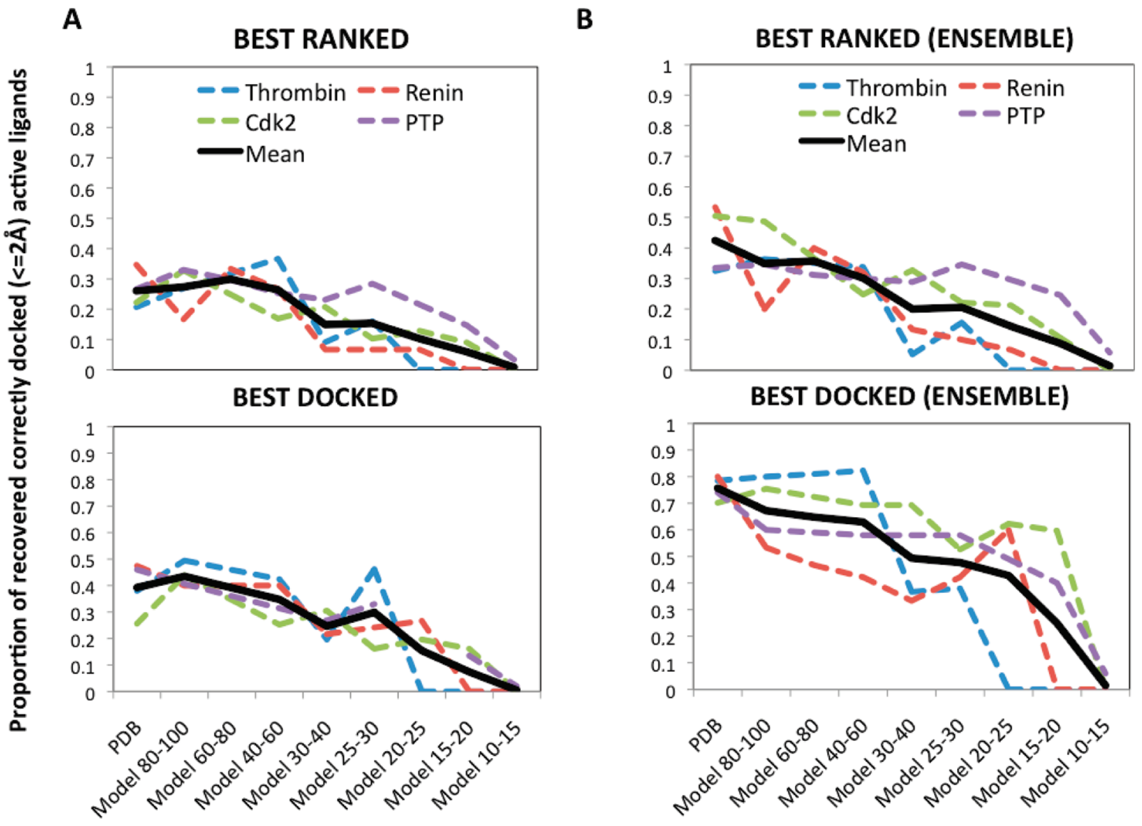

Figure 3. Recovery of correctly docked ligands versus sequence identity of the models. The recovery is defined as the fraction of correctly active docked ligands—less than 2 Å rmsd from the crystal structure—with respect to the total active docked and scored ligands. (A) In the upper plot, the *best-ranked* active ligand pose is chosen from all the proposed poses by using a score-based selection, whereas in the lower plot the *best-docked* ligand pose is chosen by using an rmsd-based selection. (B) Recovery of correctly docked ligands versus sequence identity when using an ensemble docking approach. Both score-based selection—i.e., best ranked—and rmsd-based selection—i.e., best docked—are shown. Each protein of the training set is labeled accordingly, and the mean value of the four training set proteins is shown in black.

template Glide is able to recover typically between 40 and 90% of the real ligands with a small number of false positives for a very strict scoring function threshold ($-$GS $=$ 8). The ratio of true positive increases about 10 percentile points for $-$GS $=$ 7 and 5$-$10 extra points for $-$GS $=$ 6, keeping still an acceptable rate of recovery of false positives; for larger $-$GS values the rate of false positives becomes unacceptable. In any case, the improvement with respect to random selection is very clear, demonstrating the performance of the Glide docking algorithm.

When homology models are used for docking, the performance of Glide is not lost (Figure 2 and Table 1), even in cases where the models are built using proteins with a modest level of homology as templates. It is especially encouraging that in some cases homology models outperform experimental structures for drug docking, a result already found by other authors[32,37] and which encourages the use of modeled protein structures for drug design experiments. The fact that homology models outperform X-ray structure for thrombin might appear surprising but is on the line of previous works with this protein[32] which demonstrated that probably the holo structure of thrombin is overspecialized for ligand binding, with problems arising in cross-docking experiments similar to those performed here. Homology models, less refined for a particular ligand binding mode, are then more successful.

Structural Quality of the Docking Poses Obtained Using Single-Structure Homology Models. The ability of the docking algorithm to capture specifically the maximum of active ligands is the major requirement for hit finding. However, to guide the optimization of the hit, there is an additional requirement: the drug needs to be correctly placed at the binding site. When using an experimental PDB structure as template, Glide is able to find poses that are very close (rmsd < 2 Å) to the bound conformation found in crystal in around 50% of cases, and in fact in more than 30% of cases the best scored poses (typically $-$GS > 8) match the experimental conformation (Figure 3A). Very interestingly, the global performance of the method does not change significantly when single homology models built from sequence identities above 40% are used, and even models built from templates with sequence identities around 25% can provide reasonable results. Again, it is remarkable that for some proteins homology models can provide more accurate binding mode predictions than the experimental structure—e.g., thrombin homology models recover on average 20% more correctly docked ligands compared to the crystallographic structures.

Ensemble Docking versus Single-Structure Docking. Proteins adapt their structure to the bound ligand, which explains the problems of docking methods to recognize active

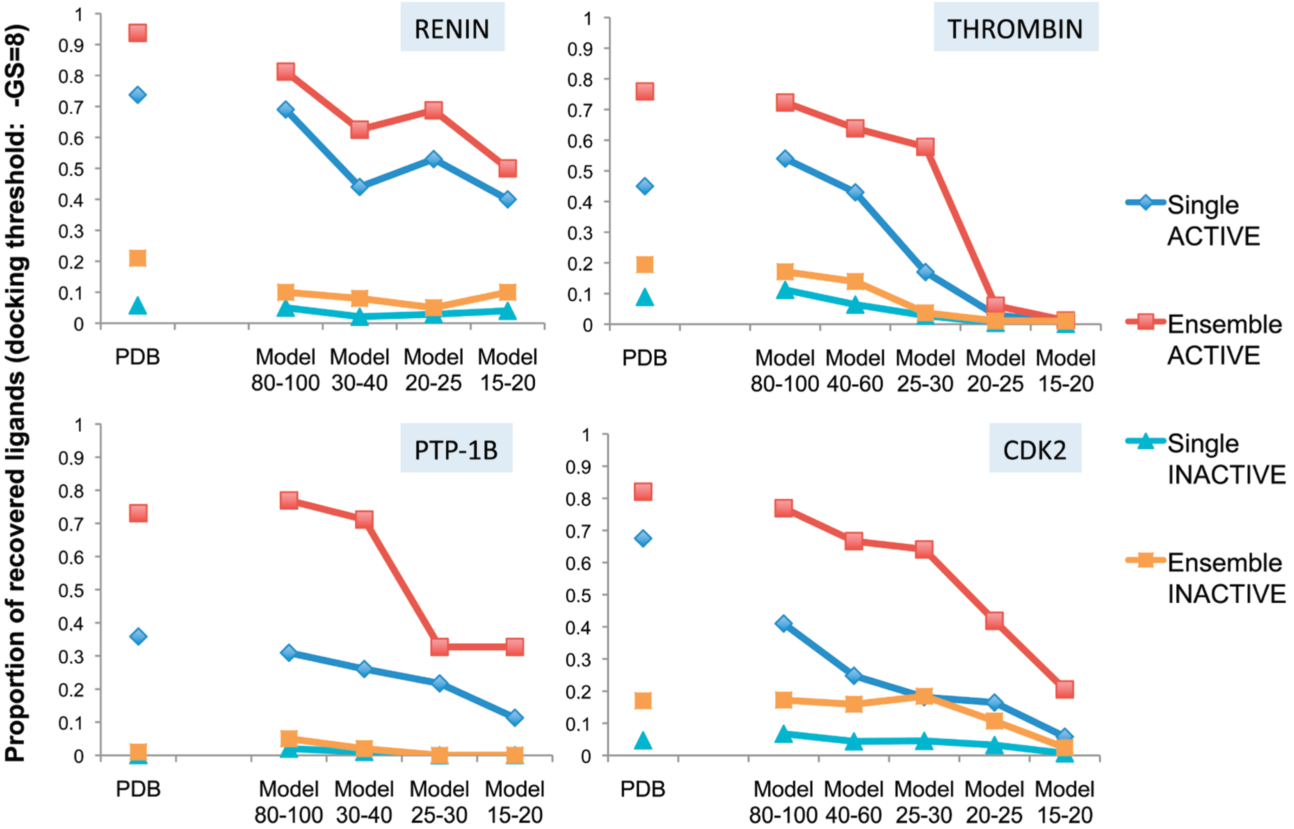

Figure 4. Ensemble docking versus single docking approach. The performance of both approaches is being compared in terms of recovered active ligands and decoys for the four proteins of the training set. The single docking approach performance is shown with blue and cyan lines, which correspond to the recovery of active and inactive ligands, respectively. Similarly, the red and orange lines correspond to the active and inactive ligand recovery, respectively, when using an ensemble docking approach. In all cases, the difference between active and inactive recovery is higher when using ensemble docking. Results shown correspond to $-\text{GS} = 8$.

ligands when the protein structure has been solved in the presence of a very different compound. This problem is graphically illustrated in Figures 3B and Supporting Information Figure S5, which show the dispersion of results that can be obtained for a given protein when different high-resolution X-ray structures are used for docking. We can alleviate this problem by docking the drug against all the protein structures, selecting then as optimal docking mode that with the best scoring. This strategy is known as multiple docking or ensemble docking, which has been used and described in previous papers.[40,71−76] An ensemble of receptor conformations provides a structural degree of freedom that cannot be achieved with other flexible-receptor docking methods, such as induced-fit docking (IFD).[77] In our ensemble docking procedure, we have used five different structures, which is in accordance with the number of receptor structures used in previous papers.[71,73] This ensemble docking procedure (using at this point only experimental structures) leads to a clear improvement with respect to the average situation found when docking was done for single structures if a restrictive GS threshold is used (see Table 1). In fact, for strict threshold values the ensemble docking approach yields in most cases better results than those obtained by using the best “dockable” experimental structure, while the performance can decay for permissive thresholds due to the retrieval of false positives. It is also worth noting that the ensemble docking approach improves also the chances to recover good structural models for lead optimization procedures (compare Figure 3A with Figure 3B, and see Supporting Information Figure S6).

Ensemble Docking from Homology Models. The preceding analysis suggests that in general better docking results are obtained if all the experimental structural information of a protein is used as input for an ensemble docking procedure. The question is now, whether or not this situation is maintained for the less accurate ensembles generated by comparative modeling. Results in Table 1 demonstrate that the use of ensembles increases very significantly sensitivity (70−100%) with respect to single models, decreasing only slightly the specificity (around 6% for $-\text{GS} \geq 8$), leading to an overall improvement in the docking results. Thus, improvement made by the use of ensemble docking is more important in cases where the initial structures have lower AUCs, such as those in homology models.

Homology-modeling based ensemble docking coupled with good structural models and strict scoring thresholds outperforms in most cases single-structure docking performed using experimental structures (Table 1 and Figure 4). In fact, the quality of the ensemble docking results for accurate homology models (sequence identity above 80%) is indistinguishable from those obtained using experimental ensembles, and on average more than 80% of active ligands are recovered with a small percentage of recovered decoys

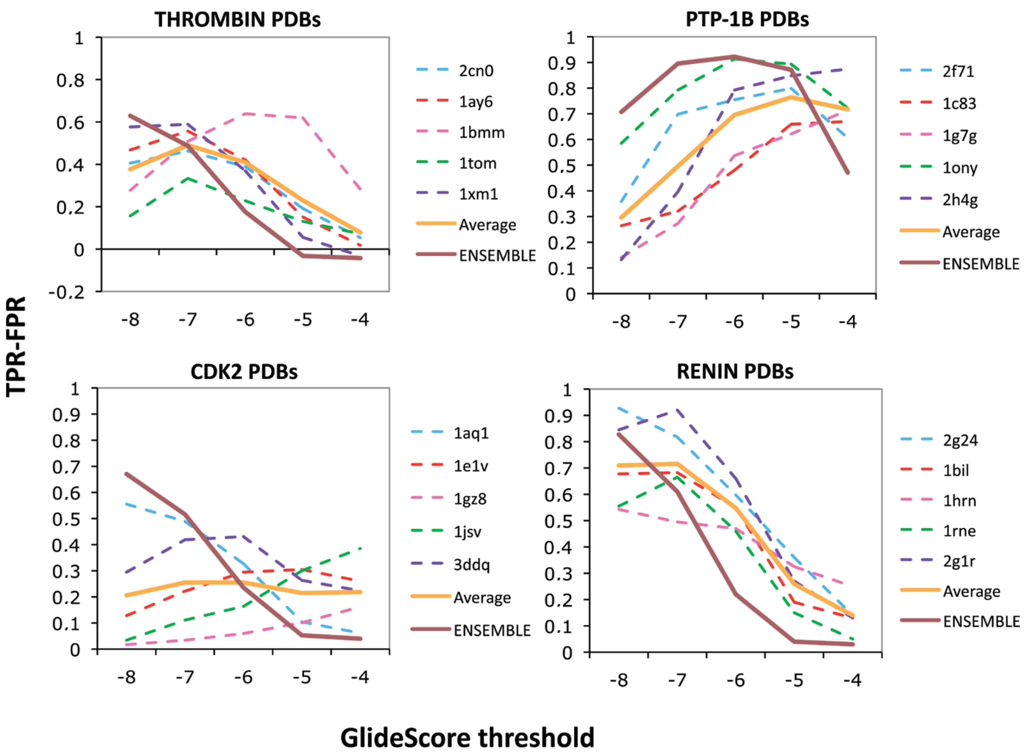

Figure 5. Performance of ensemble PDBs versus single PDBs. Performance (*y*-axis) is measured as the difference between the true positive rate (TPR) and the false positive rate (FPR). The dashed lines correspond to single PDBs, whereas the nondashed lines correspond to the average and ensemble of the PDBs, labeled with orange and brown, respectively. In all cases except for renin, the ensemble performs better than any single PDB at strict GS thresholds. However, in all proteins of the training set, the ensemble's performance decreases more rapidly—i.e., has higher slope—than any of the single PDBs.

(Figure 4) when homology ensembles are used. The ensemble docking protocol is very robust to the decrease in sequence identity, given that models with sequence identities in the range of 30−40% still provide good results. On the contrary, the protocol outlined here is very sensitive to the scoring threshold used, and less strict GS values increase excessively the recovery of false positives (Table 1 and Supporting Information Figures S7 and S8).

Finally, it is worth noting the large structural quality of the complexes obtained in homology-derived ensemble docking even when templates were not very homologous (Figure 5 and Supporting Information Figure S6). In fact, in most cases docking using ensembles of homology models outperform single experimental structure docking (Figure 4).

Validation of Results. Analysis on four proteins for which a large amount of structural data exist suggested (see above) that ensemble docking using homology models with sequence identity above 30−40% displayed a good ability to specifically recover active ligands when used as input for Glide calculations. Furthermore, the suggested complexes were in general reasonably close to the experimental binding modes, suggesting that the derived poses could be safely used in lead-optimization procedures. Analysis of the data suggests that the best balance between sensitivity and specificity is obtained when strict Glide scoring values were used to discriminate between active and decoy complexes. It is however unclear whether these results are general or specific for the proteins considered up to now. To analyze this point, we studied the ability of Glide on homology modeling ensembles of six unrelated proteins (see Methods).

Results summarized in Figure 6 demonstrate the good screening performance of the ensemble-docking approach performed with homology models also in the completely unrelated set of proteins used for validation. It is difficult to extend results of this small set of proteins to the entire proteome, but results suggest that docking performed using ensembles of homology models created using templates with sequence identity in the range of 30−40% leads to results which are of similar quality (according to most metrics) than those obtained using a single experimental structure. The screening performance of docking using ensembles of high-quality homology models is in general superior to that of docking using a single experimental structure and similar to docking procedures using an ensemble of experimental structures. Finally, Figure 7 confirms the geometrical quality of the complexes resulting from the homology model based docking procedure and accordingly its potential use in lead optimization processes. Our results indicate that the use of ensembles of homology models—built with Modeller—as input for Glide—using strict scoring thresholds—improves both the retrieval of active ligands from a chemical library and also the recovery of good structural complexes for lead optimization processes.

Gain in the Coverage of the Dockable Proteome. Results above suggest that an identity range of 30−40% is enough to build ensembles of homology models which can significantly enrich chemical libraries in active ligands. These results allow us to expand the applicability of structure-based drug design to a large universe of targets. Thus, while only 19% of (20332−Swissprot-annotated) human proteins can

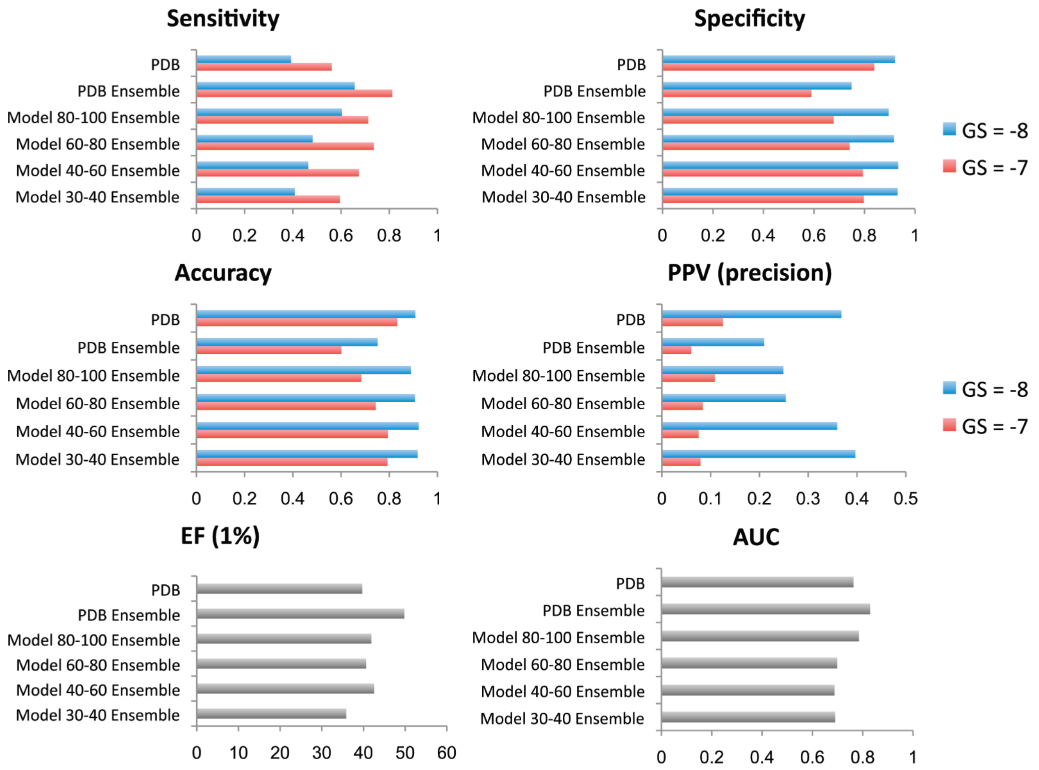

Figure 6. Enrichment descriptors for the test set. Only ensemble results for sequence identities >30% are shown for simplification. In the four top plots (sensitivity, specificity, accuracy, and PPV), enrichment descriptors are computed for $-GS = 8$ (blue) and $-GS = 7$ (red).

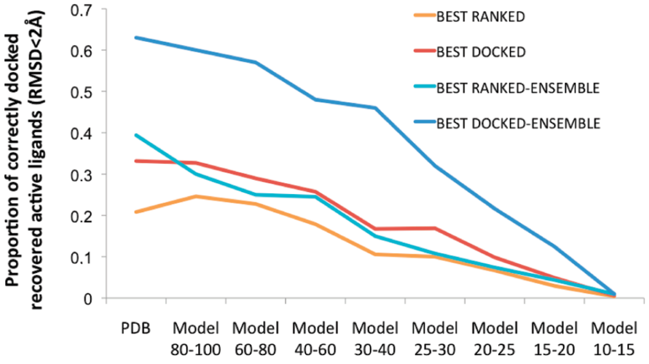

Figure 7. Recovery of correctly docked active ligands of the test set. A ligand is considered as correctly docked when its rmsd with the crystallographic ligand is below 2 Å. Both score-based selection—i.e., best ranked—and rmsd-based selection—i.e, best docked—are shown. Single docking averages are shown in red and orange, whereas ensemble docking averages are shown in blue and cyan.

be subjected to docking experiments using experimental structures, around 55% of (Swissprot) known human proteins can be studied by ensemble docking using homology models built from templates with 40% identity (Supporting Information Figure S9). Furthermore, less than 50% of human proteins of pharmacological interest have crystal structure available (DrugBank[78]). This coverage increases 41%—i.e., covering over 75% of the human drug targets—when using homology models up to 30% identity (see Supporting Information Figure S10).

With all the required cautions needed in the use of homology models for docking purposes (related mostly to the problems in finding good templates and in determining “a priori” the quality of the model), we suggest that the use of comparative models can enlarge dramatically the universe of applicability of small-molecule docking approaches, opening the possibility to analyze all potential cross-interactions of drug candidates, warning on potential adverse effects, opening new horizons both in the development of “dirty” drugs and in the determination of new indications for already annotated drugs.

ABBREVIATION. A, actives; Acc, accuracy; AUC, area under ROC curve; CDK2, cyclin-dependent kinase 2; EF, enrichment factor; ENS, ensemble; FN, false negatives; FP, false positives; FPR, false positive rate; GS, glide score; HIV, human immunodeficiency virus; HSP90, heat shock protein 90; HTD, high throughput docking; IFD, induced-fit docking; MACCS, molecular access system; PPV, positive predictive value; PTP-1B, protein tyrosine phosphatase 1B; rmsd, root-mean-square deviation; ROC, receiver operating characteristic; seq id, sequence identity; SVL, scientific vector language; TN, true negatives; TNR, true negative rate; TP, true positives; TPR, true positive rate.

Acknowledgment. This work has been supported by the Spanish Ministry of Science (Grants BIO2009-10964 and SAF2009-08811), the Instituto Nacional de Bioinformática, the Consolider E-science project, the ISCIII- COMBIOMED project, and the Fundación Marcelino Botin.

Supporting Information Available: Figures S1−S10 showing global rmsd between the homology models and the reference pdb, the correlation between the percentages of sequence identity and their sequence coverage, correlation between the binding site sequence conservation and the percentage of sequence identity of the model, rmsd of the binding site, ROC curve plots for thrombin pdbs, similarity between ligand and protein contact maps, ensemble versus single docking approach, coverage of the human proteome, and structural coverage of human targets of pharmaceutical interest. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Schneider, G.; Bohm, H. J. *Drug Discovery Today* **2002**, *7*, 64–70.

(2) Alvarez, J. C. *Curr. Opin. Chem. Biol.* **2004**, *8*, 365–370.

(3) Lyne, P. D. *Drug Discovery Today* **2002**, *7*, 1047–1055.

(4) Mohan, V.; Gibbs, A. C.; Cummings, M. D.; Jaeger, E. P.; DesJarlais, R. L. *Curr. Pharm. Des.* **2005**, *11*, 323–333.

(5) Cozzini, P.; Kellogg, G. E.; Spyrakis, F.; Abraham, D. J.; Costantino, G.; Emerson, A.; Fanelli, F.; Gohlke, H.; Kuhn, L. A.; Morris, G. M.; Orozco, M.; Pertinhez, T. A.; Rizzi, M.; Sotriffer, C. A. *J. Med. Chem.* **2008**, *51*, 6237–6255.

(6) Jacobson, M. P.; Sali, A. *Annual Reports in Medicinal Chemistry*; Academic Press: London, 2004; pp 259−276.

(7) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. *Proteins* **2006**, *65*, 15–26.

(8) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. *J. Med. Chem.* **2006**, *49*, 5912–5931.

(9) Abagyan, R.; Totrov, M. *Curr. Opin. Chem. Biol.* **2001**, *5*, 375–382.

(10) Cavasotto, C. N.; Orry, A. J. *Curr. Top. Med. Chem.* **2007**, *7*, 1006–1014.

(11) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. *Nat. Rev. Drug Discovery* **2004**, *3*, 935–949.

(12) Shoichet, B. K. *Nature* **2004**, *432*, 862–865.

(13) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. *J. Med. Chem.* **2006**, *49*, 5851–5815.

(14) Jorgensen, W. L. *Science* **2004**, *303*, 1813–1818.

(15) Brooijmans, N.; Kuntz, I. D. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335–373.

(16) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. *Proteins* **2002**, *47*, 409–443.

(17) Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 151–166.

(18) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(19) O'Donovan, C.; Apweiler, R.; Bairoch, A. *Trends Biotechnol.* **2001**, *19*, 178–181.

(20) Bairoch, A.; Apweiler, R. *Nucleic Acids Res.* **2000**, *28*, 45–48.

(21) Pruitt, K. D.; Tatusova, T.; Maglott, D. R. *Nucleic Acids Res.* **2005**, *33*, D501–D504.

(22) Clark, F.; Thanaraj, T. A. *Hum. Mol. Genet.* **2002**, *11*, 451–464.

(23) Zhang, Y. *Curr. Opin. Struct. Biol.* **2008**, *18*, 342–348.

(24) Marti-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sanchez, R.; Melo, F.; Sali, A. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 291–325.

(25) Koehl, P.; Levitt, M. *Nat. Struct. Biol.* **1999**, *6*, 108–111.

(26) Marti-Renom, M. A.; Madhusudhan, M. S.; Fiser, A.; Rost, B.; Sali, A. *Structure* **2002**, *10*, 435–440.

(27) Eswar, N.; Sali, A. Comparative Modeling of Drug Target Proteins. In *Computer-Assisted Drug Design, Comprehensive Medicinal Chemistry II*; Taylor, J., Triggle, D., Mason, J. S., Eds.;Elsevier: Oxford, U.K., 2007; Vol. 4, pp 215−236.

(28) Sanchez, R.; Pieper, U.; Melo, F.; Eswar, N.; Marti-Renom, M. A.; Madhusudhan, M. S.; Mirkovic, N.; Sali, A. *Nat. Struct. Biol.* **2000**, *7*, 986–990.

(29) Eramian, D.; Eswar, N.; Shen, M. Y.; Sali, A. *Protein Sci.* **2008**, *17*, 1881–1893.

(30) Cavasotto, C. N.; Phatak, S. S. *Drug Discovery Today* **2009**, *14*, 676–683.

(31) Baker, D.; Sali, A. *Science* **2001**, *294*, 93–96.

(32) McGovern, S. L.; Shoichet, B. K. *J. Med. Chem.* **2003**, *46*, 2895–2907.

(33) Diller, D. J.; Li, R. *J. Med. Chem.* **2003**, *46*, 4638–4347.

(34) Oshiro, C.; Bradley, E. K.; Eksterowicz, J.; Evensen, E.; Lamb, M. L.; Lanctot, J. K.; Putta, S.; Stanton, R.; Grootenhuis, P. D. *J. Med. Chem.* **2004**, *47*, 764–767.

(35) Kairys, V.; Fernandes, M. X.; Gilson, M. K. *J. Chem. Inf. Model.* **2006**, *46*, 365–379.

(36) Ferrara, P.; Jacoby, E. *J. Mol. Model.* **2007**, *13*, 897–905.

(37) Fan, H.; Irwin, J. J.; Webb, B. M.; Klebe, G.; Shoichet, B. K.; Sali, A. *J. Chem. Inf. Model.* **2009**, *49*, 2512–2527.

(38) Huang, N.; Shoichet, B. K.; Irwin, J. J. *J. Med. Chem.* **2006**, *49*, 6789–6801.

(39) Shoichet, B. K.; Bodian, D. L.; Kuntz, I. D. *J. Comput. Chem.* **1992**, *13*, 380–397.

(40) Totrov, M.; Abagyan, R. *Curr. Opin. Struct. Biol.* **2008**, *18*, 178–184.

(41) B-Rao, C.; Subramanian, J.; Sharma, S. D. *Drug Discovery Today* **2009**, *14*, 394–400.

(42) Paulsen, L. P.; Anderson, A. C. *J. Chem. Inf. Model.* **2009**, *49*, 2813–2819.

(43) Armen, R. S.; Chen, J.; Brooks, C. L. *J. Chem. Theory Comput.* **2009**, *5*, 2909–2923.

(44) Rao, S.; Sanschagrin, P. C.; Greenwood, J. R.; Repasky, M. P.; Sherman, W.; Farid, R. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 621–627.

(45) Huang, S. Y.; Zou, X. *Proteins* **2007**, *66*, 399–421.

(46) Rueda, M.; Bottegoni, G.; Abagyan, R. *J. Chem. Inf. Model.* **2009**, *50*, 186–193.

(47) Craig, I. R.; Essex, J. W.; Spiegel, K. *J. Chem. Inf. Model.* **2010**, *50*, 511–524.

(48) Damm, K. L.; Carlson, H. A. *J. Am. Chem. Soc.* **2007**, *129*, 8225–8235.

(49) Huang, S. Y.; Zou, X. *Protein Sci.* **2007**, *16*, 43–51.

(50) Sali, A.; Blundell, T. L. *J. Mol. Biol.* **1993**, *234*, 779–815.

(51) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. *J. Med. Chem.* **2004**, *47*, 1739–1749.

(52) Finn, R. D.; Mistry, J.; Schuster-Bockler, B.; Griffiths-Jones, S.; Hollich, V.; Lassmann, T.; Moxon, S.; Marshall, M.; Khanna, A.; Durbin, R.; Eddy, S. R.; Sonnhammer, E. L.; Bateman, A. *Nucleic Acids Res.* **2006**, *34*, D247–D251.

(53) Eddy, S. R. *Bioinformatics* **1998**, *14*, 755–763.

(54) Thompson, J. D.; Higgins, D. G.; Gibson, T. J. *Nucleic Acids Res.* **1994**, *22*, 4673−4680.

(55) Russell, R. B.; Barton, G. J. *Proteins* **1992**, *14*, 309–323.

(56) *Molecular Operating Environment (MOE)*, Version 2007 09; Chemical Computing Group: Montreal, Quebec, Canada, 2007.

(57) *MACCS Structural Keys*; Symyx Software: San Ramon, CA, 2002.

(58) Reynolds, C. H.; Druker, R.; Pfahler, L. B. *J. Chem. Comput. Sci.* **1998**, *38*, 305–312.

(59) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug. Delivery Rev.* **2001**, *46*, 3–26.

(60) Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. *J. Med. Chem.* **2002**, *45*, 2615–2623.

(61) Hann, M.; Hudson, B.; Lewell, X.; Lifely, R.; Miller, L.; Ramsden, N. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 897–902.

(62) *LigPrep*, Version 2.2; Schrödinger: New York, NY, 2008.

(63) *Maestro*, Version 8.5; Schrödinger: New York, NY, 2008.

(64) Feig, M.; Karanicolas, J.; Brooks, C. L. 3rd. *J. Mol. Graphics Modell.* **2004**, *22*, 377–395.

(65) Langer, T.; Hoffmann, R. D., *Pharmacophores and Pharmacophore Searches*; Wiley-VCH: Weinheim, Germany, 2006.

(66) Kirchmair, J.; Markt, P.; Distinto, S.; Wolber, G.; Langer, T. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 213–228.

(67) Truchon, J. L.; Bayly, C. I. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.

(68) Nicholls, A. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239–255.

(69) Jain, A. N.; Nicholls, A. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 133–139.

(70) Witten, I. H.; Frank, E. Credibility: Evaluating what's been learned. In *Data mining—Practical machine learning tools and techniques*, 2nd ed.; Morgan Kaufmann: San Francisco, CA, 2005; pp 161−176.

(71) Knegtel, R. M.; Kuntz, I. D.; Oshiro, C. M. *J. Mol. Biol.* **1997**, *266*, 424–440.

(72) Yoon, S.; Welsh, W. J. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 88–96.

(73) Cavasotto, C. N.; Abagyan, R. A. *J. Mol. Biol.* **2004**, *12*, 209–225.

(74) Duca, J. S.; Madison, V. S.; Voigt, J. H. *J. Chem. Inf. Model.* **2008**, *48*, 659–668.

(75) Sperandio, O.; Mouawad, L.; Pinto, E.; Villoutreix, B. O.; Perahia, D.; Miteva, M. A. *Eur. Biophys. J.*, in press.

(76) Barril, X.; Morley, S. D. *J. Med. Chem.* **2005**, *48*, 4432–4443.

(77) Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. *J. Med. Chem.* **2006**, *49*, 534–553.

(78) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. *Nucleic Acids Res.* **2008**, *36*, D901–D906.

CT100246Y