Chapter 8

Virtual Ligand Screening Against Comparative Protein Structure Models

Hao Fan, John J. Irwin, and Andrej Sali

Abstract

Virtual ligand screening uses computation to discover new ligands of a protein by screening one or more of its structural models against a database of potential ligands. Comparative protein structure modeling extends the applicability of virtual screening beyond the atomic structures determined by X-ray crystallography or NMR spectroscopy. Here, we describe an integrated modeling and docking protocol, combining comparative modeling by MODELLER and virtual ligand screening by DOCK.

Key words: Comparative modeling, Virtual screening, Ligand docking

1. Introduction

Structure-based methods have been widely used in the design and discovery of protein ligands (1–4). Given the structure of a binding site on a receptor protein, its ligands can be predicted among a large library of small molecules by virtual screening (1, 5–11): Each library molecule is docked into the binding site, then scored and ranked by a scoring function. High-ranking molecules can be selected for testing in the laboratory. Virtual screening methods can significantly reduce the number of compounds to be tested, thus increasing the efficiency of ligand discovery (12–16).

Many protein structures are relatively flexible, and can adopt different conformations when binding to different ligands. Docking a ligand to a protein structure with current methods is most likely to be successful when the shape of the binding site resembles that found in the protein-ligand complex. Therefore, the protein structure for docking is best determined in complex with a ligand that is similar to the ligand being docked, by X-ray crystallography or NMR spectroscopy. Induced fit and differences between protein conformations bound to different ligands limit the utility of

Riccardo Baron (ed.), *Computational Drug Discovery and Design*, Methods in Molecular Biology, vol. 819,
DOI 10.1007/978-1-61779-465-0_8, © Springer Science+Business Media, LLC 2012

the unbound (apo) structure and even complex (holo) structures obtained for dissimilar ligands. The problem of the protein conformational heterogeneity is especially difficult to surmount in virtual screening, which involves docking of many different ligands, each one of which may in principle bind to a different protein conformation (17).

An even greater challenge is that many interesting receptors have no experimentally determined structures at all, especially in the early phases of ligand discovery. During the last 7 years, the number of experimentally determined protein structures deposited in the Protein Data Bank (PDB) increased from 23,096 to 67,421 (November 2010) (18). In contrast, over the same period, the number of sequences in the Universal Protein Resource (UniProt) increased from 1.2 million to 12.8 million (19). This rapidly growing gap between the sequence and structure databases can be bridged by protein structure prediction (20), including comparative modeling, threading, and de novo methods. Comparative protein structure modeling constructs a three-dimensional model of a given target protein sequence based on its similarity to one or more known structures (templates). Despite progress in *de novo* prediction (21, 22), comparative modeling remains the most reliable method that can sometimes predict the structure of a protein with accuracy comparable to a low-resolution, experimentally determined structure (23).

Comparative modeling benefits from structural genomics (24). In particular, the Protein Structure Initiative (PSI) aims to determine representative atomic structures of most major protein families by X-ray crystallography or NMR spectroscopy, so that most of the remaining protein sequences can be characterized by comparative modeling (http://www.nigms.nih.gov/Initiatives/PSI/) (25, 26). Currently, the fraction of sequences in a genome for whose domains comparative models can be obtained varies from approximately 20% to 75%, increasing the number of structurally characterized protein sequences by two orders of magnitude relative to the entries in the PDB (27). Therefore, comparative models in principle greatly extend the applicability of virtual screening, compared to using only the experimentally determined structures (28).

Comparative models have in fact been used in virtual screening to detect novel ligands for many protein targets (28), including G-protein coupled receptors (GPCR) (29–41), protein kinases (42–45), nuclear hormone receptors, and a number of different enzymes (14, 15, 46–57). The relative utility of comparative models versus experimentally determined structures has been assessed (17, 29, 42, 43, 58–60). Although the X-ray structure of a ligand-bound target often provides the highest enrichment for known ligands, comparative models yield better enrichment than random selection and sometimes performs comparably to the holo X-ray structure. Recently, we assessed our automated

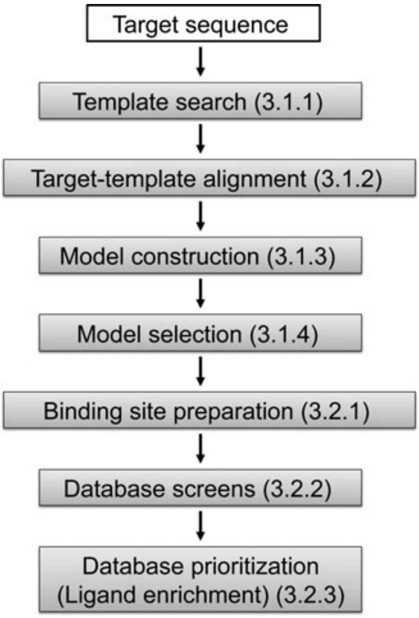

Target sequence
Template search (3.1.1)
Target-template alignment (3.1.2)
Model construction (3.1.3)
Model selection (3.1.4)
Binding site preparation (3.2.1)
Database screens (3.2.2)
Database prioritization
(Ligand enrichment) (3.2.3)

Fig. 1. The automated modeling and docking pipeline. Numbers in parentheses indicate the corresponding section in the text.

modeling and docking pipeline (17) based on MODELLER (61) for comparative modeling and DOCK (62, 63) for virtual screening. We demonstrated that when multiple target models are calculated, each one based on a different template, the "consensus" enrichment for multiple models is better or comparable to the enrichment for the apo and holo X-ray structures in 70% and 47% cases, respectively; the consensus enrichment is calculated by combining the docking results of multiple structures — for each docked compound, the best docking score across all structures was used for ranking the compound — thus, the ranking relied on optimizing the protein conformation as well as protein-ligand complementarity. Another similar criterion for ligand ranking was also described (64).

The modeling and docking protocol is carried out in seven sequential steps (Fig. 1). Steps 1–4 correspond to comparative modeling: (1) template search finds known structures (templates) related to the sequence to be modeled (target), (2) target-template alignment aligns the target sequence with the templates, (3) model construction computes multiple target models based on the input alignment, (4) model selection identifies the best-scoring model. Steps 5–7 correspond to virtual screening: (5) binding site preparation involves creating input files for generating spheres and scoring grids used in docking, (6) database screening docks database molecules into the binding site, and (7) database prioritization scores and ranks the docking poses of the database molecules. Comparative modeling is carried out by program MODELLER that implements comparative modeling by satisfaction of spatial restraints derived from the target-template alignment, atomic statistical potentials, and the CHARMM molecular mechanics force field (61). The spatial restraints are combined into an objective function that is optimized by a combination of conjugate gradients and molecular dynamics with simulated annealing; this model-building procedure is formally similar to structure determination by NMR spectroscopy. Virtual screening is performed by the DOCK suite of programs (63, 65, 66). DOCK uses a negative image of the receptor – spheres that fill the receptor site – to describe the space into which docked molecules should fit. Docking poses are generated by matching the atoms of a small molecule with the centers of the spheres. The generated poses are evaluated using a grid-based approach in which interactions between the docked molecules and the receptor are precomputed at each grid point.

2. Materials

2.1. Software for Comparative Modeling

1. The MODELLER 9v8 program can be downloaded from http://salilab.org/modeller/.
2. A typical operation in MODELLER consists of (1) preparing an input Python script, (2) ensuring that all required files (e.g., files specifying sequences, structures, alignments) exist, (3) executing the input script by typing' mod9v8 input-script-name', and (4) analyzing the output and log files. A tutorial for the use of MODELLER 9v4 or newer is available at http://salilab.org/modeller/tutorial/.

2.2. Database for Comparative Modeling

1. Sequence database (UniProt90) contains all sequences from UniProt (clustered at 90% to remove redundancy), and can be downloaded from http://salilab.org/modeller/supplemental.html.
2. Template sequence database (pdball) contains the sequence for each protein structure in PDB, and can be downloaded from http://salilab.org/modeller/supplemental.html.

2.3. Software for Virtual Screening

1. DOCK 3.5.54 (62, 63) is available under the UCSF DOCK license http://dock.compbio.ucsf.edu/Online_Licensing/dock_license_application.html (see Note 1). Documentation for DOCK 3.5 is provided at http://wiki.bkslab.org/index.php/Image:Dock3_5refman.pdf.

2. Third party applications. DMS is a program that calculates the solvent-accessible molecular surface of the protein binding site (67), and can be downloaded at http://www.cgl.ucsf.edu/Overview/ftp/dms.shar. SYBYL is a commercial molecular modeling program that can build and manipulate molecules (68). In our study, SYBYL is used to add hydrogen atoms to polar atoms in a protein receptor (in the PDB format) that contains only non-hydrogen atoms; it can be downloaded from http://tripos.com/index.php?family=modules,General.DownloadPortal,Home. Delphi is a program that computes numerical solutions of the Poisson-Boltzmann equation for molecules of arbitrary shape and charge distribution (69); a request for access to this program can be made at http://luna.bioc.columbia.edu/honiglab/software/cgi-bin/software.pl?input=DelPhi.

2.4. Docking Database of Small Molecules

1. The Directory of Useful Decoys (DUD) is a docking database designed to help test docking algorithms by providing challenging decoys (70). DUD contains a total of 2,950 compounds that bind to a total of 40 targets; in addition, for each ligand, it also contains 36 "decoys" with similar physical properties (e.g., molecular weight, calculated LogP) but dissimilar chemical topology. DUD can be downloaded from http://dud.docking.org/r2/.

3. Method

The automated modeling and docking pipeline will be illustrated with one example taken from our benchmark study (17), adenosine deaminase (ADA, EC 3.5.4.4). ADA is a metalloenzyme in whose binding pocket one catalytic zinc ion is coordinated by three histidine residues and one aspartic acid residue (71, 72). The bovine ADA has been co-crystalized with a non-nucleoside inhibitor (PDB code 1NDW). The DUD database was screened against comparative models and the ligand-bound (holo) crystal structure of the bovine ADA, to compare the utility of comparative models and holo crystal structures for virtual screening.

3.1. Comparative Modeling of Protein Structures

3.1.1. Template Search

First, a file with the bovine ADA sequence in the MODELLER "PIR" format is prepared (Fig. 2; see Note 2). Then the ADA sequence is scanned against all sequences in the PDB (stored in file "pdball") to identify suitable templates, with the MODELLER "profile.build" routine (Fig. 3; see Note 3). In this example, one holo structure (PDB code 1UIO) (73) with 85% sequence identity to the target and one apo structure (PDB code 2AMX) (74)

```
>P1;ADA
sequence:ADA:::::::-1.00:-1.00
TPAFDKPKVELHVHLDGAIKPETILYYGKRRGIALPADTPEELQNIIGMDKPLTLPDFLAKFDYYMPAIAGCRDA
IKRIAYEFVEMKAKDGVVYVEVRYSPHLLANSKVEPIPWNQAEGDLTPDEVVSLVNQGLQEGERDFGVKVRSILC
CMRHQPSWSSEVVELCKKYREQTVVAIDLAGDETIEGSSLFPGHVQAYAEAVKSGVHRTVHAGEVGSANVVKEAV
DTLKTERLGHGYHTLEDTTLYNRLRQENMHFEICPWSSYLTGAWKPDTEHAVIRFKNDQVNYSLNTDDPLIFKST
LDTDYQMTKKDMGFTEEEFKRLNINAAKSSFLPEDEKKELLDLLYKAYR/.*
```

Fig. 2. File “ADA.ali” in the “PIR” format. This file specifies the target sequence. See the MODELLER manual for the detailed description of the format.

```
from modeller import *
log.verbose()
env = environ()

#-- Read in the template sequence database
sdb = sequence_db(env)
sdb.read(seq_database_file='pdball.pir', seq_database_format='PIR',
         chains_list='ALL')

#-- Write the sequence database in binary form
sdb.write(seq_database_file='pdball.bin', seq_database_format='BINARY',
          chains_list='ALL')

#-- Now, read in the binary database
sdb.read(seq_database_file='pdball.bin', seq_database_format='BINARY',
         chains_list='ALL')

#-- Read in the target sequence/alignment
aln = alignment(env)
aln.append(file='ada.ali', alignment_format='PIR', align_codes='ADA')

#-- Convert the input sequence/alignment into profile format
prf = aln.to_profile()

#-- Scan sequence database to pick up homologous sequences
prf.build(sdb, matrix_offset=-450, rr_file='${LIB}/blosum62.sim.mat',
         gap_penalties_1d=(-500, -50), n_prof_iterations=5,
         check_profile=False, max_aln_evalue=0.01, gaps_in_target=False)

#-- Write out the profile
prf.write(file='search_templates.prf', profile_format='TEXT')

#-- Convert the profile to alignment
aln = prf.to_alignment()

#-- Write out the alignment
aln.write(file='search_templates.ali', alignment_format='PIR')
```

Fig. 3. File “search_templates.py.” This script searches for potential template structures in a database of nonredundant PDB sequences.

with 27% sequence identity are selected as templates (see Note 4), to be used independently for calculating two models of ADA.

3.1.2. Target-Template Alignment

For each target-template pair (i.e., ADA-1UIO and ADA-2AMX), the target and template sequences are scanned against all sequences in UniProt90 independently with the “profile.build” routine,

```
>P1;ADA
sequence:ADA:1::+350::::-1.00:-1.00
TPAFDKPKVELHVHLDGAIKPETILYYGKRRGIALPADTPEELQNIIGMDKPLTLPDFLAK----FDYYMPAIAG
CRDAIKRIAYEFVEMKAKDGVVYVEVRYSPHLLANSKVEPIPWNQAEGDLTPDEVVSLVNQGLQEGERDFGVKVR
SILCCMR---HQPSWSSEVVELCKKYREQTVVAIDLAGDETIEGSSLFPGHVQAYAEAVKSGVHRTVHAGE---V
GSANVVKEAVDTLKTERLGHGYHTLEDTTLYNRLRQENMHFEICPWSSYLTGAWKPDTEHAVIRFKNDQVNYSLN
TDDPLIFKSTLDTDYQMTKKDMGFTEEEFKRLNINAAKSSFLPEDEKKELLDLLYKAYR/.*

>P1;2AMX
structure:2AMX:38::365::::-1.00:-1.00
------PKVELHCHLDLTFSAEFFLKWARKYNLQPNMSDDEILDHYLFTKEGKSLAEFIRKAISVSDLYRD----
-YDFIEDLAKWAVIEKYKEGVVLMEFRYSPTFVSSSY-----------GLDVELIHKAFIKGIKNATELLNNKIH
VALICISDTGHAAASIKHSGDFAIKHKHD-FVGFDHGGRE-ID----LKDHKDVYHSVRDHGLHLTVHAGEDATL
PNLNTLYTAINILNVERIGHGIRVSESDELIELVKKKDILLEVCPISNLLLNNVKSMDTHPIRKLYDAGVKVSVN
SDDPGMFLSNINDNYEKLYIHLNFTLEEFMIMNNWAFEKSFVSDDVKSELKALYF----/.*
```

Fig. 4. File “align.ali” in the “PIR” format. The file specifies the alignment between the sequences of ADA and 2AMX (A chain).

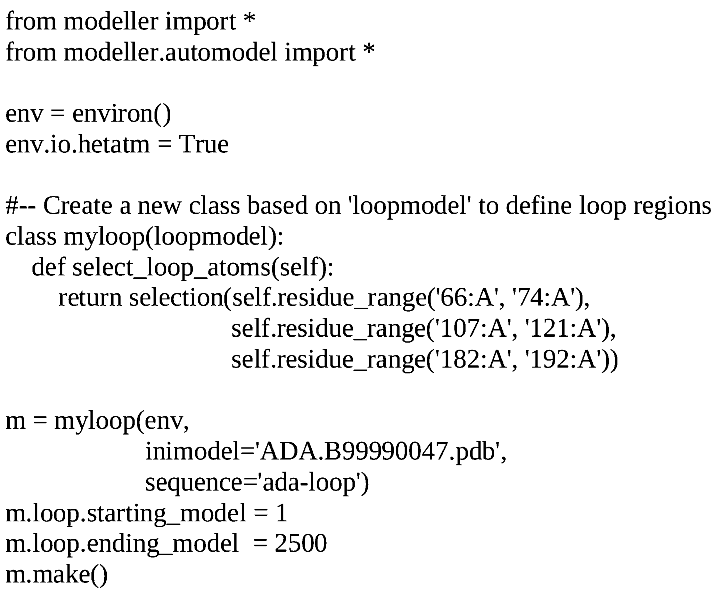
```
from modeller import *
from modeller.automodel import *

env = environ()
env.io.hetatm = True

a = automodel(env, alnfile='align.ali',
              knowns='2AMX', sequence='ADA')

a.starting_model = 1
a.ending_model = 500
a.make()
```

Fig. 5. File “build_model.py.” The script generates 500 models of ADA based on 2AMX with “automodel” routine.

resulting in the target profile and the template profile, respectively. Next, the target profile is aligned against the template profile with the “profile.scan” routine (a sample script is given at http://salilab.org/modeller/examples/commands/ppscan.py). The resulting alignment is presented in Fig. 4, for the 2AMX template (see Note 5; the ADA-1UIO alignment is not shown).

3.1.3. Model Construction

Once the target-template alignment is generated, MODELLER calculates 500 models of the target completely automatically, using its “automodel” routine (Fig. 5; see Note 6). The best model (defined in Subheading 3.1.4) is then subjected to a refinement of binding site loops (see Note 7) with the “loopmodel” routine (Fig. 6). All three binding site loops were optimized simultaneously, resulting in 2,500 conformations of ADA (see Note 8).



```
from modeller import *
from modeller.automodel import *

env = environ()
env.io.hetatm = True

#-- Create a new class based on 'loopmodel' to define loop regions
class myloop(loopmodel):
    def select_loop_atoms(self):
        return selection(self.residue_range('66:A', '74:A'),
                         self.residue_range('107:A', '121:A'),
                         self.residue_range('182:A', '192:A'))

m = myloop(env,
           inimodel='ADA.B99990047.pdb',
           sequence='ada-loop')
m.loop.starting_model = 1
m.loop.ending_model  = 2500
m.make()
```

Fig. 6. File "loop_model.py." Input script file that generates 2,500 models with the "loopmodel" routine.

3.1.4. Model Selection

When multiple models are calculated for the target based on a single template (by "automodel," and "loopmodel," if there are binding site loops), it is practical to select the model or a subset of models that are judged to be most suitable for subsequent docking calculations (see Note 9). In this example, for each template, we select the model with optimized loops that has the lowest value of the MODELLER objective function (ada-loop.BL16340001.pdb for 2AMX), which is reported in the second line of the model file (see Note 10). The most suitable model can also be selected by the Discrete Optimized Protein Energy (DOPE) (75), which is calculated using the "assess_dope" routine (see Note 11).

3.2. Virtual Screening Against Comparative Models

As described in the previous section, a single comparative model of bovine ADA is selected from models calculated based on the 2AMX template. Another model is selected from models based on the 1UIO template. The DUD database is then screened against each of the two models independently. We will only describe the docking to the ADA model based on 2AMX.

3.2.1. Binding Site Preparation

Prepare input files for the automated docking pipeline. The file containing the ADA model based on 2AMX is renamed to "rec.pdb," followed by (1) removing all lines that do not contain coordinates of non-hydrogen atoms; (2) replacing "HETATM" in the line containing the coordinates of the zinc ion by "ATOM"; and (3) removing all chain identifiers (see Note 12). Next, the file "xtal-lig.pdb" is created, containing the binding site specification in the same format as that of "rec.pdb". In this example, the ligand observed in the holo crystal structure of the target is given in "xtal-lig.pdb"; this ligand is transferred into the model by superposing the crystal structure on the model using the binding site residues (see Note 13).

Automated spheres and scoring grids generation. First, the environment variable "DOCK_BASE" is defined to be the "dockenv" directory of the DOCK 3.5.54 installation. Second, file "Makefile" from "dockenv/scripts/" is copied to the current working directory, which also contains the "rec.pdb" and "xtal-lig.pdb" files. Third, file ".useligsph" is generated. Finally, command "make" is executed to generate the spheres and scoring grids (see Note 14).

3.2.2. Database Screening

The DUD database contains 2950 annotated ligands and 95,316 decoys for 40 diverse targets (70); the DUD database is stored in 801 DOCK 3.5 hierarchy database files (DUD 2006 version) (63). Eight hundred and one sub-directories corresponding to the 801 hierarchy database files are created. In each sub-directory, two files are required for docking. One is file "INDOCK" that contains the input parameters for DOCK 3.5.54 (Fig. 7) (see Note 15). Another file, "split_database_index," contains the location and name of the corresponding database file. In file "INDOCK," "split_database_index" is given as the value for the parameter with the keyword "ligand_atom_file." Docking is performed by running the DOCK executable "dockenv/bin/Linux/dock" in each sub-directory. Two output files are produced: (1) the compressed file "test.eel1.gz" contains the docking poses of database molecules in the extended PDB format and (2) the compressed file "OUTDOCK.gz" contains the docking scores for the database molecules as well as the input file names and parameter values.

3.2.3. Database Prioritization

First, the conformations of database molecules are filtered for steric complementarity using the DOCK contact score. The conformations

```
#                         INPUT
#
mode                      search
receptor_sphere_file      ../sph/match2.sph
ligand_atom_file          split_database_index
#
#                         MATCHING
#
distance_tolerance        1.5
ligand_binsize            0.4
ligand_overlap            0.3
receptor_binsize          0.4
receptor_overlap          0.3
```

Fig. 7. A section of file "INDOCK" containing some input parameters for DOCK 3.5.54.

```
mol#    id_num    matched   nscored  nhvy  nconfs  part.fcn     Time
   E    id_num    shape      elect    + vdW + polsol + apolsol  = Total
106  C03814312    773      145642    18     4536      662.0     0.24
   E C03814312    121      -16.44    -18.51     7.06       4.77   -23.12
107  C03814313    825       98854    19      405      136.4     0.21
   E C03814313    134      -50.06    -17.63    15.69       5.54   -46.46
108  C03814313    825      101057    19      405      144.9     0.21
   E C03814313    122      -48.88    -16.30    15.19       5.19   -44.80
```

Fig. 8. A section of file "OUTDOCK.gz" containing docking scores of two DUD molecules.

that do not clash with the receptor are then scored by the DOCK energy function (the DOCK contact score is not included):

$$E_{\text{score}} = E_{\text{vdW}} + E_{\text{elec}} + \Delta G^{\text{lig}}_{\text{desolv}}, \tag{1}$$

where E_{vdW} is the van der Waals component of the receptor-ligand interaction energy based on the AMBER united-atom force field, E_{elec} is the electrostatic potential calculated by DelPhi, and $\Delta G^{\text{lig}}_{\text{desolv}}$ is the ligand desolvation penalty computed by solvmap, as described in Subheading 3.2.2. For each ligand conformation, the total energy and all the individual energy terms are written out to file "OUTDOCK.gz" (Fig. 8; see Note 16). The single conformation with the best total energy is saved in file "test.eel1.gz" as the docking pose of the database molecule. The docking pose of one ADA ligand—1-deazaadenosine (PubChem ID: 159738, ZINC ID: C03814313)—is shown in Fig. 11b. After the virtual screening, the best total energy of each database molecule and the corresponding molecule ID are extracted from the "OUTDOCK.gz" files in all sub-directories. The molecules in the docking database are ranked by their total energies. The top 500 ranked molecules are then inspected visually. Molecules forming favorable interactions with the receptor (e.g., a docking pose is similar to the binding mode found in crystal structures of proteins in the same family) can be chosen for subsequent experimental testing.

In this benchmark example, we can quantify the accuracy of modeling and docking by computing the enrichment for the known ADA ligands among the top scoring ligands:

$$\text{EF}_{\text{subset}} = \frac{(\text{ligand}_{\text{selected}}/N_{\text{subset}})}{(\text{ligand}_{\text{total}}/N_{\text{total}})}, \tag{2}$$

where $\text{ligand}_{\text{total}}$ is the number of known ligands in a database containing N_{total} compounds and $\text{ligand}_{\text{selected}}$ is the number of ligands found in a given subset of N_{subset} compounds. $\text{EF}_{\text{subset}}$ reflects the ability of virtual screening to find true positives among the decoys in the database compared to a random selection. An enrichment curve is obtained by plotting the percentage of actual ligands found (*y*-axis) within the top ranked subset of all database compounds (*x*-axis on logarithmic scale). To measure the enrichment independently of the arbitrary value of N_{subset}, we also

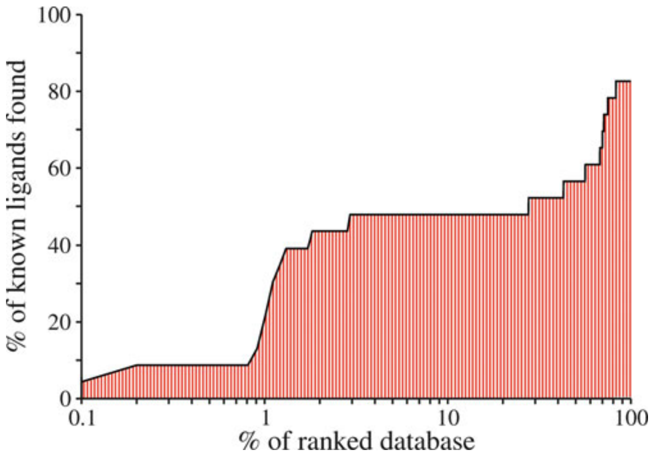


Fig. 9. The enrichment curve for virtual screening of the DUD database against the ADA model based on 2AMX. The ligand enrichment is quantified by the logAUC of 40.3.

calculated the area under the curve (logAUC) of the enrichment plot:

$$\log \mathrm{AUC} = \frac{1}{\log_{10} 100/\lambda} \times \sum_{\lambda}^{100} \left\{ \frac{\mathrm{ligand_{subset}}}{\mathrm{ligand_{total}}} \bullet \left(\lambda \bullet \log_{10} \frac{N_{\mathrm{subset}}}{N_{\mathrm{total}}} \right) \right\}, \quad (3)$$

where λ is arbitrarily set to 0.1. A random selection ($\mathrm{ligand_{selected}}/\mathrm{ligand_{total}} = N_{\mathrm{subset}}/N_{\mathrm{total}}$) of compounds from the mixture of true positives and decoys yields a logAUC of 14.5. A mediocre selection that picks twice as many ligands at any N_{subset} as a random selection has logAUC of 24.5 ($\mathrm{ligand_{selected}}/\mathrm{ligand_{total}} = 2 \times N_{\mathrm{subset}}/N_{\mathrm{total}}$; $N_{\mathrm{subset}}/N_{\mathrm{total}} \leq 0.5$). A highly accurate enrichment that produces 10 times as many ligands than the random selection has logAUC of 47.7 ($\mathrm{ligand_{selected}}/\mathrm{ligand_{total}} = 10 \times N_{\mathrm{subset}}/N_{\mathrm{total}}$; $N_{\mathrm{subset}}/N_{\mathrm{total}} \leq 0.1$). In this example, the ADA model based on 2AMX yielded the logAUC of 40.3 (Fig. 9). When multiple structures are available (either models or experimental structures), consensus enrichment can be calculated (Introduction).

4. Notes

1. The DOCK 3.5.54 source distribution contains four items: the "dock", the "dockenv" and the "test" directories, as well as the "README" file. The DOCK source code and

executable are in the "dock" directory. Scripts used in the automated docking pipeline are in the "dockenv" directory. The binary executable "dock" in "dockenv/bin/Linux/" is used in the docking calculations.

2. The target protein sometimes contains modified residues, such as carboxylated lysine (KCX) and selenomethionine (MSE). These modified residues need to be replaced by standard residues with similar physical and chemical properties (e. g., KCX by glutamic acid and MSE by methionine).
3. MODELLER script for template search.

The environ routine initializes the environment for the modeling run, by creating a new environment object, called env. Almost all MODELLER scripts require this step, because the new environment object is needed to build most other useful objects.

The sequence_db routine creates a sequence database object sdb that is used to contain large databases of protein sequences.

The sdb.read and sdb.write routines read and write a database of sequences, respectively, in the PIR, FASTA, or BINARY format.

The second call to the sdb.read routine reads the binary format file for faster execution.

The alignment(env) routine creates a new "alignment" object (aln). The aln.append routine reads the target sequence ADA from the file ada.ali, and converts it to a profile object (prf).

The prf.build routine scans the target profile (prf) against the sequence database (sdb). Matching sequences from the database are added to the profile.

4. In general, a sequence identity value above ~25% indicates a potential template, unless the alignment is too short (i.e., <100 residues). A better measure of the alignment significance is the *E*-value of the alignment (the lower *E*-value, the better; a conservative cut-off is 0.001). Besides the sequence similarity, template structures can also be chosen on the basis of other criteria, such as the accuracy of the structures (e.g., resolution of X-ray structures), conservation of active-site residues, and presence of bound ligands.
5. Different alignment methods vary in terms of the scoring function that is being optimized. When the target-template sequence identity is above 30–40%, different methods tend to produce very similar alignments. When similarity decreases, different methods tend to produce widely varying alignments. An accurate alignment is indicated when different methods,

such as MUSCLE (76), CLUSTALW (77) and T-coffee (78), produce similar alignments.

6. Model building with the "automodel" routine.

   In the input script build_model.py (Fig. 5), an automodel object is first created, specifying the alignment file ("align.ali"), the target (ADA), and the template (2AMX). The models are calculated by the "make" routine. Five hundred models for ADA are written out in the PDB format to files called ADA.B9990[0001-0500].pdb.

   Ligands, ions, and cofactors in the template structures are copied to the target models and treated as rigid bodies, using the "BLK" functionality of MODELLER.

   Models are computed by optimizing the MODELLER objective function in the Cartesian space. The optimization begins by the variable target function approach, deploying the conjugate gradients method, followed by a refinement by molecular dynamics with simulated annealing. The default optimization protocol can be adjusted (a sample script is given at http://salilab.org/modeller/examples/automodel/model-changeopt.py).

7. The binding site loops are defined as those binding site residues in the vicinity of the binding site that were not aligned to the template structure. The binding site residues may be chosen based on the prior experimental information (e.g., mutagenesis data) and/or sequence conservation within a family of homologous proteins. In this study, binding site residues are defined as the residues with more than one non-hydrogen atom within 10 Å of any ligand atom in the target structure. Thus, three insertions in the ADA-2AMX alignment are defined as binding site loops (neighboring residues within two positions of each insertions are also included) (Fig. 4).

8. *Loop optimization with the "loopmodel" routine*. In the input script "loop_model.py" (Fig. 6), the best-scoring model generated by "automodel" (ADA.B99990047.pdb) is used as the starting conformation, thus defining the loop environment. Loop regions defined by the "select_loop_atoms" routine are randomized, followed by optimization with a combination of conjugate gradients and molecular dynamics with simulated annealing. Two thousand five hundred models are written out in the PDB format to files called ada-loop.BL[0001-2500]0001.pdb. Calculating multiple loop models allows for better conformational sampling of the unaligned regions. Typically, for a single 8-residue loop, 50–500 independent optimizations are recommended (79).

9. Most proteins are flexible, often adopting different conformations when binding to different ligands. Besides the single best model, it might be helpful to select several sub-optimal models that are structurally diverse (e.g., selecting the best model from each conformational cluster of models). When no target ligand is known, the docking database can be screened against each of these representative models independently, followed by combining the screening results. However, when some target ligands are already known, the best single model could be selected based on its ability to rank these known ligands most highly in virtual screening.
10. The MODELLER objective function is a measure of how well the model satisfies the input spatial restraints. Lower values of the objective function indicate a better fit with the restraints. Models (of the same sequence) can only be ranked by the same objective function, consisting of the same restraints, usually derived from the same alignment.
11. The DOPE is an atomic distance-dependent statistical potential based on a physical reference state that accounts for the finite size and spherical shape of proteins (75). By default, the DOPE score is not included in the model building routine, and thus can be used as an independent assessment of the accuracy of the output models. DOPE considers the positions of all non-hydrogen atoms, with lower scores corresponding to models that are predicted to be more accurate. A sample script for generating a DOPE score is given at http://salilab.org/modeller/examples/assessment/assess_dope.py.
12. All lines in "rec.pdb" should start with "ATOM." If the receptor contains a cofactor that has not been defined in the DOCK force field, a dictionary of parameters needs to be provided for the cofactor. "Structural" water molecules in the receptor should be renamed as "TIP".
13. The binding site can be specified either using a modeled ligand or residues surrounding the binding pocket. In the latter case, at least three binding site residues should be defined in the file "xtal-lig.pdb"; the center of mass of these residues defines the center of the binding pocket.
14. Eleven tasks are accomplished by "make" (Fig. 10). (1) Copies of file "filt.params" (the input file for program FILT) as well as the "sph" and "grids" directories (containing input files and parameter files for sphere and scoring grids generation, respectively) are copied from directory "dockenv/scripts/". (2) Program FILT located in "dockenv/bin/Linux" is used to identify binding site residues that are within 10 Å of any atom in the file "xtal-lig.pdb". The result is stored in file "rec.site". (3) Given the receptor coordinates in "rec.pdb" and the binding site

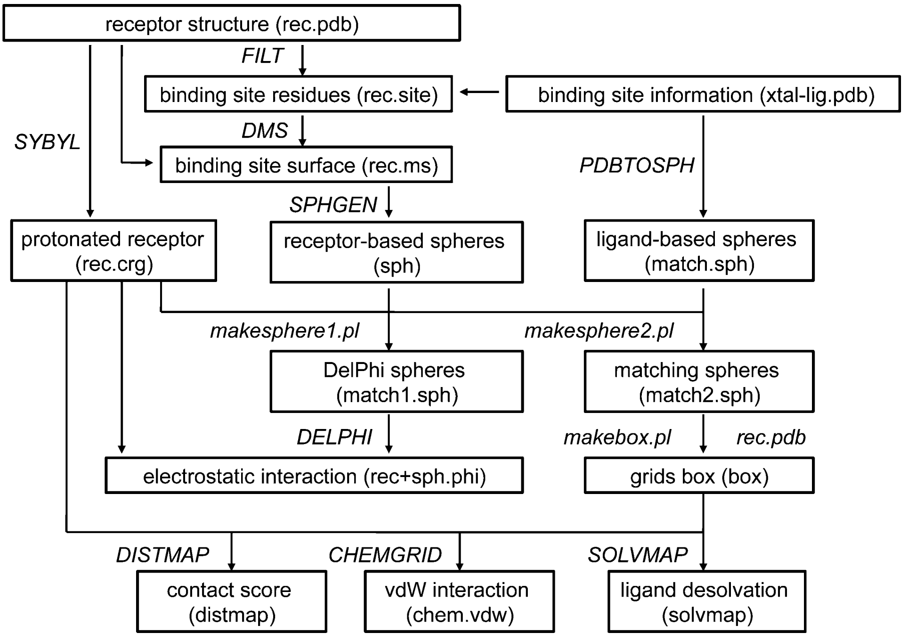

receptor structure (rec.pdb)
FILT
binding site residues (rec.site)
binding site information (xtal-lig.pdb)
SYBYL
DMS
binding site surface (rec.ms)
PDBTOSPH
SPHGEN
protonated receptor (rec.crg)
receptor-based spheres (sph)
ligand-based spheres (match.sph)
makesphere1.pl
makesphere2.pl
DelPhi spheres (match1.sph)
matching spheres (match2.sph)
DELPHI
makebox.pl
rec.pdb
electrostatic interaction (rec+sph.phi)
grids box (box)
DISTMAP
CHEMGRID
SOLVMAP
contact score (distmap)
vdW interaction (chem.vdw)
ligand desolvation (solvmap)

Fig. 10. Schematic description of the automated preparation of receptor binding site, including sphere and scoring grids generation.

definition in "rec.site", the solvent-accessible molecular surface of the receptor binding site is calculated by the program DMS. The result is written in the file "rec.ms". (4) The program SYBYL is used to add hydrogens on polar atoms to the receptor. The atomic coordinates of the protonated receptor are written to the file "grids/rec.crg". All lines that do not contain atomic coordinates are removed manually; all lines in "rec.crg" should start with "ATOM". (5) The program pdbtosph in "dockenv/bin/Linux" is used to derive spheres from atom positions in "xtal-lig.pdb". The ligand-based spheres are stored in the file "sph/match.sph". (6) Spheres in contact with the binding site surface are generated by the script "rec.ms" relying on the program sphgen (80) in "dockenv/bin/Linux". These receptor-based spheres are stored in the file "sph/sph". (7) Two perl scripts "makespheres1.pl" and "makespheres2.pl" in "dockenv/scripts" are used to generate spheres for the binding site electrostatic potential calculation with DelPhi (DelPhi spheres, named as "match1.sph") and the spheres required for orienting database molecules in the binding site (matching spheres, named "match2.sph"), respectively. For both scripts, the ligand-based spheres "match.sph", receptor-based spheres "sph", and the protonated receptor "rec.crg" need to be provided as input files. DelPhi spheres occupy a greater volume than the matching spheres (Fig. 11a). Spheres that are exposed to bulk water should be removed by hand. (8) The perl script

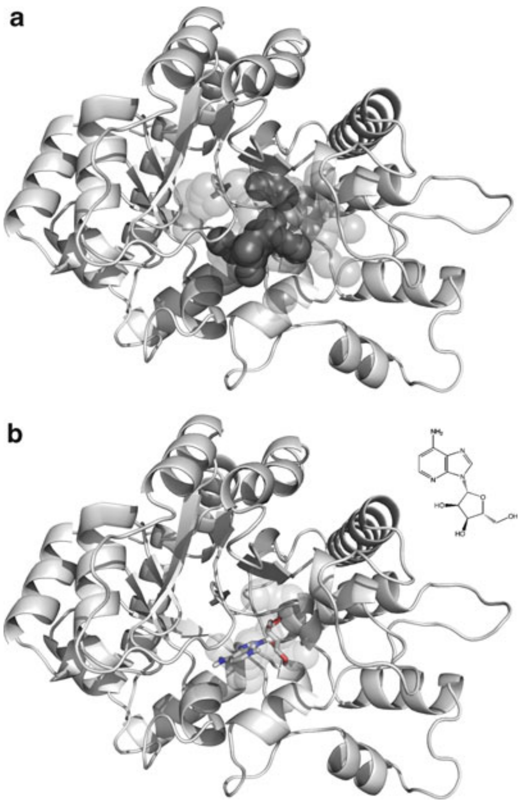


Fig. 11. (**a**) The matching spheres (*dark grey*) and DelPhi spheres (*light grey*) generated for the binding site of the ADA model (cartoon) based on 2AMX. (**b**) The docking pose (stick) and the 2D structure of one ADA ligand—1-deazaadenosine (PubChem ID: 159738, ZINC ID: C03814313)—as well as the matching spheres (*light grey*).

"makebox.pl" in "dockenv/scripts" is used to determine the location and dimensions of the region in which the scoring grids will be calculated. This region should enclose the volume that the ligands are likely to occupy (described by "match2.sph"). The resulting rectangular box is written out in the file "grids/box". (9) The contact score is a summation of the number of non-hydrogen atom contacts between a database molecule and the receptor (a contact is any intermolecular distance smaller than 4.5 Å), providing an assessment of shape complementarity. The program distmap (66) in "dockenv/bin/Linux" produces the grids for contact scoring. Three files are required for distmap, including the input file "INDIST", the protonated receptor "rec.crg", and the volume of the grids "box". The contact grid is produced in the file "grids/distmap" by running the command "distmap". (10) The DOCK's force field score is the van der Waals interaction energy. The parameters are taken from the AMBER united-atom force field (81). The program chemgrid (66) in "dockenv/bin/Linux" produces the grids for force field scoring. The force field grid is written into the file "grids/chem.vdw" by running the command "chemgrid". All receptor residues and atoms need to be defined in the parameter files "grids/prot.table.ambcrg.ambH" and "grids/vdw.parms.amb.mindock", respectively. (11) The electrostatic potential grid is generated by DelPhi (69). The receptor coordinates in "rec.crg" and the Delphi spheres in "match1.sph" are combined into the file "grids/rec+sph.crg". The DelPhi map is calculated using a relative dielectric constant of 2 for the volume defined by the receptor atoms and the spheres in the binding site, and a relative dielectric constant of 78 for the external solvent environment. The DelPhi grid is written to the file "grids/rec+sph.phi" by running the command "./delphi.com>delphi.log" in the "grids" directory. All receptor residues and atoms need to be defined in the parameter file "grids/amb.crg.oxt". (12) The solvent occlusion grid is calculated by the program solvmap, for subsequent calculation of the ligand desolvation penalty (82). Three files are required for solvmap, including the input file "INSOLV", the protonated receptor "rec.crg", and the volume of the grids "box". The solvent occlusion grid is written into the file "grids/solvmap" by running the command "solvmap". The grid file "grids/solvmap" should not contain any blank lines.

15. Several examples of file "INDOCK" are provided in the directory "dockenv/scripts/calibrate/". A detailed description of the parameters used in INDOCK can be found in the manual of DOCK 3.5. Here, we describe several parameters that are often modified to achieve an optimal docking performance (Fig. 7). The parameter "mode" should be specified as "search". In the "search" mode, DOCK generates positions and orientations for each molecule in the database (virtual screening). The parameter "receptor_sphere_file" specifies the file that contains the matching spheres for ligand orientation in the binding site. Matching spheres can be manually scaled or relocated to achieve satisfying sampling in the desired region (e.g., catalytic residues suggested by experiments). During docking, sets of atoms from database molecules match sets of matching spheres, if all the internal distances match within a tolerance value in Ångstroms specified by the parameter "distance_tolerance" (65). The choice of the tolerance value depends on the reliability of the matching sphere sizes and positions, which in turn is determined by the accuracy of the binding site conformation. We suggest a tolerance value of 1.5 Å when docking to comparative models. The sampling of the ligand positions and orientations is controlled by four parameters, including "ligand_binsize", "ligand_overlap", "receptor_binsize", and "receptor_overlap" (65). "ligand_binsize" and "receptor_binsize" define the width of the bins containing ligand

atoms and matching spheres, respectively. "ligand_overlap" and "receptor_overlap" define the overlap between the bins of ligand atoms and matching spheres, respectively. The increase of either the width of bins or the overlap between bins will result in more atoms/spheres in each bin. As a consequence, a greater number of matches will be found. Extensive sampling is achieved by setting the bin size for both ligand and receptor to 0.4 Å, and the overlap to 0.3 Å.

16. As shown in Fig. 8, for each conformation of a database molecule, two lines are written out in the file "OUTDOCK.gz". The scoring results are written in the second line starting with the letter "E", followed by the molecule identifier, contact score, electrostatic score, van der Waals score, polar solvation correction, apolar solvation correction, and total energy. The total energy is a sum of contact score, electrostatic score, van der Waals score, polar solvation correction, and apolar solvation correction.

Acknowledgement

This article is partially based on the MODELLER manual, the DOCK 3.5 manual, and the "DISI" wiki pages (http://wiki.bkslab.org). We also acknowledge funds from Sandler Family Supporting Foundation and National Institutes of Health (R01 GM54762 to AS; R01 GM71896 to BKS and JJI; P01 GM71790 and U54 GM71790 to AS and BKS). We are also grateful to Ron Conway, Mike Homer, Hewlett-Packard, IBM, NetApp, and Intel for hardware gifts.

References

1. Kuntz, I. D. (1992) Structure-Based Strategies for Drug Design and Discovery, *Science 257*, 1078–1082.
2. Klebe, G. (2000) Recent developments in structure-based drug design, *J. Mol. Med. 78*, 269–281.
3. Dailey, M. M., Hait, C., Holt, P. A., Maguire, J. M., Meier, J. B., Miller, M. C., Petraccone, L., and Trent, J. O. (2009) Structure-based drug design: From nucleic acid to membrane protein targets, *Exp. Mol. Pathol. 86*, 141–150.
4. Ealick, S. E., and Armstrong, S. R. (1993) Pharmacologically relevant proteins, *Curr. Opin. Struct. Biol. 3*, 861–867.
5. Gschwend, D. A., Good, A. C., and Kuntz, I. D. (1996) Molecular docking towards drug discovery, *J. Mol. Recognit. 9*, 175–186.
6. Hoffmann, D., Kramer, B., Washio, T., Steinmetzer, T., Rarey, M., and Lengauer, T. (1999) Two-stage method for protein-ligand docking, *J. Med. Chem. 42*, 4422–4433.
7. Stahl, M., and Rarey, M. (2001) Detailed analysis of scoring functions for virtual screening, *J. Med. Chem. 44*, 1035–1042.
8. Charifson, P. S., Corkery, J. J., Murcko, M. A., and Walters, W. P. (1999) Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional

structures into proteins, *J. Med. Chem.* ***42***, 5100–5109.

9. Abagyan, R., and Totrov, M. (2001) High-throughput docking for lead generation, *Curr. Opin. Chem. Biol.* ***5***, 375–382.
10. Klebe, G. (2006) Virtual ligand screening: strategies, perspectives and limitations, *Drug Discov. Today* ***11***, 580–594.
11. Sperandio, O., Miteva, M. A., Delfaud, F., and Villoutreix, B. O. (2006) Receptor-based computational screening of compound databases: The main docking-scoring engines, *Curr. Protein Peptide Sci.* ***7***, 369–393.
12. Hermann, J. C., Marti-Arbona, R., Fedorov, A. A., Fedorov, E., Almo, S. C., Shoichet, B. K., and Raushel, F. M. (2007) Structure-based activity prediction for an enzyme of unknown function, *Nature* ***448***, 775-U772.
13. Kolb, P., Rosenbaum, D. M., Irwin, J. J., Fung, J. J., Kobilka, B. K., and Shoichet, B. K. (2009) Structure-based discovery of beta (2)-adrenergic receptor ligands, *P Natl Acad Sci USA* ***106***, 6843–6848.
14. Song, L., Kalyanaraman, C., Fedorov, A. A., Fedorov, E. V., Glasner, M. E., Brown, S., Imker, H. J., Babbitt, P. C., Almo, S. C., Jacobson, M. P., and Gerlt, J. A. (2007) Prediction and assignment of function for a divergent N-succinyl amino acid racemase, *Nat. Chem. Biol.* ***3***, 486–491.
15. Kalyanaraman, C., Imker, H. J., Federov, A. A., Federov, E. V., Glasner, M. E., Babbitt, P. C., Almo, S. C., Gerlt, J. A., and Jacobson, M. P. (2008) Discovery of a dipeptide epimerase enzymatic function guided by homology modeling and virtual screening, *Structure* ***16***, 1668–1677.
16. Rakus, J. F., Kalyanaraman, C., Fedorov, A. A., Fedorov, E. V., Mills-Groninger, F. P., Toro, R., Bonanno, J., Bain, K., Sauder, J. M., Burley, S. K., Almo, S. C., Jacobson, M. P., and Gerlt, J. A. (2009) Computation-Facilitated Assignment of the Function in the Enolase Superfamily: A Regiochemically Distinct Galactarate Dehydratase from Oceanobacillus iheyensis, *Biochemistry-Us* ***48***, 11546–11558.
17. Fan, H., Irwin, J. J., Webb, B. M., Klebe, G., Shoichet, B. K., and Sali, A. (2009) Molecular Docking Screens Using Comparative Models of Proteins, *J. Chem. Inf. Model.* ***49***, 2512–2527.
18. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank, *Nucleic Acids Res.* ***28***, 235–242.
19. Bairoch, A., Bougueleret, L., Altairac, S., Amendolia, V., Auchincloss, A., Puy, G. A., Axelsen, K., Baratin, D., Blatter, M. C., Boeckmann, B., Bollondi, L., Boutet, E., Quintaje, S. B., Breuza, L., Bridge, A., Saux, V. B. L., deCastro, E., Ciampina, L., Coral, D., Coudert, E., Cusin, I., David, F., Delbard, G., Dornevil, D., Duek-Roggli, P., Duvaud, S., Estreicher, A., Famiglietti, L., Farriol-Mathis, N., Ferro, S., Feuermann, M., Gasteiger, E., Gateau, A., Gehant, S., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Hulo, N., Innocenti, A., James, J., Jain, E., Jimenez, S., Jungo, F., Junker, V., Keller, G., Lachaize, C., Lane-Guermonprez, L., Langendijk-Genevaux, P., Lara, V., Le Mercier, P., Lieberherr, D., Lima, T. D., Mangold, V., Martin, X., Michoud, K., Moinat, M., Morgat, A., Nicolas, M., Paesano, S., Pedruzzi, I., Perret, D., Phan, I., Pilbout, S., Pillet, V., Poux, S., Pozzato, M., Redaschi, N., Reynaud, S., Rivoire, C., Roechert, B., Sapsezian, C., Schneider, M., Sigrist, C., Sonesson, K., Staehli, S., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Veuthey, A. L., Vitorello, C., Yip, L., Zuletta, L. F., Apweiler, R., Alam-Faruque, Y., Barrell, D., Bower, L., Browne, P., Chan, W. M., Daugherty, L., Donate, E. S., Eberhardt, R., Fedotov, A., Foulger, R., Frigerio, G., Garavelli, J., Golin, R., Horne, A., Jacobsen, J., Kleen, M., Kersey, P., Laiho, K., Legge, D., Magrane, M., Martin, M. J., Monteiro, P., O'Donovan, C., Orchard, S., O'Rourke, J., Patient, S., Pruess, M., Sitnov, A., Whitefield, E., Wieser, D., Lin, Q., Rynbeek, M., di Martino, G., Donnelly, M., van Rensburg, P., Wu, C., Arighi, C., Arminski, L., Barker, W., Chen, Y. X., Crooks, D., Hu, Z. Z., Hua, H. K., Huang, H. Z., Kahsay, R., Mazumder, R., McGarvey, P., Natale, D., Nikolskaya, A. N., Petrova, N., Suzek, B., Vasudevan, S., Vinayaka, C. R., Yeh, L. S., Zhang, J., and Consortium, U. (2008) The Universal Protein Resource (UniProt), *Nucleic Acids Res.* ***36***, D190–D195.
20. Baker, D., and Sali, A. (2001) Protein structure prediction and structural genomics, *Science* ***294***, 93–96.
21. Baker, D. (2000) A surprising simplicity to protein folding, *Nature* ***405***, 39–42.
22. Bonneau, R., and Baker, D. (2001) Ab initio protein structure prediction: Progress and prospects, *Annu. Rev. Biophys. Biomol. Struct.* ***30***, 173–189.
23. Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. (2000) Comparative protein structure modeling of genes and genomes, *Annu. Rev. Biophys. Biomol. Struct.* ***29***, 291–325.

24. Sali, A. (1998) 100,000 protein structures for the biologist, *Nat. Struct. Biol.* **5**, 1029–1032.
25. Chandonia, J. M., and Brenner, S. E. (2006) The impact of structural genomics: Expectations and outcomes, *Science* **311**, 347–351.
26. Liu, J. F., Montelione, G. T., and Rost, B. (2007) Novel leverage of structural genomics, *Nat. Biotechnol.* **25**, 850–853.
27. Pieper, U., Eswar, N., Webb, B., Eramian, E., Kelly, L., Barkan, D. T., Carter, H., Mankoo, P., Karchin, R., Marti-Renom, M. A., Davis, F. P., Sali., A., and Sanchez, R. (2009) MODBASE, a database of annotated comparative protein structure models, and ssociated resources, *Nucleic Acids Res. 37, D347–354.*
28. Jacobson, M., and Sali, A. (2004) Comparative protein structure modeling and its applications to drug discovery, *Annu. Rep. Med. Chem.* **39**, 259–276.
29. Bissantz, C., Bernard, P., Hibert, M., and Rognan, D. (2003) Protein-based virtual screening of chemical databases. II. Are homology models of G-protein coupled receptors suitable targets?, *Proteins: Struct. Funct. Genet.* **50**, 5–25.
30. Cavasotto, C. N., Orry, A. J. W., and Abagyan, R. A. (2003) Structure-based identification of binding sites, native ligands and potential inhibitors for G-protein coupled receptors, *Proteins: Struct. Funct. Genet.* **51**, 423–433.
31. Evers, A., and Klebe, G. (2004) Ligand-supported homology modeling of G-protein-coupled receptor sites: Models sufficient for successful virtual screening, *Angewandte Chemie-International Edition* **43**, 248–251.
32. Evers, A., and Klebe, G. (2004) Successful virtual screening for a submicromolar antagonist of the neurokinin-1 receptor based on a ligand-supported homology model, *J. Med. Chem.* **47**, 5381–5392.
33. Evers, A., and Klabunde, T. (2005) Structure-based drug discovery using GPCR homology modeling: Successful virtual screening for antagonists of the Alpha1A adrenergic receptor, *J. Med. Chem.* **48**, 1088–1097.
34. Moro, S., Deflorian, F., Bacilieri, M., and Spalluto, G. (2006) Novel strategies for the design of new potent and selective human A(3) receptor antagonists: An update, *Curr. Med. Chem.* **13**, 639–645.
35. Nowak, M., Kolaczkowski, M., Pawlowski, M., and Bojarski, A. J. (2006) Homology modeling of the serotonin 5-HT1A receptor using automated docking of bioactive compounds with defined geometry, *J. Med. Chem.* **49**, 205–214.
36. Chen, J. Z., Wang, J. M., and Xie, X. Q. (2007) GPCR structure-based virtual screening approach for CB2 antagonist search, *J. Chem. Inf. Model.* **47**, 1626–1637.
37. Zylberg, J., Ecke, D., Fischer, B., and Reiser, G. (2007) Structure and ligand-binding site characteristics of the human P2Y(11) nucleotide receptor deduced from computational modelling and mutational analysis, *Biochem. J.* **405**, 277–286.
38. Radestock, S., Weil, T., and Renner, S. (2008) Homology model-based virtual screening for GPCR ligands using docking and target-biased scoring, *J. Chem. Inf. Model.* **48**, 1104–1117.
39. Singh, N., Cheve, G., Ferguson, D. M., and McCurdy, C. R. (2006) A combined ligand-based and target-based drug design approach for G-protein coupled receptors: application to salvinorin A, a selective kappa opioid receptor agonist, *J. Comput.-Aided Mol. Des.* **20**, 471–493.
40. Kiss, R., Kiss, B., Konczol, A., Szalai, F., Jelinek, I., Laszlo, V., Noszal, B., Falus, A., and Keseru, G. M. (2008) Discovery of novel human histamine H4 receptor ligands by large-scale structure-based virtual screening, *J. Med. Chem.* **51**, 3145–3153.
41. de Graaf, C., Foata, N., Engkvist, O., and Rognan, D. (2008) Molecular modeling of the second extracellular loop of G-protein coupled receptors and its implication on structure-based virtual screening, *Proteins: Struct. Funct. Bioinform.* 71, 599–620.
42. Diller, D. J., and Li, R. X. (2003) Kinases, homology models, and high throughput docking, *J. Med. Chem.* 46, 4638–4647.
43. Oshiro, C., Bradley, E. K., Eksterowicz, J., Evensen, E., Lamb, M. L., Lanctot, J. K., Putta, S., Stanton, R., and Grootenhuis, P. D. J. (2004) Performance of 3D-database molecular docking studies into homology models, *J. Med. Chem.* 47, 764–767.
44. Nguyen, T. L., Gussio, R., Smith, J. A., Lannigan, D. A., Hecht, S. M., Scudiero, D. A., Shoemaker, R. H., and Zaharevitz, D. W. (2006) Homology model of RSK2 N-terminal kinase domain, structure-based identification of novel RSK2 inhibitors, and preliminary common pharmacophore, *Bioorg. Med. Chem.* 14, 6097–6105.
45. Rockey, W. M., and Elcock, A. H. (2006) Structure selection for protein kinase docking and virtual screening: Homology models or crystal structures?, *Curr. Protein Peptide Sci.* 7, 437–457.

46. Schapira, M., Abagyan, R., and Totrov, M. (2003) Nuclear hormone receptor targeted virtual screening, *J. Med. Chem.* 46, 3045–3059.

47. Marhefka, C. A., Moore, B. M., Bishop, T. C., Kirkovsky, L., Mukherjee, A., Dalton, J. T., and Miller, D. D. (2001) Homology modeling using multiple molecular dynamics simulations and docking studies of the human androgen receptor ligand binding domain bound to testosterone and nonsteroidal ligands, *J. Med. Chem.* 44, 1729–1740.

48. Kasuya, A., Sawada, Y., Tsukamoto, Y., Tanaka, K., Toya, T., and Yanagi, M. (2003) Binding mode of ecdysone agonists to the receptor: comparative modeling and docking studies, *J. Mol. Model.* 9, 58–65.

49. Li, R. S., Chen, X. W., Gong, B. Q., Selzer, P. M., Li, Z., Davidson, E., Kurzban, G., Miller, R. E., Nuzum, E. O., McKerrow, J. H., Fletterick, R. J., Gillmor, S. A., Craik, C. S., Kuntz, I. D., Cohen, F. E., and Kenyon, G. L. (1996) Structure-based design of parasitic protease inhibitors, *Bioorg. Med. Chem.* 4, 1421–1427.

50. Selzer, P. M., Chen, X. W., Chan, V. J., Cheng, M. S., Kenyon, G. L., Kuntz, I. D., Sakanari, J. A., Cohen, F. E., and McKerrow, J. H. (1997) Leishmania major: Molecular modeling of cysteine proteases and prediction of new nonpeptide inhibitors, *Exp. Parasitol.* 87, 212–221.

51. Enyedy, I. J., Ling, Y., Nacro, K., Tomita, Y., Wu, X. H., Cao, Y. Y., Guo, R. B., Li, B. H., Zhu, X. F., Huang, Y., Long, Y. Q., Roller, P. P., Yang, D. J., and Wang, S. M. (2001) Discovery of small-molecule inhibitors of bcl-2 through structure-based computer screening, *J. Med. Chem.* 44, 4313–4324.

52. de Graaf, C., Oostenbrink, C., Keizers, P. H. J., van der Wijst, T., Jongejan, A., and Vemleulen, N. P. E. (2006) Catalytic site prediction and virtual screening of cytochrome P450 2D6 substrates by consideration of water and rescoring in automated docking, *J. Med. Chem.* 49, 2417–2430.

53. Katritch, V., Byrd, C. M., Tseitin, V., Dai, D. C., Raush, E., Totrov, M., Abagyan, R., Jordan, R., and Hruby, D. E. (2007) Discovery of small molecule inhibitors of ubiquitin-like poxvirus proteinase I7L using homology modeling and covalent docking approaches, *J. Comput.-Aided Mol. Des.* 21, 549–558.

54. Mukherjee, P., Desai, P. V., Srivastava, A., Tekwani, B. L., and Avery, M. A. (2008) Probing the structures of leishmanial farnesyl pyrophosphate synthases: Homology modeling and docking studies, *J. Chem. Inf. Model.* 48, 1026–1040.

55. Rotkiewicz, P., Sicinska, W., Kolinski, A., and DeLuca, H. F. (2001) Model of three-dimensional structure of vitamin D receptor and its binding mechanism with 1 alpha,25-dihydroxyvitamin D-3, *Proteins: Struct. Funct. Genet.* 44, 188–199.

56. Que, X. C., Brinen, L. S., Perkins, P., Herdman, S., Hirata, K., Torian, B. E., Rubin, H., McKerrow, J. H., and Reed, S. L. (2002) Cysteine proteinases from distinct cellular compartments are recruited to phagocytic vesicles by Entamoeba histolytica, *Mol. Biochem. Parasitol.* 119, 23–32.

57. Parrill, A. L., Echols, U., Nguyen, T., Pham, T. C. T., Hoeglund, A., and Baker, D. L. (2008) Virtual screening approaches for the identification of non-lipid autotaxin inhibitors, *Bioorg. Med. Chem.* 16, 1784–1795.

58. Fernandes, M. X., Kairys, V., and Gilson, M. K. (2004) Comparing ligand interactions with multiple receptors via serial docking, *J. Chem. Inf. Comput. Sci.* 44, 1961–1970.

59. Kairys, V., Fernandes, M. X., and Gilson, M. K. (2006) Screening drug-like compounds by docking to homology models: A systematic study, *J. Chem. Inf. Model.* 46, 365–379.

60. McGovern, S. L., and Shoichet, B. K. (2003) Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes, *J. Med. Chem.* 46, 2895–2907.

61. Sali, A., and Blundell, T. L. (1993) Comparative Protein Modeling by Satisfaction of Spatial Restraints, *J. Mol. Biol.* 234, 779–815.

62. Lorber, D. M., and Shoichet, B. K. (1998) Flexible ligand docking using conformational ensembles, *Protein Sci.* 7, 938–950.

63. Lorber, D. M., and Shoichet, B. K. (2005) Hierarchical docking of databases of multiple ligand conformations, *Curr. Top. Med. Chem.* 5, 739–749.

64. Novoa, E. M., de Pouplana, L. R., Barril, X., and Orozco, M. (2010) Ensemble Docking from Homology Models, *J Chem Theory Comput* 6, 2547–2557.

65. Shoichet, B. K., Bodian, D. L., and Kuntz, I. D. (1992) Molecular Docking Using Shape Descriptors, *J. Comput. Chem.* 13, 380–397.

66. Meng, E. C., Shoichet, B. K., and Kuntz, I. D. (1992) Automated Docking with Grid-Based Energy Evaluation, *J. Comput. Chem.* 13, 505–524.

67. Ferrin, T. E., Huang, C. C., Jarvis, L. E., and Langridge, R. (1988) The Midas Display System, *J. Mol. Graphics* 6, 13–27.

68. SYBYL, 6.7 ed., Tripos International, 1699 South Hanley Rd., St. Louis, Missouri, 63144, USA.

69. Nicholls, A., and Honig, B. (1991) A Rapid Finite-Difference Algorithm, Utilizing Successive over-Relaxation to Solve the Poisson-Boltzmann Equation, *J. Comput. Chem.* 12, 435–445.

70. Huang, N., Shoichet, B. K., and Irwin, J. J. (2006) Benchmarking sets for molecular docking, *J. Med. Chem.* 49, 6789–6801.

71. Terasaka, T., Kinoshita, T., Kuno, M., and Nakanishi, I. (2004) A highly potent non-nucleoside adenosine deaminase inhibitor: Efficient drug discovery by intentional lead hybridization, *J. Am. Chem. Soc.* 126, 34–35.

72. Terasaka, T., Nakanishi, I., Nakamura, K., Eikyu, Y., Kinoshita, T., Nishio, N., Sato, A., Kuno, M., Seki, N., and Sakane, K. (2003) Structure-based de novo design of non-nucleoside adenosine deaminase inhibitors (vol 13, pg 1115, 2003), *Bioorg. Med. Chem. Lett.* 13, 4147–4147.

73. Sideraki, V., Wilson, D. K., Kurz, L. C., Quiocho, F. A., and Rudolph, F. B. (1996) Site-directed mutagenesis of histidine 238 in mouse adenosine deaminase: Substitution of histidine 238 does not impede hydroxylate formation, *Biochemistry-Us* 35, 15019–15028.

74. Vedadi, M., Lew, J., Artz, J., Amani, M., Zhao, Y., Dong, A. P., Wasney, G. A., Gao, M., Hills, T., Brokx, S., Qiu, W., Sharma, S., Diassiti, A., Alam, Z., Melone, M., Mulichak, A., Wernimont, A., Bray, J., Loppnau, P., Plotnikova, O., Newberry, K., Sundararajan, E., Houston, S., Walker, J., Tempel, W., Bochkarev, A., Kozieradzki, L., Edwards, A., Arrowsmith, C., Roos, D., Kain, K., and Hui, R. (2007) Genome-scale protein expression and structural biology of Plasmodium falciparum and related Apicomplexan organisms, *Mol. Biochem. Parasitol.* 151, 100–110.

75. Shen, M. Y., and Sali, A. (2006) Statistical potential for assessment and prediction of protein structures, *Protein Sci.* 15, 2507–2524.

76. Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res.* 32, 1792–1797.

77. Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., and Thompson, J. D. (2003) Multiple sequence alignment with the Clustal series of programs, *Nucleic Acids Res.* 31, 3497–3500.

78. Notredame, C., Higgins, D. G., and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment, *J. Mol. Biol.* 302, 205–217.

79. Fiser, A., Do, R. K. G., and Sali, A. (2000) Modeling of loops in protein structures, *Protein Sci.* 9, 1753–1773.

80. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., and Ferrin, T. E. (1982) A Geometric Approach to Macromolecule-Ligand Interactions, *J. Mol. Biol.* 161, 269–288.

81. Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S., and Weiner, P. (1984) A New Force-Field for Molecular Mechanical Simulation of Nucleic-Acids and Proteins, *J. Am. Chem. Soc.* 106, 765–784.

82. Mysinger, M. M., and Shoichet, B. K. (2010) Rapid Context-Dependent Ligand Desolvation in Molecular Docking, *J. Chem. Inf. Model.* 50, 1561–1573.