

Assessment of Docking Poses: Interactions-Based Accuracy Classification (IBAC) versus Crystal Structure Deviations

Romano T. Kroemer,^{*,†} Anna Vulpetti,[†] Joseph J. McDonald,[‡] Douglas C. Rohrer,[‡]
Jean-Yves Trosset,[†] Fabrizio Giordanetto,^{△,‡,§} Simona Cotesta,^{†,||} Colin McMartin,[⊥]
Mats Kihlén,[#] and Pieter F. W. Stouten[†]

Computational Sciences, Pharmacia Italia, Pfizer Group, Viale Pasteur 10, 20014 Nerviano, MI, Italy,
Computational Chemistry, Pfizer Global Research & Development, 700 Chesterfield Parkway North,
St. Louis, Missouri 63017, Department of Chemistry, Queen Mary, University of London,
Mile End Road, E1 4NS, UK, and ThistleSoft, 603 Colebrook Road, Colebrook Connecticut 06021

Received January 13, 2004

Six docking programs (FlexX, GOLD, ICM, LigandFit, the Northwestern University version of DOCK, and QXP) were evaluated in terms of their ability to reproduce experimentally observed binding modes (poses) of small-molecule ligands to macromolecular targets. The accuracy of a pose was assessed in two ways: First, the RMS deviation of the predicted pose from the crystal structure was calculated. Second, the predicted pose was compared to the experimentally observed one regarding the presence of key interactions with the protein. The latter assessment is referred to as interactions-based accuracy classification (IBAC). In a number of cases significant discrepancies were found between IBAC and RMSD-based classifications. Despite being more subjective, the IBAC proved to be a more meaningful measure of docking accuracy in all these cases.

INTRODUCTION

Discovering potential small-molecule drugs by assessing if, where, and how well they fit to a target receptor has become increasingly important over the years.^{1–3} In general, this process consists of two steps: first a computer program is used to place representations of small molecules in a macromolecular target structure (or to a user-defined part thereof, e.g., the active site of an enzyme). This is referred to as “docking”. Second, the binding enthalpies of the docked molecules are estimated by evaluating their complementarity to the target in terms of shape and properties such as electrostatics. Often entropic effects of binding are also assessed. This prediction of the binding free energy (affinity) is called “scoring”. A molecule with a good score is potentially a good binder. Although simply termed “docking programs”, all programs used nowadays carry out both the docking and scoring tasks. The distinction between docking and scoring defines also the two major technical challenges faced by docking programs: to predict the binding mode of a molecule correctly (herewith also referred to as “pose prediction”, where “pose” refers to orientation and conformation of a molecule in the receptor binding site)⁴ and to predict the binding affinity of compounds or to produce a relative rank-ordering for a number of compounds in a reliable manner.⁵

Determining the correct binding mode of a molecule involves finding the correct orientation and—as most ligand molecules are flexible—the correct conformation of the docked molecule. This implies that the degrees of freedom to be searched include translational and rotational degrees of freedom of the ligand as a whole as well as the internal degrees of freedom of the molecule, i.e., predominately the torsions. To this end a number of different search algorithms have been developed. In some of these algorithms the ligand is built up incrementally, starting from a docked “base fragment”. Representatives of this approach are Hammerhead,⁶ DOCK,⁷ and FlexX.⁸ In other approaches, such as AutoDock,⁹ GOLD,¹⁰ ICM-Dock,¹¹ and QXP,¹² the ligand is treated in its entirety. In addition to ligand flexibility it may be desirable to keep at least part of the receptor flexible, to reproduce the so-called “induced fit”, although relatively few docking programs take this into account so far. Notable examples of programs that incorporate receptor flexibility are the latest versions of AutoDock,¹³ FlexE,¹⁴ QXP, and Affinity.¹⁵ The way and degree to which this is implemented differs from program to program.

Searching for the correct binding mode of a molecule is typically carried out by performing a number of trials and keeping those poses that are energetically favorable. The search stops once a certain number of trials have been carried out and/or a sufficient number of poses have been found for a molecule. To explore a large search space, algorithms have been developed that keep track of previously discovered minima and guide the search into new regions.^{16–18} The decision to keep a trial pose is based on the computed ligand–receptor interaction energy (“score”) of that pose. To identify and rank-order many different poses of a molecule during the search in a reasonable time, in many

* Corresponding author phone: +39-02-4838 5221; fax: +39-02-4838 3965; e-mail: romano.kroemer@pharmacia.com.

[†] Pharmacia Italia.

[‡] Pfizer Global Research & Development.

[§] Queen Mary, University of London.

^{||} Current affiliation: Molecular Structure and Design, Pharmaceuticals Division, F. Hoffmann - La Roche Ltd., CH-4070 Basel, Switzerland.

[⊥] ThistleSoft.

[#] Current affiliation: Computational Chemistry & Informatics, Biovitrum AB, SE-112 76 Stockholm, Sweden.

[△] Current affiliation: AstraZeneca R&D Mölndal, S-431 83 Mölndal, Sweden.

Table 1. Summary and Main Characteristics of Docking Programs Used

program	algorithm	ligand BU ^a /E ^b	search algorithm	objective function	scoring function(s)	protein flexible	references
QXP	MCDOCK+	E	Monte Carlo, minimization	grid E/modified amber FF + contact	same as objective function	user defined parts move during minimization	latest algorithms not published yet, older reference: 12
	FULLDOCK+	E	tree pruning, Monte Carlo, minimization	as above	as above	as above	as above
NWU-Dock		E	several orientations (via matching features) of precalculated conformational ensemble	energy + desolvation	same as objective function	no	44
ICM-Dock		E	Monte Carlo, minimization	grid energy /force field	among others solvation and entropic terms	no	32, 33
FlexX		BU	orient rigid fragment first, build-up of ligand guided by docking score	Böhm-like molecular interactions	many	no	8, 39
Gold		E	genetic algorithm	fitness function	same as objective function	OH, NH2	10
LigandFit		E	random conformers/shape match/docking	grid energy, soft vdW and soft electrostatics or PLP1 score	many	no	35

^a BU – the ligand is built up incrementally during the docking. ^b E – the ligand is treated in its entirety during the docking.

cases a relatively simple energy function (e.g., a force field with an electrostatic term and repulsive and attractive van der Waals terms) is employed. This function, also referred to as “objective function”, can be evaluated very rapidly. A more sophisticated function (the “scoring function”) is subsequently used to calculate the final score for that molecule. These scores can then be used to rank-order different molecules.

A large variety of scoring schemes exists, the review of which is beyond the scope of this paper. The reader is therefore referred to a number of publications on scoring functions^{19–25} and consensus scoring schemes.^{26–28} For a comprehensive overview of prediction of binding affinity a recent review by Gohlke and Klebe provides an excellent read.²⁹

As outlined above, there are two major challenges faced by docking programs: first, to predict the binding mode of a molecule correctly, and second, to predict binding affinities reliably. The first point is a prerequisite for the second: if the ligand is not docked correctly, it is unlikely that the calculated score is meaningful, apart from fortuitous cases where calculated affinities for different poses are close in energy. Therefore, in the course of our in-house evaluation of docking programs and their assessment for use as virtual screening tools, we decided to investigate first the ability of docking programs to produce correct binding modes for a number of ligand-protein complexes, although it is recognized that already at this stage the quality of the scoring function is important as it is used to accept or reject poses.

The primary goal of our evaluation studies was to identify the program(s) that deliver the best results to Pharmacia discovery projects. Therefore, we selected programs for evaluation that were either industry standards or that had delivered promising results (as evidenced by papers in the scientific literature and presentations at scientific meetings) or that we had significant prior experience with and had worked well in our hands. Here we will report results obtained with the following set of programs: LigandFit (ccO

version, Accelrys), ICM-Dock (the docking module in ICM2.8, MolSoft), NWU-Dock (the Northwestern University version of UCSF’s DOCK, version 5.2), FlexX (version 1.10.1.L, Tripos), QXP (2003 version, ThistleSoft), and GOLD (version 1.2, CCDC). References to the programs and a brief description of their main features are provided in Table 1.

During our evaluation studies we investigated a variety of different protein targets, and results regarding two of our data sets are reported here. One of the targets was CDK2/Cyclin A, a serine/threonine kinase that is involved in cell-cycle regulation. It had been chosen because of its in-house relevance and the abundance of structural and inhibitory data. For this data set, we selected 9 CDK2 inhibitors with known CDK2/Cyclin A binding modes and inhibition data. The second data set contained a variety of protein–ligand complexes from the Protein Data Bank (PDB)^{30,31} that span a variety of protein families. The results on a subset of this data set containing 10 structures are reported here. Both data sets are summarized in Tables 2 and 3.

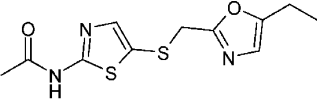
For each complex investigated, data preparation and docking were performed as required by the programs used. Despite the practical appeal of using RMSDs from a crystal structure to assess pose prediction accuracy, they do not do justice to the complex interactions ligands make with proteins. For that reason, we devised a novel way to evaluate pose prediction accuracy. Here the correctness of a pose was determined by visually comparing the (hydrogen-bonded and other) ligand–protein interactions for that pose with the experimentally observed interactions. In particular this interactions-based accuracy classification (IBAC) scheme was introduced because of the following considerations:

- Differences in the force fields implemented in the docking programs may lead to variations in the predicted poses, resulting in relatively large RMSD values without changing the overall binding modes and interactions.
- If a molecule contains a very flexible moiety (e.g., the hydroxypropyl chain in compound **7** of data set 1) that is

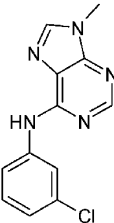
Table 2. Data Set 1 – CDK2/Cyclin A Ligands

Ligand name	Compound ID	Reference	PDB code
BMS compound 1	1		
CKP	2	48	1CKP
P38-MAPK inhibitor	3	49	1DI8
Hymenialdisine	4	50	1DM2
Oxindole	5	51	1FVT
Indirubine	6	52	
Novartis CGP60474	7		
Warner-Lambert	8		
Staurosporine	9	53	1AQ1

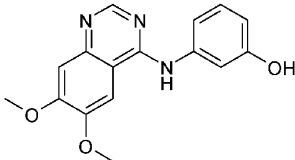
Ligand structures



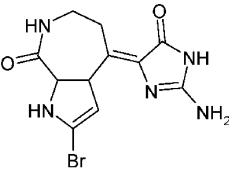
1



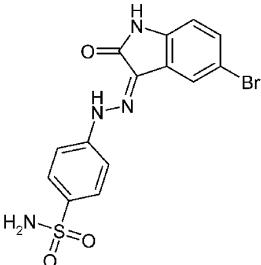
2



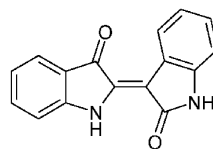
3



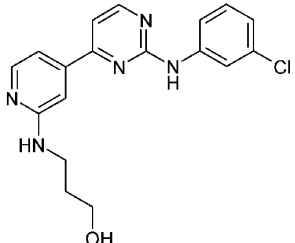
4



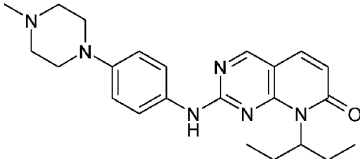
5



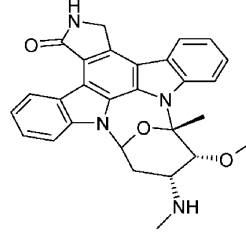
6



7



8



9

not involved in key interactions, the differences between docked and experimental binding mode for this part might lead to a high RMSD for the entire molecule, although the overall binding mode of the docked molecule is correct.

- A large, almost symmetric molecule may adopt a nearly correct pose (i.e., correct with the exception of the symmetry-breaking moieties) but may have a very large RMSD.

The exact criteria for evaluation are listed in the methods section, and the differences between the two evaluation methods are reported and discussed thereafter.

The purpose of this study was not to compare different docking programs with each other and to establish the superiority of any one over the others but to determine the

best way to assess correctness of docking poses. A comparison of different programs will be described elsewhere.

METHODS

General. The results one obtains with various docking programs can depend very sensitively on the active site definitions and docking protocols. To obtain the best possible results with each program we had to optimize the binding site definitions and docking protocols. Since we are interested in the intrinsic quality of the programs and we recognize that few if any people are expert users of all 6 programs we consider, we involved an expert-operator of each of the programs to obtain the best possible results: we set up and

Table 3. Data Set 2 – Different Ligand–Protein Complexes

PDB ID	protein	ligand	res [Å]
183L	lysozyme mutant (C54T,C97A,L99A)	indene	1.8
186L	lysozyme mutant (C54T,C97A,L99A)	N-butylbenzene	1.8
1ABE	L-arabinose-binding protein	L-arabinose	1.7
1BBP	bilin-binding protein	biliverdin IX	2.0
1LIC	adipocyte lipid-binding protein	hexadecanesulfonic acid	1.6
1MRK	α -trichosanthin	formycin	1.6
1SRH	streptavidin	3',5'-dimethoxy-HABA	2.2
1TNL	trypsin	tranylcypromine	1.9
3CPA	carboxypeptidase A	glycyl- ³ H-L-tyrosine	1.5
7TIM	triosephosphate isomerase	phosphoglycolohydroxamate	1.9

executed the docking runs in close collaboration with the developers/vendors of the programs, and in a few rare cases where this was not feasible, we shared details of our proposed docking experiments with them and implemented any changes they proposed.

All ligand–receptor complexes were visually inspected before the docking runs. Hydrogen bonds were established with the ligands in place, to ensure correct orientations of the hydrogens involved. The ligands were checked for correct tautomeric forms and bond orders. Protonation states were generated corresponding to an aqueous environment at pH 7, i.e., carboxylic acids were replaced by carboxylates and amines were protonated. Hydrogens were added where appropriate. Cominimizations of the protein–ligand complexes were performed with the appropriate force fields, to alleviate bad contacts that would prevent correct redocking. Subsequently, the ligands were extracted, and the free ligands were randomized with respect to their bound conformations, energy-minimized, and randomly translated and rotated to eliminate any possibility of biasing the docking experiments. Ligand files were usually generated in mol (SD) format and supplied as input. By default, the proteins were prepared by deleting the crystallographic water molecules and by generating the appropriate protonation states for charged residues, as implemented in the programs used. The procedures to define location and extent of the binding sites vary significantly between the docking programs investigated, and it is therefore impossible to generate identical binding site definitions. Great care was taken, however, to construct equivalent binding sites. This was achieved by expanding the ligand-binding pocket to include the entire concave region around the ligand(s) and further extending it into the solvent region (typically, by approximately 3 Å). For data set 1 the crystal structure of 1QMZ was used as a target for all ligands. K89, which partially obstructs access to the ATP binding site in the structure, was mutated to alanine in order to guarantee free access to the ATP binding pocket for all ligands.

Criteria for Assessment of Pose Prediction Accuracy.

The classification protocol for data set 1 was as follows: first, the pose of the scaffold of a docked compound (e.g., the indolenone moiety of compound **5**) was compared to the experimentally determined binding mode. If its position was comparable and it made the same interactions with the protein (e.g., the two hydrogen bonds with the backbone of the so-called hinge region, residues E81 and L83), this part of the molecule was deemed correctly docked. Second, the remaining parts of the molecules were assessed and they were checked for the presence of key interactions with the protein that occur in the experimentally determined reference structure.

Classification criteria for data set 2 were different in that three classes were defined for pose prediction quality: correct, nearly correct, and incorrect. To be classified as correct, the docked compound had to be in the correct orientation and conformation, and all key interactions such as hydrogen bonds had to be present. A pose was regarded as nearly correct when the overall orientation and conformation of the ligand was correct, but some (up to a quarter) of the interactions deemed to be relevant were not present. All docked molecules not classified as correct or nearly correct were considered incorrect. The “correct” class of data set 1 corresponds approximately to the combined “correct” and “nearly correct” classes of data set 2.

A precise description of the assessment of prediction accuracy is provided in the Results section using an example.

Description of Docking Programs and Settings. The docking algorithm implemented in **ICM** (version 2.8) optimizes the entire ligand in the receptor field, using a multistart Monte Carlo minimization procedure in internal coordinate space. During docking, the energy is evaluated using a precalculated grid for computational efficiency. By default only the highest-ranking docked conformer of each compound, as evaluated by the energy function (the dock score), is scored using the VLS scoring function (the affinity score). More details can be found in references 32 and 33.

Receptor grid maps were calculated with a grid spacing of 0.5 Å. The grid was defined in such a way that it included the ligand and key residues plus 3–5 Å in each of the x,y,z directions. Docking was carried out using the default parameters set in ICM. The Metropolis temperature for accepting or rejecting a new pose was set to 600 K. The number of Monte Carlo steps and the number of iterations of the local energy minimization were determined automatically by an adaptive algorithm, depending on the size and number of flexible torsions in the ligand.

In **LigandFit** the docking procedure starts with the generation of random ligand conformations. During docking the shape of a given conformation (defined by the principal axes of inertia) is compared with the shape of the active site. Scoring of the docked ligand conformations is done with a variety of different scoring functions implemented in the Cerius2 program.³⁴ The docking algorithm is described in detail in ref 35.

The site models for docking were derived in the following manner: an initial binding site was defined based on the position of the ligands in the crystallographic structures under investigation. The grid resolution was set to 0.5 Å (default), and the ligand-accessible grid was defined such that the minimum distance between a grid point and the protein is 2.0 Å for hydrogen and 2.5 Å for heavy atoms. The initial

binding site was expanded away from the protein into the solvent region by a triple expansion. Each expansion involves the addition to the binding site of all free grid points adjacent to all grid points already part of this then current binding site. Site partitioning was applied with a value of 5.

FlexX uses an incremental construction algorithm where ligands are docked starting with a “base fragment”. After placement of a base fragment (in different positions) the complete ligand is constructed by adding the remaining components back on. At each step the interactions are evaluated, and the best solution is selected according to the docking score. The docking score uses the model of molecular interactions developed by Böhm and Klebe.^{36–38} For more details, refer to refs 8 and 39.

For data set 1 the binding site was defined as all CDK2 residues within a radius of 9 Å of staurosporine in the 1AQ1 complex. Staurosporine was selected for this purpose because it is the largest ligand in data set 1, and it occupies the ATP binding pocket to a very large extent. For data set 2 the binding sites were defined by including all residues of the protein within a radius of 8 Å of the respective ligand. Standard parameters, as set in the FlexX implementation of Sybyl6.8,⁴⁰ were used throughout the docking runs.

NWU-Dock is an incarnation of the DOCK program,⁴¹ developed at Northwestern University. It differs from the original DOCK in two aspects: desolvation of the ligand is treated explicitly with the AMSOL 6.5.3 program⁴² and conformations of ligands are precalculated with Omega.⁴³ NWU-Dock docks flexible ligands using conformational ensembles.^{44,45}

For data set 1, the optimization of the DOCK spheres was done by hand, since the number and placement of spheres generated by SPHGEN using the Connolly surface was not sufficient, as indicated by the absence of automatically generated spheres in regions of the site where important interactions with the ligands could be expected or had been observed in some complexes. The heavy atom positions of the ATP molecule in 1QMZ were used for defining the locations of the spheres. Additional spheres were added in the phosphate and ribose binding regions of the pocket, using the crystal structures of other ligands bound to CDK2/cyclin A as a reference.

GOLD (Genetic Optimization for Ligand Docking) utilizes a genetic algorithm (GA) described by Jones et al.,⁴⁶ which mimics the process of evolution by applying genetic operators to a collection of putative poses for a single ligand (in GA terms, a population of chromosomes). For a detailed description of the GOLD algorithm refer to references 46, 47, and 10.

Three genetic operators were applied: point mutation of a chromosome, crossover (i.e., mating of two chromosomes), and migration of a population member from one island to another. Each genetic operation involved selection of an operator with probabilities as follows: mutation 47.5%, crossover 47.5%, and migration 5%. After the application of 100 000 genetic operations the algorithm terminated, saving the poses with the highest scores.

QXP (Quick Explore) is part of the Flo+ program. It contains two conceptually different docking algorithms: MCDOCK and FULLDOCK. Both were evaluated in this study. For each ligand the MCDOCK algorithm applies a user-defined number (we use 1000) of repeated cycles of

Monte Carlo (MC) followed by energy minimization (EM) to generate and refine an ensemble of 50 low-energy ligand poses. MCDOCK is evolutionary: it randomly selects a member from the ensemble and applies a mutation (i.e., a random change of the docking pose) prior to each MC/EM cycle and after each cycle the worst pose in the ensemble is replaced by the MC/EM pose, provided it meets certain energetic and positional diversity criteria. However, unlike a Genetic Algorithm, MCDOCK does not employ crossover (splicing or hybridization of existing docking pose geometries, e.g., across rotatable bonds). Further information on the MCDOCK procedure can be found in reference 12. The FULLDOCK algorithm attempts to address the issue that random Monte Carlo methods do not provide a systematic search for a globally best solution. First it explores the binding site in a near-exhaustive fashion with a single tree that represents a very large numbers of conformers. It also carries out up to 10 independent dockings to generate a diverse set of solutions. The poses thus generated are used to seed the ensemble for a subsequent MCDOCK run of typically 500 MC/EM cycles. A detailed description of the FULLDOCK algorithm is in preparation. QXP allows binding site flexibility, but we did not use that feature in this work. QXP has two scoring functions, a molecular mechanics force field and an empirical “plus” version, optimized simultaneously for pose and affinity prediction. We use the plus version throughout.

RESULTS AND DISCUSSION

The de facto literature procedure for assessing how good a pose is involves calculating an RMSD from the crystal structure. For the reasons alluded to in the Introduction, to determine pose prediction quality we devised the interactions-based accuracy classification (IBAC) scheme. The comparison between RMSD values and IBAC classification is summarized in Figures 1 and 2 for data sets 1 and 2, respectively. In general there is a fairly good correlation between the RMSD values and the IBACs: the higher the RMSD, the more likely the corresponding pose has been classified as incorrect. However, in a significant number of cases this is not true. The arrows in Figures 1 and 2 point to the clearest examples.

For data set 1, compound **1** is the most striking example of a discrepancy between RMSD and IBAC classification. Despite a low RMSD of 1.62 Å from the experimentally determined binding mode, the pose generated by ICM is deemed incorrect according to the IBAC. The reason for this is illustrated in Figures 3 and 4. In the methods section we described the overall criteria for assessing pose prediction accuracy with the IBAC method. In the following we will use compound **1** in order to illustrate how the assessment was done precisely.

First the crystal structure was inspected for key interactions of the ligand with the receptor, specifically hydrogen bonds, salt bridges, and hydrophobic contacts. For molecule **1**, the hydrogen bond pattern is shown in Figure 3, whereas the hydrophobic contacts are displayed in Figure 4. To define the latter contacts, the following procedure had been applied: Residues that were in close contact with the ligand were identified. The distance between a given residue and the closest ligand atom was defined as a close contact. In

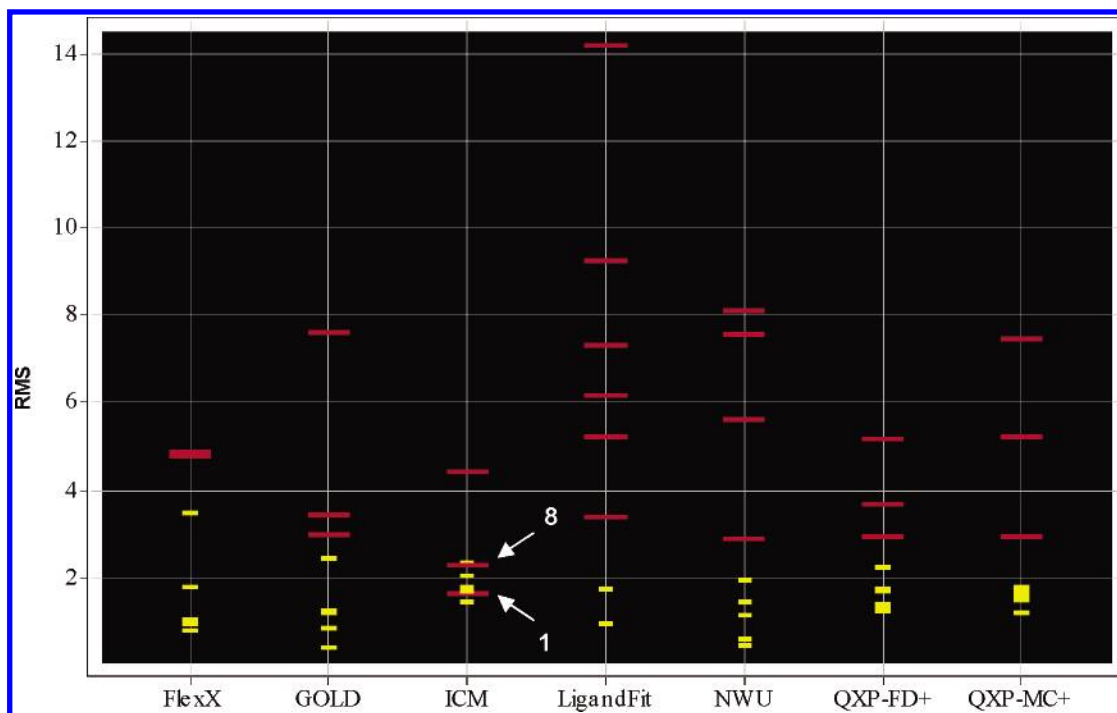


Figure 1. Comparison between calculated RMSD and assigned correctness according to IBAC for data set 1. Yellow bars indicate correct poses and red bars incorrect poses. Arrows point to entries whose IBAC and RMSD do not match (see text for explanation).

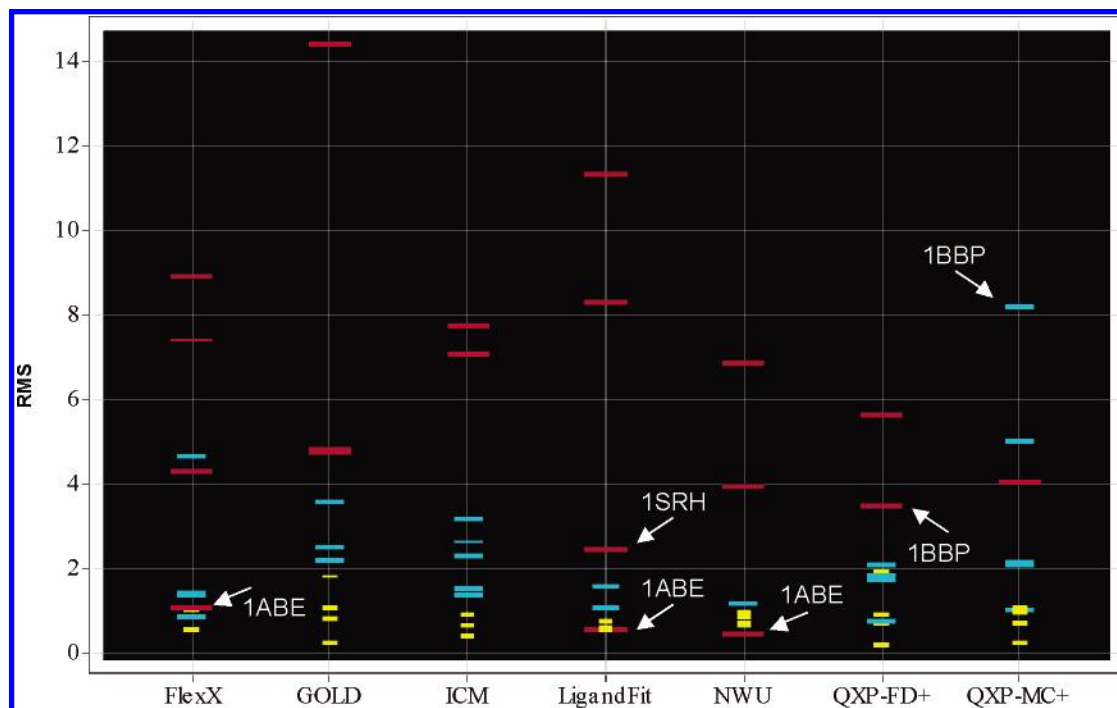


Figure 2. Comparison between calculated RMSD and assigned correctness according to IBAC for data set 2. Yellow bars indicate correct poses, blue bars indicate nearly correct poses and red bars incorrect poses. Arrows point to entries whose IBAC and RMSD do not match (see text for explanation).

case the same residue was also in close contact with a different moiety of the ligand, another hydrophobic contact was assigned for that residue, too (e.g. L134).

The crystallographic binding mode of compound **1** involves two hydrogen bonds with L83 (part of the CDK2 hinge region): the ligand's thiazole nitrogen interacts with L83-NH and its amide-NH with L83-CO (Figure 3). In the ICM pose, at first glance it seems that the entire compound has just moved away slightly from the hinge region. However, the hydrogen bond interaction pattern is

entirely different. None of the crystallographic hydrogen bonds are retained, while the ligand's amide group has flipped 180° in order to make a single new hydrogen bond interaction between its CO and L83-NH. The thiazole moiety is no longer predicted to interact with the hinge region at all.

For the crystal structure of molecule **1** six hydrophobic contacts were identified (Figure 4A). As indicated in Figure 4B, of these six contacts three are maintained by docking, whereas for three contacts the distances have been increased

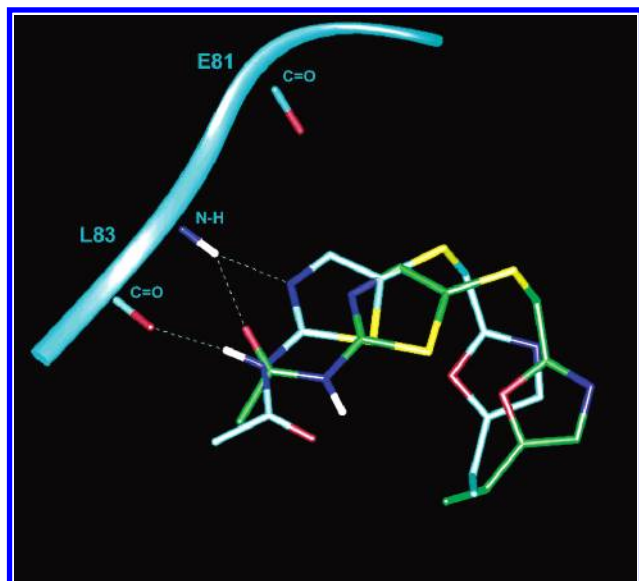


Figure 3. Comparison between experimental binding mode and the top-scoring ICM pose for compound **1**. The CDK2 hinge region is displayed as a ribbon with hydrogen bond donor and acceptors. Cyan carbon atoms: experimental binding mode. Green carbon atoms: ICM pose, classified as incorrect, RMSD = 1.62 Å.

significantly, i.e., these contacts are no longer present. Overall for compound **1** a total of eight key interactions had been identified in the crystal structure: two hydrogen bonds and six hydrophobic contacts. Examination of the docked pose revealed that both hydrogen bonds and three hydrophobic contacts had been lost. Despite the fact that the docked molecule establishes some new interactions with the protein, of the original key interactions more than 50% are not present anymore. As a consequence, the IBAC for this compound was assigned as incorrect.

Despite a rather low overall RMSD value (2.30 Å), ICM does not correctly predict the hydrogen bond pattern of compound **8** with the hinge region either (Figure 5). Again, the docked molecule has not moved very far, and its overall docking mode is rather similar to the crystal structure. Closer inspection reveals some significant differences, however: in the crystal structure the molecule forms two hydrogen bonds with the hinge region of CDK2, neither of which are present in the predicted pose.

Data set 2 contains three cases with significant RMSD-IBAC discrepancies. 1ABE (with the sugar β -arabinose as ligand) is similar to the two cases described for data set 1. All programs generate a top-scoring pose with very low RMSD values for all heavy atoms (0.20–1.04 Å), implying that the overall position of the ligand is in agreement with experiment. However, the FlexX, LigandFit, and NWU-Dock poses are classified as incorrect (see Figure 2) as each of them lacks at least 2 of the 4 hydrogen bonds observed in the crystal structure. Figure 6 shows a comparison between the correct structure and the poses generated by LigandFit and NWU-Dock. NWU-Dock gets 2/4 hydrogen bonds right, and LigandFit predicts 1/4 of the hydrogen bonds correctly. Interestingly, while the RMSDs of poses generated by FlexX (1.04 Å), LigandFit (0.54 Å), and NWU-Dock (0.44 Å) are low in absolute terms, they are higher than those of the correct poses generated by the other programs.

The second data set 2 example is provided by 1SRH. In the crystal structure, the carboxylate of the ligand's benzoate

moiety makes three hydrogen bonds with the protein (Figure 7), while its phenyl group is packed in a hydrophobic pocket made up of a threonine and three tryptophans. Since close contacts can be observed in this case in the crystal structure, care had been taken to eliminate these errors by performing a cominimization of the protein and the ligand prior to docking. Despite these efforts, in the LigandFit pose the benzoate flips 180°, putting the carboxylate in this hydrophobic pocket, where it accepts a single geometrically awkward hydrogen bond from T90-O γ only. Although this pose has a relatively low RMSD of 2.46 Å, it is clearly incorrect.

The most dramatic example of how RMSDs can be misleading is provided by 1BBP as docked by QXP/MCDOCK+ and QXP/FULLDOCK+ (Figure 8). For QXP/MCDOCK+ the predicted pose has a very high RMSD (8.17 Å) but is regarded nearly correct. The crystallographically observed ligand is rather planar and almost C₂ symmetric with respect to its main features. When docked by QXP/MCDOCK+ it is flipped about this pseudo-C₂-axis but overall adopts a very similar pose. The main difference compared to the experimental structure is that the methyl and vinyl groups are swapped. Therefore, the docked pose was classified as nearly correct. By contrast, the pose generated by QXP/FULLDOCK+ is very different, despite a much lower RMSD (3.46 Å). Although the molecule is not flipped about the C₂-axis, none of the heterocycles overlap with the corresponding experimental ones. The same is true for the two carboxyl groups, which do not engage in the same interactions as in the experimental binding mode.

The five cases discussed here illustrate that a small RMSD does not necessarily imply a correctly docked molecule and, conversely, that a large RMSD does not mean the compound is incorrectly docked. If it is at all possible to set a RMSD threshold below which a molecule is guaranteed to be docked correctly (as defined by its IBAC), our results indicate that this threshold is very low: for data set 2 we found numerous examples of poses with RMSD values around or below 1 Å that were deemed only nearly correct or even incorrect. Conversely, we found examples with rather or very high RMSD values that can still be considered correct or nearly correct (most notably 1BBP when docked with QXP/MCDOCK+).

The IBAC procedure presented here depends on the assignment of the so-called key interactions, which are hydrogen bonds, salt bridges, and hydrophobic contacts. Although the simple set of rules for the definition of certain interactions as “key” and the fact they all carry the same weight may be perceived as subjective, the IBAC method is deemed to be superior to RMSD calculations for the following reasons: in RMSD calculations no distinction is made between different atoms or atom types and they all enter the RMSD formula with the same weight, irrespective of whether they are involved in interactions or not. The IBAC method, however, through a set of strict, albeit simple, rules focuses only on the relevant features and thereby increases the signal-to-noise ratio in the assessment of pose prediction quality. Furthermore, the IBAC method will adequately account for those cases where near-symmetry can lead to a very similar binding mode.

A caveat is that the method is not yet automated and requires case-by-case evaluation. The aim of the present

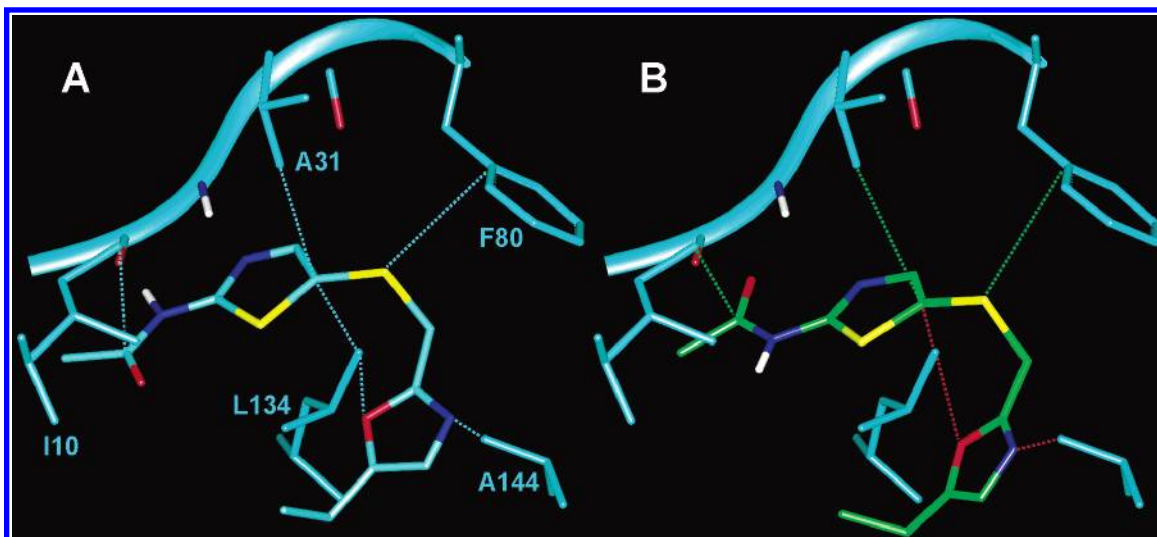


Figure 4. Pane A: Key hydrophobic contacts as identified in the crystal structure of compound **1**, represented by dotted lines. Pane B: Contacts, corresponding to the ones in Pane A, for the docked compound. Color coding is as follows: green — the interatomic distances have increased only slightly (2–14%), red — the interatomic distances have increased significantly (28–37%).

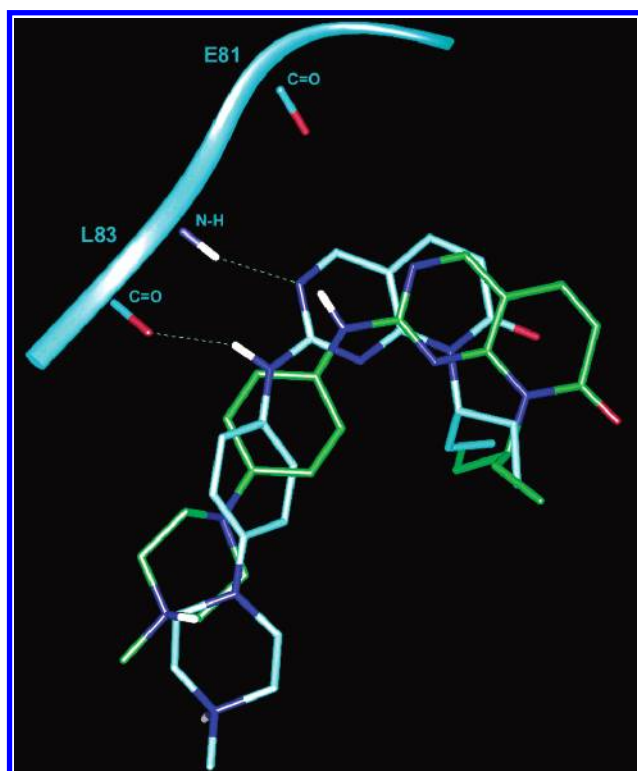


Figure 5. Comparison between experimental binding mode and the top-scoring ICM pose for compound **8**. The CDK2 hinge region is displayed as a ribbon with hydrogen bond donor and acceptors. Cyan carbon atoms: experimental binding mode. Green carbon atoms: ICM pose, classified as incorrect, RMSD = 2.30 Å.

study is to highlight the differences between RMSD classification and IBAC, but nevertheless we would like to propose potential ways to automate the method, which will be the subject of further studies. Automated calculation of key interactions, e.g. hydrogen bonds or hydrophobic contacts, could be carried out similar to the LIGPLOT implementation.⁵⁴ Interaction patterns for experimental and docked poses could subsequently be compared as it is being done in a recent procedure for identification of hydrogen bond signature patterns in proteins.⁵⁵ Based on the ligands alone, rather than a simple RMSD calculation between the same

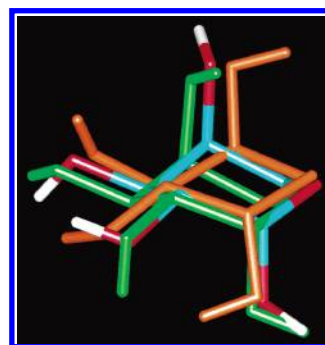


Figure 6. Comparison between experimental binding mode of β -arabinose (PDB entry 1ABE) and the top-scoring poses generated by NWU-Dock (green) and LigandFit (brown). Atom colors for the experimentally observed molecule are cyan, red, and white for carbons, oxygens, and hydrogens, respectively.

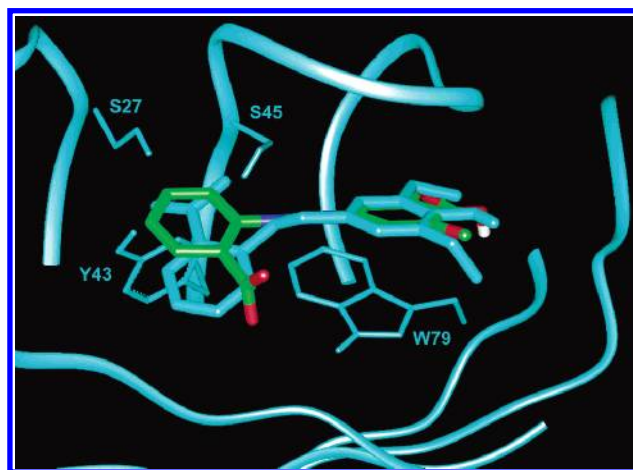


Figure 7. Comparison between experimental binding mode of 3',5'-dimethoxy-HABA (PDB entry 1SRH) and the top-scoring pose generated by LigandFit. The crystal structure of 1SRH after minimization of the ligand in the binding site is shown in cyan. Selected residues are displayed. Atom colors for the docked molecule are green, red, and white for carbons, oxygens, and hydrogens, respectively.

atoms of two different poses (crystal structure and docking mode), one might match the shapes of the two poses. A simple way of doing this would be to determine for each

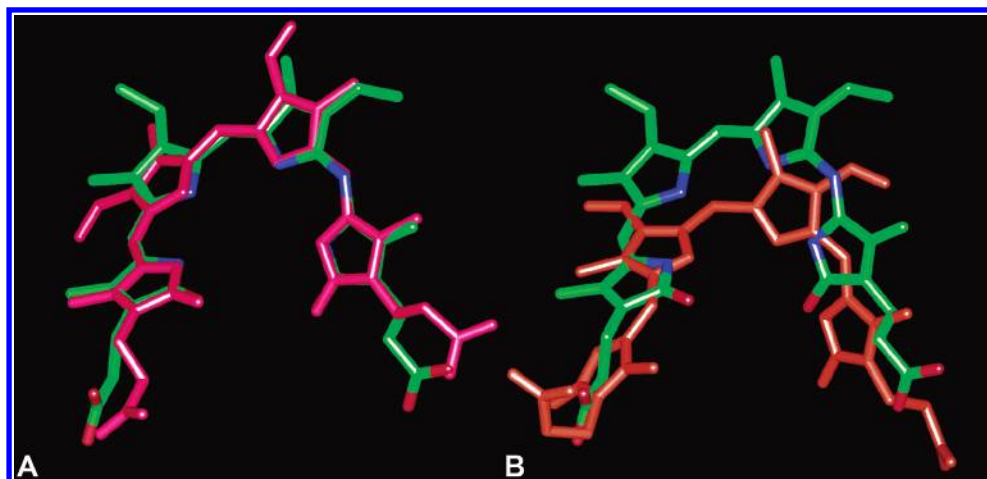


Figure 8. Comparison between QXP pose prediction results and the crystallographic pose for 1BBP. The crystallographically observed ligand is colored by atom types (carbon – green, nitrogen – blue, oxygen – red). Pane **A**: QXP/MCDOCK+ result (colored red), RMSD = 8.17 Å, classified as nearly correct. Pane **B**: QXP/FULLDOCK+ result (colored orange), RMSD = 3.46 Å, classified as incorrect.

atom of one pose the atom nearest in space of the second pose and calculate RMSDs between the resulting atom pairs. A more sophisticated way of comparing the shapes of experimental and docked poses could entail the calculation of molecular field-based similarity,⁵⁶ grid-based volume overlap,⁵⁷ or molecular similarity indices.⁵⁸ However, in these cases it would be important to determine whether the procedures have the potential to blur the results similar to an RMSD calculation. One could also devise a combination of shape matching with an interactions count.

We have demonstrated that IBACs are more meaningful measures of pose prediction quality than RMSDs, but using RMSDs to compare different docking tools applied to many different complex structures has a practical appeal, because it can be automated. If average or median RMSDs track with the IBAC, however, then we might use RMSDs for the comparison after all. Unfortunately, this is not the case as illustrated by comparing the QXP/MCDOCK+ results with the results of other programs for data set 2. For this (admittedly very limited) data set the mean/median RMSD values for QXP/MCDOCK+ are of similar magnitude as those of several of the other programs. Considering the number of incorrectly docked molecules, however, QXP/MCDOCK+ fares much better in this particular example, as it incorrectly predicts only one molecule.

CONCLUSIONS

Analysis of pose prediction results is an important issue. Although in general RMSDs from the crystal structure correlate with our interactions-based accuracy classification, we identified some striking examples where that is not true. These include, but are not limited to, cases with different hydrogen bond interaction patterns between the crystal structure and the docking pose. Of course, in some cases (e.g. the sugar molecule in this study) the hydrogen bond interactions could be corrected in hindsight upon close inspection of the docking mode, but in other instances this would not be possible.

Only considering RMSD values can lead to the misclassification of both correct and incorrect poses, and we therefore recommend close inspection of binding modes when assessing pose prediction accuracy.

One potential application of docking programs is the prediction of binding modes for novel compounds, based on which decisions can be made for further design and/or synthesis or testing. If docking predictions are used in this manner, then it is important that the key interactions are correctly predicted. Therefore, a docking program that is better at this task would be our preferred choice. The comparison of different docking programs based on median/average RMSDs or based on IBACs lead to very different conclusions. Notwithstanding the practical appeal of RMSDs, IBACs or equivalent metrics that properly account for key interactions are more meaningful.

Despite the observation that RMSD calculations appear to be less meaningful than IBAC when it comes to assessment of pose prediction quality, the advantage of RMSD values is that they can be calculated in an automated fashion for the assessment of a large number of binding modes. To implement an automated IBAC version, one could devise an automated protocol to calculate and count key interactions, such as hydrogen bonds or hydrophobic contacts. Alternatively, one could resort to shape matching algorithms or even to a combination of both interaction count and shape matching. If implemented in a suitable manner these algorithms would eliminate the completely erroneous assessments based on RMSD values described in this paper and would yield a measure of similarity to the crystal structure that better reflects the important aspects of binding than mere RMSD values.

ACKNOWLEDGMENT

We thank the following people for their help and guidance in setting up and using their docking programs: Shashidhar Rao (Accelrys; LigandFit), Brian Shoichet and John Irwin (UCSF; NWU-Dock), Maxim Totrov (MolSoft; ICM), and Bruno Cherel (Tripos; FlexX). We are also very grateful to Brian Shoichet for hosting one of us (A.V.) during a minisabbatical.

REFERENCES AND NOTES

- (1) Schneider, G.; Böhm, H.-J. Virtual Screening and Fast Automated Docking Methods. *Drug Discovery Today* **2002**, 7, 64–70.
- (2) Waszkowycz, B. Structure-Based Approaches to Drug Design and Virtual Screening. *Curr. Opin. Drug Discovery Dev.* **2002**, 5, 407–413.

- (3) Toledo-Sherman, L. M.; Chen, D. High-Throughput Virtual Screening for Drug Discovery in Parallel. *Curr. Opin. Drug Discovery Dev.* **2002**, *5*, 414–421.
- (4) Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W. Deciphering Common Failures in Molecular Docking of Ligand-Protein Complexes. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 731–751.
- (5) Stahl, M.; Rarey, M. Detailed Analysis of Scoring Functions for Virtual Screening. *J. Med. Chem.* **2001**, *44*, 1035–1042.
- (6) Welch, W.; Ruppert, J.; Jain, A. J. Hammerhead: Fast, Fully Automated Docking of Flexible Ligands to Protein Binding Sites. *Chem. Biol.* **1996**, *3*, 449–462.
- (7) Ewing, T. J. A.; Kuntz, I. D. Critical Evaluation of Search Algorithms for Automated Molecular Docking and Database Screening. *J. Comput. Chem.* **1997**, *18*, 1176–1189.
- (8) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. **1996**, *261*, 470–489. *J. Mol. Biol.* **1996**, *261*, 470–489.
- (9) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19*, 1639–1662.
- (10) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267*, 727–748.
- (11) Totrov, M.; Abagyan, R. Flexible Protein–Ligand Docking by Global Energy Optimization in Internal Coordinates. **1997**, *SI*, 215–220. *Proteins: Struct., Funct., Genet.* **1997**, *SI*, 215–220.
- (12) McMartin, C.; Bohacek, R. S. QXP: Powerful, Rapid Computer Algorithms for Structure-Based Drug Design. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 333–344.
- (13) Osterberg, F.; Morris, G. M.; Sanner, M. F.; Olson, A. J. Automated Docking to Multiple Target Structures: Incorporation of Protein Mobility and Structural Water Heterogeneity in AutoDock. *Proteins: Struct., Funct., Genet.* **2002**, *46*, 34–40.
- (14) Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. FlexE: Efficient Molecular Docking Considering Protein Structure Variations. *J. Mol. Biol.* **2001**, *308*, 377–395.
- (15) Luty, B. A.; Wasserman, Z. R.; Stouten, P. F. W.; Hodge, C. N.; Zacharias, M.; McCammon, J. A. A Molecular Mechanics/Grid Method for Evaluation of Ligand–Receptor Interactions. *J. Comput. Chem.* **1995**, *16*, 454–464.
- (16) Abagyan, R.; Argos, P. Optimal Protocol and Trajectory Visualization for Conformational Searches of Peptides and Proteins. *J. Mol. Biol.* **1992**, *225*, 519–532.
- (17) Smellie, A.; Teig, S. L.; Towbin, P. Poling: Promoting Conformational Variation. *J. Comput. Chem.* **1995**, *16*, 171–187.
- (18) *Tabu Search*; Kluwer Academic Publishers: Boston, 1998.
- (19) Ajay; Murcko, M. A.; Stouten, P. F. W. Recent Advances in the Prediction of Binding Free Energy. In *Practical Application of Computer-Aided Drug Design*; Charifson, P. S., Ed.; Marcel Dekker: New York, 1997; p 355.
- (20) Gohlke, H.; Klebe, G. Statistical Potentials and Scoring Functions Applied to Protein–Ligand Binding. *Curr. Opin. Struct. Biol.* **2001**, *11*, 231–235.
- (21) Böhm, H.-J.; Stahl, M. Rapid Empirical Scoring Functions in virtual Screening Applications. *Med. Chem. Res.* **1999**, *9*, 445–462.
- (22) Stahl, M. Modifications of the Scoring Function in FlexX for Virtual Screening Applications. *Persp. Drug Discov. Des.* **2000**, *20*, 83–98.
- (23) Muegge, I.; Martin, Y. C. A General and Fast Scoring Function for Protein–Ligand Interactions: A Simplified Potential Approach. *J. Med. Chem.* **1999**, *42*, 791–804.
- (24) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-Based Scoring Function to Predict Protein–Ligand Interactions. *J. Mol. Biol.* **2000**, *295*, 337–356.
- (25) Totrov, M.; Abagyan, R. Derivation of Sensitive Discrimination Potential for Virtual Ligand Screening. *Proc. Third Annual Intl. Conf. Comput. Mol. Biol.* **1999**, 312–320.
- (26) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem.* **1999**, *42*, 5100–5109.
- (27) Wang, R.; Wang, S. How Does Consensus Scoring Work for Virtual Library Screening? An Idealized Computer Experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422–1426.
- (28) Terp, G. E.; Johansen, B. N.; Christensen, I. T.; Jørgensen, F. S. New Concept for Multidimensional Selection of Ligand Conformations (MultiSelect) and Multidimensional Scoring (MultiScore) of Protein–Ligand Binding Affinities. *J. Med. Chem.* **2001**, *44*, 2333–2343.
- (29) Gohlke, H.; Klebe, G. Approaches to the Description and Prediction of the Binding Affinity of Small-Molecule Ligands to Macromolecular Receptors. *Angew. Chem., Int. Ed.* **2002**, *41*, 2644–2676.
- (30) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- (31) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **1977**, *112*, 535–542.
- (32) Abagyan, R. A.; Totrov, M. M.; Kuznetsov, D. A. ICM: A New Method for Structure Modeling and Design. *J. Comput. Chem.* **1994**, *15*, 488–506.
- (33) Totrov, M.; Abagyan, R. Protein–Ligand Docking as an Energy Optimization Problem. In *Drug-Receptor Thermodynamics: Introduction and Applications*; Raffa, R. B., Ed.; John Wiley & Sons: New York, 2001; pp 603–624.
- (34) Cerius2, version 4.7ccO. Accelrys Inc., San Diego, CA, 2001.
- (35) Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M. LigandFit: A Novel Method for the Shape-Directed Rapid Docking of Ligands to Protein Active Sites. *J. Mol. Graphics Modell.* **2003**, *21*, 289–307.
- (36) Böhm, H.-J. The computer program LUDI: A new method for the de novo design of enzyme inhibitors. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 61–78.
- (37) Böhm, H.-J. LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 593–606.
- (38) Klebe, G. The use of composite crystal-field environments in molecular recognition and the de-novo design of protein ligands. *J. Mol. Biol.* **1994**, *237*, 221–235.
- (39) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FlexX Incremental Construction Algorithm for Protein–Ligand Docking. *Proteins: Struct., Funct., Genet.* **1999**, *37*, 228–241.
- (40) Sybyl, version 6.8. Tripos Inc., St. Louis, MO, 2002.
- (41) Kuntz, I. D. Structure-based strategies for drug design and discovery. *Science* **1992**, *257*, 1078–1082.
- (42) AMSOL, version 6.8. University of Minnesota, Minneapolis, MN, 2002.
- (43) Omega, version 0.9.9. OpenEye, Santa Fe, NM, 2001.
- (44) Lorber, D. M.; Shoichet, B. K. Flexible ligand docking using conformational ensembles. *Prot. Sci.* **1998**, *7*, 938–950.
- (45) Shoichet, B. K.; Leach, A. R.; Kuntz, I. D. Ligand Solvation in Molecular Docking. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 4–16.
- (46) Jones, G.; Willett, P.; Glen, R. C. Molecular Recognition of Receptor Sites Using a Genetic Algorithm with a Description of Desolvation. *J. Mol. Biol.* **1995**, *245*, 43–53.
- (47) Jones, G.; Willett, P.; Glen, R. C. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 532–549.
- (48) Gray, N. S.; Wodicka, L.; Thunnissen, A. M. W. H.; Norman, T. C.; Kwon, S. J.; Espinoza, F. H.; Morgan, D. O.; Barnes, G.; LeClerc, S.; Meijer, L.; Kim, S.-H.; Lockhart, D. J.; Shultz, P. G. Exploiting chemical libraries, structure, and genomics in the search for kinase inhibitors. *Science* **1998**, *281*, 533–538.
- (49) Shewchuk, L.; Hassell, A.; Wisely, B.; Rocque, W.; Holmes, W.; Veal, J.; Kuyper, L. F. Binding Mode of the 4-Anilinoquinazoline Class of Protein Kinase Inhibitor: X-ray Crystallographic Studies of 4-Anilinoquinazolines Bound to Cyclin-Dependent Kinase 2 and P38 Kinase. *J. Med. Chem.* **2000**, *43*, 133–138.
- (50) Meijer, L.; Thunnissen, A. M.; White, A. W.; Garnier, M.; Nikolic, M.; Tsai, L. H.; Walter, J.; Cleverley, K. E.; Salinas, P. C.; Wu, Y. Z.; Biernat, J.; Mandelkow, E. M.; Kim, S.-H.; Pettit, G. R. Inhibition of Cyclin-Dependent Kinases, Gsk-3 β and Cdk1 by Hymenialdisine, a Marine Sponge Constituent. *Chem. Biol.* **2000**, *7*, 51–63.
- (51) Davis, S. T.; Benson, B. G.; Bramson, H. N.; Chapman, D. E.; Dickerson, S. H.; Dold, K. M.; Eberwein, D. J.; Edelstein, M.; Frye, S. V.; Gampe Jr., R. T.; Griffin, R. J.; Harris, P. A.; Hassell, A. M.; Holmes, W. D.; Hunter, R. N.; Knick, V. B.; Lackey, K.; Lovejoy, B. L. M.; Murray, D.; Parker, P.; Rocque, W. J.; Shewch, L. Prevention of Chemotherapy-Induced Alopecia in Rats by Cdk Inhibitors. *Science* **2001**, *291*, 134–137.
- (52) Hoessel, R.; LeClerc, S.; Endicott, J.; Noble, M.; Lawrie, A.; Tunnah, P.; Leost, M.; Damiens, E.; Marie, D.; Marko, D.; Niederberger, E.; Tang, W.; Eisenbrnd, G.; Meijer, L. Indirubin, the active constituent of a chinese antileukemia medicine, inhibits cyclin-dependent kinases. *Nature Cell Biol.* **1999**, *1*, 60–67.
- (53) Lawrie, A. M.; Noble, M. E. M.; Tunnah, P.; Brown, N. R.; Johnson, L. N.; Endicott, J. A. Protein kinase inhibition by staurosporine revealed in details of the molecular interaction with CDK2. *Nature Struct. Biol.* **1997**, *4*, 796–801.

- (54) Wallace, A. C.; Laskowski, R. A.; Thornton, J. M. LIGPLOT: a program to generate schematic diagrams of protein–ligand interactions. *Protein Eng.* **1995**, *8*, 127–134.
- (55) Prasad, T.; Prathima, M. N.; Chandra, N. Detection of hydrogen-bond signature patterns in protein families. *Bioinformatics* **2003**, *19*, 167–168.
- (56) Mestres, J.; Rohrer, D. C.; Maggiora, G. M. A molecular field-based similarity approach to pharmacophoric pattern recognition. *J. Mol. Graph. Model.* **1997**, *15*, 114–121.
- (57) Perkins, T. D.; Mills, J. E.; Dean, P. M. Molecular surface-volume and property matching to superpose flexible dissimilar molecules. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 479–490.
- (58) Parretti, M. A.; Kroemer, R. T.; Rothman, J. H.; Richards, W. G. Alignment of Molecules by the Monte Carlo Optimization of Molecular Similarity Indices. *J. Comput. Chem.* **1997**, *18*, 1344–1353.

CI049970M