Title

An improved relaxed complex scheme for receptor flexibility in computer-aided drug design

Journal

Journal of Computer-Aided Molecular Design, 22(9)

ISSN

0928-2866

Authors

Amaro, Rommie E
Baron, Riccardo
McCammon, J Andrew

Publication Date

2008-09-01

DOI

10.1007/s10822-007-9159-2

Peer reviewed

J Comput Aided Mol Des (2008) 22:693-705
DOI 10.1007/s10822-007-9159-2

An improved relaxed complex scheme for receptor flexibility in computer-aided drug design

Rommie E. Amaro · Riccardo Baron · J. Andrew McCammon

Received: 12 July 2007 / Accepted: 21 November 2007 / Published online: 15 January 2008

Abstract The interactions among associating (macro)molecules are dynamic, which adds to the complexity of molecular recognition. While ligand flexibility is well accounted for in computational drug design, the effective inclusion of receptor flexibility remains an important challenge. The relaxed complex scheme (RCS) is a promising computational methodology that combines the advantages of docking algorithms with dynamic structural information provided by molecular dynamics (MD) simulations, therefore explicitly accounting for the flexibility of both the receptor and the docked ligands. Here, we briefly review the RCS and discuss new extensions and improvements of this methodology in the context of ligand binding to two example targets: kinetoplastid RNA editing ligase 1 and the W191G cavity mutant of cytochrome *c* peroxidase. The RCS improvements include its extension to virtual screening, more rigorous characterization of local and global binding effects, and methods to improve its computational efficiency by reducing the receptor ensemble to a representative set of configurations. The choice of receptor ensemble, its influence on the predictive power of RCS, and the current limitations for an accurate treatment of the solvent contributions are also briefly discussed. Finally, we outline potential methodological improvements that we anticipate will assist future development.

Keywords Clustering · Docking · Ensemble-based docking · Kinetoplastid RNA editing ligase 1 · Molecular dynamics · Non-redundant ensemble · Protein–ligand binding · Relaxed complex method · Representative ensemble · Virtual screening · W191G cytochrome *c* peroxidase

Rommie E. Amaro and Riccardo Baron contributed equally to this work.

R. E. Amaro (✉) · R. Baron (✉) · J. A. McCammon
Department of Chemistry and Biochemistry, University of California at San Diego, La Jolla, CA 92093-0365, USA
e-mail: ramaro@mccammon.ucsd.edu

R. Baron
e-mail: rbaron@mccammon.ucsd.edu

R. E. Amaro · R. Baron · J. A. McCammon
Center for Theoretical Biological Physics, University of California at San Diego, La Jolla, CA 92039-0365, USA

J. A. McCammon
Department of Pharmacology, University of California at San Diego, La Jolla, CA 92093-0365, USA

J. A. McCammon
Howard Hughes Medical Institute, University of California at San Diego, La Jolla, CA 92093-0365, USA

Abbreviations

GA	Genetic algorithm
KREL1	Kinetoplastid RNA editing ligase 1
MD	Molecular dynamics
RCS	Relaxed complex scheme
RMSD	Root-mean-square deviation
W191G	W191G cavity mutant of cytochrome *c* peroxidase

Introduction

A full understanding of molecular recognition presents a problem of intense interest to the field of computer-aided drug design and molecular sciences in general. The interactions between ligand molecules and their corresponding receptors are dynamic and complex. Techniques that best

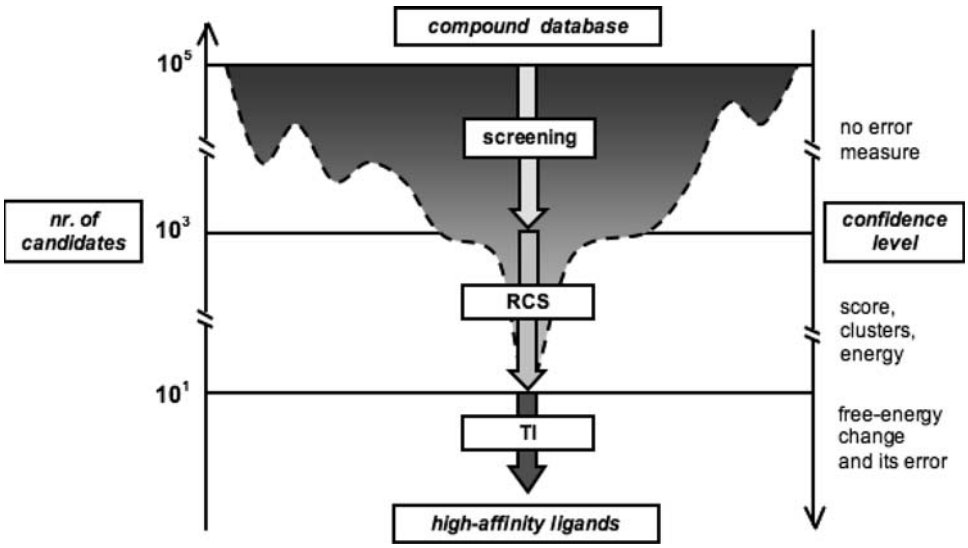

Fig. 1 The problem: how to distill a few good binders and characterize their binding propensity out of a vast database of compounds

address these issues must account for the conformational flexibility of both the ligand and the receptor and do so in an accurate and efficient manner. While the ability to explore ligand flexibility is well established, computer-aided drug design methodologies have only recently begun to take receptor flexibility into account when searching for and optimizing functional inhibitors. Since it is widely accepted that ligands may bind to receptor conformations that occur infrequently in the receptor's dynamics, and that the local motions of active site residues can drastically alter the binding and specificity of ligands to their target, the ability to efficiently sample these rare dynamics and furthermore, to incorporate the resulting conformations into the drug design protocol, remains an important challenge (reviewed in references [1–5]).

A closely related challenge is the development of effective methods to predict the binding propensity for series of compounds or flexible peptides to a given receptor [6–11]. Approaches based on scoring functions or compound libraries require large amounts of data to be available a priori. These methods include computational virtual screening [12–18], docking [18–23], and similarity searching [24, 25]. More advanced treatments of receptor–ligand binding can be achieved using molecular dynamics (MD) simulations. Free energy changes can then be estimated based on coupling-parameter approaches, such as thermodynamic integration (TI) and free energy perturbation (FEP) [11, 26–36], which describe a higher physical complexity of the binding process and include an extensive sampling of receptor, ligand, and solvent phase spaces. Importantly, the latter methods provide not only binding free energy estimates, but also a reliable measure of their accuracy. Yet, they are typically too computationally expensive to be applied to extensive sets ($\sim 10^5$) of compounds, which is the usual scenario for newly discovered biological targets.

Hybrid methods, which are faster but more approximate, have been developed with the aim of reducing a large initial set of potential binders ($\sim 10^3$ or more), to a reduced set of promising molecules ($\sim 10^1$), for which the binding properties can then be investigated in a second phase that uses more rigorous methods to predict binding free energies (Fig. 1). Examples of these hybrid methods are linear interaction energy (LIE) [37–39], single-step perturbation [40–45], docking to MD structures [44, 46], docking to relevant normal modes structures [47–49], induced-fit docking (IFD) [50], the dynamic pharmacophore model [51, 52] and the relaxed complex scheme (RCS) [53–55]. Alternatively, the receptor ensemble can be gathered from a collection of independent X-ray or NMR experimentally derived structures [56, 57]. These hybrid approaches encompass different levels of accuracy, predictive power, and level of a priori knowledge required. For example, LIE and single-step perturbation can be considered "non-empirical" in that they are derived from free energy type approaches and based on more complex physical models of the binding process. Yet, in practice they need precise information on the location of the binding site to be available a priori. Conversely, docking-based techniques are appealing for screening purposes because they do not necessarily require any information on the location of the binding site, and can therefore be employed to predict binding site locations. In principle, these docking-based approaches should also be more easily extendable to biologically relevant systems with increasing size and complexity, such as protein–nucleotide and protein–protein association. However, they cannot supply accurate estimates of free energy changes upon binding.

The RCS is a promising computational methodology that combines the advantages of docking algorithms with dynamic structural information provided by MD

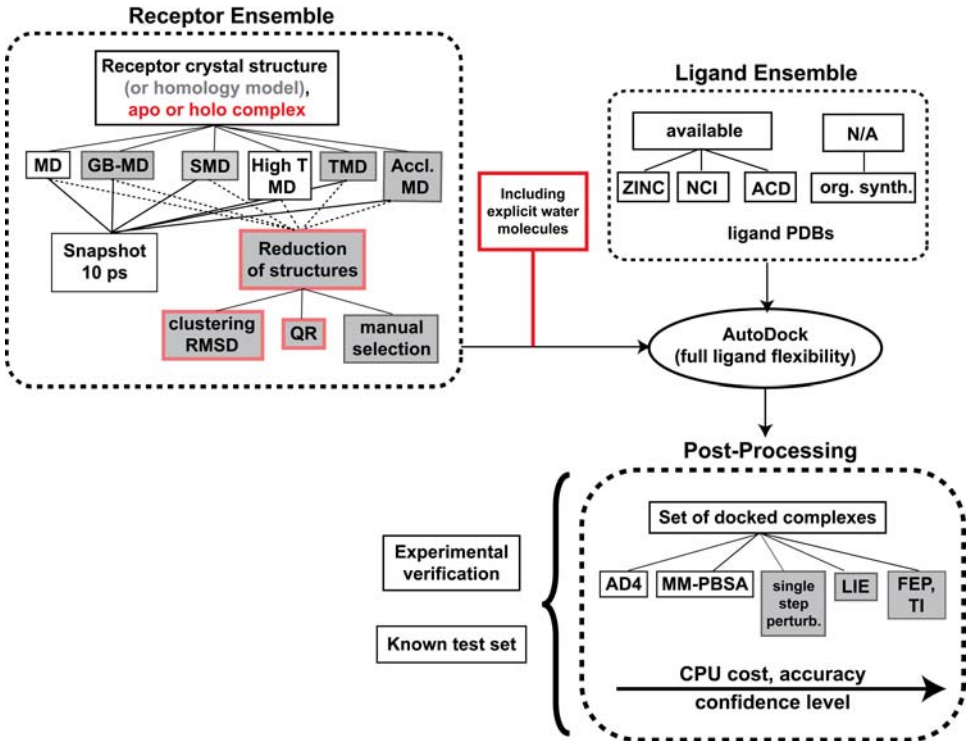

Fig. 2 An overview of the RCS. Improvements to the RCS are shown in gray background and those specifically presented in this paper are outlined in red. In the "Receptor Ensemble" box (top left), the structures can be generated with classical MD, or a variety of simulation techniques could be considered in order to enhance the sampling of the receptor configurational space, including: Generalized-Born MD (GB-MD), steered MD (SMD), high temperature MD (High T MD), targeted MD (TMD), and accelerated MD (Accl. MD). In the "Ligand Ensemble" box (top right), commercially or publicly available ligands can be found in the Zinc Is Not Commercial (ZINC), National Cancer Institute (NCI), and Available Chemicals Database (ACD), among others. AutoDock is then used to dock the ligand database into the receptor ensemble. In the "Post-Processing" stage, the docked complexes can be rescored or reevaluated using more rigorous protocols than the AutoDock version 4.0 scoring function (AD4), including molecular-mechanics Poisson–Boltzmann surface area (MM-PBSA), single step perturbation, LIE, and FEP or TI techniques

simulations, explicitly accounting for the flexibility of both the receptor and docked ligands. This procedure is appealing as a large variety of conformational changes may characterize ligand binding processes of biochemical and medical interest and, more generally, molecular recognition. The RCS has been developed in combination with various MD software packages and AutoDock for the ligand docking. Although other docking programs can be considered, all RCS applications to date have employed AutoDock, a widely distributed and tested docking program that has been shown to be successful in a variety of docking studies [20–22]. The RCS was first applied to the FKBP binding protein [53] and tested using improved rescoring functions based on MM-PBSA models [54]. Applications of the RCS identified a novel-binding trench in HIV integrase [55]. In this work, we sketch the philosophy underlying the RCS, describe new improvements, and present recent applications to exemplify the type of problems that can be tackled with this computational scheme.

Materials and methods

Relaxed complex scheme: short overview

In the typical RCS (Fig. 2), all-atom MD simulations are carried out for the target biomolecule of interest, with a substrate or inhibitor bound in the active site, starting from the crystal structure with a bound ligand (i.e., the holo complex). Typical simulation lengths range from 2 ns to tens of ns, and snapshots of the biomolecule are extracted at a predetermined time interval (e.g., every 10 ps). RCS calculations based on explicit solvent MD simulations of two different systems are presented in this work. First, the kinetoplastid RNA editing ligase 1 (KREL1), which uses the NAMD2.6 MD software [58] (freely available at http://www.ks.uiuc.edu/Research/namd/) with the Charmm27 force field [59]. Second, the W191G cavity mutant of cytochrome *c* peroxidase (W191G), based on simulations performed with the GROMOS05 software for biomolecular simulation [60] (available at http://www.igc.ethz.ch/gromos/) using the 45A4

parameter set [61] of the GROMOS force field [62]. Details of the MD for each system are described in References [63] and [64], respectively. The resulting set of structures, generated with a physically based MD force field, represents the receptor ensemble and can be conceptually thought of as a set of structures defining approximately its thermodynamic equilibrium state in solution. This receptor ensemble is subsequently used in the docking experiments, in which a reduced set of small molecules are docked into the active site and the corresponding binding affinities are evaluated.

AutoDock is used to carry out the docking experiments and full ligand flexibility is employed. One of the major advantages of AutoDock is its use of a hybrid genetic algorithm (GA) to perform an efficient and effective global search for the ligand [65]. Genetic algorithms are optimization schemes that use the language of natural genetics and evolution, and in the case of AutoDock, the optimization problem is molecular docking between a ligand and a receptor. Typically, the receptor is fixed and the translation, orientation, and conformation of the ligand are explored. Genetically derived terms such as the AutoDock "chromosome," which describes the ligand state, define its "genotype" and the atomic coordinates of the ligand, which describes its "phenotype," undergo genetic events such as "selection, crossover, and mutation" during the optimization procedure.

The AutoDock chromosome consists of a string of real-valued genes containing three cartesian coordinates for ligand translation, four variables defining a quaternion that specifies the ligand orientation, and one real-value for each ligand torsion [65]. The global search is carried out on the genotype level and performed with the GA, which allows selection, crossover, and mutation. The ligand-receptor fitness is evaluated based on a semi-empirical scoring function including an empirical estimate for the ligand configurational entropy [66]. The global search is followed by an adaptive-stepping local search that performs energy minimizations on the atomic coordinates. Afterwards, the optimized phenotype is fed back to the genotype, in accordance with "Lamarckian" genetics, from which the algorithm derives its name. Ultimately, solutions better suited to specific interactions have a better score, therefore reproduce and persist, whereas poorer suited ones die off.

The re-docking of the ligands across the ensemble of receptor structures results in a range of predicted binding affinities for each ligand, based on the AutoDock scoring function. The resulting "binding spectrum" for each ligand is then used to reorder the ligands and better predict relative affinity. Various post-processing options can be considered beyond the initial affinity estimate provided by AutoDock, including the application of MM-PBSA, single-step perturbation, LIE, FEP, or TI (Fig. 2). Although more rigorous free energy estimates increase the confidence in the predicted binding energies, they can be prohibitively computationally expensive.

Improved relaxed complex scheme

A first set of improvements involves the docking algorithm itself as implemented in AutoDock version 4.0: (i) a more complete thermodynamic cycle, where the unbound (gas phase) ligand enthalpy is computed, (ii) an improved desolvation term that accounts for a larger number of atom types than in the previous versions, and (iii) a charge model that allows fast calculation of the charge distribution [67] and compatibility of partial charges between the ligand and the receptor structures [66]. The studies presented here employed this new and improved version of AutoDock (freely available at http://www.autodock.scripps.edu/).

A second set of improvements involves the RCS methodology itself, as described in the following, based on recent applications. The first extension to the RCS we present here is the application of the method for virtual screening, which involves an essential enzyme for the protozoan parasite *Trypanosoma brucei*. The discovery of several new inhibitors was the result of a streamlined RCS method, providing a concrete example of its success when trying to discover new inhibitors from a large database of compounds [68]. The second methodological advancement for RCS involves accounting for both local-induced and global effects of ligand binding. This is shown with the well-characterized binding of a set of heterocyclic cation ligands to the W191G cavity mutant of cytochrome *c* peroxidase [69]. The third improvement attempts to define two general algorithms to reduce the number of MD trajectory snapshots for the docking experiments, which increases computational efficiency by orders of magnitude without decreasing its accuracy. First, the KREL1 application uses the QR factorization method available in the MultiSeq plugin in VMD [70]. Second, the W191G cavity mutant of cytochrome *c* peroxidase (W191G) uses an atom-positional root-mean-square deviation (RMSD) clustering algorithm [71] as implemented in the rmsdmat2 and cluster2 programs of the GROMOS++ analysis software [60]. Last, we discuss the importance and the difficulties of including explicit water molecules within the binding sites in the RCS docking experiments.

New applications

RCS as a tool for enhanced virtual screening

Given a novel protein target, the goal of identifying a new set of potential inhibitors with drug-like properties can be achieved using virtual screening type approaches. Typically

these large-scale virtual screens are carried out by evaluating the predicted affinities of thousands of molecules against a single static crystal structure [15, 72]. Here we report on the success of porting the RCS into a virtual screen type application in the search for inhibitors against an essential kinetoplastid RNA editing ligase 1 (KREL1) in *T. brucei*, the parasite responsible for the devastating tropical disease African sleeping sickness [68]. KREL1 is required for survival of both the insect and bloodstream forms of the parasite [73], and it is a particularly attractive drug target as there are no known human homologues. The high-resolution crystal structure [74] provides an excellent platform for computer-aided drug design as well as for MD simulations and the RCS application.

The KREL1 crystal structure revealed a deep active site pocket with several water molecules coordinated to the ATP substrate and the protein. Two 20 ns simulations in explicit solvent were carried out with KREL1, both with and without the bound ATP in order to generate the receptor ensembles [63]. A screen of the crystal structure against the NCI diversity set (containing 1,900 compounds) using AutoDock version 4.0 was performed, and the top twenty-five compounds that obeyed most of Lipinski's "rules of 5" [75] were selected for application of the RCS method. The top 25 ligands were then re-docked into the full receptor ensemble as well as a reduced representative set (discussed in further detail below) and these compounds were then re-ranked based on their average binding energy of the most populated cluster.

The results of the RCS virtual screen with KREL1 are particularly promising. Several new inhibitors have been identified with the RCS and an in vitro inhibition assay of the first step in the binding reaction, the adenylation step, was used to verify the computational predictions. These experiments confirmed two of the eight tested compounds found in the initial screen were inhibitory [68]. Importantly, the RCS method resulted in a reordering of the twenty-five compounds that identified inhibitors that would not have otherwise been tested, based on their rank from the static crystal structure screen. Specifically, the best hit as experimentally verified was initially ranked fifteenth, and after RCS reordering became first. In the case of limited resources and low-throughput experimental procedures, where only a handful of the best compounds identified in the screen could be experimentally tested, the application of the RCS method provided a measurable and important enrichment of the initial ranked set.

Accounting for induced fit and the global effects of ligand binding

In nature, a great number of protein–ligand recognition processes are only possible when accompanied by local (i.e., the reorganization of residues upon induced-fit binding) or global (i.e., larger scale conformational changes occurring also in remote structural elements of the receptor upon binding) effects. Our current structural knowledge of biomolecular association phenomena is predominantly based on X-ray crystallography ensemble-averaged structures. Although these experiments provide critical binding information, they typically capture only one state involved in the binding process, which may be a dominant configuration, but not necessarily exclusive. Dynamic information at the atomistic level, as provided by MD simulations, is of fundamental importance and may reveal binding modes and relevant biophysical information otherwise inaccessible to standard experimental techniques.

A relevant example of the importance of predicting receptor-flexibility effects resulted from the application of RCS to HIV integrase. MD simulations of the integrase protein bound to a known inhibitor revealed a new cavity adjacent to the active site [55]. RCS docking of ligands into this newly discovered pocket indicated favorable binding of ligands to this area. This new structural insight was exploited in the development of raltegravir (MK-0518), the first of a new class of antiretroviral agents active against the enzyme integrase that has recently been approved by the FDA [76].

As the binding propensity defining a given molecular association reflects the relative stabilities of the possible conformations of the receptor, effective drug-design protocols should be based on a distribution of receptor conformations. In this respect, RCS has the advantage of requiring the generation of only one MD ensemble per receptor macro-state (e.g., the open or closed state of a loop gating the binding pocket). This has recently been systematically investigated in the case of the W191G cavity mutant of cytochrome *c* peroxidase by analyzing the docking of small ligands into alternative ensembles of receptor conformations [69].

The binding of heterocyclic cation ligands into the W191G engineered cavity has been characterized experimentally [77–81]. Mutation of this key tryptophan in the active site creates a ligand-binding cavity and also appears to increase local flexibility, which opens a loop-gated pathway for ligands to reach the buried cavity. Recently, MD simulations suggested the importance of induced-fit effects in the W191G cavity for binding of 2-amino-5-methylthiazole (2a5mt) [64]. X-ray crystallography experiments have elucidated the structures of several ligand-protein complexes, including those for which the loop rearrangement is more pronounced and causes a shift between the closed- and open-gate structural ensembles. Benzimidazole (bzi) was suggested to produce a full opening of the cavity [78]. MD simulations starting from different initial configurations characterized the

conformational sampling and dominant configurations of the closed and open alternate states (Fig. 3a).

RCS calculations were performed on different gating-loop and binding states: the closed-gate apo, the closed-gate holo (i.e., the complex with the best binder), and the open-gate apo structure, allowing the investigation of the correlation between each compound's binding affinity and the closed/open state of the gating loop (Fig. 3b) [69]. Additional in silico experiments evaluated the benefits of using non-standard MD trajectories to enhance the conformational sampling (e.g., simulations at high temperature using atom-positional restraining potentials) or simulate an unphysical generalized-ligand interaction encompassing the characteristics of all potential binders [69]. In the case of 2a5mt, the optimal binding spectrum occurs when docked into the receptor conformations from the holo ensemble. Although both the holo and apo receptor ensembles generate ligand-binding poses (i.e., the geometry of a docked ligand into the binding site) that are similar to those determined experimentally, the 2a5mt binding affinities are closer to the experimental results for the holo ensemble (Fig. 3c). This illustrates that the holo ensemble is the best choice to perform RCS calculations for the 2a5mt ligand, and suggests the same is likely the case for other ligands with similar chemical and electrostatic properties [69].

A different picture emerges when bzi binds to the same cavity. In this case, the best agreement between RCS affinities and the experimental free energies is found when using the apo-open receptor ensemble. This agrees with the experimental observation that bzi shifts the propensity of the gating-loop configurations towards the open-gate state.

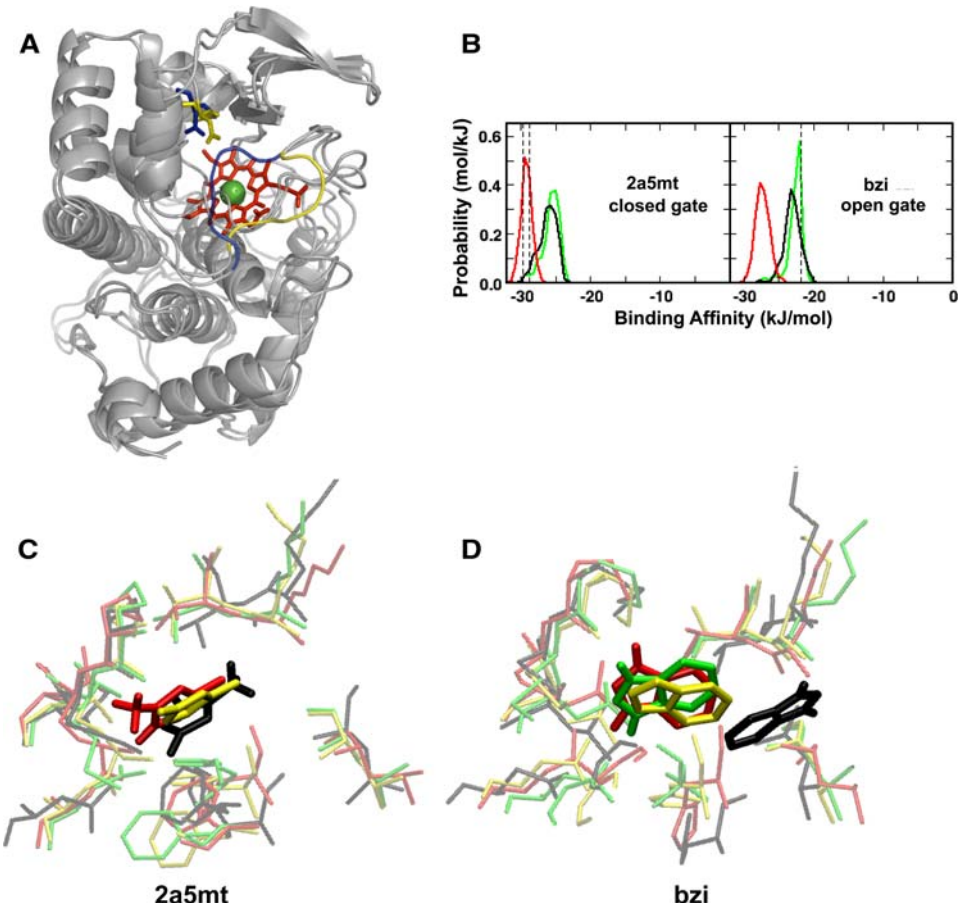

Fig. 3 (**a**) The W191G cavity mutant of cytochrome *c* peroxidase and its two dominant configurations extracted using an RMSD conformational clustering analysis for the gating-loop and MD simulations of the separate states. The closed (blue) and open (yellow) gate states are highlighted, together with Asp 235, the residue determining the orientation of the binders in the cavity. The heme cofactor is shown in red. (**b**) Binding propensities of the best binder (2a5mt) and of the binder suggested to induce the full opening of the gating loop (bzi) are shown [69]. For each of the two conformational states of the gating loop the probability distributions of the binding affinities from RCS calculations are shown as based on the apo (black), holo (red), and apo open-gate (green) receptor ensemble simulations. The dashed-vertical lines correspond to the experimental free energies of binding. Docking poses for 2a5mt (**c**) and bzi (**d**) are displayed from corresponding crystal structures (yellow) and the RCS calculations based on MD simulations of the apo (black), holo (red), and apo open-gate (green) receptors

Again, the ligand-binding poses (i.e., the relative orientation of the docked ligand into the W191G artificial cavity) are very similar between the RCS method and the crystallographic complexes (Fig. 3c). Although a false negative is found when docking bzi to the ensemble of apo receptor structures, bzi binds favorably and with a binding mode similar to experiment when using the closed-gate holo ensemble (Fig 3d). These promising results suggest that it may be possible to capture different binding propensities depending on both local and global receptor rearrangements upon binding.

Effective reduction of the receptor ensemble

In the original RCS, the computational docking experiments were carried out using snapshots extracted at equal time intervals from the MD trajectories. As the simulations are carried out for several nanoseconds, this typically sums up to $\sim 10^4$ to 10^5 receptor structures, many of which may be conformationally redundant. Two recent studies have investigated alternative methods to distill the structural information to a reduced, yet meaningful set.

A novel method to distill the ensemble of structures to a non-redundant set is the so-called "QR factorization" method. This technique was originally developed to remove inherent bias in structure databases and distill, from a vast quantity of redundant information, a minimal basis set of protein structures that accurately spans the evolutionary phase space of a particular protein [82]. It has also been applied to an ensemble of NMR structures in order to determine a small, representative subset of structures from a larger experimental dataset [83] and to create non-redundant sequence alignments [84]. This technique has most recently been incorporated into the RCS, where a multiple structural alignment of the receptor ensemble is performed with STAMP [85]. This alignment algorithm operates progressively: all possible pair-wise alignments are computed, followed by a hierarchical clustering analysis based on a structural similarity measure to build the multiple structural alignment. The measure of structural similarity applied here is Q_H, which essentially measures the distance between all pairs of C^α atoms among all aligned structures. Although the development of Q_H was motivated by the need to include gaps in order to build a similarity measure for more distantly related proteins, in the case of aligning the receptor ensemble, the gap term is unnecessary as structures of the same protein are aligned. The structural alignment is stored in a multidimensional matrix of dimension $m_{\text{aln}} \times n_{\text{receptor structures}} \times d$, where d encodes the rotated C^α atoms coordinates. In this matrix, each receptor structure is represented in a column and the rows represent the multiple alignment. Finally, a multi-dimensional QR factorization algorithm is applied to the encoded alignment of receptor structures, which results in a reordering of the structures based on increasing linear dependence. This reordering subsequently allows the construction of non-redundant sets of structures at some user-defined cutoff, representing a certain dynamical configuration space.

The application of this method to the receptor ensemble of RNA editing ligase 1 resulted in the initial set of 400 structures (extracted every 50 ps from a 20 ns simulation) being reduced to 33, with essentially no loss of binding spectrum information (Fig. 4a). When docking a large set of ligands, as is required in a virtual-screen type application, the reduction of receptor structures for the computational dockings can make a significant difference in computational cost. For example, for RNA editing ligase, the number of dockings was reduced from 11,200 to 924, resulting in a 90% reduction of computational cost [68].

An alternative method is clustering based on a matrix of all pair-wise RMSD of the aligned structures in the receptor ensemble. If the binding region of the receptor is known a priori the clustering algorithm can focus on this particular subset of residues that constitute the binding site. The employed algorithm was originally developed to capture the dominant configurations of an ensemble of structures for flexible peptides [71] and its application has been further extended to flexible molecules [86, 87] and protein surface loops [64]. After removing overall rotation and translations, the atoms of the binding region are superimposed using their C^α atoms coordinates. A matrix that contains all pair-wise RMSD values among all the structures in the trajectory is created. Next, the matrix is divided into batches corresponding to similar structures using the RMSD values, with a clustering algorithm [71] and a user-defined cutoff. This clustering allows docking trials to be performed on a reduced number of significant conformations, while retaining the dominant characteristics of the entire spectrum of binding modes.

In the case of the W191G cavity mutant of cytochrome *c* peroxidase, re-docking into the entire ensemble of structures would require docking to 10^4 snapshots. However, when these trajectory snapshots were clustered into groups of similar configurations with a RMSD similarity criterion of 0.1 nm for the cavity residues, the resulting two most dominant clusters of trajectory structures represented 36 and 16%, 81 and 10%, 48 and 23% of the structures for the apo, holo, and apo-open ensembles, respectively (Fig. 4b). This RMSD clustering resulted in a 99% reduction of computational cost for the RCS docking stage [69].

In addition to improving the computational efficiency of the RCS, clustering analyses can also supply useful information about the flexibility of the receptor, by analyzing

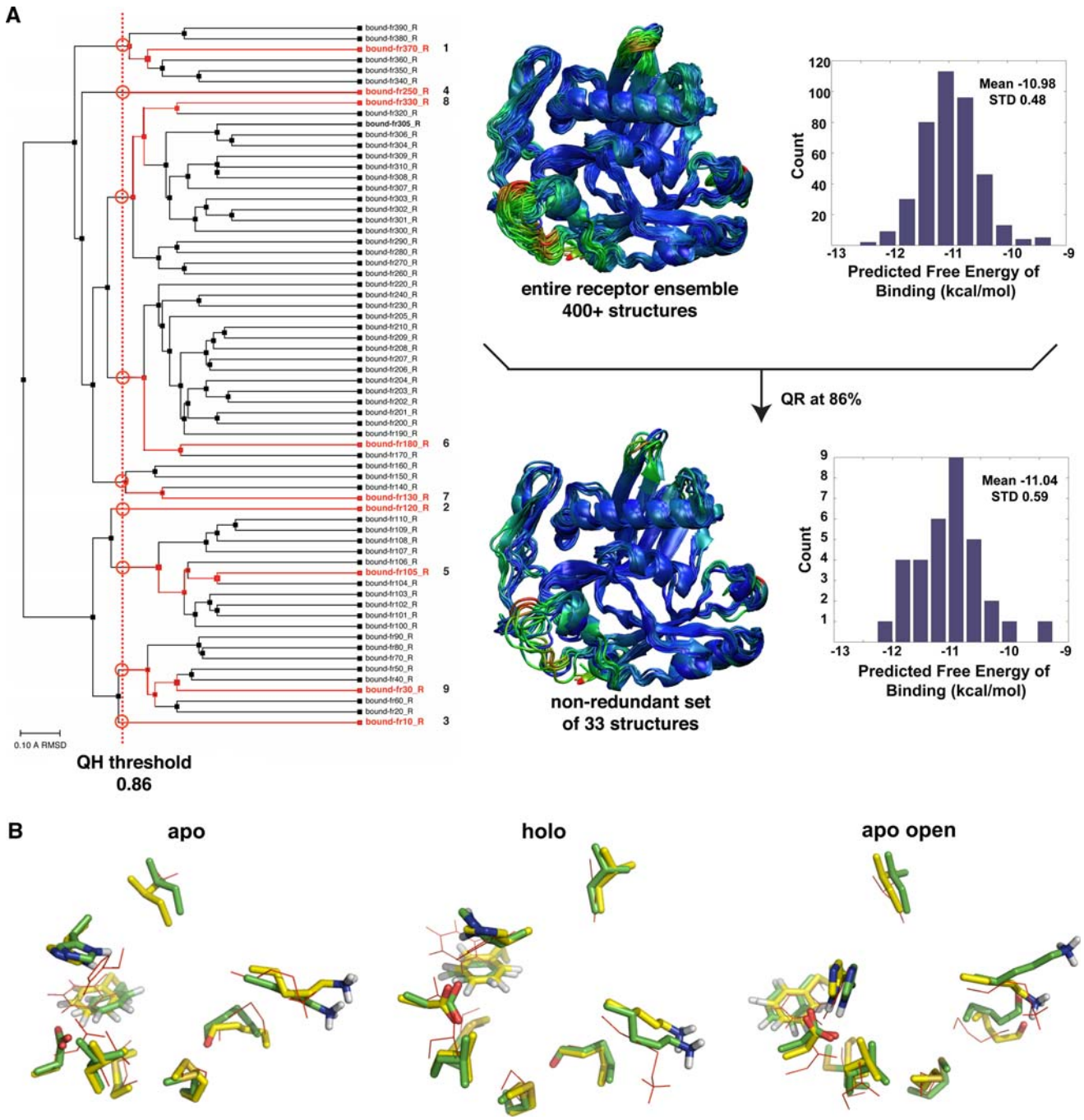

Fig. 4 Reducing redundancy in the receptor ensemble. (**a**) Left panel: Multidimensional QR factorization of KREL1 determines the distance relationship among all pairs of proteins (according to RMSD) and then reorders them based on increasing linear dependence, allowing the distillation of a reduced, representative set of structures for docking. At any particular Q_H threshold (indicated by red dotted line at Q_H 0.86), at each point of intersection of a branch, the most linearly independent structure is chosen from the group to the right of the dotted line (each red open circle drawn at the branch intersection indicates the choice of one structure to represent all structures to the right of the node). For clarity, the structure tree shown here is reduced (not all KREL1 structures are shown). Right panel: the initial (top) set of structures with the corresponding binding spectrum and the reduced set (bottom) is shown. The similarity between the full and reduced binding spectrums indicates that there is virtually no loss of information. (**b**) Dominant configurations of the W191G cavity region as extracted from RMSD conformational clustering. For each separate MD ensemble the corresponding reference crystal configuration is displayed (red thin lines) superimposed on the central member structures of the first (yellow licorice) and second (green licorice) most populated clusters.

the number of structures representing a certain Q_H or RMSD threshold cutoff (in the QR factorization method) or the cluster population versus the number of clusters populated (in the RMSD clustering method). This type of information gives quantitative insight about the local and global flexibility of the receptor. The computational gain due to these types of clustering schemes seems particularly useful when screening large compound databases.

Choice of MD receptor ensemble and RCS predictive power

One of the major challenges for hybrid docking techniques is the possibility to screen large compound databases and extract potential binders based on very limited a priori knowledge of the binding process itself. In this context, the W191G cytochrome *c* peroxidase system is used as a platform to investigate how the choice of MD receptor ensemble for RCS calculations affects the predictive power of this methodology [69]. To reflect the different amounts of knowledge that may be available on the binding process, different typical scenarios in drug discovery were considered, including cases in which: (i) no information is available on the location of the binding site, (ii) X-ray structures of the protein–ligand complexes and knowledge on potential binders is not available, and (iii) the X-ray structures known for the protein–ligand complex do not define unique ligand-binding poses. Corresponding to the above scenarios: (i) the RCS technique using the holo-receptor ensemble finds true positives using a docking grid that encompasses the entire W191G cytochrome *c* peroxidase structure; (ii) the number of true positives and true negatives can be significantly increased versus the number of false positives and false negatives by employing an MD receptor ensemble containing a generalized type of unphysical ligand that reflects the main structural properties of the compounds in the database, when compared to equivalent docking calculations performed on the apo MD receptor ensemble; and (iii) the multiple binding orientations characterizing the true positives do not necessarily correspond to non-accurate docking results when compared to the raw electron density data from X-ray crystallography. The quality of the binding poses can be judged by a combined evaluation of (i) the distribution of the docked-ligand cluster populations versus the cluster number, (ii) RMSD from the corresponding experimental complex after superimposing the structures as described above, and (iii) the comparison of the different poses for a same ligand. These results open new possibilities for enhancing the predictive power of RCS calculations in so-called "blind" test cases (when information is missing concerning the binding process), and they also suggest that the best possible choice of MD ensemble may depend on the amount of knowledge available case by case. Additional scenarios can be considered as well, for example including homology-modeling type of approaches to generate the initial receptor structural configuration.

Accurate description of solvent contributions

An effective representation of the solvent during the docking trials is a major factor limiting the accuracy of docking calculations. Until recently, RCS calculations have been performed using only the receptor structure and ligand, even when the MD trajectory structures were generated in combination with explicit water models. Accounting for the specific role of conserved water molecules is highly relevant as they may perturb the flexibility of a bound ligand, significantly alter the electrostatic environment experienced by a small molecule, and even occlude potential areas of binding. The explicit inclusion of these contributions will certainly improve the accuracy of the ligand-binding description, similarly to what recently reported for protein–protein docking [88]. Here, we present two new applications of the RCS that tested docking of the ligands into MD generated receptor structures with and without cavity water molecules. What emerges from both examples is the important (thermo)dynamic role of specific waters for the binding process.

In the case of the RNA editing ligase, the KREL1 crystal structure suggested three buried water molecules in the deep end of the ATP binding pocket. Explicitly solvated MD simulations of the holo complex (i.e., ATP bound) allowed us to predict the dynamics of the crystal water molecules. The simulations indicated that these water molecules have different exchange rates, and that one of the water molecules in particular, the one directly interacting with both the protein and ATP, persists in its original location for the duration of the 20 ns simulation [63]. In terms of the receptor structure, this single coordinated water molecule acts as a structural scaffold that prevents the localized collapse of surrounding residues while interacting with the bound ATP. By extracting the water molecules from the structure before ligand docking, an additional small cavity was open to the ligand. The RCS dockings were performed both with and without the three conserved water molecules. Interestingly, the best inhibitors were identified when the water molecules were not included in the RCS dockings. The predicted docking poses and binding affinities differ significantly depending on whether the three explicit conserved water molecules are included in the dockings or not (data not shown). Importantly, at least two of the experimentally verified inhibitors were predicted to substitute a functional group into the

location where one of the crystal water molecules was located (Fig. 5a) [68].

In the W191G cytochrome *c* peroxidase application, several docking calculations were performed to investigate the influence of the crystallographic water sites on AutoDock binding affinities and ligand poses. The tests were performed with rigid-protein docking calculations on different receptor models, alternatively including or excluding water site 308, which was suggested to be a highly conserved location for a bound water based on X-ray structures for a large group of compounds [81]. The results from the docking calculations are in agreement with what was previously suggested by similar tests based on AutoDock version 3.0, which showed that the introduction of even a single water molecule can significantly perturb the binding propensity of most of the ligands [89].

While the effect on the predicted binding affinities and poses is at variance with the specific ligand, a systematic consequence of the introduction of one water molecule into the W191G cavity is the significant reduction of the configurational space available to the ligand during the docking trials. MD simulations of the W191G cavity predict the location of the highly favorable water sites (within X-ray crystallography resolution and refinement assumptions) compared with the available experimental data (Fig. 5b). Additionally, MD simulations reveal a larger number of favorable water sites, and allow the description of the dynamic behavior of the solvent, including the swapping of water molecules between highly favorable regions [64]. Based on these observations we suggest that the significant effects of including explicit water in the static dockings may be an artificial consequence of the introduction of bound (static) water molecules in the cavity, whereas locally disordered (dynamically swapping among the favorable sites) water molecules should be considered instead. We note that an accurate sampling of receptor, ligand, and solvent phase spaces is, in principle, reached by more expensive free energy calculations [11, 26–36].

Future methodological improvements

The development of computational tools for computer-aided drug design depends on the critical compromise between accuracy and computational costs. Ideally, the most reliable prediction of molecular affinity can be obtained through rigorous free energy calculations of the ligand-binding process [11, 26–36]. In practice, however, the CPU time typically required to perform such free energy calculations on a few candidates (bottom of the funnel diagram in Fig. 1) is comparable to that involved in rough geometrical recognition over a pool of molecules more than five orders of magnitude larger (top of the funnel diagram in Fig. 1). Although the theory and methods are well established for calculating free energy in practice, they are still prohibitively expensive to be employed in high-throughput screening of drug-like databases. The future development of hybrid techniques, and especially the RCS, is therefore twofold.

First, it will certainly involve the refinement of the underlying physical models describing ligand-binding

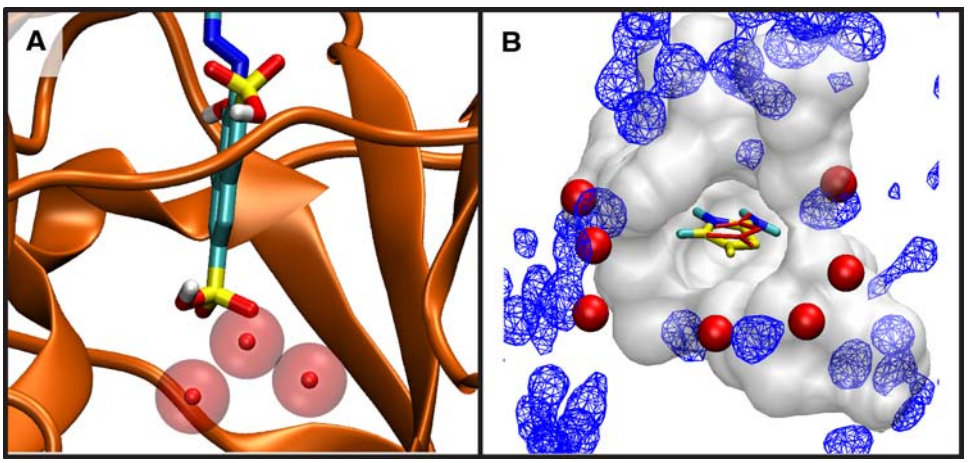

Fig. 5 Solvent contributions in protein–ligand binding. (**a**) The KREL1 active site with one of the newly discovered inhibitors in a predicted docked conformation. KREL1 is shown in orange cartoon, with the novel inhibitor shown docked in the active site (licorice, atom type colors). The three crystallographic water sites (not included in the docking calculation) are shown in licorice with their van der Waals surface in transparent. Note that the sulfonic acid group of the inhibitor replaces the location of a crystal water molecule. (**b**) For the W191G cavity (gray surface), the crystallographic water sites (solid red spheres; diameter corresponding to X-ray resolution) are compared to the highly favorable average density regions of water molecules in the MD simulations (blue wireframe isosurfaces), for the best binder 2a5mt (yellow licorice) and from the 1AEN crystal structure (red licorice)

(thermo)dynamics in increased detail, especially during the docking stage. Although the results presented here include the unbound (gas phase) ligand enthalpy term at the docking stage [66], a complete description of the thermodynamic cycle of binding is still far from being explicitly treated in RCS. Previous studies investigated the benefits of rescoring the docked complexes using a more accurate (implicit solvent) description of the solvent contributions [54]. More recently, the role of ligand entropy in the refinement of protein–ligand docking predictions has been evaluated [90, 91]. Additionally, accurate configurational entropy calculations from MD simulations and a complete quasi-harmonic analysis have demonstrated that the thermodynamic role of receptor flexibility is generally underestimated [92]. Alternative strategies based on MM-PBSA-type thermodynamic estimates, which involve the implicit description of the solvation and desolvation thermodynamic effects involved in protein–ligand binding, are being pursued. These terms are currently implemented in the AutoDock 4.0 scoring functions on empirical basis only. Using a more generally parameterized MD-type force field to evaluate and rescore the docked complexes should lead to more accurate estimates of the binding affinities, as well as allow for increased transferability of the RCS to a more diverse set of systems.

Second, concerning the final refinement procedure (bottom of the funnel in Fig. 1): the application of accurate (explicit solvent) free energy calculations for a larger number of ligands and receptors of increasing size will primarily be influenced by force field accuracy, the ability to attain extensive sampling, and an improved description of the enthalpy–entropy compensation thermodynamics. Although the computational determination of free energy changes has become a standard procedure for which a variety of techniques have been developed, absolute entropies and their differences are still rarely computed. The rapid development of computer resources accompanied by force field refinement and improved simulation algorithms will naturally extend the range of problems that free energy calculations can directly assess.

Conclusions

Accounting for receptor flexibility in computer-aided drug design is still a major challenge. Recent examples illustrate the importance of predicting and including induced-fit effects upon receptor–ligand binding. MD simulations of receptors in complex with known and potential inhibitors provide relevant biochemical insights, which are otherwise not accessible through standard experimental techniques. Despite this, a general and highly transferable procedure that reliably and efficiently accommodates receptor flexibility is still lacking. The extensions and methodological improvements to the RCS presented here take important steps toward offering such a streamlined procedure. Our examples indicate that alternative choices of receptor ensembles can significantly alter the predictive power of RCS calculations, and that it is possible to reduce the receptor ensemble to a non-redundant set of configurations by various techniques without losing relevant binding information. Furthermore, both example systems indicate that the role of explicit water molecules in molecular association remains one of the key components of computer-aided drug design methods to be further investigated. A summary of the crucial points that we anticipate will help drive the future development of the RCS was presented and what emerges is that a clear and pressing challenge, closely coupled to receptor flexibility, is the development of methods to better estimate ligand and receptor entropy. Their subsequent application to molecular association thermodynamics will allow an increased accuracy in the description of enthalpy-entropy compensation effects during the ligand binding process.

Acknowledgements We thank Justin Gullingsrud, William "Lindy" Lindstrom, and Jung-Hsin Lin for fruitful discussions. R. E. A. is funded by National Institutes of Health (NIH) Grant 1F32 GM077729 and National Science Foundation (NSF) MRAC Grant CHE060073N. Funding by NIH GM31749, NSF MCB-0506593 and MCA93S013 (to J. A. M.) also supports this work. Additional support from the Howard Hughes Medical Institute, San Diego Supercomputing Center, Accelrys, Inc., the W. M. Keck Foundation, the National Biomedical Computational Resource and the Center for Theoretical Biological Physics is gratefully acknowledged.

References

1. Carlson HA (2002) Curr Opin Chem Biol 6:447
2. Wong CF, McCammon JA (2003) Annu Rev Pharmacol Toxicol 43:31
3. Davis AM, Teague SJ (1999) Angew Chem Int Ed Engl 38:736
4. Ma B, Shatsky M, Wolfson HJ, Nussinov R (2002) Protein Sci 11:184
5. May A, Zacharias M (2005) Biochim Biophys Acta 1754:225
6. Lybrand TP, McCammon JA, Wipff G (1986) Proc Natl Acad Sci USA 83:833
7. Gilson MK, Given JA, Bush BL, McCammon JA (1997) Biophys J 72:1047
8. Åqvist J, Luzhkov VB, Brandsdal BO (2002) Acc Chem Res 35:358
9. Gohlke H, Klebe G (2002) Angew Chem Int Ed Engl 41:2644
10. Jorgensen WL (2004) Science 303:1813
11. Adcock SA, McCammon JA (2006) Chem Rev 106:1589
12. Blake JF (2000) Curr Opin Biotechnol 11:104
13. Egan WJ, Walters WP, Murcko MA (2002) Curr Opin Drug Discov Dev 5:540
14. Irwin JJ (2006) Curr Opin Chem Biol 10:352
15. Li C, Xu L, Wolan DW, Wilson IA, Olson AJ (2004) J Med Chem 47:6681

16. Fonovic M, Bogyo M (2007) Curr Pharm Des 13:253
17. Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Nat Rev Drug Discov 3:935
18. Brooijmans N, Kuntz ID (2003) Annu Rev Biophys Biomol Struct 32:335
19. Taylor RD, Jewsbury PJ, Essex JW (2002) J Comput Aided Mol Des 16:151
20. Schulz-Gasch T, Stahl M (2003) J Mol Model 9:47
21. Cole JC, Murray CW, Nissink JWM, Taylor RD, Taylor R (2005) Proteins Struct Funct Bioinf 60:325
22. Sousa SF, Fernandes PA, Ramos MJ (2006) Proteins 65:15
23. Sperandio O, Miteva MA, Delfaud F, Villoutreix BO (2006) Curr Protein Pept Sci 7:369
24. Whittle M, Gillet VJ, Willett P, Alex A, Loesel J (2004) J Chem Inf Comput Sci 44:1840
25. Hajduk PJ, Greer J (2007) Nat Rev Drug Discov 6:211
26. King PM (1993) Free energy via molecular simulation: a primer. Escom, Leiden
27. Beveridge DL, DiCapua FM (1989) Annu Rev Biophys Biophys Chem 18:431
28. Straatsma TP, McCammon JA (1991) Methods Enzymol 202: 497
29. Lamb ML, Jorgensen WL (1997) Curr Opin Chem Biol 1:449
30. Chipot C, Rozanska X, Dixit SB (2005) J Comput Aided Mol Des 19:765
31. Reinhardt WP, Miller MA, Amon LM (2001) Acc Chem Res 34:607
32. van Gunsteren WF, Bakowies D, Baron R, Chandrasekhar I, Christen M, Daura X, Gee P, Geerke DP, Glättli A, Hünenberger PH, Kastenholz MA, Oostenbrink C, Schenk M, Trzesniak D, van der Vegt NF, Yu HB (2006) Angew Chem Int Ed Engl 45:4064
33. van Gunsteren WF, Daura X, Mark AE (2002) Helv Chim Acta 85:3113
34. van Gunsteren WF, Beutler TC, Fraternali F, King PM, Mark AE, Smith PE (1993) Computation of free energy in practice: choice of approximations and accuracy limiting factors. Escom, Leiden
35. Hünenberger PH, Helms V, Narayana N, Taylor SS, McCammon JA (1999) Biochemistry 38:2358
36. Chipot C, Pohorille A (2007) Free energy calculations. Springer, Berlin
37. Åqvist J, Medina C, Samuelsson JE (1994) Protein Eng 7:385
38. Almlof M, Ander M, Åqvist J (2007) Biochemistry 46:200
39. Almlof M, Brandsdal BO, Åqvist J (2004) J Comput Chem 25:1242
40. Mark AE, Xu Y, Liu H, van Gunsteren WF (1995) Acta Biochim Pol 42:525
41. Liu H, Mark AE, van Gunsteren WF (1996) J Phys Chem 100:9485
42. Oostenbrink C, van Gunsteren WF (2003) J Comput Chem 24:1730
43. Zagrovic B, van Gunsteren WF (2007) J Chem Theory Comput 3:301
44. Kua J, Zhang Y, McCammon JA (2002) J Am Chem Soc 124(28):8260
45. Oostenbrink C, van Gunsteren WF (2005) Proc Natl Acad Sci USA 102:6750
46. Frembgen-Kesner T, Elcock AH (2006) J Mol Biol 359:202
47. Cavasotto CN, Kovacs JA, Abagyan RA (2005) J Am Chem Soc 127:9632
48. Kovacs JA, Cavasotto CN, Abagyan R (2005) J Comput Theor Nanosci 2:354
49. Orry AJW, Abagyan RA, Cavasotto CN (2006) Drug Discov Today 11:261
50. Sherman W, Day T, Jacobson MP, Friesner RA, Farid R (2006) J Med Chem 49:534
51. Carlson HA, Masukawa KM, Rubins K, Bushman FD, Jorgensen WL, Lins RD, Briggs JM, McCammon JA (2000) J Med Chem 43:2100
52. Meagher KL, Carlson HA (2004) J Am Chem Soc 126:13276
53. Lin JH, Perryman AL, Schames JR, McCammon JA (2002) J Am Chem Soc 124:5632
54. Lin JH, Perryman AL, Schames JR, McCammon JA (2003) Biopolymers 68:47
55. Schames JR, Henchman RH, Siegel JS, Sotriffer CA, Ni H, McCammon JA (2004) J Med Chem 47:1879
56. Huang SY, Zou XQ (2007) Proteins Struct Funct Bioinf 66:399
57. Damm KL, Carlson HA (2007) J Am Chem Soc 129:8225
58. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten K (2005) J Comput Chem 26:1781
59. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M (1998) J Phys Chem B 102:3586
60. Christen M, Hünenberger PH, Bakowies D, Baron R, Bürgi R, Geerke DP, Heinz TN, Kastenholz MA, Kräutler V, Oostenbrink C, Peter C, Trzesniak D, van Gunsteren WF (2005) J Comput Chem 26:1719
61. Soares TA, Hünenberger PH, Kastenholz MA, Kräutler V, Lenz T, Lins RD, Oostenbrink C, Van Gunsteren WF (2005) J Comput Chem 26:725
62. van Gunsteren WF, Billeter SR, Eising AA, Hünenberger PH, Krüger P, Mark AE, Scott WRP, Tironi IG (1996) Biomolecular simulation: the GROMOS96 manual and user guide, vdf Hochschulverlag AG an der ETH Zürich BIOMOS b.v. Groningen, Zürich Groningen
63. Amaro RE, Swift RV, McCammon JA (2007) PLoS Negl Trop Dis 1(2):e68. doi:10.1 371/journal.pntd.0000068
64. Baron R, McCammon JA (2007) Biochemistry 46:10629
65. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ (1998) J Comput Chem 19:1639
66. Huey R, Morris GM, Olson AJ, Goodsell DS (2007) J Comput Chem 28:1145
67. Gasteiger J, Marsili M (1980) Tetrahedron 36:3219
68. Amaro R, Schnaufer A, Hol WG, Stuart K, McCammon JA (2007) in preparation
69. Baron R, Mccammon JA (2007) in preparation
70. Humphrey W, Dalke A, Schulten K (1996) J Mol Graph 14:33
71. Daura X, van Gunsteren WF, Mark AE (1999) Proteins 34:269
72. Park H, Lee J, Lee S (2006) Proteins 65:549
73. Schnaufer A, Panigrahi AK, Panicucci B, Igo RP Jr, Wirtz E, Salavati R, Stuart K (2001) Science 291:2159
74. Deng J, Schnaufer A, Salavati R, Stuart KD, Hol WG (2004) J Mol Biol 343:601
75. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Adv Drug Deliv Rev 46:3
76. Markowitz M, Nguyen B-Y, Gotuzzo F et al (2006) XVI International AIDS conference, Toronto, Canada
75. Fitzgerald MM, Churchill MJ, McRee DE, Goodin DB (1994) Biochemistry 33:3807
78. Fitzgerald MM, Musah RA, McRee DE, Goodin DB (1996) Nat Struct Biol 3:626
79. Fitzgerald MM, Trester ML, Jensen GM, McRee DE, Goodin DB (1995) Protein Sci 4:1844
80. Musah RA, Goodin DB (1997) Biochemistry 36:11665
81. Musah RA, Jensen GM, Bunte SW, Rosenfeld RJ, Goodin DB (2002) J Mol Biol 315:845
82. O'Donoghue P, Luthey-Schulten Z (2003) Microbiol Mol Biol Rev 67:550

83. O'Donoghue P, Luthey-Schulten Z (2005) J Mol Biol 346:875
84. Sethi A, O'Donoghue P, Luthey-Schulten Z (2005) Proc Natl Acad Sci USA 102:4045
85. Russell RB, Barton GJ (1992) Proteins 14:309
86. Baron R, Bakowies D, van Gunsteren WF, Daura X (2002) Helv Chim Acta 85:3872
87. Baron R, Bakowies D, Van Gunsteren WF (2005) J Pept Sci 11:74
88. van Dijk AD, Bonvin AM (2006) Bioinformatics 22:2340
89. Rosenfeld RJ, Goodsell DS, Musah RA, Morris GM, Goodin DB, Olson AJ (2003) J Comput Aided Mol Des 17:525
90. Ruvinsky AM (2007) J Comput Chem 28:1364
91. Salaniwal S, Manas ES, Alvarez JC, Unwalla RJ (2007) Proteins 66:422
92. Baron R, McCammon JA (2007) Chem Phys Chem, in press