J Comput Aided Mol Des (2010) 24:1023–1033
DOI 10.1007/s10822-010-9393-x

Ligand-guided optimization of CXCR4 homology models for virtual screening using a multiple chemotype approach

Marco A. C. Neves · Sérgio Simões · M. Luisa Sá e Melo

Received: 21 July 2010 / Accepted: 11 October 2010 / Published online: 20 October 2010

Abstract CXCR4 is a G-protein coupled receptor for CXCL12 that plays an important role in human immunodeficiency virus infection, cancer growth and metastasization, immune cell trafficking and WHIM syndrome. In the absence of an X-ray crystal structure, theoretical modeling of the CXCR4 receptor remains an important tool for structure–function analysis and to guide the discovery of new antagonists with potential clinical use. In this study, the combination of experimental data and molecular modeling approaches allowed the development of optimized ligand-receptor models useful for elucidation of the molecular determinants of small molecule binding and functional antagonism. The ligand-guided homology modeling approach used in this study explicitly re-shaped the CXCR4 binding pocket in order to improve discrimination between known CXCR4 antagonists and random decoys. Refinement based on multiple test-sets with small compounds from single chemotypes provided the best early enrichment performance. These results provide an important tool for structure-based drug design and virtual ligand screening of new CXCR4 antagonists.

Electronic supplementary material The online version of this article (doi:10.1007/s10822-010-9393-x) contains supplementary material, which is available to authorized users.

M. A. C. Neves (✉) · M. L. Sá e Melo
Centro de Estudos Farmacêuticos, Laboratório de Química Farmacêutica, Faculdade de Farmácia, Universidade de Coimbra, Pólo das Ciências da Saúde, 3000-548 Coimbra, Portugal
e-mail: mneves@ff.uc.pt

S. Simões
Centro de Neurociências, Laboratório de Tecnologia Farmacêutica, Faculdade de Farmácia, Universidade de Coimbra, Pólo das Ciências da Saúde, 3000-548 Coimbra, Portugal

Keywords CXCR4 · CXCR4 antagonists · Homology modeling · Virtual screening · Drug discovery

Introduction

The C-X-C chemokine receptor type 4 (CXCR4), a seven-transmembrane (7TM) G-protein coupled receptor (GPCR), is a member of class A rhodopsin-like GPCR family that mediates signaling from chemokine CXCL12 (stromal cell-derived factor 1, SDF-1). Chemokines are small proteins with potent chemoattractant activity for immune system cells and a key role in their recruitment to the sites of inflammation [1]. CXCR4 and CXCL12 are constitutively expressed in several tissues which is consistent with their involvement in a broad spectrum of physiological activities such as cardiovascular and central nervous system cell migration during embryonic development and hematopoietic stem cell homing to the bone marrow [2, 3]. Besides the physiological roles, CXCR4 is also related to pathological conditions. Fusion of human immunodeficiency virus (HIV) with the host cell membrane initiates with binding of the viral envelope glycoprotein gp120 to both the CD4 cell surface receptor and one of the CXCR4 or CCR5 chemokine co-receptors [4]. CXCR4 signaling was shown to be involved in cancer growth and invasion, promoting metastasization of the most common types of human carcinomas including breast, prostate and lung cancers [5, 6]. CXCL12 plays an important role in the recruitment of immune cells to acute and chronic inflammation sites, e.g. during rheumatoid arthritis and atherosclerosis [7, 8]. Therefore, small compound antagonists of the CXCR4 receptor might be an interesting new therapeutic approach for treatment of HIV [9], tumor metastasization [10] and inflammatory diseases

[11]. WHIM syndrome, a congenital immunodeficiency characterized by CXCR4 mutations that cause its inability to downregulate after stimulation, is another potential application for CXCR4 antagonists [12, 13].

The development of small molecule CXCR4 antagonists was initially focused on bicyclam-containing compounds such as AMD3100 (Fig. 1) [14, 15]. Bearing an overall +2 charge per cyclam ring at physiological pH, AMD3100 was found to interact with carboxylic acid groups of Asp171, Asp262 and Glu288 located at the CXCR4 binding pocket [16, 17]. Poor oral bioavailability and significant cardiotoxicity limited however its clinical use [18, 19]. Monocyclam derivatives have lower molecular weight, less charge and improved oral bioavailability. AMD3465 (Fig. 1) fully preserves the biological properties of AMD3100 [20]. Considerable improvement in terms of potency and pharmacokinetic properties was achieved with the optimization of series of tetrahydroquinolinamines [21–23]. AMD070 (Fig. 1), a low molecular weight CXCR4 antagonist, is well tolerated and orally available [24]. Another different class of low molecular weight CXCR4 antagonists was designed by replacement of the AMD3100 cyclam moieties with alkyl or aromatic nitrogen-containing groups. MSX-122 (Fig. 1) is the lead compound of this class [25, 26]. Orally bioavailable functionalized isothioureas were found to be potent and selective CXCR4 antagonists (Fig. 1) [27]. Their potential use in WHIM syndrome treatment was recently patented by Novartis Pharmaceuticals. Another series of chemical diverse small-molecules with potent anti-CXCR4 activity was disclosed by ChemoCentryx Inc and patented for treatment of CXCR4-mediated diseases (Fig. 1) [28].

Structural information on individual GPCRs is limited to the rhodopsin receptor [29, 30], the β_1 and β_2 adrenergic receptors (β_1AR and β_2AR) [31, 32], and the adenosine A_{2A} receptor [33]. Therefore, homology modeling remains an important tool for structure–function analysis of the GPCR family. Modeling of small compound-CXCR4 complexes is useful to understand the molecular mechanisms of ligand antagonism and to provide a basis for structure-based drug design and virtual ligand screening (VLS). Typically, with more than 50% sequence identity, alignments are straightforward and the resulting models are of sufficient quality to be used in the prediction of detailed protein–ligand interactions. With more distantly related targets, however, such as modeling peptide-activated GPCRs based on the structure of (rhod)opsin, amine or nucleotide-like GPCR receptors, sequence identity drops below 30% and the overall model quality becomes critically depend on available experimental data. In such cases, conventional alignment methods might not provide a sufficiently reliable outcome if not validated with site-directed mutagenesis data or family profile-based information. Even when using an acceptable and well validated sequence alignment, the binding site shape, volume and side-chain orientation of important residues for small-molecule binding, will likely be incorrectly predicted if model and template GPCRs bind to very different ligand chemotypes. In such cases, model quality can be improved by sampling alternative binding pocket conformations and selecting the most predictive models for known active ligands. Successful applications of binding site ensembles for homology modeling refinement of GPCR receptors include the melanin-concentrating hormone receptor 1 (MCH-R1) [34], the α1A adrenergic receptor (α1AAR) [35], the neurokinin-1 receptor (NK1) [36] and the adenosine A_{2A} receptor [37].

In this work we describe the application of ligand-guided homology modeling to the CXCR4 receptor using a

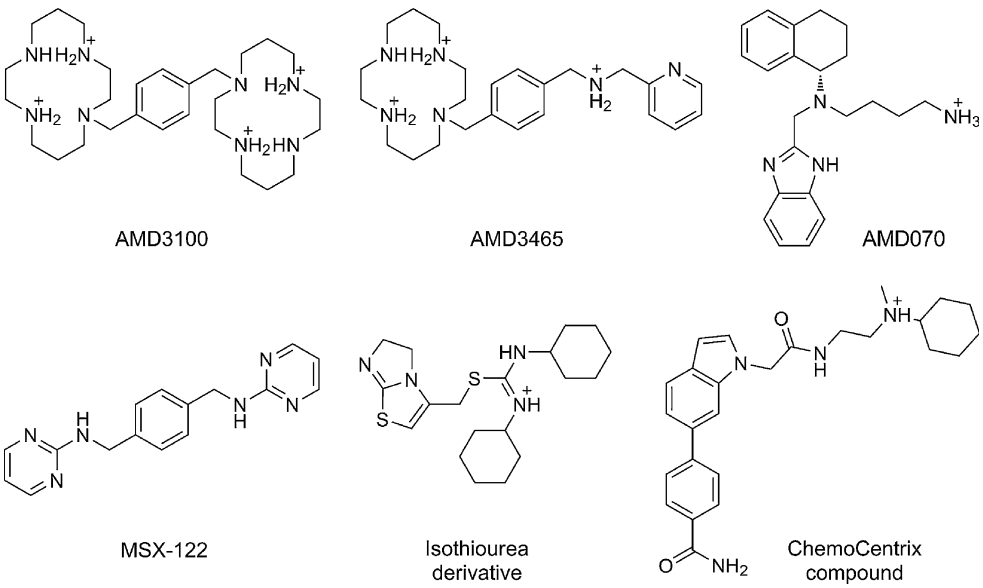

Fig. 1 Potent small-compound CXCR4 antagonists from five different chemotypes, i.e. cyclams (AMD3100 and AMD3465), tetrahydroquinolinamines (AMD070), cyclam-based molecules (MSX-122), isothioureas and ChemoCentrix compounds. The dominant protonation state at pH 7.4 defined for docking is represented

test-set with antagonists from multiple chemotypes. Model quality for virtual screening purposes was validated based on the ability to discriminate known CXCR4 antagonists from random decoys. Putative binding modes for representative CXCR4 antagonists are described and early enrichments obtained with single or multiple receptor representations are discussed.

Computational methods

Homology modeling

The amino acid sequence of human CXCR4 was retrieved from the UniProtKB database (accession code P61073) and aligned to the sequence of human β_2-adrenergic receptor extracted from the X-ray crystal structure of β_2AR (pdb entry 2RH1) without the T4 lysozyme fusion [31]. CXCR4 shares 22% amino acid sequence identity with β_2AR. Sequence alignment was carried out in ICM version 3.6–1 h using the default comparison matrix and a zero end-gap global alignment algorithm (ZEGA method) described by Abagyan and Batalov [38], with gap opening and extension penalties of 2.4 and 0.15, respectively. Manual adjustments were made to correctly align highly preserved class A GPCR Ballesteros residues and conserved sequence motifs such as the "ionic-lock" on TM3 (transmembrane domain 3) and the NPXXY motif on TM7 [39]. Cysteines participating in a disulfide bridge between TM3 and EL2 (extracellular loop 2), i.e. Cys109 and Cys186, were aligned to the corresponding residues in β_2AR and a preliminary homology model was obtained with ICM. Briefly, the ICM method aligned CXCR4 residues into the β_2AR template and the energy of the system was globally optimized by a series of global Monte Carlo random moves and gradient local minimizations in the internal coordinates space [40]. Less conserved and missing template loops were further modeled by searching through loop templates extracted from the PDB, followed by Monte Carlo sampling of the side chains, and energy minimization of the backbone.

Residues within the putative binding pocket for CXCR4 antagonists were visually inspected and compared with known site-directed mutagenesis data [41]. Although most of the important residues known to be involved in stabilizing interactions with CXCR4 antagonists were correctly predicted to be located in the binding pocket, in some cases, i.e. Trp94 and Asp97 on TM2 and Asp171 on TM4, these residues were pointing towards the outside of the TM bundle. Single-residue gaps introduced at TM2 and TM4 shifted the alignment and redirected these residues to the CXCR4 binding pocket. Following alignment correction, a new homology model was built and refined by energetic minimization under distance restraints to the starting template. EL2 was removed from the final homology model.

The stereochemical quality of the model was evaluated using PROCHECK [42] and the interactive Ramachandran plot tool of ICM. Six out of 262 Phi/Psi angle pairs were found to be outside the expected Ramachandran regions; however, deviating residues were located in loop regions and away from the putative binding site. Omega angle check revealed a single deviation outside a 180 ± 30 degree range. The residues involved, i.e. Ile169 and Pro170, were found to deviate from the ideal trans configuration as a result of the intentional one-residue gap introduced in TM4.

Small-compound database construction

A test-set of structurally diverse CXCR4 antagonists was compiled from the available literature and patent databases. Compounds were selected based on low molecular weight (MW $\leq$ 600 Da) and strong potency ($IC_{50} \leq 1$ μM) criteria and divided into five different chemotypes (Fig. 1). Discrimination between potent active compounds and random decoys was used to access the usefulness of the initial model in virtual ligand screening and to guide homology modeling refinement. In order to get reliable results and avoid bias, decoys were selected resembling the active compounds in terms of physicochemical properties, being consistent with binding to GPCR transmembrane domains, but different enough to ensure they are true non-CXCR4 binders [43]. The GPCR ligand database (GLIDA) version 2.03 with about 24000 unique ligand entries was used [44]. Decoys were selected from the GLIDA database through several steps: (1) GLIDA molecules with CXCR4 agonist or antagonist annotations were not considered. (2) Molecular weight, clog P, formal charge at pH 7.4 and number of hydrogen bonding groups was limited to the same range of the known CXCR4 antagonists used in this study. (3) Binary fingerprints were calculated for all molecules using the ICM software. The Tanimoto coefficient, defined as $T = C/(A + B–C)$, where A is the number of non-zero bits for molecule A (GLIDA compound), B is the number of non-zero bits for molecule B (CXCR4 antagonist), and C is the number of non-zero bits common to A and B, was calculated between each individual GLIDA compound and each test-set active compound. The chemical similarity distance was defined as 1-*T*. GLIDA compounds with a Tanimoto distance bellow 0.5 were excluded. (5) One thousand decoys were selected randomly from the remaining compounds on the database.

Three-dimensional models for the molecules were built using ICM Molecule Editor and prepared for docking by adding hydrogens, assigning formal charges at pH 7.4, selecting the most populated tautomeric form, assigning

Merck Molecular Force Field (MMFF) [45] atom types and partial charges and fixing omega-like angles to 180°. One initial minimization with MMFF was performed for all compounds in study imposing chirality restraints for chiral centers.

Binding pocket conformational sampling

All heavy-atom elastic network normal modes analysis (EN-NMA) was used to sample the conformational flexibility within the CXCR4 binding pocket. Normal modes are harmonic oscillations about local energy minima where interaction energies between pairs of atoms are described by the Hook's potential. Initial distances are taken to be at the energy minimum and the spring constant assumes an inverse exponential relationship with the distance. The force constant matrix of the system is described by the Hessian matrix (a 3 N × 3 N matrix, being N the number of atoms), obtained as the partial second derivatives of the potential with respect to the coordinates. Diagonalization of the Hessian matrix yields 3 N-6 eigenvectors ranked according to their corresponding eigenvalues. The eigenvectors contain the amplitude and direction of motion for each atom, and the eigenvalues give the energy cost of deforming the system along the eigenvectors [46].

Elastic network normal mode analysis was performed using Abagyan's lab EN-NMA server v1.0 (http://abagyan.ucsd.edu/MRC/) [37, 47]. Starting from the non-optimized CXCR4 homology model as initial coordinates, one hundred receptor conformations were generated by sampling backbone and side chain heavy atoms within a 2 Å amplitude, and the chemical distortions associated to the displacements were corrected with 25 steps of energy minimization. The models were then used to dock known CXCR4 antagonists and random decoys using the virtual ligand screening protocol described bellow in detail and the most predictive models were selected based on maximal VLS enrichments using either the full test-set of compounds (55 CXCR4 antagonists + 1000 decoys) or 5 separate test-sets with individual chemotypes (11 CXCR4 antagonists + 1000 decoys). Therefore, 6 different models were selected as starting coordinates for the next EN-NMA cycles (3 more cycles were performed in total) and an ensemble of 600 conformations was created during each generation.

Fully flexible ligand and receptor docking was used to optimize the best final EN-NMA derived models. Residues within a 5 Å cutoff from the docked CXCR4 antagonists were allowed to randomly move using the ICM biased-probability Monte Carlo algorithm described previously [48], followed by a full local energy minimization performed with ICM. Geometrically diverse low-energy conformations were saved in the conformational stack and benchmarked using the virtual ligand screening protocol described bellow. Normalized square root area under the curve (NSQ_AUC) in receiver operating characteristic (ROC) curves, described in more detail bellow, was used as VLS benchmark metric representing the probability that a randomly selected active compound scores higher than a randomly selected decoy early in a rank-ordered list of molecules.

Stereochemical quality assessment for EN-NMA derived models reveled similar number of residues inside the expected Ramachandran regions but a slightly lower number of residues within the most favorable areas of the plot. These results were expected due to the more dynamic nature of the EN-NMA-derived conformations.

Virtual ligand screening protocol

The multiple receptor conformations derived from EN-NMA were imported into ICM version 3.6–1 h and optimized by a multistep procedure involving global optimization of hydrogens for the best hydrogen bonding network, formal and partial charge assignment and refinement of histidine, aspargine and glutamine residues. Monte Carlo conformations were used without further refinement. The models were superimposed at the putative binding pocket and grid maps accounting for hydrogen bonding potential (gb map), van der Waals potential with a carbon-like probe (gc map), van der Waals potential with a sulphur-like probe (gl map), van der Waals potential with a hydrogen probe (gh map), hydrophobic potential (gs map) and electrostatic potential (ge map), were calculated within a common 20 Å × 20 Å × 15 Å grid box centered at the binding cavity.

Test-set docking into each individual receptor was performed using the pre-calculated potential grid maps and default parameters. Five independent docking runs were performed and the best scores were selected from the multiple answers. A ranked list was created by sorting the compounds by increasing ICM binding score, i.e. decreasingly favorable interactions.

VLS benchmark metrics

CXCR4 models were benchmarked for their ability to discriminate potent low molecular weight antagonists from random decoys using normalized square root AUC in ROC plots [49]. In a linear ROC curve the true positives rate is plotted as a function of the false positives rate for all positions of the ranked score list. Homology models with perfect discrimination, i.e. scoring all true positives at the top ranked positions, have ROC plots that pass through the upper left corner and AUCs equal to 100. Therefore the higher the AUC in a ROC curve, the better the discrimination of the model. Successful VLS runs rank active

compounds early in a large score list, however, it was shown that in some cases AUC fails to address this early recognition problem [50]. The normalized square root AUC (NSQ_AUC) is a new metric described recently by Abagyan and coworkers that combines the overall selectivity of ROC AUC with early enrichments [37]. Briefly, a square root transformation is applied to the x-axis (x = square root of true negatives rate) and the linear AUC (AUC*) is normalized according to the following equation:

$$\mathrm{NSQ\_AUC} = 100 \times \frac{\mathrm{AUC}^* - \mathrm{AUC}^*_{\mathrm{random}}}{\mathrm{AUC}^*_{\mathrm{perfect}} - \mathrm{AUC}^*_{\mathrm{random}}}$$

At each optimization step NSQ_AUC was used to select the most predictive binding pocket models generated by EN-NMA and Monte Carlo refinement.

Results and discussion

Sequence alignment and homology modeling

The high-resolution X-ray crystal structure of human β_2-adrenergic receptor bound to carazolol was used as a template for homology modeling the human CXCR4 chemokine receptor [31]. Alternative X-ray crystal structures that could have been used as templates include the rhodopsin receptor with and without co-crystallized retinal [29, 30], the β_1-adrenergic receptor bound to cyanopindolol [32] and the A_{2A} adenosine receptor co-crystallized with ZM241385 [33]. β_2AR was preferred in this study for the following reasons: (1) The rhodopsin receptor has lower sequence identity to the CXCR4 receptor than β_2AR. (2) While rhodopsin receptor is covalently linked to retinal in the inactive state, CXCR4 and β_2AR bind to diffusible ligands. (3) Structural changes in the retinal-free rhodopsin receptor have been linked to the GPCR active-state which makes it less suitable to model antagonist-bound conformations. (4) With over 65% of amino acid sequence identity, β_1AR and β_2AR X-ray crystal structures are structurally very similar and expected to give equivalent CXCR4 homology models. (5) Although A_{2A} adenosine receptor shows similar sequence identity to CXCR4 compared to β_2AR, the position and orientation of their transmembrane helixes define an antagonist-binding cavity with only limited access to residues in TM3 and TM4, known to be important for binding of CXCR4 antagonists.

Despite a relatively low sequence identity of 22%, β_2AR and CXCR4 share significant number of conserved residues and sequence patterns which can be used to guide sequence alignment. Ballesteros residues such as Asn56 on TM1, Asp84 on TM2, Arg134 on TM3, Trp161 on TM4, Pro211 on TM5, Pro254 on TM6 and Pro299 on TM7 were used to match equivalent transmembrane helixes (Fig. 2). Other conserved residues include the "ionic-lock" (DRY motif) at the cytoplasmic end of TM3, the NPXXY motif on TM7 and the "rotamer toggle switch" (Trp252) on TM6. Cys109 at the extracellular end of TM3 and Cys186 located at the extracellular loop 2 are conserved among the GPCR family and involved in disulfide bond formation [51]. These residues were aligned with equivalent cysteines in β_2AR. Sequence alignment was further refined with available site-directed mutagenesis data indentifying Tyr45 on TM1, Trp94 and Asp97 on TM2, Tyr116 on TM3, Asp171 on TM4, Asp262 on TM6 and His281 and Glu288 on TM7, as critical residues involved in stabilizing interactions with known CXCR4 antagonists [16, 17, 41]. These residues were correctly pointing towards the binding pocket in the final unrefined homology model and, with the exception of

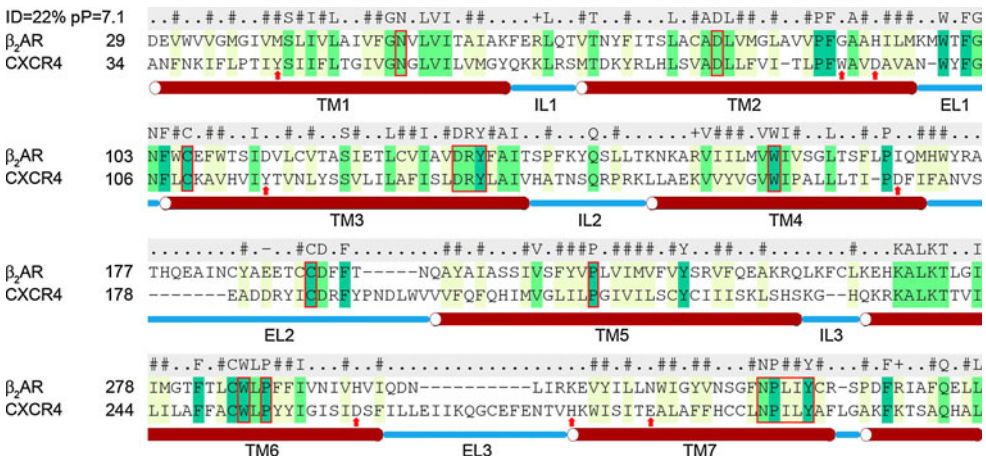

Fig. 2 Pairwise sequence alignment of β_2AR and CXCR4 GPCR receptors. Transmembrane helices (TM), intracellular loops (IL) and extracellular loops (EL) are shown below the alignment and the consensus sequence is given above ("1-letter codes" = conserved, "#" = hydrophobic, "." = not conserved, "±" = conserved charge). Highly conserved residues and sequence motifs among family A GPCRs are highlighted with a red frame and putative CXCR4 binding pocket residues identified with a *red arrow* below the alignment

Tyr45, fully accessible to interact with potential CXCR4 antagonists. Contacts with Tyr45 were limited due a displacement of the extracellular end of TM1 outwards the TM helix bundle and by a bulky tryptophan residue (Trp94) on TM2. The extracellular loop 2 was removed from the final model due to the very low sequence identity and large gaps in the alignment. Molecular recognition can be improved in homology models where highly uncertain areas are removed [52], however, some artifacts might result from this procedure such as: (1) Lower docking scores of known ligands due to a reduced number of protein-small molecule interactions. (2) Higher number of false positives in virtual ligand screening experiments due to a more open binding cavity where ligands can extend into areas otherwise not accessible. The second artifact can be avoided by limiting the binding pocket volume during the docking runs. Furthermore, due to the nature of its endogenous ligand, CXCR4 binding pocket is expected to be larger and more solvent exposed than β_2AR, with ligands establishing less contacts with EL2 residues.

The final homology model obtained at this stage was in good agreement with known site-directed mutagenesis data. Its usefulness for virtual screening and structure-based drug design was further evaluated using a ligand-guided approach.

Test-set construction

A test-set of 55 well known small molecule CXCR4 antagonists was collected from the literature, patents and publicly available resources such as the GLIDA [44] database (Online Resource 1). Compounds were selected based on criteria of low molecular weight (peptides and other CXCR4 antagonist classes with MW above 600 were not considered in this study), strong potency (limited to 1 μM IC_{50} measured in either a binding assay or a calcium mobilization assay), 3D structure confidence (racemic compounds or unknown cis–trans isomers were discarded) and chemical diversity. Representative structures of compounds from five distinct chemotypes are shown in Fig. 1. The number of compounds per class was set to eleven, i.e. the number of chemical entities available for the least populated chemotype (isothioureas).

Test-set compounds showed significant chemical diversity in terms of scaffold type and molecular properties. Molecular weights ranged from 582 g/mol (molecules with 42 atoms) to 290 g/mol (molecules with 22 atoms), calculated logP from −3.9 to 5.7, hydrogen bond donors from 1 to 10 and hydrogen bond acceptors from 0 to 6. Important for binding and functional CXCR4 antagonism is the overall charge of the compounds at physiological pH. CXCR4 antagonists have been shown to interact with carboxylic acid groups within the binding pocket of the receptor [16, 17]. Formal charges were assigned at pH = 7.4, ranging from 0, such as with some cyclam-based compounds, to +4 with most bicyclams. Distinct chemotypes will likely generate diverse binding modes to the CXCR4 binding pocket as shown in a recent study for bicyclams, monocyclams and noncyclam antagonists [41].

One thousand compounds with similar physicochemical properties were collected from the GLIDA database and used as decoys [44].

Binding pocket conformational ensemble and VLS benchmarking using a single test-set with multiple chemotypes

The homology modeling protocol described above keeps both modeled backbone and side chains in a very closely related position to the template protein, which is clearly not expected for GPCRs from different sub-families, with low sequence identity and binding to different type of ligands such as β_2AR and CXCR4. Despite that, initial enrichments for the full test-set of 55 compounds in VLS receiver operating characteristic (ROC) plots were encouraging (linear area under the curve, AUC = 69.0, normalized square root area under the curve, NSQ_AUC = 50.9, Fig. 3a) but limited for use in structure-based drug design. The ranked list of scores showed a non-homogeneous distribution of known antagonists. While chemotypes such as the ChemoCentrix molecules and isothioureas appeared at higher ranked positions (NSQ_AUC = 77.3 and 67.3, respectively), other classes of compounds such as tetrahydroquinolinamines, cyclam-based compounds and cyclams showed poor enrichments (NSQ_AUC = 21.3, 44.9 and 48.2, respectively, Fig. 3a).

The limited recognition properties identified for the non-optimized CXCR4 model prompted us to explore the conformational flexibility at the binding pocket using cycles of elastic network normal modes analysis (EN-NMA). Starting from the initial model coordinates, an ensemble of 100 conformations was generated and benchmarked using virtual ligand screening runs of the full test-set with 55 known antagonists and 1000 decoys. During the first EN-NMA iteration a new CXCR4 model with NSQ_AUC = 57.2 was obtained (Fig. 3a), with significant improvements for tetrahydroquinolinamines and cyclam-based compounds (NSQ_AUC = 51.1 and 57.4, respectively). Optimized models identified during the first EN-NMA iteration, were used as initial coordinates for a second cycle and this procedure was repeated two more times. During the second generation, improvement was achieved due to better recognition properties for cyclams, cyclam-based compounds, isothioureas and ChemoCentrix compounds (NSQ_AUC = 80.0, 74.8, 70.8 and 87.6, respectively), however, none of the two additional iterations improved the

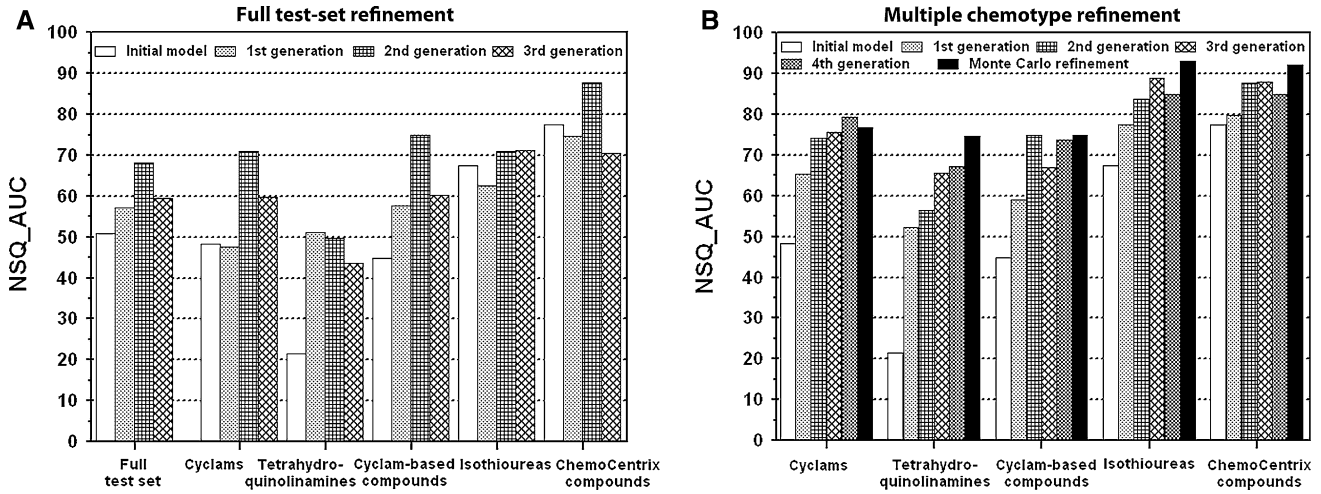

Fig. 3 Virtual ligand screening enrichments of an initial, non-refined, CXCR4 homology model, and optimized models obtained at first, second, third and fourth iterations of EN-NMA, and Monte Carlo refinement. Normalized square root area under the curve (NSQ_AUC) was used as the VLS benchmark metric. **a** Refinement of a single model using the full test-set with 55 CXCR4 antagonists from 5 distinct chemotypes (11 compounds per chemotype). Individual performances per chemotype are also shown. **b** Refinement of individual models using five test-sets with 11 compounds from single chemotypes. In both approaches one thousand GLIDA ligands were used as decoys. Selected EN-NMA models have maximal enrichments in terms of NSQ_AUC for the test-sets used

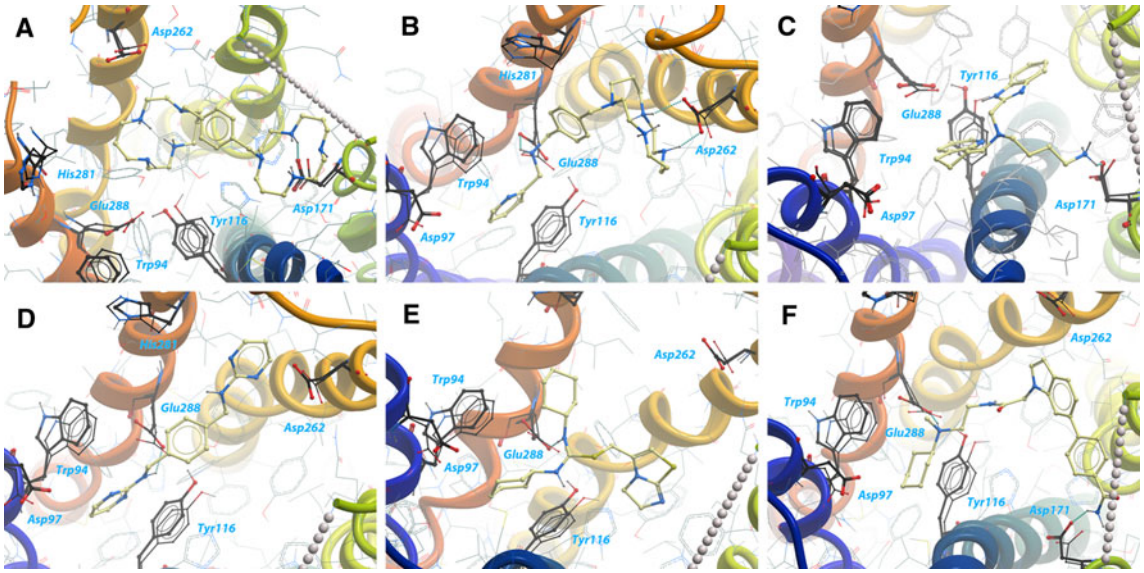

Fig. 4 Predicted binding poses of representative CXCR4 antagonists within the binding pocket of optimized CXCR4 models. Compounds represented are: **a** AMD3100, **b** AMD3465, **c** AMD070, **d** MSX-122, **e** Isothiourea derivative, **f** ChemoCentrix compound. Known binding site residues are labelled and represented with a 'ball and sticks' model (protein carbon, *black*; small compound carbon, *yellow*; nitrogen, *blue*; oxygen, *red*; polar hydrogens, *white*). For comparison, the initial, unrefined conformations of labelled active site residues are shown with a thinner model. Figures 4a, b, d, e and f represent the same CXCR4 model, optimized with the full test-set of 55 antagonists. Figure 4c represents a CXCR4 model optimized for the individual tetrahydroquinolinamine chemotype

model quality for the full test-set (Fig. 3a; iteration 4 not shown). A total of 1900 conformations were generated during the 4 cycles of EN-NMA.

The best overall model found at iteration 2 had linear AUC equal to 88.3 and NSQ_AUC equal to 68.2 (Fig. 3a). Visual inspection of predicted binding modes for representative compounds revealed important conformational changes, redefining the shape of the binding pocket, creating extra space for ligand binding and optimizing protein–ligand complementarity (Fig. 4). Ligands were found to bind deeper to the optimized CXCR4 model establishing multiple contacts with a large binding site that can be divided into two buried sub-pockets defined by TM2, TM3 and TM7 (pocket P1) and TM4, TM5 and TM6 (pocket P2).

Particular important is a 1.5 Å displacement of the extracellular end of TM7 that optimized Glu288 side chain for hydrogen bonding and charged interactions, along with rotations of Trp94 and Tyr116 creating extra space for a more buried binding at sub-pocket P1. These conformational changes were important to improve binding scores with most chemotypes. On the other end of the TM bundle, carboxylic acid side chains of Asp171 on TM4 and Asp262 on TM6 were displaced by 1.3 Å and 0.7 Å, respectively, optimizing hydrogen bonds and charged interactions with the active compounds in study.

Bicyclams bind to the optimized CXCR4 homology model in good agreement to known mutagenesis data and previous homology modeling studies [16, 17, 53, 54]. AMD3100 binds to pocket P2 establishing a hydrogen bond and charged interactions between one of the cyclam rings and Asp171 (Fig. 4a). The second positively charged cyclam ring binds between Asp262 and Glu288. Most binding pocket residues known to be important for AMD3100 binding [41] are within a 4 Å distance (Table 1) and a similar binding mode was found for the remaining bicyclams in study. AMD3465, a mono-cyclam derivative binds to pockets P1 and P2, and establishes charged interactions with Asp262 and Glu288, as well as hydrophobic contacts with Trp94, Tyr116 and His281 (Fig. 4b). Although distinct from bicyclams, this binding mode is in good agreement with known mutagenesis data (Table 1) [41]. AMD070 interacts with Trp94, Asp97 and Glu288 which is in good agreement with site-directed experiments, but binds distant from Asp171 whose mutation is known to decrease binding potency in more than 3 orders of magnitude (Table 1, predicted binding pose not shown in Fig. 4). Cyclam-based compounds such as MSX-122 bind to CXCR4 with similar poses to AMD3465 (Fig. 4d). Hydrogen bonding with Glu288 is important for binding. Isothioureas bind deep into pockets P1 and P2 with strong interactions between the positively charged nitrogen and Glu288 (Fig. 4e). ChemoCentrix compounds combine the binding modes of mono-cyclams and bicyclams, establishing a charged interaction with Glu288 and extending the molecular framework from TM2 to TM6 and from TM6 to TM4 (Fig. 4e).

VLS benchmarking using multiple test-sets with single chemotypes

While no further improvement in the recognition properties for the full test-set of 55 CXCR4 antagonists was obtained after two cycles of EN-NMA (NSQ_AUC = 68.2), model refinement based on enrichments for a single chemotype gives a better performance per class, in most cases with improvements along the 4 EN-NMA generations. Further Monte-Carlo (MC) refinement was also performed in this study, leading to models with better recognition properties per class, in most cases. Final optimized models for isothioureas and ChemoCentrix compounds presented the higher enrichments, NSQ_AUC = 93.1 and NSQ_AUC = 92.1, respectively, whereas cyclams, tetrahydroquinolinamines and cyclam-based compounds, showed NSQ_AUC = 79.4, NSQ_AUC = 74.5 and NSQ_AUC = 74.7, respectively (Fig. 3b). Homology models refined using a single chemotype of CXCR4 antagonists showed poor recognition properties for the full test-set (NSQ_AUC ~50) due to overfit for a particular molecular scaffold, however, combined together, the five models have better enrichments than any single

Table 1 Comparison between known site-directed mutagenesis data and atomic contacts predicted for binding site models optimized with a full test-set of CXCR4 antagonists or multiple test-sets with individual chemotypes

Mutation	AMD3100		AMD3465		AMD070	
	K_i/K_iwt[a]	Contacts:[b] full test-set/ multiple chemotypes	K_i/K_iwt[a]	Contacts:[b] full test-set/ multiple chemotypes	K_i/K_iwt[a]	Contacts:[b] full test-set/ multiple chemotypes
Y45A	10.0	No/No	21.6	No/No	39.8	No/No
W94A	14.0	No/Yes	89.5	Yes/Yes	107.5	Yes/Yes
D97 N	3.5	No/No	3.6	Yes/No	289.0	Yes/Yes
Y116A	9.6	No/Yes	15.7	Yes/Yes	3.5	Yes/Yes
D171 N	11.4	Yes/Yes	27.2	No/No	1248.6	No/Yes
D262 N	17.8	Yes/Yes	172.0	Yes/Yes	12.5	No/No
H281A	0.2	No/No	131.4	Yes/Yes	2.2	Yes/No
E288 N	35.0	Yes/Yes	68.7	Yes/Yes	460.3	Yes/Yes

[a] K_i fold increase compared with wild type [41]

[b] Predicted inter-atomic contacts between optimized binding poses of known CXCR4 antagonists and wild type residues (distance cut-off = 4 Å)

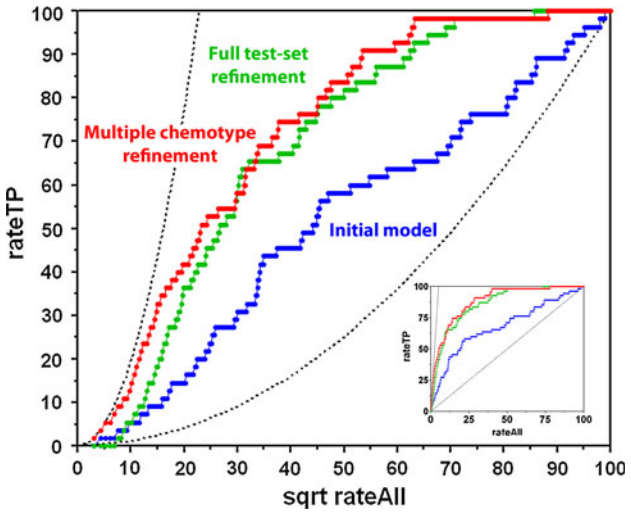

Fig. 5 Normalized square root ROC plots of an initial, non-refined, CXCR4 homology model and optimized models based on a full test-set of 55 antagonists or a multiple chemotype approach (5 test-sets with 11 compounds each). The inset shows a linear version of the ROC plot. Selected EN-NMA models have maximal enrichments for the test-sets. One thousand GLIDA ligands were used as decoys

model refined with the full test-set (linear AUC = 91.3 and NSQ_AUC = 72.7, Fig. 5). Early enchainment up to 15% of the full database is very steep with compounds from all chemotypes at the top positions of the ranked score list. Therefore, this model might be used for virtual screening of new CXCR4 antagonists.

The optimal virtual ligand screening performance obtained in this study with a multiple receptor representation can be related to a shift in the CXCR4 binding site upon ligand binding (commonly referred as the "induced fit" mechanism), highlighting the importance of flexibility and dynamics in protein–ligand docking [55]. According to this mechanism, different chemotypes can induce alternative conformational changes to the binding pocket, each one of these representing only a fraction of the total chemical molecular recognition properties. Although the use of multiple binding site representations in virtual screening is an attractive approach with potential better enrichments when compared to a single receptor representation, this method should be validated carefully because using a large number of receptor structures might also have the opposite effect and lead to an increased number of false positives [56].

Binding pockets in models optimized based on single chemotypes showed similar binding modes for the active compounds in study but larger conformational changes compared to the model obtained based on a full test-set optimization. CXCR4 antagonists establish similar contacts but are more buried within the binding pocket and surrounded by a more refined hydrogen bonding network. Slight differences with the binding mode of AMD3100 include a shorter binding distance to Trp94 and Tyr116, both residues known to be important for binding, whereas bound AMD3465 is located more distant from Asp97, also in agreement with know site-directed mutagenesis data (Table 1). Tetrahydroquinolinamines, represented in Fig. 4e with compound AMD070, presented however a different binding mode. The ring moieties are buried in a similar mode within the CXCR4 binding pocket, however, the positively charged alkyl amine points towards TM7, hydrogen bonding the negatively charged side chain of Asp288. This resulted in a better binding score and good agreement with known mutagenesis data (Table 1).

In summary, CXCR4 homology model refinement based on multiple test-sets with single chemotypes provided better enrichments in VLS runs, when combined in a multiple receptor representation, and binding poses in better agreement to known site directed mutagenesis data.

Conclusions

In this study we have described the optimization of a three dimensional homology model for the CXCR4 receptor using a ligand-guided multiple chemotype approach. We showed that known CXCR4 antagonists can be used to reshape the binding site of an initial unrefined model leading to alternative conformations with improved recognition properties which can be used for structure-based drug design. This method can be used for homology modeling refinement in the low sequence identity range.

Elastic network normal mode analysis extensively sampled the conformational flexibility within the CXCR4 binding pocket. Small compound binding to peptide-activated GPCRs differs from amine GPCRs due to larger, more flexible and solvent exposed binding pockets. The optimized models, benchmarked for early recognition properties of known CXCR4 antagonists, showed better recognition properties and provided predicted binding poses in good agreement to know site-directed mutagenesis data. Salt bridges with negatively charged residues within the CXCR4 binding pocket were important for binding of most compounds in study along with a network of hydrogen bonds and hydrophobic interactions.

The size and chemical diversity of the test-set used in this study was critically important for molecular recognition of the final CXCR4 models. Better early enrichments were obtained by combining five models optimized for single CXCR4 chemotypes which might indicate that different classes of compounds bind to slightly different conformations of the CXCR4 receptor. Binding poses described in this work can be used to guide the lead optimization of know antagonists, whereas, the optimized

CXCR4 homology model will be useful for virtual ligand screening of new, chemically diverse chemotypes.

Acknowledgments M.A.C. Neves thanks Fundação para a Ciência e a Tecnologia (FCT), Portugal, for a Post Doctoral grant (SFRH/BPD/64216/2009) and the Fulbright Scholar Program for financial support. The authors are grateful to Dr. R. Abagyan, Dr. I. Kufareva, Dr. V. Katritch, Dr. P. Lam and Dr. M. Rueda for their stimulating discussions and support.

References

1. Baggiolini M (1998) Nature 392:565
2. Zou YR, Kottmann AH, Kuroda M, Taniuchi I, Littman DR (1998) Nature 393:595
3. Tachibana K, Hirota S, Iizasa H, Yoshida H, Kawabata K, Kataoka Y, Kitamura Y, Matsushima K, Yoshida N, Nishikawa S, Kishimoto T, Nagasawa T (1998) Nature 393:591
4. Doranz BJ, Berson JF, Rucker J, Doms RW (1997) Immunol Res 16:15
5. Muller A, Homey B, Soto H, Ge NF, Catron D, Buchanan ME, McClanahan T, Murphy E, Yuan W, Wagner SN, Barrera JL, Mohar A, Verastegui E, Zlotnik A (2001) Nature 410:50
6. Taichman RS, Cooper C, Keller ET, Pienta KJ, Taichman NS, McCauley LK (2002) Cancer Res 62:1832
7. Nanki T, Hayashida K, El-Gabalawy HS, Suson S, Shi KR, Girschick HJ, Yavuz S, Lipsky PE (2000) J Immunol 165:6590
8. Abi-Younes S, Sauty A, Mach F, Sukhova GK, Libby P, Luster AD (2000) Circ Res 86:131
9. Steen A, Schwartz TW, Rosenkilde MM (2009) Mini-Rev Med Chem 9:1605
10. Burger JA, Peled A (2009) Leukemia 23:43
11. Lukacs NW, Berlin A, Schols D, Skerlj RT, Bridger GJ (2002) Am J Pathol 160:1353
12. Hernandez PA, Gorlin RJ, Lukens JN, Taniuchi S, Bohinjec J, Francois F, Klotman ME, Diaz GA (2003) Nature Genet 34:70
13. Balabanian K, Lagane B, Pablos JL, Laurent L, Planchenault T, Verola O, Lebbe C, Kerob D, Dupuy A, Hermine O, Nicolas JF, Latger-Cannard W, Bensoussan D, Bordigoni P, Baleux F, Le Deist F, Virelizier JL, Arenzana-Seisdedos F, Bachelerie F (2005) Blood 105:2449
14. Bridger GJ, Skerlj RT, Padmanabhan S, Martellucci SA, Henson GW, Struyf S, Witvrouw M, Schols D, De Clercq E (1999) J Med Chem 42:3971
15. Este JA, Cabrera C, De Clercq E, Struyf S, Van Damme J, Bridger G, Skerlj RT, Abrams MJ, Henson G, Gutierrez A, Clotet B, Schols D (1999) Mol Pharmacol 55:67
16. Rosenkilde MM, Gerlach LO, Jakobsen JS, Skerlj RT, Bridger GJ, Schwartz TW (2004) J Biol Chem 279:3033
17. Gerlach LO, Skerlj RT, Bridger GJ, Schwartz TW (2001) J Biol Chem 276:14153
18. Hendrix CW, Flexner C, MacFarland RT, Giandomenico C, Fuchs EJ, Redpath E, Bridger G, Henson GW (2000) Antimicrob Agents Chemother 44:1667
19. Hendrix CW, Collier AC, Lederman MM, Schols D, Pollard RB, Brown S, Jackson JB, Coombs RW, Gleshy MJ, Flexner CW, Bridger GJ, Badel K, MacFarland RT, Henson GW, Calandra G (2004) JAIDS 37:1253
20. Rosenkilde MM, Gerlach LO, Hatse S, Skerlj RT, Schols D, Bridger GJ, Schwartz TW (2007) J Biol Chem 282:27354
21. Skerlj RT, Bridger GJ, Caller A, McEachern EJ, Crawford JB, Zhou YX, Atsma B, Langille J, Nan S, Veale D, Wilson T, Harwig C, Hatse S, Princen K, De Clercq E, Schols D (2010) J Med Chem 53:3376
22. Gudmundsson KS, Sebahar PR, Richardson LD, Miller JF, Turner EM, Catalano JG, Spaltenstein A, Lawrence W, Thomson M, Jenkinson S (2009) Bioorg Med Chem Lett 19:5048
23. Gudmundsson KS, Boggs SD, Catalano JG, Svolto A, Spaltenstein A, Thomson M, Wheelan P, Jenkinson S (2009) Bioorg Med Chem Lett 19:6399
24. Moyle G, DeJesus E, Boffito M, Wong RS, Gibney C, Badel K, MacFarland R, Calandra G, Bridger G, Becker S (2009) Clin Infect Dis 48:798
25. Zhan WQ, Liang ZX, Zhu AZ, Kurtkaya S, Shim H, Snyder JP, Liotta DC (2007) J Med Chem 50:5655
26. Pettersson S, Perez-Nueno VI, Ros-Blanco L, de La Bellacasa RP, Rabal MO, Batllori X, Clotet B, Clotet-Codina I, Armand-Ugon M, Este J, Borrell JI, Teixido J (2008) Chem Med Chem 3:1549
27. Thoma G, Streiff MB, Kovarik J, Glickman F, Wagner T, Beerli C, Zerwes HG (2008) J Med Chem 51:7915
28. Thomas WD, Leleti MR, Pennell AMK (2007) US Patent 2007/0275965
29. Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Le Trong I, Teller DC, Okada T, Stenkamp RE, Yamamoto M, Miyano M (2000) Science 289:739
30. Park JH, Scheerer P, Hofmann KP, Choe HW, Ernst OP (2008) Nature 454:183
31. Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SGF, Thian FS, Kobilka TS, Choi HJ, Kuhn P, Weis WI, Kobilka BK, Stevens RC (2007) Science 318:1258
32. Warne T, Serrano-Vega MJ, Baker JG, Moukhametzianov R, Edwards PC, Henderson R, Leslie AGW, Tate CG, Schertler GFX (2008) Nature 454:486
33. Jaakola VP, Griffith MT, Hanson MA, Cherezov V, Chien EYT, Lane JR, IJzerman AP, Stevens RC (2008) Science 322:1211
34. Cavasotto CN, Orry AJW, Murgolo NJ, Czarniecki MF, Kocsi SA, Hawes BE, O'Neill KA, Hine H, Burton MS, Voigt JH, Abagyan RA, Bayne ML, Monsma FJ (2008) J Med Chem 51:581
35. Evers A, Klabunde T (2005) J Med Chem 48:1088
36. Evers A, Klebe G (2004) J Med Chem 47:5381
37. Katritch V, Rueda M, Lam PCH, Yeager M, Abagyan R (2010) Proteins 78:197
38. Abagyan RA, Batalov S (1997) J Mol Biol 273:355
39. Rosenbaum DM, Rasmussen SGF, Kobilka BK (2009) Nature 459:356
40. Cardozo T, Totrov M, Abagyan R (1995) Proteins 23:403
41. Wong RSY, Bodart V, Metz M, Labrecque J, Bridger G, Fricker SP (2008) Mol Pharmacol 74:1485
42. Laskowski RA, Macarthur MW, Moss DS, Thornton JM (1993) J Appl Crystallogr 26:283
43. Huang N, Shoichet BK, Irwin JJ (2006) J Med Chem 49:6789
44. Okuno Y, Tamon A, Yabuuchi H, Niijima S, Minowa Y, Tonomura K, Kunimoto R, Feng CL (2008) Nucleic Acids Res 36:D907
45. Halgren TA (1996) J Comput Chem 17:490
46. Hinsen K (1998) Proteins 33:417
47. Rueda M, Bottegoni G, Abagyan R (2009) J Chem Inf Model 49:716
48. Totrov M, Abagyan R (1997) Proteins 215
49. Triballeau N, Acher F, Brabet I, Pin JP, Bertrand HO (2005) J Med Chem 48:2534
50. Truchon JF, Bayly CI (2007) J Chem Inf Model 47:488
51. Strader CD, Fong TM, Tota MR, Underwood D, Dixon RAF (1994) Annu Rev Biochem 63:101
52. Reynolds KA, Katritch V, Abagyan R (2009) J Comput -Aided Mol Des 23:273

53. Perez-Nueno VI, Ritchie DW, Rabal O, Pascual R, Borrell JI, Teixido J (2008) J Chem Inf Model 48:509
54. Liang XY, Parkinson JA, Weishaupl M, Gould RO, Paisey SJ, Park HS, Hunter TM, Blindauer CA, Parsons S, Sadler PJ (2002) J Am Chem Soc 124:9105
55. Morra G, Genoni A, Neves MAC, Merz KM, Colombo G (2010) Curr Med Chem 17:25
56. Rueda M, Bottegoni G, Abagyan R (2010) J Chem Inf Model 50:186