J Comput Aided Mol Des (2011) 25:1–12
DOI 10.1007/s10822-010-9397-6

PERSPECTIVE

Challenges in the determination of the binding modes of non-standard ligands in X-ray crystal complexes

Alpeshkumar K. Malde · Alan E. Mark

Received: 8 October 2010 / Accepted: 25 October 2010 / Published online: 4 November 2010

Abstract Despite its central role in structure based drug design the determination of the binding mode (position, orientation and conformation in addition to protonation and tautomeric states) of small heteromolecular ligands in protein:ligand complexes based on medium resolution X-ray diffraction data is highly challenging. In this perspective we demonstrate how a combination of molecular dynamics simulations and free energy (FE) calculations can be used to correct and identify thermodynamically stable binding modes of ligands in X-ray crystal complexes. The consequences of inappropriate ligand structure, force field and the absence of electrostatics during X-ray refinement are highlighted. The implications of such uncertainties and errors for the validation of virtual screening and fragment-based drug design based on high throughput X-ray crystallography are discussed with possible solutions and guidelines.

Keywords X-ray crystallography · Ligand design · Molecular dynamics simulations · Free energy calculations · Binding mode

Abbreviations

aaRSs	Aminoacyl-tRNA synthetases
ATB	Automated Topology Builder
CDK	Cyclin Depdendent Kinase
CNS	Crystallography and NMR System
CR6	1-deoxy-1-acetylamino-β-D-gluco-2-heptulopyranosonamide
FE	Free Energy
GLG	α-D-glucopyranosyl-2-carboxamide
GPb	Glycogen Phosphorylase b
HIV-1	Human Immunodeficiency Virus-1
JG-365	Ac-Ser-Leu-Asn-Phe-Ψ[CH(OH)CH_2N]-Pro-Ile-Val-OMe
L-Ser	L-Serine
LIGA	(4-{4H-indeno[1,2-c]pyrazol-3-yl}pyridine)
LIGB	(4-{1H,4H-indeno[1,2-c]pyrazol-3-yl}pyridine)
MD	Molecular Dynamics
MM	Molecular Mechanics
Pab-NTD	N-terminal editing domain of *Pyrococcus abyssi* threonyl-tRNA synthetase
PDB	Protein Data Bank
PDE4B	Phosphodiesterase 4B
PNMT	Phenylethanolamine N-methyltransferase
QM	Quantum Mechanics
ROL	Rolipram
SC558	1-phenylsulfonamide-3-trifluoromethyl-5-*p*-bromophenylpyrazole
SKF	1,2,3,4-tetrahydro-isoquinoline-7-sulphonicacidamide
TetR	Tet repressor protein

Electronic supplementary material The online version of this article (doi:10.1007/s10822-010-9397-6) contains supplementary material, which is available to authorized users.

A. K. Malde · A. E. Mark (✉)
School of Chemistry and Molecular Biosciences,
University of Queensland, St. Lucia, QLD 4072, Australia
e-mail: a.e.mark@uq.edu.au

A. E. Mark
Institute for Molecular Bioscience, University of Queensland,
St. Lucia QLD 4072, Australia

Introduction

X-ray crystallography is an indispensable tool in structural biology and rational drug design. However, while the overall structure of the protein component within a given complex can be resolved in near atomic detail; the position, orientation and conformation of heteromolecules (small molecular ligands such as cofactors, substrates, inhibitors, drug molecules, etc.) are often much less certain [1]. In medicinal chemistry it is precisely these heteromolecules that are of primary interest and even slight errors in their structure, stereochemistry, tautomeric state, orientation or conformation can readily lead to the misinterpretation of biochemical mechanisms and/or the failure of computational drug design efforts [2–4].

Determining the structure of small heteromolecules bound within a large protein structure is challenging for two reasons. First, small non-covalently bound heteromolecules can show a higher degree of thermal motion or conformational disorder than the surrounding protein leading to less well-defined density. Second, during refinement the local conformation of the residues in the protein is primarily determined by geometric constraints such as imposed by the highly optimized parameters of Engh and Huber [5]. However, equivalent geometric constraints are not available for most heteromolecules. Instead refinement is generally based on a molecular mechanics (MM) description of the molecule(s). Frequently electrostatic interactions are neglected, as are possible alternative geometries. For this reason methods describing the molecules quantum mechanically (QM) using a mixed QM/MM [6] approach or a combined force field and shape potential approach are increasingly being proposed as alternatives to current approaches to facilitate the refinement of the geometries of heteromolecules.

Once the structure of the heteromolecule-protein complex has been refined, detecting whether the protonation state, tautomeric state, stereochemistry, orientation and/or conformation of the ligand is appropriate is a major challenge [7]. Standard crystallographic assessment tools cannot be used as the structures may fit within the density. Figure 1 shows an example in which a chiral drug Rolipram bound to the enzyme phosphodiesterase 4B [PDB 1RO6 (2.0 Å)] can be placed within the electron density in four different ways comprising of two stereoisomers each in two alternative orientations [8]. Such ambiguity in assigning the correct stereoisomer will always arise when a racemic mixture of the ligand is used for co-crystallization and either the relative binding energy of enantiomers is similar or unknown, as is the case in the example shown in Fig. 1. Furthermore, errors within the ligands cannot always be detected based on geometric and/or simple energetic criteria, as the structures are usually self-consistent. The upper panels in Fig. 2a, b shows an example in which two different crystal structures containing alternate conformations of the same ligand, specifically 1-deoxy-1-acetylamino-β-D-gluco-2-heptulopyranosonamide (CR6) bound to the enzyme glycogen phosphorylase b (GPb) have been published by the same authors in a 2 year period (PDB 1FU8 and 1P4H in 2003 and 2005) [9, 10]. As can be seen the alternate orientations of the amide group of the ligand are associated with alternate conformations of the amide group of the Asn side chain forming an equivalent hydrogen bond. While either appears reasonable, one structure is most likely preferred. In the other, an incorrect hydrogen bond is formed which will potentially have propagated an incorrect hydrogen bond network throughout the protein. The second example shown in the lower panels of Fig. 2c, d relates to the assignment of ‘O’ and ‘N’ in a sulphonamide group where two crystal structures of the same ligand SC558 [1-phenylsulfonamide-3-trifluoromethyl-5-*p*-bromophenylpyrazole] with the enzyme cyclooxygenase II were reported by the same group in a single 1996 manuscript [PDB 1CX2 (3.0 Å) and 6COX (2.8 Å)] [11]. The structures exhibit different crystallographic symmetry and were reported at a low resolution. At this resolution it is inherently difficult to assign the specific atoms and this example shows how such inconsistencies can easily be overlooked.

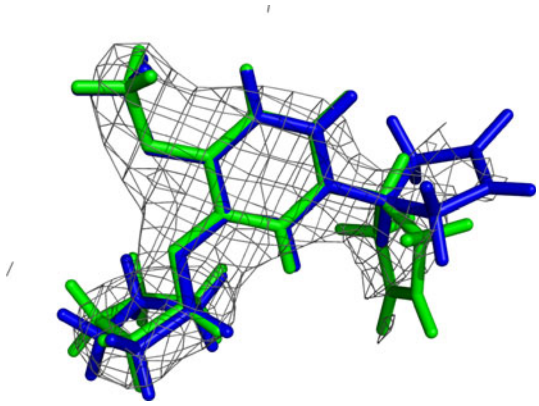


Fig. 1 2Fo-Fc map contoured at 1.0σ for the ligand rolipram (ROL) in a complex with phosphodiesterase 4B (PDE4B) from PDB 1RO6 (http://eds.bmc.ii.se/eds/). The *R* enantiomer of ROL in two different orientations (*blue* and green) is shown as reported in the PDB structure. The *S* enantiomer can also be fitted within the same density in both orientations

As illustrated above, for certain ligands, the thermodynamically preferred tautomer, stereoisomer, binding orientation and/or conformation cannot easily be distinguished based on the examination of the electron density or based on simple geometric or energetic criteria. If the ligand can be placed within the electron density changes in the binding modes will not affect global indicators such as R and R_{free} significantly. Furthermore, alternate binding modes can have complementary interactions with the surrounding environment ruling out the use of a simple energy based criteria to identify the preferred binding mode. Instead to identify the preferred binding state one must determine the state which corresponds to the lowest free

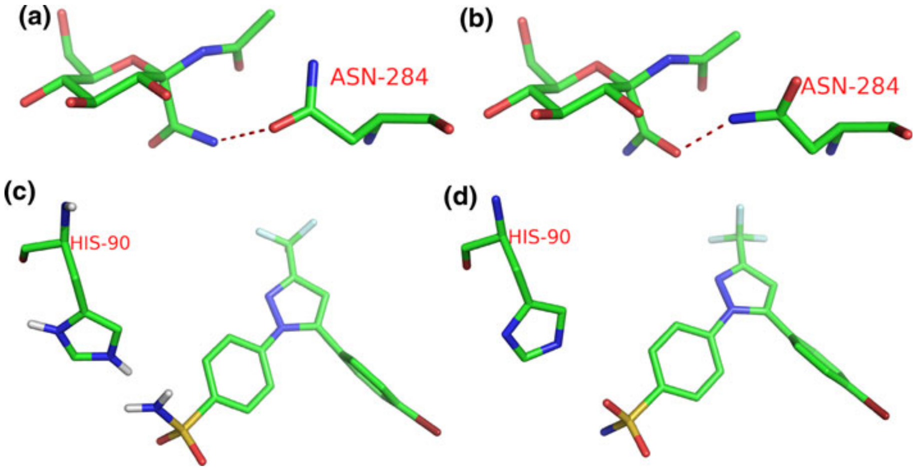


Fig. 2 Alternative conformations of the ligand CR6 (1-deoxy-1-acetylamino-β-D-gluco-2-heptulopyranosonamide) in two crystal structures **a** 1FU8 and **b** 1P4H reported by the same authors. The side chain of Asn284 is also shown illustrating compensating changes within the protein. Two alternative conformations of the ligand SC558 (1-phenylsulfonamide-3-trifluoromethyl-5-*p*-bromophenylpyrazole) in the crystal structure **c** 1CX2 and **d** 6COX reported in a single manuscript. His90, which forms a part of the binding site of the protein, is also shown. The alternate conformations reflect the uncertainty in the assignment of the 'O' (*red*) and 'N' (*blue*) atoms in the ligand. ('C' *green*, 'S' *yellow*, hydrogen atoms are not shown)

energy (FE), for example by using free energy perturbation approaches in association with molecular simulation techniques in which one calculates directly the difference in free energy between alternative states of the system. As the chemical properties of the ligand are unchanged in such calculations, force field considerations play a minor role [12]. In this perspective various examples taken from the literature as well as our own studies are used to illustrate the potential difficulties when determining the tautomeric state, stereochemistry, orientation and/or conformation of ligands within crystal complexes and strategies that can be used to determine the most appropriate solution.

Where the stereochemistry of the ligand is uncertain

One of the earliest studies in which free energy calculations were used to identify and validate the preferred stereoisomer of a ligand was the case of the interaction of the human immunodeficiency virus 1 (HIV-1) protease with the peptidomimetic inhibitor JG-365 (Ac-Ser-Leu-Asn-Phe-Ψ[CH(OH)CH_2N]-Pro-Ile-Val-OMe) in the structure PDB 7HVP (2.4 Å) [13]. Here, a racemic mixture of the ligand JG-365 was used containing both the *R* and *S* diastereomers at the chiral hydroxyethylamine carbon during crystallization. The authors modelled both the diastereomers in the experimental electron density during the refinement and based on an analysis of difference electron density maps, it was proposed that the protease exclusively bound the *S* diastereomer. Experimentally, the relative binding free energy between the *S* and *R* diastereomers was later shown to be 10.9 kJ/mol. The relative free energy of binding of these two diastereomers of JG-365 to HIV-1 protease was also calculated using the thermodynamic perturbation method by two different research groups employing different force fields and molecular dynamics programs [14, 15]. In both cases, the values calculated were in agreement with experiment and the study serves as a validation of the use of free energy calculations to distinguish between the bindings of alternative stereoisomers.

Where the orientation of the ligand is uncertain

In cases where the ligand is small or shows pseudo-symmetry the orientation after refinement can be easily biased by the initial placement of the model and care must be taken to examine a full range of possible alternatives. An example of such a case involves the preferred binding mode of the ligand L-Serine (L-Ser) in the binding pocket of the N-terminal editing domain of *Pyrococcus abyssi* threonyl-tRNA synthetase (Pab-NTD) which we reported recently [16]. The editing domain of aminoacyl-tRNA synthetases (aaRSs) prevents the misincorporation of noncognate amino acids and is essential for maintaining high fidelity in regard to both amino acid type and enantiomeric selectivity during the process of translation in protein biosynthesis. Pab-NTD binds to L-Ser, L-Cysteine and all D-amino acids. The binding mode of ligand L-Ser proposed in the X-ray crystal structure complex [PDB 2HKZ (2.1 Å), deposited in 2006] could explain the preferential binding of L-Ser over the structurally similar L-Thr but could not explain the enantiomeric selectivity of the enzyme [17]. In order to study the enantiomeric selectivity of aaRSs towards free amino acids, molecular dynamics (MD) simulations and FE calculations were used to examine the binding of L-Ser to Pab-NTD. The study revealed that the proposed orientation of L-Ser in the structure PDB 2HKZ was unstable. An alternative orientation of L-Ser within the binding site was suggested by the simulations. This orientation, in which the ligand was rotated by ~150° and translated slightly, was also compatible with the electron density. Not only was this alternate binding mode thermodynamically stable but it also could account for the fact that the binding of free amino acids is enantiomeric selective [16].

Where stereochemistry as well as orientation of the ligand is uncertain

In some cases both the stereochemistry and the orientation of the ligand will be unknown. In such cases the assumption of a specific stereochemistry or orientation may lead to the incorrect placement of the molecule within the complex. A case in point involves the chiral ligand noradrenochrome [(3R/3S)-3-hydroxy-2,3-dihydro-1H-indole-5,6-dione] binding to the enzyme phenylethanolamine N-methyltransferase (PNMT) in the structure PDB 3HCB (2.4 Å) (deposited in 2009) [18]. In this case, a racemic mixture of the ligand was used during crystallization and the relative binding energy between the enantiomers was unknown. The enzyme catalyzes the conversion of noradrenaline to adrenaline using the cofactor S-adenosyl-L-methionine. Specific inhibitors of PNMT are of therapeutic importance within the central nervous system [19].

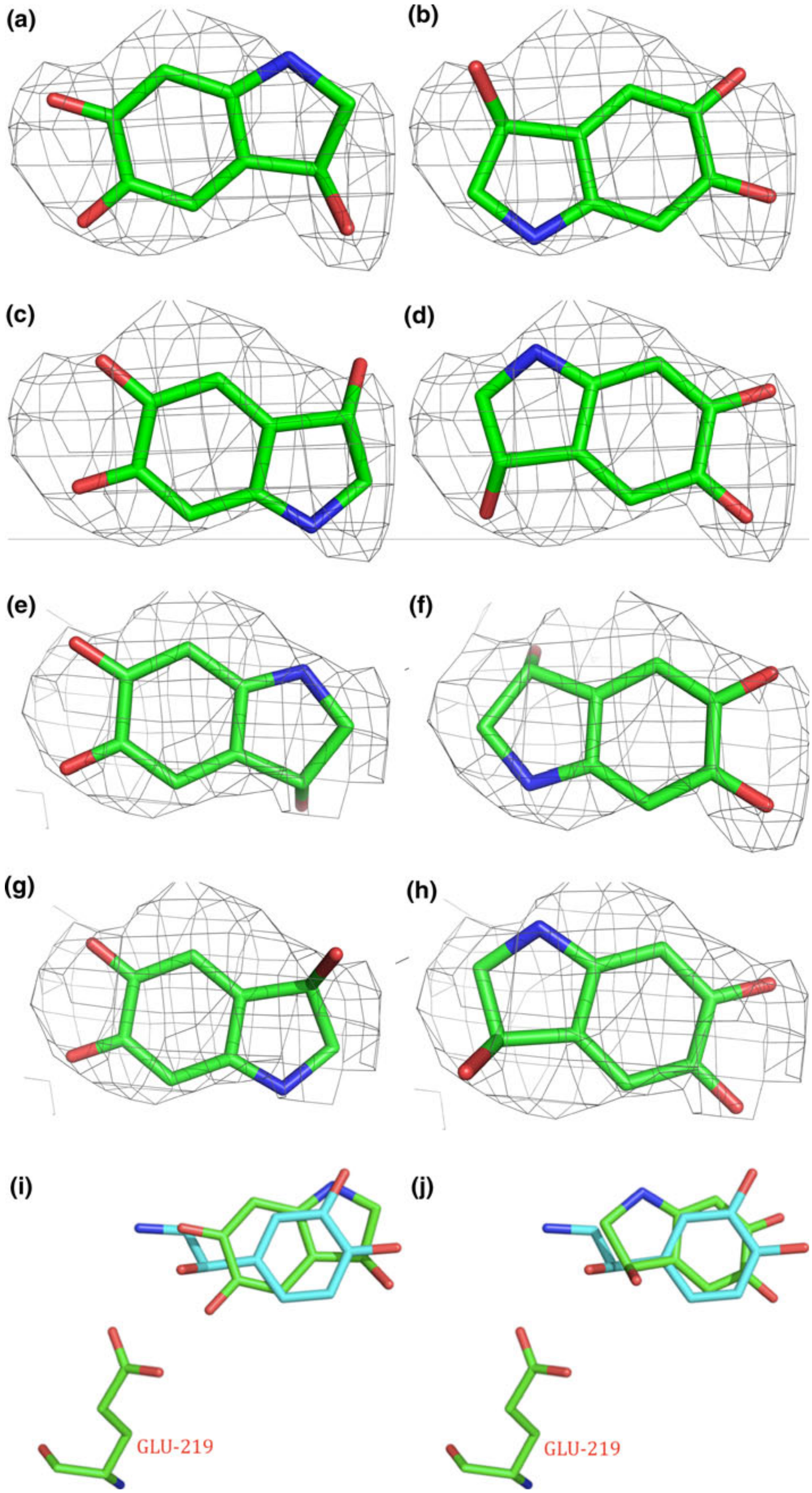
Noradrenochrome can be placed within the electron density in eight possible orientations (four for each enantiomer, Fig. 3). These will be referred to as AS (Fig. 3a), BS (Fig. 3b), CS (Fig. 3c), DS (Fig. 3d), AR (Fig. 3e), BR (Fig. 3f), CR (Fig. 3g) and DR (Fig. 3h) where 'S' and 'R' represent the stereochemistry at the single chiral center and A, B, C and D represent different binding modes. Although it was stated in the original manuscript that it was possible to fit both the enantiomers into the electron density, only the AS orientation (Fig. 3a) was deposited in the PDB. Nevertheless, the authors of the structure specifically discussed the significance of the fact that the proposed orientation is rotated by 180° about the short axis of its fused ring (Fig. 3i) compared to the binding mode of the related ligand *R*-noradrenaline as reported in the structure PDB 3HCD (2.39 Å) [18] and that one of the two quinone oxygens is in close proximity to the side chain of Glu219. In fact all eight binding modes of noradrenochrome can be placed in the electron density, and it is difficult to identify the preferred binding mode from the electron density alone. Using MD simulations and FE calculations we therefore attempted to confirm AS as the preferred binding mode of noradrenochrome in PNMT. The MD simulations of eight different PNMT:noradrenochrome complexes revealed that four complexes (AS, AR, CS and CR) including the one with the X-ray crystal binding mode (AS) were unstable. This is not surprising as the interaction between the quinone group of the ligand and the side chain carboxylate of Glu219 proposed in 3HCB would be expected to be unfavourable. The oxygen atoms of the quinone moiety of the ligand and of carboxylate of Glu219 are within hydrogen bonding distance but all four oxygen atoms are hydrogen bond acceptors, none of them are expected to be hydrogen bond donors under the conditions used during crystallization (pH 5.5–6.0). Thus, the binding mode given in the X-ray complex is likely to be inappropriate.

The remaining four binding modes (BS, BR, DS and DR) were stable with either the amine (BS and BR) or the hydroxyl group (DS and DR) of the ligand forming a stable hydrogen bond with the side chain carboxylate of Glu219. To identify the preferred binding mode amongst the stable complexes, the difference in FE between the alternative binding modes was determined by calculating all the legs of the thermodynamic cycle DR → DS → BS → BR → DR. As the free energy is a state function, the free energy for any closed thermodynamic cycle should be zero. The overall ΔΔG for the cycle was 1.8 kJ/mol indicating that the calculations are well converged. The relative binding FE difference between two modes (B and D) of *S*-noradrenochrome is relatively small to that of *R*-noradrenochrome. The calculations indicate that DR is the thermodynamically preferred binding mode with a free energy ~8 kJ/mol lower than the next lowest mode (BS) (Fig. 4). Thus the calculations suggest that the thermodynamically stable binding mode of noradrenochrome (DR) has the same stereochemistry as that of *R*-noradrenaline and that the binding modes of both compounds are similar (Fig. 3j) as would be expected based on their chemical structures.

Where the tautomer of the ligand is uncertain

One of the most commonly overlooked aspects while studying protein:ligand interactions is the consideration of possible tautomers, isomers of the molecule with alternate positions of hydrogen atoms in equilibrium, of the ligand molecule. The difficulty in the identification of the appropriate tautomer in the case of the imidazole ring of histidine is well known especially in medium to low resolution X-ray crystal structures. To assist in the selection of the appropriate tautomer of histidine tools such as MolProbity [20] have been developed which attempt to predict the appropriate tautomer based on the nature of local environment, specifically the presence of steric clashes and

Fig. 3 2Fo-Fc map contoured at 1.0σ for the ligand noradrenochrome in various binding modes: **a** AS (from PDB 3HCB, http://eds.bmc.ii.se/eds/), **b** BS, **c** CS, **d** DS, **e** AR, **f** BR, **g** CR and **h** DR, where 'S' and 'R' represent the stereochemistry at the single chiral center and A, B, C and D represent different binding orientations. ('C' *green*, 'O' *red*, 'N' *blue*; hydrogen atoms are not shown). **e** The binding mode of *S*-noradrenochrome (AS) as proposed in the crystal structure 3HCB ('C' *green*) as compared to the binding mode of *R*-noradrenaline as reported in the structure 3HCD ('C' cyan). **f** The thermodynamically stable binding mode of noradrenochrome (DR) as predicted based on FE calculations ('C' *green*) which is similar to that of *R*-noradrenaline in 3HCD ('C' *cyan*)



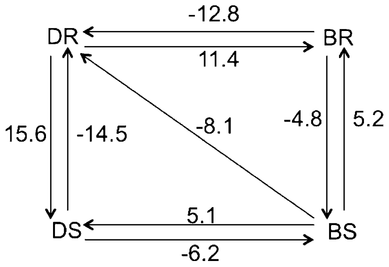


Fig. 4 The relative free energy of binding (kJ/mol) for the alternate modes of noradrenochrome to the protein phenylethanolamine N-methyltransferase (PNMT). The thermodynamically most stable binding mode is DR

potential hydrogen bonding interactions. Tools such as Marvin (http://www.chemaxon.com) are also available which attempt to enumerate all possible tautomers within a ligand molecule and estimate their relative stability. However, to truly determine the preferred tautomer either free in solution or bound to the protein one must again determine the relative free energy. A simple example to demonstrate how this can be achieved involves the interaction of cyclin-dependent kinase (CDK) with a pyrazole analogue which can exist in two tautomeric forms (Fig. 5): LIGA (4-{4H-indeno[1,2-c]pyrazol-3-yl}pyridine) and LIGB (4-{1H,4H-indeno[1,2-c]pyrazol-3-yl}pyridine) in the PDB 1JVP (1.53 Å, deposited in 2001). In the X-ray complex, both the tautomers could be fitted within the electron density and based on the apparent occupancies it was suggested that the both the tautomers were equally populated and both were deposited in the PDB [21]. The difference in FE between the two tautomers LIGA and LIGB was calculated in water and CDK. The relative FE of binding between LIGA and LIGB was estimated to be 7.0 ± 0.3 kJ/mol, indicating that the tautomer LIGA is preferred over LIGB in the bound state and that the occupancies could reflect motion of the ligand in the pocket as opposed to the two tautomers having equal probability.

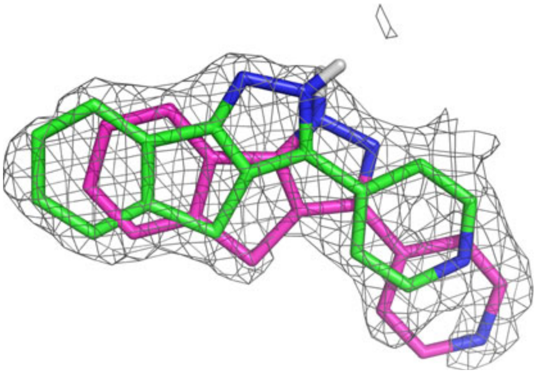


Fig. 5 2Fo-Fc map contoured at 1.0σ for the active site of cyclin-dependent kinase (CDK) enzyme with tautomer LIGA (4-{4H-indeno[1,2-c]pyrazol-3-yl}pyridine) and tautomer LIGB (4-{1H, 4H-indeno[1,2-c]pyrazol-3-yl}pyridine) as reported in PDB 1JVP (http://eds.bmc.ii.se/eds/). ('C' *green* for LIGA and magenta for LIGB, 'N' *blue*, polar 'H' *white*)

Where the protonation state of the ligand is uncertain

The protonation state and the overall charge of a molecule with titratable groups will vary depending on the pH of the medium and the pKa of titratable group in the given medium. Again the protonation state of a ligand is a thermodynamic property and cannot always be inferred from a given structure [22]. Many ligand molecules exist as different protomers. One well-studied example is the antibiotic tetracycline. The binding of tetracycline to Tet Repressor protein (TetR) in gram-negative bacteria is associated with antibiotic resistance. Tetracycline exists in two main protonation states at neutral pH, a neutral form and a zwitterionic form. To identify which was the most thermodynamically stable protomer Aleksandrov et. al. [23] performed free energy calculations in which the protonation state and the conformation of residues in the binding pocket of TetR in the X-ray crystal complex PDB 2TRT (2.5 Å) were varied. The study revealed that the zwitterionic form of tetracycline is thermodynamically the more stable both in free state as well as when bound to TetR. The results from these FE calculations were later used to facilitate the refinement of the X-ray crystal complex of TetR with doxycycline, a structural isomer of tetracycline in which a hydroxyl group is in an alternative position (PDB 2O7O, 1.89 Å). The study illustrates how a systematic evaluation of the thermodynamic properties of a given system can reduce the uncertainties in the placement of a related ligand prior to X-ray refinement.

Where the conformation of the ligand is uncertain

In a medium to low resolution X-ray crystal structure, it is difficult to distinguish similar atoms such as 'O' and 'N' based on the electron density alone. This leads to a degree of uncertainty in identifying the preferred conformations of simple groups such as an amide or a sulfonamide. There are ~1,000 structures in the PDB where the ligand contains a free amide group and ~200 structures where the ligand contains a free sulphonamide group. All have resolutions between 1.5 and 3.0 Å where the assignment of the 'O' and 'N' atoms is potentially problematic and the surrounding hydrogen-bond network uncertain. That it is difficult to correctly assign the 'O' and 'N' atoms in the side chains of Asn and Gln in protein crystal structures is well known [20, 24–26]. As mentioned previously, tools such as MolProbity [20] have been developed which attempt to predict the orientation of the amide groups in Asn and Gln in proteins based on local interactions. However, similar tools for heteromolecules are not widely available. The importance of the correct assignment of the ‘O’ and ‘N’ atoms in free amide and free sulphonamide groups will be illustrated using two test systems: glycogen phosphorylase b (GPb) and phenylethanolamine N-methyltransferase (PNMT) which contain ligands with a free amide and a free sulphonamide group, respectively.

Case 1: the conformation of an amide group

Multiple crystal structures of glycogen phosphorylase b (GPb), an important target in the treatment of diabetes, complexed with a wide variety of glucose-based inhibitors have been reported and many of these are deposited in the Protein Data Bank (PDB). Here we will just consider the conformation of the amide group of α-D-glucopyranosyl-2-carboxamide (GLG) in PDB 1GG8 (2.31 Å, deposited in 2000) [27]. GLG contains a free amide group attached to the C2β atom of glucose and the problem is analogous that shown in Fig. 2a and b.

The quantum mechanical (QM) potential energy profile at the B3LYP/6-31G* level of theory for rotation around the ring_O – C2β – C(O)NH$_2$ – amide_O dihedral angle [θ] of an isolated GLG molecule in implicit solvent is given in Fig. 6a. As can be seen from Fig. 6a, the QM profile suggests there are three possible minima on the potential energy surface. In the global minimum (designated QM minimum 1) the amide group lies co-planar with the C2β-O bond of the sugar ring and amide ‘O’ and ring ‘O’ are *syn*-periplanar with $\theta = 0°$. A second minimum (QM minimum 2) with $\theta = 300°$ lies about 5 kJ/mol higher and a third minimum (QM minimum 3) with $\theta = 240°$ lies about 12 kJ/mol higher than the global minimum. The value of this angle in the crystal structure 1GG8 which was solved at 2.4 Å resolution is $\theta = 120°$ meaning the ligand sits at about 25 kJ/mol energy higher than the global minimum for the isolated ligand with the plane of the amide group dividing the sugar ring into two halves.

In an attempt to validate the conformation proposed in the crystal structure, the structure of the ligand GLG in the crystal structure 1GG8 was submitted to the ValLigURL [28] server. ValLigURL is linked to the HicUp [4] server and is designed to provide refined geometries for heteromolecules reported in the PDB as well as topology and parameter files which can be used for X-ray refinement. Specifically, the ValLigURL server aims to provide an optimal geometry for the heteromolecule. However, the conformer provided in this case sits even higher on the potential energy surface. The structure from the ValLigURL server is similar to the global minimum except that the ‘O’ and ‘N’ of the amide group are interchanged.

The potential energy profile depicted in Fig. 6a reflects the conformational preference of the free ligand using an implicit solvent model and based on this alone it is not possible to distinguish which, if any, of the five different conformations shown in Fig. 6a actually bind to protein GPb. However, since the conformer obtained from the ValLigURL server sits at >40 kJ/mol higher than the global minimum of the free ligand and had not been fitted to the electron density, the subsequent work will focus on the three minima on the QM potential energy surface and the conformation proposed in the crystal structure. A series of free energy perturbation calculations using the thermodynamic integration approach were performed to determine which of these four conformations was the preferred binding mode. The difference in binding free energy between pairs of conformers was determined by calculating

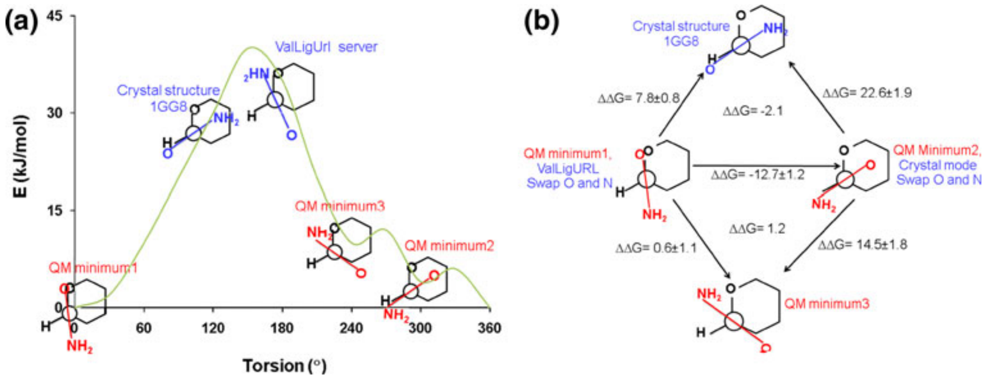


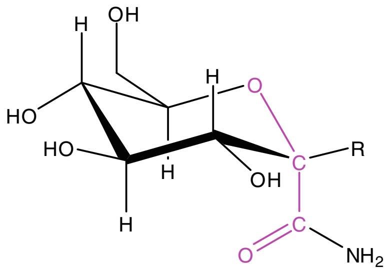
Fig. 6 **a** The quantum mechanical potential energy profile at the B3LYP/6-31G* level of theory in implicit water for the rotation around the C2β-C(O)NH$_2$ dihedral angle in the molecule GLG [glucose-2β-carboxamide]. The structures corresponding to the different minima and as found in specific crystal structures are also shown. Note, the hydroxyl groups and hydrogen atoms of the ring have been omitted for clarity. **b** The relative free energy of binding (kJ/mol) for the alternate conformations of the ligand GLG to the protein Glycogen Phosphorylase b (GPb). The excess FE in each clock-wise thermodynamic cycle is indicated

the difference in free energy in explicit water and in the protein as described in the methods.

The relative free energy of binding ΔΔG between each of the pairs of conformations of the ligand GLG is depicted in Fig. 6b. The fact that the thermodynamic cycles shown in Fig. 6b close to within −2.1 kJ/mol suggests that the calculations are well converged. The relative free energy of binding of the X-ray conformer is between 8 and 23 kJ/mol higher than the three minima observed in the free state. This clearly indicates that the conformer observed in the crystal structure is inappropriate. The preferred bound conformation is in fact 'QM minimum 2'. Free in solution the difference in potential energy between QM minima 2 and 3, which differ by a rotation of the dihedral angle θ of only 60° is approximately 7 kJ/mol. However, the difference in binding free energy is −14.5 kJ/mol. In addition, despite QM minimum 3 being approximately 12 kJ/mol higher in energy than the global minimum the difference in binding FE between global minimum and QM minimum 3 is negligible. In this case the global energy minimum of the free ligand in vacuum is not the preferred conformation when bound to the active site of the protein. In fact, the preferred bound conformation of GLG when bound to the protein is equivalent to the crystal structure but with the 'O' and 'N' atoms of the amide group interchanged. The difference in free energy associated with the miss-assignment of the 'O' and the 'N' in this case is about 23 kJ/mol.

In the case of GPb, illustrated in Fig. 2a and b, the structures of a series of related glucose analogues each containing a free amide group attached to the C2β carbon bound to GPb have been solved. In these structures (1FS4, 1B4D, 1FU8, 1GG8, 1P4H and 1P4G) a range of alternative conformations are proposed for the amide group as shown in Table 1. As can be seen three structures show the ligand in QM minimum 1 ($\theta = 0°$), one as proposed by the ValLigURL server which is equivalent to QM minimum 1 but with the 'O' and the 'N' atoms interchanged ($\theta = 180°$), one in QM minimum 3 ($\theta = 240°$) and the high energy structure found in 1GG8 ($\theta = 120°$). None were in the minimum free energy state identified in this study.

Table 1 A list of PDB structures containing glucose analogue inhibitors of GPb containing amide groups in various conformations



No.	*R*	Dihedral θ° [O_C_C_O]	PDB
1.	$-NHCO_2CH_3$	0°	1FS4, 1B4D
2.	$-NHCOCH_3$	0°	1FU8
		180°	1P4H
3.	–H	120°	1GG8
4.	$-N_3$	240°	1P4G

Case 2: the conformation of a sulphonamide group

This example involves the conformation of the sulphonamide group in the inhibitor 1,2,3,4-tetrahydro-isoquinoline-7-sulphonicacidamide (SKF) bound to the enzyme phenylethanolamine N-methyltransferase (PNMT) in the structure PDB 1HNN (2.4 Å, deposited in 2001) [29]. SKF is comprised of a tetrahydroisoquinoline ring substituted with a sulphonamide group at the 7-position. The conformational preferences of aromatic sulphonamide groups have been studied previously using gas phase electron diffraction and high-level QM calculations [30]. Benzene sulphonamide has two degenerate energy minima in which the S–N bond lies perpendicular to the aromatic ring with the dihedral angle being either 90° or −90°. The barrier to the rotation about the C–S bond is ~8.0 kJ/mol with the highest energy conformation (0° and 180°) being when the S–N bond is coplanar with aromatic ring. Since the two minima are degenerate free in solution, SKF was placed in the active site of PNMT in both conformers.

From Fig. 7a and b it can be seen that that the two alternative conformations fit easily within the experimental electron density (PDB: 1HNN, 2.4 Å resolution). As based on the density alone it is not possible to determine the preferred conformation, free energy calculations were again performed. The difference in the free energy of binding between the conformation observed in the crystal (conformation A) and an alternate conformation (conformation B) in which the sulphonamide group was rotated by 180° was estimated to be −38.7 ± 2.2 kJ/mol in favour of conformation B. In the crystal structure (Fig. 7c), the sulphonamide group of SKF lies such that the S–N bond projects beneath the plane of the ring. This orientation is compatible with the formation of hydrogen bonds between the 'O' and the 'N' of the ligand and side chain amino group of Lys57 as well as a number of water-mediated hydrogen bonds (not shown). However, one of the sulphonamide 'O' atoms is in close proximity (3.2 Å) to the side chain 'O' of Asn39, potentially an unfavourable interaction. In the alternate conformation (conformation B) given in Fig. 7d, the two sulphonamide 'O' atoms can form hydrogen bonds with Lys57 side chain while the 'N' can form a hydrogen bond with the 'O' atom of the Asn39 side chain. The other interactions are essentially unaffected. In

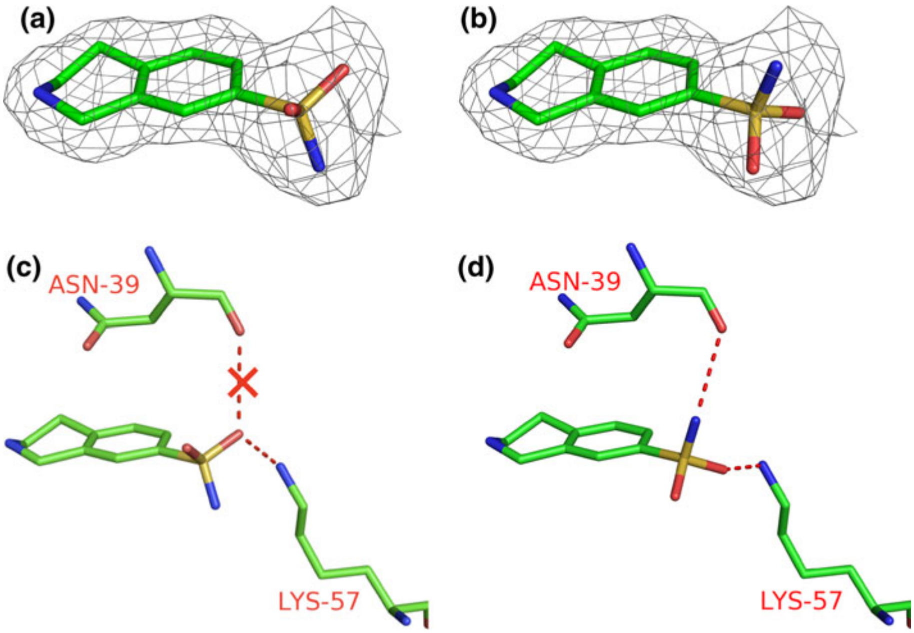


Fig. 7 **a** 2Fo-Fc map contoured at 1.0σ for the ligand SKF (1,2,3,4-tetrahydro-isoquinoline-7-sulphonicacidamide) in (**a**) PDB 1HNN (http://eds.bmc.ii.se/eds/) and **b** an alternate conformation suggested by molecular dynamics simulations and free energy calculations. **c** The proposed binding mode of the ligand SKF complexed with PNMT in the structure 1HNN. **d** The preferred binding mode of SKF in PNMT as suggested by the MD and FE calculations. (‘C’ *green*, ‘O’ *red*, ‘N’ *blue*, ‘S’ *yellow*)

fact the ligand was observed to spontaneously convert from conformation A, as proposed in 1HNN, to conformation B during equilibrium MD simulations. This alternate conformation of the ligand was stable during multiple MD simulations performed using different initial conditions.

Perspective and outlook

All the examples discussed in this work are relatively simple. In each case the alternative binding modes that should have been considered are obvious and where careful examination of the structures proposed might have alerted the authors to potential problems. It must also be stressed that the aim of this work is not to cast doubt on specific structures but to illustrate strategies that might be used to avoid potential errors. The systems investigated were chosen specifically because the differences in the binding modes were trivial and because a series of crystal structures of closely related compounds bound to the same protein were publicly available. It could also be argued that the consequences of any errors in these structures would be small. However, the simplicity of the cases serves to underline the ease with which errors can be made when it is assumed, for example, that the bound conformation of the ligand corresponds to the free energy minimum of the ligand in vacuum (or free in solution).

Structures of the ligands that are used during X-ray refinement are commonly generated using automatic procedures based on a very crude description of the molecule concerned and often exhibit unrealistic strain energy [31]. Normally such procedures provide a single conformation even if there are degenerate energetic states. Consideration of alternative tautomers, stereoisomers, orientations and conformations that are compatible with the density is of course critical. The accuracy of the proposed ligand will ultimately depend on the force field used during crystallographic refinement to describe not only the protein and the ligand but also the protein:ligand interactions. Whereas the parameters used for the refinement of the protein may be highly optimized this is not the case for small heteromolecules. Furthermore, the consideration of long-range electrostatic interactions between the protein and the ligand and not just immediate contacts is essential if the types of errors highlighted in this work are to be avoided.

Force field for heteromolecules

Various tools are available to generate force field descriptions for heteromolecules that can be used in crystallographic refinement. These include Hess2FF [32], PRODRG [33], XPLO2D [34], Antechamber [35], GENRTF [36], ATB (http://compbio.biosci.uq.edu.au/atb/), etc. Hess2FF provides parameters and a topology file for use with the program CNS (Crystallography and NMR System) based on a Hessian (force constant) matrix derived from molecular mechanics, semi-empirical or quantum mechanical calculations after geometry optimization at a given level of theory. The force field in this case is intended to describe local fluctuations around a specific geometry. It does not provide terms to model the electrostatic and van der Waals interactions between the ligand and the protein and is unsuitable for molecular simulations. PRODGR provides parameters and topologies for use with a variety of programs based on a several alternative force fields using a rule-based approach. However, little information in regard to how the force-field parameters are actually selected is provided. XPLO2D also uses a rule-based approach. It has a small set of default values for the force constants and generates a force field and topology for use with the program CNS based on a geometry supplied by the user. Antechamber generates topologies compatible with the GAFF (General AMBER Force Field, all-atom) based on a set of rules. In this case, the atomic charges are derived by fitting to the restrained electrostatic potential obtained from QM calculations. GENRTF generates parameters and a topology compatible with the CHARMM force field again based on a geometry supplied by the user using a rule based approach. The Automated Topology Builder (ATB) generates a topology and parameter set based on the GROMOS force field in a variety of formats including GROMOS, GROMACS and CNS. The ATB uses a QM optimized geometry and combines QM calculations with a rule based approach. Initial atomic charges are derived by fitting to the QM electrostatic potential but these are later adjusted to account for molecular symmetry and scaled to better reproduce solvation free energies as well as to ensure compatibility with the remainder of the force field. A QM derived Hessian matrix is also used to facilitate the selection of bonded parameters from a list of allowed types.

Alternatively, QM based methods can be used to describe directly the ligand molecule and surrounding interactions in a QM/MM protocol during refinement. For example, a systematic analysis of the preferred conformation of benzamidinium, the protonated form of benzamidine, bound to various proteins has been reported by Xue Li et. al. [37] using a combination of QM/MM based X-ray refinement and MD simulations. Benzamidinium and its derivatives are inhibitors of a wide range of serine proteases. The conformation of the benzamidinium group is a critical determinate of the interaction of the inhibitor with the protein as it forms a salt-bridge with the side chain of either an Asp or Glu residue within the binding cavity of the protein. The strength of this interaction is governed by the relative orientation of guanidinium group with respect to benzene ring. There are 87 crystal structures with a total of 153 benzamidinium moieties in the PDB. An analysis of the PDB indicated that the majority of the structures with benzamidinium-containing compounds have the benzamidinium group lying in plane with the ring. However, the QM profile of free benzamidinum and the results of QM/MM refinement of crystal structure complexes containing a benzamidinium derivative suggest that a range of twisted non-planar conformations are in fact preferred. This study further highlights the importance of the correct representation of the ligand as well as correct treatment of the environment during crystallographic refinement.

As demonstrated above, the geometric parameters and the force field used during crystallographic refinement play a major role in determining the final structure of the complex. As such there is a pressing need to publish the force field used to describe the heteromolecule together with the crystal structure in order that the structure can be validated. Equally important is the need to describe how specific physical interactions are modelled. Electrostatic interactions are commonly ignored during refinement despite the fact that they are critical to determine the correct orientation of water networks and catalytically relevant polar hydrogens (Asn, Gln and His side chains) [38]. Several of the errors highlighted in this study could have been avoided simply by the inclusion of electrostatic interactions during refinement.

Ligand structure vs. electron density

As X-ray refinement becomes ever more automated and dominated by the use of particular programs and protocols, the potential for critical aspects, such as the choice of force field and whether or not electrostatic interactions are included during refinement, to be ignored grows. When dealing with heteromolecules consideration of following four aspects could greatly reduce the uncertainty in structure fitted to the electron density: (i) whether the geometry including stereochemistry, protonation state and tautomeric state of the molecule is appropriate (ii) whether there are alternate conformations, orientations, stereoisomers and/or tautomers that could fit within the density (iii) the quality of the force field description of the molecule, and (iv) whether the interactions of the heteromolecule with the surrounding environment (macromolecule including crystallographic water and other heteromolecules) are described appropriately. In addition, in medium to low resolution X-ray structures, it is difficult to identify locations of the hydrogen atoms. A consideration of possible alternative arrangements of the hydrogen atoms on the ligand as well as on the surrounding residues in the protein could assist to identify alternative hydrogen bonding patterns and alternate tautomeric states. The consideration of such alternatives could have helped avoid the discrepancies in the orientation (PDB 3HCB, ligand noradrenochrome) and conformation (PDB 1HNN, ligand SKF) of the ligands in the PNMT X-ray crystal complexes. In particular it should be noted that the preferred tautomeric state [39] as well as the preferred conformation [37] of the bound ligand may differ significantly from that of the isolated ligand free in solution. Also, the ligand may exhibit multiple thermodynamically stable binding modes with a small difference in relative binding FE [40], as shown for the *S*-noradrenchrome:PNMT complex (Fig. 4).

Implications for virtual screening

In the case of the amide and the sulphonamide containing ligands examined in this study the conformations proposed in the crystal structures were approximately 20 and 40 kJ/mol higher than the preferred conformation, respectively. Clearly such large deviations would adversely affect attempts to use these structures in computational drug design. There can be little question that the variation and the associated uncertainty in the conformations of the ligands in the X-ray structures of protein:ligand complexes highlighted in this work have wide-spread implications for the development and validation of docking algorithms as well as the statistical analysis of ligand preferences based on structures in the protein databank. For example as electrostatics are frequently ignored during refinement it is not surprising that electrostatic terms are also neglected in empirical scoring functions such as LUDI [41], ChemScore [42], X-Score [43], AIScore [44], etc. as these algorithms are validated in part based on their ability to predict the structures of crystal complexes. However, in tests of the power of a given model to discriminate between native and decoy structures the inclusion of electrostatic terms in the scoring function e.g. QM Score [45] improves the predictive power. Docking algorithms in general are validated based on their ability to reproduce the binding modes of ligands observed in crystal structures. If the comparison between the theoretical model and the experimentally derived structure is based on atom and geometry specific measures such as the RMSD (the root mean-square distance) between the docked conformation and the crystal structure, any uncertainty or bias in the crystal structure will be reflected in the docking algorithm. This problem can be partially avoided by comparing the predicted structures more directly to an experimental observable [46] such as the real space R-factor (RSR) [47] which is measure of how well a ligand fits the electron density. However, this approach will not completely avoid difficulties associated with correlations between the structure of the ligand and the structure of the surrounding protein leading, for example, to inappropriate hydrogen bond networks.

Outlook

Uncertainties in the binding mode of, in particular, small ligands has major implications for fragment-based drug design based on X-ray crystallography. The medium to low affinity of the compounds combined with their small size and low to medium resolutions at which the structures are solved make it especially difficult to identify the correct orientation, conformation, stereochemistry, protomer and tautomeric state of the fragment molecules. Thus, even though specific fragments can be identified, their precise binding mode may be uncertain impacting on subsequent design studies.

An alternate way to deal with uncertainties in the binding modes of ligands would be to identify an ensemble of structures with alternate tautomeric states, protonation states, stereochemistry, orientations and conformations, compatible with the experimental electron density that could be deposited with the coordinates of the protein. For example this was done in the case of the drug Rolipram (ROL) bound to the enzyme phosphodiesterase 4B (Fig. 1), where multiple structures of ROL, with alternate stereochemistry and orientations, were deposited in the PDB. Likewise, in the case of cyclin-dependent kinase (PDB 1JVP) two tautomeric forms of the ligand were deposited (Fig. 5). The computational time and effort required to generate an ensemble of ligand structures will in many cases be trivial. The range of structures in the ensemble would in such cases represent the range of structures that should be considered when interpreting other sets of experimental data or theoretical predictions.

Conclusions

X-ray crystallography plays central role in structure based drug design. This said the determination of the structure, orientation and conformation of small heteromolecular ligands in protein:ligand complexes based on medium resolution X-ray diffraction data can be highly challenging. In this work a series of cases in which alternative binding modes of the ligand molecules in X-ray complexes may have been overlooked have been highlighted. In addition we have shown that the free energy cost associated with the inappropriate ligand binding modes can easily be in the order of 10’s of kJ/mol. Even in these simple cases, the preferred stereoisomer, orientation, tautomer, protomer and conformation could not be determined based on the electron density or simple energetic criteria once the structure was refined. The work underlines the importance of using an appropriate description of the heteromolecule (including electrostatic interactions) during refinement and illustrates how strategies based on FE calculations in conjunction with MD simulations can be used to identify the preferred binding modes of heteromolecules in X-ray crystal complexes. Finally we have proposed a simple set of guidelines that can be used to facilitate the correct identification of structures used for the interpretation of biochemical mechanisms and for structure-based drug design, in particular fragment-based drug design and virtual screening.

Acknowledgment The computational resources provided by NCI National Facility through projects m72 and n63 are greatly acknowledged.

References

1. Kleywegt GJ, Henrick K, Dodson EJ, van Aalten DMF (2003) Structure 11:1051
2. Davis AM, Teague SJ, Kleywegt GJ (2003) Angew Chemie-Int Ed 42:2718
3. Davis AM, St-Gallay SA, Kleywegt GJ (2008) Drug Discov Today 13:831
4. Kleywegt GJ (2007) Acta Crystallogr D Biol Crystallogr 63:94
5. Engh RA, Huber R (1991) Acta Crystallogr A 47:392
6. Yu N, Li X, Cui GL, Hayik SA, Merz KM (2006) Protein Sci 15:2773
7. Wlodek S, Skillman AG, Nicholls A (2006) Acta Crystallogr D Biol Crystallogr 62:741
8. Xu RX, Rocque WJ, Lambert MH, Vanderwall DE, Luther MA, Nolte RT (2004) J Mol Biol 337:355
9. Chrysina ED, Oikonomakos NG, Zographos SE, Kosmopoulou MN, Bischler N, Leonidas DD, Kovacs L, Docsa T, Gergely P, Somsak L (2003) Biocatal Biotransformation 21:233
10. Watson KA, Chrysina ED, Tsitsanou KE, Zographos SE, Archontis G, Fleet GWJ, Oikonomakos NG (2005) Proteins 61:966
11. Kurumbail RG, Stevens AM, Gierse JK, McDonald JJ, Stegeman RA, Pak JY, Gildehaus D, Miyashiro JM, Penning TD, Seibert K, Isakson PC, Stallings WC (1996) Nature 384:644
12. Villa A, Zangi R, Pieffet G, Mark AE (2003) J Comput Aided Mol Des 17:673
13. Swain AL, Miller MM, Green J, Rich DH, Schneider J, Kent SBH, Wlodawer A (1990) Proc Natl Acad Sci USA 87:8805
14. Ferguson DM, Radmer RJ, Kollman PA (1991) J Med Chem 34:2654
15. Tropsha A, Hermans J (1992) Protein Eng 5:29
16. Malde AK, Mark AE (2009) J Am Chem Soc 131:3848
17. Hussain T, Kruparani SP, Pal B, Dock-Bregeon AC, Dwivedi S, Shekar MR, Sureshbabu K, Sankaranarayanan R (2006) EMBO J 25:4152
18. Drinkwater N, Gee CL, Puri M, Criscione KR, McLeish MJ, Grunewald GL, Martin JL (2009) Biochem J 422:463
19. Gee CL, Drinkwater N, Tyndall JDA, Grunewald GL, Wu Q, McLeish MJ, Martin JL (2007) J Med Chem 50:4845
20. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB, Snoeyink J, Richardson JS, Richardson DC (2007) Nucleic Acids Res 35:W375
21. Furet P, Meyer T, Strauss A, Raccuglia S, Rondeau JM (2002) Bioorg Med Chem Lett 12:221
22. Donnini S, Villa A, Groenhof G, Mark AE, Wierenga RK, Juffer AH (2009) Proteins 76:138
23. Aleksandrov A, Proft J, Hinrichs W, Simonson T (2007) Chembiochem 8:675
24. McDonald IK, Thornton JM (1995) Protein Eng 8:217
25. Weichenberger CX, Sippl MJ (2006) Structure 14:967
26. Word JM, Lovell SC, Richardson JS, Richardson DC (1999) J Mol Biol 285:1735
27. Watson KA, Mitchell EP, Johnson LN, Son JC, Bichard CJF, Orchard MG, Fleet GWJ, Oikonomakos NG, Leonidas DD, Kontou M, Papageorgioui A (1994) Biochemistry 33:5745
28. Kleywegt GJ, Harris MR (2007) Acta Crystallogr D Biol Crystallogr 63:935
29. Martin JL, Begun J, McLeish MJ, Caine JM, Grunewald GL (2001) Structure 9:977
30. Petrov V, Petrova V, Girichev GV, Oberhammer H, Giricheva NI, Ivanov S (2006) J Org Chem 71:2952
31. Perola E, Walters WP, Charifson PS (2004) Proteins 56:235
32. Nilsson K, Lecerof D, Sigfridsson E, Ryde U (2003) Acta Crystallogr D Biol Crystallogr 59:274
33. Schuttelkopf AW, van Aalten DMF (2004) Acta Crystallogr D Biol Crystallogr 60:1355
34. Kleywegt GJ, Jones TA (1998) Acta Crystallogr D Biol Crystallogr D54:1119
35. Wang JM, Wang W, Kollman PA, Case DA (2006) J Mol Graph Model 25:247
36. Miller BT, Singh RP, Klauda JB, Hodoscek M, Brooks BR, Woodcock HL (2008) J Chem Inf Model 48:1920
37. Li X, He X, Wang B, Merz K (2009) J Am Chem Soc 131:7742
38. Fenn TD, Schnieders MJ, Brunger AT, Pande VS (2010) Acta Cryst 98:2984
39. Martin YC (2009) J Comput Aided Mol Des 23:693
40. Cui GL, Li X, Merz KM (2007) Biochemistry 46:1303
41. Böhm H-J (1994) J Comput Aided Mol Des 8:243
42. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP (1997) J Comput Aided Mol Des 11:425
43. Wang R, Lai L, Wang S̖ (2002) J Comput Aided Mol Des 16:11
44. Raub S, Steffen A, KaÌmper A, Marian CM (2008) J Chem Inf Model 48:1492
45. Raha K, Merz KM (2005) J Med Chem 48:4558
46. van Gunsteren WF, Dolenc J, Mark AE (2008) Curr Opin Struct Biol 18:149
47. Yusuf D, Davis AM, Kleywegt GJ, Schmitt S (2008) J Chem Inf Model 48:1411