Journal of Computer-Aided Molecular Design **18:** 235–250, 2004.

Molecular modelling prediction of ligand binding site flexibility

Ami Yi-Ching Yang[1], Per Källblad[2] & Ricardo L. Mancera[2,*]
[1]*Department of Pharmacology, University of Cambridge, Tennis Court Road, Cambridge CB2 1HQ, UK*
[2]*De Novo Pharmaceuticals, Compass House, Vision Park, Chivers Way, Histon, Cambridge CB4 9ZR, UK*

Received 7 January 2004; accepted in revised form 27 May 2004

Key words: drug design, protein, rotamer library, sidechain flexibility

Summary

We have investigated the efficacy of generating multiple sidechain conformations using a rotamer library in order to find the experimentally observed ligand binding site conformation of a protein in the presence of a bound ligand. We made use of a recently published algorithm that performs an exhaustive conformational search using a rotamer library to enumerate all possible sidechain conformations in a binding site. This approach was applied to a dataset of proteins whose structures were determined by X-ray and NMR methods. All chosen proteins had two or more structures, generally involving different bound ligands. By taking one of these structures as a reference, we were able in most cases to successfully reproduce the experimentally determined conformations of the other structures, as well as to suggest alternative low-energy conformations of the binding site. In those few cases where this procedure failed, we observed that the bound ligand had induced a high-energy conformation of the binding site. These results suggest that for most proteins that exhibit limited backbone motion, ligands tend to bind to low energy conformations of their binding sites. Our results also reveal that it is possible in most cases to use a rotamer search-based approach to predict alternative low-energy protein binding site conformations that can be used by different ligands. This opens the possibility of incorporating alternative binding site conformations to improve the efficacy of docking and structure-based drug design algorithms.

Introduction

Current ligand-protein docking methods and structure-based drug design approaches make use of protein structures obtained mainly from X-ray crystallographic determinations [1]. In some cases, protein structures determined by NMR techniques and homology modelling are also used. However, the use of a single, rigid protein structure (usually from an X-ray determination) is still the standard practice in most applications [2]. The main goal of drug design is to propose new ligands and predict the geometries and energies of the complexes that they make with their target protein. The choice of using a single protein structure simplifies the problem but is not necessarily warranted in view of the dynamic processes that occur in proteins, ranging from sidechain rotamer changes in their ligand binding sites to large backbone and sidechain conformational changes affecting their overall tertiary structure.

It is probably necessary to consider sidechain flexibility in ligand-protein docking and drug design. The conformation of sidechains may change upon or prior to ligand binding, which can alter significantly the shape and properties of a binding site and, consequently, the binding mode of ligands. Koshland pioneered the idea of an induced fit phenomenon, suggesting that the conformation of a protein undergoes complementary changes in the presence of a ligand [3]. It has been suggested that relatively few conformational changes could be observed in a binding site upon ligand binding [4]. However, the conformational rearrangement of even a single sidechain can have major consequences for ligand-protein docking and structure-based drug design. The use of multiple conformations of the ligand binding site of a protein

*To whom correspondence should be addressed. E-mail: Ricardo.Mancera@denovopharma.com

has been found to improve the prediction of binding site conformations and ligand affinities [5]. There are several indications that a given compound or a set of related compounds can bind in different ways to the same binding site of a protein [6, 7]. An increasing number of studies have investigated the implementation of conformational flexibility in virtual screening [8–11], although there has not been any validation of this approach.

Another reason for considering sidechain flexibility could lie in the need to compensate for the uncertainty in the sidechain conformation assignments that exists when a homology model is used for ligand docking or drug design. For certain protein families, only a few structural templates are available for homology model building, making it difficult to produce sufficiently good models. Previous studies have shown that for proteins sharing around 25% sequence identity with their template structures, around 50% of their Cα atoms have an RMSD of up to 2 Å [12]. It has also been noted that the prediction of sidechain conformations is no better than random when the RMSD of the Cα backbone is greater than 2 Å [13].

It is possible that incorporating ligand binding site flexibility in proteins will have to go beyond modelling sidechain reorientations only. The use of multiple protein structures with different conformations to generate an ensemble of conformational states could be the best way of accounting for the full flexibility of the receptor. This is supported by the observation that, in general, multiple protein structures can cover more conformational space than can be sampled in a molecular dynamics simulation [14, 15]. However, for protein structures that do not undergo conformational changes and thus have limited backbone motions, it is possible that sidechain reorientations alone can account for the distinct binding modes observed for different ligands.

Rotamers represent the sidechain conformations observed in crystal structures of proteins, and are widely used in the prediction of protein structures. The original definition of rotamer stemmed from the early observation that sidechains tend to exist in certain energetically favoured conformations [16]. This concept was later expanded when the clustering of conformations around χ angle space was investigated [17–21]. Rotamer libraries are usually derived from the statistical analysis of experimental structures and consist of a list of sidechain conformations and their observed frequencies. In most cases, these conformations correspond to local minima on the potential energy surface of the sidechains [22]. It has also been noticed that rotamer statistics are highly dependent on protein structure factors such as the local backbone conformation [23, 24], the secondary structure of the residue [25] and tertiary steric restrictions [26]. A number of protein modelling applications have made use of rotamer libraries with some degree of success [27–31].

Here we validate a method for the efficient identification of a set of representative conformations of the binding site of a protein that are low in energy and high in diversity, with the aim of covering the most relevant portions of the conformational space of a given ligand binding site. We use a recently introduced algorithm, DYNASITE, which makes use of a backbone-dependent rotamer library to generate a comprehensive set of ligand binding site conformations [32]. This algorithm has been used to find the low energy conformations and reproduce the experimentally observed ligand binding mode in MMP-1 [32]. We use this algorithm to test the efficacy of a method for predicting alternative binding site conformations that are observed experimentally in the presence of a ligand for a number of proteins. For this purpose, we have analysed proteins that have been determined by X-ray and NMR methods. The proteins selected exhibit little backbone movement across different structures and hence constitute an ideal test set to examine the effectiveness of this method for predicting the alternative low-energy binding site conformations observed experimentally in the presence of bound ligands for those proteins that do not exhibit major conformational changes.

Materials and methods

An outline of the methodology that we have followed to generate alternative low-energy ligand binding site conformations is sketched in Figure 1. In brief, each experimental structure was energy minimised and a template structure was selected for computational multiple conformer generation. A list of sidechains was chosen on the basis of their location and ability to affect the shape and the ligand-binding properties of each binding site. The DYNASITE algorithm was then used to generate an exhaustive list of binding site conformers, followed by an energy minimisation of each conformer. Since this procedure tends to generate a large number of conformers, their numbers were reduced by performing a statistical cluster analysis

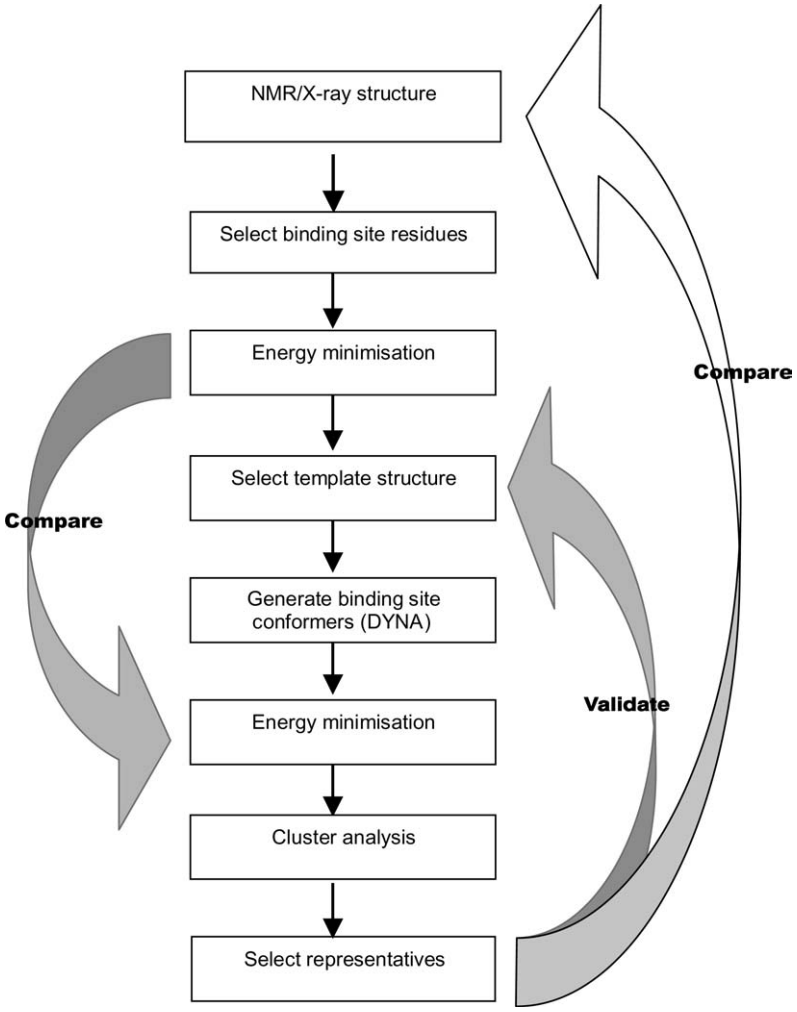

Figure 1. Outline of the method used to predict and validate alternative conformations of the binding site of a protein.

and the corresponding representative conformers were then compared to the experimental structures. We now proceed to explain the method in detail.

Selection and preparation of protein test sets

All NMR and X-ray protein structures were taken from the Brookhaven National Laboratory Protein Databank (PDB) [33]. Separate test sets of NMR and X-ray structures were obtained according to a number of criteria.

NMR structures

NMR protein structures were chosen when alternative binding site conformations could be observed. NMR determinations are usually deposited as an ensemble of structures (and sometimes, also, an average structure), reflecting the various solutions that were found to satisfy the determination constraints. Consequently, several conformations of a protein (and of its ligand binding site) become available from a single PDB entry. Although the number and nature of these conformations do not necessarily correspond to the true conformations of a protein, they portray its underlying conformational variability. Protein structures were selected according to the following rules: (1) there should be more than one solution, (2) the protein should have a non-covalently bound ligand, (3) the ligand should not be a peptide, a cofactor, or a metal ion, and (4) the protein backbone around the binding site should have little variation across the various solutions, as measured by an RMSD lower than 2.0 Å. These selection rules left us with nine structures that were considered for this study and which are listed in

Table 1. For clarity, we will refer to each individual protein according to its PDB code while its corresponding template structure (see below) will be referred to using a *PDBtemp* format.

X-ray structures

X-ray protein structures were also selected on the basis of more than one binding site conformation being available. This usually involved proteins in both free and ligand-bound forms, although a number of selected proteins did not have a free form available but had two or more bound-ligand forms. Since there are more than 23,000 protein crystal structures in the PDB, proteins were selected as broadly as possible on the basis of the structural class they belonged to, according to the SCOP [34] classification. Protein structures were selected according to the following rules: (1) the X-ray resolution should be 2.0 Å or better, (2) the protein structures should have no mutations in the binding sites, (3) the protein structures should have no cofactors and/or metals in the binding site, and (4) the protein backbone around the binding site should have little variation across the various forms of the protein, as measured by an RMSD lower than 2.0 Å. For those proteins that had more than one binding site, only one of them was considered. These selection rules left us with sixteen different proteins, with a number of different forms for each one, which are listed in Table 2. For clarity, we will refer to each individual protein using its abbreviation (see Table 2), while the crystal structure of each form of each protein will be referred to using its corresponding PDB code.

Selection of a template structure

For each protein test case a template structure was chosen. This template structure was then used to generate alternative ligand binding site conformations to try to reproduce the binding site conformation of the other forms of the protein. The template structure of each protein was selected in different ways for the NMR and X-ray test sets.

NMR structures

The NMR solutions contained in each PDB entry were initially filtered: any observed non-geometrically conserved regions (where the RMSD was larger than 2.0 Å) were excluded as long as the ligand binding site of the structure was not affected and the remaining protein sequence stayed as a whole chain. This was done in order to reduce the RMSD variance and increase the significance of the clustering. Each solution was then energy minimised without removing the ligand (further details of the energy minimisation protocol are provided in the following subsections). All non-hydrogen backbone atoms, except for those of the ligand, were then used for the superposition and RMSD calculations.

The program NMRCLUST [35] was then used to cluster the NMR solutions for a given protein determination according to their geometric variation so that a single template solution could be selected. The program superimposes each solution of the determination onto each of the other solutions in a pairwise manner, determining the corresponding RMSD value between each generated binding site conformation. It then clusters the solutions into a group of conformationally related subsets and selects a representative structure from each cluster. The method of average-linkage [36] is used by the program, which also performs a normalisation of the average spread and contains a penalty function that combines the normalised average spread and the number of clusters at each step of the clustering procedure. This combined approach is used to generate clusters that are as highly populated as possible whilst simultaneously maintaining the smallest spread (which maximizes the similarity of the conformations of the members of any cluster) [35]. The average linkage method used in NMRCLUST has been shown to outperform other clustering methods such as single linkage and complete linkage [35]. The method is also deemed to be superior to the Jarvis-Patrick method, due to the high level of user intervention needed for such a method [35].

A representative backbone template structure with its sidechains was now obtained from the largest cluster as the one closest to the centroid of each cluster in terms of RMSD. This is carried out in the NMRCLUST program using eigen analysis [35, 37]. Since other clusters tended to be scarcely populated, they were not included in the analysis.

X-ray structures

For each protein test case, each of its X-ray structures was energy minimised (those structures with a bound ligand were minimised in the presence of the ligand). If a ligand was present, the ligand was then removed and the energy of the binding site (backbone and sidechains) was re-calculated. The structure with the lowest energy was chosen as the template structure.

Table 1. NMR-determined protein data set.

PDB code	Description	Source	Number of residues	NMR solutions	Binding site residues with observed movements
1AO8	Dihydrofolate reductase	*Lactobacillus casei*	162	21	L4, W5, L19, D26, L27, F30, R31, S48, F49, T116
1AP8	RNA cap/translation initiation factor Eif4E	*Saccharomyces cerevisiae*	213	20	W 58, K 90, D 92, H 94, E103, RP104, E105, R157
1B2I	Human plasminogen	*Homo sapiens*	83	20	Y(A)36, D(A)55, E(A)57, W(A)62, F(A)64, R(A)71, W(A)72
1BM6	Metalloprotease	*Homo sapiens*	173	20	F157, N162, L164, L197, V198, L218, M219, Y220, P221, L222
1BZF	Dihydrofolate reductase	*Lactobacillus casei*	162	22	L4, D16, L19, D26, L27, F30, F49
1DIU	Dihydrofolate reductase	*Lactobacillus casei*	162	18	L4, L19, D26, L27, Y29, F30, F49, P50, L54, T116
1MUX	Calcium binding protein	*Xenopus laevis*	148	30	E83, E84, E87, F92, L105, F141, M144, M145,
1RML	Acid fibroblast growth factor	*Homo sapiens*	155	26	N32, K126, K127, T137, H138, Y139, Q141, K142, I144
2SRT	Human stromelysin	*Homo sapiens*	173	30	N162, V163, L164, L197, V198, L218, Y220, L222, Y223, H224

Since all protein test cases had two or more experimentally determined structures available (either different crystal structures or various NMR solutions), one template structure was chosen as the reference structure for multiple conformer generation. The template structure with the lowest energy was chosen (see below for energy calculations). In the case of the X-ray test cases, this meant that either the free or the bound form of the protein could be selected.

Multiple conformer generation

Once appropriate template structures had been identified, the algorithm DYNASITE [32] was applied to generate all possible conformations of a set of user-defined residues by using a backbone-dependent rotamer library [30, 38]. The algorithm performs an exhaustive search and generates all possible rotameric state combinations of a binding site. For each crystal structure considered, the ligand binding site is defined by taking all atoms of any residue with at least one atom within 6.0 Å of the ligand. All the residues selected were then combined to provide a single list of residues that would encompass the binding sites of all the crystal structures for each protein. All mainchain atoms are kept fixed during the multiple conformer search. The algorithm goes through the residues in an exhaustive manner and tries to assign them a new rotameric state while avoiding steric clashes with the surrounding protein atoms. A van der Waals clash test is performed before accepting a new conformation. The clash tolerance distance (CTD) represents the cutoff for the van der Waals clash test. Two atoms A_1 and A_2 are considered to clash if

$$\text{Distance } |A_1 - A_2| < (\text{vdW radius } A_1 + \text{vdW radius } A_2 - \text{CTD})$$

The CTD is used to control the severity of steric clashes during a multiple residue conformation generation. Since hydrogens are not explicitly treated, it is important to consider whether it should be compensated for by decreasing the CTD. In our experience, the most appropriate value of the CTD varies according to the size of the binding site of the protein and the nature of the sidechains. This is due to the fact that different binding site geometries allow for the generation of alternative sidechain conformations with varying degrees of difficulty as a consequence mainly of steric crowding. In this study, the CTD used ranged from 0.5 to 1.3 Å, which allowed for the successful generation of conformers.

The rotamer Φ/ψ acceptance threshold (RPAT) is another parameter that allows the user to control the degree of flexibility of the backbone. It represents the angular space of the backbone Φ/ψ angles of the selected residues to be allowed in the rotamer search. The default value of RPAT used was 5°. For given

Table 2. X-ray-determined protein data set.

Protein	Abbrev.	PDB code	Source	Number of residues	Form/ ligand	Binding site residues observed movements
Periplasmic lysine-, arginine, ornithine-binding protein	LAO	1LAH, 1LST, 1LAF*	*Salmonella typhimurium*	238	ORN, LYS, ARG	D11, Y14, S69, S70, S72, R77, L117, S120, T121, D161
Immunoglobin (B1-8 Fv fragment)	IMG	1A6W*, 1A6U	*Mus musculus*	229	Free, NIP	Y(L)34, W(L)93, W(L)98, W(H)333, H(H)335, R(H)350, K(H)359, Y(H)399, Y(H)401, S(H)405
Carboxypeptidase	CBP	1ARM*, 1CBX, 2CTB, 2CTC	*Bos taurus*	309	TRS, BZS, Free, LOF	H69, R71, E72, R1279, R145, I247, Y248
Transport protein receptor	GBP	1GLG*, 2GBP	*Escherichia coli*	309	GAL, GLC	D14, N91, H152, D154, R158, N211, D236, N256
Glutathione S-transferase	GS2	1GLP, 1GLQ, 2GLR*	*Mus musculus*	418	GTS, GTB, GTX	H231, L328, N319, H331, E364, H375, D376, L447
Human methionine aminopeptidase-2	IVN	1B59, 1B6A, 1BOA*	*Homo sapiens*	370	OVA, TNP, FUM	H231, L328, N329, H331, H375, D376, L447
E. coli thymidylate synthase	MTF	1KCE, 1SYN*, 1TSD, 2TSC	*Escherichia coli*	528	UMP, UMP, UMP, UMP	S54, E58, V77, T78, I79, W80, W83, L143, F176, N177, Y209
Influenza virus neuraminidase	PPD	2QWA, 2QWC* 2QWH	Influenza A virus	388	Free, DAN, G39	R118, E119, D151, R152, W178, I222, R224, E276, K292, Y406
Staphylococcal nuclease	PAB	1SNC, 1STG, 1STH*	*Staphylococcus aureus*	149	PTP, THP, THP	N36, Y38, H40, G74, R77, H92, G97, N98, F100
Human transthyretin (prealbumin)	RTL	1ETB*, 1TTB	*Homo sapiens*	254	T44, Free	M13, K15, L17, E51, E54, H56
Cellular retinoic acid binding proteins	RTB	1CBS*, 3CBS	*Homo sapiens*	137	REA, R12	L19, V24, L28, T54, V58, R59, V76, D77, R132, Y134
Plasma retinol-binding protein	RBN	1ERB*, 1FEN, 1HBP	*Bos taurus*	183	ETR, AZE, RTL	L35, F36, L37, F45, V61, L63, M73, M88, Q98, H104
Hydrolase phosphoric diesterase	RBN2	1RBX*, 1RNC, 1RND, 1ROB	*Bos taurus*	124	Free, CPG, DCG, C2P	Q11, H12, K41, V43, N44, T45, K66, R85, F120, D121
Ribonuclease T1	SPN	6RNT, 9RNT*	*Aspergillus oryzae*	104	2AM. Free	N36, Y38, H40, E58, R77, H92, N98, F100
Trypsin	TPN	1TNG, 1TNI, 1TNL*, 5PTP	*Bos taurus*	229	AMC, PBN, TPA, Free	D189, S190, Q192, S195, V213, W215
Protein tyrosine phosphatase 1B	PTP	1C83*, 1C88, 1ECV	*Homo sapiens*	298	OAI, OTA, 878	Y(A)46, D(A)48, V(A)49, K(A)120, D(A)181, F(A)182, S(A)216, I(A)219, R(A)221, Q(A)262

*Chosen template structure.

Φ/ψ angles, it allowed the algorithm to search for any rotamers that are in the Φ/ψ angles $\pm 5°$ range.

The procedure that we followed made use initially of default values for CTD and RPAT and then gradually increased their value (in 0.1 Å steps for CTD and 0.5° steps for RPAT) until all the sidechains were selected and one or more alternative conformations were suggested. The value was then increased one further and final step. The conformers generated from the last two steps were then energy minimised for 500 iterations using steepest descents followed by the method described below for generating representatives of the conformers. Two sets of clusters were compared: if the number of clusters was the same then the procedure was stopped and the lower value of the parameters used was reported. If the number of clusters increased as the value of CTD/RPAT increased, the value of the parameters was again increased by one step and the whole procedure was repeated again.

Comparative analysis of conformers

Energy minimisation

As described earlier, both the experimentally determined and computer-generated structures were energy minimised. This was done using the same protocol. The energy minimisations were carried out using InsightII/Discover3 [39] with the CFF forcefield [40] and default settings unless otherwise stated. Hydrogen atoms were added assuming a pH of 7.4, and a fixed dielectric constant of 4.0 was used to mimic the dielectric properties of the protein environment. Each structure was subjected to 500 iterations of steepest descents (SD) followed by a maximum of 10,000 iterations of conjugate gradients (CG). The minimisation was terminated when the gradient of the energy converged to a value of less than 0.1 kcal/mol/Å. During the minimisation, all the atoms of the sidechains in the binding site (defined as all atoms in any residue with at least one atom within a 6.0 Å radius of any atom of the ligand) were fully flexible, while the backbone and the rest of the protein outside of the binding site were kept fixed.

Energy calculation

All molecular mechanics energies reported in this study were calculated relative to that of a given template structure. These relative energies were thus obtained as

$$E_{bind} = E_{total} - E_{bb} \quad (1)$$

$$E_{diff} = E_{bind}[i] - E_{ref}[templ] \quad (2)$$

First, the total energy of the solution/conformer (E_{total}) was calculated. For each minimised structure the sidechains were removed and hydrogen atoms were added to fill valences, and a second energy calculation was performed (E_{bb}). The difference between the two, the binding site energy E_{bind}, corresponds only to the energy of the sidechains of that solution/conformer. The energy of each binding site conformer (E_{bind} $[i]$) was then compared with the energy of the sidechains of the template structure (E_{ref} [templ]) (calculated in the same manner as shown in Equation 1) to obtain the relative energy or energy difference E_{diff}, as shown in Equation 2.

Selection of representatives

After the above energy minimisation procedure, the NMRCLUST program was again used to identify geometrical subsets in the minimised binding site conformers (since many of them had very similar energies), and the same procedure as described above for selecting unique representative structures was followed. This set of representatives covered the geometrical variation seen in the whole set of minimised conformers.

The clustering method in NMRCLUST allowed us to compare each conformer from their pairwise RMSD values. For each test case, the entire backbone of the binding site (as defined above: all atoms in any residue within 6.0 Å radius of the ligand) was used to superimpose all the conformers. The cluster analysis was then carried out on all the non-hydrogen atoms of the sidechains previously used to generate conformers in DYNASITE. For each identified cluster, the minimised conformer structure with the lowest relative energy was selected as the cluster representative. These representative binding site conformers were then compared with the experimental binding site conformation(s). This was done, for each protein test case, by clustering all the binding site conformers together with the corresponding experimental structure. This allowed us to rapidly screen each experimental structure against the computer-generated conformers to see whether the experimental structure could be found in any of the resulting clusters. The results of this clustering procedure were then also checked visually for consistency.

Results and discussion

Binding site conformer generation

We tested our procedure on 9 NMR and 16 X-ray protein test cases. These proteins had sizes ranging from 66 to 300 residues. The generation of raw binding site conformers using DYNASITE was very fast, taking an average of 91.3 s on an SG Octane R10000 processor. The energy minimisations were relatively time consuming, taking an average of 287.5 s for each structure.

The final number of clusters of minimised binding site conformers that were generated for each protein test case can be found in Table 3 (column 2). Prior to this, DYNASITE generated between 28 and 192 binding site conformers for each protein in the NMR test set, although in most proteins (the only exception being 2SRT), less than 100 conformers were generated. In the case of the X-ray test set, the number of binding site conformers generated ranged from 48 to 2450. IVN was the only test case where DYNASITE generated less than 100 conformers, while for the majority of test cases around 1000 conformers were generated. In addition, the number of conformers generated depends on the number (the more side chains, the more possible combinations) and residue type (some residues, such as arginine, have significantly more rotamers) used. It should not be surprising that DYNASITE generated more conformers with the X-ray test cases than with the NMR ones. This is due to the fact that the rotamer library that we used contains rotameric states that were determined from an analysis of X-ray structures only. In fact, there is no available rotamer library that incorporates rotameric states of sidechains from NMR determinations [41, 42].

Cluster analysis and the selection of representatives

In most test cases hundreds, even thousands of raw binding site conformers were generated by DYNASITE. The energy minimisation procedure resulted in many binding site conformers having very similar energies. This is to be expected, as many of the non-minimised conformers generated by DYNASITE were high in energy (due mostly to steric clashes) and an energy minimisation could make them converge to the same local energy minimum. As a consequence, many minimised conformers had very similar binding site conformations and hence a selection of a representative subset of minimised unique conformers was performed as described in the Methods section.

Table 3 lists the number of minimised cluster representative conformers generated in each protein test case. The number of cluster representatives generated for the NMR test cases ranged from 4 to 11, whereas for the X-ray test cases the number of clusters ranged from 8 to 20 (column 2). The larger number of observed clusters of minimised conformers for the X-ray test cases is probably due to the larger number of binding site conformations generated, as described above. Individual RMSD values within each cluster for each test case protein are also reported in Table 3 (column 3). The efficiency of the clustering method was revealed by the fact that the average RMSD *within* each cluster was sufficiently low: for the X-ray test cases it was 0.15 ± 0.06 Å and for the NMR test cases it was 0.27 ± 0.11 Å. In a number of X-ray test cases, some clusters contained more than 200 conformers, which nonetheless exhibited a significant degree of similarity.

We can also see from Table 3 (column 4) that the mean pairwise RMSD of the NMR representative conformers is smaller than the X-ray representative conformers, revealing the smaller conformational diversity obtained with the NMR test cases when compared with the X-ray test cases. This is exemplified by the larger values of the highest RMSD (column 5) observed for the X-ray test cases. The RMSD spreads (standard deviation in column 4) between each representative conformer in the X-ray test cases are larger than those of the NMR test cases. This agrees with our observation that the collection of X-ray representative conformers is more conformationally diverse than that of the NMR representative conformers. A comparison between the RMSD within each cluster (column 3) and the RMSD between representative conformers (column 4) for each protein test case confirmed that the representative conformers were significantly different from one another in relation to the degree of similarity within each cluster.

The clustering method efficiently reduced the number of minimised conformers to a small subset of unique structural representatives that covered the conformational variation observed in the minimised binding site conformers. These representatives were taken as the final output of our method for predicting alternative protein binding site conformations.

Table 3. Summary of representative conformers.

Protein test case	Number of cluster representatives	Average RMSD (Å) within each cluster	Average RMSD (Å) between representative conformers	Lowest RMSD (Å) between representative conformers	Highest RMSD (Å) between representative conformers
NMR					
(1AO8) Dihydrofolate reductase	5	0.26	0.89 ± 0.21	0.29	1.01
(1AP8) RNA cap/translation initiation factor Eif4E	4	0.29	0.66 ± 0.31	0.31	0.99
(1B2I) Human plasminogen	8	0.28	0.94 ± 0.44	0.31	1.09
(1BM6) Metalloprotease	6	0.22	0.56 ± 0.24	0.25	0.87
(1BZF) Dihydrofolate reductase	5	0.13	0.51 ± 0.22	0.16	0.84
(1DIU) Dihydrofolate reductase	9	0.26	0.81 ± 0.26	0.29	1.10
(1MUX) Calcium binding protein	9	0.25	0.62 ± 0.33	0.27	0.90
(1RML) Acid fibroblast growth factor	7	0.33	0.88 ± 0.35	0.36	1.01
(2SRT) Human stromelysin	11	0.28	0.75 ± 0.34	0.31	0.19
X-ray					
Periplasmic lysine-, arginine, ornithine-binding protein	9	0.18	0.80 ± 0.29	0.26	1.51
Immunoglobin (B1-8 Fv fragment)	19	0.17	1.43 ± 0.42	0.25	1.98
Carboxypeptidase	11	0.21	2.25 ± 0.48	0.31	2.71
Transport protein receptor	10	0.19	0.81 ± 0.39	0.35	1.16
Glutathione S-transferase	9	0.18	1.02 ± 0.48	0.22	1.41
Human methionine aminopeptidase-2	7	0.08	0.97 ± 0.41	0.12	1.72
E. coli thymidylate synthase	6	0.17	1.14 ± 0.58	0.38	2.05
Influenza virus neuraminidase	7	0.15	1.01 ± 0.32	0.25	1.51
Staphylococcal nuclease	17	0.14	1.08 ± 0.29	0.23	1.31
Human transthyretin (prealbumin)	13	0.18	1.81 ± 0.45	0.44	2.35
Cellular retinoic acid binding protein	19	0.16	1.27 ± 0.68	0.44	2.53
Plasma retinol-binding protein	11	0.13	1.17 ± 0.45	0.31	1.72
Hydrolase phosphoric diesterase	11	0.13	0.95 ± 0.40	0.26	1.75
Ribonuclease T1	20	0.10	1.07 ± 0.36	0.18	1.44
Trypsin	12	0.09	0.94 ± 0.29	0.15	1.43
Protein tyrosine phosphatase 1B	18	0.11	1.41 ± 0.57	0.31	2.14

Reproducing the experimental binding site conformations

A crucial test for our method is to determine whether any of the predicted low-energy binding site conformations is actually observed experimentally. This test becomes more difficult given the fact that we were interested in reproducing the binding site conformation of proteins when bound to a ligand.

Prior to such analysis, it is important to show that the experimental template structure and the experimental alternative conformations are significantly different to warrant the application of the method that we report in this paper. Table 4 reports the average and individual RMSD between the sidechains of the chosen template structures and the sidechains of each of the alternative conformations that we aimed to reproduce. We can see that the RMSD values are sufficiently large to reflect the fact that there are significant structural differences between the template and alternative binding site conformations. Furthermore, these RMSD values are in general higher than the corresponding RMSD values between the alternative binding site conformations and the best matching generated representative conformers (columns 4–7 in Table 5, see below), confirming the need to consider alternative binding site conformations. In those cases where the latter RMSD values were low, the relevant sidechains that are responsible for the differences between the template and alternative binding site conformations are identified and listed in Table 4 (footnote).

In order to verify that our method could indeed generate experimentally observed alternative conformations of a ligand binding site, we performed a comparative analysis of each protein structure against the corresponding binding site conformer representatives. We decided to cluster the representative binding site conformers for each test case together with the corresponding experimental structure. This allowed us to rapidly screen each experimental structure against the computer-generated conformers. The RMSD of the sidechains between the relevant binding site template structure and the best matching representative conformer is reported in Table 5 (column 5). All experimental template structures were reproduced with an RMSD ranging from 0.09 to 0.62 Å (with an average of 0.38 Å), with the exception of 1AP8. We discuss this failed test case further on. Figure 2 shows an example (1BZF) where the binding site of the template structure was accurately reproduced. We also noticed that in those cases where we successfully reproduced the template structure, more than half of the reproduced conformers were lower in energy compared to the template structure (details not shown).

We also investigated whether our method could reproduce alternative low-energy binding site conformations of either an alternative ligand-bound form or the free form of the protein. Since we selected template structures on the basis of their energies (the one with the lowest energy was always chosen), it could be either a bound or the free form of the protein (for X-ray test cases), depending on which one had the lowest energy. Consequently, we were in a position to examine whether our method could predict the free form as well as a bound form conformation for X-ray test cases, or just an alternative conformation of the bound form for NMR test cases. The RMSD between the relevant alternative binding site template structure and the best matching representative conformer is also reported in Table 5 (columns 4–7). We can see that, using only one template structure, our method is able to reproduce in 22 out of 25 cases an alternative binding site conformation observed experimentally by either X-ray or NMR determinations. The RMSD ranged from 0.16 to 0.86 Å (with an average of 0.49 Å). Figure 3 shows an example where the binding site conformation of the free form of the protein is accurately reproduced with our method using the conformation of the bound form as the template structure.

Despite the high success rate of our method, there were a few test cases (two from the NMR test set and one from the X-ray test set) where it was not possible to reproduce the experimental conformation of the binding site from the template structure. We now proceed to examine these test cases to try to explain why our method occasionally failed.

The first failed test case is the mRNA cap binding protein (1AP8), from the NMR data set. In this test case eight sidechains (Trp58, Lys90, Asp92, His94, Glu103, Trp104, Glu105 and Arg157) were chosen for the initial rotameric conformational search in DYNASITE. Thirty-six conformers were generated, minimised and divided into four clusters with an RMSD cutoff of 0.57 Å. A structural examination revealed that the four representative conformers had an RMSD ranging from 0.61 to 1.06 Å. We verified that most conformations of each of the sidechains had been sampled. However, in the case of the sidechain of Arg157, DYNASITE could only find two rotameric conformations. Neither of these rotamers reproduced the experimental conformation. We would like to point out that our method involved M7GDP being included in the energy minimisation of the template structure but not in the energy minimisation of the conformers. Figure 4 illustrates the conformational space occupied by the bound ligand M7GDP. The terminal NH_2 of Arg157 establishes hydrogen bonds with M7GDP in the NMR-determined structure. In the absence of the ligand, Arg157 was predicted to either occupy the space where M7GDP sits or turn away from the bind-

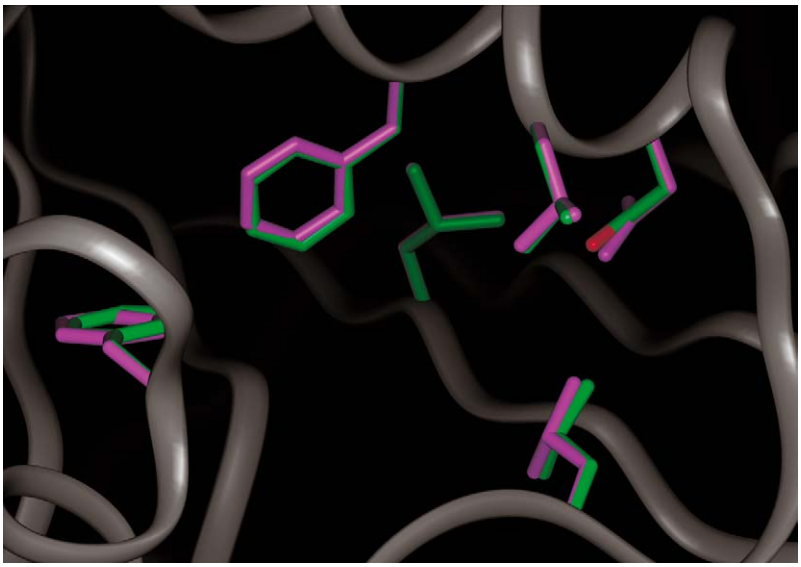

Figure 2. Superposition of the experimental template conformation of 1BZF and the closest conformer that reproduces this conformation. The template structure is coloured according to atom type and the closest minimised conformer is shown in magenta.

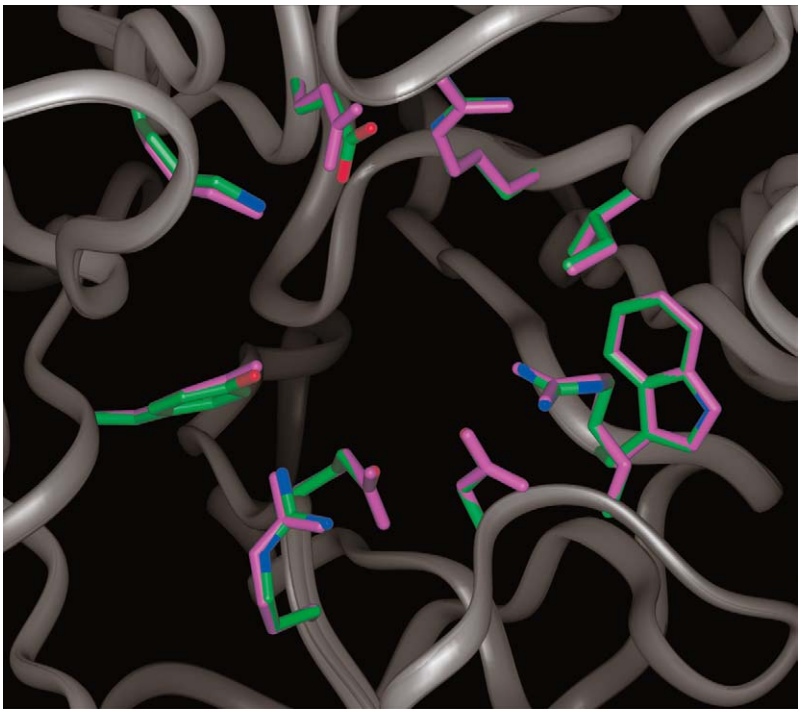

Figure 3. Superposition of the alternative experimental conformation of PPD and the closest conformer that reproduces this conformation. The alternative structure (2qwa) is coloured according to atom type and the closest minimised conformer is shown in magenta.

Table 4. Summary of comparative RMSD values between template and alternative binding site conformations.

Subset	PDB code or abbreviation	Number of experimental NMR/X-ray conformations	Average RMSD (Å) between template and all alternative conformations	RMSD (Å) between template and alternative conformations		
NMR	1AO8	4	0.89	0.50[1]	0.81[2]	0.94
	1AP8	3	2.16	1.71	2.50	–
	1B2I	3	1.22	1.01	1.73	–
	1BM6	4	1.34	1.05	1.63	1.77
	1BZF	4	1.01	0.81	0.93	1.09
	1DIU	3	0.75	0.63[3]	0.96[4]	–
	1MUX	3	1.69	1.24	1.77	–
	1RML	3	1.40	1.19	1.73	–
	2SRT	4	1.70	1.37	1.97	2.01
X-ray	LAO	3	0.76	1.05	0.29[5]	–
	IMG	2	3.52	3.52	–	–
	CBP	4	2.20	1.51	1.85	1.49
	GBP	2	1.23	1.23	–	–
	GS2	3	0.51	0.49[6]	0.61[7]	–
	IVN	3	1.30	1.77	1.73	–
	MTF	4	1.56	1.73	1.32	2.16
	PPD	3	1.09	0.83	1.19	–
	PAB	3	1.43	1.50	1.45	–
	RTL	2	2.21	2.21	–	–
	RTB	2	0.88	0.88	–	–
	RBN	3	0.77	0.61	0.94	–
	RBN2	4	1.29	1.40	1.24	1.23
	SPN	2	0.88	0.88	–	–
	TPN	4	0.57	0.62[8]	0.51[9]	0.82
	PTP	3	0.82	0.99	0.57[10]	–

*The sidechains that exhibited significant movement are: [1]L19, T116; [2]F49; [3]D26, T116; [4]L19, L27, P50; [5]D11, E14, Y108; [6]R13, E97; [7]Y49; [8]I138, D189; [9]Q192, F182; [10]K120.

ing site with both of these sidechain conformations being lower in energy than the template structure.

The second failed test case in the NMR data set was human thiotransferase (1B2I). In this test case, residues Tyr36, Asp55, Glu57, Trp62, Phe64, Arg71, and Trp72 have been reported to interact directly with ligand AMH and hence all of them were selected for the rotameric conformational search [43]. DYNASITE generated 42 raw conformers and, after an energy minimisation and clustering, eight representative binding site conformers were found. In particular we noticed that only three alternative conformations of Arg71 were predicted. Figure 5 shows two of these alternative conformations of Arg71. One of these conformations reproduced the template structure (RMSD = 0.44 Å) and the other one is the alternative conformation suggested by our method. The second alternative conformation observed in the NMR determination (coloured according to atom type) was not found during the conformational search. In this conformation, Arg71 forms a hydrogen bond with the carboxyl group of AMH. When the conformers were minimised (in the absence of the ligand), Arg71 cannot form such a hydrogen bond and, as a consequence, the sidechain moves further away to produce an alternative conformation.

The only test case from the X-ray data set where our method failed to reproduce an alternative binding site conformation is the periplasmic lysine-, arginine-, ornithine-binding protein (LAO). In all the crystal structures the superposition of the backbones, apart from the free form, yielded an RMSD of 0.18 Å. It is known that several sidechains (Asp11, Tyr14 and Ser72) appear to have different conformations in the presence of different ligands [44]. The chosen template structure was the free form of LAO. Our method

Table 5. Summary of results for all protein test cases.

Subset	PDB code or abbreviation	Number of NMR/X-ray conformations	Best RMSD (Å) with respect to template conformation*	Best RMSD (Å) with respect to alternative experimental conformations*		
				Conformation 1	Conformation 2	Conformation 3
NMR	1AO8	4	0.31	0.47	0.58	0.59
	1AP8	3	Arg57	Arg57	Arg57	–
	1B2I	3	0.44	Arg71	Arg71	–
	1BM6	4	0.35	0.61	0.74	0.80
	1BZF	4	0.12	0.29	0.41	0.57
	1DIU	3	0.35	0.60	0.86	–
	1MUX	3	0.29	0.51	0.77	–
	1RML	3	0.40	0.44	0.52	–
	2SRT	4	0.09	0.32	0.5	0.56
X-ray	LAO	2	0.12	Asp11	Asp11	–
	IMG	2	0.15	0.27 (1A6U)	–	–
	CBP	4	0.34	0.28 (2CTC)	0.35 (1CBX)	0.58 (2CTB)
	GBP	2	0.32	0.33 (2GBP)	–	–
	GS2	3	0.44	0.41 (1GLP)	0.34 (1GLQ)	–
	IVN	3	0.31	0.48 (1B6A)	0.5 (1B59)	–
	MTF	4	0.27	0.26 (2TSC)	0.41(1TSD)	0.23 (1KCE)
	PPD	3	0.39	0.49 (2QWA)	0.31 (2QWH)	–
	PAB	3	0.30	0.22 (1SNC)	0.41 (1STG)	–
	RTL	2	0.04	0.16 (1TTB)	–	–
	RTB	2	0.25	0.43 (3CBS)	–	–
	RBN	3	0.29	0.27 (1HBP)	0.33 (1FEN))	–
	RBN2	4	0.08	0.25 (1RNC)	0.47 (1ROB)	0.49 (1RND)
	SPN	2	0.22	0.28 (6RNT)	–	–
	TPN	4	0.21	0.44 (5PTP)	0.40 (1TNG)	0.58 (1TNI)
	PTD	3	0.17	0.34 (1C88)	0.28 (1ECV)	–

*In those cases where our method failed, the residue whose conformation was not predicted correctly is identified.

was able to reproduce the free form of LAO with an RMSD of 0.28 Å, but it failed to reproduce the alternative conformation of Asp11 observed in the LAO-ornithine (ORN) complex. There are several water molecules in the binding site of the free form of LAO. It was shown that the water could be replaced with appropriate polar groups by an incoming ligand [44]. In the case of ORN, the water molecules were replaced by both ORN and Asp11. The associated conformational change would be impossible without the simultaneous rotation of the peptide bond between Asp11 and Thr12, due to a steric clash between the carbonyl oxygen and the carboxyl group of Asp11 [44]. Figure 6 illustrates the 122° rotation of the peptide bond that takes place: the Φ/ψ angles of Asp11 ($-94.5/112.7$) change to $-88.3/-15.1$, while the Φ/ψ angles of Thr12 ($-85/32.2$) change to 58.3/9.8 [44]. As a consequence, this particular sidechain conformation could not be observed without this backbone torsion angle change between the two residues. This test case illustrates the fact that a local backbone movement can have a significant effect on the prediction of the associated sidechain orientations.

We can see that the three failed test cases are associated with particular ligand-induced conformations. It becomes clear that our method failed in those situations where a high-energy conformation of the binding site was stabilised by the binding of a ligand. This is due to the fact that our method involves all computer-generated conformers being obtained and energy minimised in the absence of any ligand. This is of course necessary to make this method a useful tool for predicting alternative binding site conformations without any ligand information, which is of particular relevance for ligand-protein docking and drug design purposes. Recently the use of alternative ligand bind-

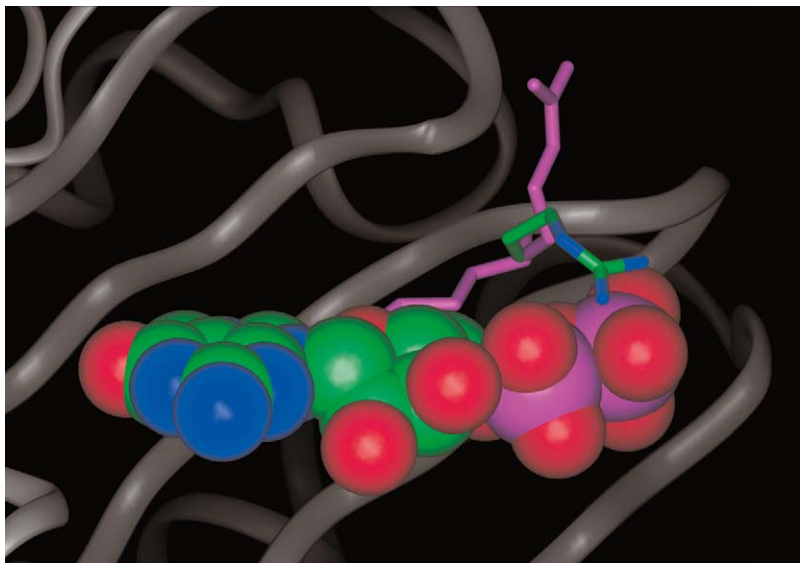

Figure 4. M7GDP-Arg157 interaction in 1AP8. The NMR-determined conformation of Arg57 was not correctly reproduced. This sidechain was predicted to adopt either of two rotameric states (shown in magenta). One pointed towards the space occupied by M7GDP while the other pointed away from the binding site. 1AP8temp is coloured according to atom type. The ligand is also coloured according to atom type and shown in a CPK representation.

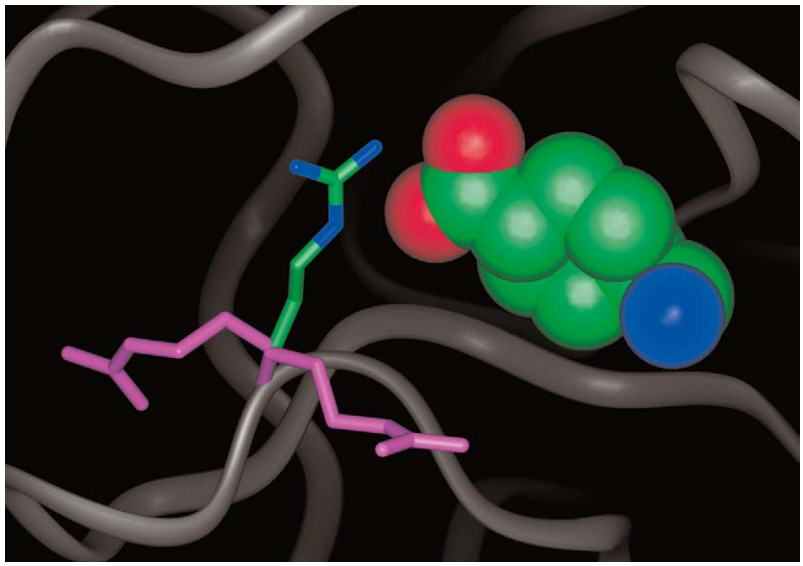

Figure 5. AMH-Arg71 interaction in 1B2I. The NMR-determined conformation of Arg71 was not correctly reproduced. This sidechain can adopt three possible rotameric states, and two of them are shown in magenta. The predicted conformations point away from the ligand. An experimentally observed alternative conformation is coloured according to atom type. The ligand is also coloured according to atom type and shown in a CPK representation.

ing site conformations generated by a similar method has been shown to outperform the efficacy of reagent virtual screening when compared to the use of a single X-ray crystal structure [45].

The results of our investigation are certainly encouraging in that a significantly high success rate is observed when our method is used to predict alternative conformations of the binding of a protein. Our method is able to generate a number of binding site conformers that are high in structural diversity and low in energy, some of which are generally seen to correspond to experimentally observed conformations.

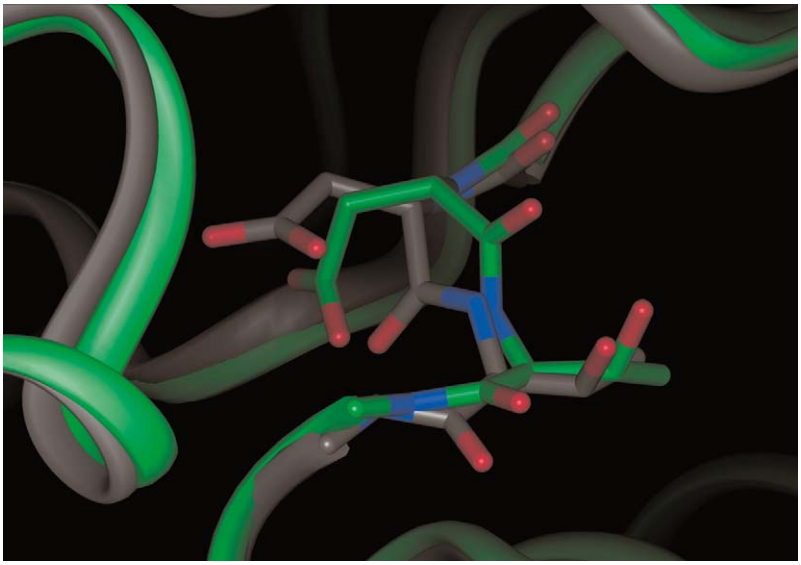

Figure 6. The superimposed structures of LAO-ORN (1LAH) and LAO-ARG (1LAF). The peptide bond of Asp11 rotates 122° in 1LAH and, as a result, the correct sidechain conformation failed to be sampled by DYNASITE. 1LAH is coloured according to atom type with its protein backbone coloured in green, while 1LAF is also coloured according to atom type but its carbon atoms and protein backbone are coloured in gray.

This should allow, for proteins that do not exhibit large conformational changes upon ligand binding, for the generation of a set of low-energy binding site conformations that can be used to design or dock ligands with different binding modes. More importantly, our results suggest that, for proteins that exhibit little backbone motion, ligands tend to bind to low-energy conformations of their binding site with only a few cases where the bound ligand induces a high-energy conformation of the binding site.

Conclusions

We have made use of a rotamer search-based method to predict alternative low-energy protein binding site conformations, for proteins that do not exhibit large conformational changes upon ligand binding. This method is able to find in most cases an experimentally observed ligand-bound binding site conformation. We applied this method to a dataset of proteins whose structures were determined by NMR and X-ray methods. All chosen proteins had two or more structures, in general involving different bound ligands. By taking one of these structures as a reference, we were able in most cases to successfully reproduce the experimentally determined conformations of the same and other structures, as well as to suggest alternative low-energy conformations of the binding site. There were a few cases where this method failed. However, it was observed that in such cases the bound ligand had induced a high-energy conformation of the binding site. Consequently, an important observation that stems from this work is that for proteins that do not exhibit any significant backbone flexibility, ligands tend to bind to low-energy conformations of the binding site.

We have seen that, in most cases, it is possible to use a rotamer search-based method to predict alternative low-energy protein binding site conformations that can be used by different ligands. This work provides evidence of the validity of generating and using alternative binding site conformations to improve the efficacy of docking and structure-based drug design methods.

Acknowledgements

A.Y.C.Y. thanks De Novo Pharmaceuticals for the award of a studentship. R.L.M. is also a Research Fellow of Hughes Hall, Cambridge, UK.

References

1. Zheng, Q. and Kyle, D.J., Drug Discov. Today, 2 (1997) 229.
2. Walters, W.P., Stahl, M.T. and Murcko, M.A., Drug Discov. Today, 3 (1998) 160.
3. Koshland, D.E., Proc. Natl. Acad. Sci. USA, 44 (1958) 98.
4. Najmanovich, R., Kuttner, J., Sobolev, V. and Edelman, M., Proteins Struct. Funct. Genet., 39 (2000) 261.
5. Frimurer, T.M., Peters, G.H., Iversen, L.F., Andersen, H.S., Moller, N.P. and Olsen, O.H., Biophys. J., 84 (2003) 2273.
6. Lewi, P.J., Jonge, M., Daeyaert, F., Koymans, L., Vinkers, M., Heeres, J., Janssen, A.J., Arnold, E., Das, K., Clark, A.D. Jr, Hughs, S.H., Boyer, P.L., Bethune, M.P., Pauwels, R., Andries, K., Kukla, M., Ludovici, D., Corte, B.E., Kavash, R. and Ho, C., J. Comput.-Aided Mol. Design, 17 (2003) 129.
7. Uytterhoeven, K., Sponer, J. and Van Meervelt, L., Eur. J. Biochem., 269 (2002) 2868.
8. Leach, A.R., J. Mol. Biol., 235 (1994) 345.
9. Anderson, A.C., O'Niel, R.H., Surti, T.S. and Stroud, R.M., Chem. Biol., 8 (2001) 445.
10. Schnecke, V., Swanson, C.A., Getzoff, E.D., Tainer, J.A. and Kuhn, L.A., Proteins Struct. Funct. Genet., 33 (1998) 74.
11. Knegtel, R.M.A., Kuntz, I.D. and Oshiro, C.M., J. Mol. Biol., 266 (1997) 424.
12. John, B. and Sali, A., Nucleic Acids Res., 31 (2003) 3982.
13. Chung, S.Y. and Subbiah, S., In Hunter, L. and Klein, T.E. (Eds.), Pac. Symp. Biocomput., World Scientific, Hawaii, 1996, pp. 126–141.
14. Philippopoulos, M. and Lim, C., Proteins Struct. Funct. Genet., 36 (1999) 87.
15. Clarage, J.B., Romo, T., Andrews, B.K., Pettitt, B.M. and Phillips, G.N. Jr., Proc. Natl. Acad. Sci. USA, 92 (1995) 3288.
16. Chandrasekaran, R. and Ramachandran, G.N., Int. J. Protein Res., 2 (1970) 223.
17. Janin, J., Wodak, S., Levitt, M. and Maigret, B., J. Mol. Biol., 125 (1978) 357.
18. Bhat, T.N., Sasisekharan, V. and Vijayan, M., Int. J. Pept. Protein Res., 13 (1979) 170.
19. Benedetti, E., Morelli, G., Nemethy, G. and Scheraga, H.A., Int. J. Pept. Protein Res., 22 (1983) 1.
20. James, M.N. and Sielecki, A.R., J. Mol. Biol., 15 (1983) 299.
21. Ponder, J.W. and Richards, F.M., J. Mol. Biol., 193 (1987) 775.
22. Gelin, B.R. and Karplus, M., Biochemistry, 18 (1979) 1256.
23. Dunbrack, R.L. and Karplus, M., J. Mol. Biol., 230 (1993) 543.
24. Dunbrack, R.L. and Cohen, F.E., Protein Sci., 6 (1997) 1661.
25. McGregor, M.J., Islam, S.A. and Sternberg, M.J., J. Mol. Biol., 198 (1987) 295.
26. Schrauber, H., Eisenhaber, F. and Argos, P., J. Mol. Biol., 230 (1993) 592.
27. Tuffery, P., Vaney, M.C., Mornon, J.P. and Hazout, S., J. Mol. Graph., 9 (1991) 175.
28. Sali, A. and Blundell, T.L., J. Mol. Biol., 234 (1993) 779.
29. Laughton, C.A., J. Mol. Biol., 235 (1994) 1088.
30. Koehl, P. and Delarue, M., J. Mol. Biol., 239 (1994) 249.
31. Bower, M.J., Cohen, F.E. and Dunbrack, R.L., J. Mol. Biol., 267 (1997) 1268.
32. Källblad, P. and Dean, P.M., J. Mol. Biol., 326 (2003) 1651.
33. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E., Nucleic Acids Res., 28 (2000) 235.
34. Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C., J. Mol. Biol., 247 (1995) 536.
35. Kelley, L.A., Gardner, S.P. and Sutcliffe, M.J., Protein Eng., 9 (1996) 1063.
36. Leach, A.R. Molecular Modelling – Principle and Applications, 2nd edition. Longman, Singapore, 1996.
37. Sutcliffe, M.J., Protein Sci., 2 (1993) 936.
38. Dunbrack, R.L. and Karplus, M., Nature Struct. Biol., 1 (1997) 334.
39. Accelrys Inc., San Diego, CA, USA.
40. Maple. J., Dinur, U. and Hagler, A.T., Proc. Natl. Acad. Sci. USA, 85 (1988) 5350.
41. Dunbrack, R.L., Curr. Opin. Struct. Biol., 12 (2002) 431.
42. West, N.J. and Smith, L.J., J. Mol. Biol., 280 (1998) 867.
43. Marti, D.N., Schaller, J. and Llinas, M., Biochemistry, 38 (1999) 15741.
44. Oh, B.H., Ames, G.F. and Kim, S.H., J. Biol. Chem., 25 (1994) 26323.
45. Källblad, P., Todorov, N.P., Willems, H.M.G. and Alberts, I.L., J. Med. Chem., 47 (2004) 2761.