Computational approaches for drug design at the Protein-Protein interface

Jonathan Christopher Fuller

Submitted in accordance with the requirements for the degree of PhD

The University of Leeds
Institute of Molecular and Cellular Biology

September 2010

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

The first results chapter entitled “Properties of small molecule protein-protein interaction inhibitors and their active volumes” contains work from the publication:

Fuller, J. C., Burgoyne, N. J., and Jackson, R. M. (2009) Predicting druggable binding sites at the protein-protein interface., Drug discovery today 14, 155-61.

JCF carried out all analysis, prepared all figures and wrote the manuscript. RMJ supervised the project and edited the manuscript. NJB wrote software to map pockets identified on apo structures onto pockets observed on bound structures.

Acknowledgements

First and foremost I would like to acknowledge Richard Jackson for his supervision for the previous four years. He has provided encouragement, ideas and critical evaluation of my work that has enabled me to produce this thesis. I would also like to thank Michael Shirts at the University of Virginia for his enthusiasm and diligent guidance for the work presented in the final three results chapters that culminated in free energy calculations performed on oligoamide compounds.

Thank you to István Kolossváry from D.E. Shaw Research for his help in coaxing Desmond to perform the calculations that I wished to perform. Additional thanks to Jeff Wereszczynski for provision of the GAFF parameters in a format compatible for use with Desmond.

I would like to thank Sean Killen and Mike Wallis for providing IT support when required. I would also like to thank Alan Real and Mark Dixon for provision and maintenance of the ARC1 computing cluster that facilitated the free energy calculations. Thanks to David Westhead and Chris Thomas for insightful comments and constructive criticism at key points in my research. Special thanks to Nick Burgoyne; Nicola Gold; Carlos Simões and Sarah Kinnings from the Jackson lab who have taken time to answer my questions and provide critical comments and sometimes useful computer code. Finally I would also like to thank members of the bioinformatics groups at the University of Leeds for their broad range of knowledge that helped with many problems and additionally for 'checking the mood' in the lab. In no particular order: Tom Forth, Matt Care; Liz Webb; John Whitaker; James Bradford; James Dalton; Archana Sharma-Oates; Joe Ward; Joel Dockray; Pete Oledzki; Phil Tedder; Jiantao Yu; Lucy Stead; Al Radford. I also wish to thank Gilly Turner for supporting and encouraging me throughout my studies.

Funding for my PhD studentship was provided in the form of a 4 year studentship from the BBSRC for which I am extremely grateful. Additional funding for my visit to University of Virginia was provided in the form of a CCPB early-stage researcher short visits scheme to the USA for which I would like to thank the organisers.

This thesis is dedicated to my parents Chris and Elizabeth and to my family.

Abstract

The ability to design drugs that disrupt formation of protein-protein interfaces is of particular interest to the pharmaceutical industry due to its promise for opening an entire new range of drug targets, many of which have already been well characterised in terms of their disease causing effect on the human body. Furthermore these interactions can be involved in many processes unique and essential to bacteria and viruses.

We show that pockets on protein-protein interface are smaller but more numerous than those of marketed drugs using a pocket finding algorithm (Q-SiteFinder). We investigate the similarities and differences between several candidate compounds designed to bind and disrupt protein-protein interfaces and compare to those of current marketed drugs designed to bind more traditional protein targets.

We ask the further question as to whether it is possible to better identify pockets on a protein surface as likely to be drug binding. We conclude that it is possible to carefully use random forest machine learning techniques to marginally improve these predictions. However, it is extremely difficult to use simple physical parameters to provide added information as to the maximal affinity that a small-molecule might be able to achieve in a given binding pocket.

Further to the above questions we then investigate the hDM2-p53 system which when disrupted can induce apoptosis in many forms of cancer, making it a target of considerable interest to the pharmaceutical industry. Molecular docking is exploited in order to generate likely structural conformations of oligoamide hDM2-p53 inhibitors which can be used as a starting point for molecular dynamics simulations. These simulations using the AMBER/GAFF force field are then further developed to perform replica-exchange alchemical free energy calculations using the Bennett Acceptance Ratio non-biased estimator. These simulations are in general shown to be very accurate and show promise in generating hypotheses for novel high-affinity oligoamide compounds.

Table of Contents

Table of Figures

Index of Tables

Abbreviations used in the thesis

ΔG – change in Gibbs free energy

ΔG_0 – change in Gibbs free energy under standard conditions

ΔH – change in enthalpy

ΔS – change in entropy

ADMET – Absorption Distribution Metabolism Excretion Toxicity

AM1 BCC – AM1 Bond Charge Correction (charge calculation method)

AMBER – Assisted Model Building with Energy Refinement

ArCO – Aromatic Carbon – Oxygen (bond)

ArNH – Aromatic Nitrogen – Hydrogen (bond)

BAR – Bennett Acceptance Ratio

CAPRI – Critical Assessment of Predicted Interactions

CHARMM – Chemistry at HARvard Molecular Mechanics

CSD – Cambridge Structural Database

DMSO – DiMethyl SulfOxide

ESP – ElectroStatic Potential

FDA – U.S. Food and Drug Administration

FEP – Free Energy Perturbation

FFT – Fast Fourier Transform

GAFF – Generalized AMBER Forcefield

GBSA – Generalized Born Surface Area

GDF – Gaussian Docking Function

GROMACS – GROningen MAchine for Chemical Simulations

HPV – Human Papilloma Virus

HSA – Human Serum Albumin

HSV – Herpes Simplex Virus

HTS – High Throughput Screening

IC_{50} – Inhibitory Constant at 50 %

IL-2 – InterLeukin-2

IL-2Rα – InterLeukin-2 Receptor α

ITC – Isothermal Titration Calorimetry

K_a – association constant

k_B – Boltzmann constant

K_d – dissociation constant

K_{eq} – equilibrium constant

K_i – inhibitory constant

L-BFGS – Low memory -

LCAO – Linear Combination of Atomic Orbitals

LINCS – Linear Constraint Solver

MC – Monte-Carlo

MCC – Matthews Correlation Coefficient

MD – Molecular Dynamics

MEP – Molecular Electrostatic Potential

MMFF94 – Merck Molecular Force Field 1994

MMPBSA – Molecular Mechanics Poisson Boltzmann Surface Area

NMR – Nuclear Magnetic Resonance (spectroscopy)

NOESY – Nuclear Overhauser Effect SpectroscopY

nPT – (constant) number of particles, pressure and temperature

nVT – (constant) number of particles, Volume, Temperature

OPLS – Optimized Potential for Liquid Simulations

PDB – Protein Data Bank

PLI – Protein Ligand Interaction

PME – Particle Mesh Ewald

PPI – Protein-protein interaction

R – molar gas constant

REDIII – RESP ESP charge Derive Version 3

RESP – Restrained ElectroStatic Potential

RETI – Replica Exchange Thermodynamic Integration

RMS(D/F) – Root Mean Squared (Deviation/Fluctuation)

SAR – Structure Activity Relationship

SASA – Solvent Accessible Surface Area

SCOP – Structural Classification of Proteins

STO – Slater Type Orbital

SVM – Support Vector Machine

TI – Thermodynamic Integration

UCSF – University of California San Francisco

WHAM – Weighted Histogram Analysis Method

1 Thesis introduction

The introduction to the work presented in this thesis is split into two main sections, the first provides technical background to many of the computational techniques used during the course of this work, whilst the biological background introduces the systems that are investigated.

1.1 Technical background

1.1.1 Molecular mechanics force fields

Molecular mechanics force fields can be derived from quantum mechanical descriptions of a group of atoms by making two major assumptions. The first is the Born-Oppenheimer approximation which states that electrons travel orders of magnitude faster than nuclei, thus they can be treated such that they will react instantaneously to any movement of the nuclei. The second is that the nuclei can be treated as particles that follow Newtonian dynamics hence the charge associated to the electrons can then be represented as a single point charge, centred on the atomic nuclei.

On making these assumptions it is possible to construct a Hamiltonian that describes the approximate motion of a group of atoms. There are several families of force fields that do this including AMBER, CHARMM, OPLS and GROMOS all of which have been applied to biological systems. All of these force fields take broadly the same shape, but we will focus on the AMBER force fields since they have been used extensively during the course of this work.

1.1.2 AMBER force field

The AMBER force field has the functional form:

$$V(r^N) = \sum_{bonds} k_r (l - l_0)^2 + \sum_{angles} k_\theta (\theta - \theta_0)^2 + \sum_{torsions} \frac{1}{2} V_n \left[1 + \cos(n\omega - \gamma)\right] + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}} \right] \quad (1)$$

Where the first term represents the contribution to the total energy of the bonds, the second term represents the angles, the third term represents the torsions and the final term represents the non-bonded terms comprised of Lennard Jones and coulomb electrostatics respectively.

More simply the bonded interaction term models a covalent bond between to atoms as if the atoms were connected with a spring using a harmonic potential with spring constant k_b, where l and l_0 are the current bond length and the equilibrium bond length. Angle bending is modelled as a harmonic potential with constant k_θ, where θ and θ_0 are the current bond angle and equilibrium bond angle. The torsional potential at angle ϕ, with amplitude determined by V_n, in the AMBER force field is the sum of cosine terms with varying periodicity (determined by the multiplicity parameter n and phase parameter γ) that can be combined as a fourier sum such that it can describe most torsional potentials with three cosines.

Non-bonded interactions are described by a Lennard-Jones 12-6 potential, where $A_{ij} = \varepsilon_{ij}^{*}(R_{ij}^{*})^{12}$ and $B_{ij} = 2\varepsilon_{ij}^{*}(R_{ij}^{*})^{6}$. Here $\varepsilon_{ij} = (\varepsilon_i \varepsilon_j)^{½}$ and $R_{ij}^{*} = R_i^{*} + R_j^{*}$. ε_i and ε_j are the depth of the Lennard Jones well for the i^{th} and j^{th} atoms respectively. R_i^{*} and R_j^{*} is the equilibrium distance of the Lennard Jones well. Electrostatic interactions are modelled by the coulomb potential, where q_i and q_j are the atomic charge of the i^{th} and j^{th} atoms, ε is the permittivity of free space and R_{ij} is the distance between the atom centres of the i^{th} and j^{th} atoms.

1.1.2.a GAFF force field

The GAFF force field was derived with the intention of allowing the extension of the parameters that allow simulation of protein and nucleic acid such that the AMBER family of force fields can be used to study small-molecules containing H, C, N, S, O, P and halogen atoms such as those typically of interest in drug discovery. To this end the authors defined 35 basic atom types alongside 22 special atom types. The atoms are necessarily general in order to allow the force field to represent the vast amount of chemical space that is present in structural databases such as over 500,000 structures in the Cambridge Structural Database(Allen 2002) (CSD), or chemical databases such as PubChem(LY Geer *et al.* 2010) or ZINC(Irwin and Shoichet 2005) which contain orders of magnitude more compounds. The GAFF force field uses the same functional form as AMBER, parameters for charge are added using the HF 6-31G* RESP charge fitting method, although the authors note that since AM1 BCC charges are parametrised in order to reproduce this charge method these may be preferred in cases such as large scale databases of compounds are used, where speed of calculation is important(Wang, Wolf, *et al.* 2004).

In a test of 74 ligand structures minimized using GAFF parameters the authors noted that they achieved a 0.25 Å RMSD from the crystal structure, which was comparable to the Tripos 5.2 force field and better than the MMFF94 and CHARMM force fields(Wang, Wolf, *et al.* 2004). In addition it is possible to modify parameters based on previously published work, or to perform quantum mechanics calculations to validate parameters if the default parameters are suspected to be unsatisfactory.

The rapid calculation of the forces related to the potentials described above is discussed in more detail in the section molecular dynamics. It is of particular importance to bear in mind several caveats that are introduced in these situations and these will be discussed in due course.

1.1.2.b GRID force field

The GRID force field was originally developed by Goodford in 1985 and is probably the first example of estimating the interaction energy of a probe with a given atom type and a protein molecule(Goodford 1985). GRID was used in the development of a potent inhibitor of neuraminidase, which is an early example of the successful application of rational drug discovery(von Itzstein *et al.* 1993). The GRID potential takes the form of the sum of Lennard Jones interactions E_{lj}, electrostatic interactions E_{el} and a hydrogen bonding potential E_{hb}. The Lennard Jones function E_{lj} is defined for all positive values of d less than a cutoff of 8 Å.

$$E_{lj}=\frac{A}{d^{12}}-\frac{B}{d^{6}} \tag{2}$$

$$E_{el}=\frac{pq}{K\zeta}\left[\frac{1}{d}+\frac{(\zeta-\epsilon)(\zeta+\epsilon)}{\sqrt{d^{2}+4s_{p}s_{q}}}\right] \tag{3}$$

where $\varepsilon = 4$, $\zeta = 80$, K is a combination of geometrical values and natural constants, s_p and s_q are the depths of the charges within the protein for atom p and q respectively, and are set to zero if they have fewer than seven protein atoms within 4 Å.

$$E_{hb}=\left[\frac{C}{d^{6}}-\frac{C}{d^{4}}\right]\cos^{m}(\theta) \tag{4}$$

where $C = 0.666D\ d_{min}^{2}$, θ is the angle subtending the bond between the donor atom and hydrogen and the acceptor atom, and $m = 2$. In cases where E_{lj} is repulsive, but E_{hb} is favoured, E_{lj} is set to zero. Additionally probes and protein

atoms are subject to constraints on the number of hydrogen bond acceptors and donors that they can make, any probe will take the most favourable values of these when the energy is calculated.

1.1.3 Parameter generation

The necessity to generate parameters that accurately describe the behaviour of the atoms in the system to be studied was briefly discussed in the section on the GAFF force field. Here we set out in more detail two possible charge calculation methods that can be used in tandem with a charge fitting technique known as RESP. Finally we discuss the generation of torsional potentials.

1.1.3.a Hartree-Fock Calculations

The Hartree-Fock method allows the approximate computation of the ground state wavefunction and energy of a quantum mechanical many body system. Carrying out these calculations then allows the computation of atomic charges for use in molecular mechanics force fields, or derivation of torsional parameters for molecules.

Whilst it is possible to calculate an analytical answer for the ground state wavefunction and energy of a hydrogen atom (and related isotopes), it is not possible to do so for atoms containing more than one electron. In the case of the Helium atom one can use perturbation theory to arrive at an approximate analytical solution, however, these approximations do not hold in the case of heavier atoms. For many body systems it is not possible to know which wavefunction is the correct wavefunction, but using the variation theorem one knows that an approximation to the true wavefunction will always have higher energy than the true wavefunction. Thus by searching for the minima where the first derivative of the energy of the wavefunction is zero one can identify the desired wavefunction. The Hartree-Fock equations are obtained by identifying this

energy minima subject to the condition that the molecular orbitals remain orthonormal. Constrained minimization can be completed using a mathematical technique known as Lagrange multipliers(Leach 2001).

Briefly the main difference between the two body system of the hydrogen atom, and a many body system, is that in the many atom system there are interactions between electrons, thus altering the spin orbital that one electron occupies will change the solution for the remaining electrons. This can be addressed by splitting the system into a fixed part χ_j (the nuclei and electrons in fixed spin orbitals) and a varied part representing a single remaining electron in spin orbital χ_i. Three operators can then be defined that represent the core Hamilitonian, Coulomb interactions and exchange interactions. Taken together we can represent the energy as the sum of these three operators H_i, J_i, K_i acting on χ_i(Leach 2001).

$$\left[-\frac{1}{2}\nabla_i^2-\sum_{A=1}^{M}\frac{Z_A}{r_{iA}}\right]\chi_i(1)+\sum_{j\neq i}\left[\int d\tau_2\chi_j(2)\chi_j(2)\frac{1}{r_{12}}\right]\chi_i(1) \\ -\sum_{j\neq i}\left[\int d\tau_2\chi_j(2)\chi_i(2)\frac{1}{r_{12}}\right]\chi(1)=\sum_j \epsilon_{ij}\chi_j(1) \tag{5}$$

and we then solve the eigenvalue problem, which does not return the orbital χ_i multiplied by a constant, but a series of orbitals χ_j:

$$\hat{F}_i(1)\chi_i(1)=\sum_j \epsilon_{ij}\chi_j(1) \tag{6}$$

we call F_i the Fock operator:

$$\hat{F}_i(1)=\hat{H}^{core}(1)+\sum_{j=1}^{N}[\hat{J}_j(1)-\hat{K}_j(1)] \tag{7}$$

for a closed shell system:

$$\hat{F}_i(1)=\hat{H}^{core}(1)+\sum_{j=1}^{N/2}[2\,\hat{J}_j(1)-\hat{K}_j(1)] \tag{8}$$

It is possible to use the method of Lagrange multipliers previously mentioned to return multipliers that are zero unless i equals j, which converts equation 8 into a standard eigenvalue equation. It is then possible to solve the equations using self-consistent field theory, by the choice of an *ansatz* used to calculate the value of the coulomb and exchange operators, followed by solution of the Hartree-Fock equations, these solutions are used in the next iteration until the results for all electrons remain unchanged(Leach 2001).

For solution of the Hartree-Fock equation of a whole molecule the Linear Combination of Atomic Orbitals (LCAO) method can be used. Typically the one electron orbitals are approximated by a linear combination of Slater type orbitals (STOs), which are in turn approximated by a linear combination of Gaussian-type orbitals for reasons of computability. These Gaussians are known as basis functions and are discussed in slightly more detail below.

1.1.3.b Basis sets

Basis sets are a set of functions that describe the molecular orbitals. One such set of functions that can do this are Slater type orbitals (STOs) which take the form(Hehre 1969):

$$R(r) = Nr^{n-1} e^{-\zeta r} \tag{9}$$

They are particularly difficult to integrate and as such are typically (as is the case in Pople basis sets) replaced by a sum of Gaussian functions that approximate the form of the STO. Gaussian functions take the form:

$$G(r) = x^a y^b z^c e^{-\alpha r^2} \tag{10}$$

In figure 1.1, sums of 1, 2, 3 and 4 Gaussian functions approximating the STO for the 1s orbital where $\zeta = 1$ are shown. As the number of Gaussian functions summed increases the approximation of the STO orbital improves.

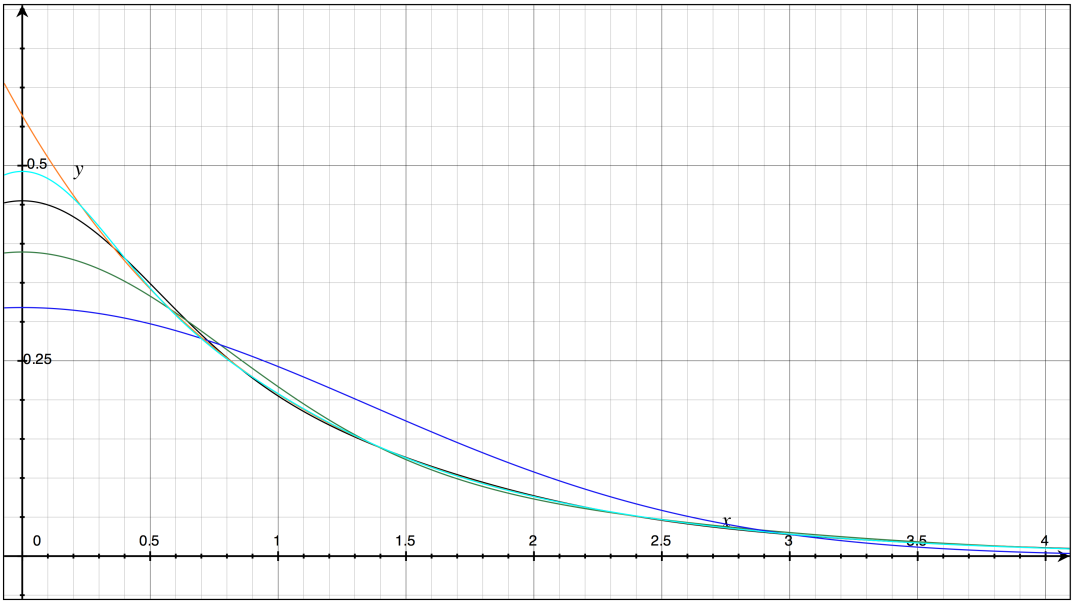

Figure 1.1: Approximation of the STO for 1s orbitals with ζ = 1 (red) by Gaussian functions with one (blue), two (green), three (black) and four (cyan) Gaussian components. The approximation of the integral increases in accuracy as more Gaussian functions are used(Hehre 1969).

Pople basis sets are commonly used for high-level quantum mechanics calculations. Pople basis sets assume that core electrons do not affect the chemical properties of the molecule very much, whilst the valence electrons do vary widely between compounds(Leach 2001). Thus a single function is used to represent the core electrons, X. Two additional sets of Gaussian functions Y and Z represent valence electrons. The value of X, Y and Z determine the number of Gaussians contained in each function. Finally the basis set can contain two additional functions that describe polarization and diffuse wavefunctions, these are represented by * and + respectively. In the studies undertaken in this work we used the 6-31G* basis set, indicating that 6, 3 and 1 Gaussian functions are used and polarization effects are taken into account.

1.1.3.c RESP charge fitting

Charge fitting methods aim to take the Molecular Electrostatic Potential (MEP) and determine a set of atomic charges for the molecule that best describe the MEP. In 1984 Singh and Kollman used a method to determine charges for the AMBER force field that they termed ElectroStatic Potential fit (ESP)(Singh and Kollman 1984). This method uses points on a series of molecular surfaces with gradually increasing van der Waals radii for the atoms. The potential is fitted to points that exist on the family of molecular surfaces. When AMBER94 was parametrised a slightly modified version of this method called Restrained ESP (RESP) was employed(Cornell *et al.* 1993). RESP uses hyperbolic restraints on non-hydrogen atoms, which tends to have the effect of reducing the charge on atoms such as buried carbons, which can sometimes be assigned erroneously high charges when using the ESP method(Leach 2001).

1.1.3.d AM1 BCC Charge Calculations

The AM1 BCC charge method aims to replicate the atomic charges produced when performing a HF 6-31G* calculation followed by ESP charge fitting(Jakalian, Jack, and Bayly 2002). Briefly the method is characterised by calculation of electronic structure and formal charge using the AM1 semi-empirical charge method. This is followed by a Bond Charge Correction (BCC) that has been parametrised by fitting to a set of > 2700 compounds for which atomic charge has been determined by HF 6-31G* calculations followed by charge fitting by ESP(Jakalian, Jack, and Bayly 2002). The method is orders of magnitude faster than standard HF 6-31G* calculations, whilst parametrised for use with the AMBER force field family making it an ideal choice for large molecules, or when dealing with large datasets.

1.1.3.e Torsional parameters

Torsional parameters have been noted to be of particular importance when parameterizing force fields. Whilst many sets of parameters for a wide variety of dihedrals exist in the GAFF force field, it is inevitable that for some compounds the parameters will either not exist or not accurately describe the potential, either in terms of the depth of the well associated with the dihedral, or in the location of the minima. In the case of a group of oligoamide compounds this has been addressed by first optimizing the geometry of a compound that carries the key torsion using HF 6-31G, MP2, BYLP, B3LYP levels of theory. The torsion in question is then scanned at 10(degree) intervals using B3LYP/6-31G(d) and B3LYP/6-311G(d,p) levels of theory, whereby the rest of the molecule (not including the torsion) is optimized at each step(Klein *et al.* 2006). These torsional profiles are then converted to a fourier series for inclusion into the GAFF force field.

1.1.4 Conformer generation

Determination of the bioactive conformation of ligands is a critical step in many forms of structure-based drug discovery. Determination of the bioactive conformation is particularly difficult due to the fact that they often exist in extended conformations(Diller and Merz 2002) with energies several kcal mol^{-1} higher than their global minima(Kirchmair *et al.* 2005). Several programs designed to identify bioactive conformations ligands exist, which broadly fall into one of two categories either generating a single low-energy ligand conformation, or generating an ensemble of conformations of which one is the bioactive conformation. Corina is an example of the first class of programs, whilst Omega, VConf, Dgeom and Balloon all fall into the second category. Previous work in the lab has identified Omega as one of the fastest and most successful in identifying bioactive conformations and for this reason it is used exclusively in this work.

Briefly Omega uses a rule-based, depth-first searching algorithm that generates ensembles. Initially Omega enumerates ring conformations using a fragment library(Agrafiotis *et al.* 2007),(Nicholls, MacCuish, and MacCuish 2004),(Kristam *et al.* 2005),(Good and Cheney 2003),(Boström, Greenwood, and Gottfries 2003),(J Boström 2001). The molecule is then disassembled into fragments of up to five contiguous rotatable bonds whereupon a library of pre-calculated torsions is used to generate different conformations for each fragment. The fragments are then reassembled based on the order of their energies, which generates a pool of alternative molecular conformations. The Merck molecular force field(Halgren 1996) (MMFF) is used to calculate the energy of each conformer, and an adjustable energy threshold, or an adjustable RMSD threshold is used to filter unfavourable conformers.

1.1.5 Adding hydrogen atoms

A related problem to that of conformer generation is the correct assignment of charges to both protonatable residues in proteins, and to protonatable species in ligands. Many of these species will have pK_as that suggest an unambiguous answer for the protonation state at a given pH, but in many cases multiple solutions may be appropriate. The OpenBabel software and OpenEye Babel both provide functionality for adding hydrogen atoms to ligands and have been used widely in this project. Additionally the molecular viewer Chimera and Maestro both have functionality for adding hydrogen atoms to both protein and ligand atoms.

1.1.6 Protein-ligand docking

The field of molecular docking could be defined as “Given the atomic coordinates of two molecules, predict their 'correct' bound association”(Halperin *et al.* 2002). This could involve the desire to determine the structure of protein-protein, protein-nucleic acid or protein-ligand complexes. Discussed below are the issues associated but not limited to protein-ligand interactions. Most often docking is approached as a two stage problem, with the first being the sampling of

conformational space to predict likely orientations and positions of ligand relative to protein. The second is the identification of the correct orientation and position from a list of several candidate poses. The challenge with the first stage is to design algorithms that can determine an ensemble of poses that contain the 'correct' solution using minimal computational power. The challenge in the second stage is to produce scoring functions that are generally applicable to complexes and can identify the 'correct' solution from the list generated by the first stage.

A wide variety of software to perform docking and scoring exist, however, we discuss particularly the software used in this study, Autodock and FRED. Both of which use distinctly different methods to generate likely conformations, and additionally use different scoring functions to attempt to identify the 'correct' binding pose.

1.1.6.a Autodock

Autodock is one of the most widely used docking programs with over 2000 citations of the methods paper, that has the additional benefit of being freely available to academics(Morris *et al.* 1998). The original Autodock software is now at version 4, and includes a genetic algorithm, and a Lamarckian genetic algorithm that both outperform the original simulated annealing method, with the Lamarckian genetic algorithm performing the best. Additionally an improved scoring function has been described(Huey *et al.* 2007). Many studies have been performed using Autodock, and it has been shown to perform well(Park, J Lee, and S Lee 2006), in many cases outperforming other popular docking software such as FlexX(Rarey *et al.* 1996) and DOCK(Ewing *et al.* 2001).

In general genetic algorithms use definitions and ideas from the field of evolution and apply them to problems in computation. In the case of a docking problem variables that describe translation, orientation and conformation of the ligand are called 'state variables'. Each of the state variables describes a 'gene' whereby the

collection of genes for each ligand determines a genotype. The specific atomic coordinates of each ligand correspond to a phenotype. The interaction energy between the protein and ligand is calculated using a force field and describes the 'fitness' of the solution, whereby solutions with better fitness are more likely to reproduce, and less fit solutions are more likely to die. New solutions can be generated by 'crossover' of genes from two parents to produce offspring. In addition mutation effects, whereby a gene is randomly mutated can also be described(Morris *et al.* 1998). The basic genetic algorithm described above can perform a global search of the possible solution space, however, it was also desirable to allow local search once a more favourable region is located and to this end the Lamarckian genetic algorithm was developed.

The Lamarckian genetic algorithm uses the same style of genetic algorithm as above for local search, but also implements a local search similar to that developed by Solis and Wets(Solis and Wets 1981). The combination of these two methods was shown to perform better than either of the methods independently(Morris *et al.* 1998).

Further to their previous work on producing a force field for use with Autodock 3(Morris *et al.* 1998), the authors also developed a new scoring function parameterized on a dataset of 188 protein-ligand complexes with known binding affinity and tested their results on a set of 100 protease inhibitors(Huey *et al.* 2007).

1.1.6.b FRED

The FRED program uses a modified Gaussian Docking Function (GDF) that is a summation of pairwise terms between protein and ligand atoms. The GDF is a function of the area overlap of the two atoms minus the volume overlap of the two Gaussian functions describing each atom multiplied by a constant λ, that minimizes the GDF(McGann *et al.* 2003). In equation 11, $d_{i,j}$ is the distance

between a pair of atom i and j, R_i and R_j are the atomic radii of the atoms i and j, $V_{i,j}$ is the intersection volume between the ligand and the negative image of the protein binding site, κ is a constant optimized during the development of the method. The first two terms correspond to the atomic area matched, whilst the final term corresponds to the overlap of molecular volume.

$$F_{i,j}(d_{i,j}) = R_i \frac{\delta V_{i,j}(d_{i,j})}{\delta R_{i,j}} + R_j \frac{\delta V_{i,j}(d_{i,j})}{\delta R_j} - \lambda \left[\left(\frac{16\kappa^3}{9\pi} \right) \left(\frac{\pi R_i^2 R_j^2}{\kappa R_i^2 + \kappa R_j^2} \right)^{\frac{3}{2}} \exp\left(\frac{\kappa}{R_i^2 + R_j^2} d_{i,j}^2 \right) \right] \quad (11)$$

The final implementation of the GDF used in FRED is slightly modified by the addition of an exponential function to the volume overlap term that acts to penalise clashes between protein and ligand atoms(McGaughey *et al.* 2007). FRED has been shown to be successful in the context of virtual screening(McGaughey *et al.* 2007).

1.1.7 Minimization of a system

Energy minimization of a system to be studied by molecular dynamics is a technique applied to ensure that the simulation is stable to numerical instabilities. An energy minimization algorithm will identify the first and second derivatives of the energy function that describes the system. The coordinates of the system are then varied until the first derivative of the energy function is zero, and the second derivative is greater than zero. This means that the system is in a local energy minimum. A variety of energy minimization routines exist for a variety of purposes. Conjugate gradient minimization and l-bgfs minimization are both suitable for use with molecular dynamics simulations(Liu and Nocedal 1989). When minimizing systems on a grid such as is often the case in molecular docking alternative algorithms such as the downhill simplex algorithm of Nelder and Mead or Solis and Wetts are required(Nelder and Mead 1965),(Solis and Wets 1981).

1.1.8 Molecular dynamics

Molecular dynamics is the process of numerically integrating Newton's laws of motion for a system of many particles interacting as described by a force field such as the AMBER/GAFF force field described in equation 1. There are many programs available to perform molecular dynamics simulations, many of which can perform very similar calculations. We focus on GROMACS and Desmond, which have both been designed to perform fast numerical integration and have support for the AMBER/GAFF force fields.

1.1.8.a Integration

Newton's second law of motion can be stated in differential form as:

$$\frac{d^2 x_i}{dt^2} = \frac{F_{x_i}}{m_i} \tag{12}$$

It is clear then that integration of the equation allows determination of the velocity of the particle i, this equation can then be integrated again to determine the position x_i of the particle at a later time. In general there are no exact solutions to the general n-body problem, so numerical integration must be used to determine the velocities and positions of the particles in the system at a later time. There are many algorithms that can do this, we discuss the leap frog and verlet algorithms since they are relatively simple and implemented in GROMACS and Desmond.

The Verlet algorithm uses the positions and accelerations at time t, and positions from the previous step $\vec{r}$ (t - dt) to calculate the new position at t + dt, $\vec{r}$ (t + dt):

$$\vec{r}(t+\delta t) = \vec{r}(t) + \delta t \vec{v}(t) + \frac{1}{2}\delta t^2 \vec{a}(t) - \ldots \tag{13}$$

$$\vec{r}(t-\delta t) = \vec{r}(t) - \delta t \vec{v}(t) + \frac{1}{2}\delta t^2 \vec{a}(t) - \ldots \tag{14}$$

Adding the two equations gives:

$$\vec{r}(t+\delta t)=2\vec{r}(t)-\vec{r}(t-\delta t)+\delta t^2\vec{a}(t) \quad (15)$$

A variation of the Verlet algorithm called the leap-frog algorithm has also been developed, which uses the relationships:

$$\vec{r}(t+\delta t)=\vec{r}(t)+\delta\vec{v}(t+\frac{1}{2}\delta t) \quad (16)$$

$$\vec{v}(t+\frac{1}{2}\delta t)=\vec{v}(t-\frac{1}{2}\delta t)+\delta t\vec{a}(t) \quad (17)$$

So the velocities at time $\vec{v}$ (t + ½dt) are calculated from the velocities at time $\vec{v}$ (t – ½dt) and the acceleration at time t, followed by the positions at time $\vec{r}$ (t + dt) from the velocities previously calculated.

Another possible method of integration is the velocity Verlet method, which has a significant advantage over the leap-frog algorithm since it can provide the positions, velocities and accelerations at the same time:

$$\vec{r}(t+\delta t)=\vec{r}(t)+\delta t\vec{v}(t)+\frac{1}{2}\delta t^2\vec{a}(t) \quad (18)$$

$$\vec{v}(t+\delta t)=\vec{v}(t)+\frac{1}{2}\delta t[\vec{a}(t)+\vec{a}(t+\delta t)] \quad (19)$$

The velocity Verlet method requires a three-step methodology since the acceleration at time t and t + dt is required, so once the positions at time t + dt is calculated the velocities at time t + ½dt can be calculated:

$$\vec{v}(t+\frac{1}{2}\delta t)=\vec{v}(t)+\frac{1}{2}\delta t\vec{a}(t) \quad (20)$$

The previous equation allows the final step of the calculation detailed in equation 19 to be carried out. Use of a velocity Verlet method is required if the use of certain barostats such as the Anderson or Martyna-Tobias-Klein barostat is desired.

1.1.8.b Temperature control

Since molecular dynamics simulations are aiming to reproduce macroscopically observable quantities of the simulated system such as the free energy of binding between a protein and a ligand, simulations will often require to be performed in the same thermodynamic ensemble as the experiment is performed. Most chemical reactions happen in the isothermal-isobaric ensemble (or nPT ensemble) meaning that the number of particles in the system n, the pressure P and the temperature T are all constant when time averaged. For this reason a thermostat and barostat for the system are required to generate the correct thermodynamic ensemble. Several methods for implementing thermostats and barostats exist.

One of the simplest methods of controlling the temperature of the system is the Berendsen thermostat, which is implemented in both Desmond and GROMACS. The Berendsen thermostat mimics weak coupling of the system to an external heat bath with temperature T_0. This has the effect of slowly correcting the temperature towards the target temperature of T_0 subject to the equation:

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau} \qquad (21)$$

Practically this means that the temperature decays exponentially towards T_0 with a time constant , which can be particularly useful for equilibration of systems. The Berendsen thermostat does not however generate the nPT thermodynamic ensemble since it suppresses fluctuations in the kinetic energy of the system(D. Van Der Spoel, E. Lindahl, B. Hess, A. R. Van Buuren, E. Apol and D. P. Tieleman, A. L. T. M. Sijbers, K. A. Feenstra 2005).

1.1.8.c Pressure control

There is a Berendsen algorithm for pressure control which uses similar ideas to the Berendsen thermostat, whereby the volume of the box is altered at each step. Here P is the current system pressure, P_0 is the reference temperature and $_P$ is the time constant.

$$\frac{dP}{dt} = \frac{P_0 - P}{\tau_P} \tag{22}$$

at each step the box volume is scaled by a factor η and the coordinates and velocities are rescaled by the cube root of η. In this equation ß is the isothermal compressibility of the system.

$$\eta = 1 - \frac{\beta \delta t}{\tau_p}(P_0 - P) \tag{23}$$

Once again the Berendsen barostat can suffer from the problem that it does not generate the correct thermodynamic ensemble, so whilst it is useful for equilibrating systems to a desired pressure it is not appropriate for use in situations whereby simulation in a well-defined thermodynamic ensemble is required(D. Van Der Spoel, E. Lindahl, B. Hess, A. R. Van Buuren, E. Apol and D. P. Tieleman, A. L. T. M. Sijbers, K. A. Feenstra 2005).

The extended ensemble Parrinello-Rahman barostat is another alternative for pressure control in MD simulations. However, we discuss the Martyna-Tobias-Klein combination since we use it in free energy calculations discussed in later chapters.

The Martyna-Tobias-Klein nPT system in Desmond is essentially using a Nose-Hoover temperature control system, with the pressure controlled by a constant pressure, constant entropy piston(Martyna, Tobias, and ML Klein 1994).

1.1.8.d GROMACS

GROMACS 4.0 has been designed to be a fast, open and scalable piece of software for performing molecular dynamics simulations(Hess *et al.* 2008). Earlier versions of GROMACS took advantage of code optimizations for specific computer architectures(Lindahl, Hess, and Spoel 2001). Version 4.0 builds on these foundations and attempts to improve the scalability of the algorithms employed such that multiple cpus can be employed to simulate a single system without loss of speed. One of the main benefits of GROMACS is the user community that has developed many analysis tools for GROMACS trajectories(Hess *et al.* 2008). One of the major disadvantages of GROMACS is that it currently employs a leap-frog algorithm as the main integrator meaning that Nose-Hoover pressure control is not available, however this is being addressed with future releases expected to support use of velocity verlet integrators.

1.1.8.e Desmond

The Desmond algorithm was originally designed as a scalable algorithm for a custom built simulation engine called ANTON. However, the Desmond code has been ported to run on x86 based processors and shows speed similar to that of other high-performance MD codes(Bowers *et al.* 2006),(Hess *et al.* 2008). Desmond has been used for performing free energy calculations and benefits from a bundled GUI that in some cases can aid in designing simulations. Furthermore it uses the leapfrog algorithm, enabling use of Nose-Hoover thermostating/barostating(Bowers *et al.* 2006).

1.1.9 Free energy calculations

Accurate determination of the free energy of protein-ligand association is of obvious interest due to the direct relationship to the binding affinity.

$$\Delta G = \Delta_r G^0 + RT \ln K_{eq} \qquad (24)$$

Here ΔG is the change in Gibbs free energy, $\Delta_r G^0$ is the change of reaction free energy under standard conditions, R is the molar gas constant, T is the temperature and K_{eq} is the equilibrium constant of the reaction which in this case will be the dissociation constant K_d. Since we want to determine the equilibrium free energy of the association, we know that ΔG is zero, so we can rearrange to show:

$$K_d = e^{-\frac{\Delta_r G^0}{RT}} \tag{25}$$

Computationally direct simulation of free energy is very demanding, since whilst thermodynamic properties such as internal energy or pressure rely on sampling of low energy regions of phase space, accurate determination of free energy depends heavily on high energy regions of phase space, which are not preferentially sampled by molecular dynamics simulations. As a result of these sampling difficulties much work has been carried out to investigate novel methods to obtain accurate, well converged values of free energy. Even more difficult to achieve is accurate, well converged values for entropy and enthalpy of the association. For constant nPT systems free energy is related to enthalpy change (ΔH) and entropy change (ΔS) by the simple thermodynamic relationship:

$$\Delta G = \Delta H - T\Delta S \tag{26}$$

So given that free energy has already been determined it would seem that determination of the enthalpy of the system will elucidate the entropy of the system. The simplest way to do this would be to determine the average enthalpy in state A and state B and subtract. However the difference is a small number and the values are large numbers that scale with system size. This means that in practice the error on the value can be an order of magnitude larger than the error on measurement of the free energy(Chipot and Pohorille 2007). From a pragmatic point of view whilst the entropy and enthalpy contributions to binding may be of interest to a medicinal chemist, the computational chemist is perhaps only

interested in the end goal of developing a compound with high binding affinity. The enthalpic and entropic contributions may be best determined by methods such as Isothermal Titration Calorimetry (ITC)(Freire 2008).

It is possible to consider two states A, and B for which we are interested in calculating the free energy difference between. In this chapter state A might be an oligoamide compound of interest, whilst state B might be a reference oligoamide compound. If we were then to introduce a third state C which represents a third different oligoamide compound, we could calculate the difference in energy between C and B, allowing us to rank compound A and C in terms of their binding affinity.

1.1.9.a Perturbation theory

We will now consider the method of thermodynamic perturbation, often called free energy perturbation (FEP). If we consider the difference in free energy between the previously mentioned states A and B.

$$\Delta G = G_B - G_A = -k_B T \ln \frac{Z_B}{Z_A} \tag{27}$$

Here we use G to represent quantities of free energies, k_B is the boltzmann constant, T is the temperature of the system and Z_A, Z_B are the partition functions of the respective states. Substituting in the equation for the two partition functions and simplifying leads to:

$$\Delta G = -k_B T \left\langle \exp - \left[\frac{(H_B - H_A)}{k_B T} \right] \right\rangle_0 \tag{28}$$

The subscript zero shows that the average is calculated over the conformations generated whilst sampling in state A. Simulating the reverse process we can swap the subscript zero for one (indicating sampling in state B), and switching the order of subtraction of Hamiltonians. This method has been attributed to

Zwanzig(Zwanzig 1954). Convergence of calculations using this method can be checked for convergence by carrying out the forward and reverse calculation to see whether they agree. The method is unlikely to converge if state A does not overlap in phase space with state B(Leach 2001).

1.1.9.b Coupling Parameter

In the case of creation/annihilation or mutation of atoms as in alchemical free energy calculations we can use a coupling parameter **λ**. When considering mutation between two states A and B we can describe state A as having **λ** = 0, and state B as having **λ** = 1. This can allow us to define intermediate states such that when state A and state B do not have good phase space overlap we can define an intermediate state Z that has overlap with both state A and state B. Whilst state A and state B are likely to have a physical reality, state Z can be an alchemical intermediate, with some degree of the interactions from both physical states. The coupling parameter **λ** describes this intermediate state.

1.1.9.c Thermodynamic integration

Thermodynamic integration (TI) is an alternative method that can be applied to calculated the free energy difference between two states. The formula that describes the method is:

$$\Delta G = \int \left\langle \frac{\partial H(p^N, r^N)}{\partial \lambda} \right\rangle_\lambda d\lambda \qquad (29)$$

Conceptually the change in the energy between two lambda states is calculated, and the area under the line that this represents is integrated using the trapezium rule(Leach 2001).

1.1.9.d Non-equilibrium methods

A counter intuitive approach to gaining estimates of the equilibrium free energy, is to approach the problem by looking at non-equilibrium simulations. Perhaps surprisingly this leads to several successful methods for predicting the equilibrium free energy. Many of these methods are reviewed in detail in the book by Chipot and Pohorille(Chipot and Pohorille 2007) and additionally in the review article by Cossins *et al.*(Cossins *et al.* 2009).

1.1.9.e Jarzynski's Identity

For a slow growth simulation that by definition cannot proceed infinitely slowly we do know that the work W( ) performed on the system during the transformation is on average greater than or equal to the free energy difference ΔG between the states at **λ** = 0 and **λ** = 1.

$$\langle W(\tau)\rangle \geqslant \Delta G \tag{30}$$

Jarzynski showed that it is possible to convert the inequality into an equality provided that a defined path **λ**(t) connects the initial and final states. Jarzynski showed that the left hand site of the equation which is the exponential average of a set of non-equilibrium work values W( ) can yield an equilibrium free energy ΔG.

$$\langle \exp[-\beta W(\tau)]\rangle = \exp(-\beta \Delta G) \tag{31}$$

In reality the exponential average here has been observed to be noisy and biased, thus the average depends strongly on behaviour at the tails of the distribution which are not as well sampled, which can lead to problems with convergence of the free energy values(Shirts *et al.* 2003).

1.1.9.f Weighted Histogram Analysis Method

Statistical mechanics tells us that the probabilities of observing a given macrostate of the system are linked to the partition function of the system. Thus observations from a simulation of the system can allow determination of the density of states,

where ß is the inverse of the boltzmann constant multiplied by the system temperature, Ω(U) is the probability distribution of macrostates with energy U, and Z(T) is the partition function of the system.

$$p(U;T)=\frac{e^{-\beta U}\Omega(U)}{Z(T)} \tag{32}$$

If we split this continuous measure of probability into finite bins, and let f(U) be the number of times an energy in the range [U,U+ΔU] is observed in a simulation we can write a formula for the normalized observed energy distribution.

$$p(U;T)=\frac{f(U)}{\Delta U\sum_{U'} f(U')} \tag{33}$$

substituting into our original equation:

$$\Omega(U)=p(U;T_0)e^{\beta_0 U}Z(T_0) \tag{34}$$

That is to say that an estimate of the density of states can be determined from a simulation at temperature T_0. It is then possible to use the density of states to calculate thermodynamic properties of interest. There are a few problems in the above method that come to light when performing a more thorough analysis. Although in principle we can measure the density of states from a simulation at any temperature, in practice the potential energies at a temperature far from the temperature of interest will lead to poor statistics(Chipot and Pohorille 2007).

Ferrenberg and Swendsen developed a procedure for incorporating data from simulations at multiple temperatures (or coupling parameters) to predict ensemble averages such as free energy(Ferrenberg and Swendsen 1989). The idea is that the contribution to the ensemble average from each histogram should be based on the error associated to that histogram. Thus histograms with lower errors will

contribute more to the final measurement. An illustrative derivation is shown in the book by Chipot and Pohorille(Chipot and Pohorille 2007), and a detailed derivation is presented in the book by Frenkel and Smit(Frenkel and Smit 2002).

1.1.9.g Bennett Acceptance Ratio

One of the major problems with WHAM is that the choice of centre and bin width of histogram can bias the free energy estimate. The Bennett Acceptance Ratio was originally shown to be an unbiased estimator of the free energy(Bennett 1976), and has subsequently been shown to be analogous to WHAM in the limit of histograms with infinitesimally narrow width(Shirts and Chodera 2008).

Writing the Crooks relation - which looks at the relationship between work and free energy when looking at the forward and reverse transformations - as an average:

$$\int f(W;\Delta G)\, e^{-\beta W} p_f(W)\, dW = \int f(W;\Delta G)\, e^{-\beta \Delta G} p_b(W)\, dW \qquad (35)$$

The function f(W; ΔG) is an arbitrary function, and p_f and p_b are the forward and backward transformation probability densities respectively. Bennett then showed that the function that minimized the mean square error of the free energy is:

$$f(W;\Delta G) = \left[\frac{e^{-\beta(W-\Delta G)}}{N_f + \frac{1}{N_b}} \right]^{-1} \qquad (36)$$

where N_f and N_b are the number of forward and backward trajectories. This yields an implicit equation that can be solved for ΔG by a Newton-Raphson method.

$$\sum_{i=1}^{N_f} \frac{1}{1 + \frac{N_f}{N_b} e^{\beta(W_i - \Delta G)}} = \sum_{i=1}^{N_b} \frac{1}{1 + \frac{N_b}{N_f} e^{\beta(W_i - \Delta G)}} \qquad (37)$$

Here W is the work in the forward direction, and $\bar{W}$ is the work in the backward direction(Chipot and Pohorille 2007).

1.1.9.h Lambda dynamics

Lambda dynamics is a non-equilibrium method where **λ** is an explicit degree of freedom. In lambda dynamics the system is slowly changed between **λ** = 0 and **λ** = 1 over the course of time. In the limit that **τ** → ∞ the transformation tends to an equilibrium simulation.

$$\Delta G = \lim_{\tau \to \infty} \int_0^{\tau} \frac{\partial H}{\partial \lambda}_{\lambda = \lambda(t)} \dot{\lambda}(t)\, dt \tag{38}$$

Lambda dynamics have been used in many situations, Khandogin and Brooks used lambda dynamics to model protonation states of protonatable residues in constant pH molecular dynamics simulations(Khandogin and Brooks 2005), whilst Michel *et al.* used water molecules with variable lambda states in monte carlo simulations to predict the location of structural waters(Michel, Tirado-Rives, and Jorgensen 2009).

1.1.10 Binding site detection

A relatively large number of pocket detection algorithms have been proposed. These pocket detection algorithms can be grouped according to the principles that they make use of in order to make their pocket predictions. The largest number of algorithms for pocket detection fall into the category of geometry based methods. These methods all use geometric considerations to define pockets. Studies have shown that the ligand binding site is commonly found in the largest geometric pocket(Laurie and Jackson, 2005). Energy based techniques have also been developed to allow the calculation of the point interaction energy between a probe molecule (e.g. a methyl, hydroxyl or amine group) and the protein(Laurie and Jackson, 2005).

The approach of using Q-SiteFinder to analyse bound protein-ligand, or protein-protein complexes can give insight into the nature of the interactions, but in many cases the structure of the bound complex may not be available, there may

however, be an unbound structure available. To this end it would be desirable to determine a method for assessing the usefulness of our predictions on unbound protein complexes.

Previously in the Q-SiteFinder paper(Laurie and Jackson, 2005) analysis of prediction of binding pockets was undertaken by superposing bound protein with unbound partner protein, and defining the ligand from the bound protein to be the unbound protein 'ligand'. A successful prediction on the unbound protein was defined with the same 25 % precision cut-off. Q-SiteFinder was shown in this case to be less successful when predicting on unbound rather than bound structures, with fewer structures in the top 3 sites. The method described above has some problems when applied to predicting protein-protein interaction pockets from unbound structures. Previous work in the lab has investigated 3 methods that may be applied to prediction of protein-protein interface pockets(Burgoyne, 2007). Briefly these were: pockets determined on the unbound structure within 5 Å of the interacting protein superposed on the bound structure; pockets determined on the unbound structure that are within 1 Å of a grid point that defines a pocket when superposed with the bound structure; clefts that are lined with the surface of interface residues in the bound complex. The best of these methods was determined to be assessing the mapping of pockets to Solvent Accessible Surface Area (SASA) on the bound protein. A comprehensive discussion of pocket detection algorithms is covered by Laurie *et al.*(Laurie and Jackson 2006) and some recent developments are included in the work by Fuller *et al.*(Fuller, Burgoyne, and Jackson, 2009).

1.2 Biochemical background

1.2.1 Protein-protein interactions

Protein-protein interactions may be defined as interactions that occur between two or more protein chains. However, we make further distinction based on the size of the interacting protein chains. We define a ‘pure’ protein-protein interaction as one occurring between two distinct protein domains. Making this size dependent distinction causes us to exclude protein-peptide interactions (such as those between SH2 domains and their interacting peptides). The small size (and associated flexibility) of peptides allows them to be treated using techniques that would usually be applied to small-molecule compounds.

Protein-protein interactions have been studied widely from both experimental and computational perspectives. Interface size, shape and hydrophobicity have been investigated. Protein-protein interfaces have been observed to be large (~1,500 - 3,000 $Å^2$) compared to protein-small-molecule interactions (~ 300 - 1,000 $Å^2$).

The concept of ‘hot-spots’ with regard to protein-protein interactions was brought to the fore by Clackson and Wells(Clackson and J A Wells 1995). Using alanine-scanning mutagenesis, whereby residues from each target are systematically mutated to an alanine and the resulting change in binding free energy is measured. They demonstrated for the 30 contacting residues in the interaction between human growth hormone (hGH) and the extra-cellular domain of its first bound receptor (hGHbp) that a central hydrophobic region dominated by two tryptophan residues accounts for three-quarters of the total binding affinity. Clackson and Wells observed that the residues surrounding the two important tryptophan residues were generally hydrophilic and partially hydrated(Clackson and J A Wells 1995). This work was vastly expanded on by Bogan and Thorn who compiled a database of 2,325 alanine scanning mutagenesis experiments(Bogan

and Thorn 1998). The resulting analysis of this database indicated that the free energy of binding is not uniformly distributed across the binding interface. They also observed that interfaces are often enriched in tryptophan, tyrosine and arginine. Their final major observation was that once again the energetically less important residues surrounding these residues appear to occlude bulk solvent from the hot-spot, which is a necessary condition for energetically favourable interactions(Bogan and Thorn 1998). An example of a 'hot-spot' is shown in figure Error: Reference source not found.

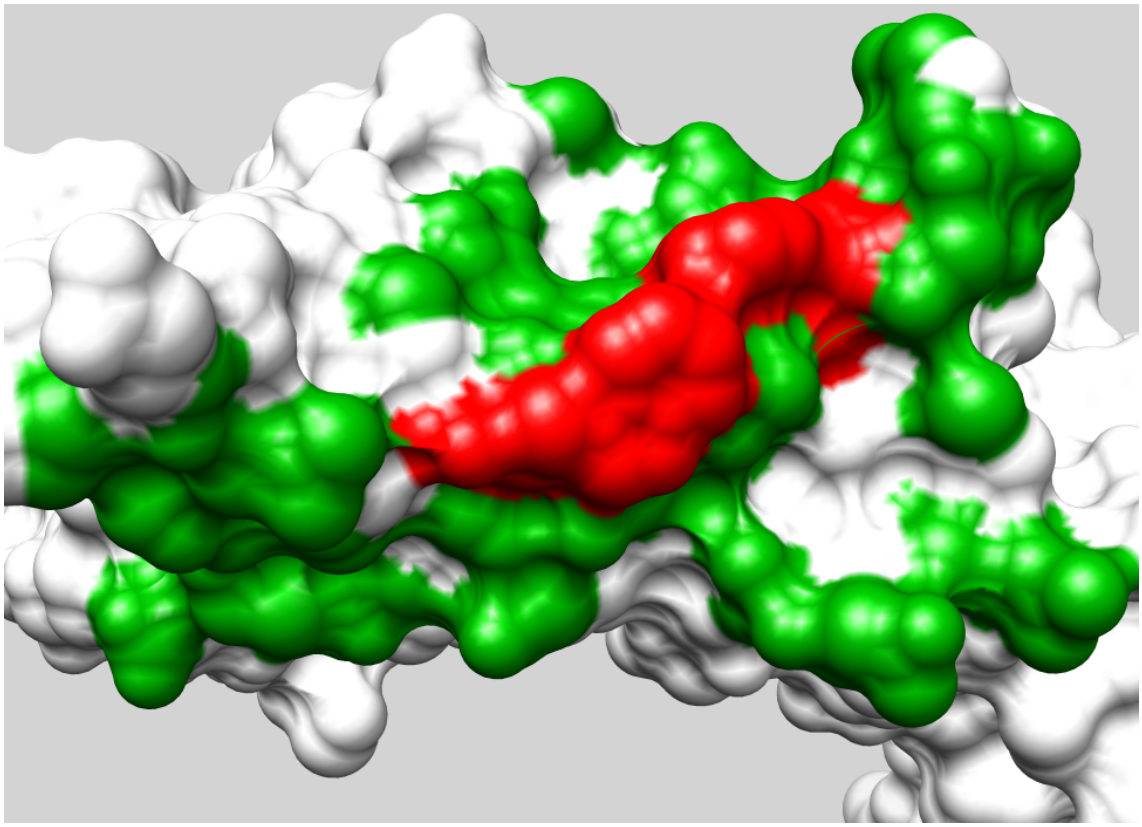

Figure 1.2: A hot-spot on the human Growth Hormone Receptor is shown in red. The two tryptophan residues contribute the most to the overall binding affinity to human Growth Hormone, with 8 out of 31 residues contributing 85 % of the total binding energy. The green region shows the area over which the human Growth Hormone protein interacts(Clackson and J A Wells 1995). Image generated with Chimera, using PDB id 1A22.

Protein-protein interactions are ubiquitous in cell biology with estimates based on model organism interactome data, protein domain structure, genome wide gene expression data and functional annotation data putting a conservative estimate on the number of interacting proteins at around 40,000 interacting protein pairs(Rhodes *et al.* 2005). Protein interaction networks contain an unprecedented level of complexity and diversity with a multitude of distinct feedback loops and control mechanisms allowing a wide regulatory impact. Malfunction of these control mechanisms can lead to disease. As a result it would be desirable to be able to target a specific protein-protein interaction that is acting as a switch for the transition between a normal, and a disease state.

An example of a specific case whereby disrupting a protein-protein interaction would give desirable therapeutic benefits is the Bcl-2 family of apoptosis regulators. The Bcl-2 family of proteins is involved in the regulation of programmed cell death (apoptosis), by controlling mitochondrial outer membrane permeabilzation. Many cancer cells express anti-apoptotic Bcl-2 family members in order to avoid apoptosis. It has been shown that selective antagonism of Bcl-2 family members is possible(Letai 2005), and effective in treating cancer cells *in-vitro* and *in-vivo*(Oltersdorf *et al.* 2005).

Chemical genetics is the study of gene-product function at the level of the cell or organism. In this approach, small molecules that bind directly to proteins are used to alter protein function. The effect of altering protein function is then analysed by observing the kinetics of the system *in-vivo*. Modulation of protein-protein interactions would also open up a whole new area of chemical genetics whereby biologists could probe specific interactions *in-vivo*.

There has been limited success in modulating protein-protein interactions so far. The vast majority of the therapeutics that have currently been marketed are biomolecules, such as antisense antibodies, and peptide therapies. These types of therapeutics have many disadvantages, such as high cost, and lack of oral bioavailability (meaning that they cannot be administered orally). As such it is desirable to target these interactions using small-molecules. Small-molecules have many advantages, particularly they are generally cheaper than other forms of therapeutics, and can be administered orally(Wells and McClendon, 2007).

1.2.2 Small-molecule protein-protein interaction inhibitors

Here we define small-molecule protein-protein interaction inhibitors (PPI inhibitors) as small-molecules that directly compete with one of the protein partners from a discontinuous protein-protein interface(Wells and McClendon, 2007). This definition specifically excludes protein-linear peptide motifs e.g. SH2 domains, and also allosteric inhibitors e.g. those targeted against TNFα(Berg 2003). As previously mentioned it would be desirable to be able to produce small-molecules that were capable of inhibiting protein-protein interactions particularly with respect to drug discovery. As such a brief overview of progress in the field of producing Protein-protein interaction (PPI) inhibitors suitable as therapeutics is presented.

The discovery of small-molecules that modulate protein-protein interactions has largely been unsuccessful(Fry 2006), with only one drug on the market falling into this class of therapeutics(D Kuritzkes, Kar, and Kirkpatrick 2008). Previously protein-protein interactions have come to be thought of as ‘difficult’, ‘high-risk’ or even ‘undruggable’(Whitty and Kumaravel 2006). However, in recent years there have been several successes allowing researchers to become more confident in protein-protein interactions as a possible therapeutic target(Wells and McClendon, 2007),(Fry 2006),(Arkin and Wells, 2004),(Chène 2006). There are many strategies that have been discussed for successfully targeted protein-protein interactions.

1.2.2.a A model for PPI inhibitor drug discovery

One such approach to determining the drugability of protein-protein interfaces is to use a decision tree to aid selection(Chène 2006). Chène makes two important points before describing a decision tree. The first is that even if an interface doesn't fit the decision tree, it may still be possible to obtain molecules that prevent interface formation. Secondly that even if it is possible to determine inhibitors with IC_{50} values in the low micro molar range, a large number of these compounds will never enter clinical use. The decision tree proposed by Chène is shown in figure 1.3 Chène argues that it is favourable to have structural information on the target allowing the drug-discovery process to be guided by this information. Having access to the structure of the target will be useful in evaluating the target prior to embarking on the costly process of drug discovery, whilst also helping to improve the potency of compounds during the optimization phase. Once an interface for which a structure is available has been selected the structure can be evaluated for the presence of cavities. The presence of a well-defined pocket across the contact region of the two proteins is the most favourable scenario, as the presence of such a pocket is more likely to allow the formation of a stable protein-inhibitor complex. However, it is important to bear in mind the fact that the contact region of some non-complexed proteins is flexible, and allows conformational changes on binding. One particular example is the Interleukin-2 (IL-2)/Interleukin-2 receptor α (IL-2Rα) complex which has been shown to have a flexible binding interface. On binding of the small-molecule SP-4206 to IL-2, the molecule accesses a pocket, which is not observed in either the apo, or the IL-2Rα bound structure(Thanos, DeLano, and JA Wells 2006). Once a suitable pocket has been identified it is most favourable if the pocket contains hydrophobic residues, which favour the design of lipophilic inhibitors. Once a suitable hydrophobic pocket has been identified the size of the pocket can be determined. The size of the pocket should be sufficient to accommodate an inhibitor. Analysis of 20 marketed drugs has shown that they have solvent-accessible surfaces that range from 150 $Å^2$ to 500 $Å^2$(Gadek and Nicholas 2003). Chène also suggests that the pocket should not be too large, such that key residues interacting with the

inhibitor are not too distant(Chène 2006). The final branch in the decision tree is shape complementarity between the two interacting subunits within the pocket. Chène suggests that the less favourable case occurs when the two interacting proteins make many densely packed contacts. This scenario would make it hard for inhibitors to make additional new contacts to enhance their potency, whilst also mimicking the interactions that are made as standard by the interacting protein partners(Chène 2006).

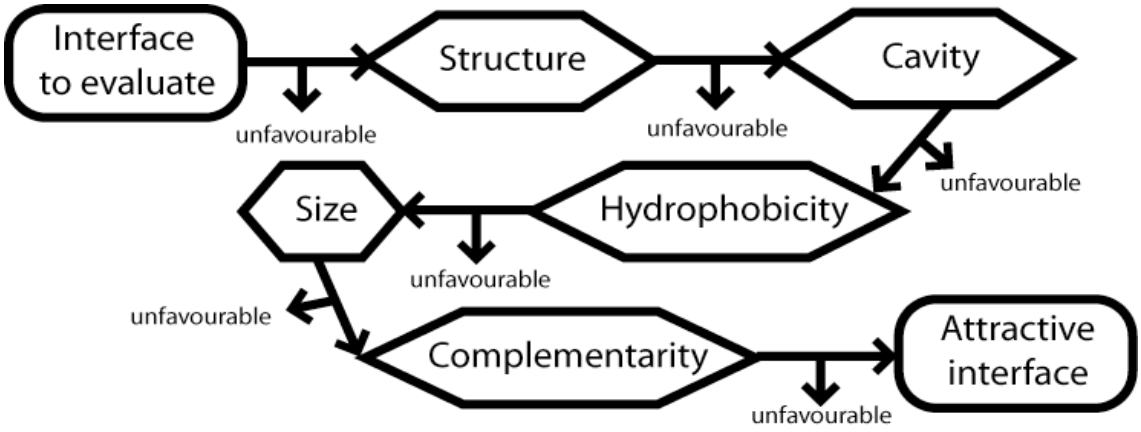

Figure 1.3: Proposed decision tree for evaluating potentially druggable protein-protein interfaces. Adapted from Chène(Chène 2006).

Several PPIs have been investigated during the course of this investigation. Each of them is discussed giving some background to their mechanism of action, the approach applied to small-molecule drug discovery, and the success of any resulting compounds from the drug discovery effort.

1.2.2.b A database of protein-protein interactions

TIMBAL is a hand-curated database of protein-protein interactions and inhibitors mined from literature. The properties of small-molecules from this database have been investigated where it was found that many of them have properties that are fairly close to those of existing drug-like compounds. Furthermore the database is linked to CREDO and PICCOLO, databases of protein-ligand, and protein-protein interactions(Higueruelo *et al.*, 2009).

1.2.3 Protein-protein interactions of interest

1.2.3.a Interleukin-2/Interleukin-2Rα

Interleukin-2 (IL-2) is a 133 residue cytokine, playing a role in growth, activation and differentiation of T cells. IL-2 binds to a heterotrimer (IL-2Rα, IL-2Rβ, IL-2Rγ) located on the T cell surface. Monoclonal antibodies that recognize IL-2Rα and block binding of IL-2 have been developed and marketed as Basiliximab and Daclizumab respectively(Waldmann 2003). The success of these antibodies in producing a clinical effect to suppress the immune response associated with organ transplant rejection validates the IL-2/IL-2Rα interaction as a therapeutic target. X-ray crystallography and NMR spectroscopy have been used to structurally characterize IL-2(Fry 2006). The pharmaceuticals company Roche attempted to find small non-peptidic compounds that would act as an inhibitor of the interaction by binding to IL-2Rα. They prepared a series of acylphenylalanine derivatives designed to mimic Arg38 and Phe42 from the IL-2 binding epitope. One compound was found to have an IC_{50} value of ~ 45 μM. This lead compound was further optimized resulting in a compound with an IC_{50} value of 3 μM. It was then discovered that whilst the acylphenyalanine derivatives were designed to bind IL-Rα the optimized lead compound actually bound to IL-2(Fry 2006). X-ray structures of IL-2/IL-2Rα complex allowed post-analysis of the design strategy that had been employed. It showed that the idea to attempt to mimic Arg38 and Phe42 appeared to be reasonable, but that the pocket on IL-2Rα was quite shallow. It also showed that the optimized compound was a successful (but lucky) mimic of IL-2Rα. Further optimization of the 3 μM compound realised an inhibitor successful to 60 nM, and have gone on to produce 600 nM inhibitors with completely non-peptidic scaffolds(Fry 2006). The IL-2/IL-2Rα case study suggests that mimicry of key side-chain interactions may be a successful strategy with which to target PPIs. As of April 2008, it appears that IL-2/IL-2Rα drug discovery is yet to bring a small-molecule inhibitor to the market.

1.2.3.b Bcl-X_L/Bak-BH3 or Bcl-2/Bak-BH3

Drug resistance has been identified as a major hurdle in developing effective chemotherapeutics(Gottesman 2002). The majority of current chemotherapeutics damage cellular components, which can lead to a wide variety of undesirable post-damage responses. One of the desirable effects of a chemotherapeutics is to induce apoptosis, a mechanism for regulated cell death. Bcl-2 and Bcl-X_L are anti-apoptotic proteins from the Bcl-2 family of apoptosis regulators, which have been frequently observed in solid tumours. Both have been linked to resistance to chemotherapy. The presumed mechanism by which the Bcl-2/Bcl-X_L proteins prevent apoptosis is by inhibiting the function of pro-apoptotic members of the Bcl-2 family such as Bax and Bak. This is shown diagrammatically in Figure 1.4. Bcl-2 or Bcl-X_L binds to the BH3 (Bcl-2-homology 3) domain activating Bcl-2 pro-apoptotic family members. These anti-apoptotic members then associate with active Bak protein (a pro-apoptotic family member) thus inhibiting its action. Inhibitors of the anti-apoptotic members of the Bcl-2 family are therefore expected to restore the function of pro-apoptotic Bcl-2 members causing release of cytochrome C which stimulates apoptosis of the cancerous cell. This is expected to increase susceptibility to chemotherapeutics, which were previously lowered due to over expression of anti-apoptotic Bcl-2 family genes.

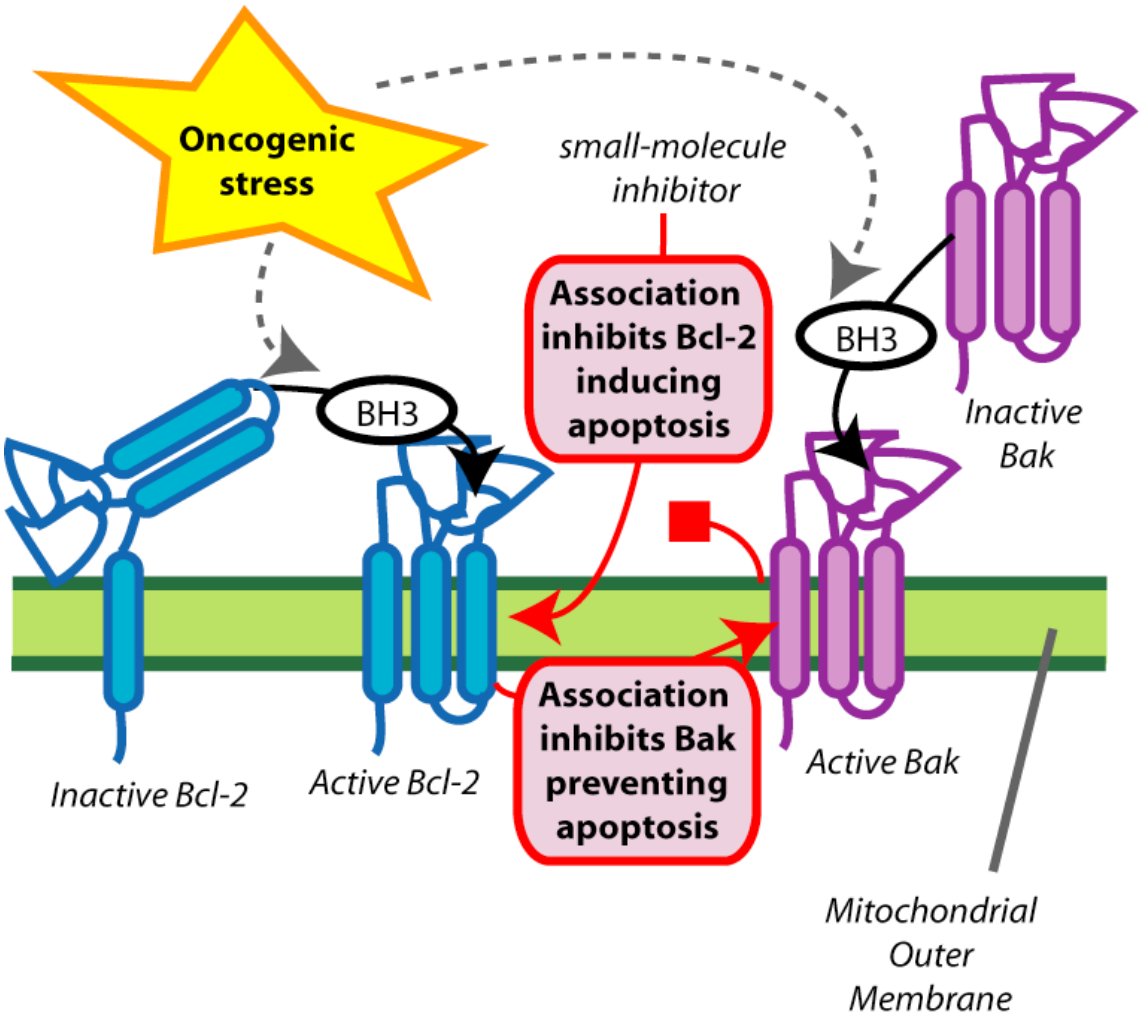

Figure 1.4: Oncogenic stress causes BH3 to bind Bcl-2 and Bak activating these proteins. Association of active Bcl-2 with Bak inhibits Bak thus preventing apoptosis. Inhibition of the Bcl-2/Bak interaction by a small molecule (such as Obatoclax or ABT-737) induces apoptosis. Figure adapted from Dlugosz *et al.*(Dlugosz et al. 2006).

Structures of both Bcl-2 and Bcl-X_L; bound to a peptide fragment of its partner protein Bak; exist. Examination of these structures suggests that the dimensions of a pocket on the protein surface are within limits suitable for efficient binding of a drug-like molecule(Fry and Vassilev 2005). Many groups have reported small-molecule inhibitors of Bcl-2 and Bcl-X_L, with the best affinity being reported having K_d < 1 nM(Oltersdorf *et al.* 2005).

Oltersdorf *et al.* describe the discovery, optimization and in vitro testing of a small-molecule inhibitor (ABT-737) of the Bcl-2 anti-apoptotic protein(Oltersdorf *et al.* 2005). The method employed was high-throughput "SAR by NMR"(Shuker *et al.* 1996) screening of a chemical library resulting in the discovery of 4'-Fluoro-biphenyl-4-carboxylic acid and 5,6,7,8-tetrahydro-napthalen-1-ol with binding affinities of K_d = 0.30 mM and K_d = 4.3 mM respectively(Oltersdorf *et al.* 2005). Both of these compounds bind to distinct adjacent sites on the hydrophobic BH3-binding groove of Bcl-X_L. SAR by NMR technology is based on the idea that once several low affinity binding compounds have been identified they might be linked to achieve high affinity binding. The two lead compounds identified by SAR by NMR were optimised to produce compound 1 with K_i = 36 nM. Compound 1 was severely attenuated in the presence of 1 % Human Serum Albumin (HSA), with its affinity reduced by a factor > 280. Oltersdorf *et al.* then used a structure-based approach to identify functional groups responsible for tight binding to HSA and substituting these groups in order to reduce affinity for HSA whilst retaining Bcl-X_L affinity. The resulting compound was ABT-737 with binding affinity K_i < 1 nM for Bcl-X_L, Bcl-2 and Bcl-w, but K_i = 0.46 μM for the less similar Bcl-B, Mcl-1 and A1 pro-apoptosis proteins.

Oltersdorf *et al.* showed that ABT-737 may be useful as part of a chemotherapy regime. They showed enhanced cytotoxicity of the chemotherapy drug paclitaxel against lung cancer cells when dosed with ABT-737. However, they note that single-agent anti-tumour activity is a more achievable clinical goal(Oltersdorf *et al.* 2005). As such they investigated the activity of ABT-737 singularly as an anti-tumour agent. They observed that ABT-737 exhibited weak activity against many solid tumour cell lines. However, they noted that ABT-737 exhibited potent activity against cell lines representing lymphoid malignancies and small-cell lung carcinoma(Oltersdorf *et al.* 2005).

Obatoclax is another small-molecule that is predicted to occupy a hydrophobic pocket within the BH3 binding groove of Bcl-2. Obatoclax has been shown to interfere with binding of Bak to Mcl-1 (an anti-apoptosis member of the Bcl-2 family), in mitochondrial outer membranes that had been extracted from the cell, and in mitochondrial outer membranes that were present in the cell. Mcl-1 has been shown to confer resistance to ABT-737 and Bortezomib (a proteasome inhibitor), however, Obatoclax has been shown to overcome this resistance(Nguyen *et al.* 2007). To this end, Obatoclax has entered stage II clinical trials as of May 2008.

1.2.3.c hDM2-p53

The murine double mutant oncogene encodes the hDM2 protein, which is a negative regulator of the transcription factor p53. The p53 transcription factor is involved in regulation of the cell cycle, and as such acts as a tumour suppressor. Over-expression of hDM2 has been observed in many human tumours, in these cases hDM2 suppresses the tumour suppressor p53. Inhibition of the hDM2 -p53 interaction can restore the effect of p53, and as such is an attractive target in cancer therapy(Vassilev *et al.* 2004).

The crystal structure of hDM2 bound to a peptide from the transactivation loop of p53 allows analysis of the hDM2 surface, revealing a relatively deep hydrophobic pocket, primarily filled by three side-chains(Kussie *et al.* 1996). Vassilev *et al.* screened a library of chemical compounds for their ability to inhibit the hDM2 -p53 interaction. They identified several lead structures, which were then optimized for selectivity and potency. One particular class of inhibitors identified was a group of *cis*-imidazoles, which they named Nutlins. Nutlins were found to displace p53 from complex with hDM2 with IC_{50} values in the range 100 nM – 300 nM(Vassilev *et al.* 2004). The crystal structure of hDM2 in complex with Nutlin -2 was determined, showing that the inhibitor largely mimics the interactions of the p53 peptide.

1.2.3.d ZipA-FtsZ

Z-interacting protein A (ZipA) is a bacterial protein involved in the formation of cell walls during cell division. Central to its role is recruitment to an organelle called the septal ring at the beginning of cell division. It is recruited by interaction with FtsZ, a protein component of the septal ring(Fry 2006). Blocking the ZipA-FtsZ interaction is seen as an attractive target for the development of novel antibiotics.

The structure of ZipA in complex with a 17-residue peptide from FtsZ has been determined(Y Zhang *et al.* 2000). Several small-molecule inhibitors of the ZipA-FtsZ interaction have been developed, although none have shown high affinity in terms of being a suitable drug candidate. High-throughput screening has identified a pyridyl-pyrimidine derivative to be a 12 μM inhibitor(Rush *et al.* 2005), which is comparable to that shown by the FtsZ 17-mer peptide (7 μM(Kenny *et al.* 2003)). Further investigation of the inhibitor suggested that there were potential toxicity issues with non-specific activity in both bacterial and yeast based assays. Additionally it is noted that pyridyl-pyrimidines are well represented in the literature in the context of their kinase inhibition properties(Furet *et al.* 2000). A second class of compounds with an indolo-quinolizinone core has also been reported to have high micromolar affinity for ZipA. This class of compounds was also developed through lead optimization of a hit determined from high-throughput screening(Jennings *et al.* 2004).

1.2.3.e XIAP-BIR3

X-linked inhibitor of apoptosis (XIAP) is a mediator of programmed cell death, which acts by a series of protein-protein interactions. XIAP binds caspases (a family of calcium dependent cysteine proteases), maintaining them in a catalytically inactive form. The ability to block the interaction of XIAP with caspases thus restoring the catalytically active form has therefore been determined a desirable drug target for oncology due to the potential to restore apoptosis in cancerous cells.

The BIR domain mediates the XIAP interaction with caspases. A naturally occurring competitor protein called SMAC is known to block XIAP-BIR interaction. A nine-mer peptide from SMAC has been shown to bind with 430 nM affinity to BIR(Z Liu *et al.* 2000). The subsequent analysis of the structure of the SMAC peptide bound to BIR showed that only the first four residues make significant contact with BIR(Z Liu *et al.* 2000). The resulting tetra-peptide was analysed and shown to exhibit similar binding affinity (480 nM)(Kipp *et al.* 2002). This tetra-peptide was further optimized to achieve 20 nM binding affinity(Kipp *et al.* 2002). Significant effort has been made to reduce the peptidic character of these inhibitors, whilst retaining the tight binding. One structure has been reported with the BIR domain co-crystallised with one of the most potent inhibitors.

1.2.3.f HPV E2-E1

Human papilloma virus (HPV) is responsible for warts and some cervical cancers. Currently both of these conditions are untreatable with small-molecules. The interaction between the transcription factor E2 and the viral helicase E1 is an essential part of the viral life cycle. As such, it is a possible target for intervention by small-molecule protein-protein interaction inhibitors.

High-throughput screening has been used to identify a class of indandiones that disrupt the HPV E2-E1 interaction with a moderate affinity (K_d = 20 μM). Further optimization of the compound allowed production of compound 23, which exhibits IC_{50} values as low of 6 nM. It has been shown that indandiones bind to the transactivation domain of E2. The X-ray structure of one indandione compound (compound 18) in complex with the transactivation domain E2 shows that one copy of the molecule binds to the three helix domain, whilst a second copy sits on top(Y Wang *et al.* 2004).

The structure of the E2-E1 complex was solved shortly afterwards, showing that the indandione compounds are in contact with far fewer residues than the E2 -E1 contact surface(Abbate, Berger, and Botchan 2004). Compound 18 accesses a cavity that is not observed in the E2-E1 protein-protein interface. Compound 23 binds with much higher ligand efficiency than the protein-protein interaction, possibly because it manages to bury its hydrophobic surface deeper than the protein-protein interaction that is spread across the E2 protein surface(Wells and McClendon, 2007).

Target	Discovery Technique	Maximal affinity	Success
IL-2/IL-2Rα	HTS - optimize	60 nM	No drug.
Bcl-2/Bak	SAR by NMR	< 1 nM	Obatoclax, ABT-737 and more in phase II clinical trials.
Bcl-X_L/Bak	SAR by NMR	< 1 nM	Obatoclax, ABT-737 and more in phase II clinical trials.
hDM2-p53	-	70 nM	No drug.
ZipA-FtsZ	Peptide	12 μM	No drug.
XIAP-BIR3	Peptide	20 nM	No drug.
HPV E2-E1	HTS	6 nM	No drug.

Table 1.1: Summary of target, discovery method, maximal affinity, and success of prospective drug.

1.2.3.g Interactions not structurally characterized

One of the major disadvantages of the structure based approach suggested by Chène is that the high-resolution protein structure is required. This excludes a relatively large number of protein-protein interactions that have been identified as targets. Since there has been little or no structural information published regarding

these interactions, one should bear in mind that they may not fit the criteria that we have specified, however, there is a reasonable chance that they will be interesting to investigate as structural information becomes available.

One such example of an interaction of interest is involved in the entry of HIV-1 into host cells. HIV-1 entry occurs in a multistep process. It involves binding of the HIV-1 envelope protein (gp120) to the host cell CD4 receptor. This is followed by interaction of chemokine receptors (CCR5 or CXCR4), which causes rearrangement of the envelope transmembrane subunit, allowing membrane fusion(Tsibris and DR Kuritzkes 2007). A small-molecule drug (Maraviroc) has been developed, which blocks entry of HIV-1 into cells, by inhibiting the gp120-CCR5 interaction(D Kuritzkes, Kar, and Kirkpatrick 2008).

Another target of significant interest is the inhibition of Herpes Simplex Virus (HSV) ribonuclease reductase dimerization. HSV is responsible for a large number of diseases, including genital herpes. Current drugs are nucleoside analogues, which are phosphorylated thus producing inhibitors of DNA polymerase. The inhibition of HSV ribonuclease reducatase dimerization has been validated as a drug target. Several peptide inhibitors derived from the C-terminus of HSV R2 have been further optimized to produce very effective inhibitors of dimerization(Chène 2006). However, there has currently been no success in developing small-molecule inhibitors.

The transcription factor c-Myc is estimated to be involved in one in seven human cancer deaths(Chène 2006),(Dang 1999). The oncogenicity of c-Myc relies on its association with its activation partner Max. Inhibitors of the c-Myc/Max interaction therefore have therapeutic potential. Both c-Myc and Max belong to the basic helix-loop-helix leucine zipper family (bHLH-LZ) family. A structure of the Max/Max homodimer exists, and the structure of the c-Myc/Max heterodimer has been

speculated on from this data. Chemical libraries have been screened in an effort to identify c-Myc/Max dimerization inhibitors. This resulted in the discovery of four small-molecule inhibitors(Chène 2006).

1.2.4 Summary

It is clear that whilst drug discovery at the protein-protein interface is a highly attractive proposition in terms of the number of new therapeutic possibilities. It is also a very difficult problem to address. However, good progress is being made in terms of elucidating small-molecules capable of disrupting protein-protein interfaces and in some cases in optimizing them for the market. Notably there are several Bcl-2/Bak and Bcl-X_L/Bak inhibitors that are currently in phase II clinical trials.

Current success in the field of small-molecule protein-protein interaction inhibitor drug discovery, seems to be mainly derived from the use of high-throughput screening techniques, peptide binding epitopes, and NMR studies.

Current attempts at elucidating PPI inhibitors have often relied on high-throughput screening methods to identify initial lead compounds. Subsequent analysis of the strategies employed shows that the small molecule in question often mimics the binding epitope of its natural partner protein. The observations that we have made may allow identification of regions on the protein surface that are likely to accommodate small molecule inhibitors. It is hoped this will help in understanding the likelihood of success for a given protein target, as well as in many cases acting as the basis for the rational targeting of specific pockets.

1.2.5 Outline of thesis aims

It is clear that protein-protein interactions are a highly desirable therapeutic target. Whilst previous efforts have shown some successes in targeting these interactions, it is clear that the methodology for doing so can be improved. The five results chapters contained in this thesis broadly split into two sections. The first two results chapters deal with databases of protein-protein and protein-ligand interactions and how one might prioritise drug discovery methods on particular proteins from these databases. The final three results chapters are a case study on techniques to aid design of novel inhibitors of the hDM2-p53 protein-protein interaction. The overall aim of these three chapters is to perform accurate alchemical free energy calculations that can distinguish high-affinity compounds from low-affinity compounds.

The first results chapter details the use of computational methods for pocket detection using the Q-SiteFinder software to study protein-protein interactions that have previously been targeted and have been structurally characterised by either X-ray crystallography or NMR experiments. The aim of this area of study is to identify key properties of protein-protein interactions that are amenable to inhibition and compare and contrast these properties to those of currently marketed drugs and to protein-protein interactions that have not yet been targeted for inhibition by small molecule compounds. Identification of properties that might distinguish pockets present on some protein-protein interfaces as suitable for inhibition has parallels to the concept of 'drugability' of a protein. The second results chapter attempts to create a structure-based method that can distinguish a pocket likely to be able to bind a small-molecule with high-affinity for a pocket that is unlikely to be able to do so.

The third results chapter is the first of the chapters that aim to perform alchemical free energy calculations. This chapter aims to use docking methods to generate conformations of oligoamide compounds bound to hDM2. Once these conformations have been generated torsional potentials for the oligoamide compounds are identified from the literature, and charge parameters are calculated using Hartree-Fock calculations and AM1 BCC charge calculations. The results from the two charge calculation methods are evaluated and the best used in further calculations. The fourth results chapter aims to perform molecular dynamics simulations of the hDM2-p53 system and current known inhibitors. Further simulations of the docked hDM2-oligoamide complexes can then be performed. These simulations can be further analysed to determine the length of alchemical free energy calculation required to give converged estimates of the free energy. The fifth and final results chapter aims to generate docked conformations for six oligoamide compounds which have been used in a previous study by Plante *et al.*(Plante et al. 2009). Relative free energy calculations for these compounds can then be performed using a three process and single process technique. Furthermore a Hamiltonian replica-exchange technique can also be applied.

1.3 References

Abbate, Eric A, James M Berger, and Michael R Botchan. 2004. The X-ray structure of the papillomavirus helicase in complex with its molecular matchmaker E2. *Genes and Development* 18, no. 16: 1981-96. doi:10.1101/gad.1220104. http://www.ncbi.nlm.nih.gov/pubmed/15289463.

Agrafiotis, Dimitris K, Alan C Gibbs, F Zhu, Sergei Izrailev, and Eric Martin. 2007. Conformational sampling of bioactive molecules: a comparative study. *Journal of Chemical Information and Modeling* 47, no. 3: 1067-86. doi:10.1021/ci6005454. http://www.ncbi.nlm.nih.gov/pubmed/17411028.

Allen, Frank H. 2002. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallographica Section B Structural Science* 58, no. 3: 380-388. doi:10.1107/S0108768102003890. http://scripts.iucr.org/cgi-bin/paper?S0108768102003890.

Arkin, Michelle R, and JA Wells. 2004. Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nature Reviews Drug Discovery* 3, no. 4 (April): 301-17. doi:10.1038/nrd1343. http://www.ncbi.nlm.nih.gov/pubmed/15060526.

Bennett, C. 1976. Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics* 22, no. 2: 245-268. doi:10.1016/0021-9991(76)90078-4. http://linkinghub.elsevier.com/retrieve/pii/0021999176900784.

Berg, Thorsten. 2003. Modulation of protein-protein interactions with small organic molecules. *Angewandte Chemie (International ed. in English)* 42, no. 22: 2462-81. doi:10.1002/anie.200200558. http://www.ncbi.nlm.nih.gov/pubmed/12800163.

Bogan, A A, and K S Thorn. 1998. Anatomy of hot spots in protein interfaces. *Journal of Molecular Biology* 280, no. 1: 1-9. doi:10.1006/jmbi.1998.1843. http://www.ncbi.nlm.nih.gov/pubmed/9653027.

Boström, J. 2001. Reproducing the conformations of protein-bound ligands: a critical evaluation of several popular conformational searching tools. *Journal of Computer-aided Molecular Design* 15, no. 12 (December): 1137-52. http://www.ncbi.nlm.nih.gov/pubmed/12160095.

Boström, Jonas, Jeremy R Greenwood, and Johan Gottfries. 2003. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *Journal of Molecular Graphics and Modelling* 21, no. 5 (March): 449-62. http://www.ncbi.nlm.nih.gov/pubmed/12543140.

Bowers, Kevin J., Federico D. Sacerdoti, John K. Salmon, Yibing Shan, David E. Shaw, Edmond Chow, Huafeng Xu, *et al.* 2006. *Molecular dynamics---Scalable algorithms for molecular dynamics simulations on commodity clusters*. *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing - SC '06*. New York, New York, USA: ACM Press. doi:10.1145/1188455.1188544. http://portal.acm.org/citation.cfm?doid=1188455.1188544.

Burgoyne, Nicholas John. 2007. The Structural Analysis and Prediction of Protein interactions. PhD Thesis, Institute of *Molecular and Cellular Biology, University of Leeds, UK*.

Chipot, Christophe, and Andrew Pohorille. 2007. *Free energy calculations: theory and applications in chemistry and biology*. Ed. Christophe Chipot and Andrew Pohorille. Springer. http://books.google.com/books?id=XTE5eejRTC0C&printsec=frontcover&dq=free+energy+calculations&hl=en&ei=Y_BGTPOOIJGUjAfX-Oj0Bg&sa=X&oi=book_result&ct=result&resnum=1&ved=0CDMQ6AEwAA#v=onepage&q&f=false.

Chène, Patrick. 2006. Drugs targeting protein-protein interactions. *ChemMedChem* 1, no. 4 (April): 400-11. doi:10.1002/cmdc.200600004. http://www.ncbi.nlm.nih.gov/pubmed/16892375.

Clackson, T, and J A Wells. 1995. A hot spot of binding energy in a hormone-receptor interface. *Science (New York, N.Y.)* 267, no. 5196 (January): 383-6. http://www.ncbi.nlm.nih.gov/pubmed/7529940.

Cornell, Wendy D., Piotr Cieplak, Christopher I Bayly, and Peter A. Kollmann. 1993. Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation. *Journal of the American Chemical Society* 115, no. 21: 9620-9631. doi:10.1021/ja00074a030. http://pubs.acs.org/doi/abs/10.1021/ja00074a030.

Cossins, Benjamin P, Sebastien Foucher, Colin M Edge, and Jonathan W Essex. 2009. Assessment of nonequilibrium free energy methods. *The Journal of Physical Chemistry B* 113, no. 16 (April): 5508-19. doi:10.1021/jp803532z. http://www.ncbi.nlm.nih.gov/pubmed/19368411.

Van Der Spoel, D., Lindahl, E., Hess, B., Van Buuren, A. R., Apol, E., Meulenhoff, P. J., Van Drunen, R., Tieleman, H. J. C. D. P., Sijbers, A. L. T. M., Feenstra, K. A., Berendsen, H.. 2005. Gromacs User Manual version 4.0. http://www.gromacs.org.

Dang, C V. 1999. c-Myc target genes involved in cell growth, apoptosis, and metabolism. *Molecular and Cellular Biology* 19, no. 1: 1-11. http://www.ncbi.nlm.nih.gov/pubmed/9858526.

Diller, David J, and Kenneth M Merz. 2002. Can we separate active from inactive conformations? *Journal of Computer-aided Molecular Design* 16, no. 2 (February): 105-12. http://www.ncbi.nlm.nih.gov/pubmed/12188020.

Dlugosz, Paulina J, Lieven P Billen, Matthew G Annis, W Zhu, Z Zhang, Jialing Lin, Brian Leber, and David W Andrews. 2006. Bcl-2 changes conformation to inhibit Bax oligomerization. *The EMBO Journal* 25, no. 11: 2287-96. doi:10.1038/sj.emboj.7601126. http://www.ncbi.nlm.nih.gov/pubmed/16642033.

Ewing, T J, S Makino, a G Skillman, and I D Kuntz. 2001. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *Journal of Computer-aided Molecular Design* 15, no. 5 (May): 411-28. http://www.ncbi.nlm.nih.gov/pubmed/11394736.

Ferrenberg, Alan, and Robert Swendsen. 1989. Optimized Monte Carlo data analysis. *Physical Review Letters* 63, no. 12 (September): 1195-1198. doi:10.1103/PhysRevLett.63.1195. http://link.aps.org/doi/10.1103/PhysRevLett.63.1195.

Freire, Ernesto. 2008. Do enthalpy and entropy distinguish first in class from best in class? *Drug Discovery Today* 13, no. 19-20 (October): 869-74. doi:10.1016/j.drudis.2008.07.005. http://www.ncbi.nlm.nih.gov/pubmed/18703160.

Frenkel, Daan, and Berend Smit. 2002. *Understanding molecular simulation: from algorithms to applications*. 2nd ed. Academic Press. http://books.google.co.uk/books?id=XmyO2oRUg0cC&dq=frenkel+&+smit&source=gbs_navlinks_s.

Fry, David C. 2006. Protein-protein interactions as targets for small molecule drug discovery. *Biopolymers* 84, no. 6 (January): 535-52. doi:10.1002/bip.20608. http://www.ncbi.nlm.nih.gov/pubmed/17009316.

Fry, David C, and Lyubomir T Vassilev. 2005. Targeting protein-protein interactions for cancer therapy. *Journal of Molecular Medicine (Berlin, Germany)* 83, no. 12: 955-63. doi:10.1007/s00109-005-0705-x. http://www.ncbi.nlm.nih.gov/pubmed/16283145.

Fuller, Jonathan C., Nicholas J. Burgoyne, and Richard M. Jackson. 2009. Predicting druggable binding sites at the protein-protein interface. *Drug Discovery Today* 14, no. 3-4 (February): 155-61. doi:10.1016/j.drudis.2008.10.009. http://www.ncbi.nlm.nih.gov/pubmed/19041415.

Furet, P, J Zimmermann, H G Capraro, T Meyer, and P Imbach. 2000. Structure-based design of potent CDK1 inhibitors derived from olomoucine. *Journal of Computer-aided Molecular Design* 14, no. 5 (July): 403-9. http://www.ncbi.nlm.nih.gov/pubmed/10896313.

Gadek, Thomas R, and John B Nicholas. 2003. Small molecule antagonists of proteins. *Biochemical Pharmacology* 65, no. 1 (January): 1-8. http://www.ncbi.nlm.nih.gov/pubmed/12473372.

Geer, LY, Aron Marchler-Bauer, RC Geer, Lianyi Han, J He, S He, C Liu, W Shi, and Stephen H Bryant. 2010. The NCBI BioSystems database. *Nucleic Acids Research* 38, no. Database issue: D492-6. doi:10.1093/nar/gkp858. http://www.ncbi.nlm.nih.gov/pubmed/19854944.

Good, A C, and D L Cheney. 2003. Analysis and optimization of structure-based virtual screening protocols (1): exploration of ligand conformational sampling techniques. *Journal of Molecular Graphics and Modelling* 22, no. 1: 23-30. doi:10.1016/S1093-3263(03)00123-2. http://linkinghub.elsevier.com/retrieve/pii/S1093326303001232.

Goodford, P J. 1985. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of Medicinal Chemistry* 28, no. 7: 849-57. http://www.ncbi.nlm.nih.gov/pubmed/3892003.

Gottesman, Michael M. 2002. Mechanisms of cancer drug resistance. *Annual Review of Medicine* 53: 615-27. doi:10.1146/annurev.med.53.082901.103929. http://www.ncbi.nlm.nih.gov/pubmed/11818492.

Halgren, Thomas A. 1996. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry* 17, no. 5-6: 490-519. doi:10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P. http://doi.wiley.com/10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P.

Halperin, Inbal, Buyong Ma, Haim Wolfson, and Ruth Nussinov. 2002. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* 47, no. 4 (June): 409-43. doi:10.1002/prot.10115. http://www.ncbi.nlm.nih.gov/pubmed/12001221.

Hehre, W. J. 1969. Self-Consistent Molecular-Orbital Methods. I. Use of Gaussian Expansions of Slater-Type Atomic Orbitals. *The Journal of Chemical Physics* 51, no. 6: 2657. doi:10.1063/1.1672392. http://link.aip.org/link/?JCP/51/2657/1&Agg=doi.

Hess, Berk, Carsten Kutzner, David Van Der Spoel, and Erik Lindahl. 2008. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation* 4, no. 3 (March): 435-447. doi:10.1021/ct700301q. http://pubs.acs.org/doi/abs/10.1021/ct700301q.

Higueruelo, Alícia P, Adrian Schreyer, G Richard J Bickerton, Will R Pitt, Colin R Groom, and Tom L Blundell. 2009. Atomic interactions and profile of small molecules disrupting protein-protein interfaces: the TIMBAL database. *Chemical Biology and Drug Design* 74, no. 5 (November): 457-67. doi:10.1111/j.1747-0285.2009.00889.x. http://www.ncbi.nlm.nih.gov/pubmed/19811506.

Huey, Ruth, Garrett M. Morris, Arthur J. Olson, and David S. Goodsell. 2007. A semiempirical free energy force field with charge-based desolvation. *Journal of Computational Chemistry* 28, no. 6 (April): 1145-52. doi:10.1002/jcc.20634. http://www.ncbi.nlm.nih.gov/pubmed/17274016.

Irwin, John J, and Brian K Shoichet. 2005. ZINC--a free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling* 45, no. 1: 177-82. doi:10.1021/ci049714+. http://www.ncbi.nlm.nih.gov/pubmed/15667143.

Itzstein, M von, WY Wu, G B Kok, M S Pegg, J C Dyason, B Jin, T Van Phan, M L Smythe, HF White, and S W Oliver. 1993. Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature* 363, no. 6428 (June): 418-23. doi:10.1038/363418a0. http://www.ncbi.nlm.nih.gov/pubmed/8502295.

Jakalian, Araz, David B Jack, and Christopher I Bayly. 2002. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *Journal of Computational Chemistry* 23, no. 16: 1623-41. doi:10.1002/jcc.10128. http://www.ncbi.nlm.nih.gov/pubmed/12395429.

Jennings, Lee D, Ken W Foreman, Thomas S Rush, Desiree H H Tsao, Lidia Mosyak, Y Li, Mohani N Sukhdeo, *et al.* 2004. Design and synthesis of indolo[2,3-a]quinolizin-7-one inhibitors of the ZipA-FtsZ interaction. *Bioorganic and Medicinal Chemistry Letters* 14, no. 6: 1427-31. doi:10.1016/j.bmcl.2004.01.028. http://www.ncbi.nlm.nih.gov/pubmed/15006376.

Kenny, Cynthia Hess, Weidong Ding, Kerry Kelleher, Susan Benard, Elizabeth Glasfeld Dushin, Alan G Sutherland, Lidia Mosyak, Ronald Kriz, and George Ellestad. 2003. Development of a fluorescence polarization assay to screen for inhibitors of the FtsZ/ZipA interaction. *Analytical Biochemistry* 323, no. 2: 224-33. doi:10.1016/j.ab.2003.08.033. http://www.ncbi.nlm.nih.gov/pubmed/14656529.

Khandogin, Jana, and Charles L Brooks. 2005. Constant pH molecular dynamics with proton tautomerism. *Biophysical Journal* 89, no. 1 (July): 141-57. doi:10.1529/biophysj.105.061341. http://www.ncbi.nlm.nih.gov/pubmed/15863480.

Kipp, Rachael a, M a Case, Aislyn D Wist, Catherine M Cresson, Maria Carrell, Erin Griner, Arun Wiita, *et al.* 2002. Molecular targeting of inhibitor of apoptosis proteins based on small molecule mimics of natural binding partners. *Biochemistry* 41, no. 23 (June): 7344-9. http://www.ncbi.nlm.nih.gov/pubmed/12044166.

Kirchmair, Johannes, Christian Laggner, Gerhard Wolber, and Thierry Langer. 2005. Comparative analysis of protein-bound ligand conformations with respect to catalyst's conformational space subsampling algorithms. *Journal of Chemical Information and Modeling* 45, no. 2: 422-30. doi:10.1021/ci049753l. http://www.ncbi.nlm.nih.gov/pubmed/15807508.

Kristam, Rajendra, Valerie J Gillet, Richard a Lewis, and David Thorner. 2005. Comparison of conformational analysis techniques to generate pharmacophore hypotheses using catalyst. *Journal of Chemical Information and Modeling* 45, no. 2: 461-76. doi:10.1021/ci049731z. http://www.ncbi.nlm.nih.gov/pubmed/15807512.

Kuritzkes, D, Santwana Kar, and Peter Kirkpatrick. 2008. Maraviroc. *Nature Reviews Drug Discovery* 7, no. 1: 15-16. doi:10.1038/nrd2490. http://www.nature.com/doifinder/10.1038/nrd2490.

Kussie, P. H., S. Gorina, V. Marechal, B. Elenbaas, J. Moreau, A. J. Levine, and N. P. Pavletich. 1996. Structure of the MDM2 Oncoprotein Bound to the p53 Tumor Suppressor Transactivation Domain. *Science* 274, no. 5289 (November): 948-953. doi:10.1126/science.274.5289.948. http://www.sciencemag.org/cgi/doi/10.1126/science.274.5289.948.

Laurie, Alasdair T R, and Jackson, R.M.. 2005. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics (Oxford, England)* 21, no. 9 (February): 1908-16. doi:10.1093/bioinformatics/bti315. http://www.ncbi.nlm.nih.gov/pubmed/15701681.

Laurie, Alasdair T R, and Jackson, R.M.. 2006. Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening. *Current Protein and Peptide Science* 7, no. 5 (October): 395-406. http://www.ncbi.nlm.nih.gov/pubmed/17073692.

Leach, Andrew R. 2001. *Molecular Modelling: Principles and Applications*. 2nd ed. Sharlow, England; New York: Prentice Hall. http://books.google.co.uk/books?hl=en&lr=&id=kB7jsbV-uhkCandoi=fnd&pg=PR11&dq=molecular+modelling+principles+and+applications+2nd+edition&ots=-XwwbVloCS&sig=6uzl6ndaBcaSVO6zlPuT-b0nBAM#v=onepage&q=molecular modelling principles and applications 2nd edition&f=false.

Letai, A. 2005. The BCL-2 network: Mechanistic insights and therapeutic potential. *Drug Discovery Today: Disease Mechanisms* 2, no. 2: 145-151. doi:10.1016/j.ddmec.2005.05.004. http://linkinghub.elsevier.com/retrieve/pii/S1740676505000118.

Lindahl, Erik, Hess, Berk, and van der Spoel, David. 2001. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *Journal of Molecular Modeling*: 306-317. doi:10.1007/s008940100045.

Liu, DC, and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming* 45, no. 1-3 (August): 503-528. doi:10.1007/BF01589116. http://www.springerlink.com/index/10.1007/BF01589116.

Liu, Z, C Sun, E T Olejniczak, R P Meadows, S F Betz, T Oost, J Herrmann, JC Wu, and S W Fesik. 2000. Structural basis for binding of Smac/DIABLO to the XIAP BIR3 domain. *Nature* 408, no. 6815: 1004-8. doi:10.1038/35050006. http://www.ncbi.nlm.nih.gov/pubmed/11140637.

Martyna, Glenn J., Douglas J. Tobias, and ML Klein. 1994. Constant pressure molecular dynamics algorithms. *The Journal of Chemical Physics* 101, no. 5 (April): 4177. doi:10.1063/1.467468. http://link.aip.org/link/JCPSA6/v101/i5/p4177/s1&Agg=doi.

McGann, Mark R, Harold R Almond, Anthony Nicholls, J Andrew Grant, and Frank K Brown. 2003. Gaussian docking functions. *Biopolymers* 68, no. 1 (January): 76-90. doi:10.1002/bip.10207. http://www.ncbi.nlm.nih.gov/pubmed/12579581.

McGaughey, Georgia B, Robert P Sheridan, Christopher I Bayly, J Chris Culberson, Constantine Kreatsoulas, Stacey Lindsley, Vladimir Maiorov, Jean-Francois Truchon, and Wendy D Cornell. 2007. Comparison of topological, shape, and docking methods in virtual screening. *Journal of Chemical Information and Modeling* 47, no. 4: 1504-19. doi:10.1021/ci700052x. http://www.ncbi.nlm.nih.gov/pubmed/17591764.

Michel, Julien, Julian Tirado-Rives, and William L Jorgensen. 2009. Prediction of the water content in protein binding sites. *The Journal of Physical Chemistry B* 113, no. 40 (October): 13337-46. doi:10.1021/jp9047456. http://www.ncbi.nlm.nih.gov/pubmed/19754086.

Morris, Garrett M., David S. Goodsell, Robert S. Halliday, Ruth Huey, William E. Hart, Richard K. Belew, and Arthur J. Olson. 1998. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry* 19, no. 14 (November): 1639-1662. doi:10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B. http://doi.wiley.com/10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B.

Nelder, Ja, and R Mead. 1965. A simplex method for function minimization. *The computer journal*. http://comjnl.oxfordjournals.org/cgi/content/abstract/7/4/308.

Nguyen, Mai, Richard C Marcellus, Anne Roulston, Mark Watson, Lucile Serfass, S R Murthy Madiraju, Daniel Goulet, *et al.* 2007. Small molecule obatoclax (GX15-070) antagonizes MCL-1 and overcomes MCL-1-mediated resistance to apoptosis. *Proceedings of the National Academy of Sciences of the United States of America* 104, no. 49 (December): 19512-7. doi:10.1073/pnas.0709443104. http://www.ncbi.nlm.nih.gov/pubmed/18040043.

Nicholls, Anthony, NE MacCuish, and JD MacCuish. 2004. Variable selection and model validation of 2D and 3D molecular descriptors>. *Journal of Computer-Aided Molecular Design* 18, no. 7-9: 451-474. doi:10.1007/s10822-004-5202-8. http://www.springerlink.com/index/10.1007/s10822-004-5202-8.

Oltersdorf, Tilman, Steven W Elmore, Alexander R Shoemaker, Robert C Armstrong, David J Augeri, Barbara A Belli, Milan Bruncko, *et al.* 2005. An inhibitor of Bcl-2 family proteins induces regression of solid tumours. *Nature* 435, no. 7042 (June): 677-81. doi:10.1038/nature03579. http://www.ncbi.nlm.nih.gov/pubmed/15902208.

Park, Hwangseo, J Lee, and S Lee. 2006. Critical assessment of the automated AutoDock as a new docking tool for virtual screening. *Proteins* 65, no. 3: 549-54. doi:10.1002/prot.21183. http://www.ncbi.nlm.nih.gov/pubmed/16988956.

Plante, Jeffrey P., Thomas Burnley, Barbora Malkova, Michael E. Webb, Stuart L. Warriner, Thomas A. Edwards, and Andrew J. Wilson. 2009. Oligobenzamide proteomimetic inhibitors of the p53–hDM2 protein–protein interaction. *Chemical Communications*, no. 34: 5091. doi:10.1039/b908207g. http://xlink.rsc.org/?DOI=b908207g.

Pophristic, Vojislava, Satyavani Vemparala, Ivaylo Ivanov, Zhiwei Liu, ML Klein, and William F DeGrado. 2006. Controlling the shape and flexibility of arylamides: a combined ab initio, ab initio molecular dynamics, and classical molecular dynamics study. *The Journal of Physical Chemistry B* 110, no. 8 (March): 3517-26. doi:10.1021/jp054306+. http://www.ncbi.nlm.nih.gov/pubmed/16494407.

Rarey, M, B Kramer, T Lengauer, and G Klebe. 1996. A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology* 261, no. 3 (August): 470-89. doi:10.1006/jmbi.1996.0477. http://www.ncbi.nlm.nih.gov/pubmed/8780787.

Rhodes, Daniel R, Scott A Tomlins, Sooryanarayana Varambally, Vasudeva Mahavisno, Terrence Barrette, Shanker Kalyana-Sundaram, Debashis Ghosh, Akhilesh Pandey, and Arul M Chinnaiyan. 2005. Probabilistic model of the human protein-protein interaction network. *Nature Biotechnology* 23, no. 8: 951-9. doi:10.1038/nbt1103. http://www.ncbi.nlm.nih.gov/pubmed/16082366.

Rush, Thomas S, J Andrew Grant, Lidia Mosyak, and Anthony Nicholls. 2005. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *Journal of Medicinal Chemistry* 48, no. 5 (March): 1489-95. doi:10.1021/jm040163o. http://www.ncbi.nlm.nih.gov/pubmed/15743191.

Shirts, Michael R., Eric Bair, Giles Hooker, and Vijay S. Pande. 2003. Equilibrium Free Energies from Nonequilibrium Measurements Using Maximum-Likelihood Methods. *Physical Review Letters* 91, no. 14 (October): 1-4. doi:10.1103/PhysRevLett.91.140601. http://link.aps.org/doi/10.1103/PhysRevLett.91.140601.

Shirts, Michael R., and John D. Chodera. 2008. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of Chemical Physics* 129, no. 12: 124105. doi:10.1063/1.2978177. http://www.ncbi.nlm.nih.gov/pubmed/19045004.

Shuker, S B, P J Hajduk, R P Meadows, and S W Fesik. 1996. Discovering high-affinity ligands for proteins: SAR by NMR. *Science (New York, N.Y.)* 274, no. 5292 (November): 1531-4. http://www.ncbi.nlm.nih.gov/pubmed/8929414.

Singh, U. Chandra, and P A Kollman. 1984. An approach to computing electrostatic charges for molecules. *Journal of Computational Chemistry* 5, no. 2 (April): 129-145. doi:10.1002/jcc.540050204. http://doi.wiley.com/10.1002/jcc.540050204.

Solis, F. J., and R. J.-B. Wets. 1981. Minimization by Random Search Techniques. *Mathematics of Operations Research* 6, no. 1 (February): 19-30. doi:10.1287/moor.6.1.19. http://mor.journal.informs.org/cgi/doi/10.1287/moor.6.1.19.

Thanos, Christopher D, Warren L DeLano, and JA Wells. 2006. Hot-spot mimicry of a cytokine receptor by a small molecule. *Proceedings of the National Academy of Sciences of the United States of America* 103, no. 42 (October): 15422-7. doi:10.1073/pnas.0607058103. http://www.ncbi.nlm.nih.gov/pubmed/17032757.

Tsibris, Athe M N, and DR Kuritzkes. 2007. Chemokine antagonists as therapeutics: focus on HIV-1. *Annual Review of Medicine* 58: 445-59. doi:10.1146/annurev.med.58.080105.102908. http://www.ncbi.nlm.nih.gov/pubmed/16958560.

Vassilev, Lyubomir T, Binh T Vu, Bradford Graves, Daisy Carvajal, Frank Podlaski, Zoran Filipovic, Norman Kong, *et al.* 2004. In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science (New York, N.Y.)* 303, no. 5659 (February): 844-8. doi:10.1126/science.1092472. http://www.ncbi.nlm.nih.gov/pubmed/14704432.

Waldmann, Thomas a. 2003. Immunotherapy: past, present and future. *Nature Medicine* 9, no. 3 (March): 269-77. doi:10.1038/nm0303-269. http://www.ncbi.nlm.nih.gov/pubmed/18179815.

Wang, J, Romain M Wolf, James W Caldwell, PA Kollman, and DA Case. 2004. Development and testing of a general amber force field. *Journal of Computational Chemistry* 25, no. 9 (July): 1157-74. doi:10.1002/jcc.20035. http://www.ncbi.nlm.nih.gov/pubmed/15116359.

Wang, Y, René Coulombe, Dale R Cameron, Louise Thauvette, Marie-Josée Massariol, Lynn M Amon, Dominique Fink, *et al.* 2004. Crystal structure of the E2 transactivation domain of human papillomavirus type 11 bound to a protein interaction inhibitor. *The Journal of Biological Chemistry* 279, no. 8: 6976-85. doi:10.1074/jbc.M311376200. http://www.ncbi.nlm.nih.gov/pubmed/14634007.

Wells, JA, and Christopher L McClendon. 2007. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature* 450, no. 7172: 1001-9. doi:10.1038/nature06526. http://www.ncbi.nlm.nih.gov/pubmed/18075579.

Whitty, Adrian, and Gnanasambandam Kumaravel. 2006. Between a rock and a hard place? *Nature Chemical Biology* 2, no. 3: 112-8. doi:10.1038/nchembio0306-112. http://www.ncbi.nlm.nih.gov/pubmed/16484997.

Zhang, Y, L Mosyak, E Glasfeld, S Haney, M Stahl, J Seehra, and W S Somers. 2000. The bacterial cell-division protein ZipA and its interaction with an FtsZ fragment revealed by X-ray crystallography. *The EMBO Journal* 19, no. 13: 3179-91. doi:10.1093/emboj/19.13.3179. http://www.ncbi.nlm.nih.gov/pubmed/10880432.

Zwanzig, Robert W. 1954. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *The Journal of Chemical Physics* 22, no. 8: 1420. doi:10.1063/1.1740409. http://link.aip.org/link/JCPSA6/v22/i8/p1420/s1&Agg=doi.

2 Properties of small molecule protein-protein interaction inhibitors and their active volumes

2.1 Abstract

The ability to identify and inhibit key protein-protein interactions (PPIs) that are involved in disease pathways is of clear therapeutic benefit. Furthermore, design and application of selective inhibitors of protein-protein interactions is of interest to the field of chemical genetics, whereby an interaction could be 'switched on/off' by application of an inhibitor, negating problems with production of knockout mutants. Here we discuss the application of the pocket detection method Q-SiteFinder to large datasets of structures of protein-ligand interactions (PLIs) and protein-protein interactions. Key differences are observed, specifically that protein-ligand interactions tend to occur in one large pocket (average volume 260 $Å^3$), whilst protein-protein interactions tend to occur in a greater number (2-8) of smaller pockets (average volume 54 $Å^3$). We then extend this to a comparison between proteins bound to marketed drugs and putative inhibitors of protein-protein interactions. The pockets observed on marketed drugs (average volume 271 $Å^3$) follow the same trend as those observed in protein-ligand interactions, whilst protein-protein interaction inhibitors tend to bind multiple (3-5) pockets that lie somewhere between that of protein-ligand/marketed drug interactions and protein-protein interactions (average volume 100 $Å^3$). Identification of these pockets and their properties may enable better screening and help in the design of potent selective inhibitors of protein-protein interactions.

2.2 Introduction

When Lipinski *et al.* published their seminal work detailing his Rule of Five he helped to guide drug discovery efforts towards compounds that were more likely to be efficacious. Lipinski's work pertained to the properties of ligands. Hopkins and Groom asked questions about the druggable genome soon after the publication of the human genome(Hopkins and Groom 2002). Several researchers have published asking similar questions making slightly different assumptions as new data becomes available on which classes of proteins have been deemed druggable(Overington, Al-Lazikani, and Hopkins 2006),(Keller, Pichota, and Yin 2006),(Billingsley 2008).

Here we detail a study that investigates the properties of protein binding sites that can bind ligands. Broadly speaking we are investigating the 'drugability' of the binding sites involved. There are a wide range of definitions (and two spellings!) of 'drugability/druggability'(Henrich *et al.* 2010). For the duration of this work we will use the definition of Egner and Hillig, who define drugability as the “likelihood of finding a selective, low-molecular weight molecule that binds with high affinity to the target.”(Egner and Hillig 2008). This definition is one of the more fundamental definitions in terms of the protein-binding site since it does not distinguish between systems whereby it may be possible to identify compounds that could bind to a site but for which it is difficult to identify compounds that bind selectively and those compounds that will bind and it may be possible to identify selective compounds. Furthermore the definition does not take into account ADMET properties which generally require different analysis to standard structural bioinformatics techniques. The concept of drugability with respect to a definition similar to that of Egner and Hillig is discussed by Hajduk *et al.*(Hajduk, Huth, and Tse 2005).

Recently researchers have begun to ask questions about the properties of protein binding sites. One of the most comprehensive studies of drugability of binding pockets uses both experimental NMR data to identify NMR hit rates to determine drugability, followed by analysis of the properties of binding pockets using structural bioinformatics techniques(Hajduk, Huth, and Fesik 2005). Broadly speaking Hajduk and co-workers use NMR hit-rate to classify protein binding pockets into one of three classes: druggable (hit-rate > 0.3 %); difficult (0.1 % < hit-rate < 0.3 %); undruggable (hit-rate < 0.1 %). Furthermore they investigated several properties of pockets such as Volume, Surface Area, Compactness (ratio of volume to surface area), apolar surface area, polar surface area, contact area, apolar contact area, polar contact area, roughness, total number of charged residues and principal component analysis to develop a drugability index.

2.2.1 Q-SiteFinder and binding sites

Q-SiteFinder was initially developed as a method to identify binding sites from protein structural information. The Q-SiteFinder program was able to identify the correct binding site in the top predicted site with a precision threshold of > 25 percent for 71 % of the dataset of 134 protein-ligand complexes from the GOLD dataset(Laurie and Jackson 2005). One of the key results from the study was that when comparing the volume of sites predicted by Q-SiteFinder to those predicted by PocketFinder (a geometric method), was that not only did Q-SiteFinder outperform PocketFinder in terms of prediction accuracy, but that Q-SiteFinder produced pockets with volumes that were not correlated to the total volume of the protein. Furthermore the volume of the Q-SiteFinder sites was more comparable to the volume of ligands that adhere to the Lipinski Rule of Five(Laurie and Jackson 2005).

For the original study the optimal cutoff of -1.4 kcal mol^{-1} was chosen so as to maximise prediction accuracy whilst retaining desirable high-precision predictions. A further study of protein-protein interactions using Q-SiteFinder used the cutoff of -1.3 kcal mol^{-1}(Nicholas J Burgoyne and Jackson 2006). Burgoyne and Jackson used Q-SiteFinder as a method to identify pockets on the protein that could potentially be hotspots for protein-ligand or protein-protein interactions. They showed that conservation, desolvation potential, electrostatics and unit-electrostatics could all help to elucidate protein-ligand binding pockets. For protein-protein interactions they saw much lower success rates, with desolvation potential being the only measure that performed reasonably well in all classes of protein-protein interaction. Generally they observed that the best predictions were made for enzymes (with desolvation potential performing well for antibodies), whilst enzyme inhibitors, antigens and other complexes were much more difficult to make high-quality predictions(Nicholas J Burgoyne and Jackson 2006).

2.2.2 Marketed drug datasets

Studies such as Q-SiteFinder and other pocket detection methods have typically used docking benchmarks to assess their performance. These datasets have the advantage that they are often well validated and well understood. Docking benchmarks can suffer from some problems since they are often aimed at only a few classes of targets such as Kinases.

There is a wealth of other structural information that can be used to identify and develop novel datasets. The most obvious is the PDB which has advanced search features that can be employed to filter results by a range of criteria(Berman *et al.* 2002). DrugBank is a resource for bioinformatic and cheminformatic information relating to drugs, targets and modes of action. It is cross-referenced with many major databases KEGG, PubChem, ChEBI, Swiss-prot, GenBank and the

PDB(Wishart *et al.* 2006),(Wishart *et al.* 2008). A manually curated database for proteins determined to < 2.5 Å resolution and related binding information is available in the form of the BindingMOAD(Benson *et al.* 2005).

Recently a powerful resource for studying protein-ligand complexes available in the PDB has been made available (CREDO)(Schreyer and T Blundell 2009), and related to this is a database (TIMBAL) that collates information on protein-protein interaction inhibitors(Higueruelo *et al.* 2009).

2.2.3 Predictions on unbound complexes

Pocket detection has been widely evaluated by removing the bound ligand from the protein-ligand complexed structure, applying the pocket detection algorithm, and then comparing the predicted pocket with the position of the bound ligand. Unfortunately there are several problems inherent with this technique. The first problem is that taking a bound structure with known ligand location, and predicting the location is of no practical use since most binding sites will undergo at least some degree of structural rearrangement on binding. For drug discovery or protein function annotation it would be desirable to be able to take the structure of an unbound protein and make an accurate prediction of the ligand binding site. One of the major problems with this is that proteins are flexible and as such may undergo large conformational change moving from the unbound to the bound conformation. Eyrisch and Helms investigated the opening and closing of pockets on protein-protein interaction partners and observed that binding pockets appeared transiently sometimes on the order of picoseconds(Eyrisch and Helms 2007). Additionally virtual screening strategies have been employed to systems such as hDM2/hDMX with the aim of improving hit-rates(Barakat *et al.* 2010).

2.2.4 Study aims

Small-molecule protein-protein interaction inhibitors are attractive to pharmaceuticals companies as they offer the opportunity to intervene therapeutically in a range of diseases in which it was not previously possible to do so. However, whilst previous ideas suggesting that this class of interactions was 'undruggable' are being challenged, there has still only been limited success in this field of drug discovery(Wells and McClendon 2007),(Arkin and Wells 2004). As such it would be desirable to better understand the similarities and differences between the interactions between: proteins and traditional marketed drugs, protein partners, and newly discovered small-molecule protein-protein interactions. To this end we define envelopes whereby it is energetically favourable for a ligand to interact with the protein and compare them between different classes of interactions. In understanding similarities and differences between these classes of interactions it is hoped that the chances of discovering novel small-molecule inhibitors of protein-protein interactions will be increased.

2.3 Methods

2.3.1 Preparation of datasets

Broadly speaking we use four datasets in this study: protein-ligand; protein-protein; protein-drug; protein-protein interaction inhibitor. We can further subdivide each dataset into bound and unbound subsets. We define the bound subset as complexes whereby a protein and ligand (or interacting protein partner) are in direct contact. We then remove the ligand from the PDB file in order to run Q-SiteFinder. The removed ligand is retained to test whether the Q-SiteFinder predicted sites are successful. We define the unbound subset as protein structures that correspond to proteins in the bound subset with a high degree of sequence similarity (> 95 % sequence identity). This sequence identity cutoff means that we are looking at very similar proteins that differ mostly because they are no-longer co-crystalized with their cognate ligand. This allows us to test the

predictive power of Q-SiteFinder in a situation similar to what would be found in a drug discovery scenario. Here a protein structure might be available, and the question is where a ligand might bind to the protein. Q-SiteFinder would then identify likely binding sites. When a protein binds a ligand it can change conformation thus altering the binding pocket. So whilst in many cases testing on the bound protein as previously described can be a useful metric, testing on the unbound protein allows a more realistic test.

Since we define the bound subset with no knowledge of unbound partners we note that the unbound subset is likely to be smaller than the bound subset. All bound and unbound protein-small-molecule datasets consist of only one protein chain, except in the rare cases where it is predicted that two or more protein chains may contribute to the ligand binding site, in these cases all relevant chains are retained prior to being passed through the following filter stage. After selection of PDBs for each dataset all datasets were filtered in order that they were suitable for input to the pocket detection algorithm Q-SiteFinder. The filtering removed all ligands that were not the intended ligand, whilst retaining any cofactors as part of the protein (such as HEM, NAD). The PDB files were then separated into two files one containing the protein (and any cofactors), the second containing the ligand file.

2.3.1.a Protein-ligand

The protein-ligand bound dataset consists of 134 complexes, which were initially described by Nissink *et al.*(Nissink *et al.* 2002). The dataset is non-redundant at the SCOP superfamily level(Murzin *et al.* 1995).

1aaq	1blh	1die	1glq	1lcp	1nis	1tka	2ack	2r07	4fab
1abe	1bma	1dr1	1hdc	1ldm	1pbd	1tmn	2ada	2sim	4phv
1acj	1byb	1dwd	1hdy	1lic	1pha	1tng	2ak3	2yhx	5p2p
1acl	1cbs	1eap	1hef	1lmo	1phd	1tni	2cgr	3cla	6abp
1acm	1cbx	1eed	1hfc	1lna	1phg	1tnl	2cht	3cpa	6rnt
1aco	1cdg	1epb	1hri	1lpm	1poc	1tph	2cmd	3gch	6rsa
1aec	1cil	1eta	1hsl	1lst	1rds	1tpp	2ctc	3hvt	7tim
1aha	1com	1etr	1hyt	1mcr	1rne	1trk	2dbl	3mth	8gch
1apt	1coy	1fen	1icn	1mdr	1rob	1tyl	2gbp	3ptb	
1ase	1cps	1fkg	1ida	1mmq	1slt	1ukz	2lgs	3tpi	
1atl	1ctr	1fki	1igj	1mrg	1snc	1ulb	2mcp	4aah	
1azm	1dbb	1frp	1imb	1mrk	1srj	1wap	2phh	4cts	
1baf	1dbj	1ghb	1ive	1mup	1stp	1xid	2pk4	4dfr	
1bbp	1did	1glp	1lah	1nco	1tdb	1xie	2plv	4est	

Table 2.1 – PDB codes of the 134 protein-ligand complexes that constitute the protein-ligand bound dataset.

The protein-ligand unbound dataset consists of 21 complexes described in Table 2.2. The protein-ligand unbound dataset is a subset of the dataset of 35 complexes described by Laurie & Jackson(Laurie and Jackson 2005) where structures that corresponded to entries in the docking benchmark of 305 complexes(Nissink *et al.* 2002), but not the bound docking benchmark of 134 complexes described in Table 2.1 were removed.

1qif	1bya	1ifb	1a4j	1ahc	1bbs	2rta	1krn	1chg	2ptn
1djb	1cge	1hsi	1ime	1phc	1stn	2ctb	2sil	6ins	3p2p
7rat									

Table 2.2 – PDB codes of the 21 protein monomers that constitute the protein-ligand unbound dataset.

2.3.1.b Protein-protein

The protein-protein bound dataset consists of the antibody-antigen and protease-inhibitor representatives of the protein-protein docking benchmark 1.0 described by Chen *et al.*(Chen et al. 2003). This dataset was expanded on by Burgoyne & Jackson and was further extended for this work, whereby the same redundancy criteria as described in the paper by Burgoyne & Jackson was applied(Nicholas J Burgoyne and Jackson 2006). The protein-protein bound dataset contained 103 pairwise-bound, non-obligate, hetero-protein complexes, corresponding to 206 protein monomers (shown in Table 2.3).

1a0o	1bql	1efu	1hcf	1im9	1kxq	1ml0	1ppe	1tab	1x86
1a2k	1brc	1eo8	1he1	1j2j	1kxt	1mlc	1pvh	1tgs	2jel
1acb	1brs	1ewy	1he8	1jhl	1kxv	1mz8	1qab	1tx4	2kai
1ahw	1bvk	1f6m	1hlu	1jwm	1kzg	1nbf	1qfu	1udi	2mta
1akj	1cgi	1f80	1hxy	1jzd	1l0y	1nca	1re0	1uex	2ptc
1am4	1cho	1fbi	1hyr	1kac	1lo5	1nmb	1s9d	1ugh	2sic
1atn	1cse	1fq1	1i4d	1kkl	1m27	1o6s	1sq0	1w1i	2sni
1avw	1de4	1fss	1i8l	1kkm	1mah	1o94	1stf	1wej	2tec
1avz	1dfj	1g4u	1iai	1ktz	1mel	1p4l	1svx	1wq1	2vir
1b6c	1dqj	1gaq	1igc	1kxp	1mi5	1p8v	1t6b	1www	4htc
1bdj	1e96	1ghq							

Table 2.3 – PDB codes of the 103 protein-protein complexes that constitute the protein-protein bound dataset.

The protein-protein unbound dataset consists of 190 monomers in Table 2.4, which fulfil the same redundancy criteria as the protein-protein bound dataset described above.

1byu	3lzt	1dqt	7rsa	1ijb	6pti	1a2p	1b3j	1d4t	1omp
1chn	1efu	1hhl	1kxv	1ppe	3ssi	5cha	1i49	2ace	1acc
1tde	1dkj	1igd	1m05	1mj0	1udi	1chg	1dr9	1pif	1ppn
1cd8	1fxa	1nkr	1dok	1sht	1poh	1cx8	1fbi	1kgc	1maa
1mh1	1keb	1naf	3lzt	1stf	2tec	5cha	1kcu	1dqq	1rgp
3dni	1l0h	1iai	1cei	1fsc	1chn	1bvl	1qqd	1m08	2ptn
1cse	2viu	1ck1	1ubi	1a2b	1bdj	1hh8	1o3y	1nb8	1jwi
1shf	1fpz	1l6p	1lza	1tab	1f6m	1d8t	1aif	1mel	2ptn
1d6o	1acb	1eaj	7nn9	1ijb	1hhj	1bql	1dlh	1nca	1mlb
2a0b	1mh1	1sph	1edh	1hpt	1rgp	1que	1jzo	1djn	1wer
1hrc	1a70	1sph	1o96	1ndw	1rdw	1eo8	1nob	1vac	1txd
1ba7	1ly2	1m9z	1ja3	1lza	5cha	1b39	1jb1	1m0z	1qfu
1aap	1wwb	1rdw	1jou	5p21	1avv	2bhn	1kw2	1sup	1akz
6pti	1mh1	1ghl	2c12	1he7	1ias	1g4w	1jhl	1bqu	2pka
1boy	5p21	1kxq	1lki	1dpf	1chn	1gaw	1pif	1nmb	2ptn
2ovo	1pne	1kxt	1brp	2viu	1qbl	1c3d	1pif	1ku1	1sup
1hpt	1enf	1an0	7nn9	1ugi	2ptn	1he9	1rj2	1pbv	1udh
1a6z	1hq8	1esf	1hur	1aan	1bra	1e8y	1bec	1m0z	2jel
2ovo	1an0	1shf	1hur	6pti	1fgn	1rdw	1dlh	2ptn	1thm

Table 2.4 – PDB codes of the 190 protein monomers that constitute the protein-protein unbound dataset.

2.3.1.c Protein-drug

The protein-drug dataset of 50 complexes was designed to reflect the diversity that can be found in marketed drugs. For the design of this dataset a marketed drug is defined as one that is labelled as ‘approved’ in the DrugBank(Wishart *et al.* 2006),(Wishart *et al.* 2008). An ‘approved’ drug is one that has been approved by the U.S. Food and Drug Administration (FDA). It should be noted that not all of these drugs are orally bioavailable, although they are all small-molecule drugs, (i.e. peptides, antibodies and other forms of therapeutics are not covered by this dataset). The dataset is non-redundant at the SCOP super family level, i.e. there are no two representatives from the SCOP superfamily included in the dataset(Murzin *et al.* 1995). It should be noted that 23 out of the 50 protein

complexes are represented at SCOP family level within the protein-ligand dataset of 134 complexes, and 27 out of 50 protein complexes are represented at the SCOP super family level. This level of redundancy is necessary to ensure a wide diversity of ligands is covered by the protein-drug dataset, but it is clear that there is still sufficient diversity at the protein level between the protein-drug and protein-ligand datasets.

1a4gA	1cqpA	1eveA	1gtbA	1jolA	1m17A	1p5zB	1rr8C
1ayvA	1dtlA	1f5lA	1h87A	1jqeA	1m4dA	1phg_	1s1xA
1b3nA	1dy4A	1fem_	1hpvA	1js3B	1mmkA	1pxxC	1s2aA
2a1hA	2bmlB	2kceA	2o7oA	1vm1A	1uofA	1upfA	1uumA
1b4e	1e7wA	1ffyA	1j78B	1lbcB	1nf7A	1qmfA	1tbfA
1a4lB	1cebA	1errB	1fkbA	1jffB	1liiA	1oq5A	1qzrB
1uzfA	1th6A						

Table 2.5 – PDB codes of the 50 protein-drug complexes that constitute the protein-drug bound dataset.

The members of the 50 bound complexes define the protein-drug unbound dataset of 33 complexes. For each bound complex the PDB(Berman *et al.* 2002) was queried, with any hits having a 90% sequence identity across the whole query sequence and no bound ligands (excepting crystallographic solvents as described by Gold & Jackson(Gold and Jackson 2006)) being marked as suitable candidates. All hits were then manually assessed for suitability, resulting in the dataset shown in Table 2.5.

132l	1bjz	1eea	1gta	1js6	1lfa	1m44	1pvg	1rxf	3phv
1a31	1ca2	1ekf	1hbq	1kas	1lio	1o8a	1qme	1shv	5cox
1aj4	1cl5	1f4b	1j8t	1kw2	1m14	1phc	1rtj	2cel	5dfr
1bd3	1d6o	1fto							

Table 2.6 – PDB codes of the 33 protein monomers that constitute the protein-drug unbound dataset.

2.3.1.d Protein-protein interaction inhibitor

The protein-protein interaction inhibitor bound dataset of 24 complexes was created by searching the literature for any references to protein-protein interaction inhibitors followed by query of the PDB for relevant structures. Due to the fact that at the time of writing only 7 classes of protein-protein interaction inhibitor have structures available in the PDB(Berman *et al.* 2002), the redundancy criteria were relaxed, and as a result the protein-protein interaction inhibitor dataset is the only dataset used in this study that is redundant at the SCOP family level.

1ysw	2o2f	1ysi	2o2m	2yxj	1m49	1py2	1rv1	1ttv	1tft
2o22	1ysg	1ysn	2o2n	1m48	1pw6	1qvn	1t4e	1tfq	1s1j
1y2g	1r6n	1s1s	1y2f						

Table 2.7 – PDB codes of the 24 protein-protein interaction inhibitor complexes that constitute the protein-protein interaction inhibitor bound dataset.

The protein-protein interaction inhibitor unbound dataset of 7 complexes was created using the same criteria as in the creation of the protein-drug unbound dataset, due to the redundancy mentioned during the creation of the protein-protein interaction inhibitor bound dataset, only 7 structures compose this dataset.

1g5m	1maz	1m47	1z1m	1f9x	1f46	1r6k

Table 2.8 – PDB codes of the 7 protein monomers that constitute the protein-protein interaction inhibitor unbound dataset.

2.3.2 Q-SiteFinder

Q-SiteFinder was used to calculate 99 pockets on the surface of each member protein of the four datasets, using the method described by Laurie and Jackson(Laurie and Jackson 2005). Briefly summarised the method has four constituent steps. Step one is to add hydrogens to the protein using the method

described by Jackson *et al.*. Step two rotates the protein about the geometric centre to minimize the bounding box volume, allowing the calculation speed to be increased as generally the fewest number of grid points are required for the calculation. Step 3 calculates the non-bonded interaction energy of a methyl probe, using the GRID force field parameters(Jackson 2002), with the protein at each position on a defined cubic grid of resolution 0.9 Å. Probes with a van der Waals interaction energy more favourable than -1.4 kcal mol^{-1} are retained for clustering. Step 4 is clustering probes to create active volumes. Clusters are defined by their spatial proximity, with any retained probes lying directly adjacent on non-cubic diagonals forming clusters. The sum of interaction energies from all probes comprising of the same cluster are then used to rank the clusters in order from the most to least favourable. The active volume is defined as the sum of cubes with sides of dimension 0.5 Å within 2 Å of the probe sites defining the cluster.

2.3.3 Analysing unbound datasets

We used the method of mapping pockets to the SASA of the unbound protein to determine whether the pockets determined in the unbound structure are likely to correspond to pockets that are occupied in the bound structure, as this method has been shown to work well for protein-protein interactions(Nicholas John Burgoyne 2007).

The method determines the SASA of the bound structure, followed by the SASA of the bound structure with the ligand removed. We then define interface residues to be those residues that are involved in a change in SASA. The SASA of the unbound protein is determined, followed by the SASA of each of the 99 pockets in turn. An unbound structure active volume is determined to be occupied when there is a change in SASA that involves greater than 95 % interface residues. This method allows us to map residues that are involved in a change of SASA on binding.

2.3.4 Optimising unbound predictions

To determine which cut-off value is the most optimal we aim to produce a distribution of pockets targeted that is close to that observed in the protein-ligand and protein-protein datasets. The cut-off value is determined by carrying out the mapping of pockets to SASA analysis described above using different cut-off values ranging between 0.5 Å and 1.0 Å at 0.05 Å intervals, for the protein-ligand and protein-protein datasets. We then measure the correspondence between the distribution of pockets targeted for the bound and unbound datasets using Fisher's exact test (fisher.test() in the R package stats), taking the best value to be our cut-off value. That is to say we use Fisher's exact test to ensure the closest match between the distribution of number of pockets targeted by ligands for the bound datasets, when compared to the same distribution as determined by our method of mapping pockets to SASA for the case of unbound datasets.

2.4 Results

We present results pertaining to several properties that allow us to distinguish noticeable similarities and differences between datasets. First we compare protein-ligand interactions to protein-protein interactions using volume of all surface pockets, active volume and percentage occupancy of pockets in Figure 2.1. We then move on to investigate noticeable similarities and differences between protein-drug interactions and protein-protein interaction inhibitors using the same properties described above in Figure 2.7. In Figure 2.3 we compare all four datasets using average active volume of bound pockets compared to average active volume of unbound pockets observed. The datasets are then analysed to show the distribution of occupied pockets in a given protein in Figure 2.6. The analysis of the predictions made on unbound datasets are then shown in Figure 2.5. Illustrations of typical protein-drug binding sites are then shown alongside representative protein-protein interaction inhibitor binding sites are then shown in Figure 2.4. Finally the average surface pocket volume and active volume are compared in the case of bound and unbound datasets in Figure 2.2.

2.4.1 Volume of protein pockets

2.4.1.a Protein-protein interactions compared to protein-ligand interactions

We first compare the protein-ligand interactions to protein-protein interactions. Several striking differences are observed, as shown in Figure 2.1. The first difference observed is that of the size of all active volumes on the protein surfaces. We define active volumes as the volume of any pocket for which Q-SiteFinder has predicted a favourable van der Waals interaction energy, regardless of whether the pocket is indeed coincident with a ligand. That is to say, the active volume is defined by a pocket where the interaction with a ligand may be favourable. Once pockets have been generated over the entire protein surface we can ask whether or not a ligand is coincident with the pocket, and in cases where this is true can determine the active volume of occupied pockets. For protein-ligand interactions the active volume for the top ranked pocket is large (508 $Å^3$), which is considerably more than the volume of protein-protein monomers (350 $Å^3$). As the rank of the pocket decreases we notice a decrease in active volume volume. It is also observed that the active volume (the area of the pocket for which it is favourable for methyl carbons to reside) of the top occupied pockets in protein-ligand interactions (471 $Å^3$) is twice that of protein-protein interaction (233 $Å^3$). We also observe a strong propensity for active volumes to be occupied if they are ranked highly (positions 1-3) in the case of protein-ligand interactions. This is however, not true for protein-protein interactions, where many occupied pockets are ranked low (positions 11-99), with only 30.6 % of occupied sites located in the top 3, compared to 85.8 % in the case of protein-ligand interactions.

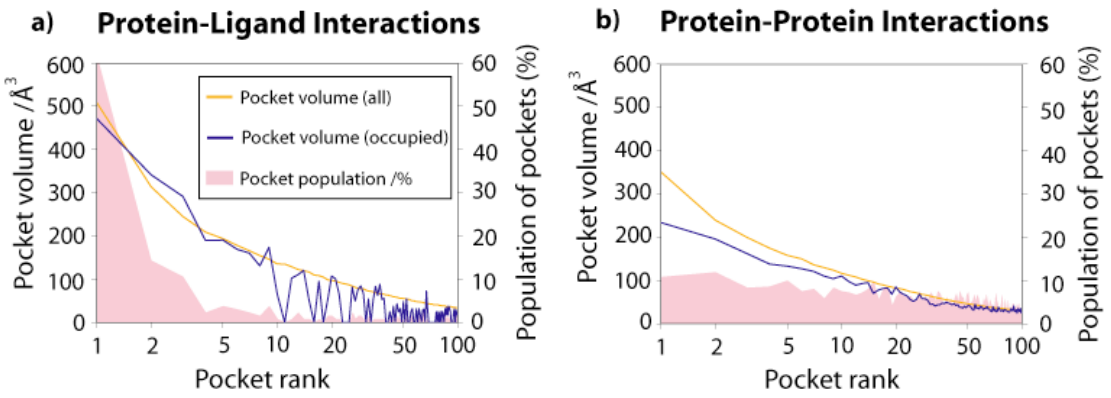

Figure 2.1: Average volume for all surface pockets (orange), occupied pockets (blue) and population of pockets (%) with occupancy for a given rank (pink fill). Both measures are compared to the pocket rank, where a pocket with the most favourable van der Waals interaction energy is ranked one. Protein-ligand interactions a) are represented by a set of 134 protein-ligand complexes, the dataset is non-redundant(Nissink *et al.* 2002), Protein-protein interactions b) are represented by a set of 97 pairwise bound complexes, all of which are non-obligate, hetero-protein complexes, a total of 194 monomers(Burgoyne and Jackson 2006). Occupied pockets (those in which Q-SiteFinder has been successful in identifying the ligand(Laurie and Jackson 2005)) are defined for each protein as those pockets that have their volume occupied by 25 % or greater by atoms from the interacting molecule (ligand or protein).

2.4.1.b Protein-protein interaction inhibitors compared to marketed drugs

Given that we have observed several differences between protein-ligand interactions and protein-protein interactions, we have gone on to investigate a dataset comprising of protein-drug interactions, with a view to determining whether these behave in a similar manner to a more general set of protein-ligand interactions. Equally we pose the question, do small-molecule protein-protein interaction inhibitor interactions share properties that link them to protein-ligand, protein-drug or protein-protein interactions? Comparison of protein-ligand interactions (Figure 2.1a) and protein-drug interactions (Figure 2.7a) seems to bear some similarities as might be expected (although the protein-ligand set contains few drug compounds it was designed as a protein-ligand docking benchmark and thus inevitably has ligands with drug-like properties). Indeed the protein-drug

dataset has a top ranked pocket with active volume slightly larger (580 Å^3) than the protein-ligand interaction (508 Å^3). In line with this slight increase in active volume of the top site, the active volume of the top occupied site also increases slightly (570 Å^3) when compared to the protein-ligand top occupied site (471 Å^3). We also see a similar pattern in the population of pockets, with high occupancy of top ranked pockets (78 % in top 3 pockets), and little occupancy of low ranked pockets (11-99). We next compare protein-protein interactions to protein-protein interaction inhibitor interactions noticing similar patterns of active volume size for both all pockets (288 Å^3) and occupied pockets (310 Å^3). However, we notice that the population of pockets seems to resemble the patterns seen in protein-ligand, protein-drug interactions (88 % in top 3 pockets). Finally we also note that whilst protein-ligand, protein-drug and protein-protein interaction inhibitor interactions all have occupied pocket active volumes that track the overall active pocket volume, however, this is not true for protein-protein interactions.

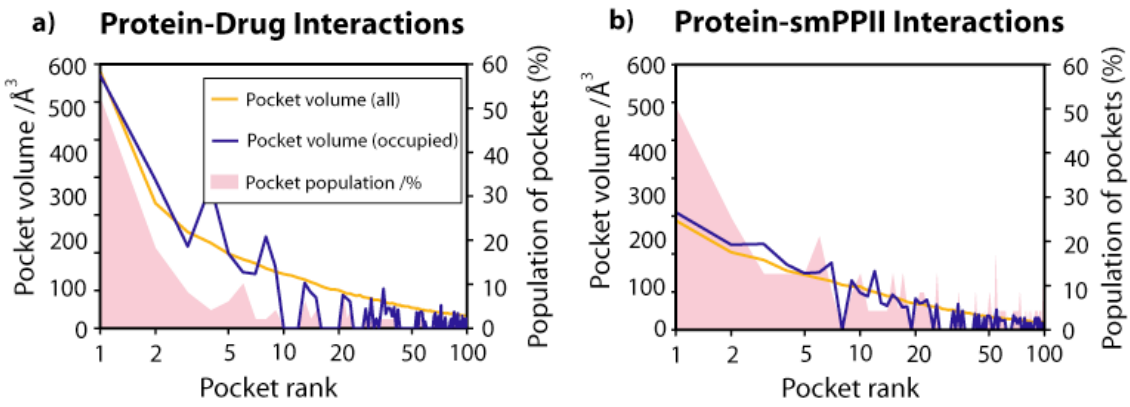

Figure 2.2: Average volume for all surface pockets (orange), occupied pockets (blue) and population of pockets (%) with occupancy for a given rank (pink fill). Both measures are compared to the pocket rank, where a pocket with the most favourable van der Waals interaction energy is ranked one. Protein-drug interactions a) are represented by a set of 50 protein-ligand complexes, the dataset is non-redundant at the SCOP superfamily level and shares only 14 superfamily relatives with the protein-ligand dataset(Fuller, Burgoyne and Jackson, 2009), Protein-small-molecule Protein-protein interaction inhibitors b) are represented by a set of 24 complexes, representing 7 distinct families of protein-protein interaction which is being blocked(Fuller, Burgoyne and Jackson, 2009). Occupied pockets (those in which Q-SiteFinder has been successful in identifying the ligand) are defined for each protein as those pockets that have their volume occupied by 25% or greater by atoms from the interacting molecule (ligand or protein).

The data about active volumes from Figure 2.1 and Figure 2.7 is summarised in terms of the average pocket volume over all pockets, and all occupied pockets in Figure 2.3. Protein-drug interactions have both the largest difference (Δ370 Å^3) between average volumes of all pockets (79 Å^3), and occupied pockets (449 Å^3). There is a similar picture for protein-ligand interactions, although the difference (Δ306 Å^3) is slightly less marked due to the average occupied volume (383 Å^3) being slightly lower, average volume of all pockets (77 Å^3), however remains the same. It is clear that for the two 'pure' protein-ligand interaction datasets, the average pocket volume is considerably lower than the average pocket volume for occupied pockets. That is to say, it appears that traditional ligands attempt to maximise the active volume of the pocket in which they bind. In the case of protein-protein interactions the difference (Δ60 Å^3) between the average volume of

all pockets (64 Å), and occupied pockets (124 Å^3) is far less marked. The protein-protein interaction inhibitor interactions lie somewhere between the two extremes. The difference (Δ176 Å^3) between the average volume of all pockets (52 Å^3) and occupied pockets (228 Å^3), is once again due mainly to the volume of all pockets remaining around a basal level, whilst the volume of occupied pockets is somewhat increased when compared to protein-protein interactions.

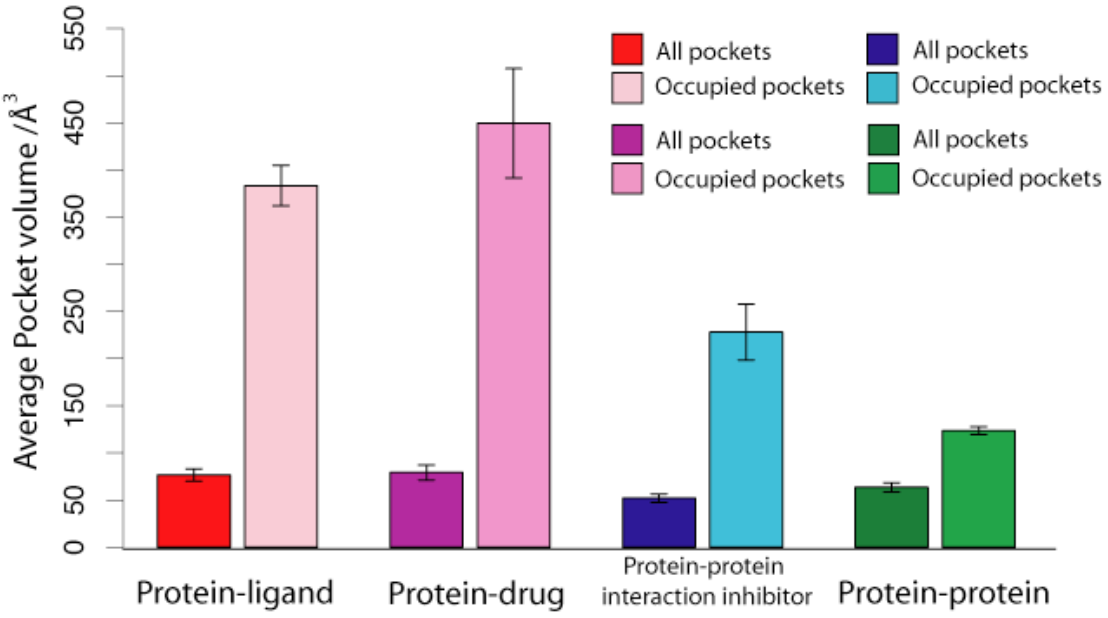

Figure 2.3: Average active volume of pockets for all sites on a protein surface compared to average active volume of occupied pockets. In all cases the occupied pockets have larger average volumes, than for a general pocket.

2.4.2 Number of pockets

After investigating the properties of the active volumes on the protein surface and observing some striking differences, we next investigate the number of these pockets that are targeted by ligands. We define a pocket to be targeted by a ligand if the pocket has > 25 percent of its volume covered by the ligand. In Figure

2.6 we observe that protein-ligand interactions shown in a) predominantly target one pocket (66 percent of cases), with the likelihood of greater than 3 pockets being targeted becoming increasingly unlikely. Protein-protein interactions follow a totally different trend, with the average number of pockets targeted being 5 (± 3). The distribution of pockets targeted is far less skewed for protein-protein interactions, when compared to protein-ligand interactions, which are strongly positively skewed. Protein-drug interactions also appear to follow the positively skewed distribution observed for protein-ligand interactions, although the peak at one site is slightly less strong (40 percent). Protein-protein interaction inhibitor interactions also target more than one pocket, with the average being around 4 (± 1).

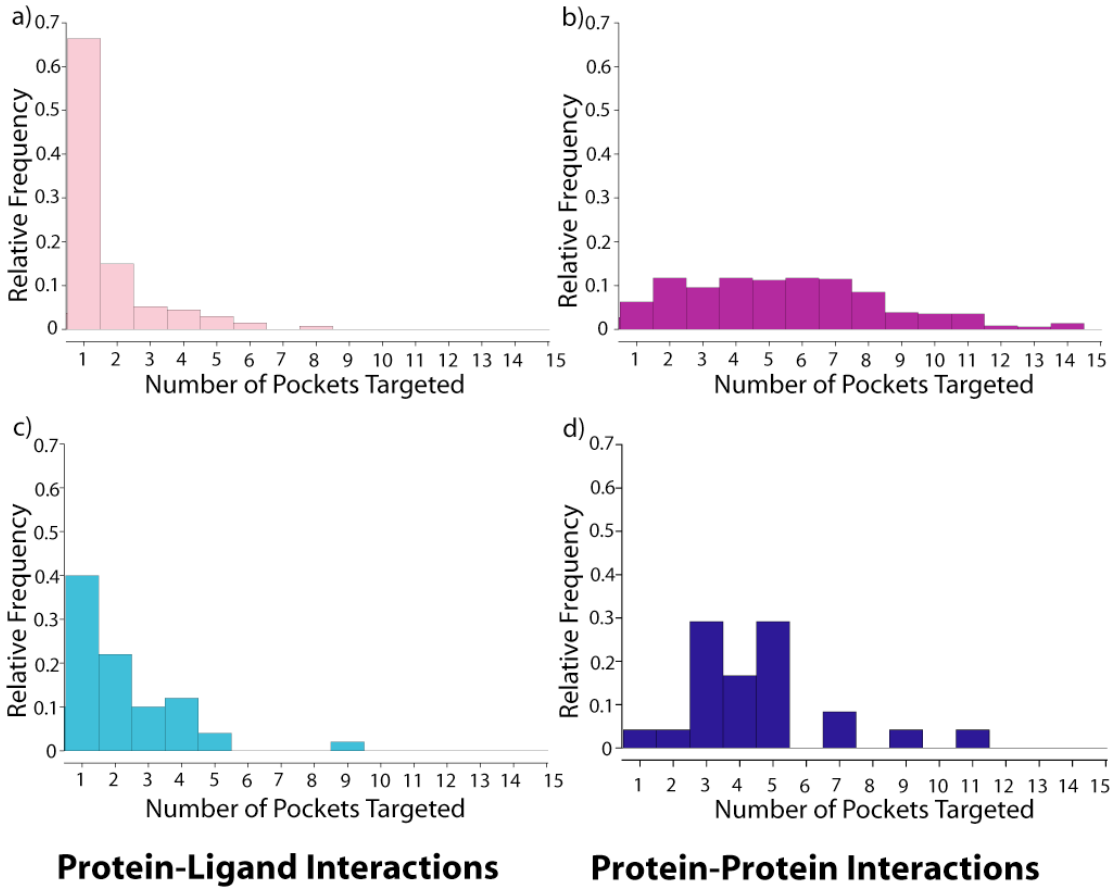

Figure 2.4: Histograms showing the relative frequency of a protein having a pocket targeted by its bound ligand. a) Protein-ligand dataset (134 complexes)(Nissink *et al.* 2002), b) Protein-protein dataset (97 pairwise-bound hetero complexes)(Burgoyne and Jackson 2006), c) Protein-drug dataset (50 complexes non-redundant at SCOP superfamily level, containing only drugs marked as approved by the FDA)(Fuller, Burgoyne and Jackson, 2009), d) Protein-protein interaction inhibitor dataset (24 complexes taken from 7 protein-protein interaction inhibitor interaction classes with structures available)(Fuller, Burgoyne and Jackson, 2009). a) and c) both show a very positively skewed distribution, whilst b) and d) both show only slight positive skew.

We have now made several observations regarding the nature of interactions of proteins with a variety of different classes of ligands, and a set of interacting protein partners. All of the previous observations have been made using bound complexes. As of June 2010, there are more than 56,000 X-ray structures held within the PDB(Berman *et al.* 2002), of which more than 40,000 are co-crystallised with a ligand although many of these are crystallographic compounds, around

10,000 are biologically relevant. This is a very small diversity of chemical space and as such it would be informative to see whether the observations made for bound structures hold true when applied to unbound structures. As discussed earlier it is necessary to have a definition of success for the unbound dataset. The results in Figure 2.6 show the results pertinent to the optimisation of the method of mapping pockets to the SASA. a) to d) show the distribution of the number of pockets targeted by ligands in the bound structures. e) to h) show the distribution of the number of pockets targeted by ligands in the unbound structures when using the 95 % cutoff. The dark blue line shows the distribution of the unbound dataset, whilst the light blue line shows the distribution of the bound dataset. The method is aiming to optimise the correspondence between the two lines. When comparing the protein-ligand distributions in Figure 2.6e, we notice that although the distribution shows strong positive skew, it does not show the strong peak at 1, which is observed in the bound distribution. Fortunately when comparing protein-drug interactions in Figure 2.6f, we once again see the strong positive skew and a higher peak for pockets ranked at position 1. We do however see that a relatively large proportion (25 %) of ligands in this distribution do not appear to be targeting any pockets. Results for the protein-protein interaction inhibitor interactions in Figure 2.6g show a reasonably strong correlation between the two datasets, with the modal peak centred around 4 for the bound dataset, and slightly nearer 5 for the unbound dataset. The results for the protein-protein interactions in Figure 2.6, also show a strong correspondence between distributions, with the modal peak once again being centred near 6 for both bound and unbound distributions.

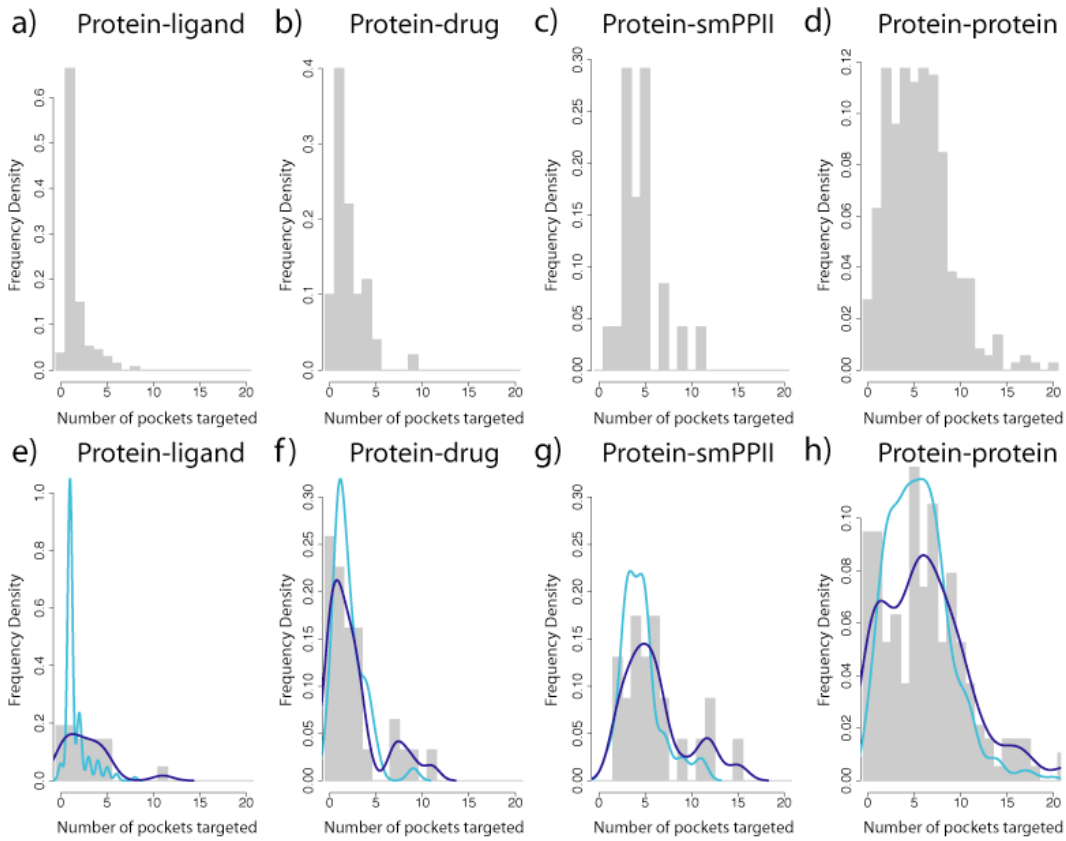

Figure 2.5: a)-d) show histograms of Frequency density vs. Number of pockets targeted for each of the investigated bound datasets. e)-h) show histograms of Frequency density vs. Number of pockets targeted for each of the investigated unbound datasets. For the unbound datasets a pocket was determined to be targeted if the change in SASA (SASA of the unbound complex plus the relevant active volume minus SASA of the unbound complex) due to interface residues was greater than 95 %. The light blue line corresponds to the distribution of the bound dataset, whilst the dark blue line corresponds to the distribution of the unbound dataset. Matching the distributions shown by the light blue and dark blue lines optimised the cut-off value of 0.95.

Figure 2.5 attempts to give an insight into the nature of the pockets that we have discussed. The left hand panel (a-f) shows some representatives from the protein-drug dataset. Many of these have top ranked pockets that accurately envelope the ligand binding site. It should be noted that these ‘representative’ sites are in many cases close to the volume that one would expect to find given the results

previously presented. The right hand panel (g-l) shows protein-protein interaction inhibitor representatives from 6 of the 7 families covered by the dataset. The Bcl-2/BH3 interaction was not chosen, as it is relatively similar to Bcl-X_L/BH3 in nature.

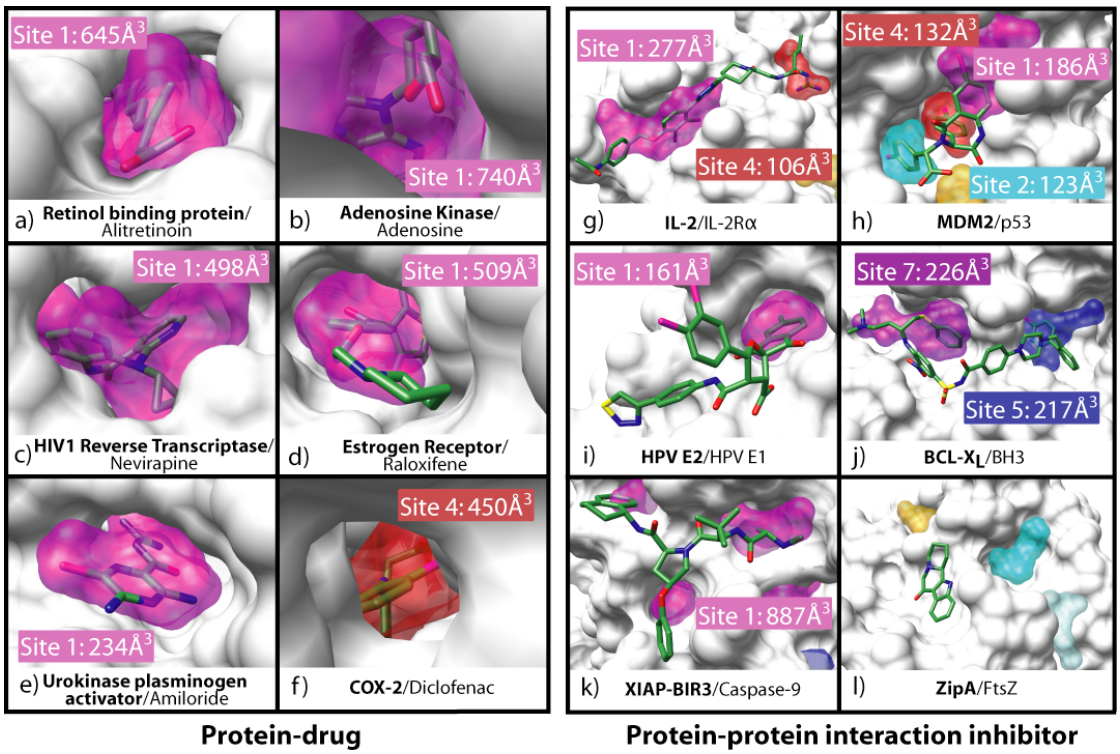

Figure 2.6: Images of the protein-ligand interface with the ligand coloured according to atom type, protein solvent accessible surface area shown in grey, and active volume of the pocket coloured according to the rank of the site. a)-f) show six representatives from the protein-drug dataset of 50 complexes, g)-l) show representatives of 6 out of 7 of the protein-protein interaction inhibitor classes represented in the protein-protein interaction inhibitor dataset of 24 complexes. PDB codes are given in order a)-f) 1fem, 1err, 1s1x, 1abe, 1f5l, 1pxx. g)-l) 1qvn, 1t4e, 1r6n, 2yxj, 1tft, 1s1j. Binding pocket images were prepared using UCSF Chimera(Pettersen *et al.* 2004).

We have previously observed (in Figure 2.4) that when looking at bound datasets, the average pocket volume of pockets on the protein surface is less than the average active volume in all cases except protein-protein interactions. The magenta bar in Figure 2.7 shows the same results as Figure 2.4, whilst the dark blue bar shows the average active volume of the top ranked pocket from each complex to allow comparison with unbound proteins. We note that in all cases the

average surface pocket volume for unbound proteins is slightly less for the unbound datasets. However, the striking difference is observed when comparing the average active volume of the top occupied pocket. For each dataset there is a significant decrease in active volume. This decrease is most noticeable for the protein-drug (Δvolume: 314 Å^3) and protein-ligand (Δvolume: 223 Å^3) datasets. However it is also significant to for protein-protein interaction inhibitors (Δvolume: 137 Å^3) and to a lesser extent protein-protein interactions (Δvolume: 78 Å^3).

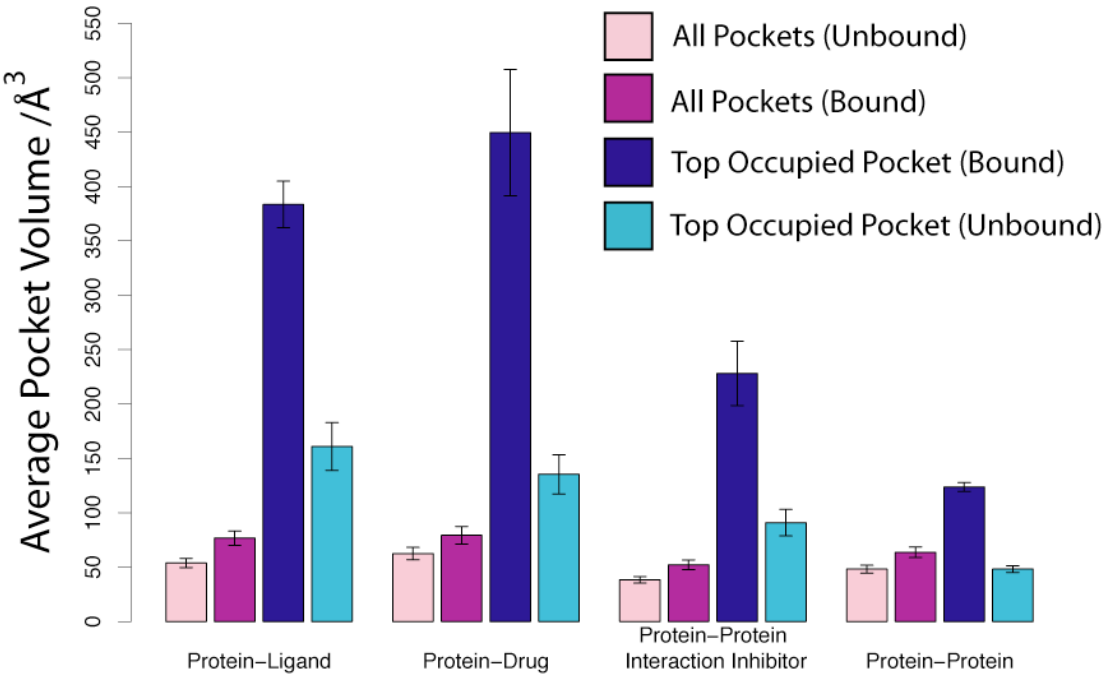

Figure 2.7: Average active volume of pockets for all sites on a protein surface in the bound (purple) and unbound (pink) states, and all occupied top ranked sites in the bound (blue) as well as those top unbound sites defined as corresponding to an occupied site in the bound protein structure (cyan). Error bars show standard error on the mean.

2.5 Discussion and Conclusion

The first major observation that we make is that PLIs have much larger active volumes than PPIs. In fact the average active volume for PLIs is more than twice that of the active volume observed with PPIs. We also see that PLIs generally

occur in a single pocket, compared to PPIs that tend to occur in multiple pockets. So ligand binding appears to favour targeting a single high-volume pocket where it can bind and optimize itself efficiently. PPIs on the other hand tend to target several lower volume pockets on the protein surface. Proteins binding to their cognate protein are able to achieve this due to their far larger size compared to binding to a cognate ligand.

We expect to see a similar picture when we transfer for looking at the general PLI dataset to the marketed drug dataset. This is indeed the case, with a slight increase in the average active volume targeted. This is likely due to variance in the data. The fact that the average volume of surface pockets is the same for both the PLI and marketed drug datasets also adds to confidence that the two datasets are very similar.

We next make comparison between the marketed drug dataset and the PPI dataset. It is observed that all three datasets have similar average surface pocket volumes. Marketed drugs have active volumes that are nearly twice that of protein-protein interaction inhibitors, which in turn have active volumes nearly 60 % larger than PPIs. Thus protein-protein interaction inhibitors lie in an intermediate, having active volumes somewhere between those of marketed drugs and PPIs. We have already made the observation that PLIs and marketed drugs tend to occupy a single high volume pocket, whilst PPIs tend to occupy several lower volume pockets. Once again protein-protein interaction inhibitors appear to occupy an intermediate position, whereby they tend to occupy more than one pocket, but fewer pockets than PPIs.

The evidence presented above suggests that current efforts at elucidating small-molecule inhibitors of protein-protein interaction inhibitors may have been successful due to the compound's abilities to target several pockets on a protein

surface. These pockets are generally smaller than those found occupied in PLIs and marketed drugs, but larger than those found in PPIs. The fact that multiple pockets are targeted perhaps confers dual benefits to any small-molecule attempting to out compete an interacting protein partner. The first is that it allows a larger proportion of the native protein interface to be shielded from the competing protein. The second is that it allows the small-molecule to increase its binding potency by targeting several energetically favourable hot-spot regions. This leads us to believe that in order to increase success levels in targeting protein-protein interactions, knowledge of the location and properties of hot-spot regions that confer stability to protein-protein interactions, yet are also favourable for the binding of ligands is required.

2.6 References

Arkin, Michelle R, and James A Wells. 2004. Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nature Reviews Drug Discovery* 3, no. 4 (April): 301-17. doi:10.1038/nrd1343. http://www.ncbi.nlm.nih.gov/pubmed/15060526.

Barakat, Khaled, Jonathan Mane, Douglas Friesen, and Jack Tuszynski. 2010. Ensemble-based virtual screening reveals dual-inhibitors for the p53-MDM2/MDMX interactions. *Journal of Molecular Graphics & Modelling* 28, no. 6: 555-68. doi:10.1016/j.jmgm.2009.12.003. http://www.ncbi.nlm.nih.gov/pubmed/20056466.

Benson, Mark L, Richard D Smith, Liegi Hu, Michael G Lerner, and Heather A Carlson. 2005. Binding MOAD (Mother Of All Databases). *Proteins* 60, no. 3 (August): 333-40. doi:10.1002/prot.20512. http://www.ncbi.nlm.nih.gov/pubmed/15971202.

Berman, Helen M., Tammy Battistuz, T. N. Bhat, Wolfgang F. Bluhm, Philip E. Bourne, Kyle Burkhardt, Zukang Feng, *et al.* 2002. The Protein Data Bank. *Acta Crystallographica Section D Biological Crystallography* 58, no. 6 (May): 899-907. doi:10.1107/S0907444902003451. http://scripts.iucr.org/cgi-bin/paper?S0907444902003451.

Billingsley, Melvin L. 2008. Druggable targets and targeted drugs: enhancing the development of new therapeutics. *Pharmacology* 82, no. 4: 239-44. doi:10.1159/000157624. http://www.ncbi.nlm.nih.gov/pubmed/18802381.

Burgoyne, Nicholas J, and Richard M Jackson. 2006. Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces. *Bioinformatics (Oxford, England)* 22, no. 11 (March): 1335-42. doi:10.1093/bioinformatics/btl079. http://www.ncbi.nlm.nih.gov/pubmed/16522669.

Burgoyne, Nicholas John. 2007. The Structural Analysis and Prediction of Protein interactions. PhD Thesis, Institute of *Molecular and Cellular Biology, University of Leeds, UK*.

Chen, Rong, Julian Mintseris, Joël Janin, and Zhiping Weng. 2003. A protein-protein docking benchmark. *Proteins* 52, no. 1 (July): 88-91. doi:10.1002/prot.10390. http://www.ncbi.nlm.nih.gov/pubmed/12784372.

Egner, Ursula, and Roman C Hillig. 2008. A structural biology view of target drugability. *Expert Opinion on Drug Discovery* 3, no. 4 (April): 391-401. doi:10.1517/17460441.3.4.391. http://www.expertopin.com/doi/abs/10.1517/17460441.3.4.391.

Eyrisch, Susanne, and Volkhard Helms. 2007. Transient pockets on protein surfaces involved in protein-protein interaction. *Journal of Medicinal Chemistry* 50, no. 15 (July): 3457-64. doi:10.1021/jm070095g. http://www.ncbi.nlm.nih.gov/pubmed/17602601.

Fuller, J.C., Burgoyne, N.J., and Jackson, R.M.. 2009. Predicting druggable binding sites at the protein-protein interface. *Drug Discovery Today* 14, no. 3-4 (February): 155-61. doi:10.1016/j.drudis.2008.10.009. http://www.ncbi.nlm.nih.gov/pubmed/19041415.

Gold, Nicola D, and Richard M Jackson. 2006. SitesBase: a database for structure-based protein-ligand binding site comparisons. *Nucleic Acids Research* 34, no. Database issue (January): D231-4. doi:10.1093/nar/gkj062. http://www.ncbi.nlm.nih.gov/pubmed/16381853.

Hajduk, Philip J, Jeffrey R Huth, and Stephen W Fesik. 2005. Druggability indices for protein targets derived from NMR-based screening data. *Journal of Medicinal Chemistry* 48, no. 7 (April): 2518-25. doi:10.1021/jm049131r. http://www.ncbi.nlm.nih.gov/pubmed/15801841.

Hajduk, Philip J, Jeffrey R Huth, and Christin Tse. 2005. Predicting protein druggability. *Drug Discovery Today* 10, no. 23-24 (December): 1675-82. doi:10.1016/S1359-6446(05)03624-X. http://www.ncbi.nlm.nih.gov/pubmed/16376828.

Henrich, Stefan, Outi M H Salo-Ahen, B Huang, Friedrich F Rippmann, Gabriele Cruciani, and Rebecca C Wade. 2010. Computational approaches to identifying and characterizing protein binding sites for ligand design. *Journal of Molecular Recognition : JMR* 23, no. 2 (March): 209-19. doi:10.1002/jmr.984. http://www.ncbi.nlm.nih.gov/pubmed/19746440.

Higueruelo, Alícia P, Adrian Schreyer, G Richard J Bickerton, Will R Pitt, Colin R Groom, and TL Blundell. 2009. Atomic interactions and profile of small molecules disrupting protein-protein interfaces: the TIMBAL database. *Chemical Biology & Drug Design* 74, no. 5 (November): 457-67. doi:10.1111/j.1747-0285.2009.00889.x. http://www.ncbi.nlm.nih.gov/pubmed/19811506.

Hopkins, Andrew L, and Colin R Groom. 2002. The druggable genome. *Nature Reviews Drug Discovery* 1, no. 9 (September): 727-30. doi:10.1038/nrd892. http://www.ncbi.nlm.nih.gov/pubmed/12209152.

Jackson, Richard M. 2002. Q-fit: a probabilistic method for docking molecular fragments by sampling low energy conformational space. *Journal of Computer-aided Molecular Design* 16, no. 1 (January): 43-57. http://www.ncbi.nlm.nih.gov/pubmed/12197665.

Keller, Thomas H, Arkadius Pichota, and Zheng Yin. 2006. A practical view of “druggability”. *Current Opinion in Chemical Biology* 10, no. 4: 357-61. doi:10.1016/j.cbpa.2006.06.014. http://www.ncbi.nlm.nih.gov/pubmed/16814592.

Laurie, Alasdair T R, and Richard M Jackson. 2005. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics (Oxford, England)* 21, no. 9 (February): 1908-16. doi:10.1093/bioinformatics/bti315. http://www.ncbi.nlm.nih.gov/pubmed/15701681.

Murzin, A, S Brenner, T Hubbard, and C Chothia. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology* 247, no. 4 (April): 536-540. doi:10.1016/S0022-2836(05)80134-2. http://linkinghub.elsevier.com/retrieve/pii/S0022283605801342.

Nissink, J Willem M, Chris Murray, Mike Hartshorn, Marcel L Verdonk, Jason C Cole, and Robin Taylor. 2002. A new test set for validating predictions of protein-ligand interaction. *Proteins* 49, no. 4 (December): 457-71. doi:10.1002/prot.10232. http://www.ncbi.nlm.nih.gov/pubmed/12402356.

Overington, John P, Bissan Al-Lazikani, and Andrew L Hopkins. 2006. How many drug targets are there? *Nature Reviews Drug Discovery* 5, no. 12 (December): 993-6. doi:10.1038/nrd2199. http://www.ncbi.nlm.nih.gov/pubmed/17139284.

Pettersen, Eric F, Thomas D Goddard, CC Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. 2004. UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of Computational Chemistry* 25, no. 13 (October): 1605-12. doi:10.1002/jcc.20084. http://www.ncbi.nlm.nih.gov/pubmed/15264254.

Schreyer, Adrian, and T Blundell. 2009. CREDO: a protein-ligand interaction database for drug discovery. *Chemical Biology & Drug Design* 73, no. 2 (February): 157-67. doi:10.1111/j.1747-0285.2008.00762.x. http://www.ncbi.nlm.nih.gov/pubmed/19207418.

Wells, James A, and Christopher L McClendon. 2007. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature* 450, no. 7172: 1001-9. doi:10.1038/nature06526. http://www.ncbi.nlm.nih.gov/pubmed/18075579.

Wishart, David S, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. 2008. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research* 36, no. Database issue (January): D901-6. doi:10.1093/nar/gkm958. http://www.ncbi.nlm.nih.gov/pubmed/18048412.

Wishart, David S, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. 2006. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research* 34, no. Database issue (January): D668-72. doi:10.1093/nar/gkj067. http://www.ncbi.nlm.nih.gov/pubmed/16381955.

3 Predicting ligand binding pockets: using physical models and machine learning techniques

3.1 Abstract

The ability to accurately determine which regions of a protein are involved in binding are of interest to several fields including molecular docking and functional annotation of proteins. Furthermore, the ability to prioritise certain binding pockets as more likely to accommodate high-affinity drug-like ligands would allow resources to be focussed towards those drug targets with better chance of success. We first analyse a large dataset of biologically relevant ligands from the PDB to determine the range of energies that they interact with the protein using the GRID force field. We then present a method that uses Q-SiteFinder to identify pockets and the GRID force field parameters to investigate whether certain probe types can give more information about the drugability of the pocket. Additional parameters identified in previous work by Hajduk *et al.*(Hajduk, Huth, and Fesik 2005) are added to create a pool of predictor variables that are used by a random forest machine learning method that aims to identify binding and non-binding pockets and furthermore to identify binding pockets that might accommodate high-affinity drug-like ligands.

3.2 Introduction

In recent years many new methods of pocket detection have been proposed, however, broadly speaking they tend to be based around one of several main concepts(Laurie and Jackson 2006). Most of these pocket detection methods can reliably determine binding sites in the top ranked sites from a dataset in at least

50 % of cases, indeed many have been shown to be far more successful. Knowledge of the location and shape of binding pockets have been used for docking studies and also as a tool for functional annotation of proteins.

One of the first discussions of the maximal affinity that a ligand might be able to develop for a protein is presented by Kuntz *et al.*(Kuntz *et al.* 1999). They performed an analysis of the binding energies of a large selection of ligands showing that the strongest binders produce on average -1.5 kcal mol^{-1} per non-hydrogen atom, with relatively little increase for greater than 15 non-hydrogen atoms(Kuntz *et al.* 1999). The average binding energy contributed per non-hydrogen atom is now known as the ligand efficiency. Cheng and co-workers took the idea of maximal affinity of ligands and asked the converse question, what is the maximal affinity that the ideal ligand might develop in a pocket on the protein surface. They developed a simple physical model based on a desolvation score for the binding site, a desolvation score for the ligand and a constant contribution from factors such as van der Waals interactions, electrostatics interactions and changes in entropy(Cheng *et al.* 2007). The solvation model for the binding site is based on the curvature of the site, whilst for the ligand it is a constant multiplied by the solvent accessible surface area of the ligand(Lee and Richards 1971). The authors then applied their maximal affinity score to a small set of proteins of which some were classed as druggable, some difficult and the remaining undruggable according to a binding affinity threshold (undruggable if predicted K_d > 100nM). They showed reasonable success in classifying their data into these classes using their model. Halgren discusses some of the problems that might be associated with using the maximal affinity model proposed by Cheng, and develops a novel method based on concepts based on pocket detection(Halgren 2009). The model is based on weighted, rescaled contributions from three variables calculated from the site. The first is the number of grid points defining the site, the second is the enclosure of the site and the third is hydrophillic score of the site. Two scores are then generated, a site score and a drug score. The former aims to distinguish

binding sites from non-binding sites, whilst the latter aims to determine whether a binding site is likely to accommodate a high-affinity ligand. The score and weights were trained using a dataset from BindingMOAD such that the score correlates with measured binding affinities(Benson *et al.* 2005). The drug score is then applied to a dataset based on the work previously described by Cheng(Cheng *et al.* 2007), where Halgren showed the ability to classify drugs into each of the three categories based on the magnitude of the drug score(Halgren 2009).

Hajduk and co-workers had previously used NMR hit-rates to approach the question of 'drugability' of a binding site. They observed that higher NMR hit-rates tend to correlate to binding sites that could accommodate a high-affinity drug-like compound(Hajduk, Huth, and Fesik 2005). They further investigated the properties of 'druggable' binding sites to determine whether there are certain properties that predispose a binding site to accommodate a high-affinity ligand. They observed that there is wide variability in many properties of pockets such as volume, surface area and number of charged residues(Hajduk, Huth, and Fesik 2005).

In order to avoid confusion several key terms used in this study are defined here to avoid ambiguity. When discussing bound pockets we are describing pockets identified using Q-SiteFinder - and a 25 % precision threshold described in the previous chapter - from the crystal structure of a protein-ligand complex. Unbound pockets are other pockets identified on the complex using Q-SiteFinder. That is to say that unbound does not describe pockets derived from the apo crystal structure and mapped to their relative location on the protein-ligand complex. It describes other pockets that do not contain a ligand but are derived from the protein-ligand bound coordinates. When discussing probes we mean specific locations at which a GRID type is located. A GRID type in this context is describing an atom or functional group for example C3 represents a methyl group. When discussing a GRID atom we mean that the energy of an atom (or group of atoms) from a specific ligand bound to a protein.

Regarding datasets we use three key datasets all of which are described in detail in the methods. We describe a large dataset of biologically relevant ligands, a dataset of proteins contained within BindingMOAD(Benson *et al.* 2005). BindingMOAD contains crystal structures of resolution better than 2.5 Å, and includes binding affinity data of the compound for the protein where available. The dataset previously described by Cheng *et al.*(Cheng *et al.* 2007) and used in work by Halgren(Halgren 2009). We refer to the datasets as: the biologically relevant ligands; BindingMOAD; the Halgren dataset, respectively.

Previously machine learning techniques have been applied to the problem of drugability prediction. Sugaya and Ikeda identified the SMAD4/SKI interaction as a candidate drugable PPI using a Support Vector Machine (SVM) that used Structural, Chemical and Functional information(Sugaya and Ikeda 2009). In this study a random forest classifier was used as it has previously been shown to perform well on unbalanced datasets and additionally on datasets with correlated learning features(Chen, Liaw, and Breiman 2004). Furthermore random forest classifiers, in contrast to SVMs, have the ability to show which learning features are used to make predictions, thus allowing evaluation of which features are most important. A random forest is a collection of decision trees. An input vector of learning features is fed into the each of the decision trees that comprise the random forest. Each decision tree outputs a classification which are then combined with the classifications from other trees. The class with the most decision trees voting for it is then chosen as the 'correct' classification.

This study aims to answer two key questions: 1) Is it possible to distinguish a binding pocket from a non-binding pocket on the protein surface? A combination of GRID point energies, a simple solvation measure, surface area and presence of donor/acceptor residues are combined and a machine learning approach is applied to determine whether a reliable classifier can be identified using these

metrics. The second more difficult problem is to ask the question 2) Is a pocket identified as a binder druggable? This has been attempted before in several papers previously detailed, and in this context we are not directly asking what the maximal affinity of a ligand for a site might be, but whether the previously derived classifier might allow druggable binding sites to be distinguished from difficult or undruggable sites.

3.3 Methods

3.3.1 Datasets

3.3.1.a Biologically relevant ligand dataset

The advanced search feature of the PDB was used to retrieve all structures from the PDB with R-free < 0.25 that were determined to less than 2.5 Å by X-ray crystallography and also contained a ligand. Biounit data from the PDB accessed on 09/02/2010 was downloaded in PDB format(Berman *et al.* 2002). A list of ligands known to be crystallographic solvents and additives was used to exclude several ligands(Strömbergsson and Kleywegt 2009). Further to this criteria used to exclude ligands were that the nearest ligand atom to the protein should be less than 5 Å in distance. The ligand should have more than 10 heavy atoms including at least one carbon and at least one nitrogen or oxygen. This produced a total of 8,861 complexes.

3.3.1.b BindingMOAD dataset

The BindingMOAD dataset was accessed on 23/06/2010 and all ligands with binding data were downloaded and processed using the same rules as described for the biologically relevant dataset to produce a total of 250 complexes.

3.3.1.c Halgren (Cheng) dataset

The PDB files described in the paper by Halgren were downloaded from the PDB and processed using the same criteria as previously defined in the section on generating the biologically relevant ligands dataset to produce a total of 62 complexes(Halgren 2009). The PDB 1T03 was excluded from the dataset as it was not clear which ligand should be used in the analysis.

3.3.2 GRID

Ligand atoms were assigned GRID atom types using the gmol2 program whilst the GRID calculations themselves were performed using liggrid, an implementation of GRID described previously by Jackson(Jackson 2002). GRID calculations for ligands were calculated over a box enclosing the ligand centred on the ligand centre of mass. A C program written by Alasdair Laurie was used to calculate the interaction energy of the ligand atoms based on the interaction energy of the nearby points. In order for this to occur the ligand was first minimized on the grid using the algorithm defined in the paper by Jackson(Jackson 2002). Calculations corresponding to the comparison of GRID energies of Q-SiteFinder sites to the energies of GRID atoms in ligands used all GRID points with spacing 0.5 Å in the box defined by the centre of mass plus the maximum dimensions of the Q-SiteFinder site. GRID calculations for OH2 probes defining a Q-SiteFinder site were performed on a GRID with spacing 0.9 Å centred on the Q-SiteFinder pocket centre of mass and grid points that did not coincide with a Q-SiteFinder site point were discarded.

3.3.3 Q-SiteFinder

Q-SiteFinder was used to calculate 99 pockets on the surface of each member protein of the four datasets, using the method described by Laurie and Jackson(Laurie and Jackson 2005). Briefly summarised the method has four constituent steps. Step one is to add hydrogens to the protein using the method described by Jackson *et al.*(Jackson 2002) Step two rotates the protein about the

geometric centre to minimize the bounding box volume, allowing the calculation speed to be increased as generally the fewest number of grid points are required for the calculation. Step 3 calculates the non-bonded interaction energy of a methyl probe, using the GRID force field parameters(Goodford 1985), with the protein at each position on a defined cubic grid of resolution 0.9 Å. Probes with a van der Waals interaction energy more favourable than -1.4 kcal mol^{-1} are retained for clustering. Step 4 is clustering probes to create active volumes. Clusters are defined by their spatial proximity, with any retained probes lying directly adjacent on non-cubic diagonals forming clusters. The sum of interaction energies from all probes comprising of the same cluster are then used to rank the clusters in order from the most to least favourable. The active volume is defined as the sum of cubes with sides of dimension 0.5 Å within 2 Å of the probe sites defining the cluster.

3.3.4 Identifying high energy probes

We hypothesised that high energy probes (identified by GRID) would indicate regions where particular atoms from a ligand could favourably reside. We developed a z-score based method that uses the data collected from probes observed in biologically relevant ligands to generate mean and standard deviation values for each probe type. Each probe contained within a pocket then uses a z-score cut-off to exclude all probe energies with a z-score less favourable than 1.7 from the mean. Since we were initially interested in determining the most likely probe type at that location, from a set of several probe types, the probe with the most favourable z-score is retained.

3.3.5 Half-sphere exposure

Buried binding pockets tend to be more druggable, so we used a method called half-sphere exposure to identify solvent accessibility. Half-sphere exposure was calculated using the algorithm implemented in the Biopython toolkit(Cock *et al.* 2009). The algorithm is described in detail in the paper by Hamelryck(Hamelryck

2005). Briefly the algorithm counts the number of atoms within 13 Å from the Cα in two half-spheres. One points along the Cα-Cß direction (hse_u, the solvent facing sphere), whilst the second half-sphere points along the Cß-Cα direction (hse_d, the protein facing sphere).

3.3.6 Machine learning

Machine learning was performed using two alternative types of Random Forest. In both cases the BindingMOAD dataset was split into four separate equally sized sets. In each of four cross-validations three sets were used for training whilst one set was used for testing. As a further investigation the Halgren dataset was used to test the predictions of each of the previously determined random forests.

The first random forest method is a standard implementation of Random Forest in the R statistical computing language called randomForest(Anon 2009). Default parameters were used in this instance. The second is an implementation of an unbiased random forest written for the weka machine learning package based on the original fast random forest package(Hall *et al.* 2009). The methods are described in detail in a technical report(Chen, Liaw, and Breiman 2004), and have been applied in a study looking to identify peptide biomarkers(Fusaro *et al.* 2009).

Rationale for inclusion of individual machine learning feature is included in the results and discussion, however, a description of each of the variables is given in table 3.1.

Learning feature	Description
C3 points	The total number of C3 points defining the pocket
C3 energy	The sum of the energy of the C3 points defining the pocket
OH2 points	The total number of OH2 points that have energy greater than the coincident C3 point.
OH2 energy	The sum of the energy of OH2 points that have energy greater than their coincident C3 point.
Pocket rank	The rank (1-99) of the Q-SiteFinder pocket
hse_u	Half-sphere exposure (up) the number of Cα atoms contained within the 11 Å half-sphere centred on the Cα atom pointing towards the Cß atom. Cα atoms are not double counted.
hse_d	Half-sphere exposure (down) the number of Cα atoms contained within the 11 Å half-sphere centred on the Cα atom pointing directly away from the Cß atom. Cα atoms are not double counted.
Near atoms	Number of atoms within 3 Å of each Cα atom defining the pocket.
Donor residues	Number of Cα atoms of TYR, THR, SER, ARG and TRP residues defining the binding site.
Acceptor residues	Number of Cα atoms of TYR, CYS, SER and THR residues defining the binding site.
Charged residues	Number of Cα atoms of LYS, ARG, HIS, ASP and GLU residues defining the binding site.
Volume	Volume of binding site as defined by Q-SiteFinder(Laurie and Jackson 2005).
Surface area	Count the number of 0.5 $Å^3$ cube faces not covered by an adjacent cube centred on a C3 point and multiply by $\sqrt{0.5}$.
Compactness	The ratio of the volume to surface area as originally defined by Hajduk *et al.*(Hajduk, Huth, and Fesik 2005).

Table 3.1: Summary of the learning features used in the machine learning section.

3.4 Results and Discussion

3.4.1 GRID applied to datasets

3.4.1.a GRID atom types observed in ligand dataset

We first investigate the distribution of GRID atom types present in the biologically relevant ligands dataset, then ask about the distribution of energies on a per GRID atom type basis.

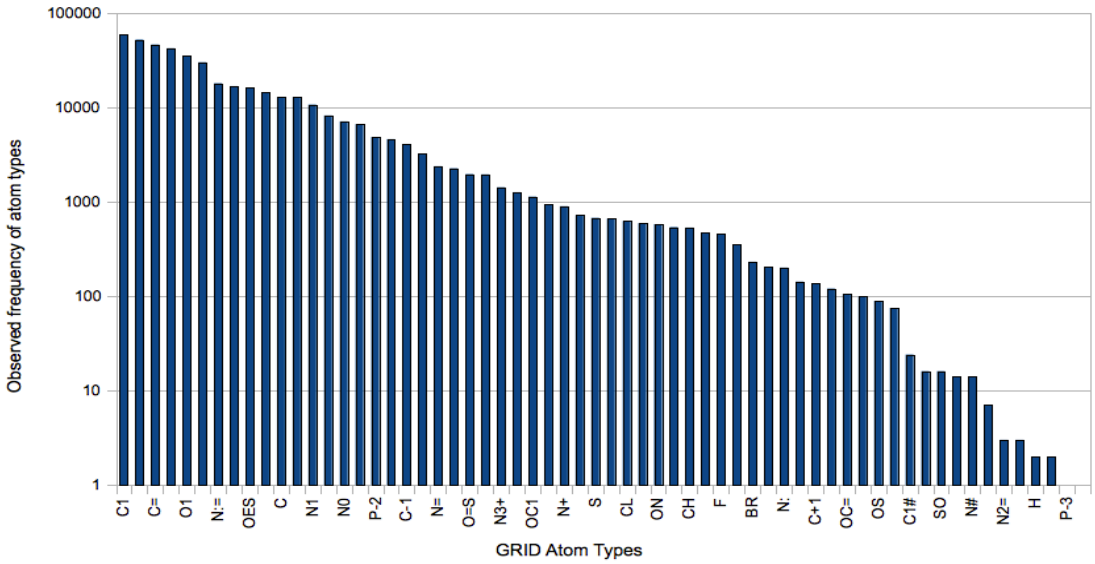

Figure 3.1: Observed frequency of GRID atom types in the biologically relevant dataset (8,861 complexes, 14,306 compounds). 425,951 total GRID atoms.

Figure 3.1 shows the observed frequency of GRID atom types from the dataset of all biologically relevant ligands in the PDB. The data is generated from 8,861 biological units and consists of a total of 14,306 compounds. In total there are 425,951 GRID atoms represented in the dataset, these comprise a total of 62 different atom types. The most populous group is C1 (an sp^3 aliphatic carbon bonded to one hydrogen) containing just short of 59000 representatives. Unsurprisingly carbon atom types are most widely represented with C2 (methylene CH2 group), C= (sp^2 carbon not bonded to hydrogen in aromatics, olefins and

esters), C1= (aromatic CH group) being next most represented, with C3 (methyl CH3 group) and C (sp^2 carbon not bonded to hydrogen in unionised carbonyl or amide) also containing more than 12000 representatives. Oxygen and nitrogen atom types are the next most populous groups followed by phosphorous, sulphur and then the halogens Cl, F and Br.

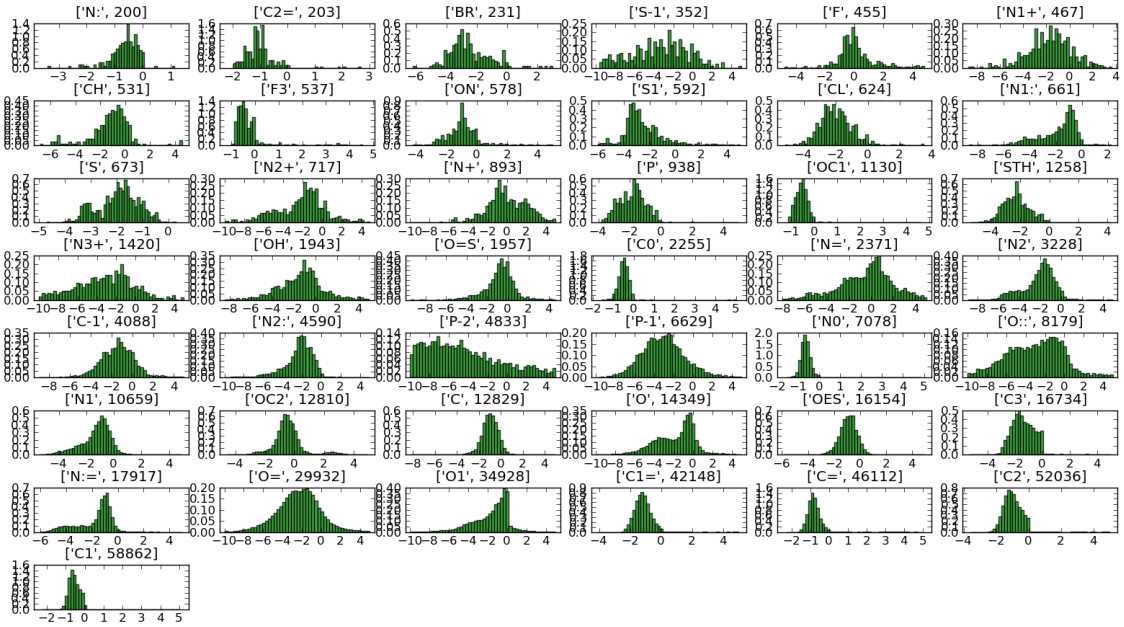

Figure 3.2: Frequency density plots showing the distribution of GRID energies for ligands observed in the PDB. Atom type and number of occurrences is shown above each histogram. All energies greater that +5 kcal mol^{-1} have been removed, and no atom types with fewer than 200 representatives are presented.

Figure 3.2 shows the distribution of energies of 43 GRID atom types for which the dataset of all biologically relevant ligands has no fewer than 200 entries. All energies greater than +5 kcal mol^{-1} have been removed. Generally carbon GRID atom types tend to form either Extreme or Gaussian like distributions with maximum values somewhere between -4 and -2 kcal mol^{-1}. This is slightly more than the -1.5 kcal mol^{-1} observed by Kuntz *et al.* when looking at per atom contributions(Kuntz *et al.* 1999) although the energies reported here cannot be directly compared with binding energies since they neglect any desolvation energy contributions. Atom types that are mediated by charge interactions tend to be

able to form more energetically favourable interactions. In many cases, such as O=, OC2, OH, P-1, OC1, F, CL, BR, N1+, the energies still form Gaussian type distributions. However many of the nitrogen atom types, N:=, N2, N1:, N=, tend to have tailed distributions skewed towards more negative energies and are more like Extreme value distributions. Also of interest is the fact that several atom types that carry a formal charge such as N+, N=, P-2 have significant numbers of atoms that register a positive energy of several kcal mol^{-1}.

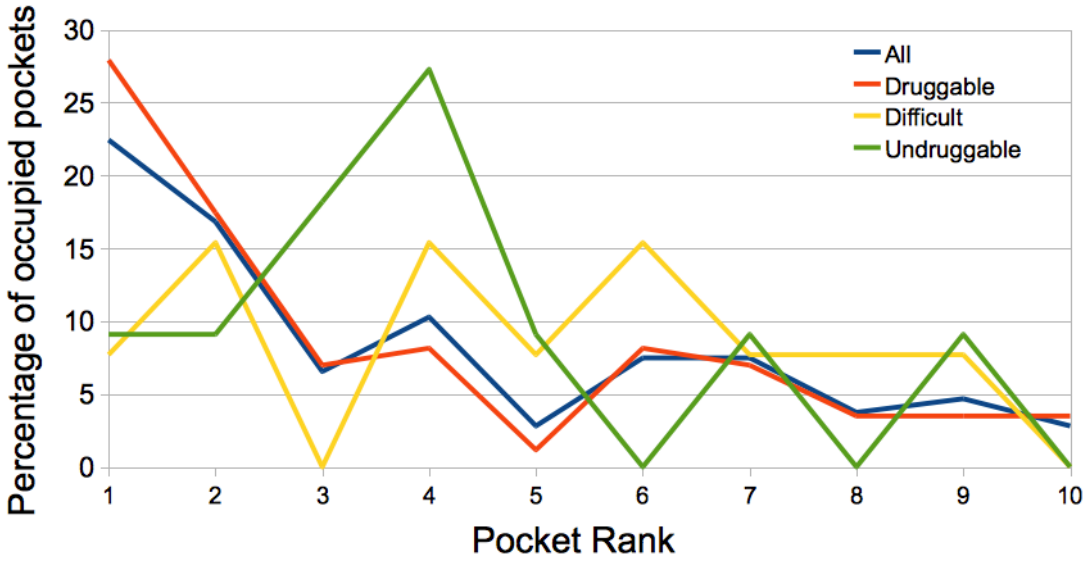

Figure 3.3: Precision of binding site predictions on the Halgren dataset. 25 % precision threshold used to define a successful prediction. Data from all 63 PDBs shown in blue, druggable subset in orange (43), difficult subset in yellow (10) and undruggable subset (10) in green.

3.4.1.b Q-SiteFinder applied to the Halgren dataset

Figure 3.3 shows that for the proteins described as 'druggable' in the Halgren dataset, there is significant enrichment of occupied pockets when looking at the top 5 ranked pockets, with the top ranked pocket showing the most enrichment. The picture is less clear when looking at the difficult and undruggable datasets, which are smaller datasets than the druggable set. These datasets are much noisier due to their relatively small size, although do still appear to show some

enrichment in the higher ranked pockets. The Halgren dataset also appears to be a more difficult challenge than datasets that have been traditionally investigated by pocket finding algorithms, as seen by the lower percentage of ligands observed in the top three pockets, compared to the GOLD docking benchmark of 134 proteins used in the original Q-SiteFinder study(Laurie and Jackson 2005). However, this might be anticipated since the GOLD benchmark contains only validated binding pockets with known inhibitors that are generally to be high affinity binders.

3.4.1.c Distribution of GRID energies

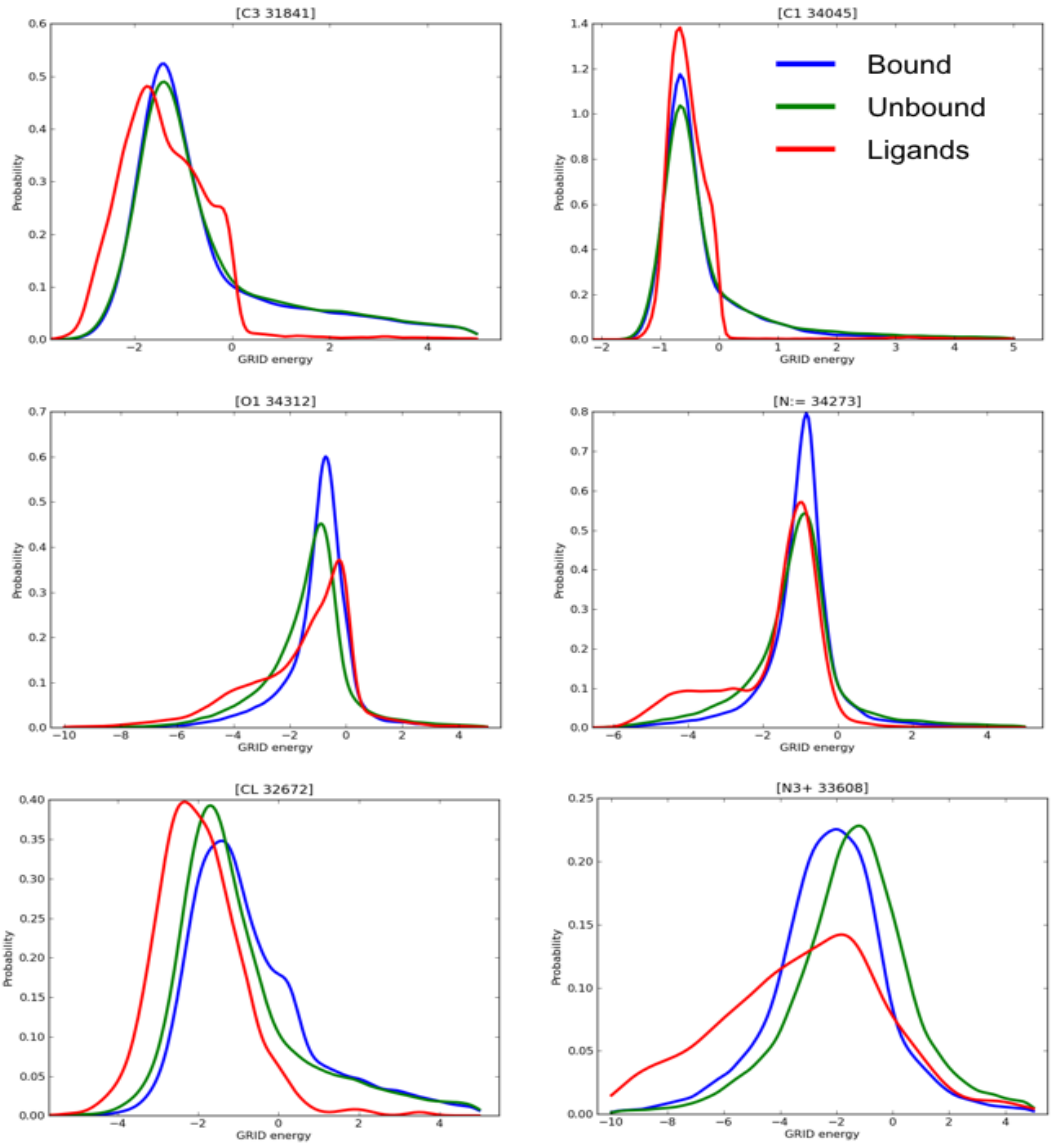

Figure 3.4: Shows distributions of GRID energies for six different atom types C3, C1, O1, N:=, CL, N3+. The red line shows the distribution of energies observed in biologically relevant ligands, whilst the blue line shows the energies observed in Q-SiteFinder pockets from the Halgren dataset that are designated as bound (precision > 25 %) whilst the green line shows the distribution of the GRID atoms in the Q-SiteFinder pockets designated as unbound.

Figure 3.4 shows the distribution of energies of GRID probes observed in pockets defined by Q-SiteFinder on proteins in the Halgren dataset (Bound and Unbound pockets) compared to the distribution of energies of GRID probes observed in ligands from the biologically relevant dataset (Ligands). The probes from the

Q-SiteFinder pockets are further split into a bound and unbound subset to allow comparison. The C3 and C1 probes both behave as van der Waals spheres with Gaussian like distribution of energies. It is noticeable that in both cases the distribution of energies for bound and unbound pockets is almost identical. Furthermore in the case of C1, the probes observed in ligands have an energy distribution almost identical to those of the bound/unbound pockets identified by Q-SiteFinder. The C3 probe type has an energy distribution that deviates from that of the Q-SiteFinder pockets, although not by a large amount, this is due to the fact that the GRID spacing is 0.5 Å rather than the 0.9 Å used for Q-SiteFinder and the fact that some additional GRID points are contained as the calculation is performed on a box placed over the Q-SiteFinder site. Whilst this has some limitations, the effect is that the right hand tail includes more positive values than expected since some of the box is outside the Q-SiteFinder site and therefore not necessarily a favourable probe site. However, the left hand tail is not significantly affected. As a result it is not expected that the difference between the bound and unbound pockets should change significantly. The initial hypothesis that the bound pockets energy distribution might be expected look more like the ligand distribution, whilst the unbound pockets would exist somewhere to the right allowing for discrimination between the two classes is not fulfilled.

The O1 and N:= probes also show few differences in distributions between bound and unbound pockets. Again there is little difference between the pocket energy distribution and that observed for the biologically relevant ligands. In the case of C3, O1 and N:= there is some tendency for slightly more negative energies to be favoured, however a majority of ligand atoms have GRID energies comparable to those observed in probes from Q-SiteFinder pockets. It would appear in most cases that there is no discernable difference between the distributions of the bound and unbound pockets of the Halgren dataset.

The largest differences between bound and unbound pockets and ligand atoms is observed in the CL and N3+ atom types. In the case of CL we see that the mean energy of ligand atoms is shifted to more negative energies. Additionally we see that the energy distribution for unbound pockets is slightly more negative than that of bound pockets. In the case of N3+ the converse is true of unbound and bound pockets, with bound pockets tending to have more negative energies than unbound pockets. Additionally we notice that N3+ atoms in ligands have a larger negative tail than either of the bound pockets or unbound pockets.

The fact that the majority of probe types do not appear to clearly distinguish between bound and unbound sites is explainable retrospectively. A Q-SiteFinder pocket demarcates a region of space for which van der Waals energy is favourable. However, a ligand may only require a region of the size several GRID points to generate a favourable interaction between an atom and the protein. Whilst the Z-score method does in several cases show regions of favourable interactions from a probe type that is similar to the atom in the cognate ligand, it is still difficult to distinguish these regions from regions of favourable interaction elsewhere on the protein.

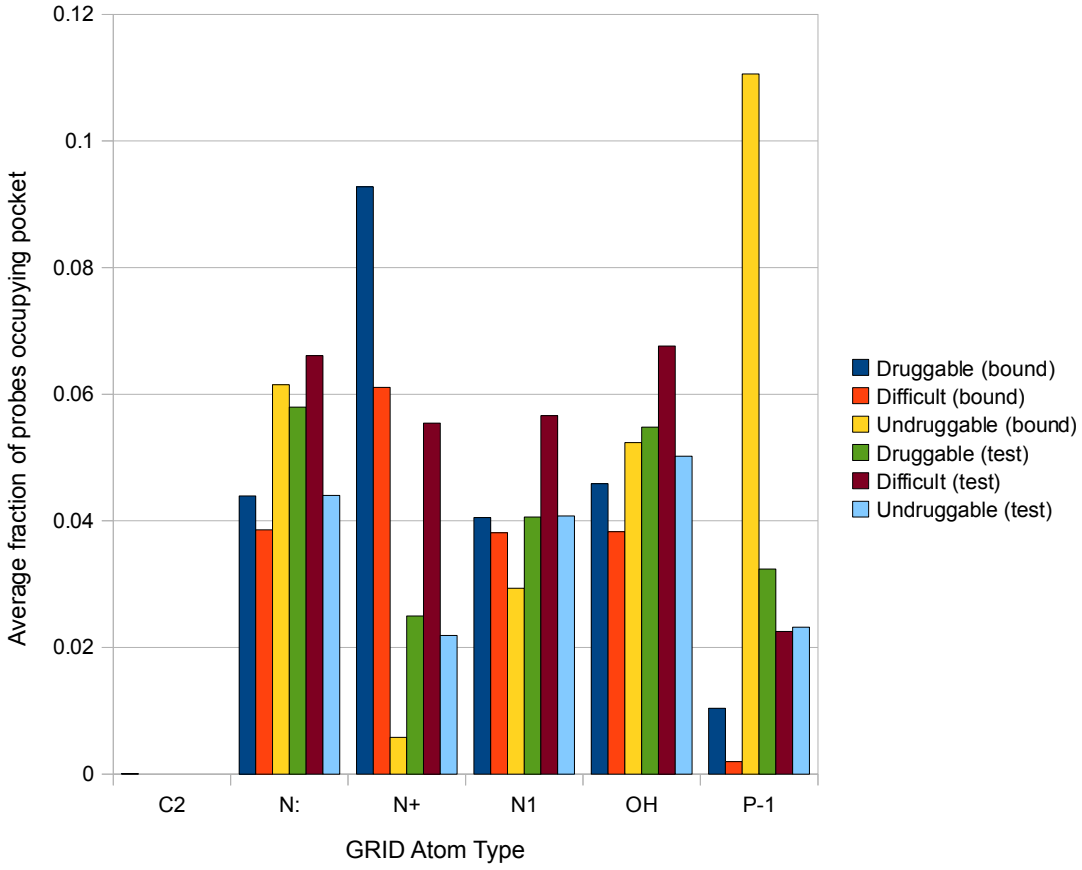

Figure 3.5: Fraction of probes selected by highest z-score for each probe position comprising a Q-SiteFinder site for druggable bound sites (dark blue) and druggable test sites (green), difficult bound sites (orange) and difficult test sites (brown), undruggable bound sites (yellow) and undruggable test sites (light blue).

3.4.1.d Regions of Favourable GRID energy

We investigated the fraction of probes that comprise Q-SiteFinder sites for their propensity to favour one of six probe types. We used the C2, N:, N+, N1, OH and P-1 probe types as they are relatively common and represent a selection of charge neutral, positive, negative and hydrogen bond donor and acceptor properties. The mean and standard deviation of each of the probe types was calculated from the ligand distributions similar to those shown in figure 3.4. From this z-scores were calculated for each probe position and the probe with the largest z-score (z-score > 1.7) at that position is retained. Figure 3.5 shows the mean fraction of

probes from a Q-SiteFinder site that are selected as having the largest z-score greater than 1.7. It is clear from figure 3.5 that the C2 probe very rarely fulfils this criteria, which is not altogether surprising given that it can only interact through van der Waals interaction which are likely to be favourable anyway given that the probe comprises a Q-SiteFinder site. Also of note is that the P-1 probe is rarely observed as the highest scoring probe in bound pockets in the druggable or difficult dataset, although it is observed as having the highest z-score for 11 % of the probes in typical bound undruggable sites. The P-1 probe is observed between 2-3 % for each of the unbound Q-SiteFinder pockets. The N+ probe is rarely observed in the undruggable bound pockets, whilst being observed in 9 % and 6 % of probes for druggable and difficult pockets respectively. The remaining N:, N1 and OH probes tend to have percentage site occupancies that don't vary by more than a couple of percent from a mean value of 5 %.

Use of the six probes (C2, N:, N+, N1, OH and P-1) enables the use of a reduced number of probes that cover some of the common properties of ligand-protein interactions. Namely: C2 (van der Waals); N: (H-bond acceptor); N+ (positive charge); N1 (H-bond donor); OH (H-bond donor/acceptor); P-1 (negative charge). It appears that the only strong discriminator here is charge, whereby positive charge is observed more in druggable/difficult sites, whereas negative charge is more often observed in undruggable sites. The use of charge as a discriminator between binding sites has been used before in work by Hajduk *et al.*, however, they observed that generally binding sites carry little formal charge(Hajduk, Huth, and Fesik 2005). It is not clear that this is a genuine distinguishing feature of druggable/undruggable binding sites. Certainly the undruggable set is very small and redundant with just four classes of protein (HIV integrase, ICE1, PTPB1 and Cathepsin K) so addition of just a few sites carrying net formal charge could bias the results heavily.

3.4.2 Drugability indices

Figure 3.6 shows a variety of pocket descriptors defined from the coordinates of the pockets. Half sphere exposure is a simple measure of solvation potential that is designed to take into account solvent accessible surface area dependent properties(Hamelryck 2005). In figure 3.6a we see that the number of atoms contained in the solvent facing half sphere (hse_u) is greater for druggable bound pockets (128) than it is for both difficult and undruggable bound pockets (70 and 65 respectively). A similar picture is observed for the protein facing half sphere (hse_d), 140 for druggable bound pockets, compared to around 100 for all other classes except undruggable unbound pockets (58).

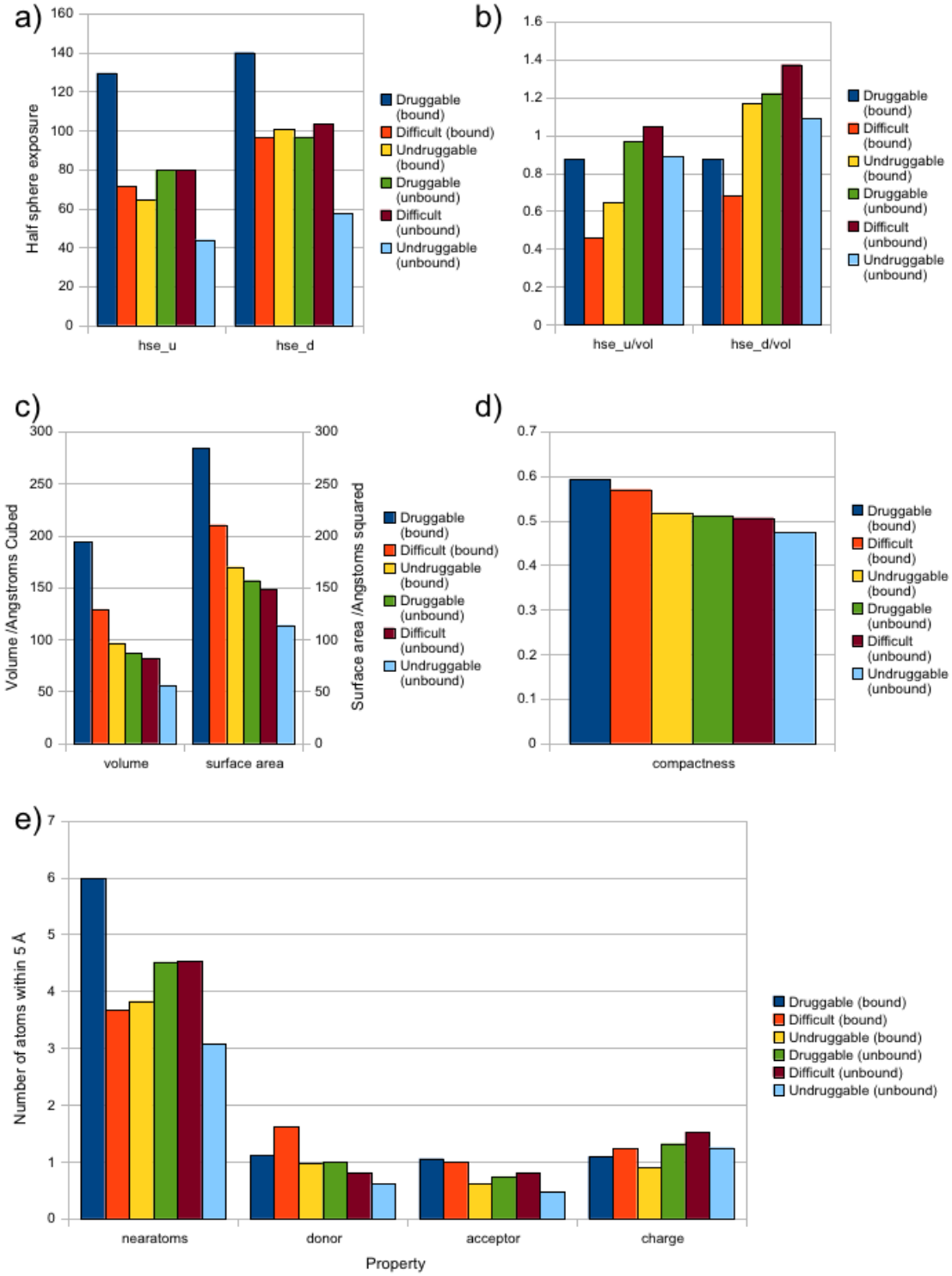

Figure 3.6: Pocket descriptor properties generated using Q-SiteFinder to define a pocket: a) half sphere exposure; b) half sphere exposure/volume; c) volume and surface area of pocket; d) pocket compactness (volume/surface area); e) number of near atoms, donor residues, acceptor residues and charged residues.

In figure 3.6b we use a measure of half sphere exposure except that we normalise for the volume of the site chosen. Normalizing by site volume somewhat alters the observed measure of half sphere exposure. We now see that the solvent facing half sphere remains larger (0.87) in the case of druggable bound sites, whereas the score for difficult and undruggable bound sites drops to around 0.4 and 0.65 respectively. The score for unbound sites is this time closer to 1 in all cases. When looking at the score for the protein facing half sphere (hse_d) we observe that the undruggable bound pockets have a score of close to 1.2 which is similar to that of unbound druggable pockets. This time druggable bound pockets score 0.87, whilst difficult bound pockets score 0.67.

We next look at the volume and surface area of pockets in figure 3.6c. A clear decreasing trend is seen when moving through druggable, difficult and undruggable, this is extended when looking between bound pockets and unbound pockets in the case of both volume and surface area. Average volume of druggable bound sites is 190 $Å^3$ with surface area being 270 $Å^2$. Difficult sites are smaller (125 $Å^3$) with correspondingly smaller surface area (210 $Å^2$). Undruggable sites are the smallest of all with volume 95 $Å^3$ and surface area 165 $Å^2$. It is noticeable that the volume of druggable and difficult unbound pockets are similar in volume to those of undruggable bound pockets, this lends some weight to the idea of using the bound/unbound classification as a proxy for the druggable/undruggable classification.

A related concept to that of volume and surface area is that of compactness which measures the ratio between the two values(Hajduk, Huth, and Fesik 2005). Figure 3.6d shows that druggable bound pockets have a compactness score of close to 0.6 with difficult sites being close in value with 0.57. Undruggable bound sites score just over 0.5, which is comparable to the scores of unbound druggable pockets and unbound difficult pockets.

Finally we investigate the average number of atoms or residues of a certain type within a radius of the points comprising the pocket in figure 3.6e. For the bound pockets we observe an average of 6, 3.65 and 3.8 atoms within a radius of 5 Å of pocket points for druggable, difficult and undruggable respectively. In the case of unbound pockets the averages are 4.4, 4.4 and 3.05 respectively. We observe that the average number of donor residues is close to one for druggable and undruggable bound pockets whilst slightly higher at one and a half for difficult bound pockets. When looking at average number of acceptor residues we observe an average of close to one for druggable and difficult bound pockets, and 0.6 for undruggable bound pockets. The average of the charged residue count varies slightly around close to one.

Several of the measures that we investigated are likely to be highly correlated, which is one reason for the choice of the random forest as a machine learning method applied to the above data, which is discussed further shortly. Examples of correlated variables are the volume, surface area and compactness, since compactness is calculated from these two measures alone it will also be correlated. The number of near atoms might also be expected to show correlation to the volume and surface area of sites, although this is not necessarily the case, as it can be seen that the number of atoms for difficult and undruggable bound sites is lower than that of unbound sites from druggable and difficult sites. This might indicate that the degree to which a ligand might be able to bury itself in the protein surface is being identified by this measure.

In the cases investigated here we notice that half-sphere exposure and volume/surface area appear to be the two measures that are most likely to distinguish between bound/unbound or druggable/undruggable. However, in most cases there doesn't seem to be a strong differential between bound/unbound pockets, meaning that it may be difficult to develop a good predictor.

We note that the paper by Halgren appeared to make good use of descriptors based on pocket size, enclosure and hydrophilicity. As previously discussed we already have several measures of pocket size in volume and surface area. Additionally we have information on the total number of C3 probes comprising a pocket, and the total energy of these probes. We have a measure of enclosure implicitly within the Q-SiteFinder energy score since more deeply buried pockets will tend to have a larger energy score, and additionally we have included the number of near atoms and the protein facing half-sphere exposure (hse_d). Furthermore we have included a hydrophilicity measure that places an OH2 probe at all probe positions and calculates the interaction energy with the protein. If the probe has energy more favourable than that of the C3 probe at that point it is retained and a measure of total OH2 energy and total number of OH2 probes in the pockets is counted. The measurements shown in figure 7 and the C3 probe count, total C3 probe energy, OH2 probe count and total OH2 probe energy along with the rank of the Q-SiteFinder site are then included as predictor variables for a machine learning technique. Since there are a large number of potentially correlated predictor variables and a small number of true positives compared to a large number of true negatives a random forest machine learning is determined to be a suitable methodology to apply.

3.4.3 Random forests to identify druggable pockets

3.4.3.a All learning features

		Training set prediction				Test set prediction			
		Bound	Unbound	Class Error	MCC	Bound	Unbound	Class Error	MCC
Observed Set 1 (101,202)	Bound	286	19	7.59 %	0.892	94	7	6.93 %	0.848
	Unbound	35	575	6.37 %		14	188	6.93 %	
Observed Set 2 (99,204)	Bound	282	20	6.67 %	0.870	95	4	4.04 %	0.884
	Unbound	38	572	6.23 %		12	192	5.88 %	
Observed Set 3 (102,204)	Bound	282	20	6.62 %	0.859	100	2	1.96 %	0.935
	Unbound	36	574	5.90 %		7	197	3.43 %	
Observed Set 4 (102,204)	Bound	283	19	6.29 %	0.864	96	6	5.88 %	0.844
	Unbound	43	570	6.56 %		16	188	7.84 %	

Table 3.2: Results from four balanced training sets, with resulting forests applied to test sets generated from the remaining data.

Table 3.2 shows results from a four-fold cross-validation of using a random forest to classify pockets as bound or unbound using the predictor variables: number of C3 points; sum of C3 point energies; number of OH2 points; sum of OH2 point energies; half-sphere exposure (hse_u); half-sphere exposure (hse_d); pocket volume; pocket surface area; pocket compactness; number of near atoms; number of donor residues; number of acceptor residues; number of charged residues; Q-SiteFinder rank of pocket. We generally observe classification error of between 5.9 % and 7.6 % in the training sets. Prediction of bound and unbound pockets in the test set are generally successful with prediction errors generally remaining low and comparable in magnitude to those observed in the training data. Use of the random forest allows us to recover details of which predictor variables are most important for making predictions. In all cases pocket rank is the most important variable, followed by total C3 energy and number of C3 points. Volume and surface area are also often important variables.

We use the Matthews Correlation Coefficient (MCC) to compare the performance of the random forest under four-fold cross-validation. Use of the MCC also measures the quality of prediction on our test sets. The MCC is a measure of the success of classifying binary variables. MCC varies between -1 (perfect classification into the opposite classes), 0 (random classifier) and +1 (perfect classification). Results for the MCC in each of the test cases are presented in Table 3.2. In our training sets we observe MCCs of between 0.859 and 0.892, whilst in our test sets we score between 0.844 and 0.935. We aim to classify the Halgren dataset into bound and unbound using our previously trained random forests. If a reasonable level of success is achieved it will then be possible to investigate whether bound sites incorrectly classified as unbound tend to be enriched with sites that are labelled by Halgren as difficult of undruggable.

		Halgren dataset predictions			
		Bound	Unbound	Class Error	MCC
Observed RF 1	Bound	84	133	61.3 %	0.209
	Unbound	433	5488	7.31 %	
Observed RF 2	Bound	85	132	60.8 %	0.214
	Unbound	426	5495	7.19 %	
Observed RF 3	Bound	85	132	60.8 %	0.200
	Unbound	476	5445	8.74 %	
Observed RF 4	Bound	85	132	60.8 %	0.206
	Unbound	454	5467	7.67 %	

Table 3.3: Results from Halgren dataset predictions using each of the four forests. 217 bound pockets and 5921 unbound pockets.

Table 3.3 shows the results from applying each of the four forests previously trained to the Halgren datasets. Here we observe that error rates for unbound sites remain between 7.19 % and 8.74 % which is very close to those observed in

Table 3.2. Error rates for classifying bound sites are far larger than previously observed, being consistent at 60-61 %, which is ten times the error rate previously observed in the cross validation set.

3.4.3.b Discarding pocket rank from learning features

		Training set prediction				Test set prediction			
		Bound	Unbound	Class Error	MCC	Bound	Unbound	Class Error	MCC
Observed Set 1 (101,202)	Bound	261	42	13.9 %	0.800	84	11	10.9 %	0.790
	Unbound	39	573	6.37 %		17	191	8.42 %	
Observed Set 2 (99,204)	Bound	261	44	14.4 %	0.800	78	10	10.1 %	0.763
	Unbound	37	573	6.06 %		21	194	10.3 %	
Observed Set 3 (102,204)	Bound	263	39	12.9 %	0.807	91	11	10.8 %	0.766
	Unbound	39	571	6.40 %		22	182	10.8 %	
Observed Set 4 (102,204)	Bound	256	46	15.2 %	0.772	92	9	8.8 %	0.860
	Unbound	46	564	7.54 %		10	195	4.90 %	

Table 3.4: Results from four balanced training sets after disregarding pocket rank as a predictor variable, with resulting forests applied to test sets generated from the remaining data.

Results from generating new random forest predictors on the same training/test set as previously without using the pocket rank predictor are shown in Table 3.4. MCC calculations for each of the cross-validations are also presented. It is clear that after removing pocket rank as a predictor variable the class error for predicting bound pockets has increased to between 12.9 % and 15.2 % from between 5.9 % and 7.6 %. Class error for predicting unbound pockets remains at between 6.1 % and 7.5 % from between 6.2 % and 6.6 %. With increasing class error we also observe decreased MCC values, with training results of between 0.77 and 0.80 for training data, and comparably 0.76 and 0.86 for test data.

When the random forest is applied to the Halgren dataset (Table 3.5) we notice that the performance is even lower than previously, with MCC scores of between 0.14 and 0.16. Once again we notice that whilst class error observed when

predicting unbound sites remains comparatively low (10.3 % - 11.7 %), class error for predicting bound sites is 61.8 % to 64.5 %, several times that observed in the training data.

		Halgren dataset predictions			
		Bound	Unbound	Class Error	MCC
Observed RF 1	Bound	77	140	64.5 %	0.141
	Unbound	647	5274	10.9 %	
Observed RF 2	Bound	81	136	62.7 %	0.156
	Unbound	620	5301	10.5 %	
Observed RF 3	Bound	83	134	61.8 %	0.148
	Unbound	691	5230	11.7 %	
Observed RF 4	Bound	81	136	62.7 %	0.158
	Unbound	609	5312	10.3 %	

Table 3.5: Results from Halgren dataset predictions using each of the four random forests trained after disregarding pocket rank as a predictor variable. 217 bound pockets and 5921 unbound pockets.

3.4.3.c Improving learning with unbalanced data

Currently applying random forest learning methods to our binding dataset has initially appeared promising, although when the methods are applied to the Halgren dataset it becomes clear that the Halgren dataset must have features that are considerably different to our binding dataset. We also note that although random forests have traditionally been thought to perform well when applied to unbalanced datasets, we note that our training appears to always perform well at predicting unbound pockets, but struggles at predicting bound sites when applied in a new context. Thus we need to consider methods to better balance our data.

Using the same datasets as above we first apply an unbiased random forest to our set of predictors after pocket rank has been excluded. We have decided to exclude pocket rank from the analysis as whilst it has clearly worked as a strong learning feature, we are worried that for difficult/undruggable pockets it may not be a reliable measure when transferred across datasets. Pocket rank may not be scale invariant, so the same pocket on a different protein may have a vastly different rank, whereas a quantitative measure such as pocket volume might distinguish pockets in a certain range or interest.

		Training set prediction			
		Bound	Unbound	Class Error	MCC
Observed (404,24295)	Bound	404	0	0 %	0.353
	Unbound	2510	21785	10.3 %	

Table 3.6: Results from four balanced training sets after disregarding pocket rank as a predictor variable, with resulting unbiased forests applied to test sets generated from the remaining data.

		Halgren dataset predictions			
		Bound	Unbound	Class Error	MCC
Observed	Bound	102	115	53.0 %	0.153
	Unbound	938	4983	15.8 %	

Table 3.7: Results from Halgren dataset predictions using each of the four unbiased random forests trained after disregarding pocket rank as a predictor variable. 217 bound pockets and 5921 unbound pockets.

The unbiased random forest performs extremely well when trained with class error for unbound pockets around 10 % which is slightly more than previously observed, but class error for bound pockets is 0 % with all predictions correct. However, when the method is transferred to pockets in the Halgren dataset, the high class error for bound pockets (53.0 %) once again becomes noticeable.

Furthermore the class error for unbound pockets increases. This results in an MCC of 0.153 compared to values in the range 0.14 and 0.15 in the case of the standard random forest.

A final test of the standard random forest was applied to the datasets, whereby pockets of volume less than 150 $Å^3$ are removed from the Halgren dataset.

There are 83 bound pockets contained in this dataset. From each of the four random forests 631 out of 5905 pockets are predicted to be bound, with 402 of these predictions common across all forests. Druggable pockets account for 66 of these pockets, with difficult pockets accounting for 9 and undruggable 8 (results not shown). When each of the random forests are applied to the Halgren dataset the vast majority of these pockets are correctly predicted although there are a large number of false negative predictions. Additionally the hypothesis that undruggable sites might be predicted as unbound sites does not appear to be the case as these sites are rarely classified as unbound.

The major issue with this method appears to be two-fold. The predictor variables contain a lot of noise, and secondly the random forests appear to be struggling with a heavily biased dataset. Therefore we proceed by using a set of unbound sites from the top 5 binding sites. This will create a ratio of bound:unbound of somewhere in the region 1:5 rather than the previous 1:50. In many ways this should be a more straightforward prediction task although it will mean that successful predictors like volume will be less predictive since top ranked sites all tend to have large volumes.

		Training set prediction				Test set prediction			
		Bound	Unbound	Class Error	MCC	Bound	Unbound	Class Error	MCC
Observed Set 1 (101,202)	Bound	154	97	38.6 %	0.472	62	39	38.6 %	0.435
	Unbound	77	430	15.2 %		37	167	18.3 %	
Observed Set 2 (99,204)	Bound	190	115	37.7 %	0.459	26	21	21.2 %	0.435
	Unbound	103	507	16.9 %		14	87	6.86 %	
Observed Set 3 (102,204)	Bound	154	96	38.4 %	0.459	69	33	32.4 %	0.471
	Unbound	83	426	16.3 %		40	162	19.6 %	
Observed Set 4 (102,204)	Bound	151	99	39.6 %	0.411	68	34	33.3 %	0.524
	Unbound	98	409	19.3 %		30	174	14.7 %	

Table 3.8: Results from four training sets using all bound Q-SiteFinder sites and only the top five unbound Q-SiteFinder sites, with resulting forests applied to test sets generated from the remaining data.

		Halgren dataset predictions			
		Bound	Unbound	Class Error	MCC
Observed RF 1	Bound	188	29	13.4 %	0.668
	Unbound	49	200	19.7 %	
Observed RF 2	Bound	184	33	15.2 %	0.678
	Unbound	42	207	16.9 %	
Observed RF 3	Bound	191	26	12.0 %	0.609
	Unbound	68	181	27.3 %	
Observed RF 4	Bound	188	29	13.4 %	0.672
	Unbound	48	201	19.3 %	

Table 3.9: Results from Halgren dataset predictions using each of the four random forests trained. 217 bound pockets and 249 unbound pockets.

3.4.4 Identifying druggable pockets

Compared to using the top ranked Q-SiteFinder site which identifies 24 out of 217 bound pockets as bound pockets if choosing the top ranked site or 48 out of 217 if choosing the top three sites (Table 3.6), the machine learning method performs

very well, identifying between 184, 188, 188 and 191 out of 217 as bound sites, with a relatively low error rate. Pockets predicted as unbound in all four forests, but bound in the difficult dataset are 1A4G, 1NNC, 1QMF, 2QWK with 1KTS identified by one forest and 1NLJ, 1BMQ, 1NNY and 1ONZ from the undruggable dataset with 1Q1M identified by one forest, which identifies Neuraminidase, Penicillin Binding Protein and Thrombin as potentially undruggable. This distinguishes the compounds Cathepsin K, ICE1 and PTP1B from HIV1-integrase as potentially undruggable. The pockets from the druggable set of 1DMP, 1H07, 1H08, 1HVR, 1HW8, 1HWR, 1KE8, 1KE9, 1KV1, 1M17, 1QBS and 1RTH are all predicted as unbound when they are actually bound pockets for each of the random forests.

There are three Neuraminidase structures whose pockets were labelled as unbound consistently. Neuraminidase drugs tend to be administered as prodrugs. Oseltamivir uses a protected carboxylic acid to form three salt bridges with the protein(Cheng *et al.* 2007). The penicillin binding protein drugs such as the ß-lactam inhibitor Clavulanic acid acts by forming a covalent disulphide bond with the protein(Poirel *et al.* 2005). This indicates that these binding modes may not be well described by the method in its current form. Thrombin has been targeted by peptidic inhibitors such as Hirudin (and derivatives), and several prodrugs such as Ximelagatran and Dabigatran. However, it has also been targeted by the small molecule Argatroban, thus the classification as a difficult target may not be truly justified.

3.5 Conclusion

The model that we originally developed has many similar measures to those described by Halgren. The three key parameters used by Halgren were size of pocket (capped at a maximum value), enclosure of pocket and hydrophilicity of pocket. Our measure also contained parameters related to the size (volume, number of C3 points, total C3 energy), furthermore we believe that use of the C3

probe also gives some measure of the enclosure of the pocket, as high energy values are more probable for a C3 probe surrounded by many nearby atoms. The C3 probe does not directly give a measure of hydrophilicity, which is why the OH2 probe was used. The data presented in the paper by Halgren shows much promise for not only increasing the quality of predictions of bound pockets, but also for stating whether they are likely to accommodate high-affinity ligands. Our model does not appear to show similar success. Although there are several key points to be considered that may explain this discrepancy.

A physical model such as that of Cheng *et al.* has several benefits over a classification into somewhat arbitrary classes. Whilst the concept of drugability will always suffer from the issue that it will never be possible to genuinely know the maximal affinity of a pocket. Presenting a numerical value for the maximal affinity is a more easily testable result, since it is possible to look at the best affinity compounds for a given protein and determine whether there are any compounds that have significantly better affinity than the stated maximal affinity. By contrast classification into druggable or undruggable classes suffers from the major problem that the distinction can be somewhat arbitrary as perhaps illustrated by the distinctions made in the paper by Cheng *et al.*(Cheng *et al.* 2007). Halgren uses two models that are individually parameterised, one to determine whether a pocket is bound or unbound, the second to determine whether a bound pocket is druggable or undruggable. We conjecture that it might be more reasonable to assume that an individual measure should be applied to determine whether a pocket is druggable or undruggable, since any pocket that is unbound should immediately be descriminated as undruggable. That is if you could identify a pocket on the protein surface and show that there are no ligands that bind selectively and with any reasonable affinity to that site, it is by definition undruggable. Using a two tiered system means that it may be possible to determine a site as unbound but druggable! Clearly the method that we employ does not meet our own criteria in that it only classifies as bound or unbound,

which is then used as a proxy for druggable or undruggable based on 'mis-classifications'. It does however, improve prediction of whether a site is bound or unbound, and identify several compounds as potentially correctly classified as undruggable when using this classification by proxy system. It is not clear whether the difficult classification is helpful since in cases like Thrombin Cheng *et al.* detemined that the maximal affinity for Thrombin may be well within the range that would make it suitable for a drug target, and indeed Argatroban is a small molecule known to target thrombin(Cheng *et al.* 2007).

In reality it appears that the unbiased random forest performs very well given the difficulty of the problem. Comparison of the results for the prediction task to those of the simplified task of identifying the bound pockets from the top 5 pockets rather than the top 99 pockets, shows a similar quality of results. In some respects this is likely to be due to the fact that the problem is more straightforward as there are fewer negative results to deal with. Conversely the pockets in the smaller set are more likely to be similar in character, due to their similar size and energy. The simpler prediction task is more likely to be a test representative of a task that a researcher may want to carry out, as many of the smaller, lower ranked pockets would be likely to be immediately discarded. One of the main problems with smaller pockets, is that several nearby small pockets may link to form a larger pseudo pocket. The current Q-SiteFinder algorithm does not take this into account. Previous work by Bridgett in a masters thesis develops a smoothing algorithm that allows smaller pockets to be represented as a more diffuse entity that can expand to encompass a larger volume. Spacing of GRID atoms in pockets might also link to the maximal affinity observed by Kuntz *et al.*(Kuntz *et al.* 1999), would distances closer to the length of C-C bonds be more appropriate, since pocket energies would then scale with number of grid points which would be similar to total number of atoms that could fit in a pocket.

Halgren uses a distance criteria to combine several smaller pockets to create a larger pocket. The main issue with this method is that as the size of pockets is increased, the probability of making a successful prediction increases. Therefore precision and coverage scores must be used to balance this improved success rate that comes from choosing a larger site, with choosing sites that more accurately describe a potential binding site. With regards to a maximal affinity model, Halgrens use of limited additional score after a certain size does marry to work of Kuntz *et al.*(Kuntz *et al.* 1999).

Research in the field of pocket detection lacks standard tests of a methods success. Fields such as ligand docking, protein-protein docking or protein structural prediction have standard test sets and in some cases blind prediction competitions that allow methods to be critically assessed. Development of a standard test would allow methods to be compared on a more even footing, and furthermore would perhaps enhance the understanding of the limitations of such tests, paralleling perhaps datasets such as DUD(Huang, Shoichet, and Irwin 2006).

In our model we show that when a very difficult test of the method is performed the success rate is low. When a more reasonable test on a dataset with fewer negative results is performed the performance improves. However, this still highlights an important shortcoming of the machine learning method. If the method can identify high-affinity sites from non-binding sites accurately the number of negative binding sites shouldn't significantly adversely affect the method.

Predicting the maximal affinity of a pocket is in general a difficult problem. It is hampered by the diverse nature of pockets. There is large variation in the size and shape of pockets on a protein surface. The simple physical models that we

applied do not appear to adequately describe the properties of the pockets. Successful representation of the effect of the presence of water in the pockets is likely to go some way to improve predictions. An improved model might do well to use methods to compare pockets on the protein surface to those of known ligand binding pockets much in the same way as functional annotation methods sometimes attempt. Furthermore design of datasets to test predictions is also hampered by addition of unspecified statistical bias.

3.6 References

Anon. 2009. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Vienna, Austria. http://www.r-project.org.

Benson, Mark L, Richard D Smith, Liegi Hu, Michael G Lerner, and Heather A Carlson. 2005. Binding MOAD (Mother Of All Databases). *Proteins* 60, no. 3 (August): 333-40. doi:10.1002/prot.20512. http://www.ncbi.nlm.nih.gov/pubmed/15971202.

Berman, Helen M., Tammy Battistuz, T. N. Bhat, Wolfgang F. Bluhm, Philip E. Bourne, Kyle Burkhardt, Zukang Feng, *et al.* 2002. The Protein Data Bank. *Acta Crystallographica Section D Biological Crystallography* 58, no. 6 (May): 899-907. doi:10.1107/S0907444902003451. http://scripts.iucr.org/cgi-bin/paper?S0907444902003451.

Chen, C, Andy Liaw, and Leo Breiman. 2004. Using Random Forest to Learn Imbalanced Data. In *Berkeley, Department of Statistics, University of California.*

Cheng, Alan C, Ryan G Coleman, Kathleen T Smyth, Qing Cao, Patricia Soulard, Daniel R Caffrey, Anna C Salzberg, and ES Huang. 2007. Structure-based maximal affinity model predicts small-molecule druggability. *Nature Biotechnology* 25, no. 1: 71-5. doi:10.1038/nbt1273. http://www.ncbi.nlm.nih.gov/pubmed/17211405.

Cock, Peter J A, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, *et al.* 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)* 25, no. 11 (June): 1422-3. doi:10.1093/bioinformatics/btp163. http://www.ncbi.nlm.nih.gov/pubmed/19304878.

Fusaro, Vincent A, D R Mani, Jill P Mesirov, and Steven a Carr. 2009. Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nature Biotechnology* 27, no. 2 (February): 190-8. doi:10.1038/nbt.1524. http://www.ncbi.nlm.nih.gov/pubmed/19169245.

Goodford, P J. 1985. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of Medicinal Chemistry* 28, no. 7: 849-57. http://www.ncbi.nlm.nih.gov/pubmed/3892003.

Hajduk, Philip J, Jeffrey R Huth, and Stephen W Fesik. 2005. Druggability indices for protein targets derived from NMR-based screening data. *Journal of Medicinal Chemistry* 48, no. 7 (April): 2518-25. doi:10.1021/jm049131r. http://www.ncbi.nlm.nih.gov/pubmed/15801841.

Halgren, Thomas A. 2009. Identifying and characterizing binding sites and assessing druggability. *Journal of Chemical Information and Modeling* 49, no. 2 (February): 377-89. http://www.ncbi.nlm.nih.gov/pubmed/19434839.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software. *ACM SIGKDD Explorations Newsletter* 11, no. 1 (November): 10. doi:10.1145/1656274.1656278. http://portal.acm.org/citation.cfm?doid=1656274.1656278.

Hamelryck, Thomas. 2005. An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. *Proteins* 59, no. 1 (April): 38-48. doi:10.1002/prot.20379. http://www.ncbi.nlm.nih.gov/pubmed/15688434.

Huang, N, Brian K Shoichet, and John J Irwin. 2006. Benchmarking sets for molecular docking. *Journal of Medicinal Chemistry* 49, no. 23 (November): 6789-801. doi:10.1021/jm0608356. http://www.ncbi.nlm.nih.gov/pubmed/17154509.

Jackson, Richard M. 2002. Q-fit: a probabilistic method for docking molecular fragments by sampling low energy conformational space. *Journal of Computer-aided Molecular Design* 16, no. 1 (January): 43-57. http://www.ncbi.nlm.nih.gov/pubmed/12197665.

Kuntz, I D, K Chen, K A Sharp, and P A Kollman. 1999. The maximal affinity of ligands. *Proceedings of the National Academy of Sciences of the United States of America* 96, no. 18 (August): 9997-10002. http://www.ncbi.nlm.nih.gov/pubmed/10468550.

Laurie, Alasdair T R, and Richard M Jackson. 2005. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics (Oxford, England)* 21, no. 9 (February): 1908-16. doi:10.1093/bioinformatics/bti315. http://www.ncbi.nlm.nih.gov/pubmed/15701681.

Laurie, Alasdair T R, and Jackson Richard M.. 2006. Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening. *Current Protein & Peptide Science* 7, no. 5 (October): 395-406. http://www.ncbi.nlm.nih.gov/pubmed/17073692.

Lee, B, and F M Richards. 1971. The interpretation of protein structures: estimation of static accessibility. *Journal of Molecular Biology* 55, no. 3 (February): 379-400. http://www.ncbi.nlm.nih.gov/pubmed/5551392.

Poirel, Laurent, Laura Brinas, Annemie Verlinde, Louis Ide, and Patrice Nordmann. 2005. BEL-1, a novel clavulanic acid-inhibited extended-spectrum beta-lactamase, and the class 1 integron In120 in Pseudomonas aeruginosa. *Antimicrobial Agents and Chemotherapy* 49, no. 9 (September): 3743-8. doi:10.1128/AAC.49.9.3743-3748.2005. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1195426&tool=pmcentrez&rendertype=abstract.

Strömbergsson, Helena, and Gerard J Kleywegt. 2009. A chemogenomics view on protein-ligand spaces. *BMC Bioinformatics* 10 Suppl 6, no. Suppl 6: S13. doi:10.1186/1471-2105-10-S6-S13. http://www.ncbi.nlm.nih.gov/pubmed/19534738.

Sugaya, Nobuyoshi, and Kazuyoshi Ikeda. 2009. Assessing the druggability of protein-protein interactions by a supervised machine-learning method. *BMC Bioinformatics* 10: 263. doi:10.1186/1471-2105-10-263. http://www.ncbi.nlm.nih.gov/pubmed/19703312.

4 Docking to identify likely conformations of novel oligoamide compounds designed to mimic helical peptides and bind in pockets on the hDM2 protein

4.1 Abstract

The design of novel α-helix mimetic inhibitors of protein-protein interactions is of interest, since an optimal scaffold that can present side-chains at a geometry analogous to that of an α-helix could be tuned to give inhibitors of high specificity and affinity. Oligoamide compounds are of specific interest since they have been shown to be synthetically accessible through a series of simple solution phase reactions allowing inclusion of a variety of side-chains to a rigid backbone with geometry similar to that of an α-helix. The hDM2-p53 interaction is a suitable model system since there is considerable structural information detailing both the wild-type p53 peptide interaction as well as several hDM2 inhibitor compounds.

Currently two structures of hDM2 bound to p53 helices, two of hDM2 bound to designed inhibitors and one NMR structure of apo hDM2 exist in the public domain, but none of hDM2 bound to novel oligoamide compounds. We use structure-based computational methods such as shape-matching and molecular docking to generate putative models for the interaction between these oligoamide compounds and hDM2 with the aim of competitively inhibiting a helix from the p53 protein. Additionally, we perform RESP charge fitting to parameterize the oligoamide compounds for further computational study using molecular dynamics simulations with the AMBER force field. Here we show that there are two putative

classes of binding modes for oligoamide compounds: the first in which the oligoamide compound lies parallel to the observed p53 helix; the second in which the oligoamide compound lies anti-parallel to the p53 helix. The side-chains from the best docking/shape matching conformations explore the same binding pockets as the p53 helix. The results presented here will allow us to perform further molecular dynamics studies on a variety of the best scoring complexes to better assess their potential for binding hDM2 and inhibiting the p53 interaction. Furthermore, the methodology can be applied in the study of any oligoamide compound designed to target the hDM2-p53 interaction.

4.2 Introduction

The interaction between the E3 ubiquitin ligase hDM2 and a helical peptide that forms part of the p53 tumour suppressor domain is of great interest as a target for protein-protein interaction inhibitor drug discovery(Dickens, Fitzgerald, and PM Fischer 2009). Several drugs have been developed that are in clinical trials, additionally the system is both well studied from a biochemical perspective(Bond *et al.* 2008), and importantly for this study there is a wealth of structural data on the system(Kussie *et al.* 1996),(Grasberger *et al.* 2005),(Vassilev *et al.* 2004).

4.2.1 The p53 pathway

The p53 pathway is complex and further work has to be done to improve our understanding of the mechanisms involved. However, many aspects of the p53 pathway are well understood. In particular the p53 pathway has been shown to modulate response to cellular stress(Bond *et al.* 2008). We are particularly interested in its involvement in the modulation of apoptosis, in combination with its negative regulator hDM2. This part of the p53 pathway has been studied in detail due in part to interest in targeting and disrupting the hDM2-p53 protein-protein interaction with the aim of inducing apoptosis in cancerous cells(Dickens, Fitzgerald, and PM Fischer 2009).

4.2.2 hDM2 p53 protein structures

The hDM2 protein structure was first solved in complex with a 15mer wild-type p53 peptide (SQET**F**SDL**W**KL**L**PEN) by Kussie *et al.*(Kussie *et al.* 1996). This structure is shown in figure 4.17. Grasberger and colleagues later determined the structure of a p53 related helix that had been optimised to bind hDM2 with higher affinity than the wild-type helix(Grasberger *et al.* 2005). The 9mer high affinity peptide (R**F**MDY**W**EG**L**) retains the key binding residues: Phe; Trp; Leu, that target the deep hydrophobic pocket present on the hDM2 surface. The wild-type helix is 15 residues long and has a calculated binding affinity (K_d) of 600 nM(Kussie *et al.* 1996). It has been shown that in general shorter helices will bind more tightly(Böttger *et al.* 1997). It appears that the optimized helix gains some of its affinity by lining the solvent exposed face of the helix (the face opposite the Phe-Trp-Leu residues) with charged residues such as Glutamine and Arginine. Experimental observation of the helix propensity for solvent exposed residues in the middle positions of **α**-helices suggests that these charged residues are generally favoured in solvent exposed helices(Pace and Scholtz 1998).

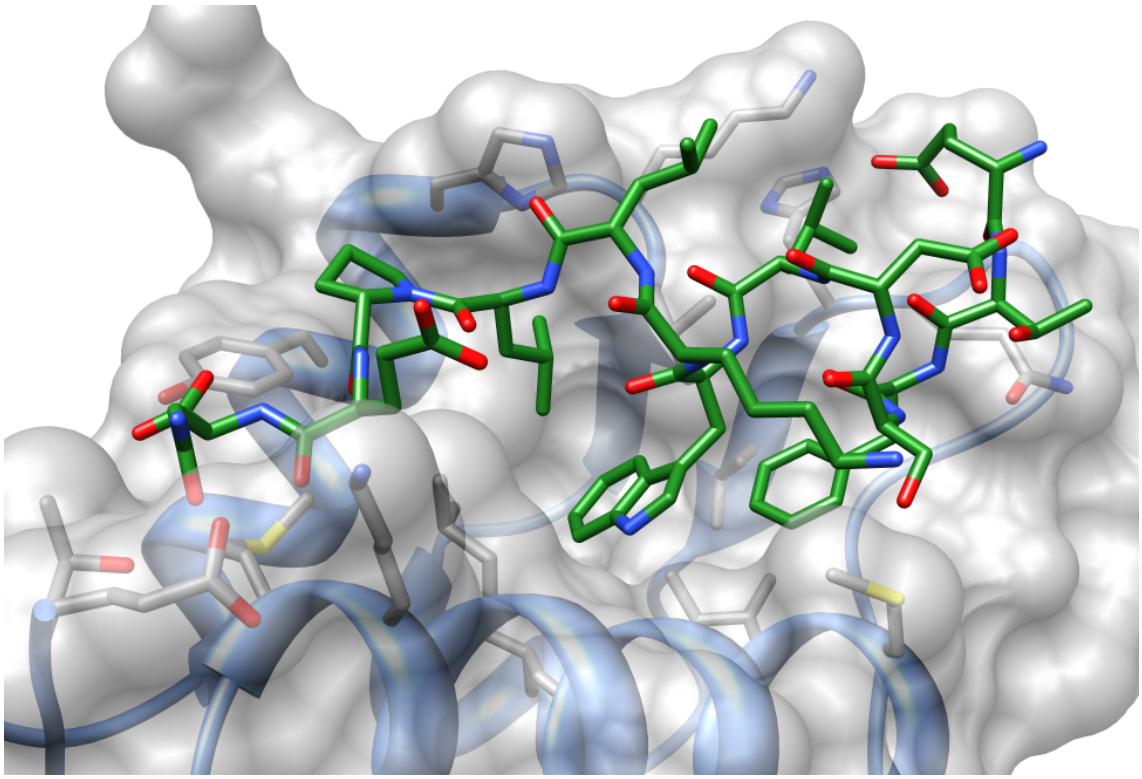

Figure 4.1: The wild-type p53 peptide is shown with green carbon, blue nitrogen and red oxygen atoms. The SASA of hDM2 is shown in transparent grey, with blue cartoon representation of the protein backbone. Contacting residues are shown in stick representation with standard atom colours and grey for carbon atoms.

In 2004, Fry *et al.* published the NMR structure of *Xenopus Laevis* hDM2 bound to a small-molecule inhibitor(Fry *et al.* 2004) which is very similar to the Nutlin compounds described in the work by Vassilev *et al.*(Vassilev *et al.* 2004). This work published the structure of hDM2 in complex with an *cis*-imidazoline compound to 2.3 Å resolution. The authors screened a diverse range of compounds identifying the *cis*-imidazoline compounds as promising lead compounds. The *cis*-imidazoline compound for which they published a structure of is known as nutlin-2 and was determined to have an IC_{50} of 0.14 μM using a surface plasmon resonance solution competition assay. An improved IC_{50} of 0.09 μM was determined from an enantiomer of their nutlin-3 compound(Vassilev *et al.* 2004).

An NMR structure detailing the structure of unbound hDM2 has also been published(McInnes *et al.* 2005). The authors of the paper detailing the NMR structure of the apo protein note the rearrangement of the region of the p53 binding site. Whilst this does not directly affect the future directions of this study it is important to note the flexibility of the site when choosing suitable structures for investigation. It also raises important questions as to whether it might be possible to design a protein-protein interaction inhibitor given only the apo structure of protein binding partners. Eyrisch and Helms previously investigated three protein-protein interactions including the hDM2-p53 interaction using molecular dynamics simulations and pocket detection(Eyrisch and Helms 2007). They concluded that pockets binding ligands were observed opening and closing on the time-scale of their simulations and in some cases on picosecond time-scales(Eyrisch and Helms 2007). Investigation into conformational changes undergone by hDM2 and a structurally related protein MDMX have been undertaken by Carotti *et al.* they used 60 ns MD simulations to investigate structural changes undergone in the bound and unbound forms of each protein(Carotti *et al.* 2009).

4.2.3 Designing hDM2 inhibitors

Much work has been undertaken in designing inhibitors for the hDM2 interaction, as discussed previously the initial inhibitors of the hDM2 interaction for which there are structures, were identified using High Throughput Screening (HTS) methods. Wang and co-workers identified spiro-oxyindole based scaffolds, using the GOLD docking program, that could inhibit the hDM2-p53 interaction with micromolar affinities(Nikolovska-Coleska *et al.* 2005). This method appears to be quite successful, although as with any structure-based design method, it is reliant on structures of the target protein being available.

Many groups have focussed on designing peptides to target the hDM2-p53 interaction, using both experimental and computational techniques to approach the problem. Peptide libraries developed using phage display technology have

been employed as an ideal experimental technique(V Böttger *et al.* 1996). Massova and Kollman studied the hDM2-p53 interaction using a technique that they developed and named computational alanine scanning(Massova and Peter A. Kollman 1999). They used an MMPBSA model of the free energy change required to mutate a side-chain to alanine. This allowed them to identify key residues that contribute to the binding energy, and suggest mutations. Similar work was performed by Kortemme and Baker, who produced a simple model using a simplified theoretical approach to the MMPBSA method(Kortemme and Baker 2002). They identified residues on both the p53 peptide and the hDM2 binding site that contribute significantly to binding energy. More recently studies have been undertaken by several other groups using similar techniques(Zhong, & Carlson 2005),(Moreira, Fernandes, & Ramos 2008),(Kalid and Ben-Tal 2009).

β-peptides are synthetically produced from β amino acids which have their amino group bonded to the β carbon instead of the α carbon. They have advantages over traditional optimized peptides since they are stable in the cell. Michel and co-workers took a previously discovered β-peptide and applied a de-novo design strategy to identify 50 candidate side-chain replacements from 10,000 structures with aromatic and non-aromatic heterocycles substituted(Michel *et al.* 2009). Binding free energies were then calculated with MC/FEP calculations for the peptides using the OPLS/AA force field and TIP4P water. A selection of 8 of the most synthetically accessible compounds were re-evaluated using a second more accurate round of MC/FEP. The study revealed novel β-peptides with affinity improved from 204 nM in the starting compound to 27.6 nM in the case of the β-peptide with the best affinity for hDM2(Michel *et al.* 2009). This study was also notable in that it also produced high-affinity β-peptides that target the related hDMX interaction with differing levels of specificity between hDM2 and hDMX(Michel *et al.* 2009). One of the disadvantages of working with β-peptides is that altering the side-chains can significantly affect the entropic cost of creating

secondary structure elements such as helices. This means that whilst it may be desirable to add a side-chain to gain interactions with the protein target, this may introduce an unfavourable entropic cost that will lower overall binding affinity.

4.2.4 Properties of oligoamide compounds

4.2.4.a Synthesis of oligoamide compounds

Previously we have discussed β-peptides, which are part of a class of molecules known as foldamers. Foldamer compounds take inspiration from nature, where polymers such as proteins and RNA can have well defined secondary and tertiary structure(Gellman 1998),(Hill *et al.* 2001). The aromatic oligoamide compounds that we are investigating are synthetically accessible using methods similar to those used in synthesis of peptides and can be designed to adopt a rod-like conformation which can present side-chains at locations similar to those at the i, i+4, i+7 locations on an α-helix(Plante *et al.* 2008). The work by Warriner *et al.* shows that X-ray structures of these oligoamide compounds show an intramolecular hydrogen bond between the amide NH and ether oxygen. This is mirrored in solution NMR of the oligoamide compound in deuterated DMSO and $CDCl_3$. 2D ^1H-^1H NOESY spectra indicate that there is free rotation about the ArCO bond, whilst the intramolecular hydrogen bonding described above restricts rotation about the ArNH bond(Plante *et al.* 2008). Further work has identified that these compounds can act as low μM inhibitors of the hDM2-p53 interaction(Plante *et al.* 2009), this results from this work are backed by elegant synthetic work that provides access to several further types of related compound(Shaginian *et al.* 2009). In the work by Plante *et al.* six trimer oligoamides were synthesised and screened against the hDM2-p53 interaction using a fluorescence polarization assay.

4.2.4.b Parameters for oligoamide compounds

In addition to synthetic work on designing oligoamide compounds there has been progress in the use of quantum mechanics and molecular dynamics to study the properties of arylamide compounds, many of which bear striking similarity to those that we intend to study. A quantum mechanics study of the torsional profile of arylamide compounds calculated the location of minima and the heights of barriers to rotation away from these minima(Vemparala *et al.* 2006). Work has also been undertaken that uses molecular dynamics to investigate the behaviour of arylamide compounds designed to mimic heparin in solution(Pophristic *et al.* 2006).

In the above works the authors use compounds with thioether bonds instead of the ether bond found in the work by Plante *et al.*(Plante *et al.* 2009), however the compounds investigated have similar ArNH bonds and ArCO bonds which prove useful since the GAFF force field(Wang *et al.* 2004) does not have parameters that agree with the crystallographic and NMR data presented by Plante *et al.*(Plante *et al.* 2009). Vemparala and co-workers note that altering the thioether to an ether group is one way in which the flexibility of the compound could be controlled. For the purposes of our investigation we use the torsional parameters presented in the work of Vemparala *et al.*(Vemparala *et al.* 2006), since we are predominantly interested in the correct location of minima in the torsions. This should allow us to well sample the thermodynamic properties of the system, even if our observation of the kinetic properties of this bond are slightly incorrect. Additional investigation of foldamer systems with arylamide bonds have been investigated, in particular the predicted response to different solvent environments has been studied by molecular dynamics and compared to experimental NMR studies(Galan *et al.* 2009).

4.2.5 Study aims

The aim was to use computational methods to guide synthesis of novel oligoamide compounds that can inhibit the hDM2-p53 interaction with high affinity. As we have seen above there are several approaches that could be applied to this problem. In particular we aimed to perform alchemical free energy calculations in order to determine the relative binding affinity of a series of oligoamide compounds and indeed this is described in the next chapters. In order to carry out alchemical free energy calculations knowledge of the 3D structure of the hDM2-oligoamide complex is necessary, this is the primary aim of this chapter. Once reasonable starting conformations have been generated that will enable further study a second aim is to generate parameters for the oligoamide complex. Given that torsional parameters have previously been identified a set of charge parameters that accurately describe the oligoamide molecules in a molecular mechanics force field is also developed.

4.3 Methods

4.3.1 Structural Superposition

Structural superposition of proteins was performed using UCSF Chimera version 1.4 on the Mac OS X operating system using the MatchMaker function with default settings(Pettersen *et al.* 2004). hDM2 chains (1Z1M-model 9, 1YCR-chain A, 1T4F-chain M, 1RV1-chain A, 1T4E-chain B) were superposed using the MatchMaker algorithm, whilst the bound ligands where present were subjected to the same rotation and translation as its partner protein. This means that the ligand is retained in the same position relative to partner protein, and all ligands can be compared in their common binding site.

4.3.2 Electrostatic surfaces

Electrostatic surfaces were calculated using DelPhi V. 4 Release 1.1(Rocchia, Alexov, and Honig 2001), with computations carried out through the DelPhi controller module of UCSF Chimera(Pettersen *et al.* 2004). An interior dielectric of 2.0, and an exterior dielectric of 80.0 and Debye-Huckel boundary conditions were used in the calculation. Results were visualised using UCSF Chimera version 1.4 on the Mac OS X operating system.

4.3.3 Hydrophobic surfaces

Hydrophobic surfaces were generated using the hydrophobic surface preset from UCSF Chimera version 1.4 on the Mac OS X operating system(Pettersen *et al.* 2004). Residues are coloured according to the Kyte-Doolittle scale, with blue showing the most hydrophillic residues, white showing a value of 0.0 and orange showing the most hydrophobic residues(Kyte and Doolittle 1982).

4.3.4 FTMap

FTMAP identifies the likely binding location of several small organic probe molecules on the surface of a protein. The authors first validated the method on elastase, for which the locations of 8 organics solvents has already been identified experimentally, followed by using the method to identify the location and 'trace out' the structure of aliskiren, the first approved renin inhibitor. FTMAP was accessed online from: http://ftmap.bu.edu/ and default settings were used(Brenke *et al.* 2009).

4.3.5 Docking

Two rounds of docking using Autodock(Morris *et al.* 1998),(Seeliger and de Groot 2010) were performed. The first round used Autodock 4.0 to perform 2.5 million evaluations for 27,000 generations with population size 300 using the compound detailed in Figure 4.11 to produce 101 compounds. The results from this set of

dockings were clustered at a 2 Å RMS cutoff. The lowest energy representative structures of these clusters were used in the initial MD simulations and are representatives from the largest low energy clusters labelled: clu1; clu2; clu3. The second round of docking calculations were performed using Autodock 4.2.1 using a Lamarkian genetic algorithm. 150 docked conformations were generated, with each using 25 million evaluations for 27,000 generations of population size 300. Random number seeds were generated from the Autodock PID and the current system time. The protein structure used was derived from the structure of hDM2 bound to a high-affinity p53 helix, with all water molecules removed, protonation states manually assigned and the high-affinity p53 helix removed (1T4F-chain M). A grid centred on 13.119, 18.969, 10.941 was used with spacing of 0.375 Å and 52, 58 and 48 points in the x, y and z directions.

In order to make a comparison the docking program FRED (OpenEye) was also used to generate 150 docked poses with default settings. FRED is a rigid body docking program, conformations for the oligoamide compounds were generated using the OMEGA conformational generator also supplied by OpenEye. OMEGA used an energy window of 25 to generate a maximum of 1 million conformers (maxconfgen), of which a maximum of 10000 with RMS less than 0.5 Å were kept (maxconfs).

4.3.6 Geometric matching

A geometric hashing algorithm was used to superpose atoms from oligoamide compounds that had been generated by OMEGA. The method has been described previously(Brakoulias and Jackson 2004), but is described here briefly. Triplets of atoms that by definition form a triangle are generated for the database molecule (p53 peptide) and query molecule (oligoamide compound). All possible pairs of triplets where each pair consists of a triplet from both query and database molecules are compared. All triplets with the same atom at each triangle vertex and similar distances between vertices are treated as a match. The resulting

triplets then define a rotation and translation matrix which will map the query molecule onto the database molecule. At this point the number of coincident atoms can be determined, and the transformation which provides the largest number of coincident atoms is treated as the best match.

4.3.7 Charge calculations

Charge calculations were performed to determine which method for charge calculation would be most appropriate for the hDM2-oligoamide system. We compared AM1 BCC semi-empirical calculations to Hartree-Fock calculations using the HF 6-31G* basis set.

4.3.7.a Generating conformers for AM1 BCC calculations

Conformers were generated for AM1 BCC calculations using OpenEye OMEGA. Parameters were selected based on those most likely to produce ligand conformations that are bioactive, thus an energy window of 25 kcal mol^{-1} was used(Kirchmair *et al.* 2006), RMS tolerance and maximum number of generated conformers was set so as to generate a wide range of conformers such that approximately 350 conformers in total were generated. This meant a value between 0.45 Å and 0.55 Å was used for the RMS cutoff and maxconfgen was set to 10,000. This resulted in: 310 Phe-Trp-Leu; 361 Phe-Nap-Leu; 380 Val-Phe-Propyl; 361 CH3-CH3-CH3, conformations.

4.3.7.b AM1 BCC calculations

Semi-empirical AM1 BCC charge calculations were performed for each of the conformers generated; using OMEGA as described in the above section; with the Antechamber program supplied with AMBER 8(35). Calculation of charge for each conformer took of the order several minutes. The mean and variance for the charge of each atom were then calculated using a custom script for the R statistical computing language(Anon 2009).

4.3.7.c Quantum calculations

Quantum calculations were performed using the Hartree-Fock level of theory and the HF 6-31G* basis set. Initial conformations of compounds were generated in Gaussview. The REDIII.1 software(Dupradeau *et al.* 2008) was used in tandem with Gaussian 03 to perform the calculations(M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Pet 2004), which consisted of a geometry optimization followed by Molecular Electrostatic Potential calculation followed by charge fitting using the RESP method. This scheme was chosen since it most closely resembles the method that was originally employed for deriving charge parameters for the AMBER force field. Calculations were performed on full oligoamide compounds with Phe-Trp-Leu substitution pattern, CH3-CH3-CH3 substitution pattern. Full molecule calculations took of the order one week to complete when carried out using a 2.2 GHz Opteron processor. Fragment compounds containing the central benzene ring, carboxylic acid, primary amine and a single substitution of CH3 or Trp were also investigated, these calculations took approximately 2 days to complete using a single processor.

4.4 Results and Discussion

The results in this chapter are presented in two distinct sections. The first deals with generation of hDM2-oligoamide complexes that are suitable for further study by molecular dynamics and might provide insight into oligoamide binding. The second section is concerned with identification and generation of parameters for the oligoamide compounds of interest.

4.4.1 Generation of hDM2 oligoamide complexes

Key to any study aimed at calculating the free energy of association of a protein-ligand complex is an accurate structure for the protein-ligand complex. Ideally this would come from X-ray or NMR structures. In the case of hDM2 we have already seen that there is an NMR structure of the free protein in addition to high-resolution X-ray structures of the protein bound to a wild-type p53 helix, a high-affinity p53 helix, a benzodiazepinedione compound and Nutlin-2. The former two are peptides whilst the latter two are small-molecules specifically designed to target this interaction. Unfortunately there are no published crystal structures of the oligoamide compounds. Therefore we must use knowledge of the behaviour of oligoamide compounds from published literature to assist in the development of a model of the bound structure in order that we can proceed with the free energy calculations. To this end we have used two molecular docking programs in addition to using an alternative superposition based method. Our key assumption in the creation of our docking model is that since the four compounds for which we have high resolution structures available all bind to the same site, then this site is where we expect the oligoamide compounds to bind. Furthermore, since the oligoamide compounds have been designed to mimic the side-chains present at positions i, i+4, i+7 on an α-helix then we expect to find oligoamide substituents bound at these sites. With this in mind we proceeded with comparing results from Autodock and FRED. Both docking programs are quite different in the way that they work and additionally use different scoring functions to rank their solutions. Since there are no structures of immediately similar compounds to the oligoamides in which we are interested, and there is no data about binding affinities of a wide range of these compounds, we are limited in the way that we can assess the quality of the results produced by these programs, and as such we must be guided by comparison to the structural data that we do have.

4.4.1.a Analysis of hDM2 binding site properties

Structural superposition of different hDM2 protein-ligand complexes is a simple way to investigate the binding site. The first thing noticed is that the two reported inhibitors target the same regions of the binding pocket as the high-affinity p53 peptide mimicking the interaction of Phe-Trp-Leu side-chains from the p53 peptide. This is interesting since both series of inhibitors were discovered through independent high-throughput screens, although they both have scaffolds that allow the presentation of their key functional groups in very similar spatial locations to the high-affinity peptide.

It can be seen in figure 4.2b that the nutlin-2 compound closely mimics the binding epitope of the high-affinity p53 helix. The two chlorophenyl groups target the Leu and Trp pockets, whilst the ethyl ether moiety binds in the Phe pocket. The crystal structure of hDM2 bound to a high-affinity helix previously described was reported at the same time as a 2.6 Å structure of hDM2 bound to a benzodiazepinedione compound(Grasberger *et al.* 2005). Once again the benzodiazepinedione compound targets the same Phe-Trp-Leu binding epitope as the p53 peptides which can be seen in figure 4.2c. The authors noted that the inhibitor interacts with the hDM2 binding pocket through non-specific van der Waals contacts. If we plan to target the hDM2 binding pocket with oligoamide based helix mimetics, it will be reasonable to assume that high-affinity compounds should also target these same regions of space.

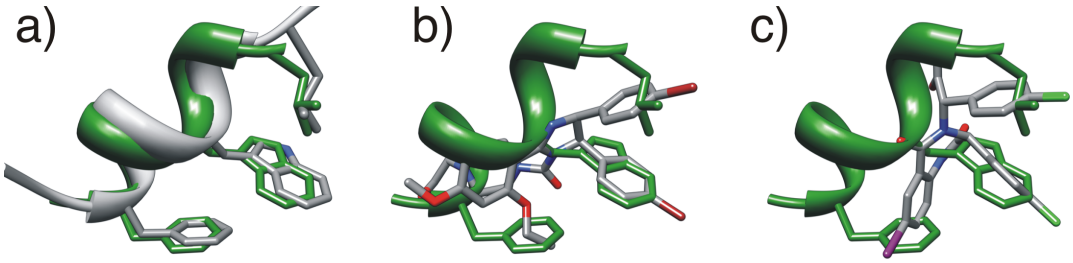

Figure 4.2: Representations of high affinity helix (green) shown relative to: a) wild type helix; b) Nutlin-2; c) Benzodiazepinedione compound. Figures were generated using the matchmaker function from Chimera to superpose hDM2 from pdb code 1T4F to pdb codes: a) 1YCR; b) 1RV1; c) 1T4E.

It is important to note that there is no hydrogen bonding of the peptides or compounds to the hDM2 binding pocket. Many successful drug compounds can gain a large amount of binding affinity by picking up hydrogen bonds to the target protein. This can also be a useful way to gain specificity for a particular member of a family of proteins.

We can see in figure 4.3 that the pocket is particularly hydrophobic, as is mentioned by Grasberger *et al.*(Grasberger *et al.* 2005). Additionally we can see from figure 4.3 that the pocket doesn't carry a strong electrostatic charge. Hydrophobic pockets are often difficult to develop compounds that bind with a high degree of specificity. Additional binding affinity can sometimes be gained by the use of halogenated functional groups, such as the chlorophenyl rings seen in Benzodiazepine compounds(Hernandes *et al.* 2010). The main issue with these elements is the decrease in solubility that is observed both experimentally and through QM/MM studies(Baum *et al.* 2009). Additionally, halogenated drug compounds can often show undesirable ADMET properties such as accumulation in fat tissue.

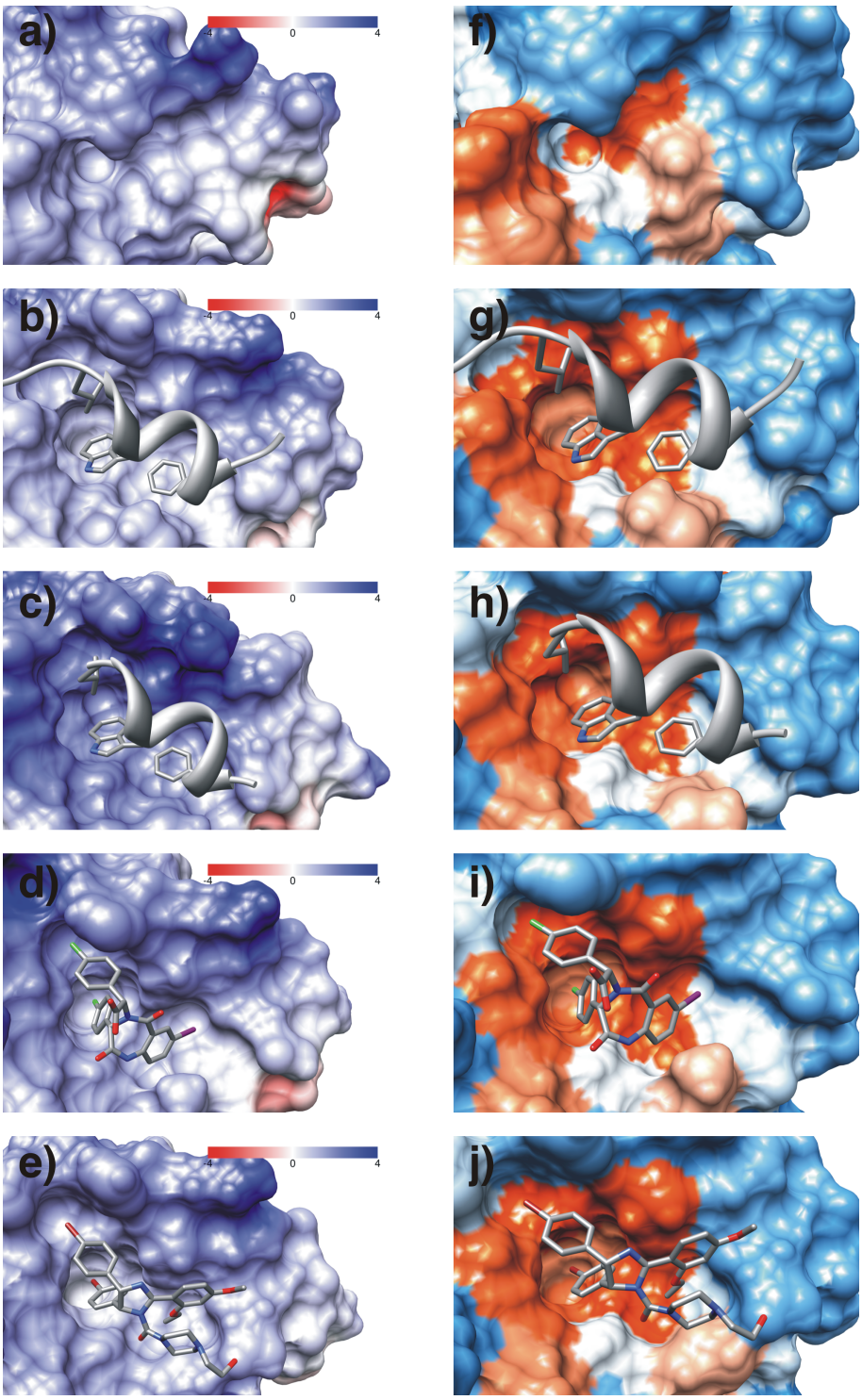

Figure 4.3: The hDM2 binding pocket shown with electrostatic surfaces (red - negative charge, blue - positive charge) a)-e) and hydrophobic surfaces (blue - hydrophillic, white - no preference, orange - hydrophobic) f)-j). a/f) hDM2 apo (1Z1M); b/g) hDM2 wild type p53 (1YCR); c/h) hDM2 high affinity p53 (1T4F); d/i) hDM2 Benzodiazepinedione (1T4E); e/j) hDM2 Nutin-2 (1RV1). Images produced using Chimera, electrostatic surfaces calculated using Delphi.

4.4.1.b Binding site properties of hDM2

We ran an analysis of the hDM2 binding site using Q-SiteFinder. We saw that Q-SiteFinder identified binding pockets that overlapped the key functional groups of the nutlin and benzodiazepine compounds. We then investigated the binding site using FTMap. In Figure 4.4 we show the predicted location of benzene fragments with respect to the high-affinity p53 helix and the benzodiazepine compound respectively. In Figure 4.4a it can be seen that the method predicts that benzene (yellow) is favoured at those locations where cyclic side-chains are observed in hDM2-helix interaction. In Figure 4.4b once again the method identifies benzene to be favoured at locations similar to those where the two halogenated benzene functional groups are observed. Successful identification of favoured benzene rings at this location on the protein is a positive outcome for several reasons. Most obviously it further supports our hypothesis that oligoamide compounds are likely to bind at this location on the protein. Secondly it identifies FTMap as a promising tool that might be useful in identifying the types of functional group suited for substitution onto oligoamides designed for higher affinity to the hDM2 protein.

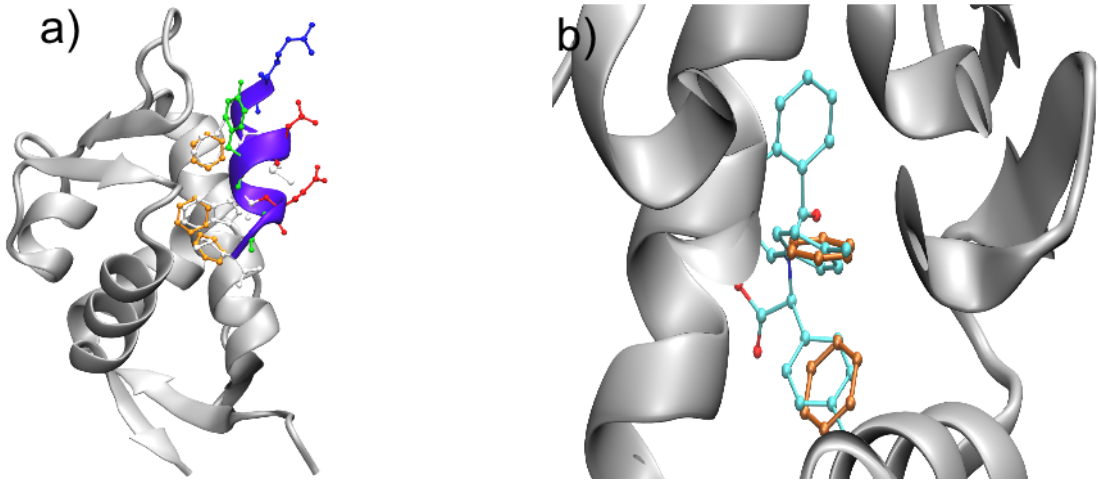

Figure 4.4: FTMap results with hDM2 represented as grey cartoon model with: a) p53 helix(purple) showing the predicted location of benzene rings (yellow) using the FT-Map algorithm; b) hDM2(grey)-benzodiazepinedione compound (cyan/red) showing the predicted location of benzene rings (yellow) using the FT-Map algorithm(Brenke *et al.* 2009).

4.4.1.c Autodock

An initial test using Autodock 4.0 was performed to provide a selection of probable conformers for further study by MD. By default Autodocktools was able to identify the amide bond present in the oligoamide compound as rigid. Plante *et al.* reported free rotation about the ArCO bond whilst the ArNH bond was observed to exist in a planar conformation(Plante *et al.* 2009). The planar conformation is facilitated by an intramolecular hydrogen bond between the ether oxygen and the NH group, stabilising the structure most favourably into this conformation, but allowing an alternative less stable conformation with the NH and ether O to exist at 180° from each other. These observations have been backed up by in-silico studies by Vemparala *et al.*(Vemparala *et al.* 2006).

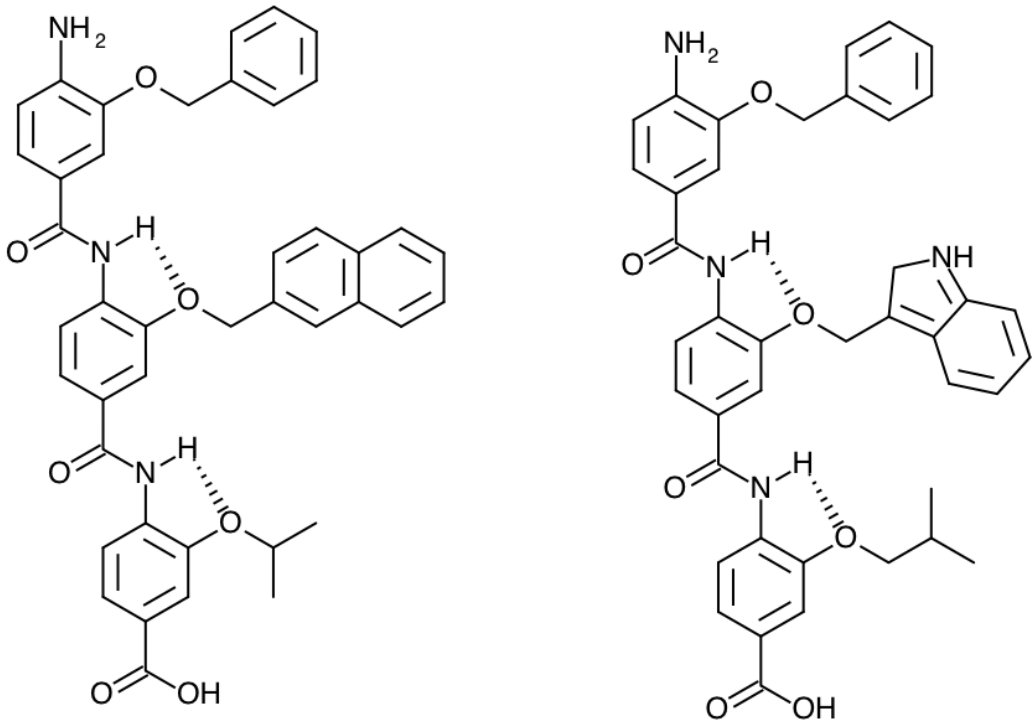

Figure 4.5: Compounds used in the docking study in this chapter. The left hand compound was synthesised by Plante *et al.* whilst the right hand compound contains the tryptophan side-chain mimic. Note the intramolecular hydrogen bond restricting the conformation of the compound. Partial sp_2 character of the ArNH bond also allows for a less stable conformation with the ArNH bond rotated by 180º causing the intramolecular hydrogen bond to be broken.

A preliminary screen was performed that did not restrict any torsional angles to determine whether the Autodock energy function could accurately describe this non-standard behaviour. Figure 4.6 shows that this is unlikely to be the case since the ArNH dihedral is 90º to the benzene ring to which it is attached. This position lies at the peak of a metastable region identified by Vemparala *et al.* and is about 6 kcal mol^{-1} greater in energy than its most stable energy minimum, thus an extremely unlikely conformation(Vemparala *et al.* 2006). As a result all further Autodock simulations also restrained the oligoamide into the favoured conformation (observed in X-ray structure and NMR data) whereby the ArNH

dihedral is oriented such that the amide hydrogen can form the intramolecular hydrogen bonds with the ether oxygen observed in the NMR study in the work of Plante *et al.*(Plante *et al.* 2009).

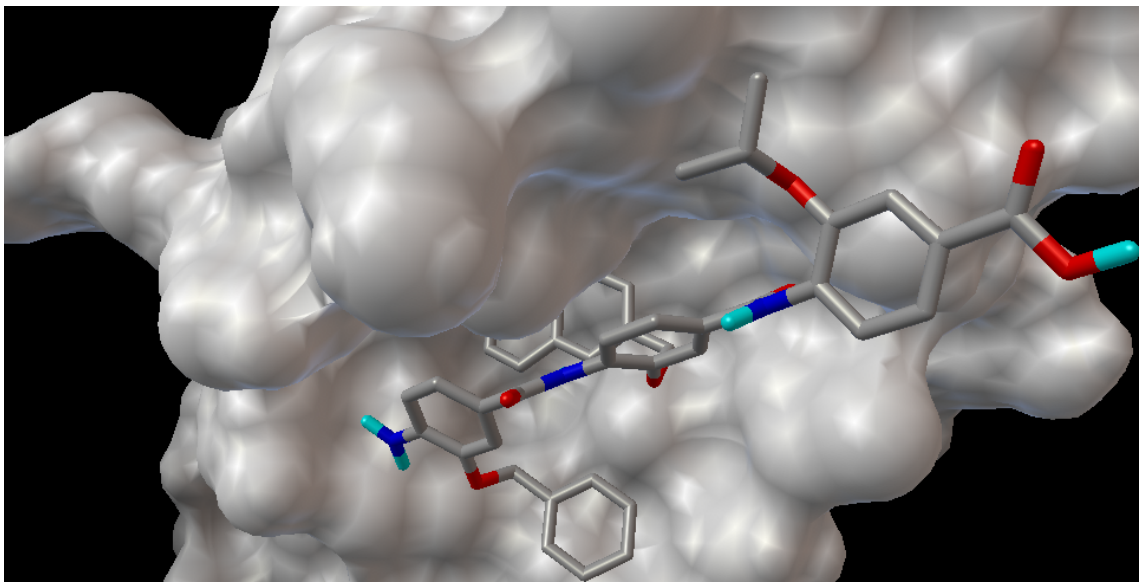

Figure 4.6: A representative for the second -most highly populated cluster- for an autodock experiment whereby the ArNH torsion was not restricted.

We performed a preliminary docking of 150 compounds with the ArNH dihedral restrained to its preferred low energy conformation previously discussed, to allow us to select a small number of compounds to use for initial MD simulations of the hDM2 binding site. We identified three possible docking modes from the top 3 low energy clusters when using a 2 Å RMS clustering threshold. The resulting structures are representatives from each cluster and are shown in figure 4.7 relative to the position of high affinity p53 helix. The figure was produced by identifying the rotation and translation that maps the hDM2 atoms used in the docking run onto the 1T4F atoms, and applying the same rotation and translation to the oligoamide compound, allowing comparison of the docked compounds to that of the high-affinity p53 peptide. This produces two possible classes of results. The first we call parallel conformations, that is those conformations which present their C-terminus spatially proximal to the location to the C-terminus of the p53 helix, and their N-terminus spatially proximal to the N-terminus of the helix such as

in figure 4.7b. The second class of conformations we call anti-parallel, that is those conformations that present their C-terminus spatially proximal to the N-terminus of the p53 helix, and their N-terminus spatially proximal to the C-terminus of the p53 helix, such as in figure 4.7a and 4.7c.

For the parallel conformation in figure 4.7b we see that it presents side-chains in a very similar way to the high-affinity helix, and as a result the wild-type helix. Figure 4.7a shows the oligoamide side-chains in the anti-parallel conformation occupying a large part of the pocket that is normally occupied by the Leucine, Tryptophan and Phenylalanine residue. However, these are not presented such that they directly map to residues presented by the helices. As such the Leucine mimic does not fill the region of the pocket that would normally be filled by the Phenylalanine of the p53 peptide. In the case of the result from cluster 3, figure 4.7c we see another anti-parallel conformation. Once again this structure presents Phenylalanine and Tryptophan residues that map to the Leucine and Tryptophan residues of the p53 peptide. However this conformation has a twisted ArCO bond at the C-terminus meaning that the Leucine side-chain does not fill the pocket normally occupied by the Phenylalanine of the p53 peptide.

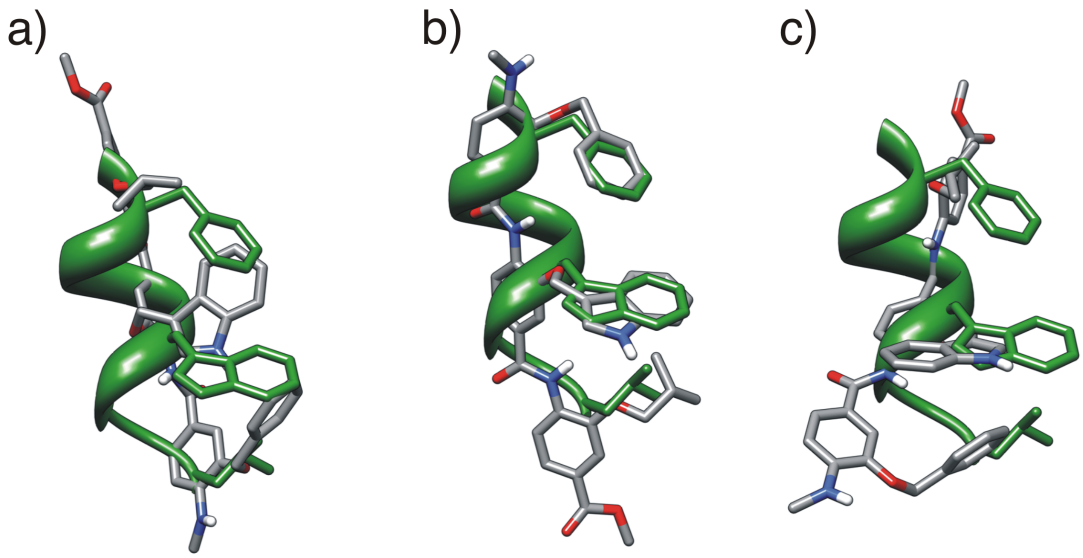

Figure 4.7: Representative structures from a preliminary docking screen that were used for initial MD studies (atom coloured), shown relative to high affinity helix from 1T4F. Representatives from: a) Cluster 1, representative 6; b) Cluster 2 representative 1; c) Cluster 3; representative 6. Figures were generated using the matchmaker function from Chimera to superpose hDM2 from docked conformations to hDM2 from 1T4F(Pettersen *et al.* 2004).

Since these docking calculations used sampling parameters that are less extensive than those suggested by the developers of Autodock for molecules with > 10 torsional degrees of freedom (the Phe-Trp-Leu compound has 12 active torsions), it is more than likely that the lowest energy conformation may not have been identified. As a result the parameters were altered and a second study was carried out, to identify whether a parallel or anti-parallel conformation might be more likely. Increasing the sampling of the torsional degrees of freedom took significantly more time thus the results were calculated on a computer cluster rather than a standalone PC. The results from the enhanced docking simulation are presented in figure 4.8, where the mean Autodock binding energy score is presented for each of the clusters generated using a 2 Å RMSD cutoff. Representative structures from each of the large low energy clusters are shown inset, alongside a representation of the high-affinity p53 helix shown in cyan (figure 4.8). Whilst there is plenty of literature to suggest that docking experiments are not

generally sufficient to predict the binding affinity of protein-ligand complexes, they are still relevant in the context of comparison amongst molecules of similar classes. So whilst comparing the Autodock score of Gleevec bound to Tyrosine Kinase ABL2 and that of Nutlin-2 bound to hDM2 is likely to be irrelevant, the comparison of scores within classes of similar proteins and ligands is likely to be useful but not quantitative. Here we compare the energy of the clusters to identify the likely binding mode with the warning from the developers that a score that has highly populated clusters within 2.5 kcal mol^{-1} of each other are unlikely to be distinguished from an incorrect binding mode. So whilst conformation 2 has the lowest energy of about -11.8 kcal mol^{-1}, conformation 1 also has a highly populated cluster with mean Autodock energy of -10.8 kcal mol^{-1}, a difference of only 1 kcal mol^{-1}.

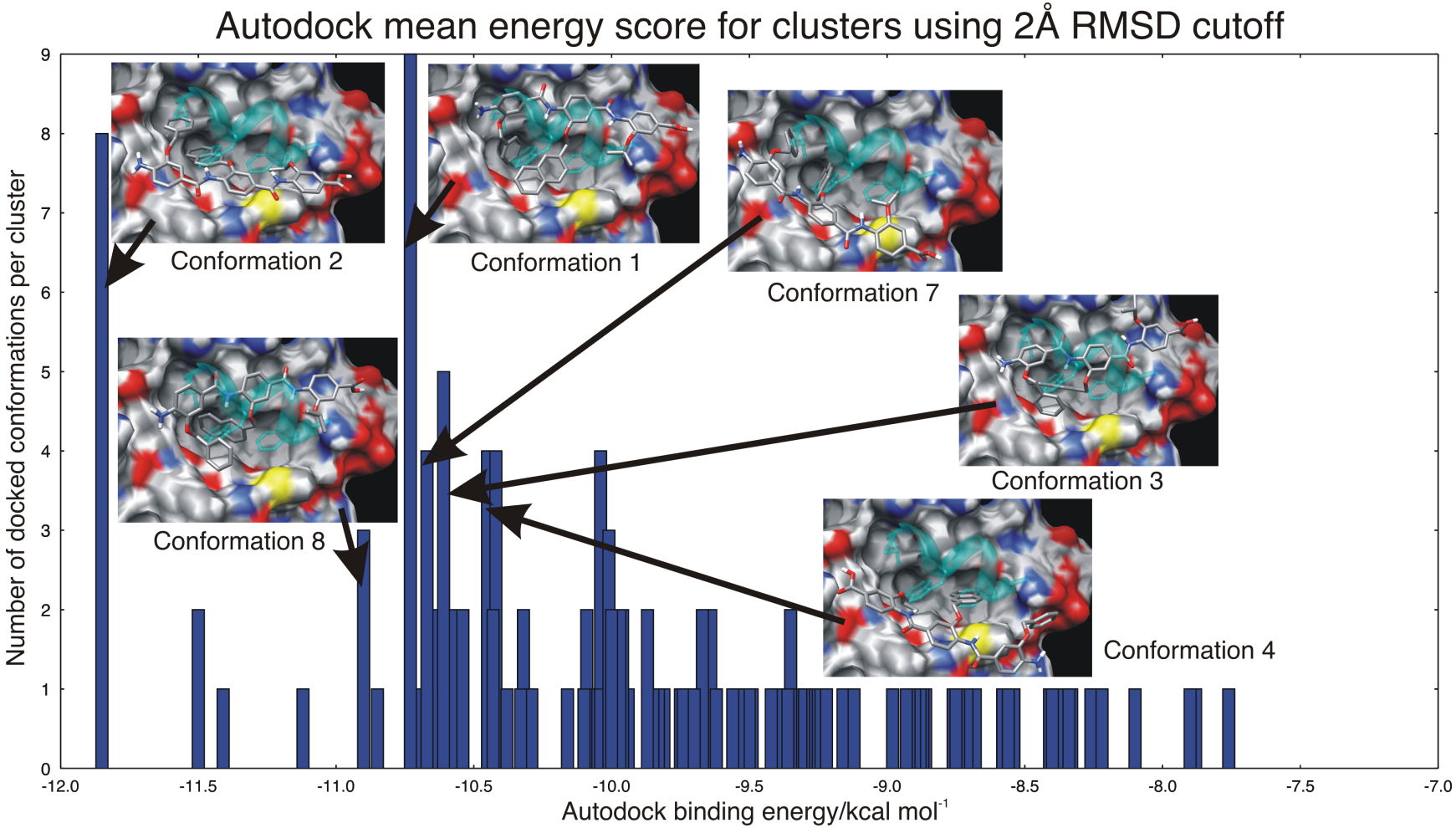

Figure 4.8: Mean autodock binding energy score and corresponding cluster occupancy created using a 2 Å RMSD cutoff. Representatives that were used as initial conformations for later MD simulations and the cluster from which they originated are highlighted.

The results in figure 4.8 suggest a bias towards anti-parallel conformations, with five of the six large low energy clusters showing this orientation. It is unclear whether this bias is due to the fact that oligoamide conformers are more stable in their anti-parallel conformation, perhaps due to steric clashes, since in figure 4.7 a and 4.7c it is possible to see the extension of the C-terminal beyond the N-terminal of the p53 helix in anti-parallel conformations allowing the possibility of steric clash with protein in this region. Perhaps the negatively charged C-terminal of the oligoamide is favoured in the region of the N-terminus of the p53 helix, since the surface potential is slightly positive in this region (see figure 4.3).

Given there are several plausible structures, it seems prudent to continue with a handful of likely structures when carrying out free energy calculations. Whilst it is of extremely low probability that parallel and anti-parallel conformations might inter-convert on the time-scale of MD simulations, one might expect likely parallel (or anti-parallel) conformations to inter-convert. Potentially this could result in convergence towards a consensus structure for parallel or anti-parallel conformations respectively. For the purposes of free energy calculations we may then decide to only use one starting conformation if there is evidence it will inter-convert between likely binding conformations on the time-scale of our calculations. To this end we continue using the five anti-parallel conformations (conformation 1, 2, 3, 7, 8) shown in figure 4.8. We also continue to use the parallel conformation determined above, whilst also selecting two additional plausible conformations to even the data set slightly (conformation 4/9, 10, 11). It may then be possible to determine the free energy difference between parallel and anti-parallel conformations of the oligoamide which would allow determination of the most likely binding mode.

4.4.1.d FRED

FRED is a rigid-body docking program that uses conformers of ligands generated by OMEGA and a static representation of the protein molecule, although it has the ability to consider flexible residues. A variety of scoring functions are available for FRED, we briefly assessed Chemgauss 3.

Figure 4.9 shows the top ranked docked position when using the Chemgauss 3 scoring function on the dataset of 150 conformations produced by FRED. Chemgauss 3 is the default scoring function for FRED and has been assessed as performing well on hydrophobic pockets from the amyloidogenic protein transthyretin in previous work performed in the lab. The score is dominated by the contribution from the steric term, with all but 3 conformations showing a favourable contribution by the steric term. All results show an unfavourable desolvation contribution to the overall score. When assessing whether the conformations are in the parallel conformation or anti-parallel conformation we observe 49 in the former and 101 in the latter. The heavy reliance of the scoring function on the steric scores seems to bias towards conformations that target the p53 Tyrosine residue (RFMD**Y**WEGL), instead of the main Phe-Trp-Leu pocket, indeed many of the conformations observed have at least one side-chain outside the main pocket region, as can be seen in figure 4.9.

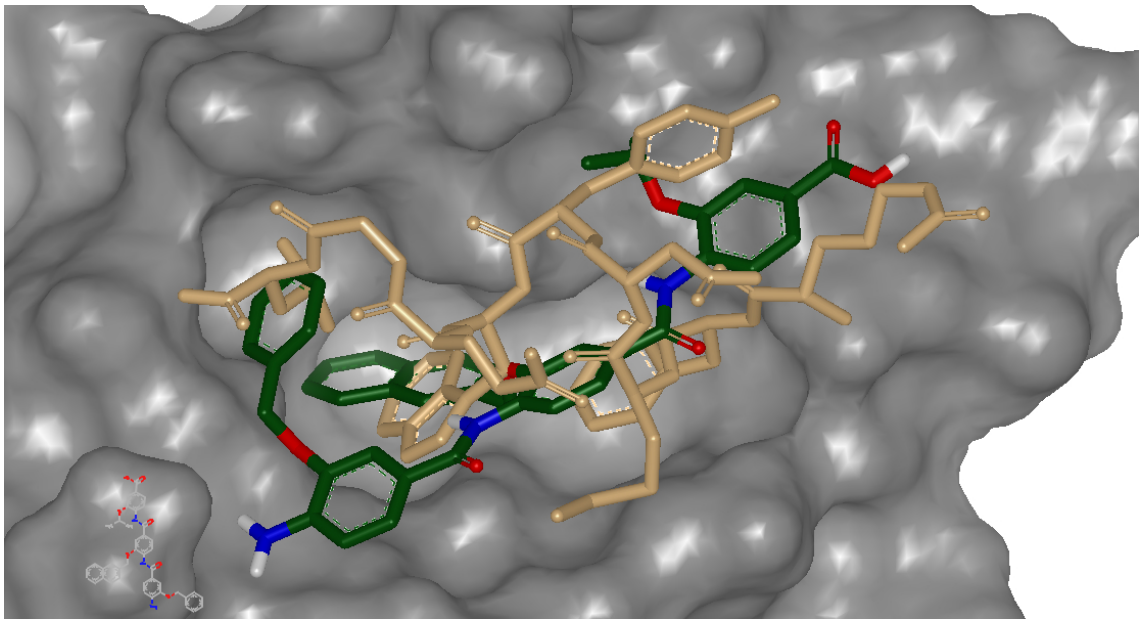

Figure 4.9: Best pose from FRED using the Chemgauss 3 scoring function (green, red, blue and white coloured atoms) compared to the Phe-Trp-Leu high-affinity p53 helix. hDM2 molecular surface shown in grey. Note the tyrosine ring from the p53 helix (beige) towards the top right hand corner of the figure.

In general we decided that whilst FRED is a fast docking program the binding poses did not look as convincing as those from Autodock. It is not possible to decide for certain which method is likely to produce the best results in terms of similarity to an X-ray structure for a given oligoamide compound, or the ability to correctly rank compounds in terms of binding affinity, due to the lack of available experimental data.

4.4.1.e Superposition Method

Since we are applying the hypothesis that side-chains from synthesised oligamide compounds directly mimic the side-chains from the p53 helix that are known hotspot residues in the hDM2 interaction, it is reasonable to assume that a simple method for generating starting conformations for free energy calculations is to simply overlay the oligoamide onto the p53 helix such that the side-chains mimic the hDM2-p53 interaction as closely as possible. We used the GH8 program to superpose oligoamide conformers generated using the OMEGA program onto the

high-affinity p53 helix. We observed that as this was carried out there was a heavy bias to matching oligoamide atoms to the helix side-chain atoms that are known to be less energetically important for the interaction. As a result we ran the superpositions again using only the side-chain atoms from the Phe-Trp-Leu residues. We now observed that whilst we no longer had the problem of matching side-chains to the wrong region, we lost information from the peptide about the preferred orientation of the side-chains with respect to the protein. That is to say that whilst we might match atoms from the oligoamide side-chains to those of the peptide well, we might then arrange the oligoamide backbone where the hDM2 protein would normally exist. We struck a balance by using only the peptide backbone atoms and all of the atoms from the Phe-Trp-Leu residues.

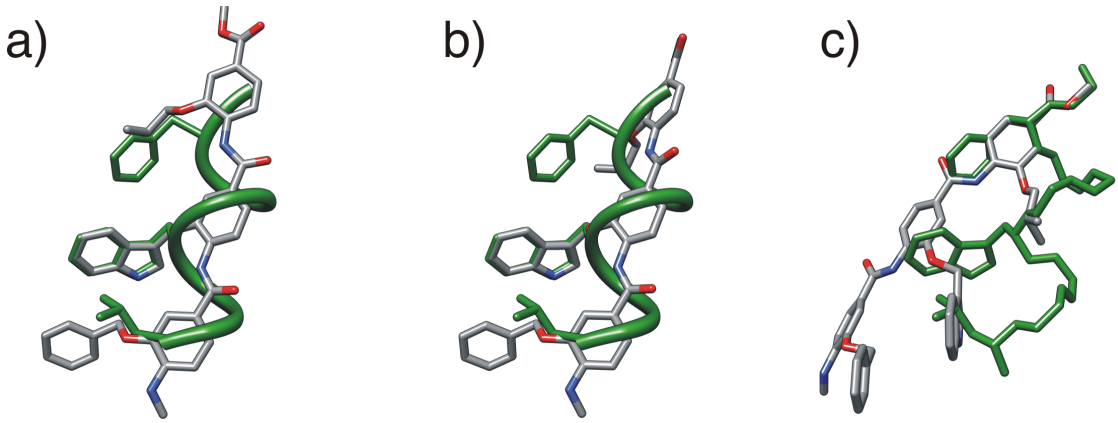

Figure 4.10: Oligoamide compounds shown coloured according to atom type superposed onto the binding Phe-Trp-Leu residues and backbone atoms from the high-affinity p53 helix shown in dark green. All compounds shown are oriented in the anti-parallel conformation with: a) showing a good match; b) a reasonable match; c) a poor match which would sterically clash with the hDM2 protein.

Representative structures from the method are shown in figure 4.10, with 4.10a showing a fairly successful match. In this case the Tryptophan rings are matched extremely well, but the method only identified anti-parallel conformations. Figure 4.10b shows a similarly successful match which is slightly worse due to the fact that the Leu residue is twisted out of alignment from the Phe residue of the

peptide. Figure 4.10c shows a bad match which would result in a steric clash with the hDM2 protein. This is similar to many of the matches that were observed in the previously described superposition of oligoamide compounds onto Phe-Trp-Leu residue side-chain atoms only.

We observed several issues with using superposition methods in the context of this system. The first is that it is difficult to score the results in such a way that it would be possible to prioritise those that are more likely to be observed in reality. For example 15 atoms are matched in both figure 4.10a and 4.10b, but when looking at the results a) is clearly a better match. Additionally figure 4.10c matches 14 atoms, only one less than 4.10a and 4.10b, but does not match the side-chains atoms very well, and would have problems with steric clashes with the hDM2 protein atoms. Whilst this would be acceptable for setting up a single system, this method would not be appropriate for a larger scale system where several different oligoamide compounds are investigated. Additionally it is not clear how well this method would fare when applied to oligoamide compounds that have side-chains designed to bind hDM2 with higher affinity, whilst perhaps having different molecular shapes. For example a napthylene ring instead of the ring from a tryptophan is likely to score less well even if all of the non ring atoms are located in exactly the same place. The final consideration is that the method is very reliant on the quality of the conformations used in the method, and whilst it appears that OMEGA is successful in generating good conformers, in the absence of a sensible scoring scheme a large amount of manual inspection is required which is unfeasible for high throughput applications.

4.4.2 Oligoamide charge calculations

Two charge calculation methods were evaluated, these were the AM1 BCC semi-empirical method(Jakalian, Jack, and Bayly 2002) implemented with Antechamber from the AMBER package(D.A. Case, T.A. Darden, T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke *et al.* 2004) and high level Hartree-Fock molecular

electrostatic potential calculation methods followed by RESP charge fitting using the REDIII.1 program(Dupradeau *et al.* 2008). Both methods calculated backbone charges for the compound shown in figure 4.11. Additionally side-chain charges for positions R_1, R_2 and R_3 were calculated for the four compounds shown in 2D to the right of figure 4.12, for comparison these values were compared to related side-chain charges from the AMBER99 force fields. The two charge calculation methods were chosen since they are consistent with the charge calculation methods used to calculate charges for the AMBER force fields. In particular Hartree-Fock calculations using the 6-31G* basis set followed by RESP charge fitting was used to derive charge parameters for the original AMBER94 force field described by Cornell *et al.*(Cornell *et al.* 1995). The AM1 BCC charge calculation is considered as an alternative to the HF 6-31G* method, since it is several orders of magnitude faster, whilst also being parameterized such that it should reproduce charges calculated using the HF 6-31G* basis set and the RESP method(Jakalian, Jack, and Bayly 2002). Comparison is made between backbone AM1 BCC charges for oligoamide compounds calculated for a large number of conformers of 4 compounds described in figure 4.12. Side-chain charges for Leucine and Phenylalanine mimics are also compared using the AM1 BCC charge method. We then make comparison of two HF 6-31G* methods for backbone charge calculation; one in which we perform the calculation for the entire oligoamide compound shown in figure 4.11, the second in which the compound is split into one of three fragments that when combined could describe the entire molecule; additionally we make comparison to the AM1 BCC method. The fragment method for calculating HF 6-31G* method would allow for increased speed of computation, since if we wanted to simulate a selection of oligoamide compounds derived from a library of three side-chains we would need to perform 27 full molecule oligoamide simulations, compared to 3 individually less expensive fragment HF 6-31G* calculations. The HF 6-31G* full molecule and fragment calculations are finally compared to the AM1 BCC charge method for side-chain mimics of alanine and tryptophan, which are additionally compared to the values provided with the AMBER force fields.

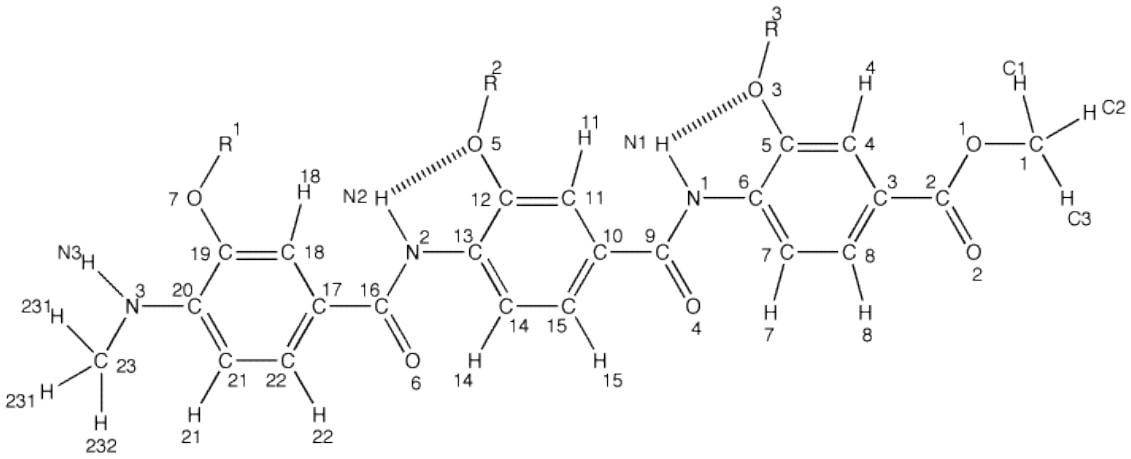

Figure 4.11: Schematic of the atom labelling scheme used in the charge calculation work for backbone atom labelling, showing the atomic element and the number used to identify specific atoms.

4.4.2.a Oligoamide Backbone charges

The first comparison made was between backbone charges calculated using the AM1 BCC charge method. Figure 4.12 shows backbone calculations for: a) -CH3-CH3-CH3; b) Phe-Trp-Leu; c) Phe-Nap-Leu; d) Val-Phe-Propyl. It is clear that there is not much variation between each of the backbones, indicating that the method provides a reasonable consensus, this is backed up by very small calculated error bars. Of note are C16, C2 and C9 which comprise the 3 carbons that form the amide bond between aryl groups that have the largest positive charges: +0.699 e; +0.655 e; +0.697 e; for carbon atoms. We also note that since the oligoamide compound is an oligomer we see a degree of symmetry in the results. That is to say that C20, C13 and C6 (carbon atoms that exist at equivalent positions in the oligomer) are all slightly positive with charges of: +0.199 e; +0.098 e; +0.097 e; respectively. For all atom types (carbon, hydrogen, nitrogen and oxygen) we do not see any cases where a charge is not calculated as all positive or all negative for the collection of 4 compounds which do not vary in backbone structure but do have varying side-chains. Nitrogens N1 and N2 carry almost identical negative charges, whilst nitrogen N3 carries a larger negative charge. We might expect more variation in the charges of the ether oxygens O3,

O5 and O7 that occupy positions that mimic positions of the Cα at positions i, i+4 and i+7 on a superposed helix. There is little variation in the charge calculated for these oxygens, however, they are less negative than the oxygen atoms that form the amide bonds.

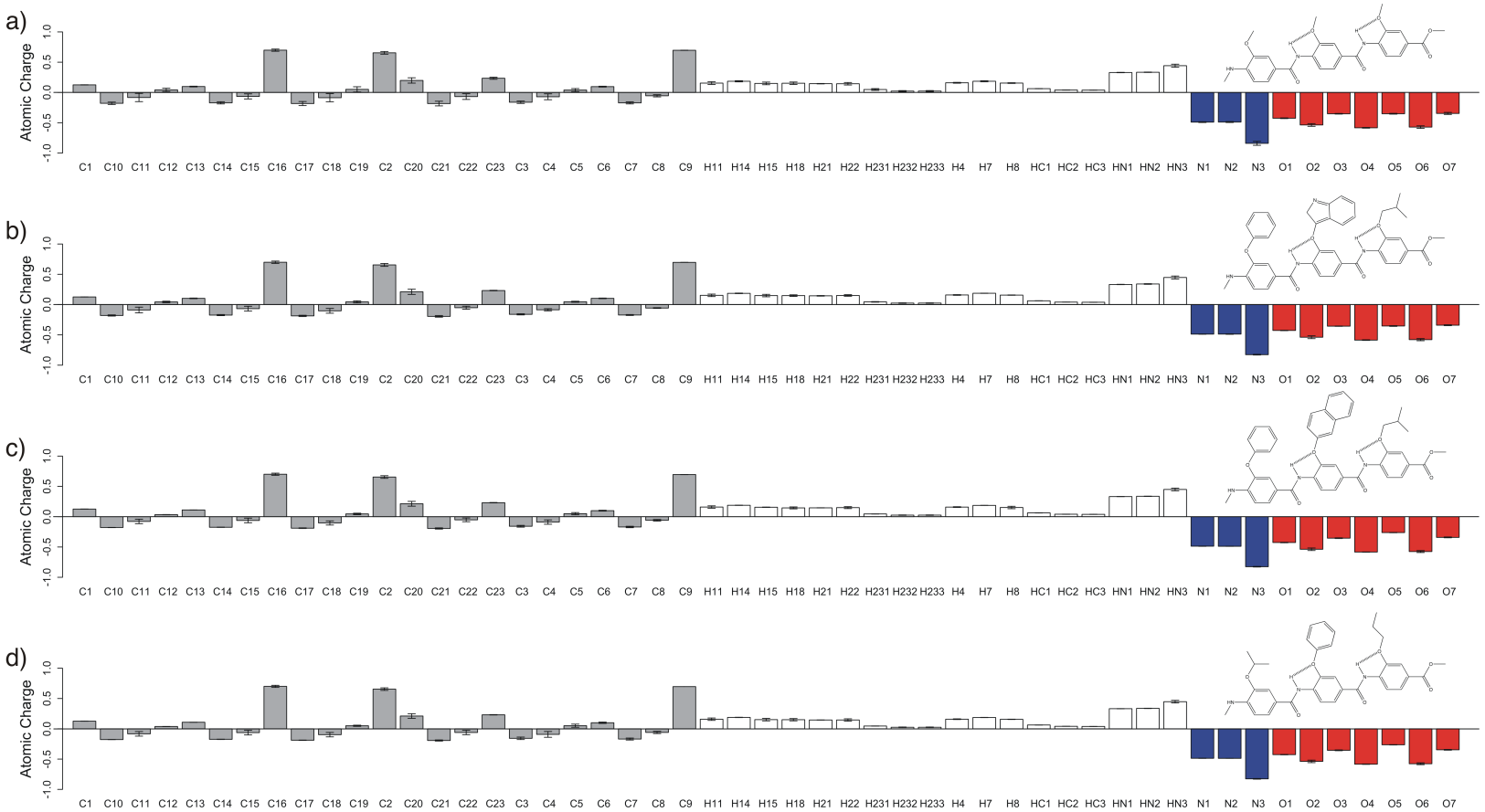

Figure 4.12: Atomic charge calculated using the AM1 BCC charge method implemented in the Antechamber program from AmberTools 1.2. Mean and 95 % confidence interval was calculated for the stated number of conformations as generated using the OMEGA package provided by OpenEye software. a) 361 structures from a triple -CH3 substituted compound; b) 310 structures from a Phe-Trp-Leu mimic; c) 361 structures from a Phe-Nap-Leu mimic; d) 380 structures from a Val-Phe-Propyl compound.

4.4.2.b Oligoamide side-chain charges

In figure 4.13 we look at how the AM1 BCC charge calculation varies; for side-chains that attempt to mimic Leucine/Valine side-chains due to the compound to which it is attached. In all cases the backbone is the same; however in the case of the Val-Phe-Propyl compound the side-chain of interest is at the N-terminus of the compound rather than the C-terminus of the compound as for the Phe-Trp-Leu and Phe-Nap-Leu compounds. These compounds are shown in 2D to the right of figure 4.12d, 4.12b and 4.12c respectively. Bars for C24, H241 and H242 are missing for the Val-Phe-Propyl compound since these atoms are not present in this compound. The strong electronegativity of the ether oxygen is clear from the large positive charge for C24 in Phe-Trp-Leu and Phe-Nap-Leu and for C25 in the case of Val-Phe-Propyl. The two carbons farthest from the attachment ether, C26 and C27 have similar positive charge values in all three compounds. As a comparison charge values for the Leucine side-chain contained in the AMBER99 force field are included. They follow the pattern C24 slightly negative, C25 large positive, C26 and C27 negative, with values: -0.110 e; +0.353 e; -0.412 e; -0.412 e; respectively. Hydrogen values in all cases are much more comparable between all data sets. AMBER99sb hydrogens attached to C26 and C27 are approximately double those calculated with the AM1 BCC charge methods.

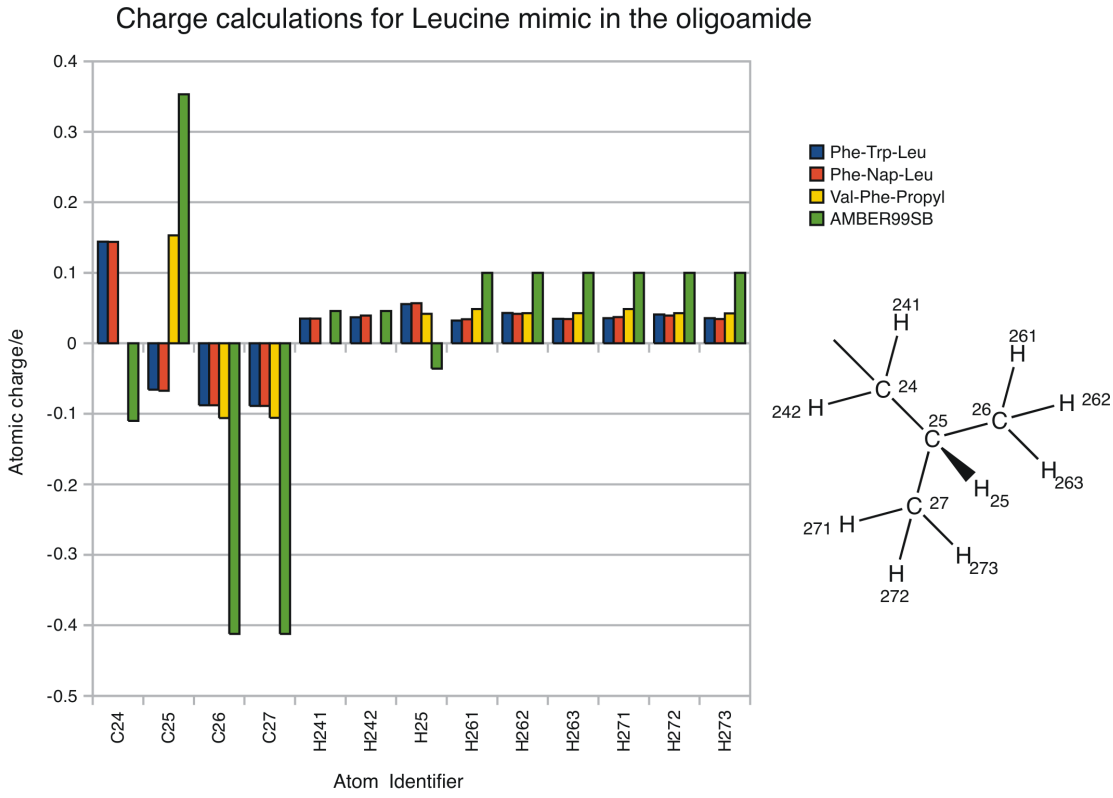

Figure 4.13: AM1 BCC charge calculations for Leucine side-chain mimics. Mean values calculated from the result of: 310 conformations from a Phe-Trp-**Leu** compound (blue); 361 conformations from a Phe-Nap-**Leu** compound (orange); 380 conformations from a **Val-**Phe-Propyl compound (yellow); AMBER99sb Leucine side-chain charges for comparison (green).

Figure 4.14 provides a similar comparison to the above Leucine side-chains for Phenylalanine mimics from the Phe-Trp-Leu, Phe-Nap-Leu and Val-Phe-Propyl compounds. The Phenylalanine mimic in the Val-Phe-Propyl compound is missing the alkyl carbon C37 and corresponding hydrogens H371, H372 as can be seen to the right of figure 4.12d. As a result carbon C37 bound to the ether oxygen has a charge of about +0.1 e compared to ~ +0.2 e for C37 and -0.1 e for C38 from Phe-Trp-Leu and Phe-Nap-Leu compounds that contain the alkyl carbon before the benzene ring. Despite this difference the charge on the other ring atoms is

generally comparable between all three compounds, and in addition phenylalanine atoms with the exception of C37 and C38 (and corresponding hydrogen atoms) also have comparable charges in the AMBER99 force field.

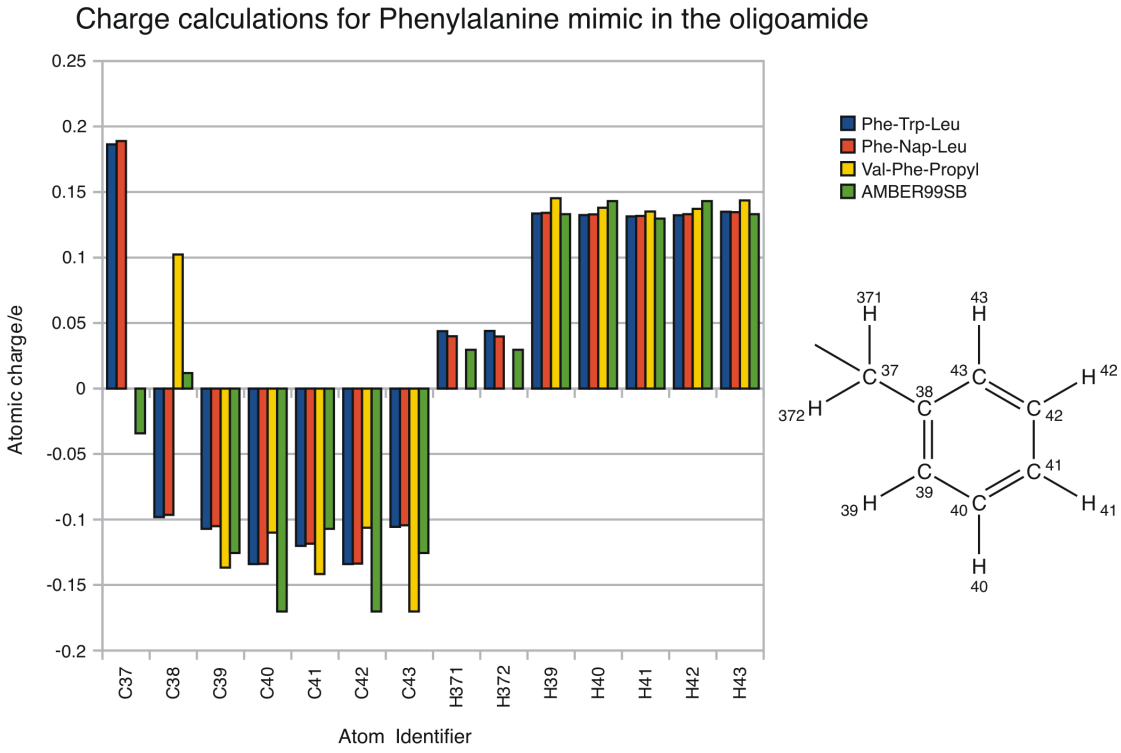

Figure 4.14: AM1 BCC charge calculations for Phenylalanine side-chain mimics. Mean values calculated from the result of: 310 conformations from a **Phe-**Trp-Leu compound (blue); 361 conformations from a **Phe-**Nap-Leu compound (orange); 380 conformations from a Val-**Phe-**Propyl compound (yellow); AMBER99sb Leucine side-chain charges for comparison (green).

AM1 BCC charge calculations as implemented here have the problem that they do not take into account symmetry of atoms in a calculation. For example in Figure 4.14 the phenylalanine side-chain calculations for C39 and C43 in the Va-Phe-Propyl compound have charge values of -0.137 e and -0.170 e respectively. Yet they are indistinguishable particles thus any model should ideally treat them as identical, meaning that they should end up with the same charge.

It is reassuring that the conformation of the molecule that OMEGA generates does not have a large effect on the charge of the molecule as shown by the small error bars observed in figure 4.12. Additionally the calculations appear to arrive at a consensus value for the charge of the backbone that is independent of the substitution of pattern for the side-chain groups. It should be noted that the side-chain mimics investigated in this study are all alkyl or aryl groups with no net charge, so some care must be taken if these compounds are investigated.

4.4.2.c Full quantum mechanical vs. semi-empirical charge calculations

We next look at the results for HF 6-31G* backbone charge calculations compared to AM1 BCC charge calculations for R_1=CH_3,R_2=CH_3,R_3=CH_3 substituted compounds. The results are presented in figure 4.15 and generally show broad agreement. The main disagreement is for C23 which is part of the N-terminal methyl cap. This capping group is not present in the final oligoamide simulations presented later in the chapter, as it was initially thought to be necessary for performing accurate free energy calculations. Thus disagreement between the two charge methods for this atom have no impact on the free energy calculations described later. Since the two charge methods agree well in the case of the backbone calculations we need to make a final comparison of charge calculation methods for side-chains before we can decide which charge method we can proceed further with.

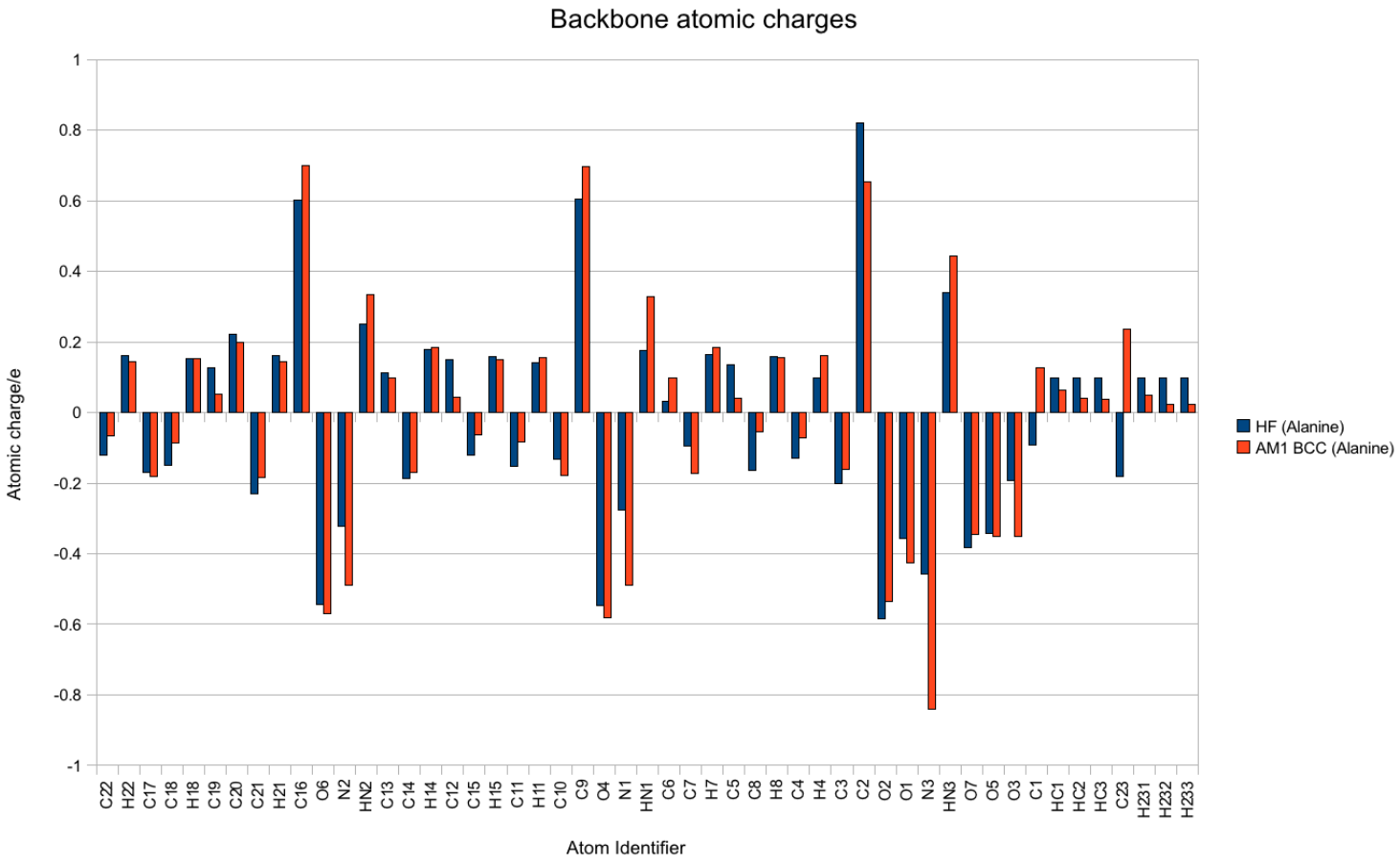

Figure 4.15: Backbone atomic charges calculated using the HF6-31G* level of theory (blue) compared to backbone atomic charges calculated using the AM1 BCC level of theory (orange). Full QM calculations were carried out using Gaussian and the REDIII.1 software package, semi-empirical QM calculations were carried out using Antechamber from AmberTools 1.2. Full details are given in the methods section.

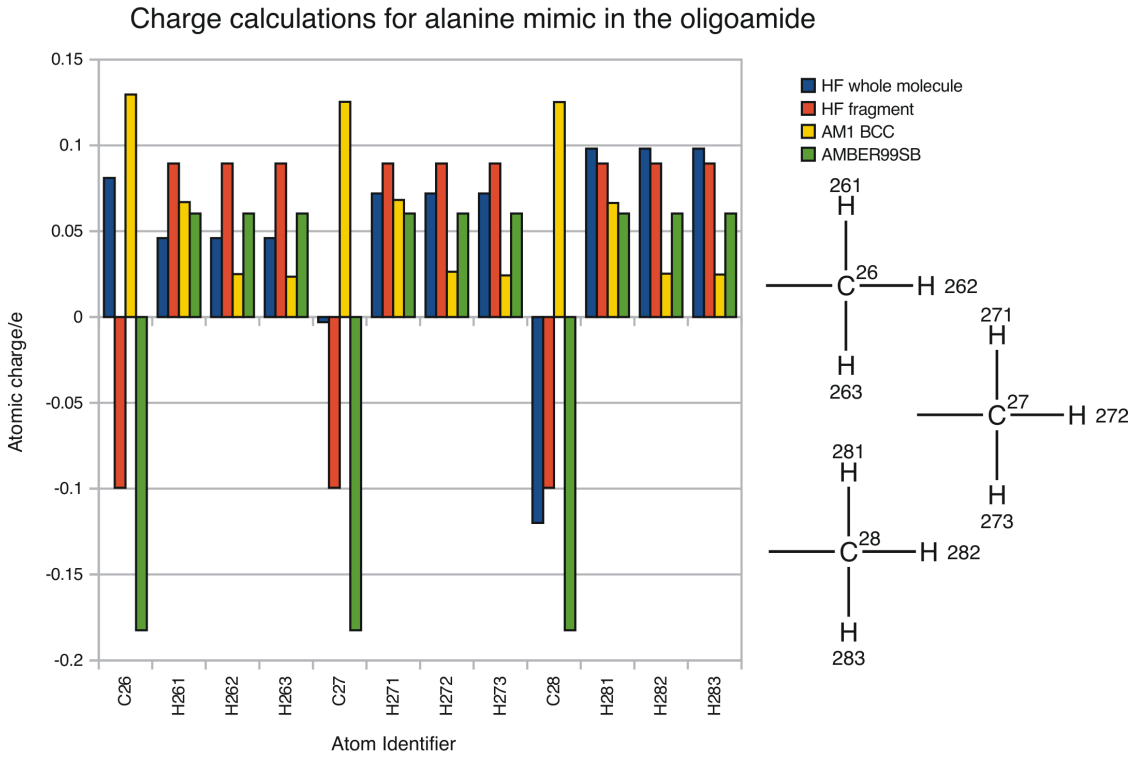

Figure 4.16: Comparison of charge calculation methods applied to the triple alanine substituted oligoamide. Results for the HF 6-31G* level of theory applied to a full molecule (blue); HF 6-31G* level of theory applied to a fragment containing the alanine side-chain (orange); the mean value from 361 conformations of the full compound using the AM1 BCC level of theory (yellow); and the charge values specified for the corresponding alanine side-chain atoms in the AMBER99sb force field (green).

In figure 4.16 we compare the two HF 6-31G* charge methods with AM1 BCC and provide comparison to AMBER charges for Alanine. Since we calculated the value for the CH_3 side-chain only once using the fragment HF 6-31G* method, we include the results against each side-chain that was calculated using the HF 6-31G* and AM1 BCC methods. We might expect that the full molecule methods for side-chain charge calculation should produce the same value for each CH_3 thus validating the fragment HF 6-31G* method as a suitable candidate. However we see that it disagrees with the full molecule side-chain calculations for the carbons comprising the two side-chains closest to the C-terminal. Full

molecule calculations in this case yield values of +0.081 e and -0.003 e compared to -0.100 e for the fragment calculation. However, the fragment and full-molecule calculations are in strong agreement with the value for the N-terminal side-chain carbon atom, -0.100 e and -0.120 e respectively. The calculations for the hydrogens in the CH_3 side-chains once again show the problem that the AM1 BCC method has with assigning differing charges to indistinguishable atoms, that is not visible in either of the HF 6-31G* methods. The CH_3 side-chain calculations is the first time that we see a disagreement between the HF 6-31G* and AM1 BCC charge methods. The AM1 BCC charge method produces very similar values for all CH_3 side-chains, whilst the HF 6-31G* method produces values that decrease from positive at the C-terminal substituent to negative at the N-terminal substituent. This in turn encourages the bound hydrogens to increase their charge values from approximately +0.05 e to +0.1 e from C to N-terminal. In this case the HF 6-31G* method appears to be taking into account the dipole moment of the molecule, since the C-terminus carries a negative charge thus is likely to have an electron donating character, whilst the N-terminus carries a positive charge thus is likely to have an electron withdrawing characteristic. Finally we note that the AMBER charges most closely resemble those of the fragment CH_3 compound with the HF 6-31G* method. This is in many ways expected since the fragment method essentially creates a compound which is most similar to those created when the original AMBER force field was parameterized.

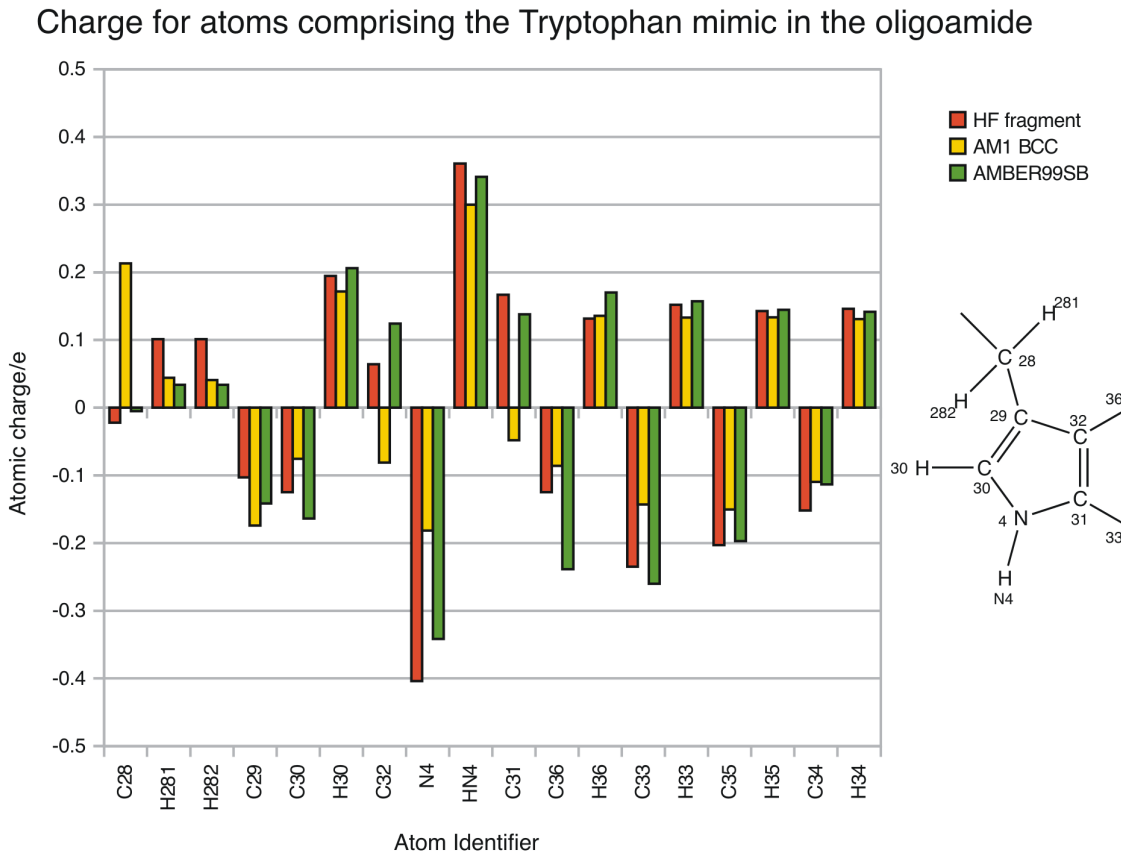

Figure 4.17: Comparison of charge calculation methods applied to the Tryptophan from a Phe-Trp-Leu substituted oligoamide. Results for the HF 6-31G* level of theory applied to a fragment containing the alanine side-chain (orange); the mean value from 310 conformations of the full compound using the AM1 BCC level of theory (yellow); and the charge values specified for the corresponding alanine side-chain atoms in the AMBER99sb force field (green).

Tryptophan side-chain charge calculations are also performed using AM1 BCC and fragment HF 6-31G* methods. Full molecule HF 6-31G* calculations are not evaluated since these calculations did not converge to an answer for over a week of CPU time. We see broad agreement with the charges calculated for atoms from both methods and the AMBER charges. There are some notable exceptions, the first is the AM1 BCC charge of +0.2 e for C28, the alkyl carbon bound to the backbone ether oxygen, compared to charges much closer to 0 e in the case of the HF 6-31G* calculation and those provided with the AMBER force field. We

also note a value close to -0.2 e for the AM1 BCC charge assigned to N4 (the tryptophan nitrogen), compared to a value slightly over -0.4 e for the HF 6-31G* calculation.

4.4.2.d Charge calculations summary

In general we have shown that the two charge calculation methods appear in broad agreement in the contexts to which they have been applied. It is likely that the HF 6-31G* method would provide more robust results, although the increased calculation time required is a very considerable additional overhead. If this type of charge calculation is to be used in an extensive modelling study it is likely that the AM1 BCC charge calculations should be sufficient and necessary in order to generate models in sufficient time. HF 6-31G* methods, or perhaps some intermediate level of theory could be used, as in the study by Vemparala *et al.* (Vemparala *et al.* 2006). These calculations are required to generate ab-initio torsional parameters for some of the compounds. It would then be possible to use these calculations and RED to do RESP charge fitting.

In addition to investigating the variation in charge parameters we can ask how much we think that incorrectly assigning charge parameters to our molecules might affect the results of a free energy calculation. Indeed, this question has been asked in the context of solvation free energies; in TIP3P and TIP4P-Ew water; of a set of 44 small neutral compounds. The GAFF force field was used to determine the bonded parameters, whilst a variety of MP2, B3LYP, HF 6-31G*, AM1 CM2 and AM1 BCC charge calculations were performed. The results from these alchemical free energy calculations were then compared to experiment. In this study all computed errors for free energies of hydration were less than 0.1 kcal mol^{-1} which is less than the reported 0.2 kcal mol^{-1} error reported in the study from which the experimental data is taken(Mobley *et al.* 2007). This report provides extra weight to support a policy of spending some time to ensure that the calculated charge values are as accurate as possible, whilst bearing in mind that

the additional speed provided by AM1 BCC methods probably outweighs any disadvantages due to inaccuracies in these charge calculations. In addition it can be noted that AM1 BCC charge methods have been used in other studies of hydration free energies to good effect(Mobley *et al.* 2009) and have additionally been used in successful calculations of protein-ligand free energy of association(Mobley *et al.* 2007).

4.5 Conclusion

The main aim of these experiments is to identify and assess a method for the preparation of plausible models for hDM2-oligoamide complexes for free energy calculations. Summarised below are the key results from this study, both in terms of those specific to carrying out free energy calculations on these particular complexes, and in terms of the methods employed to make these decisions.

In terms of docking and conformer generation, the results from both Autodock and FRED led us to believe that Autodock was producing more plausible docked conformations. However, it seems clear that the oligoamide compounds have a large degree of conformational flexibility and the Autodock scoring function predicts that many of these low energy clusters will exist within a small energy range meaning that it is not possible to identify with high probability a single lowest energy docked conformer. Indeed, it was not even possible to conclusively suggest whether the anti-parallel or parallel conformation is preferred. There are several lines of evidence to support the anti-parallel conformation as the most likely, this includes the larger number of low energy clusters for the anti-parallel conformation, secondly, the larger cluster sizes are also mainly in the anti-parallel conformation. Further evidence comes from the molecular superposition, in that the most convincing visual alignments are for the anti-parallel conformation. As a result several likely conformers will need to be taken forward for free energy

calculations including both parallel and anti-parallel conformations, where we believe that it will be possible to determine which conformations are the most favourable.

Regarding generation of charge parameters, the first key result is that there appears to be little variation in the charge values calculated for the oligoamide backbones when using the AM1 BCC charge calculation methods as compared to HF 6-31G* methods. The results from the two methods are in less agreement when it comes to side-chain calculations, however, the calculations nearly always agree on the sign of the charge, and often agree on the magnitude of the charge. Besides the fact that the AM1 BCC method is a lower level of theory than the HF 6-31G* method, the key disadvantage of the AM1 BCC method that we noted is that of the variation in charges assigned to indistinguishable atoms. However, both methods are designed to be consistent with the AMBER family of force fields including GAFF, and we conclude that it is appropriate to begin calculation of free energies using the AM1 BCC charge method since it is much faster than the Hartree-Fock method. It should also be noted that the acpype software front-end for Antechamber is considerably less involved to use than REDIII.1. This combined with the results from the study by Mobley *et al.* is enough to convince us that the AM1 BCC method is robust and accurate enough for the simulations reported in later chapters.

The results of the binding site analysis and the docking suggest that whilst the suggested poses can be quite variable (parallel and anti-parallel conformations), there is no decisive reason to expect one over the other, and there does appear to be reasonable correlation between the docking score and how well the docked conformation explores the binding pocket. Furthermore we showed that it is possible to generate charges for oligoamide compounds that are compatible with the GAFF force field. Since we have satisfied two of the requirements for MD

simulation it should now be possible to investigate the stability of the hDM2-oligoamide complexes generated here using MD simulations, and indeed these simulations are presented in the next chapter.

4.6 References

Anon. 2009. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Vienna, Austria. http://www.r-project.org.

Baum, Bernhard, Menshawy Mohamed, Mohamed Zayed, Christof Gerlach, Andreas Heine, David Hangauer, and Gerhard Klebe. 2009. More than a simple lipophilic contact: a detailed thermodynamic analysis of nonbasic residues in the s1 pocket of thrombin. *Journal of Molecular Biology* 390, no. 1 (July): 56-69. doi:10.1016/j.jmb.2009.04.051. http://www.ncbi.nlm.nih.gov/pubmed/19409395.

Bond, EE, Alexei Vazquez, Arnold J Levine, and GL Bond. 2008. The genetics of the p53 pathway, apoptosis and cancer therapy. *Nature Reviews Drug Discovery* 7, no. 12 (December): 979-87. doi:10.1038/nrd2656. http://www.ncbi.nlm.nih.gov/pubmed/19043449.

Brakoulias, Andreas, and Richard M Jackson. 2004. Towards a structural classification of phosphate binding sites in protein-nucleotide complexes: an automated all-against-all structural comparison using geometric matching. *Proteins* 56, no. 2: 250-60. doi:10.1002/prot.20123. http://www.ncbi.nlm.nih.gov/pubmed/15211509.

Brenke, Ryan, Dima Kozakov, Gwo-Yu Chuang, Dmitri Beglov, David Hall, Melissa R Landon, Carla Mattos, and Sandor Vajda. 2009. Fragment-based identification of druggable “hot spots” of proteins using Fourier domain correlation techniques. *Bioinformatics (Oxford, England)* 25, no. 5 (March): 621-7. doi:10.1093/bioinformatics/btp036. http://www.ncbi.nlm.nih.gov/pubmed/19176554.

Böttger, A, V Böttger, C Garcia-Echeverria, P Chène, H K Hochkeppel, W Sampson, K Ang, S F Howard, S M Picksley, and D P Lane. 1997. Molecular characterization of the hdm2-p53 interaction. *Journal of Molecular Biology* 269, no. 5 (June): 744-56. doi:10.1006/jmbi.1997.1078. http://www.ncbi.nlm.nih.gov/pubmed/9223638.

Böttger, V, A Böttger, S F Howard, S M Picksley, P Chène, C Garcia-Echeverria, H K Hochkeppel, and D P Lane. 1996. Identification of novel mdm2 binding peptides by phage display. *Oncogene* 13, no. 10 (November): 2141-7. http://www.ncbi.nlm.nih.gov/pubmed/8950981.

Carotti, Andrea, Antonio Macchiarulo, Nicola Giacchè, and Roberto Pellicciari. 2009. Targeting the conformational transitions of MDM2 and MDMX: insights into key residues affecting p53 recognition. *Proteins* 77, no. 3 (April): 524-35. doi:10.1002/prot.22464. http://www.ncbi.nlm.nih.gov/pubmed/19507240.

Cornell, Wendy D., Piotr Cieplak, Christopher I. Bayly, Ian R. Gould, Kenneth M. Merz, David M. Ferguson, David C. Spellmeyer, Thomas Fox, James W. Caldwell, and Peter A. Kollman. 1995. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* 117, no. 19: 5179-5197. doi:10.1021/ja00124a002. http://pubs.acs.org/doi/abs/10.1021/ja00124a002.

D.A. Case, T.A. Darden, T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R., J. Luo, K.M. Merz, B. Wang, D.A. Pearlman, M. Crowley, S. Brozell, V. Tsui, H. Gohlke, And Mongan, V. Hornak, G. Cui, P. Beroza, C. Schafmeister, J.W. Caldwell, W.S. Ross, and P.A. Kollman. 2004. AMBER. University of California, San Francisco.

Dickens, Michael P, Ross Fitzgerald, and PM Fischer. 2009. Small-molecule inhibitors of MDM2 as new anticancer therapeutics. *Seminars in Cancer Biology*. doi:10.1016/j.semcancer.2009.10.003. http://www.ncbi.nlm.nih.gov/pubmed/19897042.

Dupradeau, François-Yves, Christine Cézard, Rodolphe Lelong, Elodie Stanislawiak, Julien Pêcher, Jean Charles Delepine, and Piotr Cieplak. 2008. R.E.DD.B.: a database for RESP and ESP atomic charges, and force field libraries. *Nucleic Acids Research* 36, no. Database issue (January): D360-7. doi:10.1093/nar/gkm887. http://www.ncbi.nlm.nih.gov/pubmed/17962302.

Eyrisch, Susanne, and Volkhard Helms. 2007. Transient pockets on protein surfaces involved in protein-protein interaction. *Journal of Medicinal Chemistry* 50, no. 15 (July): 3457-64. doi:10.1021/jm070095g. http://www.ncbi.nlm.nih.gov/pubmed/17602601.

Fry, David C, S Donald Emerson, Stefan Palme, Binh T Vu, C-M Liu, and Frank Podlaski. 2004. NMR structure of a complex between MDM2 and a small molecule inhibitor. *Journal of Biomolecular NMR* 30, no. 2 (October): 163-73. doi:10.1023/B:JNMR.0000048856.84603.9b. http://www.ncbi.nlm.nih.gov/pubmed/15557803.

Galan, Jhenny F, Jodian Brown, Jayme L Wildin, Z Liu, D Liu, Guillermo Moyna, and Vojislava Pophristic. 2009. Intramolecular hydrogen bonding in ortho-substituted arylamide oligomers: a computational and experimental study of ortho-fluoro- and ortho-chloro-N-methylbenzamides. *The Journal of Physical Chemistry B* 113, no. 38 (September): 12809-15. doi:10.1021/jp905261p. http://www.ncbi.nlm.nih.gov/pubmed/19722486.

Gellman, Samuel H. 1998. Foldamers: A Manifesto. *Accounts of Chemical Research* 31, no. 4 (April): 173-180. doi:10.1021/ar960298r. http://pubs.acs.org/doi/abs/10.1021/ar960298r.

Grasberger, Bruce L, T Lu, Carsten Schubert, Daniel J Parks, Theodore E Carver, Holly K Koblish, Maxwell D Cummings, *et al.* 2005. Discovery and cocrystal structure of benzodiazepinedione HDM2 antagonists that activate p53 in cells. *Journal of Medicinal Chemistry* 48, no. 4 (February): 909-12. doi:10.1021/jm049137g. http://www.ncbi.nlm.nih.gov/pubmed/15715460.

Hernandes, Marcelo Zaldini, Suellen Melo T Cavalcanti, DRM Moreira, Walter Filgueira de Azevedo Junior, and Ana Cristina Lima Leite. 2010. Halogen atoms in the modern medicinal chemistry: hints for the drug design. *Current Drug Targets* 11, no. 3 (March): 303-14. http://www.ncbi.nlm.nih.gov/pubmed/20210755.

Hill, D J, M J Mio, R B Prince, T S Hughes, and J S Moore. 2001. A field guide to foldamers. *Chemical Reviews* 101, no. 12 (December): 3893-4012. http://www.ncbi.nlm.nih.gov/pubmed/11740924.

Jakalian, Araz, David B Jack, and Christopher I Bayly. 2002. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *Journal of Computational Chemistry* 23, no. 16: 1623-41. doi:10.1002/jcc.10128. http://www.ncbi.nlm.nih.gov/pubmed/12395429.

Kalid, Ori, and Nir Ben-Tal. 2009. Study of MDM2 binding to p53-analogues: affinity, helicity, and applicability to drug design. *Journal of Chemical Information and Modeling* 49, no. 4: 865-76. doi:10.1021/ci800352c. http://www.ncbi.nlm.nih.gov/pubmed/19323449.

Kirchmair, Johannes, Gerhard Wolber, Christian Laggner, and Thierry Langer. 2006. Comparative performance assessment of the conformational model generators omega and catalyst: a large-scale survey on the retrieval of protein-bound ligand conformations. *Journal of Chemical Information and Modeling* 46, no. 4: 1848-61. doi:10.1021/ci060084g. http://www.ncbi.nlm.nih.gov/pubmed/16859316.

Kortemme, Tanja, and David Baker. 2002. A simple physical model for binding energy hot spots in protein-protein complexes. *Proceedings of the National Academy of Sciences of the United States of America* 99, no. 22 (October): 14116-21. doi:10.1073/pnas.202485799. http://www.ncbi.nlm.nih.gov/pubmed/12381794.

Kussie, P. H., S. Gorina, V. Marechal, B. Elenbaas, J. Moreau, A. J. Levine, and N. P. Pavletich. 1996. Structure of the MDM2 Oncoprotein Bound to the p53 Tumor Suppressor Transactivation Domain. *Science* 274, no. 5289 (November): 948-953. doi:10.1126/science.274.5289.948. http://www.sciencemag.org/cgi/doi/10.1126/science.274.5289.948.

Kyte, J, and R F Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* 157, no. 1 (May): 105-32. http://www.ncbi.nlm.nih.gov/pubmed/7108955.

M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Pet, and J. A. Pople. 2004. Gaussian 03. Wallingford CT: Gaussian, Inc. Wallingford CT.

Massova, Irina, and Peter A. Kollman. 1999. Computational Alanine Scanning To Probe Protein–Protein Interactions: A Novel Approach To Evaluate Binding Free Energies. *Journal of the American Chemical Society* 121, no. 36 (September): 8133-8143. doi:10.1021/ja990935j. http://pubs.acs.org/doi/abs/10.1021/ja990935j.

McInnes, Campbell, Stanislava Uhrinova, Dusan Uhrin, Helen Powers, Kathryn Watt, Daniella Zheleva, P Fischer, and Paul N Barlow. 2005. Structure of free MDM2 N-terminal domain reveals conformational adjustments that accompany p53-binding. *Journal of Molecular Biology* 350, no. 3 (July): 587-98. doi:10.1016/j.jmb.2005.05.010. http://www.ncbi.nlm.nih.gov/pubmed/15953616.

Michel, Julien, Elizabeth A Harker, Julian Tirado-Rives, William L Jorgensen, and Alanna Schepartz. 2009. In Silico Improvement of beta3-peptide inhibitors of p53 x hDM2 and p53 x hDMX. *Journal of the American Chemical Society* 131, no. 18 (May): 6356-7. doi:10.1021/ja901478e. http://www.ncbi.nlm.nih.gov/pubmed/19415930.

Mobley, David L., Christopher I. Bayly, Matthew D. Cooper, Michael R. Shirts, and Ken A. Dill. 2009. Small molecule hydration free energies in explicit solvent: An extensive test of fixed-charge atomistic simulations. *Journal of Chemical Theory and Computation* 5, no. 2

(February): 350-358. doi:10.1021/ct800409d. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2701304&tool=pmcentrez&rendertype=abstract.

Mobley, David L., Elise Dumont, John D. Chodera, and Ken A. Dill. 2007. Comparison of charge models for fixed-charge force fields: small-molecule hydration free energies in explicit solvent. *The Journal of Physical Chemistry B* 111, no. 9 (March): 2242-54. doi:10.1021/jp0667442. http://www.ncbi.nlm.nih.gov/pubmed/17291029.

Mobley, David L., AP Graves, John D. Chodera, Andrea C. McReynolds, Brian K. Shoichet, and Ken A. Dill. 2007. Predicting absolute ligand binding free energies to a simple model site. *Journal of Molecular Biology* 371, no. 4 (August): 1118-34. doi:10.1016/j.jmb.2007.06.002. http://www.ncbi.nlm.nih.gov/pubmed/17599350.

Moreira, IS, Pedro A. Fernandes, and Maria J. Ramos. 2008. Protein–protein recognition: a computational mutagenesis study of the mdm2–p53 complex. *Theoretical Chemistry Accounts* 120, no. 4-6 (July): 533-542. doi:10.1007/s00214-008-0432-9.

Morris, Garrett M., David S. Goodsell, Robert S. Halliday, Ruth Huey, William E. Hart, Richard K. Belew, and Arthur J. Olson. 1998. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry* 19, no. 14 (November): 1639-1662. doi:10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B. http://doi.wiley.com/10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B.

Nikolovska-Coleska, Zaneta, K Ding, Y Lu, Su Qiu, Y Ding, Wei Gao, Jeanne Stuckey, *et al.* 2005. Structure-based design of potent non-peptide MDM2 inhibitors. *Journal of the American Chemical Society* 127, no. 29: 10130-1. doi:10.1021/ja051147z. http://www.ncbi.nlm.nih.gov/pubmed/16028899.

Pace, C N, and J M Scholtz. 1998. A helix propensity scale based on experimental studies of peptides and proteins. *Biophysical journal* 75, no. 1 (July): 422-7. http://www.ncbi.nlm.nih.gov/pubmed/9649402.

Pettersen, Eric F, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. 2004. UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of Computational Chemistry* 25, no. 13 (October): 1605-12. doi:10.1002/jcc.20084. http://www.ncbi.nlm.nih.gov/pubmed/15264254.

Plante, JP, Thomas Burnley, Barbora Malkova, Michael E. Webb, Stuart L. Warriner, Thomas A. Edwards, and Andrew J. Wilson. 2009. Oligobenzamide proteomimetic inhibitors of the p53–hDM2 protein–protein interaction. *Chemical Communications*, no. 34: 5091. doi:10.1039/b908207g. http://xlink.rsc.org/?DOI=b908207g.

Plante, J, Fred Campbell, Barbora Malkova, Colin Kilner, Stuart L. Warriner, and Andrew J. Wilson. 2008. Synthesis of functionalised aromatic oligamide rods. *Organic & Biomolecular Chemistry* 6, no. 1 (January): 138-46. doi:10.1039/b712606a. http://www.ncbi.nlm.nih.gov/pubmed/18075658.

Pophristic, Vojislava, Satyavani Vemparala, Ivaylo Ivanov, Z Liu, ML Klein, and William F DeGrado. 2006. Controlling the shape and flexibility of arylamides: a combined ab initio, ab initio molecular dynamics, and classical molecular dynamics study. *The Journal of Physical Chemistry B* 110, no. 8 (March): 3517-26. doi:10.1021/jp054306+. http://www.ncbi.nlm.nih.gov/pubmed/16494407.

Rocchia, W., E. Alexov, and B. Honig. 2001. Extending the Applicability of the Nonlinear Poisson–Boltzmann Equation: Multiple Dielectric Constants and Multivalent Ions. *The Journal of Physical Chemistry B* 105, no. 28 (July): 6754-6754. doi:10.1021/jp012279r. http://pubs.acs.org/doi/abs/10.1021/jp012279r.

Seeliger, Daniel, and Bert L de Groot. 2010. Conformational transitions upon ligand binding: holo-structure prediction from apo conformations. *PLoS Computational Biology* 6, no. 1: e1000634. doi:10.1371/journal.pcbi.1000634. http://www.ncbi.nlm.nih.gov/pubmed/20066034.

Shaginian, Alex, Landon R Whitby, Sukwon Hong, Inkyu Hwang, Bilal Farooqi, Mark Searcey, Jiandong Chen, Peter K Vogt, and Dale L Boger. 2009. Design, Synthesis, and Evaluation of an alpha-Helix Mimetic Library Targeting Protein-Protein Interactions. *Journal of the American Chemical Society*, no. 4: 5564-5572. doi:10.1021/ja810025g. http://www.ncbi.nlm.nih.gov/pubmed/19334711.

Vassilev, Lyubomir T, Binh T Vu, B Graves, Daisy Carvajal, Frank Podlaski, Zoran Filipovic, Norman Kong, *et al.* 2004. In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science (New York, N.Y.)* 303, no. 5659 (February): 844-8. doi:10.1126/science.1092472. http://www.ncbi.nlm.nih.gov/pubmed/14704432.

Vemparala, Satyavani, Ivaylo Ivanov, Vojislava Pophristic, Katrin Spiegel, and ML Klein. 2006. Ab initio calculations of intramolecular parameters for a class of arylamide polymers. *Journal of Computational Chemistry* 27, no. 6 (April): 693-700. doi:10.1002/jcc.20382. http://www.ncbi.nlm.nih.gov/pubmed/16634095.

Wang, J, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. 2004. Development and testing of a general amber force field. *Journal of Computational Chemistry* 25, no. 9 (July): 1157-74. doi:10.1002/jcc.20035. http://www.ncbi.nlm.nih.gov/pubmed/15116359.

Zhong, Haizhen, and Heather A Carlson. 2005. Computational studies and peptidomimetic design for the human p53-MDM2 complex. *Proteins* 58, no. 1 (January): 222-34. doi:10.1002/prot.20275. http://www.ncbi.nlm.nih.gov/pubmed/15505803.

5 Molecular dynamics simulation of novel Arylamide compounds bound to hDM2

5.1 Abstract

Molecular dynamics (MD) simulations are carried out in order to understand the behaviour of hDM2-p53 systems for which high-quality structures are already available. Following this further MD simulations are performed in order to refine simulation protocols and to determine the suitability of structures of oligoamide complexes generated using the computational methods described in the previous chapter. Recent methodologies for carrying out alchemical free energy calculations have shown that there are several particularly important features that need to be accurately sampled in order to be confident in the quality of the calculated free energy differences. Of particular importance is the sampling of dihedral angles that form the binding site and the dihedral angles present in the bound ligand. We investigate dihedral angles by analysis of the angles sampled, and additionally in using autocorrelation functions that will allow the length and number of MD simulations required to perform accurate alchemical free energy calculations. Finally we investigate the spatial sampling of ligands forming hDM2-oligoamide complexes to determine whether some starting conformations can interconvert and the time-scales over which this might occur. We show that generally some spatial sampling does not occur on the time-scale of typical MD simulations (> 20 ns), however, many do interconvert. Overall, the analysis described here increases our confidence that we can use a reduced number of starting conformations to proceed with alchemical free energy calculations and find quantitative relationships for hDM2-ligand binding.

5.2 Introduction

Molecular dynamics has been used as a tool to study both protein-ligand interactions and protein-protein interactions, with a variety of techniques being developed to improve sampling and ask a variety of questions(Woods, King, and Essex 2001),(Woo and Roux 2005),(Im, Feig, and Brooks 2003),(Lawrenz *et al.* 2010). A brief overview of some of the key techniques is provided in the thesis introduction, here we focus on application of these techniques to the field of protein-ligand and protein-protein interactions and if applicable to the hDM2-p53 system.

5.2.1 Molecular dynamics for studying protein-ligand interactions

Calculation of protein-ligand binding affinities has been performed using both Monte Carlo and molecular dynamics techniques. One of the earliest examples of free energy calculations was by Wong and McCammon who used the GROMOS molecular dynamics program to simulate for up to 64 ps and then used exponential averaging to compute free-energies of two benzamidine inhibitors of trypsin, and of benzamidine for wild-type and mutant trypsin(Wong and McCammon 1986). Essex *et al.* applied Monte Carlo FEP calculations to obtain relative free-energies of four Trypsin-Benzamidine complexes(Essex *et al.* 1997). Both studies showed success in predicting accurate free-energies. Essex *et al.* attributed worse performance on one complex to deficiency in partial charges of this complex. In addition Monte Carlo simulations have also been used to study the specificity of a series of non-peptidic inhibitors of the SH2 protein(Price and Jorgensen 2000). More recently Mobley *et al.* have applied alchemical free energy calculations to the T4 lysozyme system where they learnt some important lessons, that will be detailed later(Mobley, Chodera, and Dill 2006).

Molecular dynamics has also been applied to the hDM2 system. Zhong and Carlson used a GBSA approach to calculate first the binding affinity of the p53 peptide to hDM2 (-7.4 kcal mol^{-1}), followed by applying the technique to a p53 mimic which has some similarity to Nutlin-2 for which there is both a structure and reported binding affinity. They also identified key hot-spot residues in the hDM2 binding site (L54, I61, M62, G58, and V93)(Zhong and Carlson 2005).

5.2.2 Molecular dynamics for studying protein-protein interactions

Computational methods have been applied to many aspects of the study of protein-protein interactions. Perhaps the most widely known is the Critical Assessment of Predicted Interactions (CAPRI)(Janin and Wodak 2007). However, computational techniques have also been applied to the investigation of interactions of proteins with peptides. An excellent review of the field is provided by Russell *et al.*(Russell *et al.* 2009). Molecular dynamics has specifically been used to look at PDZ interactions(Basdevant, Weinstein, and Ceruso 2006) as well as SH2 interactions(Gan and Roux 2009). Inspired by Kollman and Massova who used a GBSA model to perform Computational Alanine Scanning Mutagenesis of the hDM2-p53 system(Massova and Kollman 1999), Kortemme and Baker have applied a relatively simple physical model of protein-interactions to the hDM2-p53 system(Kortemme and Baker 2002). Kollman and Massova performed 400 ps simulations in TIP3P water and took 400 snapshots at 1 ps intervals. They then used a continuum solvent model to apply free energy calculations, which allowed them to successfully correlate their results with observed changes in IC_{50}s. Kortemme and Baker show that their results are comparable in accuracy to that of Kollman and Massova, but the computation has the benefit of being easier to perform and requiring less computing power. The method of Kollman and Massova has the benefit with respect to that of Kortemme and Baker in that it has the potential to be more widely applicable to hDM2-ligand interactions since it is not trained on a particular set of data(Massova and Kollman 1999). More recently

Kalid and Ben-Tal applied a GBSA technique based on sampling conformations from implicit solvent simulations to the hDM2 system and a range of p53 peptidomimetic compounds which is very similar to that initially performed by Kollman and Massova(Kalid and Ben-Tal 2009). They report a high correlation between their values for ΔG and experimentally reported pK_d for p53 peptidomimetics. The authors report that peptide models were built using the p53 peptide in 1YCR as a model with which to modify who used 100 data points sampled at 2 ps intervals between 200 ps and 400 ps of simulation. A more comprehensive study of the hDM2 and related MDMX system has been performed by Carotti *et al.* who used several tools such as Essential Dynamics and Linear Discriminant analysis to try to identify key residues involved in p53 binding(Carotti *et al.* 2009).

5.2.3 Dihedral sampling for alchemical free energy calculations

The overall aim of the project is to use the technique of alchemical free energy calculations to the hDM2-oligoamide system. Previously the technique has been successfully applied to the T4 lysozyme system as detailed by Mobley *et al.* (Mobley, Chodera, and Dill 2007). Mobley *et al.* identified several key factors that can adversely affect the result of an alchemical free energy calculation. These will be discussed in further detail in the proceeding chapter detailing the application of free energy calculations to the hDM2-oligoamide system. Immediately the key important result is that inadequate sampling of dihedral angles can adversely affect the free energy calculation.

5.2.4 Study Aims

Previously we used docking methods to develop starting conformations of an oligoamide compound bound to hDM2. In this chapter we will continue to use the nine conformations that have been previously been developed. We use the same nomenclature in this chapter to refer to these conformations.

We aim to identify a set of force field parameters that can adequately describe hDM2 systems for which structural data currently exists. These parameters should be suitable for transfer to the hDM2 oligoamide systems previously developed, and additionally be compatible with planned free energy calculations. These initial simulations will provide a benchmark with which to compare hDM2 oligoamide simulations.

Finally the sampling of certain parameters (particularly dihedral angle sampling) will be studied to ensure that the system is converging. With the overall aim of identifying a protocol that will allow free energy calculation to be applied to the system using the AMBER99sb/GAFF force fields.

5.3 Methods

5.3.1 Constructing systems for MD simulation

5.3.1.a Preparation of structures

All structures were taken from the Protein Data Bank (PDB)(Berman *et al.* 2002), in cases where multiple chains were present a single hDM2 chain was selected: 1T4E-B(Grasberger *et al.* 2005); 1T4F-M(Grasberger *et al.* 2005); 1YCR-A(Kussie *et al.* 1996); 1RV1-A(Vassilev *et al.* 2004); 1Z1M(McInnes *et al.* 2005) model 9. Also selected was the corresponding bound ligand where appropriate: 1T4E-A(Grasberger *et al.* 2005); 1T4F-P(Grasberger *et al.* 2005); 1YCR-B(Kussie *et al.* 1996); 1RV1-A(Vassilev *et al.* 2004). Water molecules were removed in all cases and protonation states were manually assigned. Ligand molecules were parameterized with GAFF(Wang *et al.* 2004) parameters and AM1BCC charges using the default settings from the acpype front end to Antechamber(D.A. Case, T.A. Darden, T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke *et al.* 2004).

The grompp program from GROMACS was used to assign AMBER99sb(Hornak *et al.* 2006) force field parameters from the ffamber ports(Sorin and Pande 2005). Version 3.3.1 of grompp and ffamber force field ports was used for initial MD simulations and version 4.0 was used for oligoamide simulations.

5.3.1.b Initial MD simulations

Initial MD simulations were performed using GROMACS 3.3.1(Lindahl, Hess, and Spoel 2001) using the 2.6.16.60-0.31-smp Linux kernel running on x86_64 hardware. All structures were minimised to a tolerance of 100 kJ mol^{-1} nm^{-1} with an initial step size of 0.01 nm for a maximum of 5000 steps of L-BFGS minimisation with 10 correction steps, followed by a maximum of 500 steps of steepest descent minimisation. Minimisation was followed by 10 ps of isothermal dynamics followed by 100 ps of isothermal/isobaric equilibration using the Berendsen algorithms. Production simulations were run for a total of 10 ns. In the latter two stages pressure coupling was performed using a Berendsen barostat with reference pressure of 1 atm, compressibility of $4.5x10^{-5}$ bar^{-1} and relaxation time of 0.5 ps.

All simulations used the stochastic integrator with reference temperature 300K and relaxation time 0.1 ps for the entire system, with an integration step size of 2 fs. PME parameters are the same as those used in the work of Mobley *et al.* (Mobley, Chodera, and Dill 2007), PME spline order of 6, relative tolerance of $1x10^{-6}$ and a Fourier spacing of 0.1 nm. A long range dispersion correction is also applied for energy and pressure, to correct for the effect of truncating the long-range dispersive interactions. A Lennard Jones function with switching between 0.8 nm and 0.9 nm was used for the van der Waals interactions. The neighbour list was set to 1 nm and was updated every 10 simulation steps. All bonds with H-atoms were constrained using the LINCS algorithm with highest order expansion of the constraint coupling matrix of 12. SETTLE is used to constrain water bonds and angles.

5.3.1.c Oligoamide MD simulations

An updated version of GROMACS, version 4.0.4(Hess *et al.* 2008) was used for all oligoamide MD simulations using conformations generated from the second round of Autodock docking carried out in the previous chapter (5 anti-parallel and 4 parallel conformers labelled conf1, 2, 3, 7, 8 and conf4, 9, 10, 11 respectively). The same simulation scheme as previously discussed was used with two minor changes. Firstly the maximum number of steepest descents minimization steps was increased from 500 to 2000 steps. Pressure coupling was altered to use Parrinello-Rahman with a relaxation time of 5.0 ps, 1 atm, compressibility $4.5x10^{-5}$ bar^{-1}.

5.3.2 Analysis of GROMACS simulations

GROMACS simulations were analyzed using 4 key tools to determine: the RMSD, throughout the time-course of the simulation, from the initial structure after two rounds of minimization as described previously (g_rms); the RMSF of individual residue Cα atoms from the initial structure after two rounds of minimization (g_rmsf); the number of pairs of atoms that are with 3.5 Å where one atom is part of the hDM2 structure and the second is part of the bound ligand structure (g_hbond); the difference in the distance between the centre of mass of the hDM2 molecule and the bound ligand molecule compared to the initial structure after two rounds of minimization (g_dist).

5.3.3 Dihedral analysis

5.3.3.a Distribution graphs and traffic light figure.

The distribution of dihedral angles over 20 ns of production simulation was plotted for each angle of all contacting residues and each dihedral present in the oligoamide compound shown in figure 5.9 for each of the 5 anti-parallel and 4 parallel starting conformations. Additionally the starting value of each dihedral was marked on the distribution. We analysed the distributions such that dihedral

angles which are sampled in most simulations but missing in others can be identified. Angles are labelled as: 'well sampled' where all simulations sample the same distribution; 'mostly well sampled' where all but one simulation samples the same distribution, or some peaks are considerably different in height but still sampled and located at the same angle; and 'possible sampling problem' where peaks are missing from more than one simulation indicating that some starting conformations can access dihedral angles that others may not be able to access.

5.3.3.b Autocorrelation analysis

Autocorrelation functions of length 10 ns (from simulations of length 20 ns) were generated for each χ angle from hDM2 binding residues shown in figure 5.10 for 5 anti-parallel and 4 parallel starting conformations of the Phe-Nap-Leu conformer, with the autocorrelation function $y(x) = < \cos(\chi(\tau)) \cdot \cos(\chi(\tau+t)) >$ being fitted to an exponential of the form $y=\exp(-x/\tau)$ using the g_chi program from GROMACS 4.0.4. Numerical integration of the exponential, also carried out using g_chi, yields the relaxation time for the χ angle.

5.3.4 Spatial sampling method

Spatial sampling was analysed by projecting the position of each of the three ether oxygen atoms from 20 ns simulations at time intervals of 10 ps onto a plane defined by the Cα atoms of Tyrosine 56, Methionine 62 and Valine 93. These three atoms lie in the periphery of the binding site and define a plane that cuts through the site at a roughly constant depth. A Python program using the Numpy toolkit was written that solves the equations:

$$l_a + (l_b - l_a)\, t \tag{39}$$

$$p_0 + (p_1 - p_0)\, u + (p_2 - p_0)\, v \tag{40}$$

Here l_a and l_b are points on a line (l_a is the position of an ether oxygen, and l_b is a point defined by following a vector along the unit normal to the plane from l_a) and p_0, p_1, p_2 are points on the plane (Cα atoms from Tyr 56, Met 62 and Val 93), and t, u, v are arbitrary real numbers.

5.3.5 Cluster analysis of Oligoamide conformations

The GROMACS 4.0.4 program g_cluster was used to generate clusters using the GROMOS clustering method to generate clusters with a minimum RMS of 1.5 Å. The GROMOS clustering method takes a structure and counts all structures within the RMS threshold, it then takes the largest cluster and chooses the structure with the largest number of neighbours, and eliminates this and additional cluster members from the structures in the pool, which is repeated until the pool of structures is empty(Daura *et al.* 1999). Cluster size (number of members of each cluster), and the id of each cluster member was generated for the pooled conformations taken at 10 ps intervals between 3 ns and 20 ns from 5 anti-parallel (1, 2, 3, 7, 8). The same was repeated for the 4 parallel (4, 9, 10, 11) starting conformations.

5.4 Results and Discussion

5.4.1 Initial hDM2 MD simulations

Initial simulations were carried out for hDM2 in complex with the four compounds for which there are crystal structures. This allows us to determine sensible simulation protocols and parameters for use when simulating oligoamide compounds. Five replicates of each simulation allowed us to determine the expected variability in our simulations. Parameters followed throughout the length of the 10 ns simulations were: RMSD from the minimized structure; RMS fluctuation of individual residue Cα from their minimized location; number of

contacting atom pairs from each of hDM2 and complexed ligand within 3.5 Å; difference in centre of mass distance of hDM2 from its complex relative to the distance observed in the minimized structure.

5.4.1.a Stability of hDM2-compound simulations

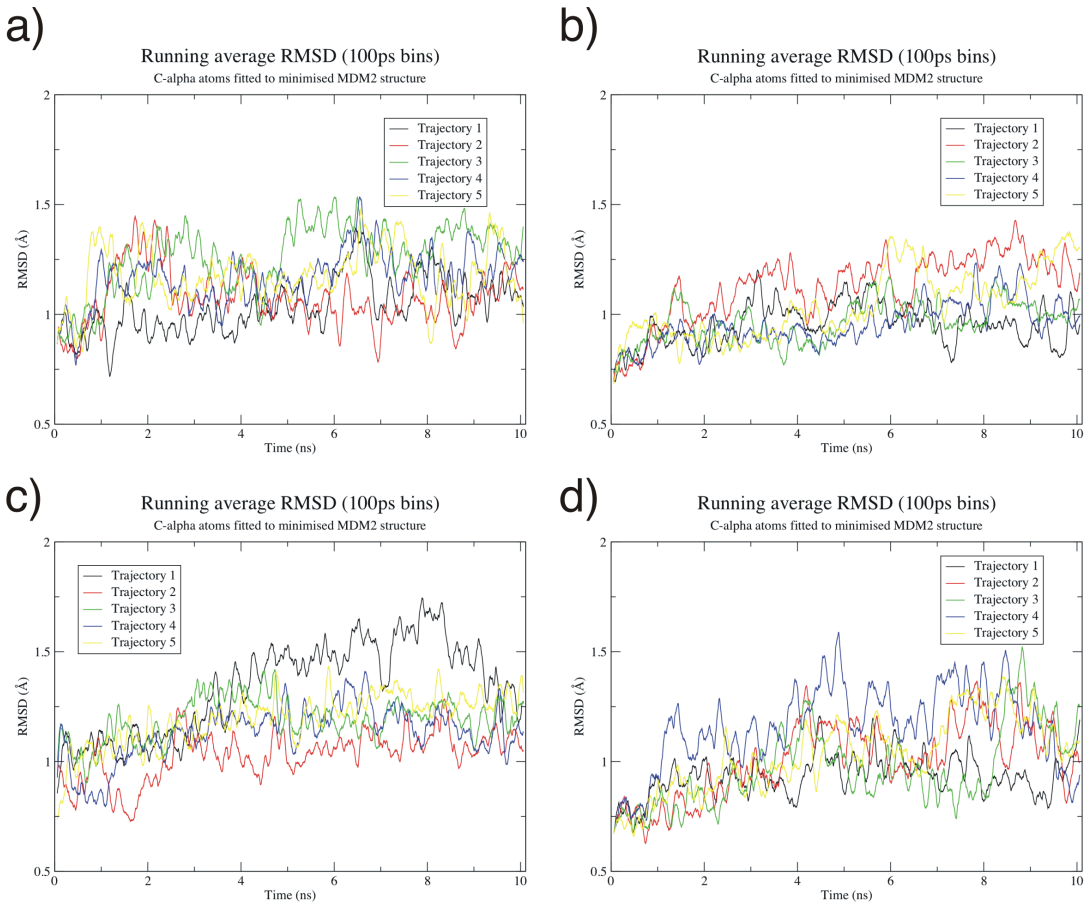

Figure 5.1: Running averages (100 ps window) of the RMSD (Å) for 5 replicates of hDM2 simulated for 10 ns in complex with: a) wild-type helix; b) high-affinity helix; c) benzodiazepinedione; d) nutlin-2.

The first property studied was the RMS distance of the structures from the coordinates of their minimized structure. The results are shown in figure 5.1, with hDM2 in complex with: a) wild-type p53 helix; b) high-affinity p53 helix; c) benzodiazepinedione; d) nutlin-2. Generally we see that the RMSD will increase from the value at t = 0 ps and approach a fluctuating but averaging constant

value. This indicates that the complex requires some time to relax from the structure that is presented in the PDB, but that the complex is well behaved. It is therefore unlikely that the protein is unfolding; the protein is likely to be in a dynamic equilibrium sampling states in a global energy minimum. In all but two cases the RMSD never exceeds 1.5 Å, typically simulations with RMSD < 2.5 Å are said to be stable. In the two cases where RMSD does exceed 1.5 Å it remains well below the 2.5 Å cutoff and returns to less that 1.5 Å. These two cases are trajectory 1 in the benzodiazepinedione simulation (figure 5.1), and trajectory 4 in the nutlin-2 simulation (figure 5.1). It is also clear from figure 5.1 that all of the replicates are behaving in a similar manner, which also indicates that we are using suitable simulation parameters.

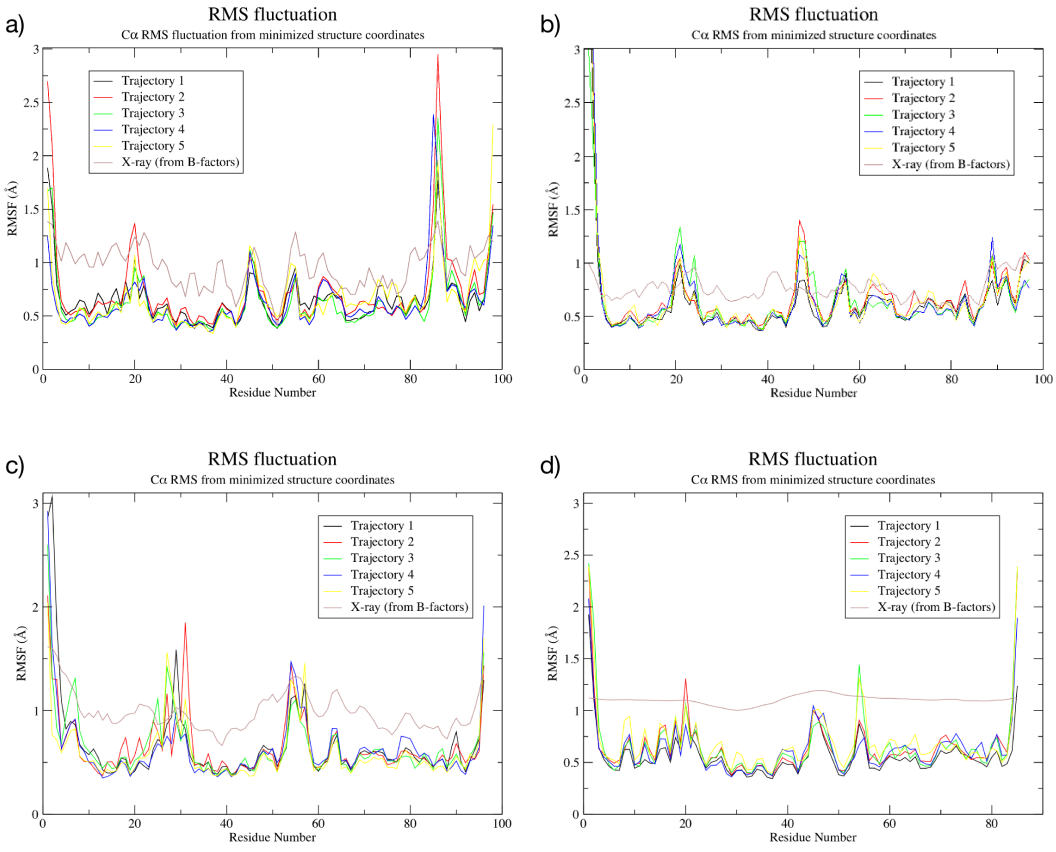

Figure 5.2: RMS fluctuation compared to experimental b-factor (as specified in the corresponding PDB file), using the relationship in equation 39 for 5 replicates of hDM2 simulated for 10 ns in complex with: a) wild-type helix (1YCR); b) high-affinity helix (1T4F); c) benzodiazepinedione (1T4E); d) nutlin-2 (1RV1).

In figure 5.2 we investigate the RMS fluctuation of Cα atoms from hDM2. Additionally we use the relationship(Willis and Pryor 1975):

$$RMSF = \sqrt{3B/8\pi^2} \tag{41}$$

Where B is the experimentally determined b-factor extracted from the atom records of the PDB coordinate file. Figure 5.2 compares the experimentally observed b-factor in beige to the RMSF observed in our MD simulations. It should be noted that there is likely to be some discrepancy between this model and experiment due to the fact that crystal structures are usually determined at less than 300 K (the temperature at which the MD simulations are carried out), as well as the possibility of influences from crystal packing.

When discussing the RMSF of residues we ignore the first and last few residues from discussion, since they are not restrained by the motion of surrounding residues they are much more free to move than one would expect to observe in most crystal structures. We see in figure 5.2a the RMSF plot for hDM2 bound to wild-type p53. We see excellent agreement between simulations, and furthermore the simulations agree well with the RMSF calculated from the experimentally determined B-factor. The maximum values from the RMSF are of the order 2.5 Å to 3 Å which is approximately double the RMSF calculated from the experimental b-factor. We observe a similar picture in figure 5.2b in the case of hDM2 bound to high-affinity p53. It is important to note that each of the five replicates are in agreement as to the value of the RMSF during the simulation. Figure 5.2c shows the RMSF for hDM2 bound to benzodiazepinedione, which shows a distribution of RMSF values that closely follows the RMSF calculated from the experimental b-factor. Figure 5.2d shows that the RMSF for hDM2 bound to nutlin-2 calculated from the experimental b-factor is nearly flat with a value of approximately 1.2 Å, this perhaps indicates that the experimental b-factors from this structure might be

less reliable than for the other three structures. The simulations of nutlin-2 show that the RMSF remains below this 1.2 Å value except for a slight deviation by two residues.

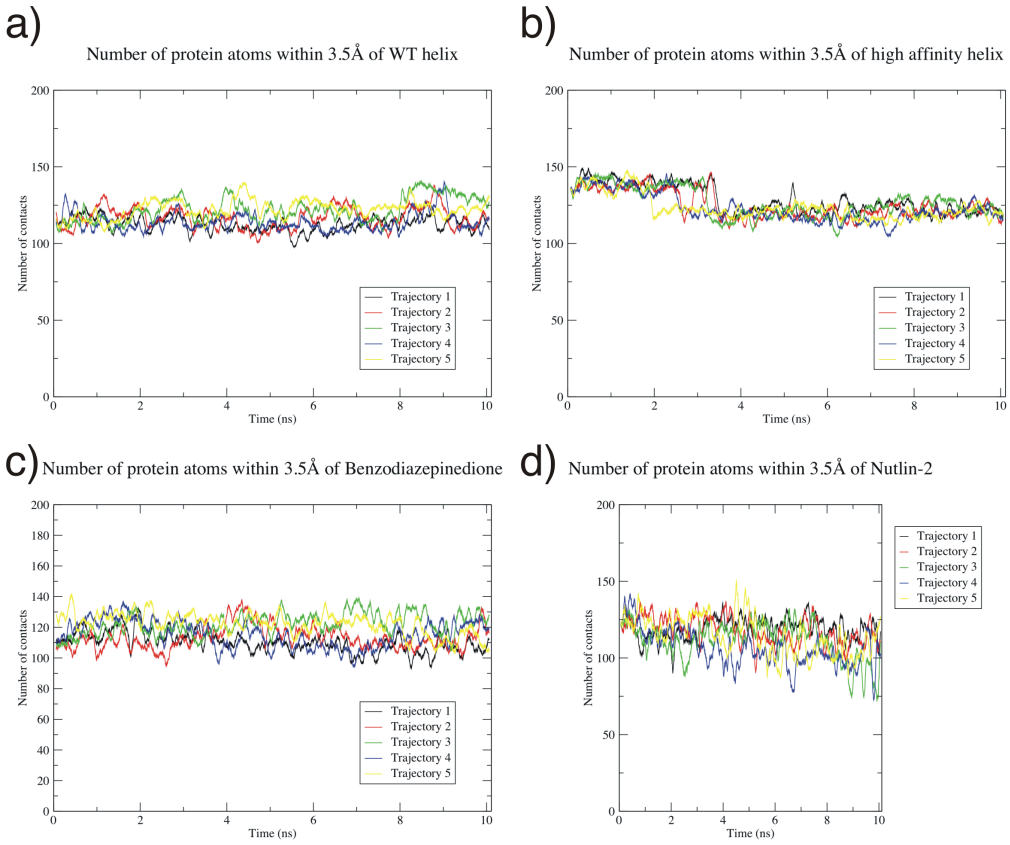

Figure 5.3: Total number of contacts, pairs of atoms that are within 3.5 Å of each other, one from hDM2 and one from complex structure: a) wild-type helix; b) high-affinity helix; c) benzodiazepinedione; d) nutlin-2.

We also looked at the total number of contacts (figure 5.3), where a contact is defined as a pair of atoms, one from hDM2 and one from the ligand which exist within 3.5 Å of each other. Since the van der Waals radius of a carbon atom is approximately 1.7 Å this means that we are essentially counting the number of contacts of the complexed molecule with the hDM2 protein. This gives us a very rough measure of how tightly bound the ligand is. In the case of co-crystallised ligands we might expect the number of contacts to stay roughly constant for the

duration of the simulations, since it should already be in a global minimum. It should be noted that in the high-affinity helix simulations in figure 5.3b all simulations start with a larger number of contacts than they finish with, moving from approximately 140 contacts to approximately 110 contacts. All simulations make this transition between 2 ns and 4 ns of simulation time. Once the high-affinity helix simulation settles to its new number of contacts this value is roughly the same as that of the wild-type helix. This is unsurprising since whilst the wild-type helix is slightly longer than the high-affinity helix they both share the same contact epitope. Much of the increase in affinity of the high-affinity helix appears to be from decreasing the overall helix length to the minimum amount required to maintain helicity, whilst substituting residues on the solvent exposed face of the peptide for those that are helix stabilisers with a high helix propensity and additionally have hydrophillic properties(Pace and Scholtz 1998). It is also of interest that the total number of contacts for both benzodiazepinedione and nutlin-2; figure 5.3c and 5.3d respectively; remain constant with values around 110. This is directly comparable to both of the p53 helices. When combined with the images of the binding site and compound superpositions from the previous chapter figure 4.2 and figure 4.1 it becomes clear why this is. All compounds are targeting essentially identical regions of the binding pocket and the 3.5 Å cutoff for contacts is very strict, meaning that peripheral atoms are not counted.

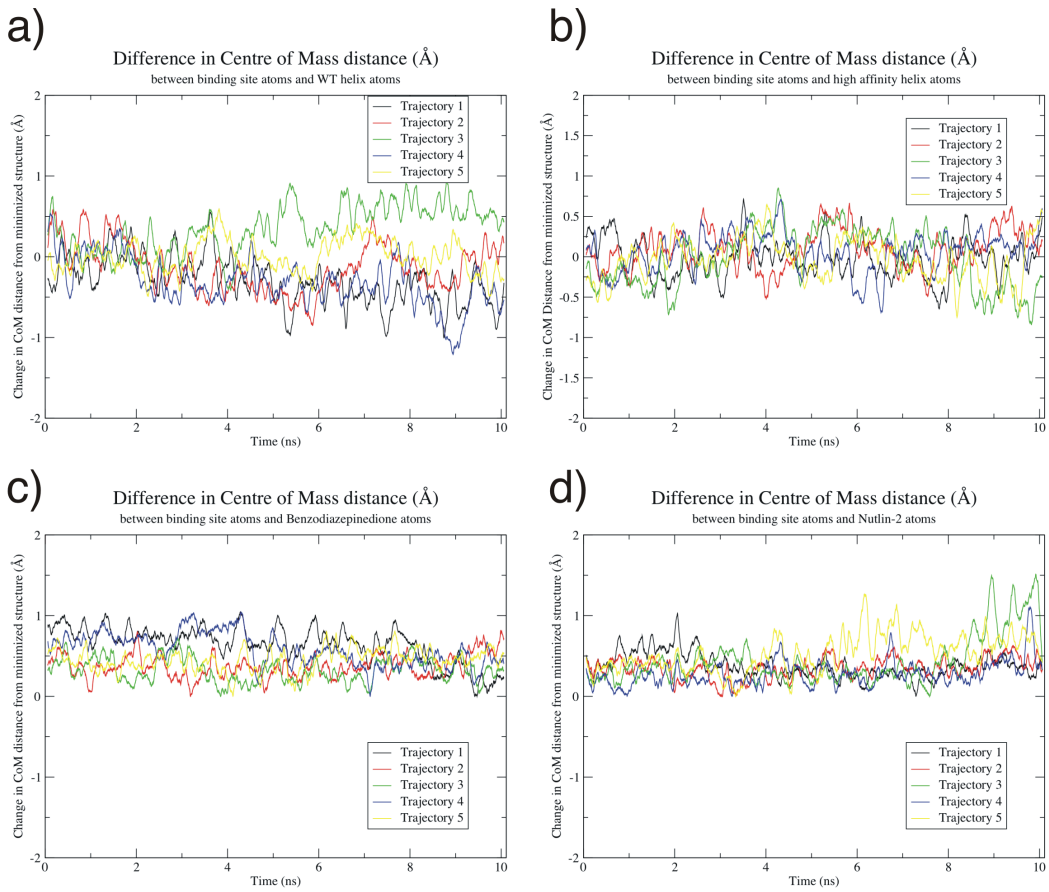

Figure 5.4: Difference in centre of mass distance (Å) from the distance measured after the initial structure undergoes the two stages of minimization described in the methods, between hDM2 and complex molecule: a) wild-type helix; b) high-affinity helix; c) benzodiazepinedione; d) nutlin-2.

In figure 5.4 we first calculate the distance between the centre of mass of the hDM2 protein and the centre of mass of the ligand binding partner. We then subtract the centre of mass distance observed in the minimized structure from all further observations. This allows us to determine whether the binding partner might be binding more deeply or alternatively less tightly in the pocket. In the case of simulations starting from known structures this value should remain close to zero since the system is likely to already be at a minima. In figure 5.4a we see that the wild-type helix distance varies between -1 Å and +1 Å and may be the most variable of all the families of simulations. The high-affinity helix in figure 5.4b shows less variability with minima and maxima closer to -0.5 Å and +0.5 Å. The

benzodiazepinedione and nutlin-2 families of simulations in figure 5.4c and 5.4d respectively, appear to slightly increase centre of mass distance. Whilst the two peptide simulations in a) and b) oscillate around a zero value, the two small-molecule simulations in c) and d) seem to oscillate around mean values of approximately +0.2 Å which is within the error range of the simulation. These values are nevertheless also very stable.

The basic analysis of initial molecular dynamics provided in the previous section serves two purposes. Firstly, it allows us to be confident that the parameters chosen for the simulation are suitable for the system in question. Secondly it allows us to gain an idea of reasonable values for observables when comparing to molecular dynamics simulations using similar parameters. Thus we have a basis from which to compare the stability of the oligoamide simulations.

5.4.2 hDM2-oligoamide simulations

A slightly altered set of parameters was used to simulate hDM2 oligoamide complexes. The change in parameters was designed to be compatible with free energy calculations in GROMACS when Verlet integrators are provided as standard. Whereas in the initial MD simulation work we simulated 5 replicates for each starting conformation, we instead choose to use just one replicate and a group of docked conformations that are similar. This is due to the reasons discussed in the section on generating starting conformations. Chiefly since we are unsure of the correct low energy conformation we use several possible conformations for study. Use of five replicates of nine conformations is computationally intractable, whereas just simulating nine conformations is still possible. The starting conformations used are shown in figure 5.5 for anti-parallel conformations and figure 5.6 for parallel conformations.

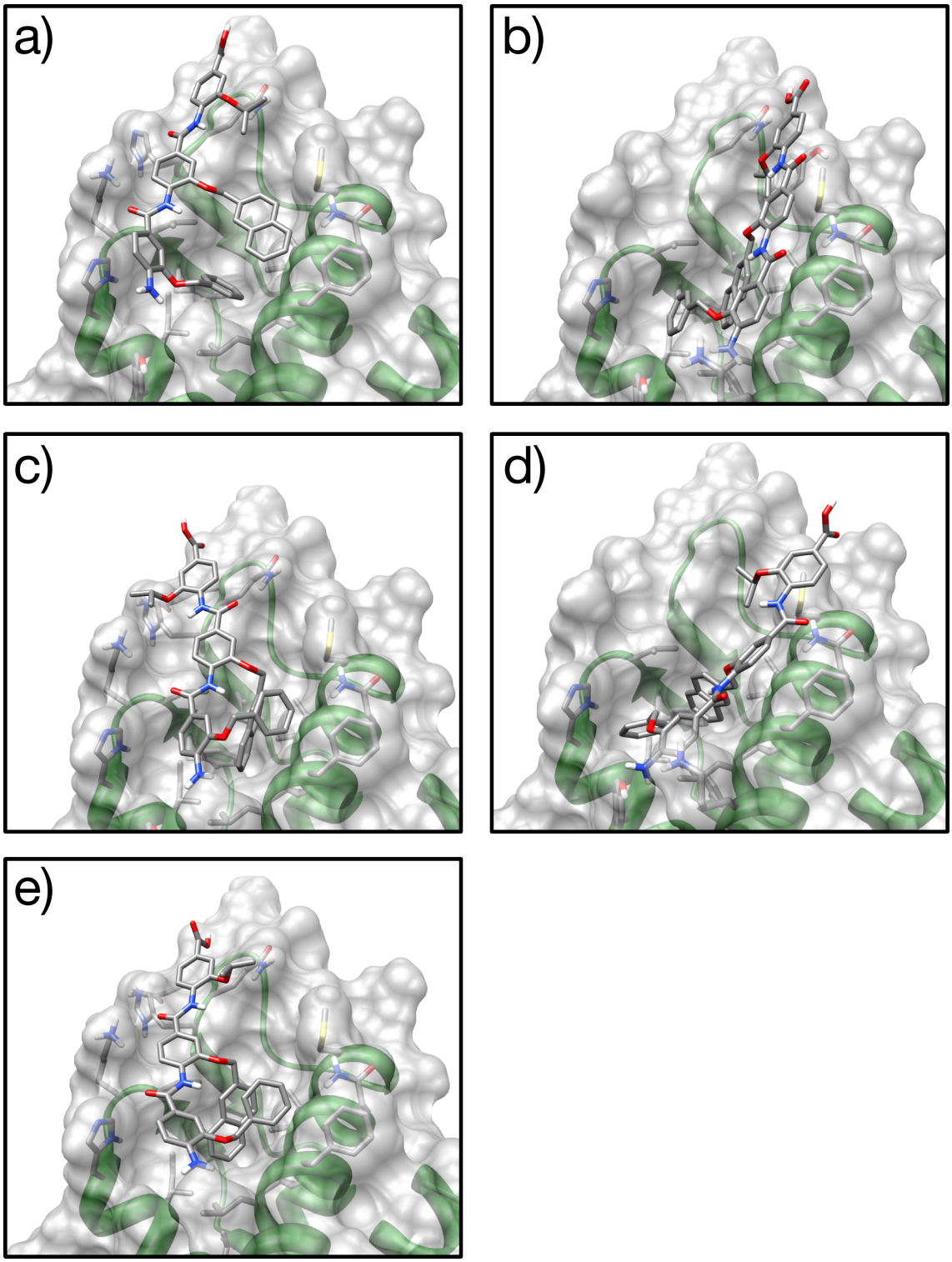

Figure 5.5: Anti-parallel docked conformations identified by Autodock in the previous chapter and used in the molecular dynamics study detailed here. In order to allow easy discussion conformations are labelled: a) Conf 1; b) Conf 2; c) Conf 3; d) Conf 7; e) Conf 8.

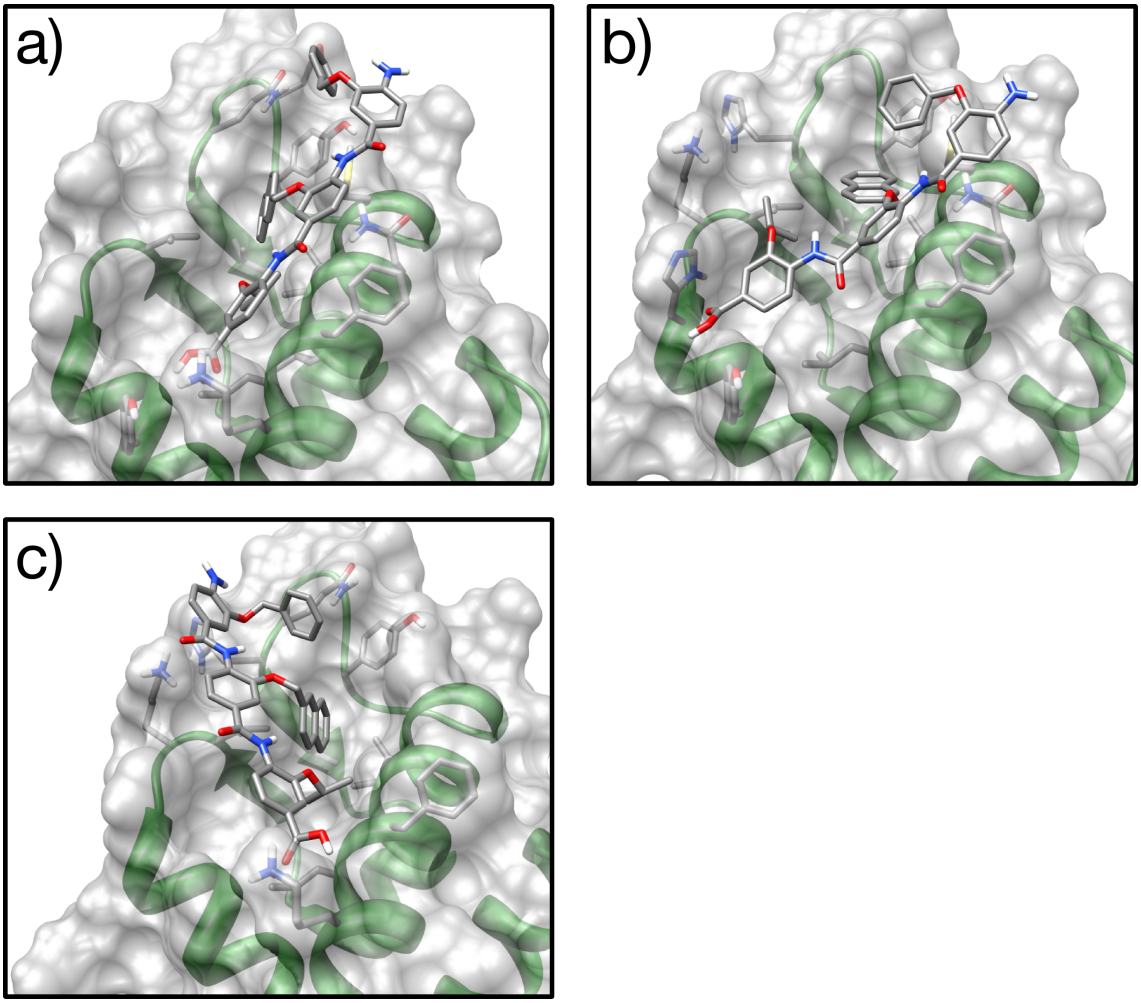

Figure 5.6: Parallel docked conformations identified by Autodock in the previous chapter and used in the molecular dynamics study detailed here. In order to allow easy discussion conformations are labelled: a) Conf 4/9; b) Conf 10; c) Conf 11.

5.4.2.a Stability of hDM2-oligoamide simulations

In figure 5.7 and figure 5.8 we see RMSD/RMSF and number of contacts/difference in centre of mass respectively. These two figures have parallel oligoamide starting configurations on the left, and anti-parallel oligoamide starting configurations on the right.

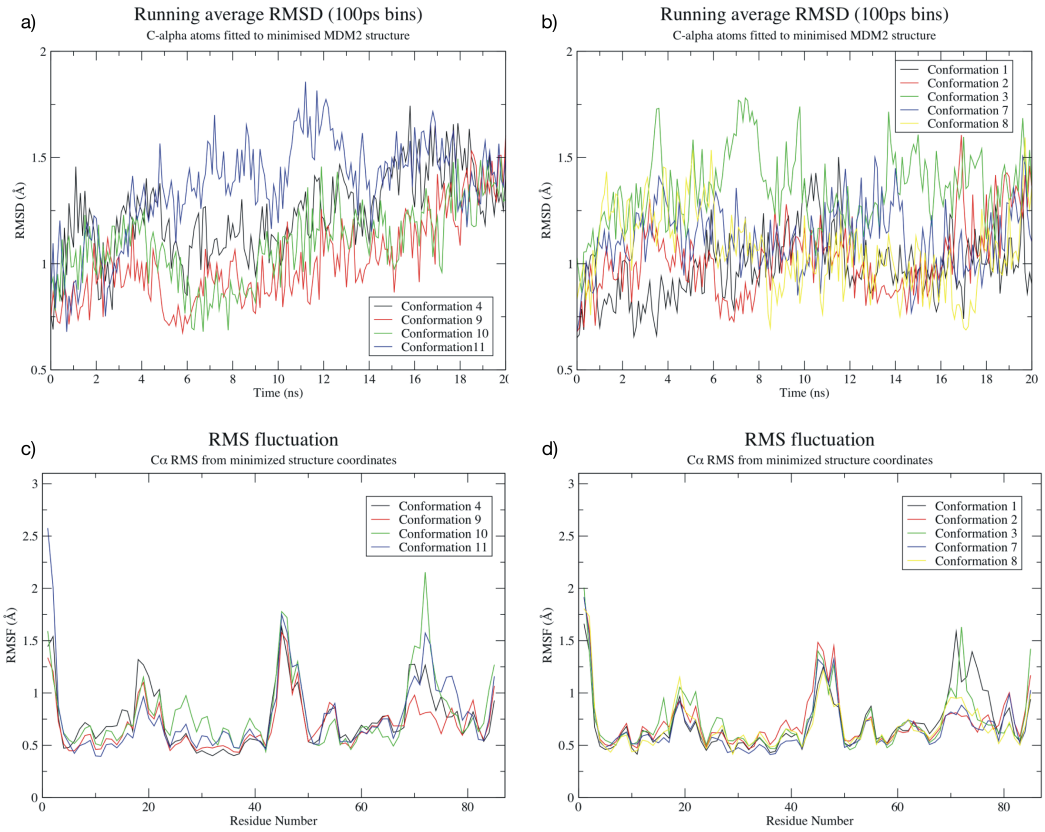

Figure 5.7: Behaviour of parallel (left) and anti-parallel (right) Phe-Nap-Leu conformations: a) RMSD relative to initial minimized parallel conformation; b) RMSD relative to initial minimized anti-parallel conformation; c) RMS fluctuation of C-alpha atoms from initial minimized parallel conformation; d) RMS fluctuation of C-alpha atoms from initial minimized parallel conformation.

Both parallel and anti-parallel starting conformations have an average RMSD of less than 1.5 Å, with conformation 4 and conformation 3 occasionally slightly exceeding this value. In all cases RMSD remains below 2 Å. RMS fluctuation behaviour is similar for both parallel and anti-parallel simulations. Fluctuation generally remains below a maximum of 2 Å. For the protein hDM2 there are areas of greater plasticity from residue 18-22, 44-47 and 69-77. The first two more plastic regions are common to all simulations, whilst parallel conformation 9 and anti-parallel conformations 2, 7 and 8 appear to be far more rigid than their counterpart simulations. We can compare the values in these simulations to the same regions of hDM2 from our initial MD simulations. The residue numbering

scheme in figure 5.7 is slightly different, so the residue range for the initial MD simulation will be compared to those from figure 5.8. It is difficult to make comparison with the RMSF for hDM2 in complex with the p53 peptides shown in figure 5.7a and 5.7b due to the large fluctuations in these simulations. However, it is possible to make comparison to the simulation of another ligand benzodiazepinedione from figure 5.7c and the nutlin-2 simulation from figure 5.7d. We see that the first plastic region is also visible for the hDM2 benzodiazepinedione complex for equivalent residues 28-32, and the hDM2 nutlin-2 complex for equivalent residues 18-22. These are residues: alanine; glutamine; lysine; asparagine and threonine, and they define a loop region on the far side of hDM2 relative to the p53 binding site. The second plastic region is also visible in the hDM2 benzodiazepinedione complex for equivalent residues 54-57, and in the hDM2 nutlin-2 complex for equivalent residues 44-47. These correspond to residues glutamic acid, lysine, glutamine, glutamine. These residues exist at the N-terminus of the high-affinity p53 peptide structure, with the glutamic acid in particular contacting the N-terminus of the helix. These residues also form a short loop between a pair of beta strands, which is likely to explain why the increased RMSF is observed in all structures from both the oligoamide simulations and the initial MD simulations. The third plastic region in the oligoamide system is between residues 69-77, however, equivalent higher RMSF regions are not visible in figure 5.7c residues 79-87 or figure 5.7d residues 69-77. These residues are lysine, glutamic acid, histidine, arginine, lysine, isoleucine, tyrosine, threonine and methionine. These residues all exist in proximity of the n-terminus of the p53 helix in the structure of hDM2 bound to the high-affinity p53 helix.

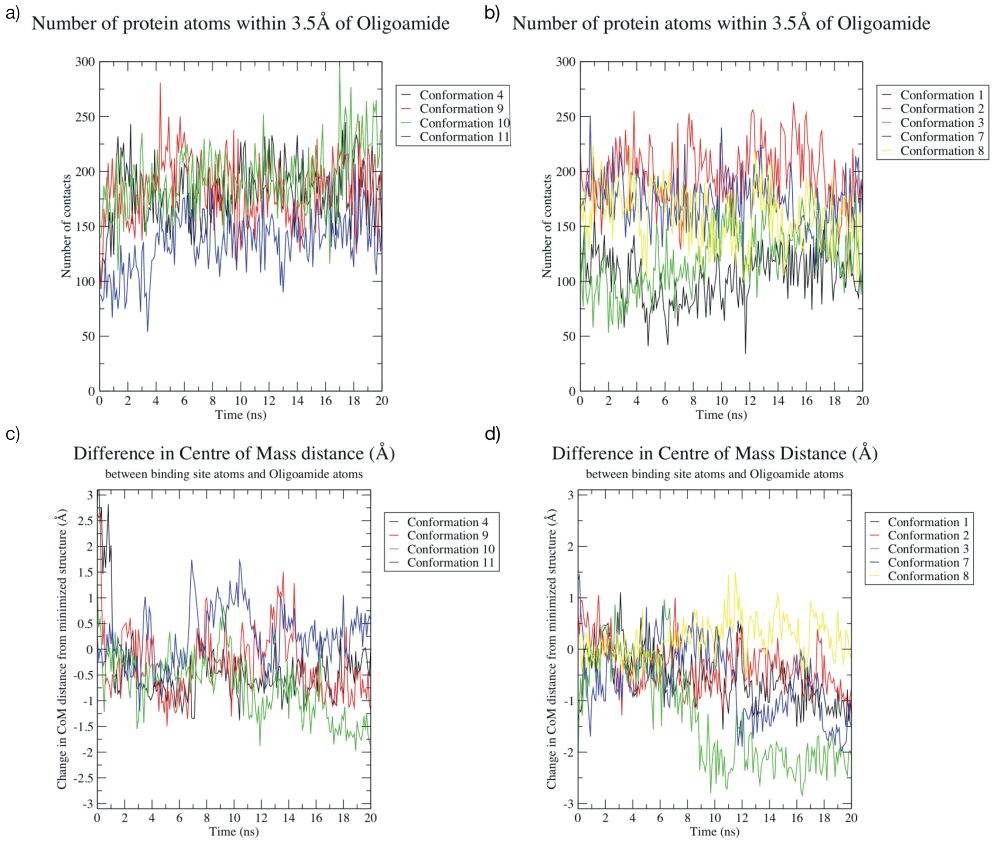

Figure 5.8: Behaviour of parallel (left) and anti-parallel (right) Phe-Nap-Leu conformations: a) RMSD relative to initial minimized parallel conformation; b) RMSD relative to initial minimized anti-parallel conformation; c) RMS fluctuation of C-alpha atoms from initial minimized parallel conformation; d) RMS fluctuation of C-alpha atoms from initial minimized parallel conformation.

There are larger fluctuations in the total number of contacts between hDM2 and oligoamide (figure 5.8) than in the simulations involving hDM2 and compounds for which there are X-ray structures (figure 5.3). We also notice that in the case of anti-parallel simulations there appears to be two clusters of contacts. Conformations 2, 3 and 7 make more contacts, averaging roughly 175. Whilst conformations 1 and 8 make fewer contacts, averaging roughly 115. In the case of the parallel conformations we see in figure 5.8 that conformation 4 and 9 take less than 1 ns and about 4 ns to reach their equilibrium values. Furthermore we observe that the total number of contacts tends to cluster towards a single average value after equilibration indicating the conformations 4 and 11 appear to

be stabilising towards a more favourable conformation. This is an interesting observation since it suggests that the starting conformations aren't the optimum in terms of number of contacts. Additionally in the case of these two conformations they are actually identical at the start of the simulation. This means that we must allow a reasonable amount of equilibration time (which could be as long as 5 ns) to allow simulations to reach their optimum values.

In the case of the initial MD simulations we observed in figure 5.4 that there was fluctuation but little deviation from the initial value of the difference in the centre of mass distance between hDM2 and complexed ligand. This was likely due to the fact that the compounds were already in their global minima. In the case of hDM2 bound to the Phe-Nap-Leu oligoamide, we already know that not all docked compounds can be in their global minima since we have a variety of low energy docked complexes. As a result it is reasonable to assume that this distance might decrease if the oligoamide manages to further optimize itself in the hDM2 binding site. Indeed figure 5.8 does show that in some simulations the average distance does tend to decrease. This is more pronounced in the anti-parallel simulations. The decreased distance between centre of masses is particularly obvious in the case of conformation 3 and is mirrored by an increased number of contacts. Conformation 3 shows a decrease in distance of 2 Å, and additional increase in the number of contacting atoms meaning that binding of the oligoamide to the protein is likely tighter as the system equilibrates.

5.4.2.b Traffic light analysis of dihedral angles

Sampling of dihedral angles can be one of the slowest observables to properly converge in molecular dynamics simulations so we investigated the timescales over which dihedral angles are sampled in our simulations. This includes dihedral angles from the hDM2 binding site side-chains and dihedrals present in the oligoamide compound. The importance of sampling dihedral angles is shown in the work by Mobley *et al.*(Mobley, Chodera, & Dill 2006). First we look at which

dihedral angles are likely to be well sampled, or poorly sampled on the timescales of our simulations. We then use the autocorrelation function of a dihedral angle to determine the relaxation time of the angle fit to a simple exponential model. The relaxation time corresponds to the average time the simulation needs to run for the a dihedral angle to ‘forgets' information about its previous value. We can then use this to guide whether we need to use several simulations with multiple starting angles, in the case of for large relaxation times, or whether a single starting conformation should allow comprehensive sampling of the relevant region of phase space. In this section we use a simple classification system to identify those dihedrals which are not sufficiently sampled. In order to do this we classify dihedral angles into one of three possible bins. Bins are determined from the results of analyzing the dihedral angle distribution over the course of the simulation (full results presented in supplementary information). The dihedral distribution is plotted for each of the simulations and compared to the other distributions. If all the distributions are in agreement the dihedral is classified as 'well sampled', which implies that the dihedral has been well sampled during the course of the simulation. If only one of the distributions differs significantly from the others, the dihedral is labelled as 'mostly well sampled'. If more than one distribution contains a region that differs significantly between simulations the dihedral is labeled as 'poorly sampled', which implies that more than one simulation does not sample a possibly important region of dihedral space. In this case it will be necessary to use multiple starting conformations to carry out free energy simulations as one of the rotameric states that is not sampled may contribute significantly to the free energy of interaction.

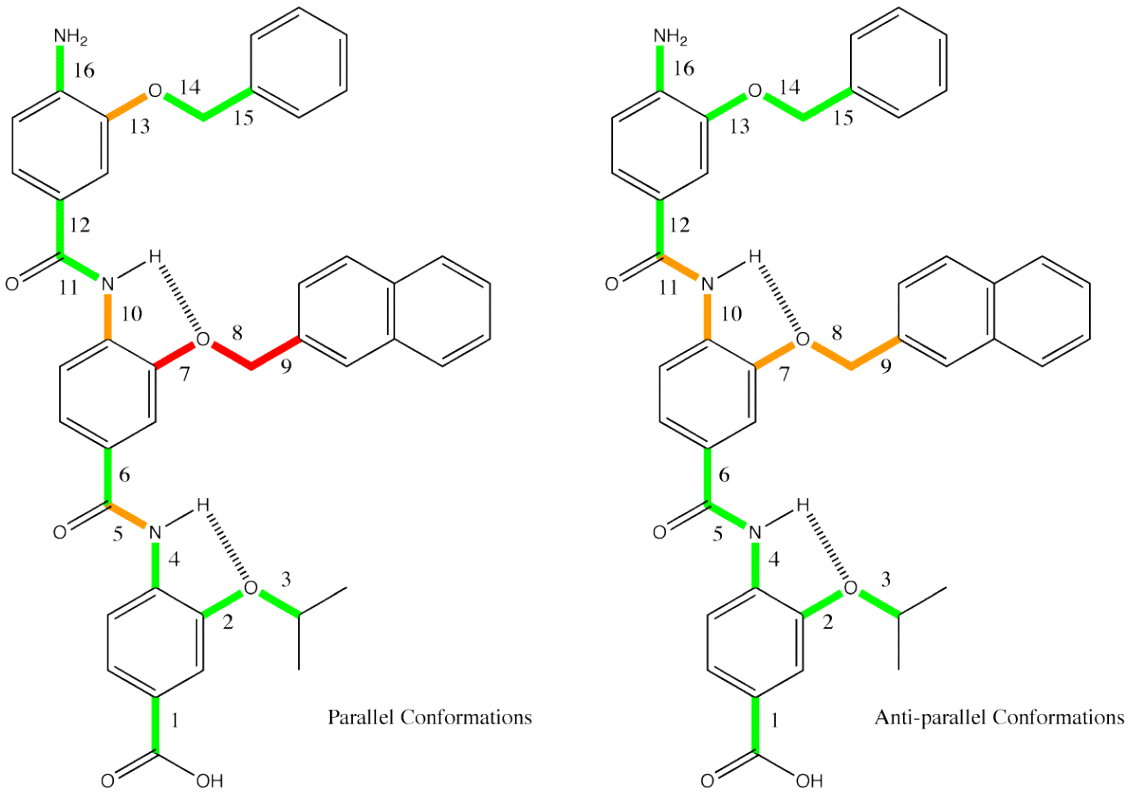

Figure 5.9: 2D representations of parallel and anti-parallel conformations of the Phe-Nap-Leu oligoamide with rotatable bonds shown in bold with colour: green (well sampled); orange (well sampled in all but one simulation); red (poorly sampled across simulations).

Figure 5.9 shows a representation of the parallel and anti-parallel Phe-Nap-Leu oligoamide compounds. Bonds with sampled dihedral angles are shown in bold, with a colour scheme representing the quality of sampling of the angle. From a total of 16 dihedral angles investigated, well sampled dihedral angles (green) are observed in 10 parallel and 11 anti-parallel dihedral angles. There are three mostly well sampled dihedral angles (orange) from parallel conformations and 5 mostly well sampled dihedral angles in the anti-parallel simulations. There are no poorly sampled dihedral angles (red) in the anti-parallel conformations, whilst the parallel conformations have poorly sampled dihedral angles for the three χ angles for the bonds attaching the 2-napthalene group. This data is derived from results not shown here.

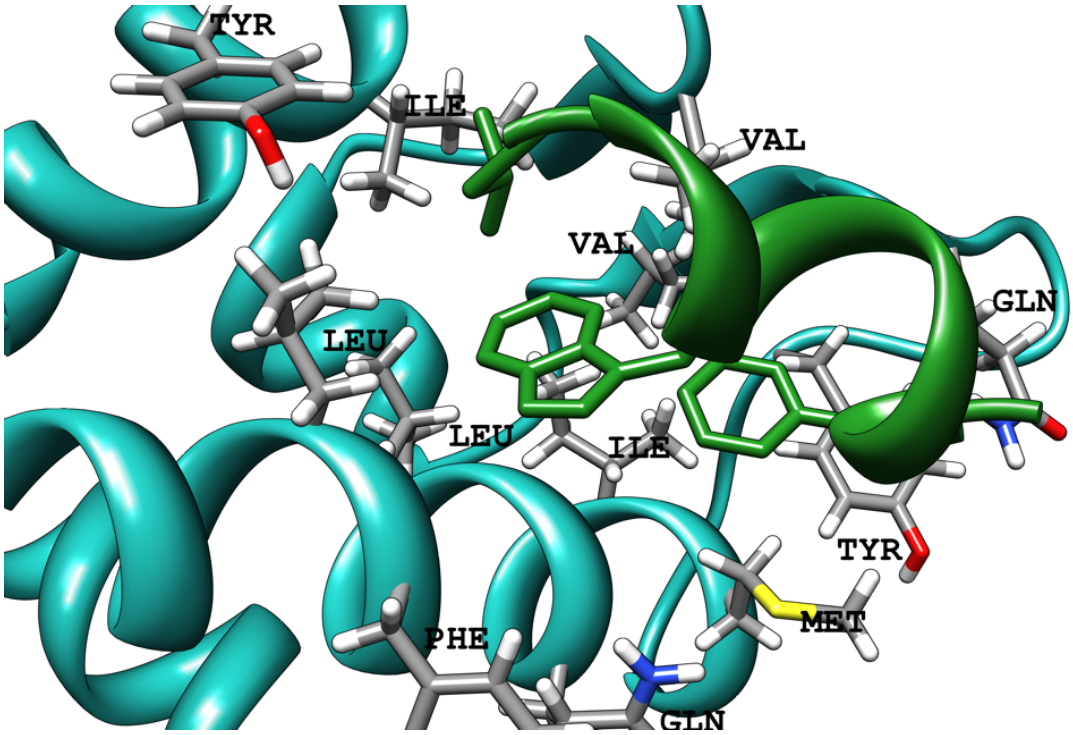

Figure 5.10: Binding site residues that are investigated in dihedral angle sampling analysis shown in table 1. hDM2 protein backbone shown in ribbon style(cyan); high-affinity p53 helix (green); residues (atom colours).

Figure 5.10 shows the binding site from hDM2 (cyan backbone, atom coloured side-chains) bound to the high-affinity p53 helix (dark green). It is immediately clear that there are no aromatic residues that might have restricted dihedral sampling due to say intermolecular π-π stacking effects, assuming that the oligoamide binding mode somewhat mimics that of the p53 peptide. However, it is still necessary to identify which residues are likely to be well sampled, and whether we expect any to be an issue. The results are shown in table 5.1, using the same three bin classification scheme as used previously for the oligoamide compounds dihedral angles.

Dihedral Identifier	Anti-parallel conformations (1, 2, 3, 7, 8)	Parallel conformations (4, 9, 10, 11)
LEU30 χ 1	Poorly sampled	Well sampled
LEU30 χ 2	Mostly well sampled	Mostly well sampled
PHE31 χ 1	Mostly well sampled	Well sampled
PHE31 χ 2	Well sampled	Well sampled
LEU33 χ 1	Well sampled	Well sampled
LEU33 χ 2	Mostly well sampled	Mostly well sampled
GLN35 χ 1	Mostly well sampled	Mostly well sampled
GLN35 χ 2	Mostly well sampled	Poorly sampled
GLN35 χ 3	Well sampled	Mostly well sampled
ILE37 χ 1	Well sampled	Well sampled
ILE37 χ 2	Mostly well sampled	Well sampled
MET38 χ 1	Well sampled	Poorly sampled
MET38 χ 2	Mostly well sampled	Mostly well sampled
MET38 χ 3	Well sampled	Well sampled
TYR43 χ 1	Poorly sampled	Well sampled
TYR43 χ 2	Poorly sampled	Well sampled
GLN48 χ 1	Mostly well sampled	Well sampled
GLN48 χ 2	Well sampled	Well sampled
GLN48 χ 3	Well sampled	Well sampled
VAL51 χ 1	Well sampled	Well sampled
VAL69 χ 1	Mostly well sampled	Well sampled
ILE75 χ 1	Poorly sampled	Mostly well sampled
ILE75 χ 2	Mostly well sampled	Mostly well sampled
TYR76 χ 1	Mostly well sampled	Well sampled
TYR76 χ 2	Mostly well sampled	Well sampled

Table 5.1: Summary of sampling of hDM2 binding site side-chain χ angles for 20 ns simulations with anti-parallel and parallel conformations of a Phe-Nap-Leu oligoamide compound. Residues in the binding site are shown in figure 5.10.

We see that dihedral angles are often well sampled or mostly well sampled in both parallel, and anti-parallel simulations. In fact from a total of 25 dihedral angles we see only two poorly sampled dihedrals in the case of parallel simulations, and four poorly sampled dihedrals in the case of anti-parallel simulations.

5.4.2.c Autocorrelation analysis of dihedrals

One of the important factors that will determine the accuracy of our free energy calculations is the length of time that we run them for, since we have already discussed that one of the major issues is ensuring proper sampling of binding site dihedral angles. We can use the autocorrelation function of a dihedral angle to determine the relaxation time of the angle if we fit it to a simple exponential model. The relaxation time allows us to find out how long a simulation would have to be run until a dihedral angle 'forgets' information about its starting value. That is we can use this to guide us as to whether we need to use several simulations with multiple starting angles (for large relaxation times), or whether a single starting conformation should allow comprehensive sampling of the relevant region of phase space.

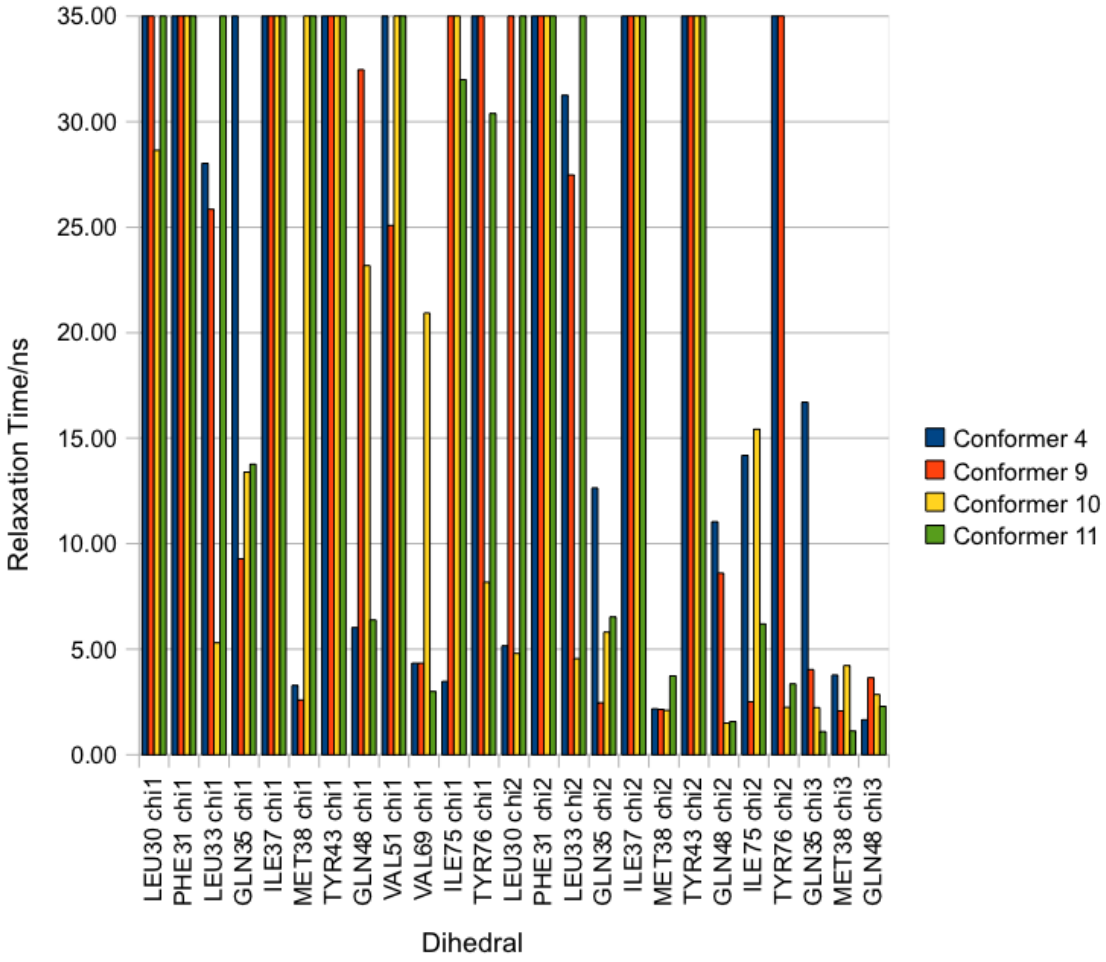

Figure 5.11: Relaxation times for dihedral angles from the hDM2 binding site for parallel conformations of bound oligoamide compounds as calculated by fitting a function of the form y=exp(-x/a) to the autocorrelation function for the dihedral angle.

In figure 5.11 we can see that the relaxation times for hDM2 binding site residues from parallel simulations is generally long with the first **X** angle from isoleucine 37 being the longest at 450 ns. However, it should be noted that any relaxation times longer than 20 ns (the length of the simulation) actually just have correlation times longer than 20 ns, but we can't say exactly how long this correlation time is. This length of simulation is intractable for all but the very longest of current molecular dynamics simulations. The second longest angle is the second isoleucine 37 χ angle with a relaxation time of 300 ns. Tyrosine 43 χ angles one and two are also

of concern with several simulations showing relaxation times in excess of 100 ns. However, we also see that many dihedral angles have relaxation times that are less than the duration of our simulations, which in this case is 20 ns.

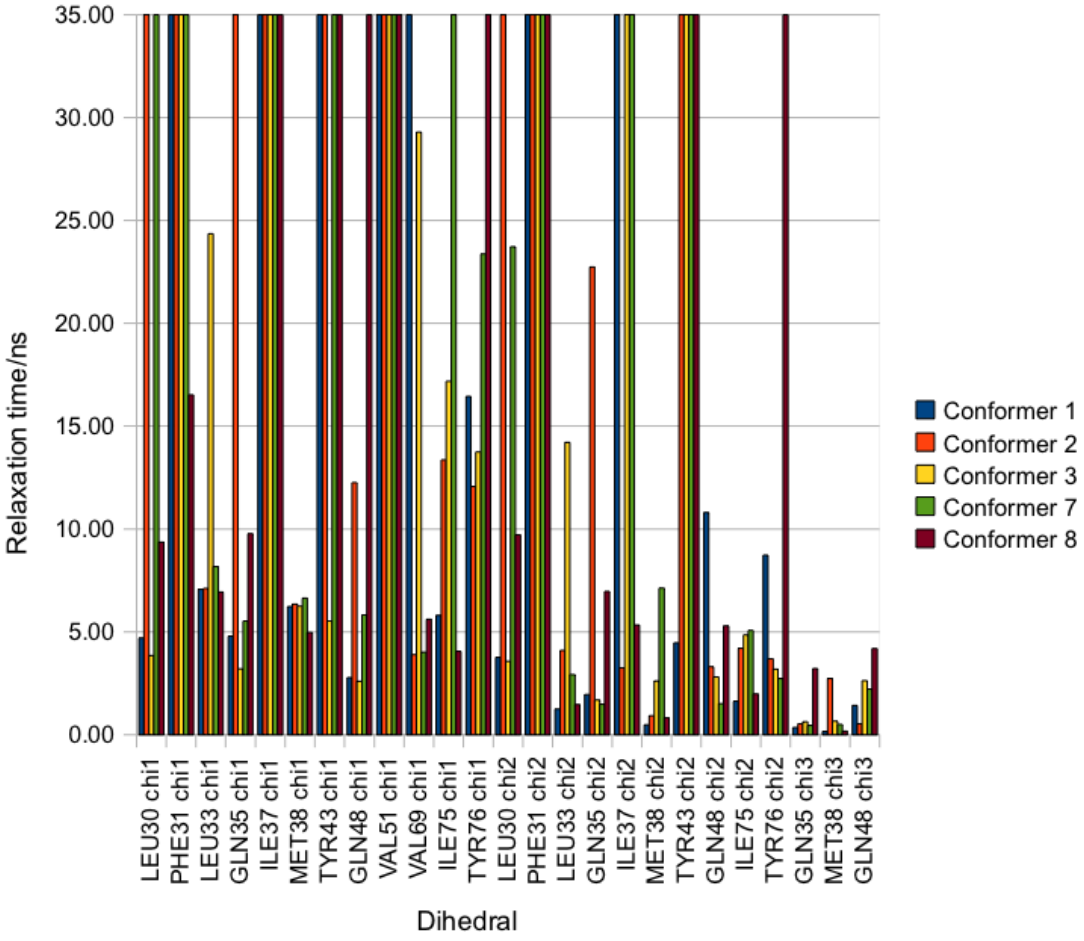

Figure 5.12: Relaxation times for dihedral angles from the hDM2 binding site for anti-parallel conformations of bound oligoamide compounds as calculated by fitting a function of the form y=exp(-x/a) to the autocorrelation function for the dihedral angle.

We note that there are some extremely long relaxation times observed for certain dihedral angles from simulations of anti-parallel oligoamide conformations (Figure 5.12). Particularly that of the first χ angle in isoleucine 37 with a relaxation time somewhere between 100 ns and 450 ns, with the majority of simulations producing results that are nearer the larger value. There are also several angles for

which the relaxation time is considerably less than 20 ns meaning that we can be confident that these dihedrals should be well sampled if simulated for the lengths of time determined by the relaxation time.

The results from the analysis of relaxation times calculated from autocorrelation functions suggests that simulations are unlikely to adequately sample dihedral space in the time that simulations are to be run. It is clearly untenable to calculate free energies from simulations of lengths approaching or exceeding half a microsecond using current mid range cluster computing. As a result it will be necessary to pick starting conformations that sample different regions of dihedral space. Additionally it would be desirable to use enhanced sampling techniques such as Hamiltonian exchange in replica space, which would allow the ligand to "pass through" the dihedrals as it mutates towards a smaller more mobile side-chain and then back towards the larger more restricted side-chain.

5.4.2.d Spatial sampling of oligoamide compounds

As with the previously discussed measures of dihedral angle sampling, we also investigated spatial sampling of the oligoamide in the vicinity of the hDM2 binding pocket. Figures 5.13 and 5.14 show ether oxygen atoms from anti-parallel and parallel starting conformations of the oligoamide respectively projected onto a plane defined by three Cα atoms from the binding site as detailed in the methods section. Inset to the graph is a representation of the starting conformation shown relative to the crystal structure of the high-affinity p53 helix. The N-terminal ether oxygen points are coloured red, central ether oxygen points are coloured black, and C-terminal ether oxygen points are coloured violet.

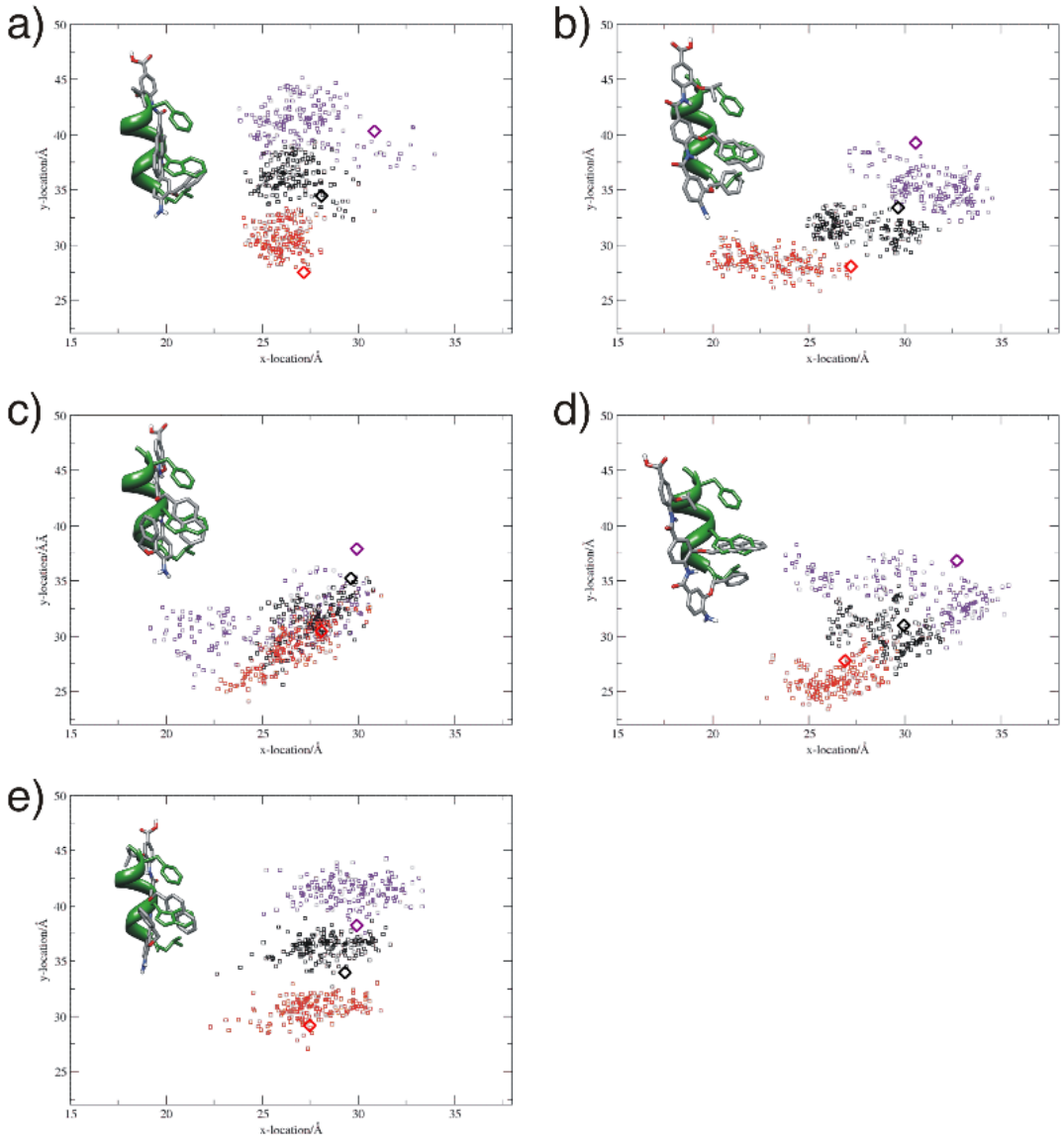

Figure 5.13: Ether oxygens from anti-parallel conformations of Phe-Nap-Leu projected onto a plane defined by Cα atoms from Tyrosine 56, Methione 62 and Valine 93. Data points are colour coded depending on which ether oxygen they belong to: R_1 (Red); R_2 (Black); R_3 (Violet). Data points were plotted at 10 ps intervals starting after 4 ns of data collection. Values at t = 0 ps are plotted with diamonds. Graphs show image of starting conformation relative to the high affinity p53 helix and data from: a) conformation 1; b) conformation 2; c) conformation 3; d) conformation 7; e) conformation 8.

In the case of anti-parallel conformations (figure 5.13) it is clear that there is some variability in the plots a)-e). However, a), b) and e) all show the same global feature of a group of vertically stacked points. This configuration is relatively stable

throughout the simulation and agrees well with the configuration of the R-groups of the high affinity p53 helix. a) and e) are not particularly skewed, although b) is skewed such that the C-terminal ether oxygens are more positive in the x-direction and the N-terminal ether oxygens are more negative in the x-direction. Since the end to end distance of the oligoamide compound is not very variable, the maximum y-distance explored is slightly less in the case of b) with respect to a) and e). Figure 5.13d shows similar behaviour to figure 5.13b although since the leucine side-chain is significantly rotated out of the binding pocket at the start of the simulation due to a rotation about the ArCO bond of the central benzene ring. This is evident in the similar behaviour of the R_1 and R_2 side-chains but the heavily skewed distribution of the leucine R_3 side-chain. Conformation 3 in figure 5.13c is initially puzzling since it doesn't appear to show behaviour similar to any of the other simulations. In actual fact, since the oligoamide is docked such that the N-terminus is pointing out of the page and the C-terminus is pointing into the page means that the angle between the oligoamide and the plane is much closer to 90 rather than the desired zero degrees. This, coupled with the fact that the R_1, R_2 and R_3 side-chains are not arranged such that they are all on the same side of the ligand, means that the graph is not as useful to observe the behaviour of the oligoamide relative to the pocket and the canonical p53 helix interaction. However, it does appear that this conformation samples the pocket quite differently to the other conformations and represents a completely different binding mode to that of the canonical helix form.

It is promising to see that in the case of figure 5.13d the conformation appears to be converging towards that of figure 5.13b which in turn does sample similar regions of the pocket to the conformations shown in figure 5.13a and 5.13e. A clear limitation of the method is shown in figure 5.13c in the case where the angle between the ligand and the plane representing the binding site is significantly

different to zero. However, in many cases this should not be an issue, for example in the case of many enzyme binding sites which are deep and usually well defined it can often be difficult for compact drug-like molecules to rotate in this pocket.

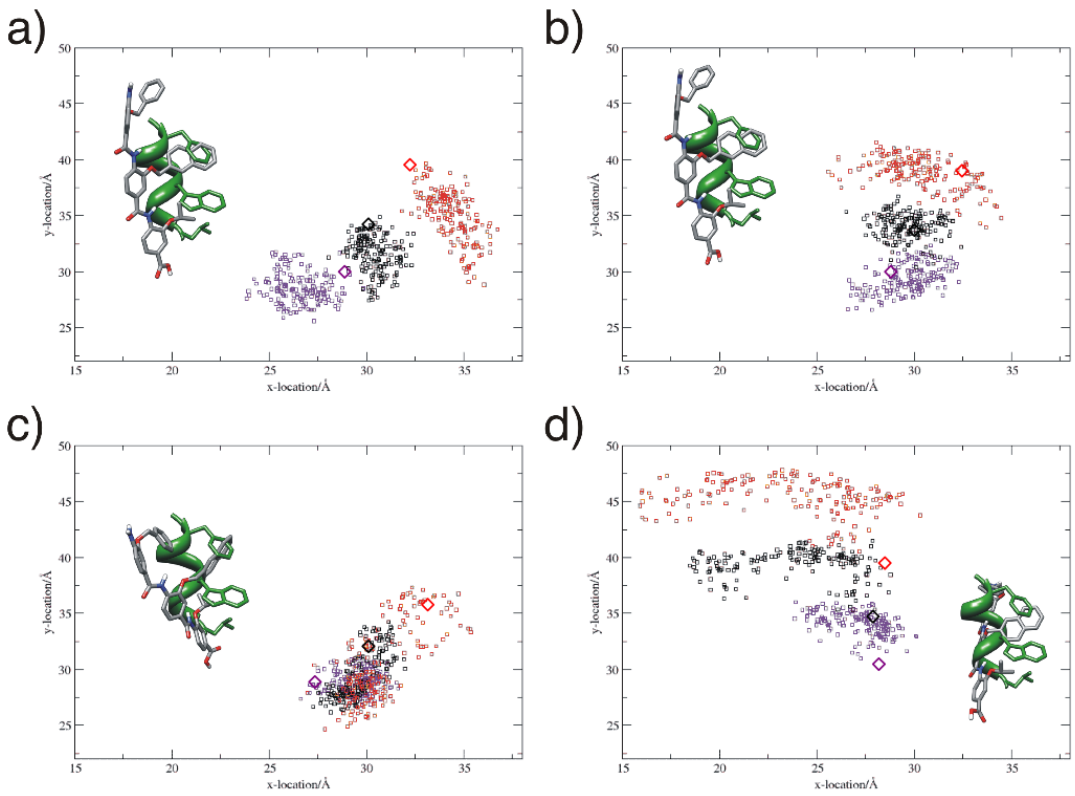

Figure 5.14: Ether oxygens from parallel conformations of Phe-Nap-Leu projected onto a plane defined by Cα atoms from Tyrosine 56, Methionine 62 and Valine 93. Data points are colour coded depending on which ether oxygen they belong to: R_1 (Red); R_2 (Black); R_3 (Violet). Data points were plotted at 10 ps intervals starting after 4 ns of data collection. Values at t = 0 ps are plotted with diamonds. Graphs show image of starting conformation and data from: a) conformation 4; b) conformation 9; c) conformation 10; d) conformation 11.

We see some further interesting observations in the case of the parallel oligoamide conformations in figure 5.14, particularly that of 5.14a and 5.14b conformations 4 and 9, which are actually the same starting conformation but describe different trajectories. The C-terminus R_3 of both simulations does sample some of the same region of space. However, it appears that the N-terminus R_1 explores a totally different region of space. Indicating perhaps that the docked conformation is

actually docked into a metastable state from which it can decay into one of two or more stable states. The graphs suggest that in figure 5.14b the N-terminal phenylalanine remains in its rotated form (ArCO dihedral such that the R_1 group is opposite the R_2, R_3 groups) somewhat similar to the conformation in figure 5.14d, whilst in figure 5.14a this dihedral relaxes such that R_1, R_2 and R_3 exist on the same side. As in figure 5.13c we see in figure 5.14c that the compound lies in slightly different orientation, that once again raises the angle between oligoamide and binding pocket plane creating a confusing graph. Visualization of molecular dynamics trajectories in order to identify important trends is a difficult task. Sometimes simple measurements such as the previously investigated RMSF and number of contacts can provide a good assessment of the behaviour of the system. Usually viewing the trajectory in a molecular viewer can also provide value, however, in this case these methods did not provide the necessary quantitative information. Projection onto the plane has several problems in that the plane varies throughout the simulation and between conformations. It does appear to show some utility in identifying the regions of space sampled more clearly. A clear improvement to the technique would be to show the time evolution of the simulations, as this might allow the method to better show whether two simulations are converging towards the same regions of space. Cluster analysis is another technique that allows a ligand centric analysis of sampling and is investigated in the next section.

5.4.2.e Cluster analysis of oligoamide compounds

Cluster analysis allows us to track those conformations of the oligoamide compound that we see regularly during the course of our simulations. We used the g_cluster tool from GROMACS to define clusters using the GROMOS clustering technique described previously in the methods. In turn this allows us to ask the question, which clusters can inter-convert? Presented in figure 5.15 is an analysis of the number of clusters observed during the course of the five anti-

parallel and three parallel oligoamide simulations each of length 20 ns. Snapshots were taken every 10 ps for the final 17 ns of simulation, resulting in a total of 8,500 and 5,100 conformations of oligoamides respectively.

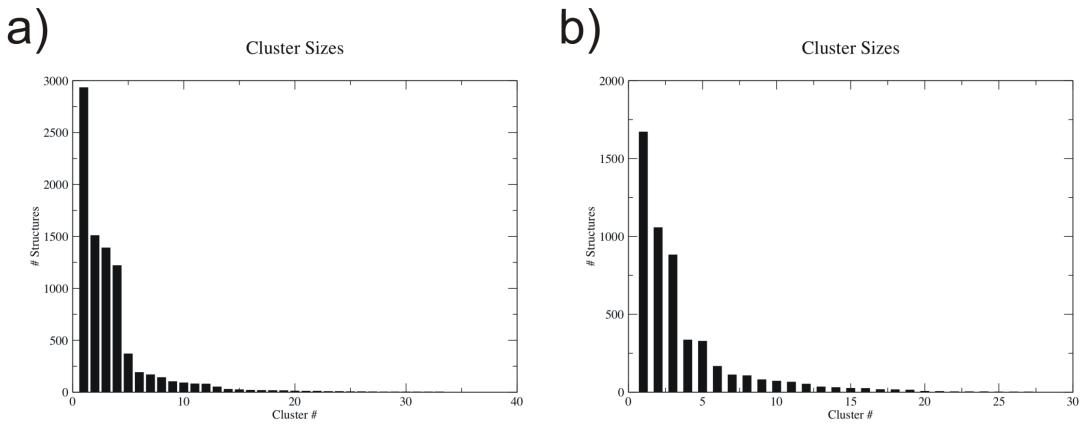

Figure 5.15: Number of conformers fitting clusters defined at an RMS threshold (of 1.5 Å) from the final 17 ns of simulation, sampled every 10 ps for a) 5 anti-parallel simulations; b) 3 parallel simulations.

In figure 5.15a we see nearly 3000 members of the most populated cluster, 1500 for the second most populated, approximately 1400 for the third most populated and just short of 1250 for the fourth most populated. The fifth most populated cluster has fewer than 500 members. There are 33 clusters of anti-parallel conformations in total. However, we have seen that the vast majority of conformations are contained in the four top ranked clusters. We see a similar picture in figure 5.15b with approximately 1700 members in the most populated cluster, slightly more than 1000 in the second most populated and just short of 900 in the third most populated cluster. There are 27 clusters in total with more than 50 % of conformers contained in the top 3 clusters. The utility of this clustering is that we can now ask the question whether these relatively diverse clusters of conformers can inter-convert between clusters on the time-scales of our simulations.

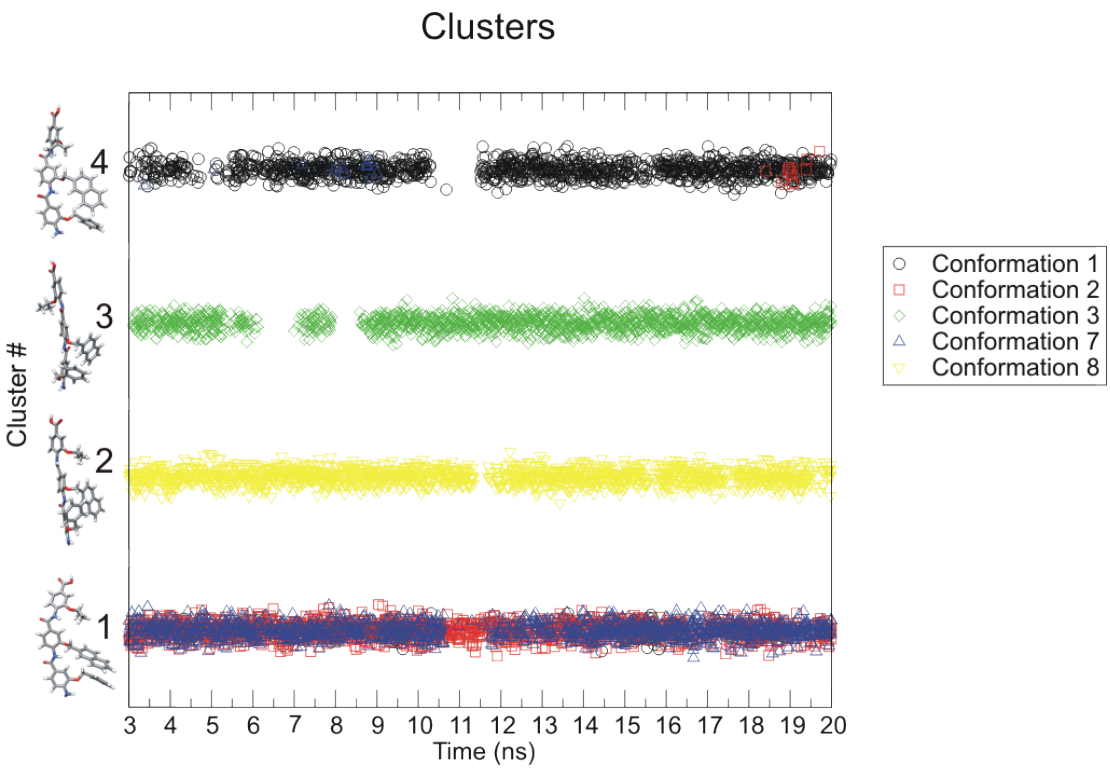

Figure 5.16: Occupancy of the top 4 anti-parallel clusters colour coded by starting conformation during the final 17 ns of the simulation.

In figure 5.16 we show the representative member (first cluster member) of each of the top 4 clusters from anti-parallel starting conformations, next to a time evolution of cluster membership during each of the five individual simulations. We observe that cluster #1 is mainly comprised of representatives from simulations with starting conformation 2 (red squares) and starting conformation 7 (blue triangles). However, it is also possible to observe a small number of conformers that emanate from starting conformation 1 (black circles). These are visible just after 9.5 ns and between 13 ns and 14 ns and between 15 ns and 16 ns. When we look at the results for the cluster #4, we see a complimentary picture to the case of cluster #1. In this cluster we see that it is mainly populated by members of conformation 1, but it is also clear that it is visited by members of starting conformation 7 around 7 ns to 8 ns, and starting conformation 2 around 18 ns to 20 ns. Thus we can conclude that there is a reasonable amount of interconversion between cluster #1 and #4, hence it may be acceptable to choose only a single

representative to sample these states sufficiently. It should be noted that these clusters appear to be quite similar with the phenyl ring in a similar orientation, the napthalene ring slightly twisted between the two conformations, and the leucine being the major difference, with a significant rotation about the leucine χ angles. Cluster #2 and #3 are both only visited by members of one starting conformation, eight and three respectively. It is noticeable that whilst they both share similar orientation of the napthalene ring, there is a large rotation about the ArCO bond from the central benzene ring, and that the χ angles for the phenylalanine are significantly different. This perhaps explains why these two clusters do not appear to interconvert on this time-scale. It is also noted that the orientation of the napthalene ring in cluster #2 and #3, is quite different to that observed in cluster #1 and #4.

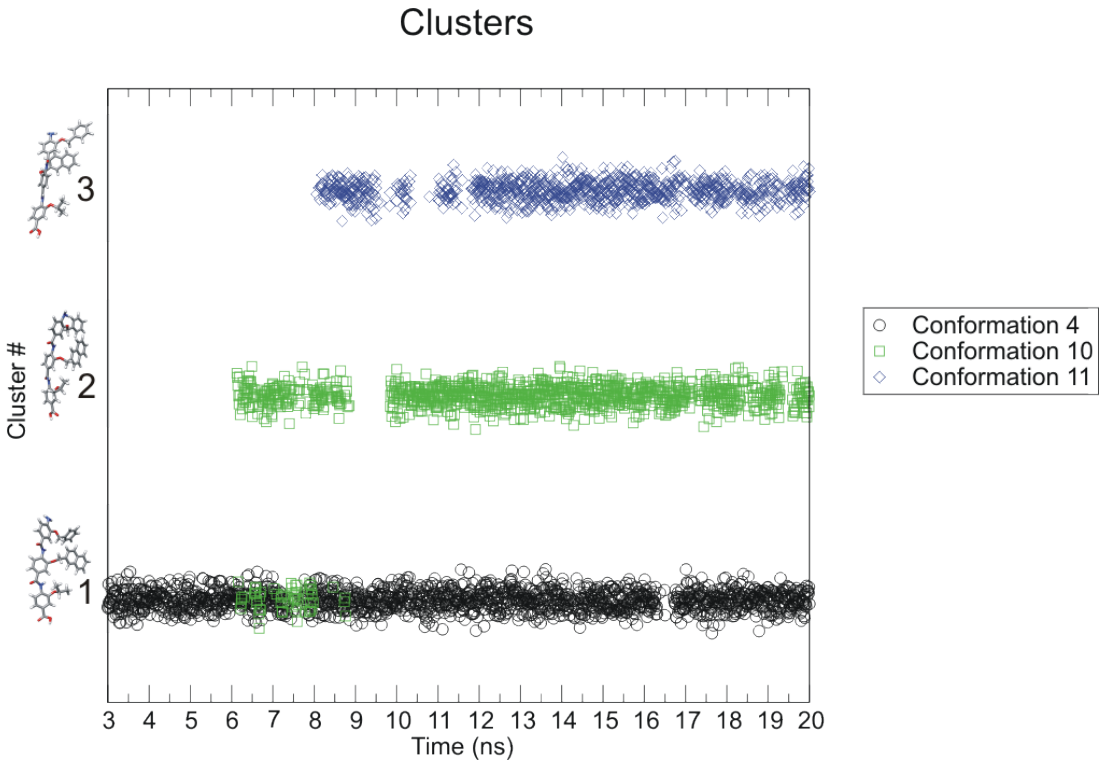

Figure 5.17: Occupancy of the top 3 parallel clusters colour coded by starting conformation during the final 17 ns of the simulation.

In figure 5.17 we show the representative member of each of the top 4 clusters from parallel starting conformations. In this case it is immediately obvious that both cluster #2 and #3 are not populated during the first 6 ns and 8 ns of the simulations. However, after this time they begin to be occupied far more often, implying that the ligand conformation converges towards these clusters during the simulation. We also note that cluster #2 and #3 do not have any members from the simulation of starting conformation 4. However, cluster #1 which is mainly populated by starting conformation 4 is also visited by conformation 10 between 6 ns and 9 ns.

5.5 Conclusion

Broadly speaking the experiments and results described in this chapter fit into two categories: a molecular dynamics study of known hDM2 complexes; a molecular dynamics study of proposed oligoamide complexes.

The initial molecular dynamics study of binders with known structure was broadly successful, showing that it is possible to simulate the hDM2 system with a set of parameters similar to those that have previously been used in accurate free energy calculations. Additionally it validates our choice of a force field designed to both simulate the properties of the protein well, whilst being compatible with the GAFF force field allowing the simulation of a large number of possible oligoamide side-chains. Finally it sets a benchmark with which to compare our future simulations, with those which were undertaken in section 2 of this chapter.

Simulation of the oligoamide compounds showed broad agreement with many of the properties such as RMSF, number of contacts and RMSD that were observed in the initial MD study of hDM2 to binders with known structure. It appears that the cluster analysis and spatial sampling best show when there is convergence behaviour from two different starting conformations. For example in the anti-

parallel simulations conformations 1, 2 and 7 exhibit convergence behaviour in the cluster analysis (figure 5.16). This convergence behaviour is somewhat mirrored in the spatial sampling graph shown in figure 5.13. In the case of the parallel simulations this is somewhat harder to observe. This may be due to the lower quality of docked conformations (remember that anti-parallel conformations were predicted to bind with higher affinity). During the course of the simulation of oligoamide compounds we also spent time evaluating the distribution of dihedral angles visited, and the relaxation time for autocorrelation functions created for each of the dihedrals featuring in the hDM2 binding pocket. It is unfortunate but not unexpected in high affinity complexes that several of the hDM2 binding pocket residues have such long relaxation times, as this has implications for the possible accuracy of our free energy calculations. It is hoped that the use of replica-exchange techniques that allow exchange of their Hamiltonian with differing values of lambda will allow enhanced sampling of these dihedral angles, thus reducing the length of these relaxation times. In principal if a sufficient number of replicas could transition from a state with all side-chains switched on, to one with all side-chains in the alanine state, these alanine side-chains should be more free to rotate than the larger more constrained side-chains. This increased flexibility of the decoupled state should allow for the correlation time of these dihedral angles to decrease. If this is not observed to be the case, then some alternative orientations of these angles with long relaxation times may have to be chosen.

We have presented a novel method to investigate the spatial sampling of the hDM2 binding site, which could be used in other protein binding studies. This spatial sampling method could be improved by accounting for the time dependence of sampling, that might better show whether several simulations are converging to the same region of space. This method as with several others suffers from the key issue that there are no defined levels of statistical significance. For example, which docking results should we choose if we were to only choose one, or how many clusters should we consider before we deem our results no

longer likely binding conformations? In the case of the comparison of distributions we may consider the use of statistical tests such as the Kolmogorov-Smirnoff test. In the case of the clustering of conformations we chose an RMSD cutoff of 1.5 Å since it generated a small number of clusters that facilitated graphical analysis. It would be ideal to first look at the variability of the RMSD of all conformations and then decide on cluster sizes that have a statistical meaning or alternatively we may be interested in picking a number of clusters that cover a large number of the total number of conformations sampled in a length of time required for all conformations to inter-convert.

In general, we have shown a variety of techniques that aim to show whether our system is likely to be adequately sampled in the time-scales that we can sample for with free energy calculations. When the evidence from: comparing oligoamide simulations to known inhibitors and the native p53 peptide; analysis of the sampling of dihedral angles in the oligoamide and the protein binding site; spatial sampling of the oligoamide relative to the binding site; and the conformations visited by the oligoamide is taken into account we show that there is a reasonable body of evidence to suggest that we can proceed with free energy calculations.

5.6 References

Basdevant, Nathalie, Harel Weinstein, and Marco Ceruso. 2006. Thermodynamic basis for promiscuity and selectivity in protein-protein interactions: PDZ domains, a case study. *Journal of the American Chemical Society* 128, no. 39 (October): 12766-77. doi:10.1021/ja060830y. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2570209&tool=pmcentrez&rendertype=abstract.

Berman, Helen M., Tammy Battistuz, T. N. Bhat, Wolfgang F. Bluhm, Philip E. Bourne, Kyle Burkhardt, Zukang Feng, *et al.* 2002. The Protein Data Bank. *Acta Crystallographica Section D Biological Crystallography* 58, no. 6 (May): 899-907. doi:10.1107/S0907444902003451. http://scripts.iucr.org/cgi-bin/paper?S0907444902003451.

Carotti, Andrea, Antonio Macchiarulo, Nicola Giacchè, and Roberto Pellicciari. 2009. Targeting the conformational transitions of MDM2 and MDMX: insights into key residues affecting p53 recognition. *Proteins* 77, no. 3 (April): 524-35. doi:10.1002/prot.22464. http://www.ncbi.nlm.nih.gov/pubmed/19507240.

D.A. Case, T.A. Darden, T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R., J. Luo, K.M. Merz, B. Wang, D.A. Pearlman, M. Crowley, S. Brozell, V. Tsui, H. Gohlke, And Mongan, V. Hornak, G. Cui, P. Beroza, C. Schafmeister, J.W. Caldwell, W.S. Ross, and P.A. Kollman. 2004. AMBER. University of California, San Francisco.

Daura, Xavier, Karl Gademann, Bernhard Jaun, Dieter Seebach, Wilfred F. van Gunsteren, and Alan E. Mark. 1999. Peptide Folding: When Simulation Meets Experiment. *Angewandte Chemie International Edition* 38, no. 1-2 (January): 236-240. doi:10.1002/(SICI)1521-3773(19990115)38:1/2<236::AID-ANIE236>3.0.CO;2-M. http://doi.wiley.com/10.1002/(SICI)1521-3773(19990115)38:1/2<236::AID-ANIE236>3.0.CO;2-M.

Essex, Jonathan W., Daniel L. Severance, Julian Tirado-Rives, and WL Jorgensen. 1997. Monte Carlo Simulations for Proteins: Binding Affinities for Trypsin–Benzamidine Complexes via Free-Energy Perturbations. *The Journal of Physical Chemistry B* 101, no. 46 (November): 9663-9669. doi:10.1021/jp971990m. http://pubs.acs.org/doi/abs/10.1021/jp971990m.

Gan, Wenxun, and Benoît Roux. 2009. Binding specificity of SH2 domains: insight from free energy simulations. *Proteins* 74, no. 4: 996-1007. doi:10.1002/prot.22209. http://www.ncbi.nlm.nih.gov/pubmed/18767163.

Grasberger, Bruce L, Tianbao Lu, Carsten Schubert, Daniel J Parks, Theodore E Carver, Holly K Koblish, Maxwell D Cummings, *et al.* 2005. Discovery and cocrystal structure of benzodiazepinedione HDM2 antagonists that activate p53 in cells. *Journal of Medicinal Chemistry* 48, no. 4 (February): 909-12. doi:10.1021/jm049137g. http://www.ncbi.nlm.nih.gov/pubmed/15715460.

Hess, Berk, Carsten Kutzner, David van der Spoel, and Erik Lindahl. 2008. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation* 4, no. 3 (March): 435-447. doi:10.1021/ct700301q. http://pubs.acs.org/doi/abs/10.1021/ct700301q.

Hornak, Viktor, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. 2006. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* 65, no. 3: 712-25. doi:10.1002/prot.21123. http://www.ncbi.nlm.nih.gov/pubmed/16981200.

Im, W, M Feig, and Cl Brooks. 2003. An Implicit Membrane Generalized Born Theory for the Study of Structure, Stability, and Interactions of Membrane Proteins. *Biophysical Journal* 85, no. 5 (November): 2900-2918. doi:10.1016/S0006-3495(03)74712-2. http://linkinghub.elsevier.com/retrieve/pii/S0006349503747122.

Janin, Joël, and Shoshana Wodak. 2007. The third CAPRI assessment meeting Toronto, Canada, April 20-21, 2007. *Structure (London, England : 1993)* 15, no. 7 (July): 755-9. doi:10.1016/j.str.2007.06.007. http://www.ncbi.nlm.nih.gov/pubmed/17637336.

Kalid, Ori, and Nir Ben-Tal. 2009. Study of MDM2 binding to p53-analogues: affinity, helicity, and applicability to drug design. *Journal of Chemical Information and Modeling* 49, no. 4: 865-76. doi:10.1021/ci800352c. http://www.ncbi.nlm.nih.gov/pubmed/19323449.

Kortemme, Tanja, and David Baker. 2002. A simple physical model for binding energy hot spots in protein-protein complexes. *Proceedings of the National Academy of Sciences of the United States of America* 99, no. 22 (October): 14116-21. doi:10.1073/pnas.202485799. http://www.ncbi.nlm.nih.gov/pubmed/12381794.

Kussie, P. H., S. Gorina, V. Marechal, B. Elenbaas, J. Moreau, A. J. Levine, and N. P. Pavletich. 1996. Structure of the MDM2 Oncoprotein Bound to the p53 Tumor Suppressor Transactivation Domain. *Science* 274, no. 5289 (November): 948-953. doi:10.1126/science.274.5289.948. http://www.sciencemag.org/cgi/doi/10.1126/science.274.5289.948.

Lawrenz, Morgan, Jeff Wereszczynski, Rommie Amaro, Ross Walker, Adrian Roitberg, and J. Andrew McCammon. 2010. Impact of calcium on N1 influenza neuraminidase dynamics and binding free energy. *Proteins: Structure, Function, and Bioinformatics* (May): n/a-n/a. doi:10.1002/prot.22761. http://doi.wiley.com/10.1002/prot.22761.

Lindahl, Erik, Berk Hess, and David Spoel. 2001. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *Journal of Molecular Modeling*: 306-317. doi:10.1007/s008940100045.

Massova, Irina, and Peter A. Kollman. 1999. Computational Alanine Scanning To Probe Protein–Protein Interactions: A Novel Approach To Evaluate Binding Free Energies. *Journal of the American Chemical Society* 121, no. 36 (September): 8133-8143. doi:10.1021/ja990935j. http://pubs.acs.org/doi/abs/10.1021/ja990935j.

McInnes, Campbell, Stanislava Uhrinova, Dusan Uhrin, Helen Powers, Kathryn Watt, Daniella Zheleva, Peter Fischer, and Paul N Barlow. 2005. Structure of free MDM2 N-terminal domain reveals conformational adjustments that accompany p53-binding. *Journal of Molecular Biology* 350, no. 3 (July): 587-98. doi:10.1016/j.jmb.2005.05.010. http://www.ncbi.nlm.nih.gov/pubmed/15953616.

Mobley, David L, John D Chodera, and Ken A Dill. 2006. On the use of orientational restraints and symmetry corrections in alchemical free energy calculations. *The Journal of Chemical Physics* 125, no. 8: 084902. doi:10.1063/1.2221683. http://www.ncbi.nlm.nih.gov/pubmed/16965052.

Mobley, David L., John D. Chodera, and Ken A. Dill. 2007. The Confine-and-Release Method: Obtaining Correct Binding Free Energies in the Presence of Protein Conformational Change. *Journal of Chemical Theory and Computation* 3, no. 4 (July): 1231-1235. doi:10.1021/ct700032n. http://www.ncbi.nlm.nih.gov/pubmed/18843379.

Pace, C N, and J M Scholtz. 1998. A helix propensity scale based on experimental studies of peptides and proteins. *Biophysical Journal* 75, no. 1 (July): 422-7. http://www.ncbi.nlm.nih.gov/pubmed/9649402.

Price, Dj, and W Jorgensen. 2000. Computational binding studies of human pp60c-src SH2 domain with a series of nonpeptide, phosphophenyl-containing ligands. *Bioorganic & Medicinal Chemistry Letters* 10, no. 18 (September): 2067-2070. doi:10.1016/S0960-894X(00)00401-7. http://linkinghub.elsevier.com/retrieve/pii/S0960894X00004017.

Russell, Robert B, Evangelia Petsalaki, Alexander Stark, and Eduardo García-Urdiales. 2009. Accurate prediction of peptide binding sites on protein surfaces. *PLoS Computational Biology* 5, no. 3 (March): e1000335. doi:10.1371/journal.pcbi.1000335. http://www.ncbi.nlm.nih.gov/pubmed/19325869.

Sorin, Eric J, and Vijay S Pande. 2005. Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. *Biophysical Journal* 88, no. 4 (April): 2472-93. doi:10.1529/biophysj.104.051938. http://www.ncbi.nlm.nih.gov/pubmed/15665128.

Vassilev, Lyubomir T, Binh T Vu, Bradford Graves, Daisy Carvajal, Frank Podlaski, Zoran Filipovic, Norman Kong, *et al.* 2004. In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science (New York, N.Y.)* 303, no. 5659 (February): 844-8. doi:10.1126/science.1092472. http://www.ncbi.nlm.nih.gov/pubmed/14704432.

Wang, Junmei, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. 2004. Development and testing of a general amber force field. *Journal of Computational Chemistry* 25, no. 9 (July): 1157-74. doi:10.1002/jcc.20035. http://www.ncbi.nlm.nih.gov/pubmed/15116359.

Willis, B. T. M., and A. W. Pryor. 1975. *Thermal Vibrations in Crystallography*. Cambridge: Cambridge University Press. http://lib.leeds.ac.uk/record=b1124792.

Wong, Chung F., and J. Andrew. McCammon. 1986. Dynamics and design of enzymes and inhibitors. *Journal of the American Chemical Society* 108, no. 13 (June): 3830-3832. doi:10.1021/ja00273a048. http://pubs.acs.org/doi/abs/10.1021/ja00273a048.

Woo, Hyung-June, and Benoît Roux. 2005. Calculation of absolute protein-ligand binding free energy from computer simulations. *Proceedings of the National Academy of Sciences of the United States of America* 102, no. 19 (May): 6825-30. doi:10.1073/pnas.0409005102. http://www.ncbi.nlm.nih.gov/pubmed/15867154.

Woods, C J, M A King, and J W Essex. 2001. The configurational dependence of binding free energies: a Poisson-Boltzmann study of Neuraminidase inhibitors. *Journal of Computer-aided Molecular Design* 15, no. 2 (February): 129-44. http://www.ncbi.nlm.nih.gov/pubmed/11272700.

Zhong, Haizhen, and Heather A Carlson. 2005. Computational studies and peptidomimetic design for the human p53-MDM2 complex. *Proteins* 58, no. 1 (January): 222-34. doi:10.1002/prot.20275. http://www.ncbi.nlm.nih.gov/pubmed/15505803.

6 Free energy calculations to determine the binding affinity of novel Arylamide compounds bound to hDM2

6.1 Abstract

Free energy calculations have long been pursued as a key objective of computational chemistry. The ability to rapidly calculate the binding affinity of a ligand for a protein of interest would bring about a paradigm shift in the rational design of compounds in drug discovery efforts, allowing focus to be shifted towards other challenging areas of drug discovery. However, whilst free energy calculations have shown much promise, with several striking applications of their success, they are difficult to calculate rapidly and consistently. Here we present an application of free energy calculations to the hDM2-p53 systems that combines published examples of best-practice calculations with the aim of illustrating a relatively straightforward way of performing consistently accurate simulations. We show that our method can achieve acceptable levels of accuracy, and more importantly, we demonstrate a methodology that could be replicated with relative ease given a relatively basic level of knowledge of molecular dynamics simulation provided that prior knowledge of the system such as that detailed over the previous two chapters is available.

6.2 Introduction

The topic of free energy calculations is a large one, with many different methodologies that have been applied to a wide variety of systems with differing levels of success. Previously, in the thesis introduction we introduced a selection of techniques that could be applied to the problem of estimating the binding affinities of a series of ligands for a specific protein. Here we review specific

techniques that are suitable for the hDM2 oligoamide system that we are studying. We then review systems where free energy calculations have been successfully applied, then focus on the hDM2-p53 system where several free energy calculation methods have already been applied. Finally, we set out the aims and techniques that will be used in this free energy study.

6.2.1 Successful application of free energy methods

The T4 lysozyme system has become a system that is regularly used to test free energy methods. The protein is relatively small, and the ligands have relatively few torsional degrees of freedom. Additionally there are a large number of high-quality crystal structures available. This allows for a relatively large amount of sampling to be performed in a reasonable amount of time. Alchemical free energy calculations have been applied to the T4 lysozyme system by Mobley *et al.*, where they developed a technique based on applying orientational restraints on ligand conformations to ensure that important conformational transitions that may not be sampled in a typical simulation can be properly accounted for(Mobley, Chodera, and Dill 2006). Application of this technique to a larger selection of ligands showed absolute binding free energies calculated to within an RMS error of 1.9 kcal mol^{-1} of experimental results(Mobley, Chodera, and Dill 2007). More recent work by Mobley *et al.* has shed some light on the use of different charge calculation techniques used for hydration free energy calculations, indicating that standard charge calculation methods tend to affect the results at a level of around 1 kcal mol^{-1} RMS error(Mobley, Chodera, and Dill 2007). In addition to the charge models for ligands, work has been done on the effect of different water models on free energy of hydration of methane. In this work it was found that the variation between water models was relatively small with SPC/E and TIP4P-Ew giving the largest variation from experiment and TIP3P performing well(Shirts and Pande 2005).

6.2.2 Dispersion corrections

Efficient simulation of molecular systems invariably requires use of cut-offs or switched potentials to facilitate rapid calculation of the behaviour of the system. Effects of using cut-off values for electrostatic calculations can have significant effects on certain parameters, however these can be mitigated by use of a technique such as Particle Mesh Ewald (PME). PME decomposes electrostatic calculations into a two summations, first of the short range interaction energy up to some cut-off; and second of the long range interaction energy calculated in Fourier space. When PME calculations in molecular dynamics are performed it is required to assume a periodic arrangement of a single box filled with water molecules which in effect represents calculation of long range interactions to an infinite range with appropriate choice of PME parameters. Lennard Jones parameters however are often cut-off after less than 1 nm where their effect is perceived to be relatively small. However, cut-off schemes perceived to contain little error can still have significant differences between observed quantities of the system. As a result it is necessary to apply accurate dispersion corrections that can eliminate the difference between different cut-off values. Dispersion corrections to correct for the pressure and energy have been applied to isotropic liquids. These methods (and the isotropic assumption) have been applied to larger non-isotropic systems such as solvated proteins. Recent work has investigated whether these assumptions are acceptable for these non-isotropic systems(Shirts *et al.* 2007). Shirts *et al.* observed that when investigating free energy of ligand binding, discrepancies of between 1-2 kcal mol^{-1} could be observed when inappropriate dispersion corrections were applied(Shirts *et al.* 2007). These techniques will be applied in this chapter in order to calculate the most accurate free energy values possible.

6.2.3 Replica-exchange

Replica-exchange techniques have been investigated as a possible method of improving sampling in free energy calculations. Traditionally replica-exchange has been applied between identical systems simulated over a ladder of temperatures. In free energy calculations the same system is used but simulated at a variety of values of lambda as per a standard FEP, TI or BAR calculation. Exchanges are made subject to a criteria defined by detailed balance allowing swaps between adjacent lambda values.(Cossins *et al.* 2009)

$$\exp\left[\beta\left[U_B(j)-U_B(i)-U_A(j)+U_A(i)\right]\right] \geqslant rand(0,1) \qquad (42)$$

Comparison of replica-exchange thermodynamic integration (RETI) to standard FEP, WHAM and several variants of TI were performed by Woods *et al.* where they showed that RETI performed the best when applied to determining the free energy of hydration of methane(Woods, Essex, and King 2003). replica-exchange methods are relatively straight forward to implement indeed it is possible to perform them in Desmond with no modifications to the standard code(Bowers *et al.* 2006). Additionally it can be a helpful tool to investigate sampling by looking at the rate of exchange between particular values of lambda. Well sampled regions with good overlap in phase space will have high rates of acceptance of replica-exchange, whereas regions with low overlap in phase space will have low rates of acceptance. This allows a graphical view of the sampling, and may be helpful to guide both the number and placement of lambda windows.

6.2.4 Applications to the hDM2/p53 system

Since the hDM2/p53 system is of significant biological and medicinal interest, it is not surprising to find that several different free energy methods have been applied to a variety of compounds designed to inhibit the interaction. In one study a very simple docking method has been applied to the system, although no attempt was made to extrapolate these results to calculate the free energy of binding for different compounds(Shaginian *et al.* 2009). MM-GBSA methods have been applied to both predict the affinity of p53 based peptides and those of known

inhibitors, they used 2000 snapshots taken every 1 ps(Zhong and Carlson 2005). Moreira *et al.* used MMPBSA methods to investigate the binding affinity of a series of hDM2 and p53 mutants, with 25 snapshots taken from the last 0.5 ns of trajectories(Moreira, Fernandes, and Ramos 2008). Generally these studies tend to use relatively short trajectories, presumably due to constraints in the amount of computational time available for the investigations. Michel *et al.* used an MC based technique to predict the binding affinity of ß-peptides designed to mimic the p53 interactions with hDM2, where they managed to achieve good accuracy in their calculations(Michel *et al.* 2009).

6.2.5 Study Aims

We aim to perform the rigorous calculation of the binding free energy of oligoamide compounds for hDM2, using alchemical free energy calculations and fully flexible protein simulations. We build on work described in the previous two chapters, docking of compounds to provide starting structures and parameters for simulation. Combining results from these chapters with alchemical techniques described in the introduction should allow for accurate and transferable free energy calculations.

6.3 Methods

6.3.1 Docking with Autodock

We used a docking procedure that utilises Autodock to dock the compounds shown in figure 6.8 into the hDM2 binding site in order to generate conformations with which to perform alchemical free energy calculations in order to determine the relative binding affinity of the compounds to the common compound shown in figure . Two rounds of docking using Autodock(Morris *et al.* 1998),(Huey *et al.* 2007) were performed. The first round used Autodock 4.2.1 to produce 300 docked conformations with a maximum of 25 million evaluations for 27000 generations with population size 300 using the compounds detailed in Figure 6.8.

The results from this set of docking and clustering at a 2 Å RMS cutoff are presented in Figure 6.3. The second round of docking calculations were performed using Autodock 4.2.1 using a Lamarkian genetic algorithm. 600 docked conformations were generated, with each using 250 million evaluations for 64,000 generations of population size 600. Random number seeds were generated from the autodock PID and the current system time. The protein structure used was derived from the structure of hDM2 bound to a high-affinity p53 helix (1T4F), with all water molecules removed, protonation states manually assigned and the high-affinity p53 helix removed from the coordinates. A grid centred on 13.119, 18.969, 10.941 was used with spacing of 0.375 Å and 52, 58 and 48 points in the x, y and z directions.

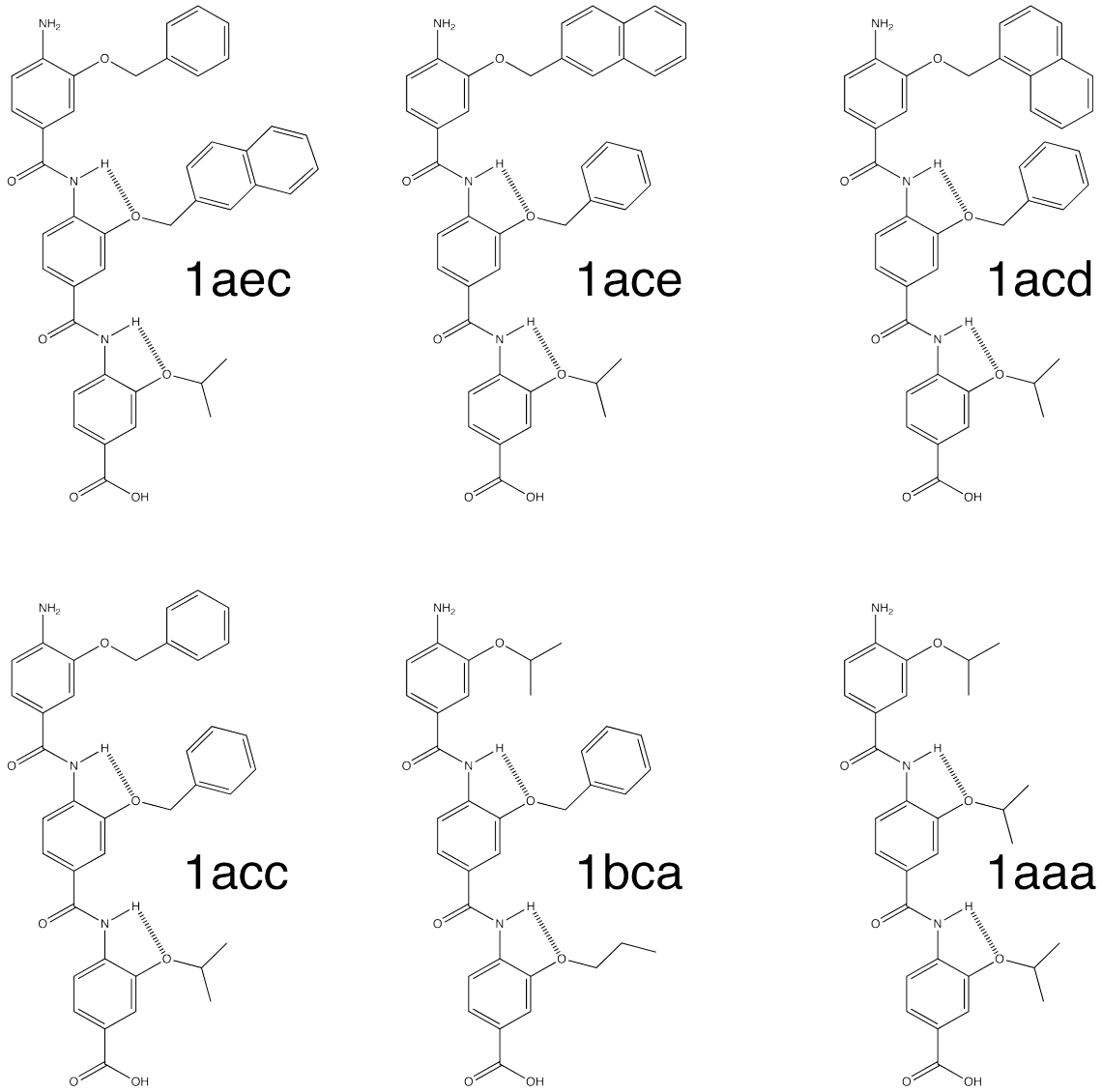

Figure 6.1: Oligoamide compounds that have previously been synthesised and tested and have been investigated further during the course of this work(Plante *et al.* 2009).

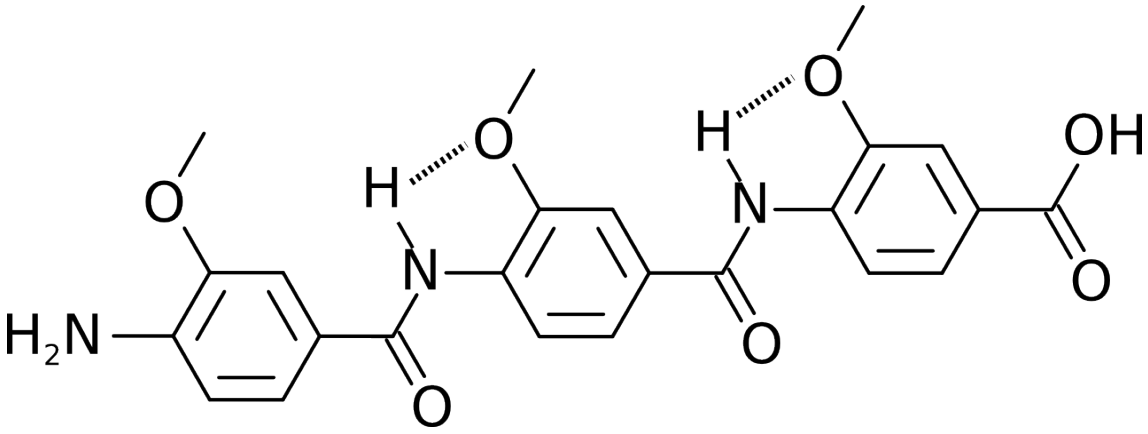

Figure 6.2: Structure to which all compounds from figure 6.8 are mutated alchemically.

6.3.2 Free energy calculations with Desmond

6.3.2.a Equilibration

All simulations performed where subjected to a six step equilibration procedure to ensure that the system was in a suitable low energy state before the free energy calculations were carried out.

A minimum of 10 steepest descent minimization steps were performed until the minimum gradient was less than 50. This was followed by a minimum of 10 steepest descent minimization steps were performed until the minimum gradient was less than 25, before continuing with a maximum of 2000 L-BFGS minimization steps until a gradient of 5 is met. 3 minimization steps were performed between each migration step and the normal of the first step was 0.005. Twelve ps of molecular dynamics simulation was then performed. It was performed in the nVT regime with Berendsen thermostat. Velocities were scaled in the range 0.85 – 1.2 and tau set to 0.1, an integration timestep of 1 fs was used. Centre of mass motion was removed. Temperature was held at 10 K. The bonded interval was 1 timestep, with non-bonded far 3 timesteps, and non-bonded near 1 timestep. A migration interval of 12 fs was used. The M-SHAKE algorithm was

used to constrain hydrogen bond lengths to a tolerace of $1.0x10^{-8}$, using a maximum of 8 iterations. This was followed by 12 ps of molecular dynamics simulation in the nPT regime, temperature was held at 10 K with a Berendsen thermostat. Velocities were scaled in the range 0.85 - 1.2 and the relaxation time was set to 0.1. Box size was varied in the range 0.95 - 1.1 per step. Pressure was scaled isotropically with a Berendsen barostat to 1.01325 bar, whilst the system compressibility was set to $4.5x10^{-5}$ bar^{-1} and a relaxation time of 50 ps were used. An integration timestep of 2 fs was used. This was followed by 24 ps of molecular dynamics simulation using the same parameters as above but at a temperature of 300 K. The final equilibration step was 24 ps of molecular dynamics simulation. In the nPT regime using the previous parameters except that the relaxation time of the barostat was decreased to 2 ps, whilst the relaxation time for the thermostat was raised to 1 ps.

6.3.2.b Free energy simulation

The majority of simulations were performed using a Hamiltonian replica-exchange methodology described below, however, comparison was made to some non replica-exchange free energy calculations. Both sets of simulations used the same simulation parameters listed in the final equilibration stage, except as noted.

We performed 5 ns of molecular dynamics simulation. The Martyna-Tobias-Klein constant pressure and temperature method was used with a piston mass of 2. The reference temperature was retained at 300 K, Two discrete updates to Nose-Hoover barostat variables per timestep and a time constant of 1 ps was used.

Replica-exchange calculations with 24 replicas with parameters described in table 6.1 were carried out for 5 ns with exchanges between neighbouring replicas attempted every 12 ps. Exchanges were accepted if both replicas fulfilled the Metropolis criteria.

$$\exp\left[\beta\left[U_B(j)-U_B(i)-U_A(j)+U_A(i)\right]\right] \geq rand(0,1) \quad (43)$$

In order to make comparison to the Hamiltonian replica exchange method, we used 24 separate simulations with standard molecular dynamics with the same simulation parameters as used for the Hamiltonian replica exchange calculations.

6.3.2.c Lambda parameters

We used the dual topology approach that is available in Desmond to perform our alchemical transformations. Alchemical free energy calculations were performed using an alpha parameter of 0.5 for the soft core potential(Shirts and Pande 2005). We used 24 windows with charge, bonded and van der Waals parameters altered as shown in Table 6.1. In the case of the bonded parameters these represent switching off the c-c bond that joins the mutated group to the oligoamide backbone. Additional lambda parameters for 12 and 40 windows are detailed in Table 6.2 and 6.3.

	Charge A	Charge B	Bonded A	Bonded B	vdW A	vdW B
0	1.0	0.0	1.0	0.0	1.0	0.0
1	0.889	0.0	1.0	0.076	1.0	0.067
2	0.778	0.0	1.0	0.143	1.0	0.119
3	0.667	0.0	1.0	0.214	1.0	0.158
4	0.556	0.0	1.0	0.286	1.0	0.190
5	0.444	0.0	1.0	0.357	1.0	0.218
6	0.333	0.0	1.0	0.429	1.0	0.247
7	0.222	0.0	1.0	0.5	1.0	0.282
8	0.111	0.0	1.0	0.571	1.0	0.325
9	0.0	0.0	1.0	0.643	1.0	0.382
10	0.0	0.0	0.929	0.714	0.827	0.456
11	0.0	0.0	0.857	0.786	0.675	0.553
12	0.0	0.0	0.786	0.857	0.553	0.675
13	0.0	0.0	0.714	0.929	0.456	0.827
14	0.0	0.0	0.643	1.0	0.382	1.0
15	0.0	0.111	0.571	1.0	0.325	1.0
16	0.0	0.222	0.5	1.0	0.282	1.0
17	0.0	0.333	0.429	1.0	0.247	1.0
18	0.0	0.444	0.357	1.0	0.218	1.0
19	0.0	0.556	0.286	1.0	0.190	1.0
20	0.0	0.667	0.214	1.0	0.158	1.0
21	0.0	0.778	0.143	1.0	0.119	1.0
22	0.0	0.889	0.071	1.0	0.067	1.0
23	0.0	1.0	0.0	1.0	0.0	1.0

Table 6.1: Parameter scaling for different values of lambda in the 24 window schedule used in the final simulations (and the lambda error calculations). Values rounded to 3 significant figures.

	Charge A	Charge B	Bonded A	Bonded B	vdW A	vdW B
0	1	0	1	0	1	0
1	0.75	0	1	0.14	1	0.12
2	0.5	0	1	0.29	1	0.19
3	0.25	0	1	0.43	1	0.25
4	0	0	1	0.57	1	0.33
5	0	0	0.86	0.71	0.67	0.46
6	0	0	0.71	0.86	0.46	0.67
7	0	0	0.57	1	0.33	1
8	0	0.25	0.43	1	0.25	1
9	0	0.5	0.29	1	0.19	1
10	0	0.75	0.14	1	0.12	1
11	0	1	0	1	0	1

Table 6.2: Parameter scaling for different values of lambda in the 12 window schedule used in the lambda error calculations.

	Charge A	Charge B	Bonded A	Bonded B	vdW A	vdW B
0	1	0	1	0	1	0
1	0.75	0	1	0	1	0
2	0.5	0	1	0	1	0
3	0.25	0	1	0	1	0
4	0	0	1	0	0.9	0
5	0	0	1	0	0.8	0
6	0	0	1	0	0.85	0
7	0	0	1	0	0.7	0
8	0	0	1	0	0.65	0
9	0	0	1	0	0.6	0
10	0	0	1	0	0.55	0
11	0	0	1	0	0.5	0
12	0	0	1	0	0.45	0
13	0	0	1	0	0.35	0
14	0	0	1	0	0.25	0
15	0	0	1	0	0.1	0
16	0	0	0.8	0	0	0
17	0	0	0.6	0	0	0
18	0	0	0.4	0	0	0
19	0	0	0.2	0	0	0
20	0	0	0	0.2	0	0
21	0	0	0	0.4	0	0
22	0	0	0	0.6	0	0
23	0	0	0	0.8	0	0
24	0	0	0	1	0	0.1
25	0	0	0	1	0	0.25
26	0	0	0	1	0	0.35
27	0	0	0	1	0	0.45
28	0	0	0	1	0	0.5
29	0	0	0	1	0	0.55
30	0	0	0	1	0	0.6
31	0	0	0	1	0	0.65
32	0	0	0	1	0	0.7
33	0	0	0	1	0	0.85
34	0	0	0	1	0	0.8
35	0	0	0	1	0	0.9
36	0	0.25	0	1	0	1
37	0	0.5	0	1	0	1
38	0	0.75	0	1	0	1
39	0	1	0	1	0	1

Table 6.3: Parameter scaling for different values of lambda in the 40 window schedule used in the lambda error calculations.

6.3.2.d Nonbonded interactions

Near van der Waals and electrostatic interactions used a cut-off of 9 Å, and a lazy migration radius of 10 Å. Lennard Jones parameters were cut-off at 9 Å. Far electrostatic interactions were calculated using PME on a cubic FFT grid with 32 points in each direction, PME interpolation order 4 and Ewald sigma coefficient 2.17. We used the soft-core potential detailed in equation 44.

$$U_{ij}^{vdW}(r_{ij};\lambda)=4\epsilon_{ij}\lambda\left\{\frac{1}{\left[\alpha_{vdW}(1-\lambda)^2+\left(\frac{r_{ij}}{\sigma_{ij}}\right)^6\right]^2}-\frac{1}{\left[\alpha_{vdW}(1-\lambda)^2+\left(\frac{r_{ij}}{\sigma_{ij}}\right)^6\right]}\right\} \quad (44)$$

In the above equation for the van der Waals potential U_{ij}, we used a value of 0.5 for the scaling parameter α_{vdW}, is the previously discussed order parameter along which the mutation is performed, $_{ij}$ is the location of the minimum of the van der Waals function, $_{ij}$ is the point at which the van der Waals function crosses the x-axis nearest to zero and r_{ij} is the distance between atoms i and j.

6.3.2.e Global cell

Simulations were partitioned such that each used 8 CPUs decomposed such that 2 partitions were located in each axis direction. The rounded clone policy was used, with an estimated number of 1 atom per particle array voxel. The clone radius was set to 11 Å (significantly larger than the default 5.00000001 Å, in order to allow multiple functional group mutation).

6.3.2.f Dispersion Correction

Snapshots from each simulation were taken from the lambda = 0 trajectories at intervals of 200 ps. The energy of the system was calculated using the same parameters as described for the free energy calculations, with the only differences being the use of a cut-off value of 25 Å instead of 9 Å, and the removal of the switched potential for the long-range calculations. A cut-off of 25 Å has been

shown to be a reasonable parameter to account for long range interactions in previous work by Shirts *et al.*(Shirts *et al.* 2007). The energy difference between the standard cut-off and the long cut-off is determined for each cutoff and then combined using equation 43:

$$dF = \frac{1}{\beta} \ln(\langle \exp(-\beta\, dE) \rangle) \tag{45}$$

The dispersion correction is applied to the free energy calculations by considering the thermodynamic cycle that converts protein and ligand in complex, and ligand in solution with short-range cut-offs to protein and ligand in complex and ligand in solution with long-range cutoffs.

6.3.2.g Overlap integrals

In Bennett's original derivation of the Bennett Acceptance Ratio a formula for the variance associated with a free energy measurement is defined:

$$\sigma^2 = \frac{2}{n}\left[\left(\int \frac{2\rho_0\rho_1}{\rho_0+\rho_1}\, dq^N\right)^{-1} - 1\right] \tag{46}$$

Here ρ_0, ρ_1 is the normalized configuration density space of state A and state B respectively. The integral containing the two configuration spaces is the overlap between the two configuration spaces. Bennett shows that the ratio between the two partition functions (hence the free energy) can only be determined accurately if a number of configurations greater than the reciprocal of the overlap between ρ_0, ρ_1.

We can define an overlap $O(x_0,x_1)$ between two regions of phase space:

$$\left(\frac{N\sigma^2}{x_0\, x_1 + 1}\right)^{-1} = \frac{O(x_{0,}\, x_1)}{x_0\, x_1} \tag{47}$$

where N is the total number of samples in each trajectory (0 and 1), x_0 is the number of samples in trajectory 0 divided by N, x_1 is the number of samples in trajectory 1 divided by N. σ^2 is the variance in the free energy moving from state 0 to state 1. In the cases presented in this chapter the samples all have x_0 equal to x_1, and N is 9982. An overlap close to 0 indicates that state 0 and 1 are not close in phase space, whereas an overlap close to 1 is the maximum and indicates two states that completely overlap in phase space. Variance can be large because of either a small overlap, or because of a small effective N. Determination of overlap can determine whether it might be more efficient to increase the number of samples taken (length of simulation) or to increase the number of lambda windows (or perhaps reorganise the spacing).

6.4 Results and Discussion

6.4.1 Generating starting conformations

6.4.1.a First round of docking

Whilst the method for generating starting conformations for compound 1aec (see figure 6.8) has been tested with some success in a previous chapter, the same method must be applied to generate more starting conformations for the five remaining compounds also detailed in Figure 6.8. It is expected that using the method again with the same parameters should produce results that look reasonable for the five new compounds. However, we expect that the method should reproduce results that are similar to those seen in the previous work when conformations for compound 1aec were generated.

The results in Figure 6.3 use the method described in the previous chapter (also described in the methods in this chapter) and appear to generate several low energy clusters for each of the compounds. We note that there appears to be a weak trend for the energies of the largest low-energy clusters to correlate to the

experimentally measured IC_{50} values(Plante *et al.*, 2009), which are also presented in figure 6.3. We attempted to improve on the previous docking work by identifying a new set of parameters to guide the docking.

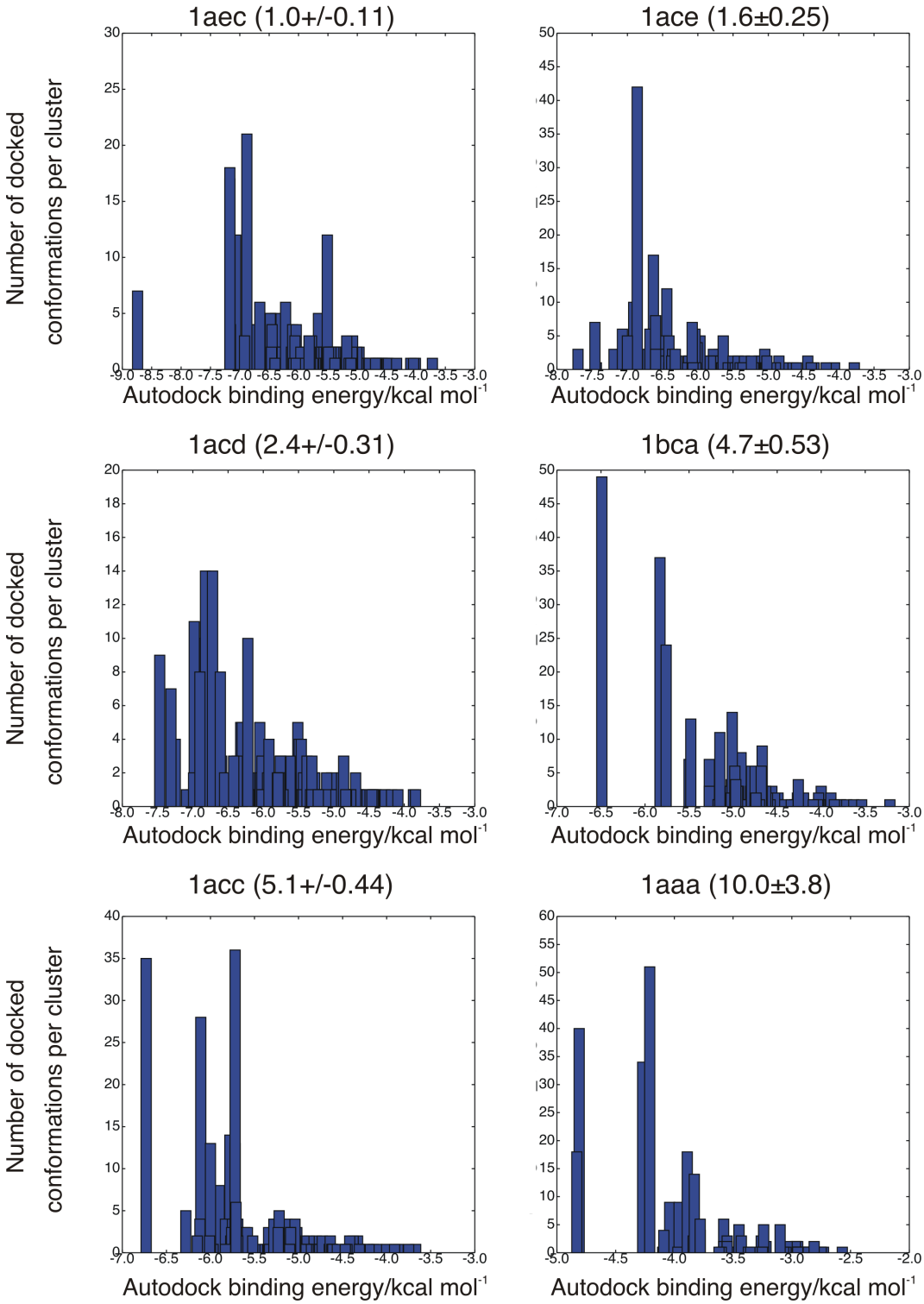

Figure 6.3: Mean Autodock binding energy score and corresponding cluster occupancy created using a 2 Å RMSD cutoff and 300 docked conformations. Measured IC_{50} values from the work by Plante *et al.* are shown above each graph with values and errors measured in nM within parenthesis(Plante *et al.* 2009). Inspection of the distributions of binding energies shows weak correlation between docked energies and experimental energies.

6.4.1.b Second round of docking

Figure 6.4 shows a structure generated from 600 docked conformations generated using the enhanced sampling parameters discussed in the methods (in green) compared to conformation 2 which was identified and used in work from the previous two chapters. Here we see that the newly generated conformation (shown in green) is very similar to that of the previously generated conformation with the RMSD between the two conformers being very low (conformer 2 from the previous two chapters, shown with atoms coloured).

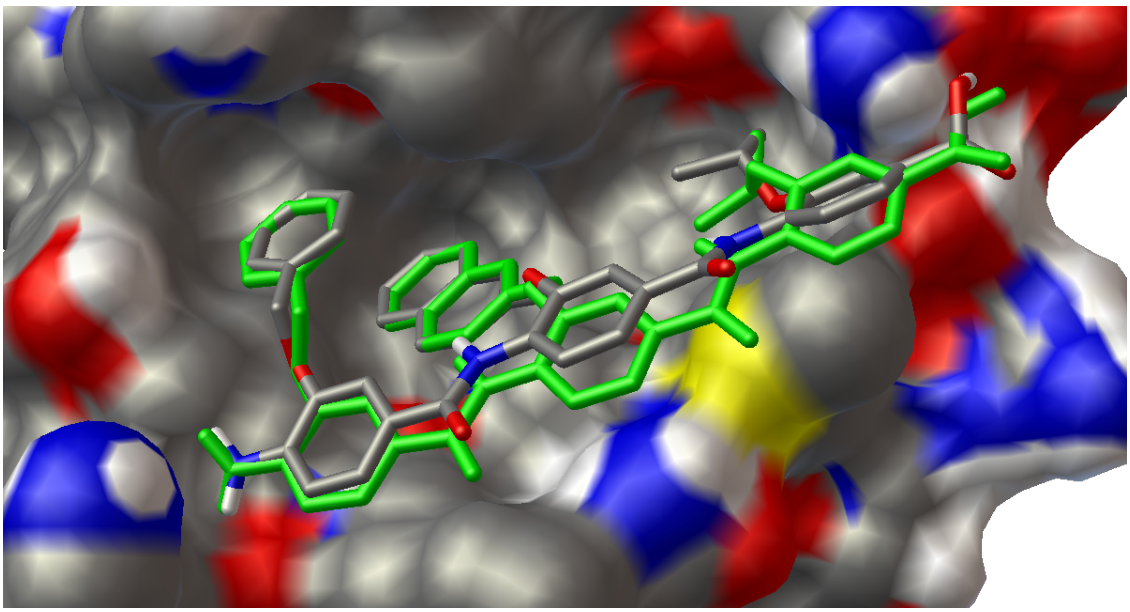

Figure 6.4: Structure of compound 1aec from a Autodock run with 600 members shown in stick representation, with atoms coloured by type compared to Conformation 2 as identified by previous work. Representative 19 from cluster 11.

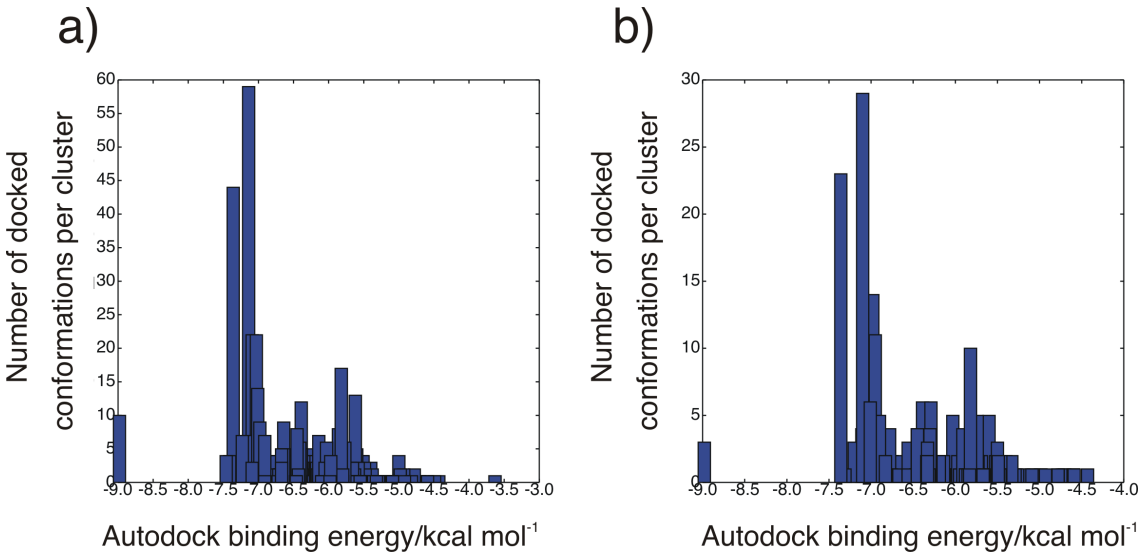

Figure 6.5: Autodock binding energy distributions for compound 1aec: a) 600 conformations; b) 300 conformations. Similar distributions of clusters are observed in each indicating that the docking experiment with fewer docked conformations appears to still sample the same low energy regions. Generally this suggests that enhanced sampling parameters are required, but that 300 docked conformations is likely be acceptable for adequate sampling.

Figure 6.5 shows the two distributions generated when using 600 and 300 conformations with enhanced sampling parameters. Since the distributions look similar it is reasonable to use 300 conformations generated using the enhanced sampling parameters. When generating conformers for the remaining five compounds the enhanced sampling parameters and 300 conformations were used.

Figure 6.6 shows the Autodock energy distributions from using the enhanced sampling parameters and generating 300 conformers for each. Furthermore, the average binding energy, calculated from an exponential average of the energies, and the energy of the largest cluster is compared to the experimentally determined IC_{50} is shown in Table 6.4. Further to this images of the docked conformations are shown in Figure 6.7.

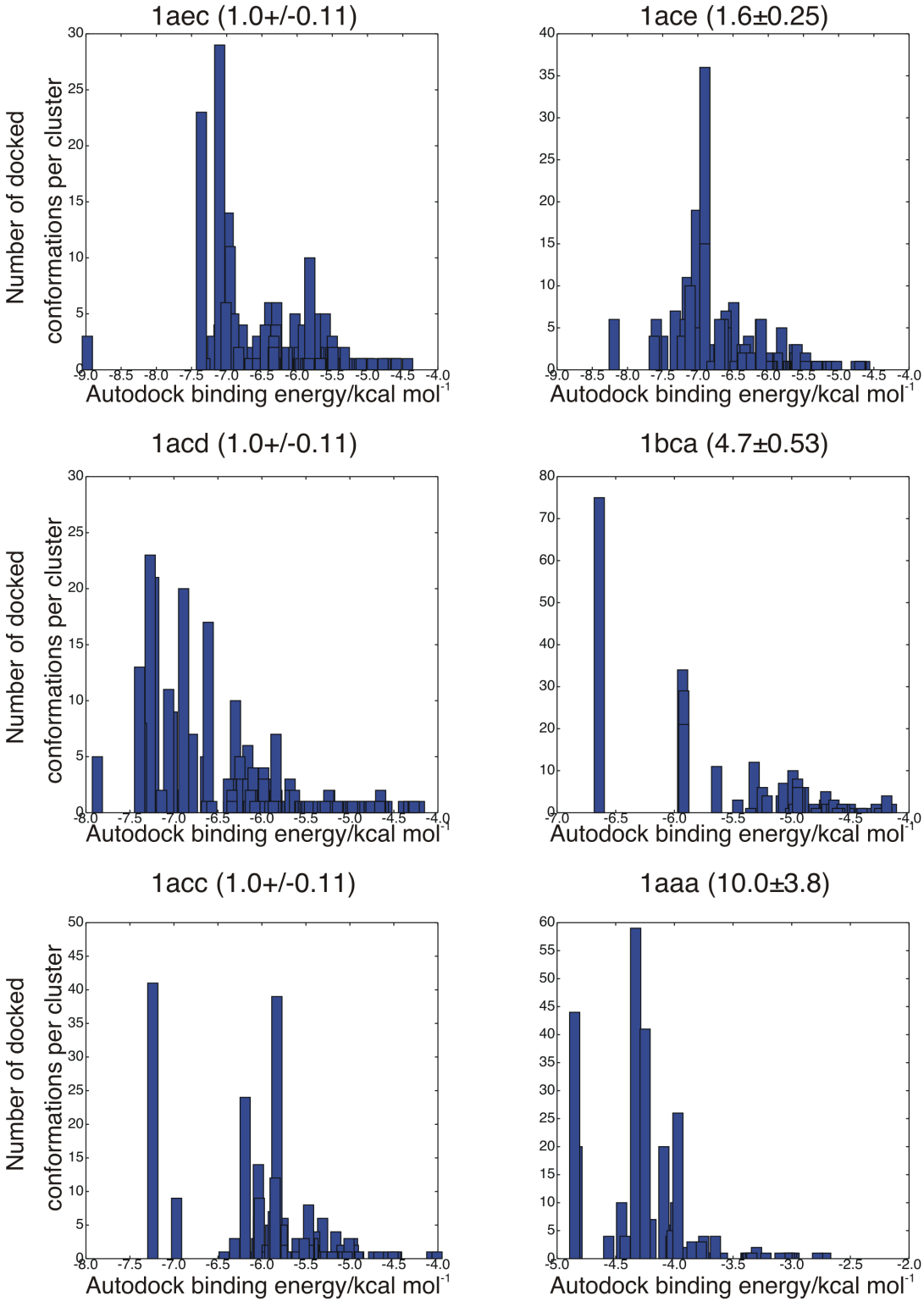

Figure 6.6: Mean autodock binding energy score and corresponding cluster occupancy created using a 2 Å RMSD cutoff and 300 docked conformations generated using better sampling parameters. Measured IC_{50} values from the work by Plante *et al.* is shown above each graph(Plante *et al.* 2009).

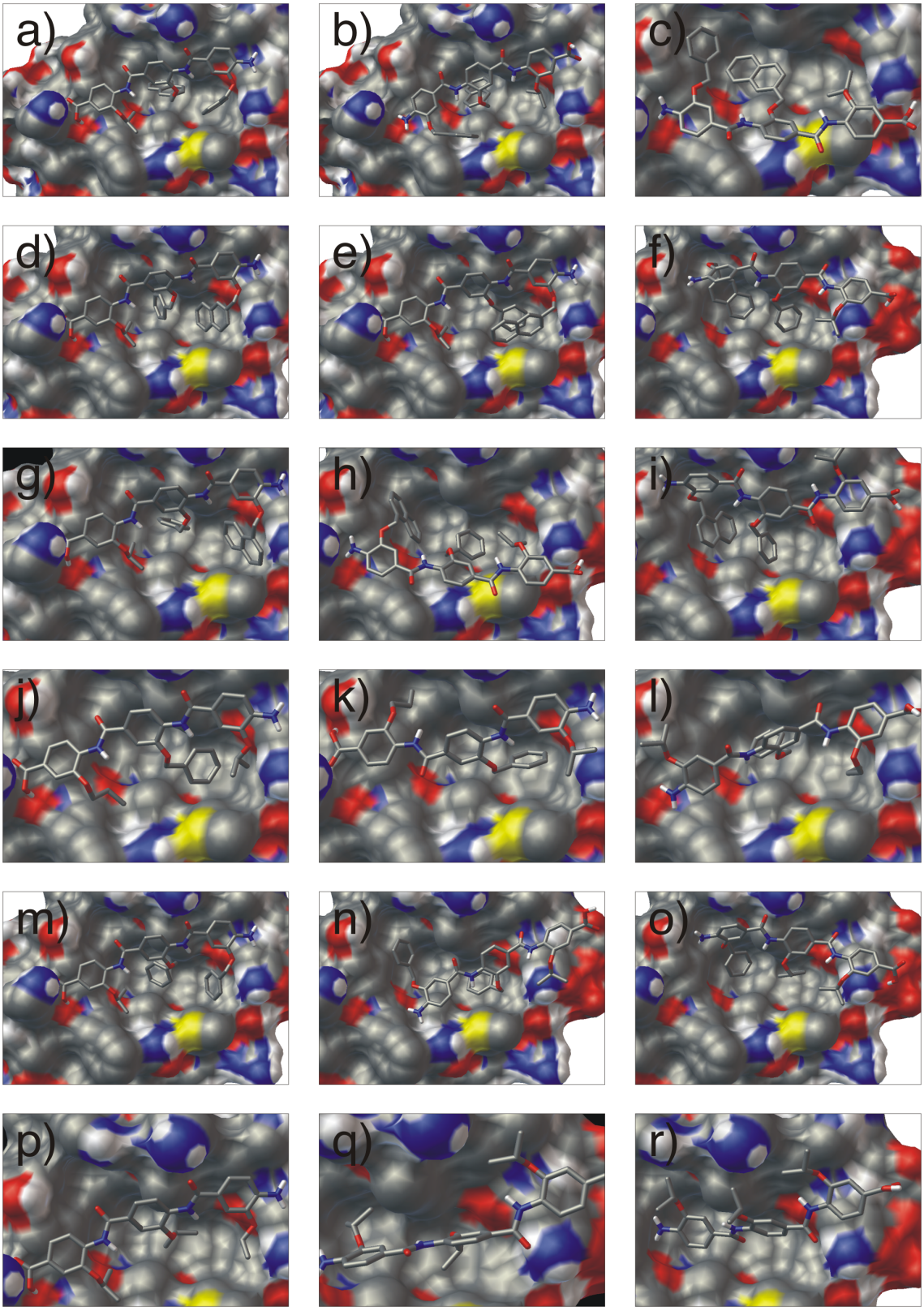

Figure 6.7: The top 3 poses from each docking are shown for illustrative purposes, these poses were carried forward to be used as starting points for free energy calculations. a-c) 1aec; d-f) 1ace; g-i) 1acd; j-l) 1bca; m-o) 1acc; p-r) 1aaa.

Compound	Autodock largest cluster energy/ kcal mol^{-1}	Autodock average binding energy/ kcal mol^{-1}	Predicted binding energy (using IC_{50})/ kcal mol^{-1}	Experimental IC_{50}/μM
1aaa	-4.4	-3.929	-6.86	10.0
1acc	-7.3	-5.419	-7.26	5.1
1acd	-7.4	-5.822	-7.71	2.4
1ace	-7.0	-5.983	-7.95	1.6
1aec	-7.2	-4.775	-8.23	1.0
1bca	-6.65	-5.062	-7.31	4.7

Table 6.4: Autodock binding energies for the largest low-energy cluster and calculated using an exponential average of energies from all docked conformations compared to the predicted binding energy calculated from experimental IC_{50}.

Table 6.4 shows that both the largest cluster energy and the average binding energy clearly distinguish compound 1aaa as the weakest binder. However a major problem occurs when considering 1aec, as this is ranked as the second weakest binder, when in fact it has the best experimental binding affinity. The average binding energy places 1bca followed by 1acc as the next weakest binder, which mixed up the experimental IC_{50}s of 4.7 and 5.1, however this is a fairly small margin to be able to detect. The average binding energy also does well to rank 1acd and 1ace correctly. When considering the largest cluster energy 1bca is placed as second weakest binder, whereas 1acc is placed as second tightest binder. Indeed the predicted range of experimental energies is very small at 1.37 kcal mol^{-1} making even an ordered ranking according to experimental energy extremely challenging. Hence, one cannot read much into these energy differences, particularly as these are simple docked poses where the scoring function is unlikely to model macroscopic thermodynamic properties to any satisfactory level. That there is some correlation is encouraging, but it must be emphasised that this is not statistically significant.

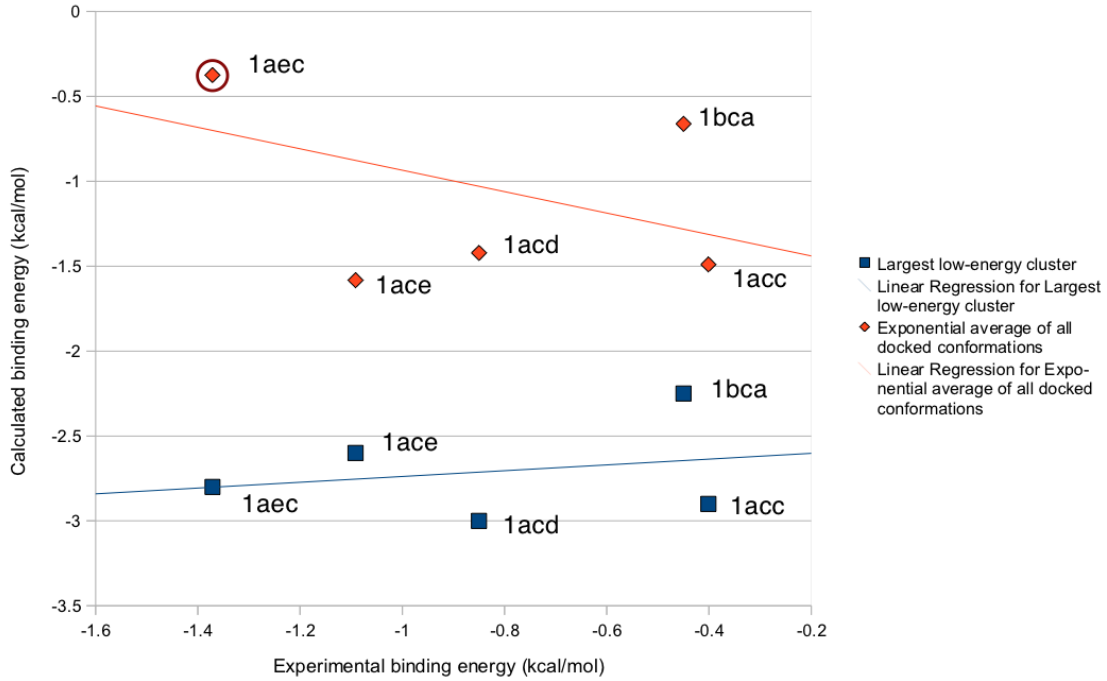

Figure 6.8: Calculated relative binding energy vs. experimental relative binding energy for exponentially averaged docked conformations from autodock (red) and energy of the largest low-energy cluster as identified by autodock (blue). There is a clear outlier in the results from the exponential averaging which is highlighted by the red circle.

6.4.2 Free energy calculations

6.4.2.a Single mutations vs. multiple mutations

We assess the difference between performing a single process mutation of all three side-chains compared to three stages of mutations that are combined by a thermodynamic cycle shown in figure 6.9. Performing a single process mutation rather than a three process mutation should provide a significant advantage. For a given amount of CPU time a single process mutation can sample for three times as long. This is particularly important in cases such as this where dihedral angles have long correlation times. We use the thermodynamic cycle to determine the total transformation from the sum of individual mutations (<1>-<8>):

$$\langle 1\rangle-\langle 8\rangle=\left[\langle 2\rangle+\langle 4\rangle+\langle 6\rangle\right]-\left[\langle 3\rangle+\langle 5\rangle+\langle 7\rangle\right] \quad (48)$$

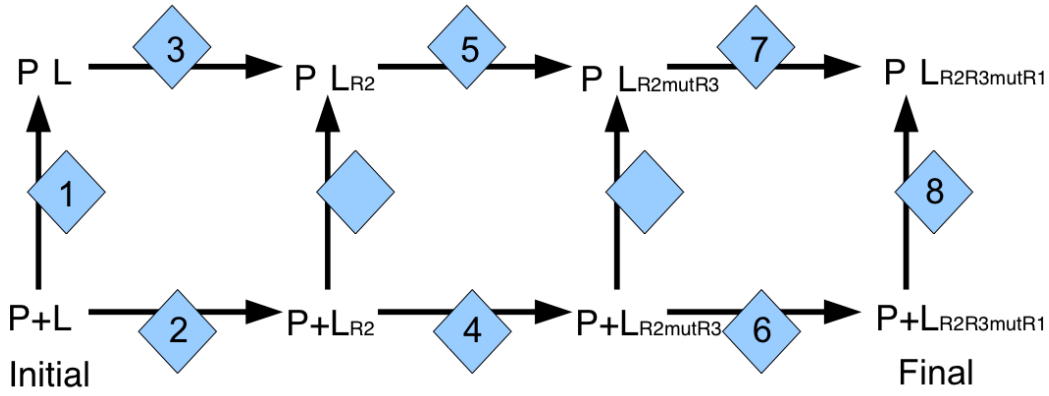

Figure 6.9: Thermodynamic cycle used to calculate the total energy change from the three triple mutations.

Results are shown in Table 6.5, which allow us to calculate the change in free energy for the single process mutation (-4.16 kcal mol^{-1}) to that of the three process mutation (-7.95 kcal mol^{-1}). This is obviously a large difference between the two methods. We can also compare the error associated with each measurement. The error associated for the single process mutation is slightly more (+/- 0.99 kcal mol^{-1} for the complex and +/- 0.124 kcal mol^{-1} for the solvent calculations, compared to +/- 0.104 kcal mol^{-1} for the error for an individual stage of the three process mutation) than the error associated with each individual stage of mutation in the three process strategy. However, when we add the error at each step (sum of squares of individual steps of the mutation) we notice a larger overall error in measurement (+/- 0.158 kcal mol^{-1} for the single process mutation and +/- 0.239 kcal mol^{-1} for the three process mutation). Further to the above, it is notable that the magnitude of the error compared to the magnitude of the result is

far larger in the case of mutating the side-chain at the third and first positions (Phenyl, and -CH-$(CH_3)_2$ respectively) compared to mutating the second side-chain (2-Napthyl), or performing the triple mutation.

To gain insight into possible causes of the discrepancy between the two calculations we first note that the solvent calculations differ by only 0.7 kcal mol^{-1} of which roughly 0.25 kcal mol^{-1} can be explained by statistical error. This implies that compared to the calculations for the complex which differ by 4 kcal mol^{-1} the solvent simulation is reasonably likely to be well converged. We know that the dihedral angles sampled in the complex calculation may not all be well sampled in the 5 ns simulations that we have performed, so extending the simulation for longer would be informative to see whether this encourages better convergence for the complex calculations. Furthermore carrying out three process mutations on some of the other complexes may help to identify which simulation method is most likely to be converged. In the case of Monte Carlo free energy calculations performed by Michel *et al.* the beta peptides used are perhaps more likely to be docked in the correct low energy conformations as they are likely to closer mimic the binding modes of the natural p53 ligand(Michel *et al.* 2009). In the case of large oligoamide compounds it is not clear that the correct binding mode is located. However, we tried to mitigate any possible problems here by using multiple starting conformations and using rigourous docking parameters. We also previously showed that it is likely that our calculations would need to be extended to allow better convergence in sampling of properties such as protein and ligand dihedral angles.

Calculation	ΔG/kcal mol^{-1}	ΔΔG/kcal mol^{-1}		(+/-) Error/kcal mol^{-1}		
1aec_conf2_complex	-20.11	-	-4.16	0.99	-	0.158
1aec_conf2_solvent	-24.27			0.124		
R2R3mutR1 complex	1.17	-3.16	-7.95	0.100	0.136	0.239
R2R3mutR1 solvent	-1.99			0.093		
R2mutR3 complex	-1.28	-0.5		0.100	0.140	
R2mutR3 solvent	-1.78			0.098		
mutR2 complex	-15.5	-4.29		0.089	0.137	
mutR2 solvent	-19.78			0.104		

Table 6.5: Three stage mutation from original compound to mutated R_2 position, to mutated R_2 and R_3 position to triple mutant R_2, R_3 and R_1 compared to mutation performed in a single step. All simulations use the replica-exchange method.

6.4.2.b Lambda schedules

The number of, and placement of lambda windows in such a way as to minimise the variance associated with each measurement is essential in allowing the desired accuracy in the calculation. It has further importance in determining the number of CPUs required to perform the calculation. In general lambda windows should be placed such that the variance associated with calculated at each window is approximately equal, since a single large variance will increase the variance of the entire calculation significantly. In figure 6.10 and 6.11 errors associated with the free energy calculated from moving from a value of lambda and the previous value of lambda. So the bar for lambda 1 represents the error for the measurement of moving from lambda 0 to lambda 1. The force field to which each lambda value corresponds to are shown in Table 6.1, 6.2 and 6.3 in the methods section. Broadly speaking in figure 8, the simulation with 12 lambda windows had the windows spaced half way between each value of the 24 window simulations. Doubling the number of windows halves the magnitude of the individual errors.

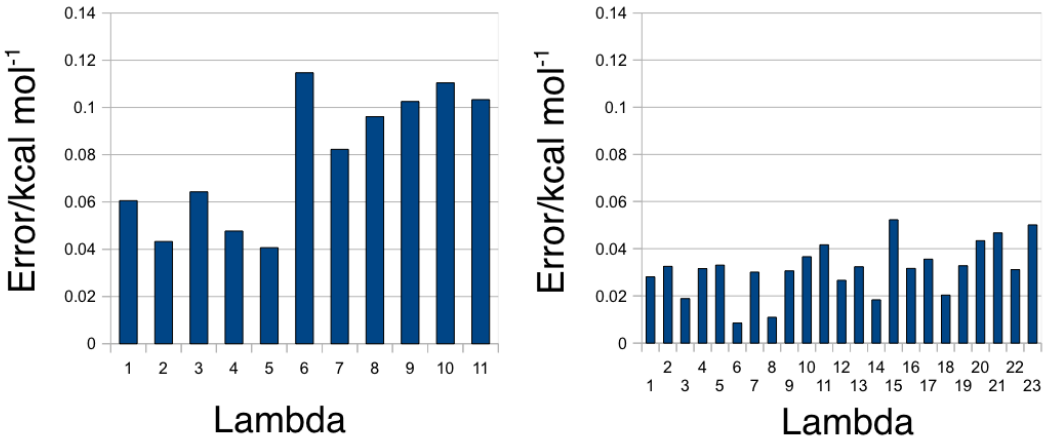

Figure 6.10: Error rate for simulation with 12 lambda windows (left) and 24 lambda windows (right), with electrostatics individually switched and van der Waals, bonded switched from on to off at the same time.

The lambda schedules used in figure 6.10 differ to those in figure 6.11, whilst they both switch off electrostatic interactions separately, the results in figure 6.11 are generated by individually switching off the van der Waals interactions, then the bonded followed by turning on the bonded and then the van der Waals interactions of the mutant (full details in Table 6.3). Unsurprisingly, switching off the electrostatic interactions produce a similar magnitude as in figure 6.10. Switching off/on the van der Waals interactions is also reasonable in terms of the magnitude of the errors. However, altering the bonded parameters has a very large error associated with it.

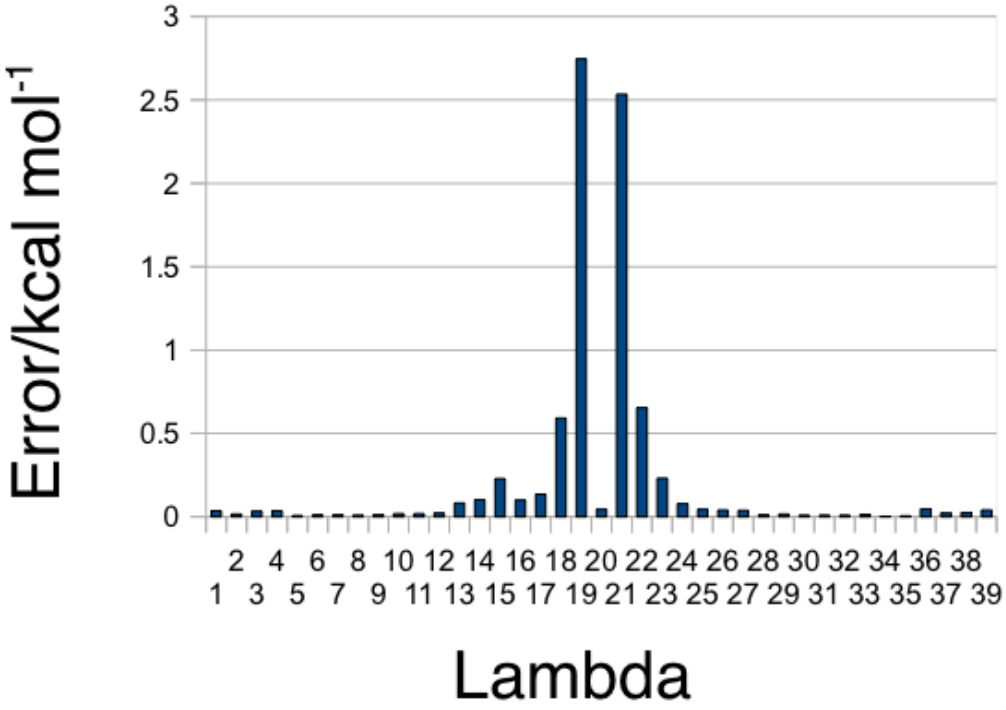

Figure 6.11: Error rate for simulation with 40 lambda windows with each parameter (electrostatics, van der Waals, bonded) individually switched from on to off.

The aim of performing the simulation shown in figure 6.11 was to identify whether particular parts of the simulation were adversely affecting the error for the total simulation. Since the total error for the simulation is somewhat related to the maximum error for a pair of lambda values it is desirable to spread lambda values such that they are all of similar magnitude. It is clear that the lambda schedules for 12 or 24 windows are both reasonable (with 24 windows producing a lower error measurement). However, it is not clear whether the results from 40 lambda windows is suggesting that bonded interactions are adding to the error, and that increasing the number of data points for bonded interactions will lower overall error, or whether this is merely an artefact of not performing the bonded interaction alterations at the same time as the van der Waals interaction alterations. We chose

to use the 24 lambda window method for further calculations since it appeared to show a reasonable overall statistical error rate whilst calculations would remain computationally tractable with the available resources.

6.4.2.c Hamiltonian exchange vs. a single long timescale simulation

Table 6.6 shows that in the case of the compound 1aec when using conformer number 2, the results from REMD and non-REMD are relatively similar. The total error associated with the calculation of the solvent steps is considerably lower in the case of the non-REMD calculation, however the values for each of the pairs of calculations are within 0.5 kcal mol^{-1} indicating that there is no major difference between the two methods. Given that the replica-exchange method has sampling benefits such as the increased likelihood of sampling dihedral angles that might otherwise be poorly sampled, it seems obvious that the replica-exchange method has clear benefits. However, we do not observe the performance benefits expected. One possible explanation is that replicas may not be exchanging in such a manner that might allow a fully coupled oligoamide (Phe-Trp-Leu) to transition to a fully decoupled state (-CH_3,-CH_3,-CH_3). Figure 6.12 shows that whilst the placement of lambda windows has been optimised to minimise statistical error this does not also necessarily encourage transition between all lambda states, and furthermore there are some states that undergo transition far more regularly than others. The error for the solvent non-remd simulation is roughly four times less than that of the equivalent replica-exchange calculation. It may be that this is due to the replica-exchange method sampling more phase space than the non-remd method resulting in an increase in the error.

Calculation	ΔG/kcal mol^{-1}	(+/-) Error/kcal mol^{-1}
1aec_conf2_complex remd	-20.11	0.099
1aec_conf2_solvent remd	-24.27	0.124
1aec_conf2_complex non-remd	-20.30	0.101
1aec_conf2_solvent non-remd	-24.90	0.032

Table 6.6: Numerical values required for the free energy calculation to compare the difference of performing a REMD simulation and a standard calculation using the same number of lambda windows, but no replica-exchange.

The exchange of replicas throughout the first 5 ns of a simulation of the conformation 2 of compound 1aec complexed with the protein is shown in Figure 6.12. It can be seen that at least 10 exchanges occur between all adjacent replicas. With many lambda values participating in far more than 10 swaps. The figure allows us to identify some important points, firstly that adequate sampling is occurring, and secondly it could be used as a tool to identify regions with many swaps that might require fewer lambda windows, and regions with few swaps that might benefit more lambda windows.

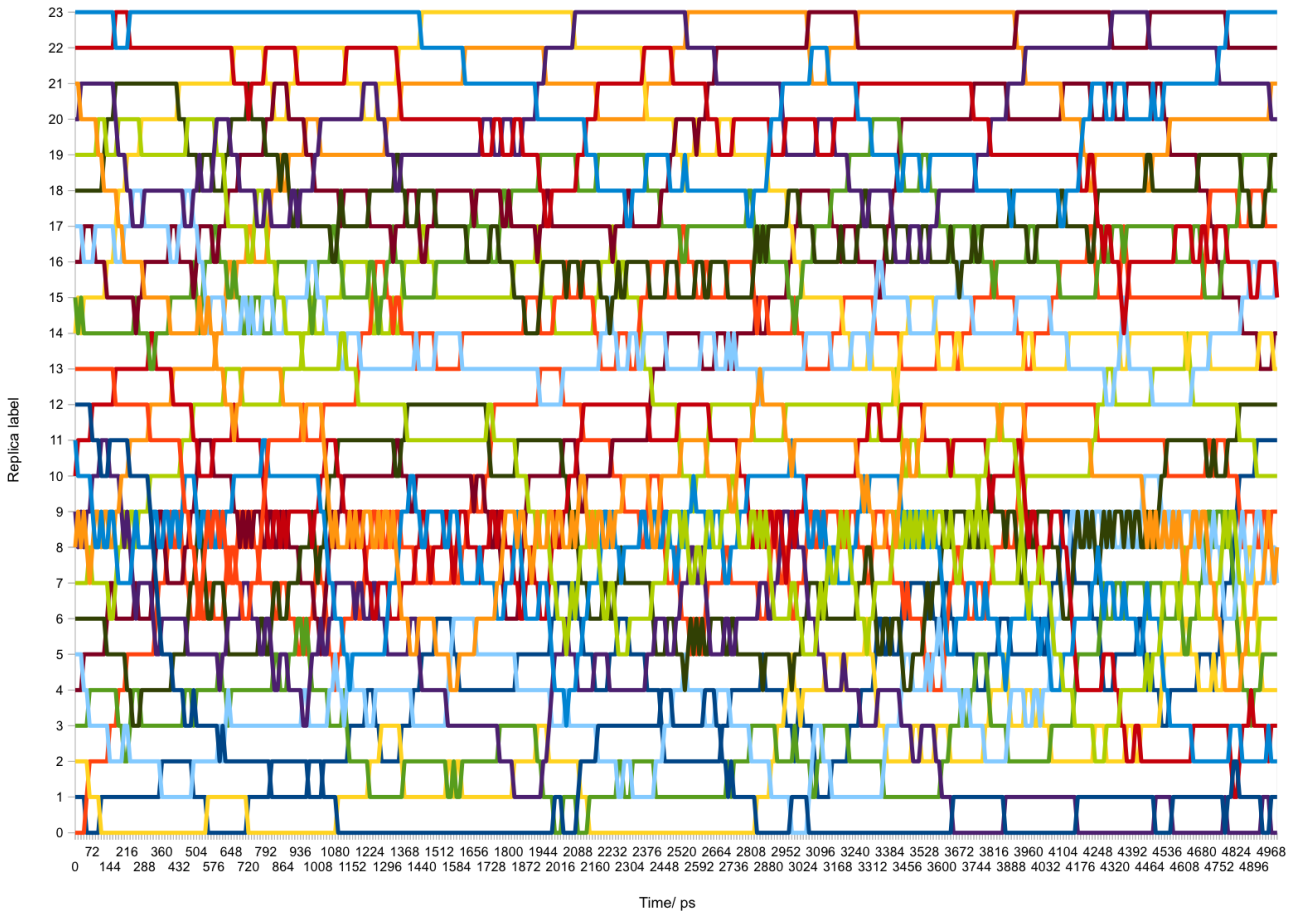

Figure 6.12: replica-exchange swaps shown over time (5 ns). Initial replicas are labelled on the y-axis and maintain their colour throughout the simulation. Swaps between neighbours are attempted every 12 ps and are subject to the detailed balance criteria discussed in the introduction.

Simulation Name	Parallel/anti-parallel	# "Mutated out" atoms	Dispersion correction Solvent (simulation) /kcal mol^{-1}	Dispersion correction Complex/kcal mol^{-1}	Solvent/kcal mol^{-1}	Complex/kcal mol^{-1}	Corrected solvent/kcal mol^{-1}	Corrected complex/kcal mol^{-1}	Relative K_d exp(-(complex-solvent)/RT)	Combined relative K_d/M	Dispersion Correction Combined relative K_d/M	Experimental IC_{50}/μM
1aaa_conf1	p		-0.41	-0.6	13.88	16.77	13.47	16.17	1.08E-02			
1aaa_conf2	ap		-0.41	-0.52	12.59	16.73	12.17	16.21	1.15E-03			
1aaa_conf3	ap	18	-0.41	-0.5	13.06	15.89	12.64	15.38	1.01E-02	4.30E-03	5.42E-03	10
1acc_conf1	p		-0.62	-0.92	-14.69	-10.81	-15.31	-11.72	2.44E-03			
1acc_conf2	ap		-0.62	-0.93	-16.17	-10.34	-16.79	-11.27	9.40E-05			
1acc_conf3	ap	24	-0.62	-0.78	-14.31	-10.21	-14.93	-10.99	1.35E-03	5.61E-04	8.61E-04	5.1
1acd_conf1	ap		-0.74	-1.01	-14.36	-8.07	-15.1	-9.08	4.05E-05			
1acd_conf2	p		-0.74	-1.07	-14.22	-9.65	-14.96	-10.72	8.06E-04			
1acd_conf3	p	30	-0.74	-1.12	-14.68	-9.82	-15.42	-10.93	5.37E-04	1.10E-04	1.83E-04	2.4
1ace_conf1	ap		-0.74	-1	-17.08	-12.18	-17.82	-13.18	4.10E-04			
1ace_conf2	ap		-0.74	-1.04	-17.02	-11.86	-17.76	-12.9	2.84E-04			
1ace_conf3	p	30	-0.74	-1.14	-17.38	-13.66	-18.12	-14.81	3.84E-03	3.46E-04	5.69E-04	1.6
1aec_conf1	p		-0.74	-1.15	-24.48	-18.43	-25.22	-19.58	7.71E-05			
1aec_conf2	p		-0.74	-0.99	-24.27	-20.11	-25.01	-21.11	1.44E-03			
1aec_conf3	p		-0.74	-1.05	-24.51	-20.08	-25.25	-21.14	9.94E-04			
1aec_conf7	p		-0.74	-1.04	-24.01	-18.83	-24.75	-19.87	2.79E-04			
1aec_conf8	p	30	-0.74	-1.09	-24.27	-21.13	-25.01	-22.21	9.13E-03	2.91E-04	5.13E-04	1
1aec_conf4	a		-0.74	-1.14	-24.15	-18.8	-24.9	-19.95	2.45E-04			
1aec_conf9	a		-0.74	-1.14	-24.15	-18.8	-24.9	-19.95	2.45E-04			
1aec_conf10	a		-0.74	-1.08	-24.42	-17.97	-25.16	-19.05	3.47E-05			
1aec_conf11	a	30	-0.74	-1.19	-23.91	-18.43	-24.65	-19.62	2.12E-04	7.10E-05	1.37E-04	1
1bca_conf1	ap		-0.51	-0.65	-7.35	-3.07	-7.86	-3.72	9.63E-04			
1bca_conf2	ap		-0.51	-0.7	-8.1	-4.62	-8.61	-5.32	3.94E-03			
1bca_conf3	p	20	-0.51	-0.71	-9.33	-4.9	-9.84	-5.6	8.18E-04	1.05E-03	1.38E-03	4.7

Table 6.7: Results from free energy calculations, including whether the starting conformation was in a parallel or anti-parallel conformation, the magnitude of dispersion correction applied, the calculated relative free energy for mutation from compound to a triple -CH_3 substituted compound for individual simulations, the average of the relative free energy for a compound and comparison to the experimental IC_{50}

6.4.2.d Overlap integrals

Transition	replica-exchange	Non replica-exchange	Mutate R2	R2 mutated, mutate R3	R2, R3 mutated, mutate R1
0 - 1	0.131	0.188	0.013	0.031	0.022
1 - 2	0.033	0.387	0.020	0.026	0.040
2 - 3	0.122	0.249	0.110	0.041	0.127
3 - 4	0.326	0.850	0.194	0.075	0.059
4 - 5	0.157	0.338	0.041	0.155	0.057
5 - 6	0.055	0.564	0.093	0.067	0.038
6 - 7	0.098	0.598	0.127	0.070	0.106
7 - 8	0.140	0.587	0.297	0.194	0.093
8 - 9	0.061	0.826	0.297	0.197	0.035
9 - 10	0.444	0.372	0.253	0.267	0.379
10 - 11	0.031	0.142	0.162	0.188	0.138
11 - 12	0.015	0.320	0.076	0.047	0.140
12 - 13	0.025	0.003	0.150	0.055	0.308
13 - 14	0.090	0.116	0.281	0.267	0.253
14 - 15	0.150	0.379	0.221	0.039	0.044
15 - 16	0.051	0.453	0.099	0.042	0.148
16 - 17	0.052	0.157	0.157	0.410	0.069
17 - 18	0.072	0.672	0.191	0.061	0.124
18 - 19	0.082	0.453	0.051	0.101	0.062
19 - 20	0.058	0.166	0.086	0.058	0.108
20 - 21	0.058	0.134	0.042	0.021	0.020
21 - 22	0.047	0.462	0.050	0.021	0.207
22 - 23	0.033	0.108	0.085	0.036	0.012

Table 6.8 Overlap integrals for a variety of simulations of the 1aec conformation 2 complex. Overlap close to zero indicates that the two states are distant in phase space, and a very large number of samples would need to be collected in order to obtain a good estimate of free energy. Overlap of one is the largest that can be obtained, and indicates that states are close in phase space, and would need correspondingly fewer samples to obtain a good estimate of free energy. Note that the overlap integral is a unit-less quantity.

The most striking result from table 6.8 is that the overlap integrals from the non replica-exchange simulations are generally considerably larger than those from the replica-exchange simulations (which include the three process mutations). We observe slightly lower variance for the non replica-exchange simulations and we sample the same number of configurations in each simulation. Hence we observe that the overlap between accessible states for the non replica-exchange simulations is larger than that of the replica-exchange simulations. When comparing to the average overlap between the three processes compared to the single process replica-exchange simulation we observe similar trends. The implication here would be that it is equally efficient to perform a single large mutation rather than three steps of smaller mutations that arrive at the same state. This would be a novel result that could benefit researchers investigating compounds with similar scaffolds or backbones, but with a diverse range of attachment points and functional groups. The ambiguity in whether to use replica-exchange methods requires further investigation. Simulations not using a replica-exchange scheme, may find it more valuable to increase the length of simulations rather than using more lambda windows, since overlap appears to be high. However, simulations using the replica-exchange methodology may perform better with more lambda windows. Whether one methodology is better than the other is unclear, since it may be that the replica-exchange method suffers from higher variance (and thus lower overlap) due to sampling more energy minima, whilst the non replica-exchange method may be stuck sampling a single energy minima.

6.4.2.e Determining best binding oligoamides

Numerical results calculated from simulation trajectories are included in table 6.7. The dispersion correction applied to simulations is detailed in the methods section. The number of mutated atoms which are present in the A state, but absent in the B state is included (“mutated out” atoms), and the free energy changes are calculated with and without dispersion correction. Results from replicate compounds are combined using the exponential summation method detailed in

the methods(Mobley, Chodera, and Dill 2006). Results are also categorised by whether they are a parallel or anti-parallel conformation as was originally discussed in the previous two chapters.

6.4.2.f Combining results from simulations

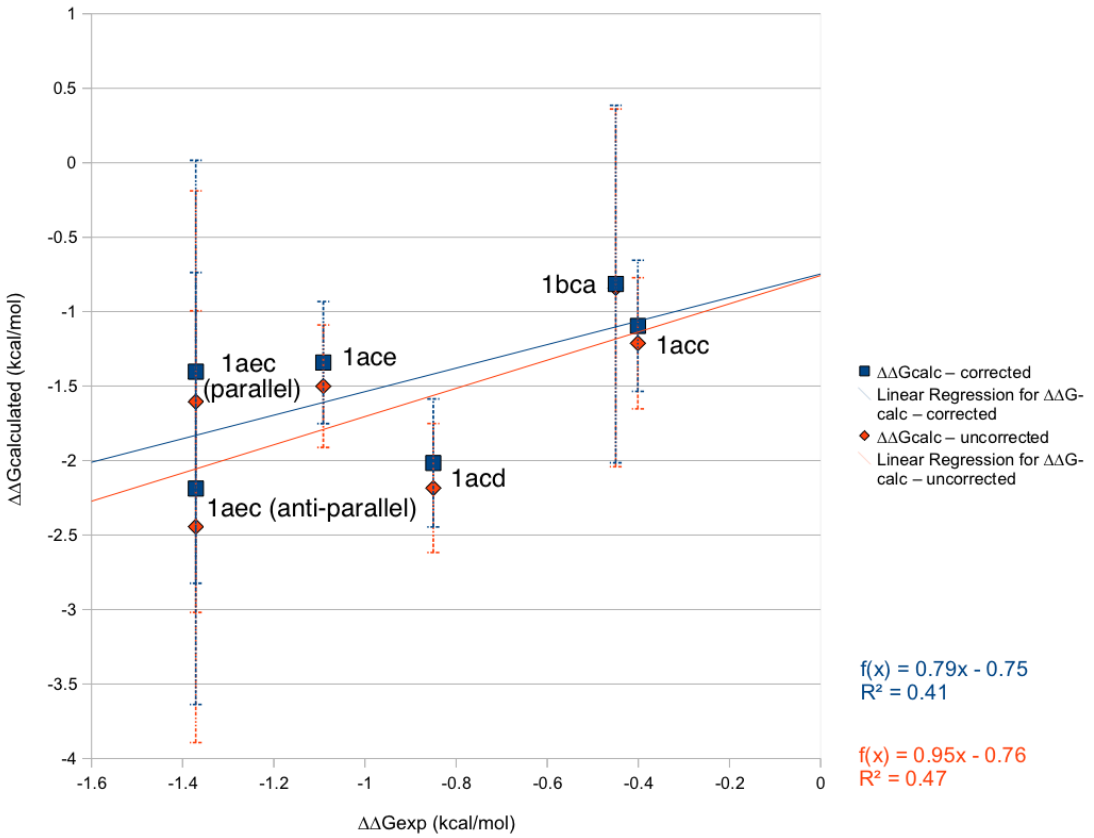

Figure 6.13: Experimental free energy change compared to calculated free energy change, with and without van der Waals corrections applied.

Figure 6.13 shows calculated ΔΔG values for each of the 5 compounds plotted against the experimental ΔΔG calculated from IC_{50}. Whilst it is not strictly correct to convert IC_{50} values into binding energies, in the absence of K_d values it is the only option, and since the values are all calculated in the same lab using the same experimental technique it is probably reasonable to assume quantitative comparisons can be made. The calculated results for both the uncorrected (red) and corrected (blue) calculated values appear to correlate well with the

experimental values. The uncorrected results actually correlate slightly better than the corrected results (R_2 = 0.47 vs. R_2 = 0.41), and both sets of results have a slope approaching the desired value of 1. The graph also shows that the calculated free energy of the parallel conformations of compound 1aec are less favourable than the anti-parallel conformations (-1.4 kcal mol^{-1} vs. -2.9 kcal mol^{-1}, for the corrected values). This appears to be weakly followed in the case of the other compounds, and may be an important consideration required in a structure-based design strategy. We have also calculated the RMSD from the experimental binding energies determining a value of 1.99 kcal mol^{-1} when the dispersion correction has not been applied, and 1.64 kcal mol^{-1} when the dispersion correction is applied.

6.5 Conclusion

We showed that the docking methods that had previously been developed may have required slightly increased sampling, and we employed the enhanced sampling to generate starting conformations for the compounds that we investigated. Therefore 300 conformers were generated for each compound rather than the previous 150 and additionally a maximum of 25 million rather than 2.5 million genetic algorithm generations were generated. We also investigated the Autodock binding energy and the energy of the largest low-energy cluster from Autodock as possible determinants of binding energy. We observed that the energy of the largest low-energy cluster showed some correlation if used to rank relative positions rather than actual energies. Further to this we showed that the energies of the relative energies of the Autodock scores were in general in good agreement with our free energy calculations and experiment. However, this interpretation can be challenged as there was an exception for the 1aec conformation which is experimentally the best binder, but was predicted by Autodock as the worst.

The difference in the three single mutations and triple mutation is very significant and it probably illustrates the requirement for good sampling. Theoretically performing the triple mutant should show some benefits over performing the three single mutations. This should allow sampling for roughly three times as long giving more probability of sampling dihedrals well. In practice due to problems with the Desmond algorithm the parameter clone radius (a cloned copy of data in the vicinity of the simulation box of the local cpu process, required for efficient parallelization) had to be extended which required addition of more water molecules thus increasing simulation time somewhat.

Use of replica-exchange methods was successful and results from these simulations agreed well with the comparison simulations when replica-exchange was not used. The main advantage is that sampling should be better. In theory dihedrals that are constrained in the A state but not in the B state could transition from the A state to the B state and rotate, after which time they could transition back to the B state. We can substitute into equation 42 the value of lambda for 15 (a state that does reach lambda 0 in simulation of 1aec conformation 2), to see what magnitude the energy barrier might be in this situation. In the case of lambda 15 van der Waals interactions for the triple CH_3 mutant are all fully switched on (meaning the atoms will have their usual van der Waals radius), whilst the Phe-Trp-Leu side-chains have their interactions scaled by 0.325. If we look at the volume with energy greater than 2 kT and ignore the attractive $1/r^6$ term we can work out the excluded volume for the Phe-Trp-Leu side-chains is still 76 % of the excluded volume for lambda 0. This means that it is still very unlikely that the side-chains would be free to rotate easily. The conclusion here is that Hamiltonian replica-exchange may not be improving sampling at all, whilst adding additional complexity to the simulations. If Hamiltonian replica-exchange is desired to be used more often then work may be needed on placement of lambda windows in order to maximise the overlap integral and thus encourage more swaps to be made that can transition between all lambda states.

In principal we don't see an individual replica managing this journey, but since there is mixing between all lambda states this will eventually occur in the limit of a longer simulation. A major issue with replica-exchange is that in order to run a simulation at a sensible speed 8 CPU cores are required for an individual replica, thus when 24 replicas are used 192 cpu cores are required. The advent of GPU co-processors might alleviate this problem somewhat as this many cores could be available on several co-processor cards, rather than requiring a relatively large computer cluster with fast interconnects. Whereas government funded research is generally likely to be able to gain access to a reasonable amount of CPU time since there are a fairly large number of high-performance computing facilities, pharmaceuticals companies (with the exception of the very largest) are likely to find this a significant barrier to access. To calculate 192 processes running in one box would require 12 core cpus with 2 threads running on each core on a 4 socket box. However, this is likely to be achievable in the near future, so it might be more reasonable to look at the fact that the method is suited for application in a parallel environment as a benefit in the longer term. It should also be noted that currently running large multiple mutations in a single step with Desmond is not officially supported, and one of the requirements is to increase the value of a parameter r_clone, which in turn requires simulation in a box that is larger than would otherwise be required. This means that whilst theoretically a single-step mutation should be more efficient than a three-step mutation this is currently not the case. However, if this issue could be overcome it would be possible to generate trajectories that are roughly three times as long when using the single-step mutation strategy, which would help to overcome the problem of adequate sampling of dihedral angles.

The results from the free energy calculations presented here appear to show good correlation with the experimental data currently available, however, the range of energies from the experimental data is within that of the error associated with the

measurement. Better validation of the method would be possible with a larger range of binding affinities of oligoamide compounds, ideally with some affinities being in the nM range. We have already discussed that comparison to free energies calculated from IC_{50} data is not ideal. However, it was done in the most reasonable way possible, where all data is from the same assay performed in the same lab. The single-step mutation strategy is a novel technique that is not officially supported in Desmond, and appears to show utility in accurate prediction of binding energies. However, further work is required to identify why overlap is less for replica-exchange simulations compared to non replica-exchange simulations.

6.6 References

Bowers, Kevin J., Federico D. Sacerdoti, John K. Salmon, Yibing Shan, David E. Shaw, Edmond Chow, Huafeng Xu, *et al.* 2006. *Molecular dynamics---Scalable algorithms for molecular dynamics simulations on commodity clusters*. *Proceedings of the 2006 ACM/IEEE conference on Supercomputing - SC '06*. New York, New York, USA: ACM Press. doi:10.1145/1188455.1188544. http://portal.acm.org/citation.cfm?doid=1188455.1188544.

Cossins, Benjamin P, Sebastien Foucher, Colin M Edge, and Jonathan W Essex. 2009. Assessment of nonequilibrium free energy methods. *The journal of physical chemistry. B* 113, no. 16 (April): 5508-19. doi:10.1021/jp803532z. http://www.ncbi.nlm.nih.gov/pubmed/19368411.

Huey, Ruth, Garrett M. Morris, Arthur J. Olson, and David S. Goodsell. 2007. A semiempirical free energy force field with charge-based desolvation. *Journal of Computational Chemistry* 28, no. 6 (April): 1145-52. doi:10.1002/jcc.20634. http://www.ncbi.nlm.nih.gov/pubmed/17274016.

Michel, Julien, Elizabeth A Harker, Julian Tirado-Rives, William L Jorgensen, and Alanna Schepartz. 2009. In Silico Improvement of beta3-peptide inhibitors of p53 x hDM2 and p53 x hDMX. *Journal of the American Chemical Society* 131, no. 18 (May): 6356-7. doi:10.1021/ja901478e. http://www.ncbi.nlm.nih.gov/pubmed/19415930.

Mobley, David L, John D Chodera, and Ken A Dill. 2006. On the use of orientational restraints and symmetry corrections in alchemical free energy calculations. *The Journal of chemical physics* 125, no. 8: 084902. doi:10.1063/1.2221683. http://www.ncbi.nlm.nih.gov/pubmed/16965052.

Mobley, David L., John D. Chodera, and Ken A. Dill. 2007. The Confine-and-Release Method: Obtaining Correct Binding Free Energies in the Presence of Protein Conformational Change. *Journal of Chemical Theory and Computation* 3, no. 4 (July): 1231-1235. doi:10.1021/ct700032n. http://www.ncbi.nlm.nih.gov/pubmed/18843379.

Moreira, Irina S., Pedro A. Fernandes, and Maria J. Ramos. 2008. Protein–protein recognition: a computational mutagenesis study of the mdm2–p53 complex. *Theoretical Chemistry Accounts* 120, no. 4-6 (July): 533-542. doi:10.1007/s00214-008-0432-9.

Morris, Garrett M., David S. Goodsell, Robert S. Halliday, Ruth Huey, William E. Hart, Richard K. Belew, and Arthur J. Olson. 1998. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry* 19, no. 14 (November): 1639-1662. doi:10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B. http://doi.wiley.com/10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B.

Plante, Jeffrey P., Thomas Burnley, Barbora Malkova, Michael E. Webb, Stuart L. Warriner, Thomas A. Edwards, and Andrew J. Wilson. 2009. Oligobenzamide proteomimetic inhibitors of the p53–hDM2 protein–protein interaction. *Chemical Communications*, no. 34: 5091. doi:10.1039/b908207g. http://xlink.rsc.org/?DOI=b908207g.

Shaginian, Alex, Landon R Whitby, Sukwon Hong, Inkyu Hwang, Bilal Farooqi, Mark Searcey, Jiandong Chen, Peter K Vogt, and Dale L Boger. 2009. Design, Synthesis, and Evaluation of an alpha-Helix Mimetic Library Targeting Protein-Protein Interactions. *Journal of the American Chemical Society*, no. 4: 5564-5572. doi:10.1021/ja810025g. http://www.ncbi.nlm.nih.gov/pubmed/19334711.

Shirts, Michael R, David L Mobley, John D Chodera, and Vijay S Pande. 2007. Accurate and efficient corrections for missing dispersion interactions in molecular simulations. *The Journal of Physical Chemistry B* 111, no. 45 (November): 13052-63. doi:10.1021/jp0735987. http://www.ncbi.nlm.nih.gov/pubmed/17949030.

Shirts, Michael R., and Vijay S. Pande. 2005. Solvation free energies of amino acid side chain analogs for common molecular mechanics water models. *The Journal of Chemical Physics* 122, no. 13: 134508. doi:10.1063/1.1877132. http://www.ncbi.nlm.nih.gov/pubmed/15847482.

Woods, Christopher J., Jonathan W. Essex, and Michael A. King. 2003. The Development of Replica-Exchange-Based Free-Energy Methods. *The Journal of Physical Chemistry B* 107, no. 49 (December): 13703-13710. doi:10.1021/jp0356620. http://pubs.acs.org/doi/abs/10.1021/jp0356620.

Zhong, Haizhen, and Heather A Carlson. 2005. Computational studies and peptidomimetic design for the human p53-MDM2 complex. *Proteins* 58, no. 1 (January): 222-34. doi:10.1002/prot.20275. http://www.ncbi.nlm.nih.gov/pubmed/15505803.

7 Final conclusions

7.1 Overview of results

7.1.1 Protein-protein interactions as drug targets

We showed that there are marked differences between the volume of pockets observed in protein-ligand interactions compared to those observed in protein-protein interactions. Furthermore we showed that protein-drug interactions mirror protein-ligand interactions, whilst protein-protein interaction inhibitors (137 $Å^3$) tend to have pocket volumes that lie somewhere between those observed in protein-ligand interactions (260 $Å^3$) and protein-protein interactions (54 $Å^3$). We also observed differences in the number of pockets targeted by protein-ligand, protein-drug interactions (one) when compared to protein-protein interactions (between five and eight). Once again we notice that protein-protein interaction inhibitors (three to five) fall somewhere between protein-ligand, protein-drug interactions and protein-protein interactions. We conclude that properties of the pockets on protein surfaces might guide as when deciding whether a particular protein-protein interaction might be amenable to inhibition.

7.1.2 Predicting protein 'drugability'

Following on from the differences that we observed in the volume and number of pockets on protein-protein interactions that have been targeted for inhibition we attempted to develop a method for determining whether a pocket on a protein surface might be 'druggable'. We based our work on ideas from previous work by Halgren and by Hajduk(Halgren 2009),(Hajduk, Huth, and Fesik 2005). We looked at the distributions of GRID atom energies within pockets that bind ligands and those that don't and additionally compared to the GRID energies of GRID atoms

centred on ligand atoms observed in the PDB. We developed a number of descriptors for each pocket on a protein surface and analysed the differences between them. We noted that many of the differences between bound and unbound pockets are not statistically significant. We attempted to apply our method to predict bound from unbound pockets with limited success, and attempted to further extend this to classifying 'druggable' from 'undruggable' pockets. We are interested to use existing methods for predicting drugability from Cheng(Cheng *et al.* 2007) or Halgren(Halgren 2009) to assess how they perform at identifying bound from unbound sites, compared to our machine learning method.

7.1.3 Alchemical free energy calculations

We performed computational docking using Autodock to generate hDM2-oligoamide conformers in the absence of X-ray crystal structures of the complexes. We attempted to better validate these models by performing molecular dynamics simulations, that remained stable using parameters similar to those used for free energy calculations. We also investigated charge parameterization methods for the oligoamide compounds where we identified the AM1 BCC charge calculation method provided with AMBER as being most suitable for our system. We then developed methods to assess whether we thought our system was likely to produce converged free energy calculations in the simulation time available to us. We particularly investigated whether oligoamide compounds converted on the time-scale of our simulations, which dihedral angles were most likely to be poorly sampled, and the spatial sampling of the binding site.

Alchemical free energy calculations were performed with Desmond using the AMBER99sb and GAFF forcefields. We performed standard simulations at 24 lambda values, which we compared to a replica-exchange method. Replica-exchange calculations conceded a performance penalty when carried out, although it may be possible to minimize this by altering the way the calculation is performed. It is not clear that in their current form replica-exchange calculations

did improve sampling in the way expected, but this may be possible with better placement of lambda windows. Replica-exchange calculations agreed with non-replica-exchange calculations to within 0.5 kcal mol^{-1}. The single process mutations and triple process mutations varied by 4 kcal mol^{-1}. They should agree to within experimental error, it is not clear why this is. Further work on calculating the overlap in phase space is needed, and quantification in the difference in calculated overlap when using Bennett Acceptance Ratio variances to compute the overlap, and when using a bootstrapped value of the variance.

7.2 Implications for drug-discovery

It is becoming more and more obvious that current pharmaceutical strategies cannot be sustainable. It is incredibly time-consuming and expensive to bring a new drug to the market. Broadly speaking this leaves two options to the industry. The first is to use existing drugs better. Examples include using currently approved drugs for new purposes(Ashburn and Thor 2004) (repurposing), or using information of the patients genetic make-up to select the drug that will work best(van't Veer and Bernards 2008) (personalized medicine). The second option is to identify and develop new entities better. Whilst Protein-protein interactions in general have been difficult to exploit thus far(Fuller, Burgoyne, & Jackson, 2009), the concept of using a relatively small number of backbone compounds that mimic naturally occurring structural motifs, may offer a much more efficient method to target protein-protein interactions.

7.3 Future Directions

7.3.1 Pocket detection

Recent years have seen a wide variety of pocket detection techniques becoming available to researchers. It is clear that they have utility in assisting the analysis of both individual protein targets, and also large datasets of proteins. It is also important that they remain available to researchers in a format that is relevant to

them. That might be as a web-server, a stand-alone program or as a web-service. Additionally integration of pocket detection algorithms into currently available databases would allow better use of data. An example could be the integration of Q-SiteFinder into SitesBase (a database of ligand binding sites) which would allow searching of identified binding sites against sites defined by the shape of ligands known to bind to an existing site(Gold and Jackson 2006). Methods such as this would be of assistance in both drug repurposing and functional annotation of proteins.

7.3.2 Drugability

The work on drugability that we previously discussed might fit in well in the space of repurposing where it could be applied in a manner similar to that of functional annotation as described in the paragraph above. That is the drugability index might allow a protein target to be assessed for whether it might bind to a list of approved drugs. This would first require existing drugability methods to be fully evaluated to highlight their strengths and weaknesses.

7.3.3 Targeting protein-protein interactions with common scaffolds

For targeting protein-protein interactions, using a single backbone (such as an oligoamide) has several key benefits: ease of synthesis; optimisation of solubility; simplicity of working with computationally. It is clear that once a backbone of interest is identified efforts can be made to develop a synthetic method that means that a library of side-chains can easily be attached. In the case of the oligoamide compounds that we worked on this has been achieved by at least two groups(Plante *et al.* 2008),(Shaginian *et al.* 2009). Secondly in the case of the oligoamides that we worked on the solubility of the compounds is a clear limiting factor. The backbone is large and greasy meaning that it is difficult to dissolve in water, this is not likely to be good for binding affinity in general. Therefore if an alternative backbone could be designed with similar properties but better solubility

the affinity of the entire library of compounds would be increased. Computationally one of the challenges of simulating these oligoamide compounds was generating parameters to accurately describe the compounds(Vemparala *et al.* 2006), (Pophristic *et al.* 2006). Fortunately in our case extensive previous work in developing dihedral parameters for our force field has been undertaken. If these parameters had not been available it may have taken far longer to complete the simulations, or they may not have achieved the same quality. Limiting the number of backbones used would make computational simulation more straightforward in this respect.

It is clear that the compounds in general are not likely to be suitable as inhibitors of the hDM2 interaction due to their relatively low affinity, part of this is probably due to the solubility of the compounds as we already discussed. It is probable that substituting alternative side-chains onto the oligoamide backbone would be a suitable way to increase affinity of the compounds to some degree. Indeed virtual screening, fragment or rapid docking based methods, may all be suitable for this. If a particularly large library was available these might even be pre-screened using a simple filter before passing to the methods described above. Each of these methods might be able to screen a dataset of several thousand compounds down to a list of a hundred or so that could be further processed before being computationally screened using the free energy methods described.

7.3.4 Alchemical free energy calculations

In the longer term it would be good to compare calculations to measured K_d, something that could be achieved by using a different assay to the fluorescence polarisation assay originally used by Plante *et al.*(Plante *et al.* 2009). A wider range of binding affinities (hundreds of μM to sub nM) of compounds with correctly measured K_ds would allow better validation of a free energy calculation method. Ideally synthesis of compounds could be guided by a high-throughput virtual screen of suitable side-chains that might produce a list of several hundred

candidate compounds. Further input could be provided by synthetic chemists to select around ten compounds for synthesis. These compounds could be used for free energy calculations which would allow post validation of the technique.

In general alchemical free energy calculations are difficult to perform and computationally expensive. However, performing simulations such as those detailed in this thesis allows the identification of which methods are appropriate for certain situations, and whether certain algorithms and techniques can be applied more generally. It is clear that whilst we didn't conclusively show Hamiltonian replica-exchange to be beneficial it is a method that is relatively easy to perform, and has benefits in the relative ease of creating a parallel implementation of the method. Developing and following best practice in the field of free energy calculations should allow for easier comparison of results between labs as well as making it easier to create tools to perform alchemical free energy calculations more easily. As computer power increases, and the quality of the tools available increases it appears very likely that free energy calculations will be of obvious utility for the development of novel ligands to target protein-protein interactions.

7.4 References

Ashburn, Ted T, and Karl B Thor. 2004. Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery* 3, no. 8 (August): 673-83. doi:10.1038/nrd1468. http://www.ncbi.nlm.nih.gov/pubmed/15286734.

Cheng, Alan C, Ryan G Coleman, Kathleen T Smyth, Qing Cao, Patricia Soulard, Daniel R Caffrey, Anna C Salzberg, and Enoch S Huang. 2007. Structure-based maximal affinity model predicts small-molecule druggability. *Nature Biotechnology* 25, no. 1: 71-5. doi:10.1038/nbt1273. http://www.ncbi.nlm.nih.gov/pubmed/17211405.

Fuller, Jonathan C, Nicholas J Burgoyne, and Richard M Jackson. 2009. Predicting druggable binding sites at the protein-protein interface. *Drug Discovery Today* 14, no. 3-4 (February): 155-61. doi:10.1016/j.drudis.2008.10.009. http://www.ncbi.nlm.nih.gov/pubmed/19041415.

Gold, Nicola D, and Richard M Jackson. 2006. SitesBase: a database for structure-based protein-ligand binding site comparisons. *Nucleic Acids Research* 34, no. Database issue (January): D231-4. doi:10.1093/nar/gkj062. http://www.ncbi.nlm.nih.gov/pubmed/16381853.

Hajduk, Philip J, Jeffrey R Huth, and Stephen W Fesik. 2005. Druggability indices for protein targets derived from NMR-based screening data. *Journal of Medicinal Chemistry* 48, no. 7 (April): 2518-25. doi:10.1021/jm049131r. http://www.ncbi.nlm.nih.gov/pubmed/15801841.

Halgren, Thomas a. 2009. Identifying and characterizing binding sites and assessing druggability. *Journal of Chemical Information and Modeling* 49, no. 2 (February): 377-89. http://www.ncbi.nlm.nih.gov/pubmed/19434839.

Plante, JP, Thomas Burnley, Barbora Malkova, Michael E. Webb, Stuart L. Warriner, Thomas A. Edwards, and Andrew J. Wilson. 2009. Oligobenzamide proteomimetic inhibitors of the p53–hDM2 protein–protein interaction. *Chemical Communications*, no. 34: 5091. doi:10.1039/b908207g. http://xlink.rsc.org/?DOI=b908207g.

Plante, J, Fred Campbell, Barbora Malkova, Colin Kilner, Stuart L. Warriner, and Andrew J. Wilson. 2008. Synthesis of functionalised aromatic oligamide rods. *Organic & Biomolecular Chemistry* 6, no. 1 (January): 138-46. doi:10.1039/b712606a. http://www.ncbi.nlm.nih.gov/pubmed/18075658.

Pophristic, Vojislava, Satyavani Vemparala, Ivaylo Ivanov, Zhiwei Liu, Michael L Klein, and William F DeGrado. 2006. Controlling the shape and flexibility of arylamides: a combined ab initio, ab initio molecular dynamics, and classical molecular dynamics study. *The Journal of Physical Chemistry B* 110, no. 8 (March): 3517-26. doi:10.1021/jp054306+. http://www.ncbi.nlm.nih.gov/pubmed/16494407.

Shaginian, Alex, Landon R Whitby, Sukwon Hong, Inkyu Hwang, Bilal Farooqi, Mark Searcey, Jiandong Chen, Peter K Vogt, and Dale L Boger. 2009. Design, Synthesis, and Evaluation of an alpha-Helix Mimetic Library Targeting Protein-Protein Interactions. *Journal of the American Chemical Society*, no. 4: 5564-5572. doi:10.1021/ja810025g. http://www.ncbi.nlm.nih.gov/pubmed/19334711.

van't Veer, Laura J, and René Bernards. 2008. Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature* 452, no. 7187 (April): 564-70. doi:10.1038/nature06915. http://www.ncbi.nlm.nih.gov/pubmed/18385730.

Vemparala, Satyavani, Ivaylo Ivanov, Vojislava Pophristic, Katrin Spiegel, and Michael L Klein. 2006. Ab initio calculations of intramolecular parameters for a class of arylamide polymers. *Journal of Computational Chemistry* 27, no. 6 (April): 693-700. doi:10.1002/jcc.20382. http://www.ncbi.nlm.nih.gov/pubmed/16634095.