Computational Studies of the
Cytochromes P450

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

Stewart Brian Kirton B.Sc. Hons (Leicester)

Department of Chemistry

Faculty of Science

University of Leicester

September 2002
In Loving Memory of my Grandfather
Frederick Burns
Statement

The work described in this thesis has been performed by the author in the Department of Chemistry at the University of Leicester, and the Department of Physical and Metabolic Science at AstraZeneca Charnwood, Loughborough, between October 1999 and September 2002. This work has not been submitted for any other degree at this or any other university.

Signed: ___________________________ Date: __________________

Stewart Brian Kirton
Department of Chemistry
University of Leicester
University Road
Leicester
United Kingdom
LE1 7RH
Abstract

Computational Studies of the Cytochromes P450

Stewart B. Kirton

Validation of standard comparative modelling techniques was carried out by comparison between several models of the mammalian P450 2C5 enzyme and experimentally determined X-ray crystal structure. Structural models of were then proposed for human cytochrome P450 2D6, a polymorphic enzyme known to be involved in the metabolism of a wide variety of pharmaceuticals. These models incorporated the only mammalian P450 crystal structure (pdb accession code 1dt6) and the impact of including this structure as a template in the modelling process was assessed using computational techniques such as active site characterisation using a number of chemical probes and principal components analysis to systematically analyse for differences between the comparative models. This analysis highlighted a single residue, Glu 216, as a residue that may be important for substrate binding, a hypothesis that was supported via docking studies of known substrates i.e. codeine and MPTP. Prediction of binding affinities was also investigated. A number of known P450 2D6 inhibitors were docked into the active site of the model and their binding affinities investigated using regression based scoring methods. The activities of 2D6 substrates were also investigated using a number of Quantitative Structure-Activity Relationship (QSAR) methods, the most successful being the grid-based Molecular Field Analysis (MFA) technique. Finally pharmacophores were used to predict the binding affinities of a number of P450 2D6 substrates. The results for these investigations compared favourably to literature precedents and served to reiterate the importance of a hydrophobic region approximately 5Å in distance from a positive ionisable group, such as a basic nitrogen atom, as an indicator that a substrate is likely to be metabolised by P450 2D6.
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statement</td>
<td>i</td>
</tr>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Contents</td>
<td>iii</td>
</tr>
<tr>
<td>Figure Legends</td>
<td>x</td>
</tr>
<tr>
<td>Table Legends</td>
<td>xx</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>xxiii</td>
</tr>
<tr>
<td><strong>1. Introduction</strong></td>
<td>1</td>
</tr>
<tr>
<td>1.1  Proteins</td>
<td>2</td>
</tr>
<tr>
<td>1.2  Protein Structure</td>
<td>2</td>
</tr>
<tr>
<td>1.3  Experimental Determination of Protein Structure</td>
<td>6</td>
</tr>
<tr>
<td>1.3.1 X-ray Crystallography</td>
<td>6</td>
</tr>
<tr>
<td>1.3.2 NMR Spectroscopy</td>
<td>8</td>
</tr>
<tr>
<td>1.3.3 Electron Microscopy</td>
<td>9</td>
</tr>
<tr>
<td>1.3.4 The Protein Data Bank (PDB)</td>
<td>11</td>
</tr>
<tr>
<td>1.4  Theoretical Determination of Protein Structure</td>
<td>11</td>
</tr>
<tr>
<td>1.4.1 Comparative Modelling</td>
<td>12</td>
</tr>
<tr>
<td>1.4.2 Fold Propensity</td>
<td>14</td>
</tr>
<tr>
<td>1.4.3 Ab Initio Modelling</td>
<td>15</td>
</tr>
<tr>
<td>1.5  Protein-Ligand Interactions</td>
<td>16</td>
</tr>
<tr>
<td>1.6  Computational Methods – Computer-Aided Drug Design</td>
<td>18</td>
</tr>
<tr>
<td>1.7  The Cytochromes P450</td>
<td>19</td>
</tr>
<tr>
<td><strong>2. Methods</strong></td>
<td>24</td>
</tr>
<tr>
<td>2.0  Preface</td>
<td>25</td>
</tr>
<tr>
<td>2.1  Comparative Modelling</td>
<td>25</td>
</tr>
<tr>
<td>2.1.1 Identification of Suitable 3D Templates</td>
<td>26</td>
</tr>
<tr>
<td>2.1.2 Alignment of Templates at Amino Acid Level</td>
<td>28</td>
</tr>
<tr>
<td>2.1.3 3D Structural Alignment of Templates</td>
<td>31</td>
</tr>
<tr>
<td>2.1.3.1 MALIGN 3D</td>
<td>31</td>
</tr>
<tr>
<td>2.1.3.2 FSSP</td>
<td>31</td>
</tr>
</tbody>
</table>
2.1.4 Alignment of Query Sequence to Profile of Aligned Template Sequences

2.1.5 Comparative Modelling

2.1.6 Model Evaluation

2.2 Docking

2.2.1 Introduction

2.2.2 GOLD

2.2.3 FlexX

2.2.4 Autodock

2.3 Characterisation of Active Sites

2.3.1 GRID

2.3.2 Principal Component Analysis (PCA)

2.3.3 MOLCAD Surfaces

2.4 Scoring Functions

2.4.1 Introduction

2.4.2 D-Score

2.4.3 Chemscore

2.4.4 Potential of Mean Force (PMF) Score

2.4.5 G-Score and GOLD Score

2.4.6 Ligscore 1 and Ligscore 2

2.4.7 Piecewise Linear Potential (PLP) 1 and 2

2.4.8 Ludi Score

2.4.9 F-Score

2.5 Quantitative Structure-Activity Relationship (QSAR)

2.5.1 General Introduction

2.5.2 Molecular Field Analysis

2.5.3 Receptor Surface Model

2.5.4 Cerius2 Descriptors

2.5.5 CoMFA

2.5.6 Hologram QSAR (HQSAR)

2.6 Pharmacophores

2.6.1 Introduction

2.6.2 Catalyst
2.7 Statistical Methods

2.7.1 Introduction

2.7.2 Ordinary Least Squares Regression (OLSR)

2.7.3 Partial Least Squares Fitting (PLS)

2.7.4 Genetic Function Approximation (GFA)

2.7.5 Genetic Partial Least Squares (G/PLS)

2.7.6 Validation Methods

2.7.7 Spearman's Ranking

2.8 Measurement of Biological Activity for Ligands of CYP 2D6

3 Comparative Modelling of Cytochrome P450 2C5

3.0 Preface

3.1 Introduction – Comparative Modelling of Cytochrome P450 2C5

3.2 Experimental

3.2.1 Identification of 3D Structural Templates

3.2.2 Alignment of Templates at Amino Acid Level

3.2.3 3D Structural Alignment of Templates

3.2.3.1 MALIGN3D

3.2.3.2 FSSP

3.2.4 Alignment of Query Amino Acid Sequence to Template Profile

3.2.5 Model Building – Comparative Modelling

3.2.6 Model Analysis

3.2.6.1 Stereochemical Analysis

3.2.6.2 Amino Acid Environment

3.2.6.3 Mainchain RMSD

3.2.6.4 Cα RMSD

3.3 Results

3.3.1 Identification of 3D Structural Templates

3.3.2 Alignment of Templates at Amino Acid and Structural Level

3.3.3 Model Analysis

3.4 Discussion

3.5 Conclusions
4 Comparative Modelling of Human Cytochrome P450 2D6
Incorporating the Mammalian P450 2C5 Structure

4.0 Preface 110
4.1 Introduction - Cytochrome P450 2D6 110
4.2 Experimental 111
4.2.1 Comparative Modelling 111
  4.2.1.1 4-Bacterial Template Approach 112
  4.2.1.2 Single Template Approach 113
  4.2.1.3 Multiple Template Approach Using FSSP 113
  4.2.1.4 Multiple Template Approach Using MALIGN3D 113
4.2.2 Model Validation 114
4.3 Results 114
4.3.1 Identification of Suitable Templates 114
4.3.2 Amino Acid Sequence Alignments 116
4.3.3 Model Analysis 120
4.4 Discussion 121
4.5 Conclusions 123

5 Impact on the Comparative Modelling of Cytochrome P450 2D6 of Incorporating the Mammalian 2C5 Crystal Structure

5.0 Preface 126
5.1 Introduction 126
5.2 Experimental 126
5.2.1 Principal Component Analysis (PCA) 126
5.2.2 Molecular Docking 127
5.3 Results 127
5.3.1 Principal Component Analysis 127
  5.3.1.1 Loadings Plots 128
  5.3.1.2 Scores Plots 128
  5.3.1.3 PCA for 2C5-Based Models 130
5.3.2 Molecular Docking of Codeine 132
5.3.3 Molecular Docking of 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) 133
5.3.4 Validation of Dockings Using GRID 135

5.4 Discussion 136
5.4.1 Principal Component Analysis 136
5.4.2 Molecular Docking 140
5.4.3 Validation of Docking Studies 142

5.5 Conclusions 147

6 Predicting Biological Activities of Cytochrome P450 2D6 Substrates Using Scoring Functions 148

6.0 Preface 149
6.1 Introduction – Scoring Functions 149
6.2 Experimental 150
6.3 Results 151
6.4 Introduction – Nitric Oxide Synthases 157
6.4.1 Experimental 158
6.4.2 Results 158
6.4.3 Investigation of Alternate Docking Algorithms 161
6.5 Discussion 164
6.5.1 D-Score 164
6.5.2 Chemscore 166
6.5.3 PMF Score 167
6.5.4 GOLD Score and G-Score 167
6.5.5 Ligscore 1 and Ligscore 2 168
6.5.6 Piecewise Linear Potential (PLP) 1 & 2 168
6.5.7 Ludi Score and F-Score 168
6.6 Conclusions 174

7 QSAR Studies on Substrates of CYP 2D6 177
7.0 Preface 178
8 Investigation into Pharmacophores of Cytochrome P450 2D6

8.0 Preface
8.1 Introduction
8.2 Experimental
8.3 Results
  8.3.1 Generation of Conformers for Catalyst
  8.3.2 Function Mapping
  8.3.3 Hypothesis 1: 2 Feature Pharmacophore
  8.3.4 Hypothesis 2: 5 Feature Pharmacophore
  8.3.5 Hypothesis 3: Protein-Based Pharmacophore
8.4 Discussion
8.5 Conclusions

9 Overall Conclusions
10 References

Appendices

Chapter 3 250
  Appendix 3.1 251
  Appendix 3.2 252
Chapter 4 253
  Appendix 4.1 254
  Appendix 4.2 255
Chapter 5 256
  Appendix 5.1 257
  Appendix 5.2 258
  Appendix 5.3 259
Chapter 6 261
  Appendix 6.1 Dataset 1 262
  Appendix 6.2 Dataset 2 267
Chapter 7 270
  Appendix 7.1 271
  Appendix 7.2 276
Publications 278
Figure Legends

Figure 1.2.1 The reaction between tyrosine and glycine to form a dipeptide, highlighting the newly formed peptide bond (grey circle).................................3

Figure 1.2.2 The various levels of protein structure for triose phosphate isomerase pdb accession code 1TIM (Banner et al 1976). A) Part of the amino acid sequence of 1TIM B) Backbone trace of an α–helix, C) A β–sheet, D) a TIM barrel formed by the association of seven β-strands (purple arrows), E) The tertiary structure of the 1TIM ‘A’ domain. F) The association of the 1TIM ‘A’ (yellow) and ‘B’ (cyan) domains as an illustration of quaternary structure.............................................4-5

Figure 1.3.1 Image produced using SPOCK (Christopher and Baldwin 1998) illustrating how the 3D structure of a protein can be determined by fitting to an electron density map (green). Image courtesy of CSB, Yale (www.csb.yale.edu/userguides/graphics/dens_stick.html)................................................7

Figure 1.4.1 Overview of theoretical techniques used to predict 3-D structure of proteins........................................................................................................16

Figure 1.6.1 Flowchart illustrating the stages leading up to the marketing of a novel and potent pharmaceutical. Blue blocks indicate stages where computational chemistry techniques can be used to assist in the process..............................................19

Figure 1.7.1 Generalised Catalytic Cycle of the Cytochromes P450 (Schlichting et al 2000)..................................................................................................................22

Figure 2.0.1 An overview of the techniques employed to complete the work carried out in this thesis.................................................................25

Figure 2.1.1 An overview of the Comparative Modelling Process.........................26
Figure 2.1.2.1 Illustration of the Gap-Open Penalties applied in the CLUSTALX algorithm. Blue indicates residues in the sequences aligned to identical residues prior to the insertion of gaps, orange indicates residues between the sequences that are aligned to identical residues as a result of incorporating a gap, and black indicates residues that are not aligned to identical residues across the sequences.

Figure 2.1.2.2. Illustration of the Gap-Extension Penalties applied in the CLUSTALX algorithm. Blue indicates residues in the sequences aligned to identical residues prior to the insertion of gaps, orange indicates residues between the sequences that are aligned to identical residues as a result of incorporating a gap, and black indicates residues that are not aligned to identical residues across the sequences.

Figure 2.1.6.1 Ramachandran plot of a comparative model of cytochrome P450 2D6. Each residue in the protein is represented by a square. The map is divided into four regions that describe the conformation of peptides in the protein. These are most favoured (red), additionally allowed (yellow), generously allowed (cream) and disallowed (white).

Figure 2.1.6.2 Verify 3D results for the P450 2C5 (1dt6) crystal structure. Poorly modelled residues have compatibility scores less than zero.

Figure 2.1.6.3 Errat plot for a comparative model of P450 2D6. The black columns identify problem areas; they occur at the N-terminus, B', F, G, and H helices and in some loops.

Figure 2.3.2.1 Projection of pattern vectors from 2 dimensions to 1 dimension. (i) shows original points (ii) shows projection onto the axes A and B.

Figure 2.4.3.1 The form of the contact term for metal ligand interaction in Chemscore.

Figure 2.4.7.1 The piecewise linear potential function used for the ligand-protein interaction energy. Values are given in Table 2.4.7.1.
Figure 2.5.6.1 Transformation of a chemical structure to its characteristic molecular hologram (So and Karplus 1999). In this example a glucose molecule is converted to a molecular hologram with a hologram length of 5. Subsequently each of the values inside the five boxes serves as an independent descriptor for the later PLS analysis.

Figure 2.7.2.1 A line fitted between five data points using the OLSR method. The minimised vertical distances between the calculated line and the observed values for the dependent variable (y) are illustrated.

Figure 2.8.1 P450 2D6 mediated demethylation of dextromethorphan to yield dextrorphan.

Figure 3.1.1 An overview of the process of comparative modelling, illustrating its iterative nature.

Figure 3.3.1.1 The Ca ribbon traces of the I-helices of P450 51 (red), P450 Terp (green) and P450 BM-3 (magenta) illustrating the significant structural differences in P450 51.

Figure 3.3.2.1 Method 1 Single template (P450 BM-3) alignment. P450 BM-3 residues in secondary structure elements in the crystal structure (DSSP; PDB code 1bu7), and P450 2C5 residues predicted (using PSIPRED) to be in secondary structure elements, have been shaded.

Figure 3.3.2.2 Method 2 FSSP-based multiple template alignment. Template residues in secondary structure elements in the crystal structure (DSSP; PDB code for BM3 is 1bu7, terp is 1cpt, NOR is 1rom, cam is 1phc and eryF is 1eup), and P450 2C5 residues predicted (using PSIPRED) to be in secondary structure elements, have been shaded.

Figure 3.3.2.3 Method 3 Malign3D multiple template alignment. Template residues in secondary structure elements, and P450 2C5 residues predicted (using PSIPRED) to be in secondary structure elements, have been shaded.
**Figure 3.3.2.4** Method 4 Manually modified Malign3D multiple template alignment. Template residues in secondary structure elements, and P450 2C5 residues predicted (using PSIPRED) to be in secondary structure elements, have been shaded. Modifications include the closing of insertions in secondary structure elements and a partial alignment of the B' helices................................................................. 95

**Figure 3.4.1 A)** Cα trace of the template BM-3 (black; PDB code 1bu7) and the ten models (grey) produced using a single template alignment (*Method 1*; Figure 3.3.2.1) and **B)** Cα trace of the most homologous template BM-3 (black; PDB code 1bu7) and the ten models produced using a multiple template alignment (*Method 2*; Figure 3.3.2.2)..............................................................................................................................100

**Figure 3.4.2** Errat plot for the lowest energy model produced using the Malign3D based alignment (Figure 3.3.2.3). The black columns identify problem areas; they occur at the N-terminus, B', F, G, and H helices and in some loops.................102

**Figure 3.4.3** Errat plot for the best energy model produced using the modified Malign3D alignment (Figure 3.3.2.4). The black columns identify problem areas; these occur in loop regions and β-strands at the C-terminus and N-terminus.......102

**Figure 3.4.4** Plot of the all Cα atom RMSD *versus* % sequence identity between the closest template to the target and the target. The results presented are for the CASP3 predictions (Jones and Kleywegt 1999) where 100% of the target residues had been modelled, together with the 2C5 models reported here..................................................105

**Figure 3.4.5** Structure-based alignment of the P450 2C5 best model against the crystal structure (PDB code 1dt6). Those regions correctly aligned are marked with a star, corresponding to 54% of the amino acids. The observed and predicted secondary structure elements are shaded grey and the substrate recognition sites (SRS) are labelled.................................................................106

**Figure 3.4.6** The P4502C5 substrate recognition sites for the crystal structure (dark grey; PDB code 1dt6) and the lowest energy model (light grey).................................107
Figure 4.3.2.1 Bacterial-template alignment (*Alignment 1*) used to generate comparative models of P450 2D6. The secondary structural elements α-helices (grey), and β-strands (black) as predicted by PSIPRED (Jones 1999) are highlighted.

Figure 4.3.2.2 Single-template alignment (*Alignment 2*) used to generate comparative models of P450 2D6. The secondary structural elements α-helices (grey), and β-strands (black) as predicted by PSIPRED (Jones 1999) are highlighted.

Figure 4.3.2.3 The FSSP determined amino acid sequence alignment (*Alignment 3*) used to generate comparative models of P450 2D6. The secondary structural elements α-helices (grey), and β-strands (black) as predicted by PSIPRED (Jones 1999) are highlighted.

Figure 4.3.2.4 MALIGN3D alignment (*Alignment 4*) used to generate comparative models of P450 2D6. The secondary structural elements α-helices (grey), and β-strands (black) as predicted by PSIPRED (Jones 1999) are highlighted.

Figure 5.3.1.1 A) GOLPE-generated loadings plot for the superimposed active sites of *Models 1-4* and the 2C5 crystal structure, analysed using GRID with 10 chemical probes. B) GOLPE-generated loadings plot corresponding to (A) but from which those GRID points for which a value of 5 kcal mol⁻¹ was returned in any structure for all ten probes investigated, have been deleted for all structures. C) GOLPE-generated concatenated scores plot to highlight the differences between the active sites between the four models and the 2C5 crystal structure. PC1 accounts for 36% of the variance, PC2 accounts for 15% of the variance.

Figure 5.3.1.3 A) GOLPE-generated loadings plot from the active sites of the crystal structure of 2C5 and those models (*Models 2, 3 and 4*) that incorporate information from 2C5. B) GOLPE-generated loadings plot corresponding to (A) but from which those GRID points for which a value of 5 kcal mol⁻¹ was returned in any structure for all ten probes investigated have been deleted for all structures. C) GOLPE-generated scores plot illustrating the differences in the active sites between *Models 2, 3 and 4* and the 2C5 crystal structure. PC1 accounts for 36% of the variance, PC2 accounts.
for 13 % of the variance D) GOLPE-generated scores plot to illustrate the differences between the most homologous bacterial template (BM3), the 2C5 crystal structure and Models 1-4. PC1 accounts for 38 % of the variance, PC2 accounts for 11 % of the variance..........................................................131

Figure 5.3.4.1 The active site of 2D6 (Model 3), illustrating the binding of codeine. There is excellent correlation between the position of the highest ranked GOLD docking of codeine and the contour map of the GRID trimethylammonium cation (NM3) probe contoured at –8 kcal mol⁻¹............................................. 135

Figure 5.4.1.1 MOLCAD surfaces (Heiden et al 1993) illustrating differences in shape and volume of the active sites of Models 1-4 and the 2C5 crystal structure: A) Model 1 generated using the 4 bacterial structural templates (Alignment 1); B) Model 2 generated using the 2C5 crystal structure (Alignment 2); C) Model 3 generated using the FSSP aligned four bacterial structural templates and the 2C5 crystal structure (Alignment 3); D) Model 4 generated using the MALIGN3D aligned four bacterial templates and 2C5 crystal structure; e) The active site of the 2C5 crystal structure. 137

Figure 5.4.1.2 The GOLPE generated GRID-like loadings contours for the trimethylammonium probe, comparing the active sites of Model 3 and the 2C5 crystal structure. This illustrates that the major difference between 2C5 and Model 3 (i.e. 2D6) results from the chemical difference between (acidic) Glu 297 in the I-helix of 2C5 (black) and the corresponding (hydrophobic) Val 308 in 2D6 (Model 3; grey) 140

Figure 5.4.3.1 The active site of 2D6, illustrating the effects on the binding of basic nitrogen-containing ligands of mutating Glu 216 and/or Asp 301: (A) Model 3, (B) the D301A mutant, (C) the E216A mutant, and (D) the E216A/D301A double mutant. In all cases, the highest ranked GOLD docking of codeine is overlaid on the contour map of the GRID trimethylammonium cation (NM3) probe contoured at –8 kcal mol⁻¹..144

Figure 5.4.3.2 Temperature factors of Asp 290 and the B-C region in the 2C5 crystal structure. These are plotted on a scale 0 (blue) to 100 (red), with 50 coloured white.
Red indicates a high temperature factor and high relative mobility. Blue indicates a low temperature factor, and low relative mobility.

**Figure 6.3.1** Linear regression plots for dataset 1 showing pIC50 vs. scoring function for A) Ludi ($r^2 = 0$), B) -PLP 2 ($r^2 = 0$), C) F-Score ($r^2 = 0.001$), D) -PLP 1 ($r^2 = 0.002$), E) G-Score ($r^2 = 0.003$), F) GOLD Score ($r^2 = 0.006$), G) Ligscore 1 ($r^2 = 0.027$), H) Chemscore ($r^2 = 0.030$), I) D-Score ($r^2 = 0.031$), J) PMF Score ($r^2 = 0.031$) and K) Ligscore 2 ($r^2 = 0.035$). 

**Figure 6.4.2.1** Linear regression plot showing pIC50 vs. F-Score predictions for Dataset 2 ($r^2 = 0.188$).

**Figure 6.4.2.2** Linear regression plot showing pIC50 vs. LUDI predictions for Dataset 2 ($r^2 = 0.071$).

**Figure 6.4.2.3** Cumulative frequency graph. This shows that a greater number of scoring functions perform to a higher level of statistical significance for Dataset 2 (triangles) as opposed to Dataset 1 (circles).

**Figure 6.4.2.4** Linear regression plot showing pIC50 vs. Ligscore1 predictions for Dataset 2 ($r^2 = 0$).

**Figure 6.4.3.1** The highest ranked docking of codeine into 2D6 active site using Autodock (Morris et al 1998) shows the docking is poor as there is no association of the methoxy group that undergoes demethylation with the Fe atom at the centre of the heme.

**Figure 6.5.3.1** The highest ranked GOLD dockings of A) corynanthine (orange) and carbamazepine (yellow) and B) Compounds 1-5 into the 2D6 model, highlighting the disparate binding orientations of the Dataset 1 compounds in comparison to those from Dataset 2.

**Figure 6.5.7.1** Linear regression plot of molecular weight versus LUDI score ($r^2 = 0.62$).
Figure 6.5.7.2. Linear regression plot of LUDI score vs. number of heavy atoms in a compound for Dataset 1 ($r^2 = 0.70$) ................................................................. 170

Figure 6.5.7.3 Linear regression plot of molecular weight vs. pIC50 for Dataset 1 ($r^2 = 0.04$) .............................................................................................................. 171

Figure 6.5.7.4 GOLD dockings for A) methyltetrahydroisoquinolinone (cyan) and B) lobeline (magenta) illustrating that larger molecules are subject to a greater possible number of steric clashes when docked into the 2D6 active site ........................................... 171

Figure 6.5.7.5 Regression of 42 compounds ($r^2 = 0.221$) from work published by Böhm (Böhm 1997) that shows over a narrower range of $K_i$ (comparable to Dataset 1) the values obtained for $r^2$ are significantly reduced ........................................ 173

Figure 7.2.1 The GOLD generated 'docked' alignments for compounds in A) Dataset 1 and B) Dataset 2 ............................................................................................................. 179

Figure 7.2.2 The 'undocked' Cerius2 generated alignments for compounds in A) Dataset 1 and B) Dataset 2 .......................................................... 179

Figure 7.3.4.1 Linear regression of CoMFA score vs. experimental biological activity for Dataset 1 compounds in their undocked conformations. $r^2 = 0.104$, $q^2 = 0.081$, Spearman Ranking coefficient $r_s = 0.36$, and probability that the correlation has occurred by random chance, $p = 0.039$ .............................................................. 190

Figure 7.3.4.2 Linear regression of CoMFA score vs. experimental biological activity for Dataset 1 compounds in their docked conformations $r^2 = 0.014$, $q^2 = -0.013$, Spearman Ranking coefficient $r_s = 0.22$, and probability that the correlation has occurred by random chance, $p = 0.447$ .............................................................. 190
Figure 7.3.4.3 Linear regression of CoMFA score vs. experimental biological activity for Dataset 2 compounds in their undocked conformations $r^2 = 0.026$, $q^2 = -0.038$, Spearman Ranking coefficient $r_s = 0.00613$, and probability that the correlation has occurred by random chance, $p = 0.534$ ..............................................................191

Figure 7.3.4.4 Linear regression of CoMFA score vs. experimental biological activity for Dataset 2 compounds in their undocked conformations $r^2 = 0.071$, $q^2 = 0.009$, Spearman Ranking coefficient $r_s = 0.25$, and probability that the correlation has occurred by random chance, $p = 0.298$ .................................191

Figure 7.3.5.1 Linear regression of HQSAR predicted pIC50 vs. experimental biological activity for Dataset 1. $r^2 = 0.068$, $q^2 = 0.044$, Spearman Ranking coefficient $r_s = 0.22$, and probability that the correlation has occurred by random chance, $p = 0.099$ ........................................................................................................192

Figure 7.3.5.2 Linear regression of HQSAR predicted pIC50 vs. experimental biological activity for Dataset 2 $r^2 = 0.729$, $q^2 = 0.711$, Spearman Ranking coefficient $r_s = 0.82$, and probability that the correlation has occurred by random chance, $p = 1.28 \times 10^{-5}$ ...............................................................................................................193

Figure 7.3.6.1 Linear regression of the mean predicted pIC50 determined from four individual QSAR experiments (MFA, RSM, Cerius$^2$ descriptor-based and HQSAR) vs. experimental biological activity for Dataset 1 compounds in their undocked conformations. $r^2 = 0.833$, $q^2 = 0.829$, Spearman Ranking coefficient $r_s = 0.88$, and probability that the correlation has occurred by random chance, $p = 8.94 \times 10^{-17}$ ...............................................................................................................................194

Figure 7.3.6.2 Linear regression of the mean predicted pIC50 determined from four individual QSAR experiments (MFA, RSM, Cerius$^2$ descriptor-based and HQSAR) vs. experimental biological activity for Dataset 1 compounds in their docked conformations. $r^2 = 0.874$, $q^2 = 0.871$, Spearman Ranking coefficient $r_s = 0.93$, and probability that the correlation has occurred by random chance, $p = 3.34 \times 10^{-19}$ ..................................................................................................................195
Figure 7.3.6.3 Linear regression of the mean predicted pIC50 determined from four individual QSAR experiments (MFA, RSM, Cerius² descriptor-based and HQSAR) vs. experimental biological activity for Dataset 2 compounds in their undocked conformations. $r^2 = 0.958$, $q^2 = 0.955$, Spearman Ranking coefficient $r_s = 0.95$, and probability that the correlation has occurred by random chance, $p = 9.32 \times 10^{-12}$

Figure 7.3.6.4 Linear regression of the mean predicted pIC50 determined from four individual QSAR experiments (MFA, RSM, Cerius² descriptor-based and HQSAR) vs. experimental biological activity for Dataset 2 compounds in their docked conformations. $r^2 = 0.957$, $q^2 = 0.917$, Spearman Ranking coefficient $r_s = 0.92$, and probability that the correlation has occurred by random chance, $p = 1.59 \times 10^{-9}$

Figure 7.4.1 Linear regression of the predicted pIC50 determined for the test set compounds in their docked conformations (Appendix 7.1) vs. experimental biological activity

Figure 8.3.5.1 The protein-based 2D6 pharmacophore. A) Interaction map illustrating hydrophobic regions (grey circles) and hydrogen-bond acceptors (blue and white cylinders) in the 2D6 active site. B) The resultant protein-based pharmacophore, comprising 3 pharmacophoric features, 2 hydrogen bond acceptors (blue and white spheres) and one hydrophobic region (white sphere)

Figure 8.4.1 Hypothesis 2A showing three hydrophobes (blue) and positive ionisable (red) features. Angle $BAC = 88.9^\circ$, $BAD = 130.2^\circ$, $CAD = 54.4^\circ$

Figure 8.4.2 Hypothesis 2C showing positive ionisable (red), two hydrophobes (blue) and a hydrogen bond acceptor (green) as features. Angle $BAC = 176.7^\circ$, $BAD = 75.8^\circ$ and $CAD = 61.5^\circ$

Figure 8.4.3 Hypothesis 2E showing positive ionisable (red), hydrophobic aliphatic (light blue), hydrophobic aromatic (dark blue) and ring aromatic (gold) features. Angle $BAC = 119.2^\circ$, $BAD = 183.3^\circ$ and $CAD = 119.0^\circ$
Table Legends

Table 2.4.7.1: Parameters of the atomic pairwise ligand-protein potentials for PLP.

Table 3.3.1.1 The P450 structures recovered by a PSI-BLAST search using the sequence of P450 2C5 (1dt6).

Table 3.3.3.1 Results of validation studies performed on the lowest energy models produced from the four different sequence alignments.

Table 3.3.3.2 Validation results for the target 2C5 structure and the five crystal structures used as templates for the modelling studies.

Table 3.4.1 Comparison of lowest energy models produced from the four different sequence alignments with 2C5 crystal structure.

Table 4.3.1.1 Sequences of P450 structures identified by a PSI-BLAST search using the sequence of P450 2D6.

Table 4.3.3.1 Results of validation studies performed on the lowest energy models produced from the three different sequence alignments.

Table 4.3.3.2 Results of validation studies performed on the 3D structural templates used to generate comparative models of 2D6.

Table 5.3.2.1 Interatomic Fe-H distances for codeine docked into the active site using GOLD compared with experimentally determined NMR distances.

Table 5.3.3.1 Interatomic Fe—H distances for MPTP docked into the active site using GOLD compared with experimentally determined NMR distances.

Table 6.3.1 Linear regression analysis of scoring functions for Dataset 1.
Table 6.4.2.1 Linear regression analysis of scoring functions for Dataset 2........158

Table 6.4.3.1 Comparison of interatomic Fe-H distances for codeine docked into CYP 2D6 using Autodock with experimentally determined interatomic Fe-H distances.. 163

Table 7.2.3.1 Cerius$^2$ descriptors used to generate the 2D6 QSAR for Datasets 1 and 2...................................................................................................................... 182

Table 7.3.6.1 Summary of $r^2$ and $q^2$ values from the QSAR investigations........197

Table 8.3.1.1 The number Catalyst generated conformers for each of the ligands in the training set..................................................................................................................208

Table 8.3.1.2 The number of Catalyst generated conformers for each of the ligands in the test set.................................................................................................................209

Table 8.3.3.1 Results for the 2-feature pharmacophores generated using the Catalyst produced conformers for the ligands in the test set................................................210

Table 8.3.3.2 Linear regression plot statistics of experimental biological activity versus predicted biological activity for hypothesis 1A, 1B and 1C..................210

Table 8.3.3.3 Results for the 2-feature pharmacophores using the GOLD generated conformers for the ligands in the test set.................................................................211

Table 8.3.3.4 Linear regression plot statistics of experimental biological activity versus predicted biological activity for hypothesis 1D, 1E and 1F.........................211

Table 8.3.4.1 Results for the 5-feature pharmacophores generated using the Catalyst produced conformers for the ligands in the test set.................................207

Table 8.3.4.2 Linear regression plot statistics of experimental biological activity versus predicted biological activity for hypothesis 2A, 2B and 2C.................213
Table 8.3.4.3 Results for the 5-feature pharmacophores generated using the GOLD produced conformers for the ligands in the test set.......................................................213

Table 8.3.4.4 Linear regression plot statistics of experimental biological activity versus predicted biological activity for hypothesis 2D, 2E and 2F...............................214

Table 8.3.5.1 Linear regression plot statistics of experimental biological activity versus predicted biological activity for the protein-based hypotheses validated using the Catalyst generated conformers for the test set.................................216

Table 8.3.5.2 Linear regression plot statistics of experimental biological activity versus predicted biological activity for the protein-based hypotheses validated using the GOLD generated conformers for the test set.................................217

Table 8.4.1 Experimentally observed and predicted pIC50 values for the 2D6 test set. Highlighted regions represent 'invalid' predictions i.e. those values > 1 residual log unit..............................................................................................................219

Table 8.4.2 Interatomic Fe—H distances for MPTP docked into the active site using GOLD compared with experimentally determined NMR distances...............224
Acknowledgements

There are several people I would like to thank for their help and input over the last three years. Firstly my sincerest thanks to my supervisor Mike Sutcliffe, for his unwavering support, patience, insight and guidance. Thanks are also especially due to Nick Tomkinson, for his role as industrial supervisor and his quiet confidence in my ability to overcome the many challenges I faced.

I would also like to thank Carol Kemp, for her support, comments and friendship during my studies, and Steve St-Gallay for his enthusiasm and gumption. These two acted as mentors of their own volition, and for their interest in my work and unconditional encouragement they have my deepest gratitude.

Past Sutcliffe group members, Fabio Zuccotto, Raj Sangar and Kam Chohan, the modelling group at AstraZeneca, Dave Marriott, Alan Tinker and Sarah Maw and the numerous members of the Biological NMR centre at the University of Leicester also deserve a mention for their friendship and advice.

Thanks are also due to the EPSRC for providing the funding which enabled me to carry out this work.

Finally, on a personal note, I would like to thank Helen Woodland, one of the most inspirational people it has been my good fortune to meet.
Chapter 1

Introduction
1.1 Proteins

The word protein is derived from the Greek word *proteios* meaning ‘of the first rank’, an indication of the importance of these macromolecules which are involved in almost all known biochemical processes in animals, plants, fungi and bacteria. A protein is defined as a large molecule consisting of fifty or more amino acid sub-units, and their roles are diverse, serving as both structural materials (e.g. hair, horns, fingernails) and as enzymes that control an organism’s biochemistry. They are also involved in the transport and storage of substances around the body, are integral in distinguishing between self and non-self in the immune response, and act as chemical messengers to regulate cell growth and reproduction. As such, the three-dimensional structure that a protein adopts is inextricably linked to its function. It is therefore of significant scientific interest to be able to elucidate the structure of proteins as this can give insight into phenomena such as drug-protein interactions and reaction mechanism.

1.2 Protein Structure

Proteins are macromolecules. In the simplest terms they can be viewed as polypeptide chains that arise from a condensation reaction between two amino acids resulting in the formation of a peptide bond (e.g. Figure 1.2.1). There are several increasingly complex levels of protein structure, which it is necessary to appreciate before insight into their function can be gained (Figure 1.2.2). These are as follows:

1) Primary structure – the amino acid sequence, a string of letters, each one representing one of the twenty naturally occurring amino acids, and the order in which they occur in the protein. Ultimately it is this sequence that will dictate the overall shape, and thereby function, of the protein.

2) Secondary Structure – the most fundamental three-dimensional arrangements of the amino acids in the polypeptide chain. There are two common secondary structural elements – the α-helix and the β-sheet.
3) Super Secondary Structure – an association of several secondary structural elements giving rise to a commonly observed structural feature within proteins e.g. the TIM barrel (Banner et al 1976).

4) Tertiary Structure – the overall three-dimensional shape of a monomeric protein

5) Quaternary structure – the association of two or more tertiary structures (sometimes referred to as domains), which may be identical or different, to give a larger more complicated protein structure. This type of protein structure is referred to as multimeric.

The nature of a protein also dictates that there will be significant interactions between the amino acid sub units that comprise it. For example, it is expected that basic amino acids would form ionic interactions with acidic amino acids in their immediate vicinity. The overall three-dimensional structure is stabilised by ionic interactions (including hydrogen bonds) and van der Waals forces, giving rise to hydrophobic and hydrophilic interactions between both the protein and the protein and its environment.

![Figure 1.2.1](image)

**Figure 1.2.1** The reaction between tyrosine and glycine to form a dipeptide, highlighting the newly formed peptide bond (grey circle).
Figure 1.2.2 The various levels of protein structure for triose phosphate isomerase pdb accession code 1TIM (Banner et al 1976). A) Part of the amino acid sequence of 1TIM B) Backbone trace of an α-helix, C) A β-sheet, D) a TIM barrel formed by the association of seven β-strands (purple arrows), E) The tertiary structure of the 1TIM ‘A’ domain. F) The association of the 1TIM ‘A’ (yellow) and ‘B’ (cyan) domains as an illustration of quaternary structure.
1.3 Experimental Determination of Protein Structure

There are three established experimental methods for determining the three dimensional structure of proteins: X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy and determination of protein topology using Electron Microscopy (EM).

1.3.1 X-ray Crystallography

In order to obtain a 3-dimensional structure via X-ray crystallography one must first obtain a pure crystalline sample of the protein under investigation. Consequently, if the protein cannot be crystallised then it is impossible to obtain a structure using X-Ray crystallography. This is the case for the vast majority of mammalian enzymes, as they tend to be membrane bound structures (Rhodes 1993).

If a crystal can be obtained it is subjected to a highly collimated beam of X-rays ($\lambda \approx 10^{-10}$ nm) These X-rays interact with the electrons in the inner orbitals of the atoms within the molecule, which scatter/diffract onto a photographic film as determined by the Bragg Equation (Equation 1.3.1.1)

$$n\lambda = 2dsin\theta \quad \text{(Equation 1.3.1.1)}$$

Where $\lambda$ is the wavelength of the incident electron beam, $d$ is the interatomic distance between layers in the crystal and $\theta$ is the angle of incidence.

Once the experiment is complete the photographic film is covered in a pattern of spots of various intensities unique to the crystal structure of the protein under investigation. However, for the image of the protein to be reconstructed the phase, as well as the intensity, of every individual reflection must be established. This is achieved by multiple isomorphous replacement (MIR), where heavy metal atoms, such as mercury and cadmium, replace atoms in the protein by soaking the protein crystal in a saturated solution of the chosen heavy metal. A second X-ray diffraction pattern is obtained. Those residues that have incorporated the heavy metal atoms will
diffraction differently to the original diffraction pattern (the extent of scattering is a function of atomic number), and by comparing and contrasting the original and MIR diffraction patterns it is possible to ascertain the phase of some of the atoms in the protein. Having achieved this it is then possible to assign the phases of the majority of the other atoms within the protein, and ultimately produce an initial electron density map of the 3D structure to which the amino acid sequence can be fitted (Figure 1.3.1.1). The structure is refined in an attempt to obtain the best model, i.e. the one that fits most closely with the experimental data. The most common way of achieving this is to employ restrained molecular dynamics refinement (Brunger et al 1987) in which a penalty term is introduced into the normal potential energy function, penalising those structures that do not fit with experimental data.

Figure 1.3.1 Image produced using SPOCK (Christopher and Baldwin 1998) illustrating how the 3D structure of a protein can be determined by fitting to an electron density map (green). Image courtesy of CSB, Yale (www.csb.yale.edu/userguides/graphics/dens_stick.html)
X-ray crystallography is often the most accurate way of determining the 3D structure of a protein and can provide a great deal of information with regards to structure and function it is not without its limitations. However, several factors must be taken into consideration when determining whether or not a protein is suitable for this type of analysis. The most fundamental of these factors is that the protein must form a regular crystal. Without a crystal the structure cannot be determined. Secondly, the scattering of electrons is proportional to the amount of electron density that surrounds them. Hence, hydrogen atoms do not diffract significantly and signals arising from these atoms are inevitably swamped by those of neighbouring heavier atoms. This means that unless the resolution of the diffraction pattern is very high signals due to hydrogen are never seen. Therefore, the position of hydrogen must be inferred by examination of nitrogen, oxygen, sulphur and carbon atoms in the structure and as such it is impossible to detect using X-ray crystallography whether or not an ionisable group in the protein is in the ionised form. Another limitation arises because the structure of the protein is determined under non-physiological conditions. This means that the way in which residues appear to interact could simply be artefacts of crystal packing (Jacobson et al 2002). Finally the resultant image is always of a static protein, belying the fact that the real system is dynamic.

1.3.2 NMR Spectroscopy

The structure of a protein or protein-ligand complex can also be solved using Nuclear Magnetic Resonance (NMR) spectroscopy. The major advantage NMR spectroscopic techniques have over X-ray crystallography is that they represent the structure of the protein in solution i.e. similar to physiological conditions. In addition it can give direct information on protein-ligand interactions, and if the protein has been radioactively labelled information with respect to protein dynamics.

NMR spectroscopy works because nuclei with a non-zero spin e.g. $^1$H, $^{13}$C and $^{15}$N, will resonate when placed in an external magnetic field. Each chemically distinct nucleus will give rise to a signal with a specific chemical shift, and it is the value of these chemical shifts that provide the information on the chemical environment of each nucleus. In contrast to X-ray crystallography this technique is best applied to
relatively small proteins (<25 kDa in size, although it is possible to study proteins at least twice this size). Larger proteins make the NMR spectrum obtained incredibly complex, and each resonance peak becomes broader because of the slower rate at which the protein tumbles in solution. As a consequence there is a greater overlap between signals, and sequential assignment becomes very difficult as molecular weight increases. For smaller proteins the assignment problem may be overcome by the use of 2D (Wemmer and Reid 1985) 3D (Bax and Grzeiszek 1993) and 4D (Clore and Gronenborn 1991) NMR techniques, as well as $^2$H, $^{15}$N and $^{13}$C labelling (Bernado et al 2002).

2D NMR experiments such as Correlation Spectroscopy (COSY) and Nuclear Overhauser Effect Spectroscopy (NOESY) yield spectra containing cross peaks. These give insight into interactions between nuclei in the protein, COSY giving through bond interactions and NOESY through space interactions. These techniques are of paramount importance as through bond and through space data can be converted into dihedral angle and distance restraints respectively. Computational methods such as distance geometry followed by simulated annealing (Havel and Wuthrich 1984) use these experimentally determined restraints to produce an ensemble of 3D structures.

1.3.3 Electron Microscopy

If it is impossible to produce a crystal large enough to be examined via X-ray crystallographic techniques, or the size of the protein is too large for NMR structure determination, an alternative way of elucidating the 3D structure of a protein is by Electron Microscopy (EM). A distinct advantage of EM over X-Ray crystallography is that direct images are obtained – so there is no need to assign phase and perform Fourier Transform calculations. Also, EM requires only a very small amount of microcrystals to obtain useful information and as the Coloumbic scattering of electrons by matter is approximately one million times stronger than the scattering of X-rays (Blundell and Johnson 1976) it takes seconds as opposed to hours to achieve appreciable electron scattering on even small samples.
The directions in which the electrons are scattered onto either a photographic film or a fluorescent tube are important in determining the structure. 90% of the incident high-energy electrons pass straight through the sample, and are not scattered. Of the remaining ten percent the majority undergo elastic scattering – whereby the high-energy electrons interact with a heavy atom and are diffracted at the same wavelength at which it was incident. There is no loss of energy from the incident electron with this type of interaction – and the scattering gives useful information for determining the 3D structure of the protein. The remainder of the interactions are classed as inelastic scattering. In this case, the incident high-energy electron interacts with a slow-moving atomic electron. Upon collision there is a transfer of energy from the high-energy particle to the low energy particle. This results in the incident electron being diffracted at a different wavelength, which leads to a blurring of the diffraction pattern – a phenomenon known as chromatic aberration.

There are disadvantages associated with this technique – the most apparent being the limited resolution associated with images determined in this manner. Despite the fact that the wavelength of an electron is much shorter than that of an X-ray, and it would therefore be reasonable to expect higher resolution images from EM in comparison to X-Ray crystallography this is not the case. The major reason for this is that it has not been possible to design lenses capable of focussing widely scattered electrons. The extensive specimen preparation, which must be undertaken to gain an electron microscopic image, is also detrimental to the resolution achievable. As such the nominal resolving power of an electron microscope is approximately 2Å. In reality, and especially for biological samples, the resolution achievable is considerably worse than this, generally between 10-15Å.

This reduction in quality of resolution can be attributed to the rigorous pre-treatment a biological sample must undergo before an image can be obtained. Because air is capable of scattering electrons EM experiments have to be carried out under high vacuum conditions. Therefore any samples have to be preserved to be able to withstand the desiccation that would occur under these conditions. Generally, the samples are chemically fixed to render them insoluble, dehydrated, embedded in an epoxy resin so they can be cut to a thickness which will facilitate electron scattering and stained by soaking them in solutions of heavy metals such as lead citrate, which
increases the contrast between the specimen and the background in the final image. At any stage of pre-treatment the architecture of the true crystal structure may be altered and as such it is unusual to achieve images much better than 15Å resolution. However, it is not impossible – and by meticulous analysis of images and correction of distortions it is possible to achieve a much lower resolution image – the most successful application of this technique in recent years being the resolution of the large trans-membrane nicotinic acetylcholine receptor protein to 4.6Å (Miyazawa et al 1999).

1.3.5 The Protein Databank (PDB)

Once the three-dimensional structure of a protein has been solved it is usually deposited in the protein databank (Berman et al 2000, http://www.rcsb.org/pdb). Hence, the PDB consists of the atomic coordinates for those proteins with structures that have been solved. It is possible to use the PDB website to search for any structure of interest and then download the coordinates of the three-dimensional structure of any protein contained in the database.

1.4 Theoretical Determination of Protein Structure

The rate at which new protein sequences are determined far exceeds the rate at which protein structures are determined experimentally. If it is impossible to ascertain the three-dimensional structure of a protein using experimental methods there are several theoretical alternatives that are available to predict the three-dimensional structure of proteins from the amino acid sequence. An overview of the theoretical techniques available for use is given in Figure 1.4.1 These approaches to molecular modelling can give remarkable insight. One of the earliest examples that demonstrates the predictive power of molecular modelling is the work carried out by Pauling (Pauling et al 1951) which predicted that the \( \alpha \)-helix would be a stable element of polypeptide structure long before the first protein structure was solved. Although this prediction was based on mechanical models built from the analysis of the geometry of the peptide unit in crystal structures, it is this premise of utilising existing information to make predictions of the unknown that underpins the modern computational
approaches. No definitive answer has yet been found to the question of how to predict protein folding, although a variety of promising approaches have been developed. The following sections consider methods that attempt to predict protein structure from first principles, techniques that use a stepwise approach in which secondary structural elements are identified and then packed together, and finally the prediction of protein structure from by comparative modelling, whereby the structure of an unknown protein is based upon the structures of related homologous proteins.

1.4.1.1 Comparative Modelling

Comparative modelling is an empirical approach to the determination of the conformations of protein structures by predicting them from the experimentally determined structures of other proteins with homologous amino acid sequences. The premise underpinning this technique is that a small change in amino acid sequence will result in a small change in three-dimensional structure (Lesk and Chothia 1986, Hubbard and Blundell 1987). To build comparative models it is necessary to identify which proteins are to be used as templates. If the biological function of the target protein is known it is often relatively straight-forward to decide which proteins could be considered as templates from which to build the model. In cases where the function of the protein is unknown it may be possible to deduce the function by scanning a sequence database for the presence of particular combinations of amino acids (called motifs) which often imply a particular function or structural feature, and this could help direct the choice of template. In other cases the template that is chosen is simply that with the closest sequence identity to the target protein, irrespective of function – all methods being equally valid.

Having identified possible templates it is necessary to generate an amino acid sequence alignment. The objective of an amino acid sequence alignment is to position the amino acid sequences so that stretches of amino acids are matched with the expectation that these correspond to common structural features (such as the secondary structure and catalytic residues). Gaps in the aligned sequences correspond to regions where polypeptide loops are deleted or inserted. If the amino acid sequences of other proteins of the same family as the target are known it is preferable to produce a multiple, as opposed to pairwise, sequence alignment. A multiple
sequence alignment is often much more reliable than a pairwise alignment between the target protein and the template as it is easier to detect clear trends; with just two sequences it would be easy to be misled by some chance correspondences. Throughout the studies of the thesis these alignments were obtained using the CLUSTALW algorithm (Thompson 1994 see section 2.1.2) via the CLUSTALX graphical user interface as it was freely available. Other algorithms such as the Needleman-Wunsch algorithm (Needleman and Wunsch 1970) will also generate amino acid sequence alignments, and would be equally valid to use in comparative modelling studies. However, these were not freely available. Unless the task at hand is exceedingly simple manual amino acid sequence alignment is not usually feasible. However, automated sequence alignments also need to be scrutinised as the results can often be improved by manual intervention. It is critical that the amino acid sequence alignment is as accurate as possible – as any inaccuracies that occur at this stage will be carried through to the resultant comparative models.

Construction of comparative models is usually achieved in one of two manners. The first is via a multi-step fragment based approach, as utilised in COMPARER (Sutcliffe et al 1987 a,b) and SWISSMODEL (Guex and Peitsch 1997) whereby the rigid secondary structural elements are built using information from the template structures, before these are joined together by modelling the loop regions between them. The second method involves a single-step approach whereby the model is built via the satisfaction of spatial and homology derived restraints. This is the approach used by the program Modeller (Sali and Blundell 1993), and was the method of choice for the studies of this thesis as the software allowed for easy parameterisation of the heme cofactor present in the P450s – unlike SWISSMODEL and COMPARER, and was also the only package available that allowed the incorporation of experimentally derived distance restraints. However, both approaches are known to produce valid models – as indicated by their relative successes in ongoing 'competitions' such as the Critical Assessment of techniques for protein Structure Prediction (CASP see e.g. Mosiman et al 1995). Indeed, the CASP experiments have served to highlight that the most important aspect of generating an accurate comparative model is the need for an accurate sequence alignment, as incorrect alignments almost always lead to incorrect structures.
Whilst comparative modelling techniques are generally not as accurate in predicting protein conformation as experimental techniques such as NMR spectroscopy and X-ray crystallography they are able to produce models with rms deviations from the crystal structure as low as 1Å, provided that they have sufficiently homologous templates of known 3D structure upon which to base the models (Topham et al 1991). In the absence of a crystal structure comparative models provide valuable testable hypotheses.

1.4.1.2 Fold Propensity

If the database search does not produce a significant template for comparative modelling the folds the query sequence are most likely to adopt in its 3D state are examined. This method is less reliable than comparative modelling from a template with a similar amino acid sequence, but is still a valid and useful tool because it is able to exploit the fact that some proteins have been shown to adopt similar folds – even though there is no significant sequence similarity (Chothia 1992).

One program used to predict the folds is TOPITS (Threading One-Dimensional Predictions in Three-Dimensional Space (Rost 199 5a,b). This technique works by fitting the query sequence on to all known folds in turn, and returning a score for each fit based on the preference for the different amino acids to reside in various environments. Logically, those folds that place an appropriate residue in an appropriate environment, e.g. a hydrophobic residue in a hydrophobic pocket, will score highly. The total score is summed across all of the amino acids in the sequence ranked, and the statistical significance of the search is given in terms of an E-Value. The highest scoring protein is the most appropriate template to use for homology modelling of the query.

If a specific fold is identified, a further search can be carried out to determine whether or not the fold belongs to an existing fold family, by scanning a database called SCOP (Murzin et al 1995), which classifies proteins into Class, Fold, Superfamily, Family and Domain. If the other proteins that are found to share the same fold are known to have similar functions to the query protein this may be indicative that the fold being used is correct.
1.4.1.3 Ab Initio Modelling

If the search yields no 3D templates then *ab initio* calculations can be carried out to elucidate the 3D structure of a protein. This is by far the most ambitious of the approaches to the protein-folding problem. *Ab initio* calculations set out to explore the conformational space of the molecule in order to identify the most appropriate structure. The total number of possible combinations is invariably very large, and as such it is usual to attempt to find only the very lowest energy structures - based on the assumption that the global minimum in the energy function corresponds to the naturally occurring structure of the molecule. Identification of such low energy conformations is usually achieved by utilising some form of empirical forcefield often augmented with a term to account for solvation (see e.g. Gibson and Scherega 1987, Li and Scherega 1988). *Ab initio* calculations can also be used to predict secondary structural elements that will form based on the query sequence, by using a program such as JPred (Cuff et al 1998). Once the secondary structural elements have been predicted how they will interact with one another is examined.

JPred has been shown to be 73% accurate in the prediction of residues in strands, helices and loops, and it is possible that these predicted secondary structures are indicative of what the 3D structure of the query protein might be. Despite the highly speculative nature of *ab initio* calculations the use of this technique is becoming more widespread, and its popularity is on the increase due to the fact that reasonably accurate models of fragments containing up to sixty residues have been elucidated using *ab initio* methods (Moult et al 1999).
Figure 1.4.1 Overview of theoretical techniques used to predict 3-D structure of proteins.

1.5 Protein - Ligand Interactions

The studies in this thesis examine the interactions between biological receptor molecules, and small organic molecules, known as ligands. There are several non-bonded receptor-ligand interaction types that are important in ligand binding. These can be assigned to one of three major categories:

1) Ionic or electrostatic interactions, of which there are two important types

   a) ion-ion interactions e.g. \( \text{NH}_3^+ ...... \text{O}_2\text{C} \)

   b) ion-dipole interactions e.g. \( \text{NH}_3^+ ...... \delta^- \text{O} = \delta^+ \)

2) The hydrophobic effect. Lipophilic groups of a ligand usually reside in lipophilic pockets formed by hydrophobic amino acid side-chains in the active/binding site of a protein. The degree of stability of these interactions is associated with the degree of order in the molecules of water surrounding the
protein-ligand complex. The hydrophobic effect is the most important factor that contributes to protein stability.

3) Hydrogen bonding interactions. These are short-range directional dipole-dipole interactions that contribute to specific interactions between the ligand and the receptor.

\[
\delta^+ - N - H^+ \ldots \ldots \cdots \delta^- - O - C^+ 
\]

The role of water is also important in protein-ligand binding – although it is incredibly difficult to model using computational techniques. Prior to forming the ligand-receptor complex both the protein and the ligand will have a solvation shell i.e. they will both be surrounded by an extensive network of hydrogen-bonded water molecules. Upon complexation part of these solvation shells are broken as the ligand competes against water for position in the active site of the receptor. Water can also bind to the protein with different relative strength. Hence, if the ligand is attempting to replace only tightly bound water molecules, without making any other significant interactions (dipole-dipole, ion-ion, ion-dipole or lipophilic) with the receptor it is likely to show a low binding affinity. Conversely a loosely bound water molecule is likely to be replaced by an incoming ligand due to entropic (the release of water to bulk solvent) and enthalpic (favourable hydrogen-bonding interactions with other water molecules) factors that favour replacement by the ligand.

Therefore, a strong protein-ligand interaction can generally be expected if:

1) There is significant lipophilic interaction. Enlargement of the lipophilic contact surface area of a receptor by placement of additional lipophilic groups of a ligand will increase binding affinity.

2) The hydrophobic effect is strong. The molecular basis for this effect is entropic, moreover it is the entropy change with respect to the solvent (water) that is important. If the loss of entropy of the ligand upon binding to the protein is smaller in magnitude than the gain in entropy due to the water
molecules being displaced into bulk solvent as the ligand binds, it is likely that the hydrophobic effect will be strong. Further stabilisation due to enthalpy changes, such as displaced water molecules being able to form favourable hydrogen bonds with other water molecules in bulk solvent will also play a role in the magnitude of the hydrophobic effect.

3) The ligand is rigid. Rigid ligands will bind more tightly than flexible ones because of the reduced loss in entropy upon complex formation – the freezing of the degrees of freedom is reduced for rigid ligands.

1.6 Computational Methods – Computer-Aided Drug Design

The process of drug-discovery in the pharmaceutical industry is a lengthy one (Figure 1.6.1), often taking as long as ten years from identification of a candidate to the marketing of the finished drug. Computational chemistry is an invaluable tool in speeding up the discovery process, and the studies contained within this thesis are aimed at reducing the time, and ultimately cost, for identifying and optimising potential lead candidates in the pharmaceutical industry. The methods employed to achieve this are detailed in Chapter 2.

In the absence of an experimental protein structure (be it from X-ray crystallography, NMR spectroscopy or electron microscopy) comparative modelling (Chapters 3 and 4) gives valuable insight into the three-dimensional structure of a protein. Docking algorithms (Chapter 5) are another useful tool, as they elucidate possible binding modes of potential drug molecules, as well as giving information on a microscopic level of the residues in the protein active site that are important for the binding of small molecules, and those that may be crucial to the catalytic activity of enzymes. Characterisation of the active site, and systematic statistical analysis yields a large amount of data that would be almost impossible to interpret without the aid of computers. This data can then be used to direct investigations such as mutation studies in a traditional 'wet' laboratory (Chapter 5).

By using techniques such as scoring functions (Chapter 6), Quantitative Structure-Activity Relationships (QSARs - Chapter 7) and pharmacophore modelling (Chapter
it is also possible to gain insight into the important chemical features of a drug candidate, assess whether or not a compound will show a therapeutic effect and predict the efficacy of a novel compound based on a knowledge of the chemical structure and biological activities of other substrates - eliminating the need to synthesise and assay a large number of organic molecules, which subsequently reduces costs.

**Figure 1.6.1** Flowchart illustrating the stages leading up to the marketing of a novel and potent pharmaceutical. Blue blocks indicate stages where computational chemistry techniques can be used to assist in the process.

1.7 The Cytochromes P450

The studies presented in this thesis concentrate on one superfamily of enzymes, the cytochromes P450 (P450s).
The cytochrome P450 superfamily comprises several families that occur in microorganisms and plants, and several families that occur in the animal kingdom. These families can be sub-divided into two classes – based on the nature of their redox partner (Ravichandran et al. 1993). Class I P450s are usually found in the mitochondrial membranes of eukaryotes and the majority of bacteria. These P450s require a flavin adenine dinucleotide (FAD) containing reductase and an iron-sulphur protein (e.g. a ferredoxin) before they can become catalytically active. In contrast, Class II P450s are found mainly in animals where they are bound directly to the endoplasmic reticulum of cells, and are dependent upon the interaction with a reductase that contains both FAD and flavin mononucleotide (FMN) to make them catalytically active.

After categorising a P450 into its correct class, it is possible to make further distinctions between the enzymes by sub-dividing them into families based on their amino acid sequence identities. The greater the sequence similarity between two proteins, the more likely it becomes that their three-dimensional structures are similar, a premise which underpins the homology modelling sections contained within this thesis. This often, although not exclusively, results in members of the same family having similar functions, for example, family 1 (CYP1) appears to have an, as yet not completely understood, role in reproduction and development (Lewis 1996). Family 2 (CYP2) is concerned with the detoxification in animals of a wide range of phytoalexins, pharmaceuticals and many other exogenous and endogenous substances (Lewis et al. 1998), whereas family 3 (CYP3) shows a higher degree of specificity as it exclusively metabolises the larger phytoalexins such as alkaloids. The remaining families also exhibit specificities for different substrates, but irrespective of substrate the function of a P450 is almost universal across all the families – the activation of molecular dioxygen (O₂) and subsequent insertion of a single atom of molecular oxygen into an organic molecule. In order for two P450s to be considered members of the same family they must share approximately 40% sequence identity.

Division of the families into sub-families e.g. 2A, 2B, 2C, 2D etc. can be used to further classify the P450s. In order for a pair of P450s to belong to the same subfamily they must share approximately 70% amino acid sequence identity. Differences arising between subfamilies are, therefore, often subtle and can account
for differences in substrate specificities and activities for different drug-like compounds. By examining the differences between enzymes contained in the same sub-family, and indeed enzymes across different sub-families, it is possible to identify regions of change in the amino acid sequences that may or may not influence enzyme activity and/or substrate specificity. Again, in terms of computer-aided molecular design, this can give insight into which residues could be important when binding a substrate in the active site, and as such can play a vital role in the rationalisation of experimentally observed metabolites for a particular substrate when used in conjunction with docking and comparative modelling studies.

In terms of comparative modelling it would be preferable to produce models based on the three-dimensional structure of a template from the same sub-family (as this has the highest amino-acid sequence identity to the target, ~70%), failing this the next best template would be one belonging to the same family (approximately 40% identity) and lastly one belonging to the same superfamily (<40% identity) could be considered, although there is speculation as to the validity of comparative models produced using templates with less than 25% sequence identity.

The P450-mediated catalytic insertion of oxygen into organic substrates is shown in Figure 1.7.1.
Figure 1.7.1 Generalised Catalytic Cycle of the Cytochromes P450 (Schlichting et al 2000).
The P450 states 1-5 have been isolated and extensively studied using a variety of absorbance, resonance and magnetic spectroscopies (Alberta and Dawson 1987, Dawson 1988). X-Ray crystal structures of the states 1, 2 and 5 have further established the structures of these P450 species (Poulos et al 1987), and there is evidence to support that the proposed intermediates 6 and 7 are correct (Schlichting et al 2000).

The resting form of the enzyme (1) is a six-coordinate low-spin ferric state with water (or hydroxide depending upon pH) as the exchangeable distal ligand trans to the proximal cysteinate. Substrate addition generates the high-spin five-coordinate ferric (Fe $^{III}$) state (2) with a vacant coordination site, ultimately available for the binding of dioxygen. Conversion of the ferric iron from low spin to high spin results in a significant increase in the redox potential of the heme group, from −330 to −170 mV for P450$^{CAM}$ (Sono et al 1996). Hence, substrate binding facilitates the transfer of an electron from the redox partner to the ferric P450 heme to generate the high-spin deoxyferrous state, (3). Dioxygen then binds to the ferrous enzyme heme iron to form the oxyferrous complex (4).

Addition of the second electron is the rate-limiting step of the cycle (Brewer and Peterson 1988) and is proposed to yield a ferric peroxide adduct which can be protonated to give a hydroperoxide complex (5). A second protonation of that same oxygen then leads to heterolytic O-O bond cleavage, releasing water and generating the proposed oxo-ferryl (O-Fe IV) porphyrin radical intermediate (6). The existence of these proposed intermediates is supported by recent work (Schlichting et al 2000) and the P450 reaction cycle is completed when an oxygen atom is transferred to the substrate (7), most probably by the oxygen rebound mechanism (Groves 1985) to give the alcohol product and regenerate (1).
Chapter 2

Methods
2.0 Preface

This chapter explains the theory behind the techniques employed to complete the work carried out in this thesis. An overview of the techniques employed is given below (Figure 2.0.1)

Figure 2.0.1 An overview of the techniques employed to complete the work carried out in this thesis.

2.1 Comparative Modelling

Whilst comparative modelling techniques are generally not as accurate in determining protein structure as experimental techniques such as NMR spectroscopy and X-ray crystallography they are able to produce models with rms deviations from the crystal structure as low as 1Å, provided that there exist sufficiently homologous templates of known 3D structure upon which to base the models (Topham et al 1991). In the absence of a crystal structure comparative models provide valuable testable hypotheses.
The studies in this thesis centre on the comparative modelling of cytochromes P450 via the satisfaction of spatial restraints using the program Modeller (Sali and Blundell 1993). An overview of the modelling process is detailed in Figure 2.1.1. There are several fragment-based modelling programs, such as SWISSMODEL (Guex and Peitsch 1997) and COMPOSER (Sutcliffe 1987a,b) that perform as well as Modeller in predicting protein structure, but these approaches do not allow the incorporation of experimentally derived distance restraints.

2.1.1 Identification of Suitable 3D Templates

Central to the generation of comparative models is the identification of proteins with homologous amino acid sequences and hence homologous three-dimensional structures on which to base them. In these studies this was achieved using the
searching algorithm PSI-BLAST (Position Specific Iterated Basic Local Alignment Search Tool) (www.ncbi.nlm.nih.gov/BLAST/) (Altschul et al 1997) that scans the sequence of the target protein, *i.e.* the protein that is the subject of the comparative modelling experiment, against the sequences of the known three-dimensional structures deposited in the protein databank (www.rcsb.org/pdb/) (Berman et al 2000).

PSI-BLAST finds weak homologues to the query sequence by performing a profile-based search. A profile, or weight matrix, allows the user to detect distant sequence relationships, where only a small number of amino acids are well conserved. It is a position-specific scoring table that defines such things as which residues are ‘allowed’ to occupy certain locations, and which regions are amenable to accommodating insertions and deletions. The sequences of proteins from the Protein Data Bank are ranked, and must be above a predefined scoring threshold. By placing this inherent limitation PSI-BLAST seeks to identify the ‘maximal segment pair’, *i.e.* the longest possible sequence pair, which is defined to be maximal if the score associated with it cannot be improved by extending or shortening either segment.

To utilise PSI-BLAST an alignment of each homologue with the query sequence is obtained using BLAST2 – a search that is not profile based. The most significant of these alignments, where significance is described by a predetermined threshold, are used as a profile in a PSI-BLAST iteration to locate weak homologues. This iteration continues until the search converges, as this aids in refining the profile and increases search sensitivity. Once the search has converged the suitability of each of the templates for comparative modelling of the target is indicated by the PSI-BLAST score and E-Value (Expect Value) returned. A high score is indicative of two sequences being very similar. The E-Value provides information on the statistical significance of the search, *i.e.* how much better than random the ‘hit’ from the database search is. However, the three-dimensional structures of the templates should also be examined carefully to ensure that there are no major structural anomalies between them and the other templates before they are approved for comparative modelling.
A drawback with searching via this method is that because of the weakness of the homology a false sequence could be unintentionally drawn into the search – which would bias the searching from that moment on towards highlighting more unrelated sequences, ‘diluting’ the desired profile (Holm 1998). PSI-BLAST does account for this fact, by highlighting sequences of low significance in an attempt to prevent profile dilution by the incorporation of false results.

2.1.2 Alignment of Templates at Amino Acid Level

Having identified all of the amino acid sequences with similarity to the query sequence it is necessary to align them to one another in order to emphasise any sequence, and as a consequence structural, relationships between them. A multiple sequence alignment is the method of choice because it is a more reliable predictor of any clear trends than a pairwise alignment. A multiple sequence alignment effectively filters out any ‘noise’ associated with the sequences under investigation.

A good sequence alignment will show that the structural elements and the binding site residues are aligned between the query and template sequences, and that there are no insertions or deletions in the conserved secondary structures.

Overall, a sequence alignment aims to maximise the chemical identity and properties of amino acids between all of the sequences. This was achieved using CLUSTALX (Thompson et al 1994)

If, for simplicity, we consider a pairwise alignment CLUSTALX identifies two similar residues at a midpoint in both sequences, and then determines points that are similar at both sides of the midpoint at an ever-expanding range until the sequences are fully aligned. CLUSTALX attempts to account for a wide range of similarities in a variety of ways. The first is concerned with the weighting of sequences in an alignment i.e. closely related sequences are assigned lower weightings than highly divergent sequences because the former often contains a substantial amount of duplicate information.
Secondly CLUSTALX offers a choice of substitution matrices, PAM (Percentage of Acceptable Point Mutations per $10^8$ years; Dayhof et al 1978) or BLOSUM (Block Substitution Matrix; Henikoff and Henikoff, 1992) dependent upon the similarity between the sequences to be aligned. For example, if an area displays a high sequence identity, a matrix which assigns a high score to identities or conservative substitutions is used e.g. PAM120, BLOSUM80.

Thirdly CLUSTALX employs gap penalties in an effort to ensure efficient and accurate multiple sequence alignments. There are two distinct types of gap penalties employed by CLUSTALX: a gap-opening penalty (GOP) and a gap-extension penalty. The more similar a pair of sequences the higher the GOP, as this helps prevent insertions into conserved areas (Figure 2.1.2.1). GEPs are important if the sequences are of differing lengths. If one sequence is much shorter than the other we see an increase in GEP to prevent the formation of extended gaps in the shorter sequence (Figure 2.1.2.2).

Despite all of these safeguards no automated multiple alignment algorithm is foolproof, and many alignments have to be altered, at least in part, manually because the amino acid sequence itself gives no indication of structure. CLUSTALX is of particular use in this area as it has a graphical user interface, which allows the user to highlight any problems with the alignment.
**Sequence A** HTCLGMMVSS  
**Sequence B** HCCCLGKMVSS

**Alignment Algorithm**

**Sequence A** HTC-LGMMVSS  
**Sequence B** H-CCLGMKVSS

Large Gap-Opening Penalty

**Sequence A** HTCLGMMVSS  
**Sequence C** FCLGTVMVMST

**Alignment Algorithm**

**Sequence A** HTCLGMMV-S  
**Sequence C** F-CLGTVMVMST

Smaller Gap-Opening Penalty

---

**Figure 2.1.2.1** Illustration of the Gap-Open Penalties applied in the CLUSTALX algorithm. Blue indicates residues in the sequences aligned to identical residues prior to the insertion of gaps, orange indicates residues between the sequences that are aligned to identical residues as a result of incorporating a gap, and black indicates residues that are not aligned to identical residues across the sequences.

**Sequence D** HMFFDST  
**Sequence E** HFDST

**Alignment Algorithm** (Gap Opening)

**Sequence D** HMFFDST  
**Sequence F** HFSRK

**Alignment Algorithm** (Gap Opening)

**Sequence D** HMF-FDST  
**Sequence E** H-FDST

**Alignment Algorithm** (Gap Extension)

**Sequence D** HMF-FDST  
**Sequence E** H-FDST

Low Gap-Extension Penalty

**Sequence D** HMF-FDST  
**Sequence F** H-FSRK

**Alignment Algorithm** (Gap Extension)

**Sequence D** HMF-FDST  
**Sequence F** H-FSRK

High Gap-Extension Penalty

---

**Figure 2.1.2.2**. Illustration of the Gap-Extension Penalties applied in the CLUSTALX algorithm. Blue indicates residues in the sequences aligned to identical residues prior to the insertion of gaps, orange indicates residues between the sequences that are aligned to identical residues as a result of incorporating a gap, and black indicates residues that are not aligned to identical residues across the sequences.
2.1.3 3D Structural Alignment of Templates

If the percentage sequence identity of the templates to the target sequence is low, as is the case for the P450s, it is likely that a more accurate sequence alignment would be obtained from structurally aligning the templates, and deriving the amino acid sequence alignment from the structural alignment. For the purpose of this thesis this was approached in two ways – using the MALIGN 3D routine in the program Modeller (Sali and Blundell 1993), and the FSSP (Fold classification based on Structure-Structure alignment of Proteins) method of Holm and Sander (Holm and Sander 1996)

2.1.3.1 MALIGN3D

The MALIGN3D routine in Modeller uses an initial sequence-based alignment of the templates generated using CLUSTALX (Thompson 1994) as its starting point. It uses this alignment to superpose the 3D structures and thus define a new alignment. This process is repeated iteratively and a sequence alignment based on the rigid body superposition of all the templates is produced.

2.1.3.2 FSSP

The FSSP database (http://www.ebi.ac.uk/dali/fssp/) is based on exhaustive all-against-all 3D structure comparison of protein structures currently in the PDB. All alignments are based on comparing the 3D structures of the proteins using Cα distance matrices – a method that compares local similarity, rather than global similarity as with MALIGN3D. Unlike MALIGN3D, FSSP does not require an initial amino acid sequence alignment as a starting point.

2.1.4 Alignment of Query Sequence to Profile of Aligned Template Sequences.

Once the templates had been aligned, the amino acid sequence of the target protein is appended using CLUSTALX. This is achieved by aligning the template profile
against a sequence alignment of the subfamily of the query protein. A subfamily sequence alignment is obtained by (i) identifying members of the subfamily using a PSI-BLAST search with the query sequence, and (ii) aligning the resulting hits using CLUSTALX. The alignment of a profile, as opposed to a single sequence, is preferable as it reduces the likelihood of misalignment, and also serves to define conserved regions across the sequences, and hence conserved structural features that are likely to be important to the function of the protein.

2.1.5 Comparative Modelling

Once a high-quality amino acid sequence alignment has been obtained it can be used to build comparative models of the target protein. In these studies, the program Modeller (Sali and Blundell 1993) was used to generate comparative models.

Modeller allows the user to generate 3D models of the query protein based on the satisfaction of spatial restraints derived from a database of 80 proteins representing 17 different families. Modeller is dependent upon good sequence alignment as a prerequisite, but manages to simultaneously model all of the components of the protein structure i.e. it is a single-step approach. Modeller achieves this, in part, via the incorporation of homology derived restraints e.g. if the Cα-Cα distances for corresponding residues in a series of templates are similar, Modeller will apply a relatively tight constraint. Conversely, if the distances between corresponding residues are different a loose restraint (i.e. a small force constant) is applied, and if the range of distances is large a penalty term is applied.

Other individual structural features, such as mainchain conformation, bond angle, bond length and dihedral angles, are described using a simple harmonic approximation (Equation 2.1.5.1).

For example, the classical harmonic model for bond length:

\[ E_b = \frac{1}{2} k_b (b-b_0)^2 \]  

(Equation 2.1.5.1)
Where $E_b = \text{potential energy}$

$k_b = \text{force constant}$

$b = \text{bond length}$

$b_0 = \text{reference bond length}$

Probability density functions are then calculated from these harmonic models for each individual structure using classical statistical mechanics (Hill 1960) (Equation 2.1.5.2).

$$P_b = \frac{1}{(\sigma_b \sqrt{2\pi}) \exp \left(-\frac{1}{2} \left(b-b_0/\sigma_b\right)^2\right)}$$

(Equation 2.1.5.2)

Where $P_b$ is the pdf for bond length

$\sigma_b = \sqrt{(k_b T/k_b)}$;

$k_B$ is Boltzmann’s constant

$k_b = \text{force constant}$

$T$ is temperature (K)

If a pdf for a structural feature is high, then there is a high probability that the particular structural feature will be present in the model of the query protein. Once the individual probabilities have been defined they are used as a set of restraints upon which the query model is constructed. Due to the fact that each feature is represented by a different pdf, as opposed to a set of single values, an ensemble of possible solutions is obtained, and the user is able to examine the ensemble as a whole to assess the significance of, for example, various ligand residue interactions.

The probability that the models initially generated by Modeller are strained is high. As a result of this Modeller attempts to refine the ‘crude’ structure using approaches such as simulated annealing and energy minimisation in an effort to create a structure which approximates more closely to the protein in its natural state. It is also possible for the models generated to have significantly better stereochemical quality than the
templates used to produce them because of the spatial restraints being derived from a database of relatively high-resolution structures.

An advantage of Modeller is that it allows the incorporation of experimentally determined restraints, such as the incorporation of protein-ligand interactions as determined by NMR, into the modelling process.

Constant violations with respect to bond length, bond angle and dihedral angle restraints across the family of models produced is usually indicative of an error in the amino acid sequence alignment in that area of the protein. This can usually be corrected by manual manipulation of the sequence alignment in conjunction with a molecular graphics package e.g. Insight II (MSI, San Diego USA) to help check topological equivalence.

2.1.6 Model Evaluation

In the absence of an experimentally determined structure it is good practice to assess the quality of the comparative models produced by using complementary, yet independent, techniques to assess the stereochemical quality, and the quality of the amino acid environments of the models.

A wide range of stereochemical checks are available through the Biotech Validation Suite for Protein Structures (http://biotech.embl-ebi.ac.uk:8400/). These assess how well the stereochemistry of the residues in a given protein structure compare with those observed in well-refined, high-resolution crystal structures. In these studies PROCHECK (Laskowski et al 1993) was used to assess the “quality” of the conformation of the polypeptide backbone and sidechains. One of the most useful stereochemical validations in PROCHECK is the Ramachandran plot, or conformational map (Ramachandran and Sasisekharan 1968, Morris et al 1992).

A Ramachandran plot shows bond angles that are permissible in an L-polypeptide chain (in this case the comparative model). As peptide bonds are planar a polypeptide bond has only two degrees of freedom per residue, the twist about the Cα – N bond
axis, denoted by the Greek letter $\phi$, and the twist about the $C_\alpha - C'$ axis $\psi$. Hence, some 3D orientations would be common due to low steric clash between atoms, and others would be impossible due to the fact that they would bring neighbouring atoms or non-bonded groups closer than their respective van der Waals radii. The quality of a model can be evaluated by considering how many of the residues occupy low-energy allowed conformations. The greater the percentage of residues occupying the allowed core regions of a Ramachandran plot, the more likely it is that the final models can be considered valid.

Ramachandran plots are empirical in nature. The four regions of the plot, described as Most Favoured, Additional Allowed, Generously Allowed and Disallowed (Figure 2.1.6.1) are based entirely on observations made by the analysis of high-resolution crystal structures (Morris et al 1992). Based on this analysis of 118 structures with resolutions $\leq 2.0$ Å a good quality model would be expected to have $>90\%$ of its residues in the most favoured regions of the plot.

**Figure 2.1.6.1** Ramachandran plot of a comparative model of cytochrome P450 2D6. Each residue in the protein is represented by a square. The map is divided into four regions that describe the conformation of peptides in the protein. These are most favoured (red), additionally allowed (yellow), generously allowed (cream) and disallowed (white).
Two programs were used to assess the quality of the amino acid environments of residues in the comparative models. Verify 3D (Luthy et al 1992) works by reducing the 3D environment of each individual amino acid in the protein to one dimension and assigning a score based on solvent accessibility, the polarisability of the residue and the 'preference' for different types of secondary structure. A polar amino acid, such as glutamate, which is exposed to solvent will be assigned a large positive score. The same residue buried in a deep hydrophobic pocket within the protein would be assigned a large negative score. The sum of the scores for each of the residues describes the overall score for the protein, and can be expressed as a line graph (Figure 2.1.6.2). Any regions that are poorly modelled will have scores less than zero, hence Verify 3D gives insight into the areas of the model that require improvement.

![Verify 3D results](image)

**Figure 2.1.6.2** Verify 3D results for the P450 2C5 (1dt6) crystal structure. Poorly modelled residues have compatibility scores less than zero.

Errat (Colovos and Yeates 1993) was also used to assess the amino acid environment in the comparative models and is statistical, as opposed to empirical, in nature. Errat
assesses the distribution of different types of atoms with respect to one another in the protein models, after having categorised them as either C, O or N and thus defining six possible interaction (CC, CO, CN, OO, ON, NN). Errat is a sensitive technique, which is good for identifying incorrectly folded regions in preliminary protein models. However, it is limited in that it does not contain any parameterisation regarding surface polarity, shown to be an important factor in determining whether or not folds in proteins are valid (Novotny et al 1988). Hence, any misfolded proteins, which display surface areas and volumes comparable to their native forms, will not be identified by Errat as having been misfolded. An example of Errat output is given in Figure 2.1.6.3

![Errat plot for a comparative model of P450 2D6. The black columns identify problem areas; they occur at the N-terminus, B', F, G, and H helices and in some loops.](image)

Figure 2.1.6.3 Errat plot for a comparative model of P450 2D6. The black columns identify problem areas; they occur at the N-terminus, B', F, G, and H helices and in some loops.

It has been shown that the Root Mean Square Deviation (RMSD), between main chain atoms in the template with the highest homology to the target and the model is a good method of validation (Venclovas et al 1997). This builds on earlier work (Chothia and Lesk 1986) that showed there was a relationship between the percentage sequence identity and the RMSD of the main chain atoms in homologous regions for different proteins. This RMSD was also used to validate the comparative models produced.
2.2 Docking

2.2.1 Introduction

The active site of a protein can be considered simplistically as a complex three-dimensional lock into which only specific keys (*i.e.*, ligands) will fit. In reality, the lock and key are both flexible and complexes form as a result of induced fit. Experimental methods have elucidated a number of protein-ligand complexes, the structures of which have been deposited in the Protein Data Bank. It would be advantageous to exploit this information by attempting to predict the binding orientation of a novel ligand when the geometry of the active site is known using computational methods. This is an active area of research colloquially known as the ‘docking problem’.

A definitive solution to the docking problem is reliant upon several factors. The first is a powerful searching technique that would allow a comprehensive exploration of the conformational space available to both the receptor and the ligand. An intimate understanding of the process of molecular recognition, so that rapid scoring functions could be developed to predict the correct binding modes of ligands in the active site, would also be necessary, as would the ability for the program to compensate for slight errors in the structure of the protein.

An early example of such a program is DOCK (Kuntz et al 1982). DOCK only considered orientational degrees of freedom, treating both the receptor and the ligand as rigid objects. This places an inherent limitation on the program, and such algorithms have now been superseded by ones that allow at least partial flexibility of the ligand and sidechains of amino acids within the protein active site.

Three docking algorithms were used in the studies of this thesis – GOLD (Genetic Optimisation for Ligand Docking) (Jones et al 1997), FlexX (Rarey et al 1996) and Autodock (Morris et al 1996).
2.2.2 GOLD

GOLD is an automated ligand docking program based on a genetic algorithm that explores the full range of ligand conformational flexibility with flexibility of the sidechains of amino acids in the active site. It also satisfies the fundamental requirement that the ligand must displace loosely bound water molecules upon binding. When tested against 100 complexes extracted from the PDB, GOLD recorded a 71% success rate for identification of the correct binding modes when the ligand was docked back into the active site of the respective proteins.

The genetic algorithm mimics the process of evolution via the manipulation of a collection of data structures termed chromosomes. Each chromosome represents a possible binding orientation of the ligand in the active site of the protein, and is assigned a 'fitness' score based on its relative merits as a solution to the docking problem. This fitness function is evaluated in six stages:

1) A conformation of the ligand and the protein active site is generated

2) The ligand is placed in the active site using a least squares fitting procedure.

3) A hydrogen-bonding energy term is evaluated for the complex. This is the sum of individual bond energies for all ligand and protein donor hydrogen and acceptor atoms across the entire complex.

4) A pairwise steric interaction energy for all of the protein ligand atoms is calculated using a softened 4-8 Lennard Jones Potential (Equation 2.2.2.1)

\[ E_{ij} = A/r_{ij}^6 - B/r_{ij}^4 \]  

(Equation 2.2.2.1)

Where \( E_{ij} \) = the interaction energy between the atom pair, and \( r \) is the distance between the atoms. Adjustments are made for atoms involved in a hydrogen
bond ($E_{ij} = 0$, $r$ is scaled by a factor of 1.43), and all pairwise interactions across the complex are summed to give an overall 'complex energy'.

5) A value for the internal energy of the ligand in the ligand receptor complex is calculated using a 6-12 Lennard-Jones potential (steric contribution Equation 2.2.2.2) and the Tripos forcefield torsional potential (torsional contribution Equation 2.2.2.3).

$$E_{ij} = C/r_{ij}^{12} - D/r_{ij}^6$$  \text{(Equation 2.2.2.2)}

Where $E_{ij}$ = the interaction energy between the atom pair, and $r$ is the distance between the atoms.

$$E_{ijkl} = \frac{1}{2} V_{ijkl} \left[ 1 + \left( \frac{n_{ijkl}}{|n_{ijkl}|} \right) \cos(n_{ijkl} \cdot \omega_{ijkl}) \right] \text{ (Equation 2.2.2.3)}$$

Where $E_{ijkl}$ is the torsional energy associated with four consecutively bonded atoms, $\omega$ is the torsional angle, $n$ is the periodicity and $V$ is the barrier to rotation.

6) The three calculated energy terms (3-5) are added together to give an overall fitness function.

2.2.3 FlexX

FlexX is a rapid docking algorithm. It uses an incremental construction algorithm to place conformationally flexible ligand structures in the active site of the receptor. Conformational flexibility of the ligand is taken into account by systematically generating low energy conformations using the conformational searching algorithm MIMUMBA (Klebe et al 1994).

A ligand core, large enough to make specific interactions with the receptor so that a definite preference for binding orientation can be determined, is selected and placed in the active site of the protein. The ligand core is then built upon. As the size of the
ligand core increases so does the number of possible interactions and hence the probability of correctly predicting the binding mode. However, there is a concomitant increase in the number of possible conformations of the ligand core that leads to an increased computational cost. The best ligand core is that one which has the most potential interaction groups and the fewest alternative conformations. The ultimate objective of ligand core placement is the generation of favourable simultaneous interactions between the ligand and the protein by having the appropriate points on each of them align in 3D space.

FlexX utilises the 'Pose Clustering' algorithm (Linnainmaa et al 1988), which is based on pattern recognition. The ligand core is held as a rigid object with defined interaction points, with the receptor also represented as a set of finite interaction points. A unique transformation of the ligand core onto the receptor is defined by superposing three interaction sites from the ligand core onto three interaction sites on the receptor. The initial step of the core placement algorithm is to determine all such triangles that are compatible.

Having identified all potential transformations they are inspected for problems with bond angles and steric clash with the receptor. All placements are then clustered according to RMS deviation between sets of two placements, and the two nearest clusters merged into one, providing the distance between them is lower than a predetermined threshold value.

Having obtained reasonable placements of the ligand core in the active site, the remaining portions of the ligand are divided into smaller fragments and incrementally grown onto the ligand core, by searching for matching interaction groups on the protein. FlexX utilises large distance tolerances of 2.0Å and angle tolerance up to 20° to ensure that no possible interactions are overlooked.

The scoring function used to rank the docked solution is an estimate of the free energies of binding (ΔG) based on the Böhm scoring function (Böhm 1994) (Equation 2.2.3.1)
\[ \Delta G_{\text{binding}} = \Delta G_0 + \Delta G_{\text{hb}} \Sigma_{\text{hbonds}} f(\Delta R, \Delta \alpha) + \Delta G_{\text{ionic}} \Sigma_{\text{ionic}} f(\Delta R, \Delta \alpha) + \Delta G_{\text{lipop}} \Sigma_{\text{lipop}} f(\Delta R) + \Delta G_{\text{aro}} \Sigma_{\text{aro}} f(\Delta R, \Delta \alpha) + \Delta G_{\text{rot}} N_{\text{ROT}} \quad (\text{Equation 2.2.3.1)} \]

Where \( N_{\text{ROT}} \) is the number of rotational bonds that are immobilized in the complex, \( \Delta G_{\text{hb}}, \Delta G_{\text{ionic}}, \Delta G_{\text{rot}} \) and \( \Delta G_0 \) are values that are calculated, \( \Delta G_{\text{aro}} \) accounts for the interaction of aromatic groups, \( \Delta G_{\text{lipop}} \) is a modified term calculated as a pairwise sum over all atom-atom contacts, \( f(\Delta R, \Delta \alpha) \) is a scaling function that penalises deviation from ideal geometry.

where \( f(\Delta R, \Delta \alpha) = f_1(\Delta R) f_2(\Delta \alpha) \)

\[
f(\Delta R) = \begin{cases} 
1 & \text{when } \Delta R \leq 0.2\text{Å} \\
1-((\Delta R-0.2)/0.4) & \text{when } 0.2\text{Å} < \Delta R \leq 0.6\text{Å} \\
0 & \text{when } \Delta R > 0.6\text{Å} 
\end{cases}
\]

\[
f(\Delta \alpha) = \begin{cases} 
1 & \text{when } \Delta \alpha \leq 30^\circ \\
1-((\Delta \alpha-30)/50) & \text{when } 30^\circ < \Delta \alpha \leq 80^\circ \\
0 & \text{when } \Delta \alpha > 80^\circ 
\end{cases}
\]

and

\[
f'(\Delta R) = \begin{cases} 
0 & \text{when } \Delta R \leq 0.6\text{Å} \\
1-((\Delta R-0.2)/0.4) & \text{when } 0.2\text{Å} < \Delta R \leq 0.6\text{Å} \\
1-(-\Delta R-0.2)/0.4 & \text{when } -0.6\text{Å} < \Delta R \leq -0.2\text{Å} \\
(\Delta R + 0.6)/0.2 & \text{when } \Delta R \leq -0.6\text{Å} 
\end{cases}
\]

Where \( \Delta R = R - R_0 \), \( R \) being the distance between atom centres, and \( R_0 \) the ideal value of the sum of the van der Waals radii of two atoms + 0.6Å
The FlexX system has been tested on more than twenty enzyme systems with the RMS deviations from X-Ray structures around the region of 1 Å (Rarey et al 1996).

2.2.4 Autodock

Autodock 3.0 (Morris et al 1998) is a docking program based on a Lamarckian genetic algorithm (LGA). The program combines a rapid grid-based method for energy evaluation (Goodford 1985, Pattabiraman et al 1985), precalculating ligand-protein pairwise interaction energies so that they can be looked up from a table during simulation with a Monte Carlo simulated annealing search (Metropolis et al 1953, Kirkpatrick et al 1983).

A genetic algorithm efficiently samples the positional, conformational and orientational space available to a ligand binding in the active site. Autodock employs three successive searching techniques in an effort to predict the correct binding mode of a ligand in the following order: a genetic algorithm, a local search method and an adaptive global-local search method based on Lamarckian genetics.

The local search method is based on the work by Solis and Wets (Solis and Wets 1981) and does not require gradient information about the local energy landscape in order to identify a local minimum. As a result, this facilitates the search of torsional space. The hybrid GA and local search method (LS) gives rise to a Lamarckian genetic algorithm (LGA), known to improve the performance of the docking algorithm relative to simulated annealing and GA alone (Hart et al 1994, Morris et al 1998).

A random population of structures is generated and docked into the active site. The production of such a 'generation' of conformations has five distinct stages:

1) Mapping and fitness evaluation. Mapping describes the conformation of the ligand with respect to the protein and allows the fitness of the solution to be evaluated. The scoring function employed is based on the work of Wesson and
Eisenberg and predicts the free energy of binding for a conformation, from which binding constants can be ascertained (Equation 2.2.4.1).

\[
\Delta G = \Delta G_{vdw} \sum (A/r^{12} - B/r^6) \\
+ \Delta G_{hb} \sum E(t) [C/r^{12} - D/r^6 + E_{hbond}] \\
+ \Delta G_{elec} \sum q_i q_j / \varepsilon (r_{ij}) r_{ij} \\
+ \Delta G_{sol} \sum S_i V_j \exp (-r_{ij}^2 / 2\sigma^2) \\
+ \Delta G_{tor} N_{tor} 
\]

(Equation 2.2.4.1)

Where \( \Delta G_{vdw} \) is the binding affinity due to van der Waals interactions

\( \Delta G_{hb} \) accounts for hydrogen bonding interactions

\( E(t) \) is a directional weight based on the angle between the probe atom and the target atom.

\( E_{hbond} \) is the estimated average energy of hydrogen bonding of a polar atom with water.

\( \Delta G_{elec} \) accounts for protein-ligand electrostatic interactions

\( q \) is the charge of the atom/probe

\( \varepsilon \) is the dielectric constant

\( r_{ij} \) is the interatomic distance between probe and target atom

\( \Delta G_{sol} \) accounts for the desolvation of the protein upon ligand binding

\( S_i \) is the atomic solvation parameter of the ligand.

\( V_j \) is the total volume of surrounding protein atoms.

\( \sigma \) accounts for the actual fragmental volume of the surrounding protein atoms.

\( \Delta G_{tor} \) accounts for the loss of rotation of side chains in the protein on ligand binding.

\( N_{tor} \) is the number of sp\(^3\) bonds in the ligand.

2) Selection. This determines which of the individuals in the generation will get to reproduce. It is described by the following function (Equation 2.2.4.2).

\[
n_0 = f_w - f_i / (f_w - \langle f \rangle) 
\]

(Equation 2.2.4.2).

\( n_0 \) is the number of offspring allocated to a particular individual.
$f_w$ is the individual with the worst fitness score

$<f>$ is the mean fitness of the population

When $f_w = <f>$ the search is considered to have converged, and the docking process is terminated.

3) Crossover and mutation. Some of the chromosomes in the original population are split into segments. Each separate segment is then combined with other segments from different chromosomes. This is termed crossover. These crossover ‘children’ then replace the parents in the population – thus ensuring a constant population size.

4) Mutation is then performed on some individuals in the population via the addition of a random number obeying the Cauchy distribution (Equation 2.2.4.3)

\[ C(\alpha, \beta, \chi) = \beta/\pi(\beta^2 + (\chi-\alpha)^2) \]  

(Equation 2.2.4.3)

$\alpha \geq 0, \beta > 0$, when $-\infty < \chi < \infty$

Where $\alpha$ and $\beta$ are parameters affecting the spread and the mean of the distribution. The Gauchy distribution has a bias towards small deviates, but unlike the Gaussian distribution it has thick tails allowing it to generate large changes occasionally.

5) Elitism. A user-defined value that determines how many of the top-ranked solutions are carried forward into the next generation.

After these five steps the LGA is applied to a user-defined proportion of the population. It is an inverse mapping function that makes it possible to finish a local search by replacing an individual in the population with the result of the local search. This is based on the Lamarckian premise that characteristics (in this case conformational, orientational and positional information with regards to a ligand in a protein active site) that have been developed over a ‘lifetime’ can become heritable.
traits. Essentially, this allows a comprehensive search of local energy minima for a given individual in any population, and facilitates the replacement of that individual should better solutions ('fitter' individuals) be identified.

This procedure is carried out on a computational loop until the fitness of the solutions for each ligand converge.

2.3 Characterisation of Active Sites.

This section describes the methods used to validate the docking studies of Chapter 4, where the unexpected binding mode of a ligand was observed using GOLD.

2.3.1 GRID

GRID (Goodford 1985) calculates the interaction energy between a chemical probe (representative of the chemical nature of the ligand) and a protein. Specific probe types correspond to functional groups commonly found in small molecules (methyl group, hydroxyl group, ether oxygen etc.).

Such characterisation is achieved by placing the active site of the protein in a user-defined grid box and measuring the interaction energy between the protein and the chemical probe(s) at discrete grid points. The non-bonded interaction energy calculated at the grid points is defined as the sum of the Lennard Jones potential, the hydrogen bond potential and the electrostatic potential (Equation 2.3.1.1)

\[ E_{xyz} = \sum E_{ij} + \sum E_{el} + \sum E_{hb} \quad \text{(Equation 2.3.1.1)} \]

The Lennard-Jones potential employed is a 6-12 potential and accounts for the steric interaction between the probe and the protein. (Equation 2.3.1.2).

\[ E_{ij} = A/d^{12} + B/d^6 \quad \text{(Equation 2.3.1.2)} \]
Where A and B are constants relating to the electronic configuration, radii and polarisability of the interacting atoms, and d is the internuclear separation between the probe and the protein. The cut-off distance for such interactions in GRID is 8Å.

The electrostatic potential is calculated as follows (Equation 2.3.1.3):

$$E_{el} = \frac{pq}{kd} \left[ \frac{1}{d} + \frac{(\zeta - \epsilon)}{(\zeta + \epsilon)} \right] \left( \frac{\sqrt{d^2 + 4spsq}}{d} \right)$$  

(Equation 2.3.1.3)

Where p and q are the charges on the probe and protein atom pair, k is a constant, d is the distance of separation and sp and sq refer to how ‘deep’ the probe and protein atom are in the protein as a whole, as determined by the number of protein atoms within 4Å. The system is assumed to have a planar interface that separates the homogeneous protein phase ($\epsilon = 80$) from the aqueous solvent phase ($\zeta = 4$).

The hydrogen-bond term is described by the following equation (Equation 2.3.1.4)

$$E_{hb} = (C/d^6 - D/d^4).\cos^m\theta$$  

(Equation 2.3.1.4)

Where C and D are predetermined constants, d is the separation between atom and probe, $\theta$ is the hydrogen bonding angle between the probe atom and the hydrogen-bond donor atom in the protein and m is a value set to 4, except when $\theta$ is less than or equal to 90° – in which case $E_{hb}$ is set to zero. A value of one is assigned to $\cos^m\theta$ in cases where the probe is acting as the hydrogen bond donor. This is due to the fact that it is assumed the probe can orient itself in the active site in any way in order to achieve the most favourable interaction.

The total probe energy at any point is the sum of these three individual terms. The information is stored on the grid, and can be contoured using a molecular graphics package, such as Sybyl (Tripos). Contour maps at low energy levels indicate favourable interactions – and as such can be used to predict where functional groups on a ligand would be positioned in the preferred binding orientation. This in turn means that GRID can be used to validate dockings. If a specific functional group in a docked ligand receptor complex can be superimposed onto a low-energy GRID.
contour map for the corresponding chemical probe it is likely that the docking is valid and accurate.

2.3.2 Principal Component Analysis (PCA)

Advances in automation, and the development of sophisticated microcomputers have made possible the rapid collection and subsequent sophisticated analysis of large amounts of data. The term chemometrics has been coined to describe mathematical and statistical techniques designed to extract useful chemical information from large amounts of chemical data – essentially techniques for multi-dimensional pattern recognition. Principal component analysis (PCA) is such a pattern recognition technique (Miller and Miller 1993)

Principal component analysis allows multi-dimensional information – in the form of pattern vectors – to be projected into 2-dimensional space so that as little information as possible is lost. Figure 2.3.2.1 shows the premise behind this method in the situation where there are only two variables.

Figure 2.3.2.1 Projection of pattern vectors from 2 dimensions to 1 dimension. (i) shows original points (ii) shows projection onto the axes A and B.

Figure 2.3.2.1 shows two distinct groups of well defined pattern vectors (black circles and white circles), and two axes A and B onto which they could be projected –
reducing the dimensionality from two to one. Projection onto axis A would preserve as much distinction as possible between the groups i.e. it accounts for the variance across the pattern vectors well. Projection onto axis B would obscure these pattern differences. Hence projection onto axis A would be preferred, where A and B are linear combinations of the original variables.

Similarly PCA allows the projection of \( n \) dimensions to 2 dimensions. In many cases only some of the original variables will feature significantly in these linear combinations, and thus the variables that contribute to the pattern vectors can be identified. The remaining variables that convey little extra information can then be rejected.

The work carried out in Chapter 5 of this thesis shows how a large amount of chemical data characterising the active sites of a number of protein structures was analysed by PCA using the computer program GOLPE (Baroni et al 1993).

2.3.3 MOLCAD surfaces

MOLCAD surfaces (Heiden et al 1993) can also be used to characterise the active sites of comparative models and crystal structures. Unlike GRID they do not probe the chemical characteristics of an active site – rather they simply generate an image of the general 3D topology associated with it. This can be invaluable as it allows the user to examine the gross differences between topological characteristics of comparative models determined in slightly different manners.

2.4. Scoring Functions

2.4.1 Introduction

Docking programs utilise the known 3D structure of a target protein to give insight into interactions between substrates and residues within the active site of a protein. An imperative prerequisite for the successful application of these tools is the availability of a scoring function to prioritise these dockings. Ideally a scoring function should be fast, able to prioritise a large list (several thousands) of structurally and functionally
diverse ligands, accurate, applicable to a broad range of problems and capable of
dealing with small uncertainties in the 3D structure of a target protein (especially
comparative models). Eleven scoring functions were used to predict the activities of
several known P450 2D6 inhibitors/substrates (Chapter 6).

2.4.2 D-Score

D-Score is a scoring function based on the work by Kuntz (Meng et al 1992). It is able
to account for electrostatic and hydrophobic protein ligand interactions, but fails to
take into account the entropic changes, such as loss of ligand flexibility, which take
place upon protein-ligand binding, and although it does contain a distance dependent
dielectric in order to attenuate charge-charge and other polar interactions there is
evidence to suggest that these interactions dominate when determining the score
(Knegtel et al 1999).

The overall scoring function is a composite of three individually calculated terms, the
contact score, the electrostatic interaction energy and a molecular mechanics
interaction energy. The active site of the receptor is encompassed in a grid, and values
for each of these terms are calculated and stored at each of the grid points.

The contact score is based on a user-specified grid resolution, and two close contact
limits. The first is for polar atoms in the receptor, and the second for non-polar
receptor atoms. For every grid point atoms within these ranges are scored as +1. The
cumulative contact score is then stored at the grid point. If the close contact limit is
violated by any atom for a grid point then a negative (-1) score is assigned. Hydrogen
atoms are ignored for contact scoring. Each heavy ligand atom in the ligand-receptor
complex is assigned the contact score value of the grid point nearest to it.

The electrostatic score is based on potentials calculated from the DelPhi program
(Gilson et al 1987), which uses a finite-difference algorithm to solve the Poisson-
Boltzmann equation (Equation 2.4.2.1)

\[ \nabla [\varepsilon(\chi) \nabla \phi(\chi)] - \kappa(\chi)^2 \sinh (\phi(\chi)) = -4\pi \rho(\chi) \]  

(Equation 2.4.2.1)
Where \( \chi \) = a specific location in space

\[ \phi(\chi) = \text{the electrostatic potential to be calculated} \]

\[ \varepsilon(\chi) = \text{spatial dielectric function} \]

\[ \kappa(\chi) = \text{Debye-Huckel parameter} \]

\[ \rho(\chi) = \text{charge distribution function} \]

\[ \sinh (\phi(\chi)) = \text{the nature of salt accumulation over regions of the protein wrt electrostatic potential.} \]

The DelPhi calculated potential of each ligand atom is calculated by trilinear interpolation of the values at the eight surrounding grid points, which is then multiplied by the point charge of the ligand atom to give electrostatic energy. The total electrostatic energy is the sum of all of the electrostatic energies.

The molecular mechanics energy is calculated as the sum of van der Waals and electrostatic components (Equation 2.4.2.2)

\[
E = \sum_{i} \sum_{j} \left[ A_{ij} / r^{12} - B_{ij} / r^{6} + 332(q_i q_j / D_{ij}) \right] \quad \text{(Equation 2.4.2.2)}
\]

Where \( A_{ij} \) and \( B_{ij} \) are van der Waals attraction and repulsion terms

\[ q = \text{point charge and} \]

\[ r = \text{interatomic distance.} \]

\[ D = \text{dielectric constant} \]

Grid-based scoring can be accomplished efficiently when the ligand and the receptor terms are separable, this is easy for the electrostatic term, but a geometric mean approximation must be adopted in order to separate the van der Waals terms. (Equation 2.4.2.3)

\[
A_{ij} = \sqrt{A_{ii} A_{jj}} \quad \text{and} \quad B_{ij} = \sqrt{B_{ii} B_{jj}} \quad \text{(Equation 2.4.2.3)}
\]

Where \( A = \epsilon(2R^{12}) \)

\[ B = \epsilon(2R^{6}) \]

\[ \epsilon = \text{Depth of the potential well and} \quad R = \text{van der Waals radius of the atom.} \]
Hence, Equation 2.4.2.3 can be substituted into Equation 2.4.2.2 to yield Equation 2.4.2.4

\[ E = \Sigma_{\text{lig}} [n^2 A_{ij} (A_{ij}/r)^2] - [n^2 B_{ij} (B_{ij}/r)^6] + 332q_i (\Sigma_{\text{rec}} q_j/D_{ij}) ] \]

(Equation 2.4.2.4)

This yields three values, \( a_{\text{val}} \), which accounts for the van der Waals interactions of the protein, \( b_{\text{val}} \), which accounts for the van der Waals interactions of the ligand, and \( e_{\text{val}} \), which accounts for the overall electrostatic interaction, that are stored at every grid point.

\[ a_{\text{val}} = \Sigma A_{ij}/r_{jk}^{12} \quad b_{\text{val}} = \Sigma B_{ij}/r_{jk}^6 \quad e_{\text{val}} = 332 \Sigma q_i/D_{jk} \]

Hence the interaction energy for a given ligand atom is described by Equation 2.4.2.5, and is arrived at by the trilinear interpolation of the values of the eight grid points surrounding the ligand atom.

\[ E = \Sigma_{\text{lig}} [n^2 A_{ij}(a_{\text{val}}) - n^2 B_{ij}(b_{\text{val}}) + q_i(e_{\text{val}})] \]  

(Equation 2.4.2.5)

### 2.4.3 Chemscore

Chemscore (Eldridge et al 1997) was developed using a diverse set of 82 training ligands and comprises four terms: a lipophilic term, a metal-ligand binding term, a hydrogen bonding term, and a term to account for loss of ligand flexibility. It is an empirical scoring function derived from databases of protein-ligand X-ray crystal structures and is an extension of the Ludi scoring function developed by Böhm (Böhm 1994).

Initially Chemscore assigns atom types in the receptor and ligand to one of six categories; lipophilic atoms, hydrogen bond donors, hydrogen bond donor/ acceptor, hydrogen bond acceptors, polar atoms and metal atoms.
Despite categorising atoms as polar, the scoring function does not assign different weights to interactions between the receptor and polar atoms and interactions between the receptor and non-polar atoms. The general form of the algorithm employed is as follows (Equation 2.4.3.1)

\[
\Delta G_{\text{binding}} = \Delta G_0 + \Delta G_{\text{hb}} \sum_{\text{hbonds}} g_1(\Delta R) g_2(\Delta \alpha) \\
+ \Delta G_{\text{metal}} \sum_{\text{metal}} f(\tau_{\text{metal}}) \\
+ \Delta G_{\text{lipid}} \sum_{\text{lipid}} f(\tau_{\text{lipid}}) \\
+ \Delta G_{\text{rot}} H_{\text{rot}}
\]

(Equation 2.4.3.1)

where \( f(\Delta R, \Delta \alpha) = f_1(\Delta R) f_2(\Delta \alpha) \)

\[
g_1(\Delta R) = \begin{cases} 1 & \text{when } \Delta R \leq 0.25\text{Å} \\ 1-(\Delta R-0.25)/0.4 & \text{when } 0.25 < \Delta R \leq 0.65\text{Å} \\ 0 & \text{when } \Delta R > 0.65\text{Å} \end{cases}
\]

\[
g_2(\Delta \alpha) = \begin{cases} 1 & \text{when } \Delta \alpha \leq 30^\circ \\ 1-(\Delta \alpha-30)/50 & \text{when } 30^\circ < \Delta \alpha \leq 80^\circ \\ 0 & \text{when } \Delta \alpha > 80^\circ \end{cases}
\]

\( \Delta G_{\text{hb}} \) accounts for the binding energy arising from the formation of hydrogen bonds between ligand and receptor. The penalty functions \( g_1(\Delta R) \) and \( g_2(\Delta \alpha) \) are enforced when there is deviation in the ligand receptor complex from the ideal hydrogen bond length of 1.85Å and the ideal O/N-H-O/N angle of 180° respectively. The severity of the penalty is dependent upon the magnitude of the deviation from ideality. Water mediated contacts are also accounted for in this scoring function, water possessing more than 2 contacts with the receptor is classed as part of the receptor. There is no distinction between ionic and non-ionic hydrogen bonds.
ΔG_{metal}Σ_{AM}(r_{AM}) accounts for the interaction between metal atoms in the receptor and acceptor atoms in the ligand. f(r_{AM}) is a contact term (Figure 2.4.3.1) where r_{AM} is the interatomic distance between the acceptor atom (a) in the ligand and the metal atom (M) in the receptor.

![Graph](image)

**Figure 2.4.3.1** The form of the contact term for metal ligand interaction in Chemscore.

The lipophilic term ΔG_{lipo}Σ_{LL} f(r_{LL}) is also a contact term – as opposed to the area term utilised in Böhm's original work. The contact function f(r_{LL}) has the same form as that used for the metal-acceptor interactions, but parameterisation of the function differs depending upon the ligand and receptor atom types coming into contact with one another and is governed by the equation

\[ R = r_{LVW} + (r_{LVW} + 0.5) \]  
\[ \text{(Equation 2.4.3.2)} \]

Where r_{LVW} is the van der Waals radius of the appropriate ligand atom, and r_{LVW} is the van der Waals radius of the receptor atom. There are two distinct advantages for calculating the lipophilic contribution in this manner. The first is that it negates the necessity for the calculation of computationally expensive area terms, and it also describes a longer-range interaction, as the nearest atoms to a ligand within a hydrophobic pocket are not necessarily the best indication of lipophilicity.
The final term, $\Delta G_{\text{rot}} H_{\text{rot}}$, deals with the reduction in entropy when ligands bind to a receptor. A bond is considered to be frozen when atoms either side of the bond of the bond contact the receptor. Terminal $\text{CH}_3$, $\text{CF}_3$, $\text{NH}_2$ and $\text{NH}_3$ groups are ignored. $H_{\text{rot}}$ is a flexibility penalty for molecules possessing frozen rotatable bonds (Equation 2.4.3.3)

$$H_{\text{rot}} = 1 + (1 - 1/N_{\text{ROT}}) \sum (P_{\text{n}_1}(r) + P'_{\text{n}_1}(r))/2 \quad (\text{Equation 2.4.3.3})$$

Where $N_{\text{ROT}}$ is the number of frozen rotatable bonds, $r$ is the interatomic distance between the atoms of the ligand and the receptor which are in contact, and $P$ and $P'$ are the percentage of non-lipophilic heavy atoms on either side of the rotatable bond.

Inherent limitations within this function are a failure to parameterise weak binders, as there is an absence of data in the databases for such complexes, and the limited application to a wide variety of chemistries and/or uncommon binding geometries which are not well-represented in the database.

2.4.4. Potential of Mean Force (PMF) Score

The PMF score (Muegge and Martin 1999) is based upon ligand-receptor atom pair interactions that are statistical, as opposed to empirical, in nature derived from crystallographic data in the protein databank (Berman et al 2000). An advantage of this scoring function is that it is able to yield free energies directly, as solvation and entropic terms are dealt with implicitly and the initial research suggested that it is particularly suited for ranking different ligands bound to a single receptor.

The premise for a statistical based method is underpinned by the success of predicting and validating protein folding using these methods (Sippl 1990, Jones et al 1992). However, there are a number of prerequisites for deriving a potential of mean force based scoring function – there must be a large cut-off for the derivation of the potential of mean force, there must be a larger (statistically significant) number of
atom pair types, there must be an appropriate reference state for purpose of comparison, and the term must include a suitable ligand volume correction. There are also two assumptions that allow the derivation of energy functions using statistical mechanics. The first is that the protein-ligand complex is in a state of thermodynamic equilibrium, and the second is that the distribution of molecules in the microscopic state obeys Boltzmann’s law.

For the purposes of the scoring function there are 16 different types of protein atom, and 34 types of ligand atom – all derived from data stored in the protein databank. If the total number of atom pairings in the protein databank was less than 1000 they were omitted from the parameterisation as they were regarded as statistically insignificant. This negated the necessity to incorporate sparse data corrections or smoothing algorithms into the scoring function. The protein ligand interaction free energy, $A$, between an atom of the protein (i) and an atom of the ligand (j) is given in Equation 2.4.5.1.

$$A_{ij}(r) = k_B T \ln \left[ \langle \rho_{\text{vol,corr}}^{ij}(r) \rangle \cdot \rho_{\text{seg}}^{ij}(r) / \rho_{\text{bulk}}^{ij}(r) \rangle \right] \quad \text{(Equation 2.4.5.1)}$$

Where $k_B$ = Boltzmann’s constant

$T$ = Absolute temperature

$\rho_{\text{vol,corr}}^{ij}$ = Ligand volume correction factor

$r$ = Interatomic distance

$\rho_{\text{seg}}^{ij}$ = Number density of pairs of atom types in the database that occur in the radius range ‘seg’

$\rho_{\text{bulk}}^{ij}$ = The distribution when no interaction occurs.

The potential of mean force score is determined by the sum over all pairwise interactions in the ligand and the protein given in equation 2.4.5.2

$$\text{PMF Score} = \Sigma A_{ij}(r) \quad \text{(Equation 2.4.5.2)}$$
The volume correction factor becomes important when interatomic distances are less than 4 Å, taking a value between 2 and 8. This serves to alter the depth of the potential mean force wells, but does not affect the extremes of the PMF values.

### 2.4.5 G-Score and GOLD Score

G-Score and GOLD Score (Jones et al 1995, Jones et al 1997) are founded on hydrogen bonding interactions, and contain terms that incorporate the desolvation of acceptors and donors in the hydrogen bond, a pairwise dispersion that accounts for non-polar or hydrophobic interactions, and a molecular mechanics equation that calculates the internal strain energy of the molecule.

The hydrogen bonding term is computed first. Each possible combination of donor H atom and acceptor atom is examined in turn and the geometrical fitting assessed. This fitting is then given a weighting between zero (bad fit) and one (perfect fit). The weight is the product of two terms – angle weight and distance weight.

The pairwise hydrogen bonding energy is given by the following (Equation 2.4.5.1)

$$ E_{\text{pair}} = (E_{\text{da}} + E_{\text{ww}}) - (E_{\text{dw}} - E_{\text{ww}}) $$  

(Equation 2.4.5.1)

Where $E_{\text{pair}}$ is the pairwise energy for the hydrogen bond

- $E_{\text{da}}$ is the pairwise H-bond energy between the donor and acceptor atoms
- $E_{\text{ww}}$ is the pairwise H-bond energy between water atoms
- $E_{\text{dw}}$ is the pairwise H-bond energy between donor atoms and water

The sum of all pairwise interactions is then calculated to give overall hydrogen bonding energy. Acceptor and donor atoms are predefined within the algorithm.

The pairwise steric ligand receptor interaction energy for those solutions which score best from this initial assessment are then calculated using a softened 4-8 Lennard Jones Potential (Equation 2.4.5.2)
Where $E_{ij}$ is the receptor ligand interaction energy between atoms $i$ and $j$

$\Delta_{ij}$ is the interatomic distance between atoms $i$ and $j$

$A$ and $B$ are predetermined constants $E_{ij}$ is a minimum when $d_{ij} = r_i + r_j$.

These pairwise interactions are summed across all of the atoms in the receptor and the ligand to give the complex energy.

Finally the internal steric and torsional energies of the ligand are calculated using the traditional 6-12 Lennard Jones potential (Equation 2.4.5.3) to calculate steric interaction and the Tripos forcefield torsional potential (Clark 1989) (Equation 2.4.5.4) to calculate torsional energies. This is referred to as the internal energy.

$$E_{ij} = C/d_{ij}^{12} - D/d_{ij}^{6} \quad \text{(Equation 2.4.5.3)}$$

Where $E_{ij}$ is the receptor ligand interaction energy between atoms $i$ and $j$

$\Delta_{ij}$ is the interatomic distance between atoms $i$ and $j$

$C$ and $D$ are predetermined constants: $E_{ij}$ is a minimum when $d_{ij} = r_i + r_j$.

$$E_{ijkl} = \frac{1}{2}V[1+n/(n\cos|\omega|)] \quad \text{(Equation 2.4.5.4)}$$

$V$ is the barrier to rotation

$n$ is the periodicity

$\omega$ is the torsional angle

The final fitness function is given by Equation 2.4.5.5

$$\text{Fitness} = -\text{H-Bond Energy} - (\text{Internal Energy} + \text{Complex Energy}) \quad \text{(Equation 2.4.5.5)}$$
2.4.6 Ligscore 1 and Ligscore 2

Ligscore 1 and Ligscore 2 (Accelrys 2002) are fast and very simple scoring functions used to predict protein-ligand binding affinities. Primarily they are intended to identify potential lead candidates from a large database of ligands.

Three descriptors are used to calculate Ligscore. The three descriptors are:

1) VdW. This is a softened Lennard-Jones 6-9 potential. The vdw descriptor is expressed in units of kcal/mol.

2) C+pol. This is a count of the buried polar surface area between a protein and ligand that involves attractive protein-ligand interactions. The C+pol descriptor is expressed in units of Å².

3) Totpol². This is a count of the buried polar surface area between a protein and ligand involving both attractive and repulsive protein-ligand interactions. The output is the square of the count. The Totpol² descriptor is expressed in units of Å⁴.

Similar equations are used for Ligscore 1 and 2 depending on the vdw descriptor being calculated. They have the following form and differ from one another only in the values of their constants (Equations 2.4.6.1 and 2.4.6.2)

\[
\text{Ligscore 1} = A - B \cdot \text{vdW\_exact} + C \cdot \text{C+pol} - D \cdot \text{Totpol}^2 \quad (\text{Equation 2.4.6.1})
\]

\[
\text{Ligscore 2} = A - B \cdot \text{vdW\_grid} + C \cdot \text{C+pol} - D \cdot \text{Totpol}^2 \quad (\text{Equation 2.4.6.2})
\]

A,B,C and D are coefficients obtained via linear regression analysis.

2.4.7 Piecewise Linear Potential (PLP) 1 and 2

Both PLP scoring functions (Gehrhaar et al 1995a, Bouzida et al 1999) are defined by intramolecular energy terms as well as intermolecular ligand-protein interaction terms
containing parameterisation for steric and hydrogen bond contributions. PLP 1 and 2 were initially designed to rank the different conformations of a single ligand docked in a variety of conformations within the active site of a protein.

PLP 1 is a minimalistic scoring function – containing only three terms – a term that accounts for steric complimentarity, a non-directional term that accounts for hydrogen bond contributions, and a rudimentary intramolecular term to account for the strain energy in the bound ligand.

The steric and hydrogen bond contributions arise from a linear potential function of the form shown in Figure 2.4.7.1. An advantage in using this functional form is that there is a finite value as the interatomic distance approaches zero – so the function is well behaved even when there are severe close contacts between the ligand and the receptor – such as at the beginning of a docking investigation. The potential parameters are modified values extracted from a potential used for the de novo design of enzyme inhibitors (Gehlhaar et al 1995b). The internal energy of the ligand is described by torsional potential and a non-bonded term (Equation 2.4.7.1)

\[ E = A[1 - \cos(n\phi - \phi_0)] \quad (\text{Equation 2.4.7.1}) \]

Where for sp³-sp³ bonds \( A = 3.0, n = 3 \) and \( \phi_0 = \pi \), and for sp³-sp² bonds \( A = 1.5, n = 6 \) and \( \phi_0 = 0 \).

The protein-ligand energy function is a pairwise sum over all heavy atoms, where the parameterisation is dependent upon interaction between two atoms of the four predetermined atom types – hydrogen bond donor, hydrogen bond acceptor, both acceptor and donor and apolar atoms. It has been found because of its loose parameterisation i.e. a lack of directionality terms, this scoring function is particularly effective for ranking ligands docked or partially docked into narrow or shallow active sites (Rarey et al 1999).
Figure 2.4.7.1: The piecewise linear potential function used for the ligand-protein interaction energy. Values are given in Table 2.4.7.1

Table 2.4.7.1: Parameters of the atomic pairwise ligand-protein potentials\(^a\) for PLP

<table>
<thead>
<tr>
<th>Interaction Type</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steric</td>
<td>3.4</td>
<td>3.6</td>
<td>4.5</td>
<td>5.5</td>
<td>-0.4</td>
<td>20.0</td>
</tr>
<tr>
<td>Hydrogen Bond</td>
<td>2.3</td>
<td>2.6</td>
<td>3.1</td>
<td>3.4</td>
<td>-2.0</td>
<td>20.0</td>
</tr>
</tbody>
</table>

\(^a\) A, B, C and D are in Angstroms (Å). E and F are in arbitrary energy units.

2.4.8 LUDI Score

The LUDI score (Böhm 1994) assesses protein-ligand interaction on the basis of a lipophilic term, a metal-ligand binding term, a hydrogen bonding term, and a term to account for loss of ligand flexibility i.e. the number of rotatable bonds which are frozen upon ligand receptor binding.
This scoring function sets out to estimate the free energy of binding $\Delta G$ and is able to accommodate small uncertainties in the protein structure. The following free energy function is used to calculate binding (Equation 2.4.8.1)

$$
\Delta G_{\text{binding}} = \Delta G_0 + \Delta G_{\text{hb}} \sum_{\text{hbonds}} f(\Delta R, \Delta \alpha) \\
+ \Delta G_{\text{ionic}} \sum_{\text{ionic}} f(\Delta R, \Delta \alpha) \\
+ \Delta G_{lipo} |A_{lipo}| \\
+ \Delta G_{\text{rot}} \text{NROT}
$$

where $f(\Delta R, \Delta \alpha) = f_1(\Delta R) f_2(\Delta \alpha)$

- $f_1(\Delta R) = \begin{cases} 
  1 & \text{when } \Delta R \leq 0.2\text{Å} \\
  1-(\Delta R-0.2)/0.4 & \text{when } 0.2\text{Å} < \Delta R \leq 0.6\text{Å} \\
  0 & \text{when } \Delta R > 0.6\text{Å}
\end{cases}$

- $f_2(\Delta \alpha) = \begin{cases} 
  1 & \text{when } \Delta \alpha \leq 30^\circ \\
  1-(\Delta \alpha-30)/50 & \text{when } 30^\circ < \Delta \alpha \leq 80^\circ \\
  0 & \text{when } \Delta \alpha > 80^\circ
\end{cases}$

Where $\Delta G_0$ is a non-specific interaction with the protein rationalised as a reduction in the translational and rotational entropy of the ligand. $\Delta G_{\text{hb}}$ is the contribution from ideal hydrogen bonds, and $\Delta G_{\text{ionic}}$ is the contribution from unperturbed ionic interactions.

$\Delta G_{lipo}$ represents the lipophilic contribution for the binding of a ligand to the receptor. It is assumed to be proportional to the contact surface area $|A_{lipo}|$ between the ligand and the receptor. $|A_{lipo}|$ is calculated using a coarse cubic grid of 1Å that encases the ligand receptor complex. Each grid point is marked as lipophilic or ionic. Any cube that is determined to overlap with the ligand i.e. any cube in which the distance between the ligand atom and the centre of the cube is less than the van der Waals radius of the atom, is highlighted. Any cubes adjacent to the ‘ligand cubes’ are then
also highlighted. Any of the second set of cubes that contain a receptor atom are then used to calculate the lipophilic contact surface area between the ligand and the receptor. The actual value obtained from this analysis is multiplied by a calibration factor of 0.71, derived from a comparison with surface areas calculated by the MS program (Connoly 1983).

$\Delta G_{\text{rot}}$ is a term defining the reduction in binding energy due to the loss of internal degrees of freedom in the ligand i.e. the number of rotatable bonds that become frozen upon ligand-receptor binding. NROT is the number of acyclic $sp^3-sp^3$ and $sp^3-sp^2$ bonds. Terminal CH$_3$,NH$_3$ and bonds in cyclic groups are not taken into account.

$AR,\Delta \alpha$ is a penalty function. $AR$ accounts for the deviation from the ideal hydrogen bond lengths between ligand and receptor (1.9Å). $\Delta \alpha$ penalises deviation from the ideal O/N-H...N/O bond angle of 180°. The exact value of these penalties varies depending upon the extent of the deviation (Equation 2.4.8.1).

There are several limitations associated with this scoring function. The most important is that it does not contain a term to penalise repulsive interactions between ligand and receptor. This can lead to an overestimation of binding affinity for larger ligands.

### 2.4.9 F-Score

F-Score (Rarey et al 1999) is based on LUDI, but incorporates changes with regards to the parameterisation of the four individual LUDI terms. The general form of the equation employed is as follows (Equation 2.4.9.1)

$$
\Delta G_{\text{binding}} = \Delta G_0 + \Delta G_{\text{hb}} \Sigma_{\text{hbonds}} f(\Delta R, \Delta \alpha)
+ \Delta G_{\text{ionic}} \Sigma_{\text{ionic-int}} f(\Delta R, \Delta \alpha)
+ \Delta G_{\text{lipophilic}} \Sigma_{\text{lipophilic-cont}} f'(\Delta R)
+ \Delta G_{\text{arom}} \Sigma_{\text{arom-int}} f(\Delta R, \Delta \alpha)
+ \Delta G_{\text{rot}} \text{NROT}
$$

(Equation 2.4.9.1)
where $f(\Delta R, \Delta \alpha) = f_1(\Delta R) f_2(\Delta \alpha)$

\[
\begin{align*}
    f(\Delta R) &= 1 \quad \text{when } \Delta R \leq 0.2\text{Å} \\
    &= 1-(\Delta R-0.2)/0.4 \quad \text{when } 0.2\text{Å} < \Delta R \leq 0.6\text{Å} \\
    &= 0 \quad \text{when } \Delta R > 0.6\text{Å} \\
    f(\Delta \alpha) &= 1 \quad \text{when } \Delta \alpha \leq 30^\circ \\
    &= 1-(\Delta \alpha-30)/50 \quad \text{when } 30^\circ < \Delta \alpha \leq 80^\circ \\
    &= 0 \quad \text{when } \Delta \alpha > 80^\circ \\
\end{align*}
\]

and

\[
\begin{align*}
    f'(\Delta R) &= 0 \quad \text{when } \Delta R \leq 0.6\text{Å} \\
    &= 1-(\Delta R-0.2)/0.4 \quad \text{when } 0.2\text{Å} < \Delta R \leq 0.6\text{Å} \\
    &= 1-(-\Delta R-0.2)/0.4 \quad \text{when } -0.6\text{Å} < \Delta R \leq -0.2\text{Å} \\
    &= (\Delta R + 0.6)/0.2 \quad \text{when } \Delta R \leq -0.6\text{Å} \\
\end{align*}
\]

Where $\Delta R = R - R_0$, $R$ being the distance between atom centres, and $R_0$ the ideal value of the sum of the van der Waals radii of two atoms + 0.6Å.

As for some of the other scoring functions (Böhm 1994, Eldridge et al 1997) the $\Delta G_0$ is rationalised by a loss of translational and rotational entropy upon ligand binding, and $\Delta G_{\text{rot NROT}}$ accounts for the loss of ligand flexibility upon ligand-receptor binding – where NROT is the number of rotatable bonds that become frozen. $\Delta G_{\text{ionic}}$ accounts for the contribution to binding energy from ionic interactions. $\Delta G_{\text{aro}}$ is a term that deals explicitly with the interactions between aromatic atoms in the ligand and receptor. Once again the penalty function $f(\Delta R)$ is used to penalise any hydrogen bonds that deviate from an ideal distance of 1.85Å and $f(\Delta \alpha)$ penalises deviations from the ideal O/N-H···N/O hydrogen bonding angle of 180°. The lipophilic term, $\Delta G_{\text{lipo}}$, is modified from Böhm's original work, and is calculated as a sum over all pairwise atom-atom contacts. The penalty term associated with it $f'(\Delta R)$ is
incorporated to penalise forbiddingly close contacts. All penalty terms are heuristic *i.e.* all of the values employed are as a result of a trial and error investigation of a number of different terms.

F-Score, like all scoring functions is limited – and in its inaugural run it only succeeded in predicting the correct binding orientation of 46.5% of the ligand-protein complexes examined (Rarey et al 1996). It was hypothesised that this was due to a lack of filter functions (Stahl and Böhm 1998) in the algorithm – *i.e.* there was no way of extracting undesirable elements in the docked complexes that were not shared with crystal structures.

**2.5. Quantitative Structure-Activity Relationship (QSAR)**

**2.5.1 General Introduction**

In drug design it is important to realise that a molecule may be required to show qualities beyond *in vitro* potency, as it is the *in vivo* efficacy of a compound that will determine its success as a therapeutic agent. A structure-activity study can aid in the understanding of which features of a molecule are responsible for activity – and how they could be modified to enable the production of enhanced compounds. Quantitative structure-activity relationships (QSARs) relate numerical properties of a molecular structure to the activity of a molecule via a mathematical model. It can be of particular use when there is no knowledge of the three-dimensional structure of the therapeutic target, as it facilitates the prediction of biological activity for novel compounds.

QSAR can be split into several different categories. The most common are classical QSAR, which uses only numerical values that can be calculated from a small molecule structure, such as molecular weight, to generate QSAR equations, 3D QSAR which also takes into consideration the 3D properties exhibited by the ligands such as their molecular volume and 4D QSAR (Ekins et al 1999) which combines 3D QSAR with some information (usually in the form of a pharmacophore) about the protein structure.

MFA, RSM, Cerius² descriptor based and CoMFA are all 3D QSAR methods, and as such are dependent upon the ligands in the datasets being in active conformations and aligned to one another. In contrast, HQSAR does not require 3D information with respect to the ligands in a dataset, and as such the dataset does not have to be in active conformations or aligned.

2.5.2 Molecular Field Analysis (Cramer et al 1988, Accelrys 2002)

Molecular field analysis calculates the interaction energy of a chemical probe over a set of aligned ligands. An energy field, also known as a probe map, is generated by placing a probe molecule at a random position and moving the probe across a set of ligands aligned within a 3D grid box. The probes used in the studies in Chapter 7 were CH₃⁺, CH₃, CH₃⁻, H⁺ and H⁻. An energy calculation, corresponding to steric, electrostatic and hydrogen bonding energies, is performed at every grid point, which allows the energy of interaction between the ligands and the probe to be determined. The intramolecular energy of the ligand is ignored during this calculation, as the atoms are held rigid in the grid box. Having calculated the probe map for each of the ligands the energy values measured at each grid point are correlated with the experimentally determined binding affinity using a robust correlation method such as Genetic Function Approximation or Genetic Partial Least Squares analysis (Section 2.6 Statistical Methods). This results in a linear equation describing those points on the grid, for a particular type of probe, that appear to be responsible for changes in biological activity.
2.5.3 Receptor Surface Model (Hahn 1995, Hahn and Rogers 1995, Hahn et al 1997)

A receptor surface model is usually generated in an attempt to predict the biological activity of a compound, by producing a hypothetical model of the receptor from known activity data. However, it also has application in QSAR, in that it can be used to relate calculated energies to experimentally determined biological activity.

A receptor surface model is non-atomistic, using explicit surfaces to characterise the shapes of the active site. It represents essential information about the hypothetical receptor site as a 3D surface with appropriate properties mapped on to the surface model. For example, the shape of the RSM will give insight into the steric nature of the hypothetical receptor and mapped across the surface will be properties such as hydrophobicity, partial charge, electrostatic potential and hydrogen bonding propensity.

An RSM is generated from an alignment of molecules of known binding affinity. Any errors in this initial alignment will significantly reduce the applicability of the receptor surface model. The surface is generated for each structure (or shape field), which are then combined to give an overall receptor shape (Wyvill 1986). In practice, a large aggregate object is formed from smaller interacting objects. This is accomplished by computing values for a set of field sources (i.e. atoms of the ligands) at different points in space, according to a specified field function. The overall shape of the receptor surface model is based on the individual contributions from each field source.

Two field functions are commonly used to generate receptor surface models. The first is the van der Waals function, which unsurprisingly characterises the van der Waals shape of the receptor surface model clearly, according to Equation 2.5.3.1

\[ V(r) = r - V_{dwr} \]  

Equation (2.5.3.1)

Where \( r \) = the interatomic distance from a given point to the centre of an atom
\( V_{dwr} \) = van der Waals radius of the atom.
The disadvantage of this function is it is very 'hard', and if the receptor surface model is being used to predict the biological activity of ligands with vastly differing structures to the ligands used to produce the RSM problems could occur.

As a result an alternative field function (Wyvill 1986) that is also distance dependent is often used (Equation 2.5.3.2)

\[ V(r) = \frac{4}{9}(r^6/R^6) - \frac{17}{9}(r^4/R^4) - \frac{22}{9}(r^2/R^2) + 1 \]  

(Equation 2.5.3.2)

For \( r > 0 \) and \( R < 2 \)

Where \( r \) = distance between atom and point

\( R \) = predetermined distance \(< 2 \text{Å} \)

Therefore \( V(0) = 1, V(R) = 0 \) and \( V(R/2) = \frac{1}{2} \), resulting in the smooth blending of the receptor surface model, meaning that the surface fit is directly related to the volume it encloses. Also, because the initial parameterisation is not as tight as for the van der Waals field function it is more likely that the RSM will be able to cope with ligands that are different to those used to generate it.

Properties associated with the putative chemical properties are then calculated and stored at each surface point, the property value at any given point being defined as the mean of the corresponding vertex volumes. The properties mapped onto the surface include partial atomic charge, electrostatic potential, hydrogen bond propensity, and hydrophobicity. These descriptors represent the binding energies of ligands bound to the receptor site.

Having mapped these properties the Receptor Surface Model can then be used in QSAR studies. By fitting ligands not used to generate the RSM to the surface and performing a QSAR followed by regression analysis using Partial Least Squares Regression or the Genetic Function Approximation (Section 2.6) it is possible to ascertain which properties associated with the ligands appear to be important in ligand-receptor binding. A limitation of RSM QSAR is that the grid points that are
generated can be very close to the ligand atoms, which results in the measured interaction energies becoming very large. This means that PLS may struggle to find useful patterns in the data, and GFA should be used in preference if possible. This will yield an equation of the form shown below (Equation 2.5.3.3). This information can then be used to direct the drug-design process.

$$\text{Activity} = a + b.\text{property1} + c.\text{property2} \ldots \quad \text{(Equation 2.5.3.3)}$$

Where a, b and c are coefficients determined by regression.

2.5.4 Cerius² Descriptors (Accelrys, San Diego, CA, USA)

A range of spatial, structural, thermodynamic, electronic and quantum mechanical descriptors can be calculated for ligands, with a view to using them to generate QSARs. The descriptors used in the study of this thesis are presented in Chapter 7. The physical and chemical properties calculated for each descriptor calculated for each ligand can be used to derive predictive QSAR models, and should ideally be uncorrelated with respect to one another, as this would prevent the QSAR being biased in any way. This is achieved by generating the QSAR using a large number of descriptors and then removing any that are cross-correlated as determined by a cross-correlation matrix. The removal of these descriptors should not be detrimental to the overall predictive ability of the QSAR. Once a stage has been reached where removal leads to a reduction in the quality of the QSAR the optimum number of descriptors has been achieved. It does not follow that a larger number of QSAR descriptors will lead to a better model (Ekins and Obach 2000). The experimental binding affinity of the ligands are correlated to chemical and physical properties calculated by the descriptors using PLS analysis.

2.5.5 CoMFA (Cramer et al 1988)

Comparative Molecular Field Analysis (CoMFA) operates in much the same manner as Molecular Field Analysis (MFA). It is based on the observation that the most relevant properties in ligand-receptor binding would be related to shape, and that by sampling the steric and electrostatic fields surrounding a molecule an understanding
of the biological activity of that molecule could be gained. An energy probe map is generated by placing a chemical probe in a random location in space (CH$_3^+$ in this case as it was the only probe available) and then moving the probe across the set of ligands aligned within a 3D grid box. A data table is constructed from the probe map at the grid lattice intersections, two values, one accounting for electrostatic properties and one accounting for steric interactions between the probe and the receptor, are calculated for each grid point. Extraction of a stable QSAR is then achieved using PLS analysis. This is in contrast to MFA, which must employ either G/PLS or GFA for proprietary reasons.

PLS or a similar method must be used because of the unusually underdetermined data table that is produced as a result of this type of analysis. Ordinary Linear Squares Regression (OLSR) would not extract a robust QSAR.


Hologram QSAR (HQSAR) is a 2D technique, requiring no 3D information from the ligands in a dataset. A chemical structure is converted to a characteristic molecular fingerprint based on the enumeration of certain types of molecular fragments present in the ligand. This numerical representation of molecules is then used as the QSAR descriptor.

This is achieved by first of all hashing the ligand to a molecular fingerprint that encodes the frequency of occurrence of a number of fragment types based on a predetermined set of rules. This fingerprint is then cut into ‘holograms’ – a string of a known length defined by a hologram length (HL) parameter. These holograms are then aligned, and the sum of each column represents the individual component of a molecular hologram at a particular length (Figure 2.5.6.1).
Figure 2.5.6.1 Transformation of a chemical structure to its characteristic molecular hologram (So and Karplus 1999). In this example a glucose molecule is converted to a molecular hologram with a hologram length of 5. Subsequently each of the values inside the five boxes serves as an independent descriptor for the later PLS analysis.

The hashed fingerprints encode the presence of all molecular fragments between four and seven atoms, and hydrogen atoms and chiral centres are included in the generation of molecular holograms. The optimal HQSAR model is determined by screening through the default HL values in the Sybyl (Tripos Inc.) HQSAR module.
2.6 Pharmacophores

2.6.1 Introduction

In drug design the term pharmacophore refers to a set of chemical features that are common across a set of biologically active compounds such as hydrogen bond donors and acceptors, positively and negatively charged groups and hydrophobic fragments. These are referred to as ‘pharmacophoric groups’. A 3D pharmacophore specifies the spatial relationships between pharmacophoric groups, often expressed as distances or distance ranges, although other geometrical features such as angles and planes may be incorporated.

The development of methods for studying the conformation of ligands has stimulated an interest in the influence of the 3D structure of molecules on their chemical and biological activity. The objective of the procedure known as pharmacophore mapping is to determine possible 3D pharmacophores for a series of active compounds. Once such a pharmacophore has been developed it can be used to find or suggest other active molecules.

When building a pharmacophore it is imperative that the conformational properties of all of the ligands under investigation are taken into account. It should also be noted that all approaches to building a pharmacophore assume a common binding orientation in the active site of a protein for even structurally disparate ligands. In these studies, pharmacophores were generated using the program Catalyst (Accelrys San Diego CA USA) for the 2D6 model (Chapter 8) based on the ligands in Dataset 1 (Appendix 6.1).

2.6.2 Catalyst

Catalyst was used to generate the pharmacophore hypotheses presented in this thesis. It provides software to generate conformers of the ligands to be studied and identify the chemical features of active molecules that can then be used in hypothesis generation. As a prerequisite to hypothesis generation, a training set of compounds with known biological activities is required. Hypothesis generation is an iterative
process that proceeds through 4-6 major cycles called phases. During major intervals in each phase catalyst writes out the ten lowest cost hypotheses, where the cost is determined by how well the known activities of the training set are predicted by the pharmacophore. When the cycles are complete the lowest score hypothesis is usually considered the best, although any other hypotheses scoring within 10-15 units of the lowest score should also be investigated visually. A null hypothesis is also generated, this is the worst i.e. highest cost pharmacophore that the program generated during its phase cycles, and can give insight into what features and spatial arrangements are important for the development of a good pharmacophore.

Another indication of how good the catalyst hypothesis is can be determined by examination of the 'Config' score associated with it. If this score is greater than 17 it is likely that the training set being used has too many degrees of freedom i.e. the structures are too dissimilar. As a result any pharmacophore generated, regardless of its overall total cost, would be unlikely to provide any valuable information.

2.7 Statistical Methods

2.7.1 Introduction

Analysis using statistical methods is an integral part of building models, estimating the predictive ability of the models and determining relationships and correlations between computed variables and biological activity data. For the studies of this thesis regression methods have been used to help interpret the data generated for Chapter 6 and Chapter 7 – relating computed variables back to experimentally determined biological data. Regression methods are used to build a model in the form of an equation that gives one or more dependent variables (such as biological activity) in terms of a number of independent variables (descriptors). The model can then be used to predict the activities of novel compounds, for example by screening a database of compounds with unknown biological activities.
2.7.2. Ordinary Least Squares Regression (OLSR)

Ordinary least squares regression methods assume that the model being fitted is linear in nature. This procedure for fitting a line through a set of \( n \) data points aims to minimise the deviations between the observed values and the corresponding points on the fitted line. This is achieved by minimizing the sum of the squares of the vertical deviations from the fitted line (Figure 2.7.2.1).

Given that the equation of a straight line can be written as shown below (Equation 2.7.2.1):

\[
y = mx + c
\]  
(Equation 2.7.2.1)

Where \( y \) is the independent variable, \( x \) is the dependent variable, \( m \) is the gradient of the fitted line and \( c \) is the intercept on the y axis, the deviation of the observed value of \( y' \) from the predicted value \( y \) of the fitted line is given as (Equation 2.7.2.2)

\[
\text{Deviation from line of best fit} = y' - y
\]  
(Equation 2.7.2.2)

Hence the sum of the squares of deviations (SSD) to be minimised (Equation 2.7.2.3) is:

\[
\text{SSD} = \sum (y' - y)^2
\]  
(Equation 2.7.2.3)

A limitation with this method of regression analysis is that it only works well for those models that are very well characterised. In the case of the QSAR work carried out in Chapter 7 the matrices that are produced are highly underdetermined \( i.e. \) the matrices that are produced contain more data in their rows in comparison to the amount of data in the columns. As such OLSR would struggle to find a robust relationship between biological activity data and the descriptors used to calculate the equation. Hence, other statistical methods are used.
Figure 2.7.2.1 A line fitted between five data points using the OLSR method. The minimised vertical distances between the calculated line and the observed values for the dependent variable (y) are illustrated.

2.7.3 Partial Least Squares Fitting (PLS)

PLS (Glenn et al 1989) provides a least squares estimate of the dependent variable, y, on the independent variable, ‘x’. The difference between PLS and OLSR is that PLS extracts latent variables that are approximately along the axes of greatest variation in x and y and those that are optimally correlated (similar to Principal Component Analysis). This exploits any colinearity that may be present in the x data, and as a result leads to more robust estimates of ‘y’ data than OLSR alone. PLS has great applicability in QSAR calculations, where there are many more variables than compounds being studied, and has a further advantage in that it can be used in cases with more than one dependent variable. OLSR dictates that only a single dependent variable can be analysed at one time.

The first component in a PLS regression is the direction of the greatest variation in the independent variables. The second describes the greatest variance across all of the
independent variables, but is also orthogonal to the first component. As more and more components are described the extent to which they account for variance between the independent variables gradually diminishes. Hence, these components are discarded in order to reduce the size of the model and to avoid over-fitting. Axes are finally chosen that maximise the retention of variance and correlate dependent and independent variables, or more specifically the covariance of a transformed independent variable with a transformed dependent variable is maximised. This represents a compromise between maximising variance and correlation with biological activity.

2.7.4 Genetic Function Approximation (GFA)

Genetic algorithms are evolutionary operations used to drive the process of computer-aided problem solving. The genetic function approximation is derived from the G/SPLINES algorithm of Rogers (Rogers et al. 1991). An individual is represented as a linear string of basis functions, and using the information in the string it is possible to reconstruct QSARs via OLSR to generate regression coefficients.

The group of initial models are generated by randomly selecting a number of features from a training set, building basis functions from these features utilising the user-specified basis function type and then constructing models from random sequences of these basis functions. These models are then rated using a modified form of Friedman's lack of fit (LOF) function (Equation 2.7.4.1)

\[
\text{LOF} = \frac{\text{LSE}}{1 - ((d + 1)c)/M}^2 \quad \text{(Equation 2.7.4.1)}
\]

Where LSE is the ordinary least square error, M is the number of components in an analysis, d is a smoothing parameter (with a default value of 1) and c is the number of basis functions in the model. LOF is sensitive if there is a small number of compounds and a large number of variables. Having rated the initial population of models the best two (those with the lowest LOF values) are selected as parents. These parents are then subject to genetic operators – the first of which is a crossover. Each parent is randomly cut into two sections and one section from each joined with one section from another. These parents are then mutated (i.e. a randomly selected basis
function is added or deleted) and the LOF of the new models ascertained. This evolutionary process ends when the average LOF score for a generation of models no longer signifies an improvement. Hence, one advantage of GFA is that it builds several different types of models. Another advantage is that GFA is a stochastic optimisation technique (as opposed to the heuristic OLSR method). As such it is able to effectively sample variable space and escape local minima in LOF.

2.7.5. Genetic Partial Least Squares (G/PLS)

G/PLS (Dunn and Rogers 1996) combines some features of GFA with PLS. Each generation of models has PLS applied to it, as opposed to OLSR, and so more terms can be incorporated into the final equation without running the risk of seriously over-fitting the data. One outstanding advantage of the G/PLS method is the reduction of noise. PLS components are extracted from 'x' and 'y' along the longest axes and optimally correlated. Therefore, any systematic variation in 'x' that is not correlated with 'y' is noise. Since the GFA component of this technique effectively filters out this noise the quality of the PLS improves.

2.7.6 Validation Methods

When a regression equation is obtained the reliability and significance is determined. Several procedures are available to achieve this, and they are used to check that the size of the model is appropriate for the quantity of data available as well as providing an indication of how well the model will be able to predict activity for new ligands.

The initial regression model has a correlation coefficient associated with it ($r^2$). For a perfect correlation between observed and predicted data $r^2 = 1$. The closer the $r^2$ value is to 0 the poorer the correlation. However, a high $r^2$ value does not necessarily indicate a predictive model. Cross-validation is a process that repeats the regression analysis many times on different subsets of the data. Each molecule, or a subset of molecules, is removed from the regression in turn and a cross-validated value for $r^2$, usually referred to as $q^2$, is computed using the predicted values of the missing molecules. If the value for $q^2$ is equivalent to that achieved for $r^2$ for all of the different models then the model generated by regression analysis can be deemed to be
predictive. However, if the value of \( q^2 \) fluctuates across the models, or is consistently low, it is an indication that the model produced is not predictive. The equation used to calculate \( r^2 \) and \( q^2 \) is given below (Equation 2.7.6.1)

\[
 r^2 \text{ or } q^2 = \frac{\sum (\text{pred}_i - \text{actual}_i)^2}{\sum (\text{actual}_i - \bar{X})^2}
\]  

(Equation 2.7.6.1)

\( N \) = number of molecules in the data set.

\( \text{pred}_i \) = predicted value of the independent variable.

\( \text{actual}_i \) = experimental value of the independent variable.

\( \bar{X} \) = mean value of the independent variables in the dataset.

### 2.7.7 Spearman’s Ranking

Spearman’s ranking is a non-parametric statistical method. The Spearman’s ranking correlation coefficient \( (r_s) \) is a measure of association that requires both variables being studied to be measured, at least in an ordinal scale, so that the objects or individuals under study may be ranked in two ordered series. The correlation coefficient can take a value between 0 and 1, the former indicating no association between the rankings of the variables and the latter representing a total agreement between the lists. The Spearman’s rank was utilised in Chapters 6, 7 and 8 of this thesis. It provides an indication of whether or not a general trend is adhered to, even in the absence of accurate absolute predictions, and was able to give insight into the strengths and weaknesses of scoring functions, QSAR studies and pharmacophores.

The equation for the calculation of the Spearman ranking correlation coefficient is as follows (Equation 2.7.1):

\[
r_s = 1 - \frac{6 \sum d_i^2}{N(N^3 - N)}
\]

(Equation 2.7.1)

Where \( d \) is the difference between the rankings in the first and second lists, and \( N \) is the number of subjects.
The results of tied rankings in the ‘x’ and ‘y’ variables (ranking lists 1 and 2 respectively) means that it is necessary to correct the sum of the squares, taking ties into account. This correction factor, T, is given by the following equation (Equation 2.7.2)

\[ T = t^3 - t/12 \] (Equation 2.7.2)

Where t is the number of tied observations at a given rank.

Taking this into account, the Spearman correlation coefficient can be calculated using the following (Equation 2.7.3)

\[ r_s = \frac{\sum x^2 + \sum y^2 - \sum d^2/2 \sqrt{(\sum x^2 \cdot \sum y^2)}}{n(n^2 - 1)/12} \] (Equation 2.7.3)

Where \( \sum x^2 = (N^3 - N/12) - \sum T_x \)
\( \sum y^2 = (N^3 - N/12) - \sum T_y \)

N is the number of subjects under investigation.

2.8. Measurement of Biological Activities for Ligands of CYP 2D6

The biological activities for the compounds in datasets 1 and 2 used in Chapters 6, 7 and 8 were measured at AstraZeneca for human CYP 2D6 (Moody et al 1999). The physiological substrate dextromethorphan primarily undergoes O-demethylation by 2D6 (Figure 2.8.1) to give dextrorphan. Evidence also exists for a minor N-demethylated product, but for the purposes of this investigation this metabolic pathway is not considered. In the assay experiments dextromethorphan is radioactively labelled, leading to the generation of labelled dextrorphan. Thus it is possible to deduce the amount of dextrorphan produced by 2D6.

Various concentrations of the ligands whose biological activities are to be determined (Datasets 1 and 2, Appendix 6.1 and 6.2) are then used to compete with the radioactively labelled dextromethorphan. It is therefore possible to calculate the percentage inhibition of the ligands versus the radioactively labelled
dextromethorphan at different concentrations, by measuring the amount of
dextrorphan produced and comparing it to the amount expected. The greater the
amount of dextrorphan synthesised the tighter the relative binding of
dextromethorphan to 2D6 compared with the respective ligand. The values obtained
for the different concentrations enable us to determine the IC\(_{50}\) value (a measure of
ccentration of the inhibitor required to reduce the binding affinity of a ligand by
half) and subsequently pIC\(_{50}\) values (where pIC\(_{50}\) = -log IC\(_{50}\)), which can then be used
in computational studies to validate the performances of scoring functions, determine
useful pharmacophores and help predict the biological activities of unknown
compounds in QSAR studies.

**Figure 2.8.1** P450 2D6 mediated demethylation of dextromethorphan to yield
dextrorphan.
Chapter 3

Comparative Modelling of Cytochrome P450 2C5
3.0 Preface

This chapter deals with the comparative modelling of mammalian cytochrome P450 2C5, the X-ray crystal structure of which has already been elucidated. By comparing the models to the X-Ray crystal structure it will be possible to critically evaluate the different approaches employed, with a view to using the most successful techniques to model an enzyme of unknown 3D structure.

3.1. Introduction – Comparative Modelling of Cytochrome P450 2C5

Comparative modelling is a predictive technique, involving the generation of the 3D structure of a protein – referred to hereafter as the ‘target’ protein – for which the amino acid sequence, but not the 3D structure, is known. The technique employs (i) the amino acid sequence of the target protein and (ii) 3D structural information from proteins of similar amino acid sequence to the target and of known 3D structure – referred to hereafter as the ‘templates’ – to generate a 3D structure for the target protein. The process of comparative modelling is illustrated in Figure 3.1.1. As indicated by the flow chart, comparative modelling is an iterative process with validation possible at almost every level. This is used to ensure that the resulting model gives as true and accurate a representation of the 3D structure of the target protein as is possible.

To illustrate how we tackle the comparative modelling problem, and validate the methods that are used in practise, a ‘worked example’ is given. This assumes that the 3D structure of rabbit P450 2C5 (Williams et al 2000) PDB code 1dt6, is not known, and produces 3D models of this. The models are validated using standard procedures and then compared with the ‘real answer’, the crystal structure.

Rabbit cytochrome P450 2C5 is known to selectively hydroxylate the 21-methyl group on the 17β side chain of progesterone (von Wachenfeldt et al 1997). It is, as yet, the only mammalian cytochrome P450 deposited in the protein data bank (Berman et al 2000), and is resolved to 3.0Å
3.2 Experimental

3.2.1 Identification of 3D structural templates

It has been shown that proteins that have similar amino acid sequences have a tendency to adopt similar 3D structures (Lesk and Chothia 1986). For this reason comparative modelling methods use structural templates that have the highest sequence homology with the target protein. Homologous proteins were found by scanning the sequence of the target (2C5) against the sequences of those 3D structures deposited in the Protein Data Bank (PDB) (www.rcsb.org/pdb/) using the algorithm
PSI-BLAST (Altschul et al 1997) (www.ncbi.nlm.nih.gov/BLAST/). Each of the templates returned was assessed visually i.e. they were examined for gross structural differences in the polypeptide backbone when compared to the other templates, using a molecular graphics package. If any template was significantly structurally dissimilar to the majority it was discarded. The scores returned by PSI-BLAST were also used to assess the suitability of the template. Appropriate templates were then carried through to the next stage.

3.2.2. Alignment of Templates at Amino Acid Level.

Multiple templates were identified as a result of the PSI-BLAST search. Several models of P450s are present in the literature (Lewis and Lake 1996, Lewis and Lake 1985, Sklarz et al 1994, Lewis and Lake 1997, Lewis et al 1999, Lewis et al 1998, Lewis et al 1997, Lewis 1999, Tan et al 1997, Dai et al 1998, Chang et al 1997, Payne et al 1999a, Payne et al 1999b, Modi et al 1996, de Groot et al 1996, Sklarz et al 1997, Chang and Loew 1999). Several of these models are generated from a single template, whilst others are derived from multiple templates. Both methods are equally valid, and in the absence of a definitive, fail-safe method for the comparative modelling of an unknown structure four different approaches were adopted. These were as follows:

1) Single template using the template with highest homology to the target protein i.e. Class II type P450BM-3.

2) Multiple (1bu7, 1cpt, 1phc, 1eup, 1rom) templates using FSSP (Holm and Sander 1996) to structurally align the templates.

3) Multiple (1bu7, 1cpt, 1phc, 1eup, 1rom) templates using MALIGN3D (Sali and Blundell 1993) to structurally align the templates.

4) Alignment 3 modified by aligning the secondary structure of the templates with the predicted secondary structure of P450 2C5.
Each approach required an amino acid sequence alignment on which to base the models. Three of the four approaches utilise more than one template (Approaches 2-4). It was therefore necessary to align the template sequences with respect to each other for these approaches. This was achieved using, ClustalX (Thompson et al 1994) (http://www-igbmc.u-strasbg.fr/BioInfo/ClustalX/Top.html). The resulting alignment of the template sequences is kept fixed (treated in CLUSTALX as a "profile") and the query sequence appended to this alignment prior to building the models. However, aligning the 3D structures and deriving the sequence alignment from this is a more accurate method for obtaining a sequence alignment, particularly in cases such as the P450s where sequence homology is relatively low.

3.2.3 3D Structural Alignment of Templates

Two methods for obtaining structure-based sequence alignments of the templates were investigated: the MALIGN3D routine in the program Modeller (Sali and Blundell 1993), and FSSP (Fold classification based on Structure-Structure alignment of Proteins)(Holm and Sander 1996).

3.2.3.1 MALIGN 3D

The MALIGN3D routine in Modeller uses an initial sequence-based alignment of the templates generated using CLUSTALX (Thompson 1994) as its starting point. This alignment is used to superpose the 3D structures and thus define a new alignment. This process is repeated iteratively and a sequence alignment based on the rigid body superposition of all the templates is produced.

3.2.3.2 FSSP

The FSSP database (Holm and Sander 1996) (http://www.ebi.ac.uk/dali/fssp/) is based on exhaustive all-against-all 3D structure comparison of protein structures currently in the PDB. All alignments are based on comparing the 3D structures of the proteins using Cα distance matrices, a method that compares local similarity, rather than
global similarity as with MALIGN3D (Section 3.2.3.1). Unlike MALIGN3D, FSSP does not require an initial amino acid sequence alignment as a starting point.

3.2.4 Alignment of Query Amino Acid Sequence to Template Profile

Once the templates had been aligned, the amino acid sequence of the target protein was appended. This was achieved by aligning the template profile against a sequence alignment of the P450 2C subfamily.

A P450 2C mammalian subfamily sequence alignment was obtained by (i) identifying members of the 2C subfamily using a PSI-BLAST search with the 2C5 sequence, and (ii) aligning (10 of) the resulting sequences using CLUSTALX.

Obtaining a subfamily alignment serves two purposes. Firstly, it serves to highlight any regions across the 2C subfamily that are highly conserved. Secondly, an appended subfamily alignment, as opposed to appending a single amino acid sequence, reduces the likelihood of misalignment of conserved residues. The resultant subfamily alignment was appended onto the structural alignment of the templates using the profile alignment option in ClustalX. This alignment is subsequently checked to ensure that:

1) All of the secondary structural elements (i.e. α-helices and β-strands) have no (or a minimum number of) insertions and/or deletions contained within them.

2) Any residues identified experimentally as being crucial to catalytic activity are consistently conserved across all of the templates (e.g. the proximal cysteinate ligand that occupies the fifth co-ordination site of the heme and is responsible for stabilisation of the high-energy oxy-ferryl species generated during the catalytic cycle of the P450 (Lewis 1986).

3) Any residues that have been shown to play an important role in the binding of substrates in the active site are conserved across the templates (e.g. Asp 301 in P450$_{2D6}$ (Koymans et al 1992, Modi et al 1996, Li et al 1997, Smith et al 1998).
The interactive molecular graphics package Insight II (Accelrys Inc, San Diego, USA) was used to examine the alignments.

Alignment of the secondary structure of the templates against the predicted secondary structure of the target protein was used to modify the amino acid sequence alignments using the DSSP database (Database of Secondary Structure in Proteins) (Kabsch and Sander 1983) (http://www.sander.ebi.ac.uk/dssp/). Secondary structure prediction was performed using PSIPRED (Jones 1999) (http://insulin.brunel.ac.uk/psipred/). The templates' secondary structure elements (defined for those residues where a particular secondary structure type is predicted with a reliability of 6 or more) were then aligned with those predicted for the target protein.

3.2.5 Model building – comparative modelling

Once the sequence alignments were produced, 3D models were generated using comparative modelling. Modeller (Sali and Blundell 1993) was used to produce the models (see Appendix 3 for scripts). Ten models were generated from each of the four separate alignments. The highest scoring models from each of the approaches, as determined by the intrinsic Modeller scoring function, were then analysed with respect to their stereochemistry, amino acid environment and Cα RMSD.

3.2.6 Model Analysis

In general (although not in this case) it is not possible to compare a theoretical comparative model to a crystal structure in order to validate it. As such, a number of independent yet complementary techniques were employed to assess the quality of the stereochemistry and amino acid environment for each of the best models from the four separate alignments. As there is a crystal structure available for the target protein it is possible to incorporate a third check, Cα RMSD, which gives a very rough indication of how well conformational space has been sampled during the model building process.
3.2.6.1 Stereochemical Analysis

A wide range of stereochemical checks are available through the Biotech Validation Suite for Protein Structures (http://biotech.embl-ebi.ac.uk:8400/). These assess how well the stereochemistry of the residues in a given protein structure compare with those observed in well-refined, high-resolution crystal structures. PROCHECK (Laskowski et al 1993) was used to assess the “quality” of the conformation of the polypeptide backbone and sidechains. For these checks, the results returned for the templates used act as a baseline against which to compare the models.

3.2.6.2 Amino Acid Environment

The compatibility between the amino acid sequence and the environment of the amino acid sidechains in the model is another validation criterion used. Verify3D (Luthy et al 1992) (http://www.doe-mbi.ucla.edu/Services/Verify_3D/) and Errat (Colovos and Yeates 1993) (http://www.doe-mbi.ucla.edu/Services/ERRAT/) were used for this. As with the stereochemical checks, we use the results returned for the templates as a baseline against which to compare the models.

3.2.6.3 Mainchain RMSD

It has been shown that the RMSD between main chain atoms in the template with the highest homology to the target and the model is a good method of validation (Venclovas et al 1997). This builds on earlier work (Chothia and Lesk 1986), which showed that for two proteins there was a relationship between the percentage sequence identity and the RMSD of the main chain atoms in homologous regions. The RMSD of the main chain atoms of residues that are correctly aligned to the most homologous template were calculated using the Structure Check option in the Biopolymer module of Insight II. The template with the highest PSI-BLAST score was deemed the most homologous.
3.2.6.4 Ca RMSD

Given that the 3D structure of P450 2C5 is known, an additional validation method which is generally not available has also been employed, the root mean square deviation of the Ca atoms (Ca RMSD) between the model and the crystal structure. This gives a measure of how close the model is to the target and was calculated using the Structure Check option in the Biopolymer module of Insight II.

3.3 Results

3.3.1 Identification of 3D Structural templates

The results from the PSI-BLAST search against the PDB (the search converged after 3 iterations), using the sequence of P450 2C5 as a target, identified the amino acid sequences of seven P450s isolated from variety of micro-organisms as possible templates. There were six Class I P450s - P450 51, P450terp, P450cam, P450eryF, P450NOR, P450 119 and one Class II P450s - P450BM-3. The sequences were ranked in order of PSI-BLAST score (Table 3.3.1.1.). The score is calculated for each "hit" sequence aligned to the sequence of P450 2C5 by looking at aligned position and gaps (Altschul et al 1997).

The fourth column in the table lists the percentage sequence identities of the structures to the target. None of the structures have a sequence identity greater than 25%. From CASP2 (Second Meeting on the Critical Assessment of Methods for Protein Structure Prediction) it was concluded that below 25% sequence identity the alignments were so poor that the models submitted were far from reality, even if the overall fold was accurate (Martin et al 1997). This evidence suggests that it will be very difficult to generate an accurate model of P450 2C5 using templates with sequence identities less than 25%.
Table 3.3.1.1
The P450 structures recovered by a PSI-BLAST search using the sequence of P450 2C5 (1dt6).

<table>
<thead>
<tr>
<th>Protein</th>
<th>PSI BLAST Score&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Sequence Identity&lt;sup&gt;b&lt;/sup&gt;</th>
<th>% Sequence Identity</th>
<th>Chain Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mammalian Cytochrome P450 2C5&lt;sup&gt;c&lt;/sup&gt;</td>
<td>692</td>
<td>473/473</td>
<td>100</td>
<td>473</td>
</tr>
<tr>
<td>Cytochrome P450 51</td>
<td>533</td>
<td>86/456</td>
<td>18</td>
<td>455</td>
</tr>
<tr>
<td>Cytochrome P450BM-3</td>
<td>480</td>
<td>100/464</td>
<td>21</td>
<td>455</td>
</tr>
<tr>
<td>Cytochrome P450terp</td>
<td>275</td>
<td>74/336</td>
<td>22</td>
<td>412</td>
</tr>
<tr>
<td>Cytochrome P450cam</td>
<td>245</td>
<td>48/276</td>
<td>17</td>
<td>414</td>
</tr>
<tr>
<td>Cytochrome P450eryF</td>
<td>225</td>
<td>53/237</td>
<td>22</td>
<td>403</td>
</tr>
<tr>
<td>Cytochrome P450NOR</td>
<td>202</td>
<td>47/257</td>
<td>18</td>
<td>403</td>
</tr>
<tr>
<td>Cytochrome P450 119</td>
<td>185</td>
<td>59/360</td>
<td>16</td>
<td>368</td>
</tr>
</tbody>
</table>

<sup>a</sup>The PSI-BLAST score for an alignment is calculated by summing the scores for each aligned position and the scores for gaps. (Altschul <i>et al.</i>, 1997)

<sup>b</sup>(The number of identical residues) / (the length of sequence fragment identified by PSI-BLAST)

<sup>c</sup>The sequence of the target is included for information; this was not used in the modelling.
A visual examination of P450 51 revealed it to be structurally very different in the I-helix region from the other P450s available in the public domain (Figure 3.3.1.1). Therefore P450 51 was rejected as a template for modelling despite its high PSI-BLAST score. P450 119 was also rejected because it is 105 residues shorter than P450 2C5. Suitable structures were found in the PDB for the remaining five sequences. Only wild-type proteins were chosen, and where more than one structure was available the structure with the highest resolution was picked. Thus, the following template structures - P450BM-3 (PDB code 1bu7 (Serioukova et al 1999)), P450terp (1cpt (Hasemann et al 1994)), P450eryF (1eup (Cupp-Vickery et al 2000)), P450cam (1phc (Poulos et al 1986)), and P450NOR (1rom (Park et al 1997)) were identified.

Figure 3.3.1.1 The Cα ribbon traces of the I-helices of P450 51 (red), P450 Terp (green) and P450 BM-3 (magenta) illustrating the significant structural differences in P450 51.

3.3.2 Alignment of Templates at Amino Acid and Structural Level

The amino acid sequence alignment used to generate models for Method 1, the single template approach, is given in Figure 3.3.2.1.
The amino acid sequence alignment used to generate comparative models for Method 2, the multiple template alignment derived from FSSP (Holm and Sander 1996), is illustrated in Figure 3.3.2.2. Method 3 the MALIGN3D based approach (Figure 3.3.2.3) and Method 4 the modified MALIGN 3D approach (Figure 3.3.2.4) are also illustrated. The ‘conserved gaps’ appearing throughout the four alignments arise due to insertions and deletions in the sequences of the 2C subfamily profile, which are not shown in the alignment.
residues predicted (using PSIPRED) to be in secondary structure elements, have been shaded.

Figure 3.3.2.2 Method 2 FSSP-based multiple template alignment. Template residues in secondary structure elements in the crystal structure (DSSP; PDB code for BM-3 is 1bu7, terp is 1cpt, NOR is 1rom, cam is 1phe and eryF is 1eup), and P450 2C5 residues predicted (using PSIPRED) to be in secondary structure elements, have been shaded.
Figure 3.3.2.3 Method 3 Malign3D multiple template alignment. Template residues in secondary structure elements, and P450 2C5 residues predicted (using PSIPRED) to be in secondary structure elements, have been shaded.
Figure 3.3.2.4 Method 4 Manually modified Malign3D multiple template alignment. Template residues in secondary structure elements, and P450 2C5 residues predicted (using PSIPRED) to be in secondary structure elements, have been shaded. Modifications include the closing of insertions in secondary structure elements and a partial alignment of the B' helices.
3.3.3 Model Analysis

The lowest energy models from each of the alignments were analysed using Ramachandran Plots (Morris et al. 1992), Errat (Colovos and Yeates 1993), Verify 3D (Luthy et al. 1992). The results are presented in Table 3.3.3.1.

Table 3.3.3.1 Results of validation studies performed on the lowest energy models produced from the four different sequence alignments.

<table>
<thead>
<tr>
<th>Method*</th>
<th>Ramachandran Plotb (%)</th>
<th>Verify 3D (total score)</th>
<th>Errat (%)</th>
<th>Main Chain RMSDc (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Single template</td>
<td>87</td>
<td>102</td>
<td>87</td>
<td>0.42</td>
</tr>
<tr>
<td>2) Multiple template, FSSP</td>
<td>81</td>
<td>130</td>
<td>66</td>
<td>1.23</td>
</tr>
<tr>
<td>3) Multiple template, MALIGN3D</td>
<td>74</td>
<td>136</td>
<td>77</td>
<td>1.22</td>
</tr>
<tr>
<td>4) Multiple template, modified, MALIGN3D</td>
<td>78</td>
<td>135</td>
<td>81</td>
<td>1.29</td>
</tr>
</tbody>
</table>

*See section 3.2.2 for more information on the alignments used.

bPercentage of residues with \( \phi, \psi \) conformation in the “most favoured” regions of the Ramachandran plot (Morris et al. 1992).

cThe RMSD between aligned main chain atoms in the model and those in the most homologous template, P450 BM-3.

The templates used to produce the models were also subject to an identical analysis, with the exception of determining the mainchain RMSD. This provided a baseline against which the models could be compared (Table 3.3.3.2)
Table 3.3.3.2  Validation results for the target 2C5 structure and the five crystal structures used as templates for the modelling studies.

<table>
<thead>
<tr>
<th>Crystal Structure (PDB code)</th>
<th>Resolution (Å)</th>
<th>Ramachandran Plot (^a) (%)</th>
<th>Verify 3D(^b)</th>
<th>Errat(^c) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P450 2C5 (1dt6)(^d)</td>
<td>3.0</td>
<td>71</td>
<td>182</td>
<td>94</td>
</tr>
<tr>
<td>P450 BM-3 (lbu7)</td>
<td>1.7</td>
<td>91</td>
<td>221</td>
<td>95</td>
</tr>
<tr>
<td>P450 terp (1cpt)</td>
<td>2.3</td>
<td>89</td>
<td>188</td>
<td>95</td>
</tr>
<tr>
<td>P450 cam (lphc)</td>
<td>1.6</td>
<td>90</td>
<td>222</td>
<td>97</td>
</tr>
<tr>
<td>P450 eryF (1eup)</td>
<td>2.1</td>
<td>91</td>
<td>184</td>
<td>96</td>
</tr>
<tr>
<td>P450 NOR (1rom)</td>
<td>2.0</td>
<td>91</td>
<td>176</td>
<td>96</td>
</tr>
</tbody>
</table>

\(^a\)Percentage of residues with \(\phi,\psi\) conformation in the “most favoured” regions of the Ramachandran plot (Morris et al., 1992).

\(^b\)The total Verify3D score summed over all residues. For a protein the same size as P450 2C5 a score of 205 is expected and a score less than 92 would indicate an incorrect structure.

\(^c\)The percentage of residues in a ‘normal’ non-bonded environment.

\(^d\)This is the target structure, and has been included for comparison purposes.

3.4 Discussion

From Table 3.3.3.2 it is apparent that the five template structures have Errat scores greater than or equal to 95% (i.e. ≥ 95% of the residues are in ‘normal’ environments)
and Verify3D scores greater than 170. Even the P450 2C5 structure, which is relatively poorly resolved (3Å), has an Errat score of 94% and a Verify3D score of 182 (normally we would not be able to compare our model with the target because the target structure would be unknown, nevertheless such a comparison is useful in the current case). The low resolution of P450 2C5 is reflected in its poor stereochemistry, only 71% of the residues lie in the “most favoured” regions of the Ramachandran plot, compared with an average of 90% for the template structures. As such, the comparable values that are obtained for the models from each of the alignments suggest that the modelling has been a relative success.

The lowest energy model produced using Method 1, a single template alignment, (Figure 3.3.2.1) has a good Ramachandran score of 87% (better than the P450 2C5 crystal structure), an Errat score of 85.6% and a Verify3D score of 102 (Table 3.3.3.1). With Verify3D, a good quality crystal structure of a protein of this size is expected to score at least 205, and a score lower than 92 would indicate an incorrectly folded structure. The lowest energy single template model has a score of 103, below that expected for a crystal structure of the same size.

In comparison, the Verify3D result (130) for the lowest energy model produced using Method 2, a multiple templates FSSP based alignment (Figure 3.3.2.2), suggests the side chain environment in the model has improved. The number of residues with good \( \phi, \psi \) stereochemistry is comparable with Method 1. However, the Errat validation check does not perform as well as for Method 1. The number of resides in a ‘normal’ non-bonded environment has fallen from 86% to 66%.

To investigate these results further, the two sets of ten models produced by Methods 1 and 2 were analysed visually. For Method 1 the polypeptide backbone of the template (P450 BM-3) and all 10 models produced are very similar (Figure 3.4.1A). The deviations across the models arise mainly due to insertions or deletions in the target sequence with respect to the template sequence (these regions are not defined by the template). Figure 3.4.1B shows the polypeptide backbone of the Method 2 models, which display a greater variability. Those regions displaying less variability in the
Method 2 models correspond to regions conserved structurally across the templates, such as the I-Helix.

Previous work (Chothia and Lesk 1986) established a relationship between the percentage sequence identity of two proteins and the RMSD in the positions of the main chain atoms in the homologous regions of the proteins. The relationship predicts that the higher the sequence identity, the lower the subsequent RMSD. For two proteins with 21% sequence identity an RMSD between the main chain atoms of 1.75Å is predicted.

The RMSD between the main chain atoms in homologous regions of the lowest energy Method 1 (single template) model and the most homologous template (P450 BM-3) is 0.42Å (Table 3.3.3.1), much lower than expected. This result suggests that, in this case, when a single template alignment is use, the amount of conformational space sampled is restricted and as a consequence the models produced are over-fitted to the three-dimensional structure of the template.

In comparison, the RMSD between the lowest energy Method 2 (multiple template) model and the most homologous template, (P450 BM-3), is 1.23Å (Table 3.3.3.1). This is much closer to the expected value. By using more than one template in the amino acid sequence alignment the degree of conformational sampling during the model building process appears to have increased. The result is that the models produced are less like the templates, and possibly more representative of the true structure. For this reason a multiple template alignment is preferable in this case.
Figure 3.4.1 A) Ca trace of the template BM-3 (black; PDB code 1bu7) and the ten models (grey) produced using a single template alignment (Method 1; Figure 3.3.2.1) and B) Ca trace of the most homologous template BM-3 (black; PDB code 1bu7) and the ten models produced using a multiple template alignment (Method 2; Figure 3.3.2.2)
Errors in the sequence alignment are the most likely cause of errors in the models and therefore we have sought to improve the quality of the models by investigating alternative methods of producing multiple template alignments. Given the structural diversity across the templates, an alternative method of structurally aligning them was investigated, using a global, rather than local, alignment method. Malign3D (Sali and Blundell 1993) was used to derive this alignment. Method 3 is shown in Figure 3.3.2.3. In contrast to Method 2 (Figure 3.3.2.2) there are many more insertions in secondary structure elements. Results of the validation checks on the lowest energy model for Method 3 are given in Table 3.3.3.1. The model has a similar Verify3D score to the Method 2 model, suggesting it is correctly folded and the Errat score has improved significantly to 77%, indicative of the model having a better non-bonded environment. However, unlike the Method 2 model a poor Ramachandran score was obtained, suggesting that there are problems with the stereochemistry. Although a Ramachandran score of 74% is disappointing it is better than the 2C5 crystal structure, indicating that the Modeller program is not necessarily limited by poor resolution structures.

The per-residue Errat plot for the Method 3 model is shown in Figure 3.4.2. The plot highlights problems with the non-bonded environment of the model in a number of regions. Residues 160-170 and 200-210 were predicted by PSIPRED to be helical and correspond to the F and G-helices of the templates, respectively. Residues 60-80 were not predicted to be helical even though they correspond to the B'-helix of the templates. The other problem regions identified by the plot occur in loop regions of the model. From Figure 3.3.2.3 it can be seen that the A, B’, F and G regions of alignment contain a number of insertions.

Figure 3.3.2.4 shows Method 4, a manually modified alignment based on Method 3 (Figure 3.3.2.3). The major modifications are the removal of insertions in secondary structure elements. Ten P450 2C5 models were produced using this modified alignment. The results of the validation studies performed on the best energy model for Method 4 are presented in Table 3.3.3.1 and show an improvement in comparison to the results achieved for Method 3, the original Malign3D alignment. The per-residue Errat plot for this model is shown in Figure 3.4.3. Alterations made to the alignment have removed the problems (as identified by Errat) from the predicted
secondary structure elements, the A, B', F and G-helices. Problems still remain in the loop regions and one new problem has been introduced in first predicted β-strand, but despite this the number of residues in a good non-bonded environments has risen from 77% to 81%.

**Figure 3.4.2** Errat plot for the lowest energy model produced using the Malign3D based alignment (Figure 3.3.2.3). The black columns identify problem areas; they occur at the N-terminus, B', F, G, and H helices and in some loops.

**Figure 3.4.3** Errat plot for the best energy model produced using the modified Malign3D alignment (Figure 3.3.2.4). The black columns identify problem areas; these occur in loop regions and β-strands at the C-terminus and N-terminus.
The results of the Errat, Verify3D, and Ramachandran checks performed on the models produced using the four alignments described in section 3.2.2 suggest that the lowest energy model produced using Method 4 (Figure 3.3.2.4) is the best model. In an attempt to validate this result the lowest energy model produced for each alignment has been compared with the 2C5 crystal structure (Table 3.4.1). Obviously this is an academic exercise because no crystal structure exists in a “normal” modelling problem. To make the comparison between the models and the crystal structure the Ca RMSD is measured. Two measurements are taken, the first across the whole structure and the second across just the secondary structure elements. Table 3.3.3.1 shows that as the validation check results improve the Ca RMSD decreases. In other words the validation checks were able to identify the model most like the crystal structure. The Ca RMSD for all residues from the best model (Method 4 model) to the corresponding atoms in the crystal structure is 4.7 Å, falling to just 3.2Å when only secondary structural elements are considered. In the summary of the findings from CASP (Martin et al 1997), three targets had sequence identities less than 25%; for these targets the all Ca RMSDs for the models submitted ranged from 6Å to 18Å. The problems associated with attempting to model proteins with low sequence identities to the closest template were also highlighted in the CASP3 studies (Jones and Kleywegt 1999). The Ca RMSDs between model and target for the predictions from CASP3 where 100% of the target’s residues had been modelled are shown in Figure 3.4.4 (extracted from http://predictioncenter.llnl.gov/casp3/Casp3.html). The results for the 2C5 models produced here have also been included. The plot illustrates that, with a sequence identity between the target and the most homologous template of less than 25%, the models are as good as can be expected.
Table 3.4.1

Comparison of lowest energy models produced from the four different sequence alignments with 2C5 crystal structure.

<table>
<thead>
<tr>
<th>Alignment*</th>
<th>All Ca Atom RMSD&lt;sup&gt;b&lt;/sup&gt; (Å)</th>
<th>2&lt;sup&gt;e&lt;/sup&gt; Structure Ca Atom RMSD&lt;sup&gt;c&lt;/sup&gt; (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Single template</td>
<td>6.3</td>
<td>4.8</td>
</tr>
<tr>
<td>2) Multiple template, FSSP</td>
<td>5.6</td>
<td>3.5</td>
</tr>
<tr>
<td>3) Multiple template, MALIGN3D</td>
<td>5.1</td>
<td>3.5</td>
</tr>
<tr>
<td>4) Multiple template, modified, MALIGN3D</td>
<td>4.7</td>
<td>3.2</td>
</tr>
</tbody>
</table>

<sup>a</sup>See section 3.2.2 for more information on the alignments used.

<sup>b</sup>The RMSD between all the Ca atoms in the model and those in the 2C5 crystal structure.

<sup>c</sup>The RMSD between all the Ca atoms in secondary structural elements in the model and the corresponding residues in 2C5 crystal structure.

As further validation, the best model has been structurally aligned against the crystal structure; 54% of the residues have been aligned correctly (Figure 3.4.5). This result is also comparable with the findings from the CASP2 study (Martin et al 1999). In terms of the secondary structure elements only four out of a total of 14 α-helices and two out of six β-strands are incorrectly aligned. Although incorrect, the positions of the E, F, and G helices, β1-4 and β3-1 have slipped by just one residue. More serious alignment problems occur at the N-terminus and in the loop regions.
Figure 3.4.4 Plot of the all Ca atom RMSD versus % sequence identity between the closest template to the target and the target. The results presented are for the CASP3 predictions (Jones and Kleywegt, 1999) where 100% of the target residues had been modeled. Also shown are 2C5 models reported here.
Figure 3.4.5 Structure-based alignment of the P450 2C5 best model against the crystal structure (PDB code 1dt6). Those regions correctly aligned are marked with a star, corresponding to 54% of the amino acids. The observed and predicted secondary structure elements are shaded grey and the substrate recognition sites (SRS) are labelled.

| 2C5 X-ray | PPGPPPFPF- -IIGNILQI-DA-K- -D1DKSLT-KPHEBCYHF- -PPPL-GMDTMMV | 31-1 | 31-2 |
| 2C5 pred. | -PPG -TPPPF-IIGNL-Q1SAXDENTS-GET- -PPPLGM-EDRTM | | |
| 2C5 X-ray | RYDVAEVAVDELGE- -EF-AGRG- -SVPILEKTSGKLGSFAIA-KL-KEIARKGLFM | | |
| 2C5 pred. | RYDVAEVAVDLGE- -EFAGRG- -SVPILECTSGKLGSFAIA-KL-KEIARKGLFM | | |
| 2C5 X-ray | TE-R-NPOGMR- -S-I-DTQCCYKELT- -NASPCDEPTT-LGACFMHTICVF | | |
| 2C5 pred. | TE-R-NPOGMR- -S-I-DTQCCYKELT- -NASPCDEPTT-LGACFMHTICVF | | |
| 2C5 X-ray | TRNPGC- -GKSIIEDRRGQACZPHEL- -KTNASPCDPTF | | |
| 2C5 pred. | TRNPGC- -GKSIIEDRRGQACZPHEL- -KTNASPCDPTF | | |
| 2C5 X-ray | HRFTFY- -KHFF- -LK- LMKLHENVELL- -TP- - - -DCWTPKPLHNA | | |
| 2C5 pred. | HRFTFY- -KHFF- -LK- LMKLHENVELL- -TP- - - -DCWTPKPLHNA | | |
| 2C5 X-ray | YSNFTDLKHKLL- -DCNPDICYLDEMQRNL- -R- -PTLLTVIANLDK | | |
| 2C5 pred. | YSNFTDLKHKLL- -DCNPDICYLDEMQRNL- -R- -PTLLTVIANLDK | | |
| 2C5 X-ray | PAGLTTSTLPLSLLLE- -PAVAKVERQTFIOMRSPSCQK- -KOMPPTDAVI | | |
| 2C5 pred. | PAGLTTSTLPLSLLLE- -PAVAKVERQTFIOMRSPSCQK- -KOMPPTDAVI | | |
| 2C5 X-ray | LQFTDLKLPTNL- -TRVVFRR-N-YP1KCTG- -LCLQDVMLDEK- -AFPNPVFDPO | | |
| 2C5 pred. | LQFTDLKLPTNL- -TRVVFRR-N-YP1KCTG- -LCLQDVMLDEK- -AFPNPVFDPO | | |
| 2C5 X-ray | MPLDSQNPXXS- -D- -YMPFSSAGAKRCV- -GLAMMRPLPASI1LQH- -GVLEV | | |
| 2C5 pred. | MPLDSQNPXXS- -D- -YMPFSSAGAKRCV- -GLAMMRPLPASI1LQH- -GVLEV | | |
| 2C5 X-ray | EPKDLDITAYV- -HGF- -SVPPSYGV- -SFPTT | | |
| 2C5 pred. | EPKDLDITAYV- -HGF- -SVPPSYGV- -SFPTT | | |

Mutagenesis studies in type II P450s identified key substrate binding residues (Gotoh et al 1992, von Wachenfeldt 1995), and together these residues form substrate recognition sites (SRSs) (Gotoh et al 1992). To identify key residues in the crystal structure of P450 2C5, a small molecule, progesterone, was docked into the binding site (Williams et al 2000). The residues that contributed to the binding of the docked molecule (L102, L113, F114, V205, L208, D290, A294, T298, L358, L359, L363, F473, V474) were in the predicted SRSs and their positions are shown in Figure 3.4.6. SRS-4 contains the residues D290, A294 and T298 (I-helix). The I-helix is correctly
aligned in our models (Figure 3.4.5) and is conserved across the five templates; hence it is no surprise that the residues in the best model are very close to the crystal structure (Figure 3.4.6). SRS-5 is predicted to be the loop between the K-helix and β1-4 and contains the residues L358, L359, and L363. The alignment is correct for L358 and L359 but has slipped by one residue for L363. Figure 3.4.5 shows that L358 and L359 are positioned close to the corresponding residues in the crystal structure but L363 is not predicted as accurately. SRS6 contains the residues F473, V474 in a loop towards the C-terminus. V474 is aligned correctly but F473 has slipped by one residue, despite this slipping both residues are positioned close to the corresponding residues in the crystal structure. Important binding residues, whose positions are not predicted well, include I102, L113 and F114 in the B-C loop and V205 and I208 in the F-helix. This result is not surprising, as the relative positions of SRS-1 and SRS-2 are not well conserved across the five templates and 2C5; for example, SRS-1, and SRS-2 have Ca RMSDs of 6.0 and 6.4Å respectively between 2C5 and BM-3 (Williams et al 2000).

![Figure 3.4.6 The P4502C5 substrate recognition sites for the crystal structure (dark grey; PDB code 1dt6) and the lowest energy model (light grey).](image-url)
3.5 Conclusions

Comparative models of the P450s are useful as, in the absence of experimentally determined structures, they provide testable hypotheses. Here an iterative cycle of model building and validation checks has been used to produce high quality models of 2C5. These studies on the P450s suggest that a single template model does not sample conformational space well and that a multiple template-based model produces better results. For the 2C5 study reported here it was found that a multiple template model, based on a structural alignment produced using Malign3D, produced the best model. Other studies have identified FSSP as the preferred method of structurally aligning templates (see Chapter 4) and therefore a combined approach is recommended. It is important that a range of independent checks, both stereochemical and environmental, are employed. Values returned for the templates are used as a baseline against which to compare the values returned for the models. If experimental data is available for the P450 being modelled, this should be used as an additional validation step. Based on the sequence identity between templates and target, the best model is as close to the crystal structure of 2C5 as can be expected. The model also predicts correctly the positions of a significant number of residues in the substrate recognition sites, reiterating the predictive value of comparative models of mammalian P450s.
Chapter 4

Comparative Modelling of Human Cytochrome P450 2D6
incorporating the Mammalian P450 2C5 Structure
4.0 Preface

Having validated several different comparative modelling methods (Chapter 3) and concluded that the best approach is a combined approach, utilising amino acid alignments that have been derived in different manners to generate models, which can then be assessed by independent means. This chapter investigates the comparative modelling of cytochrome P450 2D6, an enzyme of unknown structure, using a number of different techniques.

4.1 Introduction - Cytochrome P450 2D6

The function of the P450s is, in general, the same across all the families, the activation of molecular dioxygen (\( \text{O}_2 \)) and subsequent insertion of a single atom of molecular oxygen into an organic molecule (Schlichting et al 2000). A small number of P450s are not \( \text{O}_2 \) reductases, e.g. allene oxide synthase and P450NOR, nevertheless their structures are similar.

Cytochrome P450 2D6 belongs to family 2 of the P450s. This family is concerned with the detoxification, usually in animals, of pharmaceuticals, phytoalexins and a wide-range of other exogenous and endogenous substances (Rendic and Dicarlo 1997, Lewis 1998). It is believed that at least 30% of the drugs in current clinical use are metabolised preferentially by the hepatic 2D6 isoform (Jones et al 1997). 2D6 also gives rise to the defect in man known as the debrisoquine/sparteine polymorphism (Mahgoub et al 1977, Eichelbaum et al 1979). Inheritance of the ‘poor metaboliser’ phenotype has been linked with an increased susceptibility to ailments such as Parkinson’s disease (Eichelbaum et al 1979) and certain cancers (Smith et al 1995). Poor metabolisers may not be able to bio-activate a parent drug to its therapeutically active metabolites (Koymans et al 1992), and more than twenty compounds have been identified which show impaired oxidation for affected individuals. As medicines they are diverse, and include \( \beta \)-adrenergic blocking agents, tricyclic antidepressants, antiarrhythmics and analgesics (Kalow et al 1986, Guengerich et al 1989, Relling et al 1989, Meyer et al 1990), but a characteristic common to many 2D6 substrates is the presence of at least one basic nitrogen atom. Many models of the active site of 2D6
have postulated the involvement of a carboxylate group in the protein forming a salt bridge with this basic nitrogen (Koymans et al 1992, Wolff et al 1985, Meyer et al 1986, Islam et al 1991), and this has been proposed both by modelling (Islam et al 1991) and by mutagenesis (Mackmann et al 1996, Ellis et al 1995) to be Asp 301, a residue in the I-helix (substrate recognition site [SRS] 4).

A number of structural models of 2D6 have been produced previously (Lewis et al 1997, Modi et al 1996, de Groot 1996) using a variety of methods. The work presented in this chapter uses comparative modelling to extend previous work by incorporating the newly published mammalian P450 2C5 structure (Williams et al 2000) as a structural template, thus investigating a number of alternative structural models of 2D6.

4.2 Experimental

4.2.1. Comparative Modelling

Previously published models of 2D6 (Modi et al 1996, Modi et al 1997) are based on the three bacterial P450 structures available at the time: P450 cam (2cpp (Poulos et al 1987)), P450 terp (1cpt (Hasemann et al 1994)) and P450 BM-3 (2hpd (Ravichandran et al 1993)). These were subsequently superseded by models (unpublished), which also incorporated P450 ery-F (1oxa (Cupp-Vickery et al 1995)). These four-template models are presented here, and compared with models generated by incorporating a mammalian P450, rabbit 2C5 (PDB accession code 1dt6 (Williams et al 2000)). It should be noted that, subsequent to producing our 4 bacterial-template models, higher resolution structures of P450cam (1phc (Poulos et al 1986)), BM-3 (1bu7(Sevrioukova et al 1999)) and ery-F (1eup(Cupp-Vickery et al 2000)) have become available. To make a fair comparison with our 4-template based models, the multiple-template based models incorporating 2C5 were generated using the original (2cpp, 1cpt, 2hpd, 1oxa) structures for the bacterial enzymes. Also, an additional template has become available – P450 NOR (1rom (Park et al 1997)). Again, in order to make a fair comparison with the earlier models, this structure was omitted from the study.
4.2.1.1 Bacterial template approach

This study arises from work carried out several years prior to the structure of cytochrome P450 2C5 being resolved. Proteins of known three-dimensional structure homologous to 2D6 (i.e. suitable structural templates) were identified by scanning the amino acid sequence of human 2D6 against the sequences of those structures in the Protein Data Bank (Berman et al 2000) using PSI-BLAST (Altschul et al 1997). At that time, this resulted in the P450s cam (PDB accession code 2cpp (Poulos et al 1985)), terp (1cpt(Hasemann et al 1994)), BM-3 (2hp4 (Ravichandran et al 1993)) and Ery-F (1oxa(Cupp-Vickery et al 1995)) being identified as suitable templates. These templates were used to generate comparative models by aligning them structurally in CLUSTALX (Thompson et al 1994). This produced a profile that was then aligned against a second profile of the 2D subfamily. This subfamily was identified in an identical manner to the 2C family in the previous chapter (Section 3.2.4). Obtaining a subfamily alignment serves two purposes, highlighting conserved areas across a subfamily and reducing the likelihood of misalignment between templates and target structures.

The resultant alignment was checked to ensure that (i) all of the secondary structural elements had a minimum number of insertions/deletions in them, (ii) any residues crucial to catalytic activity, such as the proximal cysteinate in the fifth co-ordination site of the heme (Lewis 1986), were conserved and (iii) any residues known to play an important role in the binding of substrates, e.g. Asp 301 in 2D6 (Koymans 1992, Modi et al 1996, Smith et al 1998) were conserved. Correlation between the secondary structural elements observed in the template (obtained from the Database of Secondary Structure in Proteins (Kabsch and Sander 1983) [DSSP]) and the secondary structure predicted for the 2D6 (using PSIPRED (Jones 1999)) was used to modify the sequence alignment. Once an acceptable alignment had been produced (denoted Alignment 1; Figure 4.3.2.1) an ensemble of fifteen models of 2D6 was generated using Modeller (Sali and Blundell 1993).
4.2.1.2 Single template approach

The sequence of 2D6 was aligned against the sequence of most homologous template, P450 2C5 (1dt6). To achieve this a sequence alignment of the 2D subfamily was aligned against a sequence alignment of the subfamily of the template (2C) using CLUSTALX (Thompson et al 1994). This alignment was analysed in an identical manner to that for Alignment 1 and the resultant alignment, denoted Alignment 2 (Figure 4.3.2.2), used to generate fifteen models using Modeller.

4.2.1.3 Multiple template approach using FSSP

An initial alignment profile of the sequences of the five structural templates (1dt6 [2C5], 2hpd [BM-3], 2cpp [cam], 1oxa [eryF], 1cpt [terp]) was generated using the FSSP (Fold classification and Secondary Structure alignment of Proteins) database (Holm and Sander 1996). The amino acid sequence of 2D6 was aligned against this by aligning the template profile against a second profile comprising the sequence alignment of the 2D subfamily using the profile alignment option within CLUSTALX. This alignment was then checked in a manner identical to that employed for Alignment 1, and the resulting alignment, denoted Alignment 3 (Figure 4.3.2.3), used to generate fifteen models using Modeller.

4.2.1.4 Multiple template approach using MALIGN3D

An initial alignment profile of the sequences of the five structural templates (1dt6, 2hpd, 2cpp, 1oxa, 1cpt) was generated using the MALIGN3D routine within Modeller to structurally align the templates. The amino acid sequence of 2D6 was aligned against this by aligning the template profile against a second profile comprising the sequence alignment of the 2D subfamily using the profile alignment option within CLUSTALX. This alignment was then checked in a manner identical to that employed for Alignment 1, and the resulting alignment, denoted Alignment 4 (Figure 4.3.2.4), used to generate fifteen models using Modeller.
4.2.2 Model Validation

The lowest energy models (as determined by Modeller) from each ensemble of 15 models (denoted Model 1, Model 2, Model 3 and Model 4 for the lowest energy model derived from Alignment 1, Alignment 2, Alignment 3 and Alignment 4 respectively) were assessed using a number of validation methods. Two types of validations were used: stereochemical quality (PROCHECK (Laskowski et al 1993)) for backbone and sidechain conformation, and sidechain environment (Errat (Colovos and Yeates 1993) and Verify 3D (Luthy et al 1992)). For both the stereochemical and environmental checks, the values returned for the templates were used as a baseline against which the models were compared. Also, it has been shown that the RMSD between the mainchain atoms in (i) the template with the greatest homology to the target and (ii) the model is another useful method of validation (Venclovas et al 1997). Backbone RMSDs were therefore used as an additional validation check; these were calculated with reference to P450 2C5, the most homologous template for Models 2-4.

4.3 Results

4.3.1. Identification of Suitable Templates.

The PSI BLAST results for a search against the Protein DataBank with Cytochrome 450 2D6 as the target amino acid sequence are presented in Table 4.3.1.1.
Table 4.3.1.1 Sequences of P450 structures identified by a PSI-BLAST search using the sequence of P450 2D6

<table>
<thead>
<tr>
<th>Cytochrome P450</th>
<th>PSI BLAST Score</th>
<th>PSI Sequence ID to 2D6</th>
<th>PSI Sequence Identity from PSI BLAST</th>
<th>Alignment 1</th>
<th>Alignment 2</th>
<th>Alignment 3</th>
<th>Alignment 4</th>
<th>Number of Amino Acid Residues</th>
</tr>
</thead>
<tbody>
<tr>
<td>2C5 (1dt6)</td>
<td>614</td>
<td>197/479</td>
<td>41</td>
<td>41</td>
<td>41</td>
<td>42</td>
<td></td>
<td>473</td>
</tr>
<tr>
<td>Cam (2cpp)</td>
<td>283</td>
<td>71/375</td>
<td>18</td>
<td>14</td>
<td>13</td>
<td>12</td>
<td></td>
<td>414</td>
</tr>
<tr>
<td>BM-3 (2hpd)</td>
<td>257</td>
<td>85/461</td>
<td>18</td>
<td>15</td>
<td>16</td>
<td>15</td>
<td></td>
<td>471</td>
</tr>
<tr>
<td>Terp (1cpt)</td>
<td>210</td>
<td>54/264</td>
<td>20</td>
<td>15</td>
<td>14</td>
<td>14</td>
<td></td>
<td>412</td>
</tr>
<tr>
<td>ery-F (1oxa)</td>
<td>175</td>
<td>64/402</td>
<td>15</td>
<td>14</td>
<td>15</td>
<td>15</td>
<td></td>
<td>403</td>
</tr>
</tbody>
</table>

*a* The PSI-BLAST score for an alignment is calculated by summing the scores for each aligned position and the scores for gaps (Altschul et al 1997).

*b* (Number of identical residues)/ (length of sequence fragment identified by PSI-BLAST).

*c* See Figure 4.3.2.1

*d* See Figure 4.3.2.2

*e* See Figure 4.3.2.3

*f* See Figure 4.3.2.4
4.3.2 Amino Acid Sequence Alignments.

The final amino acid sequence alignments used to generate comparative models for Alignment 1 (Figure 4.3.2.1); Alignment 2 (Figure 4.3.2.2), Alignment 3 (Figure 4.3.2.3) and Alignment 4 (Figure 4.3.2.4) are presented below. The ‘conserved gaps’ appearing throughout the four alignments arise due to insertions and deletions in the sequences of the 2D subfamily profile, which are not shown in the alignment.

Figure 4.3.2.1 Bacterial-template alignment (Alignment 1) used to generate comparative models of P450 2D6. The secondary structural elements α-helices (grey), and β-strands (black) as predicted by PSIPRED (Jones 1999) are highlighted.
Figure 4.3.2.2 Single-template alignment (Alignment 2) used to generate comparative models of P450 2D6. The secondary structural elements α-helices (grey), and β-strands (black) as predicted by PSIPRED (Jones 1999) are highlighted.
Figure 4.3.2.3 The FSSP determined amino acid sequence alignment (Alignment 3) used to generate comparative models of P450 2D6. The secondary structural elements α-helices (grey), and β-strands (black) as predicted by PSIPRED (Jones 1999) are highlighted.
Figure 4.3.2.4 MALIGN3D alignment (Alignment 4) used to generate comparative models of P450 2D6. The secondary structural elements α-helices (grey), and β-strands (black) as predicted by PSIPRED (Jones 1999) are highlighted.
4.3.3 Model Analysis

The lowest energy models from each of the Alignments 1-4 were analysed using Ramachandran Plots (Morris et al 1992), Errat (Colovos and Yeates 1993), Verify 3D (Luthy et al 1992). The results are presented in Table 4.3.3.1.

Table 4.3.3.1 Results of validation studies performed on the lowest energy models produced from the three different sequence alignments.

<table>
<thead>
<tr>
<th>Model</th>
<th>Ramachandran Plot (%)</th>
<th>Verify 3Db (total score)</th>
<th>Errat (%)</th>
<th>Mainchain RMSDc (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 Template (Bacterial)</td>
<td>70</td>
<td>139</td>
<td>68</td>
<td>N/A</td>
</tr>
<tr>
<td>(Model 1d)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single Template</td>
<td>87</td>
<td>173</td>
<td>75</td>
<td>0.8</td>
</tr>
<tr>
<td>(Model 2e)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple Template</td>
<td>83</td>
<td>191</td>
<td>79</td>
<td>2.3</td>
</tr>
<tr>
<td>FSSP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Model 3f)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple Template MALIGN3D</td>
<td>80</td>
<td>182</td>
<td>74</td>
<td>2.9</td>
</tr>
<tr>
<td>(Model 4g)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

aPercentage of residues with φ,ψ conformation in the most favoured regions of the Ramachandran plot.

bThe total Verify 3D score summed over all of the residues. For a protein the same size as P450 2D6 a score of 207 is expected, and a score of less than 97 would indicate an incorrect structure.

cThe RMSD between aligned mainchain atoms in the model, and those in the most homologous template P450 2C5 (1dt6).

dModel 1, Model 2, Model 3 and Model 4 refer to the lowest energy model derived from Alignment 1 (Figure 4.3.2.1), Alignment 2 (Figure 4.3.2.2), Alignment 3 (Figure 4.3.2.3) and Alignment 4 (Figure 4.3.2.4), respectively.
The templates used to produce the models were also subject to an identical analysis, with the exception of determining the mainchain RMSD. This provides a baseline against which comparison with the models can be based (Table 4.3.3.2).

Table 4.3.3.2. Results of validation studies performed on the 3D structural templates used to generate comparative models of 2D6.

<table>
<thead>
<tr>
<th>Cytochrome</th>
<th>Ramachandran Plot (%)</th>
<th>Verify 3D (Total Score)</th>
<th>Errat (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2C5 (ldt6)</td>
<td>71</td>
<td>182</td>
<td>94</td>
</tr>
<tr>
<td>Cam (2cpp)</td>
<td>90</td>
<td>222</td>
<td>97</td>
</tr>
<tr>
<td>BM-3 (2hpd)</td>
<td>91</td>
<td>221</td>
<td>95</td>
</tr>
<tr>
<td>Terp (1cpt)</td>
<td>89</td>
<td>188</td>
<td>95</td>
</tr>
<tr>
<td>ery-F (1oxa)</td>
<td>91</td>
<td>184</td>
<td>96</td>
</tr>
</tbody>
</table>

4.4 Discussion

The percentage of residues with conformations in the most favoured regions of the Ramachandran plot (Table 4.3.3.1) suggests that Model 2 has the best stereochemistry as it has the greatest number of residues in the most favoured region (87%), followed by Model 3 (83%), Model 4 (80%), and finally Model 1 (71%). This can be put into context by examining the scores achieved by the 3D structural templates (Table 4.3.3.2). It is apparent that all of the 2D6 models perform significantly better than the 2C5 crystal structure (which returns a value of 71% due to the fact that it is a low resolution structure, 3.0 Å), and compare well with the values obtained from the other template structures. The observation that all of the models perform better than the 2C5 crystal structure can likely be attributed to Modeller being parameterised to
produce models with good stereochemistry, despite poor stereochemistry in the templates.

For a protein the size of 2D6, a Verify 3D score of 215 would indicate a valid structure, whereas a score lower than 95 would indicate that the model is incorrectly folded. None of the models achieved the expected score (Table 4.3.2.3), with Model 3 performing best (191). Again, this can be put into context by examining the crystal structures. Both P450 2C5 (182) and P450 terp (188) fail to achieve their expected scores (209 and 204, respectively). All of the amino acid sequence alignments used to generate the models utilise either 2C5, or 2C5 plus P450 terp in conjunction with the other templates. Therefore it is perhaps not surprising that the 2D6 models have Verify 3D scores lower than expected for a structure of this size, because of the inherent limitations placed upon them by the templates used.

All of the templates have Errat scores > 90%. The 2D6 models have Errat scores below this, with Model 3 once again returning the best result (79%). It is probable that the high Errat scores for the templates are due to the fact that the crystal structures are well defined with regards to their electrostatics. This is not the case with the models, as it is difficult to justify electrostatic refinement on a comparative model. Whilst every precaution is taken to ensure that the comparative model produced is as accurate a representation of the true system as is possible, it is highly likely that there will be errors in the model, particularly in regions which are not well conserved across the templates, such as the β-sheets in the N and C termini, the B'-C loop and the F-G loop. Refining these ambiguous regions in the protein with regards to the electrostatics would serve to increase the Errat score, but may not increase the quality of the model.

The RMSD between aligned mainchain atoms in the model and those in the most homologous template (2C5) were measured (Table 4.3.3.1). Model 2 (0.8 Å) does not deviate (sample conformational space) as much as would be expected, given the percentage sequence identity, from the crystal structure of 2C5. This is likely the result of using a single structural template (see Chapter 3); work carried out by Chothia and Lesk (Chothia and Lesk 1986) indicates that for the crystal structures
they studied a value of > 1Å for a template with 41% sequence identity is expected for the Cα RMSD. Model 3 and Model 4 perform better (RMSDs of 2.3Å and 2.9Å, respectively), illustrating that substantially more of conformational space has been explored during the modelling process.

Model 3 has the highest Verify 3D and Errat scores, and the second highest Ramachandran score (Table 4.3.3.1) indicating that it is the “best”, in terms of mainchain stereochemistry and amino acid environment, of all of the models.

4.5 Conclusions

In the absence of crystal structures, comparative modelling gives invaluable insight into the three-dimensional structure of a protein. The four sets of models of 2D6 generated illustrated that, in the case of 2D6, use of multiple structural templates produce better quality models than a single (most homologous) template, and in contrast to the results obtained for comparative models of P450 2C5 (Chapter 3) shows that an FSSP based multiple template alignment as opposed to a MALIGN3D based multiple template alignment results in the best quality model. This serves to reiterate that no two comparative modelling problems are identical and that there is no definitive method for the production of a high-quality model. As such a combined approach, utilising several different methods is the best way to produce comparative models.

The models which are ultimately produced can only be as good as the templates used to produce them. The Modeller program is able to generate models of significantly better stereochemistry than the poor-resolution templates they are based on due to the parameterisation within the program, but it appears that any errors in the amino acid environments of the templates, as highlighted by the results of the Verify 3D investigation, are carried across from the templates to the models.

Incorporation of information from the mammalian 2C5 crystal structure has a profound effect on the modelling process on a macroscopic level. By increasing the percentage sequence identity of the most homologous template there is a significant
improvement with regards to the quality of the 'macroscopic' properties of the comparative models \textit{i.e.} amino acid environment and overall stereochemistry. Therefore, inclusion of the 2C5 crystal structure into the comparative modelling process appears to have a profound effect.
Chapter 5

Impact on Comparative Modelling of Cytochrome P450 2D6 of Incorporating the Mammalian 2C5 Crystal Structure.
5.0 Preface

Having produced comparative models of cytochrome P450 2D6 (Chapter 4), and determined that the inclusion of the P450 2C5 crystal structure has a profound effect on the modelling process – improving the quality with respect to amino acid environment and overall stereochemistry – it was of interest to assess how the different approaches to modelling affect the P450 2D6 active site at the atomic level.

5.1 Introduction

Many models of the active site of 2D6 have postulated the involvement of a carboxylate group in the protein forming a salt bridge with this basic nitrogen (Koymans et al 1992, Wolff et al 1985, Meyer et al 1986, Islam et al 1991), and this has been proposed to be Asp 301, a residue in the I-helix (substrate recognition site [SRS] 4), both by modelling (Islam et al 1991) and by mutagenesis (Mackmann et al 1996, Ellis et al 1995).

The active sites of the four lowest energy structural models of 2D6 produced previously – *Model 1, Model 2, Model 3 and Model 4* (Chapter 4) – have been analysed using principal component analysis (PCA) in an effort to highlight the differences in the active sites of the different models. Molecular docking of two substrates with known binding orientations has also been carried out to explore the proposed salt bridge between the basic nitrogen in the substrate and a carboxylate group in the active site.

5.2 Experimental

5.2.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was used to characterise the active sites of each of the lowest energy models. *Models 1-4* and the 2C5 crystal structure were superimposed on their heme moieties. A GRID (Goodford 1985) box (18 x 18 x 18 Å³, GRID spacing 0.66 Å) large enough to encompass the superposed active sites of all of the three-dimensional structures was generated and the active site of each model...
characterised using ten GRID probes representative of the atom types likely to be found in P450 2D6 substrates [methyl (C3), carbonyl (O), carboxylate oxygen (O::), NH-amide like (N1), sp3 amine cation (N3+), phenolic OH (OH), alkyl OH (O1), water (OH2), trimethylammonium cation (NM3), and chlorine (Cl)]. This generated ten values for each grid point for each of the five structures. GOLPE (Baroni et al 1993) was used to analyse these data, generating loadings and scores plots. These plots gave insight into the differences in topology and chemical identity between the active sites of the different models. Scores plots were produced with one data point per structure by combining the data for all the probes for each grid point.

5.2.2 Molecular Docking

To evaluate the predictive value of the 2D6 models, docking studies were carried out on some of the known substrates of 2D6: codeine (Wrighton and Stevens 1992) and 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) (Modi et al 1997). Docking calculations were performed using the program GOLD (Jones et al 1997). The distinct advantage of using these substrates is that their interactions with 2D6 have been studied and the products of metabolism are well documented (Modi et al 1996, Modi et al 1997).

To investigate the validity of the dockings, the active site of the model was examined using GRID (Goodford 1985), producing contour maps that highlighted the favourable interactions of various chemical probes within the active site. The parameters for the GRID box and the probes used are given in the previous section. The contour maps generated by GRID were analysed visually using Sybyl 6.5 (Tripos).

5.3 Results

5.3.1 Principal Component Analysis

PCA was used to identify, on a macroscopic scale, the differences between the topology and chemistry of the active sites of the Models 1-4 and the most homologous crystal structure, 2C5. There are two pieces of important data generated when
performing PCA, the loadings plot and the scores plot. A loadings plot contains all of
the data for each calculated GRID point for each of the ten GRID probes used across
all five structures. The absolute value for any particular point is unimportant, although
in its simplest terms the position of a point along the principal components of a
loadings plot determines how important it is in describing the characteristics of that
principal component.

5.3.1.1 Loadings Plots

Due to the nature of the study the loadings plot is complex and we see the appearance
of five distinct lines (Figure 5.3.1.1A). Each of these lines is due to one of the five
structures being investigated (Models 1-4 and the 2C5 crystal structure), and occurs
because of the different shapes of the active sites. The reasons for the appearance of
these lines is due to a correlation between the data on the first and second principal
component axes, a phenomenon which is unhelpful to the analysis. This phenomenon
is described in greater detail in the discussion section of this chapter (Section 5.4.1).
The data responsible for the appearance of these lines was computationally removed
(Figure 5.3.1.1B).

5.3.1.2 Scores Plots

The scores plot (Figure 5.3.1.1C) produces a more easily interpretable picture of the
differences between individual models, by taking multi-dimensional information and
reducing it into two dimensions. The relative distance between individual data points
along each of the principal components is a measure of the variation between those
point.; in this case the distance along the principal components is an indication of the
extent of the differences between the active sites of Models 1-4 and the 2C5 crystal
structure. It shows a large difference between the bacterial P450-based model (Model
1), and the remainder of the structures. This structure was removed from the analysis
in an effort to explore the more subtle differences between the different 2C5-based
models.
Figure 5.3.1.1 A) GOLPE-generated loadings plot for the superimposed active sites of Models 1-4 and the 2C5 crystal structure, analysed using GRID with 10 chemical probes. B) GOLPE-generated loadings plot corresponding to (A) but from which those GRID points for which a value of 5 kcal mol\(^{-1}\) was returned in any structure for all ten probes investigated, have been deleted for all structures. C) GOLPE-generated concatenated scores plot to highlight the differences between the active sites between the four models and the 2C5 crystal structure. PC1 accounts for 36% of the variance, PC2 accounts for 15% of the variance.
5.3.1.3 PCA for 2C5-based models

Once again, the initial loadings plot for the analysis of Models 2-4 and 2C5 shows distinct lines radiating from the origin (Figure 5.3.1.3A), one for each of the structures. This reiterates that the shape of the active site for each structure is different, and to gain further information it is necessary to eliminate the data responsible for the appearance of these lines (Figure 5.3.1.3B). The scores plot generated (Figure 5.3.1.3.C) once again gives a more easily interpretable picture of the differences between the models.

It is also important to show that the multiple-template models (Model 3 and Model 4) differ significantly from the bacterial templates (2cpp, 2hpd, 1cpt, 1oxa) used. If principal component analysis was to show these models to be significantly more similar to the bacterial templates than the 2C5 template it would raise concerns with regards to the inclusion of the 2C5 template in the modelling process *i.e.* if the structure of the active site of the models is significantly different to the 2C5 structure, but not significantly different to the structure of the bacterial templates then there is little scientific value in incorporating the 2C5 crystal structure into the modelling process. A further investigation into the differences between the models and the most homologous bacterial template (BM-3) was carried out in a manner identical to those detailed above (Section 5.3.2). The scores plot (Figure 5.3.1.3D) incorporating data for the most homologous bacterial template, 2hpd (P450 BM-3), is presented below.
Figure 5.3.1.3 A) GOLPE-generated loadings plot from the active sites of the crystal structure of 2C5 and those models (Models 2, 3 and 4) that incorporate information from 2C5. B) GOLPE-generated loadings plot corresponding to (A) but from which those GRID points for which a value of 5 kcal mol\(^{-1}\) was returned in any structure for all ten probes investigated have been deleted for all structures. C) GOLPE-generated scores plot illustrating the differences in the active sites between Models 2, 3 and 4 and the 2C5 crystal structure. PC1 accounts for 36 % of the variance, PC2 accounts for 15 % of the variance. D) GOLPE-generated scores plot to illustrate the differences between the most homologous bacterial template (BM3), the 2C5 crystal structure and Models 1-4. PC1 accounts for 38 % of the variance, PC2 accounts for 11 % of the variance.
5.3.2 Molecular Docking of Codeine

The validation studies discussed previously (Chapter 4) indicated that Model 3 was the "best" of the P450 2D6 models generated. Hence, the subsequent analyses of the active site of the 2C5-inclusive model were carried out using Model 3.

It has been postulated that a carboxylate group in 2D6 forms a salt bridge with the basic nitrogen of the substrate; this has been proposed both by modelling and by mutagenesis to be Asp 301. This was not observed for any of the most highly ranked dockings, although an analysis of the distances from the heme iron to the hydrogen atoms of the substrate showed that, for these dockings, the Fe—H interatomic distances were similar to those obtained from NMR studies of the codeine complex (Modi et al 1996) (Table 5.3.2.1)

Table 5.3.2.1 Interatomic Fe-H distances for codeine docked into the active site using GOLD compared with experimentally determined NMR distances (Modi et al 1996).

<table>
<thead>
<tr>
<th>Proton</th>
<th>Experimental Distance (Å)</th>
<th>Average Distance in Models* (Å)</th>
<th>Difference Between Experimental and Average Distance in Models (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1H</td>
<td>7.5 ± 0.2</td>
<td>6.3 ± 0.3</td>
<td>1.2 ± 0.4</td>
</tr>
<tr>
<td>C2H</td>
<td>5.0 ± 0.1</td>
<td>4.1 ± 0.3</td>
<td>0.9 ± 0.3</td>
</tr>
<tr>
<td>-OCH3</td>
<td>3.1 ± 0.1</td>
<td>2.9 ± 0.1</td>
<td>0.2 ± 0.1</td>
</tr>
<tr>
<td>C4H</td>
<td>9.1 ± 0.1</td>
<td>8.6 ± 0.2</td>
<td>0.5 ± 0.2</td>
</tr>
<tr>
<td>C5H</td>
<td>10.0 ± 0.2</td>
<td>8.4 ± 0.1</td>
<td>1.6 ± 0.2</td>
</tr>
<tr>
<td>C6H</td>
<td>9.3 ± 0.3</td>
<td>8.5 ± 0.1</td>
<td>0.8 ± 0.3</td>
</tr>
<tr>
<td>C7H</td>
<td>10.2 ± 0.2</td>
<td>9.7 ± 0.1</td>
<td>0.5 ± 0.2</td>
</tr>
<tr>
<td>C8H</td>
<td>11.2 ± 0.2</td>
<td>10.4 ± 0.2</td>
<td>0.8 ± 0.3</td>
</tr>
<tr>
<td>C10H2</td>
<td>9.8 ± 0.1</td>
<td>8.8 ± 0.2</td>
<td>1.0 ± 0.2</td>
</tr>
<tr>
<td>C15H4</td>
<td>9.3 ± 0.1</td>
<td>8.5 ± 0.1</td>
<td>0.8 ± 0.1</td>
</tr>
<tr>
<td>C15H5b</td>
<td>10.8 ± 0.2</td>
<td>9.6 ± 0.1</td>
<td>1.2 ± 0.2</td>
</tr>
<tr>
<td>NCH3</td>
<td>12.1 ± 0.2</td>
<td>11.8 ± 0.3</td>
<td>0.3 ± 0.4</td>
</tr>
</tbody>
</table>

* The mean distance and standard deviation derived from fifteen dockings of codeine into the 2D6 active site.
5.3.3 Molecular Docking of 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP)

Docking investigations were carried out into another substrate of 2D6, 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP), which has also been studied by NMR (Modi et al 1997) to determine whether a different substrate would show association with Asp 301 in the 2D6 active site. Two products have been observed when MPTP is metabolised by 2D6. These are the N-demethylated and para-hydroxylated products, and are thought to arise because of the different orientations this small molecule is able to adopt in the active site of 2D6. For the first investigation GOLD predicted that only hydroxylation of the benzene ring would occur. None of the twenty dockings produced displayed an orientation that would facilitate N-demethylation.

However, only two of the twenty dockings oriented MPTP in such a manner as to facilitate para-hydroxylation, with the remainder implying that meta-hydroxylation occurred. Comparison of the interatomic iron-proton distances from the dockings with those determined by NMR shows that the only atom with significantly different Fe-H interatomic distances in comparison to the experimental values is the para-hydrogen (C9; Table 5.3.3.1). To investigate this further, a docking experiment was carried out which restrained the C9 atom of MPTP to within 3.0 Å of the Fe atom of the heme, thus ensuring that the orientation of the substrate in all of the subsequent dockings would facilitate para-hydroxylation.

The results for all of the restrained dockings place the substrate molecule in an orientation where para-hydroxylation can occur. The interatomic Fe-H distances for the restrained dockings compare favourably to the experimental results (Table 5.3.3.1). For both the restrained and the unrestrained dockings of MPTP there was no association of the basic nitrogen group in the substrate with the carboxylate group of the Asp 301 residue.
Table 5.3.3.1 Interatomic Fe—H distances for MPTP docked into the active site using GOLD compared with experimentally determined NMR distances (Modi 1997)

<table>
<thead>
<tr>
<th>Proton</th>
<th>Experimental Distance (Å)</th>
<th>Average Distance Dockings (Å)</th>
<th>Average Distance Constrained Dockings (Å)</th>
<th>Difference Between Experimental and Docking Values (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCH₃</td>
<td>12.4 ± 0.6</td>
<td>12.4 ± 0.3</td>
<td>12.7 ± 0.1</td>
<td>0.0 ± 0.7 / 0.3 ± 0.6</td>
</tr>
<tr>
<td>C2H⁴</td>
<td>11.1 ± 0.5</td>
<td>12.0 ± 0.5</td>
<td>10.8 ± 0.3</td>
<td>0.9 ± 0.7 / 0.3 ± 0.6</td>
</tr>
<tr>
<td>C2H⁵</td>
<td>10.9 ± 0.7</td>
<td>11.0 ± 0.3</td>
<td>10.3 ± 0.3</td>
<td>0.1 ± 0.8 / 0.6 ± 0.4</td>
</tr>
<tr>
<td>C3H⁴</td>
<td>9.1 ± 0.4</td>
<td>10.3 ± 0.5</td>
<td>8.8 ± 0.1</td>
<td>1.2 ± 0.6 / 0.3 ± 0.5</td>
</tr>
<tr>
<td>C3H⁵</td>
<td>8.9 ± 0.2</td>
<td>10.0 ± 0.7</td>
<td>8.3 ± 0.3</td>
<td>1.1 ± 0.7 / 0.6 ± 0.8</td>
</tr>
<tr>
<td>C5H</td>
<td>8.3 ± 0.6</td>
<td>7.9 ± 0.2</td>
<td>8.6 ± 0.2</td>
<td>0.4 ± 0.6 / 0.3 ± 0.3</td>
</tr>
<tr>
<td>C6H⁴</td>
<td>10.5 ± 0.5</td>
<td>10.2 ± 0.3</td>
<td>10.7 ± 0.2</td>
<td>0.3 ± 0.6 / 0.2 ± 0.4</td>
</tr>
<tr>
<td>C6H⁵</td>
<td>10.6 ± 0.5</td>
<td>10.2 ± 0.2</td>
<td>10.8 ± 0.2</td>
<td>0.4 ± 0.5 / 0.2 ± 0.3</td>
</tr>
<tr>
<td>C7H, C11H</td>
<td>6.9 ± 0.4</td>
<td>6.6 ± 0.2</td>
<td>6.6 ± 0.5</td>
<td>0.3 ± 0.4 / 0.3 ± 0.5</td>
</tr>
<tr>
<td>C8H, C10H</td>
<td>4.5 ± 0.2</td>
<td>4.8 ± 0.3</td>
<td>4.1 ± 0.4</td>
<td>0.3 ± 0.4 / 0.4 ± 0.5</td>
</tr>
<tr>
<td>C9H</td>
<td>3.0 ± 0.1</td>
<td>5.2 ± 1.0</td>
<td>2.9 ± 0.1</td>
<td>2.2 ± 1.0 / 0.1 ± 1.0</td>
</tr>
</tbody>
</table>
5.3.4 Validation of Dockings using GRID

The contour maps generated by GRID for the trimethylammonium cation (NM3) probe, the probe most representative of the basic nitrogen atom in both codeine and MPTP, overlap the position occupied by the tertiary nitrogen atom of the substrates in the docked structures (Figure 5.3.4.1) at a relatively low contour level. This is an indication that the dockings are valid, and not produced as a result of limitations with regards to the amount of available conformational space in the active site.

Figure 5.3.4.1 The active site of 2D6 (Model 3), illustrating the binding of codeine. There is excellent correlation between the position of the highest ranked GOLD docking of codeine and the contour map of the GRID trimethylammonium cation (NM3) probe contoured at -8 kcal mol\(^{-1}\).
5.4 Discussion

5.4.1 Principal Component Analysis (PCA)

PCA was used to identify the differences between the topology and chemistry of the active sites of the models and the most homologous crystal structure, 2C5. There are two pieces of important data generated when performing PCA, the loadings plot and the scores plot. The scores plot represents the grid data in a reduced dimensionality with the most significant differences between the models showing up in the lower components. The loadings plot describes the components in terms of the original variables (grid points).

The loadings plot is complex and we see the appearance of five distinct lines radiating from the origin (Figure 5.3.1.1A). Each of these lines is due to one of the five structures being investigated (Models 1-4 and the 2C5 crystal structure), and occurs because of the different shapes of the active sites. The different approaches to modelling will produce models of differing shapes, and it is not possible to superimpose all of the residues in the active site of one structure perfectly onto those of another structure. The active sites of all of the structures are distinctly different shapes, an observation supported by examining the MOLCAD (Tripos Inc., St. Louis, MO, USA) surfaces generated for the active site of each of the models (Figure 5.4.1.1). The major differences in shape are due to the relative positions of inherently flexible regions of the protein, in particular the B–C region and the F-G region and, to a lesser extent, secondary structural elements that vary in position across the P450 family, predominantly the F helix. Therefore, some of the GRID points will be placed in locations that, because of structural differences in the models, are partially or wholly within a protein structure. As a result, the energy of interaction calculated at this point would be phenomenally high; parameterisation within GRID returns the maximum allowed value (5 kcal/mo1\(^{-1}\)). These variables will dominate the PCA and swamp out more interesting, subtler differences. Thus, this data was eliminated by assigning a value of 99.99 to those GRID points across the structures with the maximum (5 kcal mol\(^{-1}\)) value for all ten probes in any single structure (i.e. setting zero variance across this data - Appendix 5). As the loadings plot shows (Figure [Diagram])...
5.3.1.1B), this eliminated the straight lines observed previously. Hence, any differences could no longer be attributed to a "trivial" alteration in the gross shape of the active site of a structure.

Figure 5.4.1.1 MOLCAD surfaces (Heiden et al 1993) illustrating differences in shape and volume of the active sites of Models 1-4 and the 2C5 crystal structure: A) Model 1 generated using the 4 bacterial structural templates (Alignment 1); B) Model 2 generated using the 2C5 crystal structure (Alignment 2); C) Model 3 generated using the FSSP aligned four bacterial structural templates and the 2C5 crystal structure (Alignment 3); D) Model 4 generated using the MALIGN3D aligned four bacterial templates and 2C5 crystal structure; E) The active site of the 2C5 crystal structure.
A scores plot produces a more easily interpretable picture of the differences between individual models, by taking multi-dimensional information and reducing it into two dimensions. The relative distance between individual data points along each of the principal components is a measure of the variation between those points; in this case (Figure 5.3.1.1C) the distance along the principal components is an indication of the extent of the differences between the active sites of Models 1-4 and the 2C5 crystal structure. There are three distinct groups, group 1: Model 1; group 2: Model 2 and the 2C5 crystal structure; and group 3: Models 3 & 4. The relatively large distance between group 1 and groups 2 and 3 along the first principal component shows that inclusion of the 2C5 structure in the modelling process causes significant differences in the active site of 2D6, not all of which can be attributed merely to a change in gross topology. Hence, Model 1 is distinctly different from any of the other structures analysed. This is not unexpected since it is the only model that does not include information from the 2C5 crystal structure. Group 2 illustrates that Model 2 and the 2C5 crystal structure are similar, as indicated by their proximity along both PC1 and PC2 in the scores plot (Figure 5.3.1.1C), but are significantly different from groups 1 and 3, as indicated by the separation of these groups along PC2. One reason for this may be the nature of the modelling process. Model 2 was constructed using information solely from the 2C5 crystal structure, and is therefore more likely to share more characteristics with 2C5 than Models 3 and 4, which were built using information from 2C5 plus the bacterial templates. In turn, Models 3 & 4 would be expected to be more similar to one another than to the other structures, and again this is apparent from the scores plot. Discarding the bacterial P450-based model (Model 1), because of the major differences between it and the remainder of the structures, allows the more subtle differences between the different 2C5-based models to be investigated.

The initial loadings plot for the analysis of Models 2-4 and 2C5 shows distinct lines radiating from the origin (Figure 5.3.1.3A), one for each of the structures. This illustrates, as when Model 1 was included in the analysis (Figure 5.3.1.1A), that the shape of the active site for each structure is different, and to gain further information it is necessary to eliminate the data responsible for the appearance of these lines (Figure 5.3.1.3B). The scores plot generated shows two distinct groups differentiated along PC1; group (i) Models 3 & 4 and group (ii) Model 2 and the 2C5 crystal
structure. Again, the models produced from the multiple template alignments, group (i), are similar, as indicated by their comparable PC1 values (Figure 5.3.1.3C). They are different to both Model 2 and the 2C5 crystal structure. In contrast to the previous scores plot (Figure 5.3.1.1C) Model 3 and Model 4 have significantly different values along PC2, which could be attributed to the differences between the amino acid sequence alignments used to generate Model 3 and Model 4 becoming more apparent with the removal of the data for the disparate Model 1.

One cause for concern raised by the scores plot was the difference between the 2C5 crystal structure and Models 3 and 4. If Models 3 and 4 were observed to be either as different, or more different, from the 2C5 crystal structure than the lower homology bacterial templates, this would imply that including the 2C5 template in the modelling process has no added value, as it simply gives an answer which is different to the previous one and not necessarily a more accurate representation of 2D6. Hence, the crystal structure of the most homologous bacterial template BM-3 (2hpd) was incorporated into the PCA. As can be seen from the resulting scores plot (Figure 5.3.1.3D), Models 2, 3 and 4 are closer to, and therefore more similar to 2C5 than BM3, whereas Model 1 has greater similarity with the bacterial crystal structure. This implies that the incorporation of the 2C5 structure into the modelling process alters the shape and/or chemistry of the active site substantially in comparison to the bacterial templates, and hence is a worthwhile exercise.

To ascertain the differences at the atomic level, which give rise to the differences observed in the PCA for the active sites of Models 1-4 and the 2C5 crystal structure, a single GRID probe was used to characterise the structures, and the resulting loadings plot mapped back onto the protein structures as a GRID contour map. The three probes investigated (OH2, C3 and N3+) indicated that the major differences arose from gross changes in relative positioning and shape of inherently flexible regions of the protein such as the F-G loop, the B–C region and the β-strands towards the C-terminus. However, the N3+ probe also highlighted chemical differences between the 2C5 crystal structure and the models, in particular the substitution of the acidic Glu 297 in the I helix of 2C5 for the neutral Val 308 in the I helix of the 2D6 models (Figure 5.4.1.2).
Figure 5.4.1.2 The GOLPE generated GRID-like loadings contours for the trimethylammonium probe, comparing the active sites of Model 3 and the 2C5 crystal structure. This illustrates that the major difference between 2C5 and Model 3 (i.e. 2D6) results from the chemical difference between (acidic) Glu 297 in the I-helix of 2C5 (black) and the corresponding (hydrophobic) Val 308 in 2D6 (Model 3; grey).

5.4.2 Molecular Docking

Codeine can bind to the active site of 2D6 in a manner consistent with the known role of 2D6 as an O-demethylase of codeine (Mortimer et al 1990, Desmeules et al 1991, Ladona et al 1991), and experimental NMR data (Modi et al 1996), with the methoxy group situated close to the iron atom of the heme. It has been postulated that a carboxylate group in 2D6 forms a salt bridge with the basic nitrogen (Koymans et al
1992, Wolff et al 1985, Meyer et al 1986, Islam et al 1991) of the substrate; this has been proposed both by modelling (Islam et al 1991) and by mutagenesis (Mackmann et al 1996, Ellis et al 1995) to be Asp 301. However, this was not observed. Instead, the basic nitrogen was seen to interact with a second acidic residue in the active site, Glu 216 (in the F-helix, SRS-2). This was true for all of the highly ranked dockings from GOLD. Analysis of the distances from the heme iron to the hydrogen atoms of the substrate showed that, for these dockings, the Fe–H distances were closely similar to those obtained from NMR studies of the codeine complex (Modi et al 1996) (Table 5.3.2.1) implicating Glu 216, as opposed to Asp 301, as the major binding determinant of substrates in the active site of P450 2D6. To investigation this further docking studies of 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) into the 2D6 model were carried out. MPTP was chosen because experimental data regarding the receptor-ligand complex was available (Modi et al 1997).

Two metabolites of MPTP have been observed when it is metabolised by 2D6, the N-demethylated and para-hydroxylated product, which are thought to arise because of the different orientations this small molecule is able to adopt in the active site of 2D6. Once again, it is widely believed that the acidic residue Asp 301 is responsible for sequestering the basic nitrogen of the substrate, but the GOLD generated dockings suggested that this was not the case, and that in all of the highest ranked dockings the basic nitrogen of the substrate was positioned much closer to Glu 216 than Asp 301.

Initially GOLD predicted that only hydroxylation of the benzene ring would occur. None of the twenty dockings produced displayed an orientation that would facilitate N-demethylation. However, only two of the twenty dockings oriented MPTP in such a manner as to facilitate para-hydroxylation, with the remainder implying that meta-hydroxylation occurred. This is not consistent with the experimentally observed products, and may be due to the relatively small size of the substrate molecule allowing it to move comparatively freely within the active site. Comparison of the interatomic iron-proton distances from the dockings with those determined by NMR shows that the only substrate atom showing significantly different Fe–H interatomic distances was the para hydrogen (C9; Table 5.3.3.1). To investigate this further, a docking experiment was carried out which constrained the C9 atom of MPTP to
within 3.0 Å of the Fe atom of the heme. This ensured that the orientation of the substrate in all of the subsequent dockings would facilitate para-hydroxylation only.

The results for all of the restrained dockings place the substrate molecule in an orientation where para-hydroxylation can occur. The interatomic Fe–H distances for the dockings compare favourably to the experimental results (Table 5.3.3.1), and the basic nitrogen atom of the substrate only associates with Glu 216 (and not Asp 301), lending further credence to the importance of this residue as a major determinant of substrate binding.

5.4.3 Validation of Docking Studies

The contour maps generated by GRID for the trimethylammonium cation (NM3) probe overlap the position occupied by the tertiary nitrogen atom in the substrates (Figure 5.3.4.1) at a relatively low contour level. This is an indication that the dockings, which place the basic nitrogen in the vicinity of Glu 216, are representative of the modelled active site. This also implies a key role for Glu 216 in binding of a substrate containing a basic nitrogen atom.

However, mutagenesis experiments have shown that Asp 301 is critical to the activity of P450 2D6 (Mackmann et al 1996, Ellis et al 1995). It was believed that this was due to a direct interaction between the basic nitrogen atoms of the substrate molecules and the acidic residue, Asp 301, for those substrates containing a basic nitrogen atom 5-7 Å from the site of oxidation. Codeine undergoing O-demethylation, and MPTP undergoing para-hydroxylation fall into this category. The role of Glu 216 as a binding determinant has been alluded to in previous studies as a way of explaining the metabolism of larger substrates, i.e. those substrates with have basic nitrogen atoms at a distance > 10 Å from the site of oxidation (Lewis 1998, Lewis et al 1997, Venhorst et al 2000). The results from the studies of this thesis suggest that Glu 216 additionally plays an important role in binding of 5-7 Å substrates. However, the proposed crucial role of Asp 301 cannot be overlooked. To address this, models were built using Alignment 3 but additionally incorporating the mutation of acidic Asp 301 to a hydrophobic alanine (D301A). Dockings of codeine and subsequent GRID analysis of the mutant were then carried out.
The docking experiments involving the D301A mutant did not position codeine in an orientation that would facilitate O-demethylation. Consequently, a distance restraint (3Å) was incorporated anchoring the methoxy group to the heme iron. The docking was repeated, all dockings placing the methoxy group of the substrate in a position amenable to O-demethylation. The GRID analysis showed two interesting things. First, the most negative interaction energies for the NM3 probe in the D301A mutant and the original structure are comparable despite the removal of a negative charge in the D301A active site (-11 kcal mol\(^{-1}\)). Second, the relative propensity for the positively charged NM3 moiety in the D301A mutant is significantly reduced when compared to the original model, signifying that there are less favourable interactions between the active site of the D301A mutant and the NM3 probe. This is exemplified by the fact that there is no longer the correlation between the GRID contour maps and the dockings of codeine (Figure 5.4.3.1B) that is apparent in the docking studies carried out using Model 3 (Figure 5.4.3.1A). This suggests that Asp 301 plays a crucial electrostatic role in the binding of basic substrates by increasing the net negative charge within the active site.

Further studies show that an E216A mutation also leads to a reduction in the volume of the NM3 contour map comparable to that of the D301A mutant (Figure 5.4.3.1C), and a structure that contains both E216A and D301A mutations illustrates virtually no favourable interaction sites for the NM3 probe (Figure 5.4.3.1D). Hence, it appears that both Glu 216 and Asp301 play important roles in the binding of substrates.
Figure 5.4.3.1 The active site of 2D6, illustrating the effects on the binding of basic nitrogen-containing ligands of mutating Glu 216 and/or Asp 301: (A) Model 3, (B) the D301A mutant, (C) the E216A mutant, and (D) the E216A/D301A double mutant. In all cases, the highest ranked GOLD docking of codeine is overlaid on the contour map of the GRID trimethylammonium cation (NM3) probe contoured at $-8\text{ kcal mol}^{-1}$.
Since the docking studies indicate that a salt bridge is formed between a basic nitrogen atom and Glu 216, and Asp 301 does not directly sequester the basic nitrogen atom in substrates, but is still crucial to catalytic activity, Asp 301 must affect substrate binding in a subtler manner. We have already seen that one such way would be electrostatic *i.e.* increasing the net negative charge in the active site. Another way would be to form a hydrogen bond with a residue in the flexible B–C region of the protein. Analysis of all fifteen of the models of 2D6 produced by *Alignment 3* showed that one of the carboxylate oxygen atoms of Asp 301 and the amide groups of Val 119 and Phe 120 are positioned appropriately to form a hydrogen bond. Thus there is a likely interaction between Asp 301 (I-helix) and the B–C region. Analysis of the most homologous structural template lends credence to this argument. The 2C5 crystal structure contains a possible hydrogen bond between Asp 290 (equivalent to Asp 301 in 2D6) and the amide of Ile 112 (equivalent to Val 119 in 2D6) of the B–C region. The temperature factors of the 2C5 crystal structure add additional support to this hypothesis, as the backbone of the inherently flexible B–C region becomes more ordered around Ile 112 (Figure 5.4.3.2). Hence, one of the carboxyl oxygen atoms of Asp 301 could anchor (to some extent) the inherently flexible B–C region. Therefore, Asp 301 could be subject to steric restraints, which would limit its ability to bind substrates directly.

![Figure 5.4.3.2](image)

**Figure 5.4.3.2** Temperature factors of Asp 290 and the B–C region in the 2C5 crystal structure. These are plotted on a scale 0 (blue) to 100 (red) with 50 coloured white. Red indicates a high temperature factor and high relative mobility. Blue indicates a low temperature factor, and low relative mobility.
Further support for the suggested role of Glu 216 in binding the basic moiety of substrate comes from analysis of the alignment of the amino acid sequences of 431 P450s (http://www.icbeg.trieste.it/~P450srv/p450apub.237.html). This showed that only 2.8% (12/431) of all of the P450s analysed have an acidic residue in the position equivalent to Glu 216 in 2D6, compared to 26% (112/431) which have an acidic residue in the position equivalent to Asp 301. There are only four sequences that contain residues equivalent to both Glu 216 and Asp 301: 2D6 (human), 2D14 (bovine), 2D4 (rat) and 2J1 (rabbit). These three 2D enzymes metabolise basic substrates (Mahgoub et al 1977, Matsunaga et al 1990, Tsuneoka et al 1992), with benzphetamine, another basic compound, the preferred substrate for the 2J1 enzyme (Kikuta et al 1991). This is important as it shows that Asp 301 is present in a large number of other P450s that do not metabolise basic substrates. For example, members of the 2C family which are known to hydroxylate steroids such as testosterone and progesterone (Oguri et al 1994, Leemann et al 1993); 2E1 which is responsible for the metabolism of a variety of substances including ethanol (Song et al 1986); 2H1 and 2H2 which metabolise steroids (Hobbs et al 1986); various mammalian orthologues of 1A1 and 1A2 (Kawajiri et al 1986), and 21A1 (Chung et al 1986), which hydroxylate steroids and fatty acids also fall into this category. The rat, bovine and human orthologues of 1B1, which play an important part in steroid regulation for the development of the eye (Sutter et al 1994), 17A1 which hydroxylates pregnenol and progesterone for a variety of vertebrates (Chung et al 1987), and P450-like subfamilies found in plants such as 71A (Bozak et al 1990) which, along with members of the 71C family (Frey et al 1995), act as trans-cinnamic acid hydroxylases also have Asp 301 equivalents. Some other members of the 2D family also contain an Asp 301, but no Glu 216 equivalent, *i.e.* 2D1, 2D2, 2D3, 2D5, 2D9, 2D10, 2D11, and 2D15. Of these only 2D2 contains an acidic equivalent (Asp) to Glu216 (Gonzalez et al 1987), and all are able to metabolise basic substrates such as debrisoquine. Despite this the preferred substrates of the remainder of the 2D family (*i.e.* excluding 2D4, 2D6, and 2D14) are fatty acids and steroids (see *e.g.* Matsunaga et al 1989), and the majority of basic substrates are metabolised preferentially by 2D6. Finally, the N-hydroxylases CYP79 (Koch et al 1995) and CYP83 (Chapple 1995) are also P450s with an equivalent of Asp301, but not Glu 216, and are not know to metabolise exclusively basic substrates.
If both a Glu 216 equivalent and an Asp301 equivalent are present in a P450, basic substrates appear to be preferred; in contrast, Glu 216 alone is not sufficient for this preference. For those systems that contain a Glu 216, but no Asp 301, equivalent the story is very different. Members of the CYP73A family that have a Glu 216, but no Asp 301, equivalent i.e., 73A1 (Teutsch et al 1993), 73A2 (Mizutani et al 1993), 73A3 (Fahrendorf and Dixon 1993), 73A4 (Hotze et al 1995) and 73A9 (Frank et al 1996) are known to preferentially metabolise cinammic acid, whereas 4F2 (Kikuta et al 1999) and 7B (Stapleton et al 1995) metabolise neurosteroids. This suggests that both Asp 301 and Glu 216 are important for the metabolism of basic substrates, and this could be due to the requirement for Asp 301 to position the B–C region in an appropriate position and Glu 216 to interact with the basic nitrogen atom of the substrate, before metabolism can occur. The hypothesis that Asp 301 is not directly involved in the binding of substrates in the 2D6 active site is supported by recent work carried out by Guengerich (Guengerich et al 2002). The role of Glu 216 as a residue crucial to catalytic activity has also been validated by experimental mutagenesis experiments (McLaughlin et al 2002).

5.5 Conclusions

Principal component analysis illustrates that the incorporation of information from the mammalian 2C5 crystal structure has a profound effect on the modelling process, altering the general topology and chemical characteristics of the active site, and that the models produced are significantly different from all of the templates. Docking studies predict that an acidic residue, Glu 216 (in the F-helix [SRS-2] of 2D6), is a major determinant in the binding of basic substrates, and suggest that Asp 301, thought previously to be a direct binding determinant, plays a key role in positioning the B–C region (SRS-1) via hydrogen bonds between its sidechain carboxylate group and mainchain amides in the B–C region. Investigation into the relationship between Asp 301- and Glu 216-containing systems suggests that both residues are necessary for metabolism of a basic substrate. If either residue is missing then the specificity of the enzyme is altered. Hence, Glu 216 does not replace Asp 301 in the binding of basic substrates; both appear necessary for the metabolism of basic substrates.
Chapter 6

Predicting Biological Activities of Cytochrome P450 2D6 Substrates Using Scoring Functions
6.0 Preface

The work in the previous chapter (Chapter 5) elucidated a possible new binding mode for codeine in the active site of 2D6 – subsequently verified by experiment – as a result of incorporating the 2C5 crystal structure into the modelling process. Whether or not the model can predict the biological activity for any ligand (substrate or inhibitor) is addressed in this chapter. In particular, experimental activity data (pIC50) for a number of known inhibitors of 2D6 were compared to values predicted from a number of scoring functions.

6.1 Introduction - Scoring Functions

In recent years a number of computer programs for the docking of ligands into proteins have been described (e.g. Blaney and Dixon 1993, Kuntz et al 1994, Desjarlais et al 1998, Meng et al 1992, Kuntz 1992, Schoichet et al 1993, Lawrence and Davis 1992, Miller et al 1994, Gehlhaar et al 1995, Jones et al 1997, Rarey et al 1996, Morris et al 1996). These programs utilise the 3D structure of a ‘receptor’ protein (determined by X-ray crystallography, NMR spectroscopy or comparative modelling) to construct models of protein-ligand complexes, and thus to give insight into interactions between functional groups in the ligand and residues lining the active site of a protein. A prerequisite for the successful application of these tools is the availability of a scoring function to accurately rank these hits; however the development of a single scoring function that is able to provide a definitive solution to this problem has yet to be discovered.

Ideally a scoring function should be fast, able to prioritise a large list (several thousands) of structurally and functionally diverse ligands, accurate, applicable to a broad range of problems and capable of dealing with small uncertainties in the 3D structure of a target protein (especially comparative models). This investigation sets out to determine whether any of the available scoring functions are capable of predicting the experimental pIC50s (the concentration of a ligand required to reduce the production of a metabolite for a marker substrate by fifty percent) of a set of
known 2D6 inhibitors. As such, the calculation time for these compounds is not of primary importance, and methods such as free energy perturbation theory and/or accurate calculations of the electrostatics (Kollman 1993, Kollman 1994, Warshel et al 1994, Honig and Nicholls 1995) could have been used to predict binding energies. However, the dataset (Appendix 6.1) is structurally diverse, and it was decided to investigate how commonly used scoring functions are able to handle the disparate nature of these compounds.

6.2 Experimental

The dataset was chosen as it represented a structurally and chemically diverse cross-section of known P450 2D6 compounds, an importantly all of the activity data for each of the fifty compounds had been determined using the same in-house standardised technique at AstraZeneca Charnwood. Each of the compounds from the dataset (Appendix 6.1) were docked into the active site of the 2D6 model using the program GOLD (Jones et al 1997). The active site was described as a sphere with radius of 15Å from the Fe atom at the centre of the heme moiety. Twenty dockings were generated for each compound. The highest ranked docking, as determined by the scoring function within the program, was identified and, if it suggested a reasonable 3D structure, the score returned for this docking recorded. If the docking suggested by the highest ranked solution was unreasonable i.e. the ligand occupied a position outside of the active site – the second most highly ranked was investigated, and so on until a reasonable structure was found.

The valid docking solutions for each of the compounds were then re-scored using a range of scoring functions. Once again, the highest ranked docking, according to the scoring function being investigated, was observed and if it suggested a reasonable 3D structure the score returned for this docking recorded. If this docking suggested an unreasonable 3D structure the remaining ranked dockings were analysed in turn until a reasonable structure was identified.

Eleven scoring functions were investigated during this study. The GOLD score (Jones 1997) inherent within the docking algorithm; LUDI score (Böhm 1994) obtained from the Biopolymer module within the program Insight II (Accelrys 2001); the individual
components of the CSCORE function within Sybyl 6.6 (Tripos 2001): namely F-score, (Rarey et al 1999), G-score, (Jones et al 1997) PMF score (Muegge and Martin 1999), D-score (Meng et al 1992) and Chemscore (Eldridge et al 1997); and four scoring functions available in Cerius\(^2\) (Accelrys 2001): Ligscore 1 (Accelrys 2002), Ligscore 2 (Accelrys 2002) PLP 1 (Gehlhaar et al 1995) and PLP 2 (Bouzida et al 1999) (see Chapter 2, Section 2.4).

Having generated this information, linear regression plots for all of the scoring functions investigated were obtained by plotting the highest score returned by a particular function, for each of the compounds, against their experimental pIC\(_{50}\), obtained from inhibition studies of radioactively labelled dextromethorphan (see Section 2.8). Linear regression analysis was then carried out on the plot using the regression module within Excel (Microsoft 2001) to obtain a value for \(r^2\) — a measurement that the data in the \(x\) direction (score from scoring function) is linearly related to that in the \(y\) direction (pIC\(_{50}\)). A high \(r^2\), \textit{i.e.} approaching unity indicates that there is correlation between the two sets of data, and hence the scoring function has been successful. A low \(r^2\), \textit{i.e.} one that approaches zero, is indicative of failure. Spearman ranking coefficients (\(r_s\)) were also calculated. In the absence of definitive linear relationships between score and pIC\(_{50}\) the Spearman coefficient gives insight into whether the relative ranking of the compounds in the dataset has been predicted well.

### 6.3 Results

The results from this investigation were disappointing. The graphs produced for LUDI score vs. pIC\(_{50}\) (Figure 6.3.1A) and PLP 2 vs. pIC\(_{50}\) (Figure 6.3.1B) show that there is no correlation between the scores returned and the experimental data with values for \(r^2\) returned as 0.000 in both cases with the probability of the correlation occurring by random returned as 0.971 and 0.974 respectively.

The results for F-score (Figure 6.3.1C), PLP 1 (Figure 6.3.1D), G-score (Figure 6.3.1E) and the GOLD score (Figure 6.3.1F) do not perform significantly better. Returning \(r^2\) values of 0.001, 0.002, 0.003 and 0.006 along with the probability of the correlation occurring is random of 0.864, 0.778, 0.722 and 0.635 respectively it is
apparent that these scoring functions are limited in their predictive ability with respect to this set of compounds.

Encouragingly the correlations improve significantly for the remainder of the scoring functions. Ligscore 1 returns an $r^2$ value of 0.027 with the probability that the correlation is random reduced to 0.308 (Figure 6.3.1G). Chemscore (Figure 6.3.1H) does slightly better with values of 0.030 and 0.282, and D-Score results are comparable (Figure 6.3.1I) returning results of 0.031 with only 0.271 probability that the correlation is due to chance. The PMF score (Figure 6.3.1J), which is based on statistical as opposed to empirical data, returns an $r^2$ identical to that for D-Score (0.031). However, the probability that the correlation has occurred by chance is slightly reduced in this case – from 0.271 to 0.268.

It is Ligscore 2 (Figure 6.3.1K) that performs best of all on this set of compounds, returning an $r^2$ value of 0.035 with only a probability of 0.24 that the correlation has occurred by chance. However, even this value cannot be considered to be predictive (values of $r^2 > 0.5$ are required (Tame1999)) and as such the results indicate that none of the scoring functions used are able to predict the pIC50s for the compounds in the test set. The results for these experiments are summarised in Table 6.3.1.

As a result of these studies, an investigation into a test set with less structural variance was carried out, to ascertain whether or not better results are achievable with ligands that are structurally more similar. A database of proprietary homologues with biological activities determined in a manner comparable to that for Dataset 1 was deemed most appropriate to use.
Table 6.3.1 Linear regression analysis of scoring functions for Dataset 1

<table>
<thead>
<tr>
<th>Scoring Function</th>
<th>$r^2$</th>
<th>Probability of correlation occurring by random</th>
<th>Spearman Ranking Coefficient ($r_s$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOLD Score</td>
<td>0.006</td>
<td>0.635</td>
<td>-0.04</td>
</tr>
<tr>
<td>LUDI Score</td>
<td>0.000</td>
<td>0.971</td>
<td>0.10</td>
</tr>
<tr>
<td>F Score</td>
<td>0.001</td>
<td>0.864</td>
<td>-0.07</td>
</tr>
<tr>
<td>G Score</td>
<td>0.003</td>
<td>0.282</td>
<td>0.18</td>
</tr>
<tr>
<td>PMF Score</td>
<td>0.031</td>
<td>0.268</td>
<td>0.19</td>
</tr>
<tr>
<td>D Score</td>
<td>0.031</td>
<td>0.271</td>
<td>0.34</td>
</tr>
<tr>
<td>Chemscore</td>
<td>0.030</td>
<td>0.282</td>
<td>0.24</td>
</tr>
<tr>
<td>Ligscore 1</td>
<td>0.027</td>
<td>0.308</td>
<td>0.06</td>
</tr>
<tr>
<td>Ligscore 2</td>
<td>0.035</td>
<td>0.240</td>
<td>0.12</td>
</tr>
<tr>
<td>PLP 1</td>
<td>0.002</td>
<td>0.778</td>
<td>0.0076</td>
</tr>
<tr>
<td>PLP 2</td>
<td>0.000</td>
<td>0.974</td>
<td>0.017</td>
</tr>
</tbody>
</table>
Figure 6.3.1 Linear regression plots for Dataset 1 showing pIC50 vs. scoring function for A) Ludi ($r^2 = 0$), B) -PLP 2 ($r^2 = 0.001$), D) -PLP 1 ($r^2 = 0.002$), E) G-Score ($r^2 = 0.003$), F) GOLD Score ($r^2 = 0.006$), G) Ligscore 1 ($r^2 = 0.027$), H) ChemD-Score ($r^2 = 0.031$), J) PMF Score ($r^2 = 0.031$) and K) Ligscore 2 ($r^2 = 0.035$).
Figure 6.3.1 Continued
Figure 6.3.1 Continued
6.4 Introduction - Nitric Oxide Synthase (NOS)

Nitric Oxide (NO) is an unstable gaseous molecule that, at high concentrations, functions as a cytotoxic agent, and at low concentrations acts as an effector molecule in the nervous, immune and cardiovascular systems (Bredt and Snyder 1994). It is synthesised by nitric oxide synthases (NOSs) – a family of heme-containing enzymes that catalyse the NADPH and O\(_2\) dependent oxidation of L-arginine to form citrullene and nitric oxide (Stuehr 1997). These enzymes are expressed as three isozymes; neuronal (nNOS), endothelial (eNOS) and inducible (iNOS).

Nitric oxide levels need to be controlled, as elevated levels result in pathologies including diabetes, bowel disease, rheumatoid arthritis, multiple sclerosis and transplant rejection, whilst insufficient levels can lead to hypertension, impotence, arteriosclerosis and immunodeficiency (Griffith and Stuehr 1995).

The prosthetic heme group is involved in both steps of the mono-oxygenation reaction required for NO synthesis (Stuehr and Ikeda-Saito 1992), and in common with the P450s the heme is coordinated to a proximal thiolate ligand (Chen et al 1994, Wang et al 1993). P450s catalyse mono-oxygenation reactions of organic compounds with O\(_2\) and electrons (Guengerich 1991), and similarly NOSs produce NO from L-arginine in the presence of O\(_2\) and NADPH. The NO product binds to the iron centre of the heme (Hurshman and Marletta 1995) and the resultant NO-ferric NOS complex is stable in the absence of O\(_2\) – again, analogous to the P450s.

It is logical to assume that any compound displaying an inhibitory effect on NOS enzymes would be likely to inhibit P450s to some extent. This was true for eighteen proprietary NOS inhibitors tested by AstraZeneca (Appendix 6.2). This dataset, in contrast to Dataset 1, contains compounds that have similar structures. It was of interest to ascertain whether the scoring functions would be able to rank the pIC50s of these compounds more accurately than those from the initial test set, and whether or not this would indicate a weakness with the scoring functions for dealing with structurally diverse compounds.
6.4.1 Experimental

Each compound from the second dataset (Appendix 6.2) was docked into the active site of the 2D6 model using the computer program GOLD (Jones et al 1997) to give twenty possible solutions. The score for the highest ranked docking, for each compound was recorded and. If the docking predicted by the highest ranked solution was unreasonable i.e. the ligand occupied a position outside of the active site, the second most highly ranked solution was investigated, and so on until a reasonable structure was found.

The 'reasonable' docking solutions for each of the compounds were then subject to being re-scored (Section 6.2). Once again, the score for the highest ranked 'reasonable' docking, according to the particular scoring function, was recorded.

Once all the data had been collected, linear regression plots were produced, as in Section 6.2.

6.4.2 Results

The results of the linear regression analyses for the NOS inhibitor dataset (Appendix 6.2) are summarised in Table 6.4.2.1

Table 6.4.2.1 Linear regression analysis for Dataset 2

<table>
<thead>
<tr>
<th>Scoring Function</th>
<th>$r^2$</th>
<th>Probability of Correlation Occurring by Random</th>
<th>Spearman Rank Coefficient ($r_s$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOLD Score</td>
<td>0.012</td>
<td>0.674</td>
<td>-0.04</td>
</tr>
<tr>
<td>LUDI Score</td>
<td>0.071</td>
<td>0.300</td>
<td>0.39</td>
</tr>
<tr>
<td>F Score</td>
<td>0.188</td>
<td>0.082</td>
<td>0.46</td>
</tr>
<tr>
<td>G Score</td>
<td>0.035</td>
<td>0.475</td>
<td>0.34</td>
</tr>
<tr>
<td>PMF Score</td>
<td>0.095</td>
<td>0.228</td>
<td>0.41</td>
</tr>
<tr>
<td>D Score</td>
<td>0.002</td>
<td>0.850</td>
<td>0.012</td>
</tr>
<tr>
<td>Chemscore</td>
<td>0.007</td>
<td>0.734</td>
<td>-0.02</td>
</tr>
<tr>
<td>Liscore 1</td>
<td>0.000</td>
<td>0.986</td>
<td>-0.24</td>
</tr>
<tr>
<td>Liscore 2</td>
<td>0.012</td>
<td>0.676</td>
<td>0.06</td>
</tr>
<tr>
<td>PLP 1</td>
<td>0.007</td>
<td>0.743</td>
<td>0.03</td>
</tr>
<tr>
<td>PLP 2</td>
<td>0.006</td>
<td>0.771</td>
<td>0.04</td>
</tr>
</tbody>
</table>

158
These results are encouraging. By reducing the structural and chemical diversity between the compounds being investigated an improvement in the $r^2$ values for seven of the eleven scoring functions is observed. The most dramatic of these are the improvement in values for F score (Dataset 1: 0.001, 0.864 Dataset 2: 0.188, 0.082) (Figure 6.4.2.1) and LUDI score (Dataset 1: 0.000, 0.971 Dataset 2: 0.071, 0.300) (Figure 6.4.2.2). This improvement is reflected in the cumulative frequency diagram (Figure 6.4.2.3), which shows that Dataset 2 has a greater number of scoring functions that are statistically significant, at the 90% level for Spearman rank correlation, when compared to Dataset 1.

Conversely, the remaining four scoring functions return values for $r^2$ and the probability of the correlation occurring by random significantly worse than those for the first data set. Of these, Ligscore 1 (Figure 6.4.2.4) shows that there appears to be no correlation at all (0.000, 0.986).

![Linear regression plot showing pIC50 vs. F-Score predictions for Dataset 2 ($r^2 = 0.188$)](image)

**Figure 6.4.2.1** Linear regression plot showing pIC50 vs. F-Score predictions for Dataset 2 ($r^2 = 0.188$)
Figure 6.4.2.2 Linear regression plot showing pIC50 vs. LUDI predictions for Dataset 2 ($r^2 = 0.071$)

Figure 6.4.2.3 Cumulative frequency graph. This shows that a greater number of scoring functions perform to a higher level of statistical significance for Dataset 2 (triangles) as opposed to Dataset 1 (circles)
6.4.3 Investigation of Alternative Docking Algorithms

To ascertain whether or not the poor results were as a result of the docking algorithm, an alternate program, Autodock (Morris et al 1996), was used. This investigation proved unsuccessful – possibly due to the limited ability to parameterise the heme group within this program. A visual inspection showed that even for the highest ranked dockings the sites of metabolism on the codeine molecule were distant from the heme iron (Figure 6.4.3.1). A comparison of the interatomic Fe-H distances for codeine docked into the active site – identical to the one carried out in Chapter 5 – confirmed that these dockings were poor as the distances did not correlate with those obtained from NMR experiments (Table 6.4.3.1). FlexX (Rarey et al 1996) was also available, but once again lack of parameterisation for the heme group resulted in a generally poor performance across both datasets. As a result of these findings, and in the absence of a suitable alternative docking algorithm, this investigation was abandoned.
Figure 6.4.3.1 The highest ranked docking of codeine into 2D6 active site using Autodock (Morris et al 1998) shows the docking is poor as there is no association of the methoxy group that undergoes demethylation with the Fe atom at the centre of the heme.
Table 6.4.3.1 Comparison of interatomic Fe-H distances for codeine docked into CYP 2D6 using Autodock with experimentally determined interatomic Fe-H distances.

<table>
<thead>
<tr>
<th>Proton</th>
<th>Experimental</th>
<th>Average</th>
<th>Difference Between Experimental and Average Distance in Models* (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1H</td>
<td>7.5 ± 0.2</td>
<td>7.3 ± 2.5</td>
<td>0.2 ± 2.5</td>
</tr>
<tr>
<td>C2H</td>
<td>5.0 ± 0.1</td>
<td>7.8 ± 2.8</td>
<td>2.8 ± 2.8</td>
</tr>
<tr>
<td>-OCH₃</td>
<td>3.1 ± 0.1</td>
<td>8.3 ± 3.2</td>
<td>5.2 ± 3.2</td>
</tr>
<tr>
<td>C5H</td>
<td>9.1 ± 0.1</td>
<td>8.2 ± 2.1</td>
<td>0.9 ± 2.1</td>
</tr>
<tr>
<td>C6H</td>
<td>10.0 ± 0.2</td>
<td>9.4 ± 1.8</td>
<td>0.6 ± 1.8</td>
</tr>
<tr>
<td>C7H</td>
<td>9.3 ± 0.3</td>
<td>9.1 ± 2.3</td>
<td>0.2 ± 2.3</td>
</tr>
<tr>
<td>C8H</td>
<td>10.2 ± 0.2</td>
<td>8.6 ± 1.6</td>
<td>1.4 ± 1.6</td>
</tr>
<tr>
<td>C9H</td>
<td>11.2 ± 0.2</td>
<td>8.2 ± 1.1</td>
<td>3.0 ± 1.1</td>
</tr>
<tr>
<td>C10H₂</td>
<td>9.8 ± 0.1</td>
<td>7.3 ± 2.1</td>
<td>2.5 ± 2.1</td>
</tr>
<tr>
<td>C15Hₐ</td>
<td>9.3 ± 0.1</td>
<td>8.5 ± 1.7</td>
<td>0.8 ± 1.7</td>
</tr>
<tr>
<td>C15Hₐ</td>
<td>10.8 ± 0.2</td>
<td>9.1 ± 1.4</td>
<td>1.7 ± 1.4</td>
</tr>
<tr>
<td>NCH₃</td>
<td>12.1 ± 0.2</td>
<td>8.0 ± 2.1</td>
<td>4.1 ± 2.1</td>
</tr>
</tbody>
</table>

* The mean distance and standard deviation derived from fifteen dockings of codeine into the 2D6 active site.
6.5 Discussion

On first inspection, the results from the investigations into the scoring functions seem disappointing, because none return values of $r^2$ that suggest that the 2D6 model is predictive. The hypothesis that the failings of the scoring functions for Dataset 1 (Appendix 6.1) are due to them being unable to cope with diverse structural nature of the compounds is not fully substantiated by the findings from Dataset 2, where a much smaller dataset of structurally similar compounds was used (Appendix 6.2) – and yet only seven of the eleven scoring functions showed an improvement for $r^2$, none approaching values that could be considered predictive ($r^2 > 0.5$ (Tame 1999)).

6.5.1. D-Score

Considered in isolation, each scoring function has its own relative strengths and weaknesses. D-Score does not take into account the entropic changes that occur when protein-ligand binding takes place. Therefore, it is feasible that D-Score will be limited because of the vast structural diversity of compounds. For example, the binding of a large flexible compound, such as ketoconazole, will result in a far greater decrease in entropy than if a smaller, more rigid compound, e.g. codeine, was to bind – a phenomenon ignored by this scoring function.

However, D-Score is the only scoring function that returns a value for the Spearman Rank coefficient significant to 95% for Dataset 1. The corresponding value for Dataset 2 shows a significance of less than 50%. This could be due to other limitations of the scoring function. Investigations (Knegtel et al 1999) suggest that there is a tendency for charge-charge and other polar interactions to dominate when D-Score is determined. Therefore, any compounds with a greater number of polarisable atoms might be expected to score more highly than those with fewer polarisable atoms. Dataset 1 – with an average of 16% polarisable atoms in each compound tends to score lower than Dataset 2 – with an average of 29% polarisable atoms per compound. Hence, the hypothesis is supported. However, Dataset 2 performs poorly compared to Dataset 1 – this is perhaps unexpected, as we would
imagine the D-Score to perform better when presented with molecules capable of a greater number of charge-charge and other polar interactions. Insight is gained when we examine the dockings of the compounds from Dataset 2 and contrast them with Dataset 1 (Figure 6.5.1.1). The orientations adopted by the structurally similar compounds of Dataset 2 are almost identical. Hence, the scores returned for each compound will be similar – as the polar interactions that dominate in D-Score will swamp the information from the subtle alterations in the structures that account for the differences in biological activity. It is feasible that it would be difficult to distinguish between them using D-Score – and so poor results are obtained from both linear regression analysis and Spearman ranking. In contrast, the structural differences between corynanthine and carbamazepine from Dataset 1 mean they adopt very different orientations in the active site. Corynanthine – which is able to achieve some polar protein-ligand interactions scores much more highly than carbamazepine, which forms none. D-Score is better able to distinguish between these structurally different compounds – resulting in higher values for both the linear regression analysis and Spearman ranking.

**Figure 6.5.3.1** The highest ranked GOLD dockings of A) corynanthine (orange) and carbamazepine (yellow) and B) Compounds 1-5 into the 2D6 model, highlighting the disparate binding orientations of the Dataset 1 compounds in comparison to those from Dataset 2.
When D-score was originally developed it was intended to be able to distinguish between tight and weak binding ligands from a database of thousands, with widely varied binding affinities. The failure with respect to Dataset 2 is therefore not unexpected as there is a limited range of pIC50s across the compounds, coupled with little structural diversity.

6.5.2. Chemscore

Chemscore (Eldridge et al 1997) performs amongst the best of the scoring functions for Dataset 1 ($r^2 = 0.030$) but is significantly worse for the smaller Dataset 2 ($r^2 = 0.007$).

It is an empirical scoring function derived from databases of X-ray structures of protein-ligand complexes. Inherent limitations of this function are a failure to parameterise weak binders, as there is an absence of data in the databases for such complexes, and application to a wide variety of chemistries and/or uncommon binding geometries which are not well-represented in the database. Although the first problem is ameliorated by using as large a database as is possible i.e. high-resolution structures representative of all types of protein-ligand complexes available in the Protein Databank, applicability of Chemscore to Dataset 1 is limited; the Spearman ranking coefficient does show that the prediction of the general trend of activity is significant to 90%.

A lack of stringent parameterisation within the scoring function may be responsible for the disappointing results for Dataset 2. In an effort to ensure Chemscore was applicable to a disparate range of chemistries with minimal human intervention the authors produced a scoring function that was not so tightly parameterised that it would fail to give meaningful results if it encountered new functional groups. However, this could have been at the expense of the function being able to distinguish between structurally and chemically similar compounds.
6.5.3 PMF Score

The PMF score (Muegge and Martin 1999) is based upon ligand-receptor atom pair interactions that are statistical, as opposed to empirical, in nature derived from crystallographic data in the protein databank (Berman et al 2000). For Dataset 1 this scoring function does relatively well ($r^2 = 0.031$, $r_s = 0.19$, significant to 70%) showing that it is able to cope with a diverse dataset. The results are also encouraging for Dataset 2 ($r^2 = 0.095$, $r_s = 0.41$ significant to 90%). Despite this relatively good performance the values returned indicate that the model is not predictive. This may be due to the nature of the datasets.

A flaw of both empirical and statistical scoring functions is that it is unclear to what extent they can be applied to protein-ligand complexes not in the initial training set. Also, as this work is based on a comparative model, as opposed to an X-ray crystal structure, it is likely that despite our best efforts to ensure as accurate a model as possible, areas that are ambiguous across the P450s, and as a result, not well modelled — such as the F-G loop, and the B–C region — will give misleading information with regards to ligand-receptor atom pair interactions, thus limiting the usefulness of the PMF Score.

6.5.4 GOLD Score and G Score

Both the GOLD score inherent in the docking algorithm and the G score component of CSCORE (Jones et al 1995, Jones et al 1997) are founded on hydrogen bonding interactions. These functions perform well if there are significant polar interactions between protein and ligand, and less well when the ligand is largely hydrophobic in nature. Since the only significant polar interactions in this case are the interactions between any basic groups (usually a nitrogen atom) on the ligands, and the acidic Glu 216 and Asp 301 residues (see Chapter 5) it is unsurprising that the GOLD/G-Score functions struggle to rank the ligands in Dataset 1 ($r^2 = 0.006$ and 0.003 respectively). For both scoring functions there is a marked improvement in the ability to rank the ligands in Dataset 2 ($r^2 = 0.012$ and 0.035), which is unsurprising when considering the dataset consists of compounds that are on average more polar than those of Dataset 1.
6.5.5. Ligscore 1 and Ligscore 2

Ligscore 1 and Ligscore 2 (Accelrys 2002) are fast and very simple scoring functions used to predict protein-ligand binding affinities. Primarily they are intended to identify potential lead candidates from a large database of ligands – and this suggests that they may be limited when it comes to ranking chemically and structurally similar ligands in order of their binding affinities. This hypothesis is supported by the scoring functions performing significantly better, although still not predictively, for the larger, more diverse Dataset 1 than the smaller more limited Dataset 2.

6.5.6. Piecewise Linear Potential (PLP) 1 & 2

PLP1 (Gehlhaar et al 1995) and PLP2 (Bouzida et al 1999) were devised to rank the docked conformations for a single ligand in an active site. Both scoring functions fail to predict the trend for the diverse Dataset 1 ($r^2 = 0.002$ and 0.000 for PLP 1 and PLP 2 respectively) – and this could be due to the diversity of the dataset. Different ligands bind in very different orientations, and it appears that these scoring functions are unable to cope with this diversity. The limited structural and chemical diversity of Dataset 2 seems better suited to these scoring functions, as there is a marked improvement in the performances of PLP 1 and PLP 2 ($r^2 = 0.007$ and 0.006 respectively). This is unsurprising, as for Dataset 2 the scoring functions are required to distinguish between the differences in binding orientations of structurally similar compounds – the task for which they were initially developed.

6.5.7 LUDI Score and F-Score

The LUDI score (Böhm 1994) assesses protein-ligand interaction on the basis of a lipophilic term, a metal-ligand binding term, a hydrogen bonding term, and a term to account for loss of ligand flexibility. F-Score (Rarey et al 1996, Rarey et al 1999) is based on LUDI, but incorporates changes with regards to the parameterisation of the four individual LUDI terms. However, both scores are amongst the worst performers for Dataset 1 – where they completely fail to rank the compounds with respect to their pIC50s. However, this is reversed for Dataset 2, where the $r^2$ and Spearman Ranking
values returned for LUDI and F Score are better than the majority of the others \( (r^2 = 0.071 \text{ and } 0.188 \text{ respectively}, r_s = 0.39 \text{ significant to } 90\% \text{ and } 0.46 \text{ significant to } 95\% \text{ respectively}) \).

The difference between the performance of Dataset 1 and Dataset 2 can be attributed to the distinction between the prerequisites of a docking algorithm and a scoring function. A docking algorithm – such as GOLD (Jones et al 1997) seeks to place a small molecule in a favourable orientation within a receptor – and as such allows small steric clashes by incorporating soft Lennard-Jones potentials which allow atoms to approach each other to distances less than the sums of their van der Waals radii – although it does penalise any dockings of this nature in terms of overall fitness.

Despite this, the most highly ranked docking may contain several atoms in the ligand that have some degree of steric clash with the receptor. Ludi score does not take this into account – merely assigning a score to any interaction between the ligand and the receptor – hence it does not penalise clashes. It is therefore feasible that ligands with a greater molecular weight – which potentially have more ligand receptor interactions due to their increased size – would receive higher Ludi scores than smaller ligands.

By performing a linear regression on Dataset 1 of Ludi score vs. molecular weight (Figure 6.5.7.1) \( r^2 = 0.62 \) and Ludi score vs. number of heavy atoms (Figure 6.5.7.2) \( r^2 = 0.70 \) we can see that this is the case; there appears to be a correlation between the size of the ligand and the Ludi score it achieves. It does not necessarily follow that larger compounds will bind more tightly to the receptor than smaller ligands - a linear regression carried out on Dataset 1 shows that there is no correlation between pIC50 and molecular weight (Figure 6.5.7.3) \( r^2 = 0.04 \).

There is no such trend for Dataset 1 with respect to F-Score. Linear regressions of F-Score vs. molecular weight, and F-Score vs. the number of heavy atoms in a compound yield \( r^2 \) values of 0.006 and 0.004 respectively. Therefore, there must be another reason for the failure of the scoring function in this instance. The major difference between Ludi and F-Score is the penalty employed by F-Score when calculating the contribution to binding from lipophilic elements (Chapter 2, Section 2.4.9). Any steric clashes between the receptor and the ligand result in a lower F-Score being returned – therefore, the pIC50 of larger ligands which are tight binders.
may be underestimated by this function because of the limited space within the active site of the model i.e. they are subject to a greater number of clashes. This appears to be the case. The weakly binding methyltetrahydroisoquinolinone from Dataset 1 (pIC50 = 4.70) is ranked as the most tightly binding molecule by F-Score due to the relatively low number of steric clashes between receptor and ligand. Conversely, lobeline (pIC50 = 6.60) performs poorly. It is a much larger molecule and suffers a greater number of steric clashes with the receptor. Hence, it is possible F-Score underestimates the binding affinity because of these clashes (Figure 6.5.7.4).

![Figure 6.5.7.1](image)

**Figure 6.5.7.1** Linear regression plot of molecular weight versus LUDI score ($r^2 = 0.62$)

![Figure 6.5.7.2](image)

**Figure 6.5.7.2**. Linear regression plot of LUDI score vs. number of heavy atoms in a compound for Dataset 1 ($r^2 = 0.70$)
Figure 6.5.7.3 Linear regression plot of molecular weight vs. pIC50 for Dataset 1 ($r^2 = 0.04$)

Figure 6.5.7.4 GOLD dockings for A) methyltetrahydroisoquinolinone (cyan) and B) lobeline (magenta) illustrating that larger molecules are subject to a greater possible number of steric clashes when docked into the 2D6 active site.
The limited variation in molecular weight and number of heavy atoms between the compounds of Dataset 2, and the common binding orientation displayed by the majority of the highest ranked dockings means that Ludi score is not able to rank compounds making vastly different numbers of ligand-receptor contacts, and any penalties due to clashes in F-Score will be of approximately the same magnitude for each compound. In these cases the differences in the scores will be as a result of subtle differences in the structure of the compounds as opposed to larger differences in the number of protein-ligand contacts. Hence, Ludi and F-Score are better able to rank these compounds.

However, the fact remains that neither function is predictive, so the scoring functions are limited even with respect to the smaller dataset. There is still a good correlation between molecular weight and the value returned by the Ludi score ($r^2 = 0.68$) for Dataset 2, which could account for the poor performance.

The range of activities (pIC50) for both datasets is limited in comparison to the datasets used to validate the scoring functions. For example, in a paper describing improvements to the original LUDI score (Böhm 1998) a set of 82 protein-ligand complexes that have log $k_i$ values distributed over a range of fifteen log units are investigated. When plotting a graph of experimental vs. predicted $k_i$ for this dataset an $r^2$ value of 0.79 results. However, it is possible that the data in this case is overfitted. By examining a smaller range of log $k_i$ values — over a range of four log units, akin to Datasets 1 and 2 — we can see that there is a significant reduction in the $r^2$ value which is obtained. By plotting a section of this data for forty-two compounds with experimental -log $k_i$ values between 4 and 8 (Böhm 1998) against the values predicted by Böhm’s experiment (Figure 6.5.7.5) and performing a linear regression we obtain a value for $r^2$ of 0.221 — a value which is indicative that the model is not predictive, and comparable to the $r^2$ value obtained for Dataset 2 with F-Score (0.188). This suggests that despite the disappointing results from the analysis of the scoring functions, it is possible they have done as well as could have been expected. Indeed, the fact that the $r^2$ values from a set of ligand-receptor crystal structures is comparable to that found from a docking experiment into a comparative model lends support to the model being of high quality.
Figure 6.5.7.5 Regression of 42 compounds ($r^2 = 0.221$) from work published by Böhm (Böhm 1997) that shows over a narrower range of $K_i$ (comparable to Dataset 1) the values obtained for $r^2$ are significantly reduced.

Scoring functions in general also make several assumptions that do not stand up to rigorous investigation (Tame 1999). To save computational time none of the scoring functions investigated treat the protein active site as flexible, and ignore possible conformational changes in the protein upon ligand binding. They also assume that each basic type of interaction – such as hydrogen bonding between donor and acceptor – are equivalent. This, obviously, is not the case as there is often cooperation between different individual interactions that increase binding affinity. Also, scoring functions treat each protein-ligand interaction as additive, despite evidence to suggest that this approach is flawed (Jencks 1981) and that in some cases the sum of contributions of individual residues greatly exceeds the free energy of binding.

However, scoring functions can only be as accurate as the affinity data used to develop them. Despite major advances in the area of scoring functions over the past ten years there has been little change in the manner in which binding affinities are
measured. The problem inherent in the development of an accurate and generally applicable scoring function lies in the lack of understanding we have of the biophysics and thermodynamics of ligand binding by biological molecules. The highly complex interaction of water is also poorly understood – solvation of a protein-ligand complex tends to be very different when compared to the individual solvation of the separate protein and ligand. It has been observed that in some cases, such as for antibody antigen binding (Bhat et al 1994, Tame et al 1994) the presence of water molecules increases the binding affinity. However, this is not a universally applicable rule, as in some cases the expulsion of water from the active site favours ligand binding (Pearlman and Connoly 1995). Until we are able to better understand these complex phenomena it will be impossible to generate a single scoring function that it is applicable and accurate across a wide-range of proteins and ligands. As such, it is possible that this investigation asked too much of the eleven scoring functions studied, and they performed as well as could be expected under the circumstances.

6.6 Conclusions

Although it is acknowledged that no single scoring function is the definitive solution to predicting the binding affinities for a diverse set of ligands, each of the scoring functions used have previously performed significantly better with diverse sets of protein ligand complexes in comparison to the results obtained for these studies.

The first major difference between these experiments, and the majority of those used to validate the scoring functions is that we are using a homology model as our protein, whereas validation techniques use crystal structures and generally set out to reproduce a known binding orientation from a high-resolution protein-ligand complex. The inherent problem that arises here is that comparative models are less accurate than crystal structures. Due to the lack of refinement in the model, important protein-ligand hydrogen bonding geometries may be strained and/or missing, the area of the protein where the ligand binds could be modelled badly due to the lack of a definitive template structure, and the active site — although a good approximation — will probably not be a totally accurate representation. All these factors could lead to errors within the scoring function, and could be used in part as a reason for the disappointing results of the experiments.
The binding orientations of the majority of the ligands investigated are also unknown – hence it is possible that we are attempting to predict correct binding affinities from incorrect binding orientations and geometries. This would be because of a flaw in the model and/or the docking algorithm.

This would lead to a conclusion, despite rigorous efforts to the contrary, that the model of CYP 2D6 produced was not a valid one. However, an analysis of the literature shows that this is not necessarily the case. It is the limited range of activities in both datasets that appears to be responsible for the poor performance. Comparison between work carried out as a validation of Ludi score (Böhm 1998) and the results obtained from these experiments show that over a limited range of activities, for one case (F-Score) the data from the models produced perform as well as the predictions for the literature validation set ($r^2$ of 0.188 and 0.221 respectively). This is indicative of the model being of high quality – hence the reason that the scoring functions are not predictive is likely not due to limitations with regards to the model, but inherent limitations due to the datasets, the scoring functions themselves, and a lack of understanding of the thermodynamics and biophysics of protein-ligand complex formation.

In retrospect, this experiment appears to be ambitious, as there are several assumptions that have been made which may not hold up under rigorous investigation. The first is with regards to the quality of the 2D6 model. Although rigorous quality checks have shown that the model is of a high standard, the fact remains that there is no direct evidence to suggest that the model is an accurate representation of the 2D6 active site. All scoring functions are developed using high-resolution crystal structures, and as such are not parameterised to account for deviations from the true structure. By using a homology model to conduct these experiments, which is highly likely to incorporate errors with regard to the three-dimensional structure of the protein, we are making it increasingly difficult for the scoring functions to operate effectively.

Added to this we have assumed that the docking algorithm has produced accurate conformations for the binding orientations of the compounds in the dataset within the
active site. Without a crystal structure to compare these docking results to we have no idea whether or not what we assume to be the correct orientation as determined by the docking algorithm approaches that which we could expect to see in an experimental structure. Incorrect binding orientations of compounds introduces another possible source of error into the experiment – the effectiveness of the scoring functions would, unsurprisingly, be severely hampered by incorporation of incorrect receptor-ligand complexes into the analysis – and highlights once again the limitations of this investigation without experimental evidence to validate the results.

Finally the limitations due to the scoring functions themselves would also introduce error into such an analysis. These are discussed below. With sources of error arising from so many sources, it is understandable that the results for the experiment were disappointing – and by reducing the number of variables by utilising information derived by experimental structures may have given greater insight into the relative strengths and weaknesses of the scoring functions – before such an ambitious experiment was attempted.

In the absence of a scoring function that is able to accurately predict activities of compounds, other methods need to be investigated to assess the predictive nature of the 2D6 model.
Chapter 7

QSAR Studies on Substrates of Cytochrome P450 2D6
7.0 Preface

In the previous chapter several scoring functions were used to predict the activity of a known set of 2D6 inhibitors/substrates. This technique focuses on the structure of the receptor-ligand complex formed from docking studies, and was largely unsuccessful. An alternative method of predicting biological activity, quantitative structure-activity relationships (QSAR), was investigated.

7.1 Introduction: Quantitative Structure-Activity Relationships (QSAR)

In drug design it is important to realise that a molecule may be required to show qualities beyond in vitro potency, as it is the in vivo efficacy that will make it successful as a therapeutic agent. A structure-activity study can aid in the understanding of which features of a molecule are responsible for activity, and how they could be modified to enable the production of enhanced compounds. Quantitative structure-activity relationships (QSARs) relate numerical properties of a molecular structure to the activity of a molecule via a mathematical model. As a result, QSAR can be focussed entirely on the ligands, as opposed to ligand-protein interaction, and may be better able to distinguish between what features make one ligand more active than another. Therefore, several QSAR investigations were carried out on Datasets 1 and 2 (Appendices 6.1 and 6.2). Five methods of QSAR investigated during these studies; Molecular Field Analysis (Cramer 1988, Accelrys 2002), Receptor Surface Model (Hahn 1995 Hahn and Rogers 1995, Hahn et al 1997), Cerius² descriptor-based QSAR (Accelrys 2002) CoMFA (Cramer 1988, Tripos 2002) and Hologram QSAR (Hurst and Heritage 1997, Hurst et al 1998, Tong et al 1997), see also Chapter 2.

7.2 Experimental

With the exception of Hologram QSAR, which is a two-dimensional technique, the 3D structural ‘consensus’ alignment of the ligands under investigation is central to the result of a QSAR. The orientations of the highest-ranked dockings for all of the compounds from Datasets 1 and 2, according to GOLD, provide invaluable insight as to where important interactions between the ligands and the active site of 2D6 occur,
hence one structural alignment for each dataset was already available, the 'docked' alignment (Figure 7.2.1). However, it is possible that the orientations of the compounds in the active site are dictated by steric and electronic influences of the protein surrounding the ligands, and that the alignment of structures generated in this manner may give misleading results. Therefore, an alternative 'undocked' alignment of the compounds in Datasets 1 and 2 was obtained (Figure 7.2.2). Using the program Cerius² (Accelrys 2002), a target molecule alignment was carried out on each dataset, structurally aligning all of the structures to that of the tightest-binding molecule (ajmalicine in Dataset 1 and ARC155688AP in Dataset 2).

Figure 7.2.1 The GOLD generated ‘docked’ alignments for compounds in A) Dataset 1 and B) Dataset 2.

Figure 7.2.2 The ‘undocked’ Cerius² generated alignments for compounds in A) Dataset 1 and B) Dataset 2.
7.2.1 Molecular Field Analysis (MFA)

Molecular field analysis (Cramer et al 1988, Accelrys 2002) calculates the interaction energy of a chemical probe over a set of aligned ligands. For both alignments of Dataset 1, probe maps with energy values between −30 and +30 kJmol⁻¹ were calculated on a cuboid grid, with a grid spacing of 2.0 Å. A total of 2490 Cartesian grid points were investigated using all the available predefined Cerius² probes, H⁺, CH₃, CH₃⁺ and CH₃⁻.

This was repeated for both alignments of Dataset 2. Again, probe maps with energy values between −30 and +30 kJmol⁻¹ were calculated on a cuboid grid, with a grid spacing of 2.0 Å. However, only a total of 512 Cartesian grid points were investigated using all of the predefined Cerius² probes, H⁺, CH₃, CH₃⁺ and CH₃⁻. The reduction in the total number of grid points is due to the structural similarities of the compounds in the second dataset. The 3-dimensional structures of these ligands align to a greater extent than those in Dataset 1, and therefore a relatively smaller grid, with fewer grid points, is required in order to encompass the aligned ligands. Once this data was generated a QSAR model was obtained using Genetic Partial Least Squares (GPLS) (Dunn and Rogers 1996) to give an r² value. This value was cross-validated by the ‘leave-one-out’ method to give a value for q², a measure of how robust the model is.

7.2.2 Receptor Surface Model (RSM)

A receptor surface model (Hahn 1995, Hahn and Rogers 1995, Hahn et al 1997) is usually generated in an attempt to predict the biological activity of a compound. However, it also has application in QSAR, in that it can be used to relate calculated energies to the biological activity (pIC50 in this case). Subsets of the most active compounds from both alignments of the two datasets were used to build the receptor surface models.

For Dataset 1 the three compounds with the highest pIC50s – ajmalicine, pentazocine and hydroquinidine – were used to construct the receptor surface model for the docked and undocked ligand orientations. A solvation correction, which applies a penalty function for placing polar atoms in hydrophobic regions, was applied to the...
receptor surface, as was a function characterising the van der Waals shape of the molecules, which helped to clearly define atom positions. The receptor surface model was then used to calculate energy values, Energy $E_{\text{strain}}$ and $E_{\text{interact}}$ (See Chapter 2, Section 2.5.3) for all of the ligands. This data was then imported into the QSAR study table. The Receptor energies descriptor in Cerius$^2$ was then added to the study table, before a QSAR using the Genetic Function Approximation (GFA) was carried out to determine whether or not there was a correlation between the 'goodness' of fit for a ligand in the receptor surface model, and the pIC50 data via determination of $r^2$ and $q^2$.

The treatment for the compounds in Dataset 2 was identical to that employed for Dataset 1. The only exception was the two compounds with the highest activity in Dataset 2, ARC155688AP and ARC154755AP, were used to generate the receptor surface models.

### 7.2.3 Cerius$^2$ Descriptor QSAR

A descriptor is a molecular property that can be used to determine a QSAR relationship. Within Cerius$^2$ they are divided into the following functional families: fragment constant; conformational; electronic; receptor; quantum mechanical; graph theoretic; topological; information-content; molecular shape analysis; spatial; structural; thermodynamic and pKa descriptors. Altogether there are over 100 different descriptors, although not every descriptor should be included in the analysis as it has been shown that over abundance of information which is non-essential can lead to a reduction in the quality of the QSAR (Ekins and Obach 2000).

Of the original 100 descriptors eighteen were selected from the QSAR database (Table 7.2.3.1) in Cerius$^2$. These were chosen via a process of elimination from an analysis of the 'undocked' Dataset 2 ligands, which removed cross-correlated descriptors, as determined by a cross-correlation matrix, up to a point where removal of descriptors had a detrimental effect on the quality of the QSAR. They represent a diverse and largely unrelated set of descriptors that can be computed very rapidly. The descriptors were calculated for all ligands in both docked and undocked orientations.
Table 7.2.3.1 Cerius2 descriptors used to generate the 2D6 QSAR for Datasets 1 and 2.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Descriptor Family</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoG</td>
<td>Radius of Gyration</td>
<td>Spatial</td>
</tr>
<tr>
<td>Area</td>
<td>Molecular Surface Area</td>
<td>Spatial</td>
</tr>
<tr>
<td>PMI</td>
<td>Principal Moment of Inertia</td>
<td>Spatial</td>
</tr>
<tr>
<td>Vm</td>
<td>Molecular Volume</td>
<td>Spatial</td>
</tr>
<tr>
<td>MW</td>
<td>Molecular Weight</td>
<td>Structural</td>
</tr>
<tr>
<td>Rot Bond</td>
<td>Number of Rotatable Bonds</td>
<td>Structural</td>
</tr>
<tr>
<td>Hbond Acceptor</td>
<td>Hydrogen Bond Acceptor Groups</td>
<td>Structural</td>
</tr>
<tr>
<td>Hbond Donor</td>
<td>Hydrogen Bond Donor Groups</td>
<td>Structural</td>
</tr>
<tr>
<td>AlogP</td>
<td>Ghose and Crippen LogP</td>
<td>Thermodynamic</td>
</tr>
<tr>
<td>FH2O</td>
<td>Desolvation Free Energy for Water</td>
<td>Thermodynamic</td>
</tr>
<tr>
<td>MR</td>
<td>Molecular Refractivity</td>
<td>Thermodynamic</td>
</tr>
<tr>
<td>Foot</td>
<td>Desolvation Free Energy for Octanol</td>
<td>Thermodynamic</td>
</tr>
<tr>
<td>HF_MOPAC</td>
<td>Heat of Formation</td>
<td>Quantum Mechanical</td>
</tr>
<tr>
<td>Apol</td>
<td>Sum of Atomic Polarisabilities</td>
<td>Electronic</td>
</tr>
<tr>
<td>Sr</td>
<td>Superdelocalisability</td>
<td>Electronic</td>
</tr>
<tr>
<td>LUMO_MOPAC</td>
<td>Lowest Unoccupied Molecular Orbital</td>
<td>Quantum Mechanical</td>
</tr>
<tr>
<td>DIPOLE_MOPAC</td>
<td>Dipole Moment</td>
<td>Quantum Mechanical</td>
</tr>
<tr>
<td>HOMO_MOPAC</td>
<td>Highest Occupied Molecular Orbital</td>
<td>Quantum Mechanical</td>
</tr>
</tbody>
</table>

for both of the datasets. Genetic Partial Least Squares fitting (GPLS) was used to determine the relationships between the descriptors and the pIC50s of the compounds. Values for $r^2$ and $q^2$ were obtained as before.

7.2.4 Comparative Molecular Field Analysis (CoMFA)

CoMFA (Cramer et al 1988, Tripos 2002) is similar to MFA in that it uses a grid/probe-based approach. However, the exact parameterisation of the probe types is different to those used in the MFA analysis. For both alignments of compounds in each dataset a grid was constructed that encompassed all the ligands in all directions,
and the interaction energy of the only available probe CH$_3^+$ calculated at each grid point. The resultant field matrix for each investigation was analysed using Partial Least Squares (PLS) (See Chapter 2, Section 2.7.3).

7.2.5 Hologram QSAR (HQSAR)

Hologram QSAR (Hurst and Heritage 1997, Hurst et al 1998, Tong et al 1997) requires no explicit 3D information from the ligands. The molecule is hashed to a molecular fingerprint, which encodes the frequency of occurrence of various molecular fragment types using a predefined set of rules. To construct a ‘hologram’ this molecular fingerprint is cut into strings at fixed intervals as specified by a hologram length parameter. This numerical representation of the molecules is then used to build a QSAR. The HQSAR module of SYBYL was used in these studies. The hashed fingerprints encode the presence of all molecular fragment types containing between four and seven atoms and the optimal HQSAR model is derived from screening through the twelve default hologram length values, a set of prime numbers ranging from 53-401. The correlation analysis was carried out using PLS.

7.3 Results

7.3.1 Molecular Field Analysis

The QSAR equation returned for the MFA analysis of the Dataset 1 compounds in their undocked conformation is as follows (Equation 7.3.1.1)

$$\text{Predicted } pIC_{50} = 4.57953 - 0.024191 \, (\text{CH}_3/368)$$
$$+ 0.020903 \, (\text{CH}_3^+/559) - 0.02264 \, (\text{CH}_3^+/638)$$
$$+ 0.070093 \, (\text{CH}_3^+/525) - 0.11576 \, (\text{H}^+/415)$$
$$- 0.025352 \, (\text{H}^+/716) - 0.008665 \, (\text{H}^+/424)$$
$$- 0.017523 \, (\text{H}^+/544) + 0.013648 \, (\text{CH}_3^+/557)$$
$$+ 0.021621 \, (\text{CH}_3^+/435) - 0.013167 \, (\text{CH}_3^+/355)$$
$$- 0.027787 \, (\text{CH}_3^+/717) + 0.020693 \, (\text{H}^+/247)$$
$$+ 0.01797 \, (\text{H}^+/521)$$

(Equation 7.3.1.1)
The figures left of the slash in the brackets represent the probe type. Those to the right of the slash indicate the grid point. For example, for Equation 7.3.1.1 there is significant variance across the ligands at point 368 with respect to interaction with a relatively bulky negative charge (CH$_3^-$ probe). The QSAR then highlights this probe and point (CH$_3$/368) as a region that could be responsible for difference in biological activity, and assigns a weighting to it. $r^2 = 0.819$, $q^2 = 0.814$, Spearman Ranking coefficient $r_s = 0.82$, and probability that the correlation has occurred by random chance, $p = 4.54 \times 10^{-16}$.

The QSAR equation returned for the Dataset 1 compounds in their docked formation is as follows (Equation 7.3.1.2)

$$\text{Predicted pIC}_{50} = 6.02238 - 0.038973 \,(\text{CH}_3^-/1826)$$

$$- 0.04759 \,(\text{CH}_3^+/1397) - 0.015254 \,(\text{CH}_3^+/2054)$$

$$- 0.025881 \,(\text{CH}_3^-/1016) - 0.013889 \,(\text{CH}_3^-/1548)$$

$$- 0.058434 \,(\text{CH}_3^-/1168) - 0.010987 \,(\text{CH}_3^-/1845)$$

$$- 0.032971 \,(\text{CH}_3^-/1229) + 0.043112 \,(\text{CH}_3^-/1411)$$

$$+ 0.014974 \,(\text{CH}_3^+/2220) - 0.021147 \,(\text{H}^+/1871)$$

$$- 0.026017 \,(\text{CH}_3^+/1983) - 0.02593 \,(\text{H}^+/1183)$$

$$+ 0.013996 \,(\text{CH}_3^-/1141)$$

(Equation 7.3.1.2)

$r^2 = 0.868$, $q^2 = 0.865$, Spearman Ranking coefficient $r_s = 0.89$, and probability that the correlation has occurred by random chance, $p = 8.72 \times 10^{-19}$.

The QSAR equation returned for the MFA analysis of the Dataset 2 compounds in their undocked conformation is as follows (Equation 7.3.1.3)
Predicted $pIC_{50}$ = 5.76912 - 0.022809 (CH$_3^+$/430)  
+ 0.007944 (H$^+$/228) + 0.015144 (CH$_3^+$/291)  
- 0.000465 (CH$_3^+$/422) + 0.000594 (H$^+$/164)  
- 0.027954 (CH$_3^+$/211) + 0.058617 (H$^+$/419)  
+ 0.003442 (CH$_3^+$/307) - 0.021394 (H$^+$/307)  
+ 0.013671 (H$^+$/372) + 0.020503 (CH$_3^+$/477)  
- 0.030084 (CH$_3^+$/295) + 0.008495 (CH$_3^+$/380)  
+ 0.00231 (CH$_3^+$/381)  
(Equation 7.3.1.3)

$r^2 = 0.994$, $q^2 = 0.993$, Spearman Ranking coefficient $r_s = 0.99$, and probability that the correlation has occurred by random chance, $p = 2.88 \times 10^{-18}$.

The QSAR equation returned for the MFA analysis of the Dataset 2 compounds in their docked conformation is as follows (Equation 7.3.1.4)

Predicted $pIC_{50}$ = 6.85098 - 0.004531 (H$^+$/186)  
+ 0.021107 (CH$_3$/332) + 0.001339 (CH$_3$/302)  
- 0.004663 (CH$_3$/30) + 0.003565 (CH$_3$/310)  
+ 0.008732 (CH$_3$/313) + 0.032397 (CH$_3$/358)  
- 0.01931 (CH$_3$/130) - 0.009127 (CH$_3$/263)  
+ 0.018878 (CH$_3$/239) - 0.019915 (CH$_3$/87)  
+ 0.03503 (CH$_3$/96) - 0.010876 (CH$_3$/212)  
+ 0.007667 (CH$_3$/325)  
(Equation 7.3.1.4)

$r^2 = 0.994$, $q^2 = 0.994$, Spearman Ranking coefficient $r_s = 0.98$, and probability that the correlation has occurred by random chance, $p = 1.36 \times 10^{-18}$.

### 7.3.2 Receptor Surface Models

The QSAR equation returned for the RSM analysis of the Dataset 1 compounds in their undocked conformation is as follows (Equation 7.3.2.1)
Predicted pIC<sub>50</sub> = 5.77475 + 0.24069 (InterVdW Energy)
- 0.218168 (Inter Energy)
- 0.543605 (Inter Elec Energy)
- 0.001998 (Intra Energy)

(Equation 7.3.2.1)

InterVdW Energy is the energy of steric interaction between the receptor and the ligand, Inter Energy is the total interaction energy between the ligand and the receptor, Inter Elec Energy accounts for the electrostatic interaction energy between the ligand and receptor, and Intra Energy accounts for the internal strain energy of the ligand upon binding to the receptor.

\[ r^2 = 0.189, q^2 = 0.168, \text{Spearman Ranking coefficient } r_s = 0.33, \text{ and probability that the correlation has occurred by random chance, } p = 0.004. \]

The QSAR equation returned for the RSM analysis of the Dataset 1 compounds in their docked conformation is as follows (Equation 7.3.2.2)

Predicted pIC<sub>50</sub> = 5.47961 + 0.02578 (InterVdW Energy) + 0.073618 (Inter Energy) - 1.02609 (Inter Elec Energy) - 0.002387 (Intra Energy)

(Equation 7.3.2.2)

\[ r^2 = 0.423, q^2 = 0.408, \text{Spearman Ranking coefficient } r_s = 0.65, \text{ and probability that the correlation has occurred by random chance, } p = 4.06 \times 10^{-6}. \]

The QSAR equation returned for the RSM analysis of the Dataset 2 compounds in their undocked conformation is as follows (Equation 7.3.2.3)
Predicted $\text{pIC}_{50} = 5.25795 + 0.01124 \text{ (InterVdW Energy)}$
+ 0.093034 \text{ (Inter Energy)}
- 0.11508 \text{ (Inter Elec Energy)}
- 0.018811 \text{ (Intra Energy)}$

(Equation 7.3.2.3)

$r^2 = 0.877$, $q^2 = 0.869$, Spearman Ranking coefficient $r_s = 0.95$, and probability that
the correlation has occurred by random chance, $p = 3.13 \times 10^{-8}$.

The QSAR equation returned for the RSM analysis of the Dataset 2 compounds in
their docked conformation is as follows (Equation 7.3.2.4)

Predicted $\text{pIC}_{50} = 6.38777 + 0.01747 \text{ (InterVdW Energy)}$
+ 0.0147 \text{ (Inter Energy)}
- 0.030956 \text{ (Inter Elec Energy)}
- 0.014323 \text{ (Intra Energy)}$

(Equation 7.3.2.4)

$r^2 = 0.944$, $q^2 = 0.940$, Spearman Ranking coefficient $r_s = 0.86$, and probability that
the correlation has occurred by random chance, $p = 8.1 \times 10^{-11}$.

7.3.3 Cerius² Descriptor-Based QSAR

The QSAR equation returned for the Cerius² descriptor-based analysis of the Dataset
1 compounds in their undocked conformation is as follows (Equation 7.3.3.1)

Predicted $\text{pIC}_{50} = 2.57908 + 0.008252 \text{ (Vm)}$
- 0.001021 \text{ (PMI-mag)} + 0.000167 \text{ (Apol)}
- 0.000203 \text{ (LUMO_MOPAC)} - 0.000329 \text{ (Rotl Bonds)}
- 0.002704 \text{ (Foct)} - 0.001285 \text{ (F_{H2O})}
- 0.000205 \text{ (Sr)} - 0.000674 \text{ (HOMO)}
- 0.000248 \text{ (DIPOLE_MOPAC)} + 0.000402 \text{ (Hbond Acceptor)}
- 0.003017 \text{ (MR)} - 0.007191 \text{ (HF_MOPAC)}
- 0.000311 \text{ (HOMO_MOPAC)}$

(Equation 7.3.3.1)
\[ r^2 = 0.431, \ q^2 = 0.416, \ \text{Spearman Ranking coefficient} \ r_s = 0.69, \ \text{and probability that the correlation has occurred by random chance,} \ p = 3.09 \times 10^{-6}. \]

The QSAR equation returned for the Cerius\textsuperscript{2} descriptor-based analysis of the Dataset 1 compounds in their docked conformation is as follows (Equation 7.3.3.2)

\[
\text{Predicted } pIC_{50} = 3.8106 + 0.067394 (\text{Hbond acceptor}) - 0.032187 (\text{Area}) - 0.05003 (\text{Foct}) + 0.034969 (\text{Sr}) - 0.007413 (\text{Rad of Gyration}) - 0.009911 (\text{Rotlbonds}) + 0.010689 (\text{Hbond donor}) + 0.036769 (\text{DIPOLE}_M\text{OPAC}) + 0.01287 (\text{LUMO}_M\text{OPAC}) + 0.062173 (\text{Vm}) + 0.014519 (\text{HOMO}_M\text{OPAC}) - 0.00844 (\text{HOMO}) - 0.050766 (\text{AlogP}) - 0.017201 (\text{MW})
\]

(Equation 7.3.3.2)

\[ r^2 = 0.293, \ q^2 = 0.275, \ \text{Spearman Ranking coefficient} \ r_s = 0.59, \ \text{and probability that the correlation has occurred by random chance,} \ p = 2.56 \times 10^{-4}. \]

The QSAR equation returned for the Cerius\textsuperscript{2} descriptor-based analysis of the Dataset 2 compounds in their undocked conformation is as follows (Equation 7.3.3.3)

\[
\text{Predicted } pIC_{50} = 17.9994 + 0.000333319 (\text{Apol}) + 0.0413546 (\text{HOMO}) - 0.127939 (\text{Sr}) + 0.437412 (\text{LUMO}_M\text{OPAC}) - 0.00164379 (\text{MR}) + 0.00281392 (\text{HF}_M\text{OPAC}) + 0.331457 (\text{HOMO}_M\text{OPAC}) - 2.23639 (\text{Rad of Gyration}) - 0.0051217 (\text{Area}) - 0.00108676 (\text{PMI-mag}) - 0.00120554 (\text{Vm}) - 0.00497358 (\text{MW}) + 0.127638 (\text{Rotlbonds}) - 0.113199 (\text{Hbond acceptor}) + 0.299345 (\text{Hbond donor}) - 0.341847 (\text{AlogP}) - 0.00593261 (\text{F}_2\text{O}) - 0.0143387 (\text{Foct}) + 0.00812653 (\text{DIPOLE}_M\text{OPAC})
\]

(Equation 7.3.3.3)
\( r^2 = 0.806, \ q^2 = 0.793, \) Spearman Ranking coefficient \( r_s = 0.82, \) and probability that the correlation has occurred by random chance, \( p = 1.01 \times 10^{-6}. \)

The QSAR equation returned for the Cerius\(^2\) descriptor-based analysis of the Dataset 2 compounds in their docked conformation is as follows (Equation 7.3.3.4)

\[
\text{Predicted } p\text{IC}_{50} = -3.89411 - 0.014337 \ (\text{LUMO}_\text{MOPAC}) \\
+ 0.094688 \ (\text{FocL}) + 0.1172 \ (\text{DIPOLE}_\text{MOPAC}) \\
- 0.018303 \ (\text{MW}) + 0.056562 \ (\text{Hbond Acceptor}) \\
- 0.008437 \ (\text{Rad of Gyration}) + 0.015697 \ (\text{Rollbonds}) \\
- 0.051904 \ (\text{AlogP}) - 0.019398 \ (\text{HOMO}_\text{MOPAC}) \\
- 0.08867 \ (\text{Sr}) - 0.102535 \ (\text{FH}_2\text{O}) \\
- 0.132405 \ (\text{HOMO}) + 0.016725 \ (\text{Hbond donor}) \\
+ 0.04688 \ (\text{Vm})
\]

(Equation 7.3.3.4)

\( r^2 = 0.703, \ q^2 = 0.684, \) Spearman Ranking coefficient \( r_s = 0.74, \) and probability that the correlation has occurred by random chance, \( p = 2.56 \times 10^{-5}. \)

7.3.4 Comparative Molecular Field Analysis

No regression equation was returned for the CoMFA QSAR analysis, with CoMFA scores for each compound returned instead. A linear regression of the CoMFA scores versus the experimental biological activities was carried out for each experiment and the results presented below.

The QSARs returned for the CoMFA analysis of the Dataset 1 and Dataset 2 compounds in their docked and undocked conformations are illustrated below (Figure 7.3.4.1 – Figure 7.3.4.4)
Figure 7.3.4.1 Linear regression of CoMFA score vs. experimental biological activity for Dataset 1 compounds in their undocked conformations. $r^2 = 0.104$, $q^2 = 0.081$, Spearman Ranking coefficient $r_s = 0.36$, and probability that the correlation has occurred by random chance, $p = 0.039$.

Figure 7.3.4.2 Linear regression of CoMFA score vs. experimental biological activity for Dataset 1 compounds in their docked conformations. $r^2 = 0.014$, $q^2 = -0.013$, Spearman Ranking coefficient $r_s = 0.22$, and probability that the correlation has occurred by random chance, $p = 0.447$. 
**Figure 7.3.4.3** Linear regression of CoMFA score vs. experimental biological activity for Dataset 2 compounds in their undocked conformations. $r^2 = 0.026$, $q^2 = -0.038$, Spearman Ranking coefficient $r_s = 0.00613$, and probability that the correlation has occurred by random chance, $p = 0.534$.

**Figure 7.3.4.4** Linear regression of CoMFA score vs. experimental biological activity for Dataset 2 compounds in their docked conformations. $r^2 = 0.071$, $q^2 = 0.009$, Spearman Ranking coefficient $r_s = 0.25$, and probability that the correlation has occurred by random chance, $p = 0.298$. 


7.3.5 Hologram QSAR

Similar to CoMFA, Hologram QSAR does not return a regression equation only predicted values for biological activity. Also, as it is a 2D technique the results are independent of the molecular alignments. Hence, only one QSAR is returned for each dataset.

The QSAR returned for the HQSAR analysis of the Dataset 1 compounds is illustrated in Figure 7.3.5.1. The optimal hologram length for this investigation = 2; the optimal number of components = 5.

![Figure 7.3.5.1](image_url)

**Figure 7.3.5.1** Linear regression of HQSAR predicted pIC50 vs. experimental biological activity for Dataset 1. $r^2 = 0.068$, $q^2 = 0.044$, Spearman Ranking coefficient $r_s = 0.22$, and probability that the correlation has occurred by random chance, $p = 0.099$. 

192
The QSAR returned for the HQSAR analysis of the Dataset 2 compounds is illustrated in Figure 7.3.5.2. The optimal hologram length for this investigation = 97; the optimal number of components = 5.

![Graph showing linear regression of HQSAR predicted pIC50 vs. experimental biological activity for Dataset 2. r² = 0.729, q² = 0.711, Spearman Ranking coefficient rs = 0.82, and probability that the correlation has occurred by random chance, p = 1.28 x 10⁻⁵.]

**Figure 7.3.5.2** Linear regression of HQSAR predicted pIC50 vs. experimental biological activity for Dataset 2. $r^2 = 0.729$, $q^2 = 0.711$, Spearman Ranking coefficient $r_s = 0.82$, and probability that the correlation has occurred by random chance, $p = 1.28 \times 10^{-5}$.

### 7.3.6 Jury Models

The results of the individual QSAR methods have been presented above. Each technique embraces different methodologies, each with their own different assumptions and limitations. It has been postulated that obtaining a mean prediction of biological activities from several different QSAR methods would lead to an increased accuracy of overall prediction (So and Karpluss 1999, Ajay 1994). The combination of QSAR techniques in this manner is referred to as a Jury model, which
assumes that any errors associated with individual techniques are evenly distributed. Several QSAR studies using this approach have been reported (Rogers and Hopfinger 1994, So and Karpluss 1996a, So and Karpluss 1996b). The results from the MFA, RSM, Cerius² descriptor-based and HQSAR experiments were combined to produce ‘jury’ models for these investigations. The results for the CoMFA analyses were omitted from the jury models due to the fact that scores, as opposed to predictions of pIC₅₀, were returned using this method.

The QSARs returned for the Jury models of the Dataset 1 and Dataset 2 compounds in their docked and undocked conformations are illustrated below (Figure 7.3.6.1 – Figure 7.3.6.4)

**Figure 7.3.6.1** Linear regression of the mean predicted pIC₅₀ determined from four individual QSAR experiments (MFA, RSM, Cerius² descriptor-based and HQSAR) vs. experimental biological activity for Dataset 1 compounds in their undocked conformations. $r^2 = 0.833$, $q^2 = 0.829$, Spearman Ranking coefficient $r_s = 0.88$, and probability that the correlation has occurred by random chance, $p = 8.94 \times 10^{-17}$. 
Figure 7.3.6.2  Linear regression of the mean predicted pIC50 determined from four individual QSAR experiments (MFA, RSM, Cerius$^2$ descriptor-based and HQSAR) vs. experimental biological activity for Dataset 1 compounds in their docked conformations. \( r^2 = 0.874, q^2 = 0.871, \) Spearman Ranking coefficient \( r_s = 0.93, \) and probability that the correlation has occurred by random chance, \( p = 3.34 \times 10^{-19}. \)

Figure 7.3.6.3  Linear regression of the mean predicted pIC50 determined from four individual QSAR experiments (MFA, RSM, Cerius$^2$ descriptor-based and HQSAR) vs. experimental biological activity for Dataset 2 compounds in their undocked conformations. \( r^2 = 0.958, q^2 = 0.955, \) Spearman Ranking coefficient \( r_s = 0.95, \) and probability that the correlation has occurred by random chance, \( p = 9.32 \times 10^{-12}. \)
Figure 7.3.6.4 Linear regression of the mean predicted pIC50 determined from four individual QSAR experiments (MFA, RSM, Cerius² descriptor-based and HQSAR) vs. experimental biological activity for Dataset 2 compounds in their docked conformations. \( r^2 = 0.957 \), \( q^2 = 0.917 \), Spearman Ranking coefficient \( r_s = 0.92 \), and probability that the correlation has occurred by random chance, \( p = 1.59 \times 10^{-9} \).

The values of \( r^2 \) and \( q^2 \) returned for all of the QSAR investigations are summarised below in Table 7.3.6.1.
<table>
<thead>
<tr>
<th>Dataset &amp; Alignment</th>
<th>MFA</th>
<th>RSM</th>
<th>Cerius&lt;sup&gt;2&lt;/sup&gt; Descriptors</th>
<th>CoMFA</th>
<th>HQS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Docked r^2</td>
<td>0.868</td>
<td>0.423</td>
<td>0.293</td>
<td>0.014</td>
<td>0.06</td>
</tr>
<tr>
<td>Undocked q^2</td>
<td>0.865</td>
<td>0.408</td>
<td>0.275</td>
<td>-0.013</td>
<td>0.04</td>
</tr>
<tr>
<td>Docked r^2</td>
<td>0.819</td>
<td>0.189</td>
<td>0.431</td>
<td>0.104</td>
<td>0.06</td>
</tr>
<tr>
<td>Undocked q^2</td>
<td>0.814</td>
<td>0.168</td>
<td>0.416</td>
<td>0.081</td>
<td>0.04</td>
</tr>
<tr>
<td>Docked r^2</td>
<td>0.994</td>
<td>0.944</td>
<td>0.703</td>
<td>0.071</td>
<td>0.72</td>
</tr>
<tr>
<td>Undocked q^2</td>
<td>0.994</td>
<td>0.940</td>
<td>0.684</td>
<td>0.009</td>
<td>0.71</td>
</tr>
<tr>
<td>Docked r^2</td>
<td>0.994</td>
<td>0.877</td>
<td>0.806</td>
<td>0.026</td>
<td>0.72</td>
</tr>
<tr>
<td>Undocked q^2</td>
<td>0.993</td>
<td>0.869</td>
<td>0.793</td>
<td>-0.038</td>
<td>0.71</td>
</tr>
</tbody>
</table>

Table 7.3.6.1 Summary of r^2 and q^2 values from the QSAR investigations.
7.4. Discussion

The first observation when examining the results of the QSARs is that the $r^2$ and $q^2$ values returned are significantly better than those for the eleven scoring functions investigated (Chapter 5). The method that performs worst for both datasets is the grid-based CoMFA method. The reason for this poor performance is not clearly understood but is unlikely to be due to the disparate nature of the chemical structures in Dataset 1 as the structurally similar compounds in Dataset 2 perform as poorly. The matrix that is produced via CoMFA is largely under-characterised, and the subtleties responsible for the differences in activity may therefore be lost, despite the fact that the regression equation is derived using partial least squares (PLS) analysis. Also, the program only considers one type of chemical probe (in this case $\text{CH}_3^+$) which means that changes in activity due to negatively charged regions in the ligands, or those involved in hydrogen bonding would not be accounted for in the initial matrix produced. The limited information returned by the program used to perform this analysis means it is impossible to dissect the results any further.

HQSAR also performs poorly for the disparate Dataset 1, but significantly better for the homologous structures in Dataset 2. It is likely that this is due to HQSAR taking no account for the shape and 3-D properties of the compounds in the dataset. The technique appears to be better suited for analysing structurally similar compounds. This may be because it is better able define these compounds in terms of molecular fragments and thus identify those fragments responsible for increases and decreases in biological activity. This is exemplified by the high $r^2$ and $q^2$ values returned for Dataset 2 (0.729 and 0.711 respectively). However, with the largely unrelated chemical structures in Dataset 1 the technique struggles to determine fragments that are responsible for changes in activity. This is most likely due to the increasing number of structurally different molecular fragments, and the likelihood that the changes in activity are due to not one change, but a combination of structural and chemical changes from compound to compound. The poor results for Dataset 1 ($r^2 = 0.068$, $q^2 = 0.044$) lend credence to this hypothesis.

A similar argument can be adopted to explain the results for the RSM QSAR studies. For Dataset 1 the receptor surface model is built upon the ligands with the highest in
vivo potency. Due to the structural diversity across the datasets the molecular volume of these ligands is unlikely to be representative of the group as a whole. The QSAR model is then built by fitting the remaining ligands into the receptor surface model, and mapping properties from the conformations of the ligands onto the receptor surface. Due to the structural diversity it is possible that the conformations adopted by some of the compounds are as a result of being unable to fit into the receptor surface model any other way. This could lead to a dilution of data that would result in a poor correlation in the regression analysis as observed for the undocked conformation of ligands in Dataset 1 ($r^2 = 0.189$, $q^2 = 0.168$). However, the docked conformation for Dataset 1 shows a significant improvement in the regression statistics ($r^2 = 0.423$, 0.408) which is encouraging as it suggests that the comparative model produced (Chapter 3) is providing useful information on the binding orientations of the structurally dissimilar ligands, and suggests that there may be a common binding orientation for at least some of the ligands.

The RSM results for Dataset 2 reflect the fact that the structures are very similar and adopt almost identical conformations in the 2D6 binding site. The regression analysis of the RSM model in this case is able to easily identify those areas that are different between the substrates and attribute them to the changes in activity, hence the predictive $r^2$ and $q^2$ values obtained (0.877 and 0.869 respectively) for the undocked conformation. Once again the value of the comparative model in elucidating a common binding mode is highlighted by an increase in the $r^2$ and $q^2$ values for the docked conformation of the ligands (0.944 and 0.940 respectively).

The results obtained for the Cerius² descriptor-based method are more complex. Once again it is expected that a structurally similar dataset will perform better than a structurally dissimilar one, and the $r^2$ and $q^2$ values for the undocked conformation of Dataset 2 outperform those from Dataset 1 (0.806 and 0.793 vs. 0.431 and 0.416 respectively). However, in contrast to the results for the RSM QSAR the docked conformations of both datasets show a reduction in the $r^2$ and $q^2$ values returned (0.703 and 0.684 for Dataset 2 and 0.293 and 0.275 for Dataset 1).

This highlights a potential weakness with regards to this method. Initially, the descriptors were chosen following an analysis of the Dataset 2 ligands in their
'undocked' structural alignment. The considerable effort placed in determining these descriptors is reflected in the predictive values returned for $r^2$ and $q^2$ for this data (0.806 and 0.793 respectively). However, altering the alignment of the structures to their 'docked' conformation shows a significant drop in the regression statistics values. It illustrates that the descriptors are no longer able to define the dataset under investigation well, and shows that the technique is limited in that it does not produce a solution that is generally applicable across a range of substrates, as reiterated by the poor performance for Dataset 1. Hence, it is likely that in order to obtain predictive QSAR values for each alignment of each dataset a great deal of time would need to be invested in determining the descriptors best suited to defining that group of molecules. Due to the highly sensitive nature of this technique, it is unlikely that the QSAR produced would be applicable to substrates other than those structurally similar to the ones in the initial dataset. Therefore, from the results of this investigation it would appear that any QSAR derived from descriptors would be of limited use in predicting the biological activities of an unknown set of compounds.

The results obtained for the grid-based MFA analysis were the most encouraging of all of the results from the individual QSAR studies. They showed that there was an improvement in $r^2$ and $q^2$ from the undocked to the docked conformation for Dataset 1, and that the results for Dataset 1 and Dataset 2 could be considered predictive. Such pleasing results can be attributed to extensive characterisation of the grid map with a number of chemical probes, and an efficient method for extracting the regression equation from the characterisation matrix.

In order to validate the findings from the MFA QSAR ten compounds, representative of the structural diversity of the dataset and the range of biological activities contained therein, were removed from Dataset 1 (Appendix 7.1) to form a test set. An MFA QSAR experiment performed on the remaining 40 compounds (the training set). The resultant QSAR regression equation was then used to predict the biological activities of the ligands in the test set, in order to observe whether the method was robust and accurate. The training set MFA regression analysis returned values for $r^2$ and $q^2$ of 0.813 and 0.807. When used to predict the pIC50 values of the test set values for $r^2$ and $q^2$ returned were 0.733 and 0.700 (Figure 7.4.1) indicating that the method is indeed reliable and robust, and that the good results obtained were not artefacts of the
Figure 7.4.1 Linear regression of the predicted pIC50 determined for the test set compounds in their docked conformations (Appendix 7.1) vs. experimental biological activity.

way in which data had been processed. Due to the limited size of Dataset 2 (18 compounds) it was not possible to validate the QSAR results in the same manner.

The QSAR jury models for both Dataset 1 and Dataset 2 appear to be predictive. As was suggested by So and Karpluss (So and Karpluss 1999) they improve the overall quality of the regression statistics by compensating for weaknesses in individual techniques to hopefully give more accurate and reliable results. For Dataset 1 the jury models out perform all of the individual techniques in both the docked and undocked conformations. This is mirrored for Dataset 2, with the exception of the results for the MFA analyses The high $r^2$ and $q^2$ values returned for the Dataset 2 MFA analyses are extremely uncommon and can be attributed to the small size of the dataset and the homologous nature of the compounds contained within it.
7.5 Conclusions

The results of these investigations suggest that QSAR experiments can be used to predict the activities of novel and potent pharmaceutical compounds. However, it also highlights that QSAR is more successful when analysing structurally related compounds with common binding orientations in the active site, as demonstrated by Dataset 2 consistently outperforming the diverse Dataset 1 in all of the QSAR experiments.

The work carried out here also indicated that each method of QSAR has strengths and limitations. HQSAR, a 2D technique that disregards shape and molecular conformation can be predictive for a set of structurally related compounds, but fails to be of use for larger sets of dissimilar structures. Descriptor-based QSAR is a sensitive technique based on the meticulous selection of appropriate descriptors. With time and attention to detail this can yield results that are useful and also predictive, particularly for datasets containing homologues, although the general applicability of any regression equation obtained in this manner is questionable. The grid-based MFA approach is a powerful and robust technique, which performs well even with structurally diverse compounds, although without a systematic characterisation of the ligands under investigation even powerful grid-based techniques, such as CoMFA, a technique shown to be useful and extremely robust in other studies (Cramer et al 1988) will perform poorly. However, as with comparative modelling techniques, a combined approach to QSAR experiments should be carried out to identify the method most appropriate for any particular dataset.

Of the individual techniques MFA performed by far the best. However, a more predictive 'jury' model can be obtained by combining the results from different QSAR methods and obtaining averages for any predictive values. This helps compensate for any weaknesses in an individual technique, and appears from the results of these investigations to be of particular use when dealing with diverse datasets.

Alignment of ligands is of paramount importance for a QSAR experiment. Encouragingly, the quality of the regression statistics improve for both the MFA and
RSM QSARs when the molecules of the datasets are considered in their docked conformations. This indicates that the comparative model is having a positive input into the drug-discovery process and was worthwhile building, and that there may be a common binding mode for some of the ligands in the diverse Dataset 1. However, alteration of the alignment also highlights the strengths and weaknesses of particular techniques. Sensitive techniques such as the Cerius\(^2\) descriptor-based QSAR suffer as a result of even minor alterations in alignment, and caution must be taken to examine all results from different QSARs carefully and objectively.

Overall, the QSAR experiments were successful in helping to identify a technique capable of predicting the biological activity of a novel compound, and for 2D6 compounds an approach based on Molecular Field Analysis, or a ‘jury’ approach combining MFA, RSM, HQSAR and descriptor-based data would be most appropriate.
Chapter 8

Investigating Pharmacophores of
Cytochrome P450 2D6
8.0 Preface

The investigations in Chapter 7 illustrate that it is possible to predict the binding affinity of a potential 2D6 substrate using QSAR methods. The studies in this chapter seek to extend this work by investigating whether or not pharmacophores, which have traditionally been used to filter databases of compounds in order to find potential lead candidates in the drug-discovery process, can be used to predict the activity of a P450 2D6 substrate.

8.1 Introduction

In drug design the term pharmacophore refers to a set of chemical features that are common across a set of biologically active compounds such as hydrogen bond donors and acceptors, positively and negatively charged groups and hydrophobic fragments. These are referred to as 'pharmacophoric groups'. A 3D pharmacophore specifies the spatial relationships between pharmacophoric groups, often expressed as distances or distance ranges, although other geometrical features such as angles and planes may be incorporated.

The development of methods for studying the conformation of ligands has stimulated an interest in the influence of the 3D structure of molecules on their chemical and biological activity. The objective of the procedure known as pharmacophore mapping is to determine possible 3D pharmacophores for a series of active compounds. Once such a pharmacophore has been developed it can be used to find or suggest other active molecules.

Several pharmacophores for P450 2D6 have been published that account for both the N-dealkylation and more usual O-dealkylation reaction mechanisms. All of the published pharmacophores have been reasonably successful in identifying, from large databases, substrates that are metabolised by P450 2D6, the consensus from these investigations being that a substrate needs to have a positive ionisable group between 5-10Å from the centre in the molecule undergoing oxidation and a large flat hydrophobic region in the vicinity of the site of oxidation (Koymans et al 1985, Meyer et al 1986, Islam et al 1991, de Groot et al 1995, de Groot et al 1997, Strobl et
al 1993, de Groot et al 1999, Abdel-Rahman et al 1998, Ekins et al 1999, Ekins et al 2001). The purpose of the investigations in this chapter was to build pharmacophores for the 2D6 substrates in Dataset 1 (Appendix 6.1), to compare and contrast these with the pharmacophores that appear in the literature, and to use the pharmacophores generated to predict the biological activities of the test set derived from Dataset 1.

8.2 Experimental

Ten molecules, deemed to be representative of the structural variance and range of pIC50s in Dataset 1, were defined as the test set of molecules. The test set (consisting of pentazocine, corynanthine, lobeline, econazole, ARC053620, promethazine, haloperidol, imipramine, clozapine and carbamazepine) was identical to the test set used in the QSAR studies (Chapter 7, Appendix 7.1).

The remaining 46 compounds of Dataset 1 (subsequently referred to as the training set) were used to generate pharmacophores for P450 2D6 using the program Catalyst (Accelrys, San Diego, USA). Initially up to 255 conformers of each individual molecule were generated over an energy range of 20 kcal mol⁻¹ with respect to the lowest energy conformer. 255 was chosen as the upper limit for the number of conformers as the Catalyst program is unable to handle more than this.

To generate hypotheses for a pharmacophore, the types of pharmacophoric features present in a tightly binding molecule need to be established. The tightest binding molecule in the training set, ajamalicine (pIC50 7.9), was analysed using the function-mapping feature within Catalyst. The pharmacophoric features returned from the analysis of this molecule used as a basis for the production of pharmacophore hypotheses.

Pharmacophore hypotheses were generated using the HypoGen module within Catalyst, utilising the experimental biological activities for all of the compounds in the training set and the conformations of these molecules as generated by Catalyst. In the first instance the only pharmacophoric features that were used were positive ionisable and hydrophobic, as it was reasoned that this would lead to the most generally applicable pharmacophore because both these pharmacophoric groups had
featured extensively in the literature. Secondly, an investigation into five possible pharmacophoric features, as determined by the function mapping analysis of ajmalicine was carried out. Although the function mapping of analysis returned more than five features, this was the maximum number supported by HypoGen.

The top three pharmacophores generated by each investigation, as determined by the total cost score returned, were then used to predict the activity of the molecules in the test set. The total cost of the null hypothesis i.e. the worst hypothesis identified by the program, was also noted. The predicted activities for the test set molecules were compared against their experimentally determined activities using linear regression analysis. The validity of the pharmacophore was ascertained by the values returned for $r^2$, $q^2$ and $r_s$ (the Spearman’s correlation coefficient) for the test set molecules.

The investigation was repeated using the GOLD docked conformations of the training set and test set molecules. This experiment was carried out to determine whether or not the inherent information about the structure of the 2D6 active site contained within the conformations of the docked compounds would improve the predictivity of any pharmacophore produced. Again the three top scoring pharmacophores from each of the experiments were used to predict the biological activities of the test set molecules and the results investigated using regression analysis techniques.

Finally a pharmacophore based on the comparative model of 2D6 was produced using the Structure-Based Drug Design module in Cerius$^2$. The active site of the protein was defined using the Structure-Based Focussing module. A model of 2D6 with codeine docked into the active site was loaded into the program, and the active site defined as a sphere with radius of 10.1Å from the centre of the codeine molecule. This incorporated the heme group, and the acidic residues Glu 216 and Asp 301 deemed to be important in earlier investigations (Chapters 4 and 5). An interaction map was then generated for the active site that highlighted areas in the active site that displayed hydrogen-bond donor, hydrogen-bond acceptor and hydrophilic characteristics. This interaction map was manually divided into clusters, which then formed the basis of the resultant pharmacophore.
The nature of this pharmacophore means that regions of space occupied by the protein can be incorporated as an additional pharmacophoric feature, termed excluded volume. Hence, multiple pharmacophores with heavy atom radii for the exclusion spheres of 1 Å, 0.8 Å, 0.6 Å, 0.4 Å, 0.2 Å and a pharmacophore without any exclusion volume were built and used to predict the activities of the test set compounds. The results of these predictions were analysed as before.

8.3 Results

8.3.1 Generation of Conformers for Catalyst.

The number of conformers generated for each of the compounds in the training set and the test set are presented in Table 8.3.1.1 and Table 8.3.1.2, respectively.

Table 8.3.1.1 The number Catalyst generated conformers for each of the ligands in the training set

<table>
<thead>
<tr>
<th>Name of Compound</th>
<th>Number of Conformers</th>
<th>Name of Compound</th>
<th>Number of Conformers</th>
</tr>
</thead>
<tbody>
<tr>
<td>10methylallylphenothiazine</td>
<td>8</td>
<td>Timolol</td>
<td>79</td>
</tr>
<tr>
<td>Fluphenazine</td>
<td>78</td>
<td>Flunarazine</td>
<td>52</td>
</tr>
<tr>
<td>ARC107230</td>
<td>4</td>
<td>ARC128343</td>
<td>16</td>
</tr>
<tr>
<td>Maybridge RJ0122</td>
<td>133</td>
<td>Propanolol</td>
<td>53</td>
</tr>
<tr>
<td>ARC119787</td>
<td>9</td>
<td>Cinnarizine</td>
<td>61</td>
</tr>
<tr>
<td>Perphenazine</td>
<td>21</td>
<td>Thioridazine</td>
<td>133</td>
</tr>
<tr>
<td>Amitryptiline</td>
<td>22</td>
<td>Ethopromazine</td>
<td>37</td>
</tr>
<tr>
<td>Spartane</td>
<td>5</td>
<td>ARC053631</td>
<td>6</td>
</tr>
<tr>
<td>ARC081384</td>
<td>53</td>
<td>Lidocaine</td>
<td>89</td>
</tr>
<tr>
<td>Maybridge 01581</td>
<td>62</td>
<td>Chlorpromazine</td>
<td>65</td>
</tr>
<tr>
<td>ARC116195</td>
<td>132</td>
<td>Tamoxifien</td>
<td>167</td>
</tr>
<tr>
<td>Perhexilene</td>
<td>8</td>
<td>Dextromethorphan</td>
<td>9</td>
</tr>
<tr>
<td>Ajmaline</td>
<td>3</td>
<td>ARC019504</td>
<td>29</td>
</tr>
<tr>
<td>Quinine</td>
<td>33</td>
<td>Ketoconazole</td>
<td>3</td>
</tr>
<tr>
<td>Codeine</td>
<td>5</td>
<td>ARC115096</td>
<td>247</td>
</tr>
<tr>
<td>Yohimbine</td>
<td>12</td>
<td>Metoprolol</td>
<td>81</td>
</tr>
<tr>
<td>ARC115632</td>
<td>3</td>
<td>Desipramine</td>
<td>24</td>
</tr>
<tr>
<td>Miconazole</td>
<td>55</td>
<td>10decylphenothiazine</td>
<td>31</td>
</tr>
<tr>
<td>Ajmalacine</td>
<td>7</td>
<td>Hydroquinidine</td>
<td>21</td>
</tr>
<tr>
<td>Quinidine</td>
<td>126</td>
<td>ARC108505</td>
<td>34</td>
</tr>
<tr>
<td>Clomipramine</td>
<td>41</td>
<td>Methyltetrahydroisoquinolinone</td>
<td>4</td>
</tr>
<tr>
<td>ARC127118</td>
<td>3</td>
<td>Propafenone</td>
<td>143</td>
</tr>
</tbody>
</table>
Table 8.3.1.2 The number of Catalyst generated conformers for each of the ligands in the test set.

<table>
<thead>
<tr>
<th>Name of Compound</th>
<th>Number of Conformers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pentazocine</td>
<td>49</td>
</tr>
<tr>
<td>Lobeline</td>
<td>255</td>
</tr>
<tr>
<td>ARC053620</td>
<td>14</td>
</tr>
<tr>
<td>Haloperidol</td>
<td>81</td>
</tr>
<tr>
<td>Clozapine</td>
<td>7</td>
</tr>
<tr>
<td>Corynanthine</td>
<td>12</td>
</tr>
<tr>
<td>Econazole</td>
<td>42</td>
</tr>
<tr>
<td>Promethazine</td>
<td>14</td>
</tr>
<tr>
<td>Imipramine</td>
<td>33</td>
</tr>
<tr>
<td>Carbamazepine</td>
<td>4</td>
</tr>
</tbody>
</table>

8.3.2 Function Mapping

The function mapping experiment for ajmalicine resulted in six different Catalyst pharmacophoric features being highlighted. These were: ‘hydrogen bond acceptor’; ‘hydrogen bond acceptor (lipid)’; ‘aliphatic hydrophobe’; ‘aromatic hydrophobe’; ‘positive ionisable’ and ‘ring aromatics’. These features were then used to generate pharmacophore hypotheses for 2D6.

8.3.3 Hypothesis 1: 2 Feature Pharmacophore

An experiment was carried out to find a general 2D6 pharmacophore using only the hydrophobic and positive ionisable pharmacophoric features. The results for the three highest ranked pharmacophores generated using the conformers for the test set produced in Catalyst are shown in Table 8.3.3.1. The hypotheses 1A, 1B, and 1C were then used to predict the biological activities for the ten compounds of the test set. The regression statistics are summarised in Table 8.3.3.2.
Table 8.3.3.1 Results for the 2-feature pharmacophores generated using the Catalyst produced conformers for the ligands in the test set.

<table>
<thead>
<tr>
<th>Hypothesis Name</th>
<th>Pharmacophoric Features</th>
<th>Total Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesis 1A</td>
<td>Positive Ionisable 3x Hydrophobe</td>
<td>206</td>
</tr>
<tr>
<td>Hypothesis 1B</td>
<td>Positive Ionisable 3x Hydrophobe</td>
<td>209</td>
</tr>
<tr>
<td>Hypothesis 1C</td>
<td>Positive Ionisable 3x Hydrophobe</td>
<td>211</td>
</tr>
<tr>
<td>Null Hypothesis 1</td>
<td>Positive Ionisable 3x Hydrophobe</td>
<td>220</td>
</tr>
</tbody>
</table>

Table 8.3.3.2 Linear regression plot statistics of experimental biological activity versus predicted biological activity for hypotheses 1A, 1B and 1C.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>$r^2$</th>
<th>$q^2$</th>
<th>Spearman Coefficient ($r_s$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A</td>
<td>0.36</td>
<td>0.29</td>
<td>0.76</td>
</tr>
<tr>
<td>1B</td>
<td>0.22</td>
<td>0.12</td>
<td>0.53</td>
</tr>
<tr>
<td>1C</td>
<td>0.28</td>
<td>0.19</td>
<td>0.58</td>
</tr>
</tbody>
</table>

The experiment was repeated using the GOLD-docked orientations of the ligands in the training set. The top-ranked pharmacophores (Table 8.3.3.3) were used to predict the biological activities of the test set. The regression statistics for hypotheses 1D, 1E and 1F are summarised in Table 8.3.3.4.
Table 8.3.3.3 Results for the 2-feature pharmacophores using the GOLD generated conformers for the ligands in the test set.

<table>
<thead>
<tr>
<th>Hypothesis Name</th>
<th>Pharmacophoric Features</th>
<th>Total Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesis 1D</td>
<td>Positive Ionisable 3x Hydrophobe</td>
<td>142</td>
</tr>
<tr>
<td>Hypothesis 1E</td>
<td>Positive Ionisable 4x Hydrophobe</td>
<td>150</td>
</tr>
<tr>
<td>Hypothesis 1F</td>
<td>4x Hydrophobe</td>
<td>156</td>
</tr>
<tr>
<td>Null Hypothesis</td>
<td>Positive Ionisable 3x Hydrophobe</td>
<td>165</td>
</tr>
</tbody>
</table>

Table 8.3.3.4 Linear regression plot statistics of experimental biological activity versus predicted biological activity for hypothesis 1D, 1E and 1F.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>r²</th>
<th>q²</th>
<th>Spearman Coefficient (rs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1D</td>
<td>0.05</td>
<td>-0.06</td>
<td>0.29</td>
</tr>
<tr>
<td>1E</td>
<td>0.005</td>
<td>-0.11</td>
<td>0.10</td>
</tr>
<tr>
<td>1F</td>
<td>0.005</td>
<td>-0.11</td>
<td>0.21</td>
</tr>
</tbody>
</table>

8.3.4 Hypothesis 2: 5 Feature Pharmacophore

A second experiment was carried out using five of the features determined by the function mapping of the most tightly binding molecule in the training set, ajmalicine. These features were ‘hydrogen-bond acceptor’, ‘hydrophobic aliphatic’, ‘hydrophobic aromatic’, ‘positive ionisable’ and ‘ring aromatics’. It was necessary to use only five of the features determined by the function mapping experiment because HypoGen is only capable of dealing with a maximum of five features. The two features that were...
omitted were 'hydrogen bond acceptor (lipid)' and 'hydrophobe'. These were chosen because it was believed that they were at least partially represented by the other features incorporated in the hypothesis generation i.e. 'hydrogen bond acceptor (lipid)' was represented by hydrogen bond acceptor, and 'hydrophobe' was represented by 'hydrophobic aliphatic' and 'hydrophobic aromatic'. The results for the three highest ranked pharmacophores generated using the conformers for the test set produced in Catalyst are shown in Table 8.3.4.1. Hypotheses 2A, 2B, and 2C were then used to predict the biological activities for the ten compounds of the test set. The regression statistics are summarised in Table 8.3.4.2. The experiments were then repeated using the GOLD generated conformations for the compounds in the test set, yielding three pharmacophores 2D, 2E and 2F summarised in Table 8.3.4.3.

**Table 8.3.4.1** Results for the 5-feature pharmacophores generated using the Catalyst produced conformers for the ligands in the test set.

<table>
<thead>
<tr>
<th>Hypothesis Name</th>
<th>Pharmacophoric Features</th>
<th>Total Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesis 2A</td>
<td>Hydrogen Bond Acceptor</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ring Aromatic</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Positive Ionisable</td>
<td></td>
</tr>
<tr>
<td>Hydrophobe (Aliphatic)</td>
<td></td>
<td>196</td>
</tr>
<tr>
<td>Hypothesis 2B</td>
<td>3x Hydrophobe (Aliphatic)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Positive Ionisable</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
</tr>
<tr>
<td>Hypothesis 2C</td>
<td>2x Hydrophobe (Aliphatic)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Positive Ionisable</td>
<td></td>
</tr>
<tr>
<td>Hydrogen Bond Acceptor</td>
<td></td>
<td>201</td>
</tr>
<tr>
<td>Null Hypothesis 2</td>
<td>3x Hydrophobe (Aliphatic)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Positive Ionisable</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>220</td>
</tr>
</tbody>
</table>
Table 8.3.4.2 Linear regression plot statistics of experimental biological activity versus predicted biological activity for hypothesis 2A, 2B and 2C.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>$r^2$</th>
<th>$q^2$</th>
<th>Spearman Coefficient ($r_s$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2A</td>
<td>0.143</td>
<td>0.036</td>
<td>0.13</td>
</tr>
<tr>
<td>2B</td>
<td>0.239</td>
<td>0.144</td>
<td>0.51</td>
</tr>
<tr>
<td>2C</td>
<td>0.305</td>
<td>0.218</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Table 8.3.4.3 Results for the 5-feature pharmacophores generated using the GOLD produced conformers for the ligands in the test set.

<table>
<thead>
<tr>
<th>Hypothesis Name</th>
<th>Pharmacophoric Features</th>
<th>Total Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesis 2D</td>
<td>Hydrophobic Aliphatic</td>
<td>135</td>
</tr>
<tr>
<td></td>
<td>Hydrophobic Aromatic</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ring Aromatic</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Positive Ionisable</td>
<td></td>
</tr>
<tr>
<td>Hypothesis 2E</td>
<td>Hydrophobic Aliphatic</td>
<td>137</td>
</tr>
<tr>
<td></td>
<td>Hydrophobic Aromatic</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ring Aromatic</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Positive Ionisable</td>
<td></td>
</tr>
<tr>
<td>Hypothesis 2F</td>
<td>Hydrophobic Aliphatic</td>
<td>141</td>
</tr>
<tr>
<td></td>
<td>Hydrophobic Aromatic</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ring Aromatic</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Positive Ionisable</td>
<td></td>
</tr>
<tr>
<td>Null Hypothesis 2'</td>
<td>Positive Ionisable</td>
<td>165</td>
</tr>
<tr>
<td></td>
<td>3x Hydrophobe (Aliphatic)</td>
<td></td>
</tr>
</tbody>
</table>
Table 8.3.4.4 Linear regression plot statistics of experimental biological activity versus predicted biological activity for hypothesis 2D, 2E and 2F.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>$r^2$</th>
<th>$q^2$</th>
<th>Spearman Coefficient ($r_s$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D</td>
<td>0.186</td>
<td>0.08</td>
<td>0.44</td>
</tr>
<tr>
<td>2E</td>
<td>0.170</td>
<td>0.07</td>
<td>0.66</td>
</tr>
<tr>
<td>2F</td>
<td>0.075</td>
<td>-0.04</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Hypotheses 2D, 2E, and 2F were then used to predict the biological activities for the ten compounds of the test set. The regression statistics are summarised in Table 8.3.4.4.

8.3.5 Hypothesis 3: Protein-Based Pharmacophore

The final pharmacophore produced was based on the 2D6 comparative model built in Chapter 4. The interaction map generated for the active site is shown in Figure 8.3.5.1A. This interaction map was edited manually to leave interaction clusters, 2 hydrogen bond donor clusters (blue) and a hydrophobic cluster (grey), from which the cluster centres were determined and the protein-based pharmacophore generated. Figure 8.3.5.1B.
Figure 8.3.5.1 The protein-based 2D6 pharmacophore. A) Interaction map illustrating hydrophobic regions (grey circles) and hydrogen-bond acceptors (blue and white cylinders) in the 2D6 active site. B) The resultant protein-based pharmacophore, comprising 3 pharmacophoric features, 2 hydrogen bond acceptors (blue and white spheres) and one hydrophobic region (white sphere).
The protein-based pharmacophore was then used to predict the biological activities of the compounds in the test set. This experiment was repeated using different radii for the excluded volume spheres (0.2, 0.4, 0.6, 0.8, 1.0 Å). The regression statistics are summarised in Table 8.3.5.1. Unlike the previous studies the production of the protein-based pharmacophore does not involve the training set, hence there is no null hypothesis. The experiment was repeated using the GOLD generated conformers for the test set. Regression statistics are summarised in Table 8.3.5.2.

Table 8.3.5.1 Linear regression plot statistics of experimental biological activity versus predicted biological activity for the protein-based hypotheses validated using the Catalyst generated conformers for the test set.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Radii of Exclusion Spheres (Å)</th>
<th>r²</th>
<th>q²</th>
<th>Spearman Coefficient (r_s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3B</td>
<td>0.8</td>
<td>0.017</td>
<td>-0.179</td>
<td>-0.03</td>
</tr>
<tr>
<td>3C</td>
<td>0.6</td>
<td>0.017</td>
<td>-0.179</td>
<td>0.006</td>
</tr>
<tr>
<td>3D</td>
<td>0.4</td>
<td>0.00017</td>
<td>-0.199</td>
<td>0.10</td>
</tr>
<tr>
<td>3E</td>
<td>0.2</td>
<td>9.17 x 10^{-5}</td>
<td>-0.199</td>
<td>0.052</td>
</tr>
<tr>
<td>3F</td>
<td>0.0</td>
<td>0.016</td>
<td>-0.105</td>
<td>0.24</td>
</tr>
</tbody>
</table>
Table 8.3.5.2 Linear regression plot statistics of experimental biological activity versus predicted biological activity for the protein-based hypotheses validated using the GOLD generated conformers for the test set.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Radii of Exclusion</th>
<th>$r^2$</th>
<th>$q^2$</th>
<th>Spearman Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>3H</td>
<td>0.8</td>
<td>0.04</td>
<td>-0.199</td>
<td>0.025</td>
</tr>
<tr>
<td>3I</td>
<td>0.6</td>
<td>$2.8 \times 10^{-5}$</td>
<td>-0.249</td>
<td>-0.13</td>
</tr>
<tr>
<td>3J</td>
<td>0.4</td>
<td>$2.8 \times 10^{-5}$</td>
<td>-0.249</td>
<td>-0.13</td>
</tr>
<tr>
<td>3K</td>
<td>0.2</td>
<td>$6.8 \times 10^{-8}$</td>
<td>-0.25</td>
<td>-0.13</td>
</tr>
<tr>
<td>3L</td>
<td>0.0</td>
<td>0.0061</td>
<td>-0.118</td>
<td>0.18</td>
</tr>
</tbody>
</table>

8.4 Discussion

Initially the results of the pharmacophore investigations appear to be disappointing. The statistics from the regression analysis show, almost universally, that there is little correlation between the experimentally determined pIC50s in the test set and those predicted using the various pharmacophores, and hence the pharmacophores would have limited use in the prediction of biological activity for a novel pharmaceutical. However, to objectively assess whether or not the experiments have been a success they need to be put into context alongside others that have attempted to achieve similar aims.

The $r$ value obtained for the correlation between the predicted activity of the training set compounds and the experimentally determined biological activities is 0.87. This compares well to studies carried out by Ekins and co-workers (Ekins et al 1999) whereby two datasets were utilised ($r$ values of 0.75 and 0.91 respectively) to generate a 2D6 pharmacophore. Similar to the work carried out in this chapter Ekins and his colleagues attempted to assess the predictive ability of the pharmacophores generated
by using a test set of known 2D6 substrates (15 molecules). These predictions were compared by means of a residual, which was calculated from the difference (in log units) between the predicted and the experimental activity values. A predicted value within one log unit of an experimentally determined value was considered to be a good estimate of activity. The results obtained by Ekins and co-workers predicted biological activity successfully for between 53-66% of molecules in the datasets. This method of analysis was used to assess the results obtained for pharmacophores 1A-F and 2A-F. Analysis of the protein based pharmacophores (3A-L) in this manner was not possible, as there are no predicted activities generated for the test set compounds from these pharmacophores. The results for hypotheses 1 and 2 are summarised in Table 8.4.1.
Table 8.4.1 Experimentally observed and predicted pIC50 values for the 2D6 test set. Highlighted regions represent those values > 1 residual log unit.

<table>
<thead>
<tr>
<th>Drug</th>
<th>Experimental pIC50</th>
<th>1A</th>
<th>1B</th>
<th>1C</th>
<th>1D</th>
<th>1E</th>
<th>1F</th>
<th>2A</th>
<th>2B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pentazocine</td>
<td>7.8</td>
<td>5.7</td>
<td>5.7</td>
<td>5.7</td>
<td>6.8</td>
<td>6.1</td>
<td>5.8</td>
<td>6.1</td>
<td>5.6</td>
</tr>
<tr>
<td>Lobeline</td>
<td>6.6</td>
<td>5.7</td>
<td>5.7</td>
<td>5.8</td>
<td>4.8</td>
<td>4.9</td>
<td>5.2</td>
<td>5.3</td>
<td>5.7</td>
</tr>
<tr>
<td>ARC053620</td>
<td>5.9</td>
<td>5.2</td>
<td>4.7</td>
<td>4.7</td>
<td>6.2</td>
<td>6.5</td>
<td>5.3</td>
<td>5.2</td>
<td>4.7</td>
</tr>
<tr>
<td>Haloperidol</td>
<td>5.5</td>
<td>5.7</td>
<td>5.9</td>
<td>5.7</td>
<td>5.0</td>
<td>4.7</td>
<td>5.2</td>
<td>5.5</td>
<td>5.9</td>
</tr>
<tr>
<td>Clozapine</td>
<td>4.7</td>
<td>5.7</td>
<td>5.7</td>
<td>5.5</td>
<td>6.0</td>
<td>5.6</td>
<td>6.2</td>
<td>5.9</td>
<td>5.7</td>
</tr>
<tr>
<td>Corynanthine</td>
<td>7.0</td>
<td>6.9</td>
<td>6.6</td>
<td>6.7</td>
<td>5.1</td>
<td>5.1</td>
<td>6.0</td>
<td>6.7</td>
<td>6.6</td>
</tr>
<tr>
<td>Econazole</td>
<td>6.3</td>
<td>5.7</td>
<td>5.9</td>
<td>5.7</td>
<td>5.7</td>
<td>5.7</td>
<td>6.0</td>
<td>5.9</td>
<td>5.9</td>
</tr>
<tr>
<td>Promethazine</td>
<td>5.6</td>
<td>5.2</td>
<td>5.0</td>
<td>6.0</td>
<td>6.4</td>
<td>6.6</td>
<td>5.3</td>
<td>5.3</td>
<td>5.0</td>
</tr>
<tr>
<td>Imipramine</td>
<td>5.2</td>
<td>5.4</td>
<td>5.3</td>
<td>5.6</td>
<td>5.8</td>
<td>5.5</td>
<td>5.3</td>
<td>5.3</td>
<td>5.3</td>
</tr>
<tr>
<td>Carbamazepine</td>
<td>4.5</td>
<td>4.3</td>
<td>4.7</td>
<td>4.7</td>
<td>5.2</td>
<td>5.2</td>
<td>5.3</td>
<td>4.8</td>
<td>4.8</td>
</tr>
</tbody>
</table>

Number of Prediction with Residual

< 1 log unit | 9 | 8 | 8 | 7 | 7 | 7 | 7 | 8
Special Note

Page 220 missing from the original
Although the actual number of compounds in the test set is less than that used by Ekins and his co-workers, the extent of structural diversity within the test sets, and the range of activities (Ekins test set spans 4 log units of activity, ours spans 3.3 log units of activity) are comparable. Hence, viewing the results in this manner indicates that the results achieved surpass precedents set in the literature (predicting between 70 and 90% of the biological activities of the molecules in the test set to within one log unit c.f. 53-66% from Ekins work) and that far from being disappointing, despite the poor regression analysis statistics, the results are as good as can be expected from utilising this technique in this manner.

There are five molecules in the test set that have incorrectly predicted biological activities returned by one or more of the pharmacophores. The most active compound in the test set, pentazocine is underpredicted by approximately 2 orders of magnitude for 9 of the 12 hypotheses. This is most likely due to a lack of characterisation for tightly binding molecules in the training set. The data for molecules at this end of the activity range is sparse, and increasing the number of the compounds in the training set with pIC50 values in this area may result in a pharmacophore better able to predict accurate values for the tightly binding compounds.

The remaining four molecules that are badly predicted by at least one pharmacophore (lobeline, clozapine, corynanthine and ARC053620) are all relatively large. A further observation is that the majority of these poor predictions (13/19) are for compounds in which the GOLD docked conformations for generation of the hypothesis was used. It is feasible that the GOLD docked orientations for these particular compounds are not fully representative of how they would bind in vivo, which would account for the poor performance of the pharmacophores. Without crystal structures for the complexes it is impossible to determine whether or not this is true. An alternative explanation is that for these compounds the type of inhibition obtained in measuring the pIC50s is non-competitive, such as omeprazole (Vandenbranden et al 1996), which shows mixed inhibition, and ketoconazole (Back et al 1989), which binds to the heme group. However, it is most likely that there may be one or more pharmacophore features that are missing from the models. Given time, this problem could have been addressed by the iterative inclusion of the test set molecules in the training set.
Of the 12 hypotheses generated the Spearman ranking coefficient suggests that three of them are significant in the prediction of the activities of the test set compounds to the 95th percentile (Kmietowicz and Yanoulis 1978). These are 1A, which comprises 3 hydrophobes and a positive ionisable (Figure 8.4.1), 2C which has a positive ionisable, two hydrophobes and a hydrogen bond acceptor (Figure 8.4.2) and 2E which has an aromatic hydrophobe, aliphatic hydrophobe, ring aromatic and positive ionisable (Figure 8.4.3). If we assume that the positive ionisable group is the basic nitrogen atom of many 2D6 compounds the hydrophobic regions are placed at distances between 4.01 and 5.25Å from the positive ionisable in the pharmacophore. Hence, the site of oxidation for these compounds will be in the same region. This is in keeping with previously produced pharmacophores (Ekins et al 2001) and as such lends weight to the theory that the results obtained are as good as can be expected.

**Figure 8.4.1** Hypothesis 2A showing three hydrophobes (blue) and positive ionisable (red) features. Angle $\text{BAC} = 88.9^\circ$, $\text{BAD} = 130.2^\circ$, $\text{CAD} = 54.4^\circ$. 
Figure 8.4.2 Hypothesis 2C showing positive ionisable (red) two hydrophobes (blue) and a hydrogen bond acceptor (green) as features. Angle BAC = 176.7°, BAD = 75.8° and CAD = 61.5°.

Figure 8.4.3 Hypothesis 2E showing positive ionisable (red), hydrophobic aliphatic (light blue) hydrophobic aromatic (dark blue) and ring aromatic (gold) features. Angle BAC = 119.2°, BAD = 183.3° and CAD = 119.0°
The results for the protein-based pharmacophore were disappointing, and failed to predict the activities of the inhibitors in the test set. However, an interesting observation was made during the construction of the pharmacophore. The interaction map generated suggested that there were two regions in the active site that could be involved as hydrogen bond acceptors. The first, unsurprisingly, was Glu 216 in the F-helix region. However, the second centred on the carbonyl oxygen atom of Ala 305 in the I-helix; this is close to the heme iron, and could account for so-called “5Å substrates.” It was postulated that the basic nitrogen atoms of the substrate molecules could form stabilising interactions in this region that facilitate the N-dealkylation pathway (Coutts et al 1994) of metabolites. To investigate this MPTP was docked into the active site using GOLD. The basic nitrogen atom was constrained to within 3Å of the carbonyl oxygen of Ala 305, and the N-methyl group tethered to within 3Å of the iron atom at the centre of the heme. The iron proton distances for twenty dockings were then measured and their average distances ascertained. These were compared with experimental results for the MPTP 2D6 complex obtained using paramagnetic NMR techniques (Modi et al 1996). The results are presented in Table 8.4.2.

Table 8.4.2 Interatomic Fe—H distances for MPTP docked into the active site using GOLD compared with experimentally determined NMR distances.

<table>
<thead>
<tr>
<th>Proton</th>
<th>Experimental Distance (Å) (Modi et al 1997)</th>
<th>Average Distance Restrained Dockings (Å)</th>
<th>Difference Between Experimental and Docking Values (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCH₃</td>
<td>3.4 ± 0.2</td>
<td>3.7 ± 0.1</td>
<td>0.3 ± 0.2</td>
</tr>
<tr>
<td>C₂H₆⁻</td>
<td>4.4 ± 0.1</td>
<td>4.0 ± 0.3</td>
<td>0.4 ± 0.3</td>
</tr>
<tr>
<td>C₂H₆⁺</td>
<td>5.8 ± 0.1</td>
<td>5.2 ± 0.3</td>
<td>0.6 ± 0.3</td>
</tr>
<tr>
<td>C₃H₆⁻</td>
<td>6.1 ± 0.1</td>
<td>5.0 ± 0.1</td>
<td>1.1 ± 0.1</td>
</tr>
<tr>
<td>C₃H₆⁺</td>
<td>7.1 ± 0.2</td>
<td>6.5 ± 0.3</td>
<td>0.6 ± 0.4</td>
</tr>
<tr>
<td>C₅H₆</td>
<td>7.7 ± 0.3</td>
<td>8.0 ± 0.2</td>
<td>0.3 ± 0.4</td>
</tr>
<tr>
<td>C₅H₆⁻</td>
<td>5.4 ± 0.1</td>
<td>5.6 ± 0.2</td>
<td>0.2 ± 0.2</td>
</tr>
<tr>
<td>C₆H₆⁻</td>
<td>6.6 ± 0.1</td>
<td>6.8 ± 0.2</td>
<td>0.2 ± 0.2</td>
</tr>
<tr>
<td>C₇H₆, C₁₁H₆</td>
<td>8.7 ± 0.2</td>
<td>8.5 ± 0.5</td>
<td>0.2 ± 0.5</td>
</tr>
<tr>
<td>C₉H₆⁻</td>
<td>10.9 ± 0.4</td>
<td>10.5 ± 0.4</td>
<td>0.4 ± 0.6</td>
</tr>
<tr>
<td>C₉H₆</td>
<td>11.9 ± 0.4</td>
<td>11.8 ± 0.4</td>
<td>0.1 ± 0.6</td>
</tr>
</tbody>
</table>
The results from the docking experiments correlate well with the NMR results. This suggests that in addition to the major stabilizing aromatic interaction with Phe 481 for the N-dealkylation pathway (de Groot et al 1999) there could be an additional interaction between the I-helix and the basic nitrogen atom of any substrates, possibly by the formation of a hydrogen bond, that helps to facilitate the reaction, or in the case of 2D6 inhibitors serves to stabilise the binding of the ligand in the active site.

8.5 Conclusions

Pharmacophores are useful and powerful tools for defining the prerequisite features of a drug compound that will be metabolised by a specified enzyme. The pharmacophores produced in these studies suggest that important features for a compound to be a substrate of cytochrome P450 2D6 include a positive ionisable group (presumably the basic nitrogen atom found in the majority of 2D6 substrates) and a hydrophobic region between 5-6 Å from the positive ionisable group. This is in agreement with previously produced pharmacophores, suggesting that the approaches to this experiment were valid.

On first inspection it would appear that using the pharmacophores produced yielded disappointing results as the regression statistics suggest that none of the twelve pharmacophores produced are predictive. However, by comparing these results to precedents found in the literature it was discovered that the pharmacophores performed as well as could be expected. The accurate prediction of the biological activities (to within one log unit) of between 70% and 90% of the structurally diverse molecules in a test set is a significant improvement on the 53-66% achieved by Ekins (Ekins et al 2002). Hence, despite the limitations of this technique for predicting experimental activities the results achieved by these studies compare as well as those presented by other groups investigating similar problems.

The protein-based pharmacophores were not predictive. This was probably due to the manner in which they were constructed, as a significant amount of user intervention is required to, somewhat arbitrarily, reduce the vast amount of data generated by an interaction map to produce a meaningful pharmacophore. However, the interaction
map gave insight into the chemical nature of the active site, allowing the identification of a residue that may act as a stabilising influence in the poorly understood N-dealkylation pathway of CYP 2D6. Hence, constructing the pharmacophore was worthwhile for the information that it provided with respect to the N-dealkylation metabolic pathway.

Unsurprisingly pharmacophores do not provide a definitive answer to the problem of accurately predicting the biological activities of possible drug candidates. In the course of these studies twelve very different pharmacophores were built that were capable of predicting the biological activity (to within one log unit) of between 70-90% of the drug molecules in a structurally diverse test set. This is encouraging, although caution must be exercised when using pharmacophores, as their general applicability must not be overrated. Although pharmacophores provide useful insight into which chemical characteristics are important in a drug molecule, the technique is sensitive and dependent upon the test and training sets used to generate the hypotheses. As such any information obtained should be used in conjunction with other computational methods, such as QSAR, as a guide to potential biological activity.
Chapter 9

Overall Conclusions
Computer-aided molecular design is an umbrella term that describes a number of techniques employed in the early stages of the drug-design process to identify, and subsequently refine, lead compounds. The premise that underpins this approach is the reduction of time, and thereby cost, involved in bringing a new pharmaceutical to the market. A range of the techniques used in computer-assisted molecular design have been used in the studies in this thesis.

It is encouraging to see that the techniques employed to produce comparative models of the (known) mammalian P450 2C5 structure (Chapter 3) are identified as correct by both standard validation techniques such as quality checks of stereochemical and amino acid environment, and by comparison with the crystal structure. The combined approach to modelling, which utilises a range of initial templates to generate the comparative models highlights the possible limitations of relying on a single template to produce an accurate model, but also serves to reiterate that there is no single approach that provides the definitive solution to the comparative modelling problem in cases like the P450s where the sequence homology between templates is relatively low. The fact that the results of the study of P450 2C5 were good gives confidence that the models produced for P450 2D6 (Chapter 4) will be as accurate as it is possible to make with the currently available structures.

The work also addresses how important it is to utilise the most up to date information available to produce comparative models (Chapter 5). The quality of a comparative model is entirely dependent on both the quality of, and the region of conformational space sampled by, the templates. Final structures can vary significantly in three-dimensional topology as a result of the different templates used to model them. This is particularly important with respect to the topology of active sites, as demonstrated by the docking of codeine and MPTP into P450 2D6 model, which highlighted the importance of Glu 216 as a residue involved in substrate binding. Subsequent mutagenesis studies confirmed the role of Glu 216 for substrate binding and specificity (McLaughlin et al 2002) The ability to successfully dock such substrates into protein structures is reliant on an efficient docking algorithm. The close correlation of interatomic iron-proton distances of both the MPTP-2D6 and codeine-2D6 complexes with those obtained by NMR indicates that the docking program,
GOLD, coped well with this task, and reiterates the good quality of the comparative models.

The work also showed that computational methods can be used to validate docking results (Chapter 5), as for the contour maps generated by the computer program GRID, and illustrated the capacity of Principal Component Analysis (PCA), which facilitated the investigation of very subtle chemical and topological differences between different comparative models, and the critical assessment of various comparative modelling methods.

The ability to successfully dock a series of ligands into the active site of a protein and quantitatively score each of the ligands to rank them according to their biological activity is one of the greatest challenges faced by the computational chemist. The work (Chapter 6) illustrates the limitations of the scoring functions available – the results suggest that the ranking of a set of structurally diverse compounds that have biological activities varying over a range of four log units is not possible. However, it is important to note that this experiment was ambitious, and errors incorporated by using an homology model, as opposed to an experimental structure, and the fact that there was no experimental evidence to support binding orientations of ligands in the 2D6 active site as suggested by the docking algorithm compounded the problem. This coupled with the acknowledged limitations of the scoring functions investigated made it almost impossible for the scoring functions to perform well. Although the failure may partly be due to problems with the comparative model it is encouraging that some of the results perform favourably when compared to a literature dataset with the same range of biological activities. Hence, it is likely that the poor results are not only due to inaccuracies in the model or dockings, but also an inherent problem arising from the small number of compounds in the dataset and the narrow range of biological activity being investigated. When we consider the limitations of the scoring functions due to an acknowledged lack of understanding of the mechanism of protein-ligand complex formation in addition to the other limitations it is unsurprising that the investigation proved unsuccessful. The scoring functions, although not predictive, performed as well as could be expected and in some cases are comparable with results achieved by other groups.
The QSAR paradigm was also used to predict the activity of two sets of compounds (Chapter 7). Despite the fact that QSAR is most successful when used to study ligands that have a known binding orientation in a receptor of known structure, several of the methods employed produced models that were predictive. The results underline that QSAR is better suited to a dataset consisting of homologous structures as opposed to a structurally diverse dataset, but highlighted the grid-based Molecular Field Analysis (MFA) method as the most appropriate single method for the prediction of biological activities of known structures. In agreement with previous studies it was found that the production of a 'Jury' model, which combines two or more individual QSAR methods and obtains average predicted activities from them, enhanced the predictive abilities of many of the QSAR experiments.

Finally the application of pharmacophores to the prediction of biological activity for substrates of cytochrome P450 2D6 was investigated. These studies (Chapter 8) produced pharmacophores for the 2D6 active site similar to those published in the literature – identifying a positive ionisable feature with a hydrophobic region approximately 5Å from it as prerequisites for a 2D6 substrate. On initial inspection the results of this analysis seemed disappointing in that none of the pharmacophores appeared to be predictive. However, comparison with literature precedents showed that the results were better than the majority of the published data, and hence as good as could be expected. This work also showed how pharmacophores can be used in conjunction with a protein structure to identify potentially important residues for substrate binding that can be overlooked in more conventional studies i.e. the identification of Ala 305 as a residue that may play an important role in the N-dealkylation pathway for substrates of P450 2D6. These studies have illustrated that by far the most powerful tool for gaining insight into the structure and function of a protein is structure-based analysis, whereby comparative models of an unknown structure are built and docking studies of known substrates carried out, although useful information can be obtained with other computational approaches such as QSAR.

Although there is no doubt that a high resolution experimental structure is always preferable to a three-dimensional comparative model, in the absence of such a structure a comparative model provides a testable hypothesis that is invaluable in the
study of a protein, the interaction of the protein with ligands and investigating the rationale of how a protein functions within its biological context. Such computational studies also provide a powerful predictive basis for subsequent experimental work.

Incorporating a new template (2C5) into the comparative modelling of cytochrome P450 2D6, highlighted a residue in the F-Helix (Glu 216) as one that could be crucial to substrate binding and specificity, for all substrates - not just those that have sites of oxidation 10-12 Å from the basic moiety as was previously believed. The results of these studies were used to direct laboratory-based mutation studies that showed when Glu 216 was mutated, catalytic turnover and substrate specificity were affected, thus proving the importance of this residue. The studies also gave insight into the role of another residue - Asp 301. It was previously assumed that this substrate was directly responsible for sequestering the basic nitrogen atoms of substrates in 2D6. However, docking studies showed little direct interaction between substrates and Asp 301, suggesting that this residue, proven to be crucial to catalytic activity by mutagenesis experiments, may play a subtler indirect role in the metabolism of substrates by P450 2D6, possibly by acting as a structural anchor for the flexible B–C region and increasing the net negative charge in the active site. This hypothesis for the indirect role of Asp 301 is supported by recent work carried out by Guengerich (Guengerich 2002). Hence, by incorporating the 2C5 crystal structure into a comparative model of 2D6 the understanding of the mechanisms involved in substrate metabolism has been significantly advanced.

In the future the number of mammalian P450 crystal structures will increase. These should be incorporated into the comparative modelling of 2D6 so that the model built over the course of these studies could be refined. Utilising structure-based analysis and cheminformatics techniques on refined models would then serve to strengthen the results of these investigations, or highlight other residues/regions in the 2D6 active site that are important for metabolism, giving greater insight into the function of the protein. With an increase in the number of mammalian P450-ligand complex crystal structures, and a greater understanding of the mechanism of protein-ligand complex formation, scoring functions can be developed which are better able to accurately rank the activities of diverse sets of structural compounds, which would speed up the drug-discovery process significantly. Also, with an ever-increasing database of ligand
structures and self-consistent biological data, techniques such as QSAR and pharmacophores, which are not without limitations at the current level of technology, will also play vital roles in determining ideal drug candidates for a range of different enzymes in the near future.
Chapter 10

References

Accelrys San Diego USA (2002)


Connolly, M.L. (1983) Science 221(4612):709-713


Hill T.L (1960) "An introduction to statistical thermodynamics." *Addison-Wesley Publishing Company, Reading MA.*


Lewis, D. F. (1996) "Cytochrome P450 Structure Function and Mechanism" *Taylor and Francis* UK pp ix


247

Chapter 3

Appendix
Appendix 3.1 Modeller input file used to generate comparative models of P450 2C5 using multiple 3-D templates

INCLUDE
SET TOPLIB = './top.heav2.lib'
SET PARLIB = './par.lib'
SET HETATM_IO = 'ON'
SET ATOM_FILES DIRECTORY = '
SET ALNFILE = '2c5_validation.pir'
SET KNOWNs = '1c5_fit' '2hpd_fit' '1oxa_fit' '2cpp_fit'
SET SEQUENCE = '2C5'

SET STARTING_MODEL = 1
SET ENDING Model = 15
SET DEVIATION = 3.0
SET R AND SEED = 950809
CALL ROUTINE = 'model'

SUBROUTINE ROUTINE = 'special_restraints'

SET RESTRAINT_TYPE = 'distance'
SET DISTANCE_RSR_MODEL = 7
SET RESTRAINT_STDEV = 0.0 0.0
SET RESIDUE_SPAN_RANGE = -9999 1 1 9999
SET MAXIMAL DISTANCE = 6.5
SET ADD_RERAINTS = ON
SET RESTRAINT_GROUP = 27

PICK_ATOMS
PICK_ATOMS_SET = 2, ATOM_TYPES = 'ALL', RES_TYPES = 'HEME'
PICK_ATOMS
PICK_ATOMS_SET = 3, ATOM_TYPES = 'CA', RES_TYPES = 'ALL'

MAKE_RERAINTS
SET MAXIMAL DISTANCE = 3.0

PICK_ATOMS
PICK_ATOMS_SET = 2, ATOM_TYPES = 'FE', RES_TYPES = 'HEME'
PICK_ATOMS
PICK_ATOMS_SET = 3, ATOM_TYPES = 'SG', RES_TYPES = 'CYS'

MAKE_RERAINTS
SET MAXIMAL DISTANCE = 4.0

PICK_ATOMS
PICK_ATOMS_SET = 2, ATOM_TYPES = 'FE', RES_TYPES = 'HEME'
PICK_ATOMS
PICK_ATOMS_SET = 3, ATOM_TYPES = 'CB', RES_TYPES = 'CYS'

MAKE_RERAINTS

DELETE_ALIGNMENT
SET PICK_ATOMS_SET = 1
SET RESTRAINT_STDEV = 0.0 1.0

END_SUBROUTINE
Appendix 3.2 Modeller input file used to generate comparative models of P450 2C5 using a single 3-D template.

INCLUDE
SET TOPLIB = './top_heav2.lib'
SET PARLIB = './par.lib'
SET HETATM_IO = 'ON'
SET ATOM_FILES_DIRECTORY = './'
SET ALNFILE = '2c5_validation_single.pir'
SET KNOWNS = '2hpd_fit'
SET SEQUENCE = '2c5'

SET STARTING_MODEL= 1
SET ENDING_MODEL = 15
SET DEVIATION = 3.0
SET RAND_SEED = 950809

CALL ROUTINE = 'model'

SUBROUTINE ROUTINE = 'special_restraints'

SET RESTRAINT_TYPE = 'distance'
SET DISTANCE_RSR_MODEL = 7
SET RESTRAINT_STDEV = 0.1 0.0
SET RESIDUE_SPAN_RANGE = -9999 -1 1 9999
SET MAXIMAL_DISTANCE = 6.5
SET ADD_RESTRAINTS = ON
SET RESTRAINT_GROUP = 27

PICK_ATOMS PICK_ATOMS_SET = 2, ATOM_TYPES = 'ALL', RES_TYPES = 'HEME'
PICK_ATOMS PICK_ATOMS_SET = 3, ATOM_TYPES = 'CA', RES_TYPES = 'ALL'
MAKERESTRAINTS
SET MAXIMAL_DISTANCE = 3.0

PICK_ATOMS PICK_ATOMS_SET = 2, ATOM_TYPES = 'FE', RES_TYPES = 'HEME'
PICK_ATOMS PICK_ATOMS_SET = 3, ATOM_TYPES = 'CB', RES_TYPES = 'CYS'
MAKERESTRAINTS
SET MAXIMAL_DISTANCE = 4.0

DELETE_ALIGNMENT
SET PICK_ATOMS_SET = 1
SET RESTRAINT_STDEV = 0.0 1.0

END_SUBROUTINE
Chapter 4

Appendix
Appendix 4.1 Modeller input file used to generate comparative models of P450 2D6 using a single 3-D template.

INCLUDE
SET TOPLIB = '/top_heav2.lib'
SET PARLIB = './par.lib'
SET HETATM_IO = 'ON'
SET ATOM_FILES_DIRECTORY = '/'
SET ALNFILE = 'single_templates.pir'
SET KNOWNS = '2c5_ong'
SET SEQUENCE = '2D6'

SET STARTING_MODEL = 1
SET ENDING_MODEL = 15
SET DEVIATION = 3.0
SET RAND SEED = 950809

CALL ROUTINE = 'model'
SUBROUTINE ROUTINE = 'special_restraints'

SET RESTRAINT_TYPE = 'distance'
SET DISTANCE_RSR_MODEL = 7
SET RESTRAINT_STDEV = 0.1 0.0
SET RESIDUE_SPAN_RANGE = -9999 -1 1 9999
SET MAXIMAL DISTANCE = 6.5
SET ADD_RESTRANTS = ON
SET RESTRAINT_GROUP = 27

PICK ATOMS PICK ATOMS SET = 2, ATOM_TYPES = 'ALL', RES_TYPES = 'HEME'
PICK ATOMS PICK ATOMS SET = 3, ATOM_TYPES = 'CA', RES_TYPES = 'ALL'
MAKE_RESTRANTS
SET MAXIMAL DISTANCE = 3.0

PICK ATOMS PICK ATOMS SET = 2, ATOM_TYPES = 'FE', RES_TYPES = 'HEME'
PICK ATOMS PICK ATOMS SET = 3, ATOM_TYPES = 'CB', RES_TYPES = 'CYS'
MAKE_RESTRANTS
SET MAXIMAL DISTANCE = 4.0

DELETE_ALIGNMENT
SET PICK_ATOMS_SET = 1
SET RESTRAINT_STDEV = 0.0 1.0

SET RESTRAINTS_FORMAT = 'USER'
READ_RESTRANTS FILE = 'nmr.rsr', ADD_RESTRANTS = on
SET RESTRAINTS_FORMAT = 'MODELLER'
RETURN

END_SUBROUTINE
Appendix 4.2 Modeller input file used to generate comparative models of P450 2D6 using multiple 3-D templates.

INCLUDE
SET TOPLIB = './top heav2.1ib'
SET PARLIB = './par.lib'
SET HETATM IO = 'ON'
SET ATOM FILES DIRECTORY = './'
SET ALNFILE = 'multiple_templates.pir'
SET KNOWNS = '2c5_orig' 'lcpt_fit' '2hpd_fit' '1oxa_fit' '2cpp_fit'
SET SEQUENCE = '2D6'
SET STARTING_Model = 1
SET ENDING_Model = 15
SET DEVIATION = 3.0
SET RAND SEED = 950809
CALL ROUTINE = 'model'

SUBROUTINE ROUTINE = 'special restraints'

SET RESTRAIN_TYPE = 'distance'
SET DISTANCE_RSR_MODEL = 7
SET RESTRAIN_STDEV = 0.1 0.0
SET RESIDUE_SPAN_RANGE = -9999 -1 1 9999
SET MAXIMAL DISTANCE = 6.5
SET ADD_RERAINTS = ON
SET RESTRAIN_GROUP = 27

PICK_ATOMS PICK_ATOMS_SET = 2, ATOM TYPES = 'ALL', RES TYPES = 'HEME'
PICK_ATOMS PICK_ATOMS_SET = 3, ATOM TYPES = 'CA', RES TYPES = 'ALL'
MAKE_RERAINTS
SET MAXIMAL DISTANCE = 3.0

PICK_ATOMS PICK_ATOMS_SET = 2, ATOM TYPES = 'FE', RES TYPES = 'HEME'
PICK_ATOMS PICK_ATOMS_SET = 3, ATOM TYPES = 'SG', RES TYPES = 'CYS'
MAKE_RERAINTS
SET MAXIMAL DISTANCE = 4.0

PICK_ATOMS PICK_ATOMS_SET = 2, ATOM TYPES = 'FE', RES TYPES = 'HEME'
PICK_ATOMS PICK_ATOMS_SET = 3, ATOM TYPES = 'CB', RES TYPES = 'CYS'
MAKE_RERAINTS

DELETE_ALIGNMENT
SET PICK_ATOMS_SET = 1
SET RESTRAIN_STDEV = 0.0 1.0

SET RESTRAIN_FORMAT = 'USER'
READ_RERAINTS FILE = 'nmr.rsr', ADD_RERAINTS = on
SET RESTRAIN_FORMAT = 'MODELLER'
RETURN

END_SUBROUTINE
Chapter 5
Appendix
Appendix 5.1 – Example of a GRIN input file

MOLECULAR DISCOVERY LIMITED

Command File for Programme GRIN

: Assign Channel Numbers and Output File Names:

LOUT  6
LOUT hpd_imposed_grinlout.gdl
KOUT  20
KOUT hpd_imposed.kout

: Assign Channel Numbers and Input File Names:

INKO  11
INKO hpd_imposed_for_grin.pdb
INAT  10
INAT /o2raid/MIM2data/GRID/grid_18/grub.dat

: Provide Control Parameters:

ALHY  0.040
NEHY  1
IHVA  0
LEVL  3
MOVE  1
QQHY  0.000
VDHY  0.600
IEND

:Record the name of the executable Programme and the name of the Default Directory as comment lines starting with a colon:

:PROG: /o2raid/MIM2data/GRID/grid_18/grin

:DEFA:
Appendix 5.2 – Example of a GRID input file

MOLECULAR DISCOVERY LIMITED

Command File for Programme GRID

Assign Channel Numbers and Output File names:

Assign Channel Numbers and Input File names:

Provide Control Parameters:

:Record the name of the executable Programme and the name of the Default Directory as comment lines starting with a colon

:PROG: /o2raid/MIM2data/GRID/grid_l8/grids

:DEFA: /o2raid/MIM2data/users/sbk/GRID/NMe3+/
Appendix 5.3. Perl script used to extract data from GRID output prior to analysis using GOLPE for the Principal Component Analysis of 2D6 comparative models.

```
#!/usr/local/bin/perl -w

$n = 0;
$m = 0;
$k = 0;

while (<>) {
    chomp();
    $n++;
    if ($n == 1) {
        Stitle = $_;
    } elsif ($n == 2) {
        Sngrids = $_;
    } elsif ($n == 3) {
        Snrows = $_;
    } else {
        $m++;
        if ($m == 1) {
            } elsif ($m == 2) {
                Sprotein = $_;
                Sprotein = "Protein_" . "Sprotein";
                $$protein++;
            } else {
                $k++;
                $$phash{Sprotein}->[$$protein -l]->[$k -1] = $_;
                if ($k == Sngrids) {
                    $m = 0;
                    $k = 0;
                }
            }
        }
    }
    $k++;
    $$phash{Sprotein}->[$$protein -1]->[$$protein -1] = $_;

    if ($k == $ngrids) {
        $m = 0;
        $k = 0;
    }
}

while (($key,$value) = each %$phash) {
    for ($columns = 0;$columns < $ngrids;$columns++) {
```
$rowtotal = 0;

foreach $row (@$value) {
    $rowtotal = $rowtotal + $row->[columns];
}
if ($rowtotal == 5 * $$protein) {
    #print "$key = $rowtotal";
    #print "$columns\n";
    $check_array[$columns] = 1;
}
}

$noGrids = 0;
foreach $f (@check_array) {
    if ($f) {
        $noGrids++;
    }
}

$newGrids = $ngrids - $noGrids;
print "$title\n";
print "$newGrids\n";
print "$rows\n";
$rowNumber = 0;
while (($key,$value) = each %$phash) {
    foreach $row (@$value) {
        $rowNumber++;
        print "$rowNumber\n";
        print "$key\n";
        for ($cols = 0;$cols < $ngrids;$cols++) {
            if ($check_array[$cols]) {
                print "-99.99\n";
            } else {
                print $row->[cols] . "\n";
            }
        }
    }
}
Appendix 6.1

Dataset 1
Ajmalicine $pIC_{50} = 7.9$

Pentazocine $pIC_{50} = 7.8$

Hydroquinidine $pIC_{50} = 7.7$

Corynanthine $pIC_{50} = 7.0$

Propafenone $pIC_{50} = 6.9$

Thioridazine $pIC_{50} = 6.8$

Perphenazine $pIC_{50} = 6.6$

Perhexilene $pIC_{50} = 6.5$

Econazole $pIC_{50} = 6.3$
Dextromethorphan $pIC_{50} = 5.1$

Amitryptiline $pIC_{50} = 5.1$

Sparteine $pIC_{50} = 5.1$

Lido
caine

ARC116195 $pIC_{50} = 5.0$

ARC115096 $pIC_{50} = 4.9$

Metoprolol $pIC_{50} = 4.8$

Methyl
dopa

Clozapine $pIC_{50} = 4.7$

10methylallylphenothiazine $pIC_{50} = 4.7$

ARC019504 $pIC_{50} = 4.6$
Carbamazepine \( pIC_{50} = 4.5 \)

ARC119787 \( pIC_{50} = 4.4 \)

10-decylphenothiazine \( pIC_{50} = 4.3 \)

Ajmaline \( pIC_{50} = 5.6 \)

Maybridge 01581 \( pIC_{50} = 6.0 \)
Appendix 6.2

Dataset 2
ARC150779AP pIC50 = 6.1

ARC150257AR pIC50 = 5.3

ARC148343AP pIC50 = 6.4

ARC151535AJ pIC50 = 5.3

ARC148343AP pIC50 = 6.4

ARC151968AJ pIC50 = 5.5

ARC151969AJ pIC50 = 7.0

ARC152722AA pIC50 = 4.5

ARC152723AA pIC50 = 5.4

ARC154704AA pIC50 = 4.4


ARC155657AP pIC50 = 6.4

ARC155658AP pIC50 = 5.7

ARC155688AP pIC50 = 7.5

ARC155797AJ pIC50 = 5.7

ARC156021AP pIC50 = 4.8
Appendix 7.1

Training Set
Ajmalicine $pIC_{50} = 7.9$
10-decylphenothiazine $pIC_{50} = 4.3$
Hydroquinidine $pIC_{50} = 7.7$
Quinidine

ARC128343 $pIC_{50} = 3.8$
Propafenone $pIC_{50} = 6.9$
Thioridazine $pIC_{50} = 6.8$

Perphenazine $pIC_{50} = 6.6$
Perhexilene $pIC_{50} = 6.5$
Maybridge 01581 $pIC_{50} = 6.0$
Chlorpro

Ct

Ajmalin
Ethopromazine $pIC_{50} = 6.2$

Yohimbine $pIC_{50} = 6.2$

ARC081384 $pIC_{50} = 4.6$

Quinine $pIC_{50} = 5.8$

Codeine $pIC_{50} = 5.8$

Desipramine $pIC_{50} = 5.6$

MaybridgeRJC012 $pIC_{50} = 5.6$

ARC019504 $pIC_{50} = 4.6$

Timolol $pIC_{50} = 5.5$
ARC05361 pIC50 = 5.5
ARC119787 pIC50 = 4.4
Fluphenazine pIC50 = 5.5
Flunarazine pIC50 = 5.3
Clomipramine
Tamoxifen pIC50 = 5.3
OEH
Propanolol
Ketoconazole pIC50 = 5.2
10-methylallylphenothiazine pIC50 = 4.7
Dextromethorphan pIC50 = 5.1
Amitryptiline pIC50 = 5.1
Sparteine pIC50 = 5.1

ARC116195 pIC50 = 5.0
ARC115096 pIC50 = 4.9
Metoprolol pIC50 = 4.8
Special Note

Page 275 missing from the original
Appendix 7.2

Test Set
Pentazocine pIC50 = 7.8
Lobeline pIC50 = 6.6
ARC053620 pIC50 = 5.9
Clozapine pIC50 = 4.7
Corynanthine pIC50 = 7.0
Econazole pIC50 = 6.3
Carbamazepine pIC50 = 4.5
Imipramine pIC50 = 5.2
Publications
