RATIONAL MOLECULAR DESIGN

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy in the Faculty of Science

1999

Susan Elizabeth Austin

School of Pharmacy and Pharmaceutical Sciences

TABLE OF CONTENTS

FIGURES 10

TABLES 15

ABSTRACT 17

ACKNOWLEDGEMENTS 21

LIST OF ABBREVIATIONS 25

CHAPTER ONE. GENERAL INTRODUCTION

1.1. INTRODUCTION TO RATIONAL MOLECULAR DESIGN 29

1.2. OVERVIEW OF THESIS 35

CHAPTER TWO. STRUCTURAL STUDIES OF AN ALKYLIMIDAZOLE BINARY OLIGONUCLEOTIDE SYSTEM

2.1. INTRODUCTION 37

2.1.1. Sequence-specific modification of nucleic acids 37

2.1.2. RNA cleavage 40

2.1.3. Aims 42

2.1.4. Applications of Nuclear Magnetic Resonance (NMR) spectroscopy in biological systems 45

2.1.5. NMR techniques used in the structure determination of biopolymers 46

2.1.5.1. Shift COrrelation SpectroscopY (COSY) 47

2.1.5.2. Quantum transitions and phase coherence 49

2.1.5.3. Double-Quantum Filtered COSY (DQF COSY) 49

2.1.5.4. TOtal Correlation SpectroscopY (TOCSY) 50

2.1.5.5. Nuclear Overhauser Effect SpectroscopY (NOESY) 51

2.1.5.6. Longitudinal relaxation 52

2.1.5.6.1. T_1 measurements 53

2.1.5.7. Transverse relaxation 55

2.1.5.7.1. T_2 measurements: the spin-echo experiment 56

2.1.6. Biomolecular structural studies 57

2.1.6.1. Assignment by isotope labelling 57

2.1.6.2. Proton assignment of DNA 58

2.1.6.2.1. Non-exchangeable protons: residue-specific assignment 58

2.1.6.2.2. Sequential assignment of DNA from NOESY spectra 60

2.1.6.2.3. Exchangeable protons 63

2.1.6.2.4. Stereospecific assignment of H2', H2", H5' and H5" 65

2.1.7. Determination of solution structures from analyses of NOESY spectra 65

2.1.7.1. Interproton distance determination 65

2.1.7.2. The isolated spin-pair approximation (ISPA) 67

2.1.7.3. COmplete Relaxation Matrix Analysis (CORMA) 69

2.1.7.4. Direct calculation of distances from experimental spectra (DIRECT) 69

2.1.7.5. Iterative Relaxation Matrix Analysis (IRMA) 70

2.1.7.6. Matrix Analysis of Relaxation for Discerning Geometry of an Aqueous Structure (MARDIGRAS) 72

2.1.8. Summary 74

2.1.9. Aims 74

2.2. MATERIALS AND METHODS 75

2.2.1. Synthesis of oligonucleotide derivatives and complex preparation for NMR spectroscopy 75

2.2.2. NMR Spectroscopy and instrumentation 75

2.2.3. Molecular modelling 77

2.2.4. Structural characterisation 78

2.3. RESULTS AND DISCUSSION 79

2.3.1. NMR Spectroscopy of free oligonucleotides **1**, **2** and **3** 79

2.3.2. Qualitative NMR analysis of the 1:1:1 complex of **1:2:3** 79

2.3.3. Qualitative NMR analysis of the alkylimidazole groups in the binary system 86

2.3.4. Investigation of base-pair hydrogen bonding by NMR studies of exchangeable imino protons 88
2.3.5. Evaluation of distance-range constraints using MARDIGRAS 90
2.3.6. Restrained molecular dynamics calculations 90
2.3.7. Structural analysis of DNA backbone and bases 90
2.3.7.1. Axis-base pair parameters 92
2.3.7.2. Intra-base pair parameters 92
2.3.7.3. Inter-base pair parameters 92
2.3.7.4. Glycosidic torsion angles and sugar puckers 93
2.3.7.5. Backbone torsion angles 93

2.4. OVERVIEW OF STRUCTURE DETERMINED FOR OLIGONUCLEOTIDE COMPLEX 100

CHAPTER THREE. COMPARATIVE MODELLING OF THE TRANSCRIPTION FACTOR HYPOXIA-INDUCIBLE FACTOR-1 WITH THE DESIGN AND SYNTHESIS OF POSSIBLE PEPTIDE INHIBITORS

3.1 INTRODUCTION 105
3.1.1. Homology modelling 105
3.1.1.1. Protein structure prediction 105
3.1.1.1.1. Predicting protein structure from first principles 105
3.1.1.1.2. Secondary structure prediction using rule based approaches 106
3.1.1.1.3. Tertiary structure prediction: fold recognition 108
3.1.1.1.4. Predicting protein structures by 'threading' 109
3.1.1.2. Comparative modelling methods 110
3.1.1.2.1. Protein sequence alignment 110
3.1.1.2.2. Constructing a comparative model 113
3.1.1.2.3. Fragment-based modelling 114
3.1.1.2.4. Restraint-based modelling 115
3.1.1.3. Modelling on the Internet 116

3.1.1.3.1. The PredictProtein (PP) server 116

3.1.1.3.2. The Swiss-Model automated protein modelling server 117

3.1.1.3.3. Structural genomics web sites 117

3.1.1.4. Energy minimisation, molecular dynamics and model evaluation 119

3.1.1.5. Summary 122

3.1.2. Helix-loop-helix transcription factors 123

3.1.2.1. Gene regulation and transcription 123

3.1.2.2. General DNA-binding proteins 123

3.1.2.3. Helix-loop-helix (HLH) proteins 124

3.1.2.3.1. The HLH DNA-binding site 126

3.1.2.3.2. Classification: structure and function 126

3.1.2.4. Hypoxia-Inducible Factor-1 (HIF-1) 131

3.1.2.5. Id3 133

3.1.3. Binding assays 133

3.1.3.1. Sodium Dodecyl Sulphate PolyAcrylamide Gel Electrophoresis (SDS-PAGE) 133

3.1.3.2. Resonant mirror biosensor 134

3.1.3.3. Fluorescence spectroscopy 135

3.1.3.3.1. Fluorescence Resonance Energy Transfer (FRET) 137

3.1.4. Aims 140

3.2. MATERIALS AND METHODS 142

3.2.1. Comparative modelling 142

3.2.1.1. Database searching and sequence alignment 142

3.2.1.2. Modelling of helix backbones 148

3.2.1.3. Modelling of loop regions 150

3.2.1.4. Side chain modelling 150

3.2.1.5. DNA-bound model 150

3.2.2. Peptide design, synthesis and binding studies 151

3.2.2.1. Materials 151

3.2.2.2. Peptide design 152

3.2.2.3. Peptide synthesis 154

3.2.2.3.1. First residue attachment 154

3.2.2.3.2. Coupling procedure 154

3.2.2.3.3. Fluorescent labelling 154

3.2.2.3.4. Side-chain deprotection 155

3.2.2.3.5. Cleavage 155

3.2.2.3.6. Characterisation 155

3.2.2.4. Peptide binding studies: SDS-PAGE assays 155

3.2.2.4.1. Gel preparation and electrophoresis conditions 155

3.2.2.4.2. Glutathione S-transferase-Id3 fusion protein, E12 and E47 production 156

3.2.2.4.3. HLH protein dimerisation controls 157

3.2.2.4.4. Acetonitrile control 157

3.2.2.4.5. Peptide assays on Id3 157

3.2.2.5. Peptide binding studies: Resonant mirror biosensor studies 158

3.2.2.5.1. Immobilisation of glutathione (GSH) onto a carboxymethyl (CM)-dextran surface 158

3.2.2.5.2. Immobilisation of Glutathione S-Transferase (GST)-Id3 onto GSH-derived-surface 160

3.2.2.5.3. E12-GST binding to immobilised Id3-GST 161

3.2.2.5.4. Immobilisation 161

3.2.2.5.5. Peptide binding 161

3.3. RESULTS AND DISCUSSION 164

3.3.1. Molecular modelling 164

3.3.1.1. Sequence alignments 164

3.3.1.2. Tertiary structure predictions of HIF-1 164

3.3.1.2.1. Hydrophobic core 170

3.3.1.2.2. Interface of helices 172

3.3.1.2.3. Loop region 174

3.3.1.2.4. Base of the four α-helix bundle: DNA-binding 176

3.3.1.3. Prediction of mutations for fluorescent labelling 182

3.3.1.4. Summary 182

3.3.2. Peptide synthesis and binding 184
3.3.2.1. Mass spectrometry 184
3.3.2.2. SDS-PAGE assays 184
3.3.2.2.1. HLH protein dimerisation controls 184
3.3.2.2.2. Acetonitrile control 184
3.3.2.2.3. Peptide assays 184
3.3.2.3. Resonant mirror biosensor studies 189
3.3.2.3.1. E12-GST binding to Id3-GST 189
3.3.2.3.2. Id3-GST immobilisation 189
3.3.2.3.3. Peptide binding 191
3.3.2.4. Summary 195

CHAPTER FOUR. A LIGAND-DOCKING STUDY OF THE ENZYME TRYPANOTHIONE REDUCTASE

4.1. INTRODUCTION 200
4.1.1. Molecular Docking 200
4.1.2. Algorithms 201
4.1.2.1. Descriptors 201
4.1.2.2. Grid searches 202
4.1.2.3. Fragment 202
4.1.2.4. Kinetic 203
4.1.2.5. Genetic 203
4.1.2.6. Non-rigid body searching by Fourier Transform docking 204
4.1.2.7. Low-resolution docking 204
4.1.3. Energy evaluation 205
4.1.4. The AUTODOCK docking program 207
4.1.4.1. AUTOTORS 207
4.1.4.2. Modelling hydrogen-bonds 207
4.1.4.3. AUTOGRID 208
4.1.4.4. AUTODOCK 210

4.1.5. Summary 211
4.1.6. Trypanothione reductase 212
4.1.6.1. Structural features of trypanothione reductase 212
4.1.7. Aims 218

4.2. MATERIALS AND METHODS 223
4.2.1. Docking procedure 223
4.2.1.1. Preparation of ligands and receptor for docking runs 223
4.2.1.2. AUTOTORS 223
4.2.1.3. AUTOGRID 223
4.2.1.4. AUTODOCK 224
4.2.1.5. Cluster analysis 224

4.3. RESULTS AND DISCUSSION 225
4.3.1. Validation of the minimisation technique 225
4.3.2. Docking studies 225
4.3.2.1. General overview of docking results 225
4.3.2.2. Phenothiazines 235
4.3.2.4. Open ring quaternary compounds 241
4.3.2.5. Imipramines 244
4.3.2.6. Mepacrine 247

4.4. CONCLUSION 249
4.4.1. Present study 249
4.4.2. Other docking studies on trypanothione reductase 250
4.4.3. Conclusions of the AUTODOCK study 251
4.4.4. Previous applications of AUTODOCK 252
4.4.5. Hydrophobic docking 254
4.4.6. Limitations of docking programs 255

CHAPTER FIVE. SUMMARY OF THESIS 257

CHAPTER SIX. REFERENCES 259

FIGURES

Figure 1.1(a) A simple molecular-level machine 30

Figure 1.1(b) An example of a molecule used in electronic and colour switches, exploited for its tautomeric properties 30

Figure 1.2 'DMP 323', the DuPont Merck compound chosen for clinical trials against HIV protease as a result of rational inhibitor design 33

Figure 2.1 The binary approach to sequence-specific nucleic acid modification 39

Figure 2.2 Mechanism of action of ribonuclease. The two imidazole rings of His12 and His119 act as the base and acid catalysts, respectively 41

Figure 2.3 Sequence-specific cleavage by a diimidazole construct mimicking the cleaving activity of ribonuclease A 43

Figure 2.4 The binary system **1:2:3**. The alkylimidazole groups on ^{18}C and ^{19}T are represented as **Im** and **Im'**, respectively 44

Figure 2.5 General pulse sequence schemes for some 2-D NMR experiments 48

Figure 2.6(a) The principle of the inversion-recovery experiment for measurement of spin-lattice or longitudinal relaxation time, T_1 54

Figure 2.6(b) The spin-echo experiment for measurement of spin-spin or transverse relaxation time, T_2 54

Figure 2.7(a) The four bases of DNA showing exchangeable (red) and non-exchangeable protons (blue) and connectivities between the latter 59

Figure 2.7(b) Intra-base connectivities 59

Figure 2.8 Sequential assignment of DNA. Intra- and inter-base connectivities are shown 61

Figure 2.9 2-D NOE spectrum of a 12-mer DNA duplex to illustrate the different regions of cross-peaks 62

Figure 2.10 Pulse sequence for the REFOPT experiment 64

Figure 2.11 Network of five protons in a molecule illustrating cross-relaxation rate effects arising from dipole-dipole interactions and non-dipolar relaxation 66

Figure 2.12 Schematic diagram of the iterative relaxation matrix analysis (IRMA). The variation introduced in the MARDIGRAS algorithm compared to IRMA is shown 71

Figure 2.13 Flow chart of the MARDIGRAS algorithm 73

Figure 2.14 ^{1}H-NOESY spectrum (600MHz, mixing time 200ms, 25°C) of complex **1:2:3** in 100mM NaCl, 10mM NaH_2PO_4/Na_2HPO_4, pH7.20, 0.1mM EDTA and 1.2mM TSP, prepared in 99.9% D_2O 81

Figure 2.15 Expanded region of the ^{1}H-NOESY spectrum (600MHz, mixing time 200ms, 25°C) of complex **1:2:3**, showing the H6/H8 - H1'/H5/H3' resonance region. Buffer composition was as described for Figure 2.14.

(a) The sequential assignment of oligonucleotide protons for the 12-mer 82

(b) The sequential assignment of oligonucleotide protons for the two hexamers 83

Figure 2.16 ^{1}H NMR variable temperature experiments recorded for the imino protons of complex **1:2:3** in 100mM NaCl, 10mM NaH_2PO_4/Na_2HPO_4, pH7.20, 0.1mM EDTA and 1.2mM TSP, prepared in 90%H_2O:10%D_2O (v/v). Peak numbers indicate the respective base pairs of the duplex 89

Figure 2.17 Starting structure for **1:2:3** complex and the final structure obtained by averaging the last 100 co-ordinate sets of the final 5ps of RMD calculations followed by minimisation 91

Figure 2.18 Helical parameters for the **1:2:3** complex obtained using CURVES 5.3.

(a) Axis-base pair parameters 94

(b) Intra-base pair parameters 95

(c) Inter-base pair parameters 97

(d) Backbone torsion angle parameters 99

Figure 2.19 Starting structure and final structure after RMD of a similar oligonucleotide binary system but bearing photosensitising pyrene and photoactivatable perfluoroazide (arylazide) groups which result in a huge distortion of the helix. (Kindly provided by Debbie Marks). 101

Figure 3.1 Processes in comparative modelling 121

Figure 3.2 The bHLHZ transcription factor Max bound to DNA 125

Figure 3.3 Structure of the resonant mirror sensing device 135

Figure 3.4 Pathways for production and de-excitation of an excited state 136

Figure 3.5 Absorption and emission spectra of two donor-acceptor pairs used in FRET studies 138

Figure 3.6 Sequence alignment of the HLH region of proteins representative of the HLH family 144

Figure 3.7 Sequence alignment of HIF-1α and HIF-1β with the five proteins whose 3-D structures have been solved to date 148

Figure 3.8 The 48 peptides designed and synthesised 153

Figure 3.9 Reaction scheme for protein immobilisation on CM-Dextran 159

Figure 3.10 Direct immobilisation of Id3-GST on a CM-Dextran cuvette 162

Figure 3.11(a) The molecular model of Hypoxia-Inducible Factor-1 (HIF-1) 165

Figure 3.11(b) HIF-1 bound to DNA 166

Figure 3.12 The bHLHZ transcription factor USF bound to DNA 167

Figure 3.13 The bHLHZ transcription factor MyoD bound to DNA 168

Figure 3.14 The bHLHZ transcription factor E47 bound to DNA 169

Figure 3.15 Residues involved in hydrophobic interactions important in the dimerisation of HIF-1 171

Figure 3.16 Inter- and intra-helical electrostatic interactions conferring stability to the HIF-1 heterodimer 173

Figure 3.17 Residues important in providing electrostatic interactions to stabilise the loops of HIF-1 175

Figure 3.18 A comparison of HIF-1 binding regions in different genes. The core sequence of the HRE is shown 177

Figure 3.19(a) HIF-1 bound to DNA. The amino acids important in forming electrostatic interactions are shown in more detail in Figure 3.19(b). The core HRE sequence of DNA (CGTG) is labelled 178

Figure 3.19(b) Some HIF-1 residues important in forming interactions with DNA 179

Figure 3.20 The two pairs of residues of HIF-1 to be mutated for FRET studies. The distances between the amino acids are shown 183

Figure 3.21(a) Hydrophobic residues of the HIF-1α monomer used in the design of complementary peptides 185

Figure 3.21(b) Hydrophobic residues of the Id3 monomer used in the design of complementary peptides 186

Figure 3.22 Mass spectrum (electrospray) of peptide 4(1) 187

Figure 3.23 Binding responses of aliquots of E12, 20μl and 30μl, (300mgml-1 in PBS), to immobilised Id3-GST (association) followed by washing with 2x50μl PBS (dissociation) 190

Figure 3.24 Peptide samples binding to Id3-GST immobilised directly on a CM-Dextran cuvette. (Peptides 9-1 to 9-4 are at a concentration of 200nM and peptides E and F are at 590nM and 480nM, respectively) 192

Figure 3.25 Binding response of peptide sample 9(4) at varying concentrations to Id3-GST directly immobilised on a CM-Dextran cuvette 192

Figure 3.26 Binding response of peptide sample 9(4) to Id3-GST directly immobilised on CM-Dextran: comparison with binding to native CM-Dextran as a negative control 193

Figure 3.27 Peptide samples binding to GST-Id3 (A) immobilised on a CM-Dextran cuvette 194

Figure 3.28(a) Possible binding orientation of peptide 5(1) to the Id3 monomer 196

Figure 3.28(b) Possible binding orientation of peptide 1(1) to HIF-1α 197

Figure 4.1 An example of a grid parameter (input) file for AUTOGRID 209

Figure 4.2(a) Glutathione substrate for glutathione reductase 213

Figure 4.2(b) Trypanothione substrate for trypanothione reductase 213

Figure 4.2(c) The alternative trypanothione reductase substrate N^1-glutathionly-spermidine disulphide, $[GspdS]_2$ 213

Figure 4.3(a) Structure of the enzyme trypanothione reductase 214

Figure 4.3(b) The active site of trypanothione reductase 215

Figure 4.4 Mepacrine bound in the active site of trypanothione reductase 216

Figure 4.5(a) Anthracene manually docked into the active site of trypanothione reductase 219

Figure 4.5(b) Representation of the cationic ammonium ion substituent of potential inhibitors manually docked into the active of trypanothione redictase, showing putative electrostatic interactions with residues E466' and E467' 220

Figure 4.5(c) Chlomipramine manually docked into active site of trypanothione reductase, showing putative hydrophobic and electrostatic interactions with residues used in the initial design steps 221

Figure 4.6 Root structures of phenothiazines and imipramines 217

Figure 4.7 Inhibitor of trypanothione reductase based on the lead tricyclic framework, but also possessing a quaternary amine site added to increase binding strength through electrostatic interactions with residues E466' and E467' 222

Figure 4.8(a) Binding mode I(a). The quaternary nitrogen of the ligand electrostatically interacts with E466' and/or E467' and the tricyclic moiety lies in the Z-site 232

Figure 4.8(b) Binding mode I(b). The quaternary nitrogen of the ligand electrostatically interacts with E466' and/or E467' and the hydrophobic aromatic group on the quaternary nitrogen lies in the Z-site 233

Figure 4.8(c) Binding mode II. The quaternary nitrogen of the ligand electrostatically interacts with S14 (occasionally with a hydrophobic moiety lying near L17/W21/M113) 234

Figure 4.9(a) Fraction of phenothiazine ligands in the top 20 clusters found in binding modes I(a) and I(b) 236

Figure 4.9(b) Fraction of phenothiazine ligands in the top 20 clusters found in binding modes I(c) and II 237

Figure 4.10 Fraction of open ring ligands in the top 20 clusters found in each binding mode 242

Figure 4.11 Fraction of imipramine ligands in the top 20 clusters found in each binding mode 245

TABLES

Table 2.1 Chemical shifts of **Im** and **Im'** protons for free alkylimidazole groups in methanol, when coupled to the respective hexamer (AGAAACp-**Im** and **Im'**-pTGATAC) and after hybridisation of conjugates **1** and **2** with target oligonucleotide **3** 79

Table 2.2 Chemical shifts of the non-exchangeable protons of the 12-mer component (**1**) in complex **1:2:3** at 25°C, pH7.20 in 100mM NaCl, 10mM NaH_2PO_4/Na_2HPO_4, 0.1mM EDTA, 1.2mM TSP, prepared in 99.9% D_2O 84

Table 2.3 Chemical shifts (ppm) of the non-exchangeable protons for components **2** and **3** in complex **1:2:3** at 25°C, pH7.20 in 100mM NaCl, 10mM NaH_2PO_4/Na_2HPO_4, 0.1mM EDTA, 1.2 mM TSP, prepared in 99.9% D_2O 85

Table 2.4 Spin-lattice (T_1) and spin-spin relaxation times (T_2 and T_2*) obtained for some well-resolved signals in complex **1:2:3** determined at 25°C in 100mM NaCl, 10mM NaH_2PO_4/Na_2HPO_4, pH 7.2, 0.1mM EDTA and 1.2mM TSP, prepared in 99.9% D_2O 87

Table 2.5 Chemical shifts of the exchangeable imino protons of complex **1:2:3** at 25°C, pH7.20 in 100mM NaCl, 10mM NaH_2PO_4/Na_2HPO_4, 0.1mM EDTA, 1.2mM TSP, prepared in H_2O. Note that the shifts of ^{1}G, ^{2}T, and ^{12}T imino protons were measured only at 11°C as these peaks were absent at 25°C 88

Table 3.1 Sequence identity and conserved residue similarity (shaded grey) for HIF-1α and β and the five proteins with crystal structures measured over helix 1 (31 residues) and helix 2 (21 residues) 149

Table 3.2 Residues in the HIF-1 dimer having the potential to form electrostatic interactions 174

Table 3.3 Electrostatic interactions with the potential for stabilising loop residues 176

Table 3.4 Proposed electrostatic interactions between DNA and HIF-1. The co-ordinates of the DNA from the crystal structure of USF were used and the bases mutated to a sequence based on the human erythropoeitin 3'

	enhancer site. The core of the HRE sequence is shown	180
Table 4.1	Examples of ligand docking programs	201
Table 4.2	Results of docking the phenothiazine family of ligands	227
Table 4.3	Results of docking the open ring family of ligands	229
Table 4.4	Results of docking the imipramine family of ligands	230
Table 4.5	Results of docking mepacrine (charged and uncharged)	230
Table 4.6	The RMS differences between four crystal structures found in the Cambridge Crystallographic Database structurally related to the docked ligands and their theoretical structures calculated in SYBYL	231
Table 4.7	The fraction of phenothiazine ligands in the top 20 clusters found in each binding mode	235
Table 4.8.	The fraction of open ring ligands in the top 20 clusters found in each binding mode	241
Table 4.9.	The fraction of imipramine ligands in the top 20 clusters found in each binding mode	244
Table 4.10.	The fraction of mepacrine conformations in the top 20 clusters found in each binding mode	247

ABSTRACT

This thesis study has applied a range of computational and molecular modelling techniques to three different problems in rational molecular design and three-dimensional structure determination.

To target uniquely a stretch of nucleic acids in the human genome by anti-sense methods would require an excessively long complementary oligonucleotide. To overcome this the binary system of a complementary-addressing nuclei acid sequence has been proposed. Based on this approach, oligonucleotides conjugated to a diimidazole construct mimicking the catalytic centre of ribonuclease A have been shown by Vlassov's group to exhibit sequence-specific RNA cleavage. The solution structure of a binary oligonucleotide system containing alkylimidazoles constructs was determined in this thesis study by high-resolution 2-D NMR spectroscopy in combination with restrained molecular dynamics. The model binary system chosen, **1:2:3**, comprised a 12-mer target sequence pdGTATCAGTTTCT (**1**) and two oligonucleotide derivatives, dAGAAACp-**Im** (**2**) and **Im'**-pdTGATAC (**3**), complementary to the adjacent hexamer sequences of **1** (where **Im** and **Im'** are β-alanylhistamine groups). Characterisation of complex **1:2:3** using melting experiments monitored by 1-D NMR spectroscopy of imino protons showed that neither the nick in the DNA backbone between the short oligonucleotides **2** and **3** in the fully formed **1:2:3** binary system duplex, nor the presence of the alkylimidazole groups, significantly destabilized the complex. Assignment of oligonucleotide and modifying group protons was performed using ^{1}H COSY and NOESY experiments. Comprehensive analysis of ^{1}H NOESY spectra of **1:2:3** showed a continuous set of intra- and inter-nucleotide interactions, typical of regular, right-handed double-stranded B-DNA. Despite the presence of the break in the DNA backbone between ^{18}C and ^{19}T, cross-peaks of normal intensities between the sugar ring protons of ^{18}C and aromatic protons of ^{19}T were observed, indicating the continuation of helical regularity in this nick region. A variable-temperature NMR experiment performed for the imino protons of **1:2:3** showed that the centre of the complex was, in fact, the most stable part of the system.

Proton-proton distance ranges were calculated using the full-relaxation matrix analysis implemented in the MARGIDRAS algorithm using the NOESY spectrum of **1:2:3**, (600MHz, 200ms), and the resulting 315 distance-ranges were used as restraints in subsequent molecular dynamics calculations. The final structure showed very slight distortions from the regular form of

B-DNA overall, with one alkylimidazole group positioned in the region of the major groove and the other in the minor groove. No favoured stable interaction between the imidazole groups and oligonucleotide residues was observed, with conformational flexibility of the modifying moieties within this binary system. This observation contrasts with the only other 3-D structural data for a binary system, incorporating pyrenyl and tetrafluoroazido groups, where the complex shows a very high distortion from regular B-DNA.

Hypoxia-inducible factor-1 (HIF-1) is a key component of a widely operative transcriptional response activated by hypoxia. Cells deficient in HIF-1 show a much-reduced ability to grow as solid tumours, providing a rationale for developing selective inhibitors of HIF-1 as anti-tumour agents. HIF-1 is a heterodimeric DNA-binding complex composed of two basic helix-loop-helix (bHLH) Per-AHR-ARNT-Sim proteins (HIF-1α and β). A 3-D structural model was constructed for the bHLH domain of human HIF-1 based on the X-ray crystal structures of the bHLH proteins MyoD and USF. The features of this model compared well with the dimerisation and DNA-binding features of other bHLH transcription factors whose X-ray structures had been determined. Mutations on each monomer were suggested which could be used in fluorescence resonance energy transfer studies to monitor protein dimerisation. Peptide inhibitors of dimerisation of HIF-1 and dimerisation inhibitors of the structurally related HLH transcription inhibitor protein Id3, important in cellular differentiation, were designed and 48 synthesised using semi-combinatorial chemistry. The peptides were tested against Id3 using gel electrophoresis and resonant mirror biosensor techniques, the latter technique detecting binding of some peptides.

The enzyme trypanothione reductase (TR), a prime target for the rational design of lead compounds against trypanosomiasis and leishmaniasis, is known to be inhibited by tricyclic molecular frameworks. A new class of inhibitors has been developed in our laboratory, namely the quaternary alkylammonium chlorpromazines containing an additional hydrophobic moiety, which are approximately 30-fold more potent than the parent lead. The rationale for this development was that the N^+ charge is needed for interaction with E466' or E467' in the enzyme active site, the tricyclic or equivalent moiety interacting with the major hydrophobic cleft and the second hydrophobic moiety on the side-chain nitrogen atom interacting with the so-called Z-site. In this thesis the docking program AUTODOCK was used to probe binding of three families of quaternary ligands (phenothiazines, imipramines and open-ring structures), to predict possible ligand binding modes and to rationalise computed relative binding energies with *in vitro* data. Two broad families of binding were found, equal in docking energy: the first with N^+ interacting with E466' or E467' and with a hydrophobic moiety sometimes in the Z-site and the second with the N^+ interacting with S14. The preferences some ligands displayed for one mode over the other could not be rationalised on the basis of structure, nor could the differences in docking energy between ligands in one family be correlated with experimental data of inhibitor strength. The determining factor for the final placement of the ligand appeared to be the satisfaction of the N^+ electrostatic interaction with an appropriate amino acid side-chain.

DECLARATION

No portion of this work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

COPYRIGHT DECLARATION

ACKNOWLEDGEMENTS

There are many people to thank for their encouragement, support and friendship during the last three years. Firstly, I would like to thank my supervisor, Ken Douglas, for his guidance, patience and advice on post-graduate work and research life. Equally important I would like to express my gratitude to a lot of good friends who have helped me in so many ways.

I have to mention and thank Christine first, for her great friendship, understanding and amusing sense of fun. I could not have wished for a more perfect housemate and I wish you every possible happiness and success in the future.

I cannot express my thanks to Harvey, a very dear friend who I am indebted to for his kindness and patience and supplying constant help and encouragement.

A great big thank you goes to Mandy who has helped me with so many problems in the lab as well as being a good friend, keeping me amused and sharing an insatiable passion for chocolate.

Endless thanks go to Elena who has been so kind in helping me with much of the NMR work and has been so patient, fun and easy to work with. I'll miss your wisdom, philosophy and humour.

An extra special thank-you goes to Debbie who has always encouraged me and taught me a great deal about this work. Thank you for all your help with the modelling, docking and graphics, your support and inspiration and being there to talk to. There's so much I couldn't have done without you. I wish you the best of luck always.

I would like to thank Tom for his endless wit and refreshingly positive attitude to everything and Bosco too for providing constant entertainment and for their reminders that life is to be laughed at and enjoyed and is not there to be worked through. It was a pleasure living with you both.

Special thanks have to go to two great chums Jyoati and Clair in Shrewsbury for their support, humour and friendship and always being there both during the last three years and in the past.

Thanks to other members of the Pharmacy Department; academic, secretarial and the support staff, especially all those in the Schorlemmer lab, both past and present. A special mention goes to Cecil and Peter, Nathalie and Ben and others, and also Hong for helping me in those vital first few weeks here. Thanks to Faruk for supplying experimental data and a special thank you to

Lai-Han for never refusing to do my secretarial work.

Thank you to Drs. Richard Deed and John Norton at the Paterson Institute for Cancer Research, Manchester, for their invaluable help with the Id3 work. Thanks also to Dr. Karen King, Experimental Oncology Group, Pharmacy Department, for producing and expressing plasmid constructs of HIF-1 and Arnt, the E.M. Unit, Biological Sciences Department, for producing many of the pictures and Mrs Valerie Boote in the mass spectrometry service, Department of Chemistry.

I would also like to thank the EPSRC-supported High-Field NMR Facility at the University of Edinburgh for instrument time and to Dr. John Parkinson there for his excellent operational assistance and advice. I am grateful to Prof. Gareth Morris and Dr. Martin Baron for assistance with the NMR studies in Manchester.

Thanks also to Dr. Phil Buckle, Tim Kinning and John Aherne from Affinity Sensors, Cambridge, for their advice and assistance with many of the binding studies and thanks also go to all the staff at Tripos, Inc., both in the U.K and abroad for assistance with our SYBYL queries.

I am grateful to Profs. Tom Ellenberger and Stephen Burley for kindly proving the X-ray co-ordinates of E47 and USF, respectively.

I gratefully acknowledge the Biological and Biotechnological Sciences Research Council for their three years of financial support.

Finally, a huge thank-you to my Mum and Dad. An appreciation of your love and support cannot be expressed in a few words.

This thesis would not have been possible without the help of many of the people mentioned above. It is as much theirs as it is mine.

For Mum, Dad

and Keith.

With eternal thanks.

Argue
for your limitations,
and sure enough,
they're
yours.

Richard Bach

LIST OF ABBREVIATIONS

Amino acids:

A, Ala	Alanine	N, Asn	Asparagine
C, Cys	Cysteine	NL	Norleucine
D, Asp	Aspartic Acid	NV	Norvaline
E, Glu	Glutamic Acid	P, Pro	Proline
F, Phe	Phenylalanine	Q, Gln	Glutamine
G, Gly	Glycine	R, Arg	Arginine
H, His	Histidine	S, Ser	Serine
I, Ile	Isoleucine	T, Thr	Threonine
K, Lys	Lysine	V, Val	Valine
L, Leu	Leucine	W, Trp	Tryptophan
M, Met	Methionine	Y, Tyr	Tyrosine

DNA Bases:

A	Adenine	G	Guanine
C	Cytosine	T	Thymine

Others:

Å	Angstrom = 0.1nm
ABNR	Adopted Basis Set Newton-Raphson
AEDANS	5-N-[(iodoacetamidoethyl)amino]naphthalene-1-sulphonic acid
6-AHA	6-Aminohexanioc acid (ε-amino-n-caproic acid)
AHR	Aromatic Hydrocarbon Receptor
AMPS	Ammonium peroxodisulphate
ARNT	Aromatic Hydrocarbon Receptor Nuclear Translocator
^{t}Bu	Tertiary butyl
CORMA	COmplete Relaxation Matrix Analysis
COSY	Shift Correlation Spectroscopy
CPMG	Carr-Purcell-Meibom-Gill pulse sequence

DCM	Dichloromethane
DIC	1,3-Diisopropylcarbodiimide
DMAP	N', N'-dimethylaminopyridine
DMF	N', N'-dimethylformamide
DNA	Deoxyribonucleic acid
DQF COSY	Double-Quantum Filtered COSY
DTT	Dithiothreitol
EDC	1-Ethyl-3-(3-dimethylaminopropyl)carbodiimide
FAD	Flavin adenine dinucleotide
Fmoc	Fluorenylmethoxycarbonyl
fs	Femtosecond
GA	Genetic Algorithm
GR	Glutathione reductase
GSH	Glutathione
GST	Glutathione S-transferase
HIF-1	Hypoxia-inducible factor-1
bHLH	Basic helix-loop-helix
bHLHZ	Basic helix-loop-helix zipper
HOBt	1-hydrozybenzotriazole
HPLC	High performance/pressure liquid chromatography
Hz	Hertz
Id	Inhibitor of differentiation
Im	Imidazole
IRMA	Iterative Relaxation Matrix Analysis
ISPA	Isolated Spin-Pair Approximation
IVT	*In vitro* translate
MARDIGRAS	Matrix Analysis of Relaxation for Dicerning Geometryof an Aqueous Structure
4MβNA	4-Methoxy-β-naphthylamide
MeOH	Methanol
NAD(P)H	Nicotinamide adenine dinucleotide (phosphate) (reduced)
NBD	7-nitrobenz-2-oxa-1,3-diazole-4-yl (nitrobenzofurazan)
NBD-Cl	4-chloro-7-nitrobenz-2-oxa-1,3-diazole-4-yl (4-chloro-7-nitrobenzofurazan)

NHS	N-hydroxysuccinimide
NMR	Nuclear Magnetic Resonance
NOESY	Nuclear Overhauser Effect Spectroscopy
NP40	Nonidet®P40 (Nonylphenylpolyethylene glycol)
ODhbt	3,4-Dihydro-3-hydroxy-4-oxo-1,2,3-benzotriazine ester
OPfp	Pentafluorophenyl ester
PAS	Per-Arnt-Sim
PBS	Phosphate buffered saline
pNA	p-Nitroanilide
ps	Picosecond
REFOPT	Refocussed Optimised water-supression pulse sequence
REFOPTNY	Refocussed Optimised Nuclear Overhauser Effect Spectroscopy
RMS	Root Mean Square
RNA	Ribonucleic Acid
RTC	Rhodamine isothiocyanate
SD	Steepest Descents algorithm
SDS-PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
T_1	Spin-lattice or longitudinal relaxation time
T_2	Spin-spin or transverse relaxation time
TES	Triethylsilane
TFA	Trifluoroacetic acid
TOCSY	Total Correlation Spectroscopy
TR	Trypanothione Reductase
TSP	Sodium 3-(trimethylsilyl-2, 2, 3, 3, H_4)-1-propionate
Tris	Tris(hydroxymethyl)methylamine
Trt	Trityl (triphenylmethyl)

CHAPTER ONE

GENERAL INTRODUCTION

1.1 INTRODUCTION TO RATIONAL MOLECULAR DESIGN

The rational design of molecular systems is found in every area of science and technology. Everywhere Man is using macro- and micro-molecular structural information to design and engineer new molecules with desired characteristics. Molecular structural engineering is being carried out in fields as diverse as computing, biology, physics, medicine, geology, aerospace and chemistry to tailor molecular properties, e.g. electrical, magnetic, thermal and catalytic, to those required. Only a tiny fraction of some examples of recent advances will be mentioned here.

In the world of chemistry, as the quest for ever smaller functional molecular systems continues, scientists are tailor-making inorganic and organic supramolecular 'cages', designed to trap specific metal ions and even organic molecules. This opens up new fields of chemistry associated with molecular recognition, ion-sensing, liquid crystals, display devices and energy conversion systems. Molecular-level mechanical machines have been developed which are composed of a macrocycle and two different thread-like compounds [1]. It is possible to choose, by means of a chemical, electrical or light energy input, which thread enters the macrocycle's cavity, thus the components display changes in their relative positions as a result of an external stimulus. Molecular electronic devices and colour switches have been produced which can be chemically controlled by exploiting the tautomerism between different conformations of a molecule and the corresponding changes in electronic and energetic properties [2]. Figure 1.1 shows a few examples of chemical structures.

An exciting and interesting example of miniature systems is the construction of carbon nanotubes. The attractive material characteristics of these tubes (e.g. electronic properties which vary as a function of diameter and chirality) have opened up doors to electronic, optical, magnetic and mechanical applications e.g. as electronic switches [3] and in superconductors [4]. Filled nanotubes leading to improved catalysts and biosensors are also being developed [5]. Molecular gears have even been designed from carbon nanotubes with benzyne teeth approximately 2nm across [6] with computer simulations suggesting that these gears can operate at up to 50-100GHz in a vacuum at room temperature.

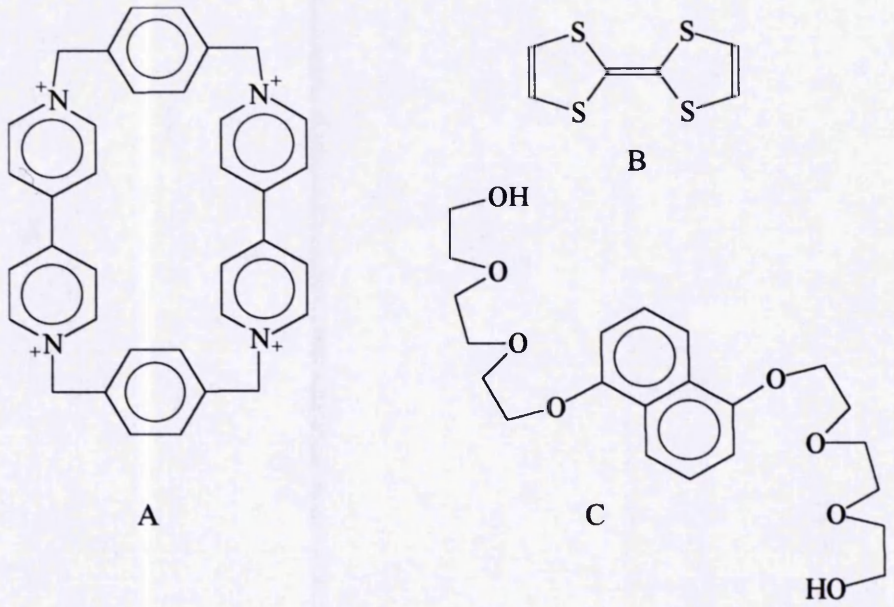


Figure 1.1(a). A simple molecular-level machine [1]. A is the macrocyclic component and B and C are the potential threads.

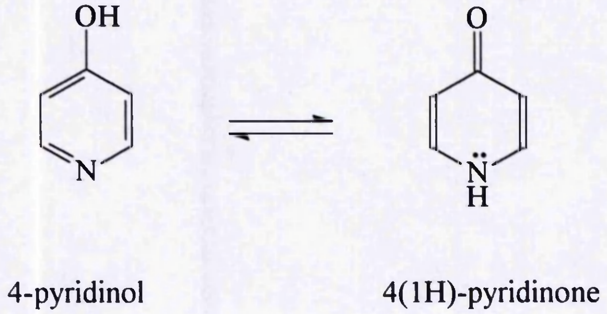


Figure 1.1(b). An example of a molecule used in electronic and colour switches exploited for its tautomeric property [2].

The biological world has been dominated by the idea of protein design. Ever since the widespread establishment of protein engineering in the early 1980s, the hope has been expressed that the systematic manipulation of protein structure and function can be driven by rational, structure-based design approaches. Over the years an impressive array of successful experiments has been carried out by applying qualitative rules of protein structure and function through the use of computer graphics. These include changes in substrate specificity [7,8], introduction of metal binding sites for affinity purification [9], allosteric control [10,11] and even the *de novo* creation of small proteins that adopt defined secondary structures, e.g. α-helical coiled coils [12,13] and helix bundles [14]. Proteins have been engineered to have increased stability, for example by the introduction of disulphide bridges, e.g. subtilisin (an enzyme used in biological washing powders) and λ-repressor protein [15] or by metal-mediated cross-linking [16]. Several advances in rational structure-based design, driven by the emergence of automated protein design programs have occurred, e.g. in the design of a hyperthermophilic variant of the Streptococcal protein Gβ1 domain [17]. This designed seven-fold mutant has a melting temperature in excess of 100°C with optimised core packing, an increased burial of hydrophobic surface area, more favourable helix-dipole interactions and improvement of secondary structure propensity. Given that the design algorithm is based on fundamental physical chemical principles, the prospect of applying the methodology to the redesign of medically and industrially important proteins is excellent.

Recently, successful attempts have been made to combine the enzyme α-chymotrypsin with plastics to produce chemically, mechanically and thermally stable biocatalytic plastic materials [18].

With these aims in mind, computational procedures have been developed for the design of ligand-binding sites of proteins of known structure, in particular metal-binding sites, e.g. zinc [19], calcium for metallobiosensors (W. Chazin, personal communication) or the introduction of metal-binding sites in proteins which do not normally possess them, e.g. thioredoxin [20-23]. Other challenges lie in rational design of protein function: the systematic structure-based engineering of substrate specificity, catalysis and control of activity.

Computational protein engineering is still in its infancy, but ultimately such approaches will lead to true *de novo* design of protein structure in which the backbone fold as well as the side-chain arrangements are designed. Automated design procedures have emerged as a powerful tool to drive manipulations of protein structure and function, both for the study of fundamental principles of structure and function, and for the development of new technologies.

Rational drug design

Discovering a lead compound with therapeutic potential has often been achieved serendipitously or by randomly screening large numbers of samples, either natural, e.g. those found in soil or plants, or through libraries of synthetic compounds. Combinatorial chemistry now provides a means of rapidly generating the large numbers of compounds required in a short time, but advances in computer-based methods means that a more rational approach to drug discovery and design can now be taken.

Designing a potential inhibitor or improving upon a lead structure of a drug using a computer-based approach requires knowledge of the three-dimensional structure of the receptor. The structure of the receptor, e.g. protein, enzyme, DNA, is sometimes known, having been solved by X-ray crystallographic or NMR spectroscopic studies. In cases where this is not so, it may be possible to construct a theoretical three-dimensional model using comparative modelling techniques, so-called homology or, more strictly, comparative or similarity modelling.

If a series of active molecules exists, but with no structural evidence of how they bind, one assumes that all the molecules bind in a common manner to the macromolecule and a pharmacophore is constructed, *i.e.* an abstract model indicating the key molecular features for binding and their spatial relationships. The pharmacophore provides the information needed to perform searches of structural databases for new compounds satisfying both the chemical and the geometrical requirements and to align compounds for use in quantitative structure-activity relationship (QSAR) studies or in the *de novo* design of new ligands. However, for most of the compounds in a typical database, no crystal structure is available so structure generation programs are used to produce one or more low-energy conformations.

Database searching, however, does not provide molecules that are truly 'novel' and many databases are biased towards particular classes of compounds, so limiting the range of structures that can be obtained. In *de novo* design, the three-dimensional structure of the receptor or the pharmacophore is used to design new molecules, either manually, in collaboration with synthetic chemists or using computer-aided automated approaches.

In a manual process, a three-dimensional structure of the target receptor with a substrate or ligand bound is ideally available to give detailed information of the conformation the ligand adopts and the interactions important in binding. This obviously suggests how they can be improved upon by molecular modelling and/or intelligent guesswork. Sometimes if only the structure of the natural substrate bound is available, this can be used to speculate on potential inhibitor structures and interactions.

In *de novo* design there are two basic types of algorithm. The first has been described as 'outside in' methods [24]. Here, the binding site is first analysed to determine where specific functional groups might bind tightly. These groups are then connected together to give molecular skeletons which are then converted into 'real' molecules, e.g. the programs SPROUT [25], LUDI [26] and CAVEAT [27]. In the 'inside out' approach, molecules are grown within the binding site using randomly selected fragments, under the control of an appropriate search algorithm. Each suggestion has its inter- and intra-molecular energy evaluated and high-energy structures are rejected, e.g. LEAPFROG (in SYBYL, Tripos). The results of any computer-aided design process should be viewed with caution and the practicality of suggestions checked for accurate binding predictions, valid geometric and steric conformations, synthetic feasibility *etc.*

Once a lead compound has been identified, a programme of chemical modification is undertaken to enhance its properties. It must be remembered that an inhibitor is not a drug, and account must be taken at an early stage of potential metabolism, chemical stability (e.g. to water, light, oxygen), toxicity, solubility and drug-delivery problems.

An impressive example of the application of structure-based methods was the design of an orally active inhibitor of the HIV protease by a group of scientists at DuPont Merck [28]. The starting point of their work was a series of X-ray crystal structures of the enzyme with a number of inhibitors bound. From these, a three-dimensional pharmacophore was generated and used to search a subset of the Cambridge Structural Database. One of the hit molecules was chosen as a lead compound and further modelling studies based on the X-ray structure were performed to predict the optimal stereochemistry and the conformation required for optimal interaction with the enzyme. A D-phenylalanine-derived cyclic urea with *p*-hydroxymethylbenzyl nitrogen substituents was eventually chosen for clinical trials (Figure 1.2).

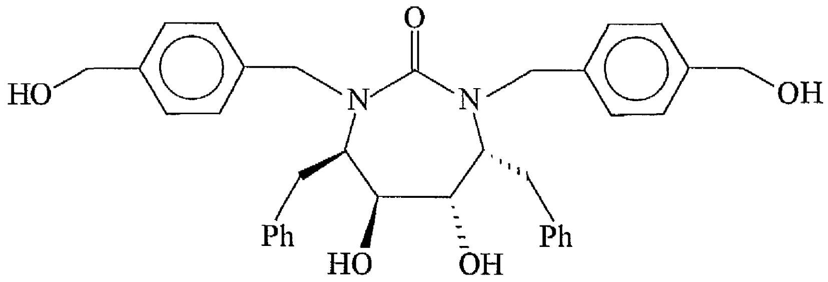


Figure 1.2 'DMP 323', the DuPont merck compound chosen for clinical trials against HIV protease as a result of rational inhibitor design [28].

In contrast to rational drug and ligand design for proteins, the situation for nucleic acid ligand design is less advanced, but is rapidly gaining momentum with therapeutic interest being focussed on ligand binding to prevent transcription and thus production of proteins, e.g. those implicated in disease. Many DNA-binding ligands exist, but there are few specific ligands described for RNA beyond metal ions, polyamines and a few organometallics directed at thiol bases. DNA ligands have been designed to bind, intercalate, damage or cleave DNA, e.g. by binding in the major or minor grooves, intercalating or acting as base analogues. Metal ions often play a part in cleavage reactions. Small molecules have also been made that mimic the destructive oxidation of DNA such as can occur with carcinogenic chemicals and ultra-violet light [29]. Some such molecules intercalate into DNA and, when exposed to high-energy light, cleave DNA by a photo-oxidation process as if it were a pair of light-activated scissors, thus revealing natural weak spots. A set of chemicals has also been built based on modified antibiotics shaped like hairpins that can switch off specific genes [29]. By changing the arrangement of ring structures in the hairpins, the molecules can bind to specific DNA sequences and even distinguish each of the four base pairs. These molecules have been designed to lock in specific ways to DNA's superstructure and have opened up new ways to study and manipulate DNA.

This thesis applies rational molecular design to three situations:

- in experimentally determining a three-dimensional structure of a potentially therapeutic nucleic acid derivative by NMR spectroscopy
- constructing a three-dimensional molecular model of a protein by molecular homology or comparative modelling and from this designing potential lead ligands
- in the use of a docking algorithm to evaluate enzyme-ligand interactions and rationalise the binding with experimental data with a view to manually designing improved inhibitors.

1.2 OVERVIEW OF THESIS

The overall objective of this thesis is to apply a range of computational and molecular modelling techniques to different problems in rational molecular design and three-dimensional structure determination. The work consists of three sections.

The first chapter involves the use of one- and two-dimensional NMR and molecular modelling techniques to elucidate the structure of a modified DNA duplex. The DNA is a model for a system designed to mimic the RNA-cleaving enzyme RNase A. This mimic has been shown to have the required RNase activity but until the studies presented in this thesis, there was no structural information available for the system on a firm three-dimensional level. The findings of this thesis can be used to suggest improvements on the nucleotide-cleavage efficiency of the system and aid in anti-sense drug design.

The second section describes the construction of a three-dimensional model of the basic-helix-loop-helix transcription factor heterodimer hypoxia-inducible factor-1, (HIF-1), by comparative modelling techniques. This part of the thesis provides an example of how rational design processes can proceed when there is not an experimentally based three-dimensional structure of the biological target, such as those provided by X-ray crystallography or NMR spectroscopic studies. HIF-1 is essential for tumour growth and development. Peptide inhibitors of dimerisation of HIF-1, and dimerisation inhibitors of the structurally related transcription inhibitor protein Id3, which is important in cellular differentiation, were designed and synthesised using knowledge-based semi-combinatorial chemistry. The peptides were tested for binding activity to the target using a range of techniques.

The final chapter probes the problem of predicting possible binding modes of families of inhibitors of the parasitic enzyme trypanothione reductase. This was approached using the ligand docking program AUTODOCK. The program is assessed in its ability to rationalise computed relative binding energies with *in vitro* data of inhibitor strength. The information obtained from these studies may be useful in the design of improved second generation or novel lead structures for anti-trypanosomal drugs.

CHAPTER TWO

STRUCTURAL STUDIES OF AN ALKYLIMIDAZOLE BINARY OLIGONUCLEOTIDE SYSTEM

2.1 INTRODUCTION

2.1.1 Sequence-specific modification of nucleic acids

A promising approach for the directed action of chemically modifying reagents to the genetic material of the cell is the sequence-specific modification of nucleic acids by reactive oligonucleotide derivatives bearing covalently attached chemical groups [30-32]. The general idea of the method is to attach a reactive group to an oligonucleotide complementary to the sequence in the target region of the nucleic acid to be modified [33,34]. It is obvious that sequence-specific modification in living cells makes it possible in principle to selectively suppress the expression of a gene, thereby preventing the biosynthesis of specific proteins, e.g. those implicated in diseased states, virus multiplication. Potential applications of the sequence-specific chemical modification of nucleic acids in oncology, virology and other branches of pharmacology have been discussed by Summerton [35] and the approach was proposed and realised with the *in vitro* mutagenesis of phages and plasmids by polyalkylating RNAs complementary to selected DNA sites [36]. Such modified oligonucleotides have found a number of applications in molecular biology and they are considered to have the potential to provide highly efficient and specific therapeutics, capable of inactivating infectious agents and of regulating biosynthetic disease-related disorders. Oligonucleotide analogues and derivatives complementary to specific messenger RNAs have been shown to inhibit expression of the corresponding genes [32,37].

A great variety of chemical compounds has been used successfully in the chemical modification of heterocyclic, sugar or phosphate moieties of nucleic acids. These include alkylating oligonucleotide derivatives [38-40], DNA-EDTA-Fe(II) derivatives [41,42], pyrimidine oligodeoxyribonucleotides [43,44], porphyrin-linked [45,46], proflavin-linked [47] and azidoproflavine-linked oligodeoxyribonucleotides [48] and DNA-psoralen mono- and di-adducts [49-51]. When trying to increase the effectiveness of reactive oligonucleotide derivatives in complicated biological systems, the most important problems to overcome are to increase the site-specificity and efficiency of target nucleic acid modification. Traditional oligonucleotide derivatives contain chemical groups capable of reacting with nucleic acids under physiological conditions or groups generating highly reactive diffusing species which cause the damage. However, the reactions of these groups are generally uncontrolled and oligonucleotide derivatives can affect non-target nucleic acid and non-nucleic acid biopolymers, thereby creating unwanted side effects. Moreover, to target uniquely a stretch of DNA in the human genome would need a complementary oligonucleotide sequence match of about 15-18 base-pairs, a very long species in terms of drug delivery. The hybridisation *in vivo* of the probing oligonucleotide sequence must also occur within a narrow window of physiological conditions. Under such conditions, the long oligonucleotides required not only have the problem of cellular uptake and survival, but can also form numerous imperfect complexes to non-target nucleic acids sequences [52]. Problems in obtaining a high yield are also associated with the synthesis of any long polymer in a pure state.

Some approaches directed towards improving site-specificity and efficiency of antisense-analogous systems designed to modify nucleic acid targets have been proposed [53-56].

(i) The first approach is based on the use of binary systems of oligonucleotides conjugated to relatively inactive precursor groups that can form an active complex when two components are located next to each other, due to simultaneous binding in the adjacent sites of the nucleic acid target (Figure 2.1).

Since the oliogonucleotides bear relatively unreactive groups (R and S), which are activated only on complex formation, they can be expected to produce fewer non-specific effects due to interaction with non-target biopolymers, thus improving site-specificty. The advantage of this system is a higher modification specificity, because it is determined by recognition of two oligonucleotide components, which bind to the target independently and it may be 'switched on' in controlled conditions when the components are correctly aligned. Each of the oligonucleotide components is long enough to avoid non-specific hybridisation under physiological conditions so increasing efficiency. Two shorter polymers also have synthesis and delivery advantages over one long chain.

This approach has been demonstrated in the photolabelling of DNA 56-60. The binary system here consists of a photosensitising group (S), e.g. pyrene and a photoreactive group (R), e.g. arylazide which can be activated by UV light or by the photosensitising group after the independent hybridisation of shorter oligomers with the DNA target. (For examples, see Figure 2.1). The photoactivated arylazide generates an arylnitrene and efficiently covalently labels the target nucleic acid strand. The yield of photo-labelling increased from 33% for a single oligonucleotide complementary to the target, to 65-68% for the full binary system [57,58]

(ii) The second approach involves equipping antisense oligonucleotides with reactive groups capable of irreversibly damaging nucleic acids to improve inhibitory potential of the compounds.

The non-enzymatic sequence-specific cleavage of single-stranded DNA has already been achieved [61]. However, no non-radical damaging methods have been described and radicals can

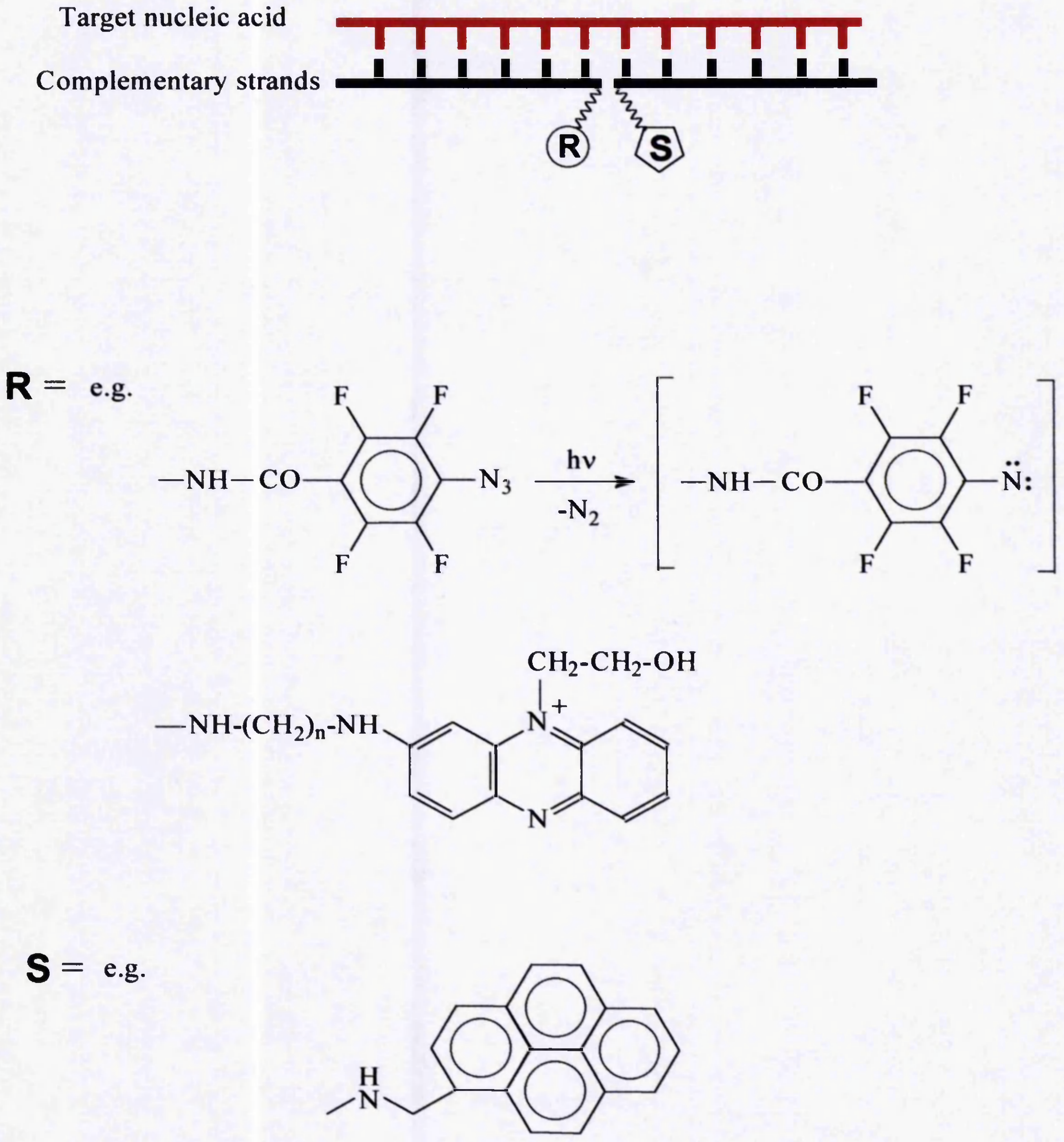


Figure 2.1. The binary system approach to sequence-specific nucleic acid modification. R and S are photoreactive and photosensitising groups, respectively.

produce unwanted reactive by-products and react with other biopolymers. DNA possesses several systems of repair enzymes that remove and replace the damaged part of the strand so a modification could be potentially reversed. RNA does not have such repair mechanisms and therefore appears a more attractive target. Retroviruses, e.g. HIV, do not possess DNA and so any potential drug-induced nuclei acid modification would have to target RNA. tRNA-binding antibiotics already exist, e.g. distamycin. The RNA part of ribosomes is also a target of antibiotic action.

2.1.2 RNA Cleavage

An ideal reactive group for antisense oligonucleotides targetted to RNA would be a group capable of cleaving RNA catalytically [62]. In this case the possibility of the reactions of the compounds with bio/polymers other than RNA could be eliminated and a catalytic turnover of the oligonucleotide derivative would provide high antisense efficiency. Ribozymes are catalytic RNA molecules that promote a variety of reactions including the hydrolytic cleavage of RNA and DNA [63-65]. This class of enzymes requires divalent metal ions for structural and catalytic purposes. RNA cleavage has been achieved by DNA-linked europium (III) texaphyrin [66] and DNA modified with a terpyridine derivative [67], designed to mimic ribozymal activity. Ethylenediamine bound to DNA has been found to selectively hydrolyse the complementary tRNA strand [68], and peptide-acridine conjugates have also been designed with ribonuclease activity [69].

Another approach to achieving irreversible damage to RNA is to design small RNA cleaving catalytic groups by mimicking the active centres of ribonucleases (RNA-cleaving enzymes) using organic molecules. RNase A is structurally well characterised [70,71] and contains two essential histidine residues in its catalytic centre [62,72](Figure 2.2).

The two imidazole rings of His-12 and His-119 act as the acid and base catalysts in RNA cleavage. The pK_a of histidine is normally close to a value of 7 so exists in partially unprotonated (imidazole) and protonated (imidazolium ion) forms in solution and these two forms must coexist for an efficient reaction. The reaction rate therefore displays a pH dependence and occurs in two stages as shown: chain cleavage to form a cyclic phosphodiester and ring opening by attack of water.

Non-specific RNA catalysis has been observed in imidazole buffer [73] and molecules have been designed to incorporate two histidine moieties to mimic RNase A activity by linking two histamine units by a single chain, or via an intercalating phenazine derivative [74,75], or by

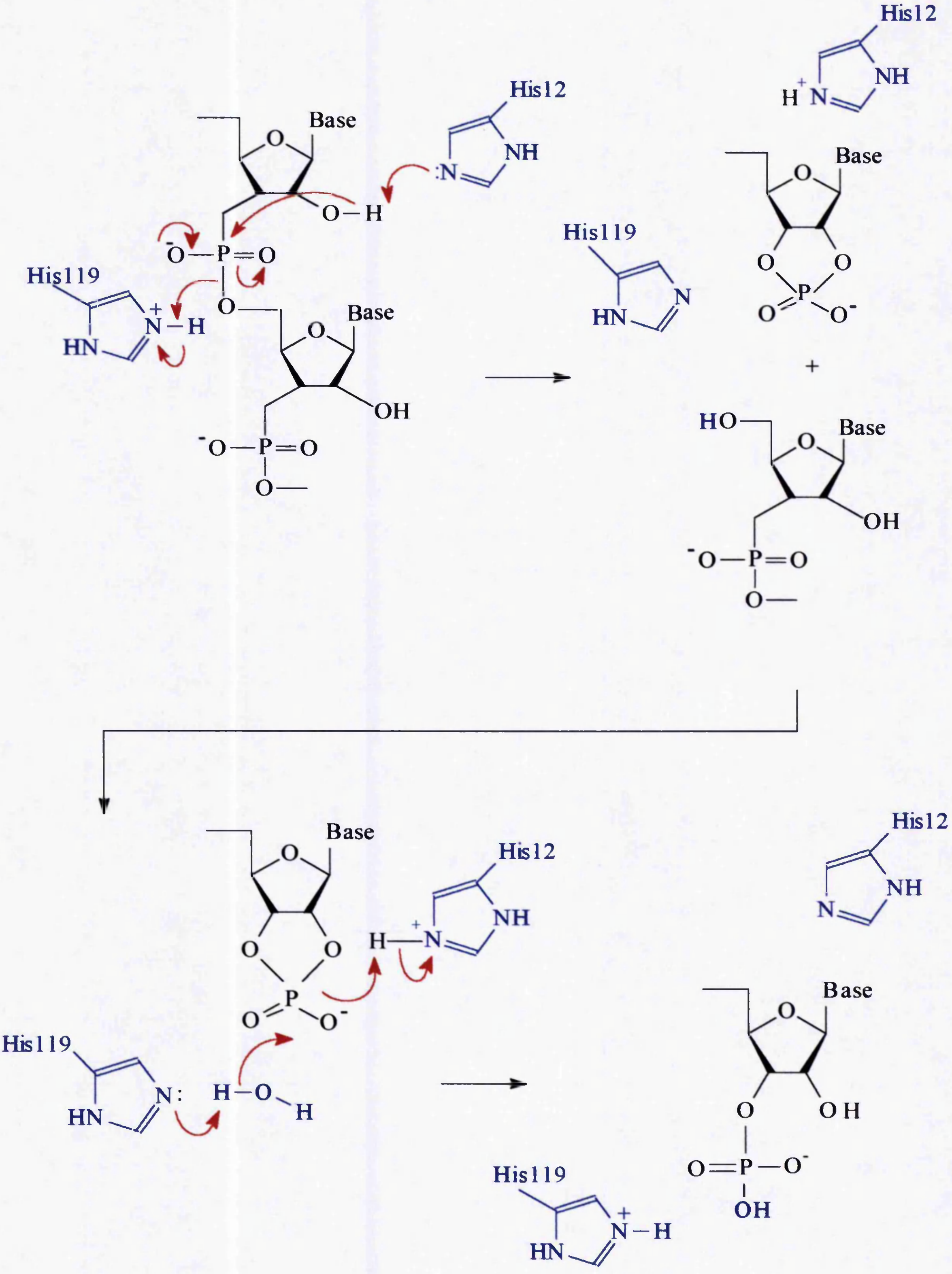


Figure 2.2. Mechanism of action of ribonuclease. The two imidazole rings of His12 and His119 act as the base and acid catalysts, respectively (see text) (adapted [62, 72]).

conjugating imidazoles to a polycationic spermidine of variable length and flexibility [75,76]. RNA cleavage has also been induced by a *bis*-alkylguanidinium receptor [77].

Specificity has been achieved by attaching an imidazole ring to a strand complementary to the target nuclei acid with the second imidazole being provided by the buffer [75,76], or having a diimidazole construct with both imidazoles on the complementary strand [76,78,79] (Figure 2.3). These modified oligonucleotides have been used to achieve specific cleavage of yeast $\mathrm{tRNA}^{\mathrm{Phe}}$ [80] and of a *Leishmania* mini-exon sequence [78]. Cationic conjugates bearing imidazole residues have also recently been described which cleave $^{\mathrm{t}}$RNA under physiological conditions [81] .

By mimicking RNase A, the problem of modifying one specific biopolymer is overcome, but the challenge of targeting a long nucleic acid sequence still remains. An obvious answer is to produce a system combining both suggestions above, *i.e.* a binary system consisting of two oligonucleotides complementary to the target sequence, each bearing an imidazole construct. This has been achieved by V. Vlassov *et al.,* using many different linker groups (personal communication). One of the most efficient is shown in Figure 2.4. The 12-mer target sequence is the universal 5' sequence of mRNA of the pathogenic parasite *Leishmania amazonensis*. The two complementary 6-mers are equipped with alkylimidazole groups which form the cleaving system by juxtaposition of the components when they are simultaneously bound at adjacent sites of the target. Specific cleavage is observed opposite the site of alkylimidazole attachment on the 12-mer target strand.

2.1.3 Aims

Insight into the mechanism of this so-called binary system requires the investigation of the mutual spatial arrangement of the components in the system. Our laboratory has already described the solution structures derived from NMR spectroscopic data of the photoactivatable pyrene-arylazide binary system [60]. However, there is no high-resolution structural information available for any binary system with cleaving potential, such as the non-aromatic alkylimidazole constructs. Therefore, in this thesis a solution structural study was undertaken by high-resolution NMR spectroscopy of a model binary cleaving system.

In the model system the target RNA (and therefore the complementary 6-mers) had to be replaced by DNA analogues to prevent undesirable cleavage and permit time-demanding two-dimensional NMR experiments. The purpose of using such a model system was to assess the potential perturbing or other effects of the alkylimidazole constructs on the nucleic acid structure

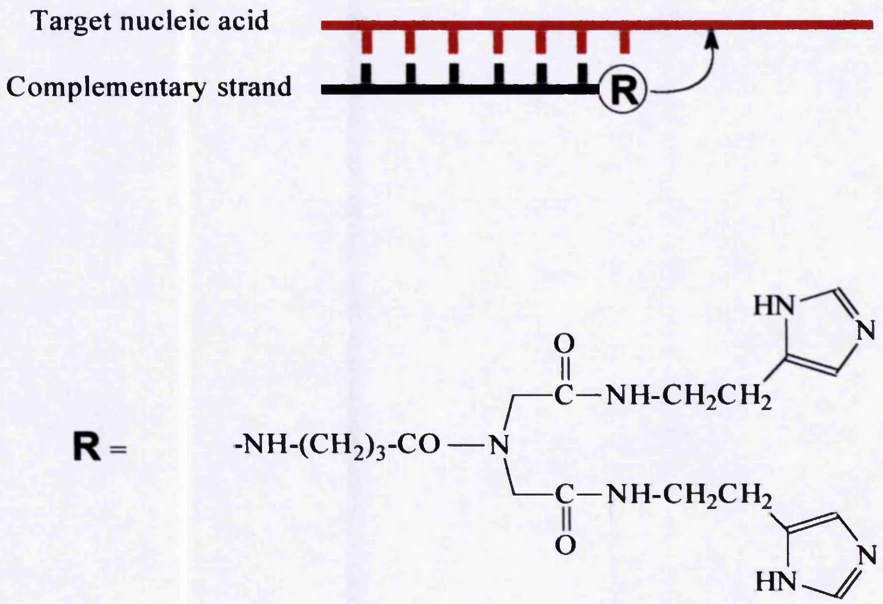
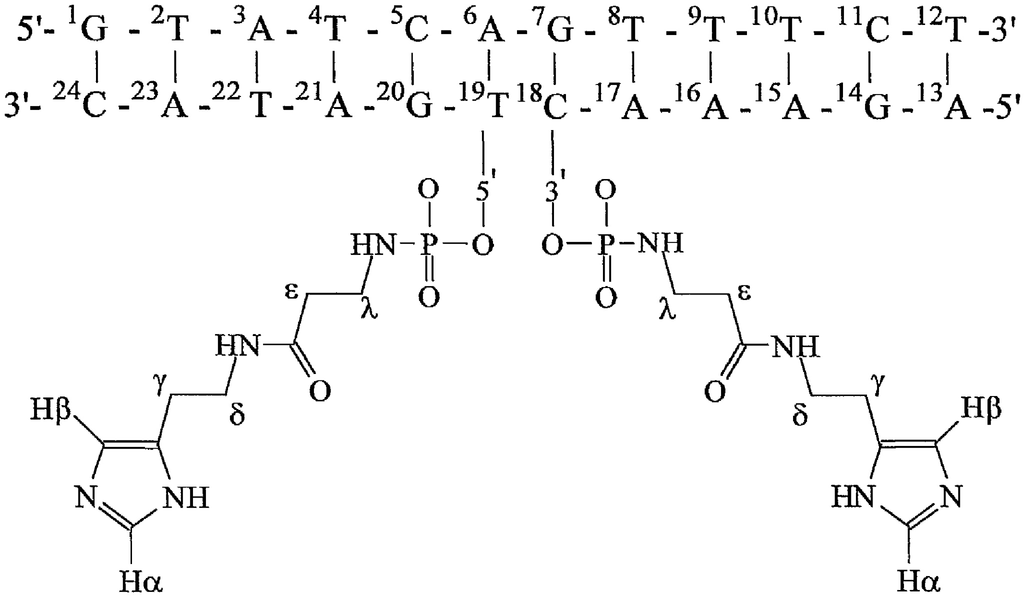


Figure 2.3. Sequence-specific cleavage by a diimidazole construct mimicking the cleaving activity of ribonuclease A.


5'- 1G - 2T - 3A - 4T - 5C - 6A - 7G - 8T - 9T - 10T - 11C - 12T -3'
3'- 24C - 23A - 22T - 21A - 20G - 19T 18C - 17A - 16A - 15A - 14G - 13A -5'
5'
3'
O
HN-P-O
O-P-NH
O
ε
λ
HN
O
NH
γ
δ
Hβ
N
NH
Hα
HN
N

d ^{1}G -^{2}T - ^{3}A - ^{4}T - ^{5}C - ^{6}A - ^{7}G - ^{8}T - ^{9}T - ^{10}T - ^{11}C - ^{12}T **(1)**

3'- d ^{13}A -^{14}G -^{15}A -^{16}A -^{17}A -^{18}Cp - **Im** **(2)**

Im' - 5'-pd ^{19}T - ^{20}G - ^{21}A - ^{22}T - ^{23}A - ^{24}C **(3)**

Figure 2.4. The binary system **1:2:3**. The alkylimidazole groups on ^{18}C and ^{19}T are represented as **Im** and **Im'**, respectively.

at the interface of the two short strands. The structural study of the binary photo-activatable system [60] showed considerable structural distortion at this point and it was important to establish the structure for the non-aromatic alkylimidazole constructs.

2.1.4. Applications of Nuclear Magnetic Resonance (NMR) spectroscopy in biological systems.

The specific interactions of biomacromolecules within themselves and with solvents, substrates and other solutes determine their biological functions in living systems. We can now systematically alter protein structure to assess structure-function relationships and identify and quantify molecular factors conferring specificity to substrate binding and important active site residues. This information is essential, for example, for the rational molecular design of proteins and the development of specific enzyme inhibitors for use in the treatment of metabolic disorders.

Although other analytical methods are available to study biopolymer structure and behaviour, NMR spectroscopy is a very powerful and versatile tool to address these problems at the molecular level. The NMR method is complemented frequently by computational molecular modelling and studies of crystal structures. In contrast to the latter methods of X-ray and neutron diffraction, however, NMR allows the investigation of biopolymer structures in solution, thus more closely simulating the *in vivo* biological system. The solution conformation and dynamics of macromolecules, which are closely linked to their biological functions, can thus be examined as a function of solvent, pH, temperature, ligand/substrate concentration and ionic strength and can provide direct, quantitative measurements of the frequencies of motional processes. NMR has diverse applications in fundamental biochemical investigations, from determining the structure of complex carbohydrates, proteins, peptides and nucleic acids, to magnetic resonance imaging and *in vivo* studies of plants and animals (reviewed [82]). Using NMR spectroscopy it is possible to follow the path of the polypeptide backbone of small proteins up to 10-12 kDa through the molecule. With ^{15}N labels, higher molecular weight systems can be studied. The combination of two-dimensional (2-D) NMR techniques with genetically engineered proteins provides one of the most powerful approaches to understanding the principles of protein folding, protein structure, protein-ligand interactions and enzyme catalysis. The conformation and dynamic behaviour of nucleic acids can also be studied and strategies have been developed for the assignment of resonances in both small proteins and oligonucleotides. NMR spectroscopy can help address many questions concerning the functions of nucleic acids which involve the effects of sequence on structure and how specific sequences of DNA are recognised by regulatory proteins. With modern spectrometers it is possible to assign duplexes containing up to 26-28 unique nucleotides (e.g. [83,84]). RNA structures have also been elucidated (e.g. influenza virus panhandle RNA of 34 nucleotides [85]). More complex *in vivo* studies of cells in suspension and of organs and tissues are possible. NMR signals are derived from the more abundant small molecule metabolites, e.g. phosphates, and determination of the relative levels of these provide information on normal and abnormal states of tissues and organs. Magnetic resonance imaging (MRI) has developed into a valuable diagnostic tool in hospitals allowing soft tissues in normal and diseased states to be studies.

NMR hardware, computer technology and experimental design have developed to produce powerful NMR spectrometers capable of probing complex biopolymer structures. Because of its application to a wide variety of systems and because of its increasing usefulness in studying a wide array of problems, NMR has great future potential in the investigations of biological systems.

2.1.5. NMR techniques used in the structure determination of biopolymers

A standard one-dimensional (1-D) NMR experiment provides information on the chemical shifts and the spin-spin coupling fine structures of individual resonances in a spectrum. To obtain additional data on through-bond or through-space connectivities between individual spins, double or multiple irradiation experiments (by selective irradiation of a particular resonance line) must be used. For work with the complex, crowded spectra of biopolymers, the use of 1-D double irradiation experiments is naturally limited. With 2-D NMR techniques, these limitations can be largely overcome. The 2-D NMR experiment, first described by the Belgian physicist J. Jeener in 1971, is now routinely used in the structural characterisation of biopolymers. In 1-D NMR experiments, the free induction decay (FID) is recorded immediately after the pulse during the detection period, t_2. If, however, the signal is not recorded immediately after the pulse but a time interval, t_1, allowed to elapse before detection, during this time interval (the evolution period) the nuclei can be made to interact with each other in various ways, depending on the pulse sequences applied. Data are collected in two different time domains: acquisition of the FID (t_2), and the t_1 delay which is successively incremented within a series of pulse cycles. The resulting signals (FIDs) thus depend on the two time variables (t_1 and t_2). For each value of t_1 a FID recorded during t_2 is stored, producing a data matrix, presented as a 2-D map in which both frequency axes describe the chemical shifts, the cross-peaks indicating which nuclei coupled. The time sequence of a typical 2-D NMR experiment (Figure 2.5(a)) includes three or four successive time periods. The preparation period usually consists of a long delay time during which thermal equilibrium is attained. After the first 90^o_x pulse and during the evolution period t_1, the nuclei may be subjected to other neighbouring spins and the spins precess under the influences of both chemical shift and spin-spin coupling. The t_1 period may be followed by a mixing period, τ_m, as in Nuclear Overhauser Enhancement Spectroscopy (NOESY) during which the magnetisation is distributed among the various spin states of the coupled nuclei. Finally, a second evolution (detection) period, t_2, follows where magnetisation is detected. There are two main types of 2-D NMR experiments: shift correlation through bonds (J-correlation, e.g. COSY) and shift correlation through space (e.g. NOESY). The following section briefly describes these methods and a few important variants.

2.1.5.1. Shift-*CO*rrelation *S*pectroscop*Y* (COSY)

Shift correlation spectroscopy (COSY) produces correlation maps (spectra) that display the connectivity of nuclei by scalar spin-spin coupling and thus provides information on proximity of nuclei along chemical bonds. Both frequency axes contain chemical shifts and cross-peaks indicate which nuclei are spin-spin coupled *i.e.* joined by a bond. The COSY pulse sequence contains only two 90^o_x pulses separated by the evolution time t_1 (Figure 2.5(b)). The first pulse produces transverse magnetisation (M_z) and the protons acquire phases that differ according to the differences in their respective chemical shifts and *J*-coupling interactions. The second 90^o_x pulse causes the mixing of spin states within a spin system and moves the *y*-component of M_z into the negative *z*-direction while the *x*-part remains in the *x,y* plane for detection. The intensity of the detected signal depends on the orientation of the vector M_z at the end of the evolution time, which is determined by the Larmor precession frequency.

In the presence of spin-spin coupling the second 90^o pulse not only affects transverse magnetisation, but also leads to population changes for the various transitions in the spin system and causes the magnetisation arising from one proton transition during t_1 to be distributed among all other transitions associated with it. Fourier transform of this data yields a spectrum with two frequency axes. The 'off-diagonal' peaks represent the spin-spin coupling between the protons and peaks on the diagonal represent unperturbed magnetisation.

COSY is an essential tool in helping to ascertain which atoms are physically joined. Protons separated by up 4 or 5 bonds can give rise to cross-peaks in the spectrum. The COSY

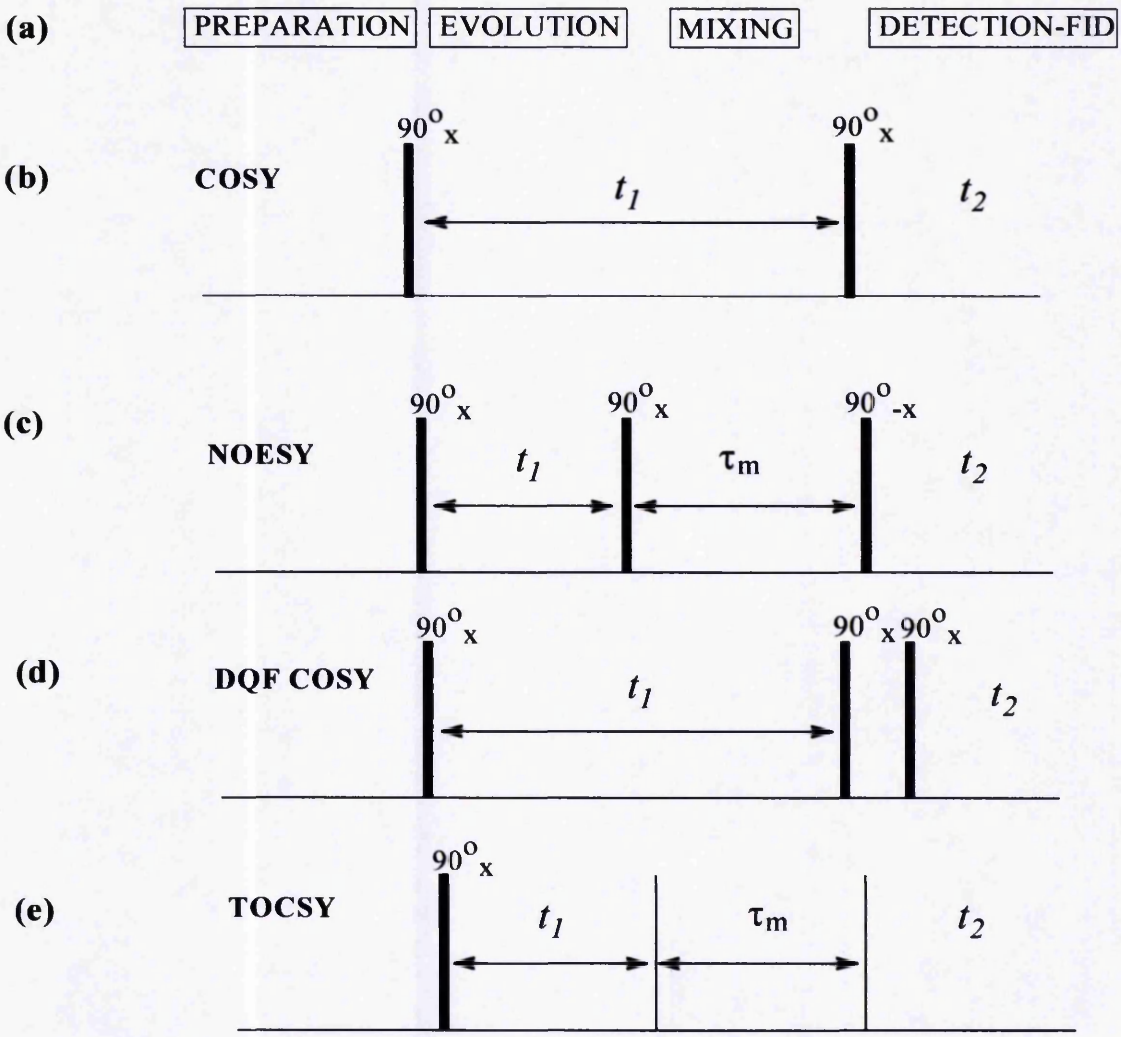


Figure 2.5. General pulse sequence schemes for some 2-D NMR experiments (adapted [86]). All consist of three or four successive time periods. Thermal equilibrium is attained in the preparation period. A 90° pulse follows and during the evolution period, t_1, the nuclei may be subjected to other neighbouring spins. A mixing period, τ_m, allows the magnetisation to be distributed among the various spin states of the coupled nuclei. Magnetisation is detected in the final period, t_2.

pulse sequence has been modified to yield different information, e.g. double-quantum filtered (DQF) COSY to eliminate singlet signals, identify spin systems and determine *J*-couplings, COSY-45° for the reduction of diagonal signals, delayed or long-range COSY (COSY-LR) to emphasize small couplings and exchange (E) COSY to only detect connected transitions and thus accurately measure *J* values. These methods are described in standard NMR texts. Only DQF COSY is outlined below.

2.1.5.2. Quantum transitions and phase coherence

Transition between energy levels of a spin system requires a change in the spin quantum number. Quantum mechanical selection rules allow single-quantum transitions, *i.e.* a change in the spin quantum number of +/-1. Double-quantum and zero-quantum transitions are not allowed, but their involvement may be considered in relaxation processes. When a 90°_x pulse is applied to a sample at equilibrium, the net longitudinal magnetisation (M_z) vanishes (*i.e.* the population difference between the α and β spin states) and transverse magnetisation is created in the *x,y* plane. A phase coherence is now said to exist between the α and β states of the nucleus, since they precess coherently with the same phase. The coherence is termed single-quantum coherence because the states are separated by a quantum number difference of 1. This causes a precessing net magnetisation of the nucleus, which can be detected in the form of a signal.

If a sequence of pulses creates two states, $\alpha\alpha$ and $\beta\beta$, which are phase coherent, double-quantum coherence is said to have been created. Similarly, in more complex spin states, coherence can be created between states differing by higher quantum numbers. The quantum mechanical selection rule stipulates that such multiple-quantum coherences do not give rise to detectable magnetisation, detectable signals resulting only from single-quantum transitions, and so for signals to be detected by the receiver, the coherence transfer must end with single-quantum coherence. Quantum transitions must therefore be converted into single-quantum coherence before detection by appropriate phase cycling procedures.

2.1.5.3. Double-Quantum Filtered COSY (DQF COSY)

One problem associated with COSY spectra is the dispersive character of the diagonal peaks, which can obliterate the cross-peaks lying near the diagonal. Moreover, if the multiplets are incompletely resolved in the cross-peaks, then because of their alternating phases an overlap can weaken their intensity or even cause them to disappear. In DQF COSY spectra, both diagonal and cross-peaks possess antiphase character, so they can be phased simultaneously to produce pure 2-D absorption line shapes in both. In practice this is done by applying a third 90°_{x} pulse immediately after the second (mixing) pulse (Figure 2.5(c)). This third pulse converts the double-quantum coherence generated by the second pulse into detectable single-quantum coherence. Only signals due to double-quantum coherence are inverted by the third pulse. If the receiver phase is correspondingly inverted only signals due to double-quantum coherence will be detected. Since all other coherences are filtered out, only the desirable double-quantum coherences will be modulated as a function of t_1 and yield cross-peaks in the 2-D spectrum. This significantly simplifies the COSY spectrum. Solvent signals which cannot generate to double-quantum coherence and signals due to triple-quantum coherence (as in a three-spin system) do not appear. Direct connectivites are represented by pairs of signal situated symmetrically on the two sides of the diagonal; remotely connected or magnetically equivalent protons will appear as lone multiplets.

Since the DQF COSY technique produces a simplified spectrum of direct connectivites, *J*-couplings can be determined, and thus torsion angles and the stereospecific assignment of protons, e.g. in sugar rings.

2.1.5.4. *TO*tal *C*orrelation *S*pectroscop*Y* (TOCSY)

Total correlation spectroscopy (TOCSY) is a powerful tool in structure elucidation since it allows scalar couplings to be observed over longer distances than COSY. The pulse sequence is shown in Figure 2.5(d). Only one 90°_{x} pulse is applied, followed by a mixing time before detection. During the mixing time, isotropic mixing occurs and the interactions with the external magnetic field and thus the chemical shifts, are practically eliminated and scalar coupling between nuclei dominates. Magnetisation transfer proceeds beyond the adjoining, directly coupled nuclei, is relayed to remoter nuclei and finally progresses through the complete scalar coupled network of nuclei in the molecule. Cross-peaks in the final 2-D spectrum are therefore seen from protons separated by more than three bonds. Short (20-50ms) mixing times yield primarily cross-peaks of strongly coupled protons, e.g. vicinal and geminal protons and a COSY-like spectrum is obtained. Longer times, (100-300ms), allow magnetisation transfer to remote protons of the spin system. (The physical basis for the Homonuclear Hartmann-Hahn experiment (HOHAHA) is quite similar).

In practice it is convenient to record spectra at mixing times of e.g. 20, 40, 60 and 100ms. Slices are then taken at specific chemical shifts at the various time intervals. This yields highly informative plots of the spread of magnetisation from key points in the molecule to adjoining regions in the same spin system, thus providing a powerful tool for structure elucidation. It is possible to divide the NMR spectrum into many sub-spectra, each corresponding to the NMR spectrum of a separate spin system, *i.e.* molecular fragment.

2-D spectra require a considerable amount of measuring time so 1-D variants are attractive alternatives. Methods with selective excitation are the most useful and sometimes are sufficient to complement the information from other sources. The selective 1-D TOCSY experiment has a large potential for applications in spectral analysis of oligosaccharides and oligonucleotides where, for example, in a complicated spectrum the signal of a single proton is observed separately and can be used as a starting point of the magnetisation transfer process.

2.1.5.5. Nuclear Overhauser Effect SpectroscopY (NOESY)

The NOESY experiment is one of the most important techniques available in biological spectroscopy, since under the correct conditions a complete set of short-range (<5Å) through-space connectivities can be obtained for a macromolecule. This is therefore the primary method for solving a complete solution structure.

Whereas COSY spectroscopy is based on shift correlation through bonds, in nuclear Overhauser spectroscopy (NOESY) cross-peaks are produced from protons interacting in space. In the three-pulse sequence for the NOESY experiment (Figure 2.5(e)), the first 90°_{x} pulse bends the longitudinal *z*-magnetisation to the *y*-axis. In the subsequent t_1 time period this magnetisation precesses in the *x,y* plane. The second 90°_{x} pulse rotates the magnetisation to the *x,z* plane. During the subsequent mixing period τ_m, *z*-magnetisation components exchange at a rate determined by cross-relaxation between the spins. The exchange of magnetisation during the mixing time is based on the nuclear Overhauser effect (*i.e.* the variation in intensity of the resonance signal of one nucleus when another nucleus lying spatially close to it is irradiated, with the two nuclei relaxing each other *via* a direct through-space coupling interaction). Since this process depends in part upon the distance between the nuclei, the cross-peaks correlate resonances which are through-space coupled and the exchanges will appear as cross-peaks on either side of the diagonal in the final spectrum. Finally, a third 90°_{-x} pulse regenerates detectable magnetisation in the *x,y* plane where each vector precesses with its characteristic Larmor frequency. The diagonal peaks which appear in the final spectrum are generated by magnetisation which fails to migrate during τ_m.

As with COSY, other techniques exist which are based on NOE techniques, e.g. ROESY which separates chemical and NOE information, MINSEY (NOESY with suppression of spin-diffusion pathways).

Multiple conformations

~~Unlike crystallography~~, NMR gives us information on multiple molecular conformations that may exist, thus, structural interpretation is not misleading by being biased towards one conformational state. Sometimes, more than one member of a conformational family can give rise to cross-peaks and their conformations can be established. In other cases, the flexibility of multiple conformations causes line broadening and no cross-peaks are found. Information on flexibility can be obtained by observing relaxation rates.

Two macroscopic magnetisations are distinguished in an NMR experiment, the longitudinal magnetisation along the *z*-axis and the transverse magnetisation along the *x,y* plane. Both are subject to relaxation phenomena, *i.e.* their magnitudes are time-dependent.

2.1.5.6. Longitudinal relaxation

A system at equilibrium has a population difference between the different spin-states. Immediately after exposing the spins of a sample to an external magnetic field, they exist in a non-equilibrium state. The build-up of the initial equilibrium magnetisation M_0, *i.e.* to re-establish the original equilibrium state between the populations of the ground and excited states, requires a time T_1. During T_1, energy is transferred from the spins to the environment, the so-called lattice, by a process called longitudinal or spin-lattice relaxation. T_1 is the longitudinal or spin-lattice relaxation time. The variation of the *z*-component of the macroscopic magnetisation obeys a first order differential equation (Equation (2.1)):

$$\mathrm{d}M_z / \mathrm{d}t = (M_0 - M_z) / T_1 \qquad \text{Equation (2.1)}$$

Longitudinal relaxation is the return of longitudinal (*z*) magnetisation of the perturbed state to its original state $M_z(0)$. $1/T_1$ is thus the rate constant for this transition. For such relaxation to occur, the nuclei must be exposed to local oscillating magnetic fields some of whose frequencies can exactly match their respective precessional frequencies, e.g. in the magnetic moments of other protons present in the same tumbling molecule. Since the molecule is undergoing various translations, rotations and internal motions, there is virtually a continuum of energy levels available and energy exchange can occur readily between the nucleus and the lattice through such dipole-dipole interactions.

Short relaxation times broaden resonance lines because the lifetime of nuclei in the excited state is decreased which causes an uncertainty in the determination of the energy difference. According to the uncertainty principle $\Delta E\Delta t \approx h$ and with $\Delta E = h\Delta\nu$ this leads to $\Delta\nu\Delta t \approx 1/2\pi$ or $\Delta\nu=1/2\pi\Delta t$ for the uncertainty in the determination of the resonance energy. The line-width is therefore proportional to $1/\Delta t$ or $1/T_1$. In organic liquids T_1 for protons is generally in the order of a few seconds or less so that spin-lattice relaxation contributes not more than 0.1Hz to the line-width, but observed line-widths are larger and may amount to several kHz in the case of solids.

It is important to determine T_1 values so that pulses can be applied for exactly the correct length of time to flip the magnetisation by the required angle. A slight deviation from the required pulse angle can adversely effect the outcome of the experiment so it is often advisable to calibrate the pulse width before the start of the experiment.

2.1.5.6.1. T_1 measurements

T_1 values for individual nuclei are significant parameters related to the dynamic properties of molecules. From the various methods available to determine T_1, the most often used is the inversion recovery experiment (Figure 2.6(a)). An initial 180° pulse brings the macroscopic magnetisation **M** into the negative *z*-direction **(b)**. As a result of spin-lattice relaxation the value of M decreases **(c)**, passes through zero **(d)**, begins to increase in the positive z-direction **(e)** and finally gains its initial value. The magnetisation can be detected by 90° pulses at τ_1 and τ_2 which align **M** along the negative or positive *y*-direction respectively.

To convert the longitudinal magnetisation into a signal, it must be brought into the *x,y* plane by the application of a 90° pulse. If this pulse is applied very soon after the original 180° pulse, it will 'catch' the magnetisation while it is still on the –*z*-axis and cause it to rotate to the –y-axis, thereby giving a negative signal. As the evolution period, τ, is increased progressively, the 90^{o}_{x} pulse will 'encounter' the z-magnetisation as it recedes along the –z-axis to a zero value. A series of spectra is produced in which the signals have decreasing negative amplitudes and then increasingly positive amplitudes.

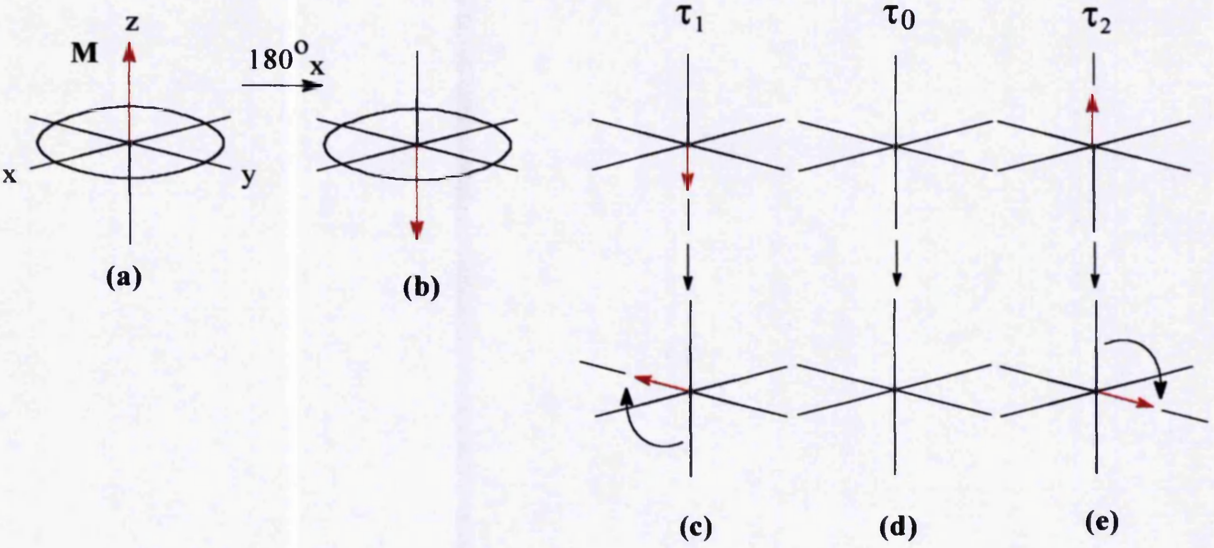


Figure 2.6(a). The principle of the inversion-recovery experiment for measurement of spin-lattice or longitudinal relaxation time, T_1 (adapted [87]).

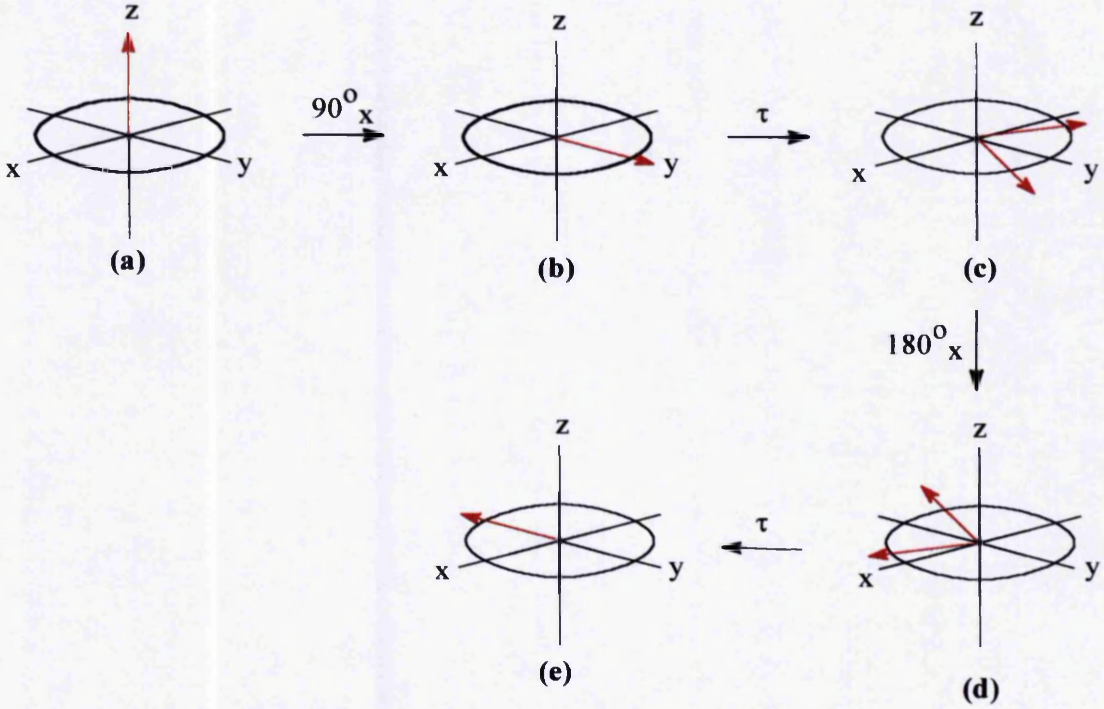


Figure 2.6(b). The spin-echo experiment for measurement of spin-spin or transverse relaxation time, T_2 (adapted [87]).

The two signals differ in phase by 180° and thus lead to an emission and absorption line respectively. At time τ_0 no signal can be observed since the sample is not magnetised. For this situation T_1 can be determined from the relation $\tau_0 = T_1 \ln 2$.

2.1.5.7. Transverse relaxation

In addition to *z*-magnetisation there is a second magnetisation in the *x,y* plane usually termed transverse or *x,y* magnetisation ($M_{x,y}$). The time-dependence of $M_{x,y}$ usually differs to that observed for M_z and reaches an equilibrium magnetisation with a time T_2. T_2 is called transverse or spin-spin relaxation time since the relaxation mechanism is based on an energy transfer within the spin system. Any transition of a nucleus between its spin states changes the local field at nuclei nearby at the correct frequency to stimulate a transition in the opposite direction. The lifetime of the spin states will be shortened by this process and therefore contribute to the NMR line width in a manner similar to the spin-lattice relaxation process.

In the simplest case $T_1 = T_2$ for liquids since, after resonance, the *x,y* component of the magnetisation vanishes at the same rate as the longitudinal magnetisation attains its previous value M_0 along the *z*-axis. However, the transverse magnetisation can be reduced without the simultaneous increase in the *z*-component ($T_1 < T_2$). As in the case of spin-lattice relaxation, fluctuating fields can interact with the transverse component $M_{x,y}$, thereby reducing its magnitude. Magnetisation can also be readily lost due to field inhomogeneity so T_2 is greatly shortened by the existence of multiple conformations. Line half-width (in hertz) is related to T_2 (in s) by Equation (2.2):

$$\Delta = \frac{1}{\pi T_2}$$

$$\Delta = \frac{1}{\pi T_2^*}$$

Equation (2.2)

where Δ is the resonance signal at half height. Since the decay of $M_{x,y}$ is caused by field inhomogeneity and natural spin-spin relaxation as well, one usually writes:

$$\frac{1}{T_2^*} = \frac{\gamma \Delta B_0}{2\pi} + \frac{1}{T_2}$$

Equation (2.3)

Where the first term is the inhomogeneity contribution to the line width. If T_2 is reduced, (e.g. by the existence of multiple conformations and exchange processes), line-width increases. Calculating T_2 from a T_2 measurement experiment and comparing it to T_2 which is calculated from observed line-widths, can indicate that multiple conformations may exist if T_2(measured) > T_2(expt).

2.1.5.7.1. T_2 measurements: the spin-echo experiment

The spin echo experiment is designed to measure T_2 relaxation times. In this pulse sequence (Figure 2.6(b)) a 90°_x pulse rotates the vector to lie along the *y*-axis **(b)**. As a result of the inhomogeneity of the external magnetic field B_0 the individual nuclear spins begin to fan out and the magnitude of the transverse magnetisation decreases **(c)**. After a certain time (τ) a 180° pulse is applied so that all vectors are turned around into the negative *y*-direction **(d)**. Now, however, their relative motion follows a course such that after a time 2τ they become focussed in the negative *y*-direction. The resultant transverse magnetisation can now be detected in the receiver coil as a signal, the so-called spin echo.

The intensity of the spin echo should depend only on the transverse relaxation rate, *i.e.* the irreversible loss of transverse magnetisation during the period 2τ, since contributions of the field inhomogeneity to the fanning out process have been eliminated by the refocussing step. The echo amplitude should therefore be proportional to $\exp(-2\tau/T_2)$. In practice, diffusion processes complicate the situation by changing the positions of the spins in the magnetic field, thereby increasing the spread of the Larmor frequencies. This complicating factor can be eliminated if, instead of using a single 180° pulse at time τ, one uses a whole sequence of such pulses at τ, 3τ, 5τ, etc. (Carr-Purcell-Meiboom-Gill pulse sequence [88,89]. The decrease in the amplitude of the spin echo which in turn is recorded at 2τ, 4τ, 6τ, etc., is now proportional to $\exp(-\tau/T_2)$ and the effect of diffusion becomes negligible if the interval between the pulses is small.

For more detail of the NMR experiments described the reader can refer to any number of excellent texts available. Other techniques not discussed here are solvent suppression techniques and 3-D NMR experiments, e.g. TOCSY-NOESY and NOESY-NOESY which provide additional separation of resonances.

2.1.6. Biomolecular structural studies

Nucleic acid and protein structure studies are often of a different nature to each other. In general for the DNA helix, we are interested in fairly small structural changes which are sequence-dependent and consequently guide protein or drug recognition. These subtle variations demand detailed knowledge of the structure and accurate inter-nuclear distance determinations. Often only an approximate, less detailed structure is required for proteins. A protein tertiary structure can probably be defined with moderate accuracy using restrained molecular dynamics (RMD) calculations without accurately determining inter-proton distances. If (nearly) all proton resonances of a small protein are resolved and assigned, and if the protein has a high content of α-helix and β-sheet, these secondary structures can be fairly easily defined by the presence or absence of certain cross-peaks. The observation of relatively few 'long-range' cross-peaks in the primary sequence will serve to orientate the secondary structures relative to one another and generate a rough idea of the structure. RMD can then improve upon this. Structures of proteins with a low content of secondary structural elements are more difficult to predict and a larger number of accurate inter-proton distances is thus needed. More importantly in protein structure determination, the conformation of the active- or ligand-binding site needs to be determined with high resolution and can be achieved using the methods described above.

Systematic methods have been developed to obtain resonance assignments in ^{1}H NMR spectra of small proteins [90-94] and nucleic acids by a combination of 2-D NOE and 2-D *J*-correlated spectroscopy.

2.1.6.1. Assignment by isotope labelling

Through isotope labelling the NMR spectra can be modified without noticeably affecting molecular structure and function: isotope labelling of a particular monomeric unit affects its spin system so that it can be uniquely recognised. For example, by ^{15}N enrichment of an amide nitrogen of an amino acid residue, the resonance could be assigned in the ^{15}NMR spectrum from its outstanding intensity. The ^{1}H spin system could be identified from the doublet fine structure arising from scalar ^{1}H-^{15}N coupling. In the ^{13}C NMR spectrum, the adjoining carbon atoms could be identified from the scalar ^{15}N-^{13}C couplings. Proteins can be produced biosynthetically containing a selected few amino acid types in perdeuterated form (*i.e.* as many protons as possible replaced with deuterium). It is also possible to partially deuterate proteins to an extent of approximately 90% to observe the complete ^{1}H NMR spectrum of the residual 10% protons. The line broadening through ^{1}H-^{1}H dipole-dipole is reduced by the isotopic dilution of the protons in the macromolecular structure and the intrinsic loss in sensitivity is partly offset by the narrower lines.

Extensive NMR assignments by the incorporation of suitably isotopically enriched monomers is too laborious and expensive to be widely practical at present. However, sequential assignment procedures can be enhanced by a combination with labelling techniques. More detailed reviews on the sequential assignment of proteins exist (mentioned above). Here, a description of resonance assignments in nucleic acids (DNA) will be discussed as the NMR spectroscopic component of this thesis was in this field.

2.1.6.2. Proton assignment of DNA

2.1.6.2.1. Non-exchangeable protons: residue-specific assignment

The ability of the 2-D NMR methods described to assign the spectra of large DNA molecules is limited only by the spectral resolution of the off-diagonal cross-peaks. The bases and sugar rings of DNA are shown in Figure 2.7 with exchangeable and non-exchangeable protons labelled. The bases constitute four independent spin systems not *J*-coupled to the sugar protons and the deoxyribose rings constitute another one. Cytidine is the only base containing vicinal protons on adjacent carbons: the CH5 and CH6 protons should be split into doublets, whereas the remaining non-exchangeable aromatic protons (AH8, GH8, TH5 and AH2) should appear as singlets. From NMR spectra of the four mononucleosides of DNA, there are 6 major spectral regions in which the protons generally resonate. The shifts reflect the different chemical and electronic environments of the nucleotide protons.

1.	7-8ppm	AH8, GH8, AH2, CH6, TH6
2.	5.2-6.2ppm	CH5, H1'
3.	4.6-5.1ppm	H3'
4.	3.6-4.6ppm	H4', H5', H5"
5.	1.7-3.0ppm	H2', H2"
6.	1.0-1.7ppm	TCH_3

It is possible to grossly assign mono- or di-nucleotides from 1-D spectra alone.

2-D COSY (and TOCSY) techniques can detect cross-peaks between TH6 and TCH_3 protons, but given the absence of sufficiently large *J*-couplings between the sugar H1', H2' and H2" and the base H6 or H8 protons, the signals of a particular sugar and those of its

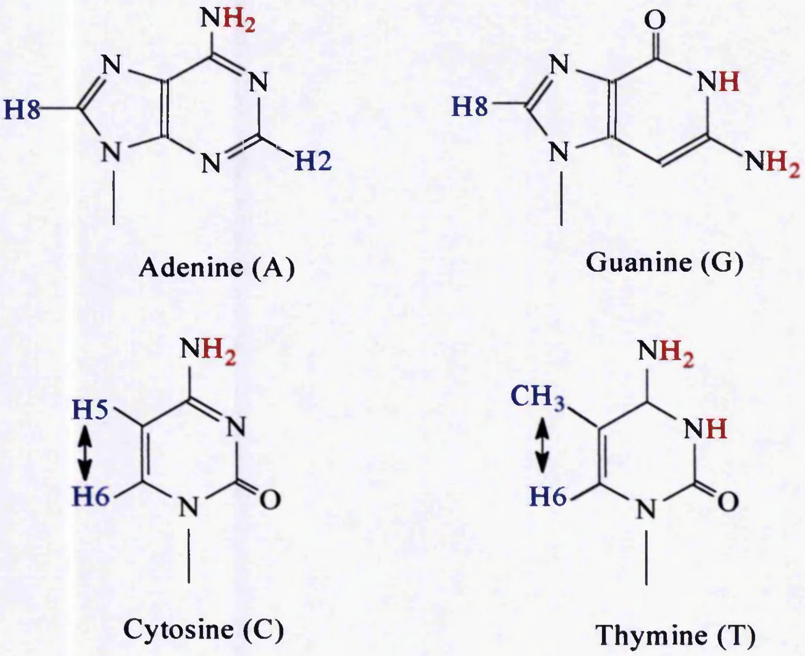


Figure 2.7(a). The four bases of DNA showing exchangeable (red) and non-exchangeable protons (blue) and connectivities between the latter.

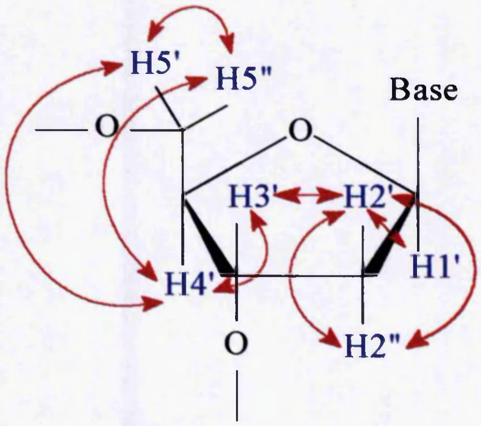


Figure 2.7(b). Intra-base connectivities.

corresponding base need to be connected *via* NOEs. These connections are made as part of the sequential assignment procedure now described.

2.1.6.2.2. Sequential assignment of DNA from NOESY spectra

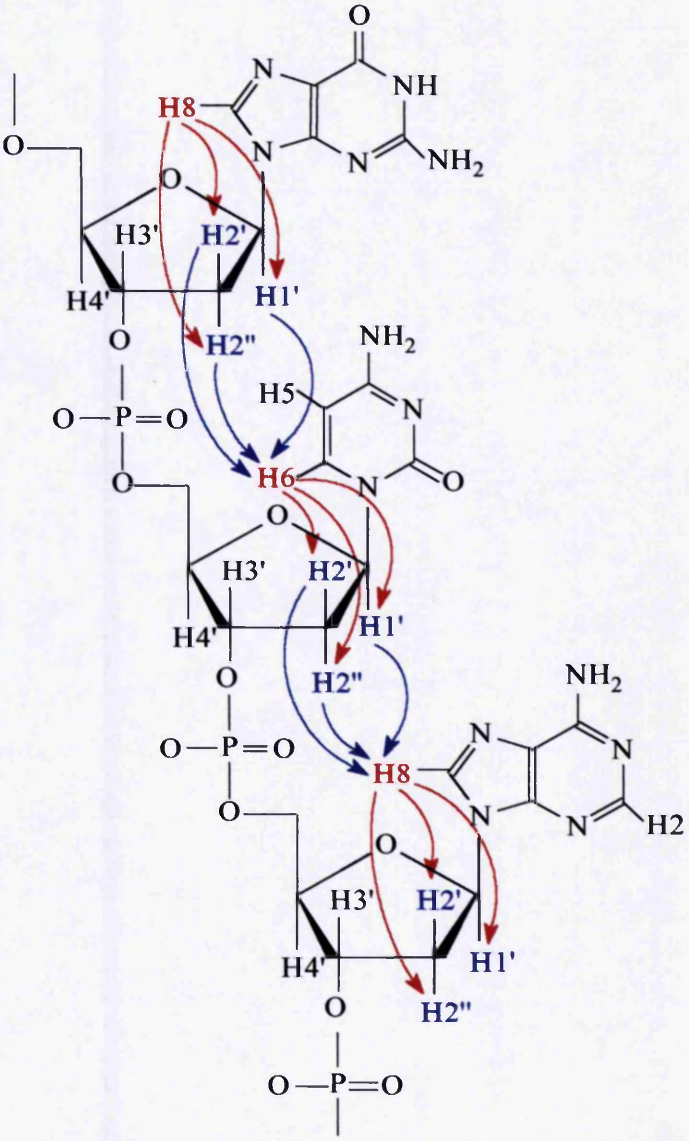
In 2-D spectra, two connectivity pathways are commonly used for assignment purposes, namely those formed by a series of cross-peaks between H6 or H8 and H1' resonances and pathways formed by cross-peaks between H6 or H8 and H2' or H2" resonances (Figure 2.8). As an example, Figure 2.9 shows a 2-D NOESY spectrum of a 12-mer duplex with three regions highlighted. **A** shows the region of cross-peaks that link the base proton (AH8, GH8, CH6, TH6) resonances with those of the CH5 and H1' protons. **B** shows the contacts between the same base protons and the TCH_3 and sugar H2', H2" protons. Region **C** indicates the corresponding H1'-H2', H2" (intra-nucleotide) cross-peaks.

Each of the purine H8 and pyrimidine H6 protons (except those at the 5' termini) is close enough to the H1', H2' and H2" protons on two sugar residues for cross-relaxation to occur, *i.e.* the sugar residue forming part of the same nucleotide as the base proton, and the sugar ring of the 5' neighbouring nucleotide. These internucleotide cross-peaks in the 2-D NOE spectrum make it possible to 'walk' from one base proton to the next in the sequence *via* its H1', H2' and H2" contacts. (See Figure 2.9 in regions A and B). The complete cross-relaxation networks in regions **A** and **B** can be checked for consistency in region **C** where the corresponding H1'- H2', H2" (intranucleotide) cross-peaks can be observed.

The most intense peaks, corresponding to the shortest proton-proton distances, can be readily identified as the intranucleotide CH6-CH5 or TH6-TCH_3 contacts which both appear in COSY and NOESY spectra. Each of the CH5 and TCH_3 resonances is linked to the purine H8 or pyrimidine H6 resonance of its 5' neighbour. These inter-base cross-peaks are found in region A. Thus, when a pyrimidine occurs in the sequence, an alternative cross-relaxation pathway exists, not involving sugar protons, but CH5 and TCH_3 protons as the linkage between purine H8 and /or pyrimidine H6 protons on adjacent nucleotides.

The systematic procedure requires a unique cross-peak at which to start the analysis supplied by the H6/H8 resonance of the 5' terminal residue which is only connected to the H1' of its own sugar and therefore exhibits only one cross-peak, in contrast to other residues in regular DNA. A cross-peak can then be found which connects the H1' of the terminal residue with the H6/H8 residue of the adjacent residue on the same strand.

5' - terminus



3' - terminus

Figure 2.8. Sequential assignment of DNA. Intra- and inter-base connectivities are shown in red and blue, respectively.

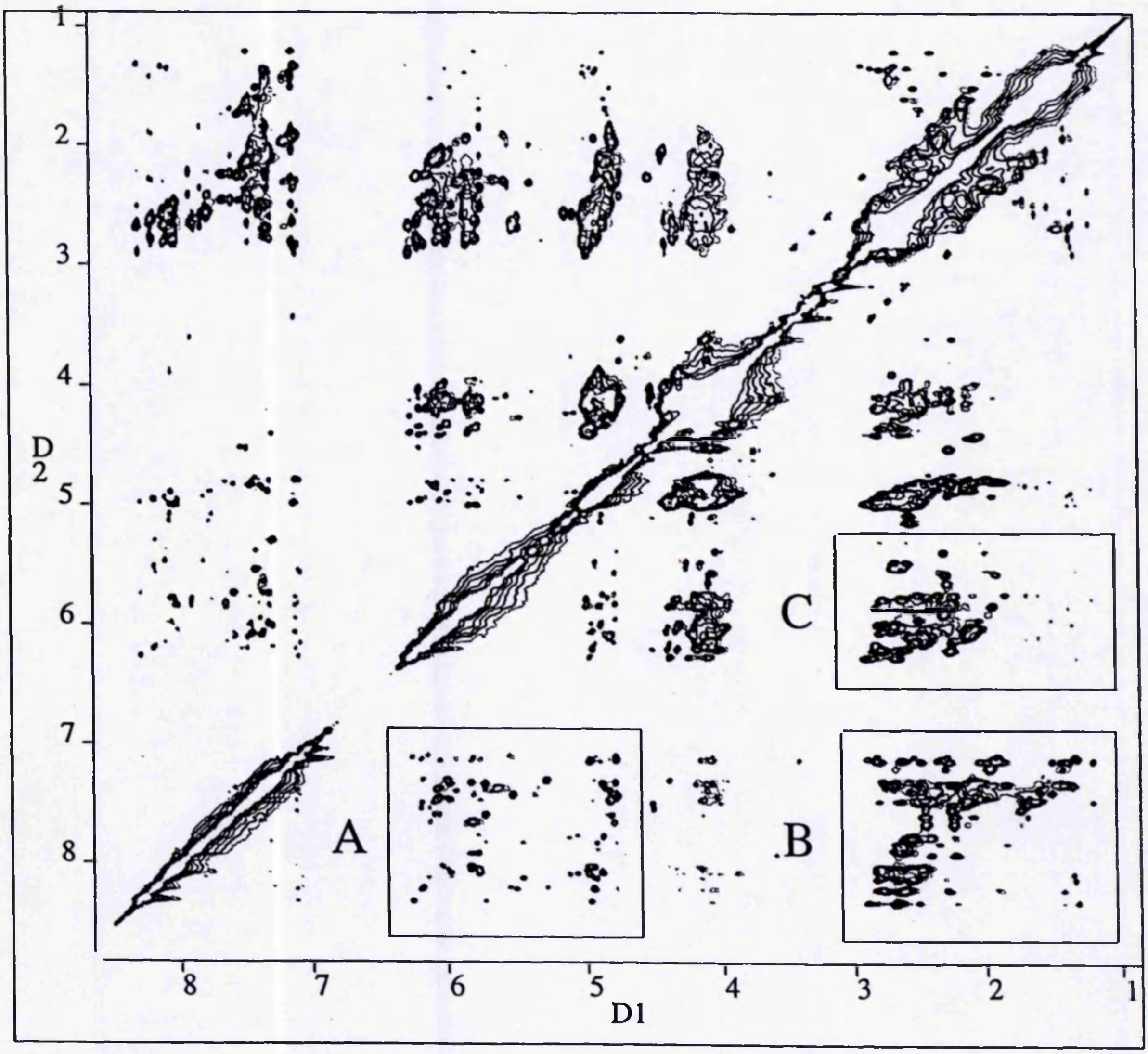


Figure 2.9. 2-D NOE spectrum of a 12-mer DNA duplex to illustrate the different regions of cross-peaks. **A** shows the region linking the base proton (AH8, GH8, CH6, TH6) resonances with those of the CH5 and H1' protons. **B** shows the cross-peaks between the same base protons and the TCH_3 and sugar H2', H2" protons. Region **C** indicates the corresponding H1'- H2', H2" (intra-nucleotide) cross-peaks. (D1 and D2 refer to the F2 and F1 frequency domains, respectively. Axes are labelled in ppm).

The complete systematic assignment of the remaining sugar resonances is impeded by strong overlap in the H4', H5' and H5" regions, even in a 2-D spectrum. The remaining aromatic resonances (A-H2) are part of another cross-relaxation network in double-stranded DNA, which also includes the exchangeable hydrogen-bridged imino protons. Systematic assignment of these resonances is also possible by using NOEs, see below [95,96].

The different forms of DNA, e.g. right-handed A-, B-, and left-handed Z-DNA with their different helix parameters (e.g. number of bases per turn, shape of major and minor groove, pitch, propellor twist) can be distinguished by NMR methods since they produce cross-peaks which have different intensities and networks of different connectivities.

2.1.6.2.3. Exchangeable protons

The amino and imino exchangeable protons of the four DNA bases are seen in Figure 2.7. The imino protons can be detected for DNA duplexes as the solvent is excluded from the core of the polymer and the protons therefore do not exchange with water. They can be observed using REFOPT and REFOPTNY pulse sequences. The REFOPTNY pulse sequence (Figure 2.10) replaces the third pulse of a conventional NOESY sequence with the REFOPT sequence [97]. REFOPT consists of a selective 90° pulse sequence followed by a selective 180° sequence. Both sequence elements contain equal numbers of 0° and 180° pulse-shifted pulses, with flip angles numerically optimised to give a null effect on and close to resonance, but approximately 90° and 180° rotations respectively for the remainder of the spectrum of interest. The overall effect is to give full excitation of the signals of interest, but efficient suppression of water signals. Signal phases are constant except for a 180° discontinuity at the transmitter frequency. The REFOPTNY pulse sequence is particularly useful on spectrometers not equipped with pulse field gradients.

The distances between the imino protons of adjacent base pairs in A- and B-type helices are typically ≤ 5Å so assignment can be achieved *via* a chain of connectivities providing a suitable starting point is available.

Connectivities between exchangeable and non-exchangeable protons

Exchangeable and non-exchangeable protons can sometimes be linked together to aid analysis. For example, the G-NH1 resonance can be connected to the C-H5 of its base pair *via* the amino protons of the cytosine base. Subsequently, through the connectivity of C-H5 with C-H6,

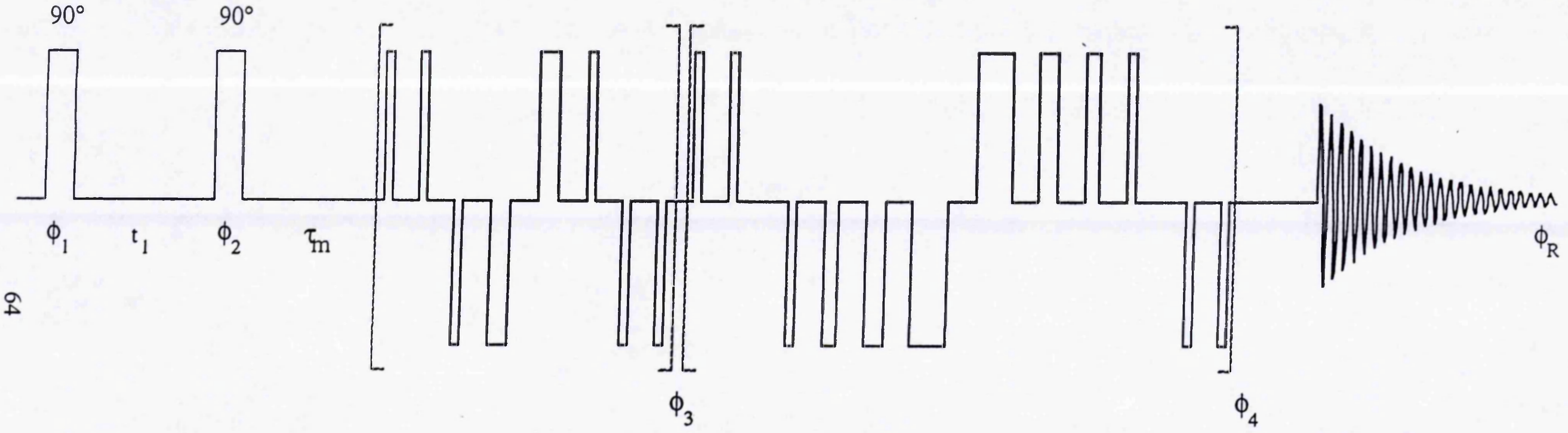


Figure 2.10. Pulse sequence for the REFOPT experiment. The initial 90^O pulse ϕ_1, evolution period t_1, second 90^O pulse ϕ_2, and mixing period τ_m, form the first part of a normal NOESY sequence, but the usual 90^O NOESY read pulse is replaced by a sequence consisting of an 8-pulse 90^O net flip angle segment ϕ_3, followed by a 12-pulse 180^O segment ϕ_4. The flip angles of the inddual pulses and the lengths of the delays between them are optimised to give flat excitation over most of the spectrum of interest, but zero excitation both at and close to the water signal.

G-NH1 is linked to the backbone proton signals. The A-H2 resonance can often be identified through NOE connectivity with the T(U)-CH_3 proton in the same base pair.

The procedure described yields the assignment of all G-H8, A-H8, T-H6, T-CH3, C-H6, C-H5, H1', H2', H2" resonances (as well as many H3' resonances indirectly) in both strands of the DNA duplex. This procedure can be used for A- and B-DNA.

2.1.6.2.4. Stereospecific assignment of H2', H2", H5' and H5"

The assignment of these resonances is very often based on their relative position in the spectrum but for irregular structures a more reliable procedure is necessary.

The H2' and H2" signals can be separately identified on the basis of the intensity of their NOEs to H1', provided spin diffusion effects are negligible or can be accounted for. The H1'-H2" cross-peaks will be more intense than the H1'-H2' cross-peaks because the interproton distance is smaller. This is basically true throughout the pseudorotation cycle of the sugar moiety.

The stereospecific assignment of the H5' and H5" resonances is intimately connected to determination of the torsion angle γ defined between atoms H4'-C4'-C5'-O5'. It can be achieved by combining measured $J_{4'5'}$ and $J_{4'5''}$ couplings with the intensities of NOE cross-peaks between the resonances of the proton pairs H3'-H5', H3'-H5", H4'-H5' and H4'-H5". If both coupling constants $J_{4'5'}$ and $J_{4'5''}$ are small, (*i.e.* ≈ 2-3Hz) and assuming the normal staggered conformations are populated, the distanceS H3'-H5" = 2.4Å and H3'-H5' = 3.8Å, so the NOE between H3' and H5" is more intense, thus identifying the H5" resonance. When the coupling constants are large, the most intense cross-peak belongs to the H3'-H5' pair.

2.1.7. Determination of solution structures from analyses of NOESY spectra

After assigning the resonances to nucleotide protons, the solution structure of the molecule must be solved. The use of distances obtained from 2-D NOE spectra as constraints in either molecular dynamics or distance geometry calculations has resulted in protein and DNA solution structures that are compatible with those obtained from X-ray crystallography.

2.1.7.1. Interproton distance determination

The effect of cross-relaxation between two neighbouring protons during the mixing time period τ_m of the 2-D NOE experiment is to transfer magnetisation between them. This results in cross-peak intensities in the spectrum that are approximately inversely proportional to the sixth power of the distance between the two neighbouring protons. Obtaining accurate distances from

2-D NOE intensities is complicated, however, by the occurrence of spin diffusion. The neighbouring protons belong to an array of all protons in the molecular structure (Figure 2.11), and the cross-relaxation between the two is part of a coupled relaxation network which should be considered as a whole. A cross-peak between correlated protons has an intensity which depends primarily on this network linking them and not simply on the interaction of the two.

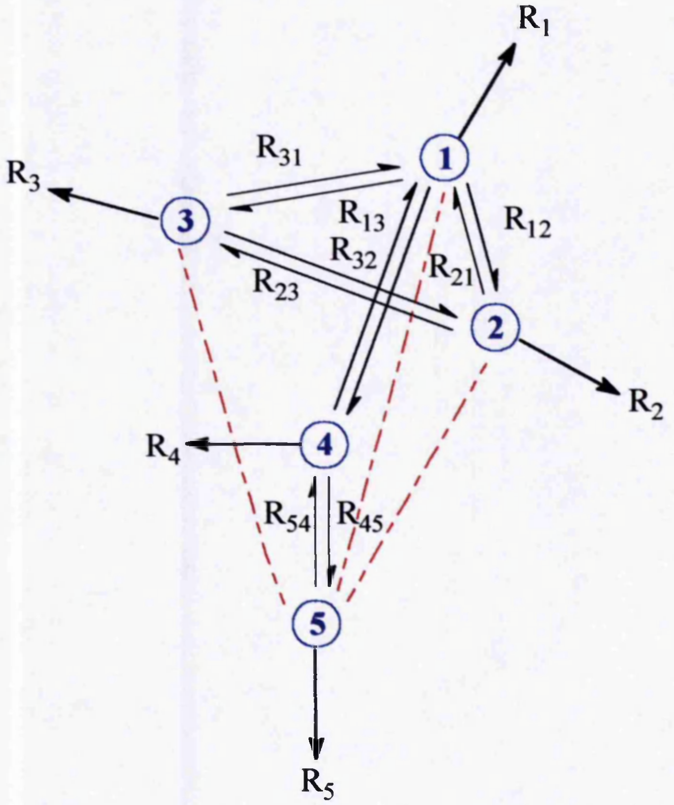


Figure 2.11. Network of five protons in a molecule illustrating cross-relaxation rate effects arising from dipole-dipole interactions (R_{ij}) and non-dipolar relaxation (R_i).

Cross-relaxation during the mixing time τ_m is described by the equation:

$$\partial M/\partial t = -RM \qquad \text{Equation (2.4)}$$

where M is the magnetisation vector describing the deviation from thermal equilibrium ($M = M_z - M_0$), and R is the matrix describing the complete dipole-dipole relaxation network. The diagonal elements of the rate matrix are the longitudinal relaxation rates (R_{ii}), while the off-diagonal elements are the cross-relaxation rates (R_{ij}) [98].

$$R_{ii} = 2q\frac{\tau_c}{r_{ii}^6}(n_i - 1)\left(\frac{3}{1+(\omega\tau_c)^2} + \frac{6}{1+4(\omega\tau_c)^2}\right)$$

$$+ q\tau_c \sum_{j\neq 1} \frac{n_j}{r_{ij}^6}\left(1 + \frac{3}{1+(\omega\tau_c)^2} + \frac{6}{1+4(\omega\tau_c)^2}\right)$$

Equation (2.5)

$$R_{ij} = q\frac{\tau_c}{r_{ij}^6}\left(\frac{6}{1+4(\omega\tau_c)^2} - 1\right)$$

Equation (2.6)

n_i is the number of equivalent spins in a group such as a methyl rotor

ω is the spectrometer frequency in radians

τ_c is the correlation time

$q = 0.1\gamma^4 (h/2\pi)^2$

Equations 2.4-2.6 have the solution:

$$M(\tau_m) = \mathbf{a}(\tau_m)M(0) = e^{-R\tau m}M(0) \qquad \text{Equation (2.7)}$$

where **a** is the matrix of mixing coefficients which are proportional to the 2-D NOE intensities. This matrix of mixing coefficients is what we wish to evaluate. The exponential dependence of the mixing coefficients on the cross-relaxation rates complicates the calculation of intensities (or the distances) so simplifications have been introduced in some of the methods of calculating of inter-proton distances. A brief description of these follows.

2.1.7.2. The isolated spin-pair approximation (ISPA)

The simplest method of obtaining distances from intensities is the isolated spin-pair approximation, (ISPA). This method ignores the effect of spin diffusion and the relaxation network of protons and assumes that a cross-peak is due to the interaction of just two protons, *i.e.* spin-spin relaxation of one proton occurs through only one other proton. Gronenborn and Clore have reviewed this approach [99]. With short experimental mixing times, a cross-peak intensity correlating a particular proton-proton pair depends on the distance separating that pair of protons alone.

Equation (2.6) can be recast into a series expansion:

$$a(\tau_m) = e^{-R\tau_m} \approx 1 - R\tau_m + \frac{1}{2}R^2\tau_m^2 - \ldots + \frac{(-1)^n}{n!}R^n\tau_m^n + \ldots$$

Equation (2.8)

Truncation after the second term results in a linear relationship between the measured intensities and the relaxation rate constants, valid for short ($\tau_m \rightarrow 0$) mixing times. This provides an easy method to calculate distances provided the intensities are obtained with a sufficiently short τ_m (e.g. 50-100ms). However, the use of short mixing times limits the signal-to-noise ratio obtainable with cross-peaks in the 2-D NOE spectrum and whenever at least one proton approaches either of the 'isolated pair' at a distance less than the distance between the pair, the approximation breaks down for practical values of τ_m for molecules with an effective correlation time greater than a nanosecond.

Typically, an inter-nuclear distance is estimated from a cross-peak intensity a_{ij} by making reference to a fixed distance r_{ref} in the molecule and its corresponding intensity a_{ref}. The H5-H6 distance in cytidine in constant at 2.45Å, and hence these cross-peaks serve as internal 'yardsticks' to which other distances can be scaled and evaluated.

Distance are calculated according to Equation (2.9):

$$r_{ij} = r_{ref}\left(\frac{a_{ref}}{a_{ij}}\right)^{1/6}$$

Equation (2.9)

(It is assumed that the correlation time for the distance to be measured is the same as the correlation time for the reference distance). However, analysis with ISPA introduces a systematic error; calculated distances are shorter than true distances. This is due to the neglect of all other protons involved in the relaxation network. Since the peak intensity is thus ascribed to two protons, their corresponding inter-proton distance is consistently underestimated. This method incurs serious errors if quantitative distances are required and as the distances become larger, the systematic deviation between true and calculated distances increases.

The implication of using ISPA-derived distances in RMD calculations is that the bounds, particularly the upper bound, may need to be relaxed more than usual. If the distances are systematically underestimated, the molecule, e.g. protein, may never be able to get to the vicinity of the global minimum.

2.1.7.3. *CO*mplete *R*elaxation *M*atrix *A*nalysis (CORMA)

Accurate intensities (incorporating the effects of spin diffusion and network relaxation effects) can be readily calculated for a known 3-D model structure after diagonalizing the relaxation matrix by the complete relaxation matrix analysis procedure. This rigorous approach uses linear algebra, and the simplifications which arise from working with eigenvalues and eigenvectors of a matrix and the mathematical theory is beyond the scope of this thesis. Comparison between these intensities calculated for the model and intensities measured experimentally then allows a determination of the validity of the model structure. A program, developed for this calculation, CORMA [100,101], has been extended to include torsions in the structural refinement in the program COMATOSE [102].

The complete relaxation matrix analysis approach is accurate, can accommodate any size spin system (computer size limitations only), is not limited to any range of mixing times, incorporates spin diffusion and can utilise any molecular motion model.

2.1.7.4. Direct calculation of distances from experimental spectra (DIRECT)

The limitation of the trial-and-error approach using CORMA is that is it governed by the choice of structural models. It is capable only of discriminating between the proposed structures and of indicating regions of good or bad fit between the model and true solution structure. The 'ideal' way to calculate distances is to directly transform the scaled intensities from the experimental 2-D NOE spectra into their associated dipole-dipole relaxation rates and then into distances [103,104]. Distances can be determined directly without making the ISPA approximation and by using a method that is not explicitly model-dependent. Rearrangement of Equation (2.7) gives Equation (2.10) which shows the fundamental logarithmic relationship between the rates and mixing coefficients:

$$R = \frac{-\ln\left(\dfrac{a(\tau_m)}{a(0)}\right)}{\tau_m}$$

Equation (2.10)

where a is the matrix of mixing coefficients which are proportional to the measured 2-D NOE intensities, $a(0)$ refers to the diagonal matrix of intensities for an experiment with mixing time zero and R is the matrix of relaxation rates.

Assuming isotropic tumbling and using Equations (2.5) and (2.6), distances can be calculated directly from the rate matrix (*via* diagonalisation of the experimental NOE matrix). For relatively small molecules yielding spectra with a very high signal-to-noise ratio, in which most of the major peaks can be resolved and accurately estimated, this is the ideal method of distance determination. The limiting factor in the calculation is frequently the paucity of usable information (2-D NOE intensities) and the need for enough of both diagonal and cross-peaks to be assigned and accurately measured. Frequently peaks are poorly resolved and hence the intensity matrix used in analysis will necessarily incorporate many errors. The resulting distance matrix will consequently also be in error and will not represent the spatial arrangement of protons leading to the intensities. A significant amount of the total magnetisation is also undetectable as it lies below the spectral noise level. As with the ISPA method, the DIRECT approach is better when used at short mixing times when spin diffusion is less significant.

2.1.7.5. Iterative Relaxation Matrix Analysis (IRMA)

A useful extension of the DIRECT calculation of distances is the method of iterative relaxation matrix analysis (IRMA) proposed by Kaptein and co-workers [105] who have shown that it is feasible to substitute into the scaled matrix of observed intensities, the corresponding intensities calculated for a reasonable model structure. This yields a full set of mixing coefficients a' which can be successfully transformed to distances *via* Equations (2.10), (2.5) and (2.6). Such an approach has been applied in an iterative manner to successfully generate solution structures [105,106].

The general scheme is shown in Figure 2.12. The critical feature is the generation of the augmented intensity matrix which contains all the experimental intensities (which have been scaled) and the intensities calculated from a suitable model structure. While straightforward back-calculation of distances will fail in the absence of all cross-and diagonal-peak intensities, experimental intensities are substituted, wherever possible, into the theoretical spectrum to yield a full spectrum suitable for direct solution of the rate and distance matrices. This process incorporates all the effects of network relaxation and spin diffusion and will be relatively accurate, depending on how close the model is to the true structure. To improve the structure, RMD can be used to generate an improved model which incorporates the distance restraints generated by the previous pass of the intensity/distance calculation.

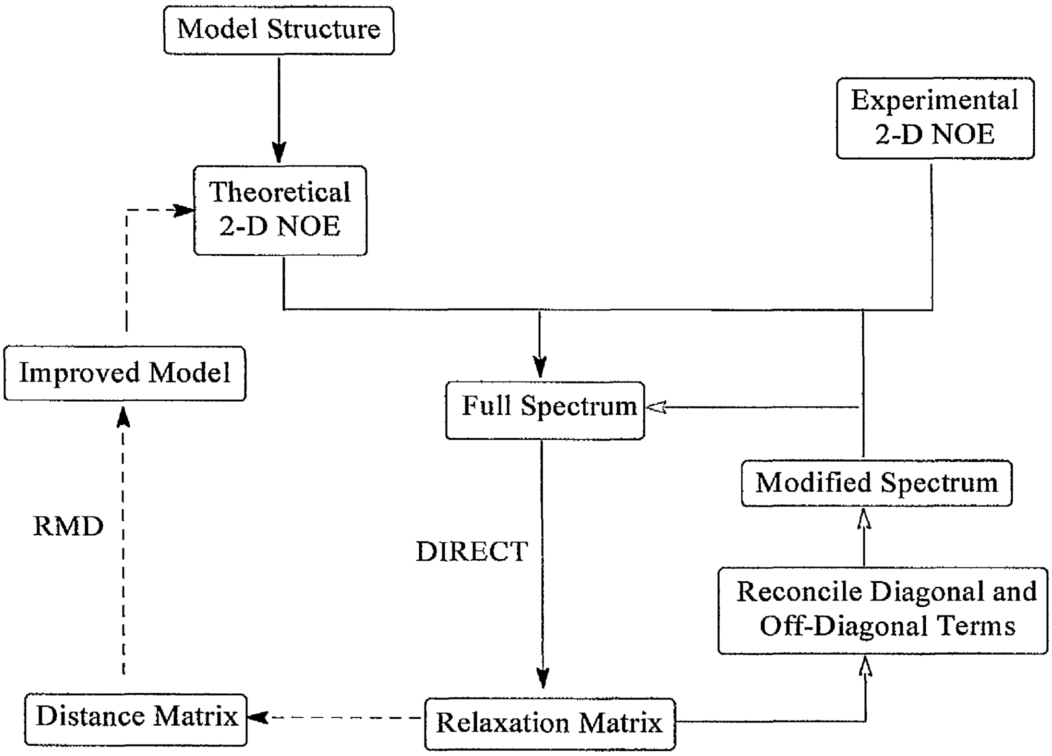


Figure 2.12. Schematic diagram of the iterative relaxation matrix analysis (IRMA). The variation introduced in the MARDIGRAS algorithm compared to IRMA is shown in red (adapted [107]).

By solving this intensity matrix for distances, a distance set is generated which is in reasonable agreement with the model, but which is also partially restrained by the experimental intensities. By iterating through the cycle of structure generation *via* distance geometry or molecular dynamics on one branch and the CORMA-type calculations on the other, a self-consistent structure is eventually reached which incorporates all the structural information inherent in the 2-D NOE intensities. RMD is used after obtaining distances from the substituted matrix to generate a new structure that is then cycled through the process again.

Use of the complete relaxation matrix methodology permits longer mixing times to be employed, with consequently larger intensities for weak cross-peaks and the possibility of measuring longer distances. However, spin-diffusion effects limit the extension to yet longer mixing times and larger cross-peak intensities.

It is possible to determine distances from the relaxation matrix analysis alone, prior to generating new structures using a program called MARDIGRAS (see below).

2.1.7.6. *M*atrix *A*nalysis of *R*elaxation for *DI*scerning *G*eomet*R*y of an *A*queous *S*tructure (MARDIGRAS)

A variant of IRMA has been developed termed MARDIGRAS [108]. The variation introduced in MARDIGRAS compared to IRMA is shown in Figure 2.12 highlighted in red and utilises requirements for internal consistency in the relaxation matrix itself rather than iterating through the computer time-consuming RMD procedures after a single pass through the relaxation matrix. A more detailed flow chart of the MARDIGRAS program is shown in Figure 2.13 [108]. With MARDIGRAS, constrained distances are used to give relaxation matrix intensities as a first approximation and diagonal and off-diagonal elements are required to be consistent. This yields a modified spectrum, which is then recombined with the experimental 2-D NOE data to produce a new augmented matrix.

A normalisation scheme has been incorporated based on the average fit of all experimental intensities to a corresponding calculated fixed-distance intensity, e.g. C-H5-C-H6 = 2.45Å. After normalisation, and during the iteration process, only cross-peak rates that have a corresponding observed intensity are allowed to change. Cross-peak rates that correspond to known distances, e.g. aromatic protons C-H5-C-H6, are not allowed to change and are reset to their known values at every iteration. Cross-peak rates that correspond to unusually short contacts (arbitrarily set to 1.5Å) are reset to the minimum allowable value.

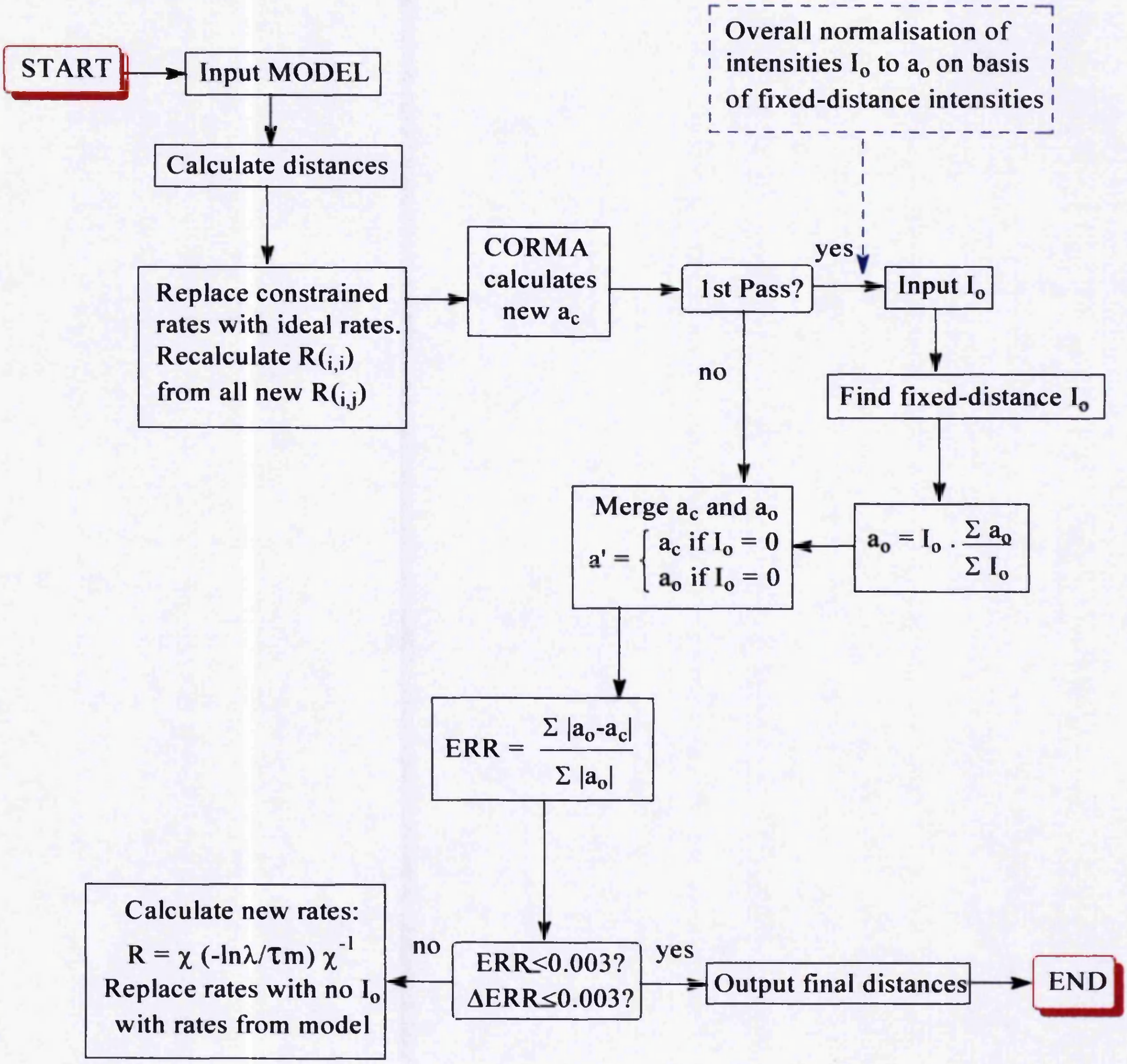

START
Input MODEL
Calculate distances
Replace constrained rates with ideal rates. Recalculate $R(_{i,i})$ from all new $R(_{i,j})$
CORMA calculates new a_c
1st Pass?
yes
no
Overall normalisation of intensities I_o to a_o on basis of fixed-distance intensities
Input I_o
Find fixed-distance I_o
$a_o = I_o \cdot \frac{\Sigma a_o}{\Sigma I_o}$
Merge a_c and a_o
$a' = \begin{cases} a_c \text{ if } I_o = 0 \\ a_o \text{ if } I_o = 0 \end{cases}$
$\text{ERR} = \frac{\Sigma |a_o - a_c|}{\Sigma |a_o|}$
ERR≤0.003?
ΔERR≤0.003?
no
yes
Calculate new rates:
$R = \chi\ (-\ln\lambda/\tau m)\ \chi^{-1}$
Replace rates with no I_o with rates from model
Output final distances
END

Figure 2.13. Flow chart of the MARDIGRAS algorithm [108]. The program is a loop with a 'forward' CORMA calculation to obtain the calculated mixing coefficients, a_c, from either the model structure (first pass) or the iterated relaxation rates (all subsequent passes), and a 'back-calculation' which generates the iterated relaxation rates after merging the calculated and experimental (observed) mixing coefficients, a_o. I_o refers to the observed intensities.

Finally, after new cross-relaxation rates are assigned to the measured intensities and all other rates are reset to their initial values, the diagonal rate constants (R_{ii}) are replaced by the appropriate sums (*cf.* Equation (2.5)) based on the calculated and constrained cross-peak rates. This step ensures that the relaxation matrix is numerically self-consistent. After the error between calculated and observed intensities reaches a minimum value, the final distances are calculated and ranges assigned.

MARDIGRAS is relatively insensitive to the choice of starting model and relatively fast[108]. Although this method is substantially more involved than ISPA, the increase in computer time is small in relation to the complete structure determination process. It has been used in many cases to solve the solution structures of nucleic acids [109-111].

2.1.8. Summary

NMR spectroscopy is a very powerful and versatile tool in the structural studies of biological macromolecules. An important selection of the multitude of techniques in use has been discussed, primarily those needed for the experimental work in this chapter of the thesis.

2.1.9. Aims

The aims of the study was to evaluate the full three-dimensional solution structure of the binary system shown in Figure 2.4 involving the following experimental stages:

1. To record, analyse and assign the 1-D and 2-D ^1H NMR spectra of the individual components and of the complete complex 1:2:3. COSY and TOCSY spectra will be used to allow determination of the connectivities of all *J*-coupling protons as members of established spin systems. The analysis of NOESY spectra will then allow these spin systems to be attributed to specific nucleotide residues and identify other aromatic protons *via* through-space dipole-dipole connectivities.
2. To investigate the thermal stability of the duplex by variable temperature NMR studies of the exchangeable imino proton region.
3. To determine proton-proton distance ranges from peak integrals using a full-relaxation matrix analysis implemented in the MARDIGRAS algorithm.
4. To use restrained molecular dynamics and minimisation calculations to refine the final structure.
5. To analyse the structural parameters of the final structure.

2.2. MATERIALS AND METHODS

2.2.1. Synthesis of oligonucleotide derivatives and complex preparation for NMR spectroscopy

The oligonucleotides pGTATCAGTTTCT (**1**), 5'-dAGAAACp- Im (**2**) and Im'-5'-pdTGATAC (**3**) were kindly provided by Dr. T. Bramova, Novosibirsk, and derivatised by Dr. E. B. Bichenkova, University of Manchester, who also prepared the NMR samples as follows. Each oligonucleotide (**1**, **2** and **3**) was lyophilised (x3) from 99.9% D_2O and the chemical structures confirmed by 1-D ^{1}H-NMR spectroscopy and by ^{1}H-COSY. An equimolar ratio mixture of the three components of the required complex (**1:2:3**) was obtained by scaling the amount of each oligonucleotide component added to the NMR sample relative to the amount of a standard compound used as an internal NMR calibrant for integration (TSP). For this purpose an aliquot of TSP (60μl of 4mM in D_2O) was added to aliquots (540μl) of each NMR sample, and the amount of each oligonucleotide component calculated from the 1-D ^{1}H-NMR spectrum relative to the integral area of the TSP ^{1}H-NMR signal. The complex 1:2:3 was prepared by dissolving 2.5μmol of each lyophilised oligonucleotide in buffer (0.6ml of 100mM NaCl, 10mM NaH_2PO_4/Na_2HPO_4, pH7.20, 0.1mM EDTA, 1.2mM TSP prepared in 99.9% D_2O). To investigate the imino proton region of the duplex **1:2:3** the sample was lyophilised (x3) from 99.9% D_2O and re-dissolved in 90% H_2O/10% D_2O.

2.2.2. NMR Spectroscopy and instrumentation

All NMR data were accumulated on Varian Unity 400 (400MHz) or Unity 600 (600MHz) NMR spectrometers equipped with Sun host computers. A dual (inverse) ^{1}H (X) probehead was used for proton detection. Data were processed using VNMR 4.3 software.

One-dimensional ^{1}H-NMR spectra of the free oligonucleotides **1**, **2** and **3** and of the 1:1:1 complex (pGTATCAGTTTCT: AGAAACp-**Im**:**Im'**-pdTGATAC, **1:2:3**) were acquired at 25°C (400MHz). Data were collected into 32K complex data points over a spectral width of 7000 Hz giving a final resolution of 0.43 Hz/point. For each spectrum 128 transients were acquired with 2.5s recycle delay during which the residual HOD resonance was suppressed by continuous, low-power irradiation.

Standard pulse sequences were used to acquire sets of two-dimensional NMR data for resonance assignment and for structural analysis including COSY, TOCSY and NOESY data using published protocols [112-116]. Pure absorption two-dimensional COSY, TOCSY and NOESY data were collected with quadrature detection into 4096 complex data points for 2x512 t_1 increments with 48-64 transients being acquired for each t_1 increment. In the case of the TOCSY experiment the residual HOD resonance was suppressed by continuous, low-power irradiation during the recycle time of 2s. In the NOESY experiment solvent suppression was achieved by continuous, low-power irradiation during both recycle and mixing times. The spectral width used was 7000 Hz giving a final ω_2 digital resolution of 1.71 Hz/point. NOESY data were acquired with a mixing time of 200ms at 15°C and 25°C (to find the optimal conditions for analysis and assignment purposes) and with mixing times of 50, 100, 150, 200 and 500ms at 25°C (for conformational analyses). TOCSY data were acquired with a mixing time of 50ms at 25°C. All two-dimensional experiments were performed in the non-spinning mode. Two-dimensional NMR data sets were transformed after zero-filling in t_2 to 2K data points, with apodization in both dimensions prior to Fourier transformation. Both 1-D and 2-D data sets were referenced internally to the singlet methyl resonance of TSP at 0ppm.

Measurement of spin-lattice relaxation times (T_1) was performed at 25°C for some resolved signals of **1:2:3** using the standard two-pulse sequence inversion recovery experiment. Data were collected into 16K complex data points over a spectral width of 5000 Hz, giving a final resolution of 0.31Hz/point. The arrayed time variable (d_2) was changed over eleven steps from 0.0625s to 64s and eight transients were acquired for each d_2 increment with the recycle delay of 40s. Measurement of the spin-spin relaxation time (T_2) for the same protons was performed at 25°C using the Carr-Purcell-Meiboom-Gill (CPMG) pulse sequence [88,89]. Data were collected into 32K complex data points over a spectral width of 5000 Hz giving a final resolution of 0.32 Hz/point. The spin-echo cycle time d_2 was 2.5ms and the total echo time (τ) was varied over thirty steps from 10ms to 300ms in 10ms increments. For each τ increment, 256 transients were acquired with the recycle delay of 10s.

REFOPTNY data were accumulated at 400MHz for the complex dissolved in H_2O/D_2O (90%/10%) and collected with quadrature detection into 4096 complex data points for 2 x 256 t_1 increments in hypercomplex phase-sensitive mode [117]. 96 transients were acquired for each free induction decay. A 1.5s recycle delay and a 175ms mixing time were used in the NOESY portion of the pulse sequence. Data were acquired over a 14kHz spectral width in both frequency dimensions.

The 1-D REFOPT pulse sequence [97] was used in variable temperature experiments (400MHz) to monitor the oligonucleotide imino proton resonance region at 400MHz. Temperatures of 11, 15, 20, 25, 30, 35, 40, 45, 50, 52 and 55°C were used in the variable temperature experiments of the 12-mer duplex. The data were collected into 32K complex data points over a spectral width of 10kHz; 60 transients were accumulated with a pre-saturation period of 2s.

2.2.3. Molecular modelling

Molecular modelling was performed using SYBYL 6.4.2, (Tripos Inc.) on a Silicon Graphics Indy R4400 workstation. All restrained molecular dynamics (RMD) and restrained minimisation calculations were performed using Kollman-All force field parameters. Initial structures for the duplex were constructed from canonical B-form [118] nucleic acid frameworks (double-stranded B-DNA). The phosphate bond between ^{18}C and ^{19}T was removed, and an additional phosphate group added to the 3' oxygen of ^{19}T. The charge distribution and structural parameters for β-alanylhistamine were calculated using MOPAC [119] and written into the Kollman-All parameter database. Two β-alanylhistamine fragments were coupled to the 3'- and 5'-phosphate groups of residues ^{18}C and ^{19}T, respectively. The starting structures were energy-minimised *in vacuo* using a conjugate gradient method with initial Simplex optimisation and Kollman-All force field parameters to avoid unfavourable van der Waals' interactions.

The lower and upper levels of distance constraints were evaluated based on complete relaxation matrix analysis [105,107,108,120] using the NMR/TRIAD/MARDIGRAS module (SYBYL 6.4.2). The ^{1}H-NOESY spectrum (τ_{mix} = 200ms, 600MHz) was used for cross-peak integration and the isotropic correlation time approach was used for the MARDIGRAS calculation. The correlation time, τ_c, for duplex **(1:2:3)**, evaluated from Equation (2.11) [121] and based on the averaged values of T_1 and T_2 obtained for some resolved signals in the ^{1}H-NMR spectrum (Table 1), was found to be 2.655ns.

$$\tau_c = 2/\omega \, [\, T_1/3T_2]^{1/2} \qquad \text{Equation (2.11)}$$

A total of 315 proton-proton distance-range constraints obtained from the MARDIGRAS calculations were incorporated into the subsequent RMD procedure. The presence of imino proton signals in the REFOPT spectra of the duplex showed that normal Watson-Crick hydrogen bonding was present for all of the base pairs of the duplex. An additional 28 distance and 36 angle restraints were therefore included to guarantee maintenance of Watson-Crick hydrogen bonds throughout the calculations (three hydrogen bonds for each of the four G-C base pairs and two for each of the A-T base pairs). To preserve the right-handed character of the DNA during the high temperature MD calculations, backbone dihedral angle constraints were introduced [122-124]. The force constants were maintained at 20kcalmol^{-1}Å^{-2}, 20kcalmol^{-1}rad^{-2} and 0.3kcalmol-1rad-1 for the distance restraints, angle restraints and torsion restraints, respectively during all calculations. The initial velocities were assigned with a Boltzmann distribution and a 1fs time step for the integrator. The effect of solvent was approximated by using a distance-dependent dielectric constant of 4ε. The system was gradually heated from 100 to 600K over 6ps (in 100K steps of 1ps each), maintained at 600K for 10ps, gradually cooled to 200K over the next 6ps (in 100K steps of 2ps each) and finally maintained at 200K for 5ps. The restraint force constants were kept at 20kcalmol^{-1}Å^{-2} and 20kcalmol^{-1}rad^{-2} during all MD calculations. The final 100 co-ordinate sets, arising from the last 5ps, were averaged and finally subjected to 1000 steps of restrained energy minimisation (an initial Simplex optimisation, followed by the Powell method with a gradient of 0.05kcalmol^{-1}, Kollman-All force field) to generate the restrained structures. Trial runs were performed to refine the restraints.

2.2.4. Structural characterisation

All backbone torsion angles, sugar conformations and helical parameters for the final structure were comprehensively analysed with the program CURVES 5.3 [125-127] to characterise the DNA structural features.

2.3. RESULTS AND DISCUSSION

2.3.1. NMR Spectroscopy of free oligonucleotides 1, 2 and 3

1-D ^1H NMR and ^1H -COSY spectra, recorded for the separate oligonucleotides **1**, **2** and **3** at 25°C, 400MHz, confirmed their nucleotide composition and the presence of **Im** and **Im'** groups in the case of **2** and **3**, respectively. The assignment of the **Im** and **Im'** protons for conjugates **2** and **3** was based on the COSY cross-signals between H_α-H_β, H_β-$CH_{2\gamma}$, $CH_{2\gamma}$-$CH_{2\delta}$, $CH_{2\varepsilon}$-$CH_{2\lambda}$ (see following section, Figure 2.4 and Table 2.1).

	Chemical Shift (ppm)					
	$CH_{2\delta}$	$CH_{2\varepsilon}$	$CH_{2\gamma}$	$CH_{2\lambda}$	H_α	H_β
Free **Im** [(a)]	3.37	3.03	2.80	2.50	8.65	7.21
AGAAACp-**Im**	3.51	3.23	2.94	2.65	8.59	7.27
Im'-pTGATAC	3.45	2.97	2.87	2.32	8.53	7.24
AGAAACp-**Im** in duplex **1:2:3**	3.50	3.24	2.92	2.65	8.40	7.12
Im'-pTGATAC in duplex **1:2:3**	3.38	2.95	2.87	2.33	8.40	7.12

[(a)] – Spectrum recorded in methanol (400MHz) at 25°C

Table 2.1. Chemical shifts of **Im** and **Im'** protons for free alkylimidazole groups in methanol, when coupled to the respective hexamer (AGAAACp-**Im** and **Im'**-pTGATAC) and after hybridisation of conjugates **1** and **2** with target oligonucleotide **3**. See Figure 2.4 for proton identities.

2.3.2. Qualitative NMR analysis of the 1:1:1 complex of 1:2:3

The first stage of the structural analysis of complex **1:2:3** was the assignment of non-exchangeable protons by means of COSY and TOCSY spectra, allowing determination of the connectivities of all *J*-coupled protons as members of established spin systems (H5 and H6 protons of cytosines, CH_3 and H6 protons of thymidines and sugar ring protons). To assign these spin systems to specific nucleotide residues and to identify other aromatic protons, through-space dipole-dipole connectivities were analysed using ^{1}H-NOESY data. Figure 2.14 shows the full 600MHz ^{1}H-NOESY spectrum of complex **1:2:3** recorded at 25°C at a 200ms mixing time. Qualitative analysis of the NOESY spectra suggested complex **1:2:3** to have the right-handed form of DNA. Thus, the assignment of aromatic, methyl and sugar-ring protons was performed using the established walk strategy [128] summarised below (see also Figure 2.8):

N_i (H6/H8/H5/CH_3) ↔ N_i (H1'/H2'/H2") ↔ N_{i+1} (H6/H8/H5/CH_3)
N_{i-1} (H6/H8) ↔ N_i (H6/H8) ↔ N_{i+1}(H6/H8)
N_i (H6/H8) ↔ T_{i+1}(CH_3) ↔ T_{i+1} (H6)
N_i (H6/H8) ↔ C_{i+1} (H5) ↔ C_{i+1}(H6)
N_i (H1'/H2'/H2") ↔ N_i (H3'/H4'/H5'/H5")
where N = any nucleotide residue, T = thymidine, C = cytidine

The expanded sections of the NOESY spectrum containing the regions for the H6/H8 - H1'/H5 protons are presented in Figure 2.15(a) and (b), with the sequential assignment shown for the 12-mer and for the two hexamers. Assignments are summarised in Tables 2.2 and 2.3, respectively. The cross-peaks corresponding to intra-nucleotide interactions H1'-H8/H6 and cytidine H5-H6 are marked by the appropriate nucleotide symbol.

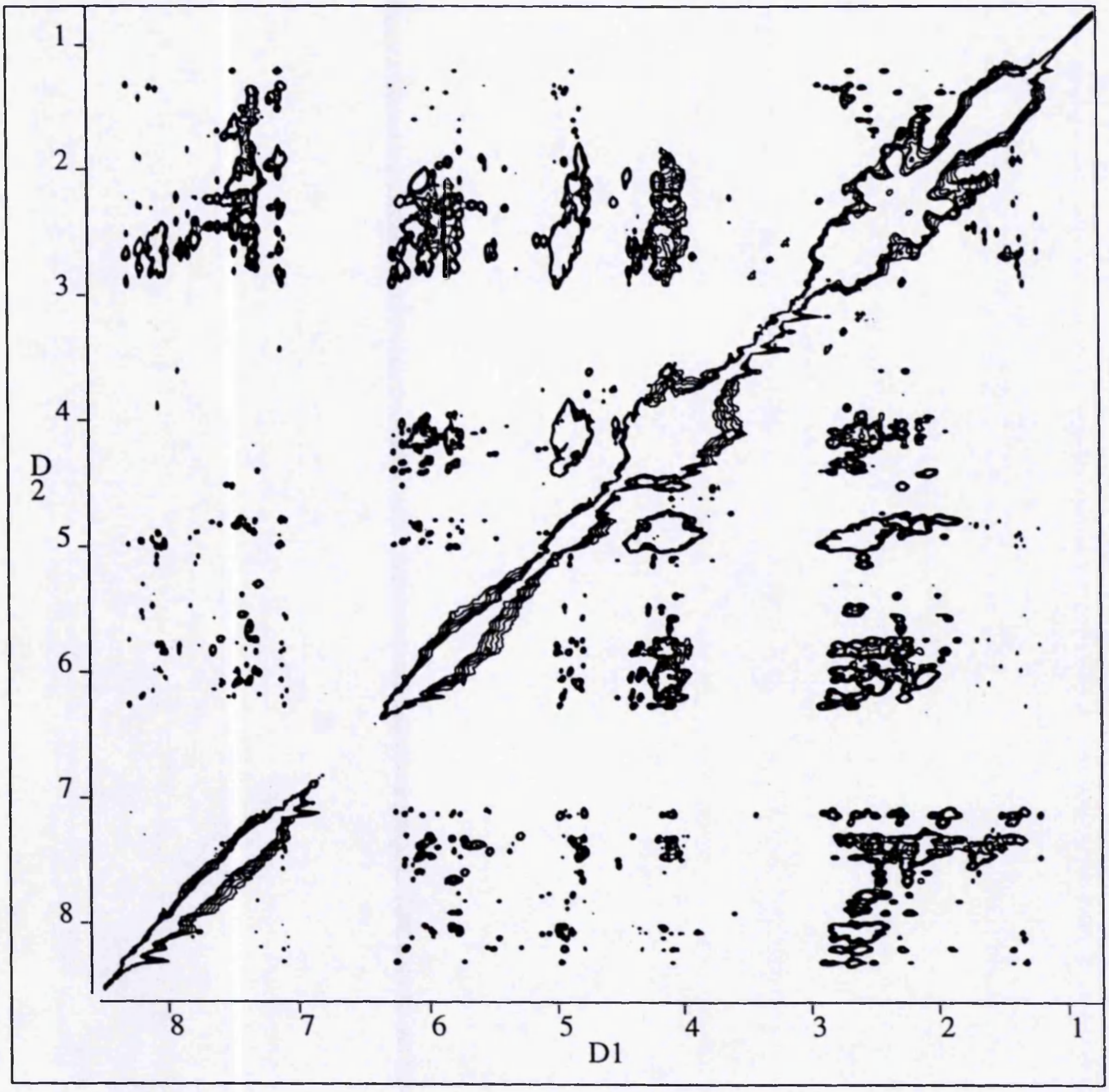


Figure 2.14. ^{1}H-NOESY spectrum (600MHz, mixing time 200ms, 25°C) of complex **1:2:3** in 100mM NaCl, 10mM NaH_2PO_4/Na_2HPO_4, pH7.20, 0.1mM EDTA and 1.2mM TSP, prepared in 99.9% D_2O. (D1 and D2 refer to the F2 and F1 frequency domains, respectively. Axes are labelled in ppm).

5' - ^{1}G - ^{2}T - ^{3}A - ^{4}T - ^{5}C - ^{6}A - ^{7}G - ^{8}T - ^{9}T - ^{10}T - ^{11}C - ^{12}T - 3'
3' - ^{12}C- ^{23}A- ^{22}T- ^{21}A- ^{20}G- ^{19}T ^{18}C - ^{17}A- ^{16}A - ^{15}A - ^{14}G - ^{13}A - 5'
| |
Im' Im

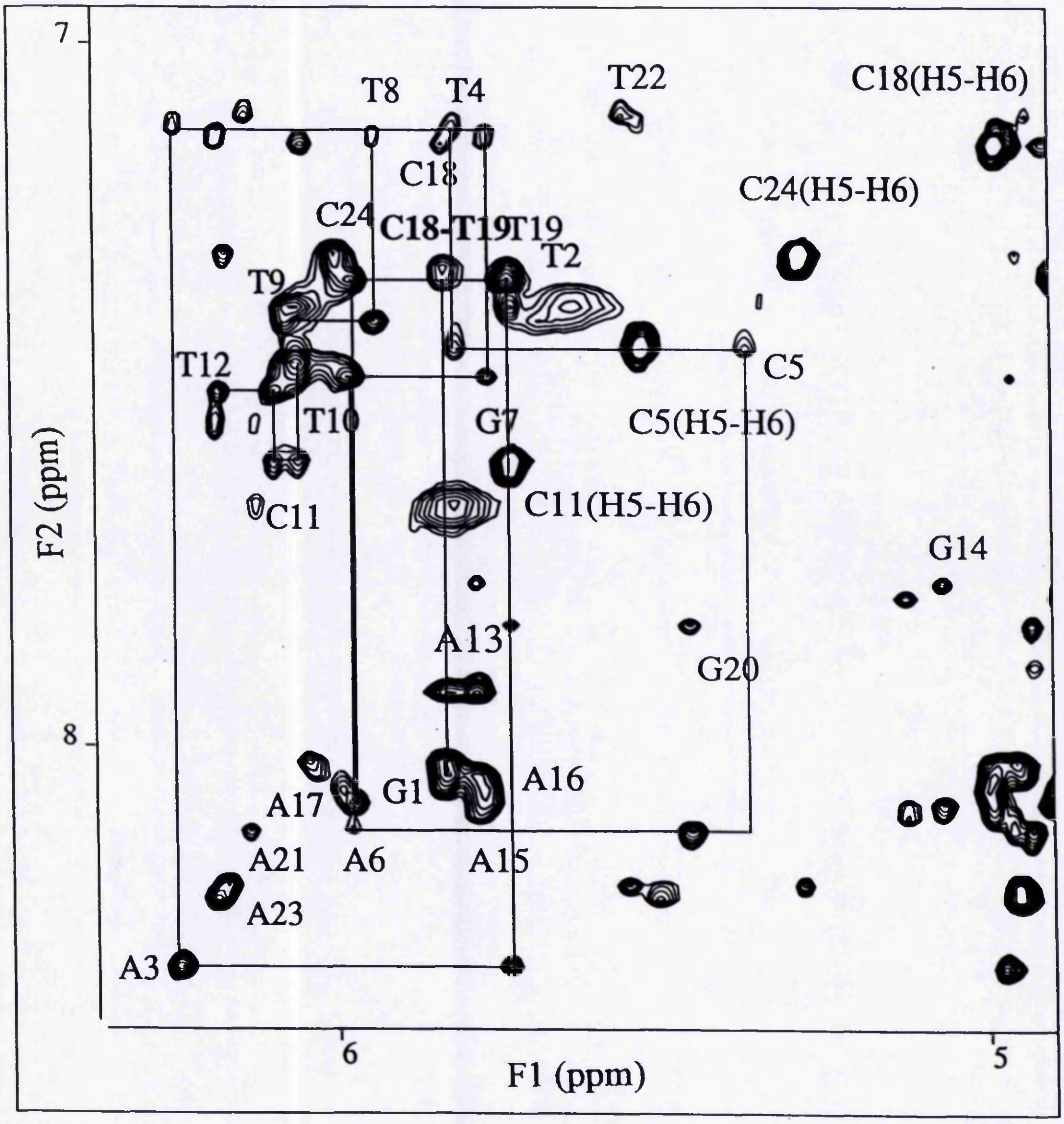


Figure 2.15(a). Expanded region of the ^{1}H-NOESY spectrum (600MHz, mixing time 200ms, 25°C) of complex 1:2:3, showing the H6/H8 - H1'/H5/H3' resonance region. Buffer composition was as described for Figure 2.14. The sequential assignment of oligonucleotide protons for the 12-mer is shown. The cross-peaks corresponding to intranucleotide interactions H1'-H8/H6 and cytidine H5-H6 are marked by the respective nucleotide symbols.

5' - ^{1}G - ^{2}T - ^{3}A - ^{4}T - ^{5}C - ^{6}A - ^{7}G - ^{8}T - ^{9}T - ^{10}T - ^{11}C - ^{12}T - 3'
3' - ^{12}C- ^{23}A- ^{22}T- ^{21}A- ^{20}G- ^{19}T ^{18}C - ^{17}A-^{16}A - ^{15}A - ^{14}G - ^{13}A - 5'

| |
Im' Im

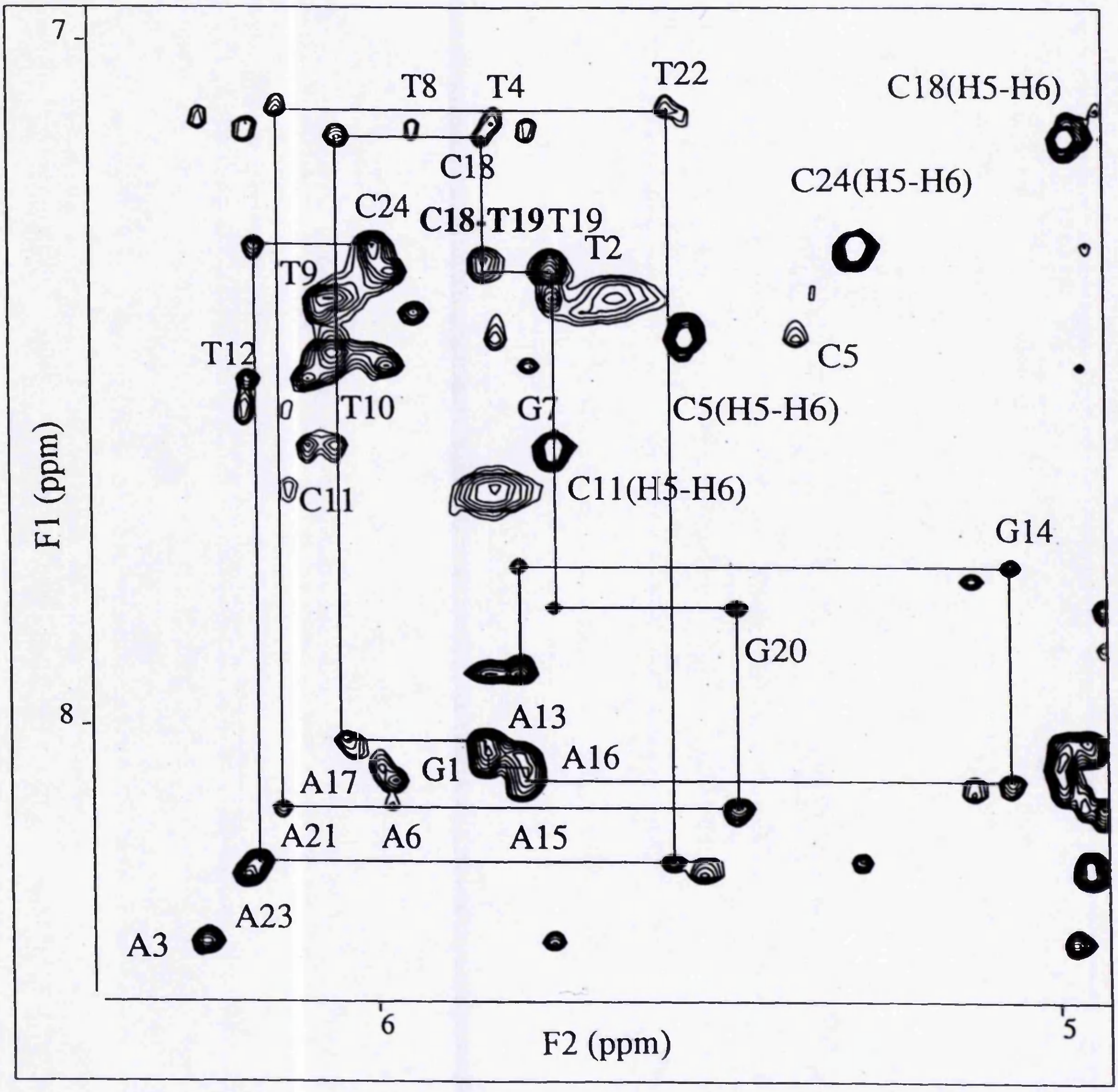


Figure 2.15(b). Expanded region of the ^{1}H-NOESY spectrum (600MHz, mixing time 200ms, 25°C) of complex **1**:**2**:**3**, showing the H6/H8 - H1'/H5/H3' resonance region. Buffer composition was as described for Figure 2.14. The sequential assignment of oligonucleotide protons for the 6-mers is shown. The cross-peaks corresponding to intra-nucleotide interactions H1'-H8/H6 and cytidine H5-H6 are marked by the respective nucleotide symbols.

	^{1}G	^{2}T	^{3}A	^{4}T	^{5}C	^{6}A	^{7}G	^{8}T	^{9}T	^{10}T	^{11}C	^{12}T
H6/ H8	8.08	7.34	8.31	7.11	7.42	8.11	7.47	7.13	7.40 *(b)*	7.46	7.59	7.50
H1'	5.97	5.73	6.25	5.82	5.38	5.98	5.77	5.94	6.07	6.06	6.10	6.19
H5/ CH$_3$	-	135	-	1.32 *(b)*	5.53	-	-	1.21	1.52	1.61	5.73	1.69 *(b)*
H2	2.69	2.24	2.67	*(c)*	1.99	*(c)*	2.66	2.04	2.12 *(b)*	2.10	2.24	2.26 *(b)*
H2"	2.76	2.51	2.90	2.37	2.32 *(b)*	*(c)*	2.81	2.49 *(b)*	2.57	*(c)*	2.44 *(b)*	2.45
H3	4.90	4.88	4.97	*(a)*	4.77	*(a)*	4.96 *(b)*	*(a)*	4.83	4.96 *(b)*	4.82	4.52
H4	4.30	4.20	4.40	4.78	*(a)*	*(a)*	4.84	4.79	*(a)*	4.42	*(a)*	4.40
H5	3.90	4.14	4.25	4.09 *(b)*	4.05	*(a)*	*(a)*	4.26	4.12	4.15	4.15	4.10
H5"	3.86 *(b)*	4.08	4.09	4.09 *(b)*	4.05	*(a)*	*(a)*	4.18	*(a)*	4.04	*(a)*	4.01

(a) The signal was not detectable

(*b*) The signal was very broad or of low intensity

(*c*) The signal was in a crowded region and therefore not assigned

Table 2.2. Chemical shifts of the non-exchangeable protons of the 12-mer component (**1**) in complex **1:2:3** at 25°C, pH7.20 in 100mM NaCl, 10mM NaH_2PO_4/Na_2HPO_4, 0.1mM EDTA, 1.2mM TSP, prepared in 99.9% D_2O.

	^{13}A	^{14}G	^{15}A	^{16}A	^{17}A	^{18}C	^{19}T	^{20}G	^{21}A	^{22}T	^{23}A	^{24}C
H6/ H8	7.92	7.77	8.08 *(b)*	8.05 *(b)*	8.02 *(b)*	7.14	7.33	7.83	8.12	7.10	8.20	7.30
H1	5.78	5.07	5.78	5.83	6.05	5.83	5.74	5.47	6.14	5.56	6.19	6.00
H5/ CH_3	-	-	-	-	-	4.98	1.39	-	-	1.35	-	5.29
H2	2.23 *(b)*	2.56	*(c)*	*(c)*	2.48	1.93 *(b)*	1.85	2.60	2.56 *(b)*	1.91 *(b)*	2.61 *(b)*	2.03 *(b)*
H2	2.42	2.56	*(c)*	*(c)*	*(c)*	2.28	2.36	2.69	2.83	2.30	2.79	2.11 *(b)*
H3	4.99	4.88	4.99 *(b)*	*(a)*	4.97	4.91	4.79	4.93	4.93	4.93	4.95	4.95
H4	4.72	4.22	4.35 *(b)*	4.36	4.35	*(a)*	*(a)*	*(a)*	4.39	4.79	4.33	4.40
H5	3.60	*(a)*	4.10	4.17	(a)	(a)	*(a)*	4.26	4.17	4.07	4.09 *(b)*	4.00
H5	3.60	2.57	4.03	4.17	3.99	*(a)*	*(a)*	4.16	4.17	4.07	4.06 *(b)*	3.97

(a) The signal was not detectable

(*b*) The signal was very broad or of low intensity

(*c*) The signal was in a crowded region and therefore not assigned

Table 2.3. Chemical shifts (ppm) of the non-exchangeable protons for components **2** and **3** in complex **1:2:3** at 25°C, pH7.20 in 100mM NaCl, 10mM NaH_2PO_4/Na_2HPO_4, 0.1mM EDTA, 1.2 mM TSP, prepared in 99.9% D_2O.

Under the experimental conditions used at 25°C, oligonucleotides **1**, **2** and **3** form a tight, specific complex without visible fraying at the ends. This conclusion is based on the high intensity of the cross-peaks originating from the terminal base-pairs, *viz.* the following interactions: ^{1}G(H8)-^{1}G(H1'/H2'/H2"), ^{1}G(H1'/H2'/H2")-^{2}T(H6), ^{11}C(H1'/H2'/H2")-^{12}T(H6), ^{12}T(H6)-^{12}T(H1'/H2'/H2"), ^{13}A(H8)-^{13}A(H1'/H2'/H2"), ^{13}A(H1'/H2'/H2")-^{14}G(H8), ^{23}A(H1'/H2'/H2")-^{24}C(H6) and ^{24}C(H6)-^{24}C(H1'/H2'/H2").

Uninterrupted connectivities were observed between aromatic and sugar ring protons along both nucleotide chains of the duplex, including in the gap from ^{18}C to ^{19}T. The cross-peaks corresponding to the interactions between ^{18}C(H1') and ^{19}T(H6), ^{18}C(H2') and ^{19}T(H6) plus ^{18}C(H2") and ^{19}T(H6) were of normal intensity indicating the regularity of the double-stranded helical structure in this region (Figure 2.16(b)).

2.3.3. Qualitative NMR analysis of the alkylimidazole groups in the binary system.

The chemical shifts of the alkylimidazole proton signals of the free groups in methanol prior to coupling, after coupling to the respective hexamer (**Im** and **Im'** respectively) and of the alkylimidazole moiety in the 12-mer duplex (**1:2:3**) are shown in Table 2.1. It is evident that the formation of the binary system from the free hexamer conjugates (**2** and **3**) and target 12-mer (**1**) does not change the chemical shifts of the methylene protons (H_γ, H_δ, H_ε and H_λ) significantly. Based on this observation, the **Im** and **Im'** methylene protons for **1:2:3** were assigned using those of the respective protons of the free conjugates.

In contrast, the chemical shifts of the imidazole ring protons, H_α and H_β, are shifted to higher field by 0.120-0.192ppm after complex formation, presumably due to interaction with the oligonucleotide part of the system and/or the shielding effect of the nearest oligonucleotide residues. This was explained by considering the differences between the experimental T_2 spin-spin relaxation times obtained directly from the CPMG experiment for a number of well-resolved oligonucleotide and imidazole signals, and apparent magnitudes of spin-spin relaxation time (T_2*) calculated from the observed line half-widths (Table 2.4). The lower magnitude of T_2* compared to T_2 suggests that the imidazole constructs have high conformational flexibility in the binary system, resulting in many rapidly exchanging contacts with the oligonucleotide residues (see Equation (2.3)). This effect is more pronounced for the **Im** and **Im'** H_β resonances indicating the dynamic behaviour of this group.

Proton	T_2 (x10^{-2}s)[a]	Error T_2 (x10^{-2}s)[a]	T_2*(calc.) (x10^{-2} s)[b]	T_1 (s)	Error T_1 (s)
His -H$_{\beta}$	9.0	0.89	5.1	1.5	0.11
^{23}A(H8)	5.9	0.39	4.4	1.1	0.08
^{21}A(H8)	4.9	0.36	4.4	1.5	0.12
^{6}A(H8)	5.7	0.41	3.9	1.5	0.06
^{15}A(H8)	6.0	0.34	4.1	1.4	0.08
^{17}A(H8)	7.0	0.44	4.8	1.3	0.07
^{13}A(H8)	9.0	0.65	5.1	1.6	0.07
^{20}G(H8)	5.4	0.21	5.3	4.1	0.22
^{14}G(H8)	5.8	0.28	5.0	1.9	0.10
^{11}C(H6)	4.0	0.28	2.7	3.5	0.37
^{10}T(H6)	5.9	0.23	2.6	1.4	0.06
^{9}T(H6)	4.4	0.12	3.2	2.7	0.06
^{24}C(H6)	5.3	0.16	[c]	1.7	0.08
^{22}T(H1')	5.1	0.37	5.1	1.8	0.07
^{5}C(H1')	4.3	0.29	2.5	1.9	0.14
^{5}C(H5)	5.1	0.44	2.9	2.3	0.37
^{12}T(CH$_3$)	5.9	0.37	4.1	1.3	0.02
^{10}T(CH$_3$)	5.3	0.24	4.4	1.5	0.04
^{9}T(CH$_3$)	4.7	0.15	4.4	1.5	0.01
^{2}T(CH$_3$)	5.1	0.33	4.5	1.5	0.04
^{22}T(CH$_3$)	4.8	0.32	4.8	1.4	0.03
^{4}T(CH$_3$)	4.5	0.29	5.3	1.7	0.08
^{8}T(CH$_3$)	5.2	0.33	5.3	1.2	0.31

[a] - values obtained from the spin-echo experiment;

[b] - values calculated from respective line half-widths; average error in T_2*(calc.) = 0.12x10^{-2}s, based on the resolution of 0.2Hz/point in the 1-D NMR spectrum.

[c] – value unobtainable due to strong signal broadening.

Table 2.4. Spin-lattice (T_1) and spin-spin relaxation times (T_2 and T_2*) obtained for some well-resolved signals in complex **1:2:3** determined at 25°C in 100mM NaCl, 10mM NaH_2PO_4/Na_2HPO_4, pH 7.2, 0.1mM EDTA and 1.2mM TSP, prepared in 99.9% D_2O.

2.3.4. Investigation of base-pair hydrogen-bonding by analysis of exchangeable imino protons

To probe duplex stability and the dynamics of the thermal denaturation process, the melting of the base-pairs structure was followed by monitoring the temperature-induced line broadening of the imino proton NMR signals involved in hydrogen bonding using the REFOPT pulse sequence. The assignment procedure for these signals was based on the assumption that the extremes of the "duplex" melted first (base-pairs ^{1}G:^{24}C, then ^{12}T:^{13}A, ^{2}T:^{23}A and so on), resulting in sequential broadening of the signals, moving towards the centre of the duplex until their eventual disappearance [129,130]. The assignments of some of the imino proton signals (^{22}T(N3H), ^{4}T(N3H), ^{8}T(N3H), ^{9}T(N3H) and ^{6}T(N3H)) were confirmed using through-space connectivities between them and their respective aromatic or/and H1' protons observed in NOESY spectra recorded in H_2O. The assignments of ^{20}G(N1H) and ^{7}G(N1H) by this approach were ambiguous. This problem was resolved by comparison with the similar binary system bearing pyrene-tetrafluoroazide moieties [60], taking into account the shielding properties of the pyrene group. The assignments are shown in Figure 2.16 and in Table 2.5.

Base-pair number	Imino proton	Chemical Shift (ppm)	
		11°C	25°C
1	^{1}G	12.80	Not detected
2	^{2}T	13.58	13.52
3	^{22}T	13.82	13.81
4	^{4}T	13.51	13.45
5	^{20}G	12.71	12.33
6	^{19}T	13.82	13.75
7	^{7}G	12.36	12.67
8	^{8}T	14.05	13.94
9	^{9}T	14.05	13.98
10	^{10}T	13.44	13.37
11	^{14}G	12.64	12.62
12	^{12}T	13.92	Not detected

Table 2.5. Chemical shifts of the exchangeable imino protons of complex **1:2:3** at 25°C, pH7.20 in 100mM NaCl, 10mM NaH_2PO_4/Na_2HPO_4, 0.1mM EDTA, 1.2mM TSP, prepared in H_2O. Note that the shifts of ^{1}G, ^{2}T, and ^{12}T imino protons were measured only at 11°C as these peaks were absent at 25°C.

5' - ^{1}G - ^{2}T - ^{3}A - ^{4}T - ^{5}C - ^{6}A - ^{7}G - ^{8}T - ^{9}T - ^{10}T - ^{11}C -^{12}T - 3'
3' - ^{12}C- ^{23}A- ^{22}T- ^{21}A- ^{20}G- ^{19}T ^{18}C - ^{17}A-^{16}A -^{15}A -^{14}G -^{13}A - 5'
| |
Im' Im

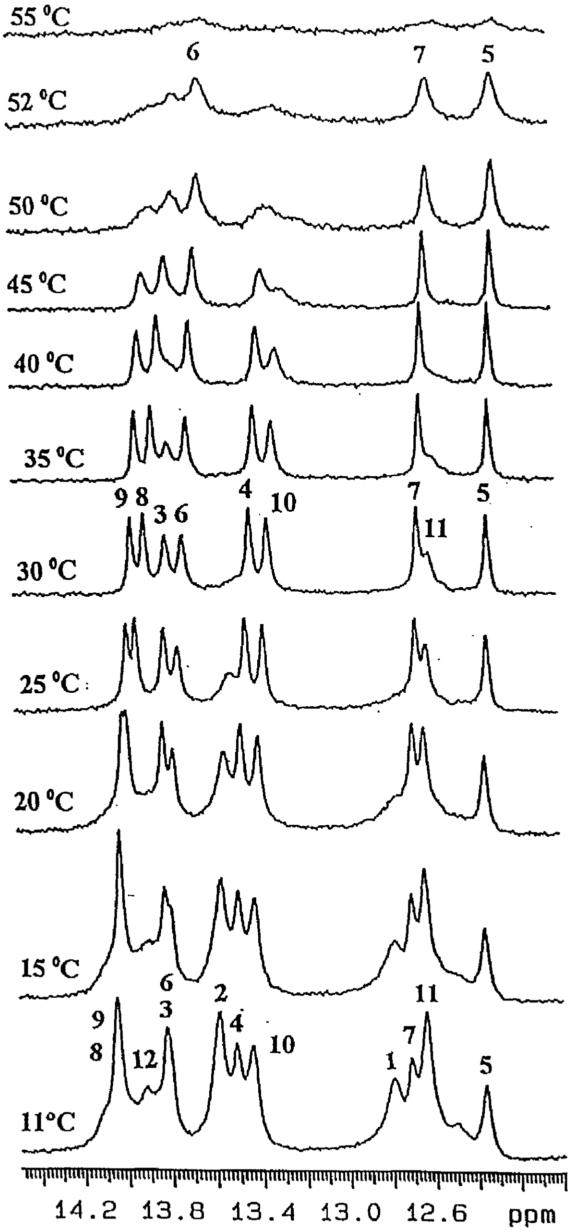


Figure 2.16. ^{1}H NMR variable temperature experiments recorded for the imino protons of complex **1:2:3** in 100mM NaCl, 10mM NaH_2PO_4/Na_2HPO_4, pH7.20, 0.1mM EDTA and 1.2mM TSP, prepared in $90\%H_2O$:$10\%D_2O$ (v/v). Peak numbers indicate the respective base pairs of the duplex.

The thermal denaturation process observed reflected the normal melting of duplex DNA, *i.e.* beginning at the termini of the duplex and working essentially monotonically towards the centre. There was no evidence from these studies that the binary complex had any melting pathway involving base-pair fraying starting from the middle of the duplex (around the gap at ^{6}A -^{19}T and ^{7}G -^{18}C). Indeed, the central region of complex **1:2:3** was essentially the most stable (see peaks 5, 6 and 7 in Figure 2.16, which correspond to ^{5}C -^{20}G, ^{6}A -^{19}T and ^{7}G -^{18}C) as would be observed in a regular, continuous DNA duplex. These results are in agreement with the structural regularity observed for non-exchangeable protons (see above).

From the above results we can confirm the presence of normal Watson-Crick hydrogen bonding for all base pairs in the complex, justifying inclusion of these 28 additional hydrogen-bond distance restraints and 36 angle restraints for molecular modelling calculations.

2.3.5. Evaluation of distance-range restraints using MARDIGRAS

Inter-proton distance range restraints were obtained using a full-relaxation matrix analysis approach implemented in the MARDIGRAS algorithm. Experimental cross-peak integrals from the NOESY spectrum (600MHz, 200ms mixing time) a correlation time of 2.655ns and co-ordinates for an initial canonic B-DNA structure were used as the input data for MARDIGRAS calculations. The 315 distance-range restraints obtained as the output data were used as restrictions in the following MD calculations.

2.3.6. Restrained molecular dynamics calculations

Restrained molecular dynamics simulations were performed starting from the canonical B-DNA structure. Figure 2.17 shows the starting structure and the refined structures, obtained from the 25ps RMD run after averaging the final 100 co-ordinate sets of the final 5ps followed by minimisation, respectively.

2.3.7. Structural analysis of DNA backbone and bases

A detailed analysis of the structural parameters of the DNA bases and backbone is given below. In the final minimised structure, the alkylimidazole group of ^{18}C was found to be located in the above major groove and that of ^{19}T was found in the lower minor groove (Figure 2.17).

Helical parameters for the final structure were analysed using CURVES 5.3 [125-127]. The structural parameters as a function of base pair position are shown in Figure 2.18(a)-(d) and those

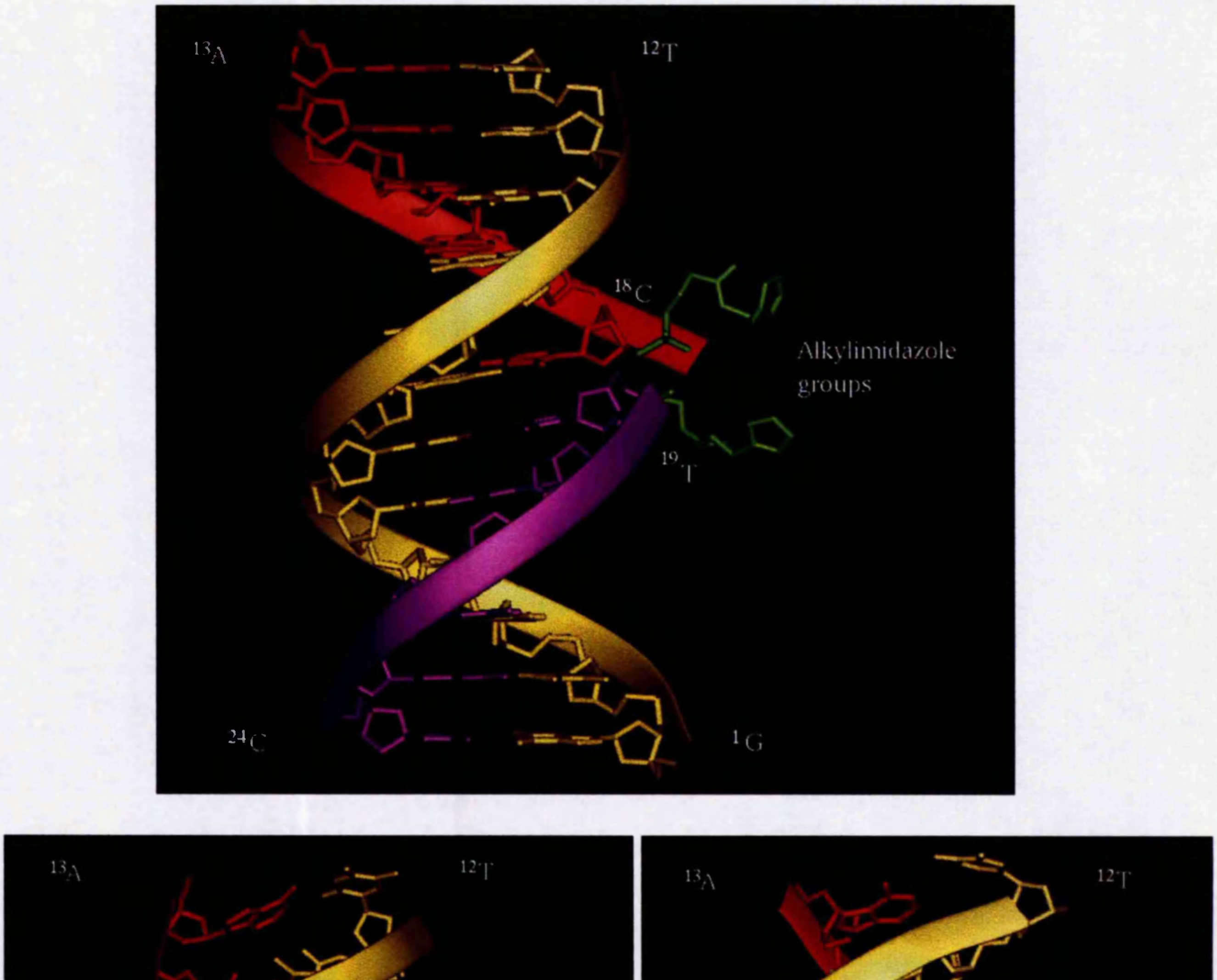


Figure 2.17. Starting structure, (above), for the 1:2:3 complex and the final structure, (below), obtained by averaging the last 100 co-ordinate sets of the final 5ps of RMD calculations followed by minimisation.

for canonic A- and B-DNA are included for comparison. The parameters reported for the terminal base pairs of the complex are probably less reliable because of dynamic processes and are included only for completeness.

2.3.7.1. Axis-base pair parameters (Figure 2.18(a))

X-axis displacement is closer to the values of B-DNA than A-DNA with positive deviations for all base pairs except for base pair 5, close to the site of modification. All base pairs in the complex exhibit a non-zero y-axis displacement, except for base pair 7, with a trend of positive to negative displacement from pairs 1 to 11. The inclination parameters for the 12 base pairs of the final structure show intermediate values between those of A- and B-DNA, although the central base pair 7 (and perhaps 6 and 8) display values closer to those of typical B-DNA. Tip values agree well with those of A- and B-DNA for base pairs 7 to 10 inclusive. For other base pairs we observe mainly negative deviations, which are most pronounced for base pairs 5 and 11.

2.3.7.2. Intra-base pair parameters (Figure 2.18(b))

The calculated parameters for shear, stretch, stagger and opening are often determined by the position of the nick in the DNA backbone and the location of the modifying groups. The shear shows positive deviations from both A- and B-DNA for most bases except around the region of the break in the backbone (pairs 6-8) and pair 11, which have strong negative deviations. The stretch values are on average slightly closer to those of B-DNA, with positive deviations for almost all base pairs, except base pairs 5 and 12. The stagger shows irregular positive and negative deviations alternating throughout the structure. For most of the duplex, the opening values are in the range of –5 to 20°, with strong distortion at base pairs 5 and 11 with values of –55 and 75° respectively. The large deviations in buckle from regular A- and B-DNA are mostly positive except for pairs 6 and 12. That for the former may be again attributed to the presence of the modifying groups in the middle of the duplex. The propeller-twist parameters for the duplex display strong negative deviations except base pair 6. Particular perturbation is observed for base pairs ^{5}A-^{19}T and ^{11}C-^{14}G.

2.3.7.3. Inter-base pair parameters (Figure 2.18(c))

The roll, slide, shift and twist values in regular A- and B-DNA conformations are essentially zero. The parameters for the modified duplex (**1:2:3**) show large deviations from regular DNA, both positive and negative, which generally correlates with the site of modification, in the vicinity of base pairs 4 to 7. The rise values are mainly a mixture of A- and B-DNA values except for base pair steps 5 and 11, which show very large distortions. The parameters for the tilt of the bases also show greatest deviations in steps 4 and 7, *i.e.* in the modified region.

2.3.7.4. Glycosidic torsion angles and sugar puckers (Figure 2.18(d))

The glycosidic torsion angles χ are well-defined by the NOE data and nucleotide residues 1, 6, 7, 18, 21 and 24 have χ values in the *anti*-domain, *i.e.* in the range expected for the *anti*-conformation (-90 to –170°). χ values for residues 5 and 11 lie in the *syn*-domain and all other residues occupy the conformational space intermediate between the *anti*- and *syn*-domains.

The sugar rings of all bases bar two are found in the S range conformation (3'-exo, 2'-endo), typical of B-DNA. For the nucleotide residues ^{18}C and ^{19}T, the sugar ring have 4'-exo and 4'-endo conformations respectively probably due to the modification at the respective 3'- and 5'-phosphate groups.

2.3.7.5. Backbone torsion angles (Figure 2.18(d))

The values for the five angles α, γ, δ, ε and ζ shows very good agreement with the values for standard B-DNA. This is actually hardly surprising since the backbone torsions were restrained in the RMD calculations. The only positive deviation is observed for γ of nucleotide residue 12 probably due to the end-effect. A negative deviation was observed for δ of residue C18 perhaps due to the modification at this point. The β values are predominantly intermediate between those of A- and B-DNA forms and do not display any regularity.

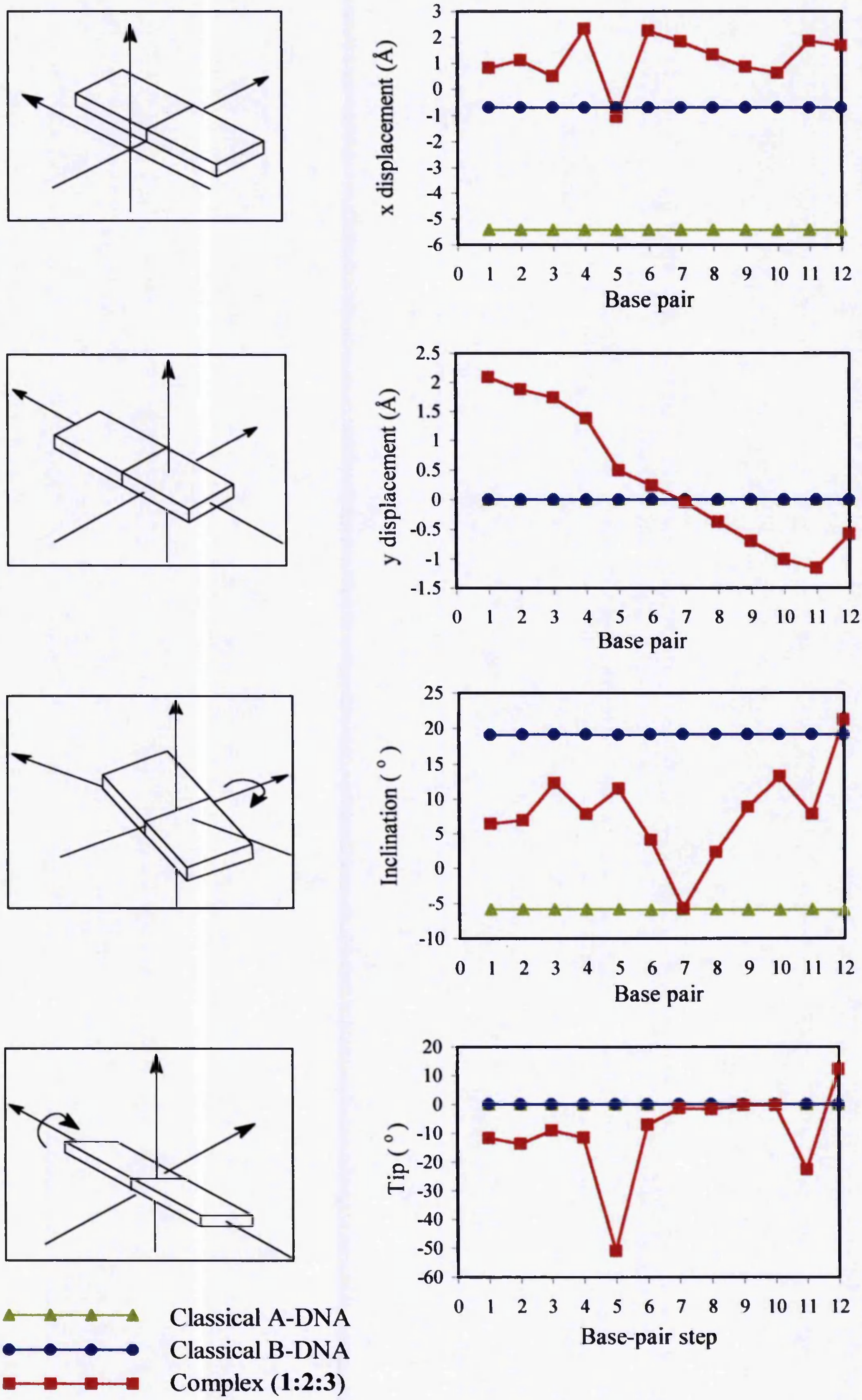


Figure 2.18(a). Axis-base pair parameters.

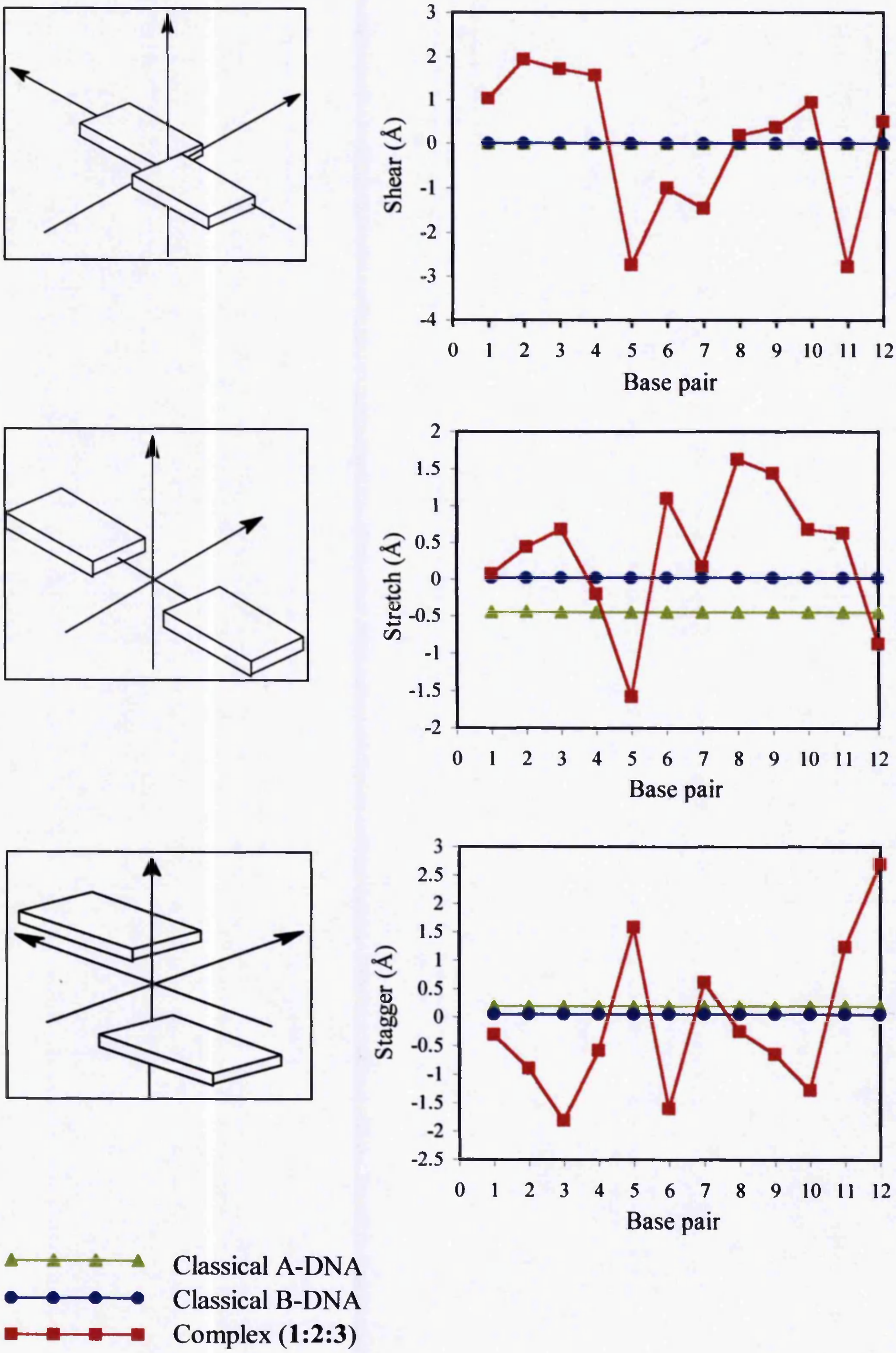


Figure 2.18(b). Intra-base pair parameters.

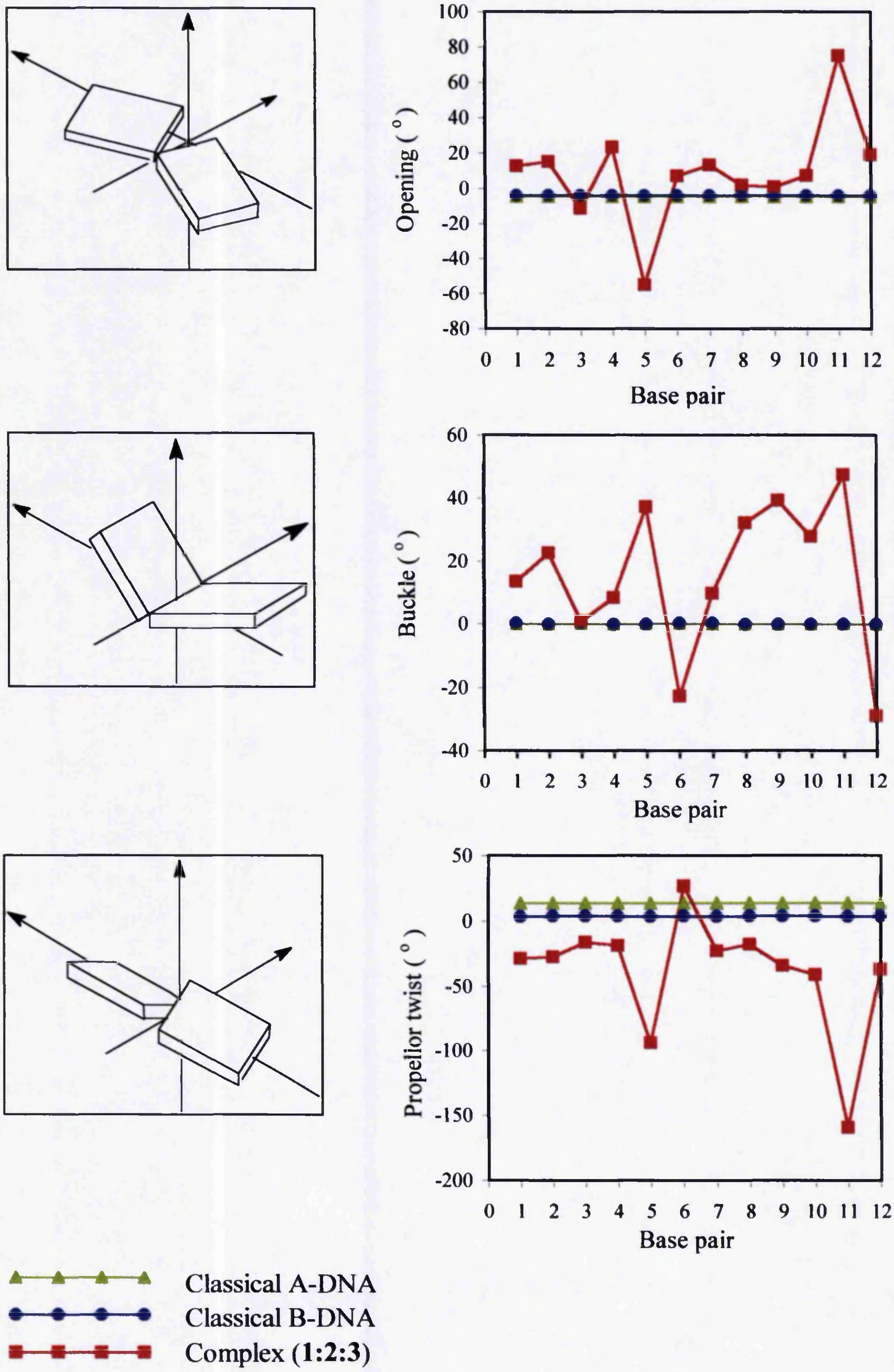


Figure 2.18(b). Intra-base pair parameters.

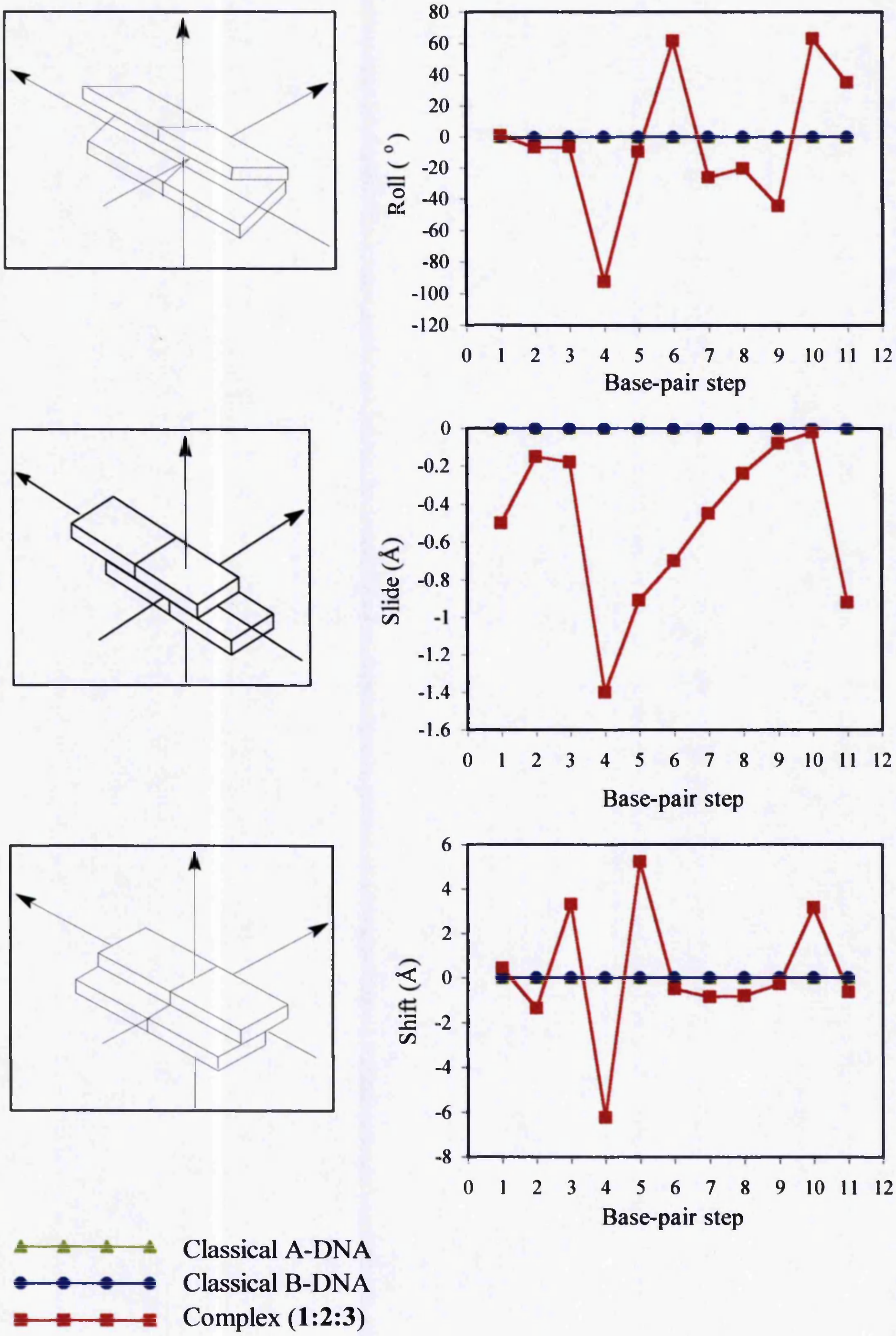


Figure 2.18(c). Inter-base pair parameters.

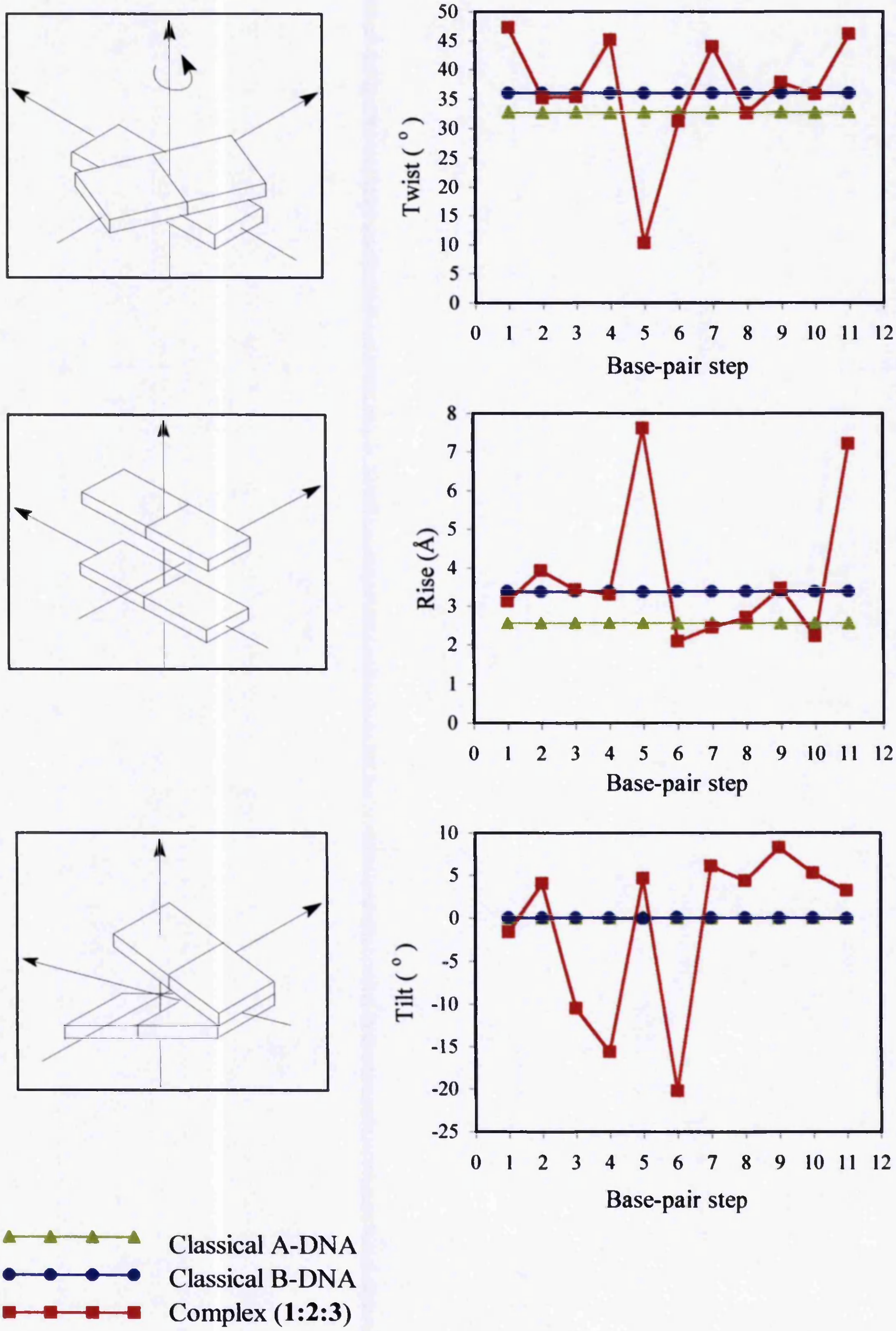


Figure 2.18(c). Inter-base pair parameters.

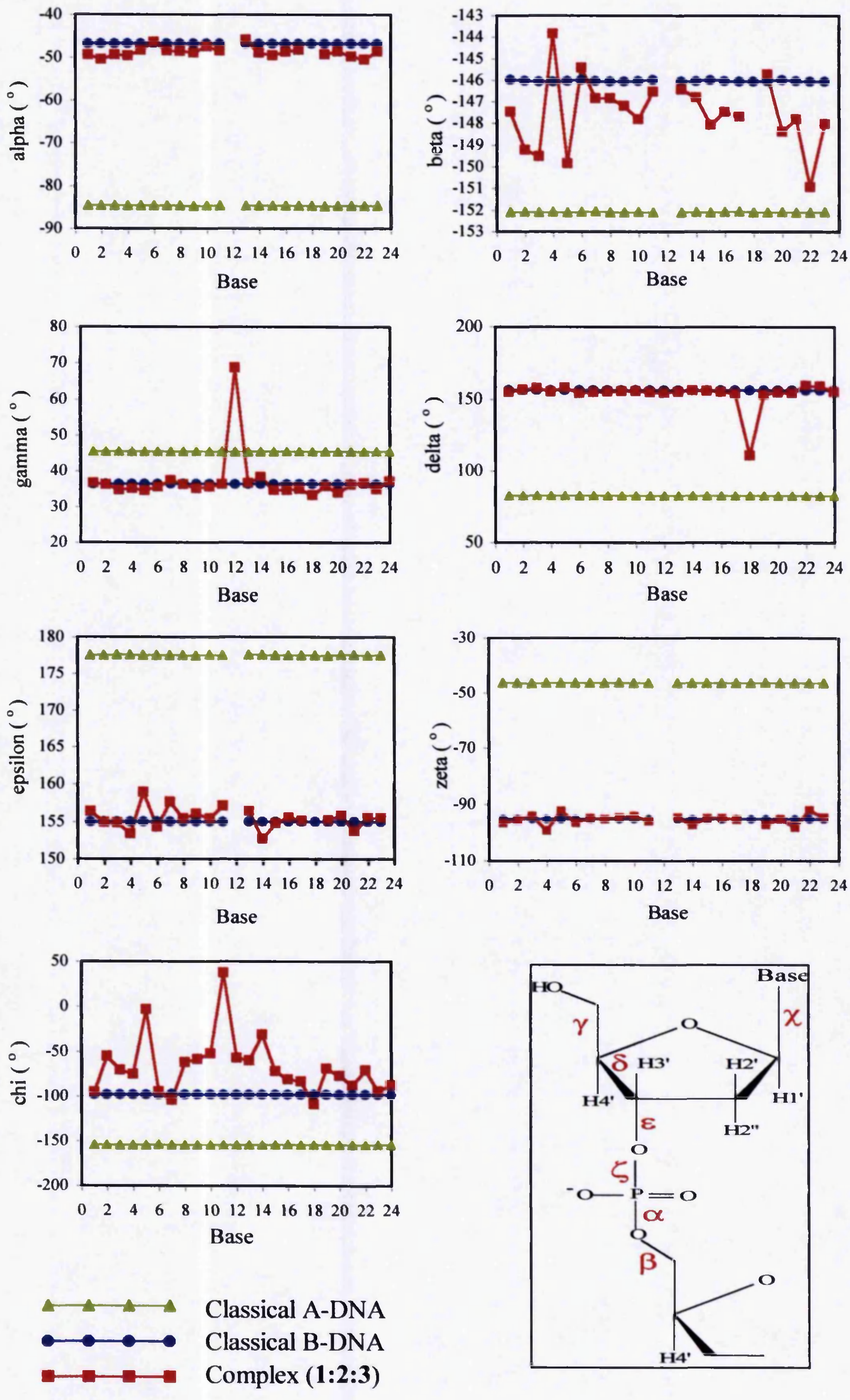


Figure 2.18(d). Torsion angle parameters.

2.4. OVERVIEW OF STRUCTURE DETERMINED FOR OLIGONUCLEOTIDE COMPLEX

The three-dimensional solution structure of an alkylimidazole-modified DNA binary complex with nucleic acid cleaving potential has been solved by NMR spectroscopy and restrained molecular dynamics, giving an essential insight into the spatial arrangement of all the components in this system. From the observations of the various experiments undertaken, it is apparent that neither the nick in the backbone between the binary components, nor the presence of the alkylimidazole cleaving groups significantly alter the regular helical structure from its regular B-DNA form. The imidazole groups do not form any special permanent contacts with the oligonucleotide part of the system and do not exist in any preferred conformation within the complex. These conclusions are justified from the following experimental observations.

From the recorded NOESY spectrum it is possible to sequentially assign aromatic, methyl and sugar ring protons to the DNA from the cross-peaks of interacting neighbouring residues, *i.e.* effectively 'walk' through the DNA sequence. The interactions observed indicate that there is no fraying at the terminii of the duplex and the regular, uninterrupted connectivities across the ^{18}C-^{19}T gap region again imply an undistorted, regular DNA structure. These observations are supported by the variable temperature NMR study of the imino protons. As the temperature increased, the melting of the DNA duplex proceeded in a manner expected for a standard continuous duplex, *i.e.* with denaturation initiating at the terminii of the duplex and proceeding sequentially towards the most stable central region. These studies gave no evidence to suggest that either the break in the backbone or the fact that two bases had been modified with alkylimidazole groups structurally distorted the helix in any way.

These results are in sharp contrast to those found for the similar binary system, with the same oligonucleotide composition, but bearing a photosensitising pyrene group at ^{19}T and a photoactivatable perfluoroazido group at ^{18}C [60]. In that case, a huge distortion in the centre of the helix was found. This can therefore be attributed not to the gap between the two hexamers, but to the presence of the strongly hydrophobic and bulky pyrene group. The structural model indicates that the pyrene moiety interacts strongly with the DNA, whereas the fluoroazide moiety does not [60] (Figure 2.19).

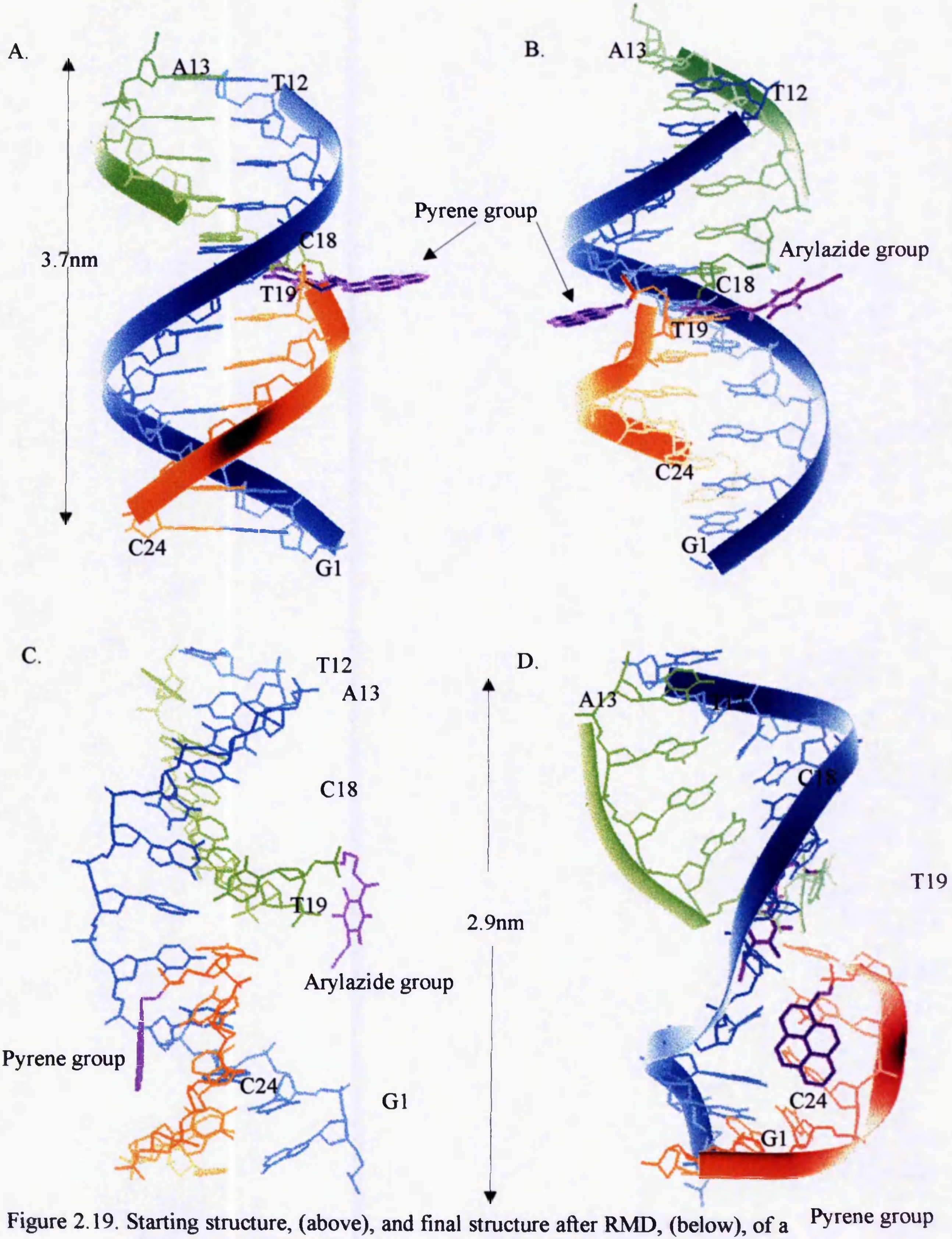


Figure 2.19. Starting structure, (above), and final structure after RMD, (below), of a similar oligonucleotide binary system but bearing photosensitising pyrene and photoactivatable perfluoroazide (arylazide) groups which result in a huge distortion of the helix [60]. (Kindly provided by Debbie Marks).

This observation is in accordance with thermal denaturation studies carried out monitored by UV-visible spectrophotometry. Melting temperature studies (carried out at concentrations of 4μM and 26μM) allowed comparison of the thermal stability of the binary system **1:2:3** with available data from other oligonucleotide systems, to determine the contribution of the nick in the DNA backbone and the presence of alkylimidazole modifying groups to the destabilisation (Drs. E.V. Bichenkova and M. I. Dobrikov, personal communication). The melting temperature obtained for the 4μM **1:2:3** duplex was found to be 37°C which is 8°C lower compared to the regular, unbroken 12-mer duplex d(CGAATTGTATGC)-d(GCATACAATTCG) investigated at the same concentration (unpublished data). This slight destabilisation effect observed for **1:2:3** can be explained by two reasons: the presence of the nick in the backbone and/or due to the alkylimidazole modifying groups in the middle of the structure.

The influence of the alkylimidazole groups on duplex stability was also assessed by comparison with the photoactivatable binary system mentioned [60]. The T_m of the photosensitizing binary system [60] and the alkylimidazole complex (**1:2:3**) at 26μM were found to be 18.3°C and 45.5°C respectively. It is seen from this data that the presence to the pyrene and azido groups greatly destabilise the binary system (by 27.2°C in comparison with **1:2:3**) which is in accordance with the pronounced structural distortion in the vicinity of the pyrene group found by 2-D NMR analysis. In contrast, the alkylimidazole groups do not appear to destabilize the duplex structure significantly, indirectly suggesting the absence of a strong structural perturbation for **1:2:3**.

The chemical shifts of the methylene protons of the alkylimidazole groups do not change significantly upon complex formation from the three components of the binary system. In contrast, the chemical shifts of the imidazole ring protons, H_α and H_β, are shifted to higher field after complex formation, presumably due to interaction with the oligonucleotide part of the system and/or the shielding effect of the nearest oligonucleotide residues. However, no cross-peaks were detected between **Im** or **Im'** and oligonucleotide protons. There are at least two possible explanations of this. Firstly, the imidazole construct may be exposed to solvent and not form any stable contacts with the oligonucleotide part of the system (but this is in disagreement with the observed shielding of the imidazole rings mentioned above). Secondly, the imidazole constructs have high conformational flexibility in the binary system, resulting in many rapidly exchanging contacts with the oligonucleotide residues. This explanation was confirmed by the CPMG T_2 experiment, the findings of which suggest that the imidazole groups have high conformational flexibility in the binary system, resulting in many rapidly exchanging contacts with the nucleotide residues. This result implies that the RNA cleavage within the parent alkylimidazole binary RNA system (V. V. Vlassov, unpublished data) is not dominated by the formation of a specific stable construction but rather by random encounter of the imidazoles to temporarily form the active intermediate. In turn, this may explain the low cleavage efficiency and low reproducibility of similar RNA-based binary systems (V. V. Vlassov, unpublished data). Because of the lack of distortion to the DNA helix, these findings can be approximated for the real binary system with a target RNA molecule and used for design of more stable reactive constructs for further improvement of cleaving ability.

In the future it may be possible to design DNA-cleaving systems based on a similar idea to the binary system discussed. At present, DNA damage is usually conferred by radical means or metal-containing complexes. This work provides evidence that modifying groups such as the alkylimidazole constructs used do not distort the DNA and it may be possible to replace these groups with other reactive moieties and provide an alternative approach to the strategy of nucleic acid modification.

CHAPTER THREE

COMPARATIVE MODELLING OF THE TRANSCRIPTION FACTOR HYPOXIA-INDUCIBLE FACTOR-1 WITH THE DESIGN AND SYNTHESIS OF POSSIBLE PEPTIDE INHIBITORS

3.1. INTRODUCTION

3.1.1. Homology modelling

3.1.1.1. Protein structure prediction

X-ray crystallography and NMR spectroscopy are the two methods used to provide detailed information about protein structures. Unfortunately, in spite of the tremendous increase in rate at which protein sequences are being determined, there is still an enormous gap between the numbers of known DNA-derived sequences and the numbers of three-dimensional structures. There is thus considerable interest in theoretical methods for predicting the three-dimensional structures of proteins from the amino acid sequence, an important, yet seemingly unattainable, goal in structural molecular biology. Protein models are very important when structure predictions are used in real world applications such as structure-based drug design or studies of mechanisms of action.

3.1.1.1.1. Predicting protein structure from first principles

The most ambitious approaches to the folding problem attempt to solve it from first principles (*ab initio*). The problem is to explore the conformational space of the molecule in order to identify the most appropriate structure. This is very much easier for peptides than proteins simply due to the very large number of possible conformations of the latter. It is therefore common to use some form of simplified model of the protein to make the problem more tractable.

Finkelstein and Reva [131] have used a simplified lattice representation of protein structure. Simple lattice models can be used to try and answer some of the fundamental questions about protein structure. For example, it may be possible to enumerate all possible conformations for a chain of a given length on the lattice and to investigate the relationship between the structure and the sequence. Chan and Dill [132] modelled a protein as a series of hydrophobic and hydrophilic monomers. The sequence is grown on a two-dimensional lattice using a self-avoiding walk and the energy of the resulting conformation calculated by summing interactions between pairs of monomers that occupy adjacent lattice sites, but are not covalently bonded. It is usually not possible to exhaustively explore the conformational space on the lattice and so methods such as Monte Carlo simulated annealing are used to generate low-energy structure(s).

Ab initio methods have been used to predict the tertiary structures of globular proteins, for example, the 'island model' of Kobayashi *et al.* [133]. This is based on the physicochemical mechanisms of folding. An island is a local structure, such as a secondary structural element formed during folding. The amino acid sequence is assigned secondary structure using any prediction method (see later) and all the side chains approximated by spheres of Van der Waals' radii. Bond lengths and angles are fixed at standard values taken from proteins in the Brookhaven Protein Data Bank [134]. The energy calculation considers the hydrophobic interactions between hydrophobic side-chains and the Lennard-Jones potentials between the atoms (or the atom groups of the side-chains represented by the spheres). The former is a driving force of folding, while the latter is mainly for avoiding collapse of the chain. The secondary structures are folded and packed through the hydrophobic binding by searching for the lowest energy conformation. Several islands pack stepwise into bigger ones to eventually reach a compact tertiary structure. The parameters of the energy function used represent the range of interactions (residue interval along the chain), but the distribution of hydrophobic-residue pairs dominates the folding process, especially its order. The conformation is refined by generating the side-chain atoms and introducing energies previously ignored, e.g. electrostatic potential energies among non-bonded atoms, followed by a full minimisation.

Ab initio modelling has proved successful in predicting the tertiary structure of NK-lysin by assembling recognised secondary structural fragments [135] and in modelling the structural details of the activation of the α_{1b}-adrenergic receptor [136]. The *ab initio* tertiary folds of proteins with different topologies have also been predicted using genetic algorithms [137,138], which can generate specific folds, such as four helical bundles. Genetic algorithms simulate evolution to solve complicated optimisation tasks and protein-folding principles can be mimicked by simple selection criteria (see below). The protein main-chain is modelled by taking ψ and ϕ backbone torsion values from a set of standard conformations. The conformations of all residues are encoded by a long bit string (chromosome). Starting from a population of random bit strings, 'offspring' are generated by chromosome mutations and recombinations. The quality of each resulting structure is judged by fitness functions (distribution of hydrophobic residues, Van der Waals' overlaps etc).

3.1.1.1.2. Secondary structure prediction using rule-based approaches

Most protein structures contain a significant amount of secondary structure (α-helices and β-sheets) or random coil. An obvious way to tackle the problem of predicting a protein's three-dimensional structure is first to determine which stretches of amino acids should adopt each type of secondary structure, and then dock these secondary structural elements together. The one-dimensional prediction of the conformational state of each residue is an old problem, yet remains of current interest, as is clear from the amount of activity in the literature (reviewed in [139]. Secondary structure prediction is important in helping to establish sequence alignments. The many methods for predicting the secondary structural motifs from the primary sequence can be roughly divided into statistical, knowledge-based and hybrid systems.

Statistical methods are based on large-scale studies of the databases of proteins of known primary and secondary structure, aimed at finding empirical relationships between sequence and secondary fold. Examples of these methods include Chou and Fasman [140], GOR (Garnier-Osguthorpe-Robson) [141,142], nearest neighbour [143] and neural networks [144-148]. The Chou-Fasman method is a statistical prediction method based on the calculation of the statistical propensities of each residue forming either an α-helix or a β-strand. The theory behind the GOR method is complex, but is based on the idea of treating the sequences of primary and secondary structure as two messages related by a translation process, regarded as a black box defined only by the relation between input and output symbols. The structure prediction process depends on measuring the amount of information residues carry about their secondary structure and other residues secondary structure.

In nearest neighbour statistical methods, the secondary structure of a new primary sequence is classified to be the same as that of the closest primary sequence to it of known secondary structure, based on the hypothesis that similar primary sequences will have similar secondary structures [143]. However, this hypothesis is often untrue. This approach has been improved upon by allowing segment length to be variable, thus allowing the algorithm to take advantage of sequence similarity that stretches beyond the small fixed length alignments of previous algorithms [149]. Neural networks mimic computational pattern recognition by neurons in living organisms. The learning system is a network of non-linear processing units that have adjustable connection strengths. The connections are not pre-programmed and learning consists of altering the weights of connections between the units in response to a teaching signal, *i.e.* amino acid sequence, which provides information about the correct classification in output terms (primary and secondary structure) [144,146,147].

Such statistical methods have the disadvantages of not fully taking into account physicochemical information known about proteins, *i.e.* knowledge of the physical and chemical basis of protein structure, and in having poor explanatory power. Lim [150] overcame this shortcoming by developing a theory of formation of α-helices and β-sheets based on the packing of polypeptide chains in native globular proteins. This theory was used to produce stereochemical rules for predicting secondary motifs based on a reasoned stereochemical theory of globular protein structure and was successfully tested on 25 proteins. The prediction method of Cohen and co-workers [151] is based on using a rule-based formalism to encode structural knowledge about proteins. The rules are lists of generalised amino acid sequences or patterns that are associated through physical and chemical theory with specific secondary structures.

Several important structural features have been recognised to be useful in a general way in predicting protein secondary structure and are commonly included in protein sequence analysis program suites, e.g. helical wheels and hydrophobicity profiles. They mainly serve to justify alignments.

SAPIENS (Secondary structure and Accessibility class Prediction Including ENvironmental-dependent Substitution tables) is a secondary structure prediction method using environment-dependent tables based on the probabilities of amino-acid substitution and conformational propensities [152]. Short-to-long range interactions can be introduced because the substitution patterns of individual amino acid residues are restricted depending on their conformational states and the local structural environment. Solvent-accessibility of the residues is also addressed.

3.1.1.1.3. Tertiary structure prediction: fold recognition

Rules exist which govern the arrangement of secondary structures into their globular folds and this information can be used to generate and improve methods for tertiary structure prediction. These methods have arisen from the observation that two structures may have very similar folds despite lacking any statistically significant sequence similarity. Although some of these protein pairs have similar structures and functions and are probably related by evolution, many show no sequence relationship. This had led to the suggestion that there may be a limited number of possible topologies or folds [153,154] and therefore a sensible approach to predicting a structure is to ask if the sequence could adopt one of the currently known set of protein folds. 'Fold recognition' attempts to detect such relationships, even of cases of sequences that share the same structure, not through divergence from a common ancestor (homologous structures), but convergence to a common structure (analogous structures). Different fold recognition strategies are required for the two scenarios, with remote homologues generally showing a much stronger degree of similarity in sequence and secondary structure and so fold recognition algorithms are generally more successful in these cases. The problem of recognition of analogy is still largely unsolved, probably because it involves the recognition of a much weaker relationship.

Most algorithms use a secondary structure prediction for the sequence as part of the input (reviewed [139]). Russell *et al.*, [155] have used substitution matrices and Bowie *et al.* [156] have attempted to match sequences to folds by describing the fold in terms of solvent accessibility (*i.e.* buried, partially buried and exposed), the degree of burial by polar rather than apolar residues and the environment of each residue located in the structure (*i.e.* α, β and coil). The environment of a particular residue thus defined tends to be more highly conserved than the identity of the residue itself, and so the method is able to detect more distant sequence-structure relationships than purely sequence-based methods. The inaccuracy of the secondary structure predictions affect fold recognition and the methods perform better when actual secondary structure is used to replace the predicted. *Ab initio* calculations and the role of genetic algorithms in the folding of secondary structural elements has already been mentioned [137,157]. Protein folding simulations of small proteins with different topologies investigated using a genetic algorithm based on sequence and experimental and predicted secondary structural knowledge were quite successful [158].

3.1.1.1.4. Predicting protein structures by 'threading'

Despite the obvious computational advantages of using residue environments, it is clear that the fold of a protein chain is governed by fairly specific protein-protein and protein-solvent atomic interactions and protein structure prediction efforts are more likely to be successful if full use is made of all available information. A given protein fold is therefore better modelled in terms of a 'network' of pair-wise inter-atomic energy terms, with the structural role of any given residue described in terms of its interactions. Classifying such a set of interactions into one environmental class such as 'buried α-helical' will inevitably result in the loss of useful information, reducing the specificity of sequence-structure matches evaluated in this way. A dynamic programming algorithm [159] has been applied to the problem of aligning a given sequence with the 'real' co-ordinates of a structure, taking into account the detailed pair-wise interactions, known as 'optimal sequence threading'. The sequence of a protein whose structure is to be determined, is 'threaded' through each known protein structure in turn. The amino acids are advanced to occupy the location occupied in the previous iteration by its predecessor and a score is calculated for each structure that is generated. The results are the structures with the lowest scoring function. Threading is increasingly being used to suggest the structures of proteins [159,160].

Rykunov *et al.* [161] studied sequence-based recognition of protein folds using the threading method and tests on 25 proteins of different structural classes showed that the native fold was often recognised as the best (most stable) fold.

3.1.1.2. Comparative modelling methods

Historically, the most successful techniques of protein structure prediction have been those based on inference from evolution. If a sequence can be shown to be sufficiently similar to another sequence of known structure, then the implied evolutionary relationship will guarantee structural similarity. There are striking similarities between the sequences and three-dimensional structures of some proteins. Comparative or homology modelling exploits these similarities by constructing a structure using the known structure of another protein as a template, after deciding the best sequence alignment of target and template.

If the biological function of the protein is known, it is often relatively straightforward to decide which protein(s) would offer a preferred template. In other cases, the function of the protein may not be known but it may be possible to deduce to which family it belongs by searching a sequence database for the presence of particular combinations of amino acids (motifs) that often imply a particular function or structural feature. The template is chosen as the protein whose sequence is the closest match for the unknown protein. The stages in comparative modelling are discussed below.

3.1.1.2.1. Protein sequence alignment

The sequence of the protein with the unknown structure is used to search protein databases to identify other proteins most closely matched in sequence. The aligning process is the most important stage in homology modelling. Incorrect alignments are the source of the most serious errors so a variety of approaches and scoring schemes should be used.

The object of a 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. Gaps in the aligned sequences correspond to regions where loops are inserted or deleted. Sequence-similarity measures generally can be classified as either global or local. Global similarity algorithms optimise the overall alignment of two sequences, which may include large stretches of low similarity [162]. Local similarity algorithms seek only relatively conserved subsequences and a single comparison may yield several distinct sub-sequence alignments; unconserved regions do not contribute to the measure of similarity [163,164].

Consider the problem of aligning just two sequences. Any alignment algorithm requires some method for 'scoring' an arbitrary alignment of the two sequences. The object is to find the alignment giving the 'best' score. Many similarity measures begin with a matrix of similarity scores for all possible pairs of residues. The simplest scoring method is *sequence identity* which gives the percentage of amino acids that are the same in the two sequences; identical pairs score 1 and all others score 0. An alternative approach is to recognise that topologically equivalent residues in two structurally homologous proteins may not be identical, but may have very similar shapes, electronic, hydrogen-bonding and hydrophobic properties. Such 'conservative' substitutions can often be made with little disruption to the three-dimensional structure of the protein and this is taken into account in the scoring scheme. Identities and conservative replacements have positive scores, while unlikely replacements have negative scores.

When there are more than two sequences to be aligned, simultaneous comparison of all the sequences cannot, in practice, be carried out since the number of segment comparisons that must be executed is of the order of the product of the sequence lengths. An alternative approach is to select the best pairwise alignments from the scores of all pairwise comparisons. Several multiple sequence alignment programs use this technique, building the final alignment by gradually aligning further sequences, according to the basic Needleman-Wunsch procedure [162]. This method attempts to maximally match sequences by finding the largest number of amino acids of one protein that can be matched with those of a second protein allowing for all possible interruptions in either of the sequences. All possible pairs of amino acids (one from each protein) are represented by a two-dimensional array and all possible comparisons are represented by pathways through the array. The maximum match is the largest number that would result from summing the cell values of every pathway. Based upon the scores of the initial alignments of all pairs of sequences, different strategies are used to determine in what order the sequences are incorporated into the final alignment, e.g. either in a sequential procedure or using clustering.

Dayhoff and co-workers analysed substitution frequencies in gapped global alignments of closely related proteins and have published tables which give the probability of mutating one amino acid to another [165]. These probabilities are usually scored as PAM matrices (Percentage of Acceptable point Mutations per 10^8 years). The number refers to evolutionary distance. Low values should be used for looking for related proteins, high values for the more distant cousins. Most often used are PAM120 and PAM250.

An approach has been developed by Henikoff and Henikoff using ungapped local multiple alignments of short regions of related sequences [166]. Substitution frequencies for all pairs of amino acids are calculated and the result used to calculate a BLOSUM (block substitution matrix) matrix and different matrices are obtained by varying the clustering threshold. The number refers to the lowest allowed sequence similarity within the alignments used. BLOSUM50 is based on alignments of less similar sequences and might be used to find more distant relations. BLOSUM62 is generally considered to be the optimal matrix of this series, which usually perform better than the PAM matrices.

BLAST (Basic local alignment tool) [167] was developed to speed up a database scan by introducing approximations. It is an algorithm to find the highest scoring locally optimal alignment and uses a maximal segment pair (MSP), which is defined as the highest scoring pair of identical length segments chosen from two sequences. The boundaries of an MSP are chosen to maximise its score, so an MSP may be of any length and a segment pair is locally maximal if its score cannot be improved either by extending or shortening both segments. The MSP score for two sequences may be computed in time proportional to the product of their lengths, but can be estimated under an appropriate random sequence model [168]. For any particular scoring matrix one can estimate the frequencies of paired residues in maximal segments too. These approximations make BLAST a quick and easy algorithm to use.

The Multalin algorithm [169] is based on the conventional dynamic-programming method of pair-wise alignment [170] using clustering to determine in what order the sequences are incorporated into the final alignment. Initially a hierarchical clustering of the sequences is performed using the matrix of the pair-wise aligned scores. The closest sequences are aligned, creating groups of aligned sequences. Close groups are then aligned until all sequences are aligned in one group. The pair-wise alignments included in the multiple alignment form a new matrix that is used to produce a hierarchical clustering. If it is different from the first one, iteration of the process can be performed.

Many multiple protein sequence alignment servers are available on the Internet. For example;

BLAST	http://www.SEQNET.dl.ac.uk
Block Maker	http://www.blocks.fhcrc.org/blockmkr/make_blocks.html

ClustalW	http://www2.ebi.ac.uk/clustalw/
Match-Box	http://www.fundp.ac.be./sciences/biologie/bms/matchbox_submit.html
MAP	http://dot.imgen.bcm.tmc.edu:9331multi-align/options/map.html
MEME	http://www.sdsc.edc/MEME/meme1.4/meme.nofeedback.html
MSA	http://alfredo.wustl.edu/msa.html
PIMA	http://dot.imgen.bcm.tmc.edu:9331multi-align/options/pima.html
Multalin	http://www.toulouse.inra.fr/multalin.html

A recent comparative analysis of seven of them evaluated them in terms of power (sensitivity) and confidence (selectivity) (ClustalW, MAP, PIMA, Block Maker, MSA, MEME and Match-Box)[171]. Results showed that any powerful method remains reliable even when the rate of identity falls. All the programs differ in terms of the emphasis they place on power and confidence so one cannot be said to be better than another: the user must select the most suitable technique according to their requirements of selectivity and sensitivity. It is a good idea for the user to collect outputs from different methods to improve the quality of the predictions by taking into account the consensus-predicted structurally conserved regions. Furthermore, any other kind of information should be referred to, such as biochemical experimental evidence or site-directed mutagenesis results, in order to validate predicted structurally conserved regions aligned by only one method.

As well as the existence of different alignment programs, there are many different protein databases available which to search. A comprehensive list can be found at the web site http://www.SEQNET.dl.ac.uk. Included here are databases of nearly every protein sequenced, proteins with known three-dimensional structures, transcription factors, gene sequences, and expressed sequence tags (ESTs). Again, it is up to the user to decide which to use, but it is generally not wise to restrict the search to one database, as not all contain the same protein sequences and some are better managed and more frequently updated that others.

3.1.1.2.2. Constructing a comparative model

After sequence alignment, one or more proteins are chosen to serve as template structures from which to build the model. Where sequence identity between the probe sequence and the sequence match from the database is high (>65-70%) comparative modelling is highly successful and greater confidence can be placed in the final model. A level of 40-50% identity will give knowledge about the overall protein fold, but no confidence can be placed in a model based on a sequence identity of 25-30% or less.

Automated and rule-based model-building procedures have been developed in recent years in order to minimise subjective manual decisions. Such modelling techniques fall into two classes:

(i) the assembly of rigid fragments from homologous and other proteins of known structure

(ii) the use of restraints such as inter-atomic distances to construct models that have the best agreement with homologues of known structure.

3.1.1.2.3. Fragment based modelling

Many approaches depend on the assembly of rigid fragments from known protein structures. Local main-chain and side-chain conformers from equivalent fragments in known homologous structures are extrapolated to the sequence of the unknown. Jones and Thirup [172] were the first to demonstrate that a protein structure can be built from a combination of segments from other proteins. The comparative modelling software COMPOSER, developed originally by Blundell and co-workers [173,174] and incorporated in the software SYBYL (Tripos Inc.), also depends on the assembly of rigid fragments.

Construction of the protein core is often relatively straightforward. In some cases, the backbone conformation can simply be transferred from the template homologue to the unknown protein. The next task is to determine the conformations of the loop regions, (*i.e.* regions of no secondary structure), such that they have a low internal energy and do not possess any unfavourable interactions with the rest of the protein. In certain cases, the loops may be restricted to a set of canonical structures. For example, the loops that connect certain types of secondary structure show distinct conformations, as seen in the β-turns that connect strands of β-sheets. Other cases require alternative methods. For loops that contain fewer than seven rotatable bonds, an algorithm devised by Go and Scheraga [175] can be used to calculate possible loop geometries directly. Using a model with fixed bond lengths and bond angles, they showed that it was possible to determine the torsion angles that would permit the end-to-end distance of the loop to achieve the desired value. More recent variants permit the bond angles to deviate slightly from their equilibrium values and so have a higher chance of finding an acceptable match [176]. Pure systematic searches can also be used to generate loop conformations. One way to alleviate the combinatorial explosion is to construct the loop from both ends simultaneously; the half-complete loops are then joined in the middle and energy minimised.

Loop conformations can be obtained by searching the Brookhaven protein databank [134] for stretches of polypeptide chain that contain the appropriate number of amino acids and also have the correct spatial relationship between the two ends [172]. Relatively short loops (approximately 10 amino acids long) can be manually built and then minimised in the context of the overall protein's structure. Longer loops pose greater problems due to their flexibility.

Once a backbone conformation has been derived for the protein, including loop regions, it is necessary to assign conformations to the side chains. The dependence of side-chain conformation on main-chain structure is now well recognised. The assessment of accuracy for side-chin placement is complex, depending upon accessibility to solvent, main-chain accuracies, resolution of structure and, in comparative modelling, the similarity of the parent structure and hence the environment will be the overriding factor. In the core region there may be a high degree of sequence identity between the unknown protein and the template, and the side-chain conformations can often be transferred directly from the template. Where there is less correspondence between the amino acid sequences, a variety of systematic and random methods can be used to predict side-chain conformations, e.g. Monte Carlo, simulated annealing and genetic algorithm methods, or *ab initio*. The concept of side-chain rotamers (that side-chains predominantly adopt a limited set of torsion angle combinations) has been extremely useful in both experimental determination and in modelling of protein structures. Rotamer libraries are obtained by collecting the frequently occurring torsion combinations for each residue type, from a database of well-determined structures [177]. Loop and side-chain modelling has been reviewed [139].

3.1.1.2.4. Restraint-based modelling

Although comparative modelling using assemblies of fragments has proved successful, procedures for modelling by satisfaction of distance or other restraints may also be used to advantage, especially where the homologues are distantly related. Such procedures derive restraints, such as inter-atomic distances involving main-chain and side-chain atoms, from homologous protein structures. Using distance geometry, an ensemble of models satisfying input restraints is built. Havel and Snow [178] proposed a method which derives distance and chirality restraints based on alignment with homologous known structure(s). Sali and Blundell [179] have developed a comparative modelling procedure (MODELLER) which arrives at a three-dimensional model by optimally satisfying restraints extrapolated from known structures homologous to the model sequence.

3.1.1.3. Modelling on the Internet

Protein structure prediction and comparative modelling are today carried out using the Internet. Numerous web sites exist which are very easily accessible, user friendly and contain all the databases, algorithms and tools required when building a homology model. Individual alignment programs and sites of protein databases are available or sites that are devoted to the whole process of comparative modelling, from sequence alignment to model validation. Two examples are briefly described below.

3.1.1.3.1.The PredictProtein (PP) Server,[147] (http://www.embl-heidelgberg.de/predictprotein/).

The PredictProtein server is an automatic service for protein database searches and the prediction of aspects of protein structure. It provides a single interface to many modelling tools (written in parentheses). An amino acid sequence of interest is submitted and PP returns:

1. a multiple sequence alignment following a BLASTP search of the SWISS-PROT database (MaxHom, a weighted dynamic programming method)
2. detection of functional motifs (PROSITE)
3. detection and assignment of protein domains (PRODOM)
4. predictions using a system of neural networks of:

(i) secondary structure (PHDsec)
(ii) residue solvent accessibility (PHDacc)
(iii) transmembrane helix location and topology (PHDhtm, PHDtopology)
(iv) protein globularity (GLOBE)
(v) a calculation of the probability of the presence of coiled-coil conformations (COILS)

Remote homologues (0-25% identity) are detected by prediction-based threading (TOPITS) which detects similar motifs of secondary structure and accessibility between a sequence of unknown structure and fold by aligning against the Brookhaven Protein Database [134].

The server can also provide evaluation of secondary structure prediction accuracy (EVALSEC).

Full explanations and details of each of the independent modelling tools used by PredictProtein can be found at the website.

3.1.1.3.2. The SWISS-MODEL automated protein modelling server, (http://expasy.hcuge.ch/swissmod/SWISS-MODEL.html).

Swiss-Model is a comparative modelling server covering the stages of sequence alignment, model building, energy minimisation and evaluation using the following procedure:

1. Search of the NRL_3D database for suitable templates using the sequence alignment algorithms BLAST, SIM and ProModII and check the sequence identity with the target.
2. Generate comparative models using the ProModII tool.
3. Energy minimisation using the Gromos96 force field.
4. Model evaluation using Swiss-PdbViewer.

Full references and details can be found at the web site.

3.1.1.3.3. Structural genomics web sites

Below are lists of other sites on the Internet useful in protein modelling containing information on many sequence alignment algorithms, secondary and tertiary structure prediction methods and classifications of protein families and topologies etc.

Databases

PDB: http://pdb.bnl.gov (Brookhaven Data Bank)

NCBI: http://www.ncbi.nlm.nih.gov (National Center for Biotechnology Information) is a comprehensive site that includes facilities for searching and analysing protein structure and sequence databases.

SCOP: http://mrclmb.cam.ac.uk/scop is a database of protein structures arranged in a hierarchy according to structural and evolutionary relatedness

CATH: http://www.biochem.ucl.ac.uk/bsm/cath/ is a classification of protein domain structures, which clusters proteins at four major levels: Class, Architecture Topology and Homologous superfamily

TOPS: http://www3.ebi.ac.uk/tops/ is a site for searching for and determining protein topologies.

DALI: http://www2.ebi.ac.uk/dali/ is a service for identifying similarities between different protein structures in the Protein Data Bank.

PRESAGE: http://presage.stanford.edu

MSD: http://www.rcsg.org/

SEQNET: http://www.SEQNET.dl.ac.uk (Daresbury Laboratories)

ModBase: http://guitar.rockefeller.edu/modbase/

GeneCensus: http://bioinfo.mbb.yale.edu/genome

Fold assignment

PhD: http://www.embl-heidelgberg.de/predictprotein/predictprotein.html

THREADER: http://globin.bio.warwick.ac.uk/"jones/threader.html

123D: http://www-lmmb.ncifcrf.gov/"nicka/123D.html

UCLA-DOE: http://www.doe-mbi.ucla.edu/people/frsvr/frsvr.html

PROFIT: http://lore.came.sbd.ac.at/

Comparative modelling

COMPOSER: http://www-cryst.bioc.cam.ac.uk/

CONGEN: http://www.cabm.rutgers.edu/"bruc

DRAGON: http://www.nimr.mrc.ac.uk/"mathbio/a-aszodi/dragon.html

Modeller: http://guitar.rockefeller.edu/modeller/modeller.html

PrISM: http://honiglab.cpmc.columbia.edu/

SWISS-MODEL: http://expasy.hcuge.ch/swissmod/SWISS-MODEL.html

WHAT IF: http://sander.embl-heidelberg.de/vriend/

Miscellaneous

FSSP: http://www2.ebi.ac.uk/dali/fssp/ (Fold classification based on Structure-Structure alignment of Proteins). This database is based on an exhaustive all-against-all 3-D structure comparison of protein structure currently in the Protein Data Bank. Similar structure are identified and superimposed.

ExPASy: http://www.expasy.ch/ (Molecular Biology Server, Swiss Institute for Bioinformatics). This is a server for the analysis of protein sequences and structures and provides many links to other molecular biology databases, servers, tools and software packages.

Pfam: http://www.sanger.ac.uk/Software/Pfam/ (Protein domain family searches) is a database of protein families and domains and contains multiple protein sequence alignments and profile-Hidden Markov Models of these families. Pfam is a semi-automatic protein family database.

Pedro's BioMolecular Research Tools: http://www.fmi.ch/biology/research_tools.html. An extensive collection of links to sites of interest to molecular and structural biologists e.g. databases, servers, alignment programs and structure prediction sites.

Roberto's List: http:/guitar.rockefeller.edu/"roberto/tools/tools.html

European Bioinformatics Institute: http://www.ebi.ac.uk provides links to other databases and servers.

No one algorithm or software package should be trusted to give definitive answers. Different computational modelling tools give different results and one should not be used exclusively over any other. They are there to be used in combination and should be thought of as complementary techniques.

3.1.1.4. Energy minimisation, molecular dynamics and model evaluation

Once a model has been constructed, energy minimisation is necessary in order to relieve any short contacts and to rectify bad geometry that may be present. This problem is particularly possible in the anchor regions where a loop is melded with the core of the protein. Before starting energy minimisation it is important to examine the model for serious flaws, (e.g. stereochemistry, unlikely distribution of amino acids (*i.e.* hydrophilic residues in the inner core and hydrophobic in the outer regions), non-planar amide bonds, steric conflicts between non-bonded atoms).

Force fields such as CHARMM [180], AMBER [181] or TRIPOS (SYBYL software, Tripos Inc.) are used. If major problems persist in a local region of the model, simulated annealing can be focused here. Subsequently, the whole model can be considered for minimisation. Initially, the electrostatic term in the force field need not be considered, as the problems with the models are expected to be due to short contacts and bad geometry alone and omitting these terms will speed up the process. When the structure to be modelled is close to a homologue of known structure, the backbone can be fixed and side-chains alone can be allowed to relax. Initially a fast optimisation procedure such as Simplex can be used followed, for example, by Powell gradient minimisation to optimise further the system to convergence. When most problems with the stereochemistry have been rectified, the electrostatic term can be invoked and the geometry of the hydrogen-bonds in the structure will be optimised. When the side-chain positions are refined, all the atoms can move freely and the minimisation terminated once all the inconsistencies in the geometry are rectified and short contacts relieved. It is still important to make a final check on the model on the factors mentioned above, even after minimisation.

A simple test is to generate a Ramachandran plot [182] in order to determine whether the amino acid residues occupy the energetically favoured conformations. The side-chains can also be examined to identify any significant deviations from those commonly observed in X-ray structures. Eisenberg's '3-D-profiles' method [156,183] calculates three properties for amino acids on the proposed structure: the local secondary structure, the total surface area of the residue that is buried in the protein and the fraction of the side-chain area that is covered by polar atoms. These parameters are used to allocate the amino acid to an environment class and the residue is given a score that reflects the compatibility of that amino acid for that environment, based upon a statistical analysis of known protein structures.

Facilities are available on the Internet to validate the final model structure. The Predict Protein and Swiss-Model servers have already been mentioned. In addition, the Biotech Validation Suite for Protein Structures can be found at http://biotech.embl-heidelberg.de:8400/. This service provides comprehensive quality checks of protein structures using three software packages:

1. PROCHECK which performs a full geometric analysis of main-chain and side-chain parameters, e.g. bond angles, RMS distances from planarity, Ramachandran plots.
2. PROVE which gives information on surface area measurements and atomic volume analyses.
3. WHAT_IF which analyses e.g. torsion angles, amino acid contacts, chirality, proline puckering, isolated water clusters and buried unsatisfied hydrogen-bond donors and acceptors.

In an ideal situation, a protein homology model should be tested experimentally, e.g. by mutagenesis studies. This will increase confidence in the model, but these experiments may not always be possible.

Regular science symposia are held to assess the current computational tools available in the subject area of protein structure prediction (the so-called CASP meetings (Critical Assessment of Structure Prediction)). Delegates are invited to submit structure prediction studies and the algorithms, software packages and techniques used are assessed in several categories, e.g. sequence alignment algorithms, secondary structure prediction, fold recognition and protein-ligand docking programs. The results of these studies are very informative with regards to the reliability of current computational modelling techniques and in vital to future developments.

Figure 3.1 shows a summary of the steps involved in the process of comparative modelling of proteins.

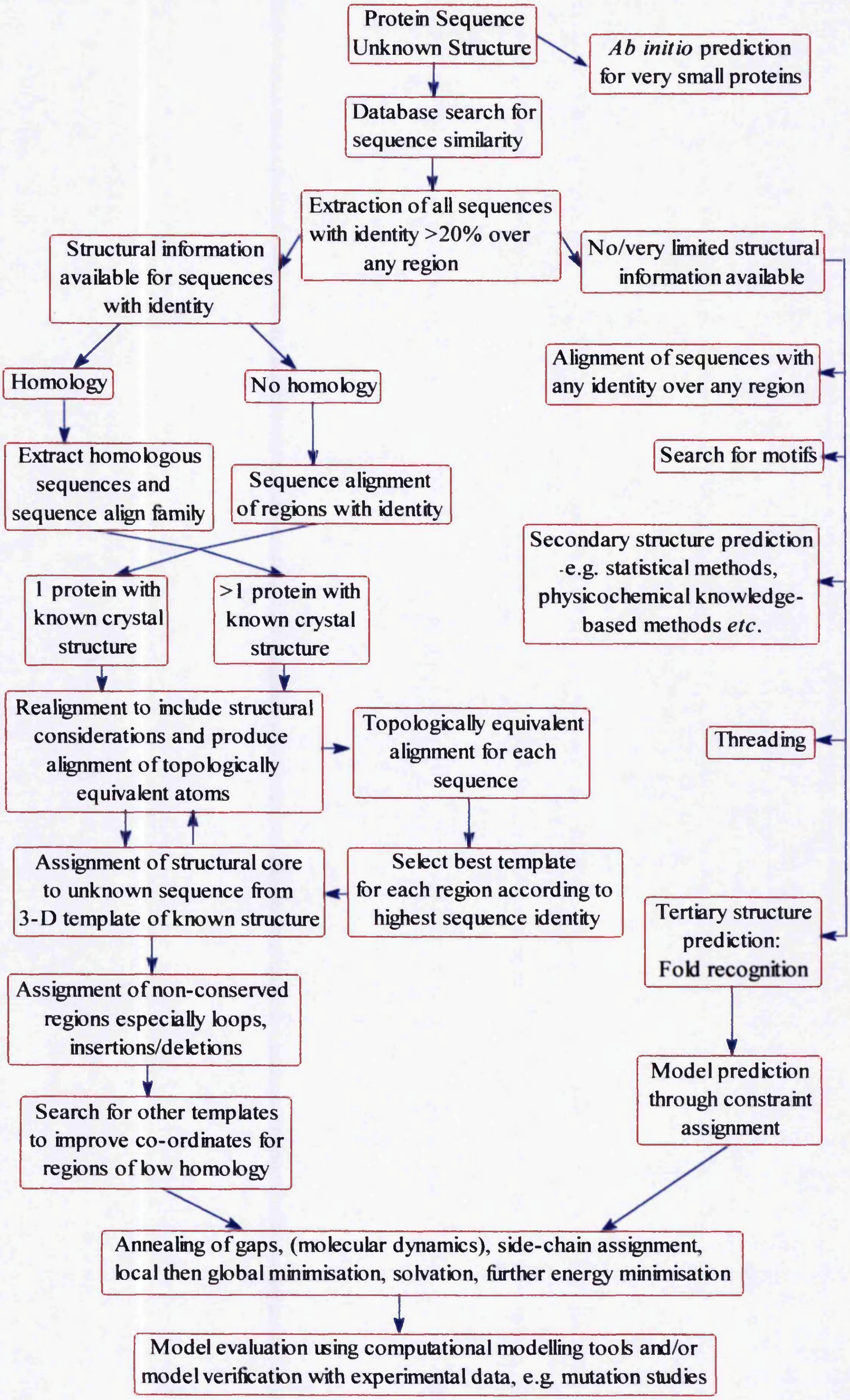

Protein Sequence
Unknown Structure
Ab initio prediction
for very small proteins
Database search for
sequence similarity
Extraction of all sequences
with identity >20% over
any region
Structural information
available for sequences
with identity
No/very limited structural
information available
Homology
No homology
Alignment of sequences with
any identity over any region
Search for motifs
Extract homologous
sequences and
sequence align family
Sequence alignment
of regions with identity
Secondary structure prediction
-e.g. statistical methods,
physicochemical knowledge-
based methods *etc*.
1 protein with
known crystal
structure
>1 protein with
known crystal
structure
Realignment to include structural
considerations and produce
alignment of topologically
equivalent atoms
Topologically equivalent
alignment for each
sequence
Threading
Assignment of structural core
to unknown sequence from
3-D template of known structure
Select best template
for each region according to
highest sequence identity
Tertiary structure
prediction:
Fold recognition
Assignment of non-conserved
regions especially loops,
insertions/deletions
Model prediction
through constraint
assignment
Search for other templates
to improve co-ordinates for
regions of low homology
Annealing of gaps, (molecular dynamics), side-chain assignment,
local then global minimisation, solvation, further energy minimisation
Model evaluation using computational modelling tools and/or
model verification with experimental data, e.g. mutation studies

Figure 3.1. Processes in comparative modelling (adapted [184]).

3.1.1.5. Summary

Biologists sometimes know a vast amount about the chemistry, biology and function of proteins, but even when the precise sequence is known, if there is no homologue or analogue, they are usually powerless to predict its final shape. The conceptual quandary is this: a protein manages to curl up into the same shape every time and while it usually has only one correct fold, it can be contorted into an enormous number of others. When two amino acids bond, they can adopt roughly 10 different orientations in relation to each other so a protein of about 60 amino acids can be in any of about 10^{60} states [185]. This means that even if a protein could try out 100 billion folds a second, it should take longer than the age of the Universe to stumble over its correct fold. But it doesn't.

Comparative modelling is currently the most popular method for predicting the three-dimensional structure of a protein from its sequence. Overall, the accuracy of a model largely depends upon the percentage sequence identity and the presence of substantial insertions or deletions between the template and target structures. Where sequence identity is high (>65-70%), comparative modelling is highly successful and the more confidence that can be placed in the final model, but 40-50% identity will only give knowledge about the overall protein fold. No confidence can be placed in a model based on a sequence identity of 25-30% or less. In some cases one is only interested in the general fold that the protein adopts and so a relatively low resolution structure is acceptable. For other applications, such as drug design, the model must be much more accurate. In these cases, a poor model may often be far worse than no model at all, as it can be seriously misleading.

If there is no obvious homologue, fold recognition can be used to search for an analogous fold. If there is neither a homologue nor an analogue, one can try to dock the secondary structural elements or attempt *de novo* folding if the protein is small.

Researchers are accumulating the structures of so many proteins that the fraction of new structures that are really unrelated in sequences or folding to anything seen before is falling rapidly. In perhaps five or ten years, scientists may have a set of archetypal proteins with folded shapes that would more or less cover the space of all possible structures. By using these guides to build better computer models, it should, in a few years, be possible to predict the shape of any new protein by looking to details of the archetypal protein that its sequence most resembles. For biology and biotechnology, this will usher in a new era. After half a century, scientists will at last be able to read the 'second half' of the genetic code.

3.1.2. Helix-loop-helix transcription factors

3.1.2.1. Gene regulation and transcription

Gene regulation can be effected at many levels of expression and in many ways from transcription (nuclear RNA synthesis) to protein synthesis. For example, RNA synthesis is mostly mediated by transcription factors which bind to DNA enhancer sequences and the accessibility of DNA for the transcriptional apparatus, which is governed by the presence of histones and chromatin packing (reviewed [186]). Differential processing of RNA primary transcripts also controls gene expression predominantly through post-transcriptional modifications of mRNA, e.g. methylation, splicing and phosphorylation. Different ways of splicing mRNA can also control the amount of active form produced. Other factors of consequence for gene regulation include transport of mRNA into the cytoplasm and its stabilisation, rate of mRNA translation, and control of protein activity (reviewed [187,188]).

Transcription is the most significant point at which gene control is exerted and is mediated and modulated though a highly intricate array of proteins and biochemical systems. In unicellular organisms transcriptional gene regulation is necessary for growth and adaptation to the environment, whereas in multicellular organisms, it is required for programming cells in cell 'determination' and for directing production of cell type-specific proteins for specialised functions in cell 'differentiation'. The helix-loop-helix (HLH) proteins are involved in this control system and form one class of the large family of proteins of transcription factors. The HLH family includes the basic-HLH and the inhibitor of differentiation (Id) proteins and regulates transcription by interacting with each other, with other proteins and with DNA. Amongst the HLH proteins are members that can effect transcription activation and members, such as Id proteins, that bind these activators to negate activity. Altered gene regulation is recognised as an event leading to transformation and generation of an oncogenic phenotype. Thus, biological interest in the HLH proteins not only stems from their fundamental role in transcription regulation, but also from the implications of transcription disregulation in development and cancer pathogenesis.

3.1.2.2. General DNA-binding proteins

The transcription of genes within eukaryotic cells is controlled by complex interactions between transcription factors and accessory proteins and specific DNA recognition sequences in target genes. These DNA-binding proteins can be classified into families, each defined by a structural motif, e.g. 'helix-turn-helix', 'helix-loop-helix', β-ribbon motifs, leucine zippers and zinc-binding proteins, e.g. containing zinc finger motifs. Many texts can be found covering the structure and function of DNA-binding motifs but only the 'helix-loop-helix' family of transcription factors is discussed in detail.

3.1.2.3. Helix-Loop-Helix (HLH) proteins

The helix-loop-helix motif is a DNA-binding motif employed by many transcription regulatory proteins which play a vital role in the complex process of controlling gene regulation, cell growth and differentiation. The HLH unit consists of two helices joined by an intervening loop and most HLH motifs are of the bHLH type, having a conserved basic, (b), region N-terminal to the HLH unit, and necessary for electrostatically interacting with the negative phosphate backbone of DNA. HLH proteins form homo- and/or hetero-dimers with other HLH proteins to effect their biological functions and mediate their cellular concentrations, one helix of each monomer being inserted into two adjacent major grooves of the DNA (see Figure 3.2). The HLH domain is used to facilitate this dimerisation and certain HLH transcription factors also possess a 'leucine zipper' region C-terminal to the HLH unit to promote dimeric interactions (bHLHZ proteins, see below). Heterodimerisation with dominant negative (dn) HLH proteins, which do not possess the basic region, negates DNA-binding. The overall effect on transcription is thus partly determined by the nature and relative strengths of the dimers formed, and on this basis gene control can be regulated by the cellular concentrations of the various proteins involved.

The regulation of bHLH proteins occurs at multiple levels. These include tissue-specific expression, differential oligomerisation (e.g. the formation of tetramers and larger multimers [189,190]), different DNA binding specificities and interaction with other bHLH proteins or negatively acting inhibitors of differentiation HLH proteins. bHLH proteins can also be post-translationally modified. Serine phosphorylation of MyoD, for example, prevents it from interacting with DNA as a homodimer *in vitro*, whereas the MyoD/E12 heterodimer is still able to bind DNA [191]. bHLH proteins can also be regulated through interaction with calcium-loaded calmodulin and S-100 proteins [192] and evidence exists to show that the activity of certain bHLH proteins can be repressed by an increased intracellular calcium ion concentration [193]. Calmodulin, S100 and other calcium-binding proteins, e.g. nucleobindin and parvalbumin contain this HLH motif (sometimes called an EF-hand) which may be the site of interaction with the HLH transcription factors. Cellular concentrations of oxygen can also affect the concentration and activity of some transcription factors, e.g. hypoxia-inducible factor-1.

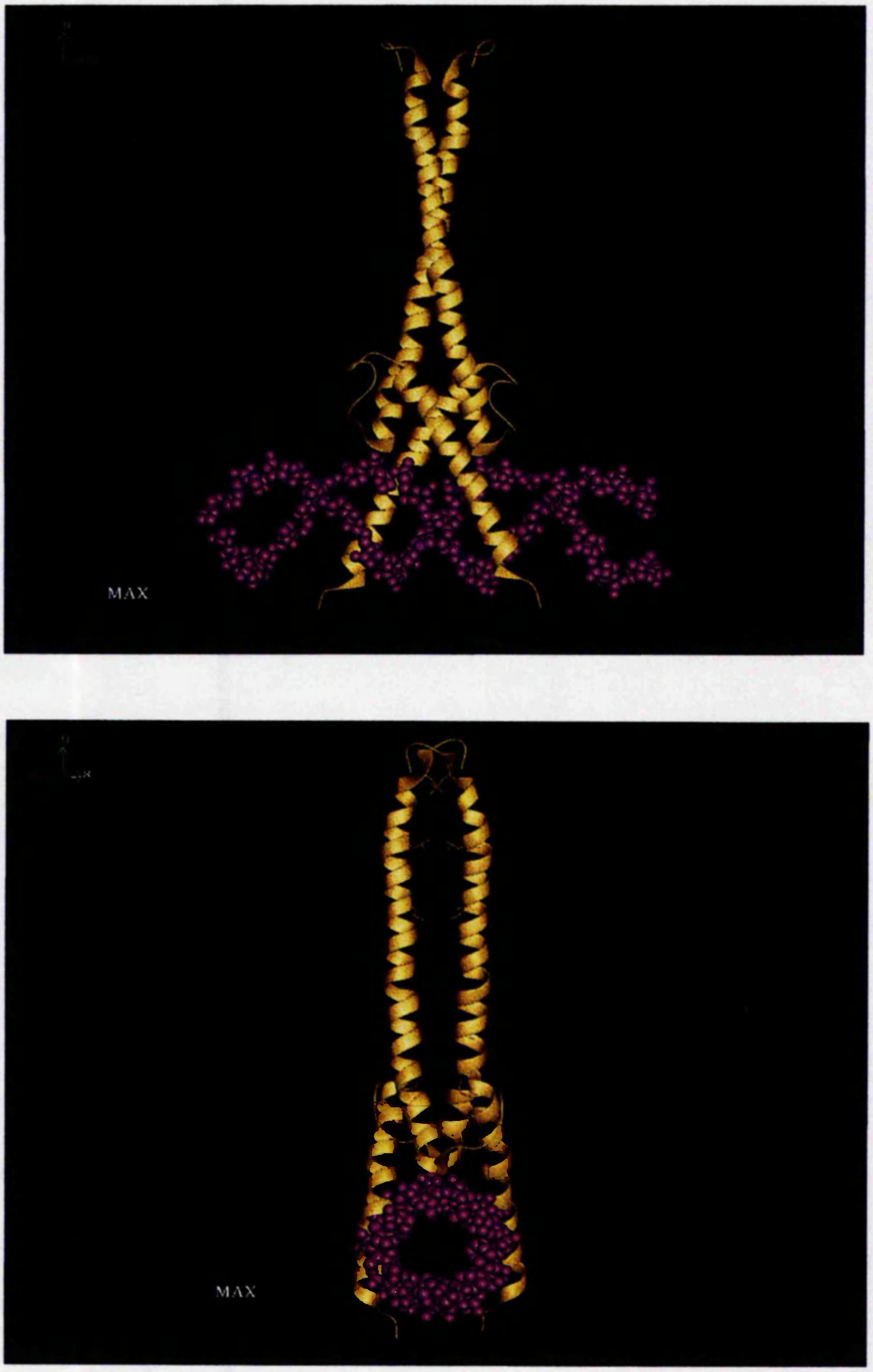


Figure 3.2. The bHLHZ transcription factor Max bound to DNA.

3.1.2.3.1. The HLH DNA-binding site

The activating protein complex, which is responsible for transcription initiation, consists of two domains, DNA-binding and transcription activating.

The DNA-binding domain of nearly all the bHLH and bHLHZ transcription factors bind to a general consensus 'E-box' recognition site 'CANNTG', where NN is the dinucleotide pair CG or GC, either as homo-, or more commonly, hetero-dimers [194-196]. E-box sites, initially identified as elements in B-cell heavy and light chain immunoglobulin gene enhancers, [197,198] are found in promoter and enhancer units in genes including insulin, chymotrypsin, muscle creatine kinase, myosin and acetylcholine receptors [195]. Each partner in the transcription factor dimer recognises a half-site sequence of the DNA binding region. Given the common occurrence of the E-box motif, HLH protein binding is made cell-specific by cooperative binding with accessory factors, adaptors and co-activators adjacent to the E-box site [199]. Such factors, e.g. ETS [200], Ets-1 [201] and LIM [202] proteins, have been identified for expression of immunoglobulin and insulin genes, respectively. The transcription activating domain of the HLH transcription factor binds other transcription factors, e.g. TFII A and TFII B, which interact with another region of DNA, (e.g. the TATA box), at the same time as the DNA-binding domain is interacting with the 'E-box'. Since the whole complex now recognises and binds two DNA sequences situated close together, higher sequence specificity is obtained than if only the 'E-box' was targetted. Regulation of transcription by hypoxia, for example, requires a multiprotein complex that includes the bHLH protein hypoxia-inducible factor-1 (HIF-1), an adjacent transcription factor CREB-1/ATF-1/HNF-4 and the general transcriptional activator p300/CBP [203-205]. Accessory proteins also interact with *myc* to modulate its function [206] and it is likely that these types of accessory proteins will exist for most cell type-specific transcription factors [199].

3.1.2.3.2. Classification: Structure and Function

The HLH proteins can be sub-classified according to their structure and function.

I. Structure

The HLH proteins can be considered in four structural classes [195]:

(i) The basic HLH (bHLH) proteins use the HLH domain for dimerisation and an N-terminal basic region to facilitate DNA-binding, e.g. MyoD and mammalian E2A proteins. This class contains the vast majority of HLH proteins.

(ii) The basic HLH zipper (bHLHZ) proteins are characterised by the possession of a second dimerisation motif, the leucine zipper. The zipper region is found C-terminal to the helix-loop-helix motif and is characterised by a leucine amino acid repeat every seven residues. Thus, when the bHLHZ protein dimerises, the leucine residues of the longer second helix form stabilising hydrophobic interactions. Figure 3.2 (lower plate) shows the three pairs of leucine residues in the transcription factor Max. The zipper region, however, means these proteins are unable to dimerise with 'zipper-less' bHLH or dnHLH members. In mammals, bHLHZ proteins are exemplified by the Max and *Myc* proteins.

(iii) The dominant-negative HLH proteins (dnHLH), lack a basic DNA-binding region and are thus able to dimerise with other HLH proteins, and inhibit DNA binding. The inhibitor of differentiation (Id) proteins have been identified in mammals in this class and there are mammalian homologues of the dnHLH Enhancer of split, (E(spl)) *Drosophila* protein.

(iv) The bHLH-PAS proteins have a dimerisation domain additional to that of the HLH unit consisting of two hydrophobic repeat motifs, PAS A and PAS B. These regions are approximately 50 amino acids in length and contain a signature His-X-X-Asp sequence in each repeat. The PAS domain was first identified in the *Drosophila* proteins Per (the circadian rhythm regulator [207]) and Sim (a neural developmental factor [208]) and in the mammalian proteins Arnt (aryl hydrocarbon receptor nuclear translocator [209,210]) and AHR (aryl hydrocarbon receptor [211]). Per only has a PAS domain and no bHLH region and thus forms abortive complexes that do not bind DNA. Since the bHLH domain of AHR displays a broad dimerisation potential and the PAS domain does not, it appears that the PAS domain is essential to confer dimerisation specificity on the AHR receptor [212]. This may be true of other PAS proteins. An example of a bHLH-PAS protein also includes HIF-1 (hypoxia-inducible factor-1) [213].

II. Function.

The bHLH group of proteins has been roughly classified according to function [195], primarily into Class I and II (or A and B [196]) but the borderlines are sometimes hard to define and distinguish. Classes II, IV and V describe other HLH proteins.

Class I proteins (A). These proteins are ubiquitously expressed and can form homo- and/or hetero-dimers [196]. These proteins direct cell-specific transcription by interacting as heterodimers with the Class II proteins. In contrast to the plethora of class B proteins (see below), only three mammalian genes encoding class A proteins have been identified: E2A, which through differential splicing encodes the proteins E12 and E47 [214]; Insulin transcription factor-1 (ITF-1), E2-2/SEF2-1 (ITF-2) [215,216] HEB/HTF4 [217,218] and daughterless proteins. E12, E47, ITF-1, and E2-2 proteins are involved in B-cell, muscle- and pancreatic-specific gene regulation [196,214,215,219]. *Daughterless* is a gene product that controls sex determination and neurogenesis in *Drosophila* [220].

Class II (B) proteins are cell-type-specific and play important roles in myogenesis, haematopoiesis and neurogenesis in mammals and sex determination in *Drosophila* :

(i) Myogenesis: Several bHLH proteins are involved in the determination of myocytes. MyoD [221], the archetypal bHLH family member, is one of a number of proteins in the MyoD family that function as muscle-determining transcriptional regulators in a heterodimeric complex with E2A proteins (E12 and E47). The heterodimers bind DNA with greater affinity that either homodimer [196]. Many MyoD regulated genes have two cooperative MyoD-binding sites [222]. Other members of the MyoD family are Myf5 [223,224], Myf-6 [225], MRF4 [226] and myogenin [227,228]. All four proteins share considerable sequence homology, which extends beyond the bHLH region, and all have the property, when expressed, of converting a range of different mesenchymal cell types into muscle [229,230]. MyoD and Myf5 share overlapping functions in generating and/or maintaining muscle cell identity and in localising myoblasts to muscle compartments, whereas myogenin appears to be required for full differentiation into myotubes [224,231].

(ii) Haematopoiesis. Several bHLH proteins have been identified as encoded by genes involved in translocations and are thus implicated in the malignant development of T cells. The Tal1 (Scl), Tal2, Lyl1 and Lyl2 genes encode homologous proteins and their disregulation is implicated in leukaemogenesis [195]. This group of proteins is proposed to be regulated by E2A and Id proteins in a similar manner to MyoD.

(iii) Neurogenesis. Whereas MyoD and related bHLH proteins specify muscles in vertebrates, *Achaete* and *Scute* [232,233] and other bHLH proteins specify the initial steps in neural development in *Drosophila*, e.g. *daughterless* [220,234], *extramacrochaetae* [235] and atonal [236]. In vertebrates, Mash1 and Mash2 [237] are known to be involved in neurogenesis and are mammalian homologues of the *Drosophila* genes *Achaete* and *Scute* (reviewed [238]) and the bHLH proteins neuroD and neurogenin [239,240] are also involved in the regulation of vertebrate neurogenesis [241].

(iv) Sex Determination. Many of the HLH proteins involved in the initial steps of *Drosophila* neurogenesis are also involved in sex determination [242]. *Drosophila* sex is determined by the ratio of X chromosomes to autosomes (X:A). The proteins *scute/sisterless-b* [243-245] and *sisterless-a,* which encodes a bZIP protein [246] act as positive transcriptional activators of the female-specific gene *sex-lethal* which promotes female development.

Class III: Basic HLH Zipper (bHLHZ) proteins. A third class of bHLH proteins can be formed including those which are involved in growth control, notably the *myc*-related proteins including c-*myc*, N-*myc* and L-*myc* [247,248]. The cellular proto-oncogene *c-myc,* encodes a multifunctional protein involved in cell-cycle entry and progression, growth, replication, development apoptosis, and is implicated in transformation and tumourigenesis when control is disturbed. The *myc* proteins heterodimerise with the bHLHZ Max proteins [249]. Max is a ubiquitously expressed protein that preferentially forms DNA-binding Max/*myc* heterodimers rather than its homodimer [250].

Four closely related bHLHZ members of the 'c-*myc* network' have been identified. These proteins are collectively referred to as the Mad family and consist of Mad1 [251], Mxi1 [252], Mad3 and Mad4 [253]. All four Mad proteins are similar in that they homodimerise poorly and thus lack intrinsic DNA binding activity. They readily form homodimers with Max, compete for c-*myc* binding sites and negatively regulate the activity of c-*myc*. The recently identified Mmip1 [254] strongly dimerises with all four Mad members, but not with c-*myc* or Max.

Other bHLHZ family members include the mammalian nuclear proteins USF, TFEB, TFE3 and AP-4. USF is implicated in the expression of several tissue-specific or developmentally regulated genes, including human growth hormone, mouse metallothionein I and rat γ-fibrinogen[255], TFEB [256] and TFE3 [257] bind to the heavy chain immunoglobulin enhancer and AP-4 [258] binds the SV40 enhancer and activates viral late gene transcription.

Class IV: Dominant Negative (dn) HLH proteins

This class contains proteins lacking a DNA-binding domain, e.g. emc (*extramacroachaetae*) [235], the Id family, Hairy [259] E(spl), [260,261], and E(spl) mammalian homologues [262]. Hairy and E(spl) proteins have a proline in their basic region which disrupts the normal helix conformation required to bind DNA. The Id proteins display differential interactions with bHLH transcription factors, prevent DNA binding, thus acting as dominant negative transcription regulators [263]. The Id proteins are involved in inhibiting cell differentiation. Id expression decreases in a variety of cell lines when they are induced to differentiate [264-266] and conversely, cell differentiation is inhibited by over-expressing Id [267-269]. They also act as positive regulators of G(1) cell cycle control [270,271]. Individual Ids also interact in distinct ways with non-bHLH proteins [272-274]. The Id family has been found to inhibit differentiation of myogenesis in myeloid [268], erythroid [275,276] and mammary epithelial cells [277]. The four Id proteins also have distinct regulatory roles during meiosis, spermatogenesis and Sertoli cell function [278]. Four distinct Id proteins have been identified, Id1 [264,279], Id2 [265,280], Id3 (also known as 1R21, HLH462) [281,282] and Id4 [283]. Id1 and Id3 heterodimerise with, and prevent DNA-binding of, the E2A proteins and MyoD [263,274,284]. These two also have a functional role in the control of proliferation and differentiation of cartilage [285]. Id1 has also been shown to regulate indirectly several non-myogenic tissue-specific promoters [286,287]. Id3 has also been identified as being involved in adipocyte differentiation [288]. Id2 promotes apoptosis by a mechanism independent of dimerisation to bHLH factors [289], enhances cell proliferation by binding to the non-bHLH unphosphorylated retinoblastoma protein (pRb) family members and abolishing their growth-suppressing function [290-292] and may have a role in human pancreatic cancer [293]. Id4 expression is restricted to neuronal cells in the developing brain and spinal cord of the mouse embryo and in adult mice expression is highest in brain, kidney and testis [294]. Expression control of cell type-specific proteins in terminal differentiation is likely to be mediated by the amount of Class I protein that is available to form functional heterodimers with Class II proteins, as regulated by dnHLH proteins. Overexpression of the Id proteins has been found to result in cell lines with similarities to malignant phenotypes [282] probably through failure to differentiate.

Class V: bHLH-PAS proteins

Transcription factors of the bHLH-PAS protein family are important regulators of developmental processes such as neurogenesis and tracheal development in invertebrates. They function in the response to environmental stimuli such as xenobiotics and hypoxia. As described above, the bHLH-PAS proteins have a PAS dimerisation domain additional to that of the HLH. Roughly a dozen bHLH-PAS proteins have been identified to date, including *Drosophila* Per (the circadian rhythm regulator [207], Clock protein (also involved in the circadian pacemaking system) [292,295], Sim (a neural developmental factor) [208], and Arnt [209,210] and AHR [211] which are involved in the xenobiotic response, i.e. the metabolism of xenobiotic compounds.

Examples of bHLH-PAS proteins also include Spineless [296], Arnt2, [297], trl (trachealess), [298], Sim-a [299], Sim-2 [300], BMAL1/MOP3 (brain and muscle Ah receptor nuclear translocator-like protein) [301,302], NPA-1 and NPA-2 (Neuronal PAS domain proteins) [303], HIF-1 (hypoxia-inducible factor-1) [213], and EPAS-1 (Endothelial PAS domain protein-1) [200], (also called MOP2, HRF (HIF-related factor)[304], and HLF (HIF-like factor) [305]). These latter two are very similar with regards to their sequences and are expressed only in hypoxic conditions. HIF-1 activates a network of genes encoding vascular endothelial growth factor (VEGF) and several glycolytic enzymes. EPAS-1 is involved in the regulation of endothelial cell gene expression.

Sequence comparison of the bHLH-PAS proteins indicates a division into two phylogenetic groups I and II. Proteins of group I include Arnt (HIF-1β), Arnt2 and Per and these can form homodimers or heterodimerise with members of group II. Group II includes HIF-1α, Sim, AHR, and EPAS-1. These will only dimerise with group I.

Proteins researched in this study

The bHLH-PAS protein HIF-1 and the dnHLH protein Id3 discussed above are studied in detail in this thesis.

3.1.2.4. Hypoxia-Inducible Factor-1 (HIF-1)

Hypoxia is an important component of many pathological processes including tumour formation, where it has been associated with resistance to radiotherapy, malignant progression and metastasis formation [306-309]. Changes in gene expression accompanying tumour hypoxia are well recognised, but the underlying mechanisms and precise consequences are not so well understood. Hypoxia-inducible factor-1 (HIF-1) is a common transcription factor that has been implicated in the hypoxic induction of a number of groups of mammalian genes. HIF-1 was first identified as a factor critical for the inducible activity of the erythropoeitin 3' enhancer [213]. It is a heterodimer with both subunits (HIF-1α and -1β (Arnt) [209,310]) containing a bHLH and a PAS domain. It is now recognised to be a key component of a widely operative transcriptional control system responding to physiological levels of cellular hypoxia. Insulin has also been shown to induce transcription HIF-1 mediates transcriptional activation of a network of genes encoding vascular endothelial growth factor (VEGF), haeme oxygenase 1, inducible nitric oxide synthase and several glycolytic enzymes (e.g. phosphofructokinase L, phosphoglycerate kinase 1, lactate dehydrogenase A) [311-316]. HIF-1 is therefore important in modulating gene expression in solid tumours, which contain hypoxic regions, influencing angiogenesis and tumour growth [317,318]. A second role of HIF-1α is in the stabilisation of p53 tumour suppressor protein by association with it [319].

To initiate transcription under hypoxic conditions, HIF-1 forms a complex with other factors. The C-terminal activation domain of the β subunit of HIF-1 binds to the transcription factor HNF-4 and the C-terminus of HIF-1α binds to p300/CBP, a general transcriptional activator [205]. A p300-binding protein (p35srj) has recently been discovered which inhibits the HIF-1α-p300 interaction thus regulating HIF-1 transactivation [320].

Two transactivation domains on HIF-1α have been identified, one at the C-terminus and the other in the middle of the polypeptide [321-323], but there is very little understanding of the steps underlying the activation of HIF-1 by hypoxia. At the protein level, Arnt is not significantly affected by oxygen. In contrast, HIF-1α is remarkably unstable in cells exposed to oxygen and is rapidly degraded, whereas hypoxia induces a striking increase in the abundance of HIF-1α protein. This suggests that accumulation of HIF-1α is a prerequisite to the activation of HIF-1, which depends primarily on hypoxia-induced stabilisation of HIF-1α. An oxygen-dependent degradation (ODD) domain within HIF-1α has recently been identified that controls its degradation by the ubiquitin-proteasome pathway [324]. This ODD domain consists of approximately 200 amino acid residues, located in the central region of HIF-1α and its removal renders HIF-1α stable, even in oxygenated cells, resulting in heterodimerisation with Arnt, DNA binding and transactivation independent of hypoxia signalling.

Inhibition of HIF-1 activity by nitric oxide in hypoxia has been demonstrated, performed by blocking an activation step of HIF-1α to a DNA-binding form [325].

Some HIF-1α-like proteins have also been identified, mentioned above. They include the hypoxia-regulated HIF-3α which also dimerises with Arnt [326], endothelial PAS-1 (EPAS1), which is expressed only in endothelial cells and is induced by hypoxia [327], neuronal PAS domain proteins (NPA) 1 and 2 [328] and HIF-related (HRF), also known as HIF-1α-like factor (HLF), detected in mouse brain capillary endothelial and bronchial epithelium cells [304,305]. A protein with a similar sequence to HIF-1β (Arnt2) also exists [297] and HIF-1α also binds the PAS protein BMAL1a (brain and muscle Ah receptor nuclear translocator-like human protein [301]).

3.1.2.5. Id3

The inhibitory Id family and the role of its members has already been briefly discussed. Id3 binds to MyoD and the E2A proteins E12 and E47, regulating myogenesis and B-cell, immunoglobulin, muscle- and pancreatic-specific genes.

Given the importance of Id proteins in the regulation of a variety of cellular processes (cell determination, differentiation and growth) and cell-type-specifc transcription, and considering the involvement of HIF-1 in tumour growth and angoigenesis, these two proteins appear attractive targets for ligand design.

3.1.3. Binding assays

3.1.3.1. Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis (SDS-PAGE)

Electrophoresis is the migration of a charged particle in an electric field. As proteins carry a net charge at any pH other than their isoelectric point, they migrate at a rate dependent on their charge density (charge/mass). Application of an electric field to a protein mixture in solution therefore results in different proteins migrating at different rates towards one of the electrodes. It is possible to carry out electrophoresis in free solution but, since all proteins were originally present throughout the whole solution, the separation is minimal. In zone electrophoresis, the mixture of molecules to be separated is placed as a narrow band at a suitable distance from the electrodes such that molecules of different mobilities travel as discrete zones which gradually separate from each other as electrophoresis proceeds. In this approach the heat produced results in convective effects that disrupt the separated protein zones and diffusion constantly broadens the protein zones. These deleterious effects can be minimised by performing the separation in a support medium which inhibits convection, e.g. gel-based media, such as starch, agarose and polyacrylamide.

Polyacrylamide gels, chemically inert non-ionic polymers are stable over a wide range of pH, temperature and ionic strength and are transparent. Moreover, they can be produced with a wide range of pore sizes, optimisable for the separation of proteins of different size ranges. In gel-based media, a pore size approximately of the same order as the size of protein molecules results in a molecular-sieving effect. Thus, the resulting separation depends on both charge density and the size of the proteins being analysed. The gel is placed in a buffer which both serves as the electrolyte and dissipates the heat produced. Molecular weight markers can also be applied as standards.

Problems can arise if any of the proteins present are insoluble or are likely to aggregate under the conditions used for separation. Suitable additives are therefore used to prevent this, e.g. dithiothreitol to cleave disulphide bonds, urea to cause protein unfolding.

SDS-PAGE is currently the most commonly used electrophoretic technique for protein analysis. The strong anionic detergent SDS, when used in the presence of disulphide bond cleaving reagents, solubilises, denatures and dissociates most proteins to produce single polypeptide chains. The resulting SDS-protein complexes can then be separated in gels containing SDS. The number of moles of anionic SDS bound to the protein is proportional to the polypeptide chain length (over a reasonable range) and hence to the subunit molecular weight.

After electrophoresis, the gel is removed, the separated proteins are immobilised and precipitated ('fixing') and stained, e.g. with Coomassie Blue. Radiolabelled proteins can be detected by using radiographic film or proteins can be labelled with a fluorescent dye prior to electrophoresis to be visualised.

3.1.3.2. Resonant mirror biosensor

The IAsys instrument (Affinity Sensors) is an optical biosensor which incorporates a stirred microcuvette for studying biomolecular interactions in real-time, revealing the dynamics as well as the strength of binding. IAsys uses the optical phenomenon of an evanescent field, which occurs when light undergoes total internal reflection and is enhanced within a patented wave-guide structure called the Resonant Mirror (Figure 3.3) [329]. This forms a resonant cavity (or wave-guide) and is almost a perfect reflector of light. In total internal reflection (which occurs at angles ≥ the 'critical angle') light waves are completely reflected at the high-to-low refractive index boundary. The component comprising the electric field penetrates approximately one wavelength into the low refractive index medium, decaying exponentially and is termed the evanescent field. Its strength is enhanced at the biosensor device's resonance angle which is highly sensitive to the refractive index at the surface. The evanescent field is used to probe the refractive index, thus conferring a surface selectivity to the sensor system: bulk changes outside of the surface region are not seen. By immobilising a receptor molecule on the sensor surface and adding a solution of ligand, it is therefore possible to measure only those molecules that bind to, or dissociate from, the receptor.

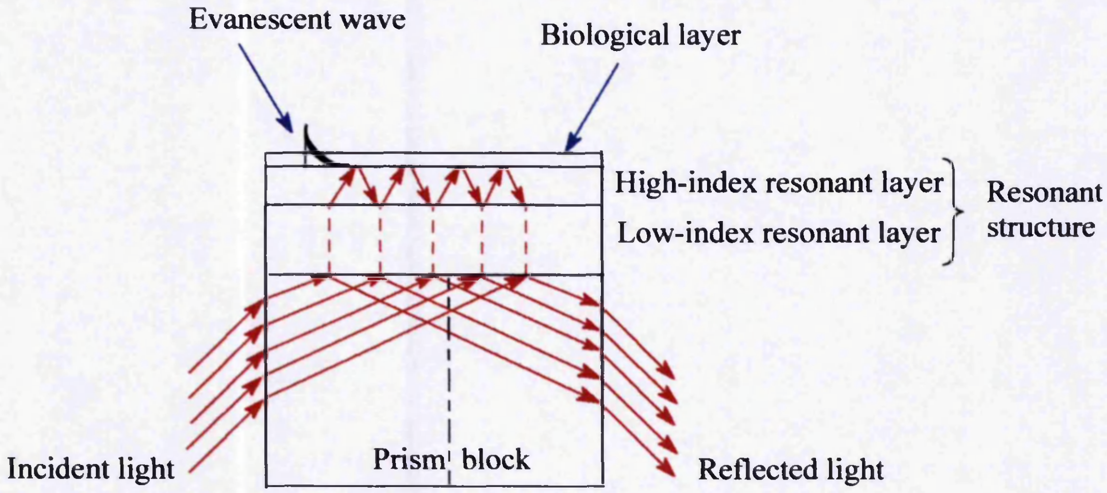


Figure 3.3. Structure of the resonant mirror sensing device (adapted [329]).

The resonant mirror consists of a high-index wave-guide (resonant layer) separated from a high-index prism block by an intervening low-index coupling layer. Changes in the refractive index at the surface of the device (to which the receptor molecules are attached) change the angle at which light can be made to propagate in the wave-guide. Laser light (670nm) is scanned in a 10° arc across the device. At the resonant angle, high intensity light passes from the prism through the coupling layer to propagate in the wave-guide as a surface evanescent wave. The light returns *via* the coupling layer, emerging to strike the detector. To properly resolve the resonance angle, the optical components are arranged to give a 90° rotation of polarisation. A polariser in front of the detector enables the angle of interest to be accurately distinguished and this is extremely sensitive to the binding reactions occurring at the device surface.

3.1.3.3. Fluorescence spectroscopy

Fluorescence spectroscopy has assumed an increasing role in the study of the dynamics and structure of biological macromolecular assemblies. Intrinsic or extrinsic fluorescent emission, extremely sensitive to local environment, can be used to monitor the kinetics and thermodynamics of the incorporation of a particular subunit or substrate into an assembly. Fluorescence can also be used to measure distances between pairs of loci in the assembled structure (discussed in the following section). The basic principles of fluorescence are now described.

When a molecule in its ground state, S_0 (Figure 3.4), is irradiated with radiation of the appropriate wavelength, energy is absorbed and the molecule excited to a singlet state S_1. This state contains many quantised energy levels and the molecule can occupy any of these, but vibrational relaxation by energy loss to the surroundings effectively immediately returns it to the lowest energy state, S_1.

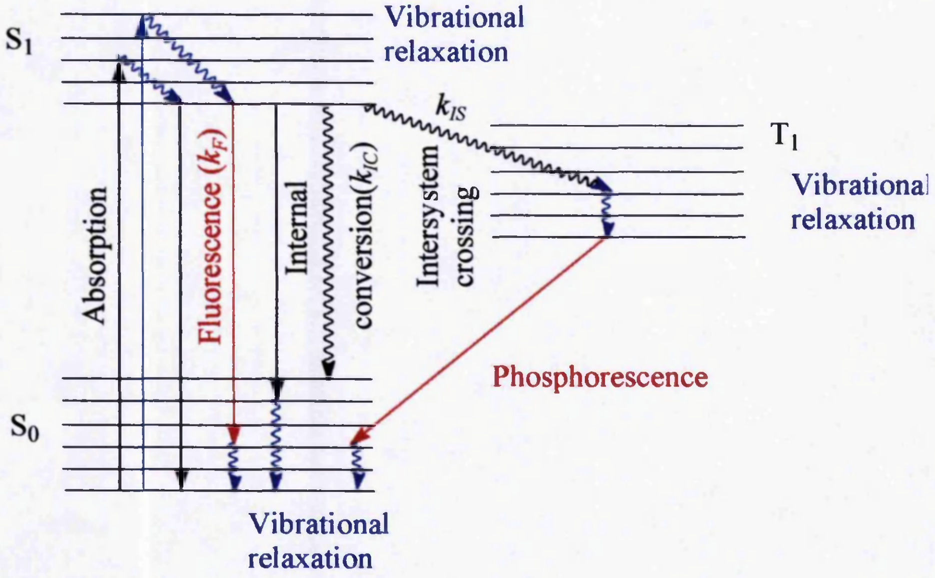


Figure 3.4. Pathways for production and de-excitation of an excited state.

The molecule can then lose energy from this state through any of the following processes, all returning the molecule to the original ground state, S_0:

(i) In *fluorescence*, (k_F), the molecule returns to S_0 by the emitting radiation at a wavelength longer than the absorption wavelength because energy loss through vibrational relaxation has already occurred in S_1, and the molecule can return to any number of energy levels of S_0 higher in energy than the ground state of S_0.

(ii) By *internal conversion*, (k_{IC}), excitation energy is lost by collision with solvent molecules, or by dissipation through internal vibrational modes. The rate (k_{IC}) of internal conversion generally increases with temperature due to the increased rate of solvent collisions. Thus, fluorescence correspondingly decreases.

(iii) In *intersystem crossing*, (k_{IS}), spin exchange converts the excited singlet (spin=0) into an excited triplet (spin=1). This can convert to the ground singlet state (S_0) either by *phosphorescence* (emission of a photon) or by *internal conversion*. A spin conversion is forbidden by the spin selection rule, so the intensity for direct singlet-triplet absorption is extremely low and the triplet state normally has an extremely long radiative lifetime, often seconds, rather than the nanoseconds found for singlets. This means that collision with quenchers (molecules which deactivate the excited state) or internal conversion can compete effectively with phosphorescence and in solution it is rarely observed.

3.1.3.3.1. Fluorescence Resonance Energy Transfer (FRET)

Excellent reviews on the basics of fluorescence resonance energy transfer (FRET) [330-333] should be consulted for more detailed discussion of the technique and its applications. A brief introduction follows.

FRET is a very sensitive technique for measuring the distance between two fluorophores separated by approximately 10-100Å. It is a distance-dependent interaction between the electronic excited states of two molecules in which excitation is transferred from a donor molecule to an˙acceptor molecule without emission of a photon. The idea behind the technique is to label the two points of interest with different dyes, a donor, which must be fluorescent, and an acceptor, which need not necessarily be fluorescent, but often is.

The three primary conditions for FRET are:

1. The absorption spectrum of the acceptor must overlap with the fluorescence emission spectrum of the donor (Figure 3.5).
2. Donor and acceptor molecules must be close (typically 10-100Å). (The exact range depends on the dyes chosen).
3. Donor and acceptor transition dipole orientations must be approximately parallel.

On photoexcitation the donor transfers energy to the acceptor, which then fluoresces at a much longer wavelength than the original absorption wavelength. FRET depends on the inverse sixth power of the intermolecular separation [334], making it useful over distances comparable with the dimensions of biological macromolecules. By measuring the amount of energy transfer, it is possible to estimate the distance between donor and acceptor [335]. The extent of energy transfer can be measured in several ways: decrease in donor intensity or quantum yield, increase in intensity of acceptor emission (sensitised emission), decrease in donor lifetime, change in lifetime

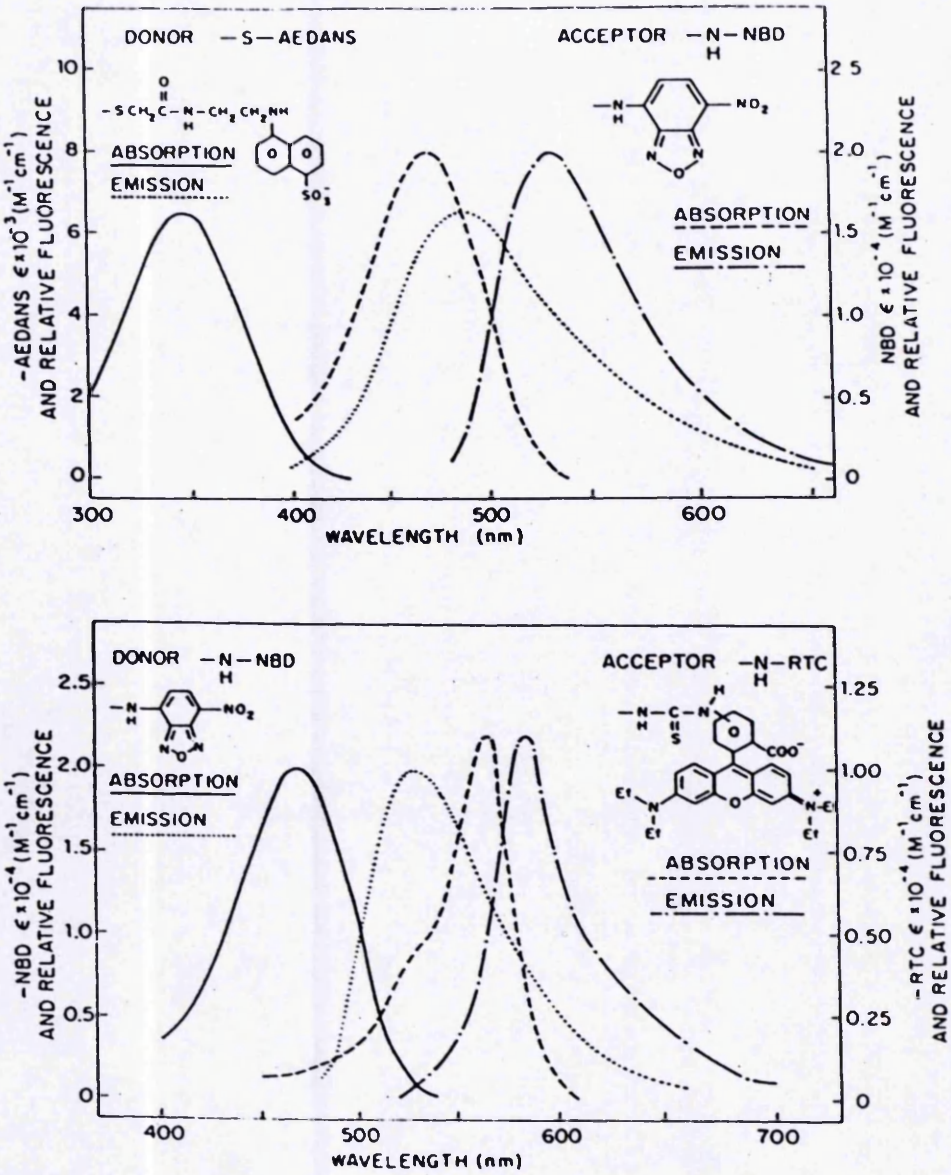


Figure 3.5. Absorption and emission spectra of two donor-acceptor pairs used in FRET studies (adapted [332]).

of excited state of the sensitised molecule. These changes in fluorescence can be measured by comparing a complex labelled with both donor and acceptor to controls labelled only with donor or acceptor. FRET is thus useful for investigating a variety of biological phenomena that produce changes in molecular proximity [336,337]. Reviews exist on the applications of FRET to the study of actin and assembly [338], nucleic acids [339], phycobiliproteins [340] and microscopy [341]. Examples of systems studied are oligonucleotides [342,343], Holliday junctions [344], nucleic acid hybridisation [345,346], oligopeptides [334], rhodopsin [347], myosin [348], calcium binders [349], receptor-ligand interactions [350,351], RNA [352], and nucleic acid-protein complexes such as nucleosomes [353] and protein-promoter interactions [354]. FRET has also been used to monitor dynamic processes, e.g. HIV protease activity [355].

Choice of dyes for FRET

Labelling the sites of interest with appropriate dyes can be difficult. The dyes must be spectrally compatible (*i.e.* the emission spectrum of the donor must overlap the absorption spectrum of the acceptor), and able to site-specifically label the molecule of interest without perturbing its initial structure. In addition, it is helpful to have an idea of the magnitudes of the distances to be measured. If the two points are less than a certain characteristic distance, R_0, (the distance at which 50% of the energy is transferred), almost all the energy is transferred, but if greater than this distance very little energy is transferred. Ideally a pair of dyes are picked with R_0 equal to the distance to be measured since small changes in the distance around R_0 lead to such large changes in signal. Because the distances are generally unknown it is wise to pick a donor-acceptor pair with a large R_0 value greater of equal to the distance to be measured. Figure 3.5 show two examples of donor-acceptor pairs. More pairs include fluorescein and tetramethylrhodamine, AEDANS and nitrobenzofurazan (NBD), tryptophan and AEDANS [332].

Problems

Whilst FRET is good for measuring relative distances, it is limited in measuring absolute distances, because the efficiency of energy transfer depends, not only on the distance between the donor and acceptor, but also on the relative orientations of the dyes, a factor not precisely known. This orientation factor can be significant, modifying the fitted-distance by a factor of 0 to 1.25. FRET is also limited for absolute distances by the uncertainty in defining the exact position of the FRET dyes because of the flexibility of the linker used to attach the dyes. The very sharp distance dependence of FRET leads to two drawbacks: (1) it is difficult to measure relatively long distances because the signal is very weak, and (2) the signal tends to be 'all or none'.

3.1.4. Aims

In many cases, fully functional transcription factors are detrimental to the survival of the cell. Knowledge of the structure of these proteins is vital in understanding and explaining their functional and dimerisation properties but the X-ray crystal structures of only a handful of the bHLH and bHLHZ domains of some transcription factors have been solved. All bHLH transcription factors bind DNA as a dimer, monomers having no transcriptional activity alone and so if dimerisation can be prevented, transcriptional activation will not occur. Site-directed mutagenesis studies on the HLH domains of the proteins Max, E47 and MyoD show that the highly conserved hydrophobic core formed in the interface of two HLH monomers is a major determinant of dimer stability [238,356,357]. An obvious way to disrupt the life-cycle of a cell in cases where it is desirable to do so, e.g. cancerous cells, is to prevent transcription of certain genes. One way of doing this is therefore to design an inhibitor to sit at the hydrophobic interface of the monomers and prevent dimerisation. There has been no report of this being undertaken, so a generic proof of principle was attempted, applied to both HIF-1 and Id proteins. Structural modelling studies had already been carried out on Id3 in our laboratory [184,358], but the three-dimensional structure of HIF-1 is unknown.

The aims of the study were therefore to build a model of HIF-1 and then design molecules to bind the hydrophobic regions of the monomers HIF-1α and Id3. Peptides were designed as inhibitors as many can be easily synthesised in a knowledge-guided combinatorial manner from just a limited number of amino acids. The binding of the synthesised peptides were to be investigated using the three techniques described above.

The SDS-PAGE pull-down assays and the IAsys biosensor were used to study the binding of peptides to Id3 and possible dimerisation inhibition potential. It was aimed to assess FRET as a potential monitor to study the peptides binding to the Arnt or HIF-1α monomer of the HIF-1 transcription factor dimer. The peptides were labelled with an acceptor molecule and the Arnt/HIF-1α protein with a fluorescent donor.

Therefore, the aims of this study were as follows:

(i) Construct a comparative molecular model of the bHLH domain of HIF-1.

(ii) Identify residues important for dimerisation in the HLH region.

(iii) Identify residues important for binding to DNA.

(iv) Design a family of peptides to inhibit dimerisation of HIF-1 and E47-Id3 by mimicking the important residues on one monomer dimerisation surface.

(v) Synthesise the peptides using solid-phase knowledge-based combinatorial chemistry.

(vi) Test the peptides for binding activity and potential dimerisation inhibition using a range of techniques e.g. pull-down assays (SDS-PAGE), resonant mirror biosensor techniques (monitoring the changes in the refractive index of free protein *versus* peptide-bound protein), fluorescence.

(vii) Suggest mutant constructs of HIF-1α and HIF-1β to be used for fluorescent labelling.

3.2. MATERIALS AND METHODS

3.2.1. Comparative modelling

3.2.1.1. Database searching and sequence alignment

A search of proteins of the Brookhaven Protein Database (http://www.pdb.pdb.bnl/gov) using the key words 'helix-loop-helix' produced five proteins: MyoD mouse protein (database entry 1mdy) [359], yeast Pho4 (entry 1a0a, [360]), Max human protein (entry 1hlo, [361] and entry 1an2, [362]), human upstream stimulatory factor 1 (USF, 1an4)[363] and a theoretical model of the human protein E47 (entry 1hlh, [364]). The structure of E47 has been solved [365], but not deposited in the Protein Database. At the time of modelling, the structure of USF was also not publicly available so the co-ordinates of these two proteins were kindly provided by personal communication from the original authors. (A crystal structure of the bHLH sterol regulator binding protein 1a, SREBP, (entry 1am9A) [366] has now been solved since the time of modelling, so is not included in this study, though it is referred to in the Discussion). NMR structures have also been determined of the bHLH region of E47 [367] and the c-Myc-Max heterodimer [368].

The amino acid sequences of the human HIF-1 monomers α and β (ARNT) were extracted from the web using SEQNET (SERC Laboratories, Daresbury U.K., http://www.SEQNET.dl.ac.uk). The HLH region of HIF-1α is found in the first 80 residues of the sequence and the HLH region of HIF-1β/ARNT is found in residues 90-160 [310]. This was confirmed using the ProDom structural analysis program [369]. These sequences were aligned with the HLH regions from the proteins with known crystal structures using the multiple sequence alignment program BLASTP, and MULTALIN [169] was also used to check the results.

A review by Littlewood and Evan [238] contains a sequence alignment of the HLH proteins known at that time and gives qualitative information on conserved residues and sequence similarity over the HLH family. To try and identify other HLH sequences to add to this alignment and possibly find more solved HLH structures, the HLH regions of HIF-1α and 1β were used as query sequences in BLASTP searches [167] of the OWL [370,371] and SWISS-PROT [372] protein sequence databases. (The following default values were used; scoring matrix = BLOSUM 62 [166], wordsize = 3, statistical significance threshold = 10). A maximum of 150 sequences was listed and a maximum of 100 sequences aligned. A search of GENBANK with the keywords 'helix-loop-helix' was also carried out and all the results edited to remove redundancy. These searches yielded only four of the proteins above whose three-dimensional structures have been determined: MyoD, Pho4, Max and USF, but it gave approximately 30 additional new sequences to add to Littlewood and Evan's alignment. The resulting alignment (Figure 3.6) gives no hierarchical order of similarity to HIF-1α or 1β, but is intended to serve as a general visual guide to conserved regions of amino acids.

Figure 3.6. Sequence alignment of the HLH region of proteins representative of the HLH family. The proteins listed until Ra were aligned by Littlewood and Evan [238] and are referenced therein. HIF-1α and the proteins from Ra onwards were found by searching the SWISS-PROT, OWL and GenBank databases using the keywords 'helix-loop-helix'.

```
                                  <           *  H1          >
HIF-1α         11          KKISSERRKEKSRDAARSRRSKESEVFYELAHQLP
HIF-1β         82         SSADKERLARENHSEIERRRRNKMTAYITELSDMVP
SIM            1                   MKEKSKNAARTRREKENTEFCELAKLLP
c-Myc(Z)       348        STDTEENVKRRTHNVLERQRRNELKRSFFALRDQIP
N-Myc(Z)       480               SERRRNHNILERQRRNDLRSSFLTLRDHVP
L-Myc(Z)       273       VSSDTEDVTKRKNHNFLERKRRNDLRSRFLALRDQVP
V-Myc(Z)       315       RTLDSEENDKRRTHNVLERQRRNELKLRFFALRDQIP
S-Myc(Z)       338       SNSDLEDIERRRNHNRMERQRRDIMRSSFLNLRDLVP
Max(Z)         19            QSAADKRAHHNALERKRRDHIKDSFHSLRDSVP
Mad(Z)         49         KSKKNNSSSRSTHNEMEKNRRAHLRLCLEKLKGLVP
Mxi1(Z)        22        SGTSNTSTANRSTHNELEKNRRAHLRLCLERLKVLIP
AP4(Z)         25          RDQERRIRREIANSNERRRMQSINAGFQSLKTLIP
USF(Z)         195           TRDEKRRAQHNEVERRRRDKINNWIVQLSKIIP
TFE3(Z)        133         ALLKERQKKDNHNLIERRRRFNINDRIKELGTLIP
TFEB(Z)        326         ALAKERQKKDNHNLIERRRRFNINDRIKELGHLIP
TFEC(Z)        104         ALAKERQKKDNHNLIERRRRYNINYRIKELGTLIP
FIP(Z)         137         RTPRDERRRAQHNEVERRRRDKINNWIVQLSKIIP
ADD1(Z)        288          AQSRGEKRTAHNAIEKRYRSSINDKIVELKDLVV
Mi(Z)          198         ALAKERQKKDNHNLIERRRRFNINDRIKELGTLIP
SREBP-1a(Z)    318          AQSRGEKRTAHNAIEKRYRSSINDKIIELKDLVV
E12            543         KAEREKERRVANNARERLRVRDINEAFKELGRMCQ
E47/ITF1       332/471       LRDRERRMANNARERVRVRDINEAFRELGRMCQ
ITF2           513         KAEREKERRMANNARERLRVRDINEAFKELGRMVQ
HEB            562         KIEREKERRMANNARERLRVRDINEAFKELGRMCQ
Da             548         KAIREKERRQANNARERIRIRDINEAFKELGRMCM
Twist          447         ETDEFSNQRVMANVRERQRTQSLNDAFKSLQQIIP
L-Sc           80             EQLPSVARRNARERNRVKQVNNGFVNLRQHLP
Scute          96             DQSQSVQRRNARERNRVKQVNNSFARLRQHIP
Achaete        23               GPSVIRRNARERNRVKQVNNGFSQLRQHIP
Asense         156            PLPQAVARRNARERNRVKQVNNGFALLREKIP
Mash1          109         LPQQQPAAVARRNERERNRVKLVNLGFATLREHVP
Mash2          117              SAAVARRNERERNRVKLVNLGFQALRQHVP
MyoD           105           TTNADRRKAATMRERRRLSKVNEAFETLKRCTS
Myogenin       80               VDRRRAATLREKRRLKKVNEAFEALKRSTL
Myf5           82               MDRRKAATMRERRRLKKVNQAFETLKRCTT
MRF4           92               TDRRKAATLRERRRLKKINEAFEALKRRTV
Lyl1/2         145/112       PQKVARRVFTNSRERWRQQHVNGAFAELRKLLP
Tal1           185             KVVRRIFTNSRERWRQQNVNGAFAELRKLIP
Tal2           1                 TRKIFTNTRERWRQQNVNSAFAKLRKLIP
Hen1/2         75/78             KYRTAHATRERIRVEAFNLAFAELRKLLP
Hes1(D)        32              SEHRKSSKPIMEKRRRARINESLSQLKTLIL
Hes2(D)        11              AELRKSLKPLLEKRRRARINESLSQLKGLVL
Hes3(D)        1                              MEKKRRARINLSLEQLRSLLE
Hes5(D)        10          LSPKEKNRLRKPVVEKMRRDRINSSIEQLKLLLE
E(spl)M5(D)    10         FVSKTQHYLKVKKPLLERQRRARMNKCLDTLKTLVA
E(spl)M7(D)    7           MSKTYQYRKVMKPLLERKRRARINKCLDELKDLMA
E(spl)M8(D)    4           TTKTQIYQKVKKPLMERQRRARMNKCLDNLKTLVA
Hairy(D)       25          ETPLKSDRRSNKPIMEKRRRARINNCLNELKTLIL
Dpn(D)         35           LSKAELRKTNKPIMEKRRRARINHCLNELKSLIL
Emc(D)         23                RIQRHPTHRCDGENAEMKMYLSKLKDLVP
Id1(D)         71             GTRLPALLDEQQVNVLLYDMNGCYSRLKELVP
Id2(D)         75                SRSKTPVDDPMSLLYNMNDCYSKLKELVP
Id3(D)         25             ARGRGKSPSTEEPLSLLDDMNHCYSRLRELVP
Id4(D)         48            ARCKAAEAAADEPALCLQCDMNDCYSRLRRLVP
Cbf1           216         TDEWKKQRKDSHKEVERRRRENINTAINVLSDLLP
Pho4           244         GALVDDDKRESHKHAEQARRNRLAVALHELASLIP
Lc             406         AQEMSGTGTKNHVMSERKRREKLNEMFLVLKSLLP
```

```
R-S        410        AQEMS--ATKNHVMSERKRREKLNEMFLVLKSLLP
Peru       378        AQE---NGAKNHVMSERKRREKLNEMFLVLKSLVP
Delila     433        AKPTADEIDRNHVLSERKRREKINERFLILKSLVP
Ra         1                           MSERRRREKLNEMFLILKSVVP
Clock      33                  KAKRVSRNKSEKKRRDQFNVLIKELGSMLP
BMAL1a     31                     REAHSQIEKRRRDKMNSFIDELASLVP
ARLC       411                 GTGTKNHVMSERKRREKLNEMFLVLKSLLP
AHR        20         KTVKPIPAEGIKSNPSKRHRDRLNTELDRLASLLP
HRF        16                  ELRKEKSRDAARCRRSKETEVFYELAHELP
HIF-3α     8                 SNTELRKEKSRDAARSRRQETEVLYQLAHTLP
trl        76                  ELRKEKSRDAARSRRGKENYEFYELAKMLP
EPAS-1     8           RSSSERRKEKSRDAARCRRSKETEVFYELAHELP
Sima       934                  KRKEKSRDAARCRRSKETEIFMELSAALP
Tango      9             KERFASRENHCEIERRRRNKMTAYITELSDMVP
Arnt2      64                 SRENHSEIERRRRNKMTQYITELSDMVP
NPA1       44                  AQRKEKSRNAARSRRGKENLEFFELAKLLP
NPA2       1        MDEDEKDRAKRASRNKSEKKRRDQFNVLIKELSSMLP
CTF4       311                 KERRMANNARERLRVRDINEAFKELGRMCQ
ATH1       158                 KQRRLAANARERRRMHGLNHAFDQLRNVIP
ESC1       333                 PELRTSHKLAERKRRKEIKELFDDLKDALP
INO2       235                 KVRKWKHVQMEKIRRINTKEAFERLIKSVR
INO4       45                   QIRINHVSSEKKRRELERAIFDELVAVVP
NDF1       91       MTKARLERFKLRRMKANARERNRMHGLNAALDNLRKVVP
NDF2       121         ARLERSKLRRQKANARERNRMHDLNAALDNLRKVVP
NDF3       91                  RSRRVKANDRERNRMHNLNAALDALRSVLP
RTG1       39                          PKDFFRDYYGISGSNDTLSESTP
POD1       80                     QRNAANARERARMRVLSKAFSRLKTTLP
Hxt        100                         PKKERRRTESINSAFAELRECIP
Hed        8                    KRRGTANRKERRRTQSINSAFAELRECIP
Id6        35                     KIPLLDEQMTMFLQDMNSCYSKLKELVP
SEF2       554       PEQKAEREKERRMANNARERLRVRDINEAFKELGRMVQ
E2FBP-1    247                    QAAAAQRAVRAQAAALEQLREKLESAEP
ABF-1      103            ECKQSQRNAANARERARMRVLSKAFSRLKTSLP
Mad3       50          QAPGALNSGRSVHNELEKRRRAQLKRCLEQLRQQMP
Mad4       50              KAPNNRSSHNELEKHRRAKLRLYLEQLKQLVP
Paraxis    68                VVRQRQAANARERDRTQSVNTAFTALRTLIP
Math5      33       PGRLESAARRRLAANARERRRMQGLNTAFDRLRRVVP
DERMO-1    61            FEELQSQRILANVRERQRTQSLNEAFAALRKIIP
Hand1      94                   RRKGSGPKKERRRTESINSAFAELRECIP
Hand2      97                 PVKRRGTANRKERRRTQSINSAFAELRECIP
```

```
                     <   LOOP        ><                 H2              >
HIF-1α        46     --------------LPHNVSSHLDKASVMRLTISYLRVRKLLDAGDLDIEDDMKAQM
HIF-1β        118    --------------TCSALARKPDKLTILRMAVSHMKSLRGTGNTSTDGSYKPSFLT
SIM           30     --------------LPAAITSQLDKASVIRLTTSYLKMRQVFPDGLG
c-Myc(Z)      384    ----------------ELENNEKPKVVILKKATAYILSVQAEEQKLISEEDLLRKRREQLKHKLEQL
N-Myc(Z)      510    ---------------ELVKNEKAAKVVILKKATEYVHSLQAEEHQLLLEKEKLQARQQQLLKKIEHA
L-Myc(Z)      310    ---------------TLASCSKAPKVVILSKALEYLQALVGAEKRMATEKRQLRCRQQQLQKRIAYL
V-Myc(Z)      352    ---------------EVANNEKAPKVVILKKATEYVLSLQSDEHKLIAEKEQLRRRREQLKHNLEQL
S-Myc(Z)      375    ---------- ----ELVHNEKAAKVVILKKATEYIHTLQTDESKLLVEREKLYERKQQLLEKIKQS
Max(Z)        52     ----------------SLQGEKASRAQILDKATEYIQYMRRKNHTHQQDIDDLKRQNALLEQQVRAL
Mad(Z)        85     ------LGPESSRHTTLSLLTKAKLHIKKLEDCDRKAVHQIDQLQREQRHLKRQLEKLG
Mxi1(Z)       59     ------LGPDCTRHTTLCLLNKAKAHIKKLEEAERKSQHQLENLEREQRFLKWRLEQLQG
AP4(Z)        60     ----------------HTDGEKLEKAAILQQTAEYIFSLEQEKTRLLQQNTQLKRFIQELSGSS
USF(Z)        228    ------------DCSMESTKSGQSKGGILSKACDYIQELRQSNHRLSEELQGLDQLQLDNDVLRQQVEDL
TFE3(Z)       168    --------------KSSDPEMRWNKGTILKASVDYIRKLQKEQQRSKDLESRQRSLEQANRSLQLRIQEL
TFEB(Z)       361    --------------KANDLDVRWNKGTILKASVDYIRRMQKDLQKSRELENHSRRLEMTNKQLWLRIQEL
TFEC(Z)       139    --------------KSNDPDIRWNKGTILKASVDYIKWLQKEQQRARELEHRQKKLEHANRQLWLRIQEL
FIP(Z)        172    ------------DCNADNSKTGASKGGILSKACDYIRELRQTNQRMQETFKEAERLQMDNELLRQQDIEL
ADD1(Z)       322    -----------------GTEAKLNKSAVLRKAIDYIRFLQHSNQKLEQENLTLRSAHKSKSLKD
Mi(Z)         233    --------------KSNDPDMRWNKGTILKASVDYIRKLQREQQRAKDLENRQKKLEHANRHLLLRVQEL
SREBP-1a(Z)   352    -----------------GTEAKLNKSAVLRKAIDYIRFLQHSNQKLKQENLSLRTAVHKSKSLKDLV
E12           578    --------------LHLNSEKPQTKLLILHQAVSVILNLEQQVRERNLNP
E47           365    --------------MHLKSDKAQTKLLILQQAVQVILGLEQQVRERNLNP
```

```
ITF2          548     -------------LHLKSDKPQTKLLILHQAVAVILSLEQQVRERNLNP
HEB           597     -------------LHLKSEKPQTKLLILHQAVAVILSLEQQVRERNLNP
Da            583     -------------THLKSDKPQTKLGILNMAVEVIMTLEQQVRERNLNP
Twist         482     ---------------TLPSDKLSKIQTLKLATRYIDFLCRMLSSSDISLLKA
L-Sc          112     ----QTVVNSLSNGGRGSSKKLSKVDTLRIAVEYIRGLQDMLDDGTASSTRH
Scute         128     ---QSIITDLTKGGGRGPHKKISKVDTLRIAVEYIRSLQDLVDDLNGGSNIG
Achaete       53      AAVIADLSNGRRGIGPGANKKLSKVSTLKMAVEYIRRLQKVLHE
Asense        188     ---EVSEAFEAQGAGRGASKKLSKVETLRMAVEYIRSLEKLLGFDFPP
Mash1         144     ---------------NGANKKMSKVETLRSAVEYIRALQQLLDEHDAVSAAFQ
Mash2         147     ---------------MGANKKLSKVETLRSAVEYIRALQRLLKEHDAVRAALS
MyoD          138     --------------S-NPNQRLPKVEILRNAIRYIEGLQALLRDQDAAP
Myogenin      110     --------------L-NPNQRLPKVEILRSAIQYIERLQALLSSLNQEER
Myf5          112     ---------------TNPNQRLPKVEILRSAIRYIESLQELLREQ
MRF4          122     ---------------ANPNQRLPKVEILRSAISYIERLQDLLHRLDQQEK
Lyl1/2        178/14 ---------------THPPDRLSKNEVLRLAMKYIGFLVRLLRDQTAVLTSGP
Tal1          215     --------------THPPDKKLSKNEILRLAMKYINFLAKLLNDQEEEGTQ
Tal2          30       --------------THPPDKKLSKNETLRLAMRYINFLVKVLGEQSLQQTT
Hen1/2        103/106--------------TLPPDKKLSKIEILRLAICYISYLNHVLDV
Hes1(D)       62      --------LDALKKDSSRHSKLEKADILEMTVKHLRNLQRAQHTAALSTDP
Hes2(D)       41      --------LPLLGAETSRYSKLEKADILEMTVRFLREQPASVCSTEAP
Hes5(D)       44      ----------QEFARHQPNSKLEKADILEMAVSYLKHSKAFAAAAGP
E(spl)M5(D)   45      -----------EFQGDDAILRMDKAEMLEAALVFMRKQVVKQQAPVS
E(spl)M7(D)   41      -----------ECVAQTGDAKFEKADILEVTVQHLRKLKESKKHVP
E(spl)M8(D)   38      -----------ELRGDDGILRMDKAEMLESAVIFMRQQKTPKKV
Hairy(D)      59      ---------DATKKDPARHSKLEKADILEKTVKHLQELQRQQAAM
Dpn(D)        69      ---------EAMKKDPARHTKLEKADILEMTVKHLQSVQRQQLNM
Emc(D)        62      --------------FMPKNRKLTKLEIIQHVIDYICDLQTELETH
Id1(D)        103     --------------TLPQNRKVSKVEILQHVIDYIRDLQLELNS
Id2(D)        104     --------------SIPQNKKVTKMEILQHVIDYILDLQIALDS
Id3(D)        57      --------------GVPRGTQLSQVEILQRVIDYILDLQVVLAEP
Id4(D)        80      --------------TIPPNKKVSKVEILQHVIDYILDLQLALETH
Cbf1          251     ------------------VRESSKAAILARAAEYIQKLKETDEANIEK
Pho4          279     ----------AEWKQQNVSAAPSKATTVEAACRYIRHLQQNGST
Lc            441     -----------------SIHRVNKASILAETIAYLKELQRRVQELES
R-S           443     -----------------SIHRVNKASILAETIAYLKELQRRVQELES
Peru          410     -----------------SIHKVDKASILAETIAYLKELQRRVQELES
Delila        468     -----------------SGGKVDKVSILDHTIDYLRGLERKVDELES
Ra            23      -----------------SIHKVDKASIFAETIAYLKELEKRV
Clock         63      ----------------GNARKMDKSTVLQKSIDFLRKHKETTAQSDASE
BMAL1a        58      -------------TCNAMSRKLDKLTVLRMAVQHMRTLRGA
ARLC          441     -----------------SIHRVNKASILAETIAYLKELQRRVQELE
AHR           55      -------------FPQDVINKLDKASVLRLSVSYLRAKSFFDVALKSSPTERN
HRF           47      -------------LPHSVSSHLDKASTMRLAISFLRTHKLLSSVCSENESEA
HIF-3α        40      -------------FARGVSAHLDKASIMRLTISYLRMHRLCAAGEWNQ
trl           106     -------------LPAAITSQLDKASIIRLTISYLKLRDFSGHGDPPWTREAS
EPAS-1        42      -------------LPHNVSSHLDKASIMRLEISFLRTHKLLSSVCSENESEAEAD
Sima          963     -------------LKTDDVNQLDKASVMRITIAFLKIREMLQFVP
Tango         42      -------------TCSALARKPDKLTILRMAVAHMKALRGTGNTS
Arnt2         92      -------------TCSALARKPDKLTILRMAVSHMKSMRGTGN
NPA1          74      -------------LPGAISIQLDKASIVRLSVTYLRLRRFAALGAP
NPA2          38      ----------------GNTRKMDKTTVLEKVIGFLQKHNEVSAQTEIC
CTF4          341     -------------LHLKSEKPQTKTLILHQAVAVILSLEQQVR
ATH1          189     --------------SFNNDKKLSKYETLQMAQIYINALSELL
ESC1(Z)       363     ---------------LDKSTKSSKWGLLTRAIQYIEQLKSEQVALEAYVKSLE
INO2          265     ------------TPPKENGKRIPKHILLTCVMNDIKSIRSANE
INO4          74      --------------DLQPQESRSELIIYLKSLSYLSWLYERNEK
NDF1          130     --------------CYSKTQKLSKIETLRLAKNYIWALSEILRSG
NDF2          150     --------------CYSKTQKLSKIETLRLAKNYIWALSEILRSGKRP
NDF3          121     --------------SFPDDTKLTKIETLRFAYNYIWALAETLRLAD
RTG1          62      ---GALGLSSKAKGTGTKDGKPNKGQILTQAVEYISHLQNQVDTQ
POD1          108     --------------WVPPDTKLSKLDTLRLASSYIAHLRQILAND
Hxt           123     --------------NVPADTKLPKIKTLRLATSYIAYLMDLDLKDAQ
Hed           37      --------------NVPADTKLSKIKTLRLATSYIAYLMDLLAKDDQNG
Id6           63      --------------TLPTNKKASKMEILQHVIDYIWDLQVELESKK
SEF2          592     -------------LHLKSDKPQTKLLILHQAVAVILSLEQQVR
E2FBP1        266     ------PEKKMALVADEQQRLMQRALQQNFLAMAAQLPM
ABF-1         136     --------------WVPPDTKLSKLDTLRLASSYIAHLRQLLQED
```

```
Mad3        86     ------LGVDCTRYTTLSLLRRARVHIQKLEEQEQQARRLKEK
Mad4        82     ------LGPDSTRHTTLSLLKRAKVHIKKLEEQDRRALSIKEQ
Paraxis     99     --------------TEPVDRKLSKIETLRLASSYIAHLANVLLLG
Math5       70     --------------QWGQDKKLSKYETLQMALSYIIALTRILAE
DERMO-1     95     ---------------TLPSDKLSKIQTLKLAARYIDFL
Hand1       123    --------------NVPADTKLSKIKTLRLATSYIAYLMDV
Hand2       128    --------------NVPADTKLSKIKTLRLATSYIAYLMDLLA
```

Key:
(Z): bHLHZ (leucine zipper) proteins.
(D): HLH proteins lacking a functional DNA-binding domain.
* - residue conferring specificity for CACGTG or CAGCTG E-box motif.

HIF-1α: Hypoxia-inducible factor-1α (*Homo sapiens*) [373]
Ra: transcriptional activator Ra homologue (*Oryza longistaminata*) [374]
Clock: Clock protein (*Mus musculus*) [290, 295]
BMAL1a: Brain and muscle Ah receptor nuclear translocator-like protein (*Homo sapiens*) [301]
ARLC: Anthocyanin regulatory LC protein, (*Zea mays*) [375]
AHR: Aromatic hydrocarbon receptor (*Homo sapiens*) [211]
HRF: HIF-related factor (*Homo sapiens*) [304]
HIF-3α: Hypoxia-inducible factor-3α [326]
trl: trachealess protein (*Drosophila*) [298]
EPAS-1: Endothelial PAS domain protein 1 (*Homo sapiens*) [327]
Sima: Similar protein (*Drosophila melanogaster*) [299]
Tango: Tango protein (*Drosophila melanogaster*) [376]
Arnt2: Arnt2 (*Mus musculus*) [297]
NPA1 and NPA2: Neuronal PAS domain proteins 1 and 2 (*Homo sapiens*) [328]
CTF4: Chicken transcription factor 4 (*Gallus gallus*) [377]
ATH: Atonal protein homologue 1 (*Homo sapiens*) [378]
ESC1: ESC1 protein (*Schizosaccharomyces pombe*) [379]
INO2: INO2 protein (*Saccharomyces cerevisiae*) [380]
INO4: INO4 protein (*Saccharomyces cerevisiae*) [381]
NDF1: Neurogenic differentiation factor 1 (*Homo sapiens*) [239]
NDF2: Neurogenic differentiation factor 2 (*Homo sapiens*) [240]
NDF3: Neurogenic differentiation factor 2 (*Homo sapiens*) [240]
RTG1: Retrograde regulation protein 1 (*Saccharomyces cerevisiae*) [382]
POD1: Mesoderm specific protein (*Mus musculus*) [383]
Hxt: Hxt protein (*Mus musculus*) [384]
Hed: Hed protein (*Mus musculus*) [384]
Id6: Id protein homologue Id6 (*Danio rerio*) [385]
SEF2: SEF2 protein (*Mus musculus*) [386]
E2FBP1: E2F binding protein (*Homo sapiens*) [387]
ABF-1: ABF-1 protein (*Homo sapiens*) [388]
Mad3 and Mad4: Mad3 (*Mus musculus*) and Mad4 (*Homo sapiens*) [253]
Paraxis: (*Mus musculus*) [389]
Math5: (*Mus musculus*) [390]
DERMO-1: Twist-related protein DERMO-1 (*Mus musculus*) [391]
Hand1: (*Homo sapiens*) [392]
Hand2: (*Homo sapiens*) [393]

Residues in **bold** are fully or semi-conserved residues of the HLH proteins: **blue** = basic, **red** = acidic, **green** = hydrophobic. The following groups of amino acids are considered similar (A,C,F,I,L,M,V); (Y,F,H); (D,E); (Q,N); (K,R); (S,T).

3.2.1.2. Modelling of helix backbones

Homology modelling was performed using SYBYL 6.4.2 (Tripos Inc.) and QUANTA 97 (QUANTA Version 97, 1997) on a Silicon Graphics Indy R4000. Figure 3.7 shows the sequence alignment of the bHLH regions of HIF-1α and 1β with the proteins with known 3-D structures and Table 1 shows their percentage sequence identities and similarities (defined in Figure 3.7).

```
                                        <         H1      >
HIF-1α        11       KKISSERRKEKSRDAARSRRSKESEVFYELAHQLP
HIF-1β        82      SSADKERLARENHSEIERRRRNKMTAYITELSDMVP
MyoD          105       TTNADRRKAATMRERRRLSKVNEAFETLKRCTS
Max           19         QSAADKRAHHNALERKRRDHIKDSFHSLRDSVP
E47(ITF1)     332       LRDRERRMANNARERVRVRDINEAFRELGRMCQ
USF           195        TRDEKRRAQHNEVERRRRDKINNWIVQLSKIIP
Pho4          244      GALVDDDKRESHKHAEQARRNRLAVALHELASLIP

                      <      L     ><       H2         >
HIF1-α        46      ---LPHNVSSHLDKASVMRLTISYLRVRKLLDAGDLDIEDD
HIF1-β        118     ---TCSALARKPDKLTILRMAVSHMKSLRGTGNTSTDGSYKPS
MyoD          138     -----SNPNQRLPKVEILRNAIRYIEGLQALLRDQDAAPPGA
Max           52      -----SLQGEKASRAQILDKATEYIQYMRRKNHTHQQDIDDLKRQ
E47(ITF1)     365     ---MHLKSDKAQTKLLILQQAVQVILGLEQQVRERNLNP
USF           228     -DCSMESTKSGQSKGGILSKACDYIQELRQSNHRLSEEL
Pho4          279     AEWKQQNVSAAPSKATTVEAACRYIRHLQQNGST
```

Figure 3.7. Sequence alignment of HIF-1α and HIF-1β with the five proteins whose 3-D structures have been solved to date.

Key: HIF1-α: Hypoxia-inducible factor-1α [213], SIM: Single-minded protein, *Drosophila Melanogaster* [208], HIF 1-β: Arnt, human aryl hydrocarbon receptor nuclear translocator protein [209,210], MyoD: Myogenic factor mouse protein [394], Max: Max human protein [250], E47: E47 Insulin transcription factor human protein [215,395], USF: Upstream stimulatory factor human protein [396], Pho4: Regulatory protein Pho4, *Saccharomyces cerevisiae* [397]. Residues in **bold**: fully or semi-conserved residues of the HLH proteins. **Blue**: basic, **red**: acidic, **green**: hydrophobic. The following groups of amino acids are considered similar (F,H,I,L,M,V,Y); (D,E); (Q,N); (K,R); (A,G); (S,T). Pair-wise identity was calculated for the number of identities between the 31 residues aligned for the H1 helices of HIF-1α and HIF-1β and the 20 residues of the H2 helices. Percentage similarity was calculated using the number of pair-wise identical and similar residues.

	HIF-1α				HIF-1β/Arnt			
	Helix 1(H1)		Helix 2(H2)		Helix 1(H1)		Helix 2(H2)	
USF	25.8	38.7	23.8	33.3	41.9	61.3	20.7	31.0
Max	19.4	32.3	18.8	34.4	32.3	45.2	23.8	47.6
MyoD	25.8	29.0	27.8	50.0	19.4	32.3	28.6	50.0
E47	22.6	29.0	4.8	38.1	25.8	35.5	33.3	38.1
Pho4	25.8	51.6	20.0	40.0	25.8	58.1	25.0	50.0

Table 3.1. Sequence identity and conserved residue similarity (shaded grey) for HIF-1α and β and the five proteins with crystal structures measured over helix 1 (31 residues) and helix 2 (21 residues).

USF has the highest sequence identity over H1 to both HIF-1α and HIF-1β (25.8% and 41.9%, respectively) and conserved residue similarity to HIF-1β (61.3%). Pho4 has the highest similarity to H1 of HIF-1α (51.6%). USF also has the highest sequence identity and conserved residue similarity to HIF-1β over H2 (20.7% and 31.0%, respectively). Max has the highest sequence identity and conserved residue similarity to HIF-1α over H2. From this analysis USF (residues D197-P227) was chosen as the basic template on which to model HIF-1α over H1 (residues S15-P45) and HIF-1β over H1 (residues E87-P117). It was thought more suitable to model both H1s on one protein rather than two separate ones, so Pho4 was not chosen for the modelling.

Even though there was high sequence similarity between the H2s of the HIF monomers and Max and USF, the latter two are zipper (b/HLH/Z) proteins possessing a heptad leucine residue-repeat motif directly following H2 which is important in dimerisation (see Figure 3.2). The presence of these leucine residues therefore constrains the H2/Z region in an interaction. HIF-1 is not a zipper protein and the two H2 regions are not so constrained to interact. Out of MyoD and E47, MyoD has the highest sequence identity to HIF-1α H2 (27.8%) and conserved residue similarity (50%) and also the highest conserved residue similarity to HIF-1β H2 (50%). The modelling of the H2s was therefore based on MyoD (MyoD residues K146-D166 used to model K56-G76 of HIF-1α H2 and K128-T148 of HIF-1β H2). Modelling was carried out by direct superposition of the amino acid sequences on the corresponding protein backbone atoms of the crystal structures after superposing similar residues matched from the alignments.

3.2.1.3. Modelling of loop regions

None of the loop regions of the HLH crystal structures have good similarity in sequence or length to the loops in the HIF-1 monomers. The loops of MyoD and Max are two residues shorter than the 10-residue loops of HIF-1 and the loops of E47 and USF are two residues longer. A fragment search of the protein structural database in QUANTA and Brookhaven Protein Database using the HIF-1 loops yielded no satisfactory hits. As the loops show no conserved sequence and considerable variability in length, (Figure 3.6), it was decided to model the loops of both monomers (P45-D55 of HIF-1α and P117-D127 of HIF-1β) on the loop of USF (P227-S239). This was achieved by superposition of residues P45-H53 of HIF-1α and P117-K125 of HIF-1β onto the α-carbons of USF P227-K235 and residues L54-D55 (HIF-1α) and P126-D127 (HIF-1β) onto the α-carbons of USF Q239-S240. The gaps were joined and the loop minimised to convergence keeping all atoms in the helix fixed (200 iterations of Steepest Descent (SD), followed by the Adopted Basis Set Newton-Raphson (ABNR) algorithm).

3.2.1.4. Side chain modelling

Side chains were added directly to the backbones of the monomers, polar hydrogens added and the monomer models minimised. Minimisation of side-chain structures (200SD) was followed by minimisation of the entire structure to convergence (200SD and ABNR). Dimerisation of the monomers was achieved by using the USF homodimer as a template and superposing the conserved hydrophobic residues in the dimer interface of HIF-1. The protein backbone of the heterodimer was fixed and the side chains relaxed again, (200SD), followed by ABNR minimisation to energy convergence. The SYBYL Biopolymer/Analyse Protein and QUANTA Protein Health options were used to validate the model by checking for close contacts, bond lengths/angles, chirality and buried/exposed residues in the dimer. The root mean square difference of the Cα trace and backbone (excluding the loop region) was calculated between the HIF-1 model and the corresponding regions of the USF and MyoD proteins.

3.2.1.5. DNA-bound model

To obtain a model structure of HIF-1 complexed to DNA, the crystal structure of DNA-bound USF was taken and the residues of the bottom helices mutated to those of the HIF-1 sequence. It was decided not to 'place' the model of the HIF-1 dimer using computer graphics onto the DNA by directly copying the DNA co-ordinates of USF, as any small irregularities which may have arisen in the model could lead to spurious results regarding DNA-binding conformations. Moreover, severe steric clashes could occur. It was more representative of the true side-chain conformations of HIF-1 if USF was mutated. The DNA sequence of USF was also mutated to a sequence based on the erythropoeitin 3' enhancer sequence (5'-GGGCCCTACGTGCTGTCTCACACAGC-3').

3.2.2. Peptide design, synthesis and binding studies

3.2.2.1. Materials

Reagents were purchased from or kindly donated by the suppliers listed below:

Affinity Sensors: Carboxymethyl-dextran cuvettes, NHS Coupling Kit (containing NHS (0.2g), EDC (1.15g) and 1M ethanolamine, pH8.5 (25ml)).

Aldrich Chemical Co. Ltd., Gillingham, Dorset, England. DIC, (99%), DMF (99.9+%, HPLC grade), 4Å molecular sieves (1/8" beads), piperidine (99.5+%, redistilled), TES.

Amersham Life Sciences. Rainbow molecular weight marker (RPN 755).

BDH Chemicals Ltd., Poole, Dorset, England. Acetic acid (glacial), AMPS, Bromophenol Blue, dichloromethane, diethyl ether, ethanol, methanol (AnalaR grade), NNN'N'-tetramethylethylenediamine (Temed, 'Electran'), sodium chloride (GPR), sodium dodecyl sulphate, tris(hydroxymethyl)methylamine (AnalaR grade)

Calbiochem-Novabiochem (UK) Ltd. OPfp esters of Fmoc-protected amino acids: L-Ala, L-Ile, Gly, L-Leu, L-Met, L-Nle, D-Phe, L-Phe, L-Pro, L-Tyr (^{t}Bu), L-Val.

Fmoc-protected amino acids: β-L-Ala-OH, L-Cys (Trt)-OH, D-Leu-OH, D-NVal-OH, L-Ser (^{t}Bu)-ODhbt, D-Tyr (^{t}Bu)-OH.

Chiron Technologies Pty Ltd., Clayton, Victoria, Australia. GAP Cleavable Multipin Peptide Synthesis Kit.

Lancaster Synthesis Ltd., Eastgate, Morecambe, Lancashire, England. NBD-Cl.

National Diagnostics, Hull, England. Ultrapure Protogel (30% (w/v) acrylamide: 0.8% (w/v) bisacrylamide stock solution (37.5:1), protein sequencing and electrophoresis grade).

Pierce: Tween 20.

Promega. *In vitro* transcription-translation kit for E12 and E47 production (consisting of DNA template (1μg/ml), RNAase inhibitors (40u/μl), DTT, (100mM), TNT buffer, reticulocyte lysate, amino acid solution minus methionine (1mM), T7 polymerase, L-[^{35}S]methionine)

Rathburn Chemicals Ltd., Walkerburn, Scotland. DMF, (peptide synthesis grade). DMF was purified before use with 4Å molecular sieves (1/8" beads), which had been activated by heating overnight and allowing to cool in an evacuated vacuum dessicator over anhydrous $CuSO_4$. The molecular sieves (approximately 200g) were then added to 2.5 litres of peptide synthesis grade DMF, left in the dark for 3 days, shaken and left for a further 4 days to stand before being decanting the solvent off for use.

Sigma Chemical Co. Ltd., St. Louis, USA. 6-AHA, GSH, PBS, HOBt, TFA.

3.2.2.2. Peptide design

The structural molecular models of HIF-1 and Id3-E47 [358] were analysed and compared to other HLH structures and sequence alignments to identify which hydrophobic residues were important in stability at the dimer interface (at the top of helix 1, bottom of helix 2, see Results section). Peptides were designed to bind to HIF-1α or Id3 with the aim of preventing dimerisation with their respective partners Arnt and E47. After identifying the important hydrophobic residues in these partners, the rest of the monomer was removed using computer graphics keeping the correct orientations and in-space conformations of the hydrophobic residues. Amino acid linker groups were selected manually to join these residues, thus forming a peptide containing all the essential hydrophobic amino acids, but without the long flexible loop linkers joining the top of helix 1 to the bottom of helix 2 in the parent protein. Once a peptide lead had been designed, variations were introduced in a knowledge-based, semi-combinatorial manner, by substituting the hydrophobic groups for related others, e.g. phe→tyr, ile→val, leu→norvaline, changing the stereochemistry and/or varying the linker, e.g. ser→gly.

Figure 3.8 lists the 48 first-generation lead-test peptides synthesised (all were 15-mers). Peptide 1(1) is the peptide most closely resembling the hydrophobic residues of Arnt. Since the hydrophobic residues of E47 are so similar to those of Arnt, variants of peptide 1(1) should also bind Id3. Two random peptides, E and F, were purchased to be used in binding studies also.

1	1(1)	D-Phe	L-Pro	D-Leu	βAla	L-Ile	L-Leu	Gly	Gly	L-Val	L-Pro	L-Val	L-Ile	L-Leu	L-Ser	L-Leu
2	2(1)	D-Phe	L-Pro	D-Leu	βAla	D-Leu	L-Leu	Gly	Gly	L-NLe	L-Pro	L-Val	L-Leu	L-Ile	L-Ser	L-Leu
3	3(1)	D-Phe	L-Pro	D-Leu	βAla	L-Leu	L-Leu	Gly	L-Ser	D-NV	L-Pro	L-Met	L-Val	L-Leu	L-Ser	L-Leu
4	4(1)	D-Phe	L-Pro	D-Leu	βAla	L-Val	L-Leu	Gly	L-Ser	L-Leu	L-Pro	L-Met	D-Leu	L-Ile	L-Ser	L-Val
5	5(1)	D-Phe	L-Pro	D-Leu	βAla	L-Met	D-Leu	Gly	L-Ser	L-Ile	L-Pro	D-NV	L-Nle	L-Leu	L-Ser	D-NV
6	6(1)	D-Phe	L-Pro	D-Leu	βAla	L-NLe	D-Leu	L-Ser	Gly	D-Leu	L-Pro	D-NV	L-Val	L-Ala	Gly	D-Leu
7	7(1)	D-Phe	L-Pro	D-Leu	βAla	L-Ile	D-Leu	L-Ser	Gly	L-Val	L-Pro	L-Tyr	L-Leu	D-NV	Gly	L-Ile
8	8(1)	D-Phe	L-Pro	L-Ile	βAla	D-Leu	D-Leu	L-Ser	Gly	D-Leu	L-Pro	L-Tyr	L-Ile	L-Ser	Gly	D-Leu
9	9(1)	D-Phe	L-Pro	L-Ile	βAla	L-Leu	D-Leu	L-Ser	L-Ser	L-NLe	L-Pro	L-NLe	D-Leu	Gly	Gly	L-Leu
1	10(1)	D-Phe	L-Pro	L-Ile	βAla	L-Val	D-Leu	L-Ser	L-Ser	L-Ile	L-Pro	L-NLe	L-NLe	D-Leu	Gly	L-Met
1	11(1)	D-Phe	L-Pro	L-Ile	βAla	L-Met	D-Leu	L-Ser	L-Ser	D-NV	L-Pro	L-Leu	L-Ile	L-Val	Gly	L-NLe
1	12(1)	L-Phe	L-Pro	L-Ile	βAla	L-NLe	D-Leu	Gly	Gly	L-Leu	L-Pro	L-Leu	L-Val	L-Ile	Gly	L-Val
1	1(2)	L-Phe	L-Pro	L-Ile	βAla	L-Ile	L-Ile	Gly	Gly	D-Leu	L-Pro	L-Ile	L-Val	L-Leu	Gly	D-NV
1	2(2)	L-Phe	L-Pro	L-Ile	βAla	D-Leu	L-Ile	Gly	Gly	L-Val	L-Pro	L-Ile	L-Leu	L-Val	Gly	L-Ile
1	3(2)	L-Phe	L-Pro	L-Ile	βAla	L-Leu	L-Ile	Gly	L-Ser	L-Leu	L-Pro	L-Val	D-Leu	L-Ser	Gly	D-Leu
1	4(2)	L-Phe	L-Pro	L-Nle	βAla	L-Val	L-Ile	Gly	L-Ser	D-NV	L-Pro	L-Val	L-NLe	D-NV	Gly	L-Leu
1	5(2)	L-Phe	L-Pro	L-Nle	βAla	L-Met	L-Ile	Gly	L-Ser	L-Ile	L-Pro	L-Met	L-NLe	L-Ala	L-Ser	L-Met
1	6(2)	L-Phe	L-Pro	L-Nle	βAla	L-NLe	L-Ile	L-Ser	Gly	L-NLe	L-Pro	L-Met	D-Leu	Gly	L-Ser	L-Val
1	7(2)	L-Phe	L-Pro	L-Nle	βAla	L-Ile	L-Ile	L-Ser	Gly	L-Val	L-Pro	D-NV	L-Ile	D-Leu	L-Ser	L-Ile
2	8(2)	L-Phe	L-Pro	L-NLe	βAla	D-Leu	L-Ile	L-Ser	Gly	D-Leu	L-Pro	D-NV	L-Leu	D-NV	L-Ser	D-NV
2	9(2)	L-Phe	L-Pro	L-Nle	βAla	D-Leu	D-NV	L-Ser	L-Ser	L-Ile	L-Pro	L-Tyr	L-Val	L-Val	L-Ser	L-NLe
2	10(2)	L-Phe	L-Pro	L-Nle	βAla	L-Val	D-NV	L-Ser	L-Ser	D-NV	L-Pro	L-Tyr	L-NLe	D-Leu	L-Ser	D-Leu
2	11(2)	L-Phe	L-Pro	L-NLe	βAla	L-Met	D-NV	L-Ser	L-Ser	L-Leu	L-Pro	L-Nle	D-Leu	L-Ser	L-Ser	L-Val
2	12(2)	D-Tyr	L-Pro	L-Leu	βAla	L-NLe	D-NV	Gly	Gly	L-NLe	L-Pro	L-Nle	L-NLe	Gly	L-Ser	L-Ile
2	1(3)	D-Tyr	L-Pro	L-Leu	βAla	L-Ile	D-NV	Gly	Gly	L-Leu	L-Pro	L-Leu	L-Leu	L-Ala	L-Ser	D-NV
2	2(3)	D-Tyr	L-Pro	L-Leu	βAla	D-Leu	D-NV	Gly	Gly	L-NLe	L-Pro	L-Leu	L-Val	D-NV	L-Ser	L-Met
2	3(3)	D-Tyr	L-Pro	L-Leu	βAla	L-Leu	D-NV	Gly	L-Ser	D-NV	L-Pro	L-Ile	L-Nle	L-Val	L-Ser	L-Leu
2	4(3)	D-Tyr	L-Pro	L-Leu	βAla	L-Val	D-NV	Gly	L-Ser	L-Val	L-Pro	L-Ile	D-Leu	L-Ile	Gly	L-NLe
2	5(3)	D-Tyr	L-Pro	L-Leu	βAla	L-Met	L-NLe	Gly	L-Ser	D-Leu	L-Pro	L-Val	L-Ile	D-Leu	Gly	L-Met
3	6(3)	D-Tyr	L-Pro	L-Leu	βAla	L-NLe	L-NLe	L-Ser	Gly	L-Ile	L-Pro	L-Val	L-Leu	L-Ala	Gly	L-NLe
3	7(3)	D-Tyr	L-Pro	L-Leu	βAla	L-Ile	L-Nle	L-Ser	Gly	L-Val	L-Pro	L-Met	L-Val	L-Ser	Gly	L-NLe
3	8(3)	D-Tyr	L-Pro	L-Val	βAla	D-Leu	L-NLe	L-Ser	Gly	L-Leu	L-Pro	L-Met	L-Nle	Gly	Gly	D-Leu
3	9(3)	D-Tyr	L-Pro	L-Val	βAla	L-Leu	L-NLe	L-Ser	L-Ser	D-NV	L-Pro	D-NV	D-Leu	L-Val	Gly	L-Ile
3	10(3)	D-Tyr	L-Pro	L-Val	βAla	L-Val	L-NLe	L-Ser	L-Ser	L-NLe	L-Pro	D-NV	L-Ile	L-Leu	Gly	L-Met
3	11(3)	D-Tyr	L-Pro	L-Val	βAla	L-Met	L-NLe	Gly	L-Ser	D-Leu	L-Pro	L-Tyr	L-Leu	L-Ile	Gly	D-NV
3	12(3)	L-Tyr	L-Pro	L-Val	βAla	L-NLe	L-NLe	Gly	Gly	L-Ile	L-Pro	L-Tyr	L-Val	D-NV	Gly	L-Val
3	1(4)	L-Tyr	L-Pro	L-Val	βAla	L-Ile	L-Let	Gly	Gly	D-NV	L-Pro	L-NLe	L-NLe	L-Ala	Gly	L-Met
3	2(4)	L-Tyr	L-Pro	L-Val	βAla	D-Leu	L-Met	Gly	Gly	L-Val	L-Pro	L-NLe	D-Leu	L-Ser	Gly	D-NV
3	3(4)	L-Tyr	L-Pro	L-Val	βAla	L-Leu	L-Met	Gly	L-Ser	L-NLe	L-Pro	L-Leu	L-Ile	Gly	L-Ser	D-Leu
4	4(4)	L-Tyr	L-Pro	L-Val	βAla	L-Val	L-Met	Gly	L-Ser	D-Leu	L-Pro	L-Leu	L-Leu	D-Leu	L-Ser	L-Ile
4	5(4)	L-Tyr	L-Pro	L-Met	βAla	L-Met	L-Met	L-Ser	L-Ser	L-Ile	L-Pro	L-Ile	L-Val	L-Ser	L-Ser	L-Val
4	6(4)	L-Tyr	L-Pro	L-Met	βAla	L-NLe	L-Met	L-Ser	Gly	L-Leu	L-Pro	L-Ile	L-NLe	L-Leu	L-Ser	L-NLe
4	7(4)	L-Tyr	L-Pro	L-Met	βAla	L-Ile	L-Met	L-Ser	Gly	D-Leu	L-Pro	L-Tyr	D-Leu	D-NV	L-Ser	L-Met
4	8(4)	L-Tyr	L-Pro	L-Met	βAla	D-Leu	L-Met	L-Ser	Gly	L-Ile	L-Pro	D-NV	L-Ile	L-Val	L-Ser	L-Leu
4	9(4)	L-Tyr	L-Pro	L-Met	βAla	L-Leu	L-Leu	L-Ser	L-Ser	L-Val	L-Pro	L-NLe	L-Leu	L-Ile	L-Ser	D-Leu
4	10(4)	L-Tyr	L-Pro	L-Met	βAla	L-Val	L-Leu	L-Ser	L-Ser	D-NV	L-Pro	L-Met	L-Val	L-Ala	L-Ser	L-Ile
4	11(4)	L-Tyr	L-Pro	L-Met	βAla	L-Met	L-Leu	Gly	L-Ser	L-Leu	L-Pro	L-Val	L-Ile	Gly	L-Ser	D-NV
4	12(4)	L-Tyr	L-Pro	L-Met	βAla	L-NLe	L-Leu	L-Ser	Gly	L-NLe	L-Pro	L-Leu	L-Ile	D-Leu	L-Ser	L-NLe

E	H-D-Phe-Pro-Arg-4MβNA
F	Z-Lys-Phe-Arg-pNA

Figure 3.8. The 48 peptides synthesised.

3.2.2.3. Peptide synthesis

Syntheses were carried out manually according to standard manufacturer's protocols on 48 crowns with the grafted surface esterified with Fmoc-glycine. The whole synthesis was carried out at room temperature

3.2.2.3.1. First residue attachment

The block of crowned pins was immersed in a bath of 20% (v/v) piperidine in DMF for 30 minutes to remove the Fmoc group, removed from the bath, excess liquid shaken off and the pins washed in a DMF bath for 2 minutes, followed by washing in a methanol bath. The methanol washing step was repeated twice with fresh methanol. The block was removed and allowed to dry in an acid-free fume cupboard for a minimum of 30 minutes.

3.2.2.3.2. Coupling procedure

Coupling was achieved using a solution of either the activated amino acid ester (OPfp) (100mM) with HOBt (120mM) or the free amino acid (100mM) with HOBt (120mM)/ DIC (100mM) in DMF. DIC serves to activate the non-esterified amino acids. Bromophenol Blue (10mM) DMF was added so that its concentration in the amino acid solution was 0.05mM. Bromophenol Blue turns blue in the presence of free amine groups, indicating the progress of the coupling. Activated amino acids (150µl) were dispensed into the appropriate wells of the polypropylene reaction trays and the block of Fmoc-deprotected pins placed in the wells. The block was placed under a polypropylene container (to minimise evaporation losses and contamination) and left for a minimum of 5 hours. After this time, if there were any crowns displaying a blue colour, the coupling step was repeated and left, if necessary, overnight.

On completion of coupling, the block was washed in a methanol bath with agitation for 5 minutes, allowed to air-dry for 2 minutes and washed in a DMF bath with agitation for 5 minutes. The Fmoc deprotection, washing, coupling and washing steps were repeated to build all 48 15-mer peptides.

3.2.2.3.3. Fluorescent labelling

The 16 peptides 1-4(1), 1-4(2), 1-4(3) and 1-4(4) (Figure 3.8) were labelled at their N-termini with the fluorogenic reagent NBD-Cl. Each was Fmoc-deprotected as above, immersed in NBD-Cl (800µl of 0.040M in 3% triethylamine in DMF) for 48hours [398], washed in DMF and of 0.040M in 3% triethylamine in DMF) for 48hours [398], washed in DMF and methanol and air-dried ready for side-chain deprotection and cleavage from the pins.

3.2.2.3.4. Side-chain deprotection

If side-chain deprotection did not proceed immediately, the blocks of peptides were stored in sealed containers with a dessicant in a cold room (4°C). Deprotection of the Cys(Trt) Tyr (^{t}Bu) and Ser (^{t}Bu) amino acids was achieved by treating the pins with DCM:TFA:TES (9:5:1) for 2.5hr at room temperature. The TES is a cation scavenger in this reaction, which produces only volatile co-products [399]. This is an improvement introduced in this laboratory to the pin methodology. On removal from the deprotection reaction bath, the pins were fully immersed in methanol for 10 min, in 0.5% glacial acetic acid in methanol/water (1:1 v/v) for 1hr and finally washed in water for 5min prior to cleavage.

3.2.2.3.5. Cleavage

The peptides were cleaved from their supports in 0.1M NaOH in 40% acetonitrile/water to produce peptides with a free acid at the C-terminus and glycine at the N-terminus. This was achieved in deep-well polystyrene, cleavage blocks with 730µl buffer per pin for 1.5-2hr. The cleavage solution was then neutralised with 2M NaH_2PO_4 (70µl), which gave a final pH of about 7, and gently mixed. The peptide solutions were transferred to tubes, sealed and stored at –80°C.

3.2.2.3.6. Characterisation

Three representative peptides were characterised by mass spectrometry.

3.2.2.4. Peptide binding studies: SDS-PAGE assays

Preliminary studies of the abilities of the peptides for Id3-binding and prevention of Id3-E47 dimerisation were carried out using two techniques, SDS-PAGE pull-down assays and a resonant mirror biosensor. The E47-related HLH protein E12 also binds to Id3 and was used in one of the biosensor assays.

3.2.2.4.1. Gel preparation and electrophoresis conditions

The following reagents were mixed to provide the gel and stacking gel:

	Running Gel:	Stacking Gel:
Sterilised H_2O:	2.3ml	6.80ml
Protogel:	5.0ml	1.70ml
1.5MTris-HCl pH8.8:	2.5ml	1.0MTris-HCl pH8.0: 1.25ml
10% SDS	100μl	100μl
10% AMPS	100μl	100μl
Temed ('Electran')	4μl	10μl

Sample Buffer (dye)	Volume (μl)
10% SDS:	2
Glycerol	1
DTT	1
1.0M Tris-HCl pH6.8	600
0.001% Bromophenol blue	5
Sterilised H_2O	5

Running gel was made first, AMPS and TEMED being added last to catalyse gel formation. An aliquot (6ml) of the resulting mixture was loaded between two glass plates with 200μl of 0.1%SDS evenly applied to the top to help settling. The gel was allowed to set, excess water poured off the top and stacking gel made in the same way. The stacking gel serves to concentrate the applied sample before it enters the running gel. Approximately 2ml stacking gel was evenly applied to the top of the running gel and dispensing combs inserted. The gel was left to set at room temperature for 30 min before use, or wrapped in wet paper towels and cling-film and stored at 0-4°C for up to 4 days. Radiolabelled samples to be run on the gel were prepared (see below), sample buffer (dye) added (20μl) and the samples incubated at 100°C for 5min. The samples (10μl) were loaded together with molecular weight reference markers, immersed in electrolyte buffer (10% Tris-glycine-SDS) and a 130V potential applied across the gel for approximately 1.5-2hr. The gel was then cut away from the glass plates, washed and 'fixed' in 10% methanol/10% acetic acid for 20-30min and finally washed in 'Amplifier' for 20min before drying. An autoradiograph was then taken.

3.2.2.4.2. Glutathione S-transferase (GST)-Id3 fusion protein, E12 and E47 production

Purified GST-Id3 fusion protein and *in vitro* translated (IVT) $[^{35}S]$-labelled E12 and E47 were kindly provided by Dr. R. Deed (Paterson Institute for Cancer Research, Manchester).

Purified IVT E12 and E47 were produced according to the manufacturers' instructions for IVT (Promega). Equimolar concentrations of each IVT protein with comparable specific activities were generated using L-[^{35}S]methionine and aliquots (5µl) analysed on Tris-glycine (12%) SDS polyacrylamide gels.

3.2.2.4.3. HLH protein dimerisation controls

To test that the assay gave reliable results on a known system, Id3-E47, Id3-E12 and MyoD-E47 binding were studied. Purified Id3-GST or MyoD-GST fusion-protein bound to glutathione-Sepharose beads (25µl) was added to buffer (50mMTris HCl pH8.0, 120mM NaCl, 0.5% NP40, 200µl) and IVT E47 or E12 (5µl) and incubated on a rotator for 1hr at 4°C. The resulting mixture (radiolabelled E47/E12 complexed with Sepharose bead-bound MyoD/Id3-GST) was washed with buffer (2x500µl), reconstituted in sample buffer (20µl) and heated to 100°C (5mins) before SDS-PAGE. Following electrophoresis, the gel was dried and autoradiographed.

The above procedure was repeated with various volumes of Id3-GST-beads (25, 10, 5, and 2.5µl) made up to a total volume of 25µl with GST-bound Sepharose beads, to determine the minimum concentration of Id3-GST-beads which gave an observable band on the autoradiograph. This concentration was chosen for use in peptide-binding studies.

3.2.2.4.4. Acetonitrile control

As the final cleavage solution of the peptides was 0.175M NaH_2PO_4 in 40% acetonitrile/water, the effect of acetonitrile on E47-Id3 binding was studied by adding acetonitrile to the incubation solution (25µl undiluted Id3-GST beads and 5µl E47) at concentrations of 40, 10, 5 and 1%(v/v) and the mixtures rotated at 4°C for 1hr.

3.2.2.4.5. Peptide assays on Id3

Some individual peptide solutions after cleavage from the solid-phase supports were divided into four aliquots (200µl) and aliquots pooled:

P1: 8(1)+10(1)+11(1)+12(1)

P2: 8(2)+10(2)+11(2)+12(2)

P3: 8(3)+10(3)+11(3)+12(3)

P4: 8(4)+10(4)+11(4)+12(4)

P5: 5(1)+6(1)+7(1)+5(4)

P6: 5(2)+6(2)+7(2)+6(4)

P7: 5(3)+6(3)+7(3)+7(4)

Peptide pools *P1-P7* (100μl) were added to Id3-GST (5μl Id3-GST-beads diluted with 20μl GST-beads) and incubated on a rotator at 4°C for 3hrs. E47 (5μl) was then added and the mixture rotated for 1hr, washed (2x500μl buffer), reconstituted in sample buffer (20μl) and heated to 100°C (5mins) before being loaded onto the gel. Following electrophoresis, the gel was dried and autoradiographed. Controls with no peptide added were also carried out and the experiment replicated.

3.2.2.5. Peptide binding studies: resonant mirror biosensor studies

The first set of the following experiments (**3.2.2.5.1-3**) was carried out while the resonant mirror biosensor instrument (IAsys) was on loan to the School of Pharmacy. In these experiments, the Id3-GST protein was immobilised on a CM-Dextran surface *via* glutathione. The aim was to observe binding of E12-GST, known to dimerise with Id3-GST as a system control. The peptide-binding experiments (**3.2.2.5.4-5**) were kindly carried out by Dr. Phil Buckle, (Labsystems, Affinity Sensors) at a later date.

3.2.2.5.1. Immobilisation of glutathione (GSH) onto a carboxymethyl (CM)-dextran surface

Reagents:

Running buffer: PBS (10mM sodium phosphate, 2.7mM potassium chloride, 138mM sodium chloride) + 0.05% (v/v) Tween 20 pH7.4.

Activation solution: 250μl of 0.26M EDC in H_2O mixed with 250μl of 0.12M NHS in H_2O.

Linker: 6-AHA.

Linker buffer: 10mM acetate, pH 7.0.

Linker solution 10mg/ml 6-AHA in linker buffer.

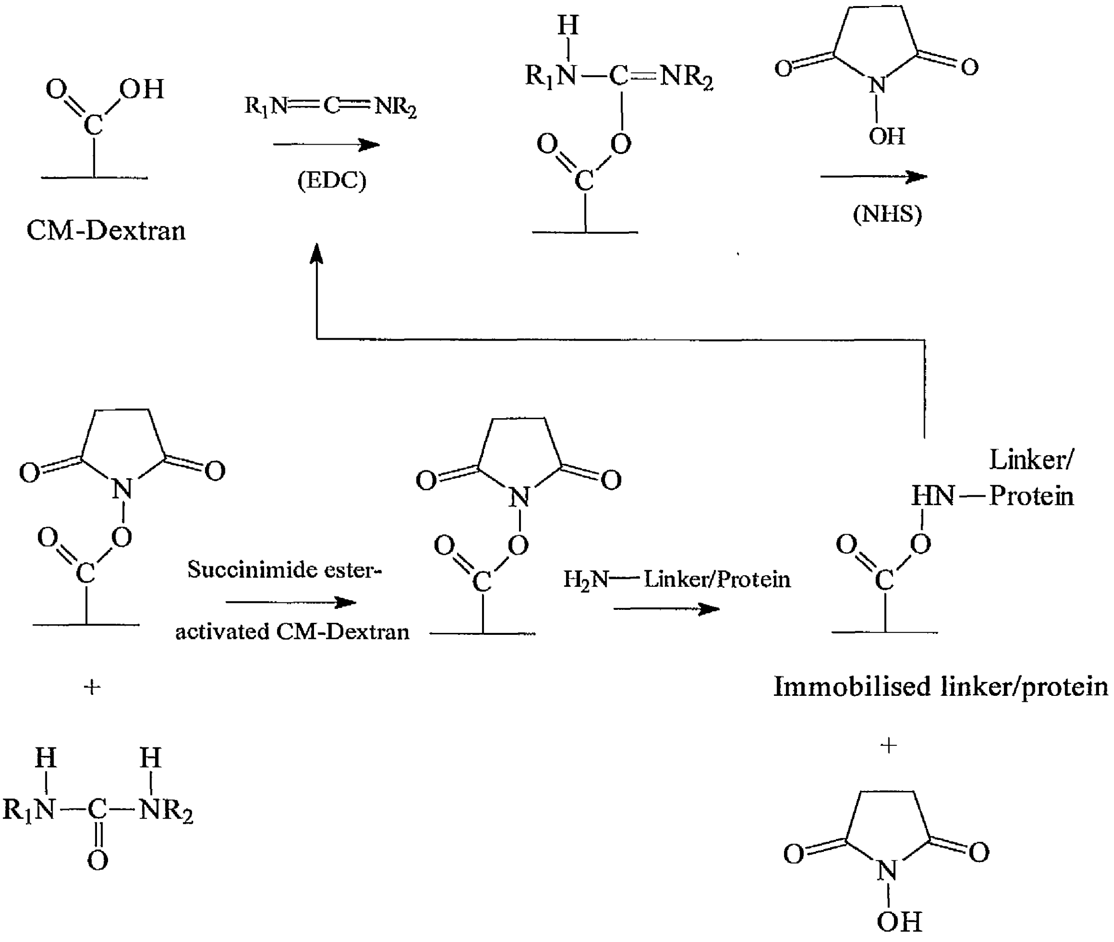
Blocking solution: 1M ethanolamine pH 8.5.

Ligand: GSH.

Ligand buffer: 10mM acetic acid, pH 7.0 containing 1mM DTT.

Ligand solution: Freshly prepared 10mg/ml GSH in ligand buffer.

Procedure: (see Figure 3.9).



$R_1 = -(CH_3)_2N(CH_2)_3$
$R_2 = -CH_2CH_3$

Figure 3.9. Reaction scheme for protein immobilisation on CM-Dextran. Linker = 6-AHA, protein = GSH or Id3-GST.

1. A carboxymethyl (CM)-Dextran cuvette was equilibrated in running buffer (50µl) for 10 minutes.

2. The carboxyl groups on the CM-Dextran were activated by replacing buffer with activation mixture (200µl) and leaving it for 7 minutes.

3. A running buffer wash (3x50µl) was carried out.

4. Steps 2 and 3 were repeated at least three more times.

5. A change was made from running buffer to linker buffer (3x50µl).

6. Linker buffer was replaced with linker solution (3x50µl) and left to allow coupling for 10 minutes.

7. The cuvette was washed with running buffer (3x50µl) to remove non-coupled ligand and left for 2 minutes.

8. Blocking solution (200µl) was added to deactivate and the system left for 3 minutes.

9. Steps 2 and 3 were repeated twice.

10. A change was made from running buffer to ligand buffer (3x50µl).

11. Running buffer was replaced with ligand solution and left to couple for 10 minutes.

12. The cuvette was washed with running buffer (3x50µl).

13. Blocking solution (200µl) was used to deactivate and the system left for 3 minutes.

14. The cuvette was washed with running buffer (3x50µl).

3.2.2.5.2. Immobilisation of Glutathione S-Transferase (GST)-Id3 onto GSH-derived-surface

Id3-GST and E12-GST were kindly provided by Dr. R. Deed (Paterson Institute of Cancer Research, Manchester).

Reagents:

Linker buffer: 10mM acetic acid, pH2.

Id3-GST in PBS (300µg/ml) diluted x20 in PBS.

Procedure:

1. A GSH-bound CM-Dextran cuvette was equilibrated in running buffer for 10 minutes (50µl) and washed (4x50µl).

2. The carboxyl groups on the CM-Dextran were activated by replacing buffer with activation mixture (2x50µl) and the mixture left for 15 minutes.

3. A running buffer wash was carried out (4x50µl).

4. The cuvette was washed with acetic acid buffer (4x50µl).

5. Id3-GST (diluted x10 in running buffer) was added (3x50µl) and left to couple for 10 minutes. By varying the volumes of Id3-GST used, immoblisation levels of 120, 390 and 930 arc seconds were achieved (a glycine buffer of pH2.2 was used for the latter immobilisation).

6. The cuvette was washed with running buffer (4x50µl) to remove non-coupled protein and left for 2 minutes.

8. Blocking solution (200µl) was added to deactivate and left for 10 minutes.

9. A running buffer wash (4x50µl) was carried out.

10. The cuvette was regenerated by washing with 20mM HCl (4x50µl).

3.2.2.5.3. E12-GST binding to immobilised Id3-GST

Procedure:

An Id3-GST-GSH-6-AHA-bound CM-Dextran cuvette was equilibrated in running buffer for 10 minutes (50µl) and washed (4x50µl).

E12-GST (20µl, 30µl) was added (association).

The cuvette was washed with running buffer (2x50µl) (dissociation).

3.2.2.5.4. Immobilisation

GST-Id3 was immobilised directly onto a CM-Dextran surface (*i.e.* no 6-AHA, linker or GSH) by coupling the -NH_2 groups of GST-Id3 directly to the -COOH groups of the dextran surface. The procedure (Figure 3.10) follows the same basic procedure of EDC/NHS activation, coupling, chloride salt washing, blocking with ethanolamine and HCl washing as described above. An immobilisation level of 5400 arc seconds was achieved.

3.2.2.5.5. Peptide binding

(a) Peptide samples 9(1), 9(2), 9(3), 9(4) were tested for binding activity to immobilised Id3-GST. Peptides were diluted from the stock solution (1µM, 0.225µg/ml in 0.175M NaH_2PO_4 in 40% acetonitrile/H_2O) to 200nM with PBS before testing.

Procedure:

1. A cuvette was equilibrated in running buffer for 10 minutes (50µl) and washed (4x50µl).
2. Peptide (20µl of 200nM solution) was added (association).
3. A wash with running buffer (4x50µl) was carried out (dissociation).

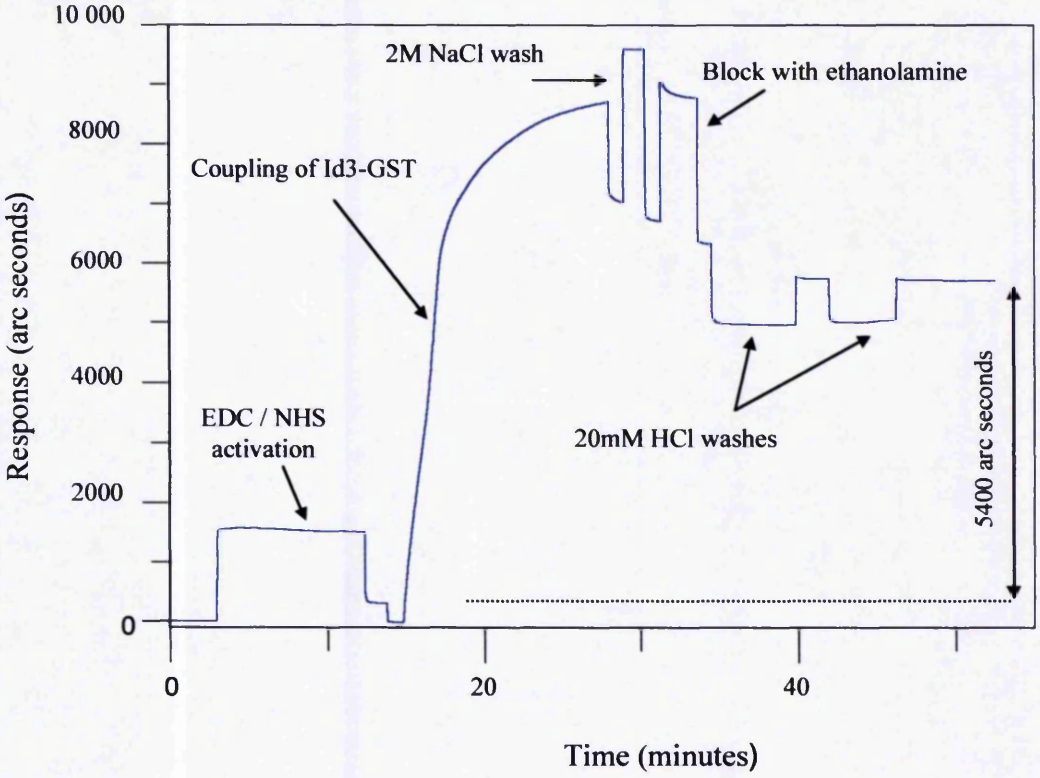


Figure 3.10. Direct immobilisation of Id3-GST on a CM-Dextran cuvette.

- Running buffer: PBS/Tween
- EDC/NHS Activation: 7 minutes
- Immobilisation buffer: 10mM acetate pH 5
- Sample dilution: 1/10 stock Id3-GST in acetate buffer

4. The cuvette was regenerated with 20mM HCl (4x50μl).
5. A wash with running buffer (4x50μl) was carried out (dissociation).

Peptide 9(4) gave the highest binding signal and was subsequently bound over a range of concentrations from 50nM to 400nM. Non-specific binding was assessed using peptide 9(4) at 20nM against both Id3 and a native CM-Dextran cuvette.

(b) Using the procedure described above, peptides 5(1), 7(1), 8(1), 5(2), 11(2), 7(3), 8(3), 11(4), the randomly chosen E and F (Figure 3.8), as well as the four peptides above, were all assessed for binding to immobilised Id3-GST and E12-GST, using a negative control of immobilised GST. These peptides were chosen because they had fairly diverse sequences. A control peptide was chosen (B1 = bradykinin) which was not expected to bind to either Id3 or E12.

3.3. RESULTS AND DISCUSSION

3.3.1. Molecular modelling

3.3.1.1. Sequence alignments

Figure 3.7 shows the alignment of HIF-1α and HIF-1β with the five proteins with solved 3-D structures. Figure 3.6 shows the alignment of HLH proteins adapted from Littlewood and Evan [238]. It is obvious from these alignments that there exist highly conserved key residues of the HLH motif. For example, basic residues at the N-terminus of helix 1 and hydrophobic residues found mostly in the latter half of helix 1 and former half of helix 2. These conserved amino acids are present in the HIF-1 monomers, giving confidence that comparative modelling gives a representative structure of HIF-1 over this region.

3.3.1.2. Tertiary structure of HIF-1

The comparative molecular model of the HIF-1 transcription factor dimer both free and bound to DNA is shown in Figures 3.11(a) and (b). The HLH proteins with solved structures are shown in Figures 3.2 (Max, bHLHZ), 3.12 (USF, bHLHZ), 3.13 (MyoD, bHLH) and 3.14 (E47, bHLH). HIF-1 displays typical HLH topology. The two N-terminal regular helices of HIF-1α and HIF-1β, which are both terminated by prolines, are of equal length (29 residues) as are the intervening loops (10 residues) and both C-terminal helices (21 residues). The structural models were analysed for satisfactory main-chain and side-chain conformations, bond lengths, angles, chirality and buried/exposed residues in SYBYL6.5- Biolpolymer/Analyse Protein (ProTable) and in QUANTA96- Protein Health and slight discrepancies were amended where necessary.

HLH proteins can have low sequence identity/similarity despite sharing the same structural motif. For example, the crystal structures of MyoD and Max (25% sequence identity) show an RMS difference of only 4.68Å for α-carbon atoms and 4.64Å for all protein backbone atoms over the HLH region. The overall RMS differences between the HIF-1 model and the USF and MyoD crystal structures on which it was modelled are 3.63Å and 3.51Å, respectively (calculated for α-carbon atoms over residues S15(1α)-T148(1β)), and 3.55Å and 3.64Å for all protein backbone atoms.

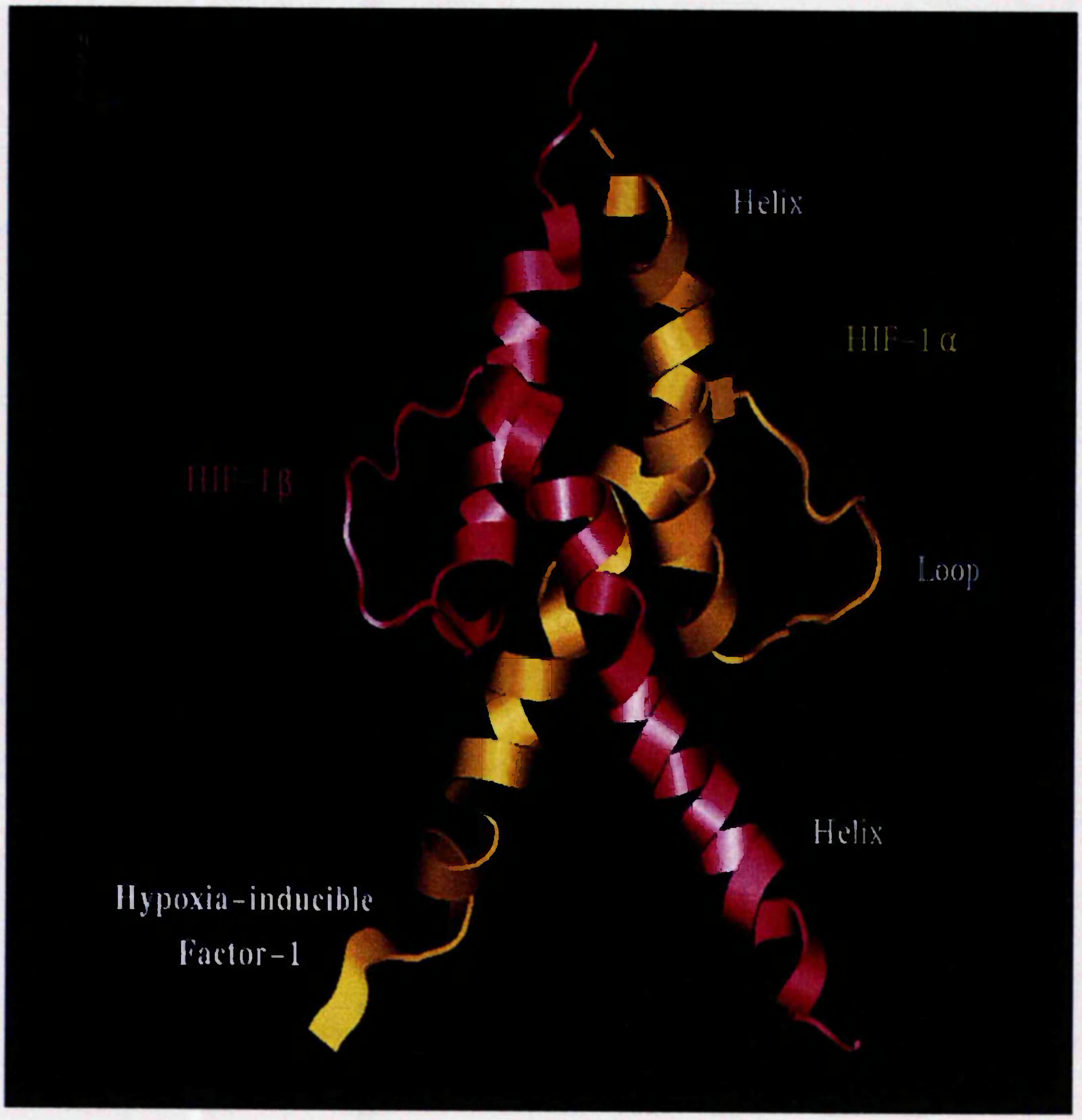


Figure 3.11(a). The molecular model of Hypoxia-Inducible Factor-1 (HIF-1).

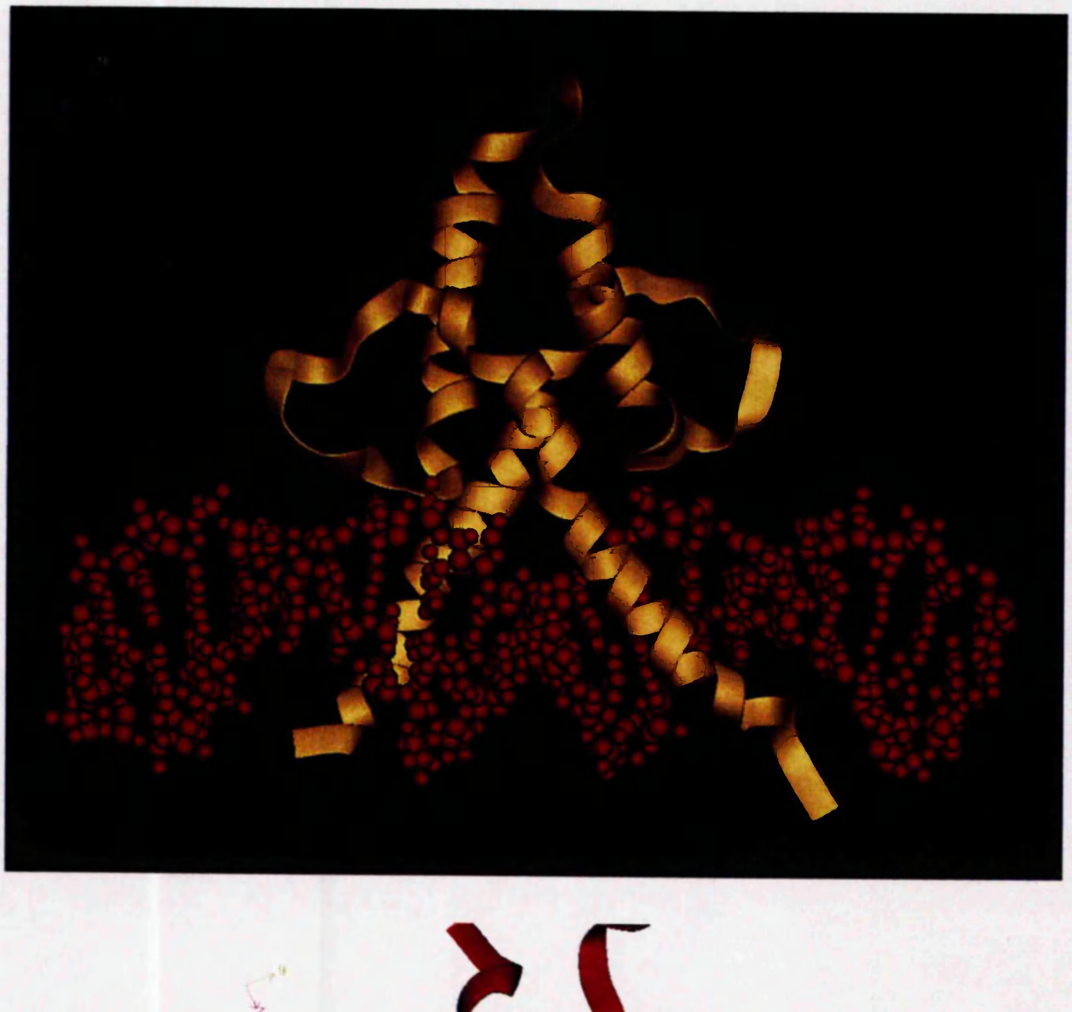


Figure 3.11(b). HIF-1 bound to DNA.

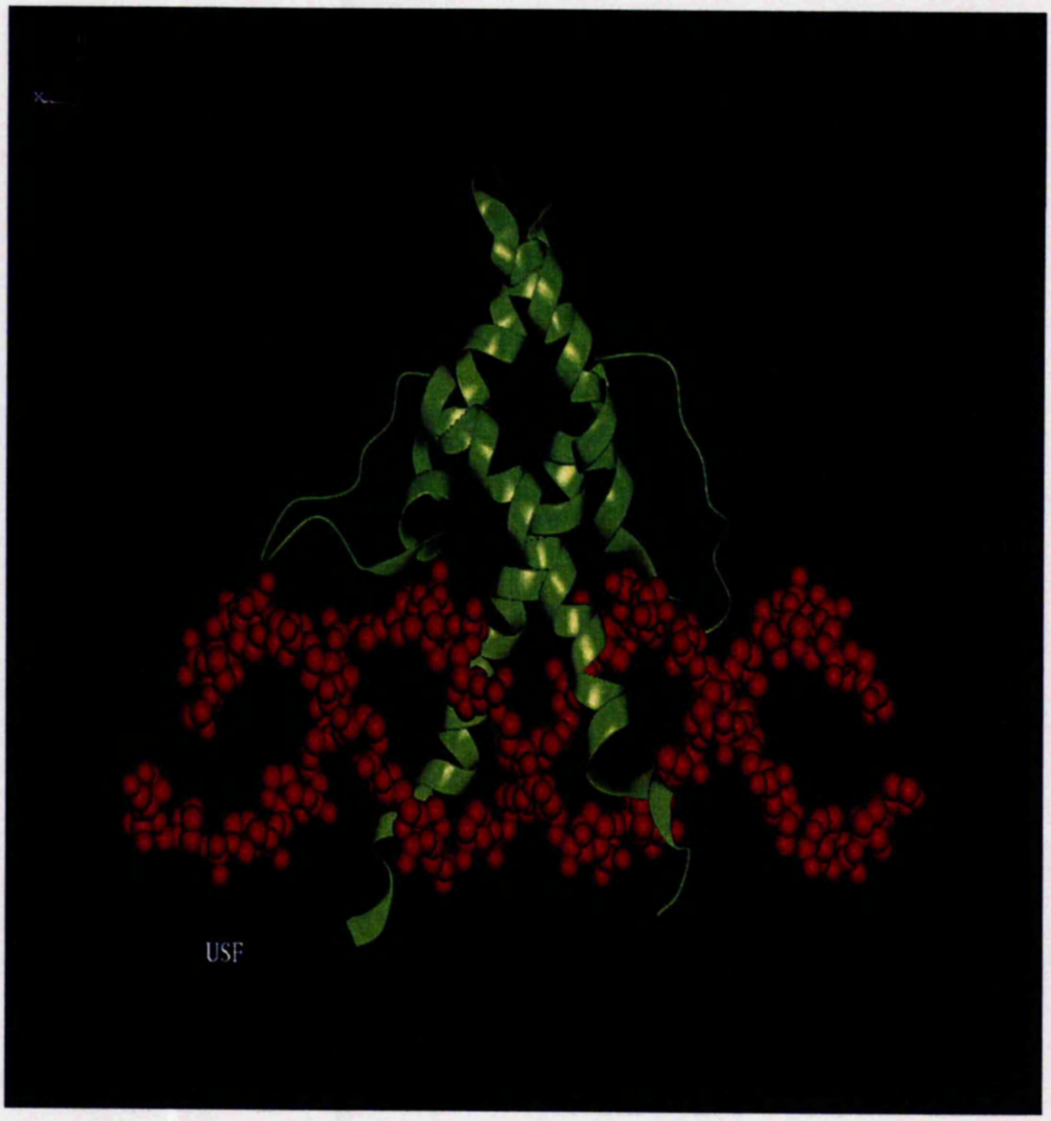


Figure 3.12. The bHLHZ transcription factor USF bound to DNA.

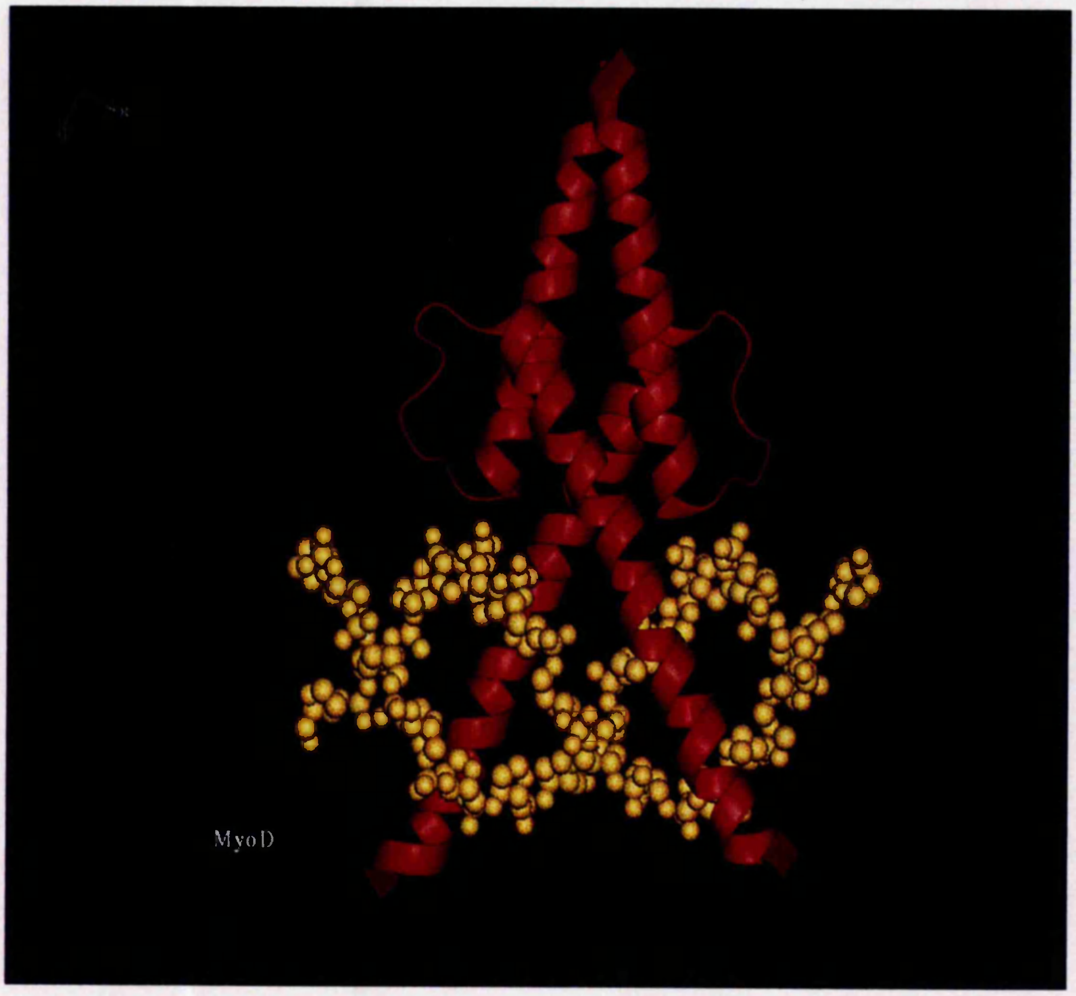


Figure 3.13. The bHLH transcription factor MyoD bound to DNA

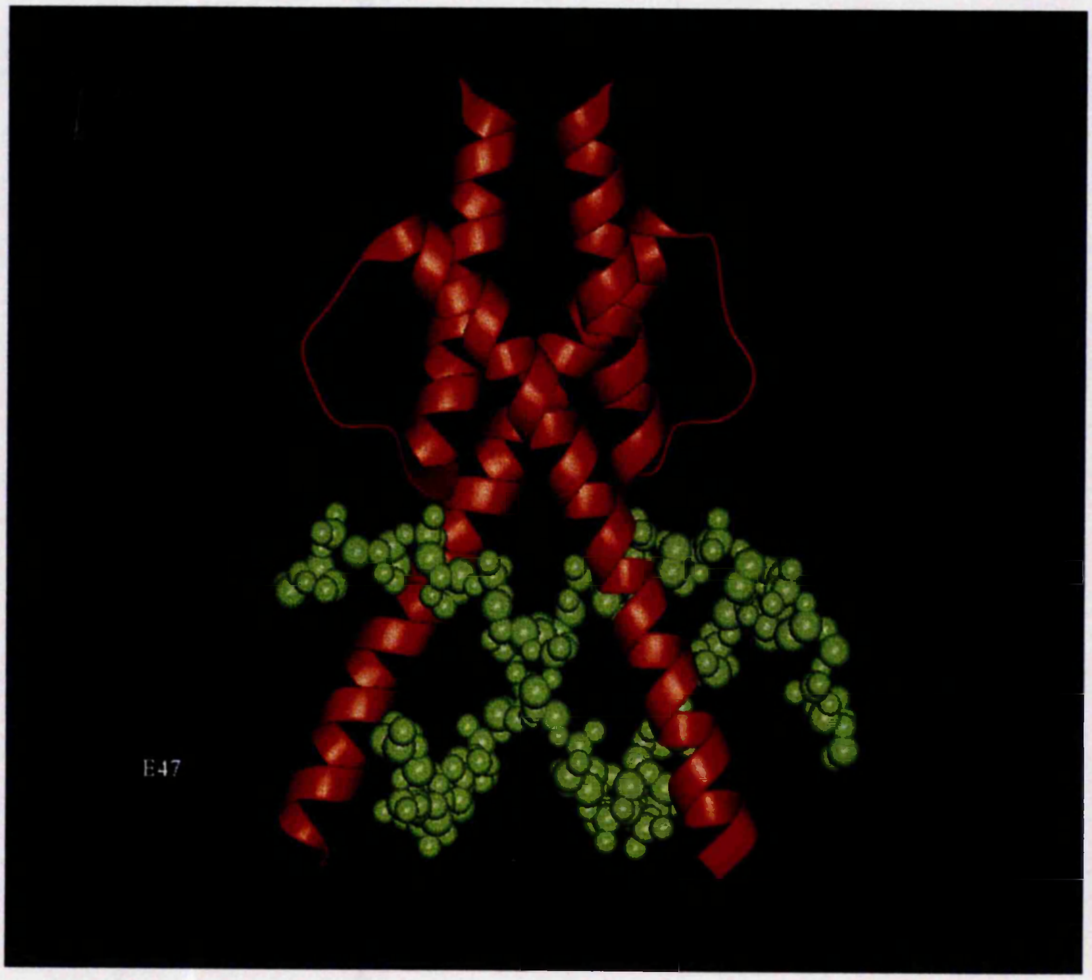


Figure 3.14. The bHLH transcription factor E47 bound to DNA

3.3.1.2.1. Hydrophobic core

The highly conserved hydrophobic amino acids (Figure 3.6) in the HLH family are essential for protein dimerisation [356,400]. These residues are found towards the top of helix 1 and the lower half of helix 2. Detailed analysis of the hydrophobic core within the HIF-1 heterodimer and comparison with the five solved HLH crystal structures confirms that the packing arrangements are conserved in HIF-1 (described in detail below). Figure 3.15 shows the residues important for forming hydrophobic interactions at the dimer interface.

MyoD, Max and E47 have a conserved phenylalanine in H1 corresponding to HIF-1α F37. In HIF-1, this makes hydrophobic contacts with L132 of HIF-1β, comparing with the Max residues F43 and L64' on different monomers. HIF-1β, like USF, has isoleucine (I109) here, and together with F37 can make van der Waals' contacts with the methylene groups of K128 of HIF-1α, as seen in USF (I219 and K240). Studies on the MyoD homodimer show that mutating F129 to A, I, L or V reduces dimerization to as low as 5% of that for the wild type [400]. Such reductions in side-chain bulk have considerable effects on stability and hydrophobic packing, as shown by the mutations of an isoleucine to valine in coiled coils/leucine zippers [401] and an isoleucine to alanine in barnase [402]. The double mutation of F356D and L359E in E47, corresponding to F37 and L40 in HIF-1α completely abolished dimerisation and DNA binding thus showing its importance in dimer stability. *It is predicted that the same mutation in HIF-1 would do the same.*

Residues V59 and M60 (HIF-1α) align perfectly with the hydrophobic residues of other HLH sequences. It is interesting to note that of all the proteins aligned, only HIF-1α has methionine at the position occupied by residue 60. The corresponding residues in other HLHs are normally occupied by isoleucine and leucine. Mutating these two residues I and L to D and E respectively in E47, abolishes DNA-binding and dimerisation activity [356]. M60 points directly into the hydrophobic core and forms hydrophobic contacts with L132 of H2 HIF-1β, (*cf.* the interacting L379 residues on the monomers of E47), and possibly with L112 (1β). V59 may interact with A41. HIF-1β has L132 at this position in the sequence, which forms close contacts with HIF-1α residues V36 and L40, the latter being another highly conserved residue in HLH sequences. *Mutation of V59 and M60 (1α) and L132 (1β) to acid residues would probably therefore prevent dimerisation too.* Y38 (1α) interacts with H42 on the same helix.

The conserved hydrophobic L44 (H1, 1α) is close enough to interact with Y66 (H2) of the same monomer. Residue P117 (H1, 1β) packs against H138 and a corresponding interaction is

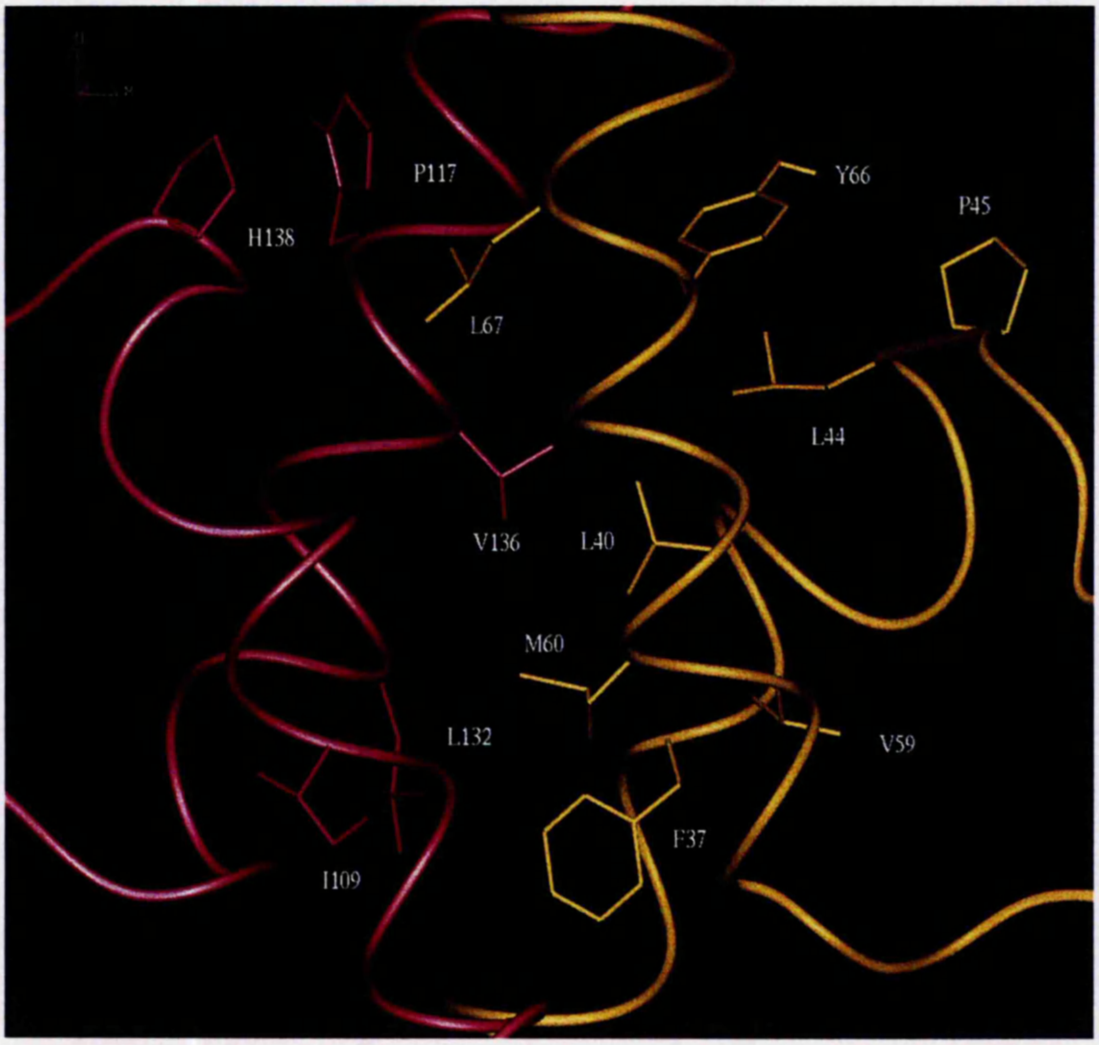


Figure 3.15. Residues involved in hydrophobic interactions important in the dimerisation of HIF-1.
HIF-1α is shown in yellow and HIF-1β is shown in pink.

found in Max (P51 and Y70), USF (P227 and Y250) and Pho4 (P28 and Y52). Residue H138 (1β) is in the position in the HLH sequence much more frequently occupied by F, Y or V in other sequences, but presumably the histidine is able to provide similar hydrophobic interactions. A basic amino acid, R70, is present in HIF-1α where a leucine is usually placed in the sequence. The methylene atoms of its side chain are in a position to possibly interact with L142 of HIF-1β (H2) and both residues sit on the top of the hydrophobic core. Mutating this leucine to lysine in E47 (L389V) together with the mutation I386D three residues earlier in the sequence abolishes dimerisation and DNA-binding [356] again highlighting the importance of the hydrophobic residues. The conserved hydrophobic residues L44 and V116 in the α and β monomers, respectively, (the penultimate H1 residues) interact with the conserved hydrophobic M139 of HIF-1β and L67 of HIF-1α, respectively (not shown in Figure 3.15)

Many of the hydrophobic interactions observed in the model have counterparts in the HLH proteins with determined crystal structures, *i.e.* Max [361,362], USF [363], MyoD [359], E47 [365], Pho4 [360] and SREBP-1a [366]. This observation, together with the fact that so many of the residues are conserved across the HLH family, provides evidence for the importance of these interactions in transcription factor dimerisation and the validity of the conformations of the hydrophobic residues in the HIF-1 model.

3.3.1.2.2. Interface of helices

Side-chains at the interface of the helices and within helices were explored and compared to the crystal structures to locate residues potentially capable of electrostatic interactions or hydrogen bonding. Table 3.2 lists potential inter- and intra-helical electrostatic interactions found in HIF-1. Most of these are shown in Figure 3.16.

It should be emphasised that the hydrophilic residues involved in electrostatic interactions are very flexible in solution and alternative interactions may exist in addition to those mentioned above due to this conformational freedom. Also, a number of the interactions observed in the model of the free heterodimer involve basic residues towards the N-termini of the lower helices. When bound to DNA, many of these residues cannot form the interactions described above because they form interactions with the DNA (see below).

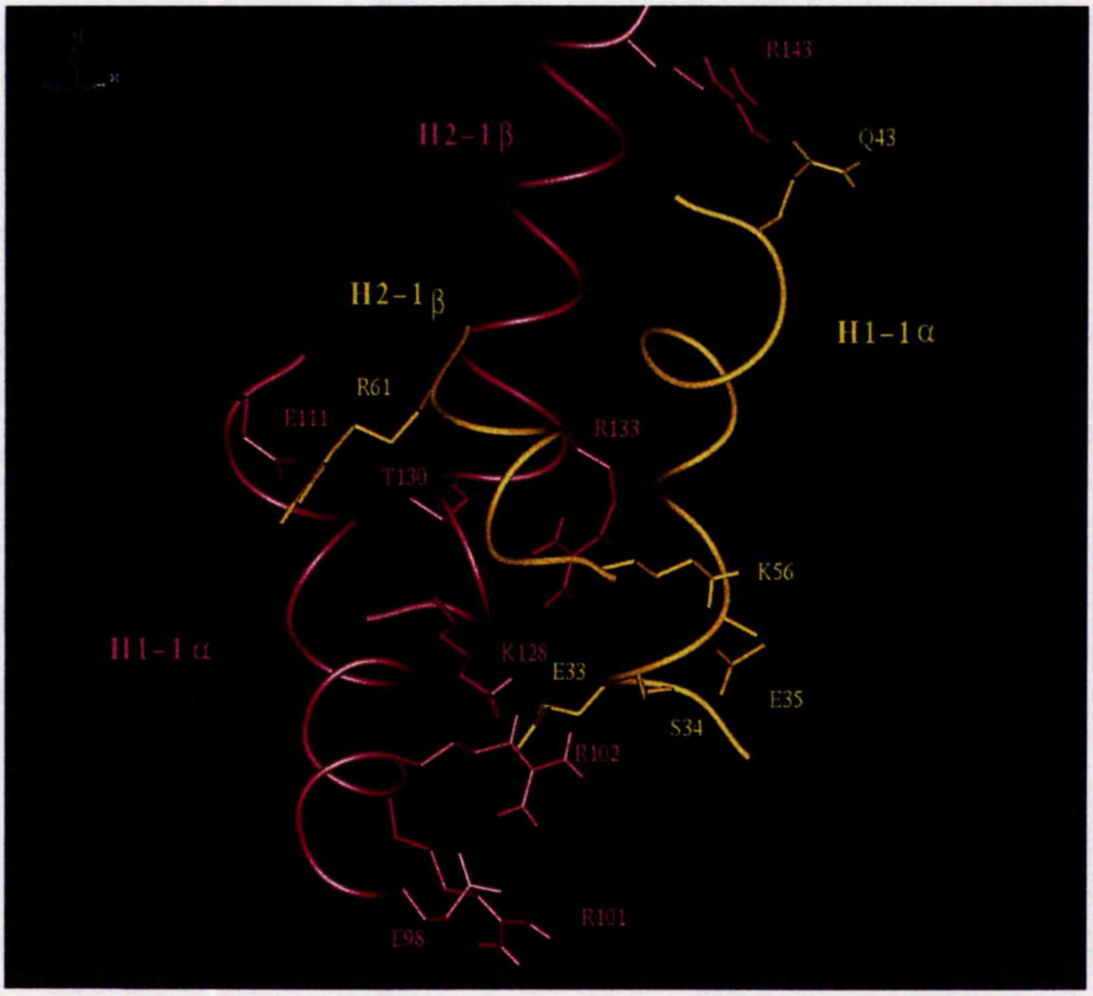


Figure 3.16. Inter- and intra-helical electrostatic interactions conferring stability to the HIF-1 heterodimer.

HIF-1α H1	**HIF-1β H2**	**HIF-1α H2**	**HIF-1β H1**
R29-E33			
R29	D127		
E33	K128		
S34		K56	
E39	K140		
	S137-K140		
		R61	E111
Q43	R143		
			E96-R99
			E98 with R101, R102
	D127-K128		
			T130 with R133

Table 3.2. Residues in the HIF-1 dimer having the potential to form electrostatic interactions.

There does not seem to be any real conservation of amino acids involved in electrostatic helix-helix interactions throughout the HLH family. Although hydrophilic amino acids do not seem to play as important a rôle in dimer stability as hydrophobic residues, salt bridges have been implicated in the dimerisation strengths and preferences of bHLH proteins [400] and bHLH-Z proteins [249,362]. Shirakata *et al.* [400,403] justified dimer specificities by assessing the number of charge-charge interactions within the MyoD-E47/E12 interface (charged-pair rule). For example, three non-hydrophobic residues in MyoD that facilitate specific dimerisation and DNA-binding with E12 were found; C135 (top H1), R155 and Q161 (both H2). The mutants C135R and R155E only formed 30% of the heterodimer complex when compared with wild-type MyoD. These residues do not really have any counterpart in HIF-1 and salt bridges in HLH dimers seem more specific to the individual proteins involved than the conserved hydrophobic residues. This was also the conclusion found when the molecular model of the HLH protein Id3 was constructed in our laboratory [358]. An inter-helical salt bridge rule has also been proposed to explain the dimerisation specificity between leucine zipper structures [404].

3.3.1.2.3. Loop region

The sequence alignment (Figure 3.6) reveals that within the loops of the HLH proteins there exists considerable variability in sequence and length which can range from 5 residues (CBF1 protein) to 20 residues long (*achaete* protein). The crystal structures also imply that no

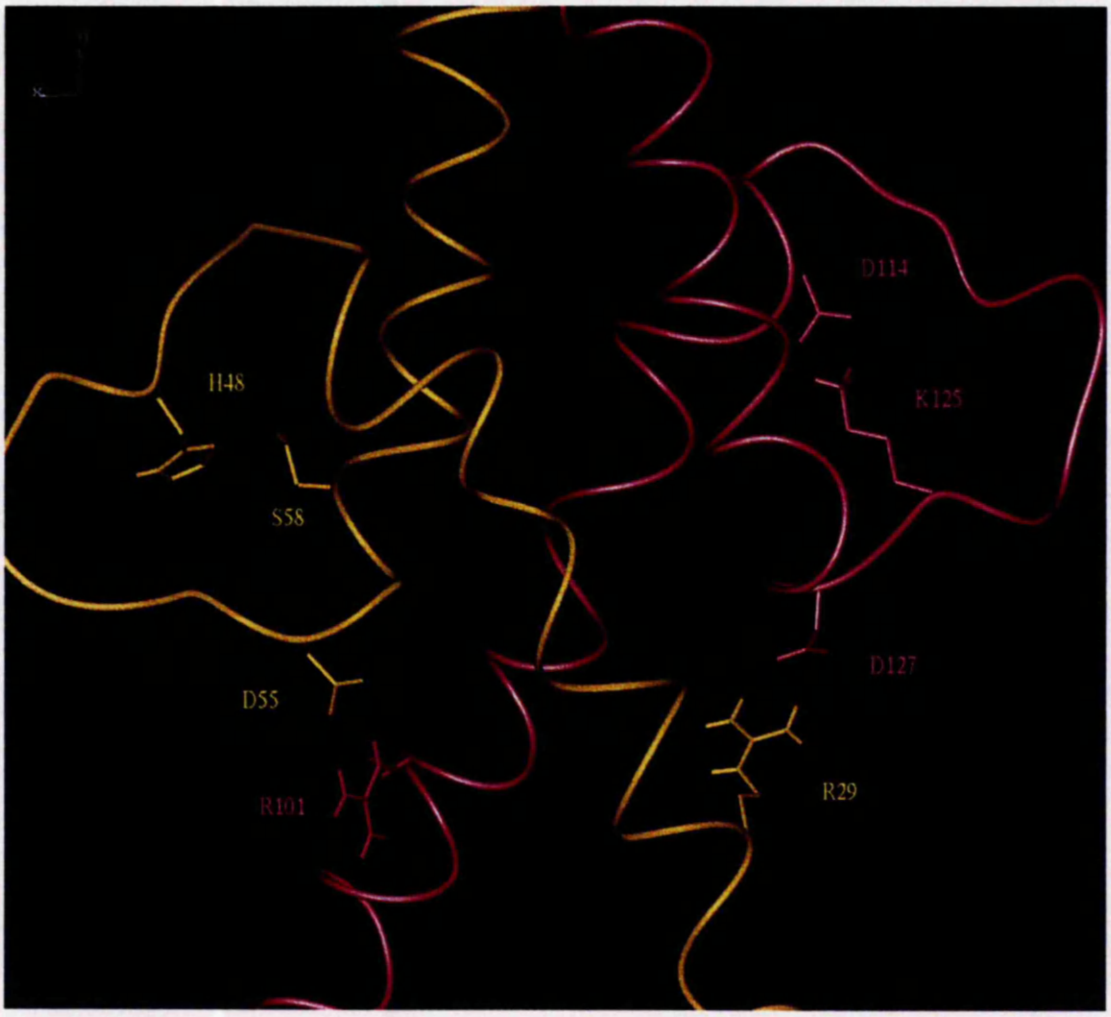


Figure 3.17. Residues important in providing electrostatic interactions to stabilise the loops of HIF-1.
HIF-1α is shown in yellow and HIF-1β is shown in pink.

underlying common loop structure exists. Investigation of the loop regions of the HIF-1 model (P45-K56 of 1α and P117-K128 of 1β) yielded a number of electrostatic interactions within the loop and between loop and helix residues and DNA, which have the potential to confer stability. These are summarised in Table 3.3 and in Figure 3.17.

HIF-1α L	**HIF-1β L**	**HIF-1α H1**	**HIF-1α H2**	**HIF-1β H1**	**DNA**
H48			S58		
D55				R101	
	D127	R28			
	K125			D114	
	R124				5' P Gua 39

Table 3.3. Electrostatic interactions with the potential for stabilising loop residues. (Residues along one row interact with each other. See Table 3.4 for position of nucleotide Gua 39).

In some HLH proteins, if the loop is long enough, a semi-conserved basic residue (lysine or arginine) can interact with the DNA phosphate backbone. R124 of HIF-1β is seen to do this in the model and also in Max (K57, 8-residue loop), MyoD (R143) and in USF (K235, 12-residue loop). The analogous lysine (K371) in E47 is too far away from the DNA to make any contacts. HIF-1α does not contain a basic amino acid and K125 of HIF-1β seems to point away from the DNA, and interacts instead with D114 of the same monomer. *Mutating R124 (HIF-1β) to D or E may reduce the strength of DNA-binding.*

3.3.1.2.4. Base of the four α-helix bundle: DNA-binding

The universal minimal DNA element required for specific binding by the HLH and bHLHZ transcription factors is the so-called "E-box", generally represented as 5'-CANNTG-3', where NN is CG or GC [194,405,406]. However, HLH proteins can also bind similar sequences. A comparison of HIF-1 binding sites in different genes is shown in Figure 3.18. In the case of the erythropoeitin and vascular endothelial growth factor genes, DNA sequences of 33 and 35 base pairs respectively have been identified that are sufficient for hypoxia-induced transcription and thus constitute hypoxia response elements (HREs) [213,315]. These HREs have in common the presence of a HIF-1 site and flanking sequences essential for function [316]. Within the HREs is an invariant core sequence of 5'-CGTG-3'.

5' 3'		
GGGCCCTA**CGTG**CTGTCTCACACAGC	Erythropoeitin	human
GGGCCCTA**CGTG**CTGCCTCGCATGGC	Erythropoeitin	mouse
TCGCTTCA**CGTG**CGGGGACCAGGGAC	Aldolase-A	human
ATTTGTCA**CGTG**CTGCACGACGCGAG	Phosphoglycerate kinase	mouse
AGTGCATA**CGTG**GGCTTCCACAGGTC	Vascular endothelial growth factor	rat
CCAGCGGA**CGTG**CGGGAACCCACGTG	Lactate dehydrogenase-A	mouse
TCCACAGG**CGTG**CCGTCTGACACGCA	Glucose transporter-1	mouse

Figure 3.18. A comparison of HIF-1 binding regions in different genes. The core sequence of the HRE is shown in red.

The sequence alignment of the HLH family shows highly conserved, non-hydrophobic residues in the basic region of the proteins, notably arginine, lysine and glutamic acid, and a lysine at the start of helix 2. The exceptions are the Id [238,264] and the E(spl) families, which lack a functional DNA-binding domain [260]. Site-directed mutagenesis of members of both the HLH and bHLHZ protein families demonstrate that these conserved amino acids in the basic region are involved in DNA binding [405,407-409] forming electrostatic interactions with the DNA phosphate backbone and hydrogen bonds with the bases.

The conserved amino acids correspond to HIF-1α residues K19, R27, R29 and K56 and HIF-1β residues R91, E98, R99, R101 and K145. The molecular model of HIF-1 shows that the residues do interact with the DNA and do so essentially in the same way as seen for the corresponding residues of MyoD [359], Max [361,362], USF [363], E47 [365], Pho4 [360] and SREBP-1a [366], thus adding validity to the computed structure.

The conserved lysine at the start of helix 2 is represented by K56 and K128 HIF-1α and 1β, respectively. Its backbone amide and the side chain interacts with the DNA phosphate backbone and almost identical interactions are seen in the structures of USF, Max, MyoD, E47, Pho4 and SREBP-1a. Table 3.4 lists the potential electrostatic interactions between the protein and DNA and many of these are shown in Figure 3.19(a) and (b).

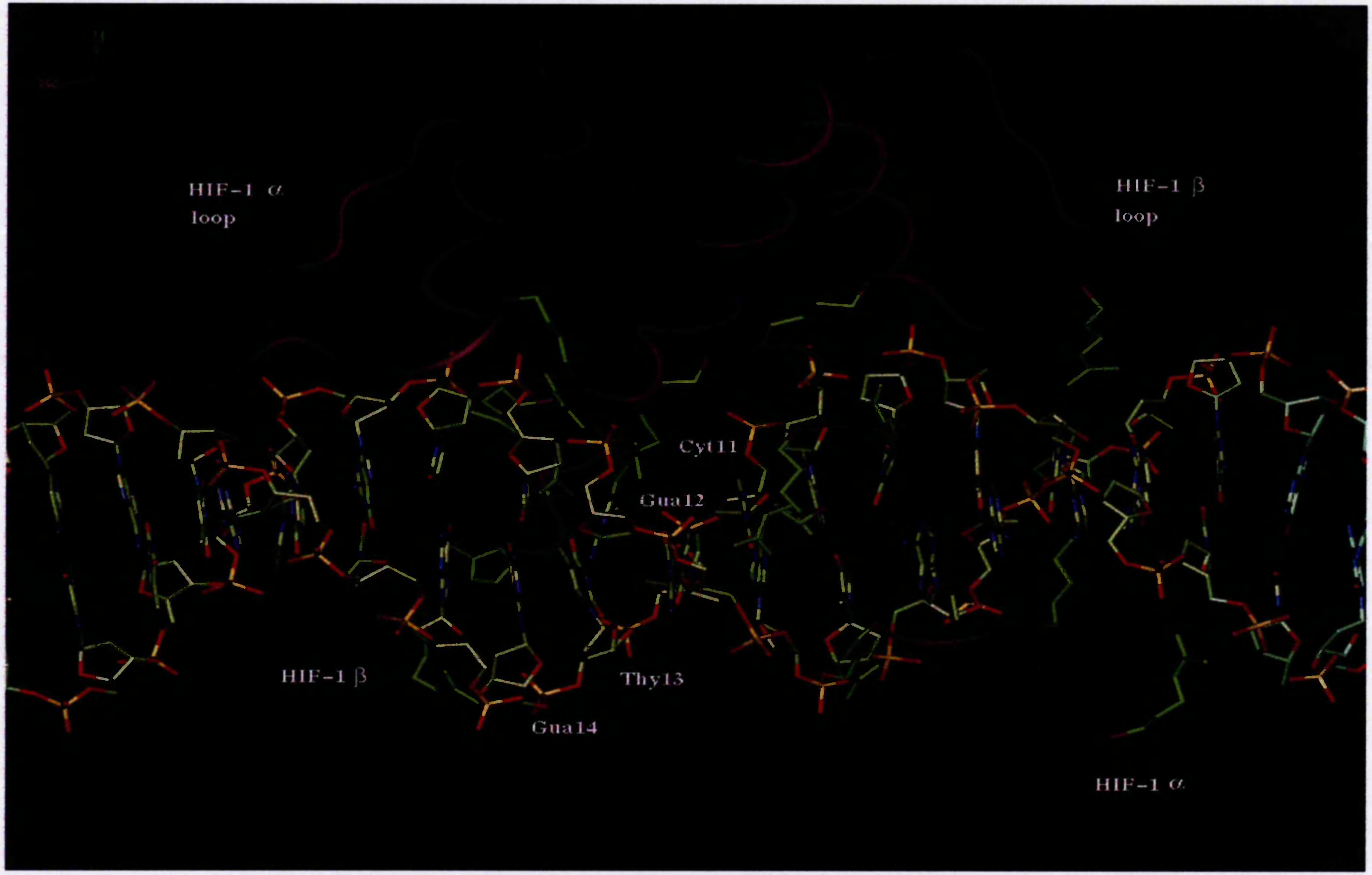


Figure 3.19(a). HIF-1 bound to DNA. The amino acids important in forming electrostatic interactions are shown in green and shown in more detail in Figure 3.19(b). The core HRE sequence of DNA (CGTG) is labelled. The protein Cα trace is shown as a purple ribbon.

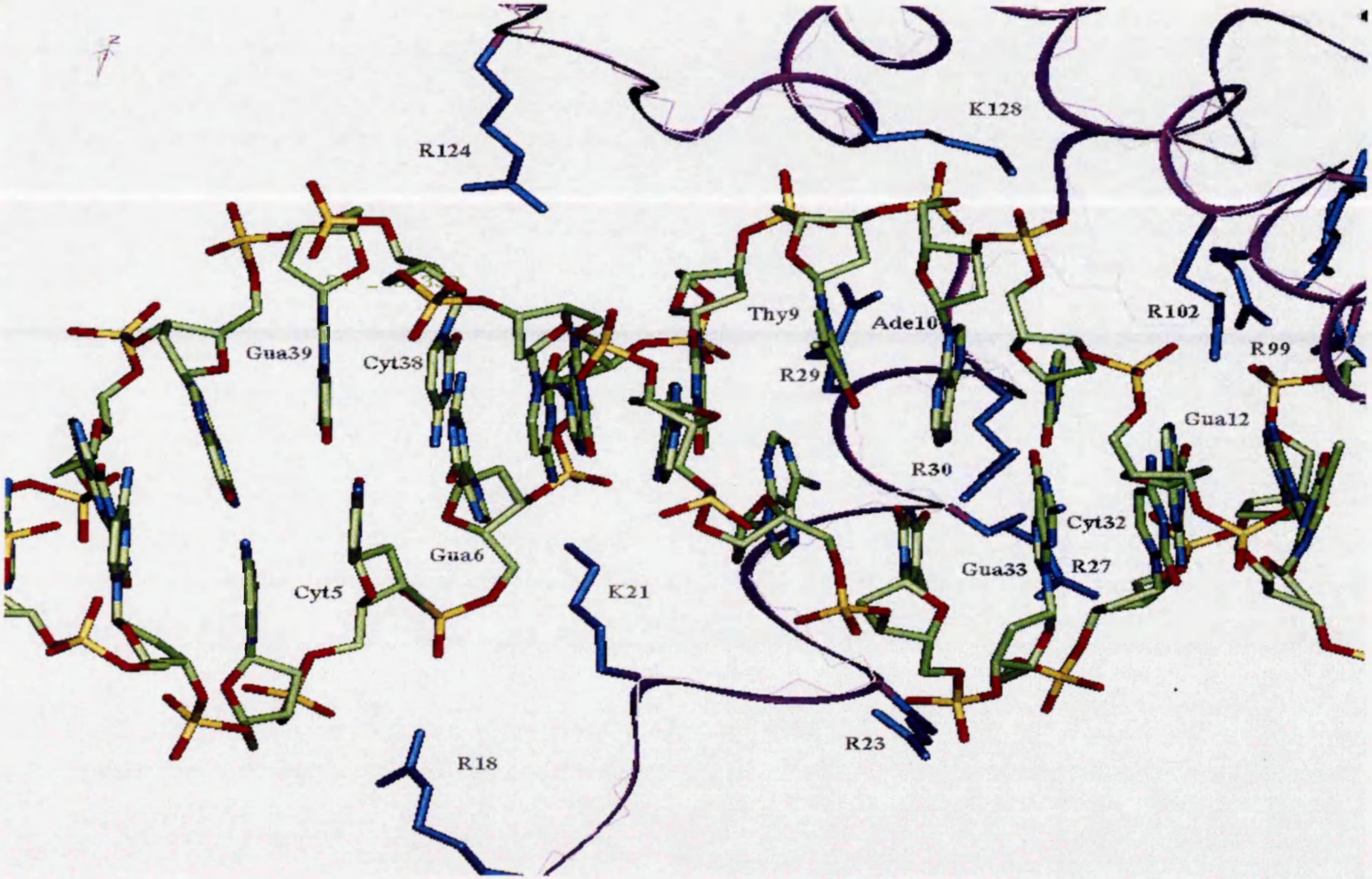


Figure 3.19(b). Some HIF-1 residues important in forming interactions with DNA. The protein Cα trace is shown as a purple ribbon and the amino acid side chains in blue.

1 10 20
5'-XXXCCGGTTACGTGGCCTACA
XXGGCCAATGCACCGGATGTX-5'
40 30 22

HIF-1α residue	DNA nucleotide	HIF-1β residue	DNA nucleotide
R18 (H1)	5' P Cyt 5, 5' P Gua 6	R91 (H1)	5' P Gua 14
K21 (H1)	5' P Gua 6	H94 (H1)	N7 Gua 14
R23 (H1)	5' P Gua 33	E98 (H1)	O4 Thy 13
*R27 (H1)	5' P Cyt 32	*R99 (H1)	5' P Gua 12
		R100 (H1)	5' P Cyt 29
*R29 (H1)	5' P Thy 9	*R101 (H1)	5' P Cyt 30
*R30 (H1)[§]	N7 and O6 Gua 33 O4 Thy34	*R102 (H1)[§]	N7 and O6 Gua 12
S34 (H1)	5' P Ade 31	R124 Nη (L)	5' P of Gua 39 5' P of Cyt 38
*K56 Nε and amide NH (H2)	5' P Cyt 30	*K128(H2)	5' P of Ade 10

Table 3.4. Proposed electrostatic interactions between DNA and HIF-1. The co-ordinates of the DNA from the crystal structure of USF were used and the bases mutated to a sequence based on the human erythropoeitin 3' enhancer site. The core of the HRE sequence is shown in red.

P – phosphate group

* - sequence-aligned residues of HIF-1α and HIF-1β.

[§] - residue conferring specificity for the central dinucleotide of the E-box (CG).

Mutation analyses of the E47 residues R338, R346 and R348 have shown their importance in DNA-binding [356]. The double mutations R346E plus R348E and R337E plus R338E completely abolished DNA-binding, and even mutating R338, R346 and R348 individually to lysines abolished DNA-binding, while still allowing dimerization. Mutation of MyoD residues R119, R120 and R121 to a triple alanine mutant [400] also prevented DNA-binding, as did the double mutants R110Q, R111Q and R119Q, R120Q [409] and mutagenesis studies on

Myc also support this view [407,410]. Nearly all of these mutations removed the basic, positively charged nature of the sequence and therefore the ability to bind the negative phosphate backbone of DNA. *Mutating these conserved residues in HIF-1 to non-basic amino acids should also prevent DNA binding.*

Just one amino acid determines which of the two DNA sequences (NN is CG or GC) the protein binds to [411] and this residue lies in the column in the alignment marked with an asterisk. An arginine here confers specificity for a CG central dinucleotide. This is due to the Nη of arginine forming an electrostatic interaction with N7 of guanosine (of the CG central pair) and is a contact pattern characteristic of HLH proteins, e.g., Max, USF and Pho4. Proteins which have a hydrophobic or polar amino acid here (commonly V, I T or Q), recognise CAGCTG sequences, e.g. MyoD (L122) and E47 (V349). These amino acids do not form any interactions with the DNA bases. The L122R mutation in MyoD has been shown to change this specificity to CG [412] and the mutation to valine in c-Myc [405] also abolished CACGTG binding activity. The corresponding positions in HIF-1 are R30 (HIF-1α) and R102 (HIF-1β) and these residues can be seen on the model to interact with the central guanosines of the central 5'-CGTG-3' (N7 of Gua 33 and Gua 12, respectively). *If these arginines were substituted by leucine for instance, HIF-1 would probably bind a 5'-GCTG-3' sequence.* It is interesting to note that SREBP-1a protein has tyrosine at this position which dramatically alters the DNA binding specificity from the typical 'E-box' to 5'-ATCACCCCAC-3'. Another anomaly is the AHR protein which has a histidine here and recognises the half-site sequence 5'-TNGC-3'. HIF-1β can form another heterodimer with AHR and this binds the sequence commonly found in dioxin-responsive enhancers 5'-TNGCGTG-3' [413]. This heterodimer has also been modelled in our laboratory and its protein and DNA interactions investigated (unpublished data).

A conserved glutamic acid (E98 of HIF-1β) is also critical for DNA-binding. Studies of the yeast HLH protein Pho4 showed that if this residue is substituted by the shorter aspartic acid or the large hydrophobic leucine, DNA-binding is abolished [408]. This is also seen in the E118D mutant of MyoD [409]. This glutamic acid also provides specificity for bases flanking the core CANNTG motif [408], e.g. in Pho4 [360]. Each monomer in an HLH transcription factor dimer recognises and binds a DNA half-sequence: HIF-1α binds the TAC/GAC/CAC half while HIF-1β recognises and binds the GTG/GTC half [413]. In HIF-1β, E98 can hydrogen-bond with T13 and contacts also made by the semi-conserved histidine HIF-1β (H94) (corresponding to Max H28 and USF H204, Pho4 H5, SREBP-1a H328) with Gua 14 define the outer E-box bases TG. HIF-

1α has A26 aligning with E98, which is too short to contact the DNA. *If E98 and H94 were mutated to alanine, HIF-1 may bind core sequences other than those ending in 5'-TG-3'.*

The central two nucleotides in the CANNTG sequence provide for discrimination in binding between different family members, but since there is a substantial number of bHLH proteins which bind both sequences, sequences flanking the core motif influence protein-binding too and provide protein-specific discrimination between otherwise identical binding sites. This has been demonstrated with USF [414] and single amino acid substitutions have been shown to alter HLH specificity for nucleotides outside the core CANNTG motif for the Pho4 protein [408].

It must also be noted that transcription factors do not bind DNA alone, but form large DNA-binding complexes with other proteins, e.g. HIF-1 is known to complex with p300/CBP protein [204] to initiate transcription. These assemblies help to recognise a specific sequence of DNA and initiate transcription of only one gene.

3.3.1.3. Predictions of mutations for fluorescent labelling

By studying the molecular model of HIF-1, residues on each monomer were chosen which could be mutated to cysteine residues and thus each easily covalently labelled with a donor or acceptor dye molecule for future FRET studies. The residues had to be directed towards each other in space and have a distance of 15-30Å between them. The residues chosen were E20 and G76 of HIF-1α intended to interact with R91 and G146 of HIF-1β respectively (Figure 3.20). Plasmid constructs are now being prepared by Dr. Karen King of the Experimental Oncology Group, the School of Pharmacy and Pharmaceutical Sciences, to test these ideas.

3.3.1.4. Summary

A structural model built for the heterodimer HIF-1 is shown in Figure 3.11. It has been validated by comparison with amino acid positions and conformations in HLH protein crystal structures available and backbone and side chain conformations have been analysed computationally using SYBYL6.5 and QUANTA96 software. Highly conserved residues agree well with those interactions observed in the determined three-dimensional structures. Further confidence in the model could be gained by carrying out the mutation predictions described and assessing their influence on dimerisation and DNA binding. The model provides a basis for understanding the interactions of the HIF-1 HLH region within the heterodimer and with DNA, predicting how these interactions could be disrupted and is a starting point for further study of

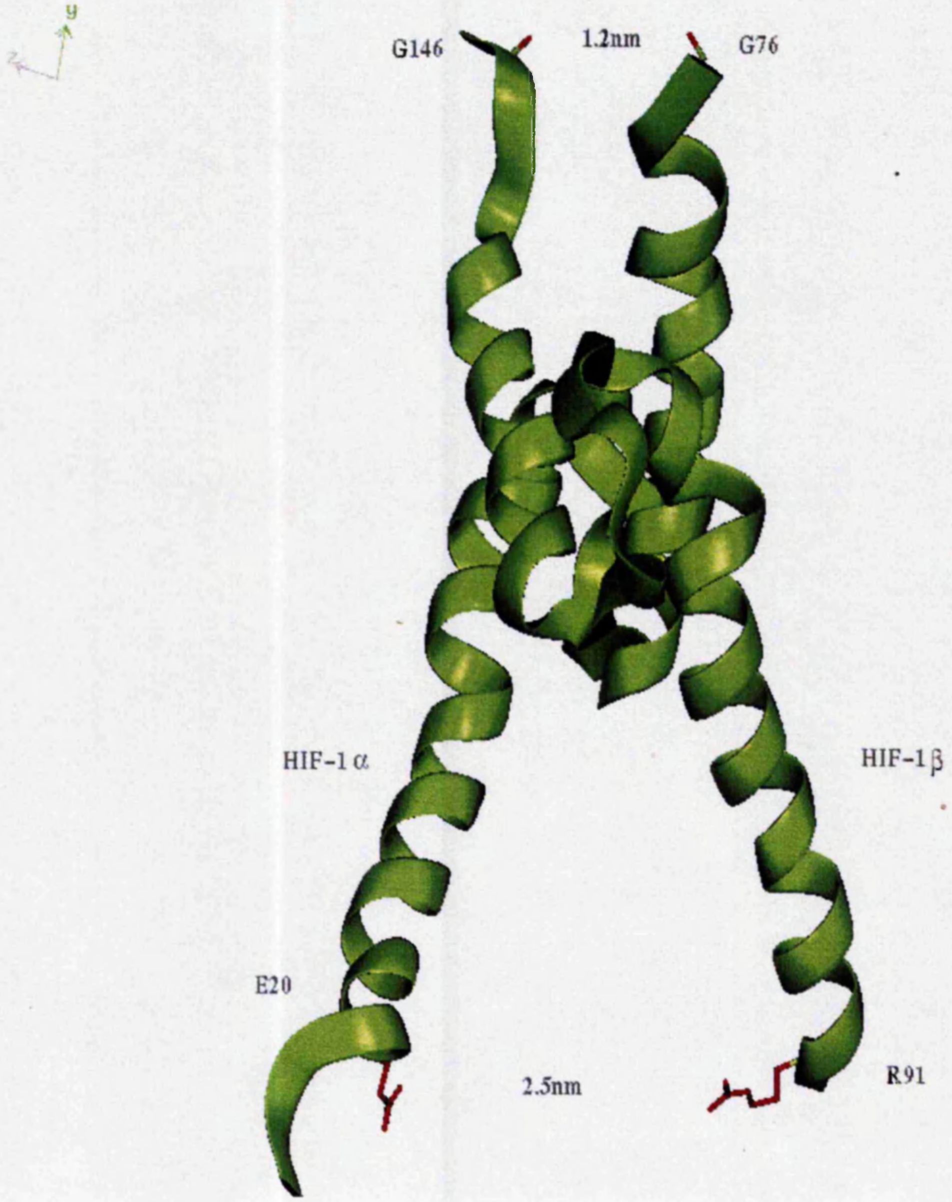


Figure 3.20. The two pairs of residues of HIF-1 to be mutated for FRET studies. The distances between the amino acids are shown

other binding partners for HIF-1α or HIF-1β. Other tests of the model's validity will emerge from the predicted peptide lead-inhibitor studies, preliminary results of which will now be presented.

3.3.2. Peptide synthesis and binding

Figures 3.21(a) and (b) show the hydrophobic surfaces of HIF-1α and Id3, respectively used as a basis for peptide design.

3.3.2.1. Mass spectrometry

It was not possible to carry out mass spectrometric analyses for all 48 peptides, so an analysis was carried out for three representative samples: 12(1), 10(4) and the labelled 4(1). The results of the latter are shown in Figure 3.22.

3.3.2.2. SDS-PAGE assays

The aim of these experiments was to study potential for inhibition of dimerisation of Id3-E47 by the peptides.

3.3.2.2.1. HLH protein dimerisation controls

Gel A shows the results of the Id3-E47, Id3-E12 and MyoD-E47 binding controls, carried out to test that the assay gave reliable results on a known system

As each peptide concentration after cleavage from the solid-phase support was only approximately 1.5μM, as little as possible of Id3 protein was used in these studies to try and observe inhibition. This concentration used in the subsequent peptide studies was determined by the amount of Id3-GST-bead still giving a visible band on the X-ray film, found to be 5μl of Id3-GST diluted with 20μl of GST-bead.

3.3.2.2.2. Acetonitrile control

Acetonitrile up to 40% (v/v) showed no effect on dimerisation of E47 with Id3 indicating that acetonitrile present in the peptide solutions would not affect binding studies (Gel B).

3.3.2.2.3. Peptide assays

The results of incubation of pools *P1-P7* with Id3 are shown in Gels C, D and E. Variation in band intensities (*i.e.* radioactivities of the loaded samples) arising from experimental

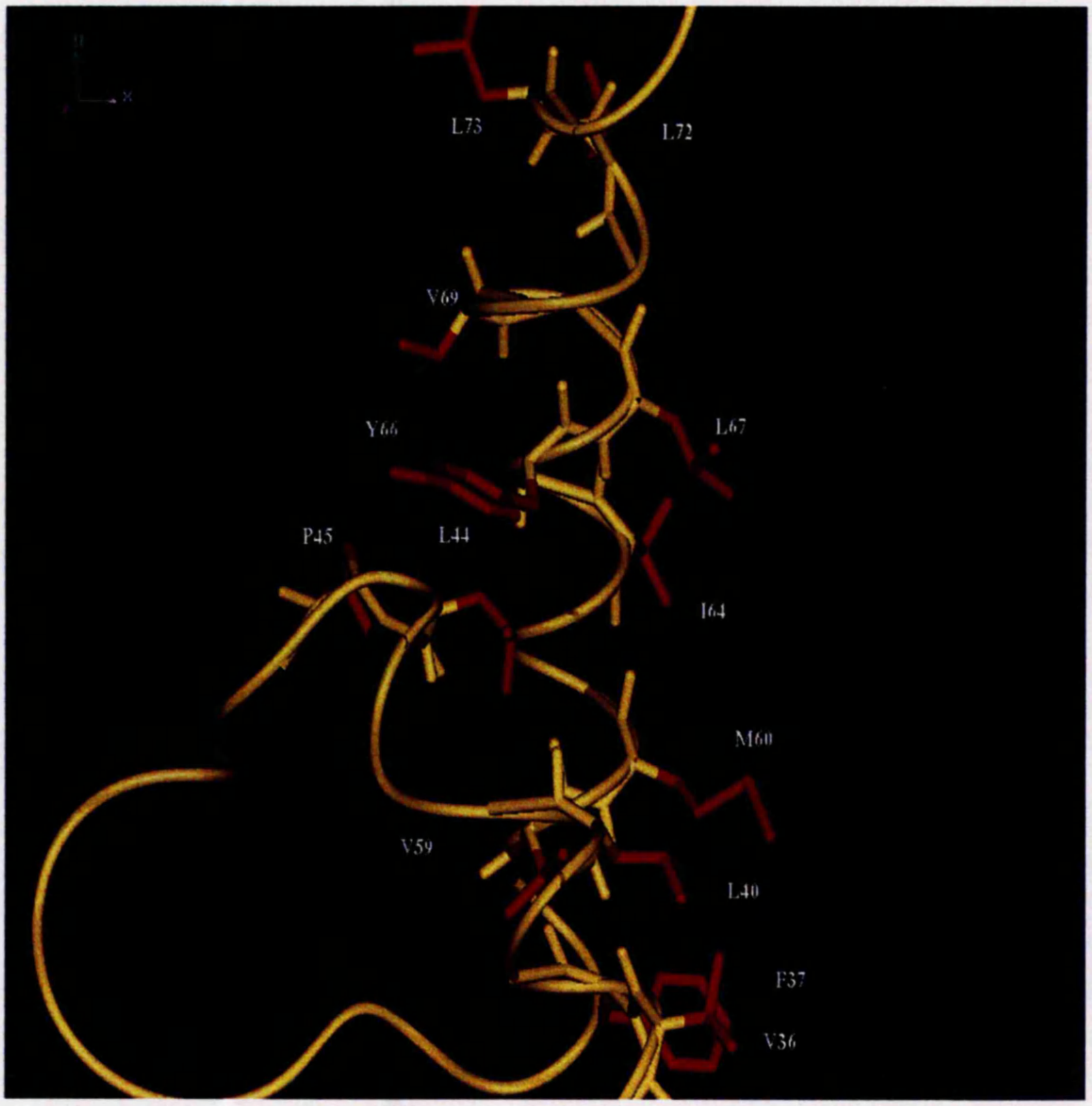


Figure 3.21(a). Hydrophobic residues of the HIF-1α monomer used in the design of complementary peptides.

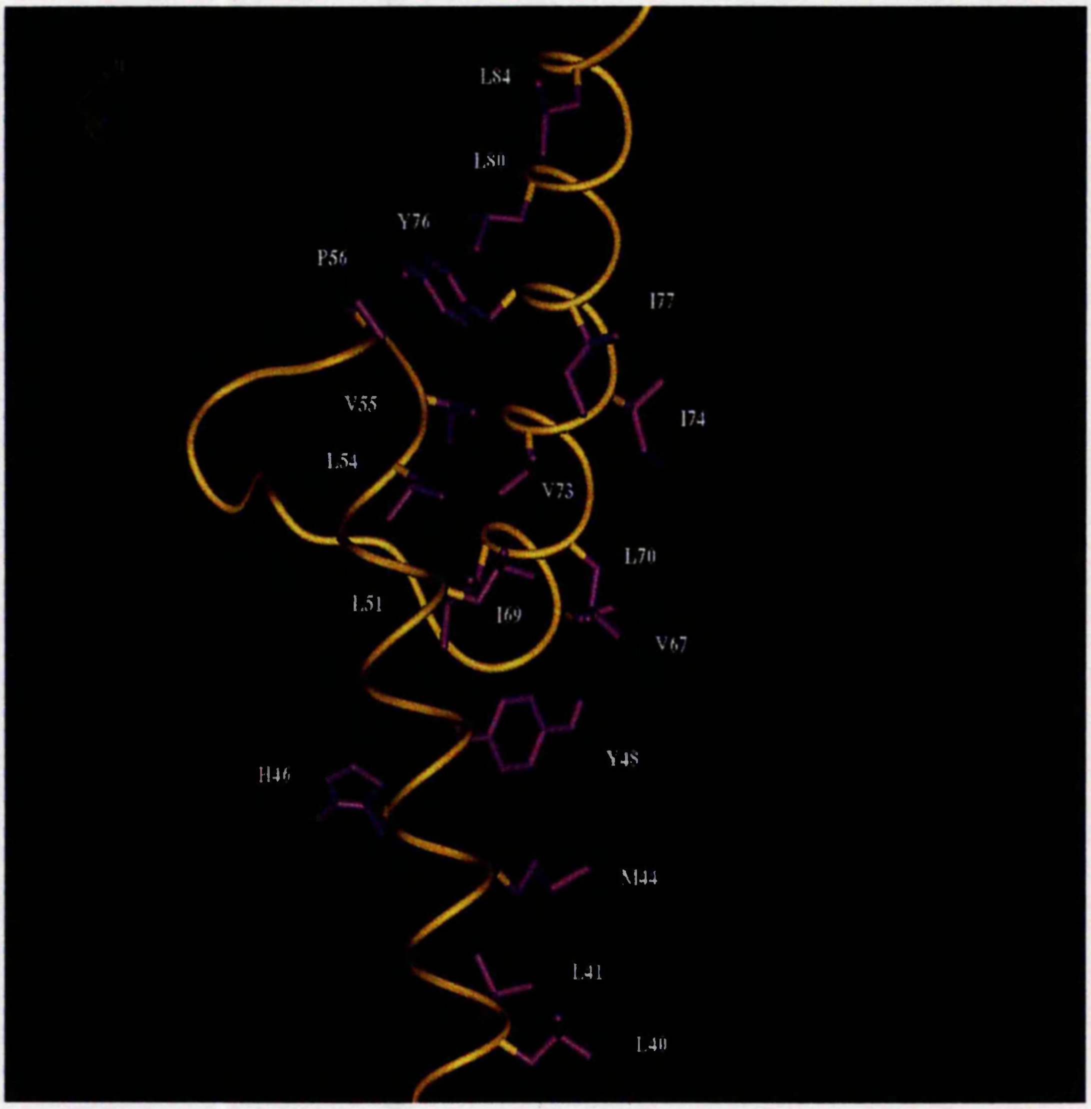


Figure 3.21(b). Hydrophobic residues of the Id3 monomer used in the design of complementary peptides.

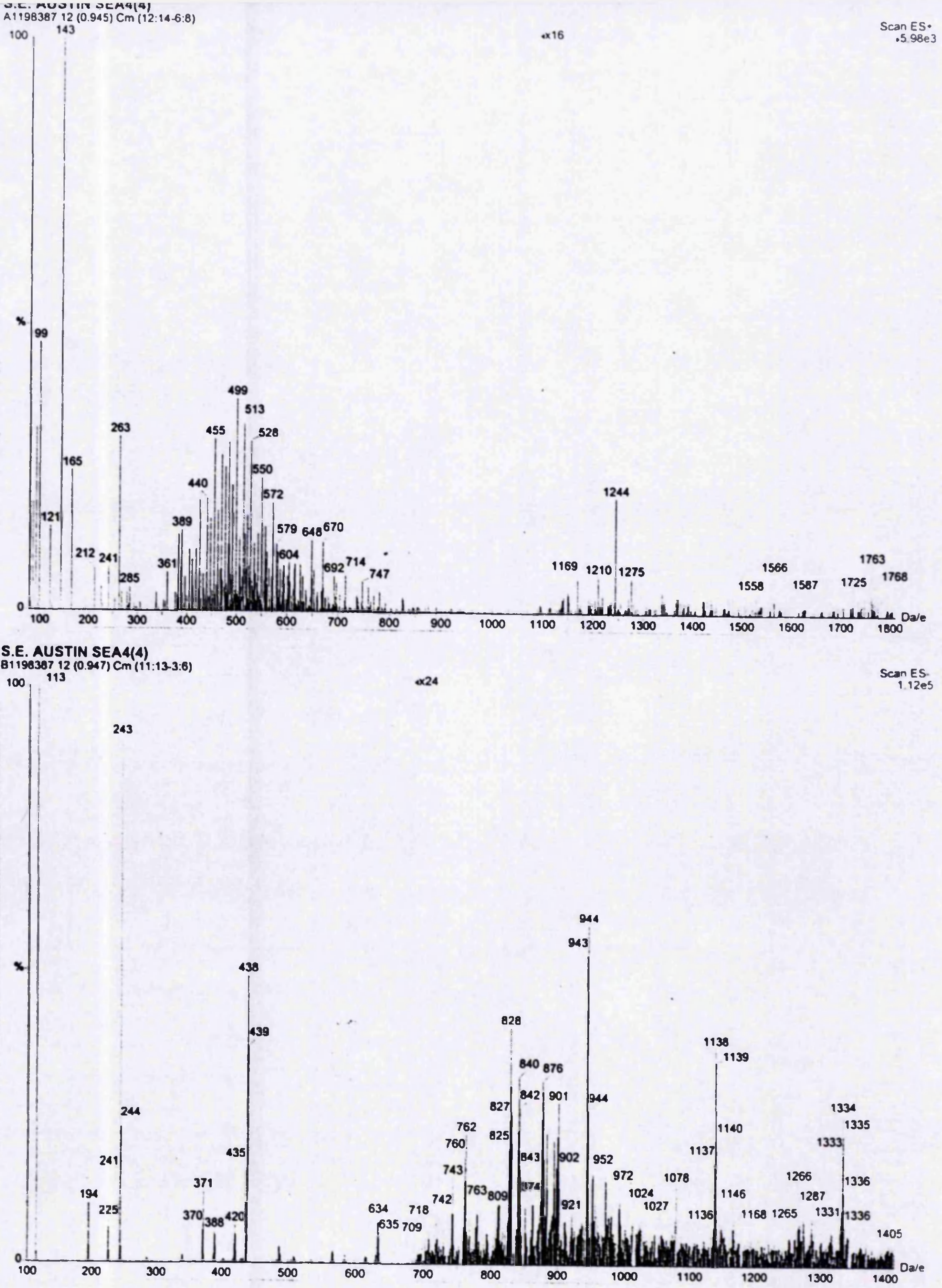


Figure 3.22. Mass spectrum (electrospray) of peptide 4(1). The results are consistent will the following suggested fragmentation patterns. The groups lost from the parent molecule are shown in red and the Da/e value given.

NBD-F-P-L-bA-V-L-G-S-L-P-M-L-I-S-V-COOH ⟶ 1558

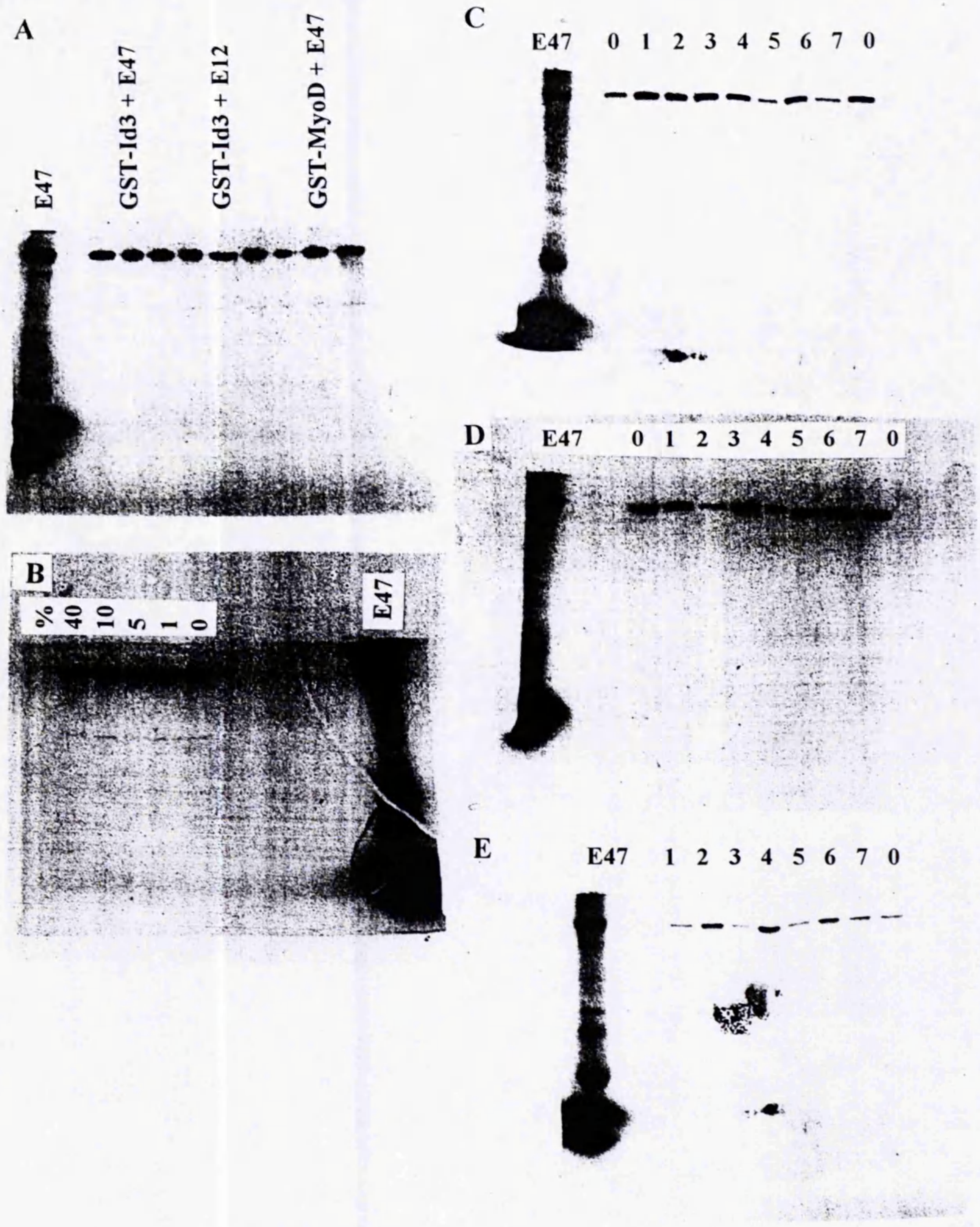
NBD-F-P-L-bA-V-L-G-S-L-P-M-L-I-S-V-COOH ⟶ 1334
|
R

NBD-F-P-L-bA-V-L-G-S-L-P-M-L-I-S-V-COOH ⟶ 1266

NBD-F-P-L-bA-V-L-G-S-L-P-M-L-I-S-V-COOH ⟶ 1244

NBD-F-P-L-bA-V-L-G-S-L-P-M-L-I-S-V-COOH ⟶ 1140
|
R

NBD-F-P-L-bA-V-L-G-S-L-P-M-L-I-S-V-COOH ⟶ 948



Results of SDS-PAGE studies.

Gel **A** shows the binding controls for three known HLH protein dimerisation systems. Samples from the same control were added to three consecutive lanes of the gel. An E12 control is not shown. **B** shows the acetonitrile controls. The percentage of acetonitrile added is labelled. Gels **C**, **D** and **E** are the results of the peptide-binding studies. The numbers refer to the peptide pools and **0** indicates the control (*i.e.* no peptide was added).

error could be confused with an inhibition in dimerisation leading to an intensity reduction. However, the repeat experiment gave a different variation in band intensities. The first experiment (Gel C) indicated that some bands were of a much lower intensity than the control, but upon repetition they were of the same intensity as the control. From these assays, no inhibition of protein dimerisation was seen. Either the peptides are not concentrated enough (1.5µM) (limited by the solid-phase synthetic procedure) or the concentration of protein used is too high (limited by the radioactivity level detectable on autoradiographs), or there really is no inhibition.

3.3.2.3. Resonant mirror biosensor studies

The purpose of these experiments was to assess the technique as a novel qualitative monitor of peptide binding to Id3, not to measure dimerisation inhibition. Preliminary tests were carried out to verify that binding of Id3 to its dimer partner E12 was observable using this technique.

3.3.2.3.1. E12-GST binding to Id3-GST

Id3-GST was immobilised to a level of 120 arc seconds on CM-Dextran *via* coupling to GSH-6-AHA as described and E12-GST binding investigated. Figure 3.23 shows the binding responses for E12-GST (20µl and 30µl) binding to Id3-GST-GSH-6AHA-CM-Dextran. The E12-GST dissociates when washed with PBS running buffer. Unfortunately, due to the limited time that the IAsys biosensor was on loan, it was not possible to carry out any control experiments to eliminate the possibility that the E12-GST was binding to the chip, the GST of Id3-GST or the GSH. In subsequent peptide-binding experiments carried out by Dr. Phil Buckle, (Labsystems, Affinity Sensors), Id3-GST was directly immobilised to the CM-Dextran, to bypass the GSH (and 6-AHA linker). Dr. Buckle repeated the E12-Id3 binding experiment, but did not observe any dimer formation. This may be due to overcrowding at such an immobilisation level.

3.3.2.3.2. Id3-GST immobilisation

The immobilisation of Id3 to a level of 5400 arc seconds directly on the CM-Dextran cuvette is shown in Figure 3.10.

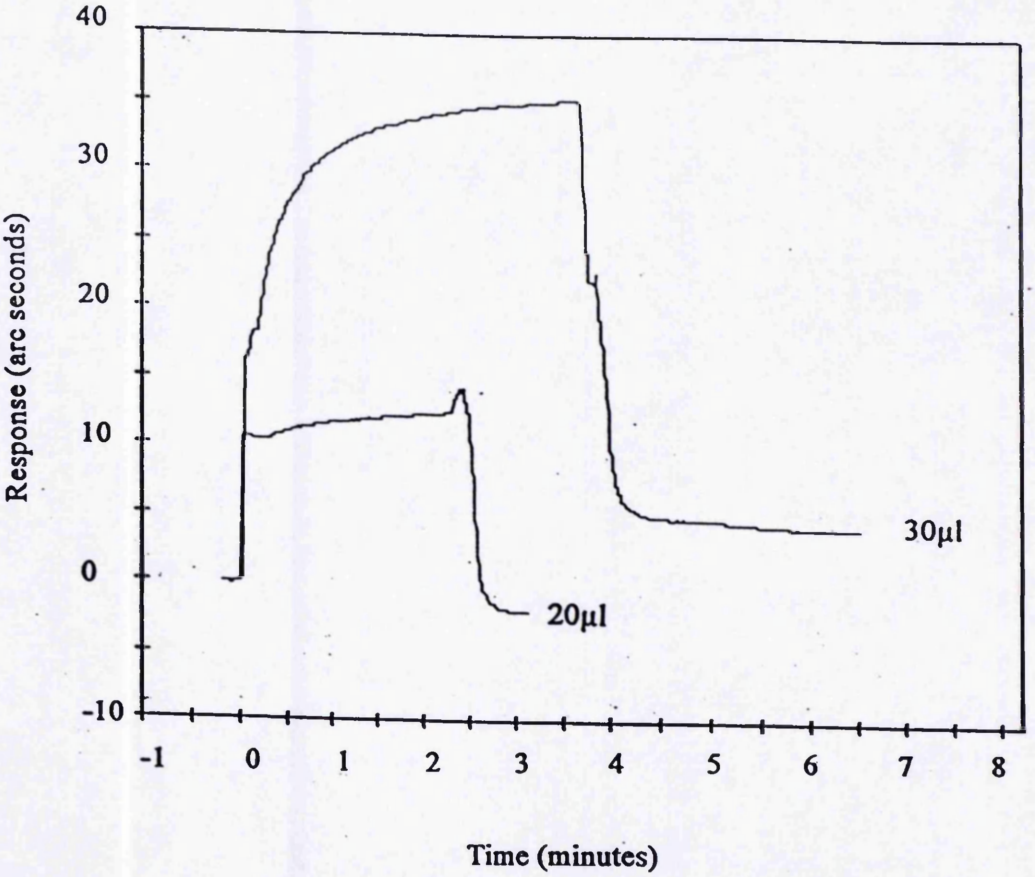


Figure 3.23. Binding responses of aliquots of E12, 20µl and 30µl, (300mgml-1 in PBS), to immobilised Id3-GST (association) followed by washing with 2x50µl PBS (dissociation).

3.3.2.3.3. Peptide binding

(a) Figure 3.24 shows the four synthesised peptides binding to immobilised Id3-GST together with two randomly chosen peptides E and F, the sequences of which are given in Figure 3.8. All four synthesised peptides showed a significant degree of binding, with peptide 9(4) binding best, *i.e.* with highest affinity. Peptides E and F show a much higher binding response because their concentrations are 2.5-3 times higher. Figure 3.25 shows the binding of 9(4) to Id3-GST at concentrations from 50nM to 400nM. The binding response increased with the concentration. It should be possible to calculate rate constants for binding from these responses at different concentrations, but it appeared that as the dextran was overloaded with monomer in the present trial run to enable observation of binding, the curves seem to be diffusion-limited preventing data analysis. If the monomer was loaded to a lower level, this might lower the peptide binding response and give less clear differences between each sample.

A comparison of peptide 9(4) binding to Id3 and the native CM-Dextran cuvette as a negative control experiment is shown in Figure 3.26. No non-specific binding (except for a bulk refractive index effect) was observed on the native CM-Dextran. This does not rule out the possibility that the peptides bind to GST and not Id3.

(b) Figure 3.27 shows the net binding of 12 peptides and peptide 'B1' to Id3-GST and E12-GST after correction with a GST-bound control. Any response is therefore due solely to binding to Id3 or E12.

From the graphs it can be seen that peptide B1 (bradykinin) displays very low affinity for either protein. Its sequence is NH_2-Arg-Pro-Pro-Gly-Phe-Ser-Pro-Phe-Arg-CO_2H and was chosen at random as a peptide not expecting to bind Id3 because of its conformationally constraining proline residues. The sequences of the other peptides are listed below. They were chosen because of their sequence variability.

5(1) DPhe-LPro-DLeu-βAla-LMet-DLeu-Gly-LSer- LIle-LPro-DNVa-LNLe-LLeu-LSer-DNVa
7(1) DPhe-LPro-DLeu-βAla-LIle-DLeu-LSer-Gly- LVal-LPro-LTyr-LLeu-DNVa-Gly- LIle
8(1) DPhe-LPro-LIle-βAla-DLeu-DLeu-LSer-Gly- DLeu-LPro-LTyr-LIle-LSer-Gly- DLeu
5(2) LPhe-LPro-LNLe-βAla-LMet-LIle-Gly- LSer-LIle-LPro-LMet-LNLe-LAla-LSer-LMet
11(2)LPhe-LPro-LNLe-βAla-LMet-DNVa-LSer-LSer-LLeu-LPro-LNLe-DLeu-LSer-LSer-LVal
7(3) DTyr-LPro-LLeu-βAla-LIle-LNLe-LSer-Gly- LVal-LPro-LMet-LVal-LSer-Gly- LNLe
8(3) DTyr-LPro-LVal-βAla-DLeu-LNLe-LSer-Gly- LLeu-LPro-LMet-LNLe-Gly- Gly- DLeu
11(4)LTyr-LPro-LMet-βAla-LMet-LLeu-Gly- LSer-LLeu-LPro-LVal-LIle-Gly- LSer-DNVa
9(1) DPhe-LPro-LIle-βAla-LLeu-DLeu-LSer-LSer-LNLe-LPro-LNLe-DLeu-Gly- Gly- LLeu
9(2) LPhe-LPro-LNLe-βAla-DLeu-DNVa-LSer-LSer-LIle-LPro-LTyr-LVal-LVal-LSer-LNle
9(3) DTyr-LPro-LVal-βAla-LLeu-LNLe-LSer-LSer-DNVa-LPro-DNVa-DLeu-LVal-Gly- LIle
9(4) LTyr-LPro-LMet-βAla-LLeu-LLeu-LSer-LSer-LVal-LPro-LNLe-LLeu-LIle-LSer-DLeu

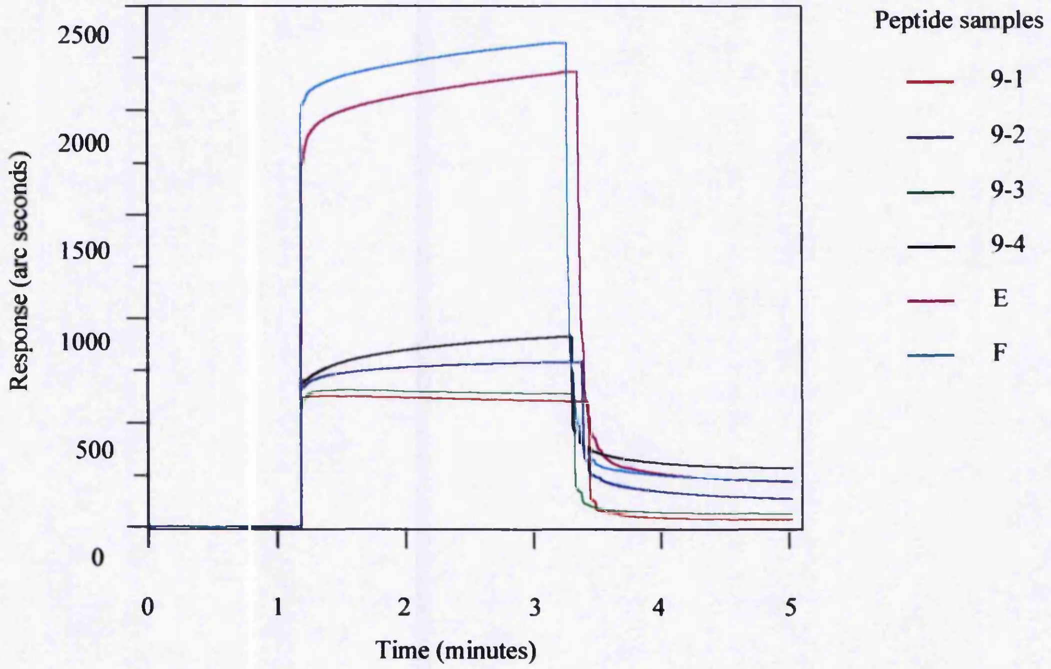


Figure 3.24. Peptide samples binding to Id3-GST immobilised directly on a CM-Dextran cuvette. (Peptides 9-1 to 9-4 are at a concentration of 200nM and peptides E and F are at 590nM and 480nM, respectively).

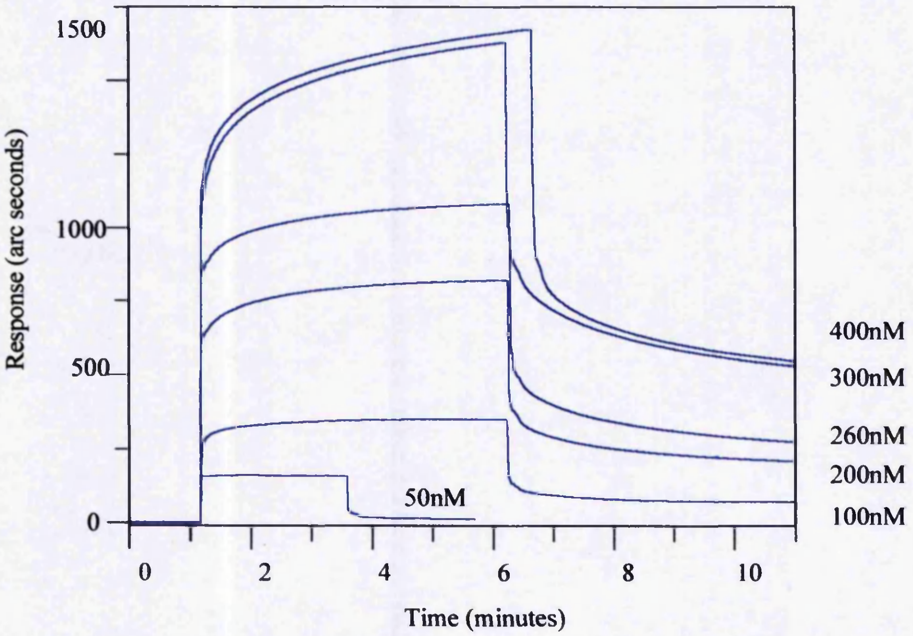


Figure 3.25. Binding response of peptide sample 9(4) at varying concentrations to Id3-GST directly immobilised on a CM-Dextran cuvette.

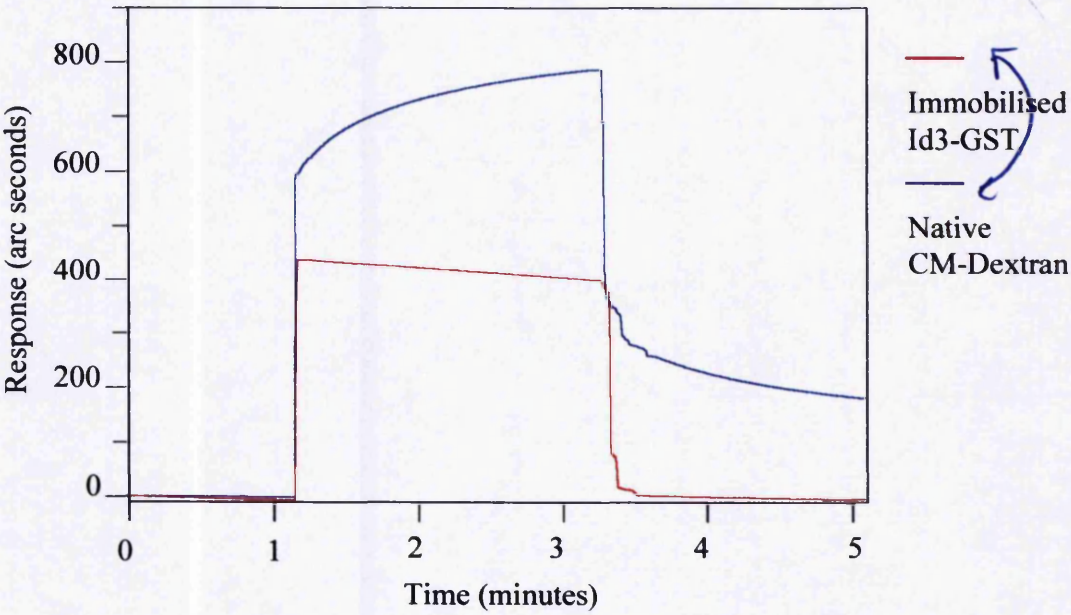


Figure 3.26. Binding response of peptide sample 9(4) to Id3-GST directly immobilised on CM-Dextran: comparison with binding to native CM-Dextran as a negative control.

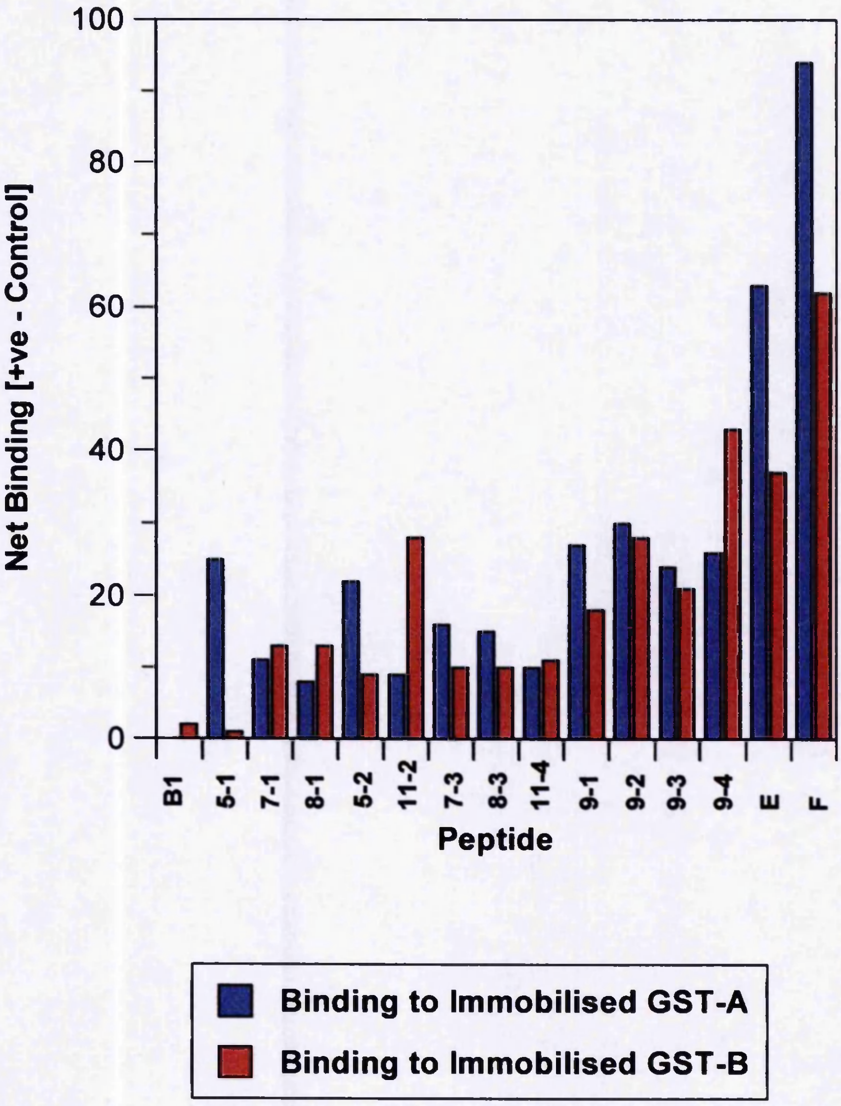


Figure 3.27. Peptide samples binding to GST-Id3 (A) and GST-E12 (B) immobilised on a CM-Dextran cuvette.

Peptides 5(1), 5(2), 7(3), 8(3) and 9(1) showed distinctly higher binding affinities for Id3 than for E12. Peptides 7(1), 8(1), 11(2) and 9(4) displayed higher affinities for E12. From the sequences alone there are no obvious differences which account for these observations. Peptides E and F were present at a concentration 2.5-3 times that of the other peptides, which explains the higher binding response. Figures 3.28(a) and (b) show the possible binding mode of peptide 5(1) to Id3 and of peptide 1(1) to HIF-1α, respectively.

3.3.2.4. Summary

Potential peptide inhibitors of dimerisation of the proteins Id3-E47 and HIF-1 have been designed, synthesised and assessed for binding and inhibition activity against Id3 using two different techniques. Although peptides were observed to bind Id3 and E12 monomers from resonant mirror biosensor experiments, it appears that binding was not strong enough to inhibit dimer formation between these two at the concentrations of protein and peptide used in SDS-PAGE studies. This may be because they were not binding strongly enough to compete with dimer formation, or that they were not binding where they were designed to, *i.e.* at the dimer hydrophobic interface. These preliminary studies give no indication of the strength of binding, (*i.e.* association/dissociation constants), but do indicate the applicability of this technique in the field.

The biosensor results also revealed that the binding of some peptides is significantly more specific for Id3 over E12, with the reverse specificity for other peptides. From just the peptide sequences there appears no explanation for this discrimination, since the peptides are very similar in length, hydrophobicity and amino acid composition. The sequences of hydrophobic residues of the HLH region of Id3 important for dimer formation and those of E12 are very similar [358] (Figure 3.6), which explains the lack of definite specificity for the peptides binding to one protein over the other. In fact, high similarity exists between these hydrophobic residues throughout all HLH proteins. Therefore, a peptide designed to bind the HLH region of one HLH protein will very likely bind to this region in another protein. Trying to target the HLH interface specifically for one protein will be very difficult. Potential dimerisation inhibitors may have to be used in combination with other inhibitors, which act in different manners to prevent transcription, either by blocking the transcription dimer forming, or blocking binding to DNA. Attempting to prevent the dimerisation of proteins by a molecule binding at the interface solely through hydrophobic interactions is unlikely to work alone.

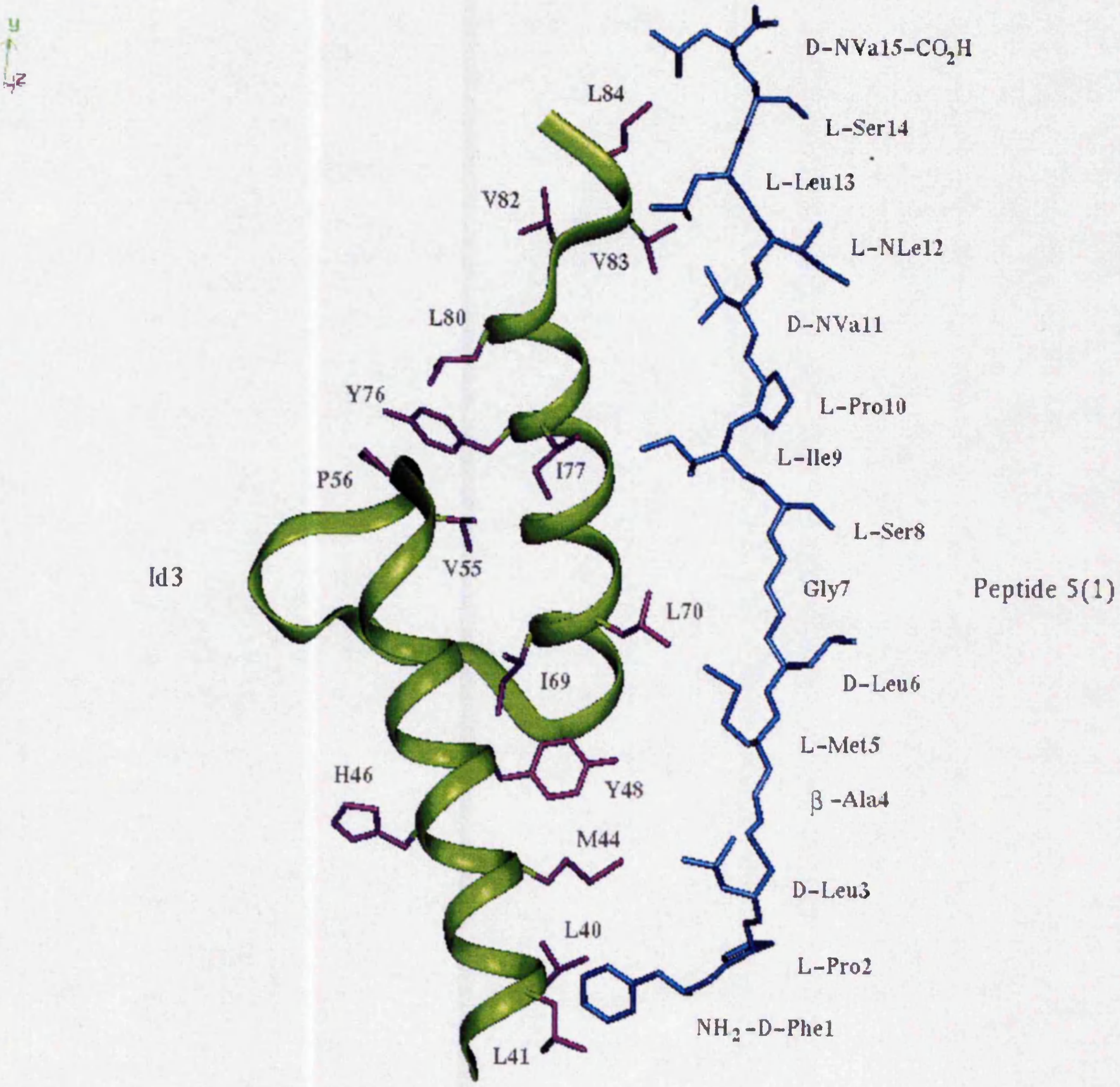


Figure 3.28(a). Possible binding orientation of peptide 5(1) to the Id3 monomer.

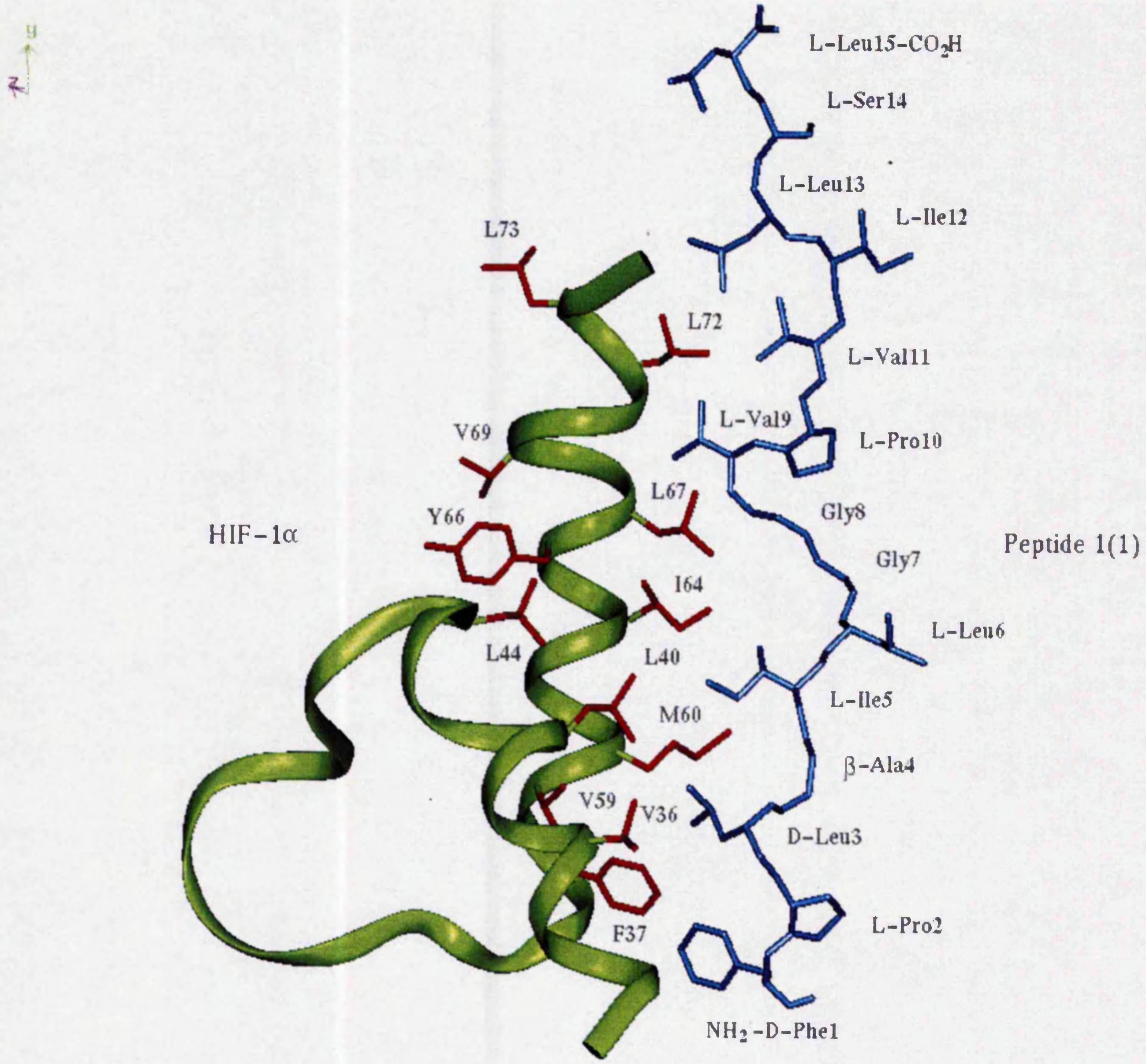


Figure 3.28(b). Possible binding orientation of peptide 1(1) to HIF-1α.

Studies on MyoD-E47 heterodimers by Wendt *et al.* [415], suggest that MyoD does not dimerise with E47 under dilute conditions in the absence of DNA and that assembly of these bHLH-DNA complexes is apparently governed by the strength of each subunit's interaction with the DNA and not by the strength of protein-protein interactions at the dimer interface. Future attempts at disrupting specific bHLH dimerisation and the initiation of transcription of certain genes could be focussed on designing longer molecules to specifically bind the DNA E-box enhancer region as well as simultaneously binding the basic DNA-binding part of the transcription factor. This will achieve higher specificity in gene targeting.

Some of the peptides were labelled in this thesis at their N-termini with the fluorogenic reagent NBD-Cl. This can be used as either a donor or acceptor dye in fluorescent studies. Colleagues have now produced mutations of Arnt and HIF-1α (Dr. Karen King personal communication) with cysteine residues at certain sites allowing for covalent attachment of donor/acceptor molecules. This provides the basis of a FRET-based assay of peptide-target interactions using an NDB-compatible acceptor (e.g. RTC) or donor (e.g. AEDANS) on the protein partner. This approach is now underway in our laboratory and should provide another qualitative method for observing peptide binding.

CHAPTER FOUR

A LIGAND-DOCKING STUDY OF THE ENZYME TRYPANOTHIONE REDUCTASE

4.1. INTRODUCTION

4.1.1. Molecular docking

A docking method seeks to find ways of fitting two molecules together in energetically favourable conformations, whether it is two proteins or a small molecule binding to a receptor, e.g. a protein. All ligand docking programs must solve three fundamental problems:

(i) where, in a relatively large site, to fit the ligand,

(ii) what conformations of ligand and receptor best complement each other, and

(iii) how to evaluate the energies of the various complexes.

Rigorously solving these problems requires elaborate energy calculations and consideration of many more conformational and configurational degrees of freedom than is now computationally feasible for a docking method. To get reasonable answers in reasonable CPU time,. docking algorithms simplify the problems. Common approximations include treating intrinsically flexible ligands and proteins as rigid objects, modelling explicit water molecules by a dielectric continuum and using enthalpy as a proxy for free energy. Ideally, any procedure should take into account the minimal amount of information relevant to recognition, leaving out less important, or less defined details which would discredit the evaluation of the fit between molecules.

Most algorithms can generate many possible docked structures and so they also require a means to rank each structure. The algorithms developed to tackle this problem can be characterised according to the number of degrees of freedom that they ignore. The simplest algorithms treat the two molecules as rigid bodies and the more sophisticated use simulated annealing to search conformational space and also allow several torsional degrees of freedom in a flexible ligand and flexible receptor site to be searched.

Docking programs may be grouped into five general families, depending on how they address these problems and what simplifications they use: descriptor, grid, fragment, kinetic and genetic methods (Table 4.1).

Program	Algorithm	Flexible ligand?	Protein-protein docking	Database search	Ligand design	Conform-ational search	Scoring
CAVEAT	Descriptor	No	No	Yes	No	No	Steric Fit
CLIX	Descriptor	No	No data	Yes	No	No	Force Field
DOCK	Descriptor	No	Yes	Yes	No	No	Force Field
FLOG	Grid Search	'Quasi'	No	Yes	No	No	Force Field
Soft docking	Grid Search	No	Yes	No	No	No	Polar/Apolar contact
GRAMM	Grid-based Fast Fourier Transform	No	Yes	No	No		Fourier correlation
FLEXX	Fragment	No	No	No	No	No	Force Field
GROW	Fragment	No	No	No	Yes	Yes	Force Field
LUDI	Fragment	No	No	Limited	Yes	Possible	Force Field
AUTODOCK	Kinetic	Yes	Yes	No	No	Yes	Force Field
FLEXIDOCK	Genetic	Yes	Yes	No	No	Yes	Force Field
GOLD	Genetic	Yes	No	No	No	Yes	Force Field
FTDOCK	Fourier Transform	Yes	Yes	No	No	Yes	Fourier correlation

Table 4.1. Examples of ligand docking programs

CAVEAT [416]; CLIX [417]; DOCK [418]; FLOG [419]; Soft docking [420]; FLEXX [421]; GRAMM [422-425]; GROW [426]; LUDI [26]; AUTODOCK [427]; FLEXIDOCK (SYBYL, Tripos Inc.); GOLD [428]; FTDOCK [429]. FLEXIDOCK is the only algorithm that allows for limited flexibility in a number of side chains of the protein receptor.

4.1.2. Algorithms

4.1.2.1. Descriptors.

The protein is first analysed for regions of likely complementarity. These 'hot-spots' are areas on the protein surface where a ligand atom might fit well. They describe the binding region. Ligand atoms are matched to receptor hot-spots and thus generate orientations of the ligand in or on the protein. For any given ligand, many orientations are sampled and evaluated for goodness of fit, typically by using a molecular mechanics-type energy function. Though not exhaustive, descriptor methods are fast and can often sample densely in a particular region of the protein. Disadvantages are that they rely on being to identify the hot-spots well and most treat proteins and ligands as rigid objects. DOCK [418] for example, uses a simple algorithm, which treats the two molecules as rigid bodies and only explores the six degrees of translational and rotational freedom. It was designed to find molecules with a high shape complementarity to the binding site. The program first derives a 'negative image' of the binding site, which consists of a collection of overlapping spheres of varying radii that touch the molecular surface at just two points. Ligand atoms are then matched to the sphere centres to find to find matching sets in which all the distances between the ligand atoms in the set are equal to the corresponding sphere centre-sphere centre distances (within some user specified tolerance). The ligand can then be orientated within the site by performing a least squares fit of the atoms to the sphere centres. An advantage of descriptor methods is that continuous space is searched, unlike grid methods (see below).

4.1.2.2. Grid searches

Grid searches fit the ligand to the protein by rotating and translating the ligand in discrete steps on a grid whilst holding the protein rigid. Energies are calculated by tri-linear interpolation of affinity values of the eight grid points surrounding each of the atoms in the ligand. Because continuous space is not searched, the accuracy of grid methods is limited to the resolution of the grid size used, but the higher the accuracy, the longer it takes to sample space. Most grid methods treat ligand and protein as rigid objects, though some allow for conformational relaxation [430]. These programs have been used extensively for single ligand docking programs [420,425,430].

FLOG (Flexible Ligands Orientated on Grid) [419] searches a database of three-dimensional coordinates to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure. Ligand flexibility is addressed by including different explicit conformations of each structure in the database. The program is similar to DOCK, using a 'match-centre' representation of the volume of the binding cavity, but also uses a grid representation of the receptor to assess the fit of each orientation.

4.1.2.3. Fragment

Fragment methods identify regions of high ligand complementarity on a protein surface by docking functional groups independently into protein cavities. Ligands are broken down into fragments so that many of the configurational and conformational issues in docking disappear. This is done at the expense of connectivity information, which fragment methods can in principle gain back by reconnection algorithms at the end of the calculation. Fragment methods can also be used for molecular elaboration of existing inhibitors. They can be useful for novel inhibitor design when working with molecules whose synthetic chemistry is as modular as the computer-generated fragments, such as peptides and oligonucleotides. The CONCEPTS [431], LUDI [26] and GROW [426] programs are good examples of this idea, as is FLEXX [421], which uses fragments defined as a connected part of a molecule containing only complete ring systems.

4.1.2.4. Kinetic

Kinetic docking techniques sample the surface of potential receptor sites using molecular dynamics or simulated annealing to fit ligands. These methods merge the configurational and conformational aspects of the docking problem smoothly. A disadvantage is that the complex topography and multiple minima of molecular potential surfaces can lead to long running times and minima traps. The *ab initio*, pseudo-Monte Carlo method of Totrov and Abagyan [432] and AUTODOCK [427] are examples. AUTODOCK uses a Metropolis Monte Carlo simulated annealing technique with energy evaluation using pre-calculated grids of molecular affinity potentials [433]. The user can specify rotatable bonds in the ligand, but the protein remains rigid [434].

4.1.2.5. Genetic

Genetic (evolutionary) algorithms, based on biological evolution, are designed to find optimal solutions to problems. Initially, a 'population' of possible ligand structures is generated. The 'fitness' of each member of the population is then calculated with respect to binding energy and a new population generated from the old one with a bias towards the fitter members, so introducing an evolutionary pressure into the algorithm. The least-fit members of the population are replaced by the 'offspring'. Each member of the population is coded for by a 'chromosome', usually stored as a linear combination of 'bits'. Each chromosome codes for the values of the rotatable bonds in the molecule (and therefore internal conformation of the ligand) and its orientation within the receptor site, thus allowing flexibility in the ligand. A new population is generated using 'operators' (commonly *reproduction*, *crossover* and *mutation*) that act on the chromosomes. As the bits are selected and changed, so both the orientation and the internal conformation of the ligand will vary as the populations evolve. The energy score of each docked structure within the site acts as the fitness function used to select the individuals for the next iteration.

GOLD (Genetic Optimisation for Ligand Docking [428]) and FLEXIDOCK (SYBYL, Tripos, Inc) are two examples of genetic algorithms used in docking applications and there are others [435-437]. FLEXIDOCK confers some torsional freedom in the receptor site as well as ligand by allowing the user to specify the flexible side-chain bonds. A traditional genetic algorithm described above maps genotypes to phenotypes (*i.e.* converts the 'chromosome' into the ligand's coordinates) by a developmental mapping function. At each generation, however, a user-defined fraction of the population can undergo a local search to explore space for lower energy structures. It is possible to inversely map the phenotypes (ligand conformations) resulting from this search to their genotype, thus further improving the fitness of the parent population. This is called the Lamarckian genetic algorithm [438], an allusion to Jean-Baptiste de Lamarck's discredited assertion that phenotypic characteristics acquired during an individual's lifetime can become inheritable traits [439].

4.1.2.6. Non-rigid body searching by Fourier Transform docking

A major problem in rigid body docking is that the algorithm must be sufficiently soft to manage conformational changes, yet specific enough to identify the correct solution. One alternative already described is non-rigid body searching, *i.e.* allowing flexibility in the ligand and/or receptor side-chains, but another approach is to use a 'soft' treatment of electrostatic interactions, e.g. in FTDOCK (Fourier Transform docking) [429]. This method uses a fast Fourier Transform to search rapidly the translational space of two rigidly rotated molecules. In this program, instead of measuring specific charge-charge interactions, point charges are measured as grid points and dispersed to simulate side-chain movement. The theory, beyond the scope of this thesis, is based on an algorithm which measures shape complementarity by Fourier correlation [425] using a fast Fourier Transform and Fourier correlation theory to scan rapidly the translational space of two rigidly rotating molecules. The GRAMM program [440] uses an algorithm of grid-based correlation by Fast Fourier Transform is low-resolution docking studies (see below).

4.1.2.7. Low-resolution docking

One of the crucial factors in molecular recognition procedures is the multiplicity of local minima of intermolecular energy, or a large amount of high-scoring false-positive matches. Existing approaches are actually designed for high-resolution structures, determined by experiment or modelling. The elaborate character of local structural details contains a huge amount of information and often interferes with the search procedure and/or demands excessive computational time. One of the best known approaches to alleviate these problems was designed by Wodak and Janin [441] who reduced atomic contacts to residue-residue interactions. Another significant problem in protein docking is the problem of structural data inaccuracy, e.g. the inaccuracy in X-ray data due to natural flexibility of atomic fragments, poor structure resolution *etc*. Important factors in docking studies are also conformation changes on complex formation. Some of the approaches include a degree of tolerance to molecular structure fluctuations [420,425] and truncation of certain side chains [442]. One approach to overcome these problems is based on ultralow (~7Å resolution) representation of molecular structures, giving an opportunity to filter out (average) all high-resolution structural details and still predict most of the structural features of the ligand-receptor complex [423]. This approach greatly improves the signal-to-noise ratio in determining best fit and moves the structure inaccuracy tolerance to the range of the macrostructure [422]. The program GRAMM, based on these approaches [422-425], has been used in the low-resolution study of a hemagglutinin-antibody complex [440].

4.1.3. Energy evaluation

A successful docking methodology requires an energy function, which is selective, *i.e.* capable of discriminating between a mis-docked structure and correct ligand-receptor structure, e.g. a low-energy conformation or one consistent with the crystallographic structure of the complex. A good energy function should also be efficient, allowing the desired minimum to be located reasonably rapidly and the landscape on which the ligand moves should be relatively smooth with no large energy barriers separating different structures.

Older docking programs simulated docking by matching shape or surface complementarity [425,443]. Some achieved this by modelling the hydrophobic effect during association from the change in solvent-accessible surface area of molecular surface area, e.g. [430]. Most recent programs have functions to model Van der Waals' and electrostatic interactions, hydrogen bonding and solvent-screening effects, e.g. AUTODOCK uses a sigmoidal distance-dependent dielectric function in the electrostatic interaction energy grid calculation to account for the solvent screening effect [444]. More recently, improvements have been made to include desolvation, free energy and entropic terms.

Horvath's [445] continuum solvent model [446] for desolvation approximates the change of the electrostatic energy of a charged atom that approaches a low-dielectric atom, which displaces the high-dielectric solvent. The hydrophobic effect is represented as a term proportional to the surface area buried during binding. In Horvath's approach, simplified 'virtual physical laws' were defined to govern the behaviour the ligands so that binding could even be characterised by a 'virtual' entropy index. Entropy is a physical quantity which has not been dealt explicitly with in many docking programs.

Different docking methods use different energy functions to evaluate the docked structures with different weightings given to the components. For example, the principle driving force of the GOLD algorithm [428] is the identification of hydrogen-bonding interactions between the ligand and the protein, as hydrogen bond motifs have been directly encoded into its genetic algorithm. It is thus unable to make binding predictions for ligands with no polar groups. GOLD describes desolvation by using a term for hydrogen bonding which takes into account the fundamental requirement that water must be displaced from both donor and acceptor before a bond forms. The term does not assume that all hydrogen bonds have a directional preference along the acceptor lone-pair, unlike some algorithms. The fact that some prefer to form bonds in the plane of the lone-pairs or have no preference in relation to the lone pair positions is taken into account.

The selectivity and efficiency of a range of energy functions has been assessed by Vieth *et al.* [447] using five different ligand-receptor complexes. The study includes the effect of dielectric constant, solvation, the scaling of surface charges, reduction of Van der Waals' repulsion and non-bonded cutoffs. It was found that energy functions displaying selectivity include a variety of distance-dependent dielectric models together with truncation at 8Å of the non-bonded interactions. A constant dielectric is disadvantageous because it selects predominantly on the basis of electrostatics and, as in the case of a low value for a dielectric constant, probably disrupts the balance between electrostatics and packing (Van der Waals' interactions) necessary for tight and specific interactions. Incorporation of the Poisson-Boltzmann continuum solvation model (which treats the solvent as a body of constant low dielectric between 2 and 4, and models the solvent as a continuum of high dielectric [448]) flattens the energy surface and reduces the gap between the energy distributions of the docked and mis-docked structures. Truncation of the non-bonded interactions obviously decreases the number of interactions to compute and hence the computational cost of docking. In addition, short values for the non-bonded interaction truncation should favour tightly packed structures and highly localised electrostatic interactions in the active site over more loosely packed structures with screened electrostatic interactions. The most selective potential, however, was found to be the most inefficient, requiring the longest CPU time.

The largest improvements in docking efficiency (number of docked structures per unit time) come from a reduction of Van der Waals' repulsion and a reduction of surface charges.

Reduction of Van der Waals' repulsion means that the resulting conformational transition barriers are much smaller than with unmodified Van der Waals' potential and allows a ligand to penetrate the interior of the protein with a relatively small energetic penalty (*i.e.* the local energy barriers are very small). Reduction of the surface side-chain charges accounts for the flexibility of the surface side-chains and resulting surface charge delocalisation due to this flexibility.

4.1.4. The AUTODOCK docking program

In the Aims section it was proposed to use a docking program to investigate the potential binding modes of inhibitors of the parasitic enzyme trypanothione reductase and to assess the ability of the program in rationalising *in vitro* inhibition data from our laboratory.

The program used was AUTODOCK [427], which uses a Monte Carlo simulated-annealing technique for configurational exploration of a flexible ligand in a rigid receptor site with a rapid energy evaluation using grid-based molecular affinity potentials based on the AMBER force field [181]. Combining the advantages of a large search space and a robust energy evaluation, it can be up to ten times faster than molecular dynamics for both large and small active site docking [449]. AUTODOCK was chosen because it has been already been applied to and proved successful in many docking studies, e.g. the docking of; isomaltose analogues, methyl α-acarviosinide and glucosyl disaccharides in the glucoamylase active site [450-452], citrate to aconitase [453], benzamidine to β-trypsin, camphor to cytochrome P-450, biotin to streptavidin, the cyclic urea protease inhibitor XK-263 to HIV-1 protease and others [454,455]. The program is easy to understand, user friendly, easy to modifiy for the users purposes and not as 'black box' as some programs can be. The results of the docking procedure can be easily viewed as files compatible with the SYBYL software run in-house. Because the receptor remains rigid, AUTODOCK is less complicated and has the potential to run more efficiently than programs dealing with flexible receptor sites.

AUTODOCK comprises several sections now outlined in the order they are used in practice.

4.1.4.1. AUTOTORS

The AUTOTORS tool allows the rotatable bonds for the ligand to be defined.

4.1.4.2. Modelling hydrogen-bonds

A molecule often contains two types of hydrogens: polar, *i.e.* those bonded to oxygen, nitrogen or sulphur and capable of forming hydrogen bonds, or non-polar. Hydrogen bonds are often important in ligand binding and these interactions can be modelled explicitly in AutoDock with polar hydrogens being allowed to freely rotate. The relatively unimportant non-polar hydrogen atoms have their partial atomic charges united with the heavy atoms (usually carbon) to which they are attached. This saves having two types of hydrogen grids (see below), thus conserving disc space and computational time. To 'unite' a non-polar hydrogen atom's partial charge, the latter is added to that of the heavy atom to which it is bonded and the hydrogen can then be deleted from the molecule. This can be done in AutoTors. Polar hydrogens remain present and retain their charges.

Pairwise-atomic interaction energies can be approximated by the general Lennard-Jones potential function (4.1).

$$V(r) \approx C_n r^{-n} - C_m r^{-m} \qquad \text{Equation (4.1)}$$

where $V(r)$ is potential energy expressed as a function of inter-nuclear separation, r, m and n are integers and C_n and C_m are constants whose values depend on the depth of the energy well and the equilibrium separation of the two atoms' nuclei. The potential is characterised by an attractive part that varies as r^{-n} and a repulsive part that varies as r^{-m}. Typically, the 12-6 Lennard-Jones parameters (n=12, m=6) are used to model the Van der Waals' forces experienced between two atoms. The 12-10 form of the expression (n=12, m=10) can be used to model hydrogen bonds more accurately. This form modulates the pairwise interaction by a function of the cosine of the hydrogen bond angle and thus takes into account the directionality of hydrogen bonds. The user must specify the appropriate 12-10 parameters for hydrogen atoms bonded to O, N or S in the ligand in the AutoGrid parameter file described below. For the Van der Waals' interaction of all the other atom types, the Lennard-Jones 12-6 potential is applied.

4.1.4.3. AutoGrid

AutoGrid detects hydrogen bond parameters in the grid parameter file and, if n in equation (1) is not 6 but 10, the pair-wise interaction is modulated by a function of the cosine of the hydrogen-bond angle. This takes into account the directionality of hydrogen-bonds. Figure 4.1 is an example of a grid parameter file. The first column of atoms represents those in the ligand and the second column contains those found in the protein. Note the 12-10 parameters for the O-H and S-H hydrogen bonding interactions.

The final energy evaluation of the docked structures is rapidly achieved by pre-calculating atomic affinity potentials for each atom type in the ligand molecule as described by Goodford [456].

```
receptor trtcks_nomep_min25.pdbq        #macromolecule
gridfld trtcks_nomep_min25_maps.fld     #grid_data_file
npts 60 60 60           #numxyzpoints
spacing .375            #spacing/Angstroms
gridcenter -24.286 -7.056 4.167         #center_of_grids or auto
types CNOSX             #atom_type_names
map trtcks_nomep_min25_C.map            #atomic_affinity_map
nbp_r_eps  4.00 0.1500  12   6  #C-C non-bond Rij & epsilonij
nbp_r_eps  3.75 0.1549  12   6  #C-N non-bond Rij & epsilonij
nbp_r_eps  3.60 0.1732  12   6  #C-O non-bond Rij & epsilonij
nbp_r_eps  4.00 0.1732  12   6  #C-S non-bond Rij & epsilonij
nbp_r_eps  3.00 0.0548  12   6  #C-H non-bond Rij & epsilonij
nbp_r_eps  4.04 0.2035  12   6  #C-X non-bond Rij & epsilonij
nbp_r_eps  4.04 0.2035  12   6  #C-X non-bond Rij & epsilonij
nbp_r_eps  4.04 0.2035  12   6  #C-X non-bond Rij & epsilonij
map trtcks_nomep_min25_N.map            #atomic_affinity_map
nbp_r_eps  3.75 0.1549  12   6  #N-C non-bond Rij & epsilonij
nbp_r_eps  3.50 0.1600  12   6  #N-N non-bond Rij & epsilonij
nbp_r_eps  3.35 0.1789  12   6  #N-O non-bond Rij & epsilonij
nbp_r_eps  3.75 0.1789  12   6  #N-S non-bond Rij & epsilonij
nbp_r_eps  2.75 0.0566  12   6  #N-H non-bond Rij & epsilonij
nbp_r_eps  3.79 0.2101  12   6  #N-X non-bond Rij & epsilonij
nbp_r_eps  3.79 0.2101  12   6  #N-X non-bond Rij & epsilonij
nbp_r_eps  3.79 0.2101  12   6  #N-X non-bond Rij & epsilonij
map trtcks_nomep_min25_O.map            #atomic_affinity_map
nbp_r_eps  3.60 0.1732  12   6  #O-C non-bond Rij & epsilonij
nbp_r_eps  3.35 0.1789  12   6  #O-N non-bond Rij & epsilonij
nbp_r_eps  3.20 0.2000  12   6  #O-O non-bond Rij & epsilonij
nbp_r_eps  3.60 0.2000  12   6  #O-S non-bond Rij & epsilonij
nbp_r_eps  1.90 5.0000  12  10  #O-H non-bond Rij & epsilonij
nbp_r_eps  3.65 0.2349  12   6  #O-X non-bond Rij & epsilonij
nbp_r_eps  3.65 0.2349  12   6  #O-X non-bond Rij & epsilonij
nbp_r_eps  3.65 0.2349  12   6  #O-X non-bond Rij & epsilonij
map trtcks_nomep_min25_S.map            #atomic_affinity_map
nbp_r_eps  4.00 0.1732  12   6  #S-C non-bond Rij & epsilonij
nbp_r_eps  3.75 0.1789  12   6  #S-N non-bond Rij & epsilonij
nbp_r_eps  3.60 0.2000  12   6  #S-O non-bond Rij & epsilonij
nbp_r_eps  4.00 0.2000  12   6  #S-S non-bond Rij & epsilonij
nbp_r_eps  2.50 1.0000  12  10  #S-H non-bond Rij & epsilonij
nbp_r_eps  4.04 0.2349  12   6  #S-X non-bond Rij & epsilonij
nbp_r_eps  4.04 0.2349  12   6  #S-X non-bond Rij & epsilonij
nbp_r_eps  4.04 0.2349  12   6  #S-X non-bond Rij & epsilonij
map trtcks_nomep_min25_X.map            #atomic_affinity_map
nbp_r_eps  4.04 0.2035  12   6  #X-C non-bond Rij & epsilonij
nbp_r_eps  3.79 0.2101  12   6  #X-N non-bond Rij & epsilonij
nbp_r_eps  3.65 0.2349  12   6  #X-O non-bond Rij & epsilonij
nbp_r_eps  4.04 0.2349  12   6  #X-S non-bond Rij & epsilonij
nbp_r_eps  3.04 0.0743  12   6  #X-H non-bond Rij & epsilonij
nbp_r_eps  4.09 0.2760  12   6  #X-X non-bond Rij & epsilonij
nbp_r_eps  4.09 0.2760  12   6  #X-X non-bond Rij & epsilonij
nbp_r_eps  4.09 0.2760  12   6  #X-X non-bond Rij & epsilonij
elecmap trtcks_nomep_min25_e.map        #electrostatic_PE_map
dielectric -1.          #distance-dep.diel=-1,constant>0
fmap trtcks_nomep_min25_f.map           #floating_grid
```

Figure 4.1. An example of a grid parameter (input) file for AUTOGRID.

In the AUTOGRID part of AUTODOCK the active site of the protein is embedded in a three-dimensional grid of user-defined size with probe atoms (defined by each atom-type in the ligand) placed at each grid point. The energy of interaction of this single atom with all included protein atoms within a non-bonded cut-off radius of 8Å is assigned to the grid point.

4.1.4.4. AUTODOCK

The docking simulation is carried out using the Metropolis method, also known as Monte Carlo simulated-annealing. The user can start the molecule in a particular conformation and location, or have each new docking run begin at a randomly chosen state. With the protein static throughout the simulation, the ligand molecule performs a random walk in the space around the protein. At each step in the simulation, a small random displacement is applied to each of the degrees of freedom of the ligand: translation of its centre of gravity, orientation and rotation around each of its flexible internal dihedral angles. This displacement results in a new configuration whose energy is evaluated by tri-linear interpolation of affinity values of the eight grid points surrounding each of the ligand atoms. If the new energy is lower than that of the preceding step, the new configuration is immediately accepted; if higher, the configuration is accepted or rejected based upon a probability expression dependent on a user-defined temperature, T. The probability of acceptance, P, is given by Equation (4.2)

$$P\ (\Delta E) = e\ \exp(-\Delta E/k_B T) \qquad \text{Equation (4.2)}$$

where ΔE is the difference in energy of the current step from the previous step and k_B is the Boltzmann constant. In each run of simulated annealing, a prescribed number of cycles is carried out, typically 50, each at a constant specified temperature, and each cycle contains a large number of individual steps, e.g. 25,000. After a specified number of acceptances or rejections, the next cycle begins with a temperature lowered by a specified schedule such as

$$T_i = gT_{(i-1)}$$

where T_i is the temperature at cycle i, and g is a constant between 0 and 1.

After all the requested runs (independent dockings) have been performed, cluster analysis is performed, based on structural root mean square (RMS) difference. The RMS deviation of any conformations generated during the docking is calculated by comparing the co-ordinates in a file specified by the user. This is useful when the experimentally determined complex structure is known. The user specifies an RMS tolerance and if two conformations have an RMS less than this, they are placed in the same cluster, being ranked by energy. Cluster analysis ranks the resulting families of docked conformations in order of increasing energy.

4.1.5. Summary

The ideal docking method would allow complete flexibility of both ligand and receptor and allow them to explore their conformational degrees of freedom such as *via* a molecular dynamics simulation of the ligand-receptor complex. The most recent docking algorithms have been developed to allow this [457-459], but such calculations are computationally very demanding and molecular dynamics only explores the range of binding modes well for very small, mobile ligands. For many systems, the energy barriers that separate one binding mode from another are often too large to be overcome. In addition, many proteins exhibit an induced fit mode of binding in reality, and a large free energy change may accompany conformational changes in the protein upon binding, in particular the surface loops.

A further problem is that, at the time of writing, no docking programs can simulate water in the active site and the displacement and rearrangement of solvent molecules as the ligand binds to the receptor. Neither can they predict protein rearrangement and simulate the high flexibility of amino acid side chains. This is due to the current limitations in computer speed (including data processing, storage and handling) which is in turn limited by present computer technology. With the rapid developments in docking programs and, more importantly, in computer technology, this is not foreseen to be a long-term problem.

It was proposed to use AUTODOCK to investigate the binding modes of families of inhibitors of the parasite enzyme trypanothione reductase, and to attempt to rationalise *in vitro* inhibition data with computer modelling predictions of relative ligand binding energies. More detailed coverage of this enzyme and its inhibitors are given in the next section.

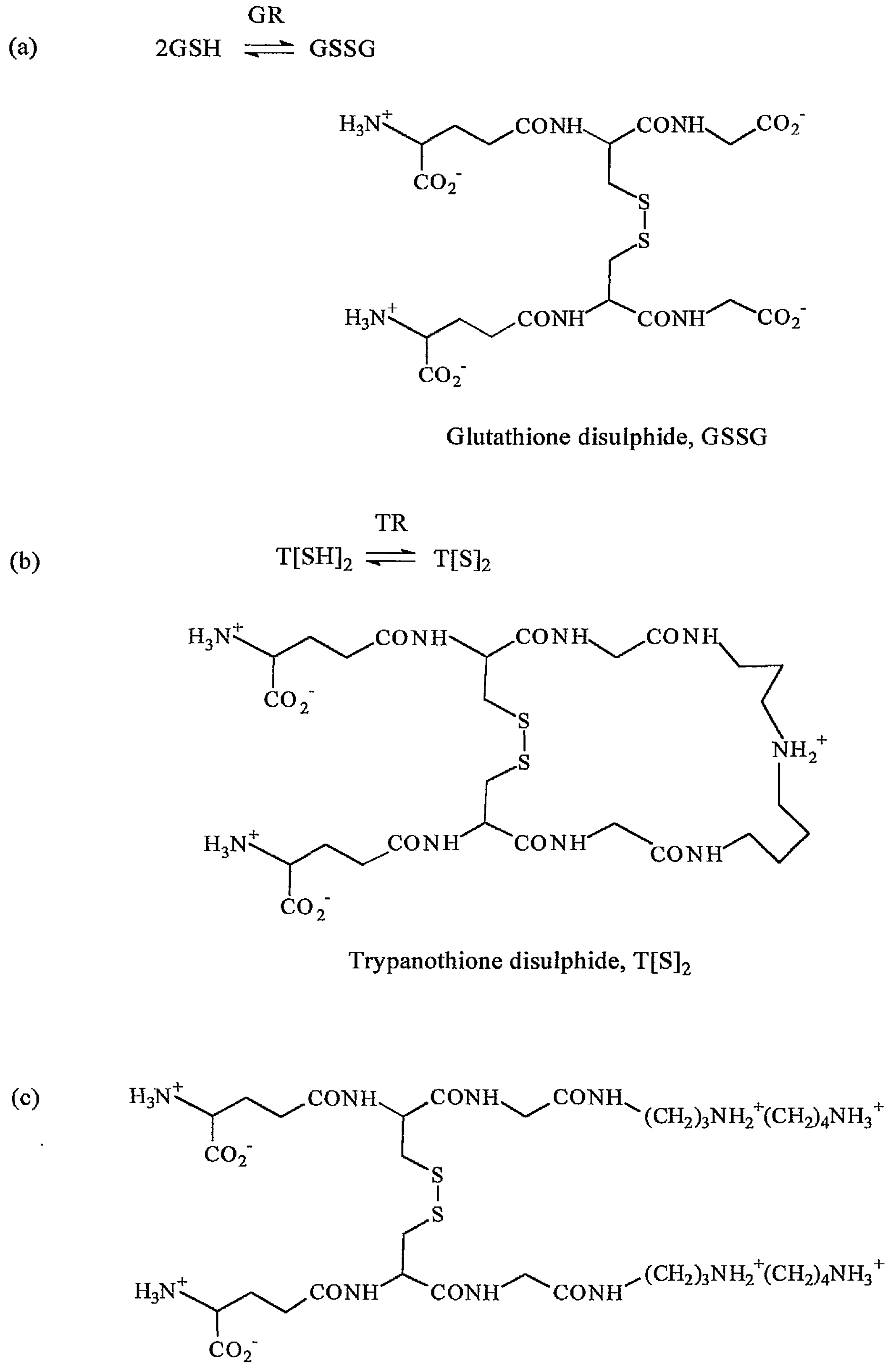
4.1.6. Trypanothione reductase

Major Third World parasitic diseases, including African Sleeping Sickness, Chagas' disease and leishmaniasis are caused by pathogenic parasites belonging to the order Kinetoplastida, (*Trypanosoma* and *Leishmania* species). Trypanothione reductase (TR), the enzyme, which in trypanosomal and leishmanial parasites catalyses the reduction of trypanothione disulphide to the redox-protective dithiol (Figure 4.2(b)), has been identified as a potential target for rational anti-parasite drug design (reviewed [460]). The analogous human enzyme is glutathione reductase (GR), which reduces glutathione disulphide to glutathione (Figure 4.2(a)). Despite gross similarities (e.g. 41% sequence identity between *Trypanosoma congelense* TR and human GR) TR and the host enzyme, human red blood cell GR, are mutually exclusive with respect to their disulphide substrates [461,462]. For instance, human GR has a 9000-fold preference for glutathione over trypanothione based on V_{max}/K_m values [463].

4.1.6.1. Structural features of trypanothione reductase

A member of the large, well-characterised protein family of FAD-dependent NAD(P)H oxidoreductases overviewed in [464-467], TR is dimeric with molecular mass 104kDa, whose subunits fold into four domains: the FAD-binding, the NADPH-binding, the central and the interface domains. Two identical active sites are formed by residues of the FAD, NADPH, and central domain of one monomer and the interface domain of the other monomer. A redox-active disulphide is involved in the formation of an enzyme-substrate mixed disulphide. The active site is rather large (approximately 20Å long, 15Å wide and 15Å deep), see Figure 4.3(a) and (b).

Crystal structures of TRs from *Crithidia fasciculata* [468,469] and *Trypanosoma cruzi* [470] have been solved. However, until now attempts to bind trypanothione to crystals of TR have been unsuccessful. Only the crystal structures of TR complexed with the alternative substrate N^1-glutathionylspermidine disulphide ($[GspdS]_2$) [471] (Figure 4.2(c)) and with the weak, but selective, inhibitor mepacrine [472] have been reported. From this crystallographic study of the TR-mepacrine complex, the hydrophobic acridine ring of mepacrine was found in the active site, close to the hydrophobic wall formed by the residues W-21 and M-113 (with L-17, I-106, Y-110 and F-114 nearby). The alkylamino chain points into the inner region of the active site, held in position by a solvent-mediated hydrogen bond to E-18 (Figure 4.4).



(a) $2GSH \overset{GR}{\rightleftharpoons} GSSG$

Glutathione disulphide, GSSG

(b) $T[SH]_2 \overset{TR}{\rightleftharpoons} T[S]_2$

Trypanothione disulphide, $T[S]_2$

(c)

Figure 4.2. (a) Glutathione substrate for glutathione reductase.

(b) Trypanothione substrate for trypanothione reductase.

(c) The alternative TR substrate N^1-glutathionylspermidine disulphide, $[GspdS]_2$.

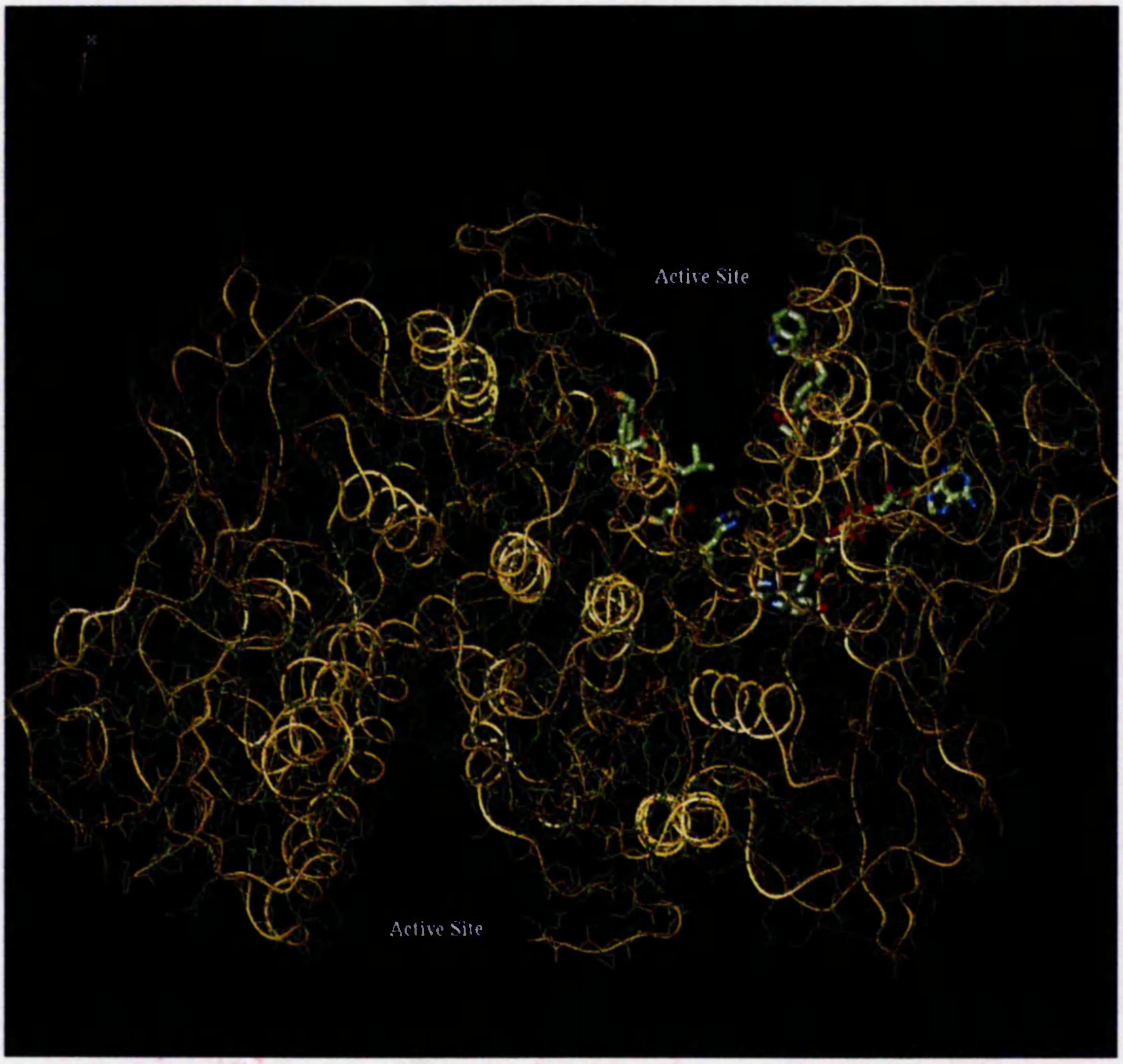


Figure 4.3(a). Structure of the enzyme Trypanothione Reductase (TR). The two large active sites of the dimer are labelled. Residues important in inhibitor design are shown in one active site as a liquorice representation and can be seen in greater detail in Figure 4.3(b).

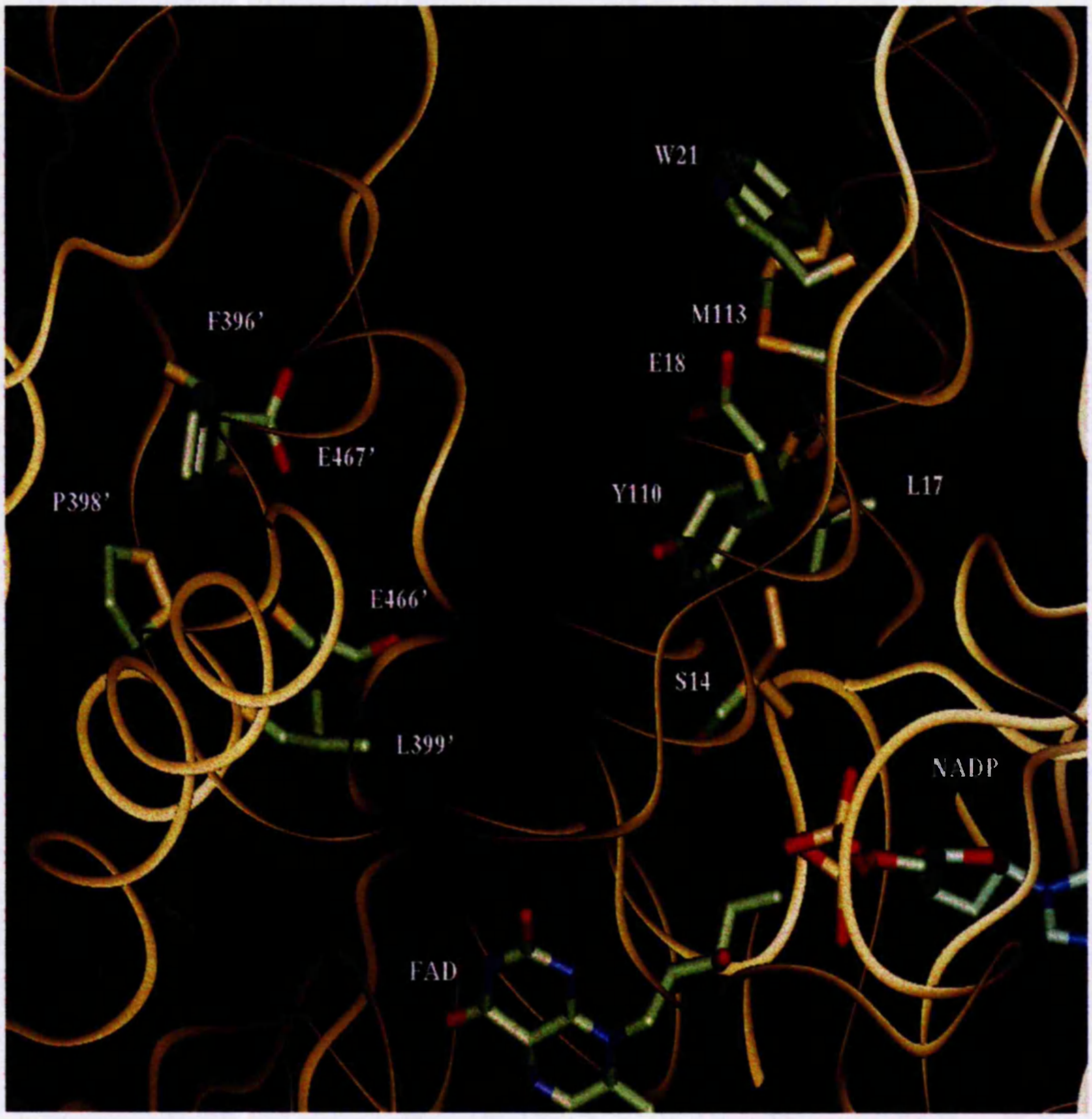


Figure 4.3(b). The active site of trypanothione reductase. The amino acids important in inhibitor design (see text) are shown as a liquorice representation, as are NADP and FAD. The Cα trace of the protein backbone is shown as a yellow ribbon.

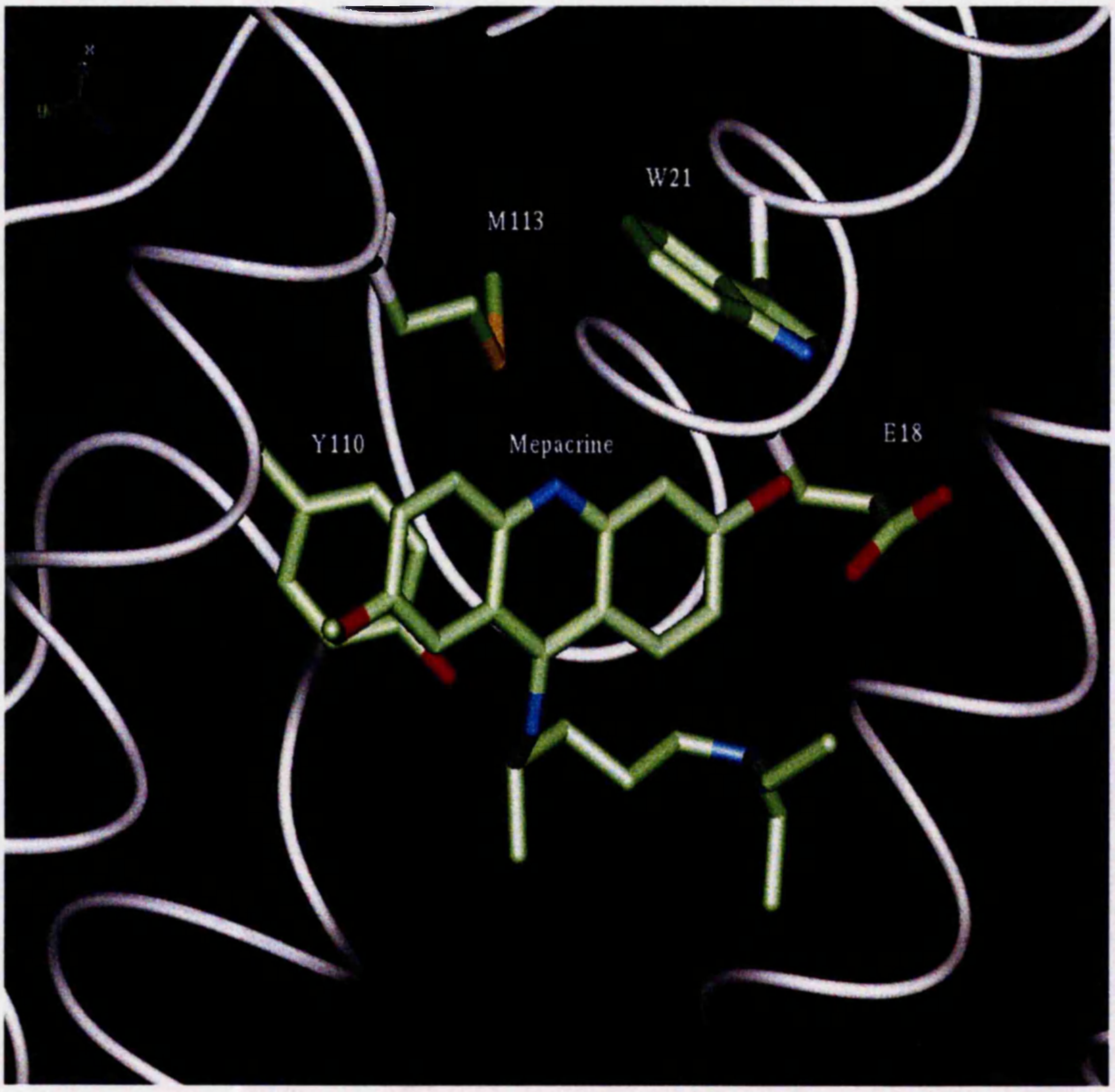


Figure 4.4. Mepacrine bound in the active site of trypanothione reductase. The important amino acid side chains are shown. The coordinates for the structure are from Jacoby *et al.* (see text).

Using this information, the rational drug design approach has been applied in our laboratory using molecular modelling, kinetic studies against the TR of *T.cruzi* and studies of *in vitro* parasitic activity against *T.brucei, T.cruzi* and *Leishmania donovani.* The initial design approach was to target the hydrophobic wall formed by W-21/M-113 *etc.* Initially this pocket was filled manually with various fused aromatic structures, e.g. naphthalene, phenanthrene, anthracene (Figure 4.5(a)). For solubility purposes and to vectorise the ligand in its putative binding site, a cationic ammonium group was introduced into the prototype inhibitors. This was located between two glutamic acid side-chains (Figure 4.5(b)). This led to the finding that phenothiazines and related tricyclic anti-depressants, e.g. saturated dibenzazepine (imipramines), (Figure 4.6) were potential lead inhibitors of TR showing strong inhibition [473-475] (Figure 4.5(c)).

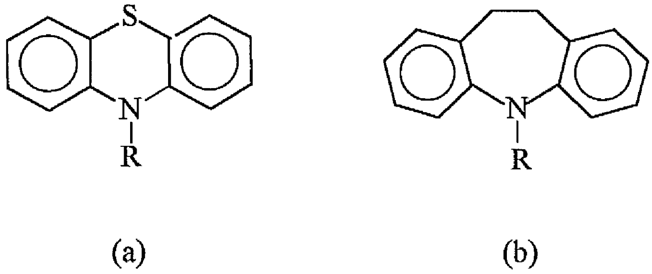


Figure 4.6. Root structures of phenothiazines (a) and imipramines (b).

The tricyclic framework has been considered to be bound in the major hydrophobic cleft of the active site used by the spermidine portion of the natural substrate, trypanothione, *i.e.* near W-21 and M-113. From analysis of a series of alternative substrates and molecular modelling studies, a second hydrophobic pocket called the Z-site was observed, located as the region near F-396' and P-398'. Using the experimental coordinate sets [472] for the bound-mepacrine site, inhibitors have been developed in our laboratory by trying to combine this putative Z-site with the hydrophobic wall containing W-21, M-113, Y-110 by further functionalising tricyclic inhibitors. Molecular modelling suggested that quaternisation of the tertiary ω-nitrogen of the side-chain on the central bridge nitrogen atom of chlorpromazine to incorporate a hydrophobic species should increase binding strength through increased hydrophobic interactions if the Z-site could be accessed (Figure 4.7).

Consequently, chlorpromazine derivatives quaternised on this nitrogen and specifically bearing substituted benzyl and other aromatic groups were designed and synthesised in our laboratory. K_i and I_{50} data for *T.cruzi* TR provided clear evidence that this additional hydrophobic group and permanent positive charge improve inhibition strength relative to the original tricyclic leads by up to 30-fold. The best of these inhibitors had a K_i value of 0.12μM, compared to 6μM for chlorpromazine [390].

4.1.7. Aims

It was proposed to attempt to rationalise the above and related sets of quantitative experimental data for three families of quaternary tricyclics with the use of a docking program and molecular models (phenothiazines, open ring phenothiazine analogues and imipramines). We aimed to use the program AUTODOCK to investigate the binding modes of the inhibitors and to address the following questions:

1. Does the calculated order of ligand binding energies correlate with observed inhibition data (experimental K_i and I_{50} values)?
2. Do the final ligand conformations show discrete families of binding locations (*i.e.* cluster into favoured areas of the active site)?
3. Do the docked conformations and sites bear any resemblance to the observed crystal structure situation for mepacrine?
4. Can AUTODOCK be used as a predictive tool to suggest structures for improved inhibitors?

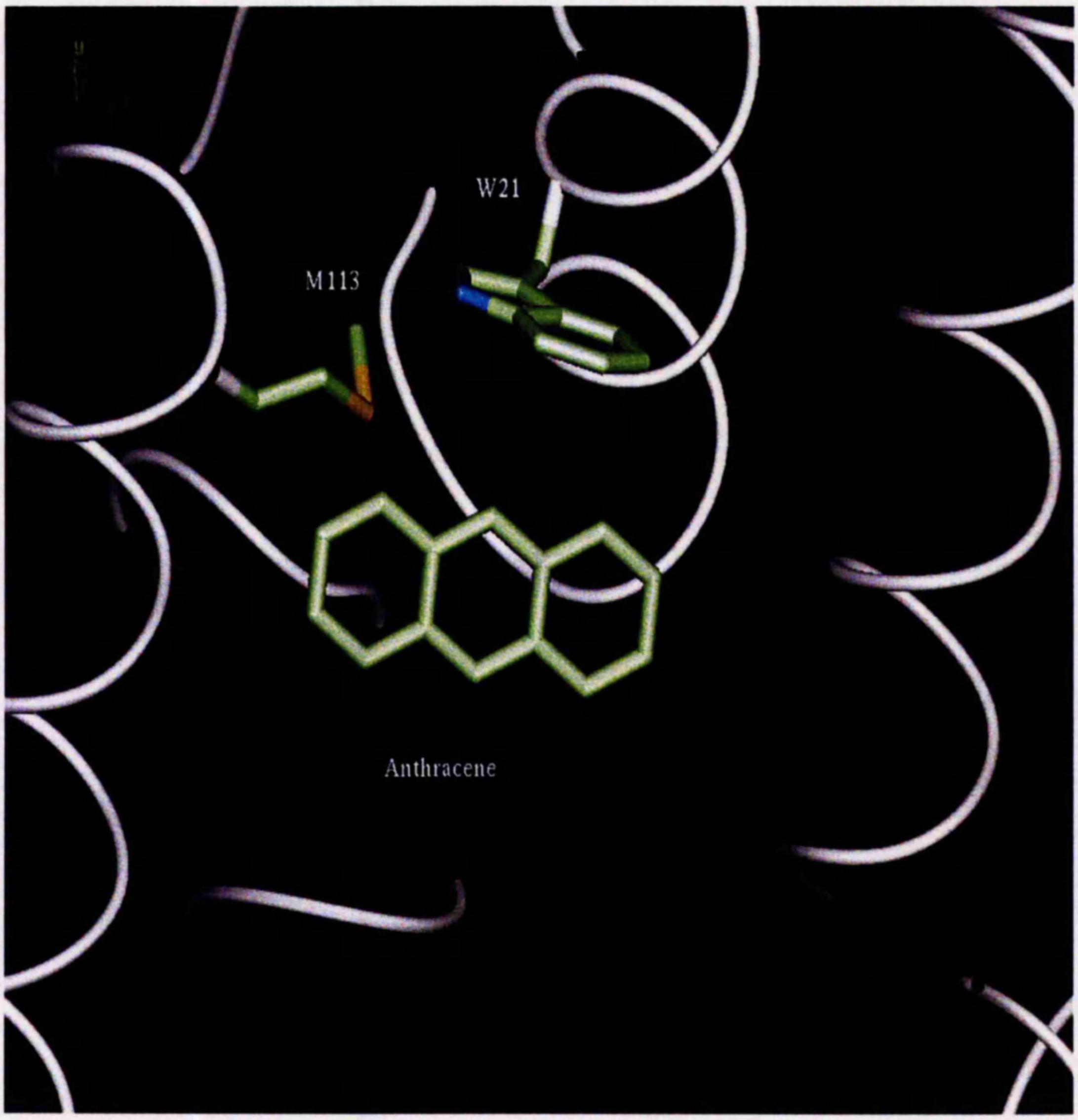


Figure 4.5(a). Anthracene manually docked into the active site of trypanothione reductase. The residues potentially involved in hydrophobic interactions and used to guide initial ligand placement are shown.

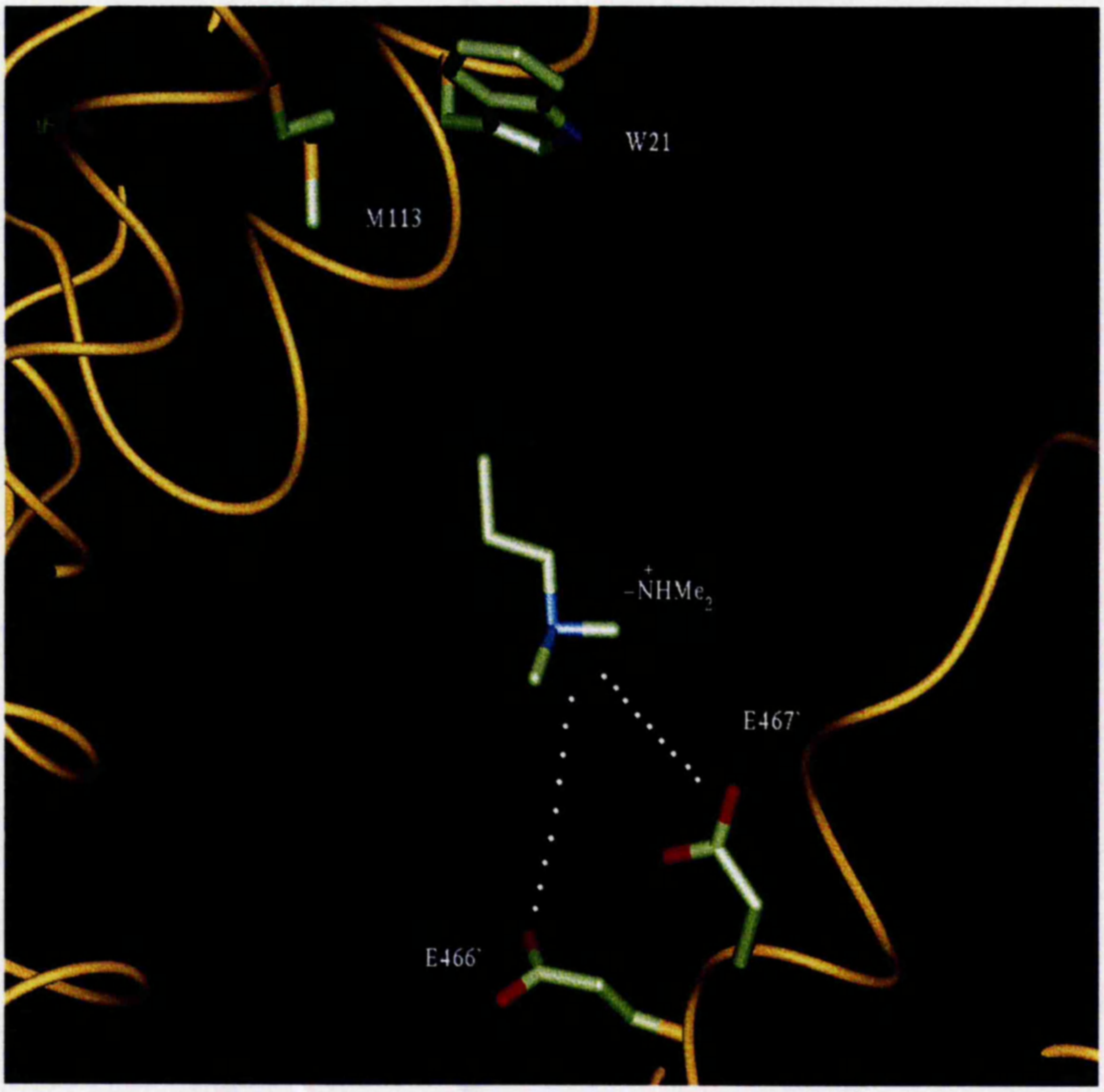


Figure 4.5(b). Representation of the cationic ammonium ion substituent of potential inhibitors manually docked into the active site of trypanothione reductase, showing putative electrostatic interactions with residues E466' and E467'.

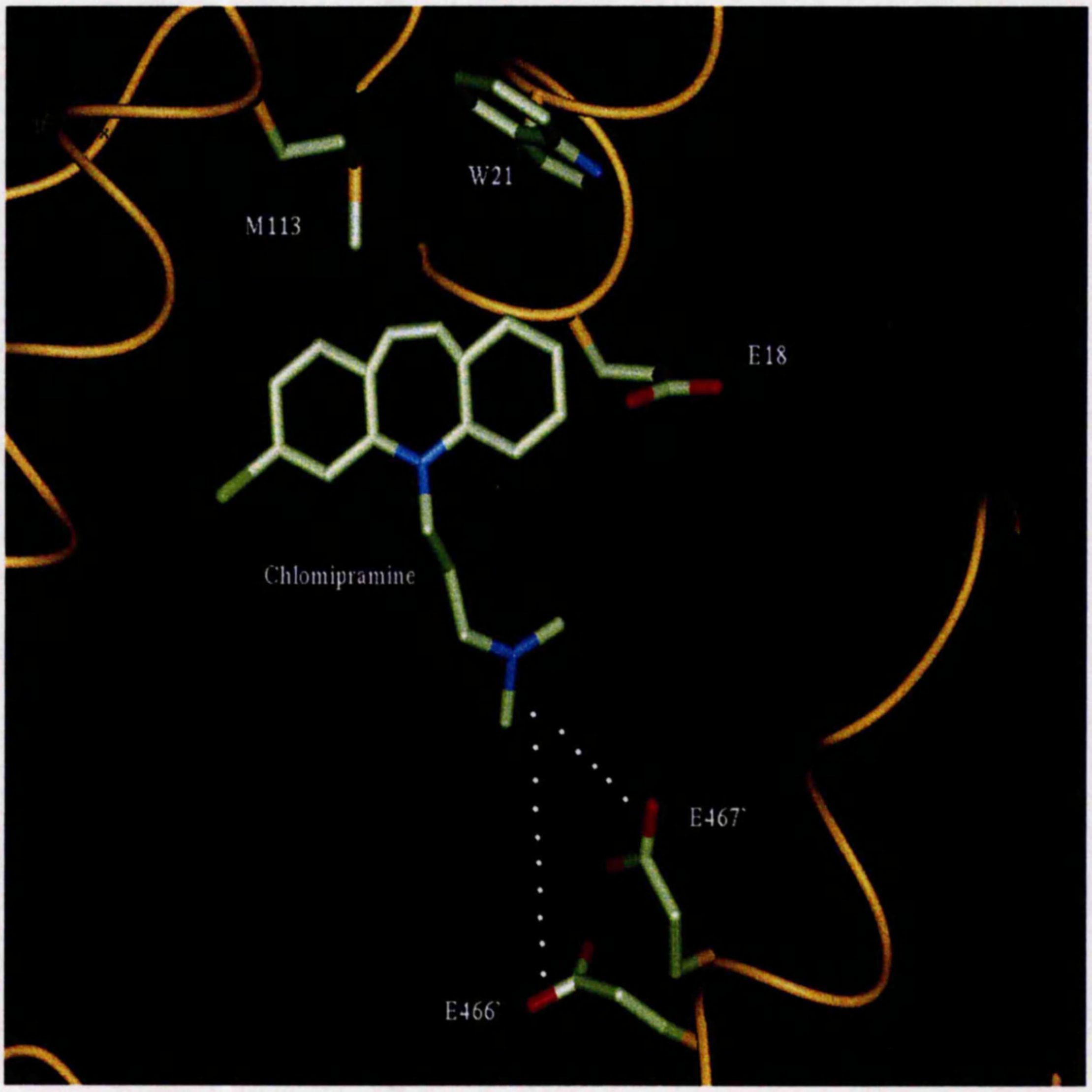


Figure 4.5(c). Chlomipramine manually docked into the active site of trypanothione reductase, showing putative hydrophobic and electrostatic interactions with residues used in the initial design steps.

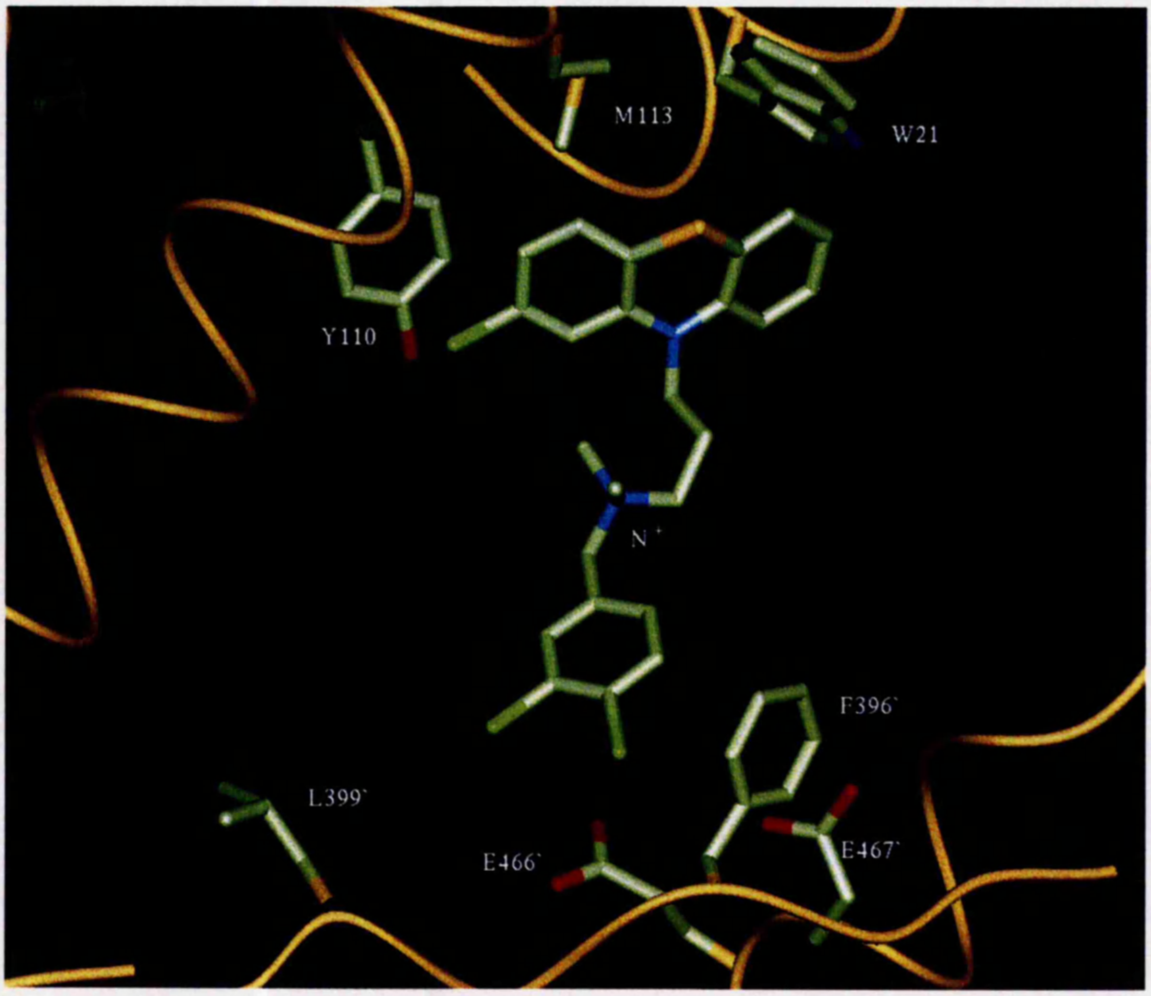


Figure 4.7. Inhibitor of trypanothione reductase based on the lead tricyclic framework, but also possessing a quaternary amine site added to increase binding strength through electrostatic interactions with residues E466' and E467'.

4.2. MATERIALS AND METHODS

Docking searches used the program AUTODOCK 2.4 [427], developed to provide a procedure for predicting the interaction of small molecules with macromolecular targets, and using a Metropolis Monte Carlo simulated annealing technique for positional and configurational exploration [433] based on the AMBER force field [181]. This approach includes terms for the Van der Waals', electrostatic and hydrogen-bonding interactions.

4.2.1. Docking procedure

4.2.1.1. Preparation of ligands and receptor for docking runs

The X-ray structure of TR complexed with mepacrine [472] was used. Mepacrine and all water molecules were removed, the charged amine and carboxyl terminii capped and essential hydrogens added using the Kollman united-atom library. The X-ray structure was briefly relaxed (25 steps Simplex) to remove any bad geometry and short contacts present, but not to allow any heavy atoms to move significantly from the calculated crystal structure positions.

Ligands to be docked (Tables 4.2-4.5) were built in SYBYL 6.4.2 (Tripos Inc.), all hydrogens added and a formal charge of +1 added to the quaternary nitrogen atom. Partial atom-centred charges were calculated using the semi-empirical method PM3 implemented in MOPAC [119] and, after an initial Simplex optimisation, ligands were minimised to convergence by the Powell method with a gradient of 0.05 kcal/mol.

To check the validity of the minimisation technique, the Cambridge Crystallographic Database was searched for phenothiazine derivatives, yielding five molecules (Table 4.6) which were built in SYBYL and minimised as described. The RMS differences between the built molecules and the crystal structures were determined.

4.2.1.2. AUTOTORS

The rotatable bonds of the ligands, defined in Tables 4.2-4.5,vary in number from 4 to 12 for the most flexible ligand. Non-polar hydrogens were united automatically using the '-h' flag.

4.2.1.3. AUTOGRID

A grid (29.625Å x 29.625Å x 30.375Å) was used for each test case, with a grid spacing of 0.375Å, centred on the active site of the protein. The number of grid points in each Cartesian direction was 40, 40 and 41, making a total of 80 x 80 x 82 points per grid. The appropriate 12-10 parameters for hydrogen atoms bonded to O, N, or S in the ligand and to O or S in the protein were specified in the AUTOGRID parameter file (e.g. Figure 4.1). (N atoms of TR were all assumed to be part of protein backbone amide bonds (or in glutamine, apsaragine, arginine, *etc*) so that the lone pair would not available to form hydrogen bonds).

4.2.1.4. AUTODOCK

As the 50 docking runs initially performed per ligand gave practically no clustering, 100 docking runs were carried out in each case. At the beginning of each simulated annealing run, the ligand was positioned randomly within the grid and movement began with a random translation, random rigid body rotation and random torsion. Each run consisted of 100 annealing cycles. A cycle terminated if the ligand made either 30,000 accepted moves, or 30,000 rejected moves. The annealing temperature, was 310K during the first cycle, and was reduced linearly at the end of each cycle by a factor of 0.95. The state with the minimum energy found during the current cycle was used to start the next cycle. For all cycles, there was a maximum translational step size of 0.2Å and a maximum torsional rotation of 5° per step.

Docking runs, performed on a Silicon Graphics Origin2000, each took between 12 and 76hours of CPU time.

4.2.1.5. Cluster Analysis

Following docking, cluster analysis of all structures generated for a single compound was performed. Once sorted by total energy of interaction, the structures were sequentially assigned to cluster families, represented by top-of-cluster structures. Structures not satisfying a tolerance of 1.5Å in the RMS deviation for all atoms to each top of cluster structure found, defined new clusters.

The final docked positions of all ligands (except mepacrine) could not be referenced against any known TR-bound three-dimensional structure but those of mepacrine were referenced to the bound conformation seen in the crystal structure.

To be able to compare final energies for different ligands, the docked ligand internal energies were referenced to the initial intermolecular undocked energy. The differences calculated between the undocked and docked energies, *i.e.* the energy gained on binding, were used for analyses.

4.3 RESULTS AND DISCUSSION

4.3.1. Validation of the minimisation technique

The five crystallographically derived structures of compounds structurally related to the docking ligands are in Table 4.6. After being built in SYBYL, charges loaded and minimised, the resulting theoretical conformations were compared with the crystallographic data. The RMS differences between all atoms only varied between 0.68 and 1.44Å, indicating that the method of ligand construction and minimisation is satisfactory.

4.3.2. Docking studies

The results of the docking studies are shown in Tables 4.2-4.5. Overall, the 35 ligands showed very little clustering into groups of docked conformations (except for chlorpromazine and chlomipramine); the energies of the final structures spanned a continuous range and were not well separated energetically. With between 28 and 89 'clusters' found out of 100 runs, there does not appear to be one distinguishable family of energetically favoured docked conformations. A 'cluster' is defined as a set of final docked conformations separated by an RMS difference of ≤1.5Å, but in many of the present cases the term actually describes only one or two conformational states and so may be mis-leading. In the following analyses, only the top 20 clusters were examined. It should be noted that although a 'mean' energy of the top clusters has been calculated, AUTODOCK does not list from which docking runs this average is from, therefore a standard deviation cannot be quoted.

4.3.2.1. General overview of docking results

A brief analysis of these docking studies shows that although the clustering was poor, two general families of preferred binding modes can be clearly seen and these are comparable in binding energy:

I The quaternary nitrogen of the ligand (N^+) electrostatically interacts with the residue E-466' (or sometimes E-467') combined with one of the two following hydrophobic interactions:

a) The tricyclic moiety lying in the hydrophobic Z-site (F-396'/P-398'/L-399'),

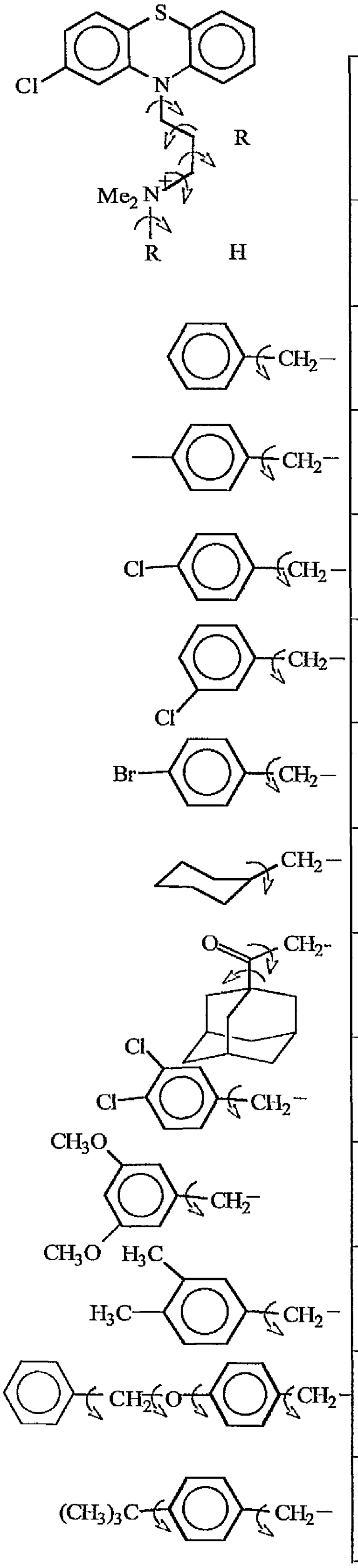
b) The hydrophobic aromatic group on the N^+ lying in the Z-site,

c) The quaternary nitrogen only interacts with E-466' or E-467' with no hydrophobic interactions,

II The quaternary nitrogen electrostatically interacts with S-14, (usually showing no hydrophobic interactions, but sometimes with the tricyclic or quaternary aromatic group lying near the hydrophobic region of L-17/W-21/M-113).

Modes I (a), I(b) and II are shown in Figure 4.8(a), (b) and (c), respectively. (For the open ring ligands, sometimes the secondary amine nitrogen participates in the same interactions with E466'/E467' seen of the N^+).

A more detailed analysis of the binding of individual ligands follows.

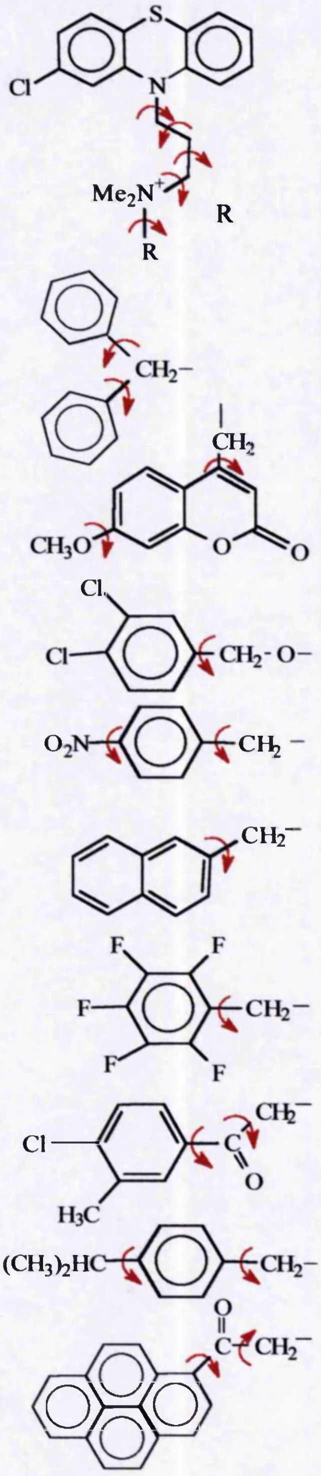


Ligand	Mean energy of top cluster (kcalmol^{-1})	% runs in top 10 clusters	% runs in top 20 clusters	K$_i$ [a] (μM)
Chlorpromazine	-66.32	61	76	10.8 [b] ± 1.1
CC289	-64.95	35	55	1.3 ± 0.2
OFK001	-64.14	40	53	1.8 ± 0.2
OFK002	-70.00	37	52	1.7 ± 0.3
OFK003	-65.48	29	47	2.4 ± 0.1
OFK004	-69.58	32	48	1.5 ± 0.1
OFK005	-67.42	35	62	2.3 ± 0.3
OFK006	-70.96	45	62	1.2 ± 0.2
OFK007	-75.45	39	53	0.12 ± 0.01
OFK008	-80.30	30	43	0.77 ± 0.12
OFK009	-66.04	27	46	0.47 ± 0.10
OFK010	-74.40	22	35	0.47 ± 0.09
OFK011	-67.36	31	62	0.68 ± 0.08

Table 4.2. Results of docking the phenothiazine family of ligands.

Rotatable bonds are defined in red.

[a]- K$_i$ values from literature [476]. [b] - Value from literature [474].



Ligand	Mean energy of top cluster (kcalmol^{-1})	% runs in top 10 clusters	% runs in top 20 clusters	K_i [(a)] (µM)
OFK012	-72.18	44	64	0.71 ± 0.12
OFK013	-69.14	19	32	0.56 ± 0.07
OFK014 [(b)]	-47.13	27	49	[(c)]
OFK015	-78.24	30	47	2.98 [(d)] ± 0.38
OFK016	-71.97	28	52	1.55 [(d)] ± 0.24
OFK017	-80.84	42	59	1.67 [(d)] ± 0.37
OFK018	-73.04	27	49	0.49 [(d)] ± 0.11
OFK024	-67.51	33	54	-
OFK028	-82.70	45	56	-

Table 4.2. Results of docking the phenothiazine family of ligands.

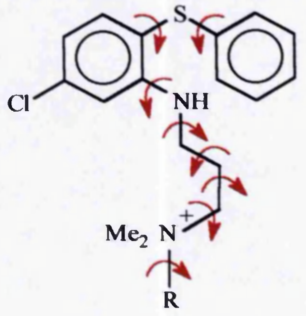
Rotatable bonds are defined in red.

[(a)] - K_i values from literature [476].

[(b)] - *N.B.* Positive quaternary nitrogen replaced by oxygen.

[(c)] - Little or no activity at 50µM. Precipitates at 100µM.

[(d)] - K_i is calculated from I_{50} value [476] using the equation $I_{50} = K_i (1 + [S_0]/K_m)$ where $[S_0]$ = 120µM and K_m = 24µM against the substrate *bis*-*N*-benzoyloxycarbonyl-L-cysteinylglycine-3-dimethylaminopropylamide disulphide.



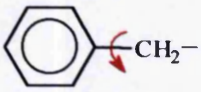
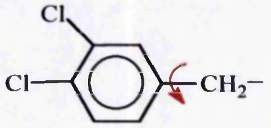
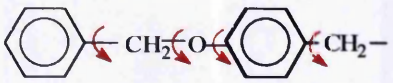
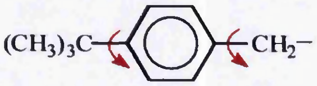
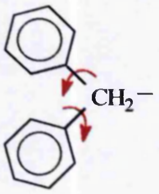
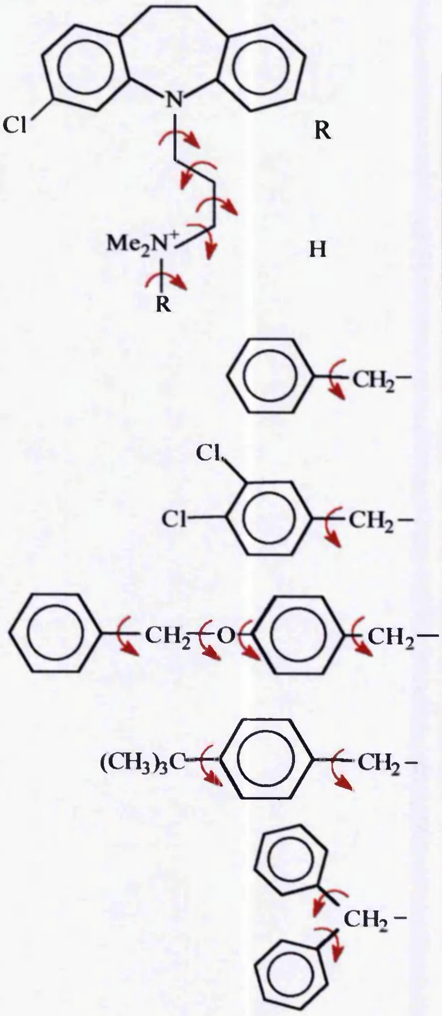
R	Ligand	Mean energy of top cluster (kcalmol^{-1})	% runs in top 10 clusters	% runs in top 20 clusters	K$_i$ [(a)] (μM)
H	OPN-00	-66.04	27	48	66 [(b)]
	OFK289b	-91.37	27	43	14.2 ± 0.1
	OFK007b	-70.04	21	37	1.69 ± 0.22
	OFK010b	-73.05	16	29	6.6 ± 0.4
	OFK011b	-70.47	17	37	6.5 ± 0.7
	OFK012b	-69.92	20	36	5.3 ± 0.9

Table 4.3. Results of docking the open ring family of ligands. Rotatable bonds are defined in red.

[(a)] - OFK ligand K$_i$ values from literature [476].

[(b)] - K$_i$ value from literature [477].

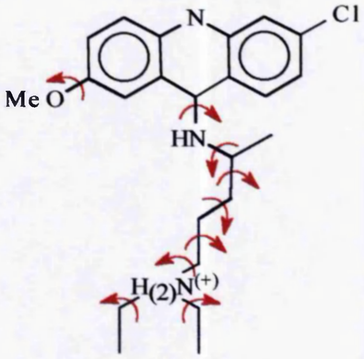


Ligand	Mean energy of top cluster (kcalmol^{-1})	% runs in top 10 clusters	% runs in top 20 clusters	K$_i$ [(a)] (μM)
Chlomipramine	-71.05	63	91	6.5 [(b)]
OFK289c	-65.06	44	72	1.9 ± 0.4
OFK007c	-66.19	27	52	0.20 ± 0.04
OFK010c	-73.95	34	53	0.66 ± .07
OFK011c	-66.60	43	66	0.39 ± 0.08
OFK012c	-77.45	44	70	0.63 ± 0.11

Table 4.4. Results of docking the imipramine family of ligands.

Rotatable bonds are defined in red.

[(a)] - K$_i$ values from liteature [476]. [(b)] - Value from literature [475].



Ligand	Mean energy of top cluster (kcalmol^{-1})	% runs in top 10 clusters	% runs in top 20 clusters	K$_i$ [(a)] (μM)
Mepacrine (charged)	-67.75	24	39	25
Mepacrine (uncharged)	-50.55	28	41	-

Table 4.5. Results of docking mepacrine (charged and uncharged).

Rotatable bonds are defined in red.

[(a)] - K$_i$ value from literature [472].

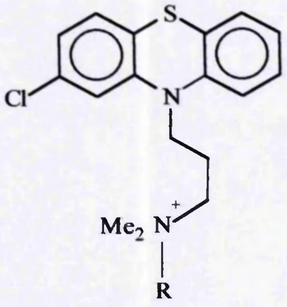
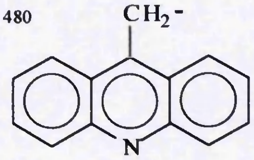
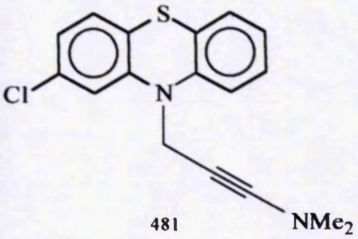
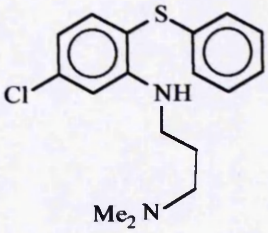
Compound		RMS difference (Å)
	R = H [478, 479]	1.20, 1.08
	R = [480] 	1.44
 481		1.06
 482		0.68

Table 4.6. The RMS differences between four crystal structures found in the Cambridge Crystallographic Database structurally related to the docked ligands and their theoretical structures calculated in SYBYL.

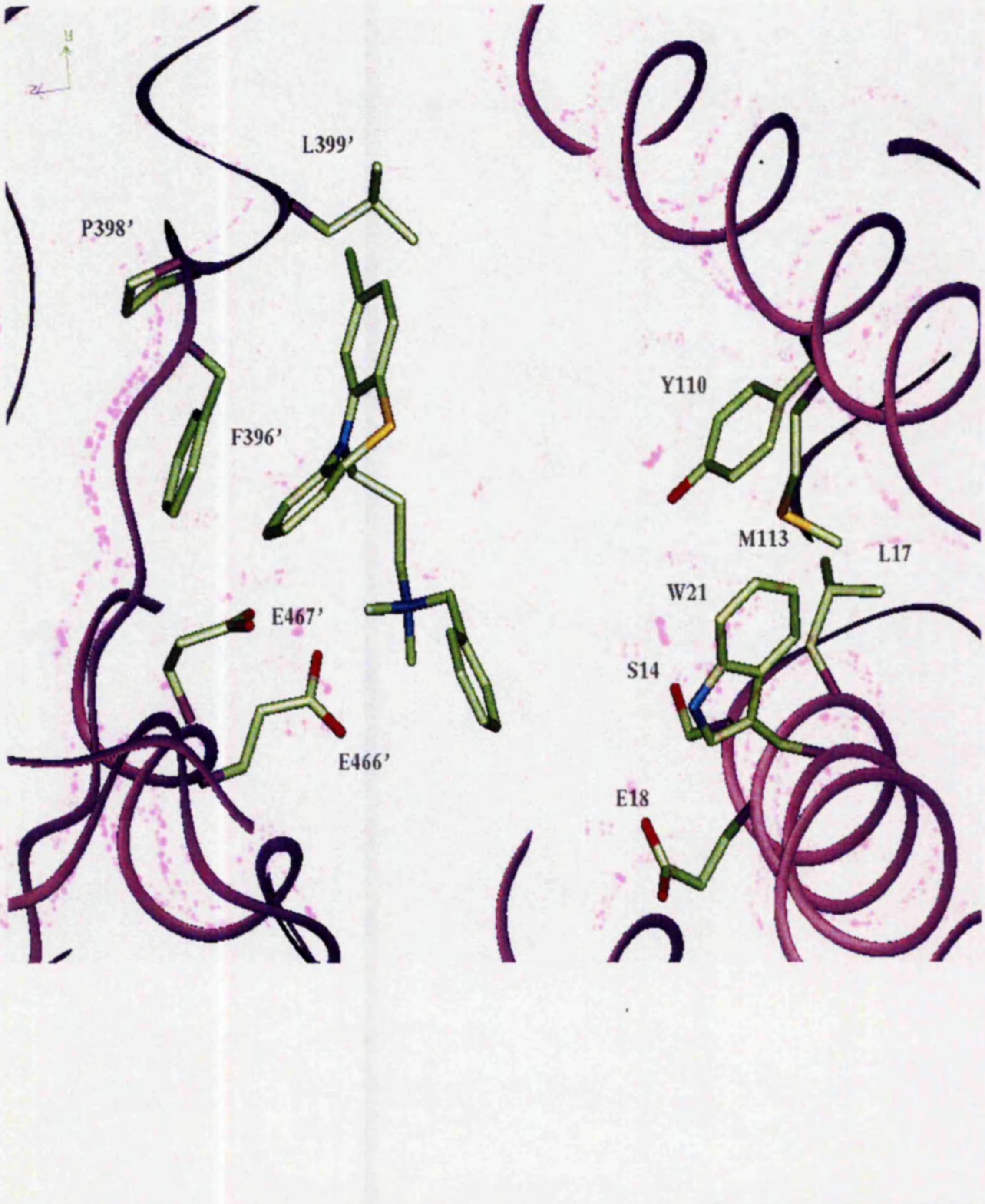


Figure 4.8(a). Binding mode I(a). The quaternary nitrogen of the ligand, here CC289, electrostatically interacts with E466' and/or E467' and the tricyclic moiety lies in the Z-site.

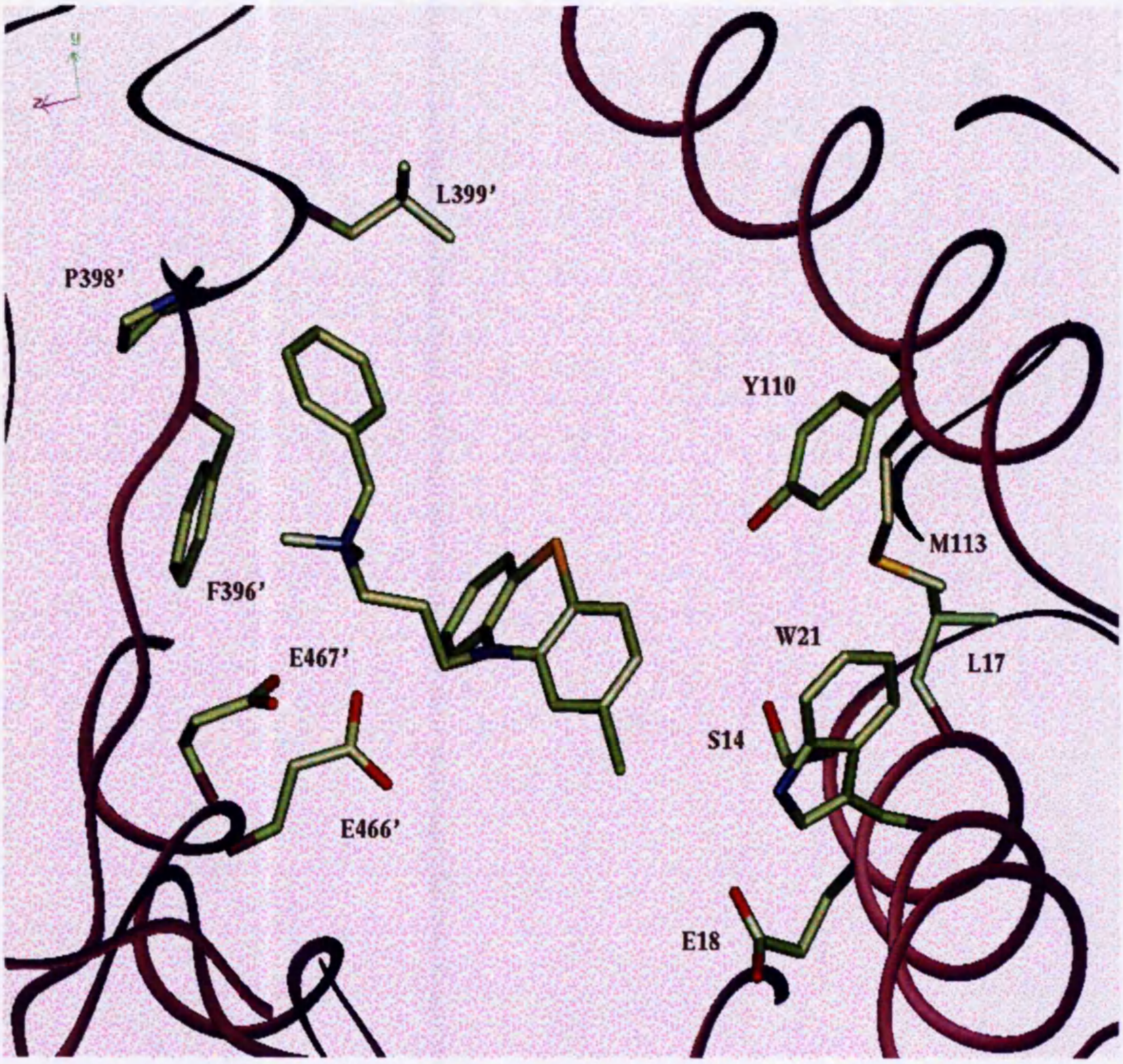


Figure 4.8(b). Binding mode I(b). The quaternary nitrogen of the ligand, here CC289, electrostatically interacts with E466' and/or E467' and the hydrophobic aromatic group on the quaternary nitrogen lies in the Z-site.

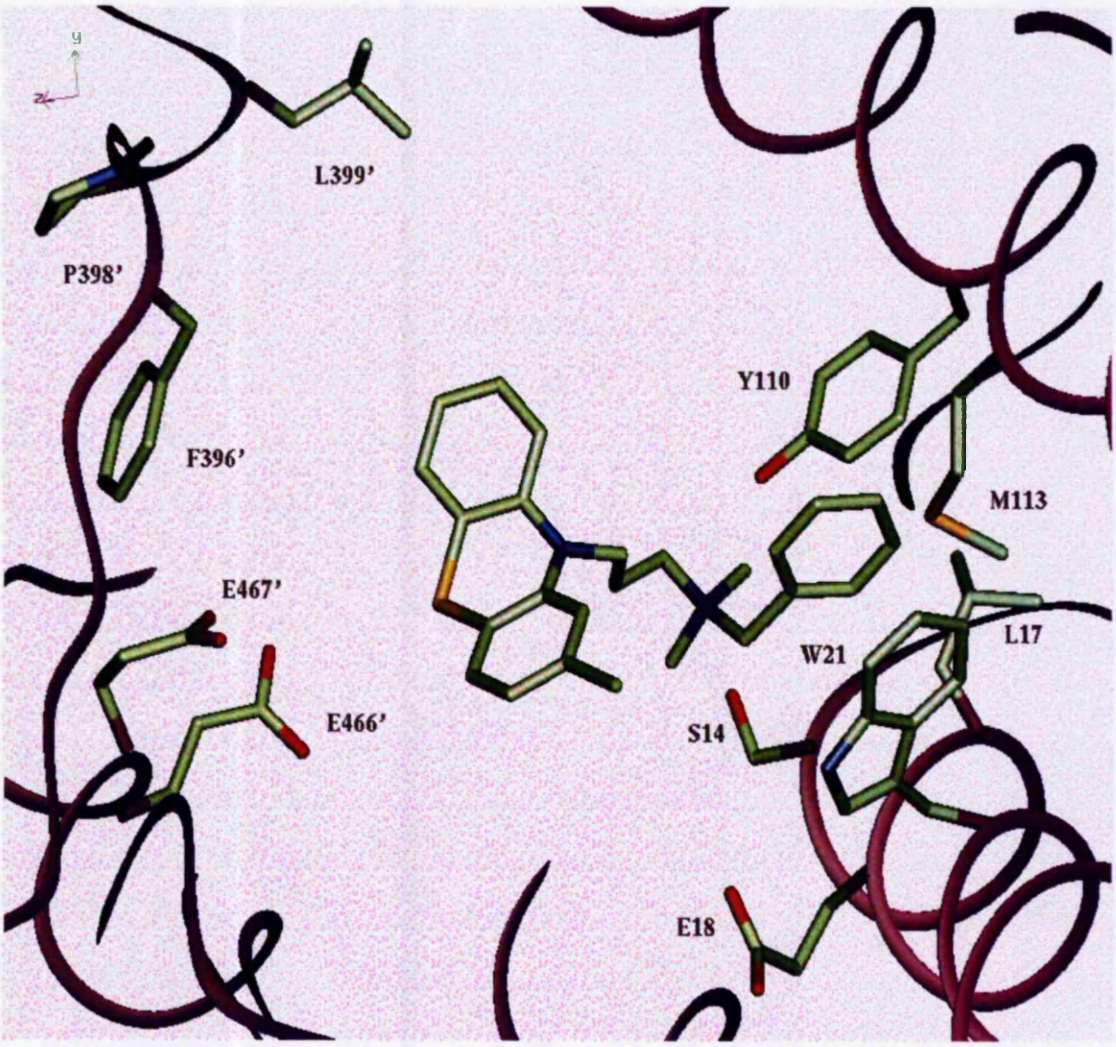


Figure 4.8(c). Binding mode II. The quaternary nitrogen of the ligand, here CC289, electrostatically interacts with S14 (occasionally with a hydrophobic moiety lying near L17/W21/M113).

4.3.2.2. Phenothiazines

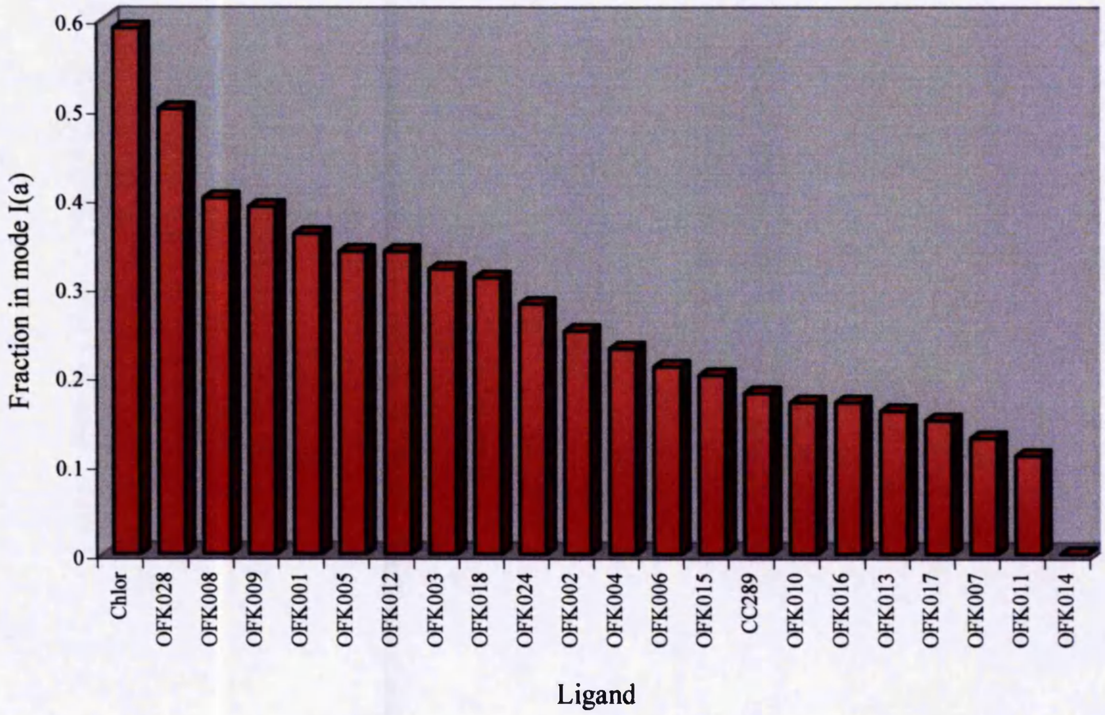
A group of 23 phenothiazine ligands was studied (22 quaternary compounds and chlorpromazine). Table 4.7 and Figure 4.9 show the fraction of ligands found in each binding mode.

Ligand	Fraction in Mode I (a)	Average Energy (kcalmol^{-1})	Fraction in Mode I(b)	Average Energy (kcalmol^{-1})	Fraction in Mode I(c)	Average Energy (kcalmol^{-1})	Fraction in Mode II	Average Energy (kcalmol^{-1})
Chlorpromazine	0.59	-63.72	-	-	0.12	-62.97	0.17	-63.68
CC289	0.18	-61.96	0.02	-67.16	0.13	-64.46	0.49	-63.67
OFK001	0.36	-64.37	0.06	-64.32	0.11	-63.78	0.38	-65.24
OFK002	0.25	-64.70	0.10	-64.59	-	-	0.54	-66.66
OFK003	0.32	-63.76	-	-	0.23	-61.92	0.32	-63.97
OFK004	0.23	-66.02	-	-	0.10	-68.53	0.58	-67.93
OFK005	0.34	-65.76	-	-	0.08	-64.76	0.15	-64.88
OFK006	0.21	-71.82	-	-	0.63	-71.22	0.03	-72.14
OFK007	0.13	-74.02	-	-	0.11	-73.91	0.68	-74.81
OFK008	0.40	-72.17	-	-	0.33	-70.12	0.02	-71.46
OFK009	0.39	-64.40	-	-	-	-	0.39	-64.68
OFK010	0.17	-72.93	0.20	-72.08	0.20	-71.93	0.11	-72.12
OFK011	0.11	-66.23	-	-	0.45	-65.19	0.39	-66.48
OFK012	0.34	-71.48	0.02	-69.75	0.09	-69.94	-	-
OFK013	0.16	-67.32	-	-	0.03	-69.24	0.03	-68.07
OFK014	-	-	-	-	-	-	-	-
OFK015	0.20	-75.95	0.08	-76.40	-	-	0.53	-76.45
OFK016	0.17	-70.40	0.19	-70.91	0.19	-70.36	0.17	-71.25
OFK017	0.15	-78.22	-	-	0.17	-79.50	0.20	-78.39
OFK018	0.31	-72.75	-	-	0.41	-72.05	0.08	-72.49
OFK024	0.28	-65.42	0.04	-66.71	0.26	-65.67	0.30	-67.18
OFK028	0.50	-78.99	0.43	-79.22	0.02	-79.93	0.02	-79.57

Table 4.7. The fraction of phenothiazine ligands in the top 20 clusters found in each binding mode.

After studying the top 20 clusters for each ligand, it was observed that for some ligands, approximately half of the final docked structures were in either conformation I(a) or II described above, e.g. CC289, OFK001, OFK002, OFK009, OFK011 and OFK017. Ligand OFK003 had approximately half its lower energy clusters in mode I(a) and one third in II. Ligand OFK004 had one third in mode I(a) and one half in mode II.

As these two modes are practically equal in energy, neither is favoured over the other. Both allow a stabilising electrostatic interaction of the quaternary nitrogen with either a charged acid group (E-466'/467') or a polar hydroxyl group (S-14) and both allow the hydrophobic



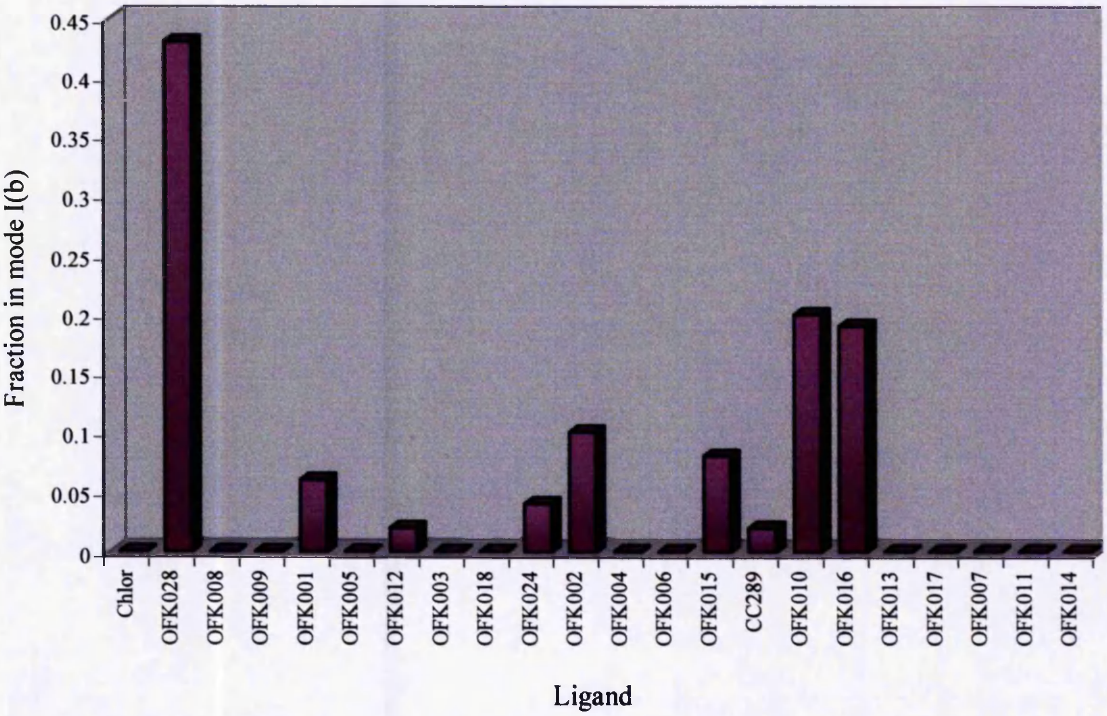
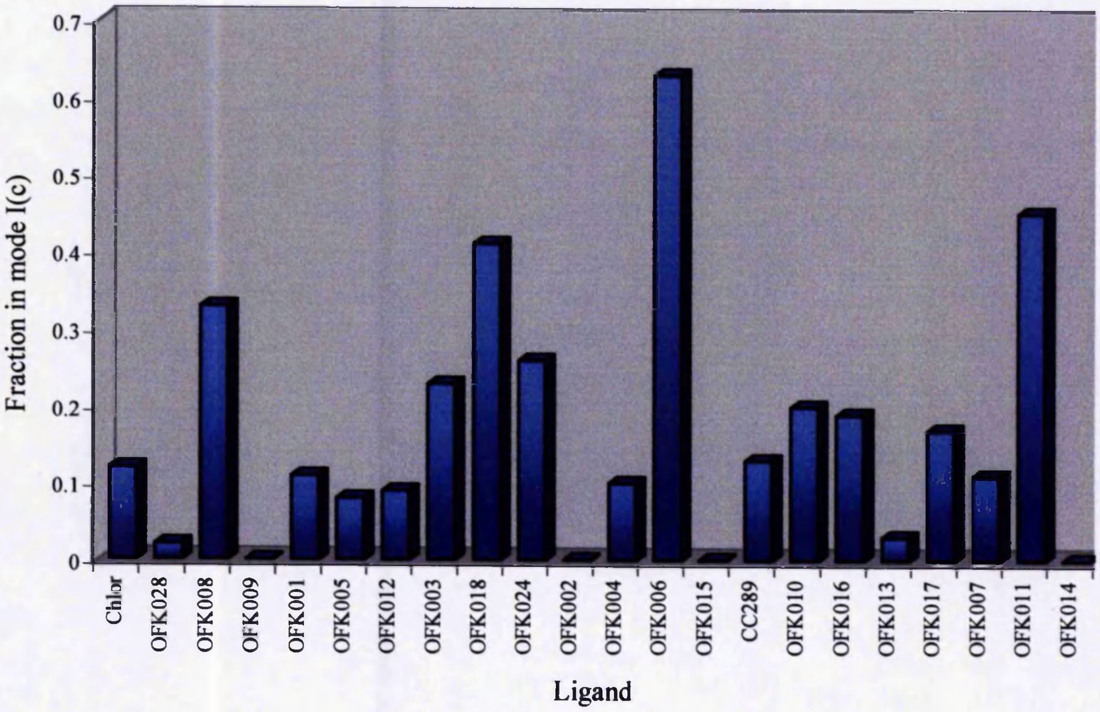


Figure 4.9(a). Fraction of phenothiazine ligands in the top 20 clusters found in binding modes I(a) and I(b). The ligands in mode I(a) are ordered in descending order of magnitude and the ligands in modes I(b) are aligned with these. (Chlor refers to chlorpromazine).



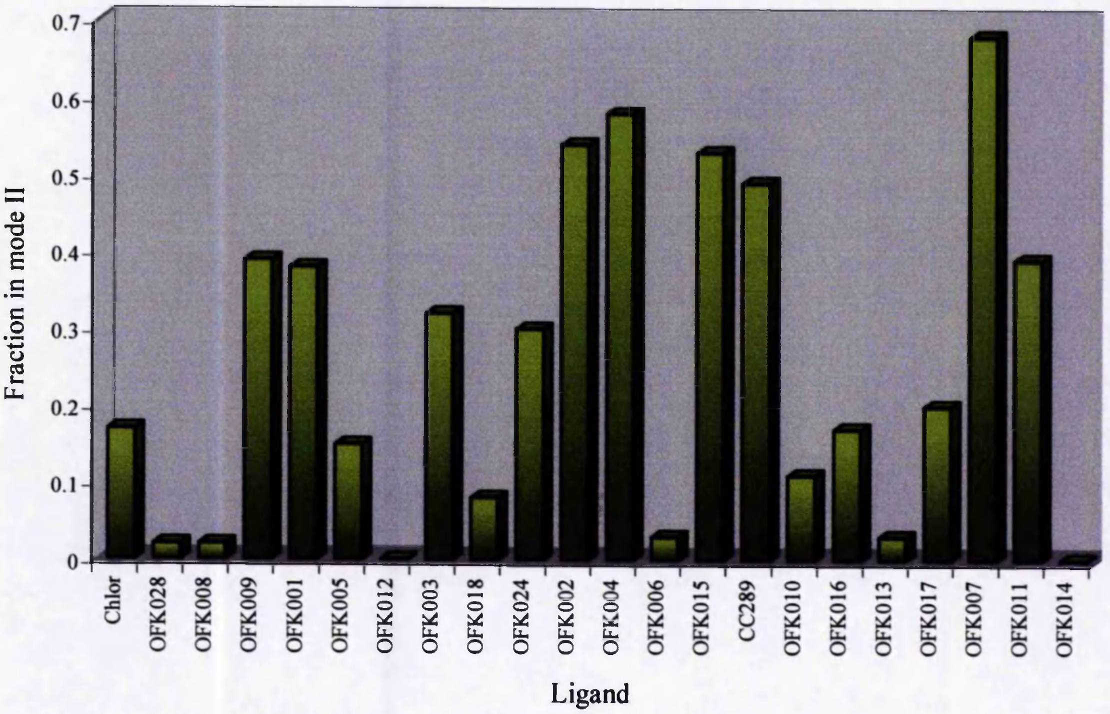


Figure 4.9(b). Fraction of phenothiazine ligands in the top 20 clusters found in binding modes I(c) and II. The ligands are ordered so as to be aligned with those shown in mode I(a). (Chlor refers to chlorpromazine).

tricyclic moiety or quaternary aromatic group to interact with neighbouring hydrophobic amino acid residues.

For OFK002 and OFK003, mode II allows a chlorine atom of the quaternary aromatic site to form a potential hydrogen bond with E-18, (if its acid group is protonated), thus contributing to the binding energy. In most cases a few individual conformations were detected which satisfy one of these interactions, *i.e.* either electrostatic or hydrophobic.

For these eight ligands, (CC289, OFK001-004, OFK009, OFK011, OFK017), the substituents on the benzyl side chain appear to make no difference to the docking preferences or to the binding energies. It is interesting to note that mode I(b) was not seen for these ligands, perhaps because the tricyclic moiety is more hydrophobic than the quaternary aryl group and therefore preferably interacts with the Z-site.

Other ligands, including chlorpromazine, did show preference for one mode:

Chlorpromazine: This ligand showed the best clustering of all the ligands docked, with 33 out of the final 100 docked structures in the top cluster alone. This cluster, together with 11 other conformations from separate clusters, puts the ligand in binding mode I(a). One quarter of this number had the N^+-S-14 interaction of binding mode II.

OFK005. Of the docked conformations in the top 20 clusters, just over half had the N^+-E-466'/467 interaction. Of these, 50% had the tricyclic in the Z-site, 25% had the cyclohexane side chain here and 25% had the tricyclic group directed towards residues L-17/M-113. Only approximately one seventh of all the conformations in the final 20 clusters showed the N^+-S-14 interaction.

OFK007 and *OFK015*. In these cases approximately three times as many final conformations were found with the N^+-S-14 interaction (mode II) than with N^+-E-466'/467' (mode I). For OFK007, in most cases a chlorine atom on the benzyl group of the side chain appeared to interact with E-18 and, for OFK015 some conformations allowed the nitro group on the benzene ring to interact with E-18. Perhaps these additional interactions account for some of the binding preference.

OFK016. Here, three times as many conformations were seen to have the quaternary nitrogen interacting with E-466' than with S-14. This binding mode preference may be because the E-466' acid group is more exposed to the active site cavity, whereas the S-14 is less accessible and surrounded by other residues so that the large naphthalene substituent sterically hinders binding to S-14. With the N^+-E-466' interaction, either the tricyclic or the naphthalene substituent interacted in the Z-pocket, but in the S-14 case, there appeared to be no hydrophobic interactions satisfied.

OFK024. Of the top 20 clusters, approximately twice as many of the runs resulted in the ligand being positioned in mode I(a) or (b), or showed just the N^+-E-466' interaction, than were in mode II.

Four other ligands show a complete preference for orientation I:

OFK006. Only two runs in the top 20 clusters resulted in the nitrogen interacting with S-14. All others had the N^+-E-466' interaction and many were in conformation I(b). One single cluster contained 15 of the final 100 docked structures with N^+-E-466', with the hydrophobic side chain positioned towards the Z-site and the tricyclic part stretching towards M-113. Both hydrophobic parts of the ligand were not close enough to the amino acid side chains to form optimum interactions though. Many conformations were also in mode I(a).

OFK012. The majority of these resulting structures had the N^+ interacting with E-466' or E-467' mostly leaving the tricyclic part to interact in the Z-site. Approximately one sixth had the N^+ interacting with E-18, but many other conformations did not seem to form any strong electrostatic or hydrophobic interactions and were sited in the centre of the cavity apparently not interacting. It may be that the size of the two benzyl substituents sterically prevents the N^+ interacting with S-14.

OFK028. The top 24 clusters contained 73 of the total 100 docked conformations. A third of these were with N^+-E-466', the tricyclic group near to the Z-site and the pyrene moiety well up in the active site cavity near I-383. 25% of the conformations all had N^+-E-466', with pyrene in the Z-site and the tricyclic spanning towards the M-113 region, but not close enough to it to form hydrophobic contacts. Other conformations were a combination of these interactions, *i.e.*:

1. pyrene in the Z-site, N^+ near E-466' and the tricyclic near I-383
2. pyrene near I-383, N^+-Glu466', with the tricyclic reaching towards the M-113 site or
3. the tricyclic in the Z-site, N^+-E-466' and the pyrene stretching to M-113.

Again, no N^+-S-14 interaction was seen because the pyrene moiety would cause steric crowding.

It is possible that these three ligands (OFK006, OFK012, OFK028), with bulky substituents are sterically prevented from forming the N^+-S-14 interaction as S-14 is less accessible in the active site than E-466' and E-467'.

OFK018. There was no real clustering observed (the top 20 or so clusters only having 1-4 members each), but of these, about half of the clusters (17 final conformations) had the N^+ interacting with E-466' or E-467' and the tricyclic near the Z-site. (Only three runs had the tricyclic near M-113/Y-110 or W-21). The remaining clusters had ligands not favouring any particular region.

OFK010. There was very little clustering for this ligand. A handful of conformations showed the N$^+$-E-466'/467' interaction and/or with the tricyclic in the Z-site, some had N$^+$-E-466' and the benzyl substituent in the Z-site and some had the N$^+$-E-466' interaction only. Only one or two had the N$^+$-S-14 interaction and the linking ether oxygen near enough to E-18 to potentially interact. This ligand had the highest number of rotatable bonds of the phenothiazine ligands and may be too flexible to cluster well.

Other ligands

OFK013. No clustering was seen for this ligand. A few lower energy conformations had N$^+$ near E-466' and the tricyclic near the Z-site, but most conformations were randomly sited. The side chain again may be too big to allow N$^+$ to interact with S-14. It appears that the oxygen atoms in the side chain ring influence binding, though it is unclear why that should prevent N$^+$ from binding E-466'.

OFK014. The ligand is structurally identical to OFK007, but with an oxygen atom replacing the quaternary nitrogen. This resulted in the loss of approximately 10kcalmol^{-1} of binding energy. This is consistent with the undetectable inhibition for OFK014 at the 50μM level. From the docking runs, one cluster (6 members) had the tricyclic moiety relatively near the Z-site. Nearly all other clusters were scattered round the active site, with many away from the protein side chains located in the open space of the cavity, and not making any favourable, stabilising contacts. By virtue of the absence of a quaternary nitrogen, this ligand serves as a control, tentatively suggesting that it is the positive nitrogen alone and not the tricyclic moiety that gives rise to families I and II. This docking result also serves to show that a large area of conformational space in the active site was sampled by the docking procedure.

OFK017. Many docked conformations for this ligand did not seem to bind preferentially anywhere and were situated in the centre of the active site or near the surface of the enzyme. Approximately 20% of the conformations in the top 20 clusters had the N$^+$-S-14 interaction and 20% had N$^+$-E-466' with the tricyclic in the Z-site or N$^+$-E-466' only. The fluorine atoms did not appear to make any difference to the electrostatic interactions of the benzyl substituent with enzyme residues.

Summary

From the phenothiazine docking results, many ligands bind the TR active site equally using modes I and II. Some have a distinct preference for mode I(a). This latter mode is favoured by ligands bearing a large side chain which would sterically hinder attempts to bind the less accessible S-14, e.g. OFK012, OFK016 and OFK028. Chlorpromazine does not appear to bind any less strongly by these calculations than the quaternary compounds with a benzyl substituent, though it did form one large cluster, perhaps due to its only having four rotatable bonds.

Table 4.2 shows the lowest relative energies of the docked structures with the corresponding K_i/I_{50} values determined experimentally [390]. Referring to Table 4.7 too, it can be surmised that these values do not correlate with each other and structure cannot account for the seeming preferences in binding modes. The binding orientations appear isoenergetic and as such, ligands should be found with equal probability in each mode.

4.3.2.4. Open ring quaternary compounds

These compounds (Table 4.3) have additional flexibility with three extra torsions per molecule compared to the phenothiazines due to the open conformation of the main ring structure. Only six ligands of this family were available to be docked. Table 4.8 and Figure 4.10 show the fraction of ligands in each binding mode.

Ligand	Fraction in Mode I(a)	Average Energy (kcalmol^{-1})	Fraction in Mode I(b)	Average Energy (kcalmol^{-1})	Fraction in Mode I(c)	Average Energy (kcalmol^{-1})	Fraction in Mode II	Average Energy (kcalmol^{-1})
OPN-00	0.40	-62.87	-	-	0.06	-63.46	0.54	-63.18
OFK289b	0.33(NH)	-87.94	-	-	0.05 (NH)	-89.15	0.21	-89.30
OFK007b	0.03	-64.96	-	-	0.19	-66.67	0.57	-66.69
OFK010b	0.07	-72.66	-	-	0.10	-70.97	0.28	-71.05
OFK011b	0.03	-63.32	-	-	0.35	-65.97	0.43	-66.00
OFK012b	0.36(NH)	-70.59	0.17(NH)	-70.82	0.06 (NH)	-71.10	-	-

Table 4.8. The fraction of open ring ligands in the top 20 clusters found in each binding mode. (NH) indicates that it is the amine nitrogen of the central ring that interacts with the protein residues and not the quaternary nitrogen.

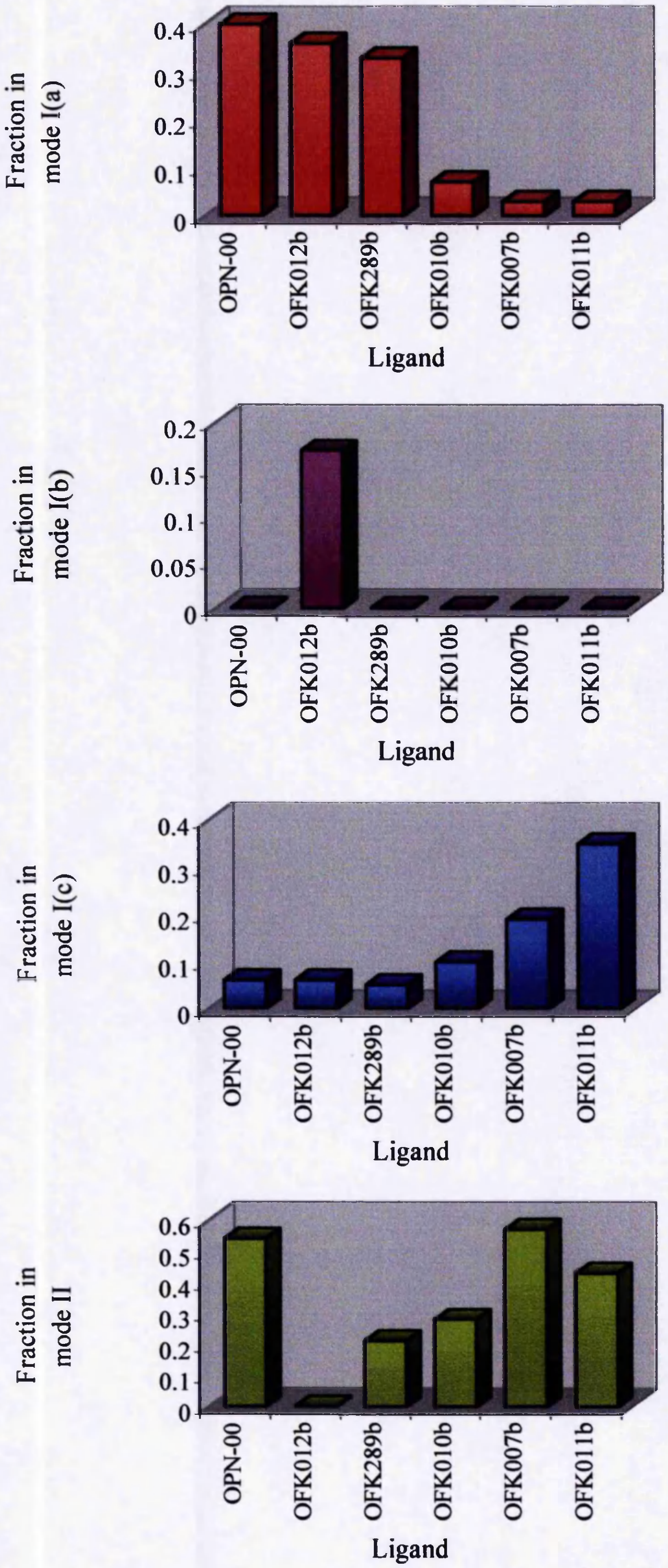


Figure 4.10. Fraction of open ring ligands in the top 20 clusters found in each binding mode. The ligands in mode I(a) are ordered in descending order of magnitude and the ligands in the other three modes are aligned with these.

The previous common interactions of the positive nitrogen with E-466'/467' and S-14 were again observed. However, the introduction of an NH group to the molecule also contributed to electrostatic interactions and binding orientation in some cases, but the conformations were still of similar energies.

OPN-00, OFK010b and OFK011b. These three ligands had roughly equal numbers of docked states in the previously described binding orientations I and II. For OPN-00, approximately 20% of the total 100 runs finished in orientation I and 25% in orientation II, (the majority with the hydrophobic groups near L-17/W-21/M-113). Some of the conformations from both families had the NH group forming hydrogen bonds with neighbouring residues, e.g. with the -OH group of Y-110, with E-467' and with backbone carbonyl groups. For OFK010b, approximately 25% of the members of the top 20 clusters had the common N^+-E-466'/E-467' interaction, but usually with the hydrophobic groups lying in the open cavity of the active site. Approximately another 25% had the N^+-S-14 interaction, some with the hydrophobic groups near W-21/M-113. Most of the other conformations had the ligand situated in the centre of the cavity or towards the top, away from the protein residues and failing to pack against any side chains. For OFK011b, a third of the conformations showed the N^+-E-466' interaction and one of the two the hydrophobic groups lay near either the Z-site or M-113/L-17/W-21. About a third of the conformations had the usual interaction of N^+ with S-14 and the benzyl group positioned near M-113/L-17. Others conformations only had the N^+-S-14 or E-466' interaction.

OFK007b. Three times as many final docked conformations were found with the N^+-S-14 interaction as with the N^+-E-466' interaction. Only one significant cluster was seen (with 7 members) with N^+-S-14 and the hydrophobic sulphide moiety in the Z-site. About the same number of conformations in other clusters also had N^+-S-14, but with the sulphide near the region containing W-21/M-113/Y-110/L-17. In some of these structures, the chlorine substituents on the benzyl group can were within interacting distance of either E-18 or E-466'. Some other structures had either a double N^+ and NH-E-466'/E-467' interaction or just the N^+- E-466' interaction.

OFK289b and OFK012b. A very commonly detected docked conformation for these two ligands was with the NH of the central ring interacting with E-466'/467' and the sulphide moiety (OFK289b) or the benzhydryl moiety (OFK012b) in the Z-site. This conformation accounted for half of the conformations observed for OFK289b. Many others had N^+-S-14, but no hydrophobic interaction. Again, over half the top low energy conformations of OFK012b were positioned in this way, but the positive nitrogen was not seen to interact with any acid residue or S-14 in any top cluster. The benzhydryl group may cause steric hindrance in trying to bind near S-14. OFK012 displayed a strong preference for the N^+ to bind E-466'/467', but OFK012b does not. This may be because OFK012b has more rotatable bonds and is less frozen entropically.

Summary

It appears that for a few compounds the opening of the central ring and the consequent prescence of an NH group influences and alters binding orientation in comparison with the phenothiazines, but on the whole, similar interactions are seen for both classes of compounds.

Again there was no correlation between the binding energies and the K_i values, e.g. OFK007b is an 8-fold better inhibitor than OFK289b based on K_i but did not bind as strongly from the docking study. OPN-00 was not found to bind differently with respect to docking energy or orientation to the quaternary compounds.

Because of the three extra rotatable bonds, the algorithm may spend a greater proportion of time searching torsional space and less time physically moving the molecule by translations for this family.

4.3.2.5. Imipramines

The only difference of these ligands from their phenothiazine counterparts is the replacement of the central sulphur-bridge of the tricyclic with two carbons. This sulphur atom contributed approximately –5.5kcalmol^{-1} to the total binding energy (~70kcalmol^{-1}) in the phenothiazine study and it is unsurprising that the same interactions as before were found to be favoured here. Table 4.4 shows the results of the docking. Table 4.9 and Figure 4.11 show the fraction in each binding mode.

Ligand	Fraction in Mode I (a)	Average Energy (kcalmol^{-1})	Fraction in Mode I(b)	Average Energy (kcalmol^{-1})	Fraction in Mode I(c)	Average Energy (kcalmol^{-1})	Fraction in Mode II	Average Energy (kcalmol^{-1})
Chlomipramine	0.47	-67.49	-	-	0.18	-65.75	0.23	-65.90
OFK289c	0.31	-62.03	-	-	0.06	-61.88	0.25	-60.75
OFK007c	0.29	-65.94	-	-	0.29	-63.45	0.25	-66.33
OFK010c	0.26	-71.47	-	-	0.02	-71.58	0.38	-73.07
OFK011c	0.17	-62.61	0.02	-62.77	0.12	-64.90	0.47	-65.40
OFK012c	0.43	-73.61	-	-	-	-	-	-

Table 4.9. The fraction of imipramine ligands in the top 20 clusters found in each binding mode.

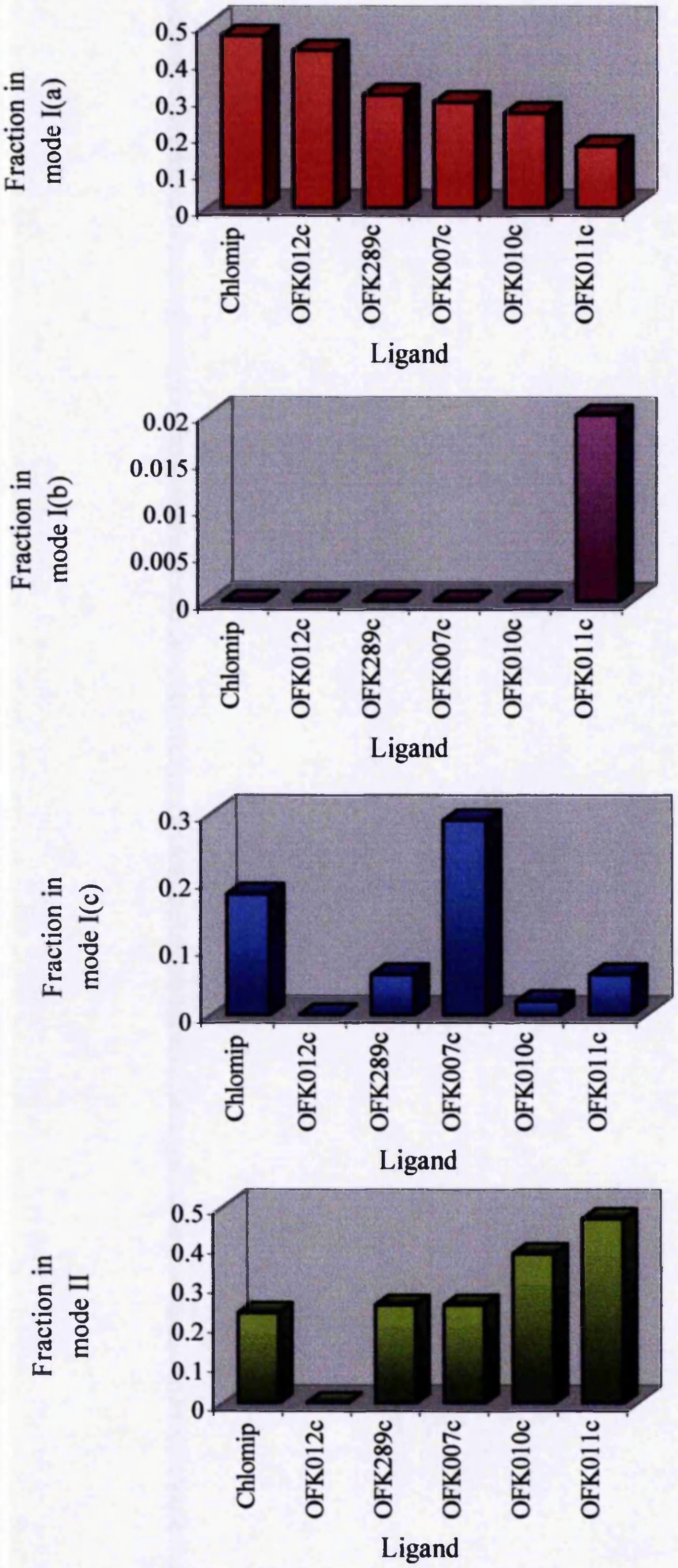


Figure 4.11. Fraction of imipramine ligands in the top 20 clusters found in each binding mode. The ligands in mode I(a) are ordered in descending order of magnitude and the ligands in the other three modes are aligned with these. (Chlomip refers to chlomipramine).

OFK010c had equal numbers of docked structures in binding modes I and II, *i.e.* half with N^+-E-466' and the tricyclic in the Z-site and half with N^+-S-14. In the latter case, the linking ether oxygen is close enough to hydrogen bond with E-18 allowing the tricyclic or benzyl substituent to point towards M-113 (W-21/L-17) in the majority of cases. For OFK010, there was a preference to bind in mode I, but mode II was also seen, again with the oxygen close to E-18.

Chlomipramine, OFK007c and OFK011c had distinct and dissimilar binding preferences.

Chlomipramine. Clustering here was quite significant with 36 of the final 100 docked conformations in the top two clusters. In total, 51 of all the conformations in the top 20 clusters were located with the N^+-E-466' interaction and the tricyclic in the Z-site. Approximately one third of this number had the N^+-S-14 interaction, sometimes with the tricyclic near L-17/M-113.

OFK007c. Twice as many final docked conformations were found with N^+-E-466'/467' as with the N^+-S-14 interaction. Of the former, 50% had N^+-E-466' with the tricyclic interacting with L-17/I-106/Y-110/M-113, and 50% had N^+-E-467' with the tricyclic in the Z-site. Quite a few structures also had single electrostatic (N^+-E-466') or hydrophobic interactions. For the phenothiazine OFK007 opposite preferences were seen, with three times the number of structures found with the N^+-S-14 interaction as with N^+-E-466'.

OFK011c. Twice as many conformations were found with N^+-S-14 and the benzyl side chain near L-17/Y-110/M-113 as with N^+-E-466'/467' and the tricyclic in the Z-site or near L-17/M-113/I-383. Again, others had a single electrostatic or hydrophobic interaction. This is in contrast to OFK011 where there was no binding preference.

OFK012c. The first 10 clusters contained 44 members between them. 40 of these were with the tricyclic in the Z-site. Only 14, however, had N^+-E-466'/E-467'. The remainder did not seemed to have N^+ a little further away from E-466' than normally seen, *i.e.* ≥ 4.2Å away compared to ~3.5Å, but the N^+ interaction still gave a large negative energy when the energy breakdown into the individual atoms was analysed. No N^+-S-14 interaction was observed again probably because of steric hindrance of the benzhydryl group. OFK012 showed a distinct preference for binding mode I as well.

OFK289c. The results for this ligand showed that 72 of the total 100 final structures were in the first 20 clusters. 30 of these conformations, with varying energies, had the tricyclic in the Z-site, and many had the quaternary nitrogen quite close to E-466', but interestingly not so close to it as is normally found (approximately 4.2Å compared to 3.5Å). However, this did not seem to affect the nitrogen atom's interaction energy. Again, if this orientation occurred *in vitro*, a water molecule may bridge the gap. 18 conformations had N$^+$-S-14, but no hydrophobic interactions were made by the rings. A handful of others had the N$^+$-E-18 interaction with 5 conformations having the tricyclic near I-106/Y-110/M-113. For the phenothiazine counterpart CC289, half the ligands were in orientation I(a) and half in II.

Summary

Although structurally very similar to their phenothiazine counterparts, most of the imipramines did not show the same binding characteristics, with two ligands (OFK007c and OFK011c) showing opposite binding preference to those seen for the sulphur-containing ligands. This may not be a property of the program: the findings may support the view that both binding orientations are of equal energy and there is an equal chance the inhibitor will bind to either site. The K_i values were not greatly different from those for the phenothiazines. Chlomipramine, with only four rotatable bonds, clustered well, but did not appear to bind any more strongly than the quaternary compounds.

4.3.2.6. Mepacrine

The docking results for mepacrine are shown in Table 4.5. Table 4.10 shows the fraction in each binding mode.

Ligand	Fraction in Mode I (a)	Average Energy (kcalmol^{-1})	Fraction in Mode I(b)	Average Energy (kcalmol^{-1})	Fraction in Mode I(c)	Average Energy (kcalmol^{-1})	Fraction in Mode II	Average Energy (kcalmol^{-1})
Mepacrine (uncharged)	-	-	-	-	-	-	-	-
Mepacrine (charged)	0.33	-65.00	-	-	0.10	-65.75	-	-

Table 4.10. The fraction of mepacrine conformations in the top 20 clusters found in each binding mode.

Charged. The crystal structure of mepacrine bound to TR shows the tricyclic in the region of L-17/W-21/Y-110/M-113 and the nitrogen interacting with E-18 *via* a water molecule. However, this was not the favoured conformation observed in the docking studies. Of the top 39 docked structures, (20 clusters), almost half were in binding mode I(a) (the tricyclic near the Z-site and

N^+-E-466'). The other structures were such that only one of these two interactions was satisfied. None of these conformations showed any close similarity to the docked conformation of mepacrine seen in the crystal structure and only one cluster (with 3 members) had the tricyclic part of the molecule near M-113 and Y-110. This location does not have the N^+ directly interacting with E-18, but it is close enough to perhaps interact *via* water. Two conformations did have N^+ interacting with E-18 but no additional stabilising hydrophobic interactions. The replacement of a benzyl substituent by ethyl groups did not make any difference to the binding energies.

Uncharged. Only 20% of the docked conformations had the tricyclic near the Z-site but the nitrogen did not display any interactions. Most of the other conformations were sited in the open space of the active site and not close to any enzyme residues at all. The lack of a positively charged nitrogen has decreased (*i.e.* made less negative) all binding energies by approximately 10kcalmol^{-1} as was seen for OFK014.

4.4 CONCLUSION

4.4.1. Present study

These studies suggest that two distinct, but energetically equivalent, binding modes exist for the quaternary ammonium compounds studied, distinguished by the positioning of the positive nitrogen and the hydrophobic tricyclic root structure.

Mode I. This family has the nitrogen atom (N$^+$) electrostatically interacting with E-466'/467' with the tricyclic (or sometimes the benzyl side chain) forming hydrophobic contacts with the Z-site of F-396'/P-398'/L-399' (Figure 4.8(a)).

Mode II structures have the N$^+$ atom interacting with S-14, which sometimes allows one of the hydrophobic groups to interact with the hydrophobic region containing residues L-17/W-21/Y-110/M-113 (Figure 4.8(b)).

There is no difference energetically between these two conformations. It does not appear that the substitution pattern on the benzene ring of the N$^+$-side chain influences binding orientation or affinity. Indeed, many ligands do not seem to discriminate between the two sites a great deal and appear capable of binding both sites under *in vitro* experimental conditions, which are thus likely to be equally populated. However, for ligands OFK007, OFK007b (with a dichloro substituent) and OFK015 (with a nitro substituent) the N$^+$-S-14 (mode II) conformation is preferred as this potentially allows an electrostatic interaction of the ring substituent with E-18.

Only a few ligands seem to display a strong preference for one of these sites, usually the E-466'/467'-Z-site orientation, e.g. OFK006, OFK008, OFK012, OFK018, OFK028. These ligands have large side chains and, since the two glutamic acid residues are more accessible than S-14, less steric hindrance occurs when the ligand binds in this conformation.

These results confirm predictions that a tricyclic molecule with a quaternary ammonium side chain could bind to the Z-site and the residues E-466' and E-467', respectively. An unexpected finding was the importance of S-14 in the docking of the inhibitors. This site has not previously been cited as a potential anchor for inhibitor design. Although none of the results appeared to show ligand binding to the residues important in catalysis, [468,469,471] these inhibitors should work competitively by preventing the binding of the natural substrate.

The natural substrate of TR is trypanothione (Figure 4.2(b)) but no X-ray structure of it complexed with TR has been solved. However, the structure of TR complexed with one of the naturally occurring substrates N^1-glutathionylspermidine disulphide ([GspdS]$_2$) has been determined (Figure 4.2(c)) [471]. This structure shows that in TR the side chains of W-21 and M-113 can provide a hydrophobic pocket positioned to bind the spermidine moiety and the carboxyl group of E-18 hydrogen bonds with the amide nitrogen of the spermidine group. Residues F-396', P-398' and L-399' form from hydrophobic interactions with part of the alkyl chains of this substrate and E-466' and E-467' form direct hydrogen bonds. By binding in orientation I, the quaternary ammonium ligands therefore disrupt many of these important protein-substrate interactions, thus inhibiting the enzyme. When the ligands bind in mode II, they often form a hydrophobic interaction with L-17/W-21/M-113, so preventing the spermidine moiety or the substrate from binding. The ligand is held in a position to do this by the interaction of the positive nitrogen with S-14. The side chain of S-14 is not seen to be important in the binding of $[GspdS]_2$, nor is it expected to be for trypanothione. It is potentially capable of forming a water-mediated hydrogen bond to $[GspdS]_2$, but otherwise only its main chain atoms are involved in interactions [468,471].

Virtually none of the docked conformations were similar to that seen of mepacrine bound in the enzyme determined by X-ray diffraction [472]. This structure showed that the tricyclic was in the L-17/W-21/M-113 site, as seen in mode II, but the side-chain nitrogen was interacting with E-18. This thesis study provides no evidence to suggest that E-18 is important in quaternary inhibitor binding, but the results do agree with the importance of the L-17/W-21/M-113 region.

These findings may be useful for the future design of anti-trypanosomal lead compounds against TR. Perhaps a quaternary ligand could be designed which is long enough to span the active site from the Z-site to the W-21/M-113 region, maintaining the N^+-Glu466'/S-14 interaction, or even a ligand which incorporates two nitrogen atoms to satisfy both electrostatic interactions.

4.4.2. Other docking studies on trypanothione reductase

A predictive algorithm of the binding affinity of various ligands to TR has already been used for the 'virtual screening' of a data base of 2500 molecular sketches and has detected several putative ligands [445]. The algorithm converts the two-dimensional molecular sketches into three-dimensional ligand structures, explores the conformational space of the latter, and performs a grid-based, rigid-body docking of the resulting family of ligand conformations into the TR site, calculating enthalpic and entropic binding indexes and predicting the binding affinity. Ligand flexibility was accounted for by docking a relevant *family of rigid conformations* of the ligand. In this algorithm a desolvation term takes into account the desolvation of the ligand by the site and *vice versa*. This is absent from AUTODOCK. A further term estimates binding entropy, since the flexibility of the TR inhibitors means that the binding *entropic* contributions are expected to be important. By comparison with experimentally obtained binding constants, this algorithm was successful in predicting the correct order of magnitude of the inhibition constants of a wide variety of compounds.

Mepacrine was also used in that study and one of its two predicted binding modes resembled that observed in the X-ray structure of it complexed with TR. This placed the aromatic moiety in the hydrophobic pocket defined by W-21, M-113 and Y-110, but with the amino group interacting with E-466' and E-467' and not with E-18. The second binding mode involved an alternative hydrophobic area defined by F-396', Thr-463', P-462', V-58, V-53, L-399', I-338 and I-106. This involves additional hydrophobic residues to add to those in the hydrophobic sites of M-113/W-21 and the Z-site and these residues also appear to be important in $[GspdS]_2$ binding [471]. The nitrogen atom does appear to interact with E-18 in the second binding mode.

4.4.3. Conclusions of the AUTODOCK study

The program AUTODOCK appeared to operate efficiently, was accurate and selective in terms of energy evaluation (discriminating between a well-docked and poorly-docked structure) and sampled conformational space well (as seen for OFK014). It successfully gave consistent results within a family of ligands in terms of binding orientation and energy. However, AUTODOCK was unable to discriminate between ligands having subtle changes in the substitution pattern of the benzyl side chain and showed no difference in binding energy between them. These ligands show significant differences in their experimental K_i values, but these could not be correlated with the docking energies. This may not be due to the nature of the docking problem and the large size of the active site (approximately 22Å x 20Å x 28Å), where such small structural alterations are relatively insignificant in the context of overall binding.

The results for ligand OFK014 (with an oxygen atom replacing the positive nitrogen) showed no particular binding location preference and no docked conformational family. Thus, the nitrogen is probably most important in determining orientation, not the hydrophobic groups. Support for this assumption comes from analysis of the contribution of the energy of each atom to the intra- and inter-molecular docking energy of the complex. The greatest contribution comes, not surprisingly, from the electrostatic interaction of the positive nitrogen with surrounding groups and this can provide between 4 and 20 times more electrostatic interaction energy than other atoms in the ligand. An interaction between this N^+ and a polar group thus drives the docking process and dominates the movement and final location of the ligand in the active site. Hydrophobic (van der Waals') interactions do not contribute much by comparison. If one wanted to study the docking of these ligands using only their hydrophobicity as the determinant of the docked location, the electrostatic energy term (or the atomic charges) would have to be switched off. It has already been suggested that simple charge characteristics, rather than differences in hydrophobicity, may account for a significant portion of the selectivity of a series of chlorpromazine analogues for TR over GR [483]. An algorithm designed to deal with hydrophobic docking is described later.

4.4.4. Previous applications of AUTODOCK

AUTODOCK has successfully been applied to the docking of many ligands prior to this study, but in cases involving active sites containing fewer hydrophobic and more directional bonding interactions, e.g. in the docking of isomaltose analogues [452], monosaccharide substrates and methyl α-acarviosinide [450] and glucosyl disaccharides [451] in the glucoamylase active site. This produced bound complexes comparable to those obtained by protein crystallography. Although grid size and number and spacing of grid points were very similar to those used in this study, the active site in glucoamylase is a few angstroms smaller in each direction than TR (approximately 20Å x 20Å x 25Å), more enclosed and contains a lot of hydrogen-bonding residues. The saccharide ligands themselves have many hydrogen-bond-donating and -accepting hydroxyl groups and oxygen atoms. Such ligands are involved in more directional binding interactions and the disaccharides especially, with flexibility around the glycosidic bond, can change conformation to satisfy as many hydrogen bonds as possible.

AUTODOCK has also been used to explore the binding of citrate to aconitase [453], the active site of which is totally enclosed allowing the enzyme to contact the substrate from all sides. It has also been used in docking benzamidine to β-trypsin, camphor to cytochrome P-450, phosphocholine to Fab McPC-603, biotin to streptavidin, sialic acid in haemagglutinin and the cyclic urea protease inhibitor XK-263 in HIV-1 protease [454]. The grid size used in each case here, was 22.875Å x 22.875Å x 22.875Å, compared to the TR situation of 29.625Å x 29.625Å x 30.375Å. The active sites were well-defined (knowing where the ligand bound) and in all cases except sialic acid/haemagglutinin, the best solution was energetically well-separated from the less-favourable conformations. Benzamidine is a small, rigid molecule and binds tightly in the specificity pocket of trypsin. The binding of phosphocholine (4 rotatable bonds) to Fab McPC-603 was predominantly electrostatic and the streptavidin-biotin complex is one of the most tightly binding complexes known because of multiple hydrogen bonds and Van der Waals' interactions. Biotin was docked with 5 rotatable bonds. The sialic acid-haemagglutinin interaction was also dominated by hydrogen bonding with sialic acid possessing 8 hydrogen bonding groups and 10 rotatable bonds. HIV-1 protease opens up *in vivo*, allowing substrates and inhibitors to insert into the active site and then closes, completely enclosing the molecule *i.e.*, once docked, the ligand is held tightly in a confined space. AUTODOCK has also been used to dock substitued 2,4-dinitrophenol and other haptens to the multispecific antibody IgE (La2) [455] and reproduce X-ray structure complexes.

Thus, all these studies have polar/charged/hydrogen-bonding groups in the ligand and protein giving some directionality to the ligand-receptor interactions and thus 'guide' orientation of the ligand. A ligand containing only hydrophobic groups relies solely on non-directional Van der Waals' interactions to orientate it. This can pose problems when trying to dock hydrophobic ligands into active sites. This fact, together with the large size of the receptor site, and hence the large conformational space to search, may also contribute to the lack of clustering seen in the study of TR. The AUTODOCK studies mentioned above often found RMS deviations of less than 1Å from the crystallographic conformation are obtained for the lowest-energy dockings, although fewer dockings find the crystallographic conformation when there are more degrees of freedom. The latter observation agrees with the present study, with more degrees of freedom, *i.e.* more rotatable bonds in the ligand increasing the time for each docking run.

The active site of TR is very large, open, easily accessible from the outside solvent and predominantly hydrophobic. Most of the hydrophobic ligands studied have few atoms capable of hydrogen bonding or electrostatic interactions, but all of them possess one positively charged nitrogen to provide electrostatic interactions and help to vectorise the ligand. The ligands also have 6 to 12 rotatable bonds adding to the complexity of the docking problem. It may be that AUTODOCK may be more suited to situations of smaller enclosed active sites with more hydrogen bonding and electrostatic interacting capabilities than predominantly hydrophobic sites.

New search methods have been introduced and tested in AUTODOCK using a new, empirical binding free energy function for calculating ligand-receptor binding affinites [438]. Three search methods have been tested (simulated annealing, traditional genetic algorithm and a Lamarckian genetic algorithm). Traditional and Lamarckian GAs were found to handle ligands with more degrees of freedom than the simulated annealing method used in AUTODOCK with the LGA being the most efficient, reliable and successful. The introduction of the LGA search method extends the power and applicability of AUTODOCK to docking problems with more degrees of freedom than could be handled by previous versions.

4.4.5. Hydrophobic docking

An enhancement to molecular recognition techniques has been proposed [424] based on the assumption that in the case of macromolecules, the surface density of hydrophobic atoms should generally be higher at the binding site than elsewhere on the surface. This is supported by the results of a systematic survey of known complexes [484] indicating that on average, a substantial complementarity between hydrophobic groups is manifested at the area of contact, and the hydrophobicity of the interface is higher than elsewhere on the surface. This method exploits the hydrophobicity of surface groups on the molecules to be matched and separates atomic groups into hydrophobic and non-hydrophobic. Geometric complementarity is looked for between molecules whose surfaces are represented by the hydrophobic atom groups only. This approach partially relieves difficulties of other molecular recognition methods, *viz.*,

1. Problems of false solutions. In view of the high occurrence of hydrophobic groups at contact sites, their contribution results in more intermolecular atom-atom contacts per unit area for correct matches than for false positive fits. The elimination of non-hydrophobic patches in the representation of the molecular surface, while decreasing the quality of the match, should affect the false positive fits more than the correct match.
2. Problems of conformational change. Assuming that non-hydrophobic groups at the surface are, in general, more exposed to the aqueous solvent, one may consider them as more flexible. Therefore, they are potentially subject to small conformational changes upon their interaction with another molecule. By taking into account only the more rigid hydrophobic groups, leaving some room at the surface would contribute to a better tolerance for local conformational changes involving more the hydrophilic groups.
3. Problems of the number of atoms. The number of atoms is reduced by eliminating non-hydrophobic groups. Therefore, the number of operations performed by a recognition procedure should be reduced accordingly.

Similar 'hydrophobic/non-hydrophobic' representations of molecules by either atoms or residues have been used and applied to the problem of peptide-receptor interaction [485] or to protein folding [486], respectively.

This method was tested on four protein complexes from the Brookhaven Data Bank [134] using 'full' and 'hydrophobic' representations of the molecules [424]. It was found that the procedure worked more efficiently with the hydrophobic representation and even though the hydrophobic docking results were obtained with the molecules represented by only one third of their surface heavy atoms, this yielded improved results compared to the previously established geometric docking approach.

The results strongly suggest that surface hydrophobic groups substantially contribute to geometric surface molecular recognition and illustrates the basic difference between inter-molecular energy calculations and recognition techniques. In energy calculations, especially in the case of strong electrostatic interactions, the necessity of taking into account the contribution of charged groups is obvious. However, in recognition procedures a more faithful representation of molecules may be detrimental in predicting the correct match by increasing the relative score of false positive matches. This approach may be considered as a potential improvement for many molecular recognition procedures.

4.4.6. Limitations of docking programs

A limitation of AUTODOCK, and indeed many docking programs, is that protein motion is not modelled and successfully predicting large-scale protein conformational changes upon binding is difficult. The AUTODOCK method works well when there is little change between the *apo* and ligand-bound forms of the protein, even if the protein undergoes significant conformational changes during the actual binding. The active site of TR is fairly rigid and there were no major alterations in side-chain conformations seen when comparing native TR and enzyme complexed with $[GspdS]_2$ [471]. The active site is also constructed from a scaffold of secondary structure, mostly α-helices, which help to promote rigidity.

Another major problem with current ligand docking programs is the inability to model the locations of specific solvent molecules and their motion and involvement in docking. *In vitro* and *in vivo* electrostatic interactions may form between N^+ and polar amino acids *via* water molecules as seen for the mepacrine amino atom and E-18 in the mepacrine-TR complex [472]. This could never have been reliably predicted using current docking programs.

CHAPTER FIVE

SUMMARY OF THESIS

5. SUMMARY OF THESIS

This thesis has provided three examples of ways in which computational and molecular modelling techniques can be applied to rational molecular design and 3-D structure determination studies: the NMR and molecular modelling of a nucleic acid complex, the comparative modelling of a protein and the use of an enzyme crystal structure with a docking algorithm to analyse binding modes of ligands.

The structure of a binary oligonucleotide system bearing alkylimidazole constructs was determined here by NMR spectroscopy and refined using restrained molecular dynamics. The complex was designed to mimic the catalytic mechanism of ribonuclease A, but the inefficiency of the cleavage that was observed by the designers (Vlassov's group) can now be attributed to the flexibility of the alkylimidazole groups which do not form any stable interaction with the target nucleic acid. These findings can thus be used to design an improved cleaving system in the future.

A molecular model of HIF-1 was constructed in this thesis and was found to compare well with other, structurally similar proteins whose X-ray structures have been determined. It has been used, along with a model of the HLH protein Id3, in the design of peptide ligands, some of which were found to show significant selective binding to Id3. Mutations were suggested to provide a basis for FRET studies based on the comparative model.

The docking program AUTODOCK has been used in this thesis to dock three families of ligands to the enzyme TR. The program reliably reproduced two orientations of binding (one being as predicted) but did not appear to distinguish well between the hydrophobic groups of the ligands. It seems that the program is more suited to smaller active sites with more electrostatic or hydrogen bonding interactions to guide the placement of the ligand. A different algorithm with more emphasis on hydrophobic interactions would be more suited to this situation. This aspect of the study indicated that the special features of the large, hydrophobic TR active site are such that unless such a specialist hydrophobic docking algorithm becomes available which also deals with a flexible ligand, manual/interactive design methods for novel ligand invention or for ligand improvement continue to be the best approach for TR.

The above examples all illustrate the increasing importance of computer-aided molecular design in helping to determine the 3-D structure of a molecular system or drug target, aiding inhibitor design and helping to suggest improvements to existing lead compounds for therapeutic purposes.

CHAPTER SIX

REFERENCES

6. REFERENCES

1. Credi, A., Montali, M., Balzani, V., Langford, S. J., Raymo, F. M., Stoddart, J. F. *New J. Chem.*, 1998, **22**, 1061-1065.

2. Reimers, J. R., Hall, L. E., Hush, N. S., Silverbrook, K. *Annals of the New York Academy of Sciences*, 1998, **852**, 38-53.

3. Dunlap, B. I. *Physical Review B*, 1992, **46**, 1933-1936.

4. Ebbesen, T. *Physics Today*, 1996, **49**, 26-32.

5. Dai, J. Y., Lauerhaas, J. M., Setlur, A. A., Chang, R. P. H. *Chem. Phys. Lett.*, 1996, **258**, 547-553.

6. Han, J., Globus, A., Jaffe, R., Deardorff, G. *Nanotechnology*, 1997, **8**, 95-102.

7. Wilks, H. M., Halsall, D. J., Atkinson, T., Chia, W. N., Clarke, A. R., Holbrook, J. J. *Biochemistry*, 1990, **29**, 8587-8591.

8. Wilks, H. M. Holbrook, J. J. *Curr. Opin. Biotech.*, 1991, **2**, 561-567.

9. Arnold, F. H. Haymore, B. L. *Science*, 1991, **252**, 1796-1797.

10. Brinen, L. S., Willett, W. S., Craik, C. S., Fletterick, R. J. *Biochemistry*, 1996, **35**, 5999-6009.

11. Higaki, J. N., Haymore, B. L., Chen, S., Fletterick, R. J., Craik, C. S. *Biochemistry*, 1990, **29**, 8582-8586.

12. Cohen, C. Parry, D. A. D. *Proteins: Structure, Function and Genetics*, 1990, **7**, 1-15.

13. Lombardi, A., Bryson, J. W., DeGrado, W. F. *Biopolymers*, 1996, **40**, 495-504.

14. Olszewski, K. A., Kolinski, A., Skolnick, J. *Proteins: Structure, Function and Genetics*, 1999, **25**, 286-299.

15. Sauer, R. T., Hehir, K., Stearman, R. S., Wiess, M. A., Jeitler-Nilsson, A., Suchanek, E. G., Pabo, C. O. *Biochemistry*, 1986, **25**, 5992-5998.

16. Muheim, A., Todd, R. J., Casimiro, D. R., Gray, H. B., Arnold, F. H. *J. Am. Chem. Soc.*, 1993, **115**, 5312-5313.

17. Malakauskas, S. M. Mayo, S. L. *Nature Struct. Biol.*, 1998, **5**, 470-475.

18. *More employable enzymes* in *Chemistry in Britain*, 1999, January, 17.

19. Regan, L. Clarke, N. D. *Biochemistry*, 1990, **29**, 10878-10883.

20. Coldren, C. D., Hellinga, H. W., Caradonna, J. P. *Proc. Natl. Acad. Sci. USA*, 1997, **94**, 6635-6640.

21. Pinto, A. L., Hellinga, H. W., Caradonna, J. P. *Proc. Natl. Acad. Sci. USA*, 1997, **94**, 5562-5567.

22. Wisz, M. S., Garrett, C. Z., Hellinga, H. W. *Biochemistry*, 1998, **37**, 8269-8277.

23. Benson, D. E., Wisz, M. S., Liu, W., Hellinga, H. W. *Biochemistry*, 1998, **37**, 7070-7076.

24. Lewis, R. A. Leach, A. R. *J. Comp. Aided Mol. Design*, 1994, **8**, 467-475.

25. Gillet, V., Johnson, A. P., Mata, P., Sike, S., Williams, P. *J. Comp. Aided Mol. Design*, 1993, 7, 127-153.

26. Bohm, H. *J. Comp. Aided Mol. Design*, 1992, **6**, 61-78.

27. Lauri, G. Bartlett, P. A. *J. Comp. Aided Mol. Design*, 1994, **8**, 51-66.

28. Lam, P. Y. S., Jadhav, P. K., Eyermann, C. J., Hodge, C. N., Ru, Y., Bacheler, L. T., Meek, J. L., Otto, M. J., Rayner, M. M., Wong, Y. N., Chang, C.-H., Weber, P. C., Jackson, D. A., Sharpe, T. R., Erickson-Viitanen, S. *Science*, 1944, **263**, 380-383.

29. Thomas, J. *New Scientist*, 1998, **160**, 36-39.

30. Cohen, I. S. *Topics in Molecular and Structural Biology*, MacMillan Press, 1983.

31. Wickstrom, E. *Prospects for Antisense Nucleic Acid Therapy of Cancer and AIDS*, 1991.

32. Knorre, D. G., Vlassov, V. V., Zarytova, V. F., Lebedev, A. V., Fedorova, O. S. *Design and Targeted Reactions of Oligonucleotide Derivatives*, CRC Press, 1994.

33. Knorre, D. G., Vlassov, V. V. *Progress in Nucleic Acids Research*, Academic Press, 1985, 291-320.

34. Belikova, A. M., Zarytova, V. F., Grineva, N. I. *Tet. Lett.*, 1967, **37**, 3557-3562

35. Summerton, J. *J. Theor. Biol.*, 1979, **78**, 77-99.

36. Salganik, R. I., Dianov, G. L., Ovchinnokova, L. P., Voronina, E. N., Kokoza, E. B., Mazin, A. V. *Proc. Natl. Acad. Sci. USA*, 1998, **77**, 2796-2800.

37. Crooke, S. T., Lebleu, B. *Antisense Research and Applications*, CRC Press, 1993.

38. Povsic, T. J. Dervan, P. B. *J. Am. Chem. Soc.*, 1990, **112**, 9428-9430.

39. Webb, T. R. Matteucci, M. D. *J. Am. Chem. Soc.*, 1986, **108**, 2764-2765.

40. Vlassov, V. V., Zarytova, V. F., Kutiavin, I. V., Mamev, S. V., Podyminogin, M. A. *Nucl. Acids Res.*, 1986, **14**, 4065-4076.

41. Dreyer, G. B. Dervan, P. B. *Proc. Natl. Acad. Sci. USA*, 1985, **82**, 968-972.

42. Boutorin, A. S., Vlassov, V. V., Kazakov, S. A., Kutiavin, I. V., Podyminogin, M. A. *FEBS Lett.*, 1984, **172**, 43-46.

43. Stobel, S. A. Dervan, P. B. *J. Am. Chem. Soc.*, 1989, **111**, 7286-7287.

44. Griffin, L. C. Dervan, P. B. *Science*, 1989, **245**, 967-971.

45. Frolova, E. I., Ivanova, E. M., Zarytova, V. F., Abramova, T. V., Vlassov, V. V. *FEBS Lett.*, 1990, **269**, 101-104.

46. Le Doan, T., Perrouault, L., Helene, C. *Biochemistry*, 1986, **25**, 6736-6739.

47. Praseuth, D., Le Doan, T., Chassingnol, M., Decout, J., Habhoub, N., Lhomme, J., Thuong, N. T., Helene, C. *Biochemistry*, 1988, **27**, 3031-3038.

48. Le Doan, T., Perrouault, L., Praseuth, D., Habhoub, N., Decout, J., Thuong, N. T., Lhomme, J., Helene, C. *Nucl. Acids Res.*, 1987, **15**, 7749-7760.

49. Shi, Y. Hearst, J. E. *Biochemistry*, 1987, **26**, 3792-3798.

50. Pieles, U. Englisch, U. *Nucl. Acids Res.*, 1989, **17**, 285-299.

51. Shi, Y., Gamper, H., Hearst, J. E. *Nucl. Acids Res.*, 1987, **15**, 6843-6854.

52. Perelroyzen, M. P. Vologodskii, A. V. *Nucl. Acids Res.*, 1988, **16**, 4693-4704.

53. Dobrikov, M. I., Gaidamakov, S. A., Koshkin, A. A., Lukyanchuk, N. P., Shishkin, G. V., Vlassov, V. V. *Doklady Akademii Nauk*, 1995, **344**, 122-125.

54. Dobrikov, M. I., Gaidamakov, S. A., Koshkin, A. A., Shishkin, G. V., Vlassov, V. V. *Dokl. Acad. Nauk. SSSR*, 1996, **351**, 547-550.

55. Vlassov, V. V., Dobrikov, M. I., Gaidamakov, S. A., Gaidamakova, E. K., Gainutdinov, T. I., Koshkin, A. A. *DNA and RNA Cleavers and Chemotherapy of Cancer and Viral Diseases*, Kluwer, Dordecht, 1996, 195-207.

56. Vlassov, V. V., Abramova, T., Godovikova, T., Giege, R., Silnikov, V. *Antisense Nucl. Drug Dev.*, 1997, 7, 39-42.

57. Dobrikov, M. I., Gaidamakov, S. A., Gainutdinov, T. I., Koshkin, A. A., Vlassov, V. V. *Antisense Nucl. Drug Dev.*, 1997, **7**, 309-317.

58. Vlassov, V. V., Dobrikov, M. I., Gaidamakov, S. A., Gaidamakova, E. K., Gainutdinov, T. I., Koshkin, A. A. *Binary systems of oligonucleotide conjugates for sequence specific energy-transfer sensitised photomodification of nucleic acids*, Kluwer, Dordecht, 1996.

59. Dobrikov, M. I., Gaidamakov, S. A., Gainutdinov, T. I., Tenetova, E. D., Shishkin, G. V., Vlassov, V. V. *Dokl. Acad. Nauk. SSSR*, 1998, **358**, 403-407.

60. Bichenkova, E. V., Marks, D. S., Lokhov, S. G., Dobrikov, M. I., Vlassov, V. V., Douglas, K. T. *J. Biomol. Struct. Dynam.*, 1997, **15**, 307-320.

61. Chu, B. C. F. Orgel, L. E. *Proc. Natl. Acad. Sci. USA*, 1985, **82**, 963-967.

62. Breslow, R. *Acc. Chem. Res.* 1991, **24**, 317-324.

63. Cech, T. R. *Science*, 1987, **236**, 1532-1539.

64. Altman, S. *Angew. Chem. Int. Ed. Engl.*,1990, **29**, 749-758.

65. Bartel, D. P. Szostak, J. W. *Science*, 1993, **261**, 1411-1418.

66. Magda, D., Miller, R. A., Sessler, J. L., Iverson, B. L. *J. Am. Chem. Soc.*, 1994, **116**, 7439-7440.

67. Bashkin, J. K., Frolova, E. I., Sampath, U. *J. Am. Chem. Soc.*, 1994, **116**, 5981-5982.

68. Komiyama, M. Inokawa, T. *J. Biochem.*, 1994, **116**, 719-720.

69. Tung, C., Wei, Z., Leibowitz, M. J., Stein, S. *Proc. Natl. Acad. Sci. USA*, 1992, **89**, 7114-7118.

70. Wlodawer, A., Bott, R., Sjolin, L. *J. Biol. Chem.*, 1982, **257**, 1325-1332.

71. Wyckoff, H. W., Tsernoglou, D., Hanson, A. W., Knox, J. R., Lee, B., Richards, F. M. *J. Biol. Chem*, 1970, **245**, 305-328.

72. Deakyne, C. A. Allen, L. C. *J. Am. Chem. Soc.*, 1979, **101**, 3951-3959.

73. Breslow, R. Labelle, M. *J. Am. Chem. Soc.*, 1986, **108**, 2655-2659.

74. Podyminogin, M. A., Vlassov, V. V., Giege, R. *Nucl. Acids Res.*, 1993, **21**, 5950-5956.

75. Vlassov, V. V., Zuber, G., Felden, B., Behr, J., Giege, R. *Nucl. Acids Res.*, 1995, **23**, 3161-3167.

76. Silnikov, V., Zuber, G., Behr, J., Giege, R., Vlassov, V. *Phosphorus, Sulfur and Silicon*, 1996, **109-110**, 277-280.

77. Smith, J., Ariga, K., Anslyn, E. V. *J. Am. Chem. Soc.*, 1993, **115**, 362-364.

78. Yurchenko, L., Silnikov, V., Godovikova, T., Shishkin, G., Toulme, J., Vlassov, V. *Nucleosides and Nucleotides*, 1997, **16**, 1721-1725.

79. Reynolds, M. A., Beck, T. A., Say, P. B., Schwartz, D. A., Dwyer, B. P., Daily, W. J., Vaghefi, M. M., Metzler, M. D., Klem, R. E., Arnold, L. J. *Nucl. Acids Res.*, 1996, **24**, 760-765.

80. Vlassov, V., Abramova, T., Godovikova, T., Giege, R., Silnikov, V. *Antisense Nucl. Drug Dev.*, 1997, **7**, 39-42.

81. Konevetz, D. A., Beck, I. E., Beloglazova, N. G., Sulimenkov, I. V., Silnikov, V. N., Zenkova, M. A., Shishkin, G. V., Vlassov, V. V. *Tetrahedron*, 1999, **55**, 503-512.

82. Schmidt, S. J., Serianni, A. S., Finley, J. W. in *Applications of NMR in Agriculture and Biochemistry*, Plenum Press, 1990, **56**, 1-6.

83. Gmeiner, W. H., Konerding, D., James, T. L. *Biochemistry*, 1999, **38**, 1166-1175.

84. Wojciak, J. M., Connolly, K. M., Clubb, R. T. *Nature Struct. Biol.*, 1999, **6**, 366-373.

85. Cheong, H. K., Cheong, C., Lee, Y. S., Seong, B. L., Choi, B. S. *Nucl. Acids Res.*, 1999, **27**, 1392-1397.

86. Rahman, A., Choudhary, M. I. *Solving Problems with NMR Spectroscopy*, Academic Press, 1996.

87. Gunther, H. *NMR Spectroscopy*, Wiley, 1992.

88. Meiboom, S. Gill, D. *The Review of Scientific Instruments*, 1958, **29**, 688-691.

89. Carr, H. Y. Purcell, E. M. *Physical Review*, 1954, **94**, 630-638.

90. Wagner, G. Wüthrich, K. *J. Mol. Biol.*, 1982, **155**, 347-366.

91. Wider, G., Lee, K. H., Wüthrich, K. *J. Mol. Biol.*, 1982, **155**, 367-388.

92. Wüthrich, K. *NMR of Proteins and Nuclei Acids*, Wiley, 1986.

93. Roberts, G. C. K. *NMR of Macromolecules: A Practical Approach*, Oxford University Press, 1993.

94. Bradbury, E. M., Nicolini, C. *NMR in the Life Sciences*, Plenum Press, 1985.

95. Patel, D. J., Kozlowski, S. A., Marky, L. A., Broka, C., Rice, J. A., Itakura, K., Breslauer, K. J. *Biochemistry*, 1982, **21**, 428-436.

96. Johnston, P. D. Redfield, A. G. *Biochemistry*, 1981, **20**, 1147-1156.

97. Morris, G. A., Silveston, A. C. T., Waterton, J. C. *J. Magn. Reson.*, 1989, **81**, 641-645.

98. Soloman, I. *Physical Review*, 1955, **99**, 559-565.

99. Gronenborn, A. M. Clore, G. M. *Progress in NMR Spectroscopy*, 1985, **17**, 1-32.

100. Keepers, J. W. James, T. L. *J. Magn. Reson.*,1984, **57**, 404-426.

101. CORMA (Complete Relaxation Matrix Analysis), Borgias, B. A., Thomas, P. D., James, T. L. 1987,

102. Borgias, B. A. James, T. L. *J. Magn. Reson.*, 1988, **79**, 493-512.

103. Mirau, P. A. *J. Magn. Reson.*, 1988, **80**, 439-447.

104. Olejniczak, E. T., Gampe, R. T., Fesik, S. W. *J. Magn. Reson.* 1986, **67**, 28-41.

105. Boelens, R., Koning, T. M. G., van der Marel, G. A., van Boom, J. H., Kaptein, R. *J. Magn. Reson.*, 1989, **82**, 290-308.

106. Boelens, R., Koning, T. M. G., Kaptein, R. *J. Mol. Struct.*, 1988, **173**, 299-311.

107. James, T. L., Borgias, B. A., Bianucci, A. M., Zhou, N. *NMR Applications in Biopolymers*, Plenum Press, 1990, 135-154.

108. Borgias, B. A. James, T. L. *J. Magn. Reson.* 1990, **87**, 475-487.

109. Gmeiner, W. H. Sahasrabudhe, P. V. *Biochemistry*, 1997, **36**, 5981-5991.

110. Spielmann, H. P., Wemmer, D. E., Jacobsen, J. P. *Biochemistry*, 1995, **34**, 8542-8553.

111. Stolarski, R., Egan, W., James, T. L. *Biochemistry*, 1992, **31**, 7027-7042.

112. Parkinson, J. A., Ebrahimi, S. E., McKie, J. H., Douglas, K. T. *Biochemistry*, 1994, **33**, 8442-8452.

113. Bax, A. Davis, D. G. *J. Magn. Reson.*, 1985, **65**, 355-360.

114. Nagayama, K., Kumar, A., Wüthrich, K., Ernst, R. R. *J. Magn. Reson.*, 1980, **40**, 321-334.

115. Jeener, J., Meier, B. H., Backmann, P., Ernst, R. R. *J. Chem. Phys.*, 1979, **71**, 4546-4553.

116. Macura, S., Huang, Y., Suter, D., Ernst, R. R. *J. Magn. Reson.*, 1981, **43**, 259-281.

117. States, D. J., Haberkorn, R. A., Ruben, D. J. *J. Magn. Reson.*, 1982, **48**, 286-292.

118. Arnott, S. Hukins, D. W. L. *J. Mol. Biol.*, 1973, **81**, 93-105.

119. Stewart, J. J. P. *J. Comp. Aided Mol. Design*, 1990, **4**, 1-105.

120. Borgias, B. A. James, T. L. *Methods Enzymol.* 1989, **176**, 169-183.

121. Suzuki, E., Pattabiraman, N., Zon, G., James, T. L. *Biochemistry*, 1986, **25**, 6854-6865.

122. Coppel, Y., Berthet, N., Coulambeau, C., Coulambeau, C., Garcia, J., Lhomme, J. *Biochemistry*, 1997, **36**, 4817-4830.

123. Baleja, J. D., Pon, R. T., Sykes, B. D. *Biochemistry*, 1990, **29**, 4828-4839.

124. Gronenborn, A. M. Clore, G. M. *Biochemistry*, 1989, **28**, 5978-5984.

125. Lavery, R. Sklenar, H. *J. Biomol. Struct. Dynam.* 1988, **6**, 63-91.

126. Lavery, R. Sklenar, H. *J. Biomol. Struct. Dynam.* 1989, **6**, 655-667.

127. CURVES 3.0, Helical analysis of irregular nucleic acids. Lavery, R., Sklenar, H. 1990,

128. Hare, D. R., Wemmer, D. E., Chou, S., Drobny, G., Reid, B. R. *J. Mol. Biol.*, 1983, **171**, 319-336.

129. Lee, S. J., Akutsu, H., Kyogoku, Y., Kitano, K., Tozuka, Z., Ohta, A., Ohtsuka, E., Ikehara, M. *J. Biochem.*, 1985, **98**, 1463-1472.

130. Boelens, R., Scheek, R. M., Dijkstra, K., Kaptein, R. *J. Magn. Reson.*, 1985, **62**, 378-386.

131. Finkelstein, A. V. Reva, B. A. *Nature*, 1991, **351**, 497-499.

132. Chan, H. S. Dill, K. A. *Physics Today*, 1993, 24-32.

133. Kobayashi, Y., Sasabe, H., Saito, N. *Fluid Phase Equilibria*, 1998, **144**, 403-413.

134. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., Tasumi, M. *Biochem. J.*, 1977, **122**, 535

135. Jones, D. T. *Proteins: Structure, Function and Genetics*, 1997, **S1**, 185-191.

136. Fanelli, F., Menziani, C., Scheer, A., Cotecchia, S., DeBenedetti, P. G. *Methods-A Companion to Methods in Enzymology*, 1998, **14**, 302-317.

137. Dandekar, T. Argos, P. *Int. J. Biol. Macromol.*, 1996, **18**, 1-4.

138. Dandekar, T. *J. Mol. Model.*, 1996, **2**, 304-306.

139. Westhead, D. R. Thornton, J. M. *Curr. Opin. Struct. Biol.*, 1998, **9**, 383-389.

140. Chou, P. Y. Fasman, G. D. *Biochemistry*, 1974, **13**, 211-245.

141. Garnier, J., Osguthorpe, D. J., Robson, B. *J. Mol. Biol.*, 1978, **120**, 97-120.

142. Levin, J. M. Garnier, J. *Biochim. Biophys. Acta*, 1988, **955**, 283-295.

143. Yi, T.-M. Lander, E. S. *J. Mol. Biol.*, 1993, **232**, 1117-1129.

144. Rost, B. Sander, C. *J. Mol. Biol.*, 1993, **232**, 584-599.

145. Stolorz, P., Lapedes, A., Xia, Y. *J. Mol. Biol.*, 1992, **225**, 363-377.

146. Chandonia, J.-M. Karplus, M. *Prot. Sci.*, 1995, **4**, 275-285.

147. Rost, B. *Methods Enzymol.*, 1996, **266**, 524-539.

149. Salamov, A. A. Soloyev, V. V. *J. Mol. Biol.*, 1997, **268**, 31-36.

150. Lim, V. I. *J. Mol. Biol.*, 1974, **88**, 873-894.

151. Cohen, F. E., Kosen, P. A., Kuntz, I. D., Epstein, L. B., Ciardelli, T. L., Smith, K. A. *Science*, 1986, **234**, 349-352.

152. Wako, H. Blundell, T. L. *J. Mol. Biol.*, 1994, **238**, 693-708.

153. Finkelstein, A. V. Ptitsyn, O. B. *Prog. Biophys. Mol. Biol.*, 1987, **50**, 171-190.

154. Chothia, C. *Nature*, 1992, **357**, 543-544.

155. Russell, R. B., Saqi, M. A. S., Bates, P. A., Sayle, R. A., Sternberg M J E. *Prot. Eng.*, 1998, **11**, 1-9.

156. Bowie, J. U., Luthy, R., Eisenberg, D. *Science*, 1991, **253**, 164-170.

157. Gochin, M. James, T. L. *Biochemistry*, 1990, **29**, 11172-11180.

158. Dandekar, T. Argos, P. *J. Mol. Biol.*, 1996, **256**, 645-660.

159. Jones, D. T., Taylor, W. R., Thornton, J. M. *Nature*, 1992, **358**, 86-89.

160. Jones, D. Thornton, J. *J. Comp. Aided Mol. Design*, 1993, 7, 439-456.

161. Rykunov, D. S., Lobanov, M. Y., Finkelstein, A. V. *Mol. Biol.*, 1998, **32**, 428-438.

162. Needleman, S. B. Wunsch, C. D. *J. Mol. Biol.*, 1970, **48**, 443-453.

163. Sellers, P. H. *Bull. Math. Biol.*, 1984, **46**, 501-514.

164. Goad, W. B. Kanehisa, M. I. *Nucl. Acids Res.*, 1982, **10**, 247-263.

165. Dayhoff, M. O., Schwatz, R. M., Drutt, B. C. *Atlas of Protein Sequence and Structure*, 1978, **5**, Suppl 3, 345-352.

166. Henikoff, S. Henikoff, J. G. *Proc. Natl. Acad. Sci. USA*, 1992, **89**, 10915-10919.

167. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. *J. Mol. Biol.*, 1990, **215**, 403-410.

168. Karlin, S. Altschul, S. F. *Proc. Natl. Acad. Sci. USA*, 1990, **87**, 2264-2268.

169. Corpet, F. *Nucl. Acids Res.*, 1988, **16**, 10881-10890.

170. Singh, S. B., Ajay, Wemmer, D. E., Kollman, P. A. *Proc. Natl. Acad. Sci. USA*, 1994, **91**, 7673-7677.

171. Briffeuil, P., Baudoux, G., Lambert, C., DeBolle, X., Vinals, C., Feytmans, E., Depiereux, E. *Bioinformatics*, 1998, **14**, 357-366.

172. Jones, T. A. Thirup, S. *EMBO J.*, 1986, **5**, 819-822.

173. Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E., Thornton, J. M. *Nature*, 1987, **326**, 347-352.

174. Blundell, T. L., Carney, D., Gardner, S., Hayes, F., Howlin, B., Hubbard, T., Overington, J., Singh, R. A., Sibanda, B. L., Sutcliffe, M. *Eur. J. Biochem.*, 1988, **172**, 513-520.

175. Go, N. Scheraga, H. A. *Macromolecules*, 1969, **3**, 178-187.

176. Bruccoleri, R. Karplus, M. *Macromolecules*, 1985, **18**, 2767-2773.

177. Lovell, S. C., Word, J. M., Richardson, J. S., Richardson, D. C. *Prot. Sci.*, 1998, **7**, 80 (148T).

178. Havel, T. F. Snow, M. E. *J. Mol. Biol.*, 1991, **217**, 1-7.

179. Sali, A. Blundell, T. L. *J. Mol. Biol.*, 1993, **234**, 779-815.

180. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., Karplus, M. *J. Comp. Chem.*, 1983, **4**, 187-217.

181. Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Weiner, P. K. *J. Am. Chem. Soc.*, 1984, **106**, 765.

182. Ramachandran, G. N. *Current Contents/Life Sciences*, 1981, **10**, 18

183. Luthy, R., Bowie, J. U., Eisenberg, D. *Nature*, 1992, **356**, 83-85.

184. Wibley, J. E. A. *PhD Thesis: DNA binding proteins: Structures and predictions*, 1995.

185. Buchanan, M. *New Scientist*, 1998, **160**, 42-46.

186. Wolffe, A. P. *Cell*, 1994, **77**, 13-16.

187. Lodish, H., Baltimore, D., Berk, A., Zipursky, S. L., Matsudaira, P., Darnell, J. *Molecular Cell Biology*, Freeman Press, 1995.

188. Latchman, D. S., *Eukaryotic Transcription Factors*, Academic Press, 1991.

189. Fairman, R., Beran-Steed, R. K., Anthony-Cahill, S. J., Lear, J. D., Stafford, W. F.,III, DeGrado, W. F., Benfield, P. A., Brenner, S. L. *Proc. Natl. Acad. Sci. USA*, 1993, **90**, 10429-10433.

190. Laue, T. M., Starovasnik, M. A., Weintraub, H., Sun, X.-H., Snider, L., Klevit, R. E. *Proc. Natl. Acad. Sci. USA*, 1995, **92**, 11824-11828.

191. Mitsui, K., Shirakata, M., Paterson, B. M. *J. Biol. Chem.*, 1993, **268**, 24415-24420.

192. Baudier, J., Bergeret, E., Bertacchi, N., Weintraub, H., Gagnon, J., Garin, J. *Biochemistry*, 1995, **34**, 7834-7846.

193. Hermann, S., Saarikettu, J., Onions, J., Hughes, K., Grundstrom, T. *Cell Calcium*, 1998, **23**, 135-142.

194. Blackwell, T. K. Weintraub, H. *Science*, 1990, **250**, 1104-1110.

195. Murre, C., Bain, G., van Dijk, M. A., Engel, I., Furnari, B. A., Massari, M. E., Matthews, J. R., Quong, M. W., Rivera, R. R., Stuiver, M. H. *Biochim. Biophys. Acta*, 1994, **1218**, 129-135.

196. Murre, C., McCaw, P. S., Vaessin, H., Caudy, M., Jan, L. Y., Jan, Y. N., Cabrera, C. C., Bushkin, J. N., Hauschka, S. D., Lassar, A. B., Weintraub, H., Baltimore, D. *Cell*, 1989, **58**, 537-544.

197. Church, G. M., Ephrussi, A., Gilbert, W., Tonegawa, S. *Nature*, 1985, **313**, 798-801.

198. Ephrussi, A., Church, G. M., Tonegawa, S., Gilbert, W. *Science*, 1985, **227**, 134-140.

199. Martin, K. *BioEssays*, 1991, **13**, 499-503.

200. Rivera, R. R., Stuiver, M. H., Steenbergen, R., Murre, C. *Mol. Cell. Biol.*, 1993, **13**, 7163-7169.

201. Sieweke, M. H., Tekotte, H., Jarosch, U., Graf, T. *EMBO J.*, 1998, **17**, 1728-1739.

202. German, M. S., Wang, J., Chadwick, R. B., Rutter, W. J. *Genes & Dev.*, 1992, **6**, 2165-2176.

203. Ebert, B. L. Bunn, H. F. *Mol. Cell. Biol.*, 1998, **18**, 4089-4096.

204. Arany, Z., Huang, L. E., Eckner, R., Bhattacharya, S., Jiang, C., Goldberg, M. A., Bunn, H. F., Livingston, D. M. *Proc. Natl. Acad. Sci. USA*, 1996, **93**, 12969-12973.

205. Huang, L. E., Ho, V., Arany, Z., Krainc, D., Galson, D., Tendler, D., Livingston, D. M., Bunn, H. F. *Kidney Int.*, 1997, **51**, 548-552.

206. Peters, M. A. Taparowsky, E. J. *Critic. Rev. Euk. Gene Exp.*, 1998, **8**, 277-296.

207. Jackson, F. R., Bargiello, T. A., Yun, S.-H., Young, M. W. *Nature*, 1986, **320**, 185-188.

208. Nambu, J. R., Lewis, J. O., Wharton, K. A. J., Crews, S. T. *Cell*, 1991, **67**, 1157-1167.

209. Wood, S. M., Gleadle, J. M., Pugh, C. W., Hankinson, O., Ratcliffe, P. J. *J. Biol. Chem.*, 1996, **271**, 15117-15123.

210. Hoffman, E. C., Reyes, H., Chu, F.-F., Sander, F., Conley, L. H., Brooks, B. A., Hankinson, O. *Science*, 1991, **252**, 954-958.

211. Hankinson, O. *Annu. Rev. Pharmacol. Toxicol.*, 1995, **35**, 307-340.

212. Pongratz, I., Antonsson, C., Whitelaw, M. L., Poellinger, L. *Mol. Cell. Biol.*, 1998, **18**, 4079-4088.

213. Semenza, G. L. Wang, G. L. *Mol. Cell Biol.*, 1992, **12**, 5447-5454.

214. Murre, C., McCaw, P. S., Baltimore, D. *Cell*, 1989, **56**, 777-783.

215. Henthorn, P., Kiledjian, M., Kadesch, T. *Science*, 1990, **247**, 467-470.

216. Corneliussen, B., Thornell, A., Hallberg, B., Grundstrom, T. *J. Virol.* 1991, **65**, 6084-6093.

217. Hu, J.-S., Olson, E. N., Kingston, R. E. *Mol. Cell. Biol.*, 1992, **12**, 1031-1042.

218. Zhang, Y., Babin, J., Feldhaus, A. L., Singh, H., Sharp, P. A., Bina, M. *Nucl. Acids Res.*, 1991, **19**, 4555

219. Bain, G., Gruenwald, S., Murre, C. *Mol. Cell Biol.*, 1993, **13**, 3522-3529.

220. Cronmiller, C., Schedl, P., Cline, T. W. *Genes & Dev.*, 1988, **2**, 1666-1676.

221. Davis, R. L., Weintraub, H., Lassar, A. B. *Cell*, 1987, **51**, 987-1000.

222. Piette, J., Bessereau, J.-L., Huchet, M., Changeux, J.-P. *Nature*, 1990, **345**, 353-355.

223. Braun, T., Buschhausen-Denker, G., Bober, E., Tannich, E., Arnold, H. H. *EMBO J.*, 1989, **8**, 701-709.

224. Braun, T., Rudnicki, M. A., Arnold, H. H., Jaenisch, R. *Cell*, 1992, **71**, 369-382.

225. Braun, T., Bober, E., Winter, B., Rosenthal, N., Arnold, H. H. *EMBO J.*, 1990, **9**, 821-831.

226. Rhodes, S. J. Konieczny, S. F. *Genes & Dev.*, 1989, **3**, 2050-2061.

227. Edmondson, D. G. Olson, E. N. *Genes & Dev.*, 1989, **3**, 628-640.

228. Wright, W. E., Sassoon, D. A., Lin, V. K. *Cell*, 1989, **56**, 607-617.

229. Li, L. Olson, E. N. *Adv. Can. Res.*, 1992, **58**, 95-119.

230. Weintraub, H. *Cell*, 1993, **75**, 1241-1244.

231. Rudnicki, M. A., Braun, T., Hinuma, S., Jaenisch, R. *Cell,* 1992, **71**, 383-390.

232. Villares, R. Cabrera, C. V. *Cell,* 1987, **50**, 415-424.

233. Jan, Y. N. Jan, L. Y. *Cell,* 1993, **75**, 827-830.

234. Caudy, M., Vassin, H., Brand, M., Tuma, R., Jan, L. Y., Jan, Y. N. *Cell,* 1988, **55**, 1061-1067.

235. Ellis, H. M., Spann, D. R., Posakony, J. W. *Cell,* 1990, **61**, 27-38.

236. Jarman, A. P., Grau, Y., Jan, L. Y., Jan, Y. N. *Cell,* 1993, **73**, 1307-1321.

237. Johnson, J. E., Birren, S. J., Anderson, D. J. *Nature,* 1990, **346**, 858-861.

238. Littlewood, T. D. Evan, G. I. *Protein Profile,* 1994, **1**, 639-709.

239. Tamimi, R., Steingrimsson, E., Copeland, N. G., Dyer-Montgomery, K., Lee, J. E., Hernandez, R., Jenkins, N. A., Tapscott, S. J. *Genomics,* 1996, **34**, 418-421.

240. McCormick, M. B., Tamimi, R., Snider, L., Asakura, A., Bergstrom, D., Tapscott, S. J. *Mol. Cell. Biol.,* 1996, **16**, 5792-5800.

241. Lee, J. E. *Curr. Opin. Neur.,* 1997, 7, 13-20.

242. Garrell, J. Campuzano, S. *Bioessays,* 1991, **13**, 493-498.

243. Erickson, J. W. Cline, T. W. *Science,* 1991, **251**, 1071-1074.

244. Parkhurst, S. M., Bopp, D., Ish-Horowicz, D. *Cell,* 1990, **63**, 1179-1191.

245. Parkhurst, S. M., Lipshitz, H. D., Ish-Horowicz, D. *Development,* 1993, **117**, 737-749.

246. Erickson, J. W., Cline, T. W. *Genes & Dev.,* 1998, 7, 1688-1702.

247. Alt, F. W., DePinho, R., Zimmerman, K., Legouy, E., Hatton, K., Ferrier, P., Tesfaye, A., Yancopoulis, G., Nisen, P. *Cold Spring Harbor Symp. Quant. Biol.,* 1986, **LI**, 931-941.

248. Marcu, K. B., Bossone, S. A., Patel, A. J. *Annu. Rev. Biochem.,* 1992, **61**, 809-860.

249. Amati, B., Brooks, M. W., Levy, N., Littlewood, T. D., Evan, G. I., Land, H. *Cell,* 1993, **72**, 233-245.

250. Blackwood, E. M. Eisenman, R. N. *Science,* 1991, **1211**, 1217

251. Ayer, D. E., Kretzner, L., Eisenman, R. N. *Cell,* 1993, **72**, 211-222.

252. Zervos, A. S., Gyuris, J., Brent, R. *Cell,* 1993, **72**, 223-232.

253. Hurlin, P. J., Queva, C., Koskinen, P. J., Steingrimsson, E., Ayer, D. E., Copeland, N. G., Jenkins, N. A., Eisenman, R. N. *EMBO J.,* 1995, **14**, 5646-5659.

254. Gupta, K., Anand, G., Yin, X., Grove, L., Prochownik, E. V. *Oncogene,* 1998, **16**, 1149-1159.

255. Gregor, P. D., Sawadogo, M., Roeder, R. G. *Genes & Dev.,* 1990, **4**, 1730-1740.

256. Carr, C. S. Sharp, P. A. *Mol. Cell. Biol.*, 1990, **10**, 4384-4388.

257. Beckmann, H., Su, L.-K., Kadesch, T. *Genes & Dev.*, 1990, **4**, 167-179.

258. Hu, Y.-F., Luscher, B., Admon, A., Mermod, N., Tjian, R. *Genes & Dev.*, 1990, **4**, 1741-1752.

259. Rushlow, C. A., Hogan, A., Pinchin, S. M., Howe, K. M., Lardelli, M., Ish-Horowicz, D. *EMBO J.*, 1989, **8**, 3095-3103.

260. Klambt, C., Knust, E., Tietze, K., Campos Ortega, J. A. *EMBO J.*, 1989, **8**, 203-210.

261. Ishibashi, M., Sasai, Y., Nakanishi, S., Kageyama, R. *Eur. J. Biochem.*, 1993, **215**, 645-652.

262. Sasai, Y., Kageyama, R., Tagawa, Y., Shigemoto, R., Nakanishi, S. *Genes & Dev.*, 1992, **6**, 2620-2634.

263. Langlands, K., Yin, X., Anand, G., Prochownik, E. V. *J. Biol. Chem.*, 1997, **272**, 19785-19793.

264. Benezra, R., Davis, R. L., Lockshon, D., Turner, D. L., Weindraub, H. *Cell*, 1990, **61**, 49-59.

265. Biggs, J., Murphy, E. V., Israel, M. A. *Proc. Natl. Acad. Sci. USA*, 1992, **89**, 1512-1516.

266. Kawaguchi, N., DeLuca, H. F., Noda, M. *Proc. Natl. Acad. Sci. USA*, 1992, **89**, 4569-4572.

267. Jen, Y., Weintraub, H., Benezra, R. *Genes & Dev.*, 1992, **6**, 1466-1479.

268. Kreider, B. L., Benezra, R., Rovera, G., Kadesch, T. *Science*, 1992, **255**, 1700-1702.

269. Murray, S. S., Glackin, C. A., Winters, K. A., Gazit, D., Kahn, A. J., Murray, E. J. *J. Bone Min. Res.*, 1992, **7**, 1131-1138.

270. Barone, M. V., Pepperkok, R., Peverali, F. A., Philipson, L. *Proc. Natl. Acad. Sci. USA*, 1994, **91**, 4985-4988.

271. Hara, E., Yamaguchi, T., Nojima, H., Ide, T., Campisi, J., Okayama, H., Oda, K. *J. Biol. Chem.*, 1994, **269**, 2139-2145.

272. Shoji, W., Inoue, T., Yamamoto, T., Obinata, M. *J. Biol. Chem.*, 1995, **270**, 24818-24825.

273. Anand, G., Yin, X., Shahidi, A. K., Grove, L., Prochownik, E. V. *J. Biol. Chem.*, 1997, **272**, 19140-19151.

274. Chen, B. Lim, R. W. *J. Biol. Chem.*, 1997, **272**, 2459-2463.

275. Lister, J., Forrester, W. C., Baron, M. H. *J. Biol. Chem.*, 1995, **270**, 17939-17946.

276. Lister, J. Baron, M. H. *Gene Expression*, 1998, 7, 25-38.

277. Desprez, P.-Y., Hara, E., Bissell, M. J., Campisi, J. *Mol. Cell. Biol.*, 1995, **15**, 3398-3404.

278. Sablitzky, F., Moore, A., Bromley, M., Deed, R. W., Norton, J. D. *Cell Growth and Differentiation*, 1998, **9**, 1015-1024.

279. Deed, R. W., Jasiok, M., Norton, J. D. *Biochim. Biophys. Acta*, 1994, **1219**, 160-162.

280. Ebrahimi, S. E. S., Bibby, M. C., Fox, K. R., Douglas, K. T. *Anti-Cancer Drug Design*, 1995, **10**, 463-479.

281. Christy, B. A., Sanders, L. K., Lau, L. F., Copeland, N. G., Jenkins, N. A., Nathans, D. *Proc. Natl. Acad. Sci. USA*, 1991, **88**, 1815-1819.

282. Deed, R. W., Bianchi, S. M., Atherton, G. T., Johnston, D., Santibanez-Koref, M., Murphy, J. J., Norton, J. D. *Oncogene*, 1993, **8**, 599-607.

283. Reichmann, V., Van Crüchten, I., Sablitzky, F. *Nucl. Acids Res.*, 1994, **22**, 749-755.

284. Surovaya, A. N. Trubitsin, S. N. *Mol. Biol.*, 1974, **7**, 403-410.

285. Asp, J., Thornemo, M., Inerot, S., Lindahl, A. *FEBS Lett.*, 1998, **438**, 85-90.

286. Pongubala, J. M. R. Atchison, M. L. *Mol. Cell. Biol.*, 1991, **11**, 1040-1047.

287. Cordle, S. R., Henderson, E., Masuoka, H., Weil, P. A., Stein, R. *Mol. Cell. Biol.*, 1991, **11**, 1734-1738.

288. Feve, B., Moldes, M., El Hadri, K., Lasnier, F., Pairault, J. *M S-Medecine Sciences*, 1998, **14**, 848-857.

289. Florio, M., Hernandez, M.-C., Yang, H., Shu, H.-K., Cleveland, J. L., Israel, M. A. *Mol. Cell. Biol.*, 1998, **18**, 5435-5444.

290. Antoch, M. P., Song, E.-J., Chang, A.-M., Vitaterna, M. H., Zhao, Y., Wilsbacher, L. D., Sangoram, A. M., King, D. P., Pinto, L. H., Takahashi, J. S. *Cell*, 1997, **89**, 655-667.

291. Lasorella, A., Iavarone, A., Israel, M. A. *Mol. Cell. Biol.*, 1996, **16**, 2570-2578.

292. Iavarone, A., Garg, P., Lasorella, A., Hsu, J., Israel, M. A. *Genes & Dev.*, 1994, **8**, 1270-1284.

293. Kleeff, J., Ishiwata, T., Friess, H., Buchler, M. W., Israel, M. A., Korc, M. *Cancer Res.*, 1998, **58**, 3769-3772.

294. Sablitzky, F., van Cruechten, I., Cinato, E., Newton, J. S. *Developmental Biology*, 1997, **186**, A218.

295. King, D. P., Zhao, Y., Sangoram, A. M., Wilsbacher, L. D., Tanaka, M., Antoch, M. P., Steeves, T. D. L., Vitaterna, M. H., Kornhauser, J. M., Lowrey, P. L., Turek, F. W., Takahashi, J. S. *Cell*, 1997, **89**, 641-653.

296. Duncan, D. M., Burgess, E. A., Duncan, I. *Genes & Dev.*, 1998, **12**, 1290-1303.

297. Hirose, K., Morita, M., Ema, M., Mimura, J., Hamada, H., Fujii, H., Saijo, Y., Gotoh, O., Sogawa, K., Fujii-Kuriyama, Y. *Mol. Cell. Biol.*, 1996, **16**, 1706-1713.

298. Wilk, R., Weizman, I., Shilo, B.-Z. *Genes & Dev.*, 1996, **10**, 93-102.

299. Nambu, J. R., Chen, W., Hu, S., Crews, S. T. *Gene*, 1996, **172**, 249-254.

300. Moffett, P., Reece, M., Pelletier, J. *Mol. Cell. Biol.*, 1997, **17**, 4933-4947.

301. Ikeda, M. Nomura, M. *Biochem. Biophys. Res. Comm.*, 1997, **233**, 258-264.

302. Hogenesch, J. B., Gu, Y.-Z., Jain, S., Bradfield, C. A. *Proc. Natl. Acad. Sci. USA*, 1998, **95**, 5474-5479.

303. Zhou, Y. D., Barnard, M., Tian, H., Li, X., Ring, H. Z., Franke, U., Shelton, J., Richardson, J., Russell, D. W., McKnight, S. L. *Proc. Natl. Acad. Sci. USA*, 1997, **94**, 713-718.

304. Flamme, I., Frohlich, T., Reutern, M. V., Kappel, A., Damert, A., Risau, W. *Mechanisms of Development*, 1997, **63**, 51-60.

305. Ema, M., Taya, S., Yokotani, N., Sogawa, K., Matsuda, Y., Fujii-Kuriyama, Y. *Proc. Natl. Acad. Sci. USA*, 1997, **94**, 4273-4278.

306. Nordsmark, M., Overgaard, M., Overgaard, J. *Radiotherapy and Oncology*, 1996, **41**, 31-39.

307. Hockel, M., Schlenger, K., Aral, B., Mitze, M., Schaffer, U., Vaupel, P. *Cancer Res.*, 1996, **56**, 4509-4515.

308. Brizel, D. M., Scully, S. P., Harrelson, J. M., Layfield, L. J., Bean, J. M., Prosnitz, L. R., Dewhirst, M. W. *Cancer Res.*, 1996, **56**, 941-943.

309. Sutherland, R. M. *Acta Oncologica*, 1998, **37**, 567-574.

310. Wang, G. L., Jiang, B., Rue, E. A., Semenza, G. L. *Proc. Natl. Acad. Sci. USA*, 1995, **92**, 5510-5514.

311. Maxwell, P. H., Dachs, G. U., Gleadle, J. M., Nicholls, L. G., Harris, A. L., Stratford, I. J., Hankinson, O., Pugh, C. W., Ratcliffe, P. J. *Proc. Natl. Acad. Sci. USA*, 1997, **94**, 8104-8109.

312. Wang, G. L. Semenza, G. L. *Proc. Natl. Acad. Sci. USA*, 1993, **90**, 4304-4308.

313. Jiang, B., Rue, E., Wang, G. L., Roe, R., Semenza, G. L. *J. Biol. Chem.*, 1996, **271**, 17771-17778.

314. Bunn, H. F. Poyton, R. O. *Physiological Reviews*, 1996, **76**, 839-885.

315. Forsythe, J. A., Jiang, B., Iyer, N. V., Agani, F., Leung, S. W., Koos, R. D., Semenza, G. L. *Mol. Cell Biol.*, 1996, **16**, 4604-4613.

316. Semenza, G. L., Agani, F., Booth, G., Forsythe, J., Iyer, N., Jiang, B., Leung, S., Roe, R., Wiener, C., Yu, A. *Kidney Int.*, 1997, **51**, 553-555.

317. Vandenbunder, B. *Bulletin du Cancer*, 1998, **85**, 843-845.

318. Semenza, G. L. *Curr. Opin. Genetics Dev.*, 1998, **8**, 588-594.

319. An, W. G., Kanekal, M., Simon, M. C., Maltepe, E., Blagosklonny, M. V., Neckers, L. M. *Nature*, 1998, **392**, 405-408.

320. Bhattacharya, S., Michaels, C. L., Leung, M.-K., Arany, Z., Kung, A. L., Livingston, D. M. *Genes & Dev.*, 1999, **13**, 64-75.

321. Pugh, C. W., O'Rourke, J. F., Nagaos, M., Gleadle, J. M., Ratcliffe, P. J. *J. Biol. Chem.*, 1997, **272**, 11205-11214.

322. Li, H., Ko, H. P., Whitlock, J. P. J. *J. Biol. Chem.*, 1996, **271**, 21262-21267.

323. Jiang, B., Zheng, J. Z., Leung, S. W., Roe, R., Semenza, G. L. *J. Biol. Chem.*, 1997, **272**, 19253-19260.

324. Huang, L. E., Gu, J., Schau, M., Bunn, H. F. *Proc. Natl. Acad. Sci. USA*, 1998, **95**, 7987-7992.

325. Sogawa, K., Numayama-Tsuruta, K., Ema, M., Abe, M., Abe, H., Fujii-Kuriyama, Y. *Proc. Natl. Acad. Sci. USA*, 1998, **95**, 7368-7373.

326. Gu, Y. Z., Moran, S. M., Hogenesch, J. B., Wartman, L., Bradfield, C. A. *Gene Expression*, 1998, 7, 205-213.

327. Tian, H., McKnight, S. L., Russell, D. W. *Genes & Dev.*, 1997, **11**, 72-82.

328. Zhou, Y.-D., Barnard, M., Tian, H., Ring, H. Z., Francke, U., Shelton, J., Richardson, J., Russell, D. W., McKnight, S. L. *Proc. Natl. Acad. Sci. USA*, 1997, **94**, 713-718.

329. Cush, R., Cronin, J. M., Stewart, W. J., Maule, C. H., Molloy, J., Goddard, N. J. *Biosensors and Bioelectronics*, 1993, **8**, 347-353.

330. van der Meer, B. W., Coker III, G., Chen, S. Y. *Resonance Energy Transfer*, VCH Press, 1994.

331. Cantor, C. R., Schimmel, P. R. *Biophysical Chemistry*, Freeman Press, 1980, **2**, 448-455.

332. Fairclough, R. H. Cantor, C. R. *Methods Enzymol.*, 1978, **XLVIII**, 347-379.

333. Stryer, L. *Ann. Rev. Biochem.*, 1978, **47**, 819-846.

334. Stryer, L. Haughland, R. P. *Proc. Natl. Acad. Sci. USA*, 1967, **58**, 719-726.

335. Carnieri, E. G. S., Moreno, S. N. J., Docampo, R. *Mol. Biochem. Parasitol.*, 1993, **61**, 79-86.

336. dos Remedios, C. G. Moens, P. D. J. *J. Struct. Biol.*, 1995, **15**, 175-185.

337. Selvin, P. R. *Methods Enzymol.*, 1995, **246**, 300-334.

338. Miki, M., O'Donoghue, S. I., dos Remedios, C. G. *J. Mus. Res. Cell Mot.*, 1992, **13**, 132-145.

339. Clegg, R. M. *Methods Enzymol.*, 1995, **211**, 353-388

340. Glazer, A. N. Stryer, L. *Methods Enzymol.*, 1995, **184**, 188-194.

341. Jovin, T. M. Arndt-Jovin, D. J. *Ann. Rev. Biophys. Biophys. Chem.*, 1989, **18**, 271-308.

342. Ozaki, H. McLaughlin, L. W. *Nucl. Acids Res.*, 1992, **20**, 5205-5214.

343. Clegg, R. M., Murchie, A. I. H., Zechel, A., Lilley, D. M. J. *Proc. Natl. Acad. Sci. USA,* 1993, **90**, 2994-2998.

344. Clegg, R. M., Murchie, A. I. H., Zechel, A., Carlberg, C., Diekmann, S., Lilley, D. M. J. *Biochemistry*, 1992, **31**, 4846-4856.

345. Parkhurst, K. M. Parkhurst, L. J. *Biochemistry*, 1995, **34**, 285-292.

346. Sixou, S., Szoka, F. C., Green, G. A., Giusti, B., Zon, G., Chin, D. J. *Nucl. Acids Res.*, 1994, **22**, 662-668.

347. Leder, R. O., Helgerson, S. L., Thomas, D. D. *J. Mol. Biol.*, 1989, **9**, 683-701.

348. Cheung, H. C., Gryczynski, I., Malak, H., Wiczk, W., Johnson, M. L., Lakowicz, J. R. *Biophys. Chem.*, 1991, **40**, 1-17.

349. Cronce, D. T. Horrocks, W. D. *Biochemistry*, 1992, **31**, 7963-7969.

350. Kubitscheck, U., Kircheis, M., Schweitzer-Stenner, R., Dreybrodt, W., Jovin, T. M., Pecht, I. *Biophys. J.* 1991, **60**, 307-318.

351. Berger, W., Prinz, H., Striessnig, J., Kang, H.-C., Haughland, R., Glossman, H. *Biochemistry*, 1994, **33**, 11875-11883.

352. Robbins, D., Odom, O. W., Lynch, J., Kramer, G., Hardesty, B., Liou, R., Ofengand, J. *Biochemistry*, 1981, **20**, 5301-5309.

353. Chung, D. G. Lewis, P. N. *Biochemistry*, 1986, **25**, 5036-5042.

354. Heyduk, T. Lee, J. C. *Proc. Natl. Acad. Sci. USA*, 1990, **87**, 1744-1748.

355. Matayoshi, E. D., Wang, G. T., Krafft, G. A., Erickson, J. *Science*, 1990, **247**, 954-958.

356. Voronova, A. Baltimore, D. *Proc. Natl. Acad. Sci. USA*, 1990, **87**, 4722-4726.

357. Winter, B., Braun, T., Arnold, H. H. *EMBO J.*, 1992, **11**, 1843-1855.

358. Wibley, J. E. A., Deed, R. W., Jasiok, M., Douglas, K. T., Norton, J. D. *Biochim. Biophys. Acta*, 1996, **1294**, 138-146.

359. Ma, P. C. M., Rould, M. A., Weintraub, H., Pabo, C. O. *Cell,* 1994, **77**, 451-459.

360. Shimizu, T., Toumoto, A., Ihara, K., Shimizu, M., Kyogoku, Y., Ogawa, N., Oshima, Y., Hakoshima, T. *EMBO J.*, 1997, **16**, 4689-4697.

361. Brownlie, P., Ceska, T. A., Lamers, M., Romier, C., Stier, G., Teo, H., Suck, D. *Structure,* 1997, **5**, 509-520.

362. Ferré-D'Amaré, A. R., Prendergast, G. C., Ziff, E. B., Burley, S. K. *Nature,* 1997, **363**, 38-45.

363. Ferré-D'Amaré, A. R., Pognonec, P., Roeder, R. G., Burley, S. K. *EMBO J.*, 1994, **13**, 180-189.

364. Gibson, T. J., Thompson, J. D., Abagyan, R. A. *Prot. Eng.,* 1993, **6**, 41-50.

365. Ellenberger, T., Fass, E., Arnaud, M., Harrison, S. C. *Genes & Dev.*, 1994, **8**, 970-980.

366. Parraga, A., Bellsolell, L., Ferré-D'Amaré, A. R., Burley, S. K. *Structure,* 1998, **6**, 661-672.

367. Fairman, R., Beran-Steed, R. K., Handel, T. M. *Prot. Sci.*, 1997, **6**, 175-184.

368. Lavigne, P., Crump, M. P., Gagne, S. M., Hodges, R. S., Kay, C. M., Skyes, B. D. *J. Mol. Biol.* 1998, **281**, 165-181.

369. ,Gouzy, J., Corpet, F., Kahn, D. *TIBS,* 1996, **21**, 493

370. Bleasby, A. J. Wooton, J. C. *Prot. Eng.,* 1990, **3**, 153-159.

371. Ackrigg, D., Bleasby, A. J., Dix, N. I. M., Findlay, J. B. C., North, A. C. T., Parry-Smith, D., Wooton, J. C., Blundell, T. L., Gardner, S. P., Hayes, F., Islam, S., Sternberg, M. J. E., Thornton, J. M. *Nature,* 1988, **335**, 745-746.

372. Bairoch, A. Apweiler, R. *Nucl. Acids Res.,* 1997, **25**, 31-36.

373. Wang, G. L. Semenza, G. L. *J. Biol. Chem.*, 1993, **268**, 21513-21518.

374. Hu, J., Anderson, B., Wessler, S. R. *Genetics,* 1996, **142**, 1021-1031.

375. Ludwig, S. R., Habera, L. F., Dellaporta, S. L., Wessler, S. R. *Proc. Natl. Acad. Sci. USA,* 1989, **86**, 7092-7096.

376. Sonnenfeld, M., Ward, M., Nystrom, G., Mosher, J., Stahl, S., Crews, S. *Development,* 1997, **124**, 4571-4582.

377. Tsay, H. J., Choe, Y. H., Neville, C. M., Schmidt, J. *Nucl. Acids Res.,* 1992, **20**, 1805

378. Ben-Arie, N., McCall, A. E., Berkman, S., Eichele, G., Bellen, H. J., Zoghbi, H. Y. *Human Moecular Genetics,* 1996, **5**, 1207-1216.

379. Benton, B. K., Read, M. S., Okayama, H. *EMBO J.,* 1993, **12**, 135-143.

380. Nikoloff, D. M., McGraw, P., Henry, S. A. *J. Biochem.*, 1994, **115**, 131-136.

381. Hoshizaki, D. K., Hill, J. E., Henry, S. A. *J. Biol. Chem.*, 1990, **265**, 4736-4745.

382. Liao, X. Butow, R. A. *Cell*, 1993, **72**, 61-71.

383. Quaggin, S. E., Vanden Heuvel, G. B., Igarashi, P. *Mech. Dev.*, 1998, **71**, 37-48.

384. Cross, J. C., Flannery, M. L., Blanar, M. A., Steingrimsson, E., Jenkins, N. A., Copeland, N. G., Rutter, W. J., Werb, Z. *Development*, 1995, **121**, 2513-2523.

385. Sawai, S. Campos-Ortega, J. A. *Mech. Dev.* 1997, **65**, 175-185.

386. Pscherer, A., Dorflinger, U., Kirfel, J., Gawles, K., Ruschoff, J., Buettner, R., Schule, R. *EMBO J.*, 1996, **15**, 6680-6690.

387. Suzuki, M., Okuyama, S., Okamoto, S., Shirasuna, K., Nakajima, T., Hachiya, T., Nojima, H., Sekiya, S., Oda, K. *Oncogene*, 1998, **17**, 853-865.

388. Massari, M. E., Rivera, R. R., Voland, J. R., Quong, M. W., Breitt, T. M., van Dongen, J. J. M., de Smit, O., Murre, C. *Mol. Cell. Biol.*, 1998, **18**, 3130-3139.

389. Burgess, R., Cserjesi, P., Ligon, K. L., Olson, E. N. *Developmental Biology*, 1995, **168**, 296-306.

390. Brown, N. L., Kanekar, S., Vetter, M. L., Tucker, P. T., Gemza, D. L., Glaser, T. *Development*, 1998, **125**, 4821-4833.

391. Tamura, M. Noda, M. *J. Cell. Biochem.*, 1999, **72**, 167-176.

392. Knofler, M., Meinhardt, G., Vasicek, R., Husslein, P., Egarter, C. *Gene*, 1998, **224**, 77-86.

393. Russell, M. W., Kemp, P., Wang, L., Brody, L. C., Izumo, S. *Biochim. Biophys. Acta*, 1998, **1443**, 393-399.

394. Braun, T., Bober, E., Buschhausen-Denker, G., Kohtz, S., Grzeschik, K., Arnold, H. H., Kotz, S. K. *EMBO J.*, 1989, **8**, 3617-3625.

395. Henthorn, P., McCarrick-Walmsley, R., Kadesch, T. *Nucleic Acids Res.*, 1990, **18**, 677-678.

396. Sawadogo, M. Roeder, R. G. *Cell*, 1985, **43**, 165-175.

397. Legrain, M., De Wilde, M., Hilger, F. *Nucl. Acids Res.*, 1986, **14**, 3059-3073.

398. Pouny, Y., Rapaport, D., Mor, A., Nicolas, P., Shai, Y. *Biochemistry*, 1992, **31**, 12417-12423.

399. Mehta, A., Jaouhari, R., Benson, T. J., Douglas, K. T. *Tet. Lett.*, 1992, **33**, 5441-5444.

400. Shirakata, M., Friedman, F. K., Wei, Q., Paterson, B. M. *Genes & Dev.*, 1993, **7**, 2456-2470.

401. Zhu, B., Zhou, N. E., Kay, C. M., Hodges, R. S. *Prot. Sci.*, 1993, **2**, 383-394.

402. Prevost, M., Wodak, S. J., Tidor, B., Karplus, M. *Proc. Natl. Acad. Sci. USA*, 1991, **88**, 10880-10884.

403. Shirakata, M. Paterson, B. M. *EMBO J.*, 1995, **14**, 1766-1772.

404. Vinson, C. R., Hai, T., Boyd, S. M. *Genes & Dev.*, 1993, **7**, 1047-1058.

405. Prendergast, G. C. Ziff, E. B. *Science*, 1991, **251**, 186-189.

406. Blackwell, T. K., Kretzner, L., Blackwood, E. M., Eisenman, R. N., Weintraub, H. *Science*, 1990, **250**, 1149-1151.

407. Fisher, D. E., Parent, L. A., Sharp, P. A. *Cell*, 1993, **72**, 467-476.

408. Fisher, F. Goding, C. R. *EMBO J.*, 1992, **11**, 4103-4109.

409. Davis, R. L., Cheng, P. F., Lassar, A. B., Weinbraub, H. *Cell*, 1990, **60**, 733-746.

410. Halazonetis, T. Kandil, A. N. *Science*, 1992, **255**, 464-466.

411. Dang, C. V., Dolde, C., Gillison, M. L., Kato, G. J. *Proc. Natl. Acad. Sci. USA*, 1992, **89**, 599-602.

412. Blackwell, T. K., Huang, J., Ma, A., Kretzner, L., Alt, F. W., Eisenman, R. N., Weintraub, H. *Mol. Cell. Biol.*, 1993, **13**, 5216-5224.

413. Swanson, H. I., Chan, W. K., Bradfield, C. A. *J. Biol. Chem.*, 1995, **270**, 26292-26302.

414. Bendall, A. J. Molloy, P. L. *Nucl. Acids Res.*, 1994, **22**, 2801-2810.

415. Wendt, H., Thomas, R. M., Ellenberger, T. *J. Biol. Chem.*, 1998, **273**, 5735-5743.

416. Bartlett, P. A., Shea, G. T., Waterman, S., Telfer, S. J. *Abstracts of Papers of the American Chemical Society*, 1991, **202**, 44-COMP.

417. Lawrence, M. C. Davis, P. C. *Proteins: Structure, Function and Genetics*, 1998, **12**, 31-41.

418. Pashley, T. V., Volpe, F., Pudney, M., Hyde, J. E., Sims, P. F., Delves, C. J. *Mol. Biochem. Parasitol.*, 1997, **86**, 37-47.

419. Miller, M. D., Kearsley, S. K., Underwood, D. J., Sheridan, R. P. *J. Comp. Aided Mol. Design*, 1994, **8**, 153-174.

420. Jiang, F. Kim, S.-H. *J. Mol. Biol.*, 1991, **219**, 79-102.

421. Rarey, M., Wefing, S., Lengauer, T. *J. Comp. Aided Mol. Design*, 1996, **10**, 41-54.

422. Vasker, I. A. *Biopolymers*, 1996, **39**, 455-464.

423. Vasker, I. A. *Prot. Eng.* 1995, **8**, 371-377.

424. Vasker, I. A. Aflalo, C. *Proteins: Structure, Function and Genetics*, 1994, **20**, 320-329.

425. Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Afalo, C., Vasker, I. A. *Proc. Natl. Acad. Sci. USA*, 1992, **89**, 2195-2199.

426. Moon, J. B. Howe, K. M. *Proteins: Structure, Function and Genetics*, 1991, **11**, 314-328.

427. Goodsell, D. S., Morris, G. M., Olson, A. J. *J. Mol. Recogn.*, 1996, **9**, 1-5.

428. Jones, G., Willett, P., Glen, R. C., Leach, A. R., Taylor, R. *J. Mol. Biol.*, 1997, **267**, 727-748.

429. Gabb, H. A., Jackson, R. M., Sternberg, M. J. E. *J. Mol. Biol.*, 1997, **272**, 106-120.

430. Cherfils, J., Duquerroy, S., Janin, J. *Proteins: Structure, Function and Genetics*, 1991, **11**, 271-280.

431. Pearlman, D. A. Murcko, M. A. *J. Comp. Chem.*, 1993, **14**, 1184-1193.

432. Totrov, M. Abagyan, R. *Nature Struct. Biol.*, 1994, **1**, 259-263.

433. Goodford, P. J. *J. Med. Chem.*, 1985, **28**, 849-857.

434. Di Nola, A., Roccatano, D., Berendsen, H. J. C. *Proteins: Structure, Function and Genetics*, 1994, **19**, 174-182.

435. Judson, R. S., Jaeger, E. P., Treasurywala, A. M. *J. Mol. Struct.*, 1994, **308**, 191-206.

436. Oshiro, C. M., Kuntz, I. D., Dixon, J. S. *J. Comp. Aided Mol. Design*, 1998, **9**, 113-130.

437. Jones, G., Willett, P., Glen, R. C. *J. Comp. Aided Mol. Design*, 1995, **9**, 532-549.

438. Morris, G. A., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., Olsen, A. J. *J. Comp. Chem.*, 1998, **19**, 1639-1662.

439. Lamarck, J. B. *Zoological Philosophy*, MacMillan Press, 1914.

440. Vasker, I. A. *Proteins: Structure, Function and Genetics*, 1997, **1**, 226-230.

441. Wodak, S. J. Janin, J. *J. Mol. Biol.*, 1978, **124**, 323-342.

442. Shoichet, B. K. Kuntz, I. D. *J. Mol. Biol.*, 1991, **221**, 327-346.

443. Sobolev, V., Wade, R. C., Edelman, M. *Proteins: Structure, Function and Genetics*, 1996, **25**, 120-129.

444. Mehler, E. L. Solmajer, T. *Prot. Eng.*, 1991, **4**, 903-910.

445. Horvath, D. *J. Med. Chem.*, 1997, **40**, 2412-2423.

446. Gilson, M. Honig, B. *J. Comp. Aided Mol. Design*, 1991, **5**, 5-20.

447. Vieth, M., Hirst, J. D., Kolinski, A., Brooks, C. L. *J. Comp. Chem.*, 1998, **19**, 1612-1622.

448. Warwicker, J. Watson, H. C. *J. Mol. Biol.*, 1982, **157**, 671-679.

449. Vieth, M., Hirst, J. D., Dominy, B. N., Daigler, H., Brooks, C. L. *J. Comp. Chem.*, 1998, **19**, 1623-1631.

450. Coutinho, P. M., Dowd, M. K., Reilly, P. J. *Proteins: Structure, Function and Genetics*, 1997, **27**, 235-248.

451. Coutinho, P. M., Dowd, M. K., Reilly, P. J. *Proteins: Structure, Function and Genetics*, 1997, **28**, 162-173.

452. Coutinho, P. M., Dowd, M. K., Reilly, P. J. *Carbohyd. Res.*, 1997, **297**, 309-324.

453. Goodsell, D. S., Lauble, H., Stout, C. D., Olsen, A. J. *Proteins: Structure, Function and Genetics*, 1993, **17**, 1-10.

454. Morris, G. M., Goodsell, D. S., Huey, R., Olson, A. J. *J. Comp. Aided Mol. Design*, 1996, **10**, 293-304.

455. Sotriffer, C. A., Liedl, K. R., Winger, R. H., Gamper, A. M., Kroemer, R. T., Linthicum, D. S., Rode, B. M., Varga, J. M. *Mol. Immun.*, 1996, **33**, 129-144.

456. Siani, M. A., Weininger, D., Blaney, J. M. *J. Chem. Inf. Comp. Sci.*, 1994, **34**, 588-593.

457. Schaffer, L. Verkhivker, G. M. *Proteins: Structure, Function and Genetics*, 1998, **33**, 295-310.

458. Sandak, B., Wolfson, H. J., Nussinov, R. *Proteins: Structure, Function and Genetics*, 1998, **32**, 159-174.

459. Lorber, D. M. Shoichet, B. K. *Prot. Sci.*, 1998, **7**, 938-950.

460. Austin, S. E., Khan, M. A. O., Douglas, K. T. *Drug Design Discov.*, 1999, In Press.

461. Shames, S. L., Fairlamb, A. H., Cerami, A., Walsh, C. T. *Biochemistry*, 1986, **25**, 3519-3526.

462. Shames, S. L., Kimmel, B. E., Peoples, O. E., Agabian, N., Walsh, C. T. *Biochemistry*, 1988, **27**, 5014-5019.

463. Bradley, M., Bücheler, U. S., Walsh, C. T. *Biochemistry*, 1991, **30**, 6124-6127.

464. Williams, C. H. *Enzymes*, Academic Press, 1976, 129-142.

465. Pai, E. F., Schulz, G. E. *Flavins and Flavoproteins*, Elsevier Press, 1982, 3-10.

466. Williams, C. H., Jr., *Chemistry and Biochemistry of Flavoenzymes*, CRC Press, 1992, 121-211.

467. Petsko, G. A. *Nature*, 1991, **352**, 104-105.

468. Kuriyan, J., Kong, X., Krishna, T. S. R., Sweet, R. M., Murgolo, N. J., Field, H., Cerami, A., Henderson, G. B. *Proc. Natl. Acad. Sci. USA*, 1991, **88**, 8764-8768.

469. Hunter, W. N., Bailey, S., Habash, J., Harrop, S. J., Helliwell, J. R., Aboagye-Kwarteng, T., Smith, K., Fairlamb, A. H. *J. Mol. Biol.*, 1992, **227**, 322-333.

470. Lantwin, C. B., Schlichting, I., Kabsch, W., Pai, E. F., Krauth-Siegel, R. L. *Proteins: Structure, Function and Genetics*, 1994, **18**, 161-173.

471. Bailey, S., Smith, K., Fairlamb, A. H., Hunter, W. N. *Eur. J. Biochem.*, 1993, **213**, 67-75.

472. Jacoby, E. M., Schlichting, I., Lantwin, C. B., Kabsch, W., Krauth-Siegel, R. L. *Proteins: Structure, Function and Genetics*, 1996, **24**, 73-80.

473. Garforth, J., Yin, H., McKie, J. H., Douglas, K. T., Fairlamb, A. H. *J. Enz. Inhib.*, 1997, **12**, 161-173.

474. Chan, C., Yin, H., Garforth, J., McKie, J. H., Jaouhari, R., Speers, P., Douglas, K. T., Rock, P. J., Yardley, V., Croft, S. L., Fairlamb, A. H. *J. Med. Chem.*, 1998, **41**, 148-156.

475. Benson, T. J., McKie, J. H., Garforth, J., Borges, A., Fairlamb, A. H., Douglas, K. T. *Biochem. J.*, 1992, **286**, 9-11.

476. Khan, O. F. *PhD Thesis: Antiparasitic drug design based on trypanothione reductase as a target*, 1999.

477. Fernandez-Gomez, R., Moutiez, M., Aumercier, M., Bethegnies, G., Luyckx, M., Ouaissi, A., Tartar, A., Sergheraert, C. *Int. J. Antimicrob. Agents*, 1995, **6**, 111-118.

478. Obata, A., Kawazura, H., Miyamae, H. *Acta Cryst., Sect. C*, 1984, **40**, 45-48.

479. Klein, C. L. Conrad III, J. M. *Acta Cryst., Sect. C*, 1986, **42**, 1083-1085.

480. Ealick, S. E., van der Helm, D., Barclay, C., Lehr, R. E. *Cryst. Struct. Comm.*, 1978, **7**, 711-717.

481. Dorignac-Calas, M.-R. Marsau, P. *C. R. Acad. Sci.*, *Ser. C*, 1972, **274**, 1806-1809.

482. Marsau, P. Cotrait, M. *Acta Cryst., Sect. B*, 1976, **32**, 3135-3137.

483. Faerman, C. H., Savvides, S. N., Strickland, C., Breidenbach, M. A., Ponasik, J. A., Ganem, B., Ripoll, D., Krauth-Siegel, R. L., Karplus, P. A. *Bioorg. Med. Chem.*, 1996, **4**, 1247-1253.

484. Korn, A. P. Burnett, R. M. *Proteins: Structure, Function and Genetics*, 1991, **9**, 37-55.

485. Galaktionov, S. G., Tseitin, V. M., Vasker, I. A., Prokhorchik, Y. V. *Biophysics*, 1988, **33**, 595-598.

486. Chan, H. S. Dill, K. A. *Annu. Rev. Biophys. Biophys. Chem.*, 1991, **20**, 447-490.