Conformational analysis of peptides and proteins for drug design using molecular simulations

A thesis submitted to the University of Manchester for the degree of

Doctor of Philosophy

in the Faculty of Medical and Human Sciences

Alessio Atzori

2014

Manchester Pharmacy School
Table of Contents

Table of Contents ........................................................................................................... 2
List of Figures .................................................................................................................. 5
List of Tables.................................................................................................................... 11
Abstract ............................................................................................................................ 12
Declaration and Copyright Statement .............................................................................. 13
Acknowledgements .......................................................................................................... 14
List of Abbreviations ....................................................................................................... 15
The Author ...................................................................................................................... 18
List of Publications ......................................................................................................... 18

1. Introduction .................................................................................................................. 19
   1.1 Molecular dynamics .............................................................................................. 20
      1.1.1 Force fields ................................................................................................... 21
      1.1.2 Equations of motions and their integration .................................................. 23
      1.1.3 Solvation models .......................................................................................... 26
         1.1.3.1 Implicit solvation models ...................................................................... 26
         1.1.3.2 Explicit solvation models .................................................................... 29
         1.1.3.3 Periodic boundary conditions and long-range interactions .......... 31
      1.1.4 Temperature and pressure regulation ............................................................ 32
      1.1.5 Limitations of molecular dynamics simulations ........................................... 34
   1.2 Advanced molecular dynamics methods ............................................................... 35
      1.2.1 Replica exchange molecular dynamics (REMD) ............................................ 36
         1.2.1.2 Application of the swarm intelligence method to molecular dynamics
                          simulations ............................................................................................ 37
      1.2.3 Swarm-enhanced sampling molecular dynamics (sesMD) .......................... 40
         1.2.3.1 Implementation and validation of sesMD .................................................. 41
   1.3 Docking .................................................................................................................... 44
      1.3.1 Ligand conformational search methods .......................................................... 45
      1.3.2 Scoring functions .......................................................................................... 47
   1.4 Retro-inverso peptides as inhibitors of p53/MDM2 interactions .......................... 49
      1.4.1 Protein-protein interactions .......................................................................... 49
      1.4.2 Hot spots ....................................................................................................... 51
      1.4.3 Peptides as drugs .......................................................................................... 53
Table of Contents

1.4.4 Peptidomimetics ................................................................. 55
1.4.5 Retro-inverso (RI) peptides .................................................. 56
1.4.6 p53 .................................................................................. 58
1.4.7 Structural characterization of the p53/MDM2 interface ............... 60
1.4.8 Pharmacological inhibitors of p53/MDM2 interaction ................. 62
1.4.9 RI peptides and p53/MDM2 ..................................................... 64

1.5 Protein kinases and DFG motif ................................................. 67
1.5.1 Protein kinases .................................................................. 67
1.5.2 Conserved structural features of protein kinases ....................... 69
1.5.3 DFG motif ..................................................................... 71

1.6 R-spondins and the canonical Wnt pathway ................................ 75
1.6.1 Canonical Wnt signalling pathway ...................................... 75
1.6.2 R-spondin proteins ............................................................. 77
1.6.3 R-spondin proteins and the cWnt signalling pathway ............... 79

1.7 References ..................................................................... 81

2. Effect of sequence and stereochemistry reversal on p53 peptide mimicry 100
2.1 Abstract ....................................................................... 101
2.2 Author Summary ................................................................. 102
2.3 Introduction ..................................................................... 103
2.4 Results/Discussion ............................................................... 107
2.5 Conclusions ..................................................................... 121
2.6 Materials and Methods ........................................................ 124
2.7 References ..................................................................... 126
2.8 Supporting Information ....................................................... 128

3. Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics ............................. 133
3.1 Abstract ....................................................................... 134
3.2 Introduction ..................................................................... 135
3.3 Methods ...................................................................... 138
3.4 Results ....................................................................... 143
3.5 Discussion and conclusions ................................................ 158
3.6 References ..................................................................... 162
3.7 Supporting Information ....................................................... 166
Table of Contents

4. Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis ................................................................. 174
   4.1 Abstract ........................................................................................................... 175
   4.2 Author Summary ............................................................................................ 176
   4.3 Introduction ..................................................................................................... 177
   4.4 Materials and Methods .................................................................................. 182
   4.5 Results and Discussion .................................................................................. 184
   4.6 Conclusions ..................................................................................................... 200
   4.7 References ....................................................................................................... 202
   4.8 Supporting Information ................................................................................. 206
5. Conclusions and outlook ................................................................................... 209

Final word count: 51701
List of Figures

Figure 1-1: The four components describing the total energy of the system in a force field ............................................................. 22

Figure 1-2: Periodic boundary conditions. The central box (in cyan) represents the system simulated surrounded by replicas which possess the same particles (represented as red dots) with the same velocities (indicated by arrows) as the simulation box. \( r_{cut} \), which represent the cut-off radius applied when calculating non-bonded interactions between particles, is also indicated in figure. ...................... 31

Figure 1-3: Comparison of forces calculated by sesMD in Amber, and forces calculated from finite difference of swarm energy (red, green and blue show the \( x \), \( y \) and \( z \) components of the forces respectively). ............................................. 42

Figure 1-4: Map of structural and functional epitopes for binding of human growth hormone (hGH) site onto the structure of hGHbp. Residues are colour-coded according to the loss of binding free energy upon the substitution with an alanine. Image adapted from [125]. ................................................................. 52

Figure 1-5: Peptides can exist in different conformations. By introducing conformational constraints into the peptide backbone (dashed line), the equilibrium can be shifted towards the conformation which leads to the desired biological effect. In addition, it might be possible to suppress undesired side effects or proteolysis. Image adapted from [135]. ................................................................. 54

Figure 1-6: Isomers of linear parent (all-L) peptide (1): (2) enantiomer, (3) retro-isomer (all-I), (4) retro-all-D-isomer, (5) end group modified retro-all-D-isomer. Carbon atoms are represented in white, while the nitrogen ones are in grey. \( \alpha \), \( \beta \), \( \gamma \) and \( \delta \) represent various side-chain residues. Image adapted from [145]. ............... 57

Figure 1-7: DNA damage or cellular stressful situations leads to the dissociation of the p53 and MDM2 complex. Upon activation, p53 can activate p21 leading to the arrest of the cellular cycle or engaging apoptosis via transcription of pro-apoptotic genes. ........................................................................................................... 59

Figure 1-8: a) Side view (left) and top view (right) of N-terminal sequence of p53 (residues 17–29, in green) bound to the MDM2 (residues 25–109, in white) derived from its crystal structure (PDBID: 1YCR[123]). The amino (N) and carboxyl (C) termini of p53 are indicated. b) Frontal view of the N-terminal sequence of p53 (green) bound to the MDM2 (white) derived from its crystal structure. Backbone atoms are represented in cartoon, while the residues Phe\(^{19}\), Trp\(^{23}\) and Leu\(^{26}\) are in sticks. ........................................................................................................ 61

Figure 1-9: Example of a phosphorylation in position 6 of glucose performed by the hexokinase to obtain glucose-6-phosphate ................................................................. 68

Figure 1-10: Crystal structure of a protein kinase (PDBID: 1CSN) in complex with ATP molecule. On the right, close-up of the ATP binding site: ATP and interacting residues of the binding site are shown in sticks. ................................................................. 69

Figure 1-11: Superposition of 1P38 (green) and 1WBT (red) X-ray crystal structures of p38\(\alpha\) MAP kinase. Backbone atoms are represented in cartoon while residues Asp168 and Phe169 are in sticks. Picture adapted from [229]. ........................................ 72
hydrogen bonds are present; (b) The binding of Wnt ligands to the Fz receptor and LRP co-receptors induces phosphorylation of LRP via Dvl and consequent disruption of the degradation complex. In this way, β-catenin can enter the nucleus, where activates the transcription of cWnt target genes. Image adapted from [243].

Figure 1-12: Overview of the canonical Wnt (cWnt) signalling pathway. a) “OFF” state: In the absence of Wnt ligands, the degradation complex (Axin, APC, GSK3β, CK1) targets β-catenin, which is phosphorylated and recognized by β-TrCP, leading to its ubiquitination and degradation. In this way, the expression of cWnt target genes via β-catenin activation is prevented by TCF-TLE/Groucho and HDAC. b) “ON” state: The binding of Wnt ligands to the Fz receptor and LRP co-receptors induces phosphorylation of LRP via Dvl and consequent disruption of the degradation complex. In this way, β-catenin can enter the nucleus, where activates the transcription of cWnt target genes. Image adapted from [243].

Figure 2-1: Peptide transformations: (a) parent peptide, (b) its inverse, (c) its sequence reversed and (d) its retro-inverso analogue.

Figure 2-2: (a) Representation of the N-terminal sequence of p53(white) bound to the MDM2(electrostatic map) derived from its crystal structure. Backbone atoms are represented in cartoon, while the residues Phe\(^{19}\), Trp\(^{23}\) and Leu\(^{26}\) are in sticks. (b) Interatomic distances measured from the crystallographic structure accordingly to the two reference sites, namely α carbon (in red) and β carbon (in green). The three key residues (Phe\(^{19}\), Trp\(^{23}\) and Leu\(^{26}\)) are represented in sticks.

Figure 2-3: Fractional helical content (ordinate) for sequences WT, I, R and RI as a function of time (ns) (abscissa) obtained from (a) MD simulations and (b) REMD simulations: maximum helicity (dashed line) and average helicity (solid line) obtained starting from extended (E), right-handed (R) and left-handed (L) helical structures.

Figure 2-4: Snapshots of the 300 K trajectory taken at 0, 10, 20, 30, 40 and 50 ns starting from an extended (E), right-handed (R) and left-handed (L) helical structure. Each configuration is ordered from N- to C-terminus.

Figure 2-5: Ramachandran plots of the distribution of the (φ, ψ) backbone dihedrals of residues 1–15 for peptides WT, I, R and RI, calculated over of the 50 ns REMD trajectories.

Figure 2-6: Circular dichroism spectra of peptides: (a) CD spectra of WT and R at 0.3 mg/mL in PBS (b) CD spectra of WT and R at 0.3 mg/mL in 60% TFE/40% PBS (c) percentage peptide α-helix estimated from circular dichroism.

Figure 2-7: Distribution of helicity (ordinate) as a function of amino acid residue for each of sequences WT, I, R and RI (abscissa), obtained from final 20 ns of REMD simulations with f99SB. Residue numbering is from N- to C-terminal direction (solid line) for WT, I, R and RI. Additionally, helicity is shown for the reverse residue numbering i.e. from C- to N-terminus for R and RI (dashed line).

Figure 2-8: Intramolecular polar contacts. (a) A helical conformation of WT when Lys10 NH...Ser6 OC and Lys10 Hζ...Asp7 Oδ hydrogen bonds are present; (b) A helical conformation of R when Ser10 NH...Lys6 OC and Ser10 Hγ...Lys6 O hydrogen bonds are present.
Figure 2-9: (top) Superimposition of Cα atoms of X-ray structure of N-terminal sequence of p53 (white) onto representative MD conformations of WT (green, a), I (red, b), R (blue, c) and RI (yellow, d). (bottom) Comparison of the orientation of the side-chains of the residues Phe, Trp and Leu for the right-handed helix of WT (green, e), right handed helix of RI (orange, f) and left handed helix of RI (yellow, g). Key residues Phe, Trp and Leu represented as sticks.

Figure 2-10: Replica exchange equilibration for the replica initiated at 270 K...

Figure 2-11: $\beta_1$ helical content (ordinate) for sequences WT, I, R and RI as a function of time (ns) (abscissa) obtained from REMD simulations.

Figure 2-12: Fractional helical content (ordinate) for sequences WT and RI as a function of time (ns) (abscissa) obtained from REMD simulation with ff96: maximum helicity (dashed line) and average helicity (solid line) obtained starting from extended (E), right-handed (R) and left-handed (L) helical structures.

Figure 2-13: Distribution of helicity (ordinate) as a function of amino acid residue for each of sequences WT and RI (abscissa), obtained from final 20 ns of REMD simulations with ff96 obtained starting from extended (E), right-handed (R) and left-handed (L) helical structures. Residue numbering is from N- to C-terminal direction (solid line) for WT and RI. Additionally, helicity is shown for the reverse residue numbering i.e. from C- to N-terminus for RI (dashed line).

Figure 3-1: Superposition of the DFG-in (PDB 1P38, green) and DFG-out (PDB 1WBT, red) X-ray crystal structures of p38α MAP kinase. Backbone atoms are shown in cartoon representation and residues Asp168 and Phe169 in stick representation.

Figure 3-2: (a, b) Distributions of pentane conformations in water defined by the central C−C dihedral angles $\varphi$ 1 and $\varphi$ 2 during (a) eight independent MD simulations of 10 ns and (b) an eight-replica sesMD simulation of 10 ns. (c, d) Distributions of alanine dipeptide conformations in water as a function of the backbone dihedral angles $\phi$ and $\psi$ during (c) eight independent MD simulations of 10 ns and (d) an eight-replica sesMD simulation of 10 ns. Blue regions indicate areas of highest density.

Figure 3-3: Ramachandran plots of the $\varphi$ and $\psi$ angles for each DFG residue sampled by 24 replicas of sesMD simulations (black) starting from p38α MAP kinase crystal structures 1WBT (a−c) and 1P38 (g−i) and by 24 independent MD simulations (d−f and j−l). Also shown are $\varphi$ and $\psi$ angles of 61 DFG-out (red), 74 DFG-in (green), and 33 unclassified (blue) p38α MAP kinase X-ray structures as defined in the MOE Kinase Explorer Database.

Figure 3-4: Projections of MD (yellow) and sesMD (cyan) ensembles of kinase and DFG-out (red), DFG-in (green), and unclassified (dark blue) p38α MAP kinase crystal structures onto the space defined by the top two principal components based on simulations starting from the (a) 1WBT and (b) 1P38 crystal structures.

Figure 3-5: (a, b, d) Time series of heavy-atom RMSDs (in Å) of the DFG motif from sesMD simulations of p38α MAP kinase for (a) replica 1$_{out}$, (b) replica 2$_{out}$, and (d) replica 1$_{in}$ calculated against the DFG-out 1WBT (red) and DFG-in 1P38 (green) crystal structures during pre-sesMD equilibration (− 0.5 to 0.0 ns) and sesMD
trajectories (0.0 to 5.0 ns). (c, e) Projections onto the first two principal components of sesMD configurations of (c) replicas 1_{out} (orange) and 2_{out} (cyan) and (e) replica 1_{in} (magenta), alongside PC values for 24 MD replicas (yellow) and DFG-in (green), DFG-out (red), and unclassified (dark blue) crystal structures. 

**Figure 3-6:** (a, c) Time series of the distance between the side-chain centroids of Asp168 and Lys53 for (a) replicas 1_{out} (orange) and 2_{out} (cyan) and (c) replica 1_{in} (magenta) of p38α MAP kinase from pre-sesMD equilibration (−0.5 to 0.0 ns) and sesMD trajectories (0.0 to 5.0 ns). (b, d) Time series of the distance between the side-chain centroids of Phe169 and Leu74 of (b) replicas 1_{out} (orange) and 2_{out} (cyan) and (d) 1_{in} (magenta) from pre-sesMD equilibration and sesMD trajectories. The X-ray distances for the 1WBT (red) and 1P38 (green) structures are also shown for reference.

**Figure 3-7:** Conformation of Phe169 in selected snapshots from (a) replica 1_{out}, (b) replica 2_{out}, and (c) replica 1_{in} from sesMD simulations of p38α MAP kinase. The conformations from sesMD are color-coded according to the simulation time/stage: 0.5 − 1 ns (orange), 1 − 1.5 ns (yellow), 1.5 − 2 ns (cyan), 2 − 5 ns (blue). Also shown are the DFG-in 1P38 (green) and DFG-out 1WBT (red) crystal structures.

**Figure 3-8:** Final conformation from sesMD of replica 2_{out} showing the Asp168 and Phe169 side-chains in the p38α MAP kinase active site (blue), compared with its initial pose from the DFG-out 1WBT X-ray structure (red). The DFG-in crystal structures 1P38 (green) and 2LGC and 3HVC (both grey) are also shown.

**Figure 3-9:** Projections onto the space defined by the first two principal components of the ensemble of 24 sesMD simulation replicas, initiated from the (a) 1WBT and (b) 1P38 crystal structures. The projections are colour-coded according to cluster: C1_{out} and C1_{in} (red), C2_{out} and C2_{in} (orange), C3_{out} and C3_{in} (yellow), C4_{out} and C4_{in} (green), C5_{out} and C5_{in} (cyan), C6_{out} (blue), C7_{out} (magenta), and C8_{out} (purple).

**Figure 3-10:** Average binding pocket topologies found in structures within conformational clusters identified from sesMD simulations of p38α MAP kinase (using PocketAnalyzerPCA), mapped onto structures nearest to the cluster centroid. Each structure is color-coded according to cluster (see the caption of Figure 3-9). The pocket is color-coded according to the weighted average (high frequency, blue; low frequency, red). As references, sets of DFG-in and DFG-out crystallographic structures were used.

**Figure 3-11:** (a, b) Superpositions of a selected p38α MAP kinase conformation from sesMD replica 1_{out} (blue) with (a) 15 p38α MAP kinase X-ray structures (purple) and (b) X-ray structures 3BV2 and 3BV3 (purple). (c, d) Superpositions of selected intermediate DFG conformations (blue) from sesMD of (c) replica 1_{out} and (d) replica 1_{in}. For reference, DFG-in (1P38, green) and DFG-out (1WBT, red) X-ray structures are also shown.

**Figure 3-12:** Dependence of biasing potential V_{ses} (in kcal/mol) on $\delta\alpha, \beta_{rms}$, the root-mean-square dihedral angle distance (in degrees) between swarm members $\alpha$ and $\beta$, for 1WBT DFG-out (black line) and 1P38 DFG-in (red line) sesMD simulations.
Figure 3-13: Heavy atom RMSD of DFG motif with respect to the first frame of the simulation (black), DFG-out (red) and DFG-in (green) X-ray structures for (a) sesMD 24 replicas from sesMD and (b) unbiased MD simulations initiated from p38α MAPK structure 1WBT. Replicas are sorted from left-to-right, top-to-bottom.

Figure 3-14: Heavy atom RMSD of DFG motif with respect to DFG-out (red) and DFG-in (green) X-ray structures for (a) sesMD 24 replicas from sesMD and (b) unbiased MD simulations from p38α MAPK structure 1P38. Replicas are sorted from left-to-right, top-to-bottom.

Figure 3-15: Superposition of the activation loop of selected conformations from sesMD replica 2_out after 0.5 ns (orange) and after 5 ns (blue) of simulation with 55 DFG-in p38α MAP kinase X-ray structures (white). For reference, DFG-in 1P38 (green) and DFG-out 1WBT (red) crystallographic structures are shown.

Figure 4-1: a) Structure of free R-spondin 1, which adopts a closed conformation (PDBID: 4BSO). The two furin domains (red, Fu1 and blue, Fu2) and the head and rod modules are defined. The eight disulfide bridges and the residues involved in binding with RNF43/ZNRF3 are shown as sticks. b) X-ray structure (PDBID: 4KNG) of the ternary complex comprised by LGR5 (orange), R-spondin 1 (red) and RNF43 (cyan). c) Superposition of X-ray structures of free, closed R-spondin 1 (4BSO, green), bound R-spondin 1 (PDBID: 4KNG, red), and unbound R-spondin 2 (PDBID: 4C8V in blue). Both 4KNG and 4C8V structures show R-spondin in an open conformation. d) Molecular surfaces calculated for closed 4BSO (left), open 4KNG (centre) and open 4C8V (right) X-ray structures. Molecular surfaces are coloured red for H-bonding regions, blue for other polar regions and white for hydrophobic regions. The potential druggable pocket enclosed between the long β-hairpin moiety of the head module and the first β-hairpin of the rod module is indicated with a red circle.

Figure 4-2: Time series of backbone atom RMSD (in Å) from 1 μs MD simulations calculated against the first frame of the simulation (black) and reference X-ray structure (red) for backbone atoms of the second β hairpin of the Fu2 of R1U (a), R1C (c) and R2U (e). Time series of angle of rotation (in degrees) between the first furin domain (aligned) and the second furin domain from 1 μs MD simulations of R1U, (b) R1C (d) and R2U (f). Root mean square fluctuations (RMSF) of backbone atoms as a function of residue number from 1 μs MD simulations of R1U, (green) R1C (red) and R2U (blue).

Figure 4-3: a) Superposition of a selected set of conformations from R1U simulation (white) with 4BSO (green) and 4KNG (red) X-ray crystal structures. b) Superposition selected set of conformations from R1C simulation (white) with 4BSO (green) and 4KNG (red) X-ray crystal structures c) Salt bridges identified between residues Arg36-Glu73. d) Salt bridges identified between pairs of residues Glu35-Lys85 (black), Arg36-Glu73 (red), Arg36-Glu70 (green), Arg36-Glu86 (blue) for 1 μs MD simulation of R1C.

Figure 4-4: a) Superposition of a selected set of conformations from R2U simulation (white) with 4C8V (blue), 4BSO (green), 4KNG (red) X-ray crystal structures and
the most similar closed conformation from R1C (yellow). b) Hydrogen bond identified between residues Arg97-Arg107 (according to the numbering defined in X-ray structure 4C8V) c) Time series of the hydrogen bond identified between residues Arg97-Arg107 (according to the numbering defined in X-ray structure 4C8V) for 1 μs MD simulation of R2U. ................................................................. 188

**Figure 4-5**: a) Superposition of selected conformations from R1U based on first furin domain (red). b) Superposition of selected conformations from R1C based on first furin domain (red). c) Superposition of the available R-spondin 2 crystal structures based on first furin domain (red). d) Superposition of selected conformations from R2U based on first furin domain (red). In order to compare the extent of rotation between the two furin domains, Hie78 of R-spondin 1 and the corresponding Arg107 of R-spondin 2 are coloured in yellow. ........................................................................... 190

**Figure 4-6**: (top) Projections onto the first two principal components of the conformations from for 1 μs MD of R1U (a), R1C (b) and R2U (c), along with 4BSO (green square) 4KNG (red square) and 4C8V (blue square) reference crystal structures. The projections are colour-coded according to the simulation time. (bottom) Distribution of R1U (d), R1C (e) and R2U (f) conformations as a function of the distance between Cα of residues Hie78 and Cys64/Lys66 (of R1U and R1C) and residues Arg107 and Cys93/Arg95 (of R2U). Blue regions indicate areas of highest density. ........................................................................................................ 191

**Figure 4-7**: a) Volume distribution of the defined binding pocket calculated for a subset of conformations for R1U (green), R1C (red) and R2U (green) sequences. b) PLB values distribution for the defined binding pocket calculated for a subset of conformations for R1U (green), R1C (red) and R2U (blue) sequences c) Average binding pockets identified by PocketAnalyzerPCA from X-ray crystal structures (top) and from selected conformations (bottom) of R1U (green), R1C (red), and R2U (blue). Pockets calculated are colour coded according to the weighted average (high frequency - blue; low frequency - red). .................................................................................. 194

**Figure 4-8**: (Top) Molecular surfaces calculated for the highest PLB scoring pocket from R1U (a), R1C (b) and R2U (c) and their top scoring molecular fragment. Molecular surfaces of conformations are coloured red for H-bonding regions, blue for other polar regions and white for hydrophobic regions). Molecular surfaces of molecular fragments are coloured green. (Centre) Top scoring molecular fragment bound to the highest PLB scoring pocket from R1U (d), R1C (e) and R2U (f). (Bottom) Consensus pharmacophores identified from the top 10 scoring docked poses from R1U (g), R1C (h) and R2U (i). ............................................................................................... 197

**Figure 4-9**: Time series of the volumes of the defined binding pocket calculated for selected conformations of R1U (green), R1C (red) and R2U (blue). ...................................................... 206

**Figure 4-10**: Representative high scoring molecular fragment bound to the highest PLB scoring pocket from R1U (a), R1C (b) and R2U (c). Binding sites as predicted by Site Finder are also shown (in grey hydrophobic regions, in red hydrophilic regions). Specifically, fragments ranked #1, #1 and #2 for R1U, R1C and R2U respectively. ................................................................. 207
List of Tables

Table 1-1: Types of ligand conformational search methods and related molecular docking software. ................................................................. 45
Table 1-2: Types of scoring functions and related examples................................................................. 47
Table 2-1: Amino acid sequences and MDM2 binding affinities (Kd) of peptides p53(15-29) (WT), its inverse (I), its reverse (R) and its retro-inverse (RI); and peptides derived from phage display (PMIα and PMIβ). D-amino acids residues are indicated with superscript “D”. ................................................................................. 106
Table 2-2: Population, p (%), of most commonly formed intramolecular hydrogen bonds involving side-chains of peptides WT, I, R and RI from last 20 ns of REMD. ........................................................................................................ 116
Table 2-3: Average interatomic distances d1 (Phe19-Trp23), d2 (Trp23-Leu26) and d3 (Phe19-Leu26) as defined by (a) α-carbons and (b) β-carbons from final 20 ns of REMD for sequences WT, I, R and RI. Standard deviations in parentheses. The d1, d2 and d3 values from crystallographic structure of N-terminal sequence of p53 bound to MDM2 also given (PDB 1YCR[11]). All distances in Å. Population (%) of trajectory of WT, I, R and RI that satisfy d1, d2 and d3 and their simultaneous combination, [d1-d2-d3] over final 20 ns of REMD. ................................................................................. 118
Table 2-4: Average number of hydrogen bonds between the backbone peptide C=O of residues i and the backbone peptide NH of the residues i + 4 and average total number of hydrogen bonds within sequence over final 20 ns of REMD for sequences WT, I, R and RI. Standard deviations in parentheses. ................................................................................. 130
Table 3-1: Population (pop) and average volume (Vav) of the binding pockets identified by PocketAnalyzerPCA for each cluster calculated in the clustering analysis along with an approximate classification of its DFG conformation (DFG conf). Standard deviations in parentheses. Propensity for ligand binding (PLB) of structures possessing the lowest distance from each cluster centroid, calculated by Site Finder. ........................................................................................................ 153
Table 4-1: Conformation (conf), residues present in the binding pocket, conformational state (state), time of the simulation (t (ns)), and volume (V (Å³)) of the three binding pockets with highest propensity for ligand binding (PLB), as calculated by Site Finder. ........................................................................................................ 195
Table 4-2: Percentages of pharmacophore matches within the top 100 and latest 10000 docking poses for each conformation. ........................................................................................................ 199
Table 4-3: Top 10 docking scores registered for the ligands docked onto the highest PLB scoring conformations coming from R1v, R1c, and R2v. ........................................................................................................ 208
Abstract

Alessio Atzori, Doctor of Philosophy, University of Manchester, September 2014

“Conformational analysis of peptides and proteins for drug design using molecular simulations”

The intrinsic plasticity of biological systems provides opportunities for rational design of selective and potent ligands. Increasingly, computational methods are being applied to predict biomolecular flexibility. However, the motions involved in these processes can be large and occur on time scales generally difficult to achieve with standard simulation methods. In order to overcome the intrinsic limitations of classical molecular dynamics, this Ph.D project focuses on the application of advanced sampling computational techniques to capture the plasticity of diverse biological systems.

The first of these applications involved the evaluation of the secondary structure of the N-terminal portion of p53 and its inverse, reverse and retro-inverso sequences by using replica exchange molecular dynamics simulations in implicit solvent. In this study, we also evaluated the effects of reversal of sequence and stereochemistry in mimicking an inhibitory pharmacophoric conformation. The results showed how the ability to mimic the parent peptide is severely compromised by backbone orientation (for D-amino acids) and side-chain orientation (for reversed sequences). Moreover, the structural information obtained from simulations showed good agreement with NMR and circular dichroism studies, confirming the validity of the combination of replica exchange molecular dynamics with the ff99SB force field and Generalized Born solvent model for computational modelling of D-peptide conformations.

In a second work, we explored conformations of the DFG motif of the p38α mitogen-activated protein (MAP) kinase. To achieve this, we employed an advanced sampling simulation method that has been developed in-house, called swarm-enhanced sampling molecular dynamics (sesMD). In contrast to multiple independent MD simulations, swarm-coupled sesMD trajectories were able to sample a wide range of DFG conformations, some of which map onto existing crystal structures. Simulated structures intermediate between DFG-in and DFG-out conformations were predicted to have druggable pockets of interest for structure-based ligand design. Overall, sesMD shows promise as a useful tool for enhanced sampling of complex conformational landscapes.

Finally, we used microsecond MD simulations to evaluate the molecular plasticity of R-spondins, a class of proteins involved in the activation of the Wnt pathway. The unbound R-spondin 1 is characterised by a closed conformation, while, when complexed to proteins LGR and RNF43/ZNRF3, assumes an open and more extended arrangement. This is true also for R-spondin 2, in both its unbound or bound forms. From our simulation, we find that the closed R-spondin 1 conformation is stable, whilst, R-spondin 1 and 2 from their open conformation explore several intermediate structures. In addition, we evaluated the druggability of a potential binding site located at the interface between the second and the third β-hairpin moiety of the first furin domain. The computational screening with small molecular fragments provided interesting insights about the druggability and the pharmacophoric features of the potential binding pockets identified, outlining promising future perspectives of structure-based design of Wnt pathway inhibitors.
Declaration and Copyright Statement

No portion of the work referred to in the 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.

Note on copyright and the ownership of intellectual property right:

I. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

II. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

III. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

IV. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487), in any relevant Thesis restriction declarations deposited in the University Library, The University’s Library’s regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University’s policy on Presentation of Theses.
Acknowledgements

I would like to thank my supervisor Dr. Richard Bryce for offering me the opportunity of being a member of his research group. I am grateful for his constant presence and guidance during the past four years. I would also like to thank my co-supervisors Dr. Pascal Bonnet and Dr. Berthold Wroblowski for their precious help and the constructive criticism. My gratitude goes also to Francesca, Pino and Aatto for introducing me to this amazing research field.

Special thanks goes to the past and present members of Dr. Bryce’s group and to the molecular modelling group of Janssen Pharmaceutica in Beerse for giving me the chance of broadening my horizons, both culturally and professionally. In particular, special thanks goes to Dr. Kepa Burusco Goñi and Irfan Alibay.

I would like to thank Biotechnology and Biological Sciences Research Council and Janssen Pharmaceutica for supporting my Ph.D studies and the University of Manchester for providing the computational resources used in this work.

My gratitude to my parents, my family, my stepfamily, and my beloved friends for their “long-range” support during these years spent away from home.

Finally, I would like to thank the woman I love, Sonia. I am grateful for her immeasurable love, endless support, and infinite patience, in spite of the distance that kept us apart in these years. Without you, this work would not have been possible.
List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Å</td>
<td>Angstrom ($10^{-10}$ m)</td>
</tr>
<tr>
<td>APC</td>
<td>Adenomatous polyposis coli gene</td>
</tr>
<tr>
<td>ATP</td>
<td>Adenosine triphosphate</td>
</tr>
<tr>
<td>β-TrCP</td>
<td>β-transducin repeat-containing protein</td>
</tr>
<tr>
<td>CD</td>
<td>Circular dichroism</td>
</tr>
<tr>
<td>CDK5</td>
<td>Cyclin-dependent kinase 5</td>
</tr>
<tr>
<td>CK1α</td>
<td>Casein kinase 1α</td>
</tr>
<tr>
<td>cWnt</td>
<td>Canonical Wnt pathway</td>
</tr>
<tr>
<td>DKK1</td>
<td>Dickkopf1 protein</td>
</tr>
<tr>
<td>Dvl</td>
<td>Dishevelled protein</td>
</tr>
<tr>
<td>ELISA</td>
<td>Enzyme-linked immunosorbent assay</td>
</tr>
<tr>
<td>FDPB</td>
<td>Finite-difference Poisson-Boltzmann</td>
</tr>
<tr>
<td>FF</td>
<td>Force field</td>
</tr>
<tr>
<td>fs</td>
<td>Femtoseconds</td>
</tr>
<tr>
<td>Fz</td>
<td>Frizzled protein</td>
</tr>
<tr>
<td>GAs</td>
<td>Genetic Algorithms</td>
</tr>
<tr>
<td>GB</td>
<td>Generalised Born</td>
</tr>
<tr>
<td>GBSA</td>
<td>Generalized Born surface area</td>
</tr>
<tr>
<td>GPGPUs</td>
<td>Graphical processing units for general purpose computing</td>
</tr>
<tr>
<td>GSK3</td>
<td>Glycogen synthase kinase 3</td>
</tr>
<tr>
<td>hGH</td>
<td>Human growth hormone</td>
</tr>
<tr>
<td>hGHBP</td>
<td>Human growth hormone receptor</td>
</tr>
<tr>
<td>H-REMD</td>
<td>Hamiltonian replica exchange molecular dynamics</td>
</tr>
<tr>
<td>HTMD</td>
<td>High temperature molecular dynamics</td>
</tr>
<tr>
<td>Kd</td>
<td>Dissociation constant</td>
</tr>
</tbody>
</table>
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>kDa</td>
<td>kiloDalton</td>
</tr>
<tr>
<td>LGR</td>
<td>Leucine-rich repeat containing G protein-coupled receptors</td>
</tr>
<tr>
<td>LRP</td>
<td>Lipoprotein receptor-related protein</td>
</tr>
<tr>
<td>μs</td>
<td>Microsecond</td>
</tr>
<tr>
<td>MAP</td>
<td>Mitogen-activated protein</td>
</tr>
<tr>
<td>MC</td>
<td>Monte Carlo</td>
</tr>
<tr>
<td>MD</td>
<td>Molecular dynamics</td>
</tr>
<tr>
<td>MDM2</td>
<td>Murine double-minute 2 protein</td>
</tr>
<tr>
<td>NMR</td>
<td>Nuclear magnetic resonance</td>
</tr>
<tr>
<td>ns</td>
<td>Nanoseconds</td>
</tr>
<tr>
<td>PB</td>
<td>Poisson-Boltzmann</td>
</tr>
<tr>
<td>PBC</td>
<td>Periodic boundary conditions</td>
</tr>
<tr>
<td>PBS</td>
<td>Phosphate buffer saline</td>
</tr>
<tr>
<td>PBSA</td>
<td>Poisson-Boltzmann surface area</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal component analysis</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein data bank</td>
</tr>
<tr>
<td>PLB</td>
<td>Propensity for ligand binding</td>
</tr>
<tr>
<td>PPIs</td>
<td>Protein-protein interactions</td>
</tr>
<tr>
<td>ps</td>
<td>Picoseconds</td>
</tr>
<tr>
<td>PSO</td>
<td>Particle swarm optimisation</td>
</tr>
<tr>
<td>QM</td>
<td>Quantum mechanics</td>
</tr>
<tr>
<td>REMD</td>
<td>Replica exchange molecular dynamics</td>
</tr>
<tr>
<td>REST2</td>
<td>Replica exchange with solute tempering</td>
</tr>
<tr>
<td>RI</td>
<td>Retro-inverso</td>
</tr>
<tr>
<td>RMSD</td>
<td>Root-mean-square deviation</td>
</tr>
<tr>
<td>RNF43</td>
<td>RING finger 43 protein</td>
</tr>
<tr>
<td>R-spondin</td>
<td>Roof plate specific-spondin</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------------------------</td>
</tr>
<tr>
<td>SesMD</td>
<td>Swarm-enhanced sampling molecular dynamics</td>
</tr>
<tr>
<td>TCF/LEF</td>
<td>T-cell factor/lymphoid enhancer factor</td>
</tr>
<tr>
<td>TFE</td>
<td>Trifluoroethanol</td>
</tr>
<tr>
<td>ZNRF3</td>
<td>Zinc and ring finger 3 protein</td>
</tr>
</tbody>
</table>
The Author

The author attained an MSc in Pharmaceutical Chemistry and Technologies from Cagliari University in September 2009. After a six-month apprenticeship at the Toxicology Department of Cagliari University, he has been undertaking the research documented in this thesis since October 2010. During these four years, he also spent a six-month period as a visiting student at Janssen Pharmaceutica, in Belgium. He is the first-author of two peer-reviewed papers, which constituted Chapters 2 and 3 of this work.

List of Publications


1. Introduction

In a growing number of biological systems, such as peptides and proteins, molecular flexibility is often correlated with biological function. The intrinsic flexibility of biomolecules also provides interesting opportunities that can be exploited in the rational design of new drug entities. Increasingly, computational methods are being applied to predict the flexibility and druggability of biomolecules. However, this has proven to be a challenging task, as not only can the motions involved be large, but they also tend to occur over time scales which are difficult to achieve with standard molecular simulation methods.

The aim of this thesis is to apply advanced sampling methods and long time scale simulations in order to define structural features and conformational changes in biomolecules and apply these to rational drug design. In the current chapter, background information about the methodology used and the biological systems studied in this work is provided. In Chapter 2, the results of simulation-based conformational analysis of peptides directed at inhibiting the p53/MDM2 interaction are presented, with the aim of defining their relation to the available albeit conflicting experimental data. In Chapter 3, the performance of an in-house developed method called swarm-enhanced sampling molecular dynamics (sesMD) is evaluated. This technique is applied to predict the flexibility of protein kinases. In Chapter 4, the plasticity of R-spondins, which play an important role in the Wnt-pathway, is evaluated via long-time scale MD simulations. Additionally, the druggability of R-spondins is evaluated using a combination of different computational techniques.
1.1 Molecular dynamics

The term molecular modelling refers to all the theoretical methods and computational
techniques which can be used to simulate the behaviour of molecular systems. Among
the different computational methods available, molecular dynamics (MD) is a
technique in which the movements of the particles of a system are described with
regards to the Newtonian laws of motion. The concept of using MD simulations was
first introduced in 1957 by Alder and Wainwright, with the study of a system
constituted by hard spheres\textsuperscript{[1]}. Simulations of a dense material were performed in 1959
by Vineyard and co-workers\textsuperscript{[2]}, and were then followed by simulation of liquid argon\textsuperscript{[3]}
in 1964 and by the first attempt at simulating liquid water by Rahman\textsuperscript{[4]}. The first MD
simulation involving a biological system was performed in 1977, where the bovine
pancreatic trypsin inhibitor protein was simulated in vacuo for 9.7 picoseconds\textsuperscript{[5]}

Increasingly, molecular modelling is applied in various fields, including chemistry,
biology and material science, for the purpose of studying different types of molecular
systems and a broad spectrum of phenomena, demonstrating the versatility of this
technique. Due to the exponential growth in computational power it is nowadays
possible to track processes occurring in the millisecond timescale\textsuperscript{[6, 7]} for systems made
of a considerable amount of particles. The information derived from molecular
dynamics simulations can provide new insights at the atomic level for different
biological systems\textsuperscript{[8–10]}, hence feeding into the present and the future of structure based
drug discovery.
1.1.1 Force fields

One of the crucial components of any MD simulation is the definition of a potential function, or rather a description of the physical forces through which the particles of the system will interact during the simulation. Quantum Mechanics (QM) are the most accurate methods for representing these interactions: however, their application to any system composed of a significant number of atoms and involving long time scales is currently impossible, although, recent studies using hardware acceleration showed promising results towards a possible future application.\textsuperscript{[11, 12]} Hence, the reduction from a QM description to a classical potential requires two main approximations. The first one is the Born-Oppenheimer approximation, which allows decoupling of the motions of nuclei and electrons: this approximation is based on the fact that the typical speed of electrons is much faster than that of the nuclei. The second approximation considers the nucleus as a point charge with an associated mass which follows the principles of classical Newtonian dynamics. This latter approximation requires a certain level of empiricism that has resulted in the creation of force fields.

A force field is constituted from a series of equations and force constants which, when combined, describe the potential energy surface of the system of interest as a function of its position and conformation. The force fields most commonly used for biological systems are GROMOS\textsuperscript{[13]}, CHARMM\textsuperscript{[14]} and AMBER\textsuperscript{[15]}. Force fields often differ in their parameterisation, especially for the electrostatic contribution, but they all share a common functional form (defined by Lifson and Warshel\textsuperscript{[16]}), and the functional form of potential energy, which is constituted by two types of contribution: bonded and non-bonded.

\begin{equation} \label{eq:potential_energy}
V_{total} = V_{bonded} + V_{non-bonded}
\end{equation}
The bonded contributions to the potential energy include harmonic type functions for the stretching and the bending of bonds, in addition to series expansions of the cosines of the potential of interaction for torsional angles (Figure 1-1).

![Diagram of bonded and non-bonded interactions](image)

**Figure 1-1:** The four components describing the total energy of the system in a force field

The non-bonded contributions include the Lennard-Jones potential for the description of van der Waals interactions, and a Coulombic potential to describe the electrostatic interactions between the point charges centred on the atoms.

\[
V_{bonded} = V_{bonds} + V_{angles} + V_{tortions} \tag{1.2}
\]

\[
V_{non-bonded} = V_{electrostatic} + V_{van der Waals} \tag{1.3}
\]

Hence, the extended form of the force field describing the total energy of the system could be written as a sum of four components: bond energy, angle energy, torsion energy, and non-bonded interaction energy.
\[ E_{\text{total}} = \sum_{\text{bonds}} \frac{k_i}{2} (l_i - l_{i,0})^2 + \sum_{\text{angles}} \frac{k_i'}{2} (\theta_i - \theta_{i,0})^2 \]

\[ + \sum_{\text{torsions}} \frac{V_n}{2} (1 + \cos(nw - r)) \]

\[ + \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left( 4\varepsilon_{i,j} \left[ \frac{\sigma_{i,j}^{12}}{r_{i,j}^{12}} - \frac{\sigma_{i,j}^6}{r_{i,j}^6} \right] + \frac{q_i q_j}{4\pi \varepsilon_0 r_{i,j}} \right) \] (1.4)

Equation 1.4 represents the AMBER force field, which was predominantly used in its derived forms (namely AMBERff99SB\textsuperscript{[17]} and AMBERff12\textsuperscript{[18]}) in all the studies included in this work.

### 1.1.2 Equations of motions and their integration

In a molecular dynamics simulation, starting from an initial system configuration, successive configurations are generated by integrating Newton’s equations of motion for each particle. Hence, trajectories, which describe changes in positions and velocity against time, are generated. For a particle of mass \( m_i \) subject to a force \( F_{r_i} \) the motion along the coordinate \( r_i \) is found by solving the differential equation constituting Newton’s second law:

\[ \frac{d^2 r_i}{dt^2} = \frac{F_{r_i}}{m_i} \] (1.5)

In this case, the force \( F_{r_i} \) acting on the particle in the system is equal to the derivative of the force field potential \( V(r_i) \) with respect to \( r_i \):

\[ F_{r_i} = -\frac{dV(r_i)}{dr_i} \] (1.6)

Knowing the force acting on the particle at position \( r_i \), it is possible to infer the acceleration, which when combined with the positions and the velocities at time \( t \), can
be used to calculate the positions and the velocities at time $t + \delta t$, where $\delta t$ is a discrete time interval called time step. Forces that act on the particles in the new configuration of the system can then be determined, from which positions at time $t + \delta t$ are determined, and so on, resulting in a trajectory.

In order to simulate a trajectory, the equations of motion are resolved numerically through finite difference methods, which approximate the integration on infinitesimal intervals between consecutive points. In the context of finite differences methods there are numerous algorithms for the integration of the equations of motion, many of which are commonly used in MD calculations. All of these algorithms use the Taylor’s series expansion of position, velocity and acceleration as described in the following equations:

$$r(t + \delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 + \cdots \quad (1.7)$$

$$v(t + \delta t) = v(t) + a(t)\delta t + \frac{1}{2}b(t)\delta t^2 + \cdots \quad (1.8)$$

$$a(t + \delta t) = a(t) + b(t)\delta t + \cdots \quad (1.9)$$

where $v$ is the velocity (or the first derivative of the atomic positions with respect to time) and $a$ is the acceleration, or the second derivative, $b$ is the third derivative of $r$ with respect to $t$.

The Verlet algorithm\textsuperscript{[19, 20]} is a simple and widely used method for integrating the equations of motion. In this method, two third-order Taylor expansions for the positions $r(t)$ are written, one forward and one backward in time.

$$r(t + \delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 \quad (1.10)$$
Introduction

\[ r(t - \delta t) = r(t) - v(t)\delta t + \frac{1}{2}a(t)\delta t^2 \]  
\hspace{1cm} (1.11)

Equations 1.10 and 1.11 can be added to obtain:

\[ r(t + \delta t) = 2r(t) - r(t - \delta t) + a(t)\delta t^2 \]  
\hspace{1cm} (1.12)

In this way, the Verlet algorithm uses positions and accelerations at time \( t \) and positions from time \( t - \delta t \) to calculate new positions at time \( t + \delta t \). As it is possible to infer from Equation 1.12, velocities, which are required to calculate and control temperatures of the simulation, are not explicitly calculated.

To overcome this limitation, the Verlet leap-frog algorithm has been developed.\(^{[21]}\) In this algorithm, velocities are calculated at time \( t + \frac{1}{2} \delta t \) and \( t - \frac{1}{2} \delta t \) as seen previously:

\[ v\left(t + \frac{1}{2} \delta t\right) = \frac{r(t + \delta t) - r(t)}{\delta t} \]  
\hspace{1cm} (1.13)

\[ v\left(t - \frac{1}{2} \delta t\right) = \frac{r(t) - r(t - \delta t)}{\delta t} \]  
\hspace{1cm} (1.14)

These velocities are used to calculate position \( r \), at time \( t + \delta t \):

\[ r(t + \delta t) = r(t) + v\left(t + \frac{1}{2} \delta t\right)\delta t \]  
\hspace{1cm} (1.15)

\[ r(t + \delta t) = r(t) - v\left(t + \frac{1}{2} \delta t\right)\delta t \]  
\hspace{1cm} (1.16)

However, velocities at half time steps can be obtained by substitution into Equation 1.12

\[ v\left(t + \frac{1}{2} \delta t\right) = v\left(t - \frac{1}{2} \delta t\right) + a(t)\delta t \]  
\hspace{1cm} (1.17)
Introduction

In this way, velocities and positions leap each other over like frogs, hence the name of the algorithm. While this method is more accurate than the classical Verlet algorithm, velocities and positions are still not calculated at the same time, resulting out of phase by half a time step.

Velocities at time $t$ can be approximated by this relationship:

$$v(t) = \frac{1}{2} \left[ v \left( t - \frac{1}{2} \delta t \right) + v \left( t + \frac{1}{2} \delta t \right) \right]$$ (1.18)

The trajectories obtained with SANDER module of AMBER simulation package\textsuperscript{[22]}, used in this thesis, are propagated using a variation of the leap-frog Verlet algorithm.

1.1.3 Solvation models

A realistic description of the dynamics and thermodynamics of the biological system of interest can be obtained only if the solvent is accurately modelled. The methods available to evaluate the solvent effect can be separated in two classes: implicit solvent models, which represent the solvent as a continuous medium, and explicit solvent models, where the solvent is composed of distinct molecules.

1.1.3.1 Implicit solvation models

Implicit solvent models are also called continuum models\textsuperscript{[23]}, due to the representation of the solvent as an uniform polarisable medium with a defined dielectric constant $\varepsilon$. The usage of implicit solvent models represent a valid expedient to reduce the computational cost of the simulation by limiting the total number of non-bonded interactions in the system, but still accounting for the impact of solvation on electrostatic interactions.
The solvation free energy ($\Delta G_{solv}$) is the cost in free energy change to transfer a molecule from vacuum to solvent and can be expressed as:

$$\Delta G_{solv} = \Delta G_{elec} + \Delta G_{vdw} + \Delta G_{cav} \quad (1.19)$$

where $\Delta G_{elec}$ is the contribution of electrostatic interactions, $\Delta G_{vdw}$ is the van der Waals contribution of interaction between solute and solvent, and $\Delta G_{cav}$ is the free energy cost required to form a cavity capable of housing the solute.

The electrostatic contribution to the solvation free energy, $\Delta G_{elec}$, can be calculated solving the Poisson equation, which treats the solute as a body with constant low dielectric inside a cavity of the solvent, which is instead characterised as a polarisable continuum with a high dielectric constant. Thus, two regions are represented, one inside and one outside the cavity, in which case the Poisson equation is properly written as:

$$\nabla \cdot \varepsilon(r) \nabla \phi(r) = -4\pi \rho(r) \quad (1.20)$$

The Poisson equation is valid solely under conditions of zero ionic strength. If mobile electrolytes are present in the solvent, the Poisson-Boltzmann (PB) equation\textsuperscript{[24]} is applied:

$$\nabla \varepsilon \cdot (r) \nabla \phi(r) - \kappa' \sinh[\phi(r)] = -4\pi \rho(r) \quad (1.21)$$

where, $\varepsilon$ is the dielectric constant of the solvent, $\phi$ is the electrostatic potential and $\rho$ is the charge density. $\kappa'$ is related to the Debye-Hückel inverse length $\kappa$, which relates to the ionic strength of the solution and measures how far the electrostatic effects extend into the solution. The Poisson-Boltzmann equation can be linearized by performing a Taylor expansion of the hyperbolic sine function, and truncating this expansion after the first term:
\[
\n\nabla \varepsilon \cdot (r) \nabla \phi(r) - \kappa' \phi(r) = -4\pi \rho(r)
\]

(1.22)

The above equation, called Linearized Poisson-Boltzmann equation (LPBE), can be solved using the finite-difference Poisson-Boltzmann (FDPB) method. \(^{[24, 25]}\) In this method, a grid is superimposed onto the solute and the solvent. Values of electrostatic potential, charge density, dielectric constant and ionic strength are assigned to each grid point. The value of the dielectric constant assigned to each point will establish which points of the grid fall into the solute (low dielectric constant) and which fall into the solvent (high dielectric constant). It should be noted that the spacing of the points of the grid is directly proportional to the accuracy and computational costs of this method. In this method, the electrostatic free energy of solvation can be calculated from the work required to move the solute charges from vacuum into the electrostatic potential calculated from the PB equation. In order to overcome limitations inherent to the complexity of the PB equation, different approximation were developed. In the Born model\(^{[26]}\) the electrostatic free energy change for the solvation of a point charge \(q\) in a spherical cavity of radius \(a\) is defined as:

\[
\Delta G_{elec} = -\frac{q^2}{2a} \left(1 - \frac{1}{\varepsilon}\right)
\]

(1.23)

The Generalised Born (GB) method\(^{[27]}\) defines the electrostatic component of the solvation free energy for cavities of any shape as:

\[
\Delta G_{elec} = \left(1 - \frac{1}{\varepsilon}\right) \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{q_i q_j}{r_{ij}} - \frac{1}{2} \left(1 - \frac{1}{\varepsilon}\right) \sum_{i=1}^{N} \frac{q_i^2}{a_i}
\]

(1.24)

where, \(\varepsilon\) is the dielectric constant of the medium, \(q_i\) is the partial charge of the \(i^{th}\) atom, and \(r_{ij}\) is the distance between the \(i^{th}\) and \(j^{th}\) atom and \(a_i\) is called the “Born radius”. The Generalised Born method can be considered as an alternative approximation to FDPB: although less computationally demanding, the precision of the GB approach is
compromised. Hence, GB is considerably a faster method than PB, whilst the latter is, however, a more rigorous and accurate method.

Non-electrostatic contributions to the solvation free energy are represented by solvent-solvent cavity terms ($G_{cav}$) and attractive van der Waals solvent-solute cavity terms ($G_{vdw}$). These terms are usually combined and are represented as:

$$\Delta G_{vdw} + \Delta G_{cav} = \gamma A + b$$

(1.25)

where $A$ is the total solvent accessible area and $\gamma$ and $b$ are constants usually taken from experimentally determined free energies for the transfer of alkanes from vacuum to water. As stated in equation 1.25, these two terms are assumed to be proportional to the total solvent accessible surface area $A$: indeed, these interaction energies will depend upon the number of solvent molecules in the first solvent shell.

The Generalized Born/surface area (GBSA) model combines the GB approach for the electrostatic contribution to the solvation free energy with cavity and van der Waals surface area terms. In the same way, FDPB calculations for the electrostatic contribution to solvation can be summed to the terms accounting for solvent accessible surface area (PBSA) to obtain a more accurate calculation to the free energy of solvation, albeit at the expense of an increased computational cost.

In this work we decided to adopt GB method to model the solvent effect in order to achieve an acceptable compromise between accuracy and computational costs of the replica exchange molecular dynamics method.

1.1.3.2 Explicit solvation models

In explicit solvent models, a defined number of water molecules surround the system of interest, leading to an increase in the number of interactions that must be taken into
account. Therefore, to reduce the size of this additional computational load, the water molecule should be modelled in the simplest way possible.

Water models can be classified according to different parameters, such as the number of points used to define the model (called site), the rigidity or flexibility of the bonds of the water molecule, and the inclusion of polarization effects. In general, the simplest and most computationally efficient water models treat the water molecule as a rigid body and rely solely on non-bonded interactions. More accurate models include multiple interaction sites, which along with the three atoms of the water molecule (three site models, such as TIP3P\textsuperscript{[30]}, SPC\textsuperscript{[31]}, SPC/E\textsuperscript{[32]}, TIPS\textsuperscript{[33]}), could include either a negative charged dummy atom (four-sites models such as BF\textsuperscript{[34]}, TIPS2\textsuperscript{[35]}, TIP4P\textsuperscript{[30]}, TIP4P-Ew\textsuperscript{[36]}, TIP4P/Ice\textsuperscript{[37]}, TIP4P/2005\textsuperscript{[38]}) or a couple a negative charged dummy atoms to represent the lone electron pairs of the oxygen atoms (five-site models, such as BNS\textsuperscript{[39]}, ST2\textsuperscript{[39]}, TIP5P\textsuperscript{[40]}, TIP5P-E\textsuperscript{[41]}) or the combination of both (six-site models\textsuperscript{[42, 43]}). Consequently, the computational cost of a simulation is increased by the number of interaction sites included in the water model adopted.

Two of the most used water models are TIP3P\textsuperscript{[30]} and SPC\textsuperscript{[31]} models. These rigid models have three interaction sites, corresponding to the three atoms of the water molecule. Positive point charges are assigned to the hydrogens along with a negative one to the oxygen atom; dispersion and repulsion forces are calculated only between oxygen atoms using a Lennard-Jones potential. In the different studies presented in this thesis, we adopted the TIP3P water model, due to the fact that AMBER force fields were optimized for this particular water model.
1.1.3.3 Periodic boundary conditions and long-range interactions

The computational time and cost required for a MD simulation increases approximately with the square of the number of particles of the system: consequently, it is necessary to adopt an approximation in order to model a system with an acceptable number of particles. Furthermore, when simulating a box, is it possible to deduce that a significant number of molecules will be close to the edges of the box. If the simulation is run using normal boundary condition, surface forces will be predominant over the bulk ones. Consequently, properties of the system will be different to the ones observed within a macroscopic container. One way to avoid these simulation artefacts is to adopt periodic boundary conditions (PBC, illustrated in Figure 1-2).

![Figure 1-2: Periodic boundary conditions. The central box (in cyan) represents the system simulated surrounded by replicas which possess the same particles (represented as red dots) with the same velocities (indicated by arrows) as the simulation box. $r_{\text{cut}}$, which represent the cut-off radius applied when calculating non-bonded interactions between particles, is also indicated in figure.](image-url)
The simulation box (called unit cell) is surrounded by replicas in order to build a periodic crystal lattice. In this way, when one particle leaves the box in one direction, it will be replaced by an identical particle coming from the opposite direction with the exact same velocity. Thus, the number of atoms in the simulation box is kept constant and particles are not experiencing surface forces. To reduce the computational cost of this method, only the non-bonded interactions between particles registered within a certain distance are evaluated ($r_{\text{cut}}$ in Figure 1-2).

In order to avoid interaction between a given particle and multiple images of another particle, the minimum image convention is used, which states that the cut-off radius must be smaller than half the width of the cell. The combination of periodic boundary conditions and long range electrostatic interactions is evaluated using the Ewald summation\(^{[44]}\), and in particular its derived method called particle-mesh Ewald\(^{[45]}\). In the latter method, short range interactions between particles are calculated using a modification of Coulomb’s Law, while the long-range ones are calculated as a reciprocal space sum using a Fourier transform to build a “mesh” of charges interpolated onto a grid.

### 1.1.4 Temperature and pressure regulation

Given a system with fixed box size and number of particles, a molecular dynamics simulation will sample data under controlled conditions (constant number of atoms, constant volume and constant energy) compatible with the microcanonical ensemble (NVE). In order to simulate other, such as the canonical (NVT) or isothermal-isobaric (NPT) ones, thermostats and barostats have been developed, which allow the regulation of the temperature and pressure of the system.

The temperature of a system is represented by its total kinetic energy via the relation:
\[
\sum_{i=1}^{N_{df}} \frac{m_i v_i^2}{2} = \frac{N_{df} k_b T}{2}
\]

where, \(N_{df}\) is the number of degrees of freedom of the system, \(k_b\) is the Boltzmann constant, and \(T\) the temperature of the system.

This relationship describes how one can control the temperature by rescaling the velocities of the particles of the system. Therefore, different thermostats, such as Berendsen\(^{[46]}\), Nosé-Hoover\(^{[47]}\), Andersen\(^{[48]}\) and Langevin\(^{[49]}\) have been proposed to regulate particle motions in order to maintain constant the temperature of the system.

In the Berendsen thermostat, the system is coupled to an external heat bath with a fixed temperature. At each time step, the velocities of the particles are scaled proportionally to the difference in temperature between the system and the heat bath according to a coupling parameter.

The Nosé-Hoover thermostat introduces an artificial variable into the Lagrangian which couples the system with a heat bath in order to regulate the temperature to the desired one to regulate the velocity on which energy is transferred.

In the case of the Andersen thermostat, the coupling between the system and the heat bath is achieved via stochastic collisions that modify the kinetic energy of the atoms. The frequency of collisions follows the Poisson distribution and the new momentum is drawn from a Boltzmann distribution of temperatures.

The Langevin thermostat is based on the Langevin equation of motion, which is a Newtonian equation of motion including frictional effects and collision effects (soft Brownian dynamics). In this method, the thermostat models the influence of a heat bath adding to the velocity of each particle a random casual force and a frictional coefficient, which is directly linked to the velocity of the particles. These two factors
are balanced in order to keep the temperature constant. In our simulations the Langevin thermostat was adopted.

In a similar fashion, a barostat is introduced to keep the pressure of the system constant (NPT). In our simulations, the isotropic Berendsen barostat\(^{[46]}\) has been used. In this method, the volume of the simulation box is rescaled at each time step according to a coefficient, which allows the pressure of the system to oscillate around the desired value.

### 1.1.5 Limitations of molecular dynamics simulations

The production of a MD simulation must take into account the available computational power. Hence, the size of the system simulated, the time step and the total duration of the simulation must be limited in order to complete a calculation a reasonable period of time. Moreover, only an extended sampling, able to cover the time scales of the phenomena studied, can lead to relevant and valid results.

Due to the constant increase of computational power, MD simulations of a system, which a decade ago could only be simulated for picoseconds, can now easily reach durations of hundreds of nanoseconds, and simulations in the order of microseconds and milliseconds are becoming increasingly common\(^{[6, 7, 50]}\). The main cause for these drastic improvements in computational power has been the advent of parallel computing, which allows for workloads to be distributed over several processors. This has been linked with the introduction of several new hardware technologies, such as graphical processing units for general purpose computing (GPGPUs), which maximise the performance of multithreaded programs.

These technical progresses lead to a major reduction of the wall-clock time in comparison to sequential computing, as documented in the recent literature\(^{[51, 52]}\).
Introduction

Besides their importance and utility in obtaining an atomic level of detail for a system of interest, the methods based purely on classical molecular dynamics are inadequate to sample rare molecular events. Indeed, a well-known issue related to the MD simulations is the local minimum problem, which consists in the inability of overtaking the multiple barriers present on the potential energy surface of a given system. This limitation confines the sampling to a restricted area of the energy landscape of the system, representing an obstacle to obtaining a wide and complete sampling of the whole conformational space. In order to reduce the simulation time spent in local energy basins and to improve the global sample, several advanced computational methods have been proposed.

1.2 Advanced molecular dynamics methods

Following the classification proposed by Adcock and McCammon in their comprehensive review on computational techniques\cite{Adcock2008}, advanced sampling methods can be classified in two separate groups:

i. methods with modified potentials  
ii. methods with modified sampling

The first group includes approaches that alter the potential energy surface using a biasing potential to avoid the obstacle represented by the local minima. These methods include metadynamics\cite{Barducci2009}, umbrella sampling\cite{Berendsen1981} and accelerated molecular dynamics\cite{Ciccotti2004}. The second group is comprised of methods in which explicit modifications of certain parameters of the system: these techniques are used to increase the chance of getting a more extensive sample. These methods include parallel replica dynamics\cite{Kirkwood1935} and temperature accelerated molecular dynamics\cite{Cerda2000}, replica exchange MD (REMD) or parallel tempering\cite{Hukushima1996}.
Introduction

A general description of two of the methods that have been used in this work, namely REMD, our in-house developed method, SWARM-MD and its implemented version, swarm-enhanced sampling MD (sesMD), will be described in the following paragraphs.

1.2.1 Replica exchange molecular dynamics (REMD)

Replica exchange molecular dynamics (REMD) is a method developed by Sugita and Okamoto in 1999\cite{59}. In this technique, a number of independent copies (replicas) of the system $q_0 = [q_{1,0}, q_{2,0}, \ldots, q_{n,0}]$ are run in parallel for a chosen number of MD steps, over a range of different temperatures $[T_1, T_2, \ldots T_n]$. The parallel simulations generate a new set of configurations $q_1 = [q_{1,1}, q_{2,1}, \ldots q_{n,1}]$, between which is attempted a swap: the outcome of the exchange $q_{i,1}$ and $q_{j,1}$ is calculated with a Metropolis acceptance criterion, given in the following equation:

$$P_{accept} = \min \left[ 1, \exp \left( -k_B \left( \frac{1}{T_j} - \frac{1}{T_i} \right) (E(q_{i,1}) - E(q_{j,1})) \right) \right]$$

(1.27)

If the exchange is possible, the temperatures of the neighbouring copies will be exchanged, and the velocity will be scaled and reassigned accordingly to the new temperature; otherwise, the two replicas will continue on their trajectories with the same temperature. In this way, replicas of the system which were trapped in local minima after the exchange have gained the kinetic energy to cross higher energy barriers, while high temperature replicas will have the chance to reach lower energy states. However, to avoid denaturation of the system of interest due to the high temperatures involved, molecular restraints or a limited temperature range needs to be adopted.
Introduction

REMD has been successfully used to sample the free energy landscape of small biological systems\(^{60-62}\), but requires a much higher computational cost when applied to larger systems. Indeed, to keep an acceptable exchange probability between the replicas, a high number of copies with a small difference between the temperatures of neighbouring copies are required\(^{63}\). Moreover, using higher temperatures to coax the trajectories to cross an energetical barrier between two states might alter the ratio between them, promoting the one with larger entropy\(^{64}\).

A possibility to overcome the limitations embedded in temperature REMD has been provided by Hamiltonian Replica Exchange methods (H-REMD).\(^{65-68}\) In these methods, the independent and non-interacting replicas are simulated using different Hamiltonians instead of using different temperatures: the variation of a reduced number of degrees of freedom allows for the inclusion of a lower number of replicas, implying a reduced request of computational resources.

1.2.2 Application of the swarm intelligence method to molecular dynamics simulations

Swarm intelligence methods are based on the observation of the behaviour of animals which tend to act cooperatively as a group like insects, birds or fish. The members of the group show an enhanced behaviour, which is more effective than the one expressed by any single individual within that group. The terms swarm, swarm intelligence and swarm optimization have been defined by Beni in 1989\(^{69, 70}\), referring to cellular robotic systems. The swarm, as it is defined in this work, possesses distinct characteristics: “decentralized control, lack of synchronicity, simple and constituted by (quasi) identical members”.

37
Introduction

The swarm intelligence methods were originally used within the field of molecular modelling along with molecular docking to have a more refined search algorithm\cite{71}, improving the performance of the already existing genetic algorithms, although sharing some common features with them. These similarities include the presence of a population with random solutions and the search for the “best solution” by the update given by the new generations.

These methods have recently been applied to MD simulations with the aim of improving the sampling of the potential energy landscape of biological systems. In 1998 Huber and van Gunsteren developed SWARM-MD, a method where a swarm of particles, represented by molecular trajectories, is used to improve the sampling of the rugged potential energy surface of a molecular system\cite{72}. In this method, each member of the swarm is subject to physical forces, represented by the force field, and to an additional potential (swarm potential), which drives any member toward the average trajectory of the entire swarm. These two sets of parameters act together modelling the dynamics of each swarm particle, both independently and as a part of an ensemble. The swarm potential which acts on all particles can be described as:

\[
V(\{\phi^j\})_{\text{swarm}} = \sum_{j=1}^{M} A \exp[-B d_{\text{rms}}(\phi^j)]
\]

where \(A\) is the maximum strength of the swarm potential (in kcal/mol), \(B\) is the range of action of the swarm potential (in rad\(^{-1}\)), \(M\) is the number of the particles in the swarm and \(d_{\text{rms}}(\phi^j)\) is the root mean square dihedral angle difference, which is used to measure the distance within each particle of the swarm \(j\) from the location of the swarm (in dihedral space), given by:
Introduction

\[ d_{rms}(\phi^j) = \left( \frac{1}{N} \sum_{i=1}^{N} (\phi_i^j - \langle \phi_i \rangle_{swarm})^2 \right)^{\frac{1}{2}} \]  

(1.29)

where \( N \) is the number of dihedral angles used to calculate \( d_{rms}(\phi^j) \), \( \phi_i^j \) is dihedral angle \( i \) of the particle \( j \) of the swarm, \( \langle \phi_i \rangle_{swarm} \) is the average value of dihedral angle \( i \) over \( M \) members of the swarm.

In the original implementation, Huber and van Gunsteren defined the average of each dihedral angle using the mean of the unsigned value:

\[ \langle \phi_i \rangle = \frac{1}{M} \sum_{j=1}^{M} \phi_i^j \]  

(1.30)

In this work, following what has been proposed in our previous implementation\textsuperscript{[73]}, we use the alternative average definition, which takes into account for the sign of the dihedral angles, and their periodic nature.

In this definition, the average value of dihedral angles is given by:

\[ \langle \phi_i \rangle = \arctan \left( \frac{\sum_{j=1}^{M} \sin \phi_i^j}{\sum_{j=1}^{M} \cos \phi_i^j} \right) \]  

(1.31)

where \( \arctan \) is the two argument arctangent function, which returns values in the range \(-180^\circ < x < 180^\circ\).

The effectiveness of the swarm MD method has been proven by simulating long chain alkanes\textsuperscript{[72]}, small peptides\textsuperscript{[72, 73]} and a small protein, Trp-cage\textsuperscript{[73]}. These studies have shown how the average structure of the swarm can reach the global minimum more easily than an independent simulation can do. For example, the results obtained for Trp-cage using SWARM-MD show an increased folding/annealing performance.
compared to the ones obtained running independent unbiased simulations and a reduced RMSD relative to the reference NMR structure.

1.2.3 Swarm-enhanced sampling molecular dynamics (sesMD)

An interesting modification of the SWARM-MD method is represented by the addition of a repulsive term to the swarm potential equation. With this approach, the replicas experience both attractive and repulsive interactions that vary with the distance of the pair in dihedral space, guaranteeing a more dynamic behaviour of the swarm of replicas.

In this work, we adapt swarm MD from an optimisation approach, as described in the previous pages, to a method that permits enhanced coverage of phase space: in this way, multiple MD simulations of a molecular system interact in a cooperative swarm-like fashion to surmount high energy barriers between conformations.

We therefore introduce the following swarm-enhanced sampling (ses) potential, $V_{ses}$:

$$V\left(\{\phi_i^j\}\right)_{ses} = \sum_{j=1}^{M} \sum_{k>j}^{M} \left( A \exp \left[-B d_{rms}^{jk} \left(\phi_i^j, \phi_i^k\right)\right] \right) + C \exp \left[-D d_{rms}^{jk} \left(\phi_i^j, \phi_i^k\right)\right] \tag{1.32}$$

where $A – D$ are suitably calibrated parameters and $d_{rms}^{jk}$ is the root-mean-square distance between swarm members $j$ and $k$. Distinct from eq. 2.28, the term in A and B in equation 1.32 is an attractive Buckingham-like potential that acts between pairs of swarm replicas. This potential is balanced by a repulsive pairwise potential (the term in C and D in equation 1.32). Together, the potential seeks to increase the spread of conformations explored, whilst promoting transitions between wells. The resulting dynamics are therefore a combination of (i) the sesMD potential applied to a selected
set of dihedral angles and acting between pairs of replicas, and (ii) the interatomic potentials of the molecular mechanics force field, which acts individually on each member of the swarm.

1.2.3.1 Implementation and validation of sesMD

The swarm potential, described in the previous paragraph, was implemented into the sander module of AMBER11 and AMBER12. The algorithm has been validated by performing a series of finite difference calculations. The $x$, $y$ and $z$ coordinates of each atom of a test system (namely a capped tyrosine residue Ac-Tyr-NMe) were altered by $\pm \Delta r$, where $\Delta r = 0.00005$. The finite difference force $F$ acting on atom $i$, in coordinate $r$, was calculated according to:

$$F_r^i = \frac{[V_{\text{swarm}}(r + \Delta r) - V_{\text{swarm}}(r - \Delta r)]}{2\Delta r}$$

(1.33)

The forces obtained by the finite difference method were compared to the analytical forces calculated within AMBER. A good agreement between the two sets of calculated forces has been achieved, as showed in Figure 1-3.

Several implementations which improved the functionality of the original SWARM-MD code have been integrated. One implementation is represented by the possibility of specifying a list of dihedrals to be included in the SWARM-MD calculation, instead of including the entire set of $\phi \psi$ dihedral angles of the system of interest, as in the original versions of the code. Furthermore, a general annealing protocol has been included in the code: in this way, each parameter of the $V_{\text{ses}}$ potential (namely A, B, C, and D from equation 1.32) can be increased and/or successively decreased to a specific value in a range of time specified by the user.
Introduction

Figure 1-3: Comparison of forces calculated by sesMD in Amber, and forces calculated from finite difference of swarm energy (red, green and blue show the $x$, $y$ and $z$ components of the forces respectively).

Whilst the concept of improving the conformational sampling by communicating replicas is not innovative, in sesMD the implementation of a potential that models attractive and repulsive interactions between replicas represents an original approach compared to other existing techniques. sesMD is an algorithm inherently suited for parallel computing architectures, hence, its performance is not affected by the required multiple copies of the system studied.

In sesMD, replicas do not require to be subjected to an energy overlap, which is a fundamental requisite of REMD. However, the exchange of information between replicas in sesMD occurs at each time step, in contrast to REMD, where neighbouring replicas could communicate in a wider time range. The sesMD potential can be applied to defined portions of the system of interest, by selecting a reduced set of dihedral angles, while, in other methods, the variation performed in temperature or potential will affect the entire system.
Similarly to other methods which apply a biasing potential (such as metadynamics, and accelerated MD), sesMD requires a careful choice of the degrees of freedom to enhance: however, in the case of sesMD, these are limited to the dihedral angles of the system. Furthermore, sesMD requires a careful parameterisation dependent on the system considered, similar to other biasing methods, such as metadynamics and accelerated molecular dynamics. For the former, if prior information on the nature of the free-energy wells are not available, the scaling parameters are chosen by performing short coarse-grained dynamic runs without bias potential. For the latter, to obtain suitable parameters, an initial estimate of the energy landscape to explore is obtained from simulation in the absence of biasing potential prior to trialling short biased simulations. Nonetheless, sesMD constitutes a promising method to enhance the conformational sampling of a given molecular system.
1.3 Docking

Molecular docking is a technique utilised to predict the preferred binding modes of ligand-receptor complexes, where usually the ligand is a small molecule or molecular fragment and the receptor is a protein. These methods consist of the sampling of the conformational space of the binding site for the protein of interest. Simultaneously, several ligand binding orientations are generated, and each of these possible poses is evaluated using a scoring function in order to identify the most energetically favourable one (the “best match”). Since the 80’s[^4], different docking software packages have been developed: the differences between these programs can be reduced to the treatment of the conformational flexibility of the ligand and/or the protein, the search algorithm employed, and the scoring function adopted to rank the poses generated.

Molecular docking calculations can involve several degrees of freedom: six degrees of rotational and translational freedom of one molecule relative to the other and the conformational degrees of freedom of each molecule. Docking algorithms can be classified according to the number of degrees of freedom considered: generally, in the simplest algorithms, ligand and receptor are treated as rigid bodies, exploring only the six degrees of translational and rotational freedom. In order to allow for the flexibility of the ligand, additional conformational degrees of freedom need to be included. Most of the available docking programs allow for both ligand and receptor to explore their conformational degrees of freedom: however, the treatment of both ligand and receptor as flexible leads to a considerable increase of the computational cost of the calculation. Therefore, when a large library of compounds is screened, the residues of the receptor are kept rigid and only the conformational space of the ligand is explored.
In the following paragraph, we will only discuss the different search algorithms which have been developed to evaluate ligand flexibility.

### 1.3.1 Ligand conformational search methods

The sampling of a ligand conformational space is the most common element featured in molecular docking programs. Ligand sampling algorithms can be divided into three different classes: systematic, stochastic and deterministic algorithms. Examples of ligand sampling algorithms used by docking software are summarized in Table 1-1.

<table>
<thead>
<tr>
<th>Ligand conformational search method</th>
<th>Molecular docking software</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systematic search algorithms</td>
<td>GLIDE(^{[75, 76]}), FRED(^{[77]}), DOCK(^{[74]}), FlexX(^{[78]}), Surflex(^{[79]}), Hammerhead(^{[80]}), eHiTS(^{[81]})</td>
</tr>
<tr>
<td>Stochastic search algorithms</td>
<td>MCDOCK(^{[82]}), ICM(^{[83]}), QXP(^{[84]}), AutoDock(^{[85]}), GOLD(^{[86, 87]}), DIVALI(^{[88]}), DARWIN(^{[89]}), PSO@AutoDock(^{[90]}), SODOCK(^{[91]}), TRIBE-PSO(^{[92]}), PLANTS(^{[93]})</td>
</tr>
<tr>
<td>Deterministic search algorithms</td>
<td>Energy minimization methods, molecular dynamics (MD) simulations</td>
</tr>
</tbody>
</table>

*Table 1-1: Types of ligand conformational search methods and related molecular docking software.*

Systematic search algorithms explore every degrees of freedom of the ligand generating all the possible binding poses. Intuitively, these methods would generate an increasing number of evaluation as the number of degrees of freedom of the ligand rises. In order to overcome this limitation, fragmentation methods, such as anchor-and-grow or incremental construction, were developed. In these methods, the rotatable bonds of the ligand are broken and the largest fragment generated (called the anchor) is fitted into the binding site. Successively, the remaining fragments are linked to the anchor in different orientations, generating binding poses.
Stochastic search algorithms sample the conformational space of the ligand by making random changes to the ligand conformation. Examples of stochastic methods are Monte Carlo (MC) methods, genetic algorithms (GAs) and swarm optimization methods. MC methods generate a random change operating rotating or translating portions of the ligand, which will be accepted or rejected according to a probabilistic criterion. If the pose generated passes the criterion, it will be conserved and further randomized to generate a new pose.

Sampling non-consecutive points in search space leads to the generation of a better binding pose more often and increases the chances of accessing conformations hindered by high energy barriers. GAs apply the principle of survival of the fittest described by the studies of Charles Darwin, in the presence of variation inducing conditions stated by Gregor Mendel. The degrees of freedom of the ligand are encoded in a binary string called gene; the totality of genes represents the chromosome or a defined binding pose. In order to generate diversity, mutations and crossover operations are performed: mutations make random changes to the genes; crossover exchanges genes between two chromosomes.

In this way, GAs generate new binding poses which will be evaluated using a fitness function which determines which members will be kept in the next generation. In a similar manner, the Particle Swarm Optimisation (PSO) utilises random changes, which however do not require genetic operators (mutations and crossovers). This approach, developed by Eberhart and Kennedy in 1995[71], tries to mimic the behaviour of flocking animals. In particular, it consists of a population (called the swarm) of binding poses, which are free to explore the conformational space following the path drawn by optimum particles (or best-ranked binding poses).
In deterministic searches, the previous conformation determines the outcome of the next generated one. Examples of deterministic methods are energy minimization and MD simulations. A well-known issue with deterministic algorithms is that they often get trapped in local minima because they cannot cross high-energy barriers. Hence, these methods are effective in predicting the best binding pose for a small molecule with a limited conformational flexibility, but may not be able to sample molecules with a higher number of rotatable bonds. However, being an effective method in local optimization, deterministic methods could be used to further refine a binding pose obtained with other techniques.

### 1.3.2 Scoring functions

Scoring functions are used to rank the binding poses predicted by the search algorithm and to estimate the binding affinity of a given predicted pose. The importance of their role in determining the accuracy of a docking study is hence obvious. An ideal scoring function should be fast and accurate in evaluating the potential of a binding pose. Several scoring functions have been developed and can be grouped into three basic classes according to their methods of derivation: force field (FF), empirical, and knowledge-based scoring functions. Scoring functions and molecular docking software which adopt them are summarized in Table 1-2.

<table>
<thead>
<tr>
<th>Scoring function</th>
<th>Molecular docking software</th>
</tr>
</thead>
<tbody>
<tr>
<td>Force-field based</td>
<td>AMBER\textsuperscript{[15]}, CHARMM\textsuperscript{[14]}</td>
</tr>
<tr>
<td>Empirical</td>
<td>LUDI\textsuperscript{[94, 95]}, PLP\textsuperscript{[96]}, ChemScore\textsuperscript{[97]}, ChemPLP\textsuperscript{[98]}, GlideScore\textsuperscript{[75, 76]}, LigScore\textsuperscript{[99]}</td>
</tr>
<tr>
<td>Knowledge-based</td>
<td>DrugScore\textsuperscript{[100, 101]}, SMoG\textsuperscript{[102, 103]}, BLEEP\textsuperscript{[104, 105]}, GOLD/ASP\textsuperscript{[106]}</td>
</tr>
</tbody>
</table>

*Table 1-2: Types of scoring functions and related examples.*
Introduction

Force field based scoring functions\cite{107, 108} sum individual atomic interaction terms (such as van der Waals energies, electrostatic energies, bond stretching/bending/torsional energies), derived from experimental data or quantum mechanical calculations, to evaluate the ligand-receptor complex binding energies. In order to increase the performance of the programs adopting FF scoring functions it is necessary to use a cut-off distance to evaluate the non-bonded interactions, which causes however a decrease in the accuracy of the results.

In empirical scoring functions, the binding affinity of a complex is calculated by the sum of a set of weighted empirical energy terms such as VDW energy, electrostatic energy, hydrogen bonding energy, desolvation, entropy, hydrophobicity term, etc. Each component is multiplied by a coefficient and then summed. Coefficients are derived from regression analysis fitted to a test set of ligand-protein complexes with known binding affinities.

Knowledge-based scoring functions are derived from the statistical analysis of protein-ligand complexes training sets. Interatomic contact and distances are derived from this sets and they are ranked according to their frequencies: a more favourable interaction should have a greater frequency of occurrence. The score is calculated assigning a positive score for preferred contacts and a negative score for the disadvantaged contacts. While being computational efficient they are not completely reliable when evaluating infrequent ligand-protein interactions.
1.4 Retro-inverso peptides as inhibitors of p53/MDM2 interactions

In the following paragraphs, we first provide background information about protein-protein interactions, peptidomimetics, and retro-inverso peptides. We also discuss the biological target of our study, p53/MDM2, along with the therapeutic strategies used to antagonize the interaction between these two proteins. This information constitutes the foundation of the work described in Chapter 2.

1.4.1 Protein-protein interactions

Interactions between two or more proteins play a fundamental role in biological systems, being involved in different vital functions, such as the cell-cell communication, signal transduction, immunological responses and the regulation of genes expression. Protein-protein interactions (PPIs) play also a structural role, constituting the quaternary structure of proteins; and a functional role, allowing the formation of receptors, ion channels and enzymes. Incorrect PPIs are known to be involved in several pathological conditions, such as Alzheimer’s disease, prion disease, Creutzfeld-Jacob disease, tumorous conditions and AIDS\cite{109-111}.

Due to the strategic role of PPIs and their connection to major diseases, pharmaceutical research is increasingly directed towards the development of suitable molecules which able to target these interactions.\cite{112} In fact, if classic drug design is mainly focused on molecules capable of targeting a well-defined binding site, nowadays the research is being extended to the development of ligands able to interact with the interface of binding between proteins.
Different reviews about the efficacy of these molecules have been published in the latest years, confirming the increasing interest for these types of targets.[113-119] Nonetheless, the development of molecules which are able to act at the interface between proteins is affected by several limitations:

- The surface of contact between proteins is typically extensive, ranging from 500 and 4900 Å$^2$[120], with an average of 800 Å$^2$ per protein; moreover, these surfaces are also flat, often lacking the grooves and pockets present at the surfaces of other molecular targets[121].

- To cover this wide area of interaction, molecules of considerable size should be designed. However, an increase in dimension, and hence in molecular weight, might negatively affect the bioavailability of such molecules.

- The interaction sites are not rigid and fixed, but they are subjected to conformational and structural variations, which may not be adequately captured by crystallography-derived structures. Moreover, “induced fit” effects linked to the interactions between proteins hinder the identification of a unique and characteristic profile of the site to be targeted.

Nonetheless, several examples of small molecules capable of disrupting the interaction between proteins, and hence presenting a therapeutic activity, are present in literature. An example is represented by paclitaxel or taxol[122], a natural anticancer molecule derived from the *Taxus brevifolia*. This molecule is capable of stabilizing the interaction between microtubules interfering with their breakdown during cell division, causing cell cycle arrest and leading to apoptosis.

However, the actual interaction between proteins often involves a reduced portion of the proteins compared to their global dimensions. An example is the complex formed
by the well-known transcription factor p53 and its physiological negative regulator double-minute 2 protein (MDM2, where M stands for murine). In this interaction, only the 19 residues constituting the N-terminal domain of p53 are involved in binding with MDM2.\textsuperscript{[123]}

In light of these facts, it is legitimate to speculate that a drug designed to target these interfaces does not have to cover the whole interaction area. In fact, on the surface of interaction between proteins, smaller regions which guarantee the essential contacts for interaction affinity have been identified. This hypothesis has been confirmed by several studies, which demonstrate that interactions within proteins are mainly due to small and limited regions, usually located at the centre of the interface area. These regions are called “hot spots”.

1.4.2 Hot spots

A hot spot can be defined as “a residue that, when mutated to alanine, gives rise to a distinct drop in the binding constant, typically tenfold or higher”\textsuperscript{[124]}. Clackson and Wells demonstrated the existence of these regions while studying the interaction of human growth hormone (hGH) with its receptor (hGHBP).\textsuperscript{[125]} In fact, when mutating each residue at the binding interface with alanine, the higher loss of binding free energy was due to only 8 (marked in red in Figure 1-4) of the 31 residues located at the interface of the hGH.
Accordingly to the findings of Bogan and Thorn\textsuperscript{126}, hot spots are characterized by the presence of two different kinds of residues: residues involved in direct binding (mainly amino acids like tryptophan, tyrosine and arginine which are capable of making different kinds of interactions at the same time) and residues that surround these ones, which are energetically less important to the binding interaction. The presence of the latter residues seems to be a necessary, but not sufficient, condition to allow the interaction between proteins by hot spots: these residues are involved in the exclusion of water molecules from the hot spot. However, additional factors, such as hydrophobic and hydrogen bond interactions, contribute to the binding energy.

Hence, broad interaction surfaces between proteins seem to be necessary to create a region where the bulk solvent is excluded, creating a suitable environment for the interaction. This way it is similar to what is observed in other proteins, where the binding site is buried in a deep hydrophobic pocket, which guarantees the exclusion of water molecules.
1.4.3 Peptides as drugs

The structural characterization of the key features of PPI has allowed medicinal chemistry to focus on the design of suitable ligands. Several endogenous small peptides have been used as scaffolds to design drugs capable of interacting with these biological targets: examples are represented by insulin, cyclosporine, oxytocin, renin inhibitors, protease inhibitors, LHRH agonists, HIV-protease inhibitors and analogues of the enkephalin\textsuperscript{[127, 128]}. Moreover, the therapeutic application of peptides has been made easier through the use of methods like solid-phase peptide synthesis (SPPS) developed by Merrifield, which allows one to synthesize peptide sequences with fifty or more amino acids.\textsuperscript{[129-133]}

In spite of the established use of polypeptide drugs as therapeutic agents, a wider application of this class of molecules is limited by several factors:

- Peptides are not structurally stable. In fact, if they are removed from the structural context of a whole protein, they generally do not maintain their structural features and resulting in a loss of biological activity.

- Peptides have a low stability due to their susceptibility to enzymatic and chemical degradation, reducing the possibility for them to be administrated orally. Moreover they are subject to an extensive liver metabolism and an efficient renal excretion.\textsuperscript{[134]}

- Whilst peptides are mildly adsorbed from the gastro-intestinal tract, they are not readily cell permeable, and pass through the blood brain barrier with difficulty. This is due to their high molecular weight or to the absence of specific transport systems.\textsuperscript{[134]}
Due to their flexibility, peptides might assume different conformations and may be able to interact with different receptor distributed along the organism, causing side effects\textsuperscript{134, 135}.

Furthermore, one of the major obstacles in designing new suitable peptidic drugs is represented by their intrinsic flexibility. Indeed, peptides typically assume different conformations in dynamic equilibrium between them: each conformation displays a given combination of torsional angles for each residue. Hence, this flexibility reduces the chance of capturing the active conformation using X-ray crystallography or NMR experiments. Moreover, the lack of relationship between different possible conformations in solution and the receptor-bound conformation presents a major obstacle in this research area (Figure 1-5).

\textbf{Figure 1-5}: Peptides can exist in different conformations. By introducing conformational constraints into the peptide backbone (dashed line), the equilibrium can be shifted towards the conformation which leads to the desired biological effect. In addition, it might be possible to suppress undesired side effects or proteolysis. Image adapted from \textsuperscript{135}.
Introduction

Indeed, the active conformation when interacting with the receptor must assume a characteristic conformation, which might be more subject to degradation or might be involved in undesired side effects interacting with other receptors.\textsuperscript{135}

Even if Ramachandran and co-workers\textsuperscript{136} demonstrated that the possible conformations assumed by peptides are limited to determined values of torsional $\phi$ and $\psi$ torsion angles, still an elevated number of possible conformations are possible for a given peptide, according to its related degrees of freedom. The active conformation might be rarely populated in solution, hence reducing the concentration of the bioactive conformer.

1.4.4 Peptidomimetics

The difficulties in designing an effective peptide to be utilised as drug, described in the previous paragraph, have driven research to the development of peptidomimetics: ligands designed to reproduce the characteristic features of a protein-protein interaction (commonly represented by secondary structure motifs such as $\alpha$-helices or $\beta$-sheets), while eliminating (as much as it is possible) the limiting features of the native peptide.

As indicated by the wide number of publications\textsuperscript{135, 137-139}, several alternative approaches have been proposed: they can be classified in three main groups:

- **Modification of the constituent amino acids of the parent sequence.**
  These modifications might involve the side-chains (β-peptides\textsuperscript{140}) or the backbone of the amino acids (azapeptides\textsuperscript{141}, N-alkylated peptides\textsuperscript{142}, peptoids\textsuperscript{143}, Cα-tetrasubstituted $\alpha$-aminoacids\textsuperscript{144}, retro-inverso peptides\textsuperscript{145}).
Introduction

- **Insertion of structural restraints.** These are modifications of the secondary structure of the peptide forcing it to assume a given conformation (head-to-tail cyclisation\[^{146}\], sidechain-to-sidechain cyclisation\[^{147}\], backbone-to-sidechain cyclisation\[^{148}\], stapled peptides\[^{149}\]).

- **Usage of a non-peptidic backbone.** These molecules contain the functional groups necessary for the ligand-receptor interaction attached to a non-peptidic scaffold, in order to reproduce the structural features of α-helices\[^{150-153}\], β-sheets\[^{154-157}\] and β-turns.\[^{158-161}\]

For the aims of our study, in the following paragraphs we provide an accurate description of retro-inverso peptides and its applications to biological targets.

**1.4.5 Retro-inverso (RI) peptides**

Retro-inverso (RI) peptides represent one of the possible options followed to design a ligand that can mimic the desired PPI and increase the bioactivity and the structural stability of regular peptide ligands. A RI peptide is defined as “an isomer of a linear peptide in which the direction of the amino acid sequence is reversed and the chirality of each amino acid is inverted”.\[^{145, 162, 163}\] In this way, the native L-peptide (peptide 1 in Figure 1-6) sequence is reversed and substituted by D-amino acids (peptide 4 in Figure 1-6). The modification might interest only selected residues or the whole sequence. From the functional point of view, RI peptides appear to be capable of mimicking the biological activity of native sequences assuming a similar orientation of side-chain residues.
Introduction

Figure 1-6: Isomers of linear parent (all-L) peptide (1): (2) enantiomer, (3) retro-isomer (all-L), (4) retro-all-D-isomer, (5) end group modified retro-all-D-isomer. Carbon atoms are represented in white, while the nitrogen ones are in grey. α, β, γ and δ represent various side-chain residues. Image adapted from [145].

The retro-inverso modification does not stabilise the sequence from a structural point of view, since the typical flexibility of a peptide chain is preserved. However, peptidases (biological enzymes responsible for proteins degradation) are not able to recognize to D-amino acids, conferring to RI peptides an augmented resistance to proteolytic degradation. Another feature of the RI modification is the inversion of the terminal charges of the sequence: if the native sequence showed a positive charge at the N-end, the RI sequence displays a negative one, and of course, this is also valid for the C-end. This issue has been solved by modification of the terminal residues, or inserting the RI sequence into larger ones.
Examples of both successful and unsuccessful applications of RI peptides can be found in literature, but the research is still ongoing to assess the potential of their therapeutic application. Their efficacy in inducing an immunological response has been assessed\cite{164, 165}, but the general opinion is that their activity is due to the lack of specific structural interaction required for the antibody-antigen recognition, hence allowing the efficacy of RI peptides. Nonetheless, RI peptides have found to be active as anti-aggregation drugs for β amyloid in Alzheimer’s disease\cite{166, 167}, as mimetics of SMAC\cite{168} and prosaposin peptides\cite{169}, and in the treatment of terminal peritoneal carcinomatosis\cite{170}.

In our study, we evaluated the ability of RI peptides to inhibit the interaction between the tumour-suppressor p53 and its natural substrate MDM2. For this reason, in the following paragraph we discuss p53, its therapeutic significance and structural information of its interaction with MDM2.

### 1.4.6 p53

The tumour suppressor p53 (where “p” stands for protein and 53 is its molecular weight in kDa) is a transcription factor deeply involved in cellular response to particular conditions, such as hypoxia, telomere shortening, DNA damage, or other cellular stressful situations\cite{171}. In these conditions, an increase of the cellular levels of p53 is registered, due to a post-translational mechanism, which intervenes to prevent the propagation of the effects of harmful conditions and/or mutations. p53 plays its role by suppressing cellular growth or engaging apoptosis. In the first process, p53 activates the transcription of the WAF1/Cip1 gene, which is followed by the expression of p21, a cyclin-dependent kinase inhibitor. The latter action is performed allowing the transcription of pro-apoptotic genes like NOXA and PUMA (Figure 1-7).
Due to its pivotal role in the response to DNA damages, p53 has been called the “guardian of the genome”.\cite{172, 173}

![Diagram](image)

**Figure 1-7**: DNA damage or cellular stressful situations leads to the dissociation of the p53 and MDM2 complex. Upon activation, p53 can activate p21 leading to the arrest of the cellular cycle or engaging apoptosis via transcription of pro-apoptotic genes.

Under physiological conditions, levels of p53 are regulated via a negative feedback mechanism by MDM2. The increase of p53 concentration causes the expression of MDM2, which binds the p53 to the α-helical transactivation domain near to its N-terminus, blocking the activity of p53 in three different steps:

i. MDM2 binds to the transactivation domain preventing the p53 to interact with DNA;

ii. MDM2 then exports p53 out of the nucleus;

iii. MDM2 is involved in its enzymatic degradation, acting as an E3 ubiquitin ligase.
Both p53 and MDM2 have a short half-life (10-20 minutes), and their levels are generally kept low by the intervention of the regulatory feedback mechanism\(^{173}\). MDM2 tends to be overexpressed in several oncogenic conditions\(^{174-176}\). This causes an excessive diminution of p53 levels and, consequently, an uncontrolled proliferation of tumoral cells. Hence, the possibility of interfering with the formation of the p53-MDM2 complex represents a critical target in cancer therapy. In fact, inhibiting this interaction might re-establish the physiological levels of tumour suppressor p53, which can bind to DNA and promote the transcription of the genes necessary to the arrest of the cell cycle or the apoptosis\(^{177}\). Besides MDM2, different proteins have demonstrated to play a role in the control of p53 recently: inhibitors of p53 like MDM4\(^{178, 179}\), and transcription activators of p53 like p300\(^{180}\). These targets might represent an alternative to the “classical” p53/MDM2 inhibition.

The physiological state of p53 limits the application in therapy of inhibitors of this interaction: in fact, almost 50% of tumorous diseases are characterised by either mutation or deletion of the p53 gene (TP53).\(^{181}\) These mutated forms are unable to perform their physiological actions, therefore making the inhibition of its interaction with MDM2 useless.\(^{173, 182}\) Nonetheless, 7.2% of 3,889 human tumours present increased levels of MDM2 without mutations in p53\(^{183}\), representing a target for inhibitors of p53/MDM2 interaction.

1.4.7  Structural characterization of the p53/MDM2 interface

The results obtained from the crystallographic study made by Pavletich and co-workers in 1995\(^{123}\), permit atomic insight of the surface of interaction between p53 and MDM2. Even if the p53/MDM2 interface includes a 1498 Å\(^2\) surface area, this complex is far away from representing a classical protein-protein interaction.
As mentioned in paragraph 1.4.1, PPIs usually rely on the presence of several hot spots, located at the flat and extended interface of the proteins. On the contrary, the presence of a single deep hydrophobic pocket on the surface of MDM2 makes this interaction an exception to this rule.

The co-crystal structure of this complex shows how the N-terminal domain of MDM2 presents a deep hydrophobic cleft where p53 binds with its N-terminal portion (residues 17-29). This interaction is granted by the complementarity of these two moieties and, in particular, to the presence of three amino acid residues of p53 (Phe\textsuperscript{19}, Trp\textsuperscript{23} and Leu\textsuperscript{26}), which fit deeply into the hydrophobic pocket of MDM2. Only the residues from 18 to 26 of the p53 peptide assume an α helical conformation, while the terminal residues possess an extended conformation (Figure 1-8).

**Figure 1-8:** a) Side view (left) and top view (right) of N-terminal sequence of p53 (residues 17–29, in green) bound to the MDM2 (residues 25–109, in white) derived from its crystal structure (PDBID 1YCR\textsuperscript{123}). The amino (N) and carboxyl (C) termini of p53 are indicated. b) Frontal view of the N-terminal sequence of p53 (green) bound to the MDM2 (white) derived from its crystal structure. Backbone atoms are represented in cartoon, while the residues Phe\textsuperscript{15}, Trp\textsuperscript{23} and Leu\textsuperscript{26} are in sticks.
The importance of the presence of Phe\(^{19}\), Trp\(^{23}\) and Leu\(^{26}\) residues has been confirmed by several studies, where the mutation of these residues is related to the unsuccessful binding with MDM2 and reduced target gene transactivation\(^{184,185}\). Even the presence of Asp\(^{21}\) has found to be relevant to the p53/MDM2 interaction, not from an interactive point of view, but from a structural one: in fact, the substitution of the Ala\(^{21}\) causes the loss of an existing hydrogen bond with Thr\(^{18}\), which guarantees the stability of the \(\alpha\) helical structure\(^{185}\).

Hence, from the characterization of the interaction site of MDM2 and the conformation assumed by the N-terminal portion of p53, several approaches have been investigated, trying to find a suitable molecule which, mimicking the position of the three amino acids, could interact with the binding site of MDM2, inhibiting the interaction with p53, and producing the desired pharmacological effect. We discuss these approaches below.

### 1.4.8 Pharmacological inhibitors of p53/MDM2 interaction

It is possible to classify the molecule capable of acting as inhibitors of p53/MDM2 interaction in three different chemical classes: natural molecules, small synthesised molecules and peptidomimetics. We will briefly discuss the most important exponent of each class.

Chalcones (1,3-diphenyl-2-propen-1-ones) were the first reported natural antagonists of the p53/MDM2 interaction\(^{186}\). Unfortunately, these molecules are characterized by low potency and inhibition of glutathione-S-transferase as main limiting side effect. Recently, novel boronic–chalcone derivatives have been reported as effective antagonists characterized by their antitumour effect\(^{187}\). Another example of natural p53/MDM2 inhibitor is chlorofusin\(^{188}\), derived from the fungus Microdochium
caepitosum, identified by Duncan and co-workers. Nonetheless, chlorofusin does not represent a candidate drug, due to its complex structure and related high molecular mass, but it can be used as a lead for the design of more efficient analogues.

The most important member of the class of synthesized inhibitors is represented by nutlins\textsuperscript{189,190}. These molecules are cis-imidazoline derivates synthesized to mimic the N-terminal portion of p53. The constituent functional groups possess the chemical and physical characteristics to interact with the binding site of MDM2. The helical backbone of the p53 peptide is substituted by the imidazoline portion which spaces with the correct geometrical orientation the three groups into the pockets normally occupied by the three key residues of p53. This occupancy is related with \textit{in vitro} and \textit{in vivo} activity. Others example of successful synthesized inhibitors include: sulfonamides\textsuperscript{191}, syc-7\textsuperscript{192}, benzodiazepinediones\textsuperscript{193, 194}, spiro-oxindoles\textsuperscript{195} and isoindolinones\textsuperscript{196}.

The last class is represented by molecules which have been designed as a scaffold capable of mimicking the three-dimensional arrangement of the residues necessary to the binding with MDM2. One of the most successful exponents of this class is represented by stapled peptides. These molecules mimic the natural peptide and show an increased cell penetration and resistance to degradation compared to other peptides.\textsuperscript{149} Stapled peptides are synthesized according to the natural sequence, except for two structurally optimized amino-acids which yield an all-hydrocarbon staple. Bernal and co-workers\textsuperscript{197, 198} demonstrated how stapled peptides possess a high affinity for the MDM2 binding site compared to the wild type p53 peptide. Hence, the increased helicity, due to the presence of the hydrocarbon staple, increases the inhibitor activity of stapled peptides. Other examples of peptidomimetics active on p53/MDM2 interaction are peptoids\textsuperscript{199, 200}, terphenyls\textsuperscript{201}, cyclic $\beta$ hairpin
Introduction

peptides\textsuperscript{[202]}, \(\beta\) peptides\textsuperscript{[203]}, p-oligobenzamides\textsuperscript{[204]} and retroinverso peptides, which represent the main target of interest in our study.

1.4.9 RI peptides and p53/MDM2

In 2004 Sakurai and co-workers\textsuperscript{[205]} investigated possible inhibitors for the p53/MDM2 system. Knowing the main structural features involved in this interaction, they hypothesised that the backbone of the p53 portion involved in binding played only the role of a scaffold, permitting the correct orientation of the three key residues and their spacing. Hence, they investigated this hypothesis by focusing on the results obtained using four different peptides: 1) the wild-type peptide, 2) the enantiomer of the parent sequence (D-peptide or \textit{retro peptide}), 3) the L-retro-peptide, (a peptide with the opposite order of residues of the parent sequence and made by L-amino acids or \textit{inverso peptide}), and 4) the retro-inverso peptide. The side-chains of peptides 1 and 3 are not superimposable with those of 2 and 4 because they are mirror images of each ones and the backbones of 1 and 2 are not superimposable with those of 3 and 4 due to the opposite orientation of the sequence.

Therefore, Sakurai and co-workers hypothesized that if 4 assumed a left-handed helical conformation, its side-chains would not align with those of the wild type peptide, which assumes a right-handed helical form, but if 4 could adopt some kind of a right-handed arrangement, the superimposition would be possible. The helical propensity of each peptide was measured by circular dichroism (CD) experiments, showing how peptides 2 and 4, composed of D-amino acids, form a right-handed helix, while and 1 and 3, made by L amino acids, form a left-handed helix. The biological activity of the set of peptides has been tested by inhibition ELISA experiments: peptides 2 and 3 showed a reduced affinity for MDM2, while the RI sequence presented an inhibitory
activity comparable to the one possessed by the natural sequence. Surprisingly, the inverted chirality of the constituent amino acids and the reversed direction of the sequence of the peptide do not reduce the ability of the RI peptide to act as an inhibitor of the p53/MDM2 interaction.

To explain this result, they hypothesized that even if RI peptides preferentially adopt a left handed helical conformation, in this particular situation, 4 must adopt a conformation constituted by two successive right-handed turns: only in this way this ligand can interact with MDM2 as the parent peptide. Nonetheless, no structural evidence to confirm their theory was provided in the publication. Trying to clarify the findings of Sakurai and co-workers, in 2010, Li and co-workers[206] performed an exhaustive investigation on the same set of peptides studied by Sakurai. In addition, to investigate if their findings apply also to other systems, two additional peptide-protein systems were analysed: the C-terminal domain of HIV-1 2 capsid protein (CA) and the Src homology 3 (SH3) domain of Abl tyrosine kinase with their respective peptide ligands. Using biochemical (inhibition ELISA), biophysical (isothermal titration calorimetry), and structural (CD, fluorescence spectroscopy, NMR spectroscopy) tools of analysis, Li and co-workers showed that RI peptides adopted a left-handed helix with a stronger propensity for assuming this conformation than the parent peptide.

Moreover, they challenged the results of Sakurai and co-workers, finding that the binding affinity of the RI peptide to the MDM2 is consistently lower compared to the wild type p53. The same considerations are valid for the other studied systems, with a consistent diminution of the binding free energy of the RI peptides against the parent sequences. In conclusion, the findings of Li and co-workers challenged the view of RI peptides as valuable peptidomimetics.
Introduction

The main aim of our study is to examine the structural features possessed by RI peptides using computational methods, in order to establish their potential activity as peptidomimetics. Hence, molecular dynamics simulations are used to assess the flexibility of the RI peptides conformation and to evaluate their ability of assuming a conformation superimposable to the one assumed by the parent peptide of p53 in the bound state with MDM2. Furthermore, this study will also examine the influence of the sampling method and the force field adopted on the prediction of the structural features of RI peptides. To this end, we will run simulations of the different sets of peptides, composed of native and retro-inverso peptide pairs, using different combinations of sampling methods (namely REMD and classical MD) and force fields (AMBERff99SB and AMBERff96), in order to observe if and how the results will be affected by the different combinations of these parameters.
1.5 Protein kinases and DFG motif

In the following paragraphs we first provide functional and structural information about protein kinases. Furthermore, one of the most important structural feature of protein kinases, the DFG motif, will be presented. The information discussed in this section delineates the underpinning for the study described in Chapter 3.

1.5.1 Protein kinases

Protein kinases (or phosphotransferases) are the largest group of enzymes available among all the living organisms: all the different sub-types of kinases belong to a family which has maintained common features for each member\(^\text{[207]}\). It is estimated that the human genome is able to encode for about 500 or more kinases (which constitute about 2\% of all human genes\(^\text{[208]}\) but not all of them have been characterized from a structural and functional point of view. Due to their important role in different biological processes and to their wide location, the deregulation, misregulation or mutation of kinases have been associated to different diseases including cancer, diabetes, metabolic disorders, inflammatory processes, vascular disorders, central nervous system and autoimmune diseases\(^\text{[209-211]}\). It is, therefore, unsurprising that kinases are a major drug target in pharmaceutical research\(^\text{[209]}\).

Protein kinases are enzymes capable of transferring a phosphate group from high energy donor molecules (like the $\gamma$-phosphate group of adenosine triphosphate, ATP) to a specific substrate in a reversible process called phosphorylation. According to the biochemical nomenclature, the correct name for a given kinase should be constituted by "[substrate name] kinase" (the name hexokinase, for example, is indicative of a kinase that catalyses the phosphorylation of a hexasaccharide).
Kinases can perform their activity on very diverse substrates, such as small molecules and proteins. Phosphorylation of small molecules (such as lipids, carbohydrates, amino acids, nucleotides, vitamins and cofactors), activates them in signalling or metabolic processes. An example is hexokinase, which acts on the first step of the glycolysis by phosphorylating a glucose molecule to give glucose-6-phosphate (Figure 1-9).

![Figure 1-9: Example of a phosphorylation in position 6 of glucose performed by the hexokinase to obtain glucose-6-phosphate](image)

However, kinases achieve their main biological role by modifying the activity of up to 30% of all proteins[212]: in fact, the phosphorylation of a target protein might cause an activation or inactivation of the target protein, the association with other proteins, or promote a change of the position of the protein inside the cell. Moreover, the phosphorylation process is involved in many crucial cellular pathways (cell cycle division, cell proliferation and growth, apoptosis).[213] The important role occupied by protein kinases implies a high degree of regulation of their activity, or rather a fine regulation of their activation or inactivation is necessary. This regulation processes mainly involve the phosphorylation of the kinases (that sometimes can be an auto phosphorylation made by the kinase itself or by other kinases) or the binding of other proteins or small molecules to the kinases.
Introduction

In order to phosphorylate a protein, the phosphate group from the ATP molecule has to be attached to particular amino acid residues that present a hydroxyl group: serine, threonine and tyrosine. Regarding their targets, kinases can be divided into two main classes: the serine-threonine kinases and the tyrosine kinases, although it has been shown that some kinases can act on all three amino acids (dual specificity kinases\textsuperscript{[214]}). Recently, kinases which can act on histidine, in spite of its lack of a hydroxyl group, have also been discovered.\textsuperscript{[215]}

1.5.2 Conserved structural features of protein kinases

All the members of the protein kinase family share a conserved common catalytic domain\textsuperscript{[216-220]}, which consist of 250-300 amino acid residues.

\textbf{Figure 1-10}: Crystal structure of a protein kinase (PDBID: 1CSN) in complex with ATP molecule. On the right, close-up of the ATP binding site: ATP and interacting residues of the binding site are shown in sticks.
Introduction

The general structure of a protein kinase (Figure 1-10) is constituted by two domains: a small N-terminal with a binding function and a C-terminal portion. The former portion is composed by antiparallel β sheets, connecting loops and one important alpha helix, called α-C helix (which plays an important role in the catalysis of the ATP molecule).

Between the connecting loops, the P-loop (phosphate binding group or Waker loop) is an important conserved region between kinase proteins; this loop consists in a glycine rich sequence\cite{221} followed by a conserved lysine residue\cite{222}, involved in the correct stereochemical coordination of the α and β phosphates of the ATP.\cite{221, 223, 224} Most of the interactions with the ATP molecule are related to this lobe.

The C-terminal portion, which possess catalytic and peptide binding functions, is mainly formed by α helices and several β strands.\cite{225} Between the beta strands are located two important conserved regions: the catalytic loop and the metal binding loop (which contains an important conserved motif). The catalytic loop is constituted by residues which are directly involved in catalysis. In this region is located a conserved aspartate residue, which is thought to be involved in receiving the hydrogen atom of the hydroxyl group of the substrate protein.\cite{226}

The metal binding loop contains an important conserved\cite{227} Asp-Phe-Gly motif (called the DFG motif) involved in the coordination of the ATP-bound Mg$^{2+}$ion, which in turns coordinates the β and γ phosphates of ATP. The DFG motif is followed by the activation loop, which contains a threonine or a tyrosine residue, involved in the activation process. These two lobes are joined by a loop called the hinge, which also plays a role in the stabilization of the adenosine of ATP via hydrogen bonds. The cleft, made up by the two lobes, forms a binding pocket which provides accommodation for the ATP molecule, with its triphosphate moiety pointing outwards of the cavity.
Protein kinases are known to adopt two different activation states: an inactive (or closed conformation and an active (or open conformation) state. The transition between these two states, and hence the activity of protein kinases, is associated with different conformational transitions of its components.

The phosphorylation of the threonine/tyrosine residue in the activation loop (via auto phosphorylation or via other activating kinases) leads to its displacement from the catalytic centre, enabling the binding of the substrate. Along with this motion, the rotation of the N-lobe allows the formation of an ion pair between a conserved glutamate residue, part of the α-C helix, and a conserved lysine residue, part of the P-loop. As it goes through the catalytic cycle, the active kinase toggles between the two available conformations.

1.5.3 DFG motif

Among the conserved structural features mentioned above, DFG motif represent an important and highly conserved sequence across the protein kinase family, playing a major role in the coordination of the ions bounded to the ATP molecule. Experimental studies have proved that the DFG motif is characterized by two different conformations called DFG-in and DFG-out. The difference between these two conformations consists in the orientation of the aromatic ring of the phenylalanine169, which is facing the ATP binding site in the “in” conformation (Figure 1-11, in green). In the “out” conformation, the phenylalanine ring is turned towards the solvent and interferes with the binding of the ATP molecule (Figure 1-11, in red) and hence characterising the inactive state of protein kinases.
Introduction

Figure 1-11: Superposition of 1P38 (green) and 1WBT (red) X-ray crystal structures of p38α MAP kinase. Backbone atoms are represented in cartoon while residues Asp168 and Phe169 are in sticks. Picture adapted from [229].

DFG motif represents an important site of interaction for class II kinase inhibitors, which bind to the ATP site and to an additional allosteric site. Indeed, the movement of the aromatic ring of the phenylalanine to the out conformation, along with the displacement of the αC-helix, creates a hydrophobic pocket which suits a particular moiety of the inhibitor molecule. Many structures of kinases have been classified in the DFG-in conformation, whereas fewer DFG-out conformations have been observed only in presence of class-II and class-III kinase inhibitors.

The nature and mechanism of the DFG flip is still unclear, and the experimental studies [230-232] available have contributed to the formulation of two main hypothesis:

i. The DFG flip is due to the binding of allosteric type-II inhibitors. The inhibitors can bind to the kinase in a DFG out conformation, or they bind in the “in” state and cause the transition to the “out” conformation.

ii. The two different conformations co-exist in a dynamic equilibrium.
Introduction

Experimental studies have also highlighted the presence of intermediate conformations\textsuperscript{[233]} between the two already known states, which can contribute to the interaction of inhibitors with protein kinases.

In order to have a more detailed insight about this molecular mechanism, the DFG loop has also been studied using several computational techniques. Fremberg-Kesner and Elcock applied high temperature molecular dynamics (HTMD) simulations to sample the DFG motif of p38α MAP kinase in both directions, identifying two intermediate conformations defined as pseudo-DFG-in and pseudo-DFG-out.\textsuperscript{[234]} Berteotti and co-workers described the “open-to-closed” transition of cyclin-dependent kinase 5 (CDK5) using a sampling method able to find the lowest free-energy pathway between an initial and final state, defined by the crystallographic structures available. The two-step mechanism identified by their simulations is characterised by the presence of a energetically stable intermediate conformation.\textsuperscript{[235]}

In another study, Filomia and co-workers performed non-equilibrium MD simulations applying steering forces to the side-chains of the DFG loop of p38α MAP kinase. The simulations drew an accurate description of the interactions involved in the DFG transition, with a primary role played by Phe169 of the DFG loop. This residue establishes cation–π interactions which stabilises intermediate conformations and facilitates the transition to a DFG-out state.\textsuperscript{[236]}

Shan and co-workers, in a combined computational and experimental study, described the conformational change of Abl kinase, emphasising the importance played by the protonation state of the aspartate residue of the DFG motif. These findings are supported by experiments which demonstrate how the kinetics of the binding of Imatinib to Abl kinase shows a pH dependence that is lost upon the mutation of the aspartate residue.\textsuperscript{[237]}
Introduction

The main objective of our study is represented by the reproduction of the previously described conformational change involving the DFG motif of protein kinases using our in-house developed method, called swarm-enhanced sampling molecular dynamics (sesMD). The results obtained from the sesMD simulations are compared with the ones from classical MD in order to prove the augmented sampling capacity of sesMD. The necessity of using biased methods comes from the general observation that rare molecular events, like the DFG conformational change, need time scales that are difficult to reach using classical MD simulations. Moreover, the augmented sampling guaranteed by sesMD allowed for the identification intermediate conformations between DFG-in and DFG-out which are predicted to have druggable pockets of interest for structure-based ligand design.
1.6 R-spondins and the canonical Wnt pathway

In the following section we first provide background information about the canonical Wnt pathway and a class of proteins involved in the regulation of its activity, R-spondins. This outlines the underlying basis for the work detailed in Chapter 4.

1.6.1 Canonical Wnt signalling pathway

Wnt proteins are a large family of conserved glycoproteins\textsuperscript{[238]} involved in the regulation of several developmental processes such as cell fate determination, cell proliferation and tissue homeostasis.\textsuperscript{[239-241]} Hence, mutations and misregulation of the Wnt pathway are typically associated with tumours and birth defects.\textsuperscript{[239]}

Wnt signalling is regulated by three different pathways: a canonical pathway (cWnt), and two non-canonical ones, namely planar cell polarity pathway and Wnt/Ca\textsuperscript{2+} pathway.\textsuperscript{[242]} cWnt, also known as β-catenin-dependent pathway, is the most characterized Wnt pathway, and it will also be the one described in the following paragraph.

In the absence of Wnt ligands, the Wnt receptor complex, composed of the transmembrane Frizzled (Fz) protein and its co-receptor lipoprotein receptor-related protein 5/6 (LRP5/6)\textsuperscript{[243]} is disrupted by the secreted protein Dickkopf1 (DKK1). DKK1 binds to the transmembrane protein Kremen2 and to LRP5/6 forming a ternary complex, which induces endocytosis and consequent removal of LRP5/6 from the membrane.\textsuperscript{[244]} In this situation, the cytoplasmatic β-catenin is degraded by a complex composed of Axin, adenomatous polyposis coli gene product (APC), casein kinase 1 α (CK1α), and glycogen synthase kinase 3 (GSK3).\textsuperscript{[243, 245-248]} GSK3 and CK1α phosphorylate the conserved serine and threonine residues of β-catenin, which is hence recognized by the β-transducin repeat-containing protein (β-TrCP) and successively...
ubiquitinated and degraded by the proteasome complex.\cite{243,249} The degradation of β-catenin precludes its entry in the nucleus: in this way, the repressor protein Groucho binds to members of the T cell factor/lymphoid enhancer factor (TCF/LEF) family of transcription factors, preventing the transcription of cWnt target genes\cite{243} (Figure 1-12).

The binding of a Wnt ligand to its receptor, composed of Fz and LRP 5/6, leads to its internalization followed by the activation of intracellular protein Dishevelled (Dvl). Dvl inhibits the degradation of β-catenin by presenting FRAT/ GBP (a cellular GSK3 inhibitor) to the degradation complex.\cite{250-252} In addition, the phosphorylation of LRP5/6 by CK1γ and GSK3 leads to the binding of Axin to the membrane, causing the disassembly of the degradation complex.\cite{253,254} Ultimately, the entirety of these events promotes the accumulation of cytoplasmatic β-catenin, followed by its entry
Introduction

into the nucleus, where it forms a complex with TCF/LEF and activates the transcription of cWnt target genes (Figure 1-12).\[255-257]\]

1.6.2 R-spondin proteins

The genome of vertebrates encode four different roof plate specific-spondin (R-spondin1–R-spondin4), which are known to act as agonists of the cWnt pathway. These proteins, with an approximately size of 35 kDa, share 40-60% of amino acid sequence identity.\[259]\ All four R-spondin family members possess a common general structure (Figure 1-13): a N-terminal signal peptide; two cysteine-rich furin-like domains, involved in the activation of cWnt signalling; one thrombospondin type 1 domain (TSP1); a C-terminal region.\[260-262]\]

![Figure 1-13: Structure of unbound R-spondin 1 (PDBID 4BSO). The two furin domains are indicated as Fu1 and Fu2. Disulphide bonds and disordered residues are shown respectively as ball and stick and dashed lines. Image adapted from [263].](image)

Kamata et al. are accountable for the identification of R-spondin1 from the screening of genes expressed during embryonic development in the central nervous system of mice.\[260]\ In this study, they were also able to observe that the expression of R-spondin1 was reduced in Wnt-1/3a double knockout mouse, providing the evidence for a link between R-spondin1 and the cWnt pathway.
In the same year, the findings of Kazanskaya and co-workers\cite{261} supported the theory formulated by Kamata. In their study, they showed how R-spondin2 and 3 act a positive modulator of Wnt/β-catenin pathway and are co-expressed along with several cWnt ligands. In the following years, other several studies have established the role of R-spondin proteins as activators of the cWnt signalling pathway and provided useful insights on the activity of each member of the R-spondin family.\cite{261, 262, 264-269}

Asides from its involvement during the developmental stages, R-spondin1 plays a major role as a regulator of sexual development and its presence is fundamental for the development of the female reproductive system.\cite{265, 270, 271} R-spondin1 has also proved to possess potential therapeutic effects via the activation of the cWnt pathway in several models of inflammation-associated diseases: R-Spondin1 enhances basal layer epithelial regeneration and accelerates the mucosal repair in chemotherapy or radiation-induced oral mucositis models\cite{272}; R-spondin1 can act as a growth factor, being a mitogen for gastrointestinal epithelial cells in normal mice, and also proved to be beneficial in an animal model of inflammatory bowel disease by “stimulating crypt cell growth, accelerating mucosal regeneration, and restoring intestinal architecture”\cite{273, 274} R-spondin1 proved to be a valid anabolic agent, capable of preserving the structure of the joints in a tumour necrosis factor alpha (TNFα)-transgenic mouse model of arthritis.\cite{275} Mutations involving the R-spondin1 gene have been associated with female-to-male sex reversal in mice\cite{265, 276} and humans\cite{277, 278}, palmoplantar hyperkeratosis and predisposition to squamous cell carcinoma of the skin in humans.\cite{277} R-spondin2 has proven to be necessary for the normal development of limbs\cite{266}, lungs\cite{279}, and hair follicles\cite{280, 281} in animal models. Additionally, the presence of R-spondin2 was found to be fundamental for the osteoblast maturation and mineralization mediated by Wnt11 through β-catenin/Wnt pathway.\cite{282} Mutations of
the gene encoding R-spondin2 in mice lead to craniofacial, limb, and lung morphogenesis defects.\cite{283} R-spondin3 is necessary for a regular development of the placenta in mice\cite{284}; R-spondin3, along with R-spondin1, are involved in embryonic vasculogenesis and angiogenesis developmental processes through their activation of the vascular endothelial growth factor (VEGF) via the Wnt pathway.\cite{285, 286} Additionally, gene fusions involving R-spondin3 and R-spondin2 have been found in 10% of primary colon cancers.\cite{287} R-spondin4 is involved in nail morphogenesis processes: autosomal recessive mutations in the furin domain of R-spondin4 are linked to complete absence or severe hypoplasia of fingernails (anonychia) and toenails (hyponychia).\cite{288-290} All four R-spondin proteins are potentially able to activate the cWnt pathway: however, according to the results presented by Kim et al., R-spondin2 and R-spondin3 are more potent than R-spondin1 and R-spondin4 in activating Wnt signalling.\cite{259}

### 1.6.3 R-spondin proteins and the cWnt signalling pathway

Several hypothesis have been formulated to elucidate the mechanism behind the interaction between R-spondins and the Wnt pathway. According to the studies made by Nam et al. and Wei et al., R-spondins bind to Wnt receptors Fz8 and LRP6, causing the phosphorylation of LRP6 and an increase of cytoplasmatic concentration of β-catenin, which leads to augmented expression of cWnt target genes.\cite{262, 269} In a different theory, R-spondin prevents the actions of DKK 1, interacting with transmembrane protein Kremen1. In this way, R-spondins prevent the internalization of LRP6, hence activating Wnt signalling.\cite{264, 267} Recent studies identified the leucine-rich repeat containing G protein-coupled receptors 4, 5 and 6 (Lgr4, Lgr5 and Lgr6) as the physiological receptors of R-spondins: this interaction mediates an increase of the Wnt pathway activity, increasing the phosphorylation of LRP6.\cite{291-294} These
findings were implemented in a recent study by Hao and co-workers\textsuperscript{[295]}, where they showed how R-spondins are able to enhance Wnt signalling interacting with the transmembrane E3 ubiquitin ligase zinc and ring finger 3 (ZNRF3) or its homologue RING finger 43 (RNF43). In the absence of R-spondins, ZNRF3/RNF43 ubiquitylates Fz receptors causing a clearance of Fz-LRP complex, hence down-regulating Wnt signalling. In the presence of R-spondins, a ternary complex including ZNRF3/RNF43, R-spondin and LGR is formed, leading to an augmented membrane clearance of ZNRF3/RNF43 via endocytosis. This results in accumulation of Fz and LRP receptors on the membrane and enhancement of the Wnt signalling.\textsuperscript{[295, 296]}

Numerous studies suggested that upregulated activation of the Wnt signalling pathways play an important role in tumorigenesis.\textsuperscript{[297-300]}, and several approaches in targeting the different components of the Wnt pathway are described in the literature.\textsuperscript{[301-304]} Given the importance of Wnt pathway in a plethora of physiological and pathological processes, we aim to explore a new way of modulating this through the elucidation of the structural dynamics of its important regulator R-spondin. Hence, in our study we are going to focus firstly on assessing the plasticity of R-spondin1 and R-spondin 2 using microsecond molecular dynamics simulations. The conformations obtained from these simulations are then used to perform a druggability analysis, aiming to highlight potential binding sites on the surface of R-spondins. Our hypothesis is that these sites could be targeted by a ligand disrupting the interactions in the ternary complex R-spondin/LGR/RNF43. In this way, we could reduce the clearance of RNF43 mediated by R-spondins and thus module a hypothetical hyper activated Wnt signalling pathway, acting on the expressions of Fz/LRP complex on the plasma membrane.
References


83
Introduction


Introduction


Introduction


Introduction


Introduction


Introduction


Introduction


Introduction


Introduction


Introduction


2. Effect of sequence and stereochemistry reversal on p53 peptide mimicry

Alessio Atzori\textsuperscript{a}, Audrey E. Baker\textsuperscript{b}, Mark Chiu\textsuperscript{b}, Richard A. Bryce\textsuperscript{a}, Pascal Bonnet\textsuperscript{c}

\textsuperscript{a}School of Pharmacy and Pharmaceutical Sciences, University of Manchester, Manchester, United Kingdom

\textsuperscript{b}Biologics Research, Janssen Research and Development Inc., Radnor, Pennsylvania, United States of America

\textsuperscript{c}Janssen Research & Development, a division of Janssen Pharmaceutica N.V., Beerse, Belgium
2.1 Abstract

Peptidomimetics effective in modulating intracellular protein-protein interactions and resistance to proteolysis have potential in therapeutic applications. An appealing yet underperforming peptidomimetic strategy is to employ D-amino acids and reversed sequences to mimic a lead peptide conformation, either separately or as the combined retro-inverso peptide. In this work, we examine the conformations of inverse, reverse and retro-inverso peptides of p53(15-29) using implicit solvent molecular dynamics simulation and circular dichroism spectroscopy. In order to obtain converged ensembles for the peptides, we find enhanced sampling is required via the replica exchange molecular dynamics method. From these replica exchange simulations, the D-peptide analogues of p53(15-29) result in a predominantly left-handed helical conformation. When the parent sequence is reversed sequence as either the L-peptide and D-peptide, higher helical content is observed across the peptides, in agreement with NMR and CD studies, along with an altered residue profile. The simulations also indicate that, while approximately similar orientations of the side-chains are possible by the peptide analogues, their ability to mimic the parent peptide is severely compromised by backbone orientation (for D-amino acids) and side-chain orientation (for reversed sequences). Thus, a retro-inverso peptide is disadvantaged as a mimic in both aspects, and further chemical modification is required to enable this concept to be used fruitfully in peptidomimetic design. Nevertheless, we find here that replica exchange molecular dynamics using the ff99SB force field and Generalized Born solvent model is an effective combination for computational modelling of D-peptide conformations, with potential as a tool in guiding structure-based design of new improved peptidomimetics.
2.2 Author Summary

Interactions between proteins determine a wide range of key biological processes. The ability to modulate these processes through the development of molecules that interfere in a specific way with these interactions has enormous therapeutic potential. One approach to designing such molecules is to create short peptide variants of the region of one protein that forms the interaction with the partner protein. These analogue peptides involve reversal of the parent amino sequence (retro peptides), inversion of sequence (inverso peptides) and both (retro-inverso peptides). In order to understand the extent to which these peptides can mimic an inhibitory conformation, we perform a combined computational and experimental study of the p53(15-29) peptide, the contact region of tumour suppressor protein p53, a transcription factor involved in cellular response to DNA damage. Here, we derive an effective computational simulation protocol to predict converged distributions of structures for p53(15-29) and its retro, inverso and retro-inverso analogues. The simulations find greater helical content for the retro and retro-inverso analogues, in agreement with NMR and CD studies. However, analysis of the computed structural distributions shows that the p53 analogues are unable to mimic the detailed backbone orientation of the parent peptide for D-amino acid analogues and side-chain orientation for reversed sequence analogues. Nevertheless, the computational protocol employed here will prove useful as a tool in predicting peptide conformation and potentially in designing new, improved peptide-based therapeutics.
2.3 Introduction

Protein-protein interactions (PPIs) are key to a range of fundamental biological functions. Conversely, erroneous PPIs are linked to pathological conditions such as Alzheimer’s disease, Creutzfeldt-Jakob disease, tumorous conditions and AIDS.[1-3] Molecules that can modulate these interactions have potential as therapeutics.[4] However, peptidic ligands are limited by several factors including chemical and conformational stability in vivo.[5-6] Strategies to surmount such issues include the use of D-amino acids and other (non-natural) amino acids, structural restraints and non-peptidic backbones.[5-6] An interesting form of peptidomimetic is the retro-inverso (RI) peptide.[7-8] The concept is based on two operations: (i) inversion of chirality of the lead peptide (Figure 2-1a) to produce its enantiomer (Figure 2-1b) and (ii) reversal of its sequence, which reverses the direction of the termini and peptide bonds. (Figure 2-1c). When operations (i) and (ii) are combined, a retro-inverso peptide is produced, which should recover the original side-chain orientation of the lead peptide (Figure 2-1d). This postulate holds for a peptide in its extended form and in helices of the same handedness. The degree of peptidomimicry by RI peptides is less clear when secondary structures are adopted by these and their parent peptides. RI peptides have been shown to elicit cross-reactivity with antibodies against pro-peptide antigens. In perhaps more exacting tests of molecular recognition, however, RI peptides have exhibited mixed success; for example, RI peptides did not perform as well as the corresponding parent peptide in inhibiting the interactions of S-peptide/S-protein[9] and p53/MDM2.[10]
Figure 2-1: Peptide transformations: (a) parent peptide, (b) its inverse, (c) its sequence reversed and (d) its retro-inverso analogue.

The p53/MDM2 is a particularly interesting case, as it highlights the role of secondary structure in determining the inhibitory activity of peptidomimetics. The tumour suppressor p53, a protein of 53 kDa, is a transcription factor involved in cellular response to DNA damage. Considerable interest has focused on the design of modulators of the p53/MDM2 interaction, as inhibition of the human E3 ubiquitin ligase MDM2 leads to reactivation of p53 activity in human glioblastoma cells. The crystal structure of the p53/MDM2 complex indicates that residues 17–29 of p53 interact with a deep hydrophobic cleft on MDM2.\[11\] In particular, p53 residues Phe19,
Effect of sequence and stereochemistry reversal on p53 peptide mimicry

Trp$^{23}$ and Leu$^{26}$ make extensive interactions with this pocket (Figure 2-2); to achieve this bound pose, p53 forms a partially α-helical conformation over these residues. Attempts to mimic the p53/MDM2 interaction using mutant p53 peptides$^{[12]}$ show a positive correlation between helicity and inhibitory effect; and peptides stapled into a helical conformation also proved inhibitory$^{[13]}$. Clearly, secondary structure is important for MDM2 interaction of peptidic ligands, particularly in orienting the key i, i+4 and i+7 side-chains of helical peptides.

Figure 2-2: (a) Representation of the N-terminal sequence of p53(white) bound to the MDM2(electrostatic map) derived from its crystal structure. Backbone atoms are represented in cartoon, while the residues Phe$^{19}$, Trp$^{23}$ and Leu$^{26}$ are in sticks. (b) Interatomic distances measured from the crystallographic structure accordingly to the two reference sites, namely α carbon (in red) and β carbon (in green). The three key residues (Phe$^{19}$, Trp$^{23}$ and Leu$^{26}$) are represented in sticks.
Effect of sequence and stereochemistry reversal on p53 peptide mimicry

The RI peptide of p53(15–29), however, has proved only weakly active against MDM2. Originally, it was postulated by experimental methods that the RI peptide adopted a right-handed helical form[14], contrary to the expected propensity of a D-peptide. Interestingly this harks back to an earlier controversy about the helical handedness of RIs.[15] However, a subsequent NMR structure of the RI of p53(15–29) indicated that a left-handed helix was found in a solution of 50% TFE; no crystals of the RI/MDM2 complex were obtainable.[10] Whilst the concept of retro-inverso peptidomimetics is compelling in its simplicity, to progress their application to structure-based design of protein-protein interactions (PPIs), a greater understanding of their ability to reproduce key structural features of the parent peptide is required. Consequently, in this work, we systematically examine the structural consequences of transformation of the parent p53(15–29) peptide (WT, Table 2-1), considering its stereochemical inverse (I), its reversed sequence (R) and finally its retro-inverso analogue (RI).

<table>
<thead>
<tr>
<th>Peptide</th>
<th>Sequence</th>
<th>$K_d$ [nM]</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>Ace-Ser Gln Glu Thr Phe Ser Asp Leu Trp Lys Leu Leu Pro Glu Asn-Nme</td>
<td>140 ± 5[29]; 255 ± 5[10]</td>
</tr>
<tr>
<td>I</td>
<td>Ace-$^D$Ser-$^D$Gln-$^D$Glu-$^D$Thr-$^D$Phe-$^D$Ser-$^D$Asp-$^D$Leu-$^D$Trp-$^D$Lys-$^D$Leu-$^D$Leu-$^D$Pro-$^D$Glu-$^D$Asn-Nme</td>
<td>n/d[29]</td>
</tr>
<tr>
<td>R</td>
<td>Ace-Asn Glu Pro Leu Leu Lys Trp Leu Asp Ser Phe Thr Glu Gln Ser-Nme</td>
<td>n/d[29]</td>
</tr>
<tr>
<td>RI</td>
<td>Ace-$^D$Asn-$^D$Glu-$^D$Pro-$^D$Leu-$^D$Lys-$^D$Trp-$^D$Leu-$^D$Asp-$^D$Ser-$^D$Phe-$^D$Thr-$^D$Glu-$^D$Gln-$^D$Ser-Nme</td>
<td>71600 ± 8600[10]</td>
</tr>
<tr>
<td>$^D$PMIα</td>
<td>$^D$Thr-$^D$Asn-$^D$Trp-$^D$Tyr-$^D$Ala-$^D$Asn-$^D$Leu-$^D$Glu-$^D$Lys-$^D$Leu-$^D$Leu-$^D$Arg</td>
<td>219 ± 11[19]</td>
</tr>
<tr>
<td>$^D$PMIβ</td>
<td>$^D$Thr-$^D$Ala-$^D$Trp-$^D$Tyr-$^D$Ala-$^D$Asn-$^D$Phe-$^D$Glu-$^D$Lys-$^D$Leu-$^D$Leu-$^D$Arg</td>
<td>34.5 ± 0.6 [19]</td>
</tr>
</tbody>
</table>

Table 2-1: Amino acid sequences and MDM2 binding affinities ($K_d$) of peptides p53(15-29) (WT), its inverse (I), its reverse (R) and its retro-inverse (RI); and peptides derived from phage display ($^D$PMIα and $^D$PMIβ). D-amino acids residues are indicated with superscript "D".

To achieve these, we first assess the ability of replica exchange molecular dynamics (MD) simulations to reproduce the experimental secondary structural trends of these peptides in aqueous solution that were observed previously[10] and performed here by
circular dichroism measurements. To our knowledge, this is the first application of replica exchange MD to characterize D-peptide and retro-inverso conformations. Secondly, we examine the ability of each peptide to mimic the key Phe\(^{19}\)-Trp\(^{23}\)-Leu\(^{26}\) pharmacophore formed in binding of p53 to MDM2.

### 2.4 Results/Discussion

In order to explore via computation the structural consequences of transformation of parent 15-mer p53-derived peptide \(\text{WT}\) into its inverse (I), reversed (R) and retro-inverso (RI) sequences, we first consider the ability of molecular dynamics simulations to provide a converged estimate of their secondary structures. To this end, extended and right-handed helical initial conformations of L-peptides \(\text{WT}\) and R were constructed; additionally, left-handed conformations of D-peptides I and RI were built. The peptides were then simulated for 50 ns in Generalised Born implicit solvent using either molecular dynamics or replica exchange molecular dynamics. For REMD, this equates to 3.2 \(\mu\)s of aggregate dynamics per peptide simulation. Good equilibration of the 64 replicas in the REMD simulations across the 270–400 K temperature range was observed (for example, see Supplementary Information, Figure 2-10 in Supporting Information).

Over 50 ns of MD, all four sequences fail to establish a converged secondary structure distribution, with a highly uneven estimate of helical content as a function of time (Figure 2-3a). Likewise, the estimates of helical content vary markedly as a function of initial conformation; this is demonstrated for example by the estimate of 60% helical content for MD starting from the right-handed helix of peptide RI, averaged over the last 20 ns, versus 40% helix initiated from an extended conformation (Figure 2-3a).
Effect of sequence and stereochemistry reversal on p53 peptide mimicry

Likewise, by contrast, REMD simulations exhibit well-converged estimates of helical content after 30 ns for all four peptides (Figure 2-3b). Similarly, the helical content is converged between replicate simulations initiated from different conformations (extended, right-handed helix and left-handed helix in Figure 2-3b). Inspection of snapshots from the time series of the retro-inverso peptide RI illustrate this convergence in structure, with extended, left-handed and right-handed helical initial structures each converging over the 50 ns to ensembles containing left-handed rather than right-handed helix (Figure 2-4). Interestingly these characteristics emerge here after only 20 ns, but it is clear from subsequent structures that an equilibrium of folded and less-folded conformations is established (Figure 2-4).
Effect of sequence and stereochemistry reversal on p53 peptide mimicry

Figure 2-4: Snapshots of the 300 K trajectory taken at 0, 10, 20, 30, 40 and 50 ns starting from an extended (E), right-handed (R) and left-handed (L) helical structure. Each configuration is ordered from N- to C-terminus.

Nevertheless, over the final 20 ns, the peptides occupy a region around φψ backbone torsion angles of (60°, 30°), in the left-handed helical domain on a Ramachandran plot (Figure 2-5). The corresponding converged α-helical content for RI is computed to be ~40% (Figure 2-3b). The Ramachandran plots for peptide I similarly indicate convergence to broadly left-handed helical conformations, with a somewhat lower α-helical content of ~20% (Figures 2-4 and 2-5). In addition to hydrogen bonds separated by four residues, the DSSP algorithm also finds a proportion of configurations of I and RI, where i→i+3 hydrogen bonds are observed, of ~18% and ~11% respectively (Figure 2-11 in Supporting Information).
Effect of sequence and stereochemistry reversal on p53 peptide mimicry

Figure 2-5: Ramachandran plots of the distribution of the (φ, ψ) backbone dihedrals of residues 1–15 for peptides WT, I, R and RI, calculated over the 50 ns REMD trajectories.

By contrast to D-peptides I and RI, for the 15-mer L-peptides WT and R, REMD simulations predict a proportion of right-handed rather than left-handed helical structures, with φψ backbone torsion angles converging to the (−60°,−30°) α-helical region (Figure 5). The α-helical content of WT reflects its enantiomer I, with a value in the vicinity of 20%; likewise, the α-helical content of R reflects its enantiomer RI (Figure 2-4). Thus, α-helical content varies with amino acid sequence but, as expected, does not vary between mirror images (this in fact is another measure of the satisfactory convergence of the REMD simulations). Correspondingly, the proportion of 3-10 helical configurations for WT and R mirror those of their enantiomers I and RI, with populations of ~18% and ~11% respectively (Figure 2-11 in Supporting Information).

We do also observe a small proportion of conformations in the β-sheet (−60°,140°) region, for both parent peptide WT and in a symmetric region (60°,−140°) for its enantiomer I (Figure 2-5). These populations are observed for all replicate simulations of WT and I; very little of this conformation is found for R and RI.
We turn now to consider experimental estimates of helicity in these peptides. Circular dichroism measurements have been performed previously\(^{10}\) for peptides WT and RI, and indicate that the peptides are largely unstructured in aqueous solution (PBS buffer). Although one must treat absolute values with caution, the helical content estimated from CD in that work was of the order of 10\%.\(^{10}\) To reveal the helical propensity of these peptides, 60% v/v TFE was then used as the helix-promoting solvent. In this solvent, peptide WT readily adopted a right-handed helix and RI a left-handed helix. Furthermore, the structures of peptides WT and RI in TFE/water were determined by NMR spectroscopy\(^{10}\); although insufficient detail was published to allow detailed comparison with our simulations, WT was reported to adopt a flexible structure, which was somewhat disordered at the N- and C-termini but with a helical structure in its centre; peptide RI formed a left-handed helix with a more ordered and helical structure.

Thus the NMR structures reflect the helicity estimated from CD spectra. In this work, we have reprised CD measurements to estimate of helicity for WT and additionally examine retro-peptide R. For spectra obtained in aqueous solution (PBS buffer), the peptides showed strong negative peaks near 195 nm typical of random coil (Figure 2-6a). The peptides also showed signals near 222 nm typical of helices. The second typical helix signal at 208 nm appears to be masked by the broad peak near 195 nm. The percentage α-helix was obtained from data fitting using CD deconvolution software (see Materials and Methods). The percentage secondary structure element obtained from this software is derived from comparisons of CD structures of globular proteins with known α-helix or β-sheet content. Since the samples examined here are peptides and not globular proteins, there may be some error inherent in the estimation. From this method, the helical contents of WT and R in PBS buffer were found to be
Effect of sequence and stereochemistry reversal on p53 peptide mimicry

similar: this was around 17% for WT and only slightly larger for R, at 19% (Figure 2-6c). When studied in 60% v/v TFE (Figure 2-6b), peptides WT and R showed negative peaks at 208 nm and 222 nm typical of alpha helices. Peptide R showed the strongest helix signal, while WT displayed ~25% of the intensity seen with R: the estimated helical content of R was 58% in 60% v/v TFE, whereas that of WT was only 20% (Figure 2-6c). Thus, the mixed TFE/water solvent exposes the greater helical propensity of the reverse sequence, common to R and RI, relative to the sequence of peptides WT and I.

Figure 2-6: Circular dichroism spectra of peptides: (a) CD spectra of WT and R at 0.3 mg/mL in PBS (b) CD spectra of WT and R at 0.3 mg/mL in 60% TFE/40% PBS (c) percentage peptide α-helix estimated from circular dichroism.
Effect of sequence and stereochemistry reversal on p53 peptide mimicry

This propensity for a larger helical content for R and RI is reflected in the α-helical populations obtained from our REMD simulations, with values of ~20% for WT and I, and ~40% for R and RI (Figure 2-3). The estimates from these aqueous solvent simulations are most comparable with the CD spectra in PBS buffer and therefore are not in quantitative agreement with CD estimates; this discrepancy might depend on a number of factors, including on the algorithm used to compute helicity from the CD spectra (as referred to above), peptide solubility or indeed on the DSSP method used to analyse the simulations.

Another factor is the force field and implicit solvent model employed. The ff99SB force field used here was derived from ff99 to improve simulation of secondary structure.\cite{18} We note that an early form of ff99SB, combined with the OBC Generalized Born solvent model, reproduced explicit solvent simulations of three proteins to within a backbone RMSD of ~1.5 Å.\cite{19} However, we also note that a recent study of a model tetrapeptide\cite{20}, using the same force field and solvent model as employed in this work, found, an overestimation of ion pair interactions and a greater propensity to form helix relative to explicit solvent simulations, which may indeed be influencing the level of helical content observed in our simulations.

We observe that comparable simulations we performed using the ff96 force field with Generalised Born solvent yielded helical content in excess of 80% for peptides WT and RI, a result that might suggest that this the ff96 over-emphasizes the α helical content of the studied sequences (see Figure 2-12 Supporting Information). Within the limitations of an implicit solvent framework, the simulations based on ff99SB provide improved agreement between experiment and calculation, although the predicted helicity remains somewhat higher than estimates obtained from CD. In terms of predicted distribution of helicity across the amino acid sequence of each peptide,
our simulations find similar profiles for mirror image pairs WT and I, and for R and RI (Figure 2-7). The higher helicities of R and RI are evident, extending over the entirety of the peptides and with a peak in helicity around Leu5-Lys6-Trp7 (numbered from the N-terminus). By contrast, peptides WT and I have lower helicity across the peptide, including at the C- and N-terminals, in qualitative agreement with the NMR structure of WT (Figure 2-7). The peaks in predicted helicity for WT and I are found around residues Asp7-Leu8-Trp9 (Figure 2-7), again illustrating the different conformational consequences by sequence reversal. On the contrary, the predicted distribution of helicity across the amino acid sequence of WT and RI using ff96 show a general lack of correlation (see Figure 2-13 Supporting Information). For example, while the profiles registered for RI from an extended and right helical conformations are comparable, the one obtained from the left helical conformation assumes a completely different trend, with a predicted percentage of 100% of helicity extended to almost all residues. This is another evidence of how ff96 over-emphasizes the α helical content of the studied sequences.

Figure 2-7: Distribution of helicity (ordinate) as a function of amino acid residue for each of sequences WT, I, R and RI (abscissa), obtained from final 20 ns of REMD simulations with ff99SB. Residue numbering is from N- to C-terminal direction (solid line) for WT, I, R and RI. Additionally, helicity is shown for the reverse residue numbering i.e. from C- to N-terminus for R and RI (dashed line).
In order to obtain further insight into the origin of helical preference of reversed sequence (R/RI) over wild type (WT/I), we perform here a detailed analysis of the hydrogen bonding formed within the four peptides over the last 20 ns of REMD simulation.

As expected, the total computed fractions of i→i+4 hydrogen bonds for R and RI sequences (1.8 and 1.7, Table 2-4 in Supporting Information) are greater than that of WT and I sequences (both 0.9).

Also, for WT and I peptides, the distribution of i→i+4 hydrogen bond population as a function of residue i reflects their profiles obtained from the DSSP algorithm (Figure 2-7), with maxima at residue Leu8 for WT and I (Table 2-4 in Supporting Information). For R and RI, more approximate agreement is found, with maxima predicted at Glu2 and Leu5 (Table 2-4 in Supporting Information). We note that no hydrogen bond is observed for Trp9–Pro13 in WT/I, but is observed for Pro3–Trp7 in R/RI (Table 2-4 in Supporting Information); in the former, Pro has no backbone NH available to act as proton donor, but in the latter, Pro has a backbone C=O available to act as proton acceptor.

Figure 2-8: Intramolecular polar contacts. (a) A helical conformation of WT when Lys10 NH…Ser6 OC and Lys10 Hζ…Asp7 Oδ hydrogen bonds are present; (b) A helical conformation of R when Ser10 NH…Lys6 OC and Ser10 Hγ…Lys6 O hydrogen bonds are present.
We also examine all hydrogen bonding interactions formed by amino acid side-chains, to other side-chain or main-chain atoms. We find that for WT and I peptides, only a single hydrogen bond is populated beyond 15% during the REMD trajectories at 300 K; this interaction is a salt-bridge formed between side-chains of Asp7 and Lys10 (Figure 2-8a), with a population of 20% for both sequences (Table 2-2). Associated with this interaction, only infrequently is a bridging i→i+4 hydrogen bond formed between the peptide groups of residues 7 and 11 (<1%, data not shown).

Conversely, for the reverse sequences (R and RI), three hydrogen bonds are significantly populated (Table 2-2).

<table>
<thead>
<tr>
<th>sequence</th>
<th>hydrogen bond atom pair</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>Asp7 Oδ – Lys10 Hζ</td>
<td>20</td>
</tr>
<tr>
<td>I</td>
<td>Asp7 Oδ – Lys10 Hζ</td>
<td>20</td>
</tr>
<tr>
<td>R</td>
<td>Lys6 O – Ser10 Hγ</td>
<td>25</td>
</tr>
<tr>
<td>RI</td>
<td>Lys6 O – Ser10 Hγ</td>
<td>25</td>
</tr>
<tr>
<td>R</td>
<td>Leu8 O – Thr12 Hγ</td>
<td>22</td>
</tr>
<tr>
<td>RI</td>
<td>Leu8 O – Thr12 Hγ</td>
<td>21</td>
</tr>
<tr>
<td>R</td>
<td>Phe11 O – Ser15 Hγ</td>
<td>16</td>
</tr>
<tr>
<td>RI</td>
<td>Phe11 O – Ser15 Hγ</td>
<td>17</td>
</tr>
</tbody>
</table>

*Table 2-2: Population, p (%), of most commonly formed intramolecular hydrogen bonds involving side-chains of peptides WT, I, R and RI from last 20 ns of REMD.*

In each case, these interactions involve a side-chain hydroxyl group as proton donor (of Ser10, Thr12 and Ser15) and in each case the side-chain OH hydrogen bonds to the corresponding i-4 residue’s peptide O atom. When this hydrogen bond is made, population of the corresponding α-helical peptide hydrogen bond ranges from 3–10%.

A representative MD conformation where the peptide carbonyl O of Lys6 hydrogen bonds to the peptide NH and OH side-chain of Ser10 (Figure 2-8b) suggests the potentially helix-stabilising nature of these interactions. From inspection, the absence
of an analogous interaction in the WT/I sequence appears to arise from an increased distance of the side-chain OH to the main-chain O atom.

**Pharmacophore.** From the preceding discussion, it is evident that the 50 ns REMD simulations provide a converged estimate of the conformational ensembles of the two L- and two D-peptides in aqueous solution that is in reasonable agreement with experiment. Based on these simulations, we now consider the ability of the right handed helix of sequences WT and R and left handed helix of sequences I and RI to mimic the relative spatial locations of the side-chains of the p53 Phe-Trp-Leu triad that is key to its interaction with MDM2 (Figure 2-2a).

We use two measures of relative spatial arrangement: the distance between the α-carbons for Phe-Trp, Trp-Leu and Phe-Leu pairs, denoted $d_a1$, $d_a2$ and $d_a3$ respectively; and the equivalent inter-β-carbon distances, denoted $d_b1$, $d_b2$ and $d_b3$ respectively (Figure 2-2b). From the crystal structure of p53/MDM2 (PDB code: 1YCR\cite{11}, resolution 2.6 Å), the distances between Phe$^{19}$ and Trp$^{23}$ and between Trp$^{23}$ and Leu$^{26}$ are in a narrow range of 5.7–6.2 Å over both α- and β-carbon measures (Table 2-3). Unsurprisingly, the inter-α- and inter-β-carbon distances between Phe$^{19}$ and Leu$^{26}$ are rather larger, at ~11.7 Å (Table 2-3).

We therefore take these two sets of crystallographic distances, defined by either α- or β-carbons, to represent pharmacophores for interaction with MDM2. These are denoted $[d_a1–d_a2–d_a3]$ and $[d_b1–d_b2–d_b3]$ respectively, and correspond to a conformation of the peptide such that all three distances are simultaneously satisfied to within ±0.5 Å of the crystallographic values of the p53 motif when MDM2-bound. Thus, we analyse the last 20 ns of the REMD simulations of WT, I, R and RI, to consider what if any population of $[d_a1–d_a2–d_a3]$ and $[d_b1–d_b2–d_b3]$ pharmacophoric motifs are present in the solution ensembles.
Effect of sequence and stereochemistry reversal on p53 peptide mimicry

<table>
<thead>
<tr>
<th>Distance/pharmacophore</th>
<th>X-ray distance p53(15-29)</th>
<th>MD average distance</th>
<th>population (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>WT</td>
<td>I</td>
</tr>
<tr>
<td>(a) inter-C(\alpha)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(d_{a1})</td>
<td>5.89</td>
<td>7.28 (1.45)</td>
<td>7.22 (1.33)</td>
</tr>
<tr>
<td>(d_{a2})</td>
<td>5.92</td>
<td>6.02 (1.27)</td>
<td>5.99 (1.29)</td>
</tr>
<tr>
<td>(d_{a3})</td>
<td>11.67</td>
<td>11.31 (2.13)</td>
<td>11.31 (1.94)</td>
</tr>
<tr>
<td>([d_{a1}-d_{a2}-d_{a3}])</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(b) inter-C(\beta)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(d_{b1})</td>
<td>5.67</td>
<td>7.67 (1.81)</td>
<td>7.56 (1.72)</td>
</tr>
<tr>
<td>(d_{b2})</td>
<td>6.22</td>
<td>7.04 (1.54)</td>
<td>7.04 (1.49)</td>
</tr>
<tr>
<td>(d_{b3})</td>
<td>11.62</td>
<td>11.50 (2.47)</td>
<td>11.56 (2.28)</td>
</tr>
<tr>
<td>([d_{b1}-d_{b2}-d_{b3}])</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**Table 2-3**: Average interatomic distances \(d_1\) (Phe19-Trp23), \(d_2\) (Trp23-Leu26) and \(d_3\) (Phe19-Leu26) as defined by (a) \(\alpha\)-carbons and (b) \(\beta\)-carbons from final 20 ns of REMD for sequences WT, I, R and RI. Standard deviations in parentheses. The \(d_1\), \(d_2\) and \(d_3\) values from crystallographic structure of N-terminal sequence of p53 bound to MDM2 also given (PDB 1YCR[11]). All distances in Å. Population (%) of trajectory of WT, I, R and RI that satisfy \(d_1\), \(d_2\) and \(d_3\) and their simultaneous combination, \([d_1-d_2-d_3]\) over final 20 ns of REMD.

Peptide WT, being directly derived from the p53 sequence, might be expected to exhibit the most highly populated pharmacophore. For WT, we find a population of 5.0% of pharmacophore \([d_{a1}-d_{a2}-d_{a3}]\) (Table 2-3). This fairly low population reflects for example that the average simulated distance \(d_1\) is larger than the crystallographic value of \(d_1\), by 1.4 and 2.0 Å for \(\alpha\)- and \(\beta\)-carbon definitions respectively (Table 2-3).

Each of the single interatomic distances considered also exhibit significant standard deviations around their mean values (1.5–2.5 Å for inter-C\(\beta\) distances in WT), reflecting the dynamic nature of the peptide. With the perhaps more exacting and accurate measure of the three inter-\(\beta\)-carbon distances, the \([d_{b1}-d_{b2}-d_{b3}]\) pharmacophore in peptide WT is only populated for 1.5% of the 300 K REMD trajectory. This is set against higher populations of individual distances (20–51%, Table 2-3). Compared to WT, a very similar pattern of populations is observed for mirror image peptide I, once again indicating convergence of sampling (Table 2-3).
Effect of sequence and stereochemistry reversal on p53 peptide mimicry

For reversed sequence \( R \) and its mirror image \( R'I \), the population of pharmacophore \([d_{a1}−d_{a2}−d_{a3}]\) is 4.7\% (Table 2-3). This value is similar to that of \( WT \) and \( I \).

The population for \( R \) and \( R'I \) is also similar to \( WT \) and \( I \) when considering \([d_{b1}−d_{b2}−d_{b3}]\), at 1.6\% for \( R \) and 2.1\% for \( R'I \) respectively. We note in all cases that there is a similar level of flexibility in the peptides, as reflected by the standard deviation in \( d_{b1} \), \( d_{b2} \) and \( d_{b3} \) distances ranging from 1.3–2.5 Å for \( R \) and \( R'I \) (Table 2-3).

Consequently, given the similar population of rudimentary pharmacophores \([d_{a1}−d_{a2}−d_{a3}]\) and \([d_{b1}−d_{b2}−d_{b3}]\) by the four peptides, one might expect the retro peptide and the retro-inverso peptide to be quite as capable of interaction with MDM2 as parent p53(15–29) peptide. For further insight, we compare the crystallographic conformation of p53(15–29) with representative MD configurations of \( WT \), \( I \), \( R \) and \( R'I \) that exhibit pharmacophore. To achieve this superposition, reversed peptides \( R \) and \( R'I \) were superimposed in their Ct→Nt orientation on the crystal structure in its Nt→Ct direction (Figure 2-9a–d). The side-chains of Phe\(^{19} \), Trp\(^{23} \) and Leu\(^{26} \) in the crystal structure overlap well with those found for peptide \( WT \) and, to a lesser extent, \( I \) (Figure 2-9a, b).

By comparison, whilst the side-chains of \( R \) and \( R'I \) exist in the same approximate location as for the crystal structure, there is essentially no overlap of the actual side-chain moieties of Phe, Trp and Leu (aromatic rings and isopropyl group). The reversed sequence \( R \) orients the \( \alpha-\beta \) bond vectors, and thus residue side-chains, in opposite quite distinct direction to p53(15–29) when \( R \) is aligned Ct→Nt against p53(15–29) in a Nt→Ct direction (Figure 2-9c). Recovery of a similar orientation for these residues would be achieved by inversion of the reverse sequence only if \( R'I \) retained a right-handed helix; indeed, a modelled hypothetical right-handed helix of \( R'I \) (Figure 2-9f) demonstrates a similar position of the side-chains to that of the crystal structure and
Effect of sequence and stereochemistry reversal on p53 peptide mimicry

peptide WT (Figure 2-9e), contrasting with left-handed RI (Figure 2-9g). Indeed, experiment and simulation finds that RI adopts a left-handed rather than right-handed helix and therefore the side-chains point into volumes quite disparate from the native structure (Figure 2-9d).

Figure 2-9: (top) Superimposition of Ca atoms of X-ray structure of N-terminal sequence of p53 (white) onto representative MD conformations of WT (green, a), I (red, b), R (blue, c) and RI (yellow, d). (bottom) Comparison of the orientation of the side-chains of the residues Phe, Trp and Leu for the right-handed helix of WT (green, e), right handed helix of RI (orange, f) and left handed helix of RI (yellow, g). Key residues Phe, Trp and Leu represented as sticks.
Alongside the differing side-chain orientations of R and RI, the sense of the left-handed helix dictates a very different volume occupied by the D-peptide backbones of peptides I and RI (Figures 2-9b,d); this lack of mimicry could present a potential problem in binding. It should also be borne in mind that the peptide bonds are reversed in peptides R and RI, which could affect peptide-receptor hydrogen bonds involving these atoms. In summary, whilst helical structures are adopted by I, R and RI, and approximate pharmacophores \([d_a1−d_a2−d_a3]\) and \([d_b1−d_b2−d_b3]\) are fractionally populated, the detailed molecular structures of I and RI do not suggest strong analogy to parent sequence WT. Indeed, retro-inverso peptide RI is found to be a weak MDM2 binder experimentally, with a Kd of 72 µM, as compared to the sub-micromolar potency of WT (Table 2-1). Although aspects of their findings have been disputed (including their observed high activity of RI), Sakurai et al.\[^{14}\] found no observable MDM2 inhibitory activity for peptides I and R using an ELISA-based assay.

2.5 Conclusions

In this study, we have examined the consequences of reversed sequence and inverted stereochemistry on structural features of a p53-derived peptide in aqueous solution using molecular dynamics simulations. As a prerequisite to this analysis, we find that, in order to obtain converged ensembles for the peptides, enhanced sampling is required via the replica exchange molecular dynamics method. From these replica exchange simulations, D-peptides I and RI result in a predominantly left-handed helical conformation, regardless of initial structure. Interestingly, when the parent sequence is reversed to give the L-peptide R and D-peptide RI, higher helical content with an altered residue profile, is observed for the peptides; this propensity for increased helicity of the reversed sequence is suggested by NMR and CD studies in TFE/water solvent.
However the predicted degree of helicity across the four peptides from our simulations in aqueous solution, based on the ff99SB/Generalized Born potential, appears somewhat high.

From the simulations, the increased predicted helicity of the reversed sequences relative to wild type appears to stem, at least in part, from stabilizing side-chain hydrogen bonding by Ser10, Thr12 and Ser15 with their corresponding i-4 peptide group; this interaction appears to be precluded in the wild type sequences, and, from inspection, may be due to a combination of the altered orientation of the Cα-Cβ bond vector and an increased distance of the side-chain OH to the main-chain O atom in the wild type sequence. From these simulations, we also consider the potential of these modifications to achieve mimicry of the natural peptide. We find that the inter-Cα distances or inter-Cβ distances of the key triad of amino acids are similar between WT and its mimics I, R and RI. However, analysis of molecular orientation underlines problems in their alignment of backbone (peptides I and RI) and their side-chains (peptides R, RI and to some extent, I also). Thus, a retro-inverso peptide is disadvantaged as a mimic in both aspects. That it can produce antigenic mimicry appears a function of the less discriminatory molecular recognition by immunoglobulins.

Nevertheless, the concept of using D-peptide mimetics has been successfully utilized in generating inhibitors of MDM2. D-peptides DPMI-α and DPMI-β (Table 2-1), generated by mirror-image phage display and native chemical ligation, were found to be sub-micromolar inhibitors of MDM2.\textsuperscript{[17]}
Effect of sequence and stereochemistry reversal on p53 peptide mimicry

Interestingly, crystallography revealed that DPMI-\(\alpha\) binds to MDM2 in a left-handed helical conformation, in a pose shifted from p53(15–29). This suggests some flexibility in the backbone (and side-chain) motifs recognized by MDM2.

Stapling strategies have been used with success in targeting a number of PPIs using peptides; one could envisage future structure-based drug design strategies employing a retro-inverso peptide which is stapled into a right-handed helical conformation (consider Figures 2-9e and 2-9f), thus preserving the conformation of backbone and side-chain whilst benefitting from greater resistance to proteolysis and improved in vivo stability. Indeed, the simulation-based approach adopted here, in particular employing replica exchange molecular dynamics, has provided insights into D-peptide conformations and may prove a useful tool in directing the future design of such peptidomimetic structures.
2.6 Materials and Methods

Computational details. Four systems were considered: p53(15–29) (WT) and its inverse (I), reversed (R) and retro-inverso (RI) sequence (Table 2-1). Each peptide was capped by an acetyl group (Ace) at the N-terminus and an N-methylamine group (Nme) at the C-terminus. In subsequent discussion, we refer to the residue number 1 to 15 of each peptide; the Ace capping group of the N-terminus is considered residue 0. For each, extended (E) and right helical (R) conformations were built. For D-peptides, an additional left helical (L) form was modelled. The extended structures of the studied sequences were built using xleap program from the AMBER 11\(^{[21]}\) suite, whilst right- and left-handed helical structures were created using MOE\(^{[22]}\). The SHAKE algorithm constrained bonds between hydrogen and heavy atoms. A 2 fs time step was used. The peptides were modelled with the AMBER ff99SB\(^{[18]}\) force field using Generalized Born (GB) implicit solvent\(^{[19]}\) with a dielectric constant of 80. Simulations were performed in the canonical ensemble NVT with a Langevin thermostat, using a collision frequency of 2 ps\(^{-1}\). An infinite cut-off for long-range non-bonded interactions was used.\(^{[23]}\) All peptide sequences were minimized prior to 200 ps equilibration, to the desired temperature for each replica in the case of the REMD\(^{[24]}\) simulations, or to 300 K for standard MD simulations. For REMD of each peptide, 64 replicas were simulated for 50 ns, using a temperature range from 270 to 400 K. The algorithm of van der Spoel and Patrikkson\(^{[25]}\) was used to generate a set of temperatures with the target exchange acceptance ratio of 30%. Exchanges were attempted every ps. REMD and MD simulations were performed using the AMBER 11\(^{[21]}\) molecular simulation package. Configurations were archived every 5 ps.
Effect of sequence and stereochemistry reversal on p53 peptide mimicry

Secondary structure analysis used the DSSP method by Kabsch and Sander\textsuperscript{[26]} as implemented in the ptraj module of AMBER 11, which bases its approach mainly on hydrogen bonding patterns.\textsuperscript{[21]} Calculations of $\alpha$-helicity were obtained performing by block averaging of the trajectories.

**Circular dichroism spectroscopy**\textsuperscript{1}. CD analyses were conducted using published protocols with slight modifications.\textsuperscript{[27, 28]} For CD analyses, peptides were diluted from stocks at 4 mg/mL in PBS, except sequence R which was prepared at 1 mg/mL due to its poor solubility at 4 mg/mL. CD spectra were collected on 0.3 mg/mL peptide solutions in PBS or in 60\% (v/v) TFE, 40\% (v/v) 1x Dulbecco’s PBS by running four scans at 1 nm intervals from 190 nm–260 nm, at 298 K with 1 sec signal averaging and using 100 ms time constant. Spectra of the PBS and PBS-60\% TFE buffers were also collected using the same parameters. Peptide and buffer spectra were averaged and represented as plots of helicity versus wavelength. $\alpha$-helical content was estimated using CD deconvolution software.\textsuperscript{[29]}

\textsuperscript{1}Circular dichroism experiments and analysis of their results were performed exclusively by Audrey E. Baker and Mark Chiu.
2.7 References


Effect of sequence and stereochemistry reversal on p53 peptide mimicry


2.8 Supporting Information

Figure 2-10: Replica exchange equilibration for the replica initiated at 270 K.
Effect of sequence and stereochemistry reversal on p53 peptide mimicry

Figure 2-11: $\beta_{10}$ helical content (ordinate) for sequences WT, I, R, and RI as a function of time (ns) (abscissa) obtained from REMD simulations.
Effect of sequence and stereochemistry reversal on p53 peptide mimicry

<table>
<thead>
<tr>
<th>i → i+4</th>
<th>i(WT/I)</th>
<th>i(R/RI)</th>
<th>WT</th>
<th>I</th>
<th>R</th>
<th>RI</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – 4</td>
<td>Ace</td>
<td>Ace</td>
<td>0.04 (0.19)</td>
<td>0.04 (0.19)</td>
<td>0.00 (0.02)</td>
<td>0.00 (0.03)</td>
</tr>
<tr>
<td>1 – 5</td>
<td>Ser</td>
<td>Asn</td>
<td>0.04 (0.21)</td>
<td>0.04 (0.20)</td>
<td>0.09 (0.28)</td>
<td>0.05 (0.22)</td>
</tr>
<tr>
<td>2 – 6</td>
<td>Gln</td>
<td>Glu</td>
<td>0.11 (0.31)</td>
<td>0.10 (0.30)</td>
<td>0.31 (0.46)</td>
<td>0.30 (0.46)</td>
</tr>
<tr>
<td>3 – 7</td>
<td>Glu</td>
<td>Pro</td>
<td>0.08 (0.26)</td>
<td>0.07 (0.25)</td>
<td>0.18 (0.39)</td>
<td>0.17 (0.38)</td>
</tr>
<tr>
<td>4 – 8</td>
<td>Thr</td>
<td>Leu</td>
<td>0.14 (0.34)</td>
<td>0.12 (0.33)</td>
<td>0.26 (0.44)</td>
<td>0.23 (0.42)</td>
</tr>
<tr>
<td>5 – 9</td>
<td>Phe</td>
<td>Leu</td>
<td>0.16 (0.37)</td>
<td>0.18 (0.39)</td>
<td>0.28 (0.45)</td>
<td>0.28 (0.45)</td>
</tr>
<tr>
<td>6 – 10</td>
<td>Ser</td>
<td>Lys</td>
<td>0.05 (0.22)</td>
<td>0.06 (0.23)</td>
<td>0.12 (0.33)</td>
<td>0.11 (0.32)</td>
</tr>
<tr>
<td>7 – 11</td>
<td>Asp</td>
<td>Trp</td>
<td>0.05 (0.21)</td>
<td>0.04 (0.20)</td>
<td>0.16 (0.37)</td>
<td>0.14 (0.35)</td>
</tr>
<tr>
<td>8 – 12</td>
<td>Leu</td>
<td>Leu</td>
<td>0.25 (0.43)</td>
<td>0.26 (0.44)</td>
<td>0.15 (0.36)</td>
<td>0.14 (0.35)</td>
</tr>
<tr>
<td>9 – 13</td>
<td>Trp</td>
<td>Asp</td>
<td>-</td>
<td>-</td>
<td>0.05 (0.22)</td>
<td>0.06 (0.23)</td>
</tr>
<tr>
<td>10 – 14</td>
<td>Lys</td>
<td>Ser</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.01)</td>
<td>0.09 (0.29)</td>
<td>0.11 (0.31)</td>
</tr>
<tr>
<td>11 – 15</td>
<td>Leu</td>
<td>Phe</td>
<td>0.02 (0.15)</td>
<td>0.01 (0.12)</td>
<td>0.15 (0.36)</td>
<td>0.14 (0.35)</td>
</tr>
<tr>
<td>total</td>
<td></td>
<td></td>
<td>0.94 (0.92)</td>
<td>0.93 (0.91)</td>
<td>1.84 (1.48)</td>
<td>1.74 (1.45)</td>
</tr>
</tbody>
</table>

*Table 2-4: Average number of hydrogen bonds between the backbone peptide C=O of residues i and the backbone peptide NH of the residues i + 4 and average total number of hydrogen bonds within sequence over final 20 ns of REMD for sequences WT, I, R and RI. Standard deviations in parentheses.*
Figure 2-12: Fractional helical content (ordinate) for sequences WT and RI as a function of time (ns) (abscissa) obtained from REMD simulation with ff96: maximum helicity (dashed line) and average helicity (solid line) obtained starting from extended (E), right-handed (R) and left-handed (L) helical structures.
Figure 2.13: Distribution of helicity (ordinate) as a function of amino acid residue for each of sequences WT and RI (abscissa), obtained from final 20 ns of REMD simulations with ff96 obtained starting from extended (E), right-handed (R) and left-handed (L) helical structures. Residue numbering is from N- to C-terminal direction (solid line) for WT and RI. Additionally, helicity is shown for the reverse residue numbering i.e. from C- to N-terminus for RI (dashed line).
3. Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics

Alessio Atzori,* Neil J. Bruce,*‡ Kepa K. Burusco,* Berthold Wroblowski,ª Pascal Bonnetc and Richard A. Brycea

a Manchester Pharmacy School, University of Manchester, Oxford Road, Manchester, M13 9PT, UK

b Janssen Research & Development, a division of Janssen Pharmaceutica N.V., Turnhoutseweg 30, 2340 Beerse, Belgium

c Structural Bioinformatics & Chemoinformatics, Institut de Chimie Organique et Analytique (ICOA), UMR CNRS-Université d’Orléans 7311, Université d’Orléans, Rue de Chartres, F-45067 Orléans cedex 02, France

‡ Current address: Heidelberg Institute for Theoretical Studies (HITS gGmbH), Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany
3.1 Abstract

Protein plasticity, while often linked to biological function, also provides opportunities for rational design of selective and potent inhibitors of their function. The application of computational methods to the prediction of concealed protein concavities is challenging, as the motions involved can be significant and occur over long time scales. Here we introduce the swarm-enhanced sampling molecular dynamics (sesMD) method as a tool to improve sampling of conformational landscapes. In this approach, a swarm of replica simulations interact cooperatively via a set of pairwise potentials incorporating attractive and repulsive components. We apply the sesMD approach to explore the conformations of the DFG motif in the protein p38α mitogen-activated protein kinase. In contrast to multiple MD simulations, sesMD trajectories sample a range of DFG conformations, some of which map onto existing crystal structures. Simulated structures intermediate between the DFG-in and DFG-out conformations are predicted to have druggable pockets of interest for structure-based ligand design.
Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics

3.2 Introduction

Crystal structures are able to provide molecular-level detail on the interactions of ligands with their target receptors. Recently, valuable additional information has been mined from computing the time evolution of an initial protein or protein–ligand crystal structure using molecular dynamics (MD) simulations. For example, MD simulations of HIV integrase discovered a previously unknown part of the binding site that was able to explain experimental binding and mutagenesis data. This finding contributed to the successful development of the drug raltegravir, now a frontline drug in AIDS treatment.[1] Similarly, potential design opportunities from MD have been found, for example, for neuraminidase[2] and in recent work on Trypanosoma cruzi trans-sialidase.[3–5] However, because of their rough free energy surfaces, observing nontrivial changes in protein structure, such as the opening and closing of pockets, loops, and grooves, is not commonly possible by MD without the application of biasing potentials.[6] These biasing methods can be equilibrium (e.g., umbrella sampling)[7] or nonequilibrium (e.g., steered MD) approaches.[8]

Unfortunately, the manner in which these biasing potentials or forces should be applied is not always apparent. Specific biased MD approaches range in their complexity of application and computational overhead and include active-site pressurization[9], locally enhanced sampling[10], accelerated molecular dynamics[11], and replica-exchange-based approaches[12] as well as the group formed by local elevation[13], conformational flooding[14], and metadynamics methods.[15,16] Here we propose an approach for exploring conformational space that utilizes a swarm-based MD scheme. This multicopy MD algorithm builds on the SWARM-MD method, which was first proposed by Huber and van Gunsteren[17] and more recently refined by us for the prediction of biomolecular structure.[18]
Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics

The SWARM-MD approach involves creating multiple copies of the molecular system (i.e. a swarm of systems) and then simulating the dynamics of each copy in parallel. The simulations of these copies mimic cooperative swarm behaviour through the addition of attractive forces that act on each member of the swarm, driving their trajectories toward the mean trajectory of the entire swarm. In this way, the additional attractive force experienced by a given replica within the swarm is directly proportional to its distance from the average swarm structure. We demonstrated that our modified version of SWARM-MD is a simple and effective approach for the optimization of peptide and protein structures.\textsuperscript{18} For example, during 20 independent 40 ns MD simulations of the Trp-cage miniprotein, none were able to fold to within a backbone root-mean-square deviation (RMSD) of 1.5 Å of the NMR conformation. By contrast, of 20 SWARM-MD replicas, 16 folded to a backbone RMSD below 1.5 Å within 40 ns.\textsuperscript{18}

In this work, we move from using a swarm of MD replicas to optimize biomolecular structure to instead harnessing a swarm to explore the conformational landscape. To promote barrier crossing into alternative energy minima, we introduce a set of attractive and repulsive pair potentials that act between swarm copies. We refer to this method subsequently as swarm-enhanced sampling MD (sesMD). We evaluate the ability of this sesMD approach to enhance the conformational exploration of a protein known to exhibit different structures, the anticancer target p38α mitogen-activated protein (MAP) kinase.

Specifically, we focus on sampling of its Asp168-Phe169-Gly170 (DFG) motif, which can be characterized by different orientations of its phenylalanine side-chain: in the “DFG-out” conformation, Phe169 points away from the nearby α C helix and projects into the ATP binding pocket, exposing an additional hydrophobic cavity, sometimes
Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics
called the allosteric or deep pocket. In the “DFG-in” conformation, access to the deep pocket is hindered by Phe169 (Figure 3-1).[19]

Figure 3-1: Superposition of the DFG-in (PDB 1P38, green) and DFG-out (PDB 1WBT, red) X-ray crystal structures of p38α MAP kinase. Backbone atoms are shown in cartoon representation and residues Asp168 and Phe169 in stick representation.

While active and inactive states of protein kinases can exhibit DFG-in conformations, the DFG-out conformation is ordinarily associated only with the inactive state. Thus, the movement of the phenylalanine side-chain to the DFG-out conformation creates this deep pocket, a cavity that has been exploited in the design of inhibitors with good specificity, i.e., by type-II kinase inhibitors such as Imatinib.[20] Experimental[21 – 24] and computational studies[25 – 32] have found that there is a significant energetic barrier between kinase DFG-in and DFG-out states. For example, molecular dynamics simulations have required strategies such as studying kinase mutants with modified kinetics[28], massively distributed MD simulations[29], or biasing methods.[25 – 27,30 – 32]
Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics

3.3 Methods

**sesMD.** The SWARM-MD method described by Huber and van Gunsteren\(^\text{17}\) involves the attraction of swarm members to the mean of the swarm via an extra potential, \(V_{\text{swarm}}\), given by:

\[
V_{\text{swarm}}(\{\phi^\alpha\}) = \sum_{\alpha} A \exp\left(-Bd_{\alpha}^{\text{rms,av}}(\phi^\alpha)\right)
\]

(3.1)

where \(d_{\alpha}^{\text{rms,av}}\) is the root-mean-square dihedral angle difference of swarm member \(\alpha\) from the average location of the swarm, \(\phi^\alpha\) is set of dihedral angles of swarm member \(\alpha\) to which the swarm potential is applied and \(M\) is the number of members in the swarm. \(A\) and \(B\) are parameters that govern the strength and range of attraction between each member and the swarm average.

Here, we adapt SWARM-MD from an optimisation approach to a method that permits enhanced coverage of conformational space, such that the multiple MD simulations of a molecular system interact in a cooperative swarm-like fashion to surmount high energy barriers between conformations. We therefore introduce the following swarm-enhanced sampling (ses) potential, \(V^{\text{ses}}\):

\[
V^{\text{ses}}(\{\phi^\alpha\}) = \sum_{\alpha} \sum_{\beta>\alpha} A \exp\left(-Bd_{\alpha\beta}^{\text{rms}}(\phi^\alpha, \phi^\beta)\right) + C \exp\left(-Dd_{\alpha\beta}^{\text{rms}}(\phi^\alpha, \phi^\beta)\right)
\]

(3.2)

where \(A\) – \(D\) are suitably calibrated parameters for attractive (A,B) and repulsive (C,D) terms; and \(d_{\alpha\beta}^{\text{rms}}(\phi^\alpha, \phi^\beta)\) is the root-mean-square dihedral angle distance between swarm members \(\alpha\) and \(\beta\). Distinct from eq. 1, the first term in eq. 3.2 is a pairwise attractive potential, acting between pairs of swarm replicas. This potential is balanced by a repulsive exponential potential (the second term in eq. 2). In this scheme, the
values of parameters A and C must be negative and positive, respectively, while both B and D should be positive. The potential seeks to increase the spread of conformations explored, whilst promoting transitions between wells. The resulting dynamics are therefore a combination of (i) the *ses* potential, $V_{ses}$, applied to a selected set of dihedral angles and acting between pairs of replicas, and (ii) the interatomic potentials of the molecular mechanics force field, which act individually on each member of the swarm.

As a simple illustration of sesMD, we consider the effect of introducing $V_{ses}$ on sampling the conformational space of two solutes - pentane and alanine dipeptide. Firstly, eight unbiased MD simulations of 10 ns were performed of each solute in explicit aqueous solvent (for computational details, see Supporting Information). Sampling of low energy regions for pentane (Figure 3-2a) and alanine dipeptide (Figure 3-2c) remains fairly localised, with only minor excursions into alternative wells on the landscapes.

Alternatively coupling together the eight replicas via the *ses* potential enables exploration of alternative low energy regions during 10 ns of sesMD: for pentane, much greater coverage of symmetry-related gauche conformations is found (Figure 3-2b); for alanine dipeptide, sampling of other regions of the Ramachandran landscape is achieved (Figure 3-2d).
However, we note that, due to the presence of the $V_{\text{ses}}$ term acting between replicas, the configurations sampled by sesMD are not drawn from a Boltzmann distribution. Nevertheless, the underlying potential energy surface topologies appear to be reasonably preserved via sesMD sampling (Figure 3-2b,d). Furthermore, due to the well-defined nature of the overall swarm Hamiltonian, it is formally possible to recover Boltzmann-weighted properties from the sesMD method where required (for more details, see Supporting Information).
Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics

**Computational Details.** As starting points for molecular dynamics simulations of p38α MAP kinase, the crystallographic structures 1P38[33] and 1WBT[34] were used, representing DFG-in and DFG-out conformations of the protein, respectively. The structures were prepared for MD using the *xleap* program from the AMBER 11 suite[35]. The systems were modelled using the AMBER ff99SB force field[36] and solvated with ~13000 TIP3P water molecules. Counterions were added in order to neutralize the net charge of the systems. The SHAKE algorithm[37] constrained bonds between hydrogen and heavy atoms. A 2 fs time step was used. Simulations were performed in the NPT ensemble with a Langevin thermostat using a collision frequency of 2 ps$^{-1}$. A Berendsen isotropic barostat was used to maintain an average pressure of 1 atm. The particle mesh Ewald method[38] was used for computing long-range electrostatic interactions and a 12 Å cutoff for van der Waals’ contributions. For sesMD, the ses potential described in equation 3.2 was applied to the φ and ψ backbone dihedral angles of the DFG motif (Asp168, Phe169, and Gly170). For sesMD simulations starting from the DFG-in 1P38 crystal structure, the sesMD parameters (A, B) and (C, D) were taken as (−150.0 kcal/mol, 0.2 rad$^{-1}$) and (150.0 kcal/mol, 0.8 rad$^{-1}$) respectively. For sesMD simulations initiated from the DFG-out 1WBT structure, the parameters (A, B) and (C, D) adopted were (−250.0 kcal/mol, 0.2 rad$^{-1}$) and (375.0 kcal/mol, 0.8 rad$^{-1}$) respectively. The two ses potentials therefore differ in their dependence on $d_{φθ}^{\text{rmsd}}$ (Figure 3-12 in the Supporting Information).

The choice of parameters for the two structures is to some degree a reflection of the local potential energy landscape of the DFG loop within the two kinase states of DFG-in and DFG-out: a stronger ses potential was applied to the DFG-out structure, whereas the same strength of potential led to distortion of peptide bonds within the DFG loop for the DFG-in structure, apparently as a result of greater steric constraints of the
Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics

occupied deep pocket within this structure. For sesMD, 24 replicas of the kinase were employed. The initial configurations of the 24 replicas were generated as follows: the final structure of a 700 ps MD simulation at 300 K was taken as the initial configuration for 24 independent replica 500 ps simulations at 300 K with different initial velocities, and sesMD simulations were then performed using these 24 final configurations. During the first 500 ps of sesMD, the parameters \( A - D \) (equation 3-2) were increased linearly from zero to their final values, followed by 500 ps of sesMD at the final \( A - D \) values; these parameters were then reduced to zero over 500 ps, and 3.5 ns of unbiased MD was then performed. The 24 independent 5 ns MD simulations used for comparison with the sesMD simulations were initiated from the configurations immediately prior to the first application of the ses potential detailed above. All of the MD and sesMD simulations were performed using a modified version of sander from the Amber11 molecular simulation package.\(^{[35]}\) Configurations were archived every 2 ps for analysis.

For principal component analysis (PCA), conformations were taken from every 20 ps of each sesMD simulation replica’s trajectory and fitted onto the \( C_\alpha \) atoms of an average conformation calculated across all replicas in that simulation. With these fitted conformations, PCA of the atomic displacements of the DFG residues was performed using the \textit{ptraj} module of Amber 11. Clustering analysis of the DFG residues was performed using the kclust algorithm from the MMTSB toolset\(^{[39]}\) with a fixed clustering radius of 4.0 Å. Pocket detection and druggability analysis were performed using PocketAnalyzer\(^{[40]}\) and Site Finder\(^{[41]}\) from MOE.\(^{[42]}\)
3.4 Results

We evaluated the ability of sesMD to sample the conformational plasticity of unliganded p38α MAP kinase initiated from either a DFG-out crystal structure (1WBT) or a DFG-in crystal structure (1P38). As described in Methods, we applied the ses potential of equation 3.2 to the six backbone torsions linking the DFG loop amino acids, performing 5 ns of sesMD simulation using 24 replicas of the kinase in explicit solvent. These 5 ns of simulation comprise 1.5 ns under the influence of the ses potential followed by 3.5 ns of unbiased MD. In the following discussion, we compare these sesMD trajectories with 24 independent 5 ns MD simulations of the kinase in explicit solvent.

Sampling from DFG-out and DFG-in Crystal Structures. We consider the ability of MD and sesMD to sample the backbone φ and ψ torsions for the three residues of the key DFG motif of the kinase. As a reference, on the basis of the annotation of the MOE Kinase Explorer Database[^42], in Figure 3-3 we plot the φ and ψ angles of Asp168, Phe169 and Gly170 from the 74 DFG-in p38α MAP kinase crystal structures (green), the 61 DFG-out structures (red), and the 33 structures denoted as “unclassified” (blue). The DFG-in and DFG-out distributions of crystallographic φ and ψ angles are largely distinct for all three residues of the loop, but for Asp168 in particular (Figure 3-3a). Thus, the φ and ψ angles of Asp168 in the DFG-out 1WBT structure are (−139° , 111°), which sit within the DFG-out cluster around the β-sheet region (Figure 3-3a); in contrast, for the 1P38 DFG-in structure the φ and ψ angles are (32° , 68°), lying within the narrow left-handed α-helical region of DFG-in structures of the map (Figure 3-3a).
Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics

Figure 3-3: Ramachandran plots of the $\phi$ and $\psi$ angles for each DFG residue sampled by 24 replicas of sesMD simulations (black) starting from p38α MAP kinase crystal structures 1WBT (a−c) and 1P38 (g−i) and by 24 independent MD simulations (d−f and j−l). Also shown are $\phi$ and $\psi$ angles of 61 DFG-out (red), 74 DFG-in (green), and 33 unclassified (blue) p38α MAP kinase X-ray structures as defined in the MOE Kinase Explorer Database.

Unbiased MD simulations initiated from the 1WBT structure appear only to explore ($\phi$, $\psi$) values in the locality of its initial backbone conformation (Figure 3-3d). However, application of the ses potential leads additionally to exploration of the DFG-out ($\phi$, $\psi$) cluster in the right-handed $\alpha$-helical region and notably around the DFG-
Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics

in cluster as well (Figure 3-3a). Broader sampling by sesMD over MD is also observed for the backbone torsions of Phe169 (Figure 3-3b vs Figure 3-3e) and Gly170 (Figure 3-3c vs Figure 3-3f); generally, the crystal $\phi$ and $\psi$ values map onto the conformations sampled by sesMD, although some simulated conformations do encompass regions of the Ramachandran map with few crystal structures (e.g., $\phi = 90^\circ$ for Phe169 and $\phi = 180^\circ$ for Gly170). Interestingly, for MD simulations initiated from the 1P38 DFG-in structure, ($\phi$, $\psi$) values of Asp168 for both DFG-in and DFG-out regions are explored (Figure 3-3j). However, once again, sesMD provides more complete coverage of ($\phi$, $\psi$) space (Figure 3-3g–i), such that, for example, DFG-in ($\phi$, $\psi$) values in the region of ($-90^\circ$, $-45^\circ$) are explored for Asp168 and Phe168 (Figure 3-3g,h) and in the vicinity of ($90^\circ$, $0^\circ$) for Gly169 (Figure 3-3i).

To provide a broader view of the sampling of the DFG loop conformation, PCAs were performed on the heavy atoms of the DFG main chain and side-chain atoms on the basis of the 168 X-ray crystal structures in combination with the 1WBT trajectory (Figure 3-4a) or the 1P38 trajectory (Figure 3-4b) from the MD or sesMD methods. The two top principal components from these PCAs account for 75% and 76% of the variance in atomic displacements of the DFG motif in the 1WBT and 1P38 simulations, respectively. Projection of the 168 X-ray crystal structures onto the axes defined by these first two principal components (Figure 3-4a,b) results in two distinct clusters, separating the DFG-out (red) and DFG-in (green) conformations. Conformations unclassified by the MOE Kinase Explorer Database (dark blue) to some degree occupy both clusters. It can also been seen that unbiased molecular dynamics simulations (yellow) explore only locally around the initial structures of the trajectories for either 1WBT (Figure 3-4a) or 1P38 simulations (Figure 3-4b).
Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics

**Figure 3-4:** Projections of MD (yellow) and sesMD (cyan) ensembles of kinase and DFG-out (red), DFG-in (green), and unclassified (dark blue) p38α MAP kinase crystal structures onto the space defined by the top two principal components based on simulations starting from the (a) 1WB and (b) 1P38 crystal structures.

By contrast, sesMD simulations of 1WB and 1P38 (cyan) explore more extensively, in both cases populating regions in the principal component space around the initial DFG geometry and the opposing DFG orientation.

Further analysis of the sesMD trajectories indicates that the PC space most remote from the initial structures is sampled by a minority of its 24 replicas. This is shown by analysis of the RMSD of the atomic positions of the DFG-loop heavy atoms of each of the 24 replicas with respect to the 1WB and 1P38 X-ray structures when aligned according to the overall protein conformation (Figures 3-13 and 3-14 in the Supporting Information). For the majority of the replicas in the 1WB DFG-out simulation, the RMSD values are \(\sim 2 - 3 \text{ Å}\) from this initial structure (Figure 3-13), although for two replicas that we have labelled 1\text{out} and 2\text{out}, the RMSD relative to 1WB rises to \(\sim 7 \text{ Å}\) (Figure 3-5a,b respectively). Similarly, for sesMD simulations from 1P38, the majority of the conformations sample locally (Figure 3-14); however, for one replica of the 1P38 simulation, which we have labelled 1\text{in}, the RMSD rises to \(\sim 7 \text{ Å}\) from 1P38 (Figure 3-5d).
Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics

**Figure 3-5**: (a, b, d) Time series of heavy-atom RMSDs (in Å) of the DFG motif from sesMD simulations of p38α MAP kinase for (a) replica 1\textsubscript{out}, (b) replica 2\textsubscript{out}, and (d) replica 1\textsubscript{in} calculated against the DFG-out 1WBT (red) and DFG-in 1P38 (green) crystal structures during pre-sesMD equilibration (−0.5 to 0.0 ns) and sesMD trajectories (0.0 to 5.0 ns). (c, e) Projections onto the first two principal components of sesMD configurations of (c) replicas 1\textsubscript{out} (orange) and 2\textsubscript{out} (cyan) and (e) replica 1\textsubscript{in} (magenta), alongside PC values for 24 MD replicas (yellow) and DFG-in (green), DFG-out (red), and unclassified (dark blue) crystal structures.

From a chart of the conformations sampled in the defined PC space by replicas 1\textsubscript{out}, 2\textsubscript{out}, and 1\textsubscript{in}, we observe that they occupy the majority of the remotely sampled regions in the PCA (orange and cyan, Figure 3-5c; magenta, Figure 3-5e). For all three of these simulation replicas, the largest change in RMSD occurs during the application of the ses potential rather than during the free MD phase of the simulations. More specifically, it occurs during the phase in which there is a gradual increase of the ses potential, i.e., during the interval from 0 to 0.5 ns (Figure 3-5a,b,d). We may also examine the RMSDs of the DFG loops of these replicas with respect to their opposing crystal structure conformations. For replica 1\textsubscript{out} the RMSD remains high with respect
Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics
to the 1P38 structure throughout the simulation, dropping only slightly from 8 to 6 Å (green, Figure 3-5a). A similar pattern is observed for replica 1_in of the 1P38 sesMD simulation (Figure 3-5d). Indeed, the high RMSDs of replicas 1_out and 1_in with respect to the DFG-in and DFG-out crystal structures is suggestive of the adoption of an intermediate DFG geometry. However, for replica 2_out, the RMSD drops over the course of the simulation to a value of ~3–4 Å (Figure 3-5b).

In order to obtain greater molecular detail of the conformational change in these replicas, we consider two inter-residue distances that have been suggested to be useful diagnostic measures of DFG-in and DFG-out conformations[^32]: the distance between the side-chain centroids of Asp168 and Lys53 (neighbouring the glycine-rich loop) and the distance between the side-chain centroids of Phe169 and Leu74 (located in the αC-helix). For the 1WBT DFG-out crystal structure, these distances are 13.5 and 17.8 Å respectively, and for the 1P38 DFG-in structure, the values are 11.8 and 6.9 Å respectively. Thus, from DFG-out to DFG-in, the considerable shortening of the Phe169 − Leu74 distance and lesser reduction of the Asp168 − Lys53 distance indicates that Phe169 now occupies the kinase deep pocket and Asp168 has correspondingly swung into the active site. For the first 2 ns of sesMD of 1WBT replicas 1_out and 2_out, we observe a transient deviation of the Asp168 − Lys53 distance from its crystallographic value (Figure 3-6a) and a drop in the Phe169 − Leu74 distance to ~ 11 Å (Figure 3-6b). However, whereas replica 1_out remains at a Phe169 − Leu74 distance of ~ 11 Å, replica 2_out further decreases to a final value of ~ 7 Å, the crystallographic value for the DFG- in structure (Figure 3-6b).
Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics

**Figure 3-6**: (a, c) Time series of the distance between the side-chain centroids of Asp168 and Lys53 for (a) replicas $1_{\text{out}}$ (orange) and $2_{\text{out}}$ (cyan) and (c) replica $1_{\text{in}}$ (magenta) of p38α MAP kinase from pre-sesMD equilibration (−0.5 to 0.0 ns) and sesMD trajectories (0.0 to 5.0 ns). (b, d) Time series of the distance between the side-chain centroids of Phe169 and Leu74 of (b) replicas $1_{\text{out}}$ (orange) and $2_{\text{out}}$ (cyan) and (d) $1_{\text{in}}$ (magenta) from pre-sesMD equilibration and sesMD trajectories. The X-ray distances for the 1WBT (red) and 1P38 (green) structures are also shown for reference.

For replica $1_{\text{in}}$ of 1P38, local conformational sampling is observed around the Asp168–Lys53 X-ray distance (Figure 3-6c). By contrast, a large change in the Phe169 – Leu74 distance is observed to stabilize in proximity to the DFG-out X-ray value (Figure 3-6d). We can visually trace the conformational motion of the DFG loop in the three replicas by considering snapshots from their trajectories (Figure 3-7).
Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics

![Figure 3-7](image)

**Figure 3-7**: Conformation of Phe169 in selected snapshots from (a) replica 1\textsubscript{out}, (b) replica 2\textsubscript{out}, and (c) replica 1\textsubscript{in} from sesMD simulations of p38α MAP kinase. The conformations from sesMD are color-coded according to the simulation time/stage: 0.5 − 1 ns (orange), 1 − 1.5 ns (yellow), 1.5 − 2 ns (cyan), 2 − 5 ns (blue). Also shown are the DFG-in 1P38 (green) and DFG-out 1WBT (red) crystal structures.

The DFG-out conformation of the 1WBT crystal structure (red) and the DFG-in pose of the 1P38 crystal structure (green) bracket conformations sampled over the sesMD simulations. These sesMD snapshots show the time progression over the 5 ns trajectory (from orange to yellow to cyan to blue). The trajectories for both replicas 1\textsubscript{out} and 2\textsubscript{out} (Figure 3-7a,b, respectively) show movement from their initial DFG-out structures (red) toward the opposing DFG-in conformations (green). However, whereas replica 1\textsubscript{out} is arrested in an intermediate structure (Figure 3-7a), replica 2\textsubscript{out} completes the transition to a DFG-in geometry with Phe169 occupying the deep pocket (Figure 3-
Explores protein kinase conformation using swarm-enhanced sampling molecular dynamics

7b) and mapping onto DFG-in orientations found in the apo (1P38) and liganded (2HVC and 3LGC)\[43,44] DFG-in crystal structures (Figure 3-8).

**Figure 3-8**: Final conformation from sesMD of replica 2\textsubscript{out} showing the Asp168 and Phe169 side-chains in the p38\textalpha MAP kinase active site (blue), compared with its initial pose from the DFG-out 1WBT X-ray structure (red). The DFG-in crystal structures 1P38 (green) and 2LGC and 3HVC (both grey) are also shown.

This structural transition is more completely illustrated by a movie of the trajectory (see the Appendix). The DFG-out to DFG-in transition contrasts with a control simulation we ran from the same initial conditions of replica 2\textsubscript{out} using 50 ns of unbiased MD. Here no transition is observed, but instead the kinase remains in a local DFG-out conformation (see the Appendix). For replica 1\textsubscript{in}, a partial transition from DFG-in to DFG-out is observed, such that there is a significant shift in the Phe169 conformation (Figure 3-7c). However, Asp168 remains in a local well around its initial structure, maintaining a polar interaction with the backbone of Asn152. For replicas 1\textsubscript{out}, 2\textsubscript{out}, and 1\textsubscript{in}, the largest movement of the DFG loop is associated with the period where the ses potential is increasing to its full strength over the interval 0.0 to 0.5 ns (orange in Figure 3-7). However, also particularly important for the replica 2\textsubscript{out} simulation is the final 3.5 ns of nonbiased MD, during which the activation loop relaxes to a structure within the manifold of DFG-in loop conformations that are observed.
Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics crystallographically (Figure 3-15 in the Supporting Information). A relaxation of the swarm potential was found to be similarly important in the final stages of folding of the Trp-cage miniprotein by SWARM-MD.\textsuperscript{[18]}

**Conformations and Druggability of Intermediate Structures of the DFG Loop.** It is clear that a range of DFG loop conformations is sampled during sesMD of p38\textalpha MAP kinase (e.g., Figure 3-7), some of which may present opportunities for molecular design. To this end, we performed conformational clustering of the combined trajectories of all 24 sesMD replicas for the 1WBT and 1P38 simulations separately.

![Figure 3-9: Projections onto the space defined by the first two principal components of the ensemble of 24 sesMD simulation replicas, initiated from the (a) 1WBT and (b) 1P38 crystal structures. The projections are colour-coded according to cluster: C1\textsubscript{out} and C1\textsubscript{in} (red), C2\textsubscript{out} and C2\textsubscript{in} (orange), C3\textsubscript{out} and C3\textsubscript{in} (yellow), C4\textsubscript{out} and C4\textsubscript{in} (green), C5\textsubscript{out} and C5\textsubscript{in} (cyan), C6\textsubscript{out} (blue), C7\textsubscript{out} (magenta), and C8\textsubscript{out} (purple).](image)

Adopting a fixed clustering radius of 4.0 Å, we identified eight clusters of distinct conformations explored by the 1WBT sesMD replicas (labelled C1\textsubscript{out} – C8\textsubscript{out}) and five clusters of 1P38 sesMD conformations (C1\textsubscript{in} – C5\textsubscript{in}). From the location of these structures in PC space, we classify the clusters into DFG-out- like and DFG-in-like conformations for 1WBT (Figure 3-9a) and 1P38 simulations (Figure 3-9b). As might be expected, the most populated clusters are localized around the parent X-ray
Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics conformation. Thus, for the 1WBT sesMD simulation, clusters $C_{1\text{out}} - C_{4\text{out}}$, representing 85% of the total sesMD configurations (Table 3-1), occupy DFG-out-like conformations (left-hand side of Figures 3-9a and 3-10); however, the less populated $C_{5\text{out}} - C_{8\text{out}}$ clusters occupy a spectrum toward a DFG-in structure (right-hand side of Figures 3-9a and 3-10).

<table>
<thead>
<tr>
<th>cluster</th>
<th>DFG conf</th>
<th>pop (%)</th>
<th>$V_{av}$ (Å$^3$)</th>
<th>PLB</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_{1\text{out}}$</td>
<td>out</td>
<td>25.2</td>
<td>1212 (347)</td>
<td>3.9</td>
</tr>
<tr>
<td>$C_{2\text{out}}$</td>
<td>out</td>
<td>23.9</td>
<td>1044 (255)</td>
<td>3.0</td>
</tr>
<tr>
<td>$C_{3\text{out}}$</td>
<td>out</td>
<td>18.8</td>
<td>1334 (294)</td>
<td>4.3</td>
</tr>
<tr>
<td>$C_{4\text{out}}$</td>
<td>out</td>
<td>17.6</td>
<td>1255 (225)</td>
<td>4.5</td>
</tr>
<tr>
<td>$C_{5\text{out}}$</td>
<td>intermed</td>
<td>6.6</td>
<td>866 (278)</td>
<td>3.8</td>
</tr>
<tr>
<td>$C_{6\text{out}}$</td>
<td>intermed</td>
<td>3.9</td>
<td>712 (277)</td>
<td>3.5</td>
</tr>
<tr>
<td>$C_{7\text{out}}$</td>
<td>in</td>
<td>1.9</td>
<td>1181 (294)</td>
<td>2.8</td>
</tr>
<tr>
<td>$C_{8\text{out}}$</td>
<td>in</td>
<td>2.0</td>
<td>1071 (291)</td>
<td>2.7</td>
</tr>
<tr>
<td>$C_{1\text{in}}$</td>
<td>in</td>
<td>79.6</td>
<td>880 (260)</td>
<td>3.0</td>
</tr>
<tr>
<td>$C_{2\text{in}}$</td>
<td>in</td>
<td>14.0</td>
<td>686 (211)</td>
<td>2.9</td>
</tr>
<tr>
<td>$C_{3\text{in}}$</td>
<td>in</td>
<td>3.5</td>
<td>955 (242)</td>
<td>4.0</td>
</tr>
<tr>
<td>$C_{4\text{in}}$</td>
<td>intermed</td>
<td>2.6</td>
<td>770 (183)</td>
<td>4.5</td>
</tr>
<tr>
<td>$C_{5\text{in}}$</td>
<td>intermed</td>
<td>0.4</td>
<td>1104 (213)</td>
<td>4.4</td>
</tr>
<tr>
<td>DFG-in X-ray av</td>
<td>in</td>
<td>-</td>
<td>736 (246)</td>
<td>-</td>
</tr>
<tr>
<td>DFG-out X-ray av</td>
<td>out</td>
<td>-</td>
<td>784 (316)</td>
<td>-</td>
</tr>
<tr>
<td>1P38</td>
<td>in</td>
<td>-</td>
<td>989</td>
<td>2.7</td>
</tr>
<tr>
<td>1WBT</td>
<td>out</td>
<td>-</td>
<td>1440</td>
<td>3.6</td>
</tr>
</tbody>
</table>

*Table 3-1*: Population (pop) and average volume ($V_{av}$) of the binding pockets identified by PocketAnalyzerPCA for each cluster calculated in the clustering analysis along with an approximate classification of its DFG conformation (DFG conf). Standard deviations in parentheses. Propensity for ligand binding (PLB) of structures possessing the lowest distance from each cluster centroid, calculated by Site Finder.

For the 1P38 simulation, 80% of the conformations fall into a $C_{1\text{in}}$ DFG-in-like conformation (left-hand side of Figure 3-9b, right-hand side of Figure 3-10, and Table 3-1). Clusters $C_{2\text{in}}$ and $C_{3\text{in}}$, representing ~17% of the conformations sampled, are characterized by a displacement of the Phe169 side-chain towards the $\alpha$C helix.
Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics

Figure 3-10: Average binding pocket topologies found in structures within conformational clusters identified from sesMD simulations of p38α MAP kinase (using PocketAnalyzer$^{PCA}$), mapped onto structures nearest to the cluster centroid. Each structure is color-coded according to cluster (see the caption of Figure 3-9). The pocket is color-coded according to the weighted average (high frequency, blue; low frequency, red). As references, sets of DFG-in and DFG-out crystallographic structures were used.
Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics

Clusters C4\textsubscript{in} and C5\textsubscript{in} group a smaller fraction of the ensemble, belonging only to replica 1\textsubscript{in}, that adopt the intermediate state described previously (Figure 3-7c). From inspection, we observe similarities in the transition structures sampled by the 1WBT sesMD simulation and crystal structures of p38\textalpha MAP kinase with intermediate DFG loop conformations. This is achieved despite the fact that in these crystal structures the kinase was complexed with ligand, whereas here we performed sesMD of the kinase in the absence of ligand. Good agreement is seen in the DFG conformations of 15 crystal structures of the kinase and that of replica 1\textsubscript{out} (Figure 3-11a). For example, the sesMD conformations of Asp168 and Phe169 residues in this replica superimpose to within an RMSD of 0.4 Å with p38\textalpha MAP kinase in cocrystals with pyrrolotriazine compounds (PDB codes 3BV2 and 3BV3; Figure 3-11b).\cite{45}

Furthermore, other p38\textalpha MAP kinase transitional structures from sesMD bear a resemblance to intermediates identified from a high-temperature restrained MD study of the kinase.\cite{27} In that work, pseudo-DFG-in and pseudo-DFG-out transition structures of p38\textalpha MAP kinase were defined, both lying at RMSDs of ∼6−8 Å from DFG-in and DFG-out X-ray structures.\cite{27} These high-temperature conformations were characterized by a β-turn-like structure, with the side-chain of Leu171 occupying the location normally occupied by Phe169 in a DFG-in (denoted pseudo-DFG-in) or DFG-out (pseudo- DFG-out) conformation. From the 1P38 simulation, we obtained sampling of structures akin to a pseudo-DFG-out conformation from sesMD, such that the side-chain of Leu171 replaces the position previously occupied by the aromatic ring of Phe169 (Figure 3-11c); however, only a partial pseudo-DFG-in structure was found from either the 1WBT or 1P38 simulation (Figure 3-11d).
Figure 3-11: (a, b) Superpositions of a selected p38α MAP kinase conformation from sesMD replica 1_out (blue) with (a) 15 p38α MAP kinase X-ray structures (purple) and (b) X-ray structures 3BV2 and 3BV3 (purple). (c, d) Superpositions of selected intermediate DFG conformations (blue) from sesMD of (c) replica 1_out and (d) replica 1_in. For reference, DFG-in (1P38, green) and DFG-out (1WBT, red) X-ray structures are also shown.

In assessing the opportunities for structure-based design against p38α MAP kinase, it is useful to characterize the pockets formed by the clustered sesMD conformations. For reference, we first consider the average pocket shapes arising from the 74 DFG-in and 61 DFG-out crystal structures: the location of the additional deep pocket is clearly shown in X-ray DFG-out structures relative to DFG-in (Figure 3-10). However, on average the total pocket volumes in the DFG-out and DFG-in crystal structure sets are very similar, with values of 784 and 736 Å³, respectively (Table 3-1). This hides considerable variability; for example, the volumes of pockets in 1WBT and 1P38 are 1440 and 989 Å³, respectively (Table 3-1). A range of pocket shapes (Figure 3-10) and volumes (Table 3-1) are exhibited by the conformational clusters of p38α MAP kinase from sesMD, i.e., C1_out−C8_out and C1_in−C5_in.
Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics

The pockets in C1\text{out}–C4\text{out} display a dual-lobed density, reflecting the volumes of the hinge and deep pockets, whereas structures representative of C1\text{in}–C3\text{in} possess a single pocket in the hinge region (Figure 3-10). These pocket shapes are akin to those found in the parent X-ray structures (Figure 3-10). Interestingly, the C5\text{out} kinase conformation has a pocket shape distinct from that of the C1\text{out}–C4\text{out} and X-ray structures. This cluster conformation shows an orientation of Phe169 comparable to that encountered on one of two transition pathways sampled via the aforementioned high-temperature MD study.\footnote{27} Clusters C6\text{out}–C8\text{out} clearly show diminished deep pockets and expanded cavities in the hinge region typical of the average crystallographic DFG-in pocket (Figure 3-10). For 1P38 sesMD conformations C4\text{in} and C5\text{in}, there is an opening up of the deep pocket region that is comparable to the topology of the DFG-out pocket (Figure 3-10). Indeed, cluster C5\text{in} has the largest average pocket volume of all the 1P38-derived clusters identified and closest in volume to that of the 1WBT structure (Table 3-1).

To quantify the potential of these pockets to bind small molecules, we performed an assessment of their druggability. Specifically, for the kinase structures closest to the cluster centroids, we computed a propensity for ligand binding (PLB) index\footnote{46} for pockets defined by Site Finder (we note that the kinase pocket topologies from Site Finder are broadly comparable to those obtained by Pocketanalyzer\textsuperscript{PCA}). The PLB index is based on the specific amino acid composition of a pocket and has proven effective in predicting drug binding propensities for known protein structures.\footnote{46}

Here, we first consider the PLB indices for 1WBT and 1P38 X-ray structures, which have values of 3.6 and 2.7, respectively (Table 3-1). These large positive PLB values indicate that both sites are predicted to be druggable, as expected; the DFG-out 1WBT structure has a larger index, in part a function of the larger pocket volume compared
Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics

with 1P38 (Table 3-1). This trend is also reproduced by the PLB scores predicted for the cluster centroid structures: sesMD structures based on 1WBT show a reduced PLB index by switching to a DFG-in structure, such that the PLB index of 3.9 for C1$_{out}$ decreases to 2.7 for C8$_{out}$ (Table 3-1); the opposite trend is observed for 1P38 sesMD structures progressing toward a DFG-out state (compare C5$_{in}$ with C1$_{in}$; Table 3-1). Interestingly, PLB indices greater than 4 are obtained from intermediate structures on either of these pathways, such as C4$_{in}$, C5$_{in}$, and C5$_{out}$ (Table 3-1). These structures could represent interesting starting points for structure-based inhibitor design.

3.5 Discussion and conclusions

From the preceding analysis, it is evident that application of swarm-enhanced sampling MD to the DFG motif of p38α MAP kinase provides a broader sampling of DFG conformational states compared with unbiased MD simulations. From sesMD of an initially DFG-out kinase conformation, we observe sampling of a range of loop conformations that includes DFG-in structures (Figure 3-8) and intermediate geometries (Figure 3-11a–c) resembling p38α MAP kinase conformations identified previously from crystallography$^{[45]}$ and computation.$^{[27]}$ From sesMD based on a DFG-in crystal structure, a range of DFG loop conformations is sampled, although to a lesser extent than for the simulation based on a DFG-out conformation. In particular, while a partial transition is observed for Phe169, Asp168 remains localized in an interaction with the backbone of Asn152. This hydrogen bond may point to the implication of the protonation state of Asp168 in conformational change, as suggested in work by Shan et al. on Abl kinase.$^{[28]}$ In both 1P38 and 1WBT sesMD simulations, we identify interesting intermediate pocket shapes, for example C4$_{in}$ and C5$_{out}$ pockets, that could be useful in identifying novel kinase inhibitors via virtual screening.
Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics

SesMD is a biased MD approach using multiple simulation replicas. While the concept of coupling replicas within a multicopy MD framework to improve sampling is not new\cite{16,10,47–52}, in sesMD we have implemented a novel form of the potential between replicas, incorporating attractive and repulsive terms. This swarm of simulation replicas could be viewed as somewhat akin to a swarm of van der Waals particles, simultaneously exploring mutually exclusive regions of phase space. An alternative multicopy MD approach to enhanced sampling is the replica exchange method (REM)\cite{12}, where at periodic intervals neighbouring replicas may be swapped between simulations. The key choice is in the proximity of the replicas in terms of temperature (for temperature REM) or potential (for Hamiltonian REM) such that efficient exchange can occur in accord with the requirement of detailed balance, without requiring too many replicas to span the desired range of temperature or potential. This has proven to be an issue for large explicitly solvated solutes, which require a large number of replicas to achieve this overlap. This problem can be mitigated by using a new implementation of the replica exchange with solute tempering (REST2) method\cite{53}, where the replica Hamiltonians are scaled in order to remove the dependence on the number of explicit water molecules.

SesMD is not subject to the energy overlap requirement of REM. However, in its current implementation, sesMD exchanges coordinate/dihedral information between all replicas at each time step, in contrast to REM, which compares only neighbouring replicas typically at every picosecond. As a result of this communication in sesMD, for a given time step, the decrease in speed of sesMD compared with the equivalent set of independent MD trajectories is $\sim 38\%$. However, for the 5 ns mixed sesMD/MD protocol applied here to p38α MAP kinase, an overall decrease in speed of only 14% was found, while the degree of sampling within the computed 5 ns was markedly
Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics

improved for sesMD relative to 24 independent MD trajectories of 5 ns or a single 50 ns trajectory. From REM methods, as with the popular single-trajectory enhanced sampling approaches of accelerated MD\cite{11} and metadynamics,\cite{15,16} it is possible to obtain Boltzmann-weighted ensemble properties directly or indirectly. This is also formally the case for sesMD: because of the well-defined Hamiltonian of the overall swarm of trajectories, one can reweight the swarm of replicas according to the approach of Torrie and Valleau\cite{7} (see the Supporting Information).

We do not apply this approach here, where we scale the ses potential in seeking to drive conformational exploration of the kinase. Nevertheless, the general topology of the underlying unbiased potential energy surface appears to be preserved for sesMD of the model butane and alanine dipeptide systems (Figure 3-2b,d), and experimentally observed intermediate DFG conformations of the kinase are obtained during sesMD (Figure 3-11a,b). We note that other biased MD approaches have been employed to study kinase conformational change: for example, targeted MD simulations have been used to study the structural reorganization of Abl kinase\cite{54} and EGFR kinase.\cite{31}

However, distinct from sesMD, a priori knowledge of the alternative protein conformation is required for targeted MD. Another biased MD study performed nonequilibrium MD simulations of p38α MAP kinase in order to explore DFG conformation.\cite{26} Steering forces were applied to the Asp168 and Phe169 side-chains of the DFG loop; the direction and magnitude of the pulling forces required recalibration at different stages in the conformational change.

As with this nonequilibrium MD approach and the metadynamics method, for sesMD a choice must be made as to the degrees of freedom to which to apply the biasing potential, although in the case of the kinase this is intuitive, i.e., DFG backbone torsions. Potentially, for a given protein crystal structure, the choice of the amino acid
Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics
dihedral set to explore by sesMD could be guided by experimental insight, for example
from active-site B-factors, or systematically mapped for active-site residues.
Accordingly, it will be important to assess the level of transferability of ses parameter
to other systems of interest, for small and larger molecules. We note that careful
system-dependent parametrization is required for other biasing methods such as
metadynamics and accelerated molecular dynamics. For the latter, to obtain suitable
parameters, an initial estimate of the energy landscape to explore is obtained from
simulation in the absence of biasing potential prior to trialling short biased simulations.
We anticipate that future work will provide insight into the required parameter ranges
and protocols for optimal sesMD applications. SesMD therefore constitutes a
straightforward, parallel, and promising approach to conformational sampling. Here
we have applied sesMD to the generation of a range of possible alternative
conformations of a protein based on a single crystal structure. Such a strategy applied
to a therapeutic target protein could reveal novel protein concavities of potential use
in providing new directions for structure-based inhibitor design.
3.6 References


Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics


Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics

Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics


Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics

3.7 Supporting Information

Pentane and dialanine simulations

The *gaff* force field was used for pentane and ff99SB force field for dialanine. The TIP3P model was used for water. SHAKE was applied and a 1 fs time step. The particle mesh Ewald method was employed for long range electrostatic interactions and a 9 Å cut-off applied for non-electrostatic long range interactions. The *ses* potential was applied to the central $\phi_1$ and $\phi_2$ angles of pentane and $\phi\psi$ backbone torsions of dialanine. For both pentane and dialanine, the *sesMD* parameters $(A,B)$ and $(C,D)$ were taken as $(-10 \text{ kcal/mol}, 0.2 \text{ rad}^{-1})$ and $(15 \text{ kcal/mol}, 0.8 \text{ rad}^{-1})$ respectively (we note these parameters are smaller than those employed in the kinase *sesMD* simulation, reflecting the covalent constraints on the DFG loop by the surrounding protein matrix). Eight replicas were used for unbiased MD and *sesMD* simulations. Equilibration of each replica, initiated with different velocities, was performed at 300 K for 100 ps; for *sesMD* this was followed by 500 ps of annealing *ses* parameters to their full values; and then by 10 ns of *sesMD*. For the independent replica simulations, this was 10 ns of unbiased MD simulation per replica.

Reweighting of *sesMD* properties

We outline below an approach to recovering Boltzmann-weighted properties from a *sesMD* simulation. The total potential acting on a swarm of $M$ simulation replicas $\alpha$, within a *sesMD* simulation, is given by:

$$V^{tot} = V^{ses} + \sum_{\alpha} V^{MM}_{\alpha}$$

(S1)

1 The reweighting algorithm has been written and tested exclusively by Dr. Richard A. Bryce and Dr. Kepa Burusco Goñi.
Exploring protein kinase conformation using swarm-enhanced sampling molecular
dynamics

that is, the sum of the molecular mechanics force field potentials acting on all replicas
and the \( ses \) potential acting on the whole swarm. As written in Eq. 2, the \( ses \) potential
\( V^{ses} \) is a function of a set of torsional angles, which ultimately depend on the Cartesian
coordinates of atoms within all replicas. As the force field potentials of each replica
are non-interacting and additive, if we define the vector \( \mathbf{r}^M \) as the \( 3NM \) dimension
vector describing the coordinates of \( N \) atoms in \( M \) replicas, we can rewrite Eq. S1 as:

\[
V^{tot}(\mathbf{r}^M) = V^{ses}(\mathbf{r}^M) + V^{MM}(\mathbf{r}^M)
\]

(S2)

where \( V^{MM}(\mathbf{r}^M) \) is the sum of force field potentials \( V^{MM}_{\alpha}(\mathbf{r}_\alpha) \).

Any extensive quantity \( X \), which can be described as a function of coordinates \( \mathbf{r} \), can
be described as a function of \( \mathbf{r}^M \) according to:

\[
X(\mathbf{r}) = \frac{1}{M} X^M(\mathbf{r}^M)
\]

(S3)

where

\[
X^M(\mathbf{r}^M) = \sum_{\alpha} X_{\alpha}(\mathbf{r}_\alpha)
\]

(S4)

The expectation value of \( X \) within the generalised ensemble of a sesMD simulation
can therefore be obtained from:

\[
\langle X \rangle_{ses} = \frac{1}{M} \int d\mathbf{r}^M X^M(\mathbf{r}^M) \exp\left(-\beta V^{tot}(\mathbf{r}^M)\right)
\]

(S5)

\[
\int d\mathbf{r}^M \left(-\beta V^{tot}(\mathbf{r}^M)\right)
\]

Substituting for \( V^{tot} \) we obtain:

\[
\langle X \rangle_{ses} = \frac{1}{M} \left[ \int d\mathbf{r}^M X^M(\mathbf{r}^M) \exp\left[-\beta(V^{ses}(\mathbf{r}^M) + V^{MM}(\mathbf{r}^M))\right]\right]
\]

(S6)
Exploring protein kinase conformation using swarm-enhanced sampling molecular
dynamics

According to the approach of Torrie and Valleau\cite{1} to recover a Boltzmann weighted
average from our generalised ensemble, we may correct for the swarm bias acting on
each replica according to:

\[
\langle X \rangle = \frac{1}{M} \int d\mathbf{r}^M \mathbf{X}^M (\mathbf{r}^M) \exp \left[- \beta (V_{\text{ses}}^M (\mathbf{r}^M) + V_{\text{MM}}^M (\mathbf{r}^M)) \right] \exp \left[ \beta V_{\text{ses}}^M (\mathbf{r}^M) \right] 
\]

(S7)

It is possible to simplify Eq. S7 by dividing the numerator and denominator by

\[
\int d\mathbf{r} \exp \left[- \beta V_{\text{tot}}^M (\mathbf{r}^M) \right] 
\]

to give:

\[
\langle X \rangle = \frac{1}{M} \frac{\int d\mathbf{r}^M \mathbf{X}^M (\mathbf{r}^M) \exp \left[- \beta (V_{\text{ses}}^M (\mathbf{r}^M) + V_{\text{MM}}^M (\mathbf{r}^M)) \right] \exp \left[ \beta V_{\text{ses}}^M (\mathbf{r}^M) \right]}{\int d\mathbf{r} \exp \left[- \beta V_{\text{tot}}^M (\mathbf{r}^M) \right]}
\]

(S8)

Using angular brackets to denote an ensemble average, we may express this as:

\[
\langle X \rangle = \frac{1}{M} \left\langle \frac{\mathbf{X}^M (\mathbf{r}^M) \exp \left[ \beta V_{\text{ses}}^M (\mathbf{r}^M) \right]}{\exp \left[ \beta V_{\text{ses}}^M (\mathbf{r}^M) \right]} \right\rangle_{\text{ses}}
\]

(S9)

We also note that a useful discussion of reweighting coupled replicas can be found in
Malevanets et al.\cite{2}.


Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics

**Figure 3-12**: Dependence of biasing potential $V_{\text{ses}}$ (in kcal/mol) on $d_{\alpha,\beta}^{\text{rms}}$, the root-mean-square dihedral angle distance (in degrees) between swarm members $\alpha$ and $\beta$, for 1WBT DFG-out (black line) and 1P38 DFG-in (red line) sesMD simulations.
Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics

Figure 3-13: Heavy atom RMSD of DFG motif with respect to the first frame of the simulation (black), DFG-out (red) and DFG-in (green) X-ray structures for (a) sesMD 24 replicas from sesMD and (b) unbiased MD simulations initiated from p38α MAPK structure 1WBT. Replicas are sorted from left-to-right, top-to-bottom.
Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics

Figure 3-14: Heavy atom RMSD of DFG motif with respect to DFG-out (red) and DFG-in (green) X-ray structures for (a) sesMD 24 replicas from sesMD and (b) unbiased MD simulations from p38αMAPK structure 1P38. Replicas are sorted from left-to-right, top-to-bottom.
Figure 3-15: Superposition of the activation loop of selected conformations from sesMD replica 2_{out} after 0.5 ns (orange) and after 5 ns (blue) of simulation with 55 DFG-in p38α MAP kinase X-ray structures (white). For reference, DFG-in 1P38 (green) and DFG-out 1WBT (red) crystallographic structures are shown.
Exploring protein kinase conformation using swarm-enhanced sampling molecular dynamics

In order to illustrate the structural transition from DFG-out to DFG-in observed in the sesMD trajectory of replica 2_{out} we included a sample movie of the above-mentioned trajectory. In this movie is it possible to observe the movement from the initial DFG-out structure toward the opposing DFG-in conformation with the aromatic ring of Phe169 occupying the deep pocket.

As reference, we also included a sample movie of a control simulation we ran from the same initial conditions of replica 2_{out} using 50 ns of unbiased MD. Here no transition is observed, but instead the DFG residues of the kinase tend to remain close to the initial DFG-out conformation.

The two movies are available at http://pubs.acs.org/doi/suppl/10.1021/ci5003334
Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis

4. Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis

Alessio Atzori, a Berthold Wroblowski, b Pascal Bonnet c and Richard A. Bryce a

a Manchester Pharmacy School, University of Manchester, Oxford Road, Manchester, M13 9PT, UK
b Janssen Research & Development, a division of Janssen Pharmaceutica N.V., Turnhoutseweg 30, 2340 Beerse, Belgium
c Structural Bioinformatics & Chemoinformatics, Institut de Chimie Organique et Analytique (ICOA), UMR CNRS-Université d’Orléans 7311, Université d’Orléans, Rue de Chartres, F-45067 Orléans cedex 02, France
4.1 Abstract

R-spondins comprise a class of proteins that play a key role in activation of Wnt signalling, a pathway responsible for tissue patterning. Existing crystallographic structures of R-spondin 1 find rotation around an interdomain hinge can lead to a range of open conformations or in one case a closed orientation. In this work, we employ microsecond molecular dynamics simulations to probe the degree of conformational flexibility in unbound isoforms R-spondin 1 and R-spondin 2 in aqueous solution. Our simulations find that the closed conformation remains stable on the microsecond timescale, whereas an open structure, adopted in the ternary complex of R-spondin 1 with its client proteins, appears to gradually relax towards more closed-like conformations when free in solution. Simulation of isoform R-spondin 2, for which no closed X-ray structure has been determined, finds that relaxation from an open towards a closed state occurs but to a lesser extent. In addition to highlighting the importance of this closed state in understanding R-spondin interactions, the conformation is also predicted to possess a particularly druggable pocket in the interdomain region, from site mapping and virtual screening analyses. Therefore, this work suggests that the R-spondin plasticity could be exploited in the pursuit of structure-based design of Wnt pathway inhibitors as anti-cancer therapeutics.
4.2 Author Summary

The Wnt pathway is involved in development and differentiation of tissue. Of the molecular interpreters featuring in the Wnt pathway, roof plate-specific spondin (R-spondin) proteins have recently been recognized to play a key role in its regulation. The targeting of the R-spondins by drug discovery methods could lead to new treatments for cancer. Structural characterisation of R-spondins by experimental methods has found a flexibility intrinsic to this family of proteins, which are found either in a closed or in an open state. In this study, we employ long molecular dynamics simulations to assess the structural plasticity of R-spondins. We find that the closed conformation of R-spondins is maintained over the course of the simulation, while trajectories initiated from open forms are able to approach closed-like structures. Furthermore, computational analysis of R-spondin structures obtained from the simulations reveal the presence of a potentially druggable binding region. Therefore, our work suggests that R-spondin flexibility be targeted in the pursuit of effective Wnt pathway inhibitors.
Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis

4.3 Introduction

A wide range of signalling pathways are involved in development, differentiation and regulation of tissue patterning. One of the most important systems regulating these processes is represented by the canonical Wnt (cWnt) pathway.[1-3] Among the interpreters implicated in the cWnt pathway, roof plate-specific spondin (R-spondin) proteins[4] have recently been recognized to play a particularly important role in its regulation. The genome of vertebrates encodes four different R-spondin isoforms (R-spondin 1 – R-spondin 4), which share 40-60% amino acid sequence identity and have an approximate mass of 35 kDa. Several studies have demonstrated the close relationship between R-spondins and the Wnt pathway: evidence ranges from the simultaneous expression of R-spondins and Wnt ligands during development[5,6] to the phenotypic similarities observed due to absence of expression of Wnt ligands and/or receptors[7-9] compared with mutations of R-spondin proteins.[10-13]

A recent study[14] has found that R-spondins increase Wnt signalling by interacting with transmembrane E3 ubiquitin ligase zinc and ring finger 3 (ZNRF3) or its homologue ring finger 43 (RNF43); and with the leucine-rich repeat containing G proteins (LGR) 4, 5, or 6. This ternary complex of LGR–R-spondin–RNF43/ZNRF3 leads to augmented membrane clearance of ubiquitin ligase; to accumulation of Frizzled (Fz) Wnt receptors and lipoprotein receptor-related protein (LRP) 5 and 6 co-receptors on the cell membrane; and to enhancement of Wnt signalling. Conversely, in the absence of R-spondins, RNF43/ZNRF3 promotes the membrane clearance of Fz-LRP complex, which downregulates Wnt signalling.[14,15] The abnormal activation of the Wnt pathway is associated with uncontrolled cell growth and consequently with oncogenesis in several tissues, including colon, skin, liver and ovary.[3,16-19]
Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis

A range of different approaches to targeting the components of the cWnt pathway has been described\cite{20-23}, e.g. focusing on members of the destruction complex (CK1\cite{24}, GSK3\cite{25} and Axin\cite{26}) or targeting the TCF/β-catenin complex.\cite{27} An alternative approach is represented by the disruption of the ternary complex, LGR–R-spondin–RNF43/ZNRF3. In this way, a hypothetical hyperactivated Wnt signalling pathway, due to overexpression of Fz receptors, is targeted by reducing the clearance of RNF43/ZNRF3.

In the pursuit of a deeper understanding of the structural basis of R-spondin action and their potential inhibition, several X-ray structures of R-spondin 1 and 2 have emerged for the uncomplexed state\cite{28,29} and bound to LGR\cite{28,30,31} or RNF43/ZNRF3.\cite{29,32}

These include a structure of the ternary complex (PDBID: 4KNG\cite{33}), providing information at the atomic level of the interactions between R-spondin 1, LGR, and RNF43/ZNRF3.

Structurally, the R-spondins are characterized by two furin-like cysteine-rich regions (Fu1 and Fu2), arranged to form a ladder-like structure composed of six β-hairpins; each domain possesses four disulfide bridges (Figure 4-1a). Interactions with LGRs and RNF43/ZNRF3 involve two different regions of R-spondins: an N-terminal region, which has been referred to as the “head module” (residues Gln8-Cys49, Figure 4-1a,b; note we use residue numbering following the R-spondin 1 sequence) interacts with RNF43/ZNRF3, while the C-terminal “rod module” (residues Pro50-Ser113, Figure 4-1a,b) is responsible for contacts with LGR.\cite{33}
Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability

Figure 4-1: a) Structure of free R-spondin 1, which adopts a closed conformation (PDBID: 4BSO). The two furin domains (red, Fu1 and blue, Fu2) and the head and rod modules are defined. The eight disulfide bridges and the residues involved in binding with RNF43/ZNRF3 are shown as sticks. b) X-ray structure (PDBID: 4KNG) of the ternary complex comprised by LGR5 (orange), R-spondin 1 (red) and RNF43 (cyan). c) Superposition of X-ray structures of free, closed R-spondin 1 (4BSO, green), bound R-spondin 1 (PDBID: 4KNG, red), and unbound R-spondin 2 (PDBID: 4C8V in blue). Both 4KNG and 4C8V structures show R-spondin in an open conformation. d) Molecular surfaces calculated for closed 4BSO (left), open 4KNG (centre) and open 4C8V (right) X-ray structures. Molecular surfaces are coloured red for H-bonding regions, blue for other polar regions and white for hydrophobic regions. The potential druggable pocket enclosed between the long β-hairpin moiety of the head module and the first β-hairpin of the rod module is indicated with a red circle.
Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis

The long central β-hairpin in the head module of R-spondin (Phe31-Leu46) is fundamental for the interaction of R-spondin with RNF43/ZNRF3, and in particular the conserved residues Leu34, Arg36 and Gln41\[^{[33]}\]. Arg36 and Gln41 form hydrophilic interactions with RNF43/ZNRF3 residues Gln84, His86 and Asp97 (which comprise part of a shallow pocket in the latter); Leu34 of R-spondin 1 is responsible for hydrophobic contacts with His86, Leu88 and Tyr89 of RNF43/ZNRF3. These three residues, Leu34, Arg36, and Gln41, are present in all four members of the R-spondin family, except that in R-spondin 4 Leu64 is replaced by Ile; hence they represent a key recognition triad for the interaction between R-spondins and RNF43/ZNRF3. A further point of contact is the R-spondin residue projecting towards the groove created by Leu82 and Met83 of RNF43/ZNRF3: this residue is an isoleucine in R-spondins 1 and 4, and a methionine in R-spondin 2 and 3. The presence of the methionine seems to be associated with greater activation of the Wnt pathway in these members of the family.\[^{[34]}\]

Interestingly, in addition to shedding light on R-spondin-ligand interactions, X-ray structures are suggestive of considerable plasticity in R-spondin conformation, enabling a range in rotation around the hinge between the two furin domains.\[^{[29]}\] For example, when their Fu1 domains are superimposed, uncomplexed R-spondin 1 (PDBID: 4BSO, Figure 4-1c, green) exhibits a ~90° rotation of its Fu2 domain with respect to its conformation in the ternary complex (PDBID: 4KNG, Figure 4-1c, red) or in the uncomplexed R-spondin 2 structure (PDBID: 4C8V, Figure 4-1c, blue). This rotation occurs around a hinge defined by residues Lys66-Ile69, conferring on uncomplexed R-spondin 1 what we denote here a distinct closed conformation (Figure 4-1c, green). In other X-ray R-spondin structures available, which are either in the ternary complex or uncomplexed, the angle of rotation between the two furin domains
Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis

is ~40°, and hence, these structures also possess a more linear open conformation, similar to that of 4KNG and 4C8V (Figure 4-1c, red and blue). However, it is not clear if this closed conformation for the uncomplexed 4BSO structure of R-spondin 1 is a consequence of the crystalline environment or is a genuinely unbound solution structure. Therefore, in this work, we apply microsecond molecular dynamics (MD) simulations to the closed form of R-spondin 1 in explicit aqueous solvent in order to evaluate its stability in solution. We also simulate the dynamics of free R-spondin 1 initiated in its LGR/RNF43-bound conformation and R-spondin 2 in its unbound conformation to assess their flexibility, and in particular the ability of these more extended structures to adopt a closed conformation.

As the R-spondins represent an interesting potential target on the Wnt pathway, we also explore the potential for ligand design into R-spondin conformations in such a way as to inhibit its binding to RNF43/ZNRF3. Inspection of the R-spondin–RNF43/ZNRF3 binding interface in X-ray structures of the R-spondins shows a potentially druggable pocket enclosed between the long β-hairpin moiety of the head module and the first β-hairpin of the rod module (circled in Figure 4-1d). The presence of a ligand capable of interacting with this region could inhibit formation of the ternary complex, LGR–R-spondin–RNF43/ZNRF3, and hence lead to a decrease in Wnt signalling via augmented activity of the ubiquitin ligases towards the complex Fz/LRP.

We therefore probe this area of the protein for sites of interaction, based on the MD ensembles of R-spondin conformations, using in silico druggability analysis and screening of small molecule fragments.
Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis

4.4 Materials and Methods

**Computational details.** Three initial structures for MD were considered: R-spondin 1 in its free form (PDBID: 4BSO\textsuperscript{[28]}, resolution 2.2 Å), R-spondin 1 in complex with LGR5 and RNF43 (PDBID: 4KNG\textsuperscript{[33]}, resolution 2.5 Å) and unbound R-spondin 2 (PDBID: 4C8V\textsuperscript{[29]}, resolution 2.2 Å). The MODELLER software package (version 9.13)\textsuperscript{[35]} was used to homology model unresolved residues of the C-termini portion of the X-ray structures 4BSO (residues Glu101-Thr109) and 4KNG (residues Ser133-Gly138). Among the predicted models, the ones with the lowest DOPE and molpdf scores were selected. The proteins were simulated using the AMBER ff12SB force field\textsuperscript{[36]} and the TIP3P solvent model.\textsuperscript{[37]} Counterions were added to the system in order to neutralize the net charge of the systems and give an ionic strength of 0.1 M. The SHAKE algorithm\textsuperscript{[38]} constrained bonds between hydrogen and heavy atoms. A 2 fs time step was used. Simulations were performed in the NPT ensemble with a Langevin thermostat\textsuperscript{[39]}, using a collision frequency of 2 ps$^{-1}$. A Berendsen isotropic barostat was used to maintain an average pressure of 1 atm. The particle mesh Ewald method\textsuperscript{[40]} was used for long range electrostatic interactions and a 10 Å cutoff for non-electrostatic interactions. The sequences followed a multistep equilibration protocol of 700 ps followed by 1 µs of production. MD simulations were performed using the AMBER 12 molecular simulation package\textsuperscript{[36]} using NVIDIA K20 graphics processing units. Configurations were archived every 10 ps for analysis.

Analysis of trajectories was performed using the ptraj and cpptraj modules of AMBER 12. The conformations of R-spondin structures were aligned to the backbone atoms of the Fu1 domain in order to calculate the RMSD of the backbone atoms of the first β hairpin of the Fu2 domain (residues Cys72-Cys84 and Cys101-Cys113 for R-spondin 1 and R-spondin 2, respectively). The angle of rotation between the Fu1 and Fu2
Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis

domains of R-spondins was calculated using DynDom.\[41,42\] For principal component analysis (PCA), the conformations were fitted onto the C α atoms of an average structure calculated across all conformations for each simulation. With these fitted conformations, PCA of the atomic displacements of the first β hairpin of the Fu2 domain was performed using the ptraj module of Amber 12. Pocket detection and druggability analysis were performed on a subset of conformations taken every 0.5 ns for each MD simulation using PocketAnalyzer\[43\] and Site Finder\[44\] from MOE\[45\]. These tools were used to characterize the protein cavity in the region of interest, defined as (i) residues Leu33, Leu34, Glu35, Arg36, Asn37, Gln41, Tyr53, Lys63, Cys64, Ile65, Lys66, Lys68, Phe76, Ser77, His78, Asn79 for the R-spondin 1 conformations; and (ii) residues Tyr62, Leu63, Arg64, Arg65, Glu66, Gln70, Tyr82, Arg92, Cys93, Ser94, Arg95, Arg97, Phe105, Ser106, Arg107, Asp108 for R-spondin 2. The conformations showing the highest PLB score for each sequence were used as a target for subsequent docking of molecular fragments.

The ligand libraries used in this study were prepared using LigPrep\[46\] from the Maestro 9.8 suite\[47\], which generated up to 10 stereoisomers, tautomers and protonation states corresponding to a pH of 7.0. The generation of the conformers was followed by energy minimisation using the OPLS2005 force field.\[48\] A library composed of 1748064 fragments obtained from ZINC\[49\] was selected for our docking study. The selection criteria of the dataset were: log P ≤ 3.5; molecular weight ≤ 250 Da; and rotatable bonds ≤ 5. From the entire dataset, we considered a subset which contained 10342 fragments, clustered at the 60% Tanimoto similarity (downloaded from ZINC on the 5/05/2014). After the LigPrep preparation stage, this library contained 15354 fragments. Docking simulations were performed using GOLD version 5.2 (50,51) and the CHEMPLP scoring function\[52\] was used. The default
Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis

parameters of the automatic settings were used to set the genetic algorithm parameters with a search efficiency of 30% (virtual screening). Based on the top 10 docked poses for each pocket, a consensus pharmacophore was derived using MOE and used for subsequent screening of the library, in order to assess the key binding features of the site.

4.5 Results and Discussion

We evaluate the conformational plasticity of R-spondins in aqueous explicit solvent using 1 μs MD simulations, considering in turn isoforms 1 and 2.

R-spondin 1. Two MD simulations to explore the conformational plasticity of isolated R-spondin 1 in solution were performed, initiated either from the uncomplexed R-spondin 1 4BSO X-ray structure (which we label R1u) or from the 4KNG X-ray structure of R-spondin 1 in complex with LGR5 and RNF43 (labelled R1c). As a measure of R-spondin flexibility, we first consider the root mean square deviation (RMSD) in atomic position of backbone atoms of a β-hairpin of the Fu2 domain when the Fu1 domains were aligned (for more details, see Materials and Methods). Over the duration of the R1u simulation, RMSD values remain in the region of the initial closed structure, at ~5 Å, Figure 4-2a, black) and somewhat removed from the open 4KNG structure, at ~11 Å (Figure 4-2a, red).

By contrast, the R1c system exhibits significant shifts in RMSD over the course of the 100 ns trajectory: for example showing transient rises from 2.5 Å to over 10 Å at 250 ns and again 370 ns (Figure 4-3c, black). This variation are accompanied by large shifts in RMSD with respect to the closed 4BSO structure, for example approaching 5 Å at 400 ns (Figure 4-2c, red).
Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis

Figure 4-2: Time series of backbone atom RMSD (in Å) from 1 μs MD simulations calculated against the first frame of the simulation (black) and reference X-ray structure (red) for backbone atoms of the second β hairpin of the Fu2 of R1U (a), R1C (c) and R2U (e). Time series of angle of rotation (in degrees) between the first furin domain (aligned) and the second furin domain from 1 μs MD simulations of R1U, R1C (d) and R2U (f). Root mean square fluctuations (RMSF) of backbone atoms as a function of residue number from 1 μs MD simulations of R1U, (green) R1C (red) and R2U (blue).

The plasticity of R-spondins can be measured as well from the calculation of the root-mean square fluctuation (RMSF) of the backbone atoms: RMSF curves for R1U (Figure 4-2g, green) and R1C (Figure 4-2g, red) assume similar values in the first furin domain (residues Gln8-Lys66), whereas the trend in the second domain (residues Cys67-Ser113) characterise the diverse flexibility which characterises the open and the closed state of R-spondins. If we exclude the C-terminal residues, we can identify for R1U the long β hairpin of the Fu1 domain (residues Leu34-Gln41) as the portion with the highest flexibility in both sequences, along with residues Asp55-Lys63. Furthermore, it is possible to highlight the different flexibility of the first β hairpin of the Fu2 domain (residues Cys72-Cys84) of R1U compared to R1C, indicating how the closed conformation hinders the elasticity of this portion of the protein.
Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis

It is also instructive to consider snapshots from the R₁U and R₁C trajectories. The 4BSO and 4KNG crystal structures provide references for the R-spondin 1 conformations sampled during the MD simulations (green and red respectively, Figure 4-3a,b).

**Figure 4-3**: a) Superposition of a selected set of conformations from R₁U simulation (white) with 4BSO (green) and 4KNG (red) X-ray crystal structures. b) Superposition selected set of conformations from R₁C simulation (white) with 4BSO (green) and 4KNG (red) X-ray crystal structures c) Salt bridges identified between residues Arg36-Glu73. d) Salt bridges identified between pairs of residues Glu35-Lys85 (black), Arg36-Glu73(red), Arg36-Glu70 (green), Arg36-Glu86 (blue) for 1 μs MD simulation of R₁C.
Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis

For the duration of the $R_{1u}$ simulation, conformations aligned on Fu2 appear to cluster around its initial 4BSO structure (green, Figure 4-3a) and do not sample in the region of the 4KNG structure (red, Figure 4-3a). This again suggests the stability of the closed conformation, which maintains a $\sim 15^\circ$ angle of rotation between the two furin domains (Figure 4-2b). By contrast, superposition of Fu2 of selected conformations from the $R_{1c}$ trajectory show considerable rotation in orientation of Fu1 from its initial 4KNG structure (red, Figure 4-3b), some of which approach the closed-like conformation of 4BSO (green, Figure 4-3b).

This is reflected by the time series for the angle of rotation between Fu1 and Fu2 measured for $R_{1c}$ conformations: this angle fluctuates considerably and latterly samples down to values of $\sim 40^\circ$ (Figure 4-2d); this value is considerably removed from the $\sim 90^\circ$ angle measured for the 4KNG X-ray structure, and approaches the value of $15^\circ$ for 4BSO in its closed conformation (Figure 4-2d). However, the profile in RMSD (Figure 4-2c) and fluctuations in the interdomain angle may be suggestive of longer time scale relaxation despite undulations.

The rise in RMSD at $\sim 250$ ns in $R_{1c}$ appears to be mainly due to the formation of four transient salt bridges: these involve residues Glu35 and Arg36 of the second $\beta$ hairpin of the Fu1 domain and residues Glu70, Lys85 and Glu86 of the Fu2 domain (Figure 4-3d). These salt bridges involve formation of a compact R-spondin 1 conformation (Figure 4-3c). However, these interactions are not observed during the $R_{1u}$ simulation or in the 4BSO or 4KNG X-ray structures: in fact, due to the characteristic structure of the closed conformation, the folding of the first $\beta$-hairpin of the second furin domain actually orientates residues Glu70-Leu90 away from the residues of the long $\beta$-hairpin of the first furin domain.
Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis

**R-spondin 2.** For comparison, we now consider MD simulation of the uncomplexed form of R-spondin 2 and specifically its ability to form a closed conformation. This simulation was initiated from the (open) X-ray structure 4C8V and is denoted R2U. We compare the R-spondin 2 trajectory with its initial geometry and the 4BSO closed R-spondin 1 structure. The RMSD for the Fu2 domain, calculated from its initial structure, remains fairly stable between 2.5 and 4 Å over the microsecond (black, Figure 4-2e). The R2U simulation also remains ~10 Å from the closed 4BSO structure, although with a dip below 7.5 Å at 890 ns (red, Figure 4-2e).

**Figure 4-4:** a) Superposition of a selected set of conformations from R2U simulation (white) with 4C8V (blue), 4BSO (green), 4KNG (red) X-ray crystal structures and the most similar closed conformation from Ric (yellow). b) Hydrogen bond identified between residues Arg97-Arg107 (according to the numbering defined in X-ray structure 4C8V) c) Time series of the hydrogen bond identified between residues Arg97-Arg107 (according to the numbering defined in X-ray structure 4C8V) for 1 μs MD simulation of R2U.
Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis

The RMSF curve for the Fu1 domain of R2u assumes a pattern more similar to R1c than R1u (Figure 4-2g, blue), while the opposite is observed for the Fu2 domain. Indeed, the residues of the first β hairpin of the Fu2 domain (residues Cys101-Cys113) are more flexible than the ones of R1u and comparable to the ones of R1c, demonstrating again how the closed conformation limits the flexibility of this portion of the sequence.

Superposition of selected conformations from the R2u trajectory (white, Figure 4-4a) indicates that sampling is localized around the initial 4C8V X-ray structure of R-spondin 2 (blue, Figure 4-4a), although with considerable fluctuation. The similarity between this X-ray structure and that of the extended bound R-spondin 1 conformation is also apparent (red, Figure 4-4a).

Although the R-spondin 2 structures sampled do not reach a conformation comparable to the closed 4BSO structure (green, Figure 4-4a), R2u shows some flexibility, with the rotation angle around the hinge within R2u conformations approaching during the second half of the simulation values of ~50° from an initial values of ~90° (Figure 4-2f). We note that the stability of the RMSD may correlate with the subsequent formation of stable interactions between Arg97, located in the first β hairpin of the rod domain, and residues Arg107 and Asp108, located at the tip of the adjacent second β hairpin (Figure 4-4b,c).

From the superposition of 22 crystallographic structures available for R-spondin 2, Zebisch et al. observed a variation in interdomain orientation ranging from 50 - 60° (Figure 4-5c). Here we find that molecular dynamics simulations also capture this high rotational flexibility between domains. Indeed, for the R2u simulation we find a comparable but more expansive behaviour, with the angle of rotation between the two domains spanning a broader range, from 45 - 125° (Figure 4-5d).
Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis

**Figure 4-5:** a) Superposition of selected conformations from $R_{1u}$ based on first furin domain (red). b) Superposition of selected conformations from $R_{1C}$ based on first furin domain (red). c) Superposition of the available R-spondin 2 crystal structures based on first furin domain (red). d) Superposition of selected conformations from $R_{2u}$ based on first furin domain (red). In order to compare the extent of rotation between the two furin domains, His78 of R-spondin 1 and the corresponding Arg107 of R-spondin 2 are coloured in yellow.
Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis

Figure 4-6: (top) Projections onto the first two principal components of the conformations from for 1 μs MD of R1U (a), R1C (b) and R2U (c), along with 4BSO (green square) 4KNG (red square) and 4C8V (blue square) reference crystal structures. The projections are colour-coded according to the simulation time. (bottom) Distribution of R1U (d), R1C (e) and R2U (f) conformations as a function of the distance between Cα of residues Hie78 and Cys64/Lys66 (of R1U and R1C) and residues Arg107 and Cys93/Arg95 (of R2U). Blue regions indicate areas of highest density.

To provide a more comprehensive representation of the sampling captured from the simulations, we performed PCAs of the backbone heavy atoms of the first β hairpin of the Fu2 domain. The two top principal components from these PCAs account for 63%, 68% and 62% of the variance in atomic displacements of the selected atoms of R1U, R1C and R2U simulations, respectively. The projection of these principal components of the X-ray crystal structures, which represent the open and the two closed states of R-spondins, are distinctly separated (Figure 4-6a,b,c). As expected, conformations from R1U (Figure 4-6a) are localised around the area defined by the initial structure of the simulation. By contrast, R1C (Figure 4-6b) and R2U (Figure 4-6c) simulations explore the principal component space more extensively, and approach the opposite conformational state represented by the 4BSO X-ray structure in both cases (Figure 4-6b,c green square).
Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis

However, the higher flexibility of the first β hairpin of the Fu2 domain of R1C compared to R2U has been captured by the PCA, mirroring the trends observed with the RMSD and RMSF analysis. Furthermore, the projections of the two first principal components for R2U are divided into two clusters: this separation reflects the difference registered in the angle of rotation between the two furin domains within the first and the second half of the simulation.

The different sampling of conformational space covered by the simulations can also be illustrated using representative interatomic distances capable of differentiating the two states: the distance between the Ca of His78 for R-spondin 1 and Arg107 for R-spondin 2, both located in the first β hairpin of the Fu2 domain, and two residues located in the more stable portion of the second β hairpin of the Fu1 domain (namely Cys64 and Lys66 for R-spondin1, and Cys9 and Arg95 for R-spondin2). As observed previously, the sampling for R1U is localised around the initial region of the landscape (Figure 4-6d), confirming the stability of the closed conformation. Oppositely, a wider coverage of the sampling space for this portion is found for R1C and R2U. In the case of R1C we can notice a shift towards the distances measured in the 4BSO X-ray structure, while the remaining conformations are localized in a region close to the starting structure. R2U conformations instead are mainly concentrated in an alternative well, which appears to lie between the initial state and 4BSO reference structure.

From these MD simulations of R-spondin 1 and 2, we may make several observations: the initial crystallographic closed conformation of R1U (Figure 4-1c, green) appears to be stable in solution on the microsecond timescale (Figures 4-3a, and 4-5a). However, the more extended initial conformation of R1C, taken from the ternary complex, demonstrates higher flexibility than R1U (Figures 4-3b and 4-5b) and also periodically samples closed-like conformations (Figures 4-2d and 4-3b). Moreover, due to
Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis

interdomain contacts, alternative compact conformations of R-spondin 1, not observed from crystallography, were observed during the dynamics (Figure 4-3c,d). Interestingly, simulation of R-spondin 2 in an extended conformation (R2U), despite considerable flexibility (Figures 4-4a and 4-5d), on the simulation time scale did not assume a conformation superimposable to the closed conformation adopted by its unbound R-spondin 1 counterpart. It did show some degree of similarity with the closed conformation (Figure 4-4a). Nevertheless, from these observations, we can conclude that R-spondins possess an intrinsic flexibility, due to the rotation of the two furin domains around the Lys66 - Ile69 hinge.

**Druggability.** Given their importance in the regulation of the Wnt pathway, the conformations sampled by both R-spondin 1 and R-spondin 2 could present opportunities for structure-based ligand design. Indeed, locking out the closed unbound conformation of R-spondin 1 or competing with the normal interactions of R-spondin 1/2 with RNF43/ZNRF3 could prevent the formation of the ternary complex leading to a down-regulation of the Wnt pathway. To explore R-spondin inhibition, we focus on the R-spondin region between the second β-hairpin of the head module and the second β-hairpin of the rod module, which in all three structures are the secondary structures that frame a concave region although most clearly pocket-like for the closed conformation (Figure 4-1c). To evaluate the druggability of this region as a function of conformation (closed versus extended) and sequence (R-spondin 1 versus R-spondin 2), we assess pocket volume, topology and druggability score.

For reference, we first consider the pocket identified in this target region from the R-spondin 1 and 2 crystal structures. Using PocketAnalyzer\textsuperscript{PCA}\textsuperscript{[43]}, the pocket volumes are estimated as 586, 175 and 470 Å\textsuperscript{3} for 4BSO, 4KNG and 4C8V structures respectively: however, despite the difference in volume, a shared portion of the pocket
Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis

enclosed between the second and third β hairpins of the first furin domain is evident for all three structures (top of Figure 4-7c). This pocket region is also identified in most of the MD simulations (99% for R1U, 98% for R1C, and 91% for R2U, Figure 4-9 in the Supporting Information). However, the volume distributions calculated for each simulation appear distinct (Figure 4-7a): the majority of the pockets identified for R1U configurations possess a volume between 600 - 800 Å³, while a smaller pocket between 100 - 300 Å³ in volume is found for most R1C and R2U structures.

Figure 4-7: a) Volume distribution of the defined binding pocket calculated for a subset of conformations for R1U (green), R1C (red) and R2U (green) sequences. b) PLB values distribution for the defined binding pocket calculated for a subset of conformations for R1U (green), R1C (red) and R2U (blue) sequences. c) Average binding pockets identified by PocketAnalyzer²PCA from X-ray crystal structures (top) and from selected conformations (bottom) of R1U (green), R1C (red), and R2U (blue). Pockets calculated are colour coded according to the weighted average (high frequency - blue; low frequency - red).
Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis

The absence of a fully closed structure for $R_{1C}$ and $R_{2U}$ leads to smaller pocket volumes (Figure 4-7a,c). Nonetheless, the average pockets identified for the three trajectories show a consensus volume located at the interface between the furin domains (blue spheres, bottom of Figure 4-7c).

<table>
<thead>
<tr>
<th>Conf</th>
<th>Residues in pocket</th>
<th>PLB</th>
<th>state</th>
<th>t (ns)</th>
<th>V (Å³)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$4BSO$</td>
<td>I32,L33,L34,E35,R36,N37,D38,Q41,Y53,</td>
<td>1.8</td>
<td>closed</td>
<td>-</td>
<td>754</td>
</tr>
<tr>
<td></td>
<td>K63,C64,I65,K66,K68,F76,S77,H78,N79</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R_{1U}$</td>
<td>L33,L34,E35,R36,Q41,Y53,K63,C64,I65,</td>
<td>2.9</td>
<td>closed</td>
<td>657</td>
<td>1345</td>
</tr>
<tr>
<td></td>
<td>K66,K68,F76,S77,H78,N79</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$4KNG$</td>
<td>I32,L33,L34,E35,R36,Q41,Y53,K63,C64,</td>
<td>-0.6</td>
<td>open</td>
<td>-</td>
<td>346</td>
</tr>
<tr>
<td></td>
<td>I65,K66,K68</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R_{1C}$</td>
<td>L33,L34,E35,R36,Q41,Y53,K63,C64,I65,</td>
<td>2.9</td>
<td>open</td>
<td>749</td>
<td>638</td>
</tr>
<tr>
<td></td>
<td>K66,K69,F76,S77,H78,N79</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$4C8V$</td>
<td>F61,Y62,L63,R64,R65,Q70,Y82,Y83,R92,</td>
<td>-1.1</td>
<td>open</td>
<td>-</td>
<td>464</td>
</tr>
<tr>
<td></td>
<td>C93,S94,R95,R97,R107</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R_{2U}$</td>
<td>Y62,L63,R64,R65,E66,Q70,Y82,R92,C93</td>
<td>2.4</td>
<td>open</td>
<td>361</td>
<td>1104</td>
</tr>
<tr>
<td></td>
<td>S94,R95,R97,R107</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4-1: Conformation (conf), residues present in the binding pocket, conformational state (state), time of the simulation (t (ns)), and volume (V (Å³)) of the three binding pockets with highest propensity for ligand binding (PLB), as calculated by Site Finder.

We then quantify the druggability of these pockets by computing their propensity for ligand binding (PLB) index via Site Finder from MOE. This score is based on the specific amino acid composition of a pocket and has proven effective in predicting drug binding propensities for known protein structures. We first consider the PLB indices for $4BSO$, $4KNG$ and $4C8V$ X-ray structures, which have values of 1.8, -0.6, and -1.1 respectively. $4BSO$ has a positive index (1.8), indicating a readily druggable pocket. Negative scores correspond to non-druggable pockets, as observed for $4KNG$ and $4C8V$. These PLB values therefore suggest that the closed conformation forms a more druggable pocket than the open structure.
Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis

Most of the conformations for each of the three trajectories are associated with a wide range of positive PLB scores (Figure 4-7b). Interestingly, R1C and R2U simulations, initiated from the negatively scoring 4KNG and 4C8V X-ray structures respectively, are able to assume positive PLB values during MD. The majority of the conformations of R1U possess PLB scores in the range of 1 - 2 (Figure 4-7b), while values below 1 are found for the majority of R1C and R2U conformations (Figure 4-7b). This again suggests the higher druggability of the closed conformation over an extended one and is most likely due to participation of additional amino acid side chains from the second hairpin rod module in the more enclosed pocket.

**Docking.** All three R-spondin MD simulations sampled pockets with positive PLB values, although to differing extents. Therefore, we further analyse the most druggable pocket conformations from each simulation (according to PLB index, Table 4-1), using fragment docking to derive insights into potential sites of interaction, within R-spondin conformations. From each of the R1U, R1C and R2U simulations, we select the pocket conformation with the highest PLB score. Then, using GOLD, we dock a set of 15354 molecular fragments, generated from an initial ZINC database of 10342 different ligands, into these pockets (Table 4-3 in the Supporting Information).

Gratifyingly, we find that the docked fragment poses map onto the binding areas predicted by Site Finder (Figure 4-10 in the Supporting Material). The top ten ranked fragments identified for each of the three protein pockets fit well into the binding pockets identified for example (Figure 4-8a,b,c), making interactions bridging the Fu1 and Fu2 domains.
Figure 4-8: (Top) Molecular surfaces calculated for the highest PLB scoring pocket from R1U (a), R1C (b) and R2U (c) and their top scoring molecular fragment. Molecular surfaces of conformations are coloured red for H-bonding regions, blue for other polar regions and white for hydrophobic regions. Molecular surfaces of molecular fragments are coloured green. (Centre) Top scoring molecular fragment bound to the highest PLB scoring pocket from R1U (d), R1C (e) and R2U (f). (Bottom) Consensus pharmacophores identified from the top 10 scoring docked poses from R1U (g), R1C (h) and R2U (i).
Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis

The most frequently formed ligand interactions to R-spondin 1 include hydrogen bonds to Leu34, Arg36, Gln41, Lys63, Lys66 and His78; and for R-spondin 2, fragment hydrogen bonds to Leu63, Arg65, Gln70, Arg92, Arg95 and Arg107. Notably, among these residues, Leu34, Arg36 and Gln41 are the key R-spondin residues that interact with RNF43/ZNRF3 in the ternary complex. For docking into all three R-spondin pockets, the central region of the binding sites are typically occupied by hydrophobic ligand moieties (Figure 4-8d,e,f), due in part to the presence of hydrophobic residues in this region for R-spondin 1 (such as Leu35 and Ile65) and R-spondin 2 (such as Leu28, Tyr62 and Tyr82).

In order to resolve the most important binding features of the three binding sites, we derive consensus pharmacophores based on the top 10 docked poses for each R-spondin pocket in turn. Interestingly, for the three pockets, the pharmacophores possess shared features: the pharmacophore points identified for R1U include three hydrogen bond acceptors, one hydrogen bond donor and three aromatic/hydrophobic features (Figure 4-8g). The R1C pharmacophore possesses one additional hydrogen bond acceptor feature, assuming, however, a different geometry from the R1U pharmacophore (Figure 4-8h). The R2U pharmacophore displays multiple aromatic/hydrophobic features, accompanied by hydrogen bond donor/acceptors points (Figure 4-8i). It is interesting to observe how, for R1U, pharmacophoric features are concentrated in the central region, while for R1C and R2U, these features are instead oriented towards the edge of the pocket. Moreover, the identified pharmacophores illustrate the different preferences of interaction for the top ten fragments with respect to the two pockets: while the fragments binding to R1C possess interactions with residues of the right hand side of the pocket, ligands docked to R2U interact with
residues on the left, although they both share a hydrophobic feature in the central portion of the pocket.

In order to verify the reliability of the structure-based pharmacophores identified, a simple screen was performed in order to test if the queries are able to discern between high and low scoring fragments from docking. To this end, the pharmacophore query was applied to search the databases of fragments, composed of the top 100 and the last 10000 scoring fragments for each of the three pockets. Encouragingly, the R1U pharmacophore matched 24% of the top 100 scoring fragments from docking (Table 4-2). Lower values of 5% and 7% are found for R1C and R2U respectively (Table 4-2).

<table>
<thead>
<tr>
<th>Conformation</th>
<th>Top 100</th>
<th>Last 10000</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1U</td>
<td>24%</td>
<td>1.82%</td>
</tr>
<tr>
<td>R1C</td>
<td>5%</td>
<td>0.07%</td>
</tr>
<tr>
<td>R2U</td>
<td>7%</td>
<td>0.48%</td>
</tr>
</tbody>
</table>

Table 4-2: Percentages of pharmacophore matches within the top 100 and latest 10000 docking poses for each conformation.

As mentioned above, for R1U, the pocket is additionally defined by the close interaction with the second furin domain, which leads to a more enclosed pocket to guide pharmacophore generation and ligand docking (Figure 4-8a). On the other hand, the wider pockets observed for R1C and R2U (Figure 4-8b,c) allow occupation of a broader range of binding poses, reflected in the lower percentages of hits found for these pharmacophores. However, much lower percentages for the screening of each pharmacophore against the low-scored poses are found (1.8%), suggesting the three pharmacophores are discriminating.
4.6 Conclusions

In this study, we have examined the plasticity of anti-cancer targets R-spondins 1 and 2 using microsecond molecular dynamics simulations in explicit aqueous solvent. We find that the closed conformation of R-spondin 1 identified from the 4BSO X-ray structure is preserved over the duration of its trajectory, and therefore does not appear to rely on crystallographic contacts for stability. For simulations of R-spondin 1 and 2 from extended, open X-ray structures, two observations can be made: these R1C and R2U simulations reveal a higher level of flexibility than for the closed conformation, exploring several transient structures, some characterized by interdomain salt bridges. Interestingly, a proportion of the sampled structures appear to be partially closed conformations, although this is observed to a greater extent for R-spondin 1 than 2. Indeed, the considerable fluctuations and progressive decrease in the angle of rotation between the two furin domains for the initially open structures suggest a possible conformational relaxation towards the closed state.

It is tempting to consider that the closed state observed for the 4BSO crystal structure might represent the resting conformation of unbound R-spondin 1 in solution, which then adapts to form the ternary complex via induced fit, as observed for the 4KNG X-ray structure. No closed form of R-spondin 2 was found either in its unbound crystallographic structures or in our simulations. Nonetheless, R-spondin 2 possess an intrinsic flexibility around the hinge between the two furin domains, showing some degree of similarity with the closed-like conformations observed from the simulations of R-spondin 1.
Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis

Based on X-ray and simulated structures, we then evaluated the druggability of a potential inhibitor binding region identified at the interface between the Fu1 and Fu2 domains of the R-spondins. Although this pocket showed considerable variability in predicted volume and druggability index over the MD trajectories and between the two R-spondin sequences, at least ~40% of the structures derived from the three simulations were characterized as druggable. The predicted fragment docking hits to this specific interdomain region of R-spondins demonstrate how this might indeed be a favourable region to target. This docking analysis demonstrates how a molecular fragment could fit in the pocket located between the second β-hairpin of the head module and the second β-hairpin of the rod module. The design of a suitable ligand could lead to conformational locking of R-spondin in a conformation unproductive for binding to RNF43/ZNRF3, such as the closed conformation observed for R-spondin 1, or simply compete with its native interaction with RNF43/ZNRF3.

The computational analysis of R-spondins in this study, using a combination of molecular dynamics, druggability evaluation and molecular docking analysis provide possible new directions for structure-based design of potential inhibitors targeting the Wnt pathway.


4.7 References


Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis

Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis


45. Molecular Operating Environment (MOE), 2013.08; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2014.


Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis


Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis

4.8 Supporting Information

Figure 4-9: Time series of the volumes of the defined binding pocket calculated for selected conformations of R1v (green), R1c (red) and R2v (blue).
Figure 4-10: Representative high scoring molecular fragment bound to the highest PLB scoring pocket from R1u (a), R1c (b) and R2u (c). Binding sites as predicted by Site Finder are also shown (in grey hydrophobic regions, in red hydrophilic regions). Specifically, fragments ranked #1, #1 and #2 for R1u, R1c and R2u respectively.
Molecular dynamics of R-spondin 1 and 2: solution conformations and druggability analysis

<table>
<thead>
<tr>
<th>ranked pose</th>
<th>ligand name</th>
<th>Gold PLP score</th>
<th>ligand name</th>
<th>Gold PLP score</th>
<th>ligand name</th>
<th>Gold PLP score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R_{1U}$</td>
<td></td>
<td>$R_{1C}$</td>
<td></td>
<td>$R_{2U}$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>ZINC95512841</td>
<td>59.3</td>
<td>ZINC71791704</td>
<td>55.2</td>
<td>ZINC82329356</td>
<td>55.6</td>
</tr>
<tr>
<td>2</td>
<td>ZINC19912137</td>
<td>58.8</td>
<td>ZINC39651147</td>
<td>54.9</td>
<td>ZINC12362733</td>
<td>54.6</td>
</tr>
<tr>
<td>3</td>
<td>ZINC31394118</td>
<td>58.7</td>
<td>ZINC82331483</td>
<td>53.6</td>
<td>ZINC05484292</td>
<td>54.1</td>
</tr>
<tr>
<td>4</td>
<td>ZINC71779426</td>
<td>57.7</td>
<td>ZINC05117632</td>
<td>53.2</td>
<td>ZINC39111167</td>
<td>54.0</td>
</tr>
<tr>
<td>5</td>
<td>ZINC41257662</td>
<td>55.3</td>
<td>ZINC95465100</td>
<td>52.8</td>
<td>ZINC67642871</td>
<td>53.6</td>
</tr>
<tr>
<td>6</td>
<td>ZINC16953306</td>
<td>54.8</td>
<td>ZINC82331483</td>
<td>52.8</td>
<td>ZINC39102910</td>
<td>53.5</td>
</tr>
<tr>
<td>7</td>
<td>ZINC67756406</td>
<td>54.6</td>
<td>ZINC05117632</td>
<td>52.8</td>
<td>ZINC22115224</td>
<td>53.4</td>
</tr>
<tr>
<td>8</td>
<td>ZINC95526757</td>
<td>54.5</td>
<td>ZINC75624901</td>
<td>52.6</td>
<td>ZINC39110914</td>
<td>53.1</td>
</tr>
<tr>
<td>9</td>
<td>ZINC01393937</td>
<td>54.5</td>
<td>ZINC39651147</td>
<td>52.6</td>
<td>ZINC71783071</td>
<td>52.6</td>
</tr>
<tr>
<td>10</td>
<td>ZINC96332108</td>
<td>54.2</td>
<td>ZINC36467611</td>
<td>52.6</td>
<td>ZINC71781886</td>
<td>52.5</td>
</tr>
</tbody>
</table>

Table 4-3: Top 10 docking scores registered for the ligands docked onto the highest PLB scoring conformations coming from $R_{1U}$, $R_{1C}$, and $R_{2U}$. 
5. Conclusions and outlook

In this study, advanced sampling techniques and long time scale molecular dynamics simulations have been used for evaluating the flexibility of biological structures. The studies in this thesis provided several examples of the applications of different computational techniques in order to assess properties of biomolecular systems.

In Chapter 2, REMD has been used to evaluate the effects of reversal of sequence and stereochemistry on the secondary structure of a p53-derived peptide. In this study, we find that, in order to obtain converged ensembles for the sequences considered, especially for the ones composed by D-amino acids, enhanced sampling via the replica exchange molecular dynamics method is required. From the simulations, we found that D-peptide sequences assume a left-handed helical conformation, regardless of their initial structure. Interestingly, we observed how reversed sequence are associated with higher helical content compared to parent peptides. The increased helicity of the reversed sequences appears to be related to the presence of characteristic hydrogen bonds, which are not registered in the parent sequences. These results showed general agreement with NMR and CD studies, suggesting how the combination of replica exchange molecular dynamics with a ff99SB force field and Generalized Born solvent model, has proven to be a suitable protocol for modelling D-peptide conformations.

From the simulations, we also evaluate the potential of sequence and chirality modifications in mimicking the parent sequence. We observed that the inter-Cα and inter-Cβ distances of the three key residues are similar between the parent sequence and its derived sequences. However, a comparison of the molecular orientation revealed how the superposition between the backbone of the inversed and its retro-inverso sequence is not possible. A similar issue is also observed for the side-chains of the reversed and the retro-inverso sequence.
Thus, a retro-inverso peptide is disadvantaged as a mimic in both aspects. Nonetheless, several examples of D-peptide sequences are able to bind to MDM2 in a left-handed helical conformation, in a pose shifted from p53, suggesting some degree of flexibility in the motifs recognized by MDM2. This could suggest future structure-based drug design strategies employing a retro-inverso peptide with the inclusion of structural restraints, in order to retain a right-handed helical conformation.

In Chapter 3, swarm-enhanced sampling molecular dynamics (sesMD) was used to explore the conformational space of p38 MAP kinases. SesMD is a method designed to enhance conformational sampling of higher lying energy wells, whilst preserving the basic shape of the potential energy surface of a given molecular systems.

In this approach, a swarm of replica simulations interact cooperatively via a set of pairwise potentials incorporating attractive and repulsive components. The application of sesMD to the DFG motif of p38α MAP kinase provided a broader sampling of DFG conformational states compared to unbiased MD simulations. Indeed, sesMD simulations started from a DFG-out conformation sampled a wide range of conformations that includes DFG-in structures.

Moreover, from sesMD simulations we identified intermediates that can be compared to p38α MAP kinase conformations identified from experimental and computational studies. Nonetheless, sesMD replicas initiated from a DFG-in crystal structure, sample a lesser range of DFG loop conformations, compared to the simulation based on a DFG-out conformation. sesMD was also able to generate a wide range of interesting pocket shapes which could represent new potential starting points for a structure-based kinase inhibitor design. The sesMD approach demonstrated to be a straightforward, parallel algorithm for increasing the sampling of the conformational space of a given molecular system.
Conclusions and outlook

Whilst the sesMD method currently is based on distance in dihedral space, in the future, one could envisage swarm-enhanced sampling applied to other degrees of freedom, for example in guiding the docking to proteins of large flexible molecules, such as peptides.

In Chapter 4, unbiased microsecond MD simulations have been used to evaluate the flexibility of R-spondins. In this study, we find that the characteristic closed conformation of unbound R-spondin 1 is preserved over the course of the simulations. Conversely, R-spondin 1 and 2 simulations initiated from extended, open structures, reveal a higher level of flexibility than the closed conformation, sampling several intermediate conformations: among them, closed-like conformations are also found, although to a greater extent for R-spondin 1 than R-spondin 2.

The druggability of a potential inhibitor binding region identified at the interface between the two furin domains of R-spondins has also been evaluated: we found that for this region, at least ~40% of the structures sampled during MD were characterized as druggable, in spite of considerable variability in predicted volume, conformational structure and sequence. Molecular docking analysis predicted that molecular fragments could fit favourably into pockets identified from both open and closed conformation and from R-spondin 1 or 2. The pharmacophores from ligands docked into these pockets showed shared as well as unique pharmacophoric features.

In this way, the computational approach adopted in this study evaluated protein plasticity of R-spondins, providing new possible directions for future structure-based drug design of active ligands capable of alternative targeting of the Wnt pathway.
Conclusions and outlook

In the chapters outlined in this thesis, we demonstrated how molecular dynamics simulations can be applied to the study of diverse biological systems in order to shed light on their plasticity, provide detailed conformational insights, and most importantly, reveal concealed binding sites. Importantly, we showed how the application of advanced sampling techniques, such as REMD and sesMD, as well as long-time scale simulations, were able to overcome the core limitation of classical MD simulations, increasing the amount of conformational space sampled and identifying other stable structures. Due to the constant improvement in computational power available and the design of new advanced algorithms, MD simulations are on a course to assume a pivotal role in the future of computer-aided drug design.