Novel Applications for Hierarchical Natural Move Monte Carlo Simulations:

From Proteins to Nucleic Acids

Samuel Demharter

St Edmund Hall

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2016

To my family.

Acknowledgements

I would like to thank my co-supervisors Prof Charlotte M Deane and Prof Bernhard Knapp for their invaluable contributions to my DPhil. As part of the Oxford Protein Informatics Group I had the privilege of interacting with many friendly and talented people. I especially enjoyed the group meetings, not only because of the cake. I would also like to thank Dr Konrad Krawczyk for a very productive collaboration. Last but not least I would like to thank my main supervisor Dr Peter Minary who has been an outstanding supervisor and an inspiring mentor.

Abstract

Biological molecules often undergo large structural changes to perform their function. Computational methods can provide a fine-grained description at the atomistic scale. Without sufficient approximations to accelerate the simulations, however, the time-scale on which functional motions often occur is out of reach for many traditional methods. Natural Move Monte Carlo belongs to a class of methods that were introduced to bridge this gap. I present three novel applications for Natural Move Monte Carlo, two on proteins and one on DNA epigenetics. In the second part of this thesis I introduce a new protocol for the testing of hypotheses regarding the functional motions of biological systems, named customised Natural Move Monte Carlo. Two different case studies are presented aimed at demonstrating the feasibility of customised Natural Move Monte Carlo.

Contents

List of Figures

Acronyms

3pt 3-point.
5caC 5-carboxylcytosine.
5fC 5-formylcytosine.
5hmC 5-hydroxymethylcytosine.
5mC 5-methylcytosine.

A Angstrom.
ANM Anisotropic Network Model.
APC Antigen Presenting Cells.

b2m β2-microglobulin.

CD4/CD8 Cluster of differentiation 4/8.
CG Coarse-Grained.
CLIP Class II-associated invariant chain peptide.
cNMMC customised Natural Move Monte Carlo.
CpG Cytosine and guanine connected by a phosphodiester bond.

DDD Dickerson-Drew dodecamer.
DFC Deterministic Full Closure.
DOF Degrees of freedom.

ED-CG Essential Dynamics Coarse Graining.
ENM Elastic Network Model.
EpoR Erythropoietin receptor.
ER Endoplasmic Reticulum.

GNM Gaussian Network Model.

HA Hemaglutinin peptide.
HLA-DM Human Leukocyte Antigen DM.
hNMMC hierarchical Natural Move Monte Carlo.

Ii Invariant Chain.

K Kelvin.

lncRNA long non-coding RNA.

m6A N6-methyladenosine-dependent.
MC Monte Carlo.
MCMC Markov Chain Monte Carlo.
MD Molecular Dynamics.

MHC Major Histocompatibility Complex.
MHCI MHC Class I.
MHCII MHC Class II.
MIIC MHCII Compartment.
miRNA microRNA.
MmCpn Methanococcus maripaludis chaperonin.
MZ Molten Zone.

ncRNA non-coding RNA.
NMA Normal Mode Analysis.
NMMC Natural Move Monte Carlo.

PCA Principal Component Analysis.
PDB Protein Data Bank.
pMHCII peptide-MHCII complex.
PT Parallel Tempering.

RMSD Root mean square deviation.
RSC Recursive Stochastic Chain Closure.
RTB Rotation-Translation Blocks.

SCOP Structural Classification of Proteins.
SPC Stochastic Partial Closure.
STSAMC: Simulated tempering, simulated annealing Monte Carlo.

Tet - Ten eleven translocation protein.

VH/VL Variable domain, heavy/light chain.

Chapter 1

Introduction

1.1 Motivation

The number of biomolecular structures that have been solved at an atomistic scale has been growing exponentially for the last two decades. This year the total entry count of biomolecular structures in the protein database has surpassed the 100,000 mark [1]. While such a vast amount of structural information provides a great insight into the general architecture of individual biomolecules, it is often limited in the dynamic information it can reveal. Due to the importance that functional motions play in the workings of a given biomolecule it is most crucial to shed light on the conformational heterogeneity that is required for its action. A range of experimental techniques are capable of capturing dynamics. Due to the limitations in the spatio-temporal scale and their tendency to generate averaged distributions of dynamical properties, however, information on intramolecular motions is still scarce.

By using the structures of the Protein Data Bank (PDB) as frameworks and exposing them to algorithms following simple biophysical and chemical rules, the field of computational structural biology has attempted to address this shortcoming for several decades.

The quantum mechanical characteristics of molecules at a subatomic level is described by the Schrödinger equation. However, applying this equation in practice is computationally impractical for macromolecules. A common approach to simulating the motions of biomolecules is molecular dynamics (MD) simulation, where the positions and velocities

of atoms change based on the laws of classical physics. The forces for these particles are calculated with the help of a force field, which is typically constructed from a combination of first-principles, parameter fitting, quantum mechanical computations and experimental tests. Despite MD simulation not modelling the exact physics, it provides a good approximation for many biochemical processes.

As the simulation proceeds through time, the forces on each atom are calculated after which Newton's laws of motion provide the new positions and velocities of all atoms. Most force fields (a force field refers to the functional form and parameter sets used to calculate the potential energy of a system of particles) describe the total force on an atom as the sum of three parts: 1. bonded forces, which include interactions between several atoms connected by covalent bonds; 2. van der Waals forces, interactions between pairs of atoms that are in close proximity; and 3. electrostatic forces, which affect interactions between all pairs of atoms and decrease slowly with distance.

There are two main factors that make MD simulations computationally expensive: 1. The calculation of forces at each iteration requires around one billion operations for a system with a size of one hundred thousand atoms; and 2. the calculation of forces needs to happen in small time-steps. Usually, time steps are do not exceed the femtosecond scale, which means that it would take almost one trillion steps to simulate one millisecond. In order to cope with this computational cost, researchers perform simulations on a large number of processors in parallel. This works to some extent, however, beyond a certain threshold there is no additional benefit to adding more processors, as the communication between cores becomes a bottle neck. Also, force calculations have to be performed in sequence as each step is dependent on the force output of the previous one. Despite these significant challenges, performance has improved faster than Moore's Law would predict. In a span of just five years, the performance of state-of-the-art simulations has increased by more than three orders of magnitude. In 2007, the longest atomistic MD simulation of a protein was 2 microseconds. In 2009, one millisecond was simulated. This jump in performance is due to a variety of factors including hardware, software, and algorithm innovations [2].

Monte Carlo (MC) simulations make use of a probability distribution inherent to the biomolecular system (usually a Boltzmann distribution) to generate new samples. Due to the complex energy landscapes that result from the all-atom representation and explicit treatment of solvent, gradient-based methods such as MD simulations were considered better suited for simulating dense molecules such as proteins. However, as MC methods and implicit solvent models have improved [3, 4], smoother energy landscapes have become available. This makes MC simulations an increasingly viable option for studying the structural plasticity of macromolecules, especially when analysing large structural changes that occur over long-time scales.

An important factor to consider when simulating biomolecular structures is the model resolution. The first molecular dynamics simulation of a protein at all-atom resolution was performed in 1977 [5]. Although at that time no significant discoveries were made due to the very limited scope of the simulation, the field has progressed along these lines towards millisecond trajectories enabled by optimised algorithms, improved models and considerable advances in computer performance [6, 7].

Milestones include the 50 nanosecond MD simulation of the full satellite tobacco mosaic virus with one million particles [8], the Folding@Home project that used over 400,000 personal computers to generate a range of interesting results [9] and a study that presented millisecond simulations to study the folding pathways of small fast folding proteins [7].

Despite these advances, the high dimensionality and complex energy surfaces when simulating functional motions in large biomolecules still pose a challenge [10, 11]. In an effort to address the gap between biologically relevant time and size scales and the capabilities of molecular simulations there have been promising developments in dimensionality reducing methods that exploit the redundancy resulting from the modularity and collective motions in biomolecules [12, 13]. For example essential dynamics coarse-graining (ED-CG) identifies CG sites that reflect the essential dynamics of an atomistic molecular dynamics trajectory [14]. Other methods based on elastic network models [15], principal component analysis [16] and normal mode analysis [17] have also been successfully used to

study functional motions in biomolecules. While these methods are not as physically accurate as classical molecular simulations, their increased sampling efficiency makes them a valuable tool to generate new hypotheses that can be tested by experiments. One of the main challenges of these methods, however, is finding a set of degrees of freedom that describes the system accurately enough to make biologically relevant claims [13]. Furthermore, NMA only produces physically relevant conformations close to the starting structure and PCA is dependent on a molecular simulation trajectory. NMA and PCA are also computationally expensive for larger systems (they scale exponentially), while MC and MD simulations scale linearly. Thus, it is of value to have a computationally cheap method that allows for the easy manipulation of degrees of freedom and testing of several different hypotheses about the functional motions of biomolecules *in silico*.

1.2 Aims

Functional motions in macromolecules often involve large-scale, slow and concerted structural changes. Experimental methods can give insight on the conformational endpoints as well as the kinetics of these movements. However, simulations are often required to elucidate more detailed transition states and to investigate the mechanism by which functional motions occur. In order to make the simulation of large macromolecular structure and their functional motions computationally feasible, a number of approximations are routinely employed. These include the reduction of particles (e.g. atoms) in the system (coarse-graining), the use of alternative degrees of freedom (torsional vs Cartesian coordinates) and the use of efficient minimisation and sampling algorithms. Natural Move Monte Carlo (NMMC) adds another layer of approximation by grouping atoms within a structure that are expected to move collectively (e.g. secondary structures, or the aromatic ring within a nucleotide) and moving them as a set of rigid bodies, thereby significantly reducing the degrees of freedom. A number of NMMC studies had already been published, however the scope of the method was still relatively unknown. The purpose of this thesis was to explore novel biological applications suitable for NMMC and to provide a robust protocol based on customised Natural Moves for the testing of hypothesis regarding the

functional motions of biological systems.

1.3 Outline

In **chapter 1** I give a brief introduction to the field of protein structure research, outline the objectives of the thesis and present a list of recent publications associated with this work.

In **chapter 2** I discuss the importance of computational methods as well as experimental techniques for the investigation of functional aspects of protein motion and function. I provide background information on the main biological systems used in this thesis and introduce the fundamentals of protein and DNA structure. I also provide a short literature review on techniques and discoveries regarding collective motions in biological molecules and discuss various aspects of molecular simulations including different types of models, degrees of freedom and algorithms.

In **chapter 3** I present detailed explanations for the various methods used in this thesis, including the stochastic chain closure algorithm central to NMMC and the physics- and knowledge-based potentials used for nucleic acids and proteins, respectively. I then introduce a novel theoretical framework that I developed, named customised Natural Move Monte Carlo (cNMMC), that makes use of the inherent customisation capabilities of NMMC. I present a formal description for the generation of different test cases aimed at representing a specific research question or hypothesis with respect to the functional motions of a biological system. I conducted case studies on two different biological systems, which are presented in chapters 7 and chapter 8. Furthermore, I outline NMMC as well as the NMA method used in one of the chapters (chapter 7) and explain the working details of MODELLER (used in chapter 5) as well as x3DNA, a popular structural analysis tool that I used to characterise conformational changes in epigenetically modified DNA molecules (chapter 8).

In **chapter 4** I present the first NMMC research application in this thesis. Here, we applied NMMC to study the detachment of peptides from a MHC class I complex. The aim of this study was to prove that NMMC can be used as a training-free method to

accurately classify MHC class I binding and non-binding peptides and provide insight into the dissociation process. The computational efficiency of NMMC allowed us to explore the MHCI/peptide interaction for over 30 peptides.

In **chapter 5** we investigated the hinge flexibility in diabodies developed to bind and activate the Erythropoietin receptor (EpoR). *In vitro*, the diabodies elicited a range of different signalling amplitudes, from full to minimal agonism. Due to concerns that flexibility in the hinge connecting the two VH/VL domains in the diabody would have an effect on this mechanism, we simulated two selected structures with NMMC to assess their dynamic range.

In **chapter 6** a study regarding the effects of different epigenetic marks (5mC, 5hmC, 5fC and 5caC) on DNA dodecamers is presented. Here, we used NMMC as an efficient computational framework to simulate a range of epigenetic makeups for different DNA structures.

In **chapter 7** I present the case study of the first biological system, the MHC class II complex. First, I investigate the structural plasticity of the empty MHCII complex then the mechanism by which the peptide-loading chaperone HLA-DM stabilises the open form of the MHCII binding groove.

In **chapter 8** I demonstrate the use of the cNMMC protocol with a nucleic acid case study. We simulated the Dickerson-Drew Dodecamer in the presence and absence of an epigenetic mark (5hmC). We use customised Natural Moves to enforce hypothesised structural effects during simulation in order to be able to detect structural changes caused by a single epigenetic mark.

The thesis ends with a conclusion in **chapter 9**.

1.4 Publications

- Knapp, B., Demharter, S., Deane, C. M., & Minary, P. (2016). Exploring peptide/MHC detachment processes using hierarchical natural move Monte Carlo. *Bioinformatics*, 32(2), 181-6.
- Moraga, I., Wernig, G., Wilmes, S., Gryshkova, V., Richter, C. P., Hong, W.-J., Sinha, R., Guo, F., Fabionar, H., Wehrman, T. S., Krutzik, P., Demharter, S., Plo, I., Weissman, I. L. Minary, P., Majeti, R., Constantinescu, S. N., Piehler, J., Garcia, K. C. (2015). Tuning Cytokine Receptor Signaling by Re-orienting Dimer Geometry with Surrogate Ligands. *Cell*, 160(6), 1196-1208.
- Krawczyk, K., Demharter, S., Knapp, B., Deane, Charlotte M., & Minary, P. (2017). Structural effects of epigenetic marks on DNA structure in silico. *Bioinformatics (accepted pending minor revisions).*
- Demharter, S., Knapp, B., Deane, C. M., & Minary, P. (2016). Modeling Functional Motions of Biological Systems by Customized Natural Moves. *Biophysical Journal*, 111(4), 710-721.
- Demharter, S., Knapp, B., Deane, C. M., & Minary, P. (2017). Stabilisation of the empty MHCII binding groove by HLA-DM: A customised Natural Move Monte Carlo study. *Scientific Reports (accepted pending minor revisions).*

Chapter 2

Background

2.1 Computational background

2.1.1 Natural Move Monte Carlo

Natural Move Monte Carlo (NMMC) was developed mainly to address the dimensionality problem [18, 19] encountered by molecular simulations. NMMC is a conformational sampling method that exploits the modular nature of biomolecules to accelerate the exploration of a structural landscape. Instead of sampling the position of each atom in the system, groups of atoms or residues that are part of a shared structural feature can be grouped into segments and moved collectively. This gives rise to a conformational sampling strategy that considers the system as a collection of segments and exclusively samples their arrangements along the user-defined degrees of freedom. Thus, this method reduces dimensionality by several orders of magnitude by sampling along generalised coordinates. While it is not aimed at revealing any kinetic information it can rapidly generate ensembles of thermodynamically feasible structures that appear according to canonical probabilities using computational resources that are readily accessible. In a recent study we showed that NMMC yields comparable results to and is three orders of magnitude faster than conventional Molecular Dynamics when simulating peptide detachment from MHCI molecules [20].

Traditionally NMMC is used to explore the conformational landscape along a partic-

ular set of degrees of freedom chosen by the researcher. Several studies have followed this approach [18, 19, 21–24]. However, the initial choice of degrees of freedom might not always be optimal. Additionally, if the objective is to investigate the causality of functional motions, it may be informative to perform NMMC simulations for a variety of sets of degrees of freedom. This aspect will be addressed with the introduction of customised Natural Move Monte Carlo in the last chapters of this thesis.

2.1.2 The history of Natural Move Monte Carlo

Natural Move Monte Carlo has emerged from a number of milestone developments over recent years. A robust conformational sampling approach was published in 2008 [21], when loop torsion angle degrees of freedom, which preserve rigid secondary structures, a 3-point per residue coarse-grained simplified model, and advanced multi-canonical sampling were combined to study the stability of a selection of protein fold classes. By clustering the trajectories and selecting the minimum energy from each cluster, this approach led to the successful reconstitution of the native folds. This proof of concept study was also key to the validation of the simplified model.

A further challenge that needed solving was the large computational cost of sealing breaks in the chain. To avoid large displacements due to the amplification of small structural changes in distal regions of the structure (lever arm effect) methods such as CONROT [25] were introduced that made local torsional moves to small segments while the remainder of the structure was kept in position. To resolve the chain breaks that may occur a number of algorithms have been used [26–28], however the computational cost for each was high, which limited the number of chain breaks that could be solved.

In 2010 this problem was addressed with the recursive stochastic chain closure algorithm (RSC) [18]. This algorithm extended an existing closure method, called deterministic full closure (DFC), by adding stochastic partial closure steps (SPC). Given that DFC is based on a single bond-angle degree of freedom its application is limited to small chain breaks, thus only enabling small moves [29]. DFC on its own used to struggle with large breaks, as the algorithm would fail if the two disconnected atoms were further apart than

the allowed bond distance, thereby limiting the method to small moves. By introducing SPC steps before the break is sealed by DFC, the atoms leading up to the break are rearranged to reduce the size of the break, therefore allowing for larger moves. In addition, the cost of the RSC algorithm scales linearly with the number of degrees of freedom while methods such as CONROT have non-polynomial complexity. Thus, the method that is now called Natural Move Monte Carlo was born.

A further iteration of this approach was introduced in 2012 when Sim et al. showed how nested moves, or hierarchical Natural Moves (hNM) could be used to accelerate sampling of large biological structures even further [19]. This method was successfully benchmarked on a symmetrical RNA four-way junction, where the hNMMC simulation (distribution of observables) converged much earlier than competing methods including NMMC. Thus, by using several layers of degrees of freedom that each describe the movement of separate structural elements such as residues, base pairs, base stacks and whole helices, the energy-surface is significantly smoothed compared to single-level rigid body simulations, leading to large increase in sampling efficiency. Similarly for protein systems, these could include domains, secondary structure elements or residues. To quote the paper: *âĂIJLess computational time is spent with exploring conformations in phase space with negligible probability of occurrence. Exploring a smaller but more important region of the conformational space solves the problem of high dimensionalityâĂİ.*

The method was also adapted to refine structural conformations against projections of single particle images measured by electron microscopy [22]. This refinement was applied to the mmCpn Chaperonin, a protein-folding nanomachine of ~950 kDa, over 60 times as large as the lysozyme protein or over 6 times as large as an IgG antibody.

In recent years, NMMC has been deployed for the training-free atomistic prediction of nucleosome occupancy [30], the assessment of diabody/receptor complexes [24], the study of MHC/peptide dissociation pathways [31] and the structural effects of epigenetic marks on DNA structure [32].

2.1.3 Setting up a molecular simulation

In the following section various aspects important to biomolecular simulations will be discussed and reviewed. A particular focus will be on three important considerations that have to be made before setting up a MCMC simulation: 1. The type of model in the sense of representation and choice of potential; 2. The degrees of freedom considered; 3. The algorithm that should be used to explore the conformational space (Figure 2.1). Depending on the research question the choices for each of these considerations might vary substantially.

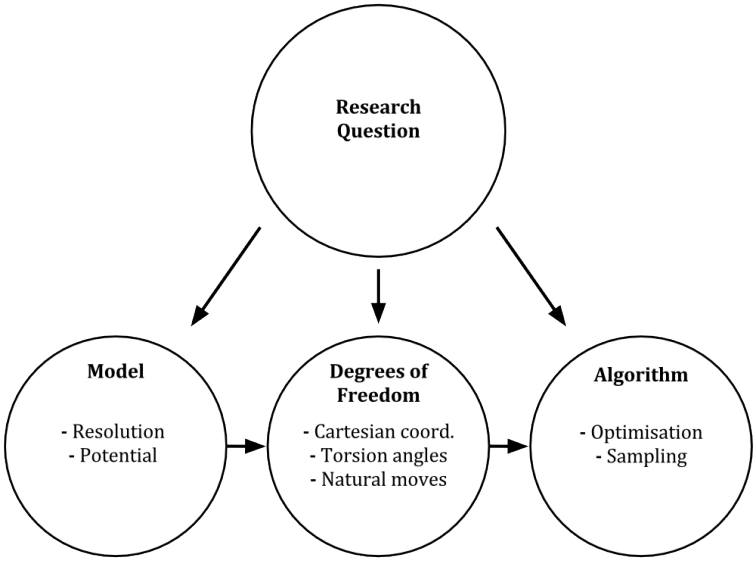

Figure 2.1: Three important considerations for biomolecular simulations. Three important considerations for biomolecular simulations. First, the level of structural detail in the model is determined. Secondly, a decision on the degrees of freedom in a system is made (e.g. torsional/cartesian sampling, Natural Moves, hierarchical Natural Moves). Finally, an algorithm is chosen for exploring the resulting phase space.

2.1.3.1 Choice of model

Resolution Due to sustained attempts to characterise structures at the atomic scale, a number of methods exist to approximate accurately the dynamics and motions of biomolecules [5]. By representing each atom in a biomolecular structure separately and treating solvent molecules explicitly good agreement with experiments can be achieved given a correctly parameterised force field [33–35]. This comes at a computational cost

that limits the scope of this highly accurate and detailed approach. Reduced complexity models offer an opportunity to broaden the scope of biomolecular simulations and have shown to perform sufficiently accurate to generate results in accordance with experimental data [36].

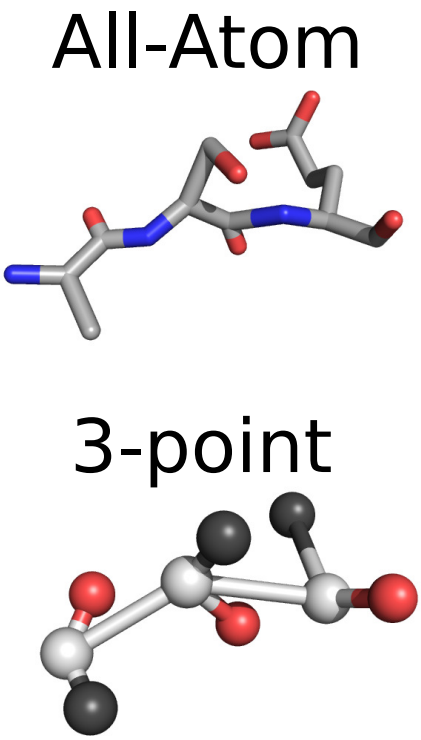

Figure 2.2: All-atom (hydrogen atoms not shown) versus 3-point coarse-grained representation of a tripeptide. Nitrogens are depicted in blue, carbon in grey and oxygens in red. The centre of mass of the side chain is shown in dark grey. The 3-point representation shown here was developed by Minary et al. [21] and was used for all protein simulations presented in this thesis.

The first coarse-grain protein MD simulation method was introduced by Levitt et al. in 1975 [37]. Driven by strong limitations in computational power a simplified protein model was developed, which stripped away all atoms and only considered two interaction sites; the C_α atom and the side chain centre. In addition, a so called united atom representation was used that "unifies" hydrogens with C_α atoms. This resulted in a four-fold reduction in degrees of freedom and cut the number of interaction centres by a factor of fifteen. Despite the loss of information due to coarse-graining, their simulations were able to fold a small polypeptide chain into a structure with a deviation of 3.37 Å when compared to the native state. Since then a range of different bead models have been developed

encompassing anything between one and four beads per residue [38–40]. The Martini model is an exception, which uses a four-to-one mapping, where four heavy atoms are represented by a single particle [41]. Another type of representation are C_α models that are commonly used in Gō models (discussed below).

Potentials Potentials for all-atom representations and explicit solvents have generally been the main focus of most of the common MD simulation packages [42–44]. The potential provides the means for determining the energy of a system while its derivatives represent the atomic forces that ultimately guide the simulation. The typical potential energy function used for biomolecular simulations relies on fixed point charges [44]. The potential energy is updated at each time step according to the sum of internal terms such as bond, valence, dihedral and backbone angles and non-bonded terms such as the Lennard-Jones term and the Coulombic interaction. The parameters for these potentials are commonly optimised using experimental data and quantum mechanical calculations [45].

In addition, implicit solvent models such as distance-dependent dielectric models [46] are commonly used to calculate pairwise Coulombic interactions. Although they are not the most accurate models, they generally achieve good results when compared to experimental data while keeping the computational cost to a minimum.

Another approach was introduced by Marrink et al. [41] who have developed the MARTINI CG model, which is now part of the Gromacs MD programme and is widely used. This mostly empirical approach does not solve or determine the CG potential precisely for a system; rather, simple functional forms such as Lennard-Jones potentials are fitted to experimental data. The MARTINI CG force field has proven especially useful for membrane simulations as the MARTINI parameterisation approach is amenable to describing amphipathic assembly forces [13].

Coarse-grained models are often combined with so called knowledge-based potentials that extrapolate their parameters from structural information derived from the wealth of information stored in large data banks such as the PDB [47, 48]. Generally these potentials are based on simple statistics of pairwise amino acid contacts (two residues

with a distance that is smaller than a certain cutoff). From this data a potential can be generated in the form of a matrix that defines the energy for each interaction [49]. RAPDF (residue-specific all-atom conditional probability discriminatory function) was one of the first potentials to be generated with this approach [50].

Some potentials reshape the energy surface in order to guide simulations and promote a certain behaviour. In 1978, Gō et al. presented a potential that mirrored a perfectly smooth energy funnel, which was especially advantageous in tackling protein-folding problems [51]. Favourable energies were restricted to pairs of residues (usually represented by their C_α atoms) known to interact in the native conformation. However, due to this in-built information the results were heavily skewed towards the native structure and thus were considered to have limited predictive power for applications other than protein folding.

To what extent are Go models useful? There is now a consensus that states that there are severe draw-backs to the perfect funnel model description. A small but significant number of residues are thought to be part of interactions that are important in the transition states but are not present or are weaker in the native conformation. Evidence that shows the limits of the perfect funnel model are examples of proteins that undergo nonnative intermediate states during folding. The majority of these are β-proteins like β-lactoglobulin [52], which is known to have some α-helical structures in its intermediate states. For this protein a simple Go model struggles to give the right answer.

Generally, the notion of nonnative interactions is an import one, and gives rise to the question how specificity arises. The energy-landscape model incorporates two concepts; it explains the folding of proteins driven by native interactions using the global funnel shape idea and describes non-native interactions by the local perturbations of the energy surface. The success of Go models in predicting folding pathways may be ascribed to the validity of the former. Experiments as well as simulations show that for fast folding small proteins the global funnel concept seems to be the main contributing factor. However, for better modelling, the potential functions have to become more accurate so that the landscape is smoothed while the depth of the native well is maintained sufficiently large.

Go models are largely employed for the modelling of protein folding pathways as well as interactions in larger biomolecular complexes [53, 54]. The physical relevance of the simulated interactions are unclear, however, as they are described by the native structure instead of physico-chemical characteristics of the system [55]. Consequently, Go models cannot predict the folding of unknown structures, and are unsuitable for exploring the conformational space away from the native structure.

Other methods approximate the vibrational fluctuations and positional uncertainties in a protein structure by a network of springs connecting a set of spheres. Tirion et al. first outlined the elastic network model (ENM) approach in 1996 [56]. Originally, a single spring constant acts on these 'bonds' with a length shorter than a certain cut-off distance. Combined with normal mode analysis (NMA) these models provided a unique insight into the low-frequency modes of biomolecules of all sizes (including the ribosome [57]), however their utility is limited due to the harmonic nature of the system, which prevents the exploration of large conformational changes when applied in the traditional manner. Instead of connecting the atoms by their covalent bonds as described by the x-ray structure, all atoms that are spatially proximal to each other are connected by springs. In the original version all particles in the system were treated identically and the springs had the same stiffness. Pairs of particles whose distance is below a certain cutoff are connected by springs with a certain stiffness. This representation can then be used to describe a system's collective motions using normal mode analysis (NMA). Tirion et al. have demonstrated that a single-parameter spring potential is sufficient to reproduce the slow elastic modes of proteins observed with sophisticated multi-parameter empirical potentials. This insight represented a milestone with regards to the simulation of protein dynamics as it made the potential robust against details of the equilibrium structure, which was defined to have minimal energy by construction (this is a disadvantage for sampling conformational space far away from the starting structure as will be discussed later). Thus, for the first time experimental structures could be used as is, without the inefficient energy minimisation step associated with atomic force fields.

A further approximation was introduced by Bahar and Hinsen, who applied the same

harmonic potential to a simplified model in which each residue was only represented by a single particle at the position of the respective $C\alpha$ atom [58]. This approach has also been shown to be effective in describing the thermal fluctuation motions of proteins.

Elastic normal modes may be regarded as a straight forward and interpretable representation of the protein's collective motion, which can be conveniently coarse-grained and are computationally inexpensive. As a result they quickly replaced molecular mechanics force fields that had been used for NMA of proteins in previous years [59–61].

It is important to note that NMA is based on the approximation of the energy function around the equilibrium conformation of the elastic network. It came as a surprise then that large-amplitude functional motions in some proteins could be accurately described by this approximation. However, in general it is not known to what extent normal mode approximation is valid, especially for conformations distant from the starting structure. It is problematic to describe large-scale motions in proteins in terms of the near-equilibrium properties of the elastic network.

Another weakness of the method is the often arbitrary choice of the cutoff-length for the construction of the network. If chosen too large the network tends to include a vast number of connections including very weakly connected residue groups or even disconnected particles, if chosen too small crucial connections are omitted. In an attempt to remove the cutoff-length as a free parameter Yan et al. [62] have proposed a parameter-free elastic network model, where interactions between all network particles are assumed, but their interactions strengths are determined by springs constants that are inversely proportional to the distance.

The effectiveness of NMA has been discussed extensively, especially in a comprehensive review by Yang, Song and Jernigan [63]. It was investigated how well normal modes of elastic networks characterised large-scale protein movements. The study concluded that ENM gave the most accurate results for structural changes that were collective in nature but struggled for non-collective conformational transitions.

Note that simplified methods have been developed since, that were inspired by ENM, namely the Gaussian Network Model (GNM) and Anisotropic Network Model (ANM).

GNM uses a N x N matrix based on inter-residue contact topology instead of the $3N$ x $3N$ force constant Hessian matrix used in traditional ENMs. The ANM was built on the GNM and includes information on the directions of individual residue fluctuations.

Despite their simplicity, the GNM and ANM have been widely used for the study of protein dynamics, for example to elucidate both the molecular machinery and structural plasticity of macromolecular structures and complexes including HIV reverse transcriptase [64], hemagglutinin A [65], F1 ATPase [66], RNA polymerase [67], GroEL-GroES [68], the ribosome [69], and viral capsids [70].

Furthermore, large scale GNM studies have revealed conserved dynamical behaviour between different biomolecules [71], the effect of topology on native structural stability [72] or the automatic domain characterisation of proteins [73].

One of the biggest advantages of these methods is that low-energy modes are robustly maintained [15]. However, most ENMs lack information on residue specificities, atomic interactions, and side-chain fluctuations.

2.1.3.2 Degrees of Freedom

Global and Local Torsional Angles In order to improve upon the efficiency of traditional approaches using Cartesian coordinates, computationally cheaper methods have been introduced. However, efforts to address the dimensionality problem often lead to an increase in complexity of the energy landscape. For instance, a popular approach to reduce the degrees of freedom is to use torsional angles instead of Cartesian coordinates to describe the structure of a biomolecular system [74]. In this method new molecular conformations are generated by gradually 'growing' the structure with new dihedral angles from a given seed residue. This approach, however, struggled to generate feasible conformations for structures larger than short peptides, due to low acceptance rates in traditional Markov Chain Monte Carlo simulations so that hardly any sampling progress was made. The limitation of this so called global torsional angle sampling method, resulted from the lever arm effect, which caused small changes at the beginning of the chain to be propagated through and amplified at more distal parts of the structure, thus leading

to energetically unfavourable conformations and poor acceptance rates.

This issue was addressed with the introduction of local torsional angle sampling. Here, torsional moves of segments are sampled in isolation while the remaining part of the structure is kept untouched [25]. Consequently, the chain connecting segment (molten zone) may be broken at every step and needs to be closed using numeric chain closure algorithms [29] at the expense of computational resources. While significant improvements in acceptance rates is achieved, the limitations of short gap closure distances and computational cost due to the closure algorithms persist.

Natural Moves Learning from the benefits and shortcomings of local torsional angle sampling methods a new concept called Natural Moves was presented in 2010 [18]. By introducing an improved stochastic chain closure algorithm that could be used in combination with any advanced sampling method, efficiency was increased, larger chain break distances could be bridged and arbitrary degrees of freedom could be introduced into the system. By these means it became possible to define collections of residues that could be propagated as unified bodies rather than separate entities. Therefore, a structure could be decomposed into a set of segments that would undergo so called Natural Moves. This allowed for the drastic reduction of the conformational space to be sampled in a system and therefore provided large improvements in performance [19]. Figure 2.3A and D show two examples systems that may be decomposed into segments differently due to their scale and according research question. Figure 2.3B and E show two different segmentation strategies for Natural Moves. The strategies shown are purely based on structural information but could be extended by more experimental or simulation data.

Hierarchical Natural Move Monte Carlo Hierarchical Natural Move Monte Carlo (hNMMC) [19] is based on the same building blocks as NMMC but adds an additional layer of hierarchy to the simulation. In addition to segments that may represent secondary structure elements, higher level entities can be defined which may encompass multiple segments, domains, subunits and event tertiary structures. These so called "regions" provide the means to simulate even larger structures without facing a significant increase in

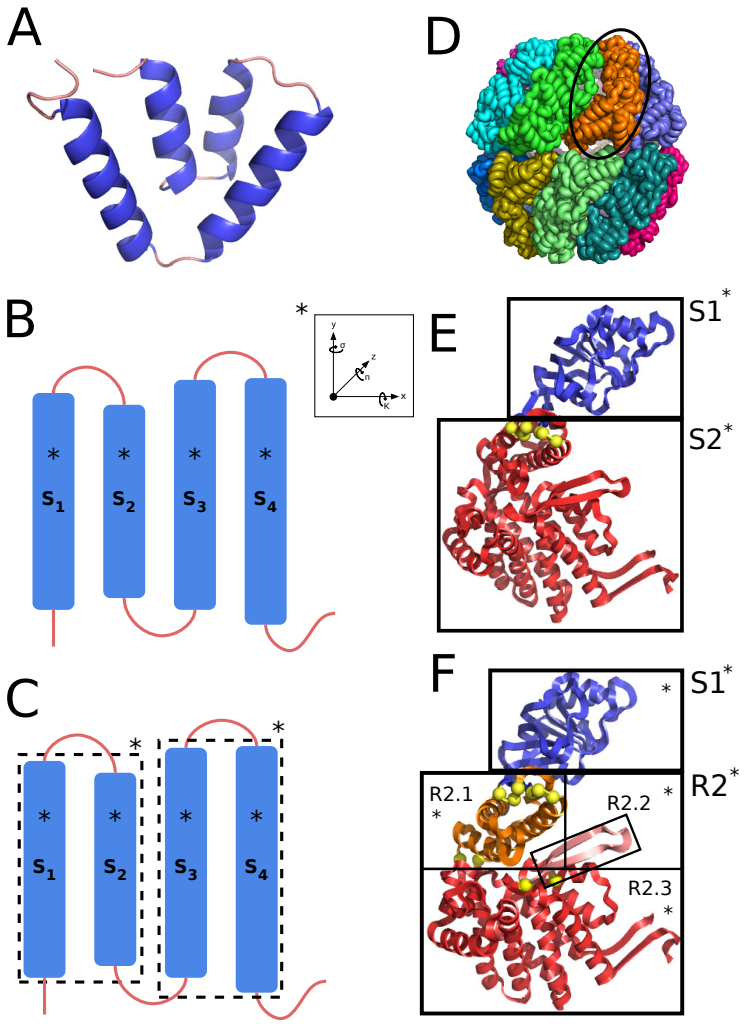

Figure 2.3: Examples of natural and hierarchical Natural Moves.**A** Cartoon representation of uteroglobin, an α-helix-bundle (PDB accession code: 1UTG). Helices are shown in blue, loops in red. **B** Schematic representation of uteroglobin and its decomposition into segments and molten zones. Each of the four segments can move independently along six degrees of freedom in a Natural Move Monte Carlo (NMMC) simulation (indicated by *). The segmentation was defined using secondary structure information. **C** The same system as shown in B but using regions, which contain segments. These are used in hierarchical Natural Move Monte Carlo (hNMMC). Regions can be moved collectively, while the individual segments within regions retain their independent movement, thereby confering 'internal flexibility'. **D** Cartoon representation of the Methonococcus maripaludis chaperonin (Mm-cpn). **E** The subunit highlighted in D is shown in more detail. In NMMC each subunit may be decomposed into two segments. **F** In hNMMC, however, using experimental data or intuition one of the segments may be further split by the user resulting in region R2 with sub-segments (R2.1, R2.2, R2.3).

complexity [18]. Therefore, this method allows for the analysis of the functional motions of very large biomolecular complexes [19, 22], as hierarchical Natural Moves can be adapted to account for the size of the molecule. A more detailed description of the method will

follow in the next chapter. Figure 2.3C and F show examples of hierarchical Natural Moves.

2.1.3.3 Algorithms

Two common objectives when performing biomolecular MC simulations are the adequate sampling of structures to explore conformational space and the locating of energy minima. There are various algorithms that facilitate either of these objectives, few of which I will briefly discuss below.

Simulated Annealing Simulated annealing is a method commonly used for finding approximate solutions to global optimisation problems [75]. By setting a high initial temperature and subsequently cooling the system gradually, high energy barriers can be overcome and local minima can be identified effectively. In order to improve upon this algorithm periodic temperature modulations can be introduced that repeatedly allow a system to evade local minima. By gradually heating and cooling the system there exists a higher probability of finding low energy conformations. Such a temperature function may take the following form:

$$T_k = A \times \sin\left(\frac{2\pi k}{\Omega}\right) + s \tag{2.1}$$

where k is the MCMC step counter, Ω is the number of steps per period, A is the amplitude and s is used to shift the minimum temperature.

Figure 2.4 illustrates a typical biomolecular simulation using the repeated simulated annealing protocol.

Parallel Tempering Parallel tempering is popular for sampling large volumes of conformational space and provide canonical distributions. Here one performs multiple simulations at increasing temperatures in order to improve the sampling of a particular system [76]. Adjacent simulations may exchange states according to the metropolis criterion, if their potential energies are sufficiently close. Figure 2.5 shows the energy distributions of different replicas during a parallel tempering simulation.

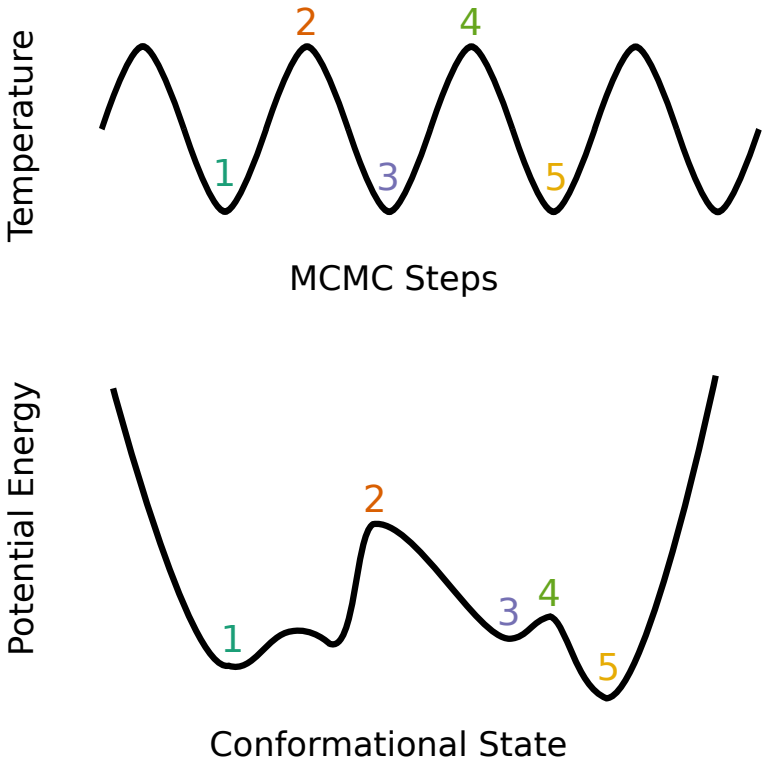

Figure 2.4: Illustration of a repeated simulated annealing simulation. The top figure represents a typical temperature profile typically used for repeated simulated annealing. Higher temperatures (states 2 and 4) facilitate the crossing of energy barriers, while lower temperatures (states 1, 3 and 5) enable the search for local energy minima.

This allows for states that result from high temperature simulations to diffuse across replicas. Since high temperature systems are able to surpass larger energy barriers than low temperature simulations, this enhances conformational sampling along low temperature replicas (usually the system of interest). Note, that the probability of state exchange between replicas is designed so that each replica samples states according to their canonical probability of occurrence at the given temperature. Therefore, this enhanced sampling method does not alter the correct canonical distribution of observables (e.g. RMSD, inter-helix distance) but only accelerates the rate of convergence to the desired distributions.

Formally, $K+1$ Markov Chain simulations are performed simultaneously: $\{X^{T_0}, X^{T_1}, ..., X^{T_K}\}$ and each trajectory samples from a Boltzmann distribution, $f_i(X, T_i)$, at a given temperature T_i where $i = 0, ..., K$. Adjacent Markov chains X^{T_i} and $X^{T_{i+1}}$ may exchange states at step n with probability $P_a = \min\{1, f_{i-1}(y)/f_{i-1}(z) \times f_i(z)/f_i(y)\}$, where $y = X_n^{(T_i)}$ and $z = X_n^{(T_{i+1})}$.

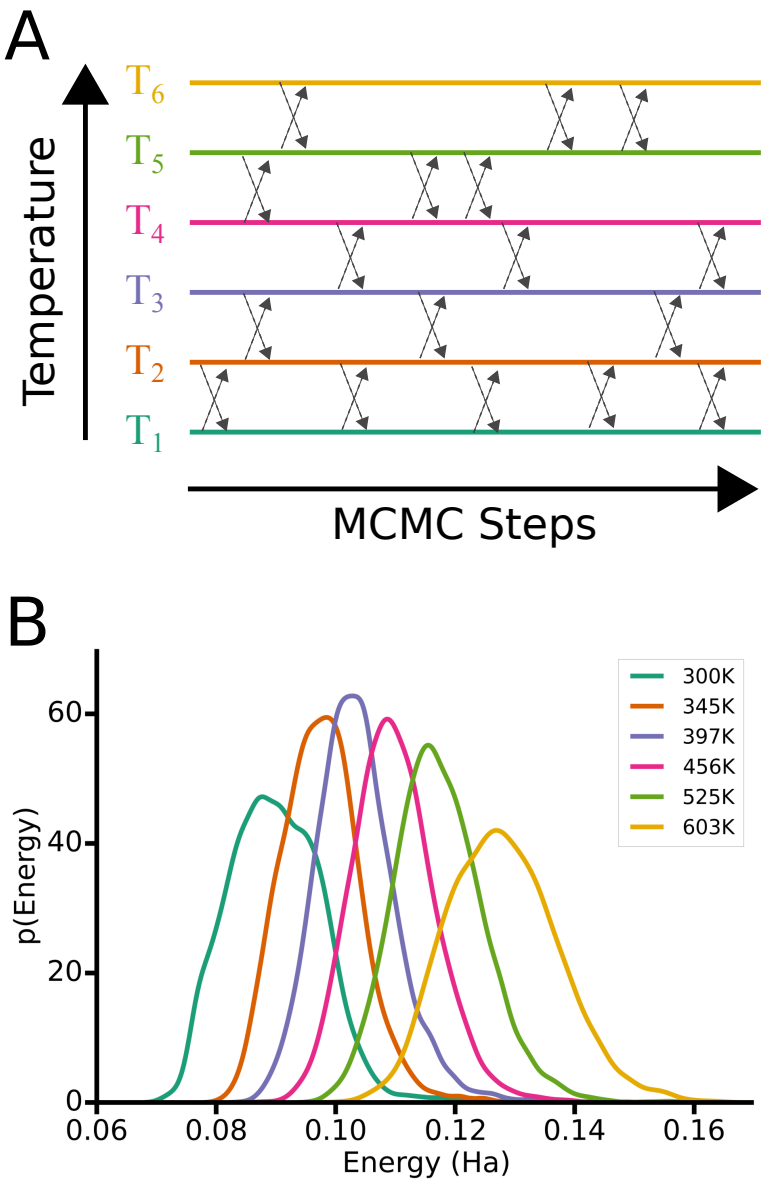

Figure 2.5: **A** The six simulations running at different temperatures in parallel are shown. State exchange between temperature levels is represented by dashed arrows. **B** The energies for replicas running at different temperatures during a typical parallel tempering simulation.

2.2 Biological background

2.2.1 Protein and DNA structure

2.2.1.1 Nucleic Acid Structure

DNA and RNA have highly similar chemical compositions, however their structures can vary greatly. DNA and RNA both consists of four different nucleotides that each contain a

phosphate group linked by a phosphoester bond to a pentose that itself is connected to an organic base. In DNA, the pentose is deoxyribose; in RNA, it is ribose. Furthermore, one of the four organic bases differs between the two molecules. The bases adenine, guanine, and cytosine are found in both DNA and RNA, whereas thymine only exists in DNA, and uracil only in RNA. RNA mostly exists as a single polynucleotide strand that can fold back on itself, while the chemical composition of DNA strongly promotes the formation of a double helix with two intertwined polynucleotide strands. This structural difference is crucial for the different functions of DNA and RNA.

The charged nature of the sugar-phosphate backbone means that nucleic acids are frequently affected by the local surroundings; for example, ligand binding often causes large structural changes [77, 78]. However, the sequence-context is critical. Pyrimidine-purine dimers are especially flexible in protein-DNA interactions [79] and GG:CC dimers are amenable to drug- and protein-induced conversion to the A-DNA form [80].

Comprehensive investigations of three-dimensional structures and interactions of nucleic acids require quantitative tools for characterising the spatial orientations of the constituent molecular units. In 1988 a group of experts developed a set of parameters for describing the orientation and displacement of bases in double helical nucleic acid structures [81]. However, due to different programmes using different standards a number of conflicting interpretations arose for identical structures [82]. Structural segments that were considered 'normal' according to one method was classified as very unique according to another. As a result it was challenging to conduct extensive comparative studies of nucleic acid structures and to identify unusual features in individual structures. To resolve these issues a standard base-centred reference frame was introduced in 2001 for use in nucleic acid structural studies [83].

The hydrogen bonding and stacking of base pairs are fundamental to DNA and RNA structure. Canonical base pairs (Watson-Crick) in double stranded DNA represent only one of many possible interactions of bases. The range of structural forms seen in RNA and DNA, e.g. (rRNA [84] and DNA tetraplexes [85]) are a result of a wide variety of alternative base pairing motifs [86].

In addition, water and ions play an important role in stabilising RNA and DNA structures. For example, the transition from B- to the A-form of DNA that occurs as a result of reduced humidity can be prevented by the addition of divalent metal ions [87].

In the following sections I describe in more detail the structural characteristics of nucleic acids.

Primary Structure As with proteins, the primary structure of nucleic acids refers to the linear form, the sequence, of the polymer. Cellular RNAs range in length from less than one hundred to many thousands of nucleotides. Cellular DNA molecules can be as long as several hundred million nucleotides.

Secondary Structure The secondary structure in nucleic acids refers to the set of interactions between bases. In DNA these interactions mostly occur between bases of two different strands, whereas in RNA these interactions occur within the same strand. The simplest secondary structures in single-stranded RNAs are formed by pairing of complementary bases. 'Hairpins' are formed by pairing of bases within 5 to 10 nucleotides of each other, and 'stem-loops' by pairing of bases that are separated by approximately 50 to several hundred nucleotides.

Tertiary Structure The tertiary structure describes the locations of the atoms in three-dimensional space, taking geometrical and steric constraints into account. There are four ways in which the structures of DNA molecules can differ.

- Handedness - right or left
- Length of the helix turn
- Number of base pairs per turn
- The ratio of the widths of major and minor groove

The tertiary structure of the DNA double helix in space can take on different forms: B-DNA, A-DNA, Z-DNA and F-DNA.

B-DNA is the most common form of DNA *in vivo*. It is a narrower, longer helix than A-DNA. It has a wide major groove, which makes it a good binding site for proteins. The

hydration of the minor groove favours B-DNA. B-DNA base pairs are almost perpendicular to the axis of the helix.

A-DNA is usually observed under dehydrating conditions. It is shorter and wider than B-DNA. RNA adopts this double helical form, and RNA-DNA duplexes are mostly in the A-form. A-DNA has a deep, narrow major groove, which reduces its accessibility to proteins. The base pairs in A-DNA are tilted relative to the axis of the helix.

Z-DNA is a rarely observed left-handed double-helix. Given the right sequence and super-helical tension, it can form *in vivo* but its function is unclear. It has a more narrow, more elongated helix than A- or B-DNA. The major groove is hardly existent, and it has a narrow minor groove. The most favourable conditions for Z-DNA are high salt concentrations.

F-DNA is a recently discovered form that occurs in GC-rich sub-regions of the DNA structure where multiple cytosines in a row carry an epigenetic (formyl) mark. Its characteristic helical underwinding makes it easy to distinguish from the other forms mentioned above [88].

In RNA, simple secondary structures can cooperate to form more complicated tertiary structures, one of which is termed a ‘pseudoknot‘. Pseudoknots are formed when nucleotides from the hairpin-loop pairs with a single stranded region outside of the hairpin to form a helical segment. Pseudoknots are diverse functional elements in RNA structure and are found in most classes of RNA.

Tertiary structures have also been shown to exist in mRNA, particularly near the termini of molecules. Similarly to proteins, they can have structured domains linked by less structured, flexible regions, and sometimes also exert catalytic activity.

Quarternary Structure The quaternary structure of nucleic acids refers to interactions with other biological molecules. A frequently seen form of quarternary structure is seen in the form of chromatin, which is formed by the interaction between DNA and small histone proteins.

2.2.1.2 Protein Structure

Primary Structure An amino acid consist of a carboxylic acid that is attached to an amino group. Apart from glycine, which has no side chain, amino acids have chirality (the molecule is not superimposable on its mirror image); the L-form is the most commonly observed. Twenty-two different amino acids are encoded by 61 codons (the remaining 3 are stop codons). Polycondensation of amino acids in the ribosome leads to proteins. Each of the amino acids has unique physicochemical characteristics, which enable them to play different functional roles within a protein. The primary structure is schematically shown in Figure 2.6.

Secondary Structure A peptide bond between two amino acids is formed by condensation under the elimination of water. Addition of more amino acids to the dipeptide results in peptide chains. Chains of 20 residues and less are called oligopeptides, while longer chains are referred to as polypeptides. Proteins are polypeptides with a biological function. Polypeptides vary in length from a few dozen to thousands of amino acids. Proteins may be made up from a single polypeptide chain, or multiple polypeptide chains (subunits). The amino acid sequence of a protein is called the primary structure, with the N-terminus defining the beginning and the C-terminus the end of the molecule; proteins are also synthesised in this direction.

The regular, repetitive folding pattern of the primary structure is referred to as secondary structure. Hydrogen bonds within the protein backbone (between the amino- and keto-groups of the peptide bonds) stabilise secondary structure. While the bond energy of an individual hydrogen bond is weak, the total of all hydrogen bond energies amounts to a significant effect on the structure. Secondary structure comes in several forms, two of which are shown in Figure 2.6.

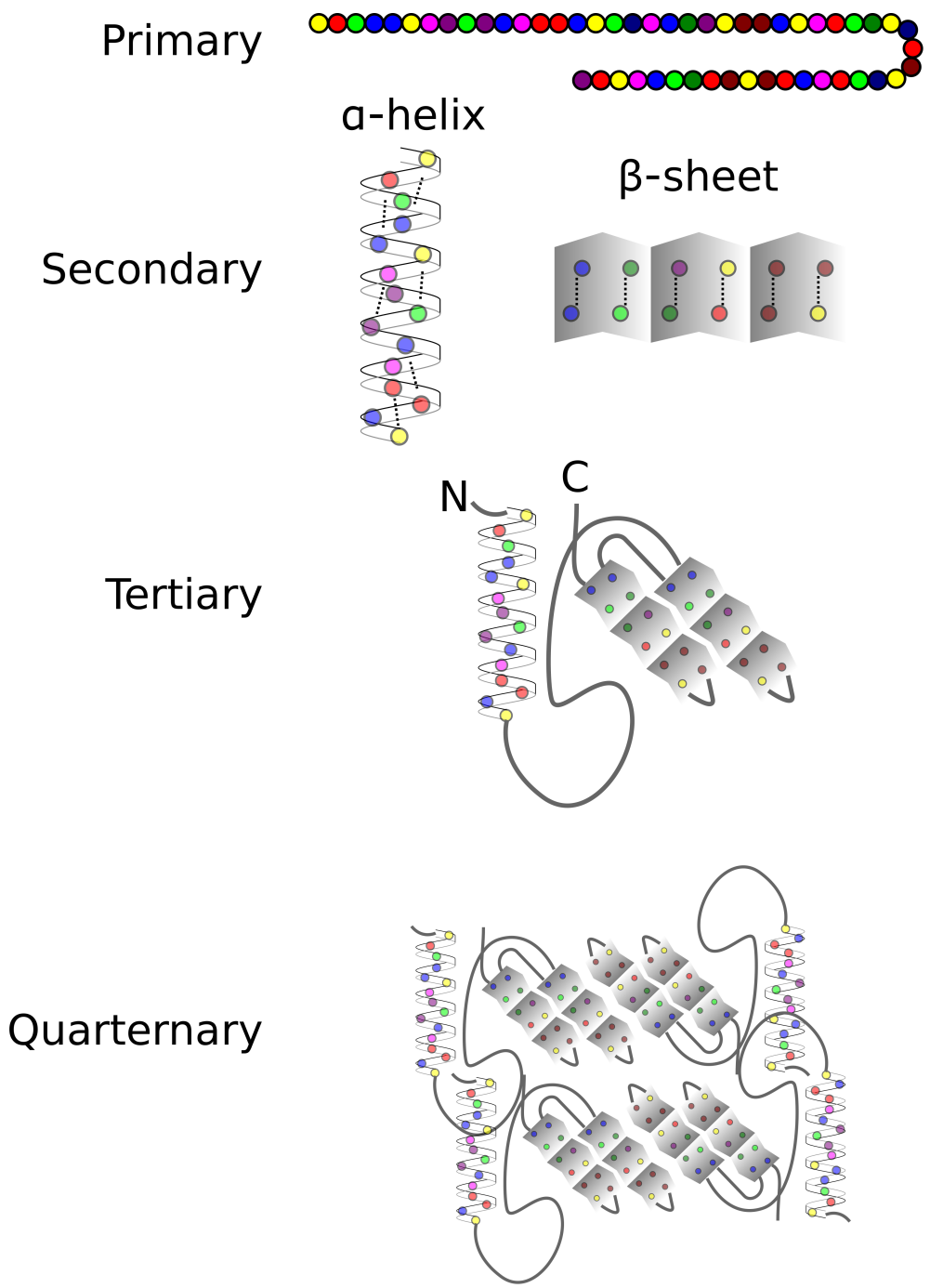

Figure 2.6: The hierarchy of protein structure. The primary structure refers to the amino acid sequence of a protein, which is represented by a string of coloured beads. Secondary structures are regions with a well defined regular structure stabilised by hydrogen bonds (black dashed lines). These include α-helices and β sheets. Residues that are part of the same secondary structure element are often moved collectively as segments in Natural Move Monte Carlo. The tertiary structure refers to the overall arrangement of a single polypeptide chain into a three-dimensional structure. Multiple structural elements (secondary structures, domains) may be grouped and moved collectively as regions (sets of segments) in Natural Move Monte Carlo. The quarternary structure describes the arrangement of multiple folded protein subunits in a multi-subunit complex. One or more protein subunits may be moved collectively in Natural Move Monte Carlo.

The most commonly observed secondary structure is a right-handed spiral with 3.6 amino acids per turn, called the α-helix. Left-handed helices are less common for L-amino acids, as the α-carbon and carbonyl-oxygen would clash. The side chains points outward and the compact conformation is stabilised by hydrogen bonds between a carboxyl-oxygen (partial negative charge) and the amino hydrogen (partial positive charge) 4 residues down the chain. A complete helical turn occurs every 3.6 amino acids on average. Proline and glycine are usually not found in α helices as their physio-chemical properties do not suit the α- helix. Proline is known to break or cause kinks in α-helices as the lack of a secondary amide prevents stabilisation through hydrogen bonding and also because its sidechain interferes sterically with the backbone of the preceding turn - inside a helix. However, prolines are often found at the N-terminal beginning of helices, which is thought to be due to its structural rigidity that can support the formation and maintenance of the secondary structure.

Proline either breaks or kinks a helix, both because it cannot donate an amide hydrogen bond (having no amide hydrogen), and also because its sidechain interferes sterically with the backbone of the preceding turn - inside a helix, this forces a bend of about 30Âř in the helix's axis.[10] However, proline is often seen as the first residue of a helix, it is presumed due to its structural rigidity.

In the β-strand, the polypeptide backbone is extended. When multiple strands are arranged either parallel (carboxyl groups are on the same side) or antiparallel, they form hydrogen bonds between each others NH and C=O groups. This arrangement is referred to as a β-pleated sheet. The main differences between the α-helix and β-strand is that the hydrogen bonds occur between residues of the same helix and between amino acids of adjacent strands, respectively.

Even though it is energetically more favourable for β-strands to interact with each other, individual β-strands are stable because the amino acids in this extended structure benefit from entropic stabilisation. The side chains alternately point up- and downwards, and is usually not entirely flat, but has a right-handed bend (e.g. β-barrels). In antiparallel β-sheets the strands run in opposite directions. They are often connected through

β-turns. In a parallel β-sheet the strands point the same way. Hydrogen bonds are less favoured in the parallel β-sheet, thus it is less stable than the antiparallel form. The strands of parallel β-sheets are commonly linked by α-helices.

Coils are structures that do not fall in the categories discussed above. While their geometry is not as well described, they still have a defined position within the protein. Coils are sometimes referred to as 'random' or 'unordered', which can be misleading. Coils play an important role, because they provide plasticity and allow for conformational changes. Their backbones do not participate in intra-molecular hydrogen bonding and thus are commonly accessible for interactions with water, ligands, or other proteins.

Tertiary Structure The position and orientation of all secondary elements in a protein structure is described as tertiary structure. One of the main factors that influences tertiary structure is the hydrophobic effect, i.e. the propensity of non-polar molecules to aggregate in solution and exclude water molecules. Generally, in globular proteins, amino acids with hydrophobic side chains are found in the core, while the hydrophilic residues are usually exposed on the surface.

Van-der-Waals-interactions are transient interactions between dipoles with a bond length of roughly 4 Å. In contrast, hydrogen bonds occur between permanent partial charges and have a bond length of about 3 Å. For interactions that are further apart, indirect hydrogen bonds occur with water acting as a bridge (water-mediated hydrogen bond). Salt bridges are formed between fully charged species and have a bond length of 2.8 Å. Disulphide bonds are found between two cysteines, usually once folding has been completed and the native structure has been formed. This oxidation, where hydrogen is removed, does not usually take place in the cytosol, which has a reducing environment. Rather, disulphide-bridge formation often occurs in the oxidising environment of the endoplasmic reticulum. Consequently, disulphide bridges are found more often on the cell's surface and in secreted proteins than in cytosolic proteins. The length of a disulphide bond is 2.2 Å.

Some proteins consist of multiple domains, modular structural entities connected by short segments. A domain is a conserved part of a protein that can undergo evolutionary

changes, perform functions, and exist separately from the remaining structure. Domains usually have a well-defined three-dimensional conformation and generally are stable and can fold on their own. Due to the increasing numbers in deposited protein structures, researchers have been able to characterise large numbers of proteins according to features such as certain recurring tertiary structure patterns. These folding patterns are referred to as motifs. Interestingly, motifs are more evolutionary conserved than amino acid sequences. Numerous proteins are structural homologues of each other, despite their sequence identity being low.

Using these motifs, protein domains have been classified into hierarchical groups. Due to the inexact definition of classifying three-dimensional structures a number of different approaches have been developed. One of the most popular is the Structural Classification of Proteins (SCOP) database (since 2014 SCOP2, http://scop2.mrc-lmb.cam.ac.uk/).

An illustration of tertiary structure is shown in Figure 2.6.

Quaternary Structure A number of proteins form multimeric complexes to perform their function. The arrangement and orientations of subunits in such assemblies is referred to as quaternary structure. As with tertiary structure, ionic and hydrophobic interactions are key factors in the formation of the quarternary structure. A schematic of quarternary structure is depicted in Figure 2.6.

2.2.2 Collective Motions

The idea of the energy landscape has gained most of its traction in the context of protein folding [89]. However, this concept had been developed before to describe the structural states involved in the function of myoglobin [90, 91]. Frauenfelder et al. found that the energy-landscape concept helped to characterise various kinetic and functional features, and were able to identify the heights of energy barriers and the multiple conformational substates [92].

Large-scale conformational transitions generally require collective, large-amplitude motions as well as small-amplitude local fluctuations on the picosecond timescale. How are

these dynamic ranges connected? The interplay between these timescales was elucidated by a NMR spectroscopy and MD simulation study of the adenylate kinases thermoAdk and mesoAdk [93]. Local backbone regions that connected the domains that moved during lid closure showed increased picosecond dynamics. Interestingly, in mesoAdk these regions were more flexible at low temperature than in thermoAdk, and a correlation between hinge dynamics and catalytic activity was observed. The authors suggested that the hinge regions were the origin of the catalytically important collective domain motions. Increased packing and hydrogen bonding in the thermoAdk hinge regions, due to variations in the sequence were thought to be responsible for the difference in their dynamics.

The characterisation of protein dynamics includes the timescale (kinetics) as well as the amplitude and directionality of motions (structure). As a result, the energy landscape for a typical protein is high-dimensional. Thereby it is important to note that a particular energy landscape is affected by temperature, pressure as well as solvent conditions. By changing these parameters the relative proportion of states and the kinetics can be modulated. Computational algorithms such as simulated annealing and parallel tempering for example make use of this phenomenon by exploring conformations in a range of different temperatures, thus allowing simulations to cross energy-barriers and explore multiple energy-minima. Theoretically, the energy landscape describes all the states that could be sampled by the system, including unfolded structures. However, dimensionality reducing simulation techniques, such as Natural Move Monte Carlo, can be used to limit this phase space (conformational space) by imprinting certain assumptions into the simulation; for example the assumption of collective motion of secondary structure elements in folded proteins may be imprinted by treating α-helices and β-sheets as rigid bodies. The following discussion among others describes how such assumptions can be justified.

Slow Timescales Protein motions can be divided into groups of slow and fast timescales. 'Slow' dynamics may be fluctuations between different states (tier-0 states) that populate distinct energy minima. These motions correspond to microseconds and longer at 310 K. Generally, these include collective motions between just a few states. The system is not

static within one of these states, it rather undergoes smaller and faster timescale motions around the average structure, thereby visiting a large number of very similar structures. Transitions between tier-0 states occur less frequently than between higher tier states, due to the large concerted effort that is required to enable the transition. Domain motions and dynamics of this size and this timescale are biologically interesting as they seem to drive functions such as enzyme catalysis, signal transduction, protein-protein interactions and cooperative/allosteric switches. Due to their relatively long half-lives, tier-0 states can be studied with a range of experimental techniques (discussed below).

Fast timescales 'Fast' dynamics, referred to as tier-1 and tier-2 dynamics, are fluctuations that occur within a tier-0 state. These are motions that do not change the global structure significantly, leading to a large number of highly similar conformations that are separated by low energy barriers and occur on the picosecond-to-nanosecond time scale at 310 K. Henzler-Wildman et al. describe tier-1 and tier-2 motions as small numbers of atoms moving collectively on the nano-/microsecond (e.g. loop reconfiguration) and atomic fluctuations on the picosecond scale (side-chain isomerization), respectively [94]. Higher tiers describe fluctuations such as femtosecond bond vibrations. It is well known that dynamics are encoded by the three-dimensional structure i.e. the geometry or topology of native contacts, with backbone atoms within secondary structures fluctuating less than atoms in unstructured regions.

Karplus and co-worker observed in their simulations that not only the motions of adjacent residues were highly correlated (covalent bonds). The motions of residues that were further apart were also correlated, albeit to a lesser degree, as long as they were located in the same secondary structure element (hydrogen bonding). For the first time their simulations suggested that it is possible that secondary structural elements can move as collective entities [95].

It has become common to group protein movements into hinge and shear motion [96, 97]. Hinge movements describe the rotation of protein substructures (usually domains) around a hinge element such as a loop. These movements commonly include a small group of residues that undergo significant conformational changes, while the rotating

protein substructures that make up the majority of residues maintain their structure. Shear movements on the other hand describe the sliding movement of substructures with respect to each other. These motions are usually more subtle, with small structural changes along the plane of movement. However, it has become apparent that the repertoire of conformational changes in proteins may be more complicated [98].

2.2.2.1 Experimental methods for studying collective motions

The ultimate goal of structure elucidation is to characterise the functionally relevant conformations in a biological system as well as its transition pathways and conversion rates. Techniques such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, cryo-electron microscopy and small-angle X-ray scattering (SAXS) provide medium to high resolution structures of tier-0 states. For X-ray crystallography, which provides atomic detail, a homogeneous crystal is required. Thus, different conformations of structures have to be induced by biochemical means [99]. A good example of such a procedure was the crystallographic characterisation of cytochrome P450 structural substates that are responsible for its enzymatic activity [100].

No crystal are required for cryo-electron microscopy or small-angle X-ray scattering, enabling the elucidation of structural ensembles present in the sample, albeit with lower resolution. However, these methods are unsuitable for determining interconversion rates. Structural data generated with these methods is often supplemented by kinetic data generated from spectroscopic methods. Furthermore, hydrogen-deuterium exchange in combination with mass spectrometry or NMR spectroscopy can be useful to identify unfolding on long time scales (milliseconds and longer) [101].

NMR can reveal rates of transitions as well as atomic resolution structures. The dynamics are inferred from observing the relaxation of nuclei after excitation [102]. The dynamics can be monitored while the structure is in solution in steady-state conditions [102], as opposed to other spectroscopic methods, where the system is perturbed to calculate the kinetics. Traditionally, NMR experiments were only possible with small, soluble proteins. More recently, however, modern spectrometers have enabled the study of larger

systems of 100 kDa and higher (and even up to the size of the ribosome) [103–105]. Solid-state NMR, in particular, now allows to resolve large membrane-protein complexes at atomic resolution [106].

About a decade ago, older techniques based on fluorescence such as circular dichroism, infrared spectroscopy, Raman spectroscopy and electron paramagnetic resonance have experienced a new surge in popularity due to their ability to measure kinetic information. These methods can examine a large range of timescales with relatively high precision and generate information for one or more sites of interest.

Single-molecule fluorescence [107] allows researcher to observe in real time conformational changes in a single molecule. The advantage of single-molecule methods lies in their ability to identify molecular heterogeneity, transient states, rare events and the order of events, all of which is often lost in ensemble-averaging methods. Furthermore, fluorescence resonance energy transfer (FRET) makes it possible to measure intermolecular distances over time [108].

Site-directed spin labelling experiments have found characteristic signals in α-helical structures that strongly suggests collective modes of motion: rigid-body motion of the helix around three orthogonal axes [109].

2.2.2.2 Computational Methods for studying collective motions

One of the main advantages of computational methods is the intricate level of detail by which systems can be studied. In theory, the exact location of all atoms over time for one or more molecules in any solvent can be simulated, and statistics such as energies and kinetics can be calculated given that at least one high-resolution structure is available. However, results should be interpreted with care as we are far from quantum-level, long-timescale simulations that give accurate results [2].

Full-detail all-atom MD simulations on the microsecond-to-millisecond timescale is still unfeasible for most research groups. In order to bridge this gap a large number of strategies have been developed that use simplified force fields including normal mode analysis [110] Gaussian network models [111], FIRST (floppy inclusion and rigid sub-

structure topography) [112], FRODA (framework rigidity optimised dynamic algorithm) [113] and Go models [114]. Other MD approaches involve accelerating the dynamic process by applying an external force (targeted, steered and accelerated molecular-dynamics simulations [115–117]), or they use prior knowledge about the conformational change (potential of mean force in umbrella sampling algorithms [118]) or the transition end points (transition-path sampling [119]). Other techniques include hyperdynamics [120], replica exchange molecular dynamics (REMD) [121], metadynamics [122] and adaptive sampling [123] among others.

REMD performs independent parallel Monte Carlo random walks across several parallel MD simulations that are performed at different temperatures. This technique is based on the parallel tempering algorithm [124] that is also used in NMMC; it allows for the efficient exploration of conformational space across energy minima. Parallel tempering was first applied in combination with MC simulations. The transfer of this algorithm to MD has introduced sampling problems resulting from the failure of the constant-temperature MD integrators to maintain certain variants. REMD algorithms cannot use the leap-frog integrator commonly used for microcanonical (constant energy) MD and resort to isothermal (constant temperature) integrators. This can have a detrimental affect to the dynamics of the system. Entropy-preserving constant-temperature integrators have been used but their application remains limited [125].

Metadynamics adds memory to the sampling by penalising the re-sampling of previously visited states, thereby 'filling the free energy wells with computational sand'. It has been used to study questions regarding protein folding, molecular docking, phase transitions and conformational changes [11].

Hyperdynamics accelerates molecular dynamics simulations by reducing the sampling time in potential energy minima in order to overcome energy barriers. The method modifies the potential energy surface by adding a bias potential to the true potential such that the potential close to the minima is increased. Most applications using hyperdynamics have been limited to roughly a thousand atoms. For large systems it becomes increasingly challenging to generate a bias potential that results in a raised potential within the well

without affecting the transition states [126].

Adaptive sampling incrementally gathers information from simulations and adjusts subsequent simulations to improve the exploration of conformational space [127]. The starting points of new simulations may be selected from conformations of the most under-sampled conformational regions detected in all previous simulations.

These enhanced sampling methods make use of some of the same techniques as NMMC, while maintaining better accuracy, however their computational cost is still orders of magnitude higher than that of dimensionality-reduced Monte Carlo techniques and the collective propagation of substructures is challenging due to the kinetic character inherent to MD methods.

Furthermore, molecular dynamics simulations generate an overwhelming amount of information contained in the trajectory of atomic coordinates. The viewing of such a trajectory on a graphics screen reveals the tremendous complexity of protein motion, but little else. In order to reveal the concerted fluctuations with large amplitudes, a principal component analysis (PCA) can be carried out on a large number of configurations chosen from the MD trajectory.

Faster dynamics (tier-1 and tier-2) are better suited to be studied by MD simulations. A benefit of MD simulations is that correlated motions can be identified, information that is usually lost in experiments looking at ensembles of molecules. MD simulations can elucidate the molecular nature of certain movements as the underlying forces and respective energies can be easily extracted from the simulation.

Long MD simulations of lysozyme have revealed that there are essentially two configurational subspaces [128]: 1. an 'essential' subspace that contains a small set of degrees of freedom, which describe the majority of the structural fluctuations (essential degrees of freedom); and 2. the remaining space in which movements exhibit a narrow range of structural changes and that can be described as 'physically constrained'. If the overall translation and rotation is accounted for, these two subspaces can be described by a linear transformation in Cartesian space, while being maintained for hundreds of picoseconds. The essential degrees of freedom appear to be able to capture the most relevant

motions for protein function, while the degrees of freedom in the physically constrained subspace capture redundant local fluctuations. This method allows for the the separation of equations of motion and enables the independent simulation and analysis in the essential subspace, thereby reducing computational cost and run-time.

While the characterisation of motions along a single essential dynamics analysis eigenvector is easier to understand than large trajectory files, there are many proteins for which approaches can be used that simplify protein dynamics even further. For example, the collective motions of many proteins can be approximated by the movement of quasi-rigid bodies. There are multiple tools that use two different conformations of a protein and describe the conformational change rigid body motions. There are two main strategies for solving this problem. The first identifies rigid substructures within the system using a multiple least-squares fitting procedure (HingeFind) [129]. The second groups residues into rigid bodies by comparing rotation vectors, interdomain screw axes and interdomain bending regions (DynDom) [130]. The conformational change described by each eigenvector can be visualised by a single structure, with arrows that represent hinge axes indicating the relative movement of the substructures [131].

A few years ago there has been an increase in the number of studies that used elastic network models (ENMs) and normal mode analysis (NMA) to investigate the structure-encoded dynamics of different protein systems (8, 17). Several easily accessible web-servers are now available to perform these types of simulations that enable the visualisation of possible collective motions ([132–135].

Partly due to computational methods it is now established that structural dynamics are an integral part to protein function.

Ultimately, the most successful methods are a combination of novel and rigorous mathematical ideas combined with biophysical intuition. The disconnect between experimental and theoretical timeframes is slowly narrowing, leading to renewed interest in computational modelling.

Examples of studies identifying collective motions An example of a large-scale collective motion is the oscillating hinging motion around the active site of lysozyme [136].

The components responsible for this motion can be clearly divided into the upper and lower parts of the protein pinching the substrate, resulting in peptidoglycan cleavage. Without substrate, this motion is constantly oscillating. This functional motion was first predicted almost three decades by normal mode analysis performed by Brooks and Karplus [137]. The calculated modes are in accordance with experimental results with respect to the root mean-squared displacements, and the predicted timescales for these motions range between picoseconds and nanoseconds.

Another classical example of a protein with a functional collective motion is hemoglobin. Its transition from the T to the R state has been shown in atom- [138] and ANM-based [139] NMAs to occur along a single slow mode. The transition has been found to be cooperative, meaning that the hemoglobin subunits undergo a conformational change in a concerted fashion, the mechanism of which is defined by the overall quaternary structure.

A further example is the bacterial protein-folding machine, GroEL, an ATP-regulated chaperonin. The structural movements that underlay its closed/open transition involve an allosteric cycle that is driven mainly by a couple of low-frequency modes [140]. It was demonstrated that these modes are insensitive to sequence variations [141].

Terahertz (THz) time domain spectroscopy in combination with MD simulations have been used to identify concerted motions that could explain the observed high physiological on-rates and affinities in the heme protein, cytochrome c ([142]).

NMR studies have shown that large-scale conformational changes are related to the activation of a bacterial signalling protein [143], and that the movement of flaps or entire domains regulates the access of substrates to the active sites of enzymes [144, 145]. Conformational switching has furthermore been identified in membrane proteins using site-directed spin labelling (SDSL) [146]. For example the helix-tilting motion that triggers rhodopsin activation [147], a pH-gated structural change in the KcsA bacterial K+ channel [148] and the local unfolding of the bacterial transporter BtuB after vitamin B12 binding [149]. The timescale of these conformational switching events is generally in the micro- to millisecond range.

The growing number of high-resolution protein structures, combined with the easy

access of various computational methods and computational resources has led to a body of computational studies of protein dynamics. However, large approximations still have to be made to access long-time scales and large motions. Given the stringent requirements for accurate energetic calculations of biological system, experimental validation is therefore still required but computational methods are a valuable tool for making predictions and generating new hypotheses.

2.2.2.3 Identification of collective motions

The success of NMMC largely depends on the quality of the initial decomposition of the structure into segments and the conformational space that can be sampled as a result. Thus, the main challenge is to choose the segments to reflect the collective motion of the molecule as closely as possible. This is where customised Natural Move Monte Carlo (cNMMC) and other methods such as NMA could work in a complementary fashion. Low-frequency modes calculated by NMA are often used to approximate collective functional motions in biomolecules [17]. In most cases, all the important modes are contained in the normal mode basis set. However, it is often unclear which modes are functionally relevant as the normal mode basis set contains a range of possible candidates [150]. Thus, NMA provides valuable information on the collective motion of biological systems that may guide the design of Natural Moves. Conversely, cNMMC simulations with different sets of Natural Moves that represent unique low-frequency modes may be used to identify functionally relevant modes. Most importantly, while NMA scales $O(N^3)$ with N particles in the system, NMMC scales linearly ($O(n)$) with n chain closure particles, making it more suitable for large systems.

Another method that provides information on functional motion is essential dynamics coarse graining (ED-CG). These essential dynamics are identified by principal component analysis (PCA) of MD simulations [14] or an elastic network model (ENM) of a single atomic structure [151].

Information on collective motions derived from NMR data [152] may also be used to generate hypotheses regarding functional motions that may be further evaluated with

cNMMC decomposition. A range of methods already exist that use NMR data to complement MD simulations [153].

2.2.3 The biological systems

A large part of this thesis focuses on two biological systems: The Major Histocompatibility Complex (class I and II) and the Dickerson-Drew Dodecamer as well as related dodecamers. In the following paragraphs I provide a short introduction to each biological system that are presented in this thesis as well as relevant background information regarding DNA and RNA epigenetics.

2.2.3.1 The Major Histocompatibility Complex

The cell-mediated adaptive immune response is regulated by the two classes of major histocompatibility complexes (MHC), which were first investigated for their role in graft rejection and tissue compatibility.

MHC class I (MHCI) and II (MHCII) complexes both present peptides at the cell surface; MHCI to CD8+ cytotoxic T-cells and CD4+ to helper T-cells. However, the sources and pathways for these peptides differ, with MHCI mostly binding peptides derived from intracellular and MHCII peptides from exogenous proteins [154].

MHC class I molecules are built from two non-covalently linked polypeptide chains: the α chain and β2-microglobulin (b2m). The α chain consists of three domains. The α1 and α2 domains together form the binding groove for peptides of 8-10 amino acids in length (Fig 2.7). The α3 domain spans the membrane and interacts with the CD8 receptor of T-cells, acting as an anchor, while peptide recognition takes place. MHCI molecules exert their function on the cell surface of all nucleated cells in the body. They present non-self peptides from within the cell to cytotoxic T-cells. This leads to the activation of an immediate immune response against the antigen presented by the MHCI molecule. This pathway is often referred to as the endogenous or cytosolic pathway [155].

Class II Major Histocompatibility Complexes (MHCII) are transmembrane proteins expressed by Antigen Presenting Cells (APCs) that are critical for the activation of the

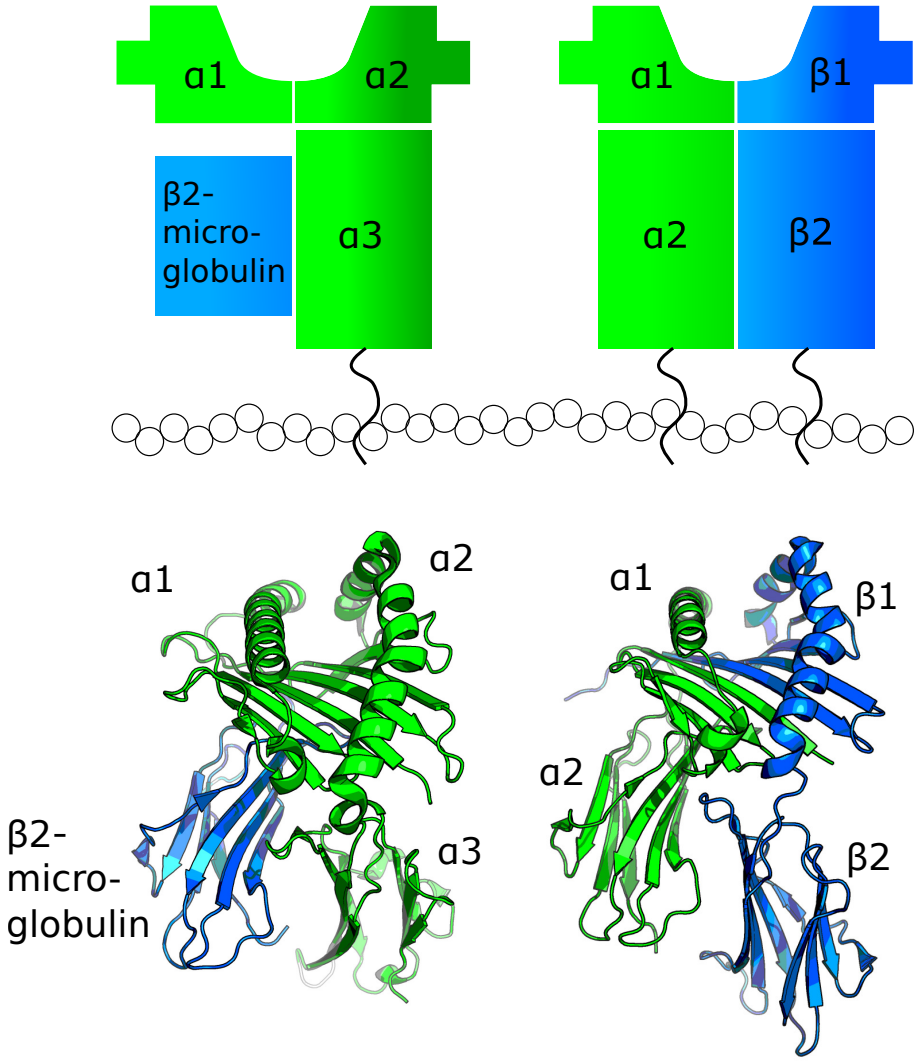

Figure 2.7: The structures of MHCI and MHCII side by side. The binding groove as well as one of the globular domains of MHCI is made up from a single chain (chain α), with the other chain (β2-microglobulin) forming the second globular domain. Only the αchain is tethered to the membrane. In MHCII the binding groove is made up from two chains (α and β), with each also forming a globular domain tethered to the membrane.

adaptive immune response in vertebrates [156, 157]. Peptides derived from extracellular proteins bind the MHCII binding groove inside the cell, are transported to the surface and are recognised by receptors on the surface of CD4+ helper T-cells [154]. While several MHCII crystal structures with high structural similarity have been solved in the presence of peptide [158], the MHCII structure devoid of peptide has not been solved to date [20].

In the absence of peptide the MHC class II binding groove can take on kinetically distinct forms that are either receptive or averse to peptide binding [159]. The receptive state mainly exists straight after peptide dissociation and has a half-life of a few minutes

after which the MHCII takes on a peptide averse state [160–163]. Structural changes in the binding groove have been implicated in this process [164, 165]. In this case study we demonstrate how customised Natural Moves may be used to investigate the plasticity of the empty MHCII binding groove. Here we follow the general steps introduced in the Methods section.

2.2.3.2 HLA-DM: An MHCII peptide editing chaperone

The diversity of MHCII molecules in any given individual is limited. Nevertheless, a large number of peptides derived from various pathogenic organisms need to be presented to the immune system via the MHCII binding groove [166].

It is helpful to view the MHCII-restricted presentation of peptides to CD4+ T cells as the result of a selection process that occurs within the cell. On their own, MHC class II molecules tend to aggregate when the binding groove is unoccupied by peptide [166]. Mutant B cell lines exhibiting a severe antigen presentation defect first led researchers to realise the importance of HLA-DM as a regulatory molecule and promoter of MHCII-peptide association [167].

During their maturation MHCII complexes are shipped, in complex with a chaperone protein called invariant chain (Ii) [168], via the endoplasmic reticulum (ER) through the Golgi to a endosomal MHCII compartment (MIIC). During transport Ii stabilises the MHCII heterodimer and prevents other peptides in the ER from attaching to the binding groove. Once in the MIIC, Ii is proteolytically cleaved leaving only a short peptide called CLIP in the MHCII binding groove. HLA-DM then binds the complex and releases CLIP, which enables the binding of peptides derived from engulfed proteins to the MHCII binding groove. HLA-DM is similarly responsible for promoting the release of weakly binding peptides, thereby facilitating the selection of kinetically stable peptide-MHCII complexes (pMHCII) [169]. For example, in the absence of HLA-DM the influenza hemaglutinin (HA) peptide (306-318) binds HLA-DR1, an MHCII serotype, with a half-life of one month. In complex with HLA-DM the half-life amounts to no more than two minutes [170].

The structure of HLA-DM is highly similar to MHCII molecules, however it is unable to bind peptides [171]. It laterally binds to MHCII and forms a stable complex following peptide dissociation. When the MHCII binding groove is occupied, however, the DM-MHCII complex dissociates [172]. Thus, HLA-DM causes CLIP removal and then stabilises empty MHCII proteins until a peptide is available. The mechanism of binding-groove by HLA-DM is still unknown. In chapter 7 I present an investigation of this mechanism using customised Natural Moves.

2.2.3.3 The Dickerson-Drew Dodecamer

Much of our knowledge regarding B-form DNA is derived from structural investigations of the DNA oligonucleotide CGCGAATTCGCG, named Dickerson-Drew dodecamer (DDD) after the researchers who first solved its structure [173]. This sequence was of interest because it contains an EcoRI restriction site, G-A-A-T-T-C, and a strong tendency to adapt a classical Watson-Crick B-form helix throughout.

Many discoveries have been made using this structure and those of related dodecamers such as the interdependence of base sequence and structure [174], backbone flexibility [175], solvation [176], bending and bendability [175] and the effects of crystallisation conditions [177] on the conformation of DNA. Interesting characteristics of the DDD double helix are the narrow minor groove in proximity of the AATT subsequence and the structured nature of the water molecules found in that groove.

Several computational studies of the DDD have been published, including the MD and continuum solvent simulations reported [178, 179] that confirmed the experimentally determined conformations and yielded insight into the molecular nature behind the relative stabilities of A- and B-helices.

The large amount of experimental data about its structure and dynamics makes the DDD a suitable benchmark system for testing force fields and protocols. In chapters 6 and 8 we investigate the effects of epigenetic marks on the structure of the DDD.

2.2.3.4 DNA and RNA Epigenetics

Epigenetics is the molecular process that regulates heritable changes in gene expression by mechanisms other than DNA sequence. Epigenetic mechanisms include chromatin remodelling through covalent modifications of 1. DNA bases or 2. histone proteins and 3. regulation of mRNA expression through non-coding RNAs (ncRNA). Given a certain genome, the epigenome is ultimately responsible for determining the gene expression patterns in a cell. This realisation in addition to its reversible nature has made epigenetic mechanisms a popular target for therapeutics and diagnostic biomarkers.

The interplay of many different epigenetic mechanisms has been found to regulate genes in a context and site-specific manner (Fig 2.8)[180].

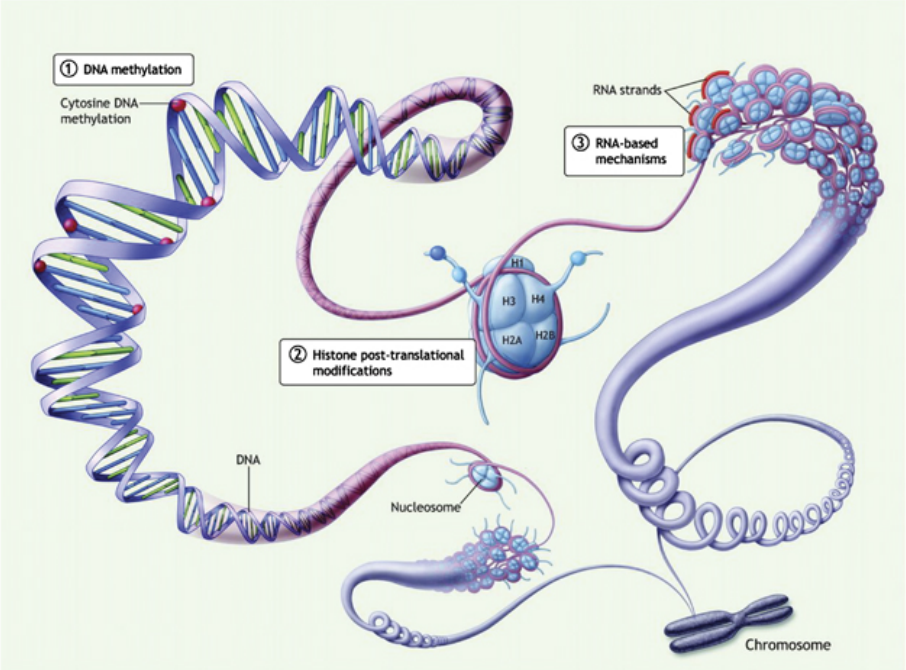

Figure 2.8: The three main mechanisms by which epigenetics regulates gene expression. 1) DNA modifications 2) Histone modifications 3) RNA-based mechanisms. Adopted from [180].

DNA Modifications The effects of DNA modifications, most prominently methylation, on gene expression were first described in 1975 by two independently working groups. They were investigating the mechanism of a molecular switch that appeared to silence genes during development. Methylation of DNA promoters is now known to be a key

factor for silencing genes within differentiated somatic cells, which is key to preventing the deleterious effect caused by the reactivation of early developmental genes [181]. DNA methylation occurs on cytosine bases located in CpG dinucleotides, predominantly in CpG repeats (where 'p' denotes the phosphodiester bond). While our understanding of the effects of DNA methylation has improved rapidly over recent years, the molecular mechanisms by which DNA modifications act is still largely unknown. In the last few years, additional DNA modifications started to emerge, when the catalytic activity of Ten-eleven translocation (Tet) proteins was discovered [182]. 5-hydroxymthelycytosine (5hmC) was first observed in mammalian DNA in the early 1970s [183], but a first biological role was only described in 2009 [184]. Tet proteins generate 5hmC from existing 5mC, which is subsequently oxidised to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) (Fig 2.9). 5fC and 5caC used to be mainly interpreted as side products of a demethylation pathway, however recent findings suggest that they may also play a functional role. By extending the chemical diversity of the nucleotide library they enable the formation of structural forms other than the well known forms B-DNA, Z-DNA and A-DNA. For instance, Raiber et al. have published the structure of a self-complementary 5fC-containing dodecamer that exhibits a new structural form, called F-DNA. An investigation of this system is presented in chapter 8.

Chromatin Remodelling A further site of epigenetic regulation is chromatin, a functional scaffold of DNA and histone proteins that enables the dense packaging of chromosomes. The post-translational modification of histones may change its interaction with other proteins and affect the binding of the transcriptional machinery, thus providing control over gene expression at this location. Over 130 modifications of histones are known in mammalian cells, with methylation and acetylation being the most studied [185].

Non-coding RNAs The vast majority of the genome ($> 90\%$) is transcribed into non-coding RNA (ncRNA), also known as genomic 'dark matter' [186]. ncRNAs are a diverse collection of RNAs that are grouped into two classes according to their size: small and long ncRNA (lncRNA), with up to 200 nucleotides and 200 to around 100 kb, respectively.

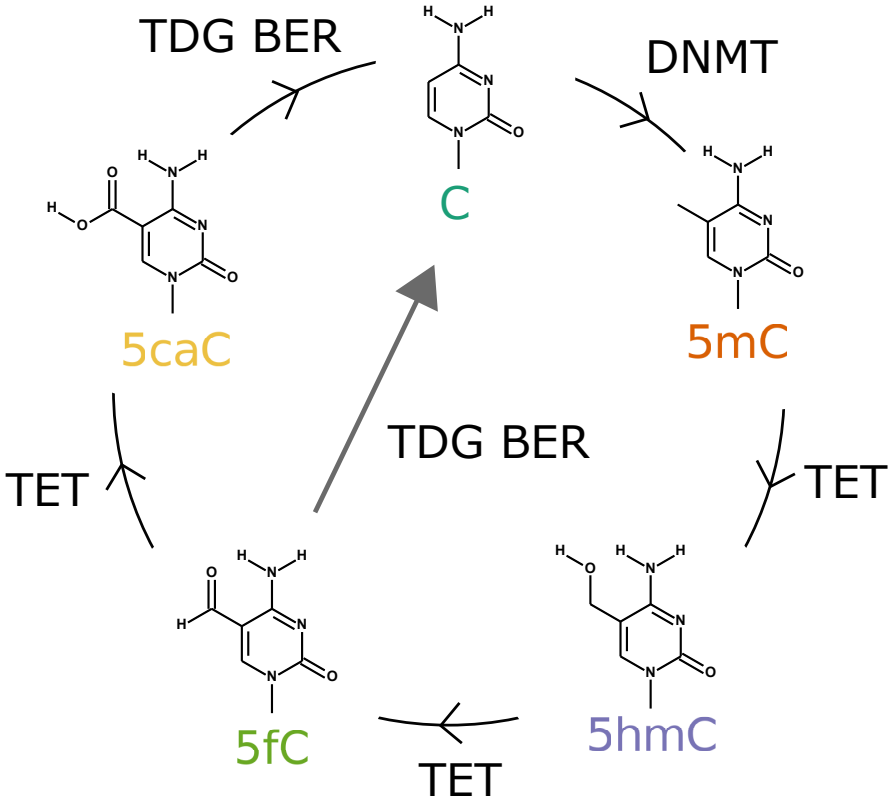

Figure 2.9: Methylation and sequential oxidation of cytosine to 5-methylcytosine (5mC) by DNA methyltransferase (DNMT), 5-hydroxymethylcytosine (5hmC) by Ten-eleven Translocation (Tet), 5-formylcytosine (5fC) by Tet and 5-carboxylcytosine (5caC) by Tet. 5caC and 5fC can be converted back to cytosine by thymine DNA glycosylase (TDG)-dependent base excision repair (BER).

The most commonly discussed small ncRNAs are microRNAs (miRNAs), which bind their target mRNA transcripts and label them for inactivation by the RNA-induced silencing complex. This mechanism is responsible for regulating a number of processes such as gene silencing and transcription, DNA imprinting and methylation as well as chromatin remodelling [187]. In contrast, lncRNAs form secondary structure and interact with other proteins; the molecular nature of these events is poorly understood but a number of roles have been identified, including cell cycle control, trafficking, transcription, translation and cell differentiation [188].

Similarly to DNA, a number of modifications in RNA have been identified [189]. One of the most common modifications is methylation, which occurs on either nitrogen or oxygen at the post-transcriptional level. With over 80% of all RNA methylations N6-methyl adenosine (m6A) is the most prevalent nucleoside in eukaryotic mRNA [190].

Recently, a N6-methyladenosine-dependent (m6A) structural switch was found to regulate an lncRNA-protein interaction [191].

A further epigenetic marker in RNA is 5-methylcytosine (5mC). Compared to DNA, m5C and its oxidative products 5-hydroxymethyl-, 5-formyl- and 5-carboxyl-cytosines (5hmC, 5fC and 5caC, respectively) have been little studied. However, a crystal structure of a 5-formylcytosine RNA duplex has recently been published [192]. It was shown that this modification promoted the base stacking between the 5fC and the adjacent nucleotides but did not affect the overall structure of the duplex.

Chapter 3

Methods

3.1 Summary

In this chapter I provide a more detailed description of the methods used in this thesis, including various aspects of NMMC and a full description of the customised Natural Moves (cNMMC) protocol that I developed during my thesis. Other methods presented here include the NMA approach used in one of the chapters (chapter 7), MODELLER (used in chapter 5) as well as x3DNA, a popular structural analysis tool that I used to characterise conformational changes in epigenetically modified DNA molecules (chapter 8).

3.2 Natural Move Monte Carlo

Natural move Monte Carlo (NMMC) aims to sample the conformational space along user defined independent degrees of freedom X_i. Given this initial choice, the method generates canonical distributions along X_i over an effective energy surface $\tilde{E}$, which is defined by equation 3.1 below. Since the proposal kernel [18] along X_i is symmetrical, we use classical Metropolis Monte Carlo [193], which satisfies detailed balance, to sample the different states of X_i. Numerical experiments [19] demonstrate the accuracy (convergence to limiting distributions) and effectiveness (rate of convergence) of this approach.

In NMMC all degrees of freedom, X are partitioned into independent (X_i) and de-

pendent (X_d) degrees of freedom (DOF). For example, X_i represent the independent orientational, translational or internal motions of structural fragments in a molecular chain, whereas X_d are the DOFs that are instantaneously minimised to facilitate exploration along X_i and preserve the integrity of the molecular chain(s) through chain closure(s). Thus, the effective potential over X_i is defined as

$$\tilde{E}(X_i) = \underset{X_d}{min}\Big\{E(X_i \cup X_d)\Big\}. \tag{3.1}$$

Therefore, Natural Move Monte Carlo is analogous to Metropolis sampling [193], exploring state space spanned by X_i over the energy surface $\tilde{E}$. The most unique feature of NMMC is how the complex moves are generated. This is described below.

3.2.1 Implementation

The basic principle is that each new configuration during a proposal step is obtained via a combined chain breakage closure algorithm. This composite proposal kernel includes a stochastic proposal to update X_i followed by finding the most optimal (with respect to the new X_i) arrangement along X_d. This scheme can accelerate the conformational search for possible arrangements of *a priori* defined structural segments or regions (e.g. groups of segments) and is also free of any limitations caused by the lever-arm effects of distant torsional changes, which leads to increasingly (by chain length) low acceptance rates of dihedral moves. Thus, NMMC can be applied for any system regardless of size.

While the above description is general, the exact definition of independent X_i and dependent X_d degrees of freedom should be custom tailored to the model of interest. For the coarse-grained protein model of Case 1 and for the all-atom DNA model of Case 2 X_i and X_d are described in detail in [18].

3.2.2 Numerical experiments

In MCMC simulations it is generally thought that an acceptance rate of ~0.4 is optimal when a single parameter (one independent variable of X_i) is updated and ~0.2 when a group of parameters (all independent variables in X_i) are updated. Given that we update X_i based on a multivariate normal distribution [18], we consider acceptance rates for Natural Moves between 0.2 and 0.3 as optimal.

If NMMC is used in combination with parallel tempering (replica exchange) then we consider acceptance rates for adjacent temperature replica exchange of ~0.2 as optimal and rates in the interval [0.1, 0.3] as acceptable. The choice for these rates are based on considerations such as the sufficient relaxation time of individual Markov chains and the probability of 'coast to coast' visits of individual replicas.

3.2.3 The stochastic chain closure algorithm

Torsional move sampling of macromolecules with Markov Chain Monte Carlo algorithms used to suffer from low acceptance rates. This was largely due to the large displacement of atoms distant from the point of rotation due to the lever-arm effect. To avoid this problem a new approach was introduced where the global structure was kept in place while the relevant torsional moves were performed locally; the most popular of these methods was called CONROT[25].

Once the local torsional move was applied, the resulting break in the chain needed to be closed. This step commonly presented a significant bottle neck as the chain closure algorithm was computationally expensive. Improvements to these algorithms were made, however they were still limited by size of torsional moves as well as the computational cost.

In 2010 Minary and Levitt introduced a novel stochastic chain closure algorithm [18] that was able to close large chain breaks at a fraction of the computational cost of previous methods. Furthermore, the algorithm was capable of running with any sampling or optimisation protocol of choice.

As this algorithm plays a key part in Natural Move Monte Carlo (NMMC) [22], the main method used throughout the studies presented in this thesis, I will describe the algorithm below.

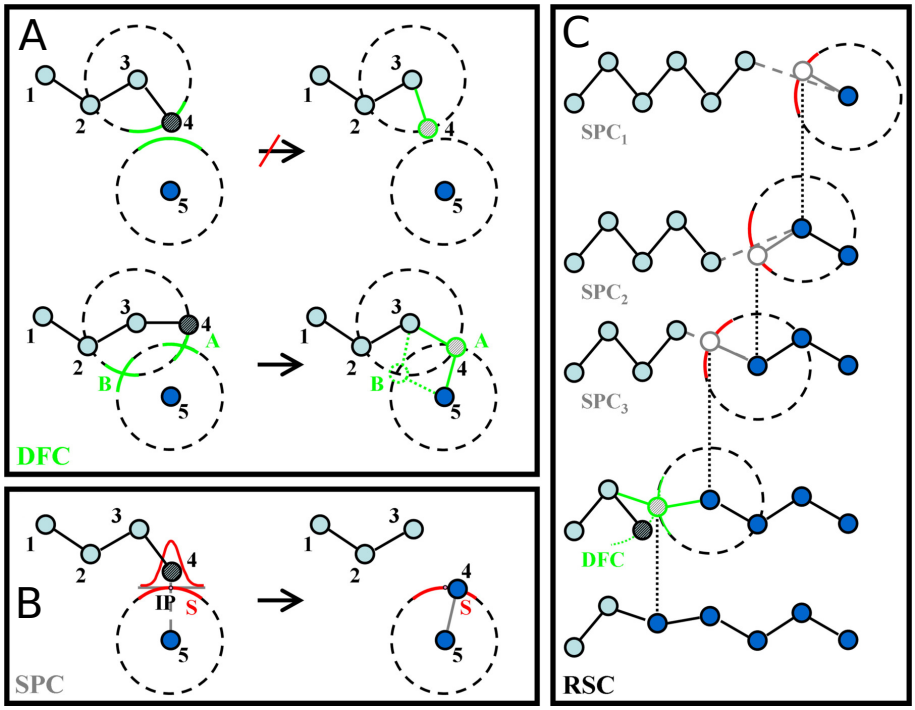

Figure 3.1: A) Schematic drawing of Deterministic Full Closure (DFC). The bond between atoms 4 and 5 in the 5-atom chain r1,..,r5 is broken. Depending on the position of atom 4 in relation to atom 5 there either is a solution or there is no solution. B) The same problem when using Stochastic Partial Closure (SPC) always has a solution. SPC always positions atom 5 on arc S according to a normal distribution centred around the intersection point of the arc and the line connect atoms 4 and 5. C) A cycle of Recursive Stochastic Closure (RSC) includes multiple iterations of SPC (dependent on the length of the molten zone, in this case three) and a final sealing stage of DFC. The right-most atom is the head atom, which is on the other side of the break; the two left-most atoms belong to the anchor. The other atoms are molten zone. Figure adopted from [18].

The algorithm

The recursive stochastic closure algorithm (RSC) introduced by Minary and Levitt [18] is split into two parts: 1. stochastic partial closure (SPC) [18] and 2. deterministic full closure (DFC) [29]. **DFC** works as follows: Given an anchor and a head atom, DFC tries to close the break by placing a connecting atom at the intersection of two spheres with a certain radius (usually the bond length) centred around the head and anchor atoms. If the two atoms are too far apart the step gets rejected as seen in Figure 3.1A. **SPC**

in contrast always gives a solution: SPC defines a) a line between the anchor and the head atoms b) an intersection point (IP) between a sphere centred around the head atom and the line c) a plane tangential to a surface centred around IP and d) a random point on the tangential surface calculated by a two-dimensional normal distribution (for details see [18]). Thus, SPC on its own does not seal the break but requires a final DFC step to complete the cycle. It serves as a 'gap reduction tool' to increase acceptance rates (Figure 3.1B). The **RSC** algorithm includes $n-1$ SPC steps, where n is the number of backbone atoms in the molten zone, and one DFC step at the end to seal the break (Figure 3.1C). The molten zones differ between systems (e.g. nucleic acids or proteins) and representations (e.g. all-atom or 3-point (3pt)) and are defined by the user.

3.2.4 The models

3.2.4.1 Nucleic Acids - Physics based

Force fields are constantly undergoing development and much research is spent on improving their performance. Modern force fields are generally developed by fitting parameters to high-level quantum mechanics data and are tested against further quantum calculations, simulations as well as experimental data. As such, they can only approximate quantum mechanical properties, which inevitably leads to inaccuracies during long simulations; consequently these force fields are updated regularly. The potential used for nucleic acids throughout this thesis is the Amber parmbsc0 force field [194]. It has been adapted from parm99 [195], which itself was an iteration of the parm98 modification of AMBERâĂŹs second generation force field [196], and improved to correct for a bias in backbone angles that had been observed in long constant temperature and pressure MD simulations of B-DNA with explicit water counterions [197]. Reportedly, it significantly outperforms parm99 for simulations that are 20 times longer. A new version of this type of potential has recently been developed, however it has not been applied here. We used the Amber parmbsc0 force field with explicit solvent, i.e. including water (TIP3 model) and ion molecules (Na+ and Cl-).

3.2.4.2 Proteins - Knowledge based

A statistical or knowledge-based potential is an energy function generated from datasets of solved protein structures, for example the Protein Data Bank.

The energy function used for proteins in this thesis evaluates pairwise amino acid contacts or distances but other features may be used such as torsion angles (Ramachandran plot), solvent exposure or hydrogen bond geometry. Pairwise knowledge-based potentials may be described as a 2-dimensional interaction array that relates an energy value to all possible residue pairs in a protein (usually pairs of residues adjacent in the chain are excluded). This allows for an 'energy' calculation of a given conformation by summing up all the values of the pairwise contacts (within a certain cutoff).

We used the knowledge-based potential introduced by Summa et al. [49], that was later adapted for the 3pt model by Minary et al. [21].

The frequencies of pairwise atomic contacts were generated from the Top500 database [198] (superseded by TOP8000), a curated dataset of high quality protein structures. More than 74,000 heavy atoms were recorded and all 167 atom types used in the popular RAPDF potential [50] were included, which resulted in a total of 250 million pairwise interactions. For each type of contact, the observed distances were binned according to their length and the counts were converted to an energy value according to the method described by Lu and Skolnick [199]. Each distance-dependent pairwise curve was subsequently fitted to a quintic spline function. The force field was successfully validated against other widely known physics-based all-atom potentials [200].

For the purposes of NMMC with 3pt per residue models the potential was adapted by Minary et al. [21]. In this model each residue is represented by the C_α, carbonyl oxygen and the atom closest to the centre of mass of the side chain (note that Glycine is the exception). This representation allowed for a straightforward mapping of the all-atom energy function to the simplified model. As intramolecular bonding and bending were disregarded in the 3pt model these terms were not needed. The only internal degrees of freedom were torsion angles ($C_\alpha - O - C_\alpha - O$ and $O - C_\alpha - O - C_\alpha$), for which no separate potential was used in order to add additional flexibility.

3.2.5 Limitations

NMMC makes a number of approximations for the benefit of sampling efficiency but at the cost of accuracy. NMMC can take many forms as the simulation parameters and algorithms can vary depending on the application. In large proteins, NMMC is generally most suitable for exploring drastic conformational changes of structural modules (domains, secondary structures, subunits, groups of subunits). However, relatively small scale systems such as the MHC/peptide complex have been simulated, where substructures as small as 3-9 residues, i.e. the peptide, were propagated [31]. While, these simulations can reveal overall biophysical trends it is not designed to reveal intricate details at the inter-atomic level. For DNA systems NMMC can reveal more detail. Due to the inherent integrity and modular characteristic of DNA structure, it is more feasible to perform Natural Move Monte Carlo at atomic resolution. Therefore, inter-atomic information can be extracted from simulations. Another obvious drawback of NMMC is the lack of a kinetic component.

3.3 Customised Natural Moves

This work as published in: Demharter, S., Knapp, B., Deane, C. M., & Minary, P. (2016). Modeling Functional Motions of Biological Systems by Customized Natural Moves. *Biophysical Journal*, 111(4), 710-721. http://doi.org/10.1016/j.bpj.2016.06.028

One of the main challenges of Natural Move Monte Carlo is finding a set of degrees of freedom that describes the system accurately enough to make biologically relevant claims [13]. In this chapter I introduce a new protocol based on the customisation capabilities of NMMC, that allows researchers to construct test cases with different sets of degrees of freedom and systematically test hypotheses regarding the functional motions of biological systems.

3.3.1 Introduction

Functional motions in biomolecules are central to many biological processes [94]. Molecular simulations are often used as a tool to investigate these dynamics and interpret [201, 202] and/or refine [153] experimental data or inspire new experiments [203].

Large improvements in computational resources and algorithms have been made since the first molecular simulation of a protein in 1977 [2, 5]. Recent milestones include the 50 nanosecond molecular dynamics (MD) simulation of the full satellite tobacco mosaic virus with one million particles [8], the Folding@Home project which used over 400,000 personal computers to study challenging problems such as protein folding [9] and a study that presented millisecond simulations to study the folding pathways of small fast-folding proteins [7].

Despite these advances, the high dimensionality and complex energy surfaces still pose a challenge for simulations of large biomolecules [10, 11]. In an effort to address these limitations there have been promising developments in dimensionality reducing methods that exploit the inherent modularity and collective motions in biomolecules [12, 13]. For example essential dynamics coarse-graining (ED-CG) identifies sites that reflect the essential dynamics of an atomistic molecular dynamics trajectory [14]. Other methods based on elastic network models, principal component analysis and normal mode analysis have also been successfully used to study functional motions in biomolecules [15–17]. While these methods are not as physically accurate as MD simulations, their increased sampling efficiency makes them a valuable tool to generate new hypotheses that can be tested by experiments. One of the main challenges of these methods, however, is finding a set of degrees of freedom that describe the system accurately enough to draw biologically relevant conclusions [13]. Thus, it is of value to have computationally cheap methods that allow for the easy manipulation of degrees of freedom to test different hypotheses about the functional motions of biomolecules *in silico*.

Here we introduced a protocol based on customised Natural Moves (cNM) to address the challenge of choosing suitable degrees of freedom and to allow for the systematic investigation of hypotheses regarding functional motions in biomolecules. We used cNMs

to modulate translations and rotations of segments as well as torsion and bend angles of bonds and compare different sets of cNMMC simulations to infer causal relationships in functional motions. We used two case studies to demonstrate its application. First we investigated functional motions in the class II major histocompatibility complex (MHCII) and second we studied the structural effects of an epigenetic mark on a DNA model system.

Natural Move Monte Carlo

As described above, a molecular system can be defined as a collection of monomers or more formally a set, $\Omega = \{m_1, ..., m_N\}$, where m_i for $i = 1, ..., N$ refers to the residues of a structure. Sequentially numbered residues can be grouped into chains, $\mathbf{C} = \{C_1, ..., C_{N_c}\}$, where $C_i = \{m_{c_h^i}, m_{c_h^i+1}..., m_{c_t^i}\}$ (e.g. c_h^i and c_t^i are indices of the 'head' and 'tail' residues of chain i) for $i = 1, ..., N_c$. The chain 'concept' can be further generalised into a group of segments, $\mathbf{S} = \{S_1, ..., S_{N_s}\}$, where $S_i = \{m_{s_h^i}, m_{s_h^i+1}..., m_{s_t^i}\}$ for $i = 1, ..., N_s$ (e.g. s_h^i and s_t^i are indices of the 'head' and 'tail' residues of segment i).

After defining the segments, we can introduce the set of residues that make up the segments.

$$\Omega_s = \bigcup_{i=1}^{N_s} S_i \quad | \quad \Omega_s \subseteq \Omega \tag{3.2}$$

At each iteration in the simulation the segments are moved in a Monte Carlo fashion along user-defined degrees of freedom (DOFs), which collectively are called Natural Moves; these may include translations and rotations of segments as well as torsion and bend angles within segments. After each propagation step, the set of atoms (or entire residues) connecting two segments are rearranged by a linear complexity chain closure algorithm [18]. We will refer to this set as molten zone or MZ. This allows for the reconstruction of chain breaks that may result from the movement of the segments.

Thus, the residues outside of the segments form the set of molten zone residues

$$\Omega_m = \Omega \setminus \Omega_s = \bigcup_{k=1}^{N_{MZ}} \Omega_{MZ}^{(k)}, \tag{3.3}$$

where N_{MZ} is the number of MZs and $\Omega_{MZ}^{(k)}$ is the set of residues in the k^{th} molten zone.

Furthermore, we define $\Omega_{MZ} = \{\Omega_{MZ}^{(1)}, ..., \Omega_{MZ}^{(k)}\}$ as the set of molten zones.

In this study, $\Omega_s \subset \Omega$ (Ω_s is a proper subset of Ω), thus $\Omega_m \neq \emptyset$ for proteins. For nucleic acids we define each residue as a segment, therefore $\Omega_m = \emptyset$. In this case we use ${}^*\Omega_{MZ}$ to denote the set of molten zones ${}^*\Omega_{MZ}^{(j)}, j = 1, ..., k$, where ${}^*\Omega_{MZ}^{(j)}$ refers to a molten zone with a set of atoms that are used by the closure algorithm to connect chain breaks that may be caused by moving adjacent nucleotides independently.

Customised Natural Moves

Customised Natural Moves (cNMs) are Natural Moves that can be modified to investigate functional motions in biomolecules. cNMs include translations and rotations of segments, which may also exhibit internal flexibility such as torsion and bend angles of bonds. cNMs can be created by grouping two segments into one so that they move as a unified segment. Instead of two independent segments that are moved separately, there is now a single set of Natural Moves that describes the collective motion of both segments. This can be useful, for example, to test whether flexibility in an α-helix kink is important for a particular functional motion or to explore how different levels of collective motion may affect a structural mechanism. Customisation may in addition occur at the level of internal flexibility of segments. When internal flexibility is disabled, segments are treated as rigid bodies. When internal flexibility is activated, some torsion angles around bonds are changed along with the movement of segments. Customised Natural Moves also allow to selectively activate or deactivate sampling of torsional rotations around specific bonds.

3.3.2 The Protocol

We devised a protocol based on cNMs to investigate functional motions in biomolecular structures. The key steps of the protocol are, **Step I:** Define a hypothesis; **Step II:** Translate hypothesis into Natural Moves; **Step III:** Activate/inactivate Natural Moves to generate test cases for the investigation of the hypothesis; **Step IV:** Perform conformational sampling on each test case and evaluate the results with respect to the hypothesis. The steps are described in more detail below.

3.3.2.1 Step I: Define a hypothesis

First, a hypothesis regarding a functional motion is defined. Experimental data, observations in the literature and/or biological intuition may be used to identify candidate functional motions with characteristics such as flexibility, rigidity or collective motion. These features may include individual bonds, residues and secondary, tertiary or quaternary structure elements.

3.3.2.2 Step II: Translate hypothesis into Natural Moves

Based on this hypothesis an initial set of Natural Moves can be defined that encapsulate all movements that the researcher specifies as important for the functional motion. These might be residues and larger segments as well as the torsion and bend angles of individual bonds.

3.3.2.3 Step III: Generate Test Cases

After the initial set of Natural Moves has been defined it is then possible to generate different sets of customised Natural Moves by selectively modifying certain degrees of freedom to study their effect on functional motions. Natural Moves may be customised by modulating the degrees of freedom that describe the movement of segments (translations and rotations) as well as their internal flexibility (torsion and bend angles of bonds).

The relative movement of two segments results in a chain-break that is closed by the rearrangement of atoms in the melting zone. When two segments are grouped into a bigger segment the relative orientation of these segments to each other is maintained throughout the simulation.

Here, we consider a molten zone $\Omega_{MZ}^{(k)}$, for $k = 1, ..., N_{MZ}$ to be active when the segments on either sides are moved independently and inactive otherwise.

Formally we can introduce a function $f : \Omega_{MZ} \rightarrow \{0, 1\}$, $f(\Omega_{MZ}^{(k)}) = 1$ if $\Omega_{MZ}^{(k)}$ is enabled, which leaves corresponding S_j and S_{j+1} as independent segments and $f(\Omega_{MZ}^{(k)}) = 0$ if $\Omega_{MZ}^{(k)}$ is disabled, which fuses the two adjacent segments into one, e.g. $S_{(j,j+1)}$.

Similarly, some segments may have internal flexibility such as bond torsion angles.

Therefore, it is possible to introduce a set of torsion angles $\Omega_{\phi}^{(l)}$, for $k = 1, ..., N_{\phi}$ and define a function $g : \Omega_{\phi} \rightarrow \{0, 1\}$ where $g(\Omega_{\phi}^{(l)}) = 1$ if $\Omega_{\phi}^{(l)}$ is active and $g(\Omega_{\phi}^{(l)}) = 0$ if $\Omega_{\phi}^{(l)}$ is inactive.

The decomposition and the internal flexibility of segments in a structure may be represented by a vector $\mathbb{D}$ in which each element refers to the state of a specific molten zone or torsion angle. $\mathbb{D}$ may be defined as:

$$\mathbb{D} = \{f(\Omega_{MZ}^{(1)}), ..., f(\Omega_{MZ}^{(N_{MZ})}), \\ g(\Omega_{\phi}^{(1)}), ..., g(\Omega_{\phi}^{(N_{\phi})})\} \tag{3.4}$$

where Ω_{MZ} and Ω_{ϕ} refer to the respective molten zones and torsion angles and $|\mathbb{D}| = N_{MZ} + N_{\phi}$.

Thus, each $\mathbb{D}(f(k), g(l))$, where $f(k) = f(\Omega_{MZ}^{(k)}$ and $g(l) = f(\Omega_{\phi}^{(l)}$ leads to a different decomposition, which we refer to as test cases that are denoted as ${}^{\mathbb{D}(f(k),g(l))}T$.

Test cases are generated by a set of functions $f_i, i = 1, ..., N_T$ where N_T is the number of test cases. If we have three molten zones then there are $2^3 = 8$ different test cases that result from 8 functions with identical domains $\{\Omega_{MZ}^{(1)}, \Omega_{MZ}^{(2)}, \Omega_{MZ}^{(3)}\}$ and co-domains $\{0, 1\}$ but unique functional maps. Thus, $f_i : \{\Omega_{MZ}^{(1)}...\} \rightarrow \{0, 1\}$, for $i = 1, ...N_T$.

A particular test case may allow for flexibility in an α-helix kink while another test case treats the helix as rigid. Similarly, a selected bond may be sampled freely in one test case, while in another test case the bond angle is maintained throughout the simulation. This capability allows the researcher to investigate causal relationships between structural features and biophysical mechanisms.

3.3.2.4 Step IV: Conformational sampling and evaluation

Each test case implies a unique set of degrees of freedom (customised Natural Moves) that can be sampled with NMMC. The resulting distributions can then be evaluated with respect to the initial hypothesis. Below we outline the method details for both of our case studies. Note that for reproducible results replica simulations are needed [20, 31, 204].

Comparing results from different test cases

Each test case of the protocol is an independent model, in which the available conformational space is a subspace of C_f, the domain that includes all functionally relevant conformational variability. C_f is usually chosen to be a proper subspace of the 'full domain', C; $C_f \subset C$, e.g. C could be spanned by the Cartesian degrees of freedom (DOFs) and C_f by dihedral angles about single bonds and bond angles between bonds that an atom forms.

Each test case features some restricted set of DOFs spanning the state space $C_i \subseteq C_f \subset C$ for all $i = 1, \hat{a}\check{A}\c{e}, N_T$, where N_T is the number of test cases. Note, that the full domain, C is equipped with an energy function (the original energy surface), $E : C \to \mathbb{R}$ and the energy surface for a given test case is given by the function, E_i, which is a restriction of E to C_i and defined as $E_i : C_i \to \mathbb{R}, E_i(x) = E(x)$, for all $x \in C_i$. Thus, each test case is an independent model featured by C_i and the corresponding energy surface, E_i.

In spite of each test case being associated with its own state space, C_i distributions (over structural observables) obtained for different test cases can be compared to assess the contribution of a particular DOF (e.g. the relative motion of two adjacent helices enabled by a central kink) to functional motions (e.g. changes in MHC-II binding groove area and width). For example, let $\alpha : C \to \mathbb{R}$ be a structural observable and let $P_i(\alpha), i = 1, .., N_T$ be the normalised numerical distributions over α we obtain for each test case via performing independent Natural Move Monte Carlo simulations covering each state space, $C_i, i = 1, .., N_T$.

For the protein study discussed in chapter 7 we assess some features (e.g. bimodal) of these distributions $P_i(\alpha)$ to identify the DOFs that are essential and ones that are less critical to produce that feature, which may be linked to important biological function. For example, if the binding-groove width distribution is bimodal then the MHC binding groove can exhibit two stable conformations (open and closed) even in the absence of the peptide. By systematically grouping all $P_i(\alpha)$ that exhibit this behaviour from those that do not, we can identify the underpinning essential DOFs responsible for this phenomenon. In a similarly qualitative but systematic approach the DNA study in chapter 8 compares

distributions for test cases to purely identify the existence and directionality of effects a chemical modification imposes on the DNA structural parameters. Our robust initial search can identify test cases or phenomena that could be further investigated by molecular dynamics to obtain refined quantitative information.

3.3.3 Discussion

cNMMC is a protocol for the testing of hypotheses regarding the functional motions in biological systems. It is based on the Natural Move Monte Carlo method that allows for the sampling of conformations given a structural decomposition defined by the researcher.

The use of both cNMMC and NMMC assume the decomposition of the molecular system into segments and molten zones. The implementation methodology [18, 19] of NMMC follows a 'segment centric' approach; if adjacent segments move with respect to each other their translational and orientational updates are independent; otherwise a larger segment including the adjacent segments is defined.

In cNMMC end users may consider each molten zone as 'active/inactive' or '1/0' so that adjacent segments may move independently or synchronously. Using this 'molten zone' centric approach each set of DOFs (each test case) is associated with a binary string so that test cases can be easily organised and annotated in a systematic and high-throughput manner. In this way cNMMC reduces the technical barrier to the use of the NMMC approach [18–20, 22] to study the 'anatomy' of necessary and sufficient sets of degrees of freedom responsible for molecular function.

The efficient chain closure algorithm [18] allows the user to introduce arbitrary degrees of freedom into a system without substantially compromising computational run-time.

We used this customisation capability as the basis for a protocol for the investigation of structural mechanisms. The protocol allows for an investigative strategy using a range of simulations with distinct sets of customised Natural Moves to test hypotheses concerning the functional motions in biological systems.

In molecular biology, a classical approach to testing hypotheses regarding the function of a certain gene is to interfere with its expression and see what happens to the organism.

Similarly, in experimental structural biology, residues can be mutated or removed to identify functional regions such as protein binding or enzyme active sites [205].

In this protocol we apply this concept of reverse engineering to the investigation of functional motions in simulations. Our customised Natural Moves protocol enables the testing of hypotheses regarding the functional motions of a biological system by allowing the user to enhance or restrict the movement of certain structural regions.

Hypotheses may be derived from biological intuition or computational and/or experimental methods such as essential dynamics coarse graining (ED-CG) [14, 151], principal component analysis (PCA) [206], elastic network models (ENM) [207], Normal Mode Analysis (NMA) [17] and Nuclear Magnetic Resonance (NMR) [152].

ED-CG provides information on essential motion by PCA of MD simulations [14] or an elastic network model (ENM) of a single atomic structure [151]. Similarly, low-frequency modes calculated by NMA are often used to approximate collective functional motions in biomolecules [17]. In most cases, all the important modes are contained in the normal mode basis set. Thus, NMA provides valuable information on the collective motion of biological systems that may guide the design of Natural Moves. However, it is often unclear which modes are functionally relevant as the normal mode basis set contains a range of possible candidates [150]. cNMMC simulations may be used to identify functionally relevant modes with different sets of Natural Moves that represent unique low-frequency modes.

Furthermore, NMA or PCA can be costly as the computational complexity associated with these methods is $O(N^3)$ (worst case), where N is the number of atoms in the system. This is due to solving the underlying eigenvalue problem associated with the Hessian matrix. Advanced solvers might produce better scaling, e.g. $O(N^c)$ with $c < 3$ but even using these advanced methods the computational cost associated with NMA or PCA will dominate $O(N)$, which is the time complexity of modern algorithms [208, 209] for calculating most statistical [21] or empirical [194] force fields. In contrast, the time complexity of NMMC is strictly linear ($O(N)$) because our chain closure algorithm [21] has linear complexity, with respect to the number of degrees of freedom used to solve the

chain closure problem. The application of Natural Moves, unlike the calculation of NMA or PCA, will never dominate the computational cost of molecular simulations. This is the main quantifiable advantage of using cNMMC instead of NMA or PCA.

A less quantifiable but still notable advantage of cNMMC compared to NMA (or PCA) is that it allows the use of highly unconventional experimentally inferred DOFs such as the hand shaking motion of adjacent subunits in a chaperonin [22]. These 'experimentally derived' moves are not necessarily associated with or dependent on a single conformation (NMA) or conformational ensembles (PCA). They can be simply defined without any limitation to test any experimental observation or intuition. Therefore, NMMC not only supports moves derived using PCA or NMA but any type of moves (e.g. move any part of the system and the rest will deform to follow the change).

The scope of NMMC also differs from the scope of NMA or PCA. NMA (PCA) takes a minimum energy conformation (conformational ensemble) as its input and outputs collective motions or deformations of the molecular system. On the contrary NMMC takes any collective motion (including but not limited to the ones derived from PCA or NMA) as input and provides distributions as output by exploring the relevant conformational space orders of magnitude more efficiently [21] than conventional methods such as Cartesian Monte Carlo or molecular dynamics. In this capacity NMMC has linear $O(N)$ scaling, so it is perfectly fitted to high-throughput testing of customised Natural Moves. NMA based Monte Carlo would require the successive recalculation of normal modes in concert with the changing molecular conformation and the computational cost would scale as $O(N^3)$ (worst case).

Thus, cNMMC should be considered as a complementary approach to NMA or PCA. For example, NMA or PCA can be used in the construction of Natural Moves and cost efficient NMMC can explore the conformational space. cNMMC can also be used to test the validity of low frequency normal mode based Natural Moves while exploring the conformational space distant from the minimum energy conformation used to generate the normal mode.

Similarly to the above discussion on how cNMMC differs from NMA or PCA we

would like to highlight the differences of cNMMC over molecular dynamics or Monte Carlo methods with imposed constraints. The latter two methods enable the user to impose constraints on certain degrees of freedom. In contrast, the use of cNMMC primarily facilitates conformational change along a set of user defined or experimentally inferred degrees of freedom (referred to as 'Natural Moves'); other degrees of freedom are treated as subordinate (but not constrained) to fully facilitate the exploration of the conformational space along Natural Moves. This is a very different strategy from Cartesian (or generalised) coordinate based exploration of the conformational space with constraints, regardless whether the exploration algorithm is molecular dynamics or Monte Carlo based. Due to the benefits of Natural Moves, where chain breakage is followed by closure, any part of the molecular assembly can be moved and the necessary subordinate degrees of freedom will be rearranged to maintain the integrity of the system. Given that the latter part is automated and hidden from the end users, this strategy provides them with the opportunity to focus only on the essential moves or molecular deformations rather than the less important degrees of freedom.

By the straightforward definition of candidate Natural Moves, cNMMC can facilitate the robust compilation of experimentally inferred [22] 'molecular motion' into a molecular simulation protocol. This advantage is particularly relevant for computational structural biology given the complexity and diversity of biomolecular architectures. Focusing on the Natural Moves as opposed to the corresponding constraints can provide a more intuitive way to describe, classify and ultimately understand the underpinning mechanisms of functionally relevant deformations. This is among the grand challenges of computational structural biology and biophysics and can only be achieved with tight collaboration of computational and experimental scientists. To tighten this partnership, the use of cNMMC can catalyse more active engagement of experimental biophysical scientists, who often have extensive experience working on a given biological system, in conducting these types of molecular simulations.

The more quantifiable advantage of cNMMC compared to molecular dynamics (or Monte Carlo) with imposed constraints is the large speed gain by reducing the number

of essential degrees of freedom. In a cNMMC protocol the investigation is commonly restricted to a few degrees of freedom (e.g. 6+; orientational and translational parameters of a structural segment plus a few internal dihedral and bond angle degrees of freedom), whereas it is less intuitive for a general user, who might not be specialised in molecular simulations, to automate the procedure for imposing constraints on the remaining degrees of freedom. In addition, the use of dependent degrees of freedom, which significantly facilitates exploration along desired motions, is another unique feature of the presented technology compared to constrained molecular dynamics or Monte Carlo methods.

With the advantage of being able to 'define moves' liberally and sample conformations along these moves very efficiently, we managed to address applications [19, 22, 24, 30, 31] that were not feasible before NMMC. For example, the latest application [31] demonstrated that we could speed up simulations by orders of magnitude compared to molecular dynamics, while still being able to reproduce experimental observables. With cNMMC these computational experiments will become more accessible to a wider scientific community including experimental laboratories.

As described above with NMA and PCA, cNMMC is best used as a complementary method to constrained molecular dynamics (or Monte Carlo), which could refine our understanding of systems with degrees of freedom that cNMMC predicts to be relevant.

As well as other computational methods, experimental information on collective motions derived from NMR data [152] may also be used to guide the design of customised Natural Moves. A range of methods already exist that use NMR data to complement MD simulations [153].

Additionally, pre-existing expert knowledge is central to generating new ideas. The cNMMC protocol presented here is a first step to bridging the gap between the biological intuition of scientists and molecular simulations by allowing the introduction of arbitrary degrees of freedom for the investigation of conformational changes and mechanisms.

3.4 Normal Mode Analysis

Normal mode analysis (NMA) is a computationally efficient method to study large structural rearrangements in proteins. The normal mode vectors define the direction in which the atoms move, as well as their relative displacement to each other. There is no absolute measure of displacement for each atom and all atoms in each normal mode vibrate with the same frequency. Even for different models with varying degrees of coarse-graining the low-frequency motions are largely conserved [210]. Thus, while normal modes often give a good indication of the type of structural motions, they yield little information on the amplitude of the movement. It has been shown that the normal modes with the lowest frequencies (soft modes) characterise the largest movements in a protein structure, a selection of which are often functionally relevant. A number of methods have been developed [60, 211, 212] and successfully used to study the concerted, large structural motions of protein systems including lysozyme [213], HIV1-protease [214], Ca-ATPase [215]], F1-ATPase [216], chaperonin GroEL [217], and the ribosome [69]. Recently, the use of NMA for the investigation of structural motions of biomolecular systems increased significantly. This is due to the low computational cost of many of these methods, which has allowed researchers to develop easily accessible web-servers, such as NOMAD [134], that perform normal mode calculations on a structure file submitted by the user.

NMA traditionally uses a set of coordinates, a force field describing atomic interactions, and a set of algorithms to perform the necessary calculations. In Cartesian coordinate space NMA involves three main steps: 1. Energy minimisation of the structure of interest 2. The calculation of the so-called 'Hessian' matrix, a matrix that contains all second derivatives of the potential energy with respect to the mass-weighted atomic coordinates 3. The diagonalisation of the Hessian matrix. This last calculation generates the eigenvalues and eigenvectors (the 'normal modes'). Each of these three steps can be computationally demanding, depending on the size and resolution of the structural model. Generally, the first and third step are the biggest bottlenecks. Note that if an Elastic Network Model is used to describe the potential, the initial energy minimisation step is not required.

The large number of variables required to build and diagonalise the Hessian, means that one of the limiting factors for NMA is the diagonalisation of the $3N \times 3N$ matrix, where N represents the number of atoms. Several methods have been developed to address this challenge [134, 218, 219]. In the following section I will describe the approach that was used in one of the studies presented in this thesis.

Rotation-Translation Blocks

Most relevant to this thesis, as it was used to calculate the normal modes of HLA-DM (see chapter 7), is a low-dimensionality algorithm implemented by Sanejouand and co-workers [220] based on decomposing the system into rigid blocks of several residues, each only having six rotation-translation degrees of freedom. The lowest-frequency normal modes of the protein are obtained as a linear combination of the rotations and translations of these blocks. This approach is named the rotation-translation of blocks (RTB) method and allows for the quick and accurate approximation of normal modes in all-atom structures.

Traditionally, the normal modes of a system are calculated by diagonalising the Hessian matrix, the $3N \times 3N$ matrix containing the second derivatives of the potential energy with respect to the mass-weighted atoms, where N refers to the total number of atoms in the system. The RTB method, however, defines H in terms of the rotations and translations of n_b blocks, resulting in a simplified Hessian matrix H_b. The purpose of the Rotation-Translation Blocks (RTB) approach is the efficient calculation of normal modes for large biological structures with atomic representation. n_b number of blocks of residues are defined and the six rotation-translation modes of each block are calculated. These $6n_b$ vectors then represent the new dimensionality-reduced basis, on which the Hessian matrix, H_b, is defined. The low-frequency normal modes of the system are approximated by diagonalising H_b, which now has a size $6n_b \times 6n_b$ instead of $3N \times 3N$.

3.5 DNA structural analysis with x3DNA

DNA structure is well defined and can be accurately described by a set of geometrical parameters. The relative orientation of nucleotides within individual base pairs, and base stacks in particular, is widely used as a way of characterising DNA structure. One of the most widely used software packages for the analysis of these parameters is x3DNA [221].

It can analyse antiparallel and parallel double helices, single-stranded nucleic acids chains, triplexes, quadruplexes and further tertiary structures of DNA and RNA. First the software creates a map of existing base pairs and establishes the double helical character of suitable base pair steps. It applies a well studied and widely agreed upon reference frame for the geometric characterisation of base pairs and a robust matrix-based scheme to calculate local conformational parameters.

3.5.1 Identification of base pairs

X3DNA begins by uniquely defining the reference frame for each base in the structure. This is done with least squares fitting of a standard base structure with an embedded reference frame [222].

The initial base pairing data is generated with a secondary programme that identifies all bases that are in close contact. This exclusively geometric method utilises the widely used standard base reference frame [83] and can be used to find all base pairs (canonical and non-canonical), higher order base interactions and double helical regions in a structure. The hydrogen bonding patterns of the identified interactions are tested, allowing for the characterisation of unconventional pairings.

All base pairs are defined by six rigid body parameters (Figure 3.2 upper left), as discussed in the following section.

3.5.2 Base pair parameters

Each base has two distinct faces, due to molecular asymmetry [223], which can be classified with the standard nucleic acid base reference frame. For canonical (Watson-Crick)

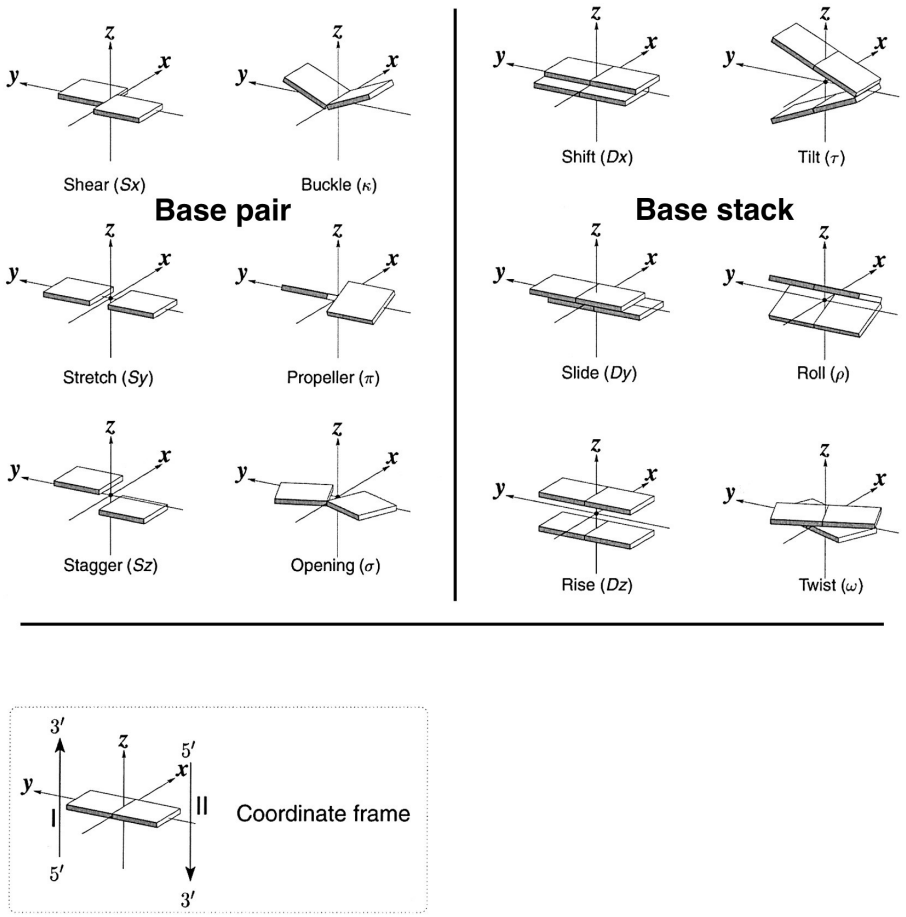

Figure 3.2: Schematic representation of rigid body parameters for the geometric characterisation of base pairs (on the left) and sequential base pair stacks (on the right). The reference frame (lower left) is defined so that the x-axis faces away from the minor groove side of a base or base pair and the y-axis faces the sequence strand (I). The relative position and orientation of successive base pair planes are described with respect to both a dimer reference frame (upper right) and a local helical frame (lower right). The schematic depicts positive values of the parameters. Figure adapted from [221].

base pairs in a double helix the strands are antiparallel and the faces of the pairs of bases are of the opposite sense. If the two bases in a pair share the same face, such as in the Hoogsteen base pair the strands are considered to be parallel. The six base pair parameters shear, stretch, stagger, buckle, propeller and opening with the two strands running in opposite directions describe the relative position and orientation of two bases. The parameters provide a simple way to classify structures and perform database searching. Only three of the six base pair parameters, Shear, Stretch and Opening are required for fully characterising key hydrogen bonding features and identifying the base pair type:

Shear and stretch describe the relative offset of the two bases and opening represents the angle between the two x-axes with respect to the average normal to the base pair plane (see upper left in Figure 3.2). Buckle, propeller and stagger are secondary parameters, which describe the imperfections, such as the non-planar orientation of a given base pair.

3.5.3 Dimer step parameters

Six rigid body parameters (three rotations and three translations) are sufficient to fully characterise the position and orientation of a base pair with respect to another. Common parameters used for this purpose are the set of local base pair stack parameters, shift, slide, rise, tilt, roll and twist, which describe the stacking geometry of two neighbouring dinucleotides. In X3DNA, the helical axis is specified at a local level. This axis is defined as the rotational axis that aligns the reference frames of successive base pairs. The calculations for most of the two sets of rigid body parameters are identical and rigorous. Thus, one set of parameters may be derived from the other without losing any information [224].

3.6 Modeller

MODELLER, as the name suggests, is a software package for the modelling of protein structures [225, 226]. Most commonly the input consists of an alignment of the target and template sequences, the structure of the template as well as a small parameter file. MODELLER then proceeds to generate a model containing all non-hydrogen atoms.

For comparative protein structure modelling MODELLER attempts to satisfy a number of spatial restraints including 1. homology-derived restraints 2. stereochemical restraints such as bond length and angle preferences 3. knowledge-based preferences for dihedral angles and non-bonded interatomic distances and 4. optional user-defined restraints derived from NMR, secondary structure characteristics, cross-linking experiments, fluorescence spectroscopy, site-directed mutagenesis and researcher intuition. The spatial restraints are defined as probability density functions and are grouped into an objective

function that is optimised by several steps including conjugate gradients and molecular dynamics with simulated annealing.

MODELLER can also perform additional tasks such as the *de novo* modelling of loops in protein structures [226]. We used this capability in chapter 5 to model the missing linker regions between the diabody VH/VL domains.

3.6.1 Modelling of loops in protein structures

Loop modelling may be regarded as a local protein folding problem. An accurate conformation of a given section of a protein needs to be generated mainly using the sequence in question. However, loops are usually not long enough to provide enough context regarding their structure. Short loops (9 residues or shorter) sometimes take on entirely unrelated folds in different proteins [227, 228]. This observation suggests that the structure of a given segment is affected by the residues in the loop itself as well as the structure of the remaining protein around the loop.

Due to the diversity of loops between homologues, predicting loop conformations is a challenging task - loops are commonly the most inaccurate parts of a protein model. Furthermore, the prediction accuracy of *ab initio* as well as knowledge based methods decreases with loop length (as the number of degrees of freedom increases). Hybrid approaches that combine *ab initio* and knowledge-based techniques have shown promising results [229].

MODELLER approaches loop modelling using an optimisation-based approach, as opposed to a database approach, in order to make use of the wide applicability and conceptual simplicity of energy minimisation and to avoid the restrictions faced by databases imposed by a limited number of experimental protein loop structures. The MODELLER loop modelling protocol optimises the positions of all non-hydrogen atoms of a loop in a fixed environment. The optimisation relies on conjugate gradients and molecular dynamics with simulated annealing. The optimised pseudo energy function is a sum of many terms, including some terms from the CHARMM-22 molecular mechanics force field [230], and spatial restraints based on distributions of distances and dihedral angles in known

protein structures.

The MODELLER method for modelling a loop in a given environment can be divided into three parts: 1. the protein representation 2. the restraints that guide the objective function; and 3. the energy minimisation method.

Representation The representation of the protein includes all non-hydrogen atoms. Explicit solvent molecules are excluded and Cartesian coordinates of the loop atoms are used as the degrees of freedom in the optimisation steps. The loop atoms 'feel' the presence of the other atoms in the system, but the positions of the atoms around the loop are not changed during optimisation.

Energy function The energy function partly relies on statistical knowledge about the atoms for different geometries as extracted from the PDB. The stereochemical features such as covalent bonds, bond angles etc. are captured by the CHARMM molecular mechanics force field [230]. Non-bonded interactions and solvation are approximated by a statistical potential of mean force for pairs of protein atoms [231]. The accuracy of the scoring function is also enhanced by the use of statistical knowledge regarding the backbone and side-chain dihedral angles [225]. The energy function is a sum of restraints, each depending on a distance, angle, dihedral angle, improper dihedral angle, or a pair of dihedral angles defined by two, three, four, or eight atoms.

Optimisation The first optimisation is the generation of an initial structure. The loop atoms are placed uniformly along the axis that connects the backbone carbonyl oxygen and amide nitrogen atoms of the N- and C-terminal anchor sites, respectively. The atoms are then randomised by adding a random number distributed uniformly from -5 to 5 Å to each of the Cartesian coordinates. A single loop prediction consists of the independent optimisation of a set of such randomised initial structures, and choosing as the final model the structure with the lowest energy value. The optimisation of a single loop contains conjugate gradient minimisation and MD simulation with simulated annealing.

First, a conjugate gradient step relaxes the system. The atoms are allowed to move

close to each other without having to pass high energy barriers. This stage is followed by by a simulated annealing MD simulation. In the last step, the optimisation is finalised by a conjugate gradient relaxation. This cycle of conjugated gradient, MD simulation, conjugated gradient occurs twice. In the first cycle only those non-bonded atom pairs are considered that exclusively contain loop atoms i.e., the loop does not interact with its environment. In the second cycle, the atom pairs that contain a maximum of one environment atom are also considered in the energy function i.e., the loop is allowed to interact with its environment.

Generally, the final loop model corresponds to the lowest energy conformation of 500 independent optimisations.

Chapter 4

MHC Class I Peptide Detachment Pathways

This work was published in: Knapp, B., Demharter, S., Deane, C. M., Minary, P. (2015). Exploring peptide/MHC detachment processes using Hierarchical Natural Move Monte Carlo. *Bioinformatics*.

4.1 Summary

The interaction between peptides and major histocompatibility complexes (MHC) is a crucial step for the activation of the adaptive immune system. A number of software tools have been developed to predict peptide/MHC (pMHC) binding. However, there has been a lack of insight regarding the process by which peptides detach from the MHC binding groove. Here, we used a simplified protein model with 3pt-representation and a knowledge potential as well as hierarchical Natural Move Monte Carlo (hNMMC) and stochastic conformational optimisation to investigate the detachment pathways of 32 different peptides from a MHC class I molecule (HLA-A*02:01, one of the most frequently observed MHC alleles in humans). We performed 100 independent replica simulations for each peptide and found that experimentally proven anchor amino acids have a significant effect on peptide detachment pathways. We assessed our results by comparing them to experimental binding affinity data. We were able to show the reliability of our approach

with an area under the receiver operating characteristic curve of 0.85. We also compared our simulations to a 1000 ns molecular dynamics simulation that we performed of a non-binding peptide (AAAKTPVIV) and HLA-A*02:01. Despite this being the longest published molecular dynamics simulation for pMHC, the peptide only partially detaches. Our approach is orders of magnitude faster and as such allows us to explore pMHC detachment processes on a scale that was not possible with all-atom molecular dynamics simulations performed previously.

4.2 Introduction

The presentation of peptides through MHC molecules at the cell surface is essential for the mounting of an effective immune response. In order to bind to the MHC molecule, unstructured peptides change to a helix-like extended configuration that is stabilised by the surrounding residues of the binding-groove [232]. Understanding this process will provide valuable information for the modulation of this medically important receptor.

A range of software for the prediction of MHC-peptide binding exists that can calculate binding free energies, predict epitopes, and distinguish binding from non-binding peptides. These can be divided into two main groups according to the type of data used: sequence-based models and structure-based methods. Sequence-based models can be either trained on qualitative data that indicate binding or non-binding of the peptide or quantitative data such as experimentally determined binding affinities. Structure-based methods include MD simulations [233, 234], peptide threading [235, 236] and peptide docking [237–239].

Molecular dynamics simulations can distinguish effectively between binders and non-binders with some attempts being successful at predicting affinity between the peptide and the MHC receptor [234, 240, 241]. The dynamics on short timescales can also be captured [242]. However, to date there has been no simulation showing the full detachment process of non-binding peptides leaving the binding groove.

In this study we introduced a protocol based on Natural Move Monte Carlo that enabled us to study the full peptide detachment process. Binders were distinguished from

non-binders by evaluating their propensity to dissociate from the binding groove during simulation. Non-binders were expected to detach from the binding groove significantly earlier than strong binders. This study provides a proof of concept for studying small scale biophysical processes with a knowledge-based, coarse-grained Natural Move Monte Carlo method.

4.3 Methods

Protein Model and Potential

The all-atom MHCII and peptide structures were converted to 3pt using gro2mat. The same knowledge potential as for all other protein applications in this thesis was used [21], however some adaptions had to be made in order to optimise the potential for peptide-protein interactions. In principal, residues in a coarse-grained model can get closer to each other than the excluded volume. In order to prevent this from happening, we uniformly scaled the energies for the pair interactions by a continuous non-linear function:

$$s(r) = \begin{cases} s_0 + (1 - s_0)\left(\frac{r}{r_0}\right)^6 & if\ r\ <\ r_0 \\ 1.0 & if\ r\ \geq\ r_0 \end{cases} \tag{4.1}$$

where r_0=0.7nm (r refers to radius) approximately represents the size of a large amino acid and s_0=0.15 (s refers to scaling factor) was chosen as a offset to allow peptides to get out of deep energy minima. All pair interactions with distances larger than 0.7nm were treated as in [21].

Natural Move Decomposition

We decomposed the pMHCI structure into seven regions, as shown in Fig 4.1 and table 4.1. Both helices (A1 and A2) are split into two segments (A1.1, A1.2 and B1.1, B1.2). The helices as a whole as well as each of the segments are allowed to move independently, which accounts for the plasticity that is expected around the conserved kinks in the

centre of the helices [243]. Similarly the peptide is propagated as one region and as two independent segments along three translational and three rotational degrees of freedom.

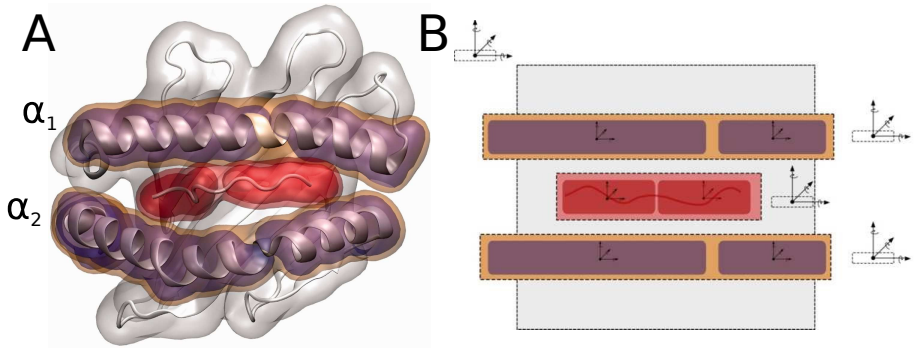

Figure 4.1: *The structural decomposition of pMHCI as used in this study.* **A** Cartoon representation of HLA-A*02:01 based on PDB accession code 3PWN. White: MHC beta-floor; red: peptide; orange: whole helices; magenta: regions interspersed by molten zones. For clarity, the $\alpha 3$ region and the $\beta 2$-microglobulin are not shown. **B** Schematic representation of the peptide and MHCI regions used in this study. The axes indicate the six degrees of freedom per region. Figures adapted from [31].

Table 4.1: Decomposition of MHC class I structure for hNMMC.

Region #	Residues	Name
MHCI		
1	A:1-54, 88-135, 182-275, B:0-99	beta-floor and globular domains
2	A:56-73	helix A1.1
3	A:75-86	helix A1.2
4	A:56-73, 75-86	helix A1.1 + A1.2
5	A:137-150	helix A2.1
6	A:152-180	helix A2.2
7	A:137-150, 152-180	helix A2.1 + A2.2
Peptide		
8	C:1-4, 6-9	peptide C1 + C2

Simulation Protocol

In order to accelerate the peptide detachment process we performed hNMMC using a temperature modulation protocol. Specifically, we used repeated simulated annealing, which throughout the simulation applies sinusoidal cycles of rising and falling temperatures to the system. This allows for the efficient exploration of energetically favourable states along the peptide detachment pathway. It was implemented as described in equation 2.1

using the following parameter values: A=600 (Kelvin), k = 100,000, Ω=5000 and s=0 (Kelvin). The equation is reproduced here:

$$T_k = A \times \sin\left(\frac{2\pi k}{\Omega}\right) + s \tag{4.2}$$

The peptide/MHC dataset

The most commonly found MHCI allele in humans, HLA-A*02:01, was chosen for this study. We selected PDB accession code 3PWN as it was the median structure when clustered with 10 other high resolution structures of allele HLA-A*02:01. The original peptide sequence was LLYGFVNYI and was not simulated in this study.

A set of 32 peptide sequences with existing experimental affinities was chosen from an investigation of T cell cross-reactive peptides [244]. We used the experimental IC50 values from this study to test our method, as these are commonly used to benchmark peptide binding affinity predictors. The peptide sequence and corresponding IC50s are shown in table 4.2.

Table 4.2: The 32 peptides and their IC50 values used for this study. Binders: $\leq$ 1000nM, Weak binders: $\leq$ 10000, Non-binders: $\leq$ 20000.

Sequence	IC50	Sequence	IC50
AAAKTPVIV	20000	LTSNCTRTT	20000
AQFLYLYAL	0.4	MLFTKFFYL	0.2
CVDNHLGAT	20000	MTFGDIPLV	1
DMPPEVVYL	20000	NLFDIPLLT	10
EAAAATCAL	20000	NLLLWPLYV	0.1
EIEKVEKYL	20000	PTPKKMNIV	20000
FIADIGIGV	0.3	RAGYSIVEL	5000
FLGGTTVCL	20000	RQQLEDIFM	10033
FLIDLAFLI	0	RQVSVKLLI	250
FVKKMLPKI	3004	SIMAFILGI	0.2
GLLQFIVFL	0	SLKRNAEGI	20000
GTVINEDIV	20000	SLPACPEII	504
HQWDIDSAI	1007	TLNRNQPAA	15190
KMIYDLNAV	0.3	VLWTVFHGA	100
KTSTLIFFV	1	WIKTISKRM	20000
LLMMTLPSI	0.1	YLTAIQDFI	1.9

Each peptide was modelled on top of the original peptide of PDB accession code 3PWN

by using the same backbone and predicting the new residue side-chain conformations with SCWRL4 [245]. This was demonstrated to be the most effective approach for altering peptides inside the MHCI binding groove [246].

Simulations

We performed 100 replica simulations (100,000 steps each) for all 32 pMHC complexes using the protocol described above. The scripts used have been included at the end of this chapter.

4.4 Results

Examples of the detachment processes of the non-binder AAAKTPVIV during MD and hNMMC simulations are shown side by side in Figure 4.2.

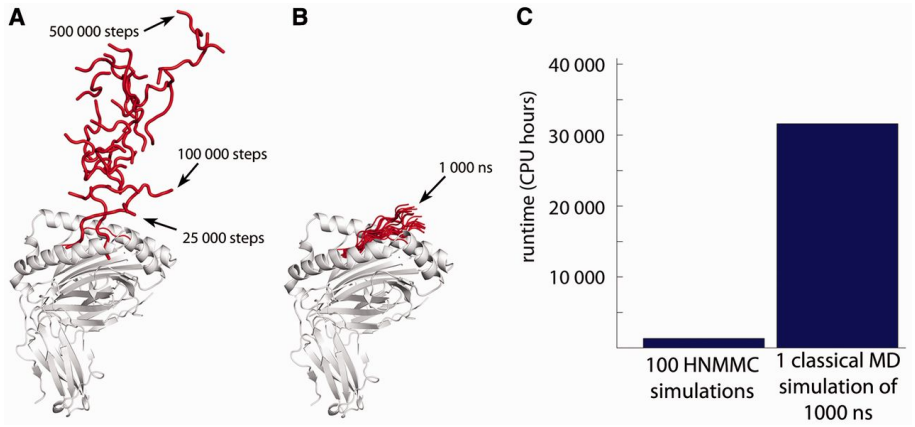

Figure 4.2: *Comparison of MD- and hNMMC-simulated peptide detachment pathway* (sequence:AAAKTPVIV, MHCI: HLA-A*02:01). **A** Equally distributed frames taken from the 500,000 step hNMMC simulation are shown. The computing time was 13h on a single core of an Intel i7-3770 3.40GHz CPU. **B** Equally distributed frames of a 1000 ns MD simulation (performed by Bernhard Knapp). The simulation took 247h using 128 Xeon cores at 2.0GHz of the Oxford Advanced Research Computing facility. **C** Comparison of computing time between hNMMC simulations and a single MD simulation. The demand for the 1000ns MD simulation was the same as for all the hNMMC simulations (n=3200) of our study combined. Figure adapted from [31].

In hNMMC the peptide starts detaching within 25,000 steps and has fully detached after 100,000 steps (shown by the arrow in Figure 4.2A. The MD simulation trajectory largely matches the first 25,000 steps of our hNMMC simulations. In both cases the

peptideâĂŹs C-terminus detaches first and starts 'flappingâĂŹ. However, during the MD simulation full detachment is never observed, despite this being the longest reported MD simulations of pMHC, with 128 cores used over 247 hours. This corresponds to the runtime it took to run 100 hNMMC replica simulations (100,000 steps) for all 32 peptides (Figure 4.2B). Thus, hNMMC simulations are capable of efficiently calculating pMHC detachment processes.

After we showed that our hNMMC protocol is suitable for simulating peptide detachment and that the results resembled those of MD simulations we performed hNMMC for 32 known binding and non-binding peptides. We performed 100 replica simulations, each with a different random seed in order to generate a wide range of stochastically different trajectories. Figure 4.4A shows the average distance between the N-terminus, center and C-terminus of known binders and non-binders, and the binding groove.

The initial peptide/MHC β-floor distance in the starting structures of all 32 peptides is 13.1 Ångstrom for the centre of the peptides (peptide Cα5:MHC Cα28), and 10.8 Ångstrom and 10.9 Ångstrom for the N- (peptide Cα1:MHC Cα99) and T-terminal (peptide Cα9:MHC Cα117) ends, respectively 4.3B. We found that there was an equal tendency for the C-terminal and N-terminal end to detach, while the centre of the peptide tends to move closer to the binding groove during the first 20,000 steps. However, individual peptides seemed to prefer one of the ends (Figure 4.4A-C). In contrast, we never found that the centre detached first or both ends detached at the same time (Figure 4.3B, bottom panel).

In the next step we attempted to use the peptide/MHC distances generated during our simulations to discriminate between binders and non-binders. Non-binders are expected to exhibit larger distances than binders i.e. they should detach earlier. We investigated this by comparing the average distance over all replicas of a peptide against the experimentally known binding affinity. This resulted in an area under the receiver operating characteristic curve (AROC) of 0.85 (Figure 4.6A) and Pearson correlation coefficient of 0.67. The difference between the pMHC-distances of all binders and all non-binders was significant (Figure 4.6B). It is important to note that single simulations could be misleading as the

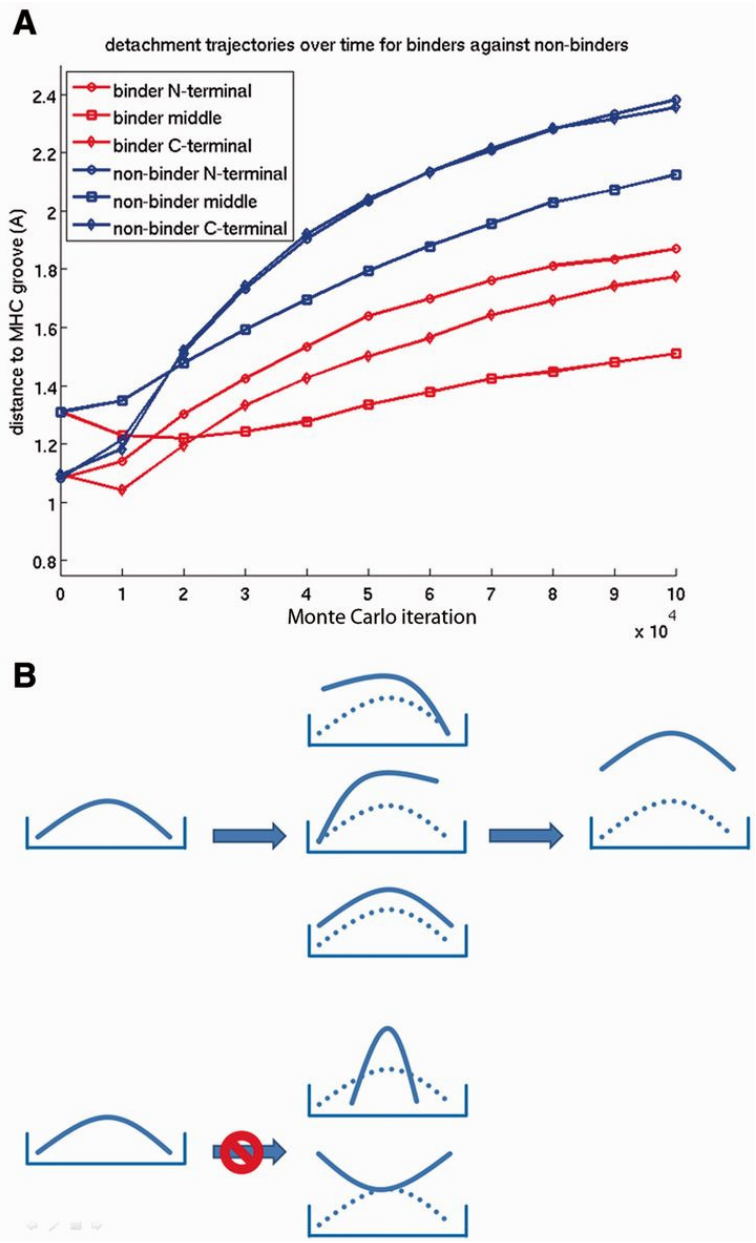

Figure 4.3: *Summary of simulated detachment pathways.* **A** Average distance during peptide detachment trajectories grouped by experimental binders and non-binders. **B** Schematic drawing of observed detachment pathways. Figure adapted from [31].

conformational exploration could be trapped in one or more local energy minima.

Bootstrapping analysis revealed that replica simulations were necessary to reduce the variability of the results. Figure 4.7A shows a sharp descent of the AROC standard deviations until 25 replicas and a slower descent until 50 replicas. This shows that our hNMMC approach can predict pMHC detachment with good accuracy and reliability if at least 25 replicas are used. The distribution of the AROC of 1 replica per peptide chosen randomly 5000 times with repetition from our 100 replicas gave a spread of AROC

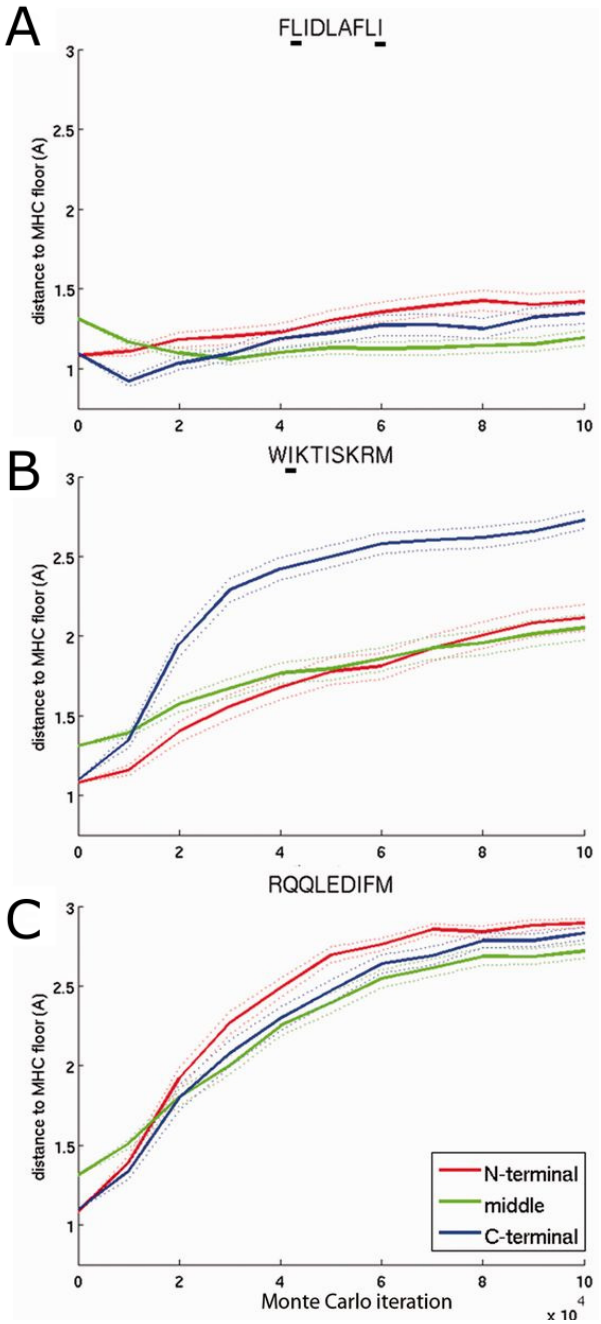

Figure 4.4: *Examples of simulated detachment pathways.* **A** Stable binding of peptide FLIDLAFLI due to favourable anchor residues at peptide positions two and nine **B** C-terminal detachment of the WIKTISKRM peptide from MHC. Position two is a known anchor residue. **C** Equal detachment of RQQLEDIFM across the peptide. This peptide contains no anchor residues. The distances were averaged over all 100 replicas. The dotted lines show the standard error of the mean. The detachment plots of all peptides are shown in Figure 4.5. Figure adapted from [31].

values between 0.53 and 0.91 and the standard deviation was 0.1 (Figure 4.7B). When 100 replicas were used the AROC consistently lied between 0.81 and 0.89 with a standard deviation of 0.01 (Figure 4.7C). This demonstrates the importance of multiple replicas,

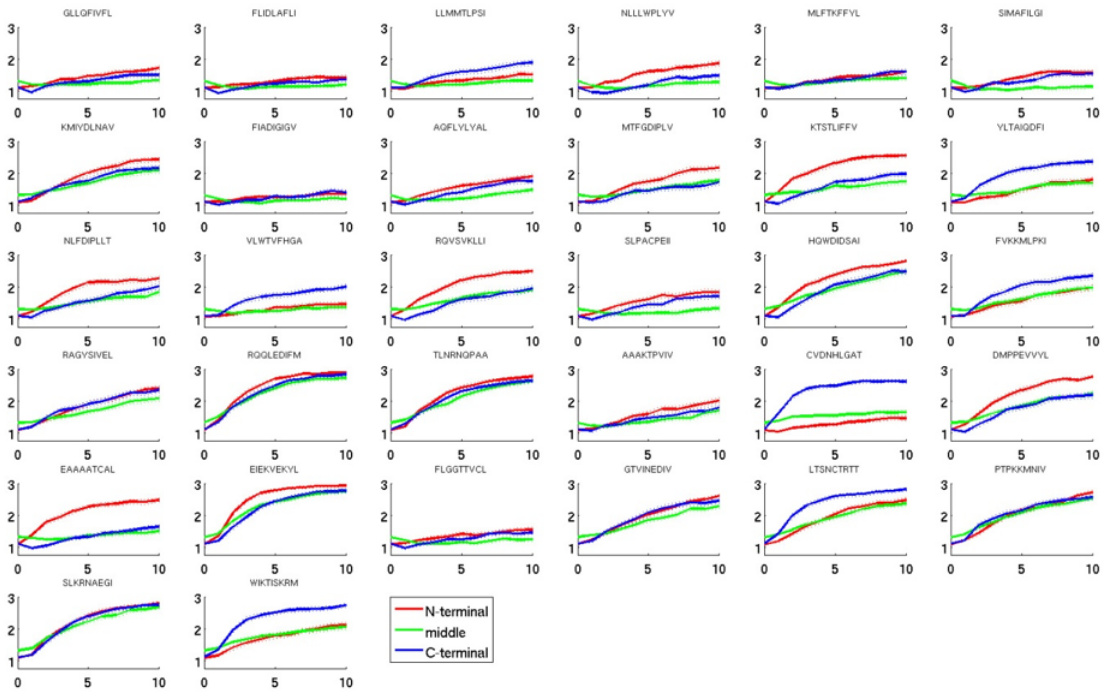

Figure 4.5: *The detachment plots of all 32 peptides.* Figure adapted from [31].

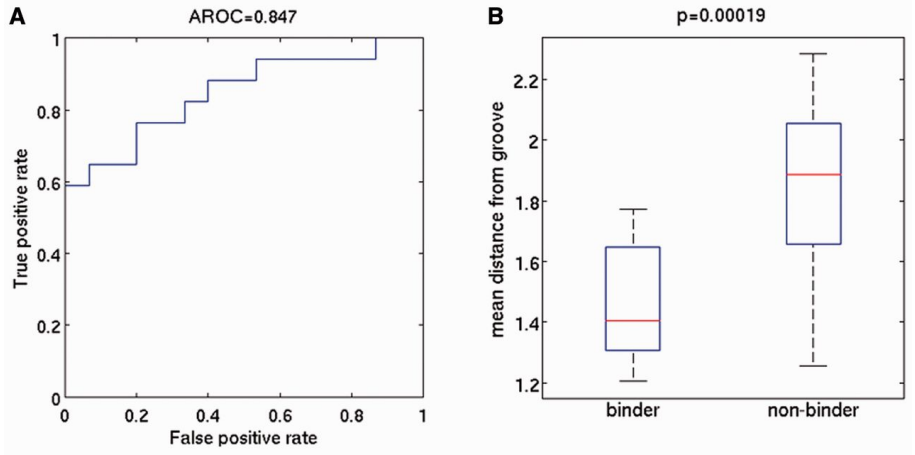

Figure 4.6: *Classification of binders and non-binders using hNMMC.* **A** ROC curve. **B** The peptide/MHCI distance averaged over all simulations and grouped by experimental binders and non-binders. Figure adapted from [31].

as a single simulation can give a wide range of different results.

4.5 Discussion

The MHC/peptide interaction has been studied in a number of different MD simulations studies [20]. While partial detachment was observed in some, full peptide dissociation has not been observed to date, not even in the longest pMHC simulation of 400 ns [247].

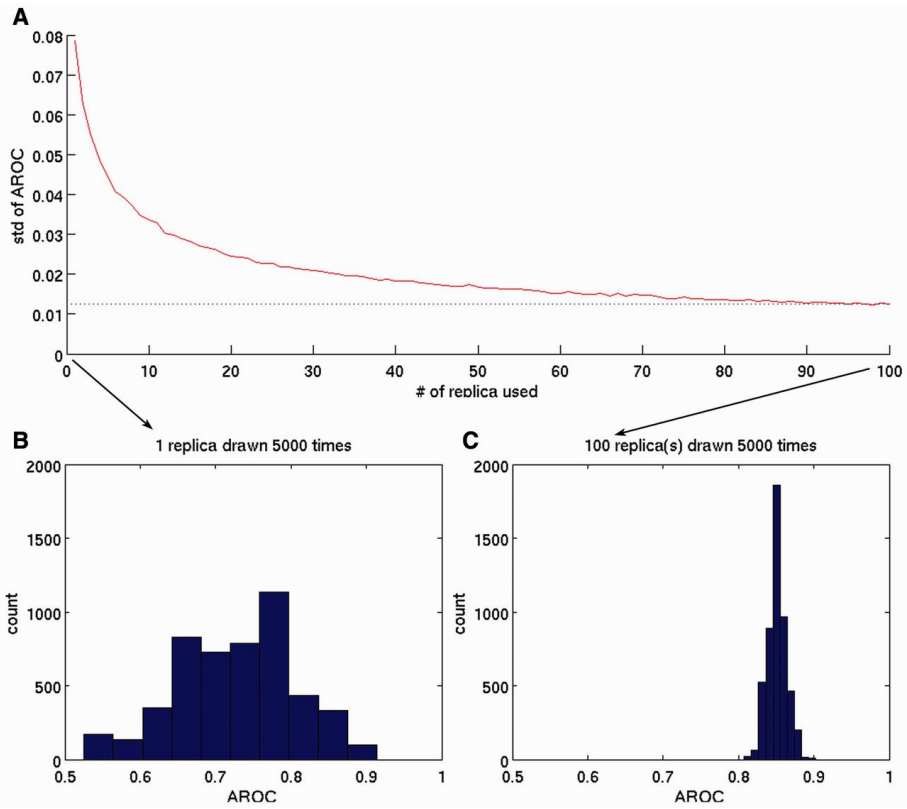

Figure 4.7: *Bootstrapping analysis.* **A** The AROC standard deviation with 5000 random selections is plotted against the number of replicas used. **B** The AROC distribution of 5000 random selections of 1 replica (with repetition) from the set of 100 replicas. The distribution ranges from 0.53 to 0.91. **C** The same as (B) but for 100 replicas chosen randomly 5000 times with repetition from our 100 replicas. Figure adapted from [31].

Here, we performed a 1000 ns simulation of a peptide that was experimentally shown to be a non-binder in complex with MHCI. Similarly to previous studies only partial detachment was achieved, which shows that classical MD simulations are not suitable to investigate pMHC detachment pathways within a reasonable time span. As a result most structural have studies have focused on bound pMHC [248], TCR/pMHC structures [204] or empty MHC molecules [158, 249]. Thus, we applied hNMMC in combination with a 3pt model and the simulating annealing algorithm to study this problem. By using these three technologies we were able to circumvent the bottleneck that was computational cost, and demonstrate how detachment processes of different peptides from the MHCI complex may take place.

In order to test how meaningful our results were, we compared them to experimental data. First, we identified preferred anchor amino acids from the IEDB [250]. While an-

chor residues are not considered to be the main contributing factor for pMHC binding, we observed good agreement between N- or C-terminal detachment and the presence/absence of preferred anchor residues (Figure 4.4A-C). Second, we checked if our detachment results can distinguish known binders from non-binders. We found good agreement between simulated detachment speed and binding data from experiments (Figure 4.6) [244]. The AROC for our training-free approach was 0.85; this is in the range of sequence-based prediction methods [251] and superior to structural docking techniques [234]. This indicates that coarse-grained hNMMC applied to pMHC has good accuracy at the biophysical scale and can detect the key features contributing to peptide binding. Even though temperature modulation and hNMMC were used, there was a finite probability that simulations become trapped in local minima or generate outlier trajectories leading to problems regarding convergence. Therefore, we performed a total of 100 replica simulations with distinct random seeds for every peptide. Using boot strapping analysis, we found that 25-50 replicas are required to reliably conclude a result. Similarly, recent studies showed that the comparison between too few MD simulations often generate misleading results [204] and that 50 replica simulations are needed to reliably predict the binding free energy of HIV drugs to HIV-1 Protease [252] and peptides to MHC [253].

4.6 Conclusion

We showed that hNMMC has the potential to give valuable insights into the peptide detachment pathways of pMHC complexes. This was the first study that analysed full peptide detachment trajectories and provided new perspective on the conformational landscape of pMHC.

4.7 Files

The following perl and bash scripts were used to run the simulations. The PDB files of the MHCI/peptide complexes with 3-pt representation were located in a folder named `testSet_32_threePoint` (see `runMultipleMosaicsSims.sh`). The script for converting

all-atom to 3-pt can be found at www.cs.ox.ac.uk/mosaics.

submitJobs.sh

```
#!/usr/bin/perl -w
# This script runs 50 replica of "runMultipleMosaicsSims.sh" (internally different seeds are used)
# B. Knapp and S. Demharter 2014-03-18


 for (my $repIter = 1; $repIter <= 5; $repIter++) {
     my $cmd = "qsub -v rep=$repIter -t 1-32 runMultipleMosaicsSims.sh";
    system $cmd;
 }
```

runMultipleMosaicsSims.sh

```
#!/bin/bash

#PBS -V
#PBS -N Mosaics
#PBS -l walltime=0:06:00
#PBS -l nodes=1:ppn=32

cd $PBS_O_WORKDIR

. enable_arcus_mpi.sh

randNumber=$(($[ 0 + $[ RANDOM % 10000000 ]]*100000))
randNumber=$(($randNumber+$[ RANDOM % 10000000 ]))
randNumber=$(($randNumber*-1))

filename=`ls -1 /data/stat-opig/bknapp/mosaicsRuns/testSet_32_threePoint/ | tail -n +${PBS_ARRAYID
filename=${filename%.*} # remove the extension

/data/stat-opig/bknapp/mosaicsRuns/runOneMosaicsSim.pl $filename $rep $randNumber
```

runOneMosaicsSim.pl

```
#!/usr/bin/perl -w

# runs one MOSAICS sim by creating the input
# B. Knapp 2014-02-04
#
# example for usage:
# ./runOneMosaicsSim.pl threePoint_AAAKTPVIV_20000 1
```

```
8
9
10  use strict;
11  use Cwd;
12
13  sub execCmd {
14      my $printOnly = 0;
15
16      my $returnVal = 0;
17      my $cmd = $_[0];
18      if (@_ == 2){
19          $printOnly = $_[1];
20      }
21      if($printOnly == 1) {
22          print $cmd."\n";
23          $returnVal = 1;
24      } else {
25          $returnVal = system $cmd;
26          # print "<<<<<<<< " . $returnVal . "\n";
27      }
28      return $returnVal;
29  }
30
31  my $bn = $ARGV[0]; # base name
32  my $repNr = $ARGV[1];# the current replica
33  my $mySeed = $ARGV[2]; # the random seed created outside
34
35  my $fn = $bn . ".pdb"; # file name
36  print "working on \"" . $bn . "\" ... \n";
37  my $srcDir = cwd();
38  $srcDir = $srcDir . "/";
39  my $cmd = "";
40
41  my $returnVal = -1;
42
43  my $fullDirName = $srcDir . $bn . "_rep".$repNr;
44
45  #$cmd = "rm -rf " . $srcDir . $bn;
46  $cmd = "rm -rf " . $fullDirName;
47  $returnVal = execCmd ($cmd);  #  success == 0
48  if ($returnVal) { # if rm was not successfull this directory did not exist. that is fine. success
49      print "Removing of directory \"" . $fullDirName  . "\" was not successful!\n";  print $! . "\r
50  } else {
51      print "Successfully removed directory \"" . $fullDirName . "\" ...\n";
52  }
53  $returnVal = execCmd ("mkdir " . $fullDirName); #  success == 0
54  if ($returnVal) { # creation of directory
55      print "Could not create directory \"" . $fullDirName . "\".\n"; print $! . "\n";
```

```
    exit;
} else {
    print "Successfully created directory \"" . $fullDirName . "\" ...\n";
}
$returnVal = chdir($fullDirName); # attention success == 1 !!!
if (not($returnVal)) { # change to directory
    print "Could not change to directory \"" . $fullDirName . "\".\n"; print $! . "\n";
    exit;
} else {
    print "Successfully changed to \"" . $fullDirName . "\"...\n";
}
$returnVal = execCmd ("cp /data/stat-opig/bknapp/mosaicsRuns/testSet_32_threePoint/" . $fn. " tmp.
if ($returnVal) { # copy the input filename
    print "Could not copy \"" . $fn . "\" to \"" . $fullDirName . "\".\n"; print $! . "\n";
    exit;
} else {
    print "Successfully copied \"" . $fn . "\" to \"" .$fullDirName . "/\"...\n";
}

# add the stride line to the top of the input pdb file
my $strideLine = "CBLC ~ABC
STRIDE ~RRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRCRRRRRRRRRRRRRRRRRRCRRRRRRRRRRRRCRRR
STRIDE ~RRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRR
STRIDE ~RRRRCRRRR
";

open my $in,  '<',  "tmp.pdb"      or die "Can not read file \"tmp.pdb\": $!";
open my $out, '>', "$fn" or die "Can not write new file \"$fn\": $!";

print $out $strideLine; # add this text

while( <$in> )
    {
    print $out $_;
    }
close $out;
print "Successfully added STRIDE line to pdb input file\" " . $fn ."\".\n";



my $regionsFileName = $bn . ".regions";
my $regionsFileContent = "_________________MHC_________________
~region[\\element_top_type{segment}
         \\dependency_type{independent}

         \\nseg{4}
         \\ncenter{4}
         \\segments_firstres{A:1,A:88,A:182,B:1}
```

```
          \\segments_lastres{A:54,A:135,A:275,B:100}

          \\segments_baseres{A:27,A:111,A:228,B:49}

          \\centers{A:27,A:111,A:228,B:49}

          \\prop_trans_sig{1.e-4}
          \\prop_rot_sig{1.e-5}
          \\prop_trans_sig_freeres{0}
          \\prop_rot_sig_freeres{0}
]

_________________helix1.1_________________
~region[\\element_top_type{segment}
          \\dependency_type{independent}

          \\nseg{1}
          \\ncenter{1}
          \\segments_firstres{A:56}
          \\segments_lastres{A:73}

          \\segments_baseres{A:64}

          \\centers{A:64}

          \\prop_trans_sig{1.e-4}
          \\prop_rot_sig{1.e-5}
          \\prop_trans_sig_freeres{0}
          \\prop_rot_sig_freeres{0}
]

_________________helix1.2_________________
~region[\\element_top_type{segment}
          \\dependency_type{independent}

          \\nseg{1}
          \\ncenter{1}
          \\segments_firstres{A:75}
          \\segments_lastres{A:86}

          \\segments_baseres{A:80}

          \\centers{A:80}

          \\prop_trans_sig{1.e-4}
          \\prop_rot_sig{1.e-5}
          \\prop_trans_sig_freeres{0}
          \\prop_rot_sig_freeres{0}
```

```
]



_________________helix2.1_________________
~region[\\element_top_type{segment}
         \\dependency_type{independent}

         \\nseg{1}
         \\ncenter{1}
         \\segments_firstres{A:137}
         \\segments_lastres{A:150}

         \\segments_baseres{A:143}

         \\centers{A:143}

         \\prop_trans_sig{1.e-4}
         \\prop_rot_sig{1.e-5}
         \\prop_trans_sig_freeres{0}
         \\prop_rot_sig_freeres{0}
]

_________________helix2.2_________________
~region[\\element_top_type{segment}
         \\dependency_type{independent}

         \\nseg{1}
         \\ncenter{1}
         \\segments_firstres{A:152}
         \\segments_lastres{A:180}

         \\segments_baseres{A:166}

         \\centers{A:166}

         \\prop_trans_sig{1.e-4}
         \\prop_rot_sig{1.e-5}
         \\prop_trans_sig_freeres{0}
         \\prop_rot_sig_freeres{0}
]



_________________peptide_________________
~region[\\element_top_type{segment}
        \\dependency_type{independent}

```

```
        \\nseg{2}
        \\ncenter{2}
        \\segments_firstres{C:1,C:6}
        \\segments_lastres{C:4,C:9}

        \\segments_baseres{C:2,C:7}

        \\centers{C:2,C:7}

        \\prop_trans_sig{1.e-3}
        \\prop_rot_sig{1.e-4}
        \\prop_trans_sig_freeres{1.e-4}
        \\prop_rot_sig_freeres{1.e-5}
]";

open (MYFILE, ">>$regionsFileName");
print MYFILE $regionsFileContent;
print "Successfully wrote region file \" " . $regionsFileName ."\".\n";
close (MYFILE);


my $paramFileName = $bn . ".params";
my $paramFileContent = "~sim_gen_def[
   \\simulation_typ{MIN}  PT EEMC SEQ_PT SEQ_EEMC NM DBFR
   \\minimize_type{stsamc}
   \\energy_report{2}
   \\num_procs{1} # of processors to be used (default is 1) use replica_number+1
   \\prop_type{tors} cart: cartesian, tors: torsional
   \\prop_tors_sig{0} 1.e-5 proposal sig 0 < number < 2 Pi, usually 1.e-5
   \\prop_rot_sig{1.e-5} 1.e-5 {0 <= radian < 2 Pi}
   \\prop_trans_sig{1.e-4} 1.e-4 {Angstrom >= 0.0}
   \\prop_clos_sig{1.e-3} 1.e-3 {Angstrom >= 0.0}
   \\replica_number{5} :5 10 number >=0 replicas:0, 1, 2, 3, 4, ....
   \\total_step_mc{100000} 100 000 :10   number > 1
   \\local_step_md{1} 10     number > 1
   \\time_step_md{0.5} :0.5
   \\statistics_freq{250} 200
   \\write_energy_unit{Ha} kcal  Ha: atomic unit, kcal: kcal/mol
   \\temperature{300} 300
   \\stsamc_type{trigonom}
   \\stsamc_period{5000}
   \\stsamc_ampl{600}
   \\stsamc_shift{0}
   \\burn_in_B{0} 2
   \\burn_in_N{0} 2
   \\postprop_minimize{clos} clos
   \\postprop_minimize_itmax{8} {integer >= 0}
   \\postprop_minimize_energy{bond_bend} {bond,bond_bend,bond_bend_tors,bond_bend_tors_onfo,all}
```

```
    \\extend_inter{3bond_conn} 3bond_conn, 4bond_conn, off:default
    \\cancel_res_inter{off} local, neighbor (includes local), off
    \\rinter_switch_length{0.0} 1.0 real inter switching length in A
    \\inter_list{none} lnk_list none
    \\EEMC_Emin{-0.3} -0.04 in Ha
    \\EEMC_Emax{0.0}  0.0 in Ha
    \\random_seed{$mySeed} {large integer} : -9378000501
]

~sim_mol_def[
  \\system_def{residue} primitive
  \\cgres_model{KB_3pt}  KB_3pt, off
  \\mol_parm_file{../bin/top_3pt_prot_na.rtf}
  \\bond_database_file{../bin/par_3pt_prot_na.prm}
  \\bend_database_file{../bin/par_3pt_prot_na.prm}
  \\tors_database_file{../bin/par_3pt_prot_na.prm}
  \\onfo_database_file{../bin/par_3pt_prot_na.prm}
  \\inter_database_file{../bin/par_3pt_prot_na.prm}
  \\pos_init_file{./".$bn.".pdb}
  \\pos_out_file{./".$bn."_sampled.pdb}
  \\atom_pos_file{./".$bn."_sampled.pos}
  \\tors_pos_file{./".$bn."_sampled.tors_pos}
  \\epot_file{./".$bn."_sampled.pot_energy}
  \\einter_file{./".$bn."_sampled.inter_energy}
  \\region_database_file{". $fullDirName . "/" . $regionsFileName ."}
  \\energy_term{bond}
  \\energy_term{bend}
  \\energy_term{tors}
  \\energy_term{onfo}
  \\energy_term{inter}
] ";

open (MYFILE, ">>$paramFileName");
print MYFILE $paramFileContent;
print "Successfully wrote param file \" " . $paramFileName ."\".\n";
close (MYFILE);



$cmd = "/data/stat-opig/bknapp/mosaicsRuns/bin/mosaics.x " . $fullDirName . "/" . $paramFileName .
print "Starting MOSAICS with \"" . $cmd . "\"\n";

$returnVal = execCmd ($cmd);  #  success == 0
if ($returnVal) { # exec MOSAICS
    print "There is a problem executing MOSAICS!"; print $! . "\n";
    exit;
} else {
    print "Successfully started MOSAICS (check with TOP or the log-file for progress).\n\n";
```

```
296 }
297
298 chdir("../");
```

Chapter 5

Assessing the dynamic range of EpoR-specific diabodies

This work was published in: I. Moraga, G. Wernig, S. Wilmes, V. Gryshkova, C. P. Richter, W.n Hong, R. Sinha, F. Guo, H. Fabionar, T. S. Wehrman, P. Krutzik, S. Demharter, I. Plo, I. L. Weissman, P. Minary, R. Majeti, S. N. Constantinescu, J. Piehler and K. C. Garcia (2015). Tuning Cytokine Receptor Signaling by Re-orienting Dimer Geometry with Surrogate Ligands. *Cell*, 160(6), 1196-1208.

5.1 Summary

The majority of cytokine and growth-factor receptors in the membrane act as homodimers. Thus it has been suggested that modulating the dimer topology should lead to a change in downstream signalling. In this study a set of four diabodies were designed from known antibodies and deployed as surrogate ligands for the Erythropoietin receptor (EpoR), that dimerised the EpoR ectodomains. The diabodies elicited a range of different signalling amplitudes, from full to minimal agonism. Interestingly, the solved structures varied in EpoR dimer configuration and the proximity of the N-terminal ends. As part of this study, in order to strengthen the conclusions that the diabodies enforce certain receptor dimer topologies that lead to different signalling outcomes, we conducted a NMMC study to investigate the dynamic range of the diabodies. This 'tuning' of receptor

signalling may have broad application across many dimeric receptors systems.

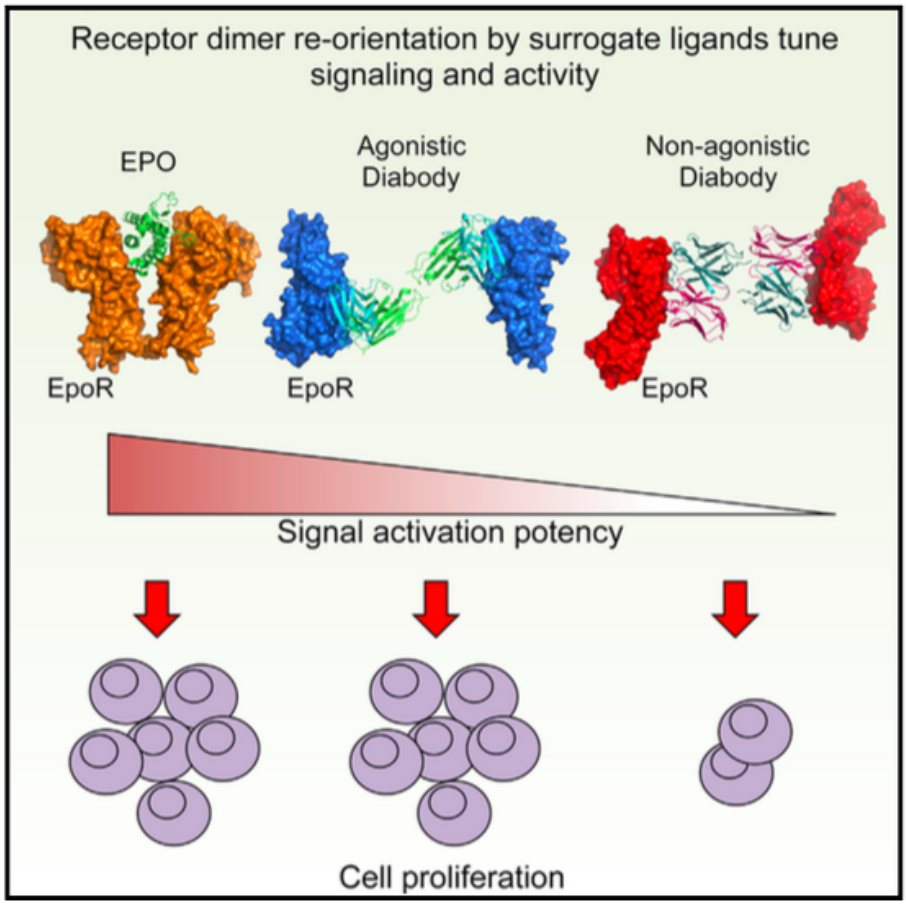

Figure 5.1: Diabodies can alter the strength of EPO signalling and inhibit oncogenic ligand-independent signalling, by reorienting receptor dimer geometry. Adopted from [24].

5.2 Introduction

The homo-dimerisation of receptors often plays a key role in the transduction of signals across the cell membrane. For example, cytokines, a class of secreted glycoproteins that participate in the regulation of cell fate and function bind to extracellular domains of their respective receptors, which triggers dimerisation and signalling [254].

Cytokine-receptor complexes exhibit a range of different structures and architectures that are suitable for signalling. This variety is further extended by evidence that showed unique dimerisation characteristics caused by the binding of engineered peptides, antibodies and other ligands. However, the structural topology of these complexes has not been

revealed to date [255]. Furthermore, the impact of these non-native ligands on signalling activity is still unclear. Previous work has shown that signalling activity is affected by structural perturbations and mutagenesis of extracellular domains [256].

Thus, the question arises: how does modulating EpoR homo-dimer geometry affect activity and does it constitute a viable strategy to 'tune' signalling?

Erythropoietin (EPO) is a cytokine that upon binding to EPO-Receptor (EpoR) ectodomains leads to their homo-dimerisation and initiation of signalling. A number of studies investigating EpoR activation by its native ligand EPO and engineered peptides seem to suggest that orientational changes between the two EpoR molecules affect signal strength [257]. It was unclear, however, if these observations were due to geometrical differences or binding affinity. It was suggested that an EPO agonist peptide (EMP-1) could be modified into a non-activating peptide (EMP-33) by chemically modifying EMP-1. X-ray structures of these peptide ligands in complex with the extracellular domains of EpoR showed homo-dimeric complexes [257], but it was identified that the non-functional EMP-33/EpoR-dimer angle differed by 15 degrees when compared to the agonistic EMP-1/EpoR-dimer [258]; and Figure S1A in [24]). The abrogation of signalling activity by EMP-33 was considered to be due to this change in EpoR ECD dimer orientation. Here, we show that this might not be the case.

This study also demonstrates that the signalling process can be mirrored and differentially modulated with diabodies that act as surrogate cytokines. It has been shown that the receptor dimer architecture, specifically the orientation of the C-terminal domains with respect to each other can be manipulated with diabodies, which in turn provides the capability to selectively enhance or taper down signalling. Here, four diabody/EPOR complexes with distinct geometries were investigated. The diabodies were tested for their ability to initiate EPO signalling and high quality structures for the most and least active diabody were solved. Interestingly, there was an inverse relationship between EPO signalling strength and the distance between EPOR C-termini in the diabody/EPOR crystal structures.

Due to concerns that flexibility in the hinge connecting the two VH/VL domains in

the diabody may have an important effect on this mechanism, we simulated two selected structures. I performed Natural Move Monte Carlo on the two diabody structures in order to investigate their dynamic range (Figure 5.2). Specifically, I assessed the flexibility in the two linkers connecting the VH/VL modules of the diabodies.

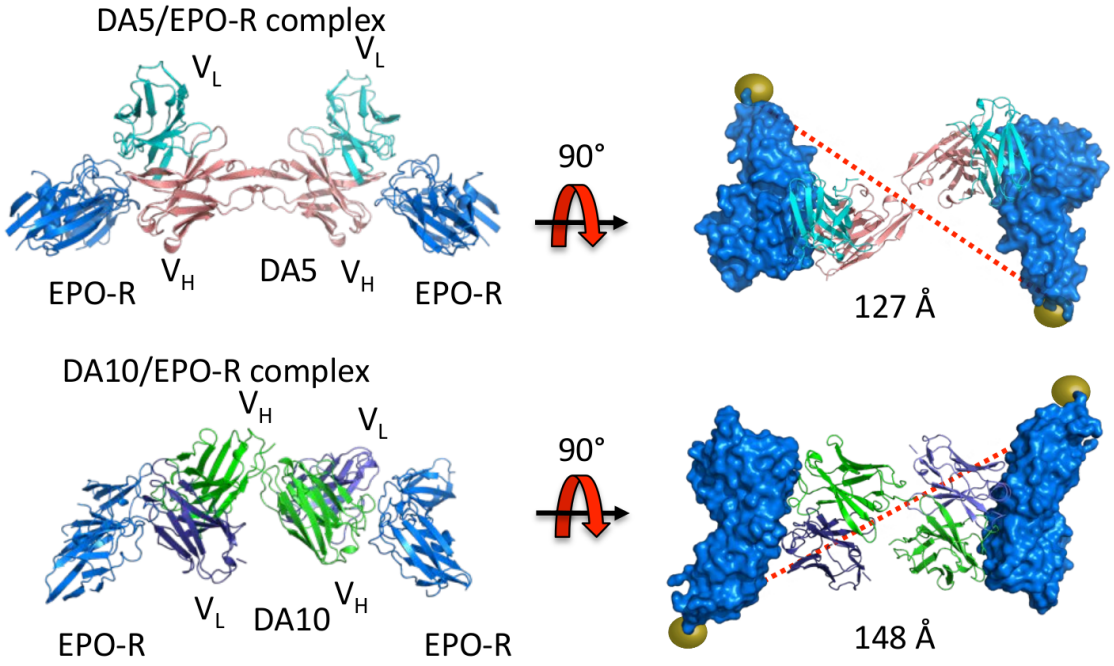

Figure 5.2: X-ray structures of the two diabody/EPOR complexes that were simulated (DA5 VH/VL in salmon/cyan, DA10 VH/VL in blue/green). The initial distances between the two EPOR c-termini (dotted red lines) are shown. Adopted from [24].

5.3 Results

5.3.1 Experimental results by Garcia lab

Moraga et al. studied the biological activity of these peptides on EpoR reporter cells, which enabled them to test signalling by receptor phosphorylation. First, they synthesised EMP-1 and EMP-33 with EpoR binding KDs of 1 mM and 50 mM for EMP-1 and EMP-33, respectively. The low binding affinity of EMP-33 raised the question whether the reduced activation was caused by the low occupancy of the receptor on the cell. The activity of both peptides regarding signalling and receptor dimerisation on cells was tested at a range of concentrations. Using 10 mM, only the EMP-1 peptide triggered

EpoR phosphorylation (and dimerisation) at concentrations similar to those exhibited by EPO (Figure 5.3). When higher concentrations of peptide were used (100 mM), EMP-33 showed a similar level of EpoR homo-dimerisation and phosphorylation as EMP-1 and EPO (Figure 5.3). Thus, when EMP-33 is applied at concentrations that homo-dimerise EpoR on cells, the dimer configuration of the EMP-33/EpoR complex is amenable to signal activation. The difference in signalling strength exhibited by the synthetic peptides seem to be mainly due to their different binding affinities to EpoR.

5.3.1.1 EPOR dimerisation and activity induced by EPO peptides and diabodies

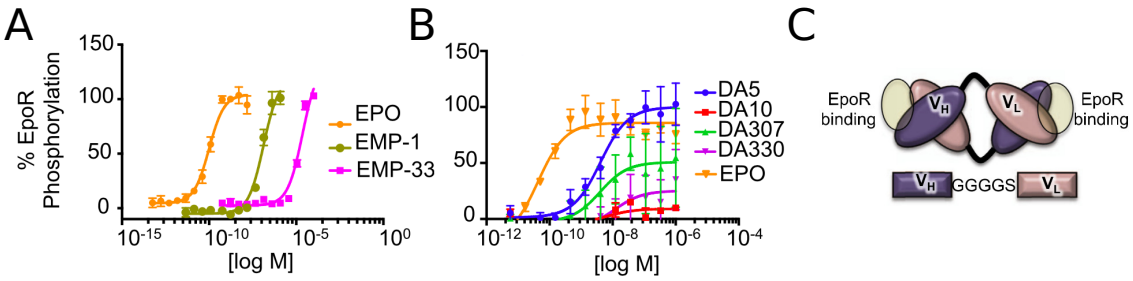

Figure 5.3: EpoR phosphorylation induced by EPO agonist peptide and diabodies. **A** Levels of phosphorylation promoted by EPO agonist peptides EMP-1 and EMP-33 at the shown concentrations. **B** Levels of phosphorylation promoted by diabodies DA5, DA10, DA307 and DA330 at the shown concentrations. **C** Schematic drawing of a bivalent diabody molecule. Adapted from [24].

Next, Moraga et al. generated diabodies that could induce much larger structural changes in the EpoR dimer to systematically test the relationship between dimer architecture and signalling. It was reasoned that these ligands, which consists of two covalently attached antibody variable domain fragments (Fvs) with two binding sites, could dimerise and activate signalling of EpoR. Furthermore, diabodies may be rigid enough to allow crystallisation with EpoR in order to visualise their respective topologies (Perisic et al., 1994). Antibodies have been shown to activate cytokine receptor signalling in many systems, probably by receptor dimerisation. However, the structural heterogeneity of native antibodies has hindered a structural analysis of antibody/EpoR complexes. Here, four known anti-EpoR antibodies were synthesised and their VH and VL domains rearranged into diabodies (Figure 5.3C). All diabodies bound EpoR with similar affinities (Figure S2

in [24]) and homo-dimerised EpoR to a similar extent, albeit less efficiently than EPO (Figure 1D in [24]). However, they induced phosphorylation of EpoR with different efficacies; from full activation (DA5) to weak partial activation (DA10) (Figure 5.3).

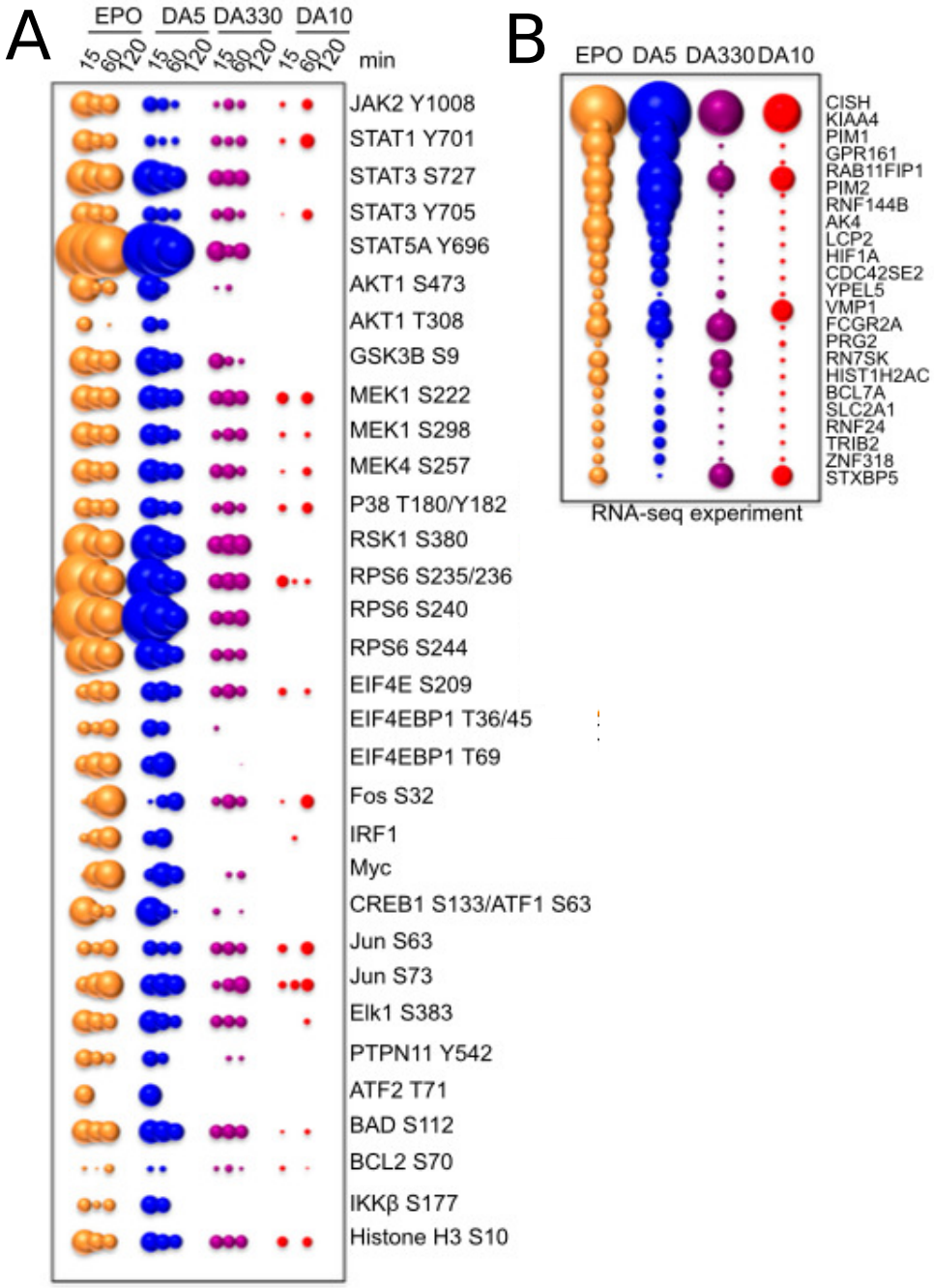

Figure 5.4: **Different EpoR diabodies induce differential signal activation. A** Bubble plot showing the pathways induced by EPO and the three diabodies at the indicated time points in UT-7-EpoR cells. The strength of the signal is represented by the size of the bubble. **B** Genes that were expressed as a result of EPO and diabody treatment of MEP cells for 2 hours. Expression levels are indicated by the size of the bubbles. Adapted from [24].

5.3.1.2 Differential signal activation

'Biased signal' activation has been shown in G-protein-coupled receptor (GPCR) ligands, where a single GPCR can differentially induce signalling pathways (e.g. beta-arrestin and G protein), depending on the ligand (Drake et al., 2008). Here, it was investigated if similar differential signalling could be detected in the EPO-EpoR complex. 78 different signalling molecules were tested (Table S1 in [24]) by phospho-flow cytometry. EPO as well as the diabodies activated 33 signalling proteins, including some of the STAT family, MAP kinase family, and PI3K family (Figure 5.4). Furthermore, upregulation of EPO-induced transcription factors such as Myc, cFos, IRF1, and Elk was observed (Figure 5.4). In concordance with the previous observations, the signalling activities of the three diabodies ranged from full activation for DA5 to partial activation for DA330 and non-activation for DA10 (Figure 5.4). The diabodies did not activate the 33 signalling molecules to the same level.

In order to investigate how different signal activation levels affected gene expression, RNA sequencing was performed (Figure 5.4D). In agreement with the signalling data shown above, the gene-induction activities demonstrated by the diabodies matched their relative signalling efficacies (DA5 > DA330 > DA10) (Figure 5.4D).

5.3.1.3 Model of diabody action on JAK interaction

EpoR dimerisation generally results in the activation of intracellular, non-covalently associated Janus kinases (JAKs). Their subsequent phosphorylation induces the STAT signalling pathway, which is responsible for regulating gene expression and determines cell fate [254]. Figure 5.5 shows a schematic depiction of the mechanism by which DA10 may increase the distance between the two JAK subunits, thereby affecting signal activation.

5.4 Methods

Computational simulations were initiated from the initial structures of the DA5/EpoR and DA10/EpoR complexes (Figure 5.2). All missing linkers have been built using MOD-

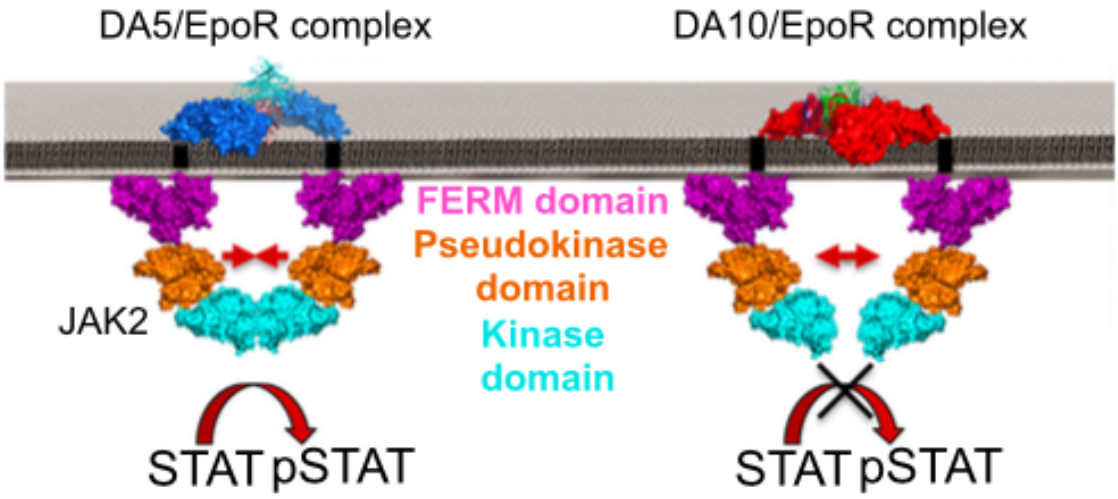

Figure 5.5: **Proposed mechanism by which the diabodies modulate EpoR signalling activity.** The large gap between the two subunits caused by diabody binding changes the location of the JAK2 molecules, thereby affecting their ability to activate each other and initiate signalling. Adapted from [24].

ELLER 9.12 [225] and the MMTSB toolset [259]. The original structures were left unchanged and the loops between the VH C-termini and VL N-termini were modelled. Two hundred loops for each of the five residue stretches (GGGGS) were generated and the models with the lowest MODELLER scores were selected. These completed structures were coarse-grained using a simplified protein model [21] that has been shown to predict all known fold topologies and was successfully applied to probe functionally relevant conformational changes in large molecular machines such as chaperonines [22]. The structures were then fed into our Natural Move Monte Carlo (NMMC) [18] protocol. In NMMC the system is partitioned into independently moving segments (that can be part of the same chain) and melting regions. The independent (translational and rotational) motion of segments may break the molecular chain that is restored by a chain closure algorithm [18] applied on residues in the melting region. This technique has been successfully used to study conformational changes on protein [22] and RNA [19] assemblies and recently for the prediction of primary chromatin structure [30]. To evaluate how diabody flexibility in the linker region will affect the distance distribution of the EpoR ligands, the melting region comprised the three center residues along the linker regions of the two diabodies. The choice of using three residues was made as a compromise between maximizing

the conservation of the initial structures (e.g., experimental information) and providing sufficient flexibility in the linker regions. To further accelerate conformational sampling efficiency, we used parallel tempering [76], in which NMMC simulations were performed in parallel using six replicas that span a temperature range from 300K to 529K (300, 336, 376, 421, 472, 529). The simultaneous propagation of the six systems was performed for 2,000,000 Monte Carlo iterations and the inter replica exchange probability was set to 0.1. The acceptance rate of propagating conformations within individual replicas and the inter replica exchange ratio were 0.3 and 0.1, respectively. All conformational statistics were collected from the system at 300K. The simulation can be reproduced by following the tutorial available at http://www.cs.ox.ac.uk/mosaics. Experimental methods can be found in [24].

5.4.1 hNMMC simulations

We performed hNMMC studies to investigate the relationship between the EPOR C-terminus/ C-terminus distance as a function of the diabody hinge angle on the full agonist DA5 and the non-signalling DA10. The results of these studies show that the diabody hinge angles appear to be in energy minima and sample only a small range of alternative conformations (i.e. distances) around those seen in the crystal structures (figure 5.6). The sampling of these alternative conformations has only minimal consequences on the inter-EPOR distances, and the distances we measured from the crystal structure correlate with signalling outcome even when flexibility is considered. However, since differences in both the EPOR/diabody docking angles, and the distances between EPOR C-termini in the dimeric complexes were observed, it cannot be certainly established whether distance or geometry/topology, or a combination of both factors, is responsible for the differences in signalling between the complexes.

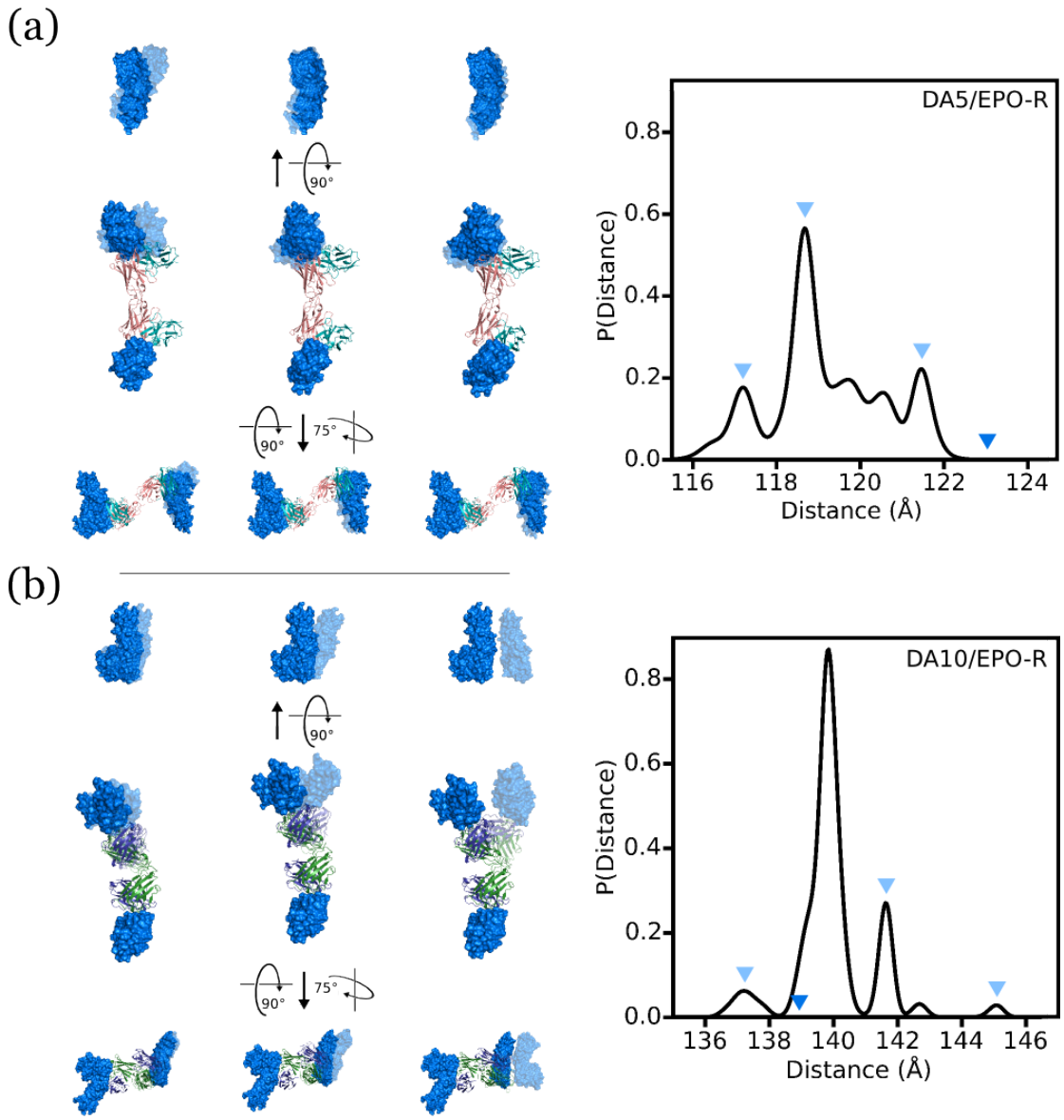

Figure 5.6: *Simulating the linker flexibility in DA5/EPOR and DA10/EPOR complexes.* Each image column shows the original X-ray structure (opaque) overlaid with the representative structure (translucent) from one out of the three major conformational clusters that were generated by Parallel Tempering Monte Carlo simulations. Each row presents a new orientation of the structures. The density plots present the probability of conformations as a function of the distances between the EPOR C-termini of diabody/EPOR. Blue and light blue arrows mark the distances in the X-ray and simulated structures, respectively. The left, middle and right image columns present the structures that have corresponding distances indicated by the left, middle and right light blue arrows on the conformational probability density plots. (a) DA5/EPOR. (b) DA10/EPOR.

5.5 Discussion

Generally, signalling by transmembrane receptors that contain ECDs with ligand binding sites occurs in response to dimerisation (Klemm et al., 1998; Stroud and Wells, 2004).

Ligand engaging ECDs are structurally independent from the intracellular signalling molecules such as Kinase domains, and linked through a transmembrane helix. Thus, the intracellular domains are likely to detect ligand binding by changes of receptor orientation and distance that are transmitted to the membrane-proximal part of the protein. However, the degree to which ligand binding can affect signalling by altering ECD orientation is unclear. In GPCRs, however, this is already known to exist; ligand binding to the transmembrane helices changes their conformation within the plane of the membrane. This mechanism has been used by the pharmaceutical companies for small-molecule drug development. In this study, we asked if ligand-induced conformational changes in the receptor dimer can perform a similar role to the diverse types of conformational changes observed in GPCRs.

The impact of orientational changes on signalling has been widely discussed but has remained speculative until now [260–262]. The effects of mutated and genetically modified receptors have revealed the structural involvement of the extracellular domain in signal modulation [256]. However, for this mechanism to be exploited effectively in therapies, surrogate ligands that induce alternative signalling outputs in native receptors are required. Here, we used diabodies as they have been shown to act as cytokine receptor agonists and are expected to cause large structural changes in dimer geometry. Even though it has been demonstrated that antibodies can elicit a number of different signalling outputs through cytokine receptors [255, 263], they have proven to be elusive structural targets due to their flexible nature. Diabody structures are more constrained than antibody structures, which makes them more amenable for X-ray crystallography.

This behaviour has also been observed in agonistic cytokine-receptor complex structures that exhibited a diverse range of dimeric ligand-receptor configurations [264]. Thus, we found that large rearrangements of EpoR dimer topology are required to modulate signalling output. This approach may present a viable strategy for the regulation of other dimeric receptor systems, where the role of the ligand is to dimerise, and reposition the receptors.

Due to the commonly encountered adverse effects of cytokines and growth factors

as therapeutic agonists, the 'tuning' of receptor signalling activity may present a viable strategy to reduce toxicity while maintaining efficacy. While the underlying molecular mechanisms by which the diabodies presented here modulate intracellular signalling remain unclear, the signal tuning effects are clearly caused by changes in extracellular receptor dimer distance and orientation.

Diabodies are practical as they can be derived from known monoclonal antibodies for human cell-surface receptors. However, a range of different types of synthetic scaffolds could be used to induce dimer re-orientation. A number of new dimerisation geometries could be tested with different dimerising agents to induce a certain signalling property or function. In principle, it appears that receptor ECD dimer orientation represents a new structure-activity parameter for drug discovery for a range of type I and II cell-surface receptors.

5.6 Conclusion

The study's overall conclusion was that a number of different diabodies were successfully used to modulate the signalling activity of EpoR activity by enforcing different dimer geometries. Our computational study of two of the diabody structures showed that their dynamic range was limited in simulation, which strengthens the thesis that the diabodies were able to enforce certain EpoR-dimer geometries, without being subject to a lot of movement.

5.7 Files

The following Mosaics parameters were used for the simulations (definitions are listed in appendix A).

nmmc.input

```
~sim_gen_def[
   \simulation_typ{PT}  PT EEMC SEQ_PT SEQ_EEMC NM DBFR
   \minimize_type{samc}
   \energy_report{2}
```

```
    \num_procs{1} # of processors to be used (default is 1) use replica_number+1
    \prop_type{tors} cart: cartesian, tors: torsional
    \prop_tors_sig{0} 1.e-5 proposal sig 0 < number < 2 Pi, usually 1.e-5
    \prop_rot_sig{0} 1.e-5 {0 <= radian < 2 Pi}
    \prop_trans_sig{0} 1.e-4 {Angstrom >= 0.0}
    \prop_clos_sig{1.e-4} 1.e-3 {Angstrom >= 0.0}
    \replica_number{5} :5 10 number >=0 replicas:0, 1, 2, 3, 4, ....
    \total_step_mc{100000} 2000000 :10   number > 1
    \local_step_md{1} 10     number > 1
    \time_step_md{0.5} :0.5
    \statistics_freq{1000} 200
    \write_energy_unit{Ha} kcal  Ha: atomic unit, kcal: kcal/mol
    \prob_eemc_jump{0.10}   :0.15   number in 0,1
    \temperature{300} 300
    \stsamc_type{trigonom}
    \stsamc_period{1000} 4000 10000
    \stsamc_ampl{1000} 2500
    \stsamc_shift{0}
    \energy_gap{1.12} :1.25 number > 1.0 defined as E_i = n^(i): n = 1.2
    \eemc_disk_size{10}  1000
    \burn_in_B{0} 2
    \burn_in_N{0} 2
    \postprop_minimize{clos} clos
    \postprop_minimize_itmax{8} {integer >= 0}
    \postprop_minimize_energy{bond_bend} {bond,bond_bend,bond_bend_tors,bond_bend_tors_onfo,all}
    \extend_inter{3bond_conn} 3bond_conn, 4bond_conn, off:default
    \cancel_res_inter{off} local, neighbor (includes local), off
    \rinter_switch_length{0.0} 1.0 real inter switching length in A
    \inter_list{none} lnk_list none
    \EEMC_Emin{-0.3} -0.04 in Ha
    \EEMC_Emax{0.0}  0.0 in Ha
    \random_seed{-751488301} {large integer}
]

~sim_mol_def[
  \system_def{residue} primitive
  \cgres_model{KB_3pt}  KB_3pt, off
  \mol_parm_file{../top_3pt_prot_na.rtf}
  \bond_database_file{../par_3pt_prot_na.prm}
  \bend_database_file{../par_3pt_prot_na.prm}
  \tors_database_file{../par_3pt_prot_na.prm}
  \onfo_database_file{../par_3pt_prot_na.prm}
  \inter_database_file{../par_3pt_prot_na.prm}
  \pos_init_file{init.pdb}
  \pos_out_file{last_frame.pdb}
  \atom_pos_file{sampled.pos.pdb}
  \tors_pos_file{sampled.tors_pos}
  \epot_file{sampled.pot_energy}
```

```
  \einter_file{sampled.inter_energy}
  \region_database_file{region.data}
  \energy_term{bond}
  \energy_term{bend}
  \energy_term{tors}
  \energy_term{onfo}
  \energy_term{inter}
]
```

Region file for diabody DA5:

region.data

```
~region[\element_top_type{segment}

    \dependency_type{independent}

    \nseg{3}
    \ncenter{3}
    \segments_firstres{A:1,B:123,C:4}
    \segments_lastres{A:119,B:234,C:189}

    \segments_baseres{A:60,B:178,C:96}

    \centers{A:60,B:178,C:96}

    \prop_trans_sig{1.e-5}
    \prop_rot_sig{1.e-6}
    \prop_trans_sig_freeres{0}
    \prop_rot_sig_freeres{0}
]

~region[\element_top_type{segment}

    \dependency_type{independent}

    \nseg{3}
    \ncenter{3}
    \segments_firstres{A:123,B:1,D:4}
    \segments_lastres{A:234,B:119,D:189}

    \segments_baseres{A:178,B:60,D:96}

    \centers{A:178,B:60,D:96}

    \prop_trans_sig{1.e-5}
    \prop_rot_sig{1.e-6}
    \prop_trans_sig_freeres{0}
```

```
    \prop_rot_sig_freeres{0}
]
```

Region file for diabody DA10:

region.data

```
~region[\element_top_type{segment}

    \dependency_type{independent}

    \nseg{3}
    \ncenter{2}
    \segments_firstres{A:1,B:128,L:3}
    \segments_lastres{A:124,B:237,L:207}

    \segments_baseres{A:62,B:182,L:105}

    \centers{A:62,B:182}

    \prop_trans_sig{1.e-5}
    \prop_rot_sig{1.e-6}
    \prop_trans_sig_freeres{0}
    \prop_rot_sig_freeres{0}
]

~region[\element_top_type{segment}

    \dependency_type{independent}

    \nseg{3}
    \ncenter{2}
    \segments_firstres{A:128,B:1,O:3}
    \segments_lastres{A:237,B:124,O:205}

    \segments_baseres{A:182,B:62,O:104}

    \centers{A:182,B:62}

    \prop_trans_sig{1.e-5}
    \prop_rot_sig{1.e-6}
    \prop_trans_sig_freeres{0}
    \prop_rot_sig_freeres{0}
]
```

Chapter 6

Structural effects of epigenetic marks on DNA structure in silico

This work was submitted as: Krawczyk, K., Demharter, S., Knapp, B., Deane, Charlotte M., & Minary, P. (2017). Structural effects of epigenetic marks on DNA structure in silico. *Bioinformatics (accepted pending minor revisions).*

6.1 Summary

The epigenetic cytosine modifications in mammals are 5mC, 5hmC, 5fC and 5caC. While the biological role of 5mC has been intensively studied, our understanding of the other marks' effects is only just emerging. Experimental as well as computational experiments indicate that isolated epigenetic marks have little influence on DNA structure but in larger numbers can lead to significant conformational changes in DNA. One such experimentally proven example is a newly discovered form, F-DNA, caused by six 5fC marks in a CpG repeat of a small model structure. However, there currently exists no systematic solution to study the full range of effects caused by all possible combinations of epigenetic marks. Here, we present an approach based on Natural Move Monte Carlo to efficiently simulate the conformations of epigenetic marks. We are able to reproduce experimental observations from two recent crystal structures that contain 5hmC and 5fC, respectively. We also show that compared to experiment our protocol correctly identifies the energetically

favourable forms for structures with different epigenetic marks; 5fC is energetically more favourable in the F-DNA form, 5hmC in the B-DNA form. The computational efficiency and straight forward application of this protocol enables comprehensive computational investigation of epigenetic systems.

6.2 Introduction

The clustering of epigenetic marks is thought to cause structural changes in chromatin structure [88]. Furthermore, a recent crystallographic study demonstrated that the presence of six formylated CpG elements can lead to significant conformational alterations, resulting in a novel structural form, referred to as F-DNA (Figure 6.1) [88]. Interestingly, a previous 5fC structure with two formylated CpG elements (PDB: 1VE8) has exhibited unusual base step parameters with similar local rotational and translational values to the ones observed within F-DNA, however the overall structure remained in the B-DNA form. This seemed to suggest that there is a critical number of 5fC marks required to shift the structure from B-DNA to F-DNA.

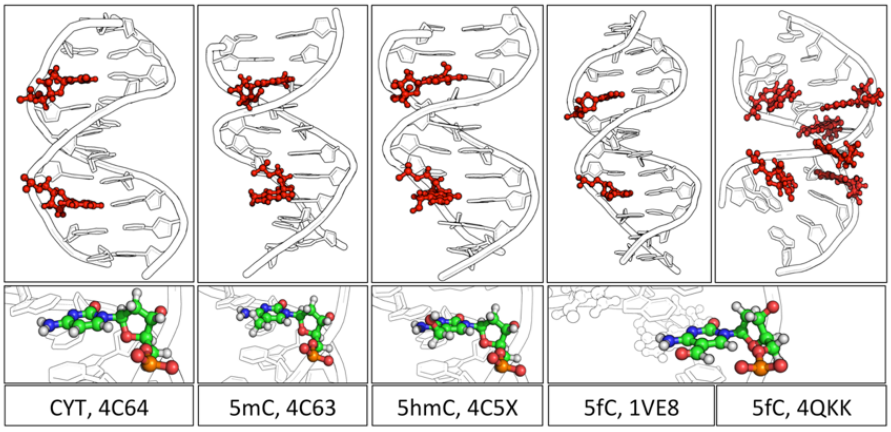

Figure 6.1: *Crystal structures of a short DNA model sequence with different epigenetic marks.* The structure remains in the B-DNA form for the epigenetic marks 5mC (PDB: 4C63), 5hmC (PDB: 4C5X) as well as for 5fC (PDB: 1VE8) with only two marks. However, in the presence of 6 marks the 5fC structure (PDB: 4QKK) takes on the F-DNA form. Figure adopted from [265].

The thorough investigation of structural effects caused by epigenetic marks requires a protocol that allows the efficient study of different collections of epigenetic marks. We

refer to these collections as âĂŸepigenetic makeupsâĂŹ; a set of epigenetic marks in a given DNA structure. The experimental testing of all combinations of epigenetic makeups in a given DNA structure for thermodynamic and structural effects would be slow and expensive as all such epigenetic makeup variants would have to be generated and analysed under identical conditions. Instead, computational approaches such as Density Functional Theory (DFT) [266–268] and Molecular Dynamics (MD) simulations [269, 270] have been applied. However, these methods are computationally very resource-intensive. Here, we present an efficient computational framework that allows for the simulation of a variety of epigenetic makeups for different DNA structures.

The protocol is based on Natural Move Monte Carlo (NMMC) as implemented in the MOSAICS software package. It has been previously used to efficiently explore the conformational space of nucleic acid structures [18, 19] and to predict nucleosome occupancy for long methylated DNA sequences [30]. Here, we use NMMC for the first time to simulate DNA structures containing different epigenetic marks, namely 5mC, 5hmC and 5fC. The simulations are based on recent crystal structures of the Dickerson-Drew Dodecamer containing 5mC, 5hmC and 5fC marks. We show that our results are in agreement with experimental observations [88, 271], including the preferred configuration of 5hmC and 5fC and the thermodynamic preference of F-DNA for multiple 5fC marks rather than a single 5fC mark or multiple non-5fC marks [272].

6.3 Methods

DNA dodecamer X-ray structures with epigenetic marks As the basis of our simulations we used the crystal structures of dodecamers with 5fC (PDB: 4QKK), 5hmC (PDB: 4C5X), 5mC (PDB: 4C63) and unmodified C (PDB: 4C64). The 5fC structure 4QKK only contains one strand, thus the complementary strand was obtained by the symexp command in Pymol [273]. Furthermore, the sequence for structure 4QKK (CTACGCGCGTAG) differed from the sequence of the remaining structures (CGCGAATTCGCG). This was due to the GC repeats that were necessary for the formylation of multiple sites. For simulations with explicit solvation, the TIP3 water model was used and NA+ and Cl-

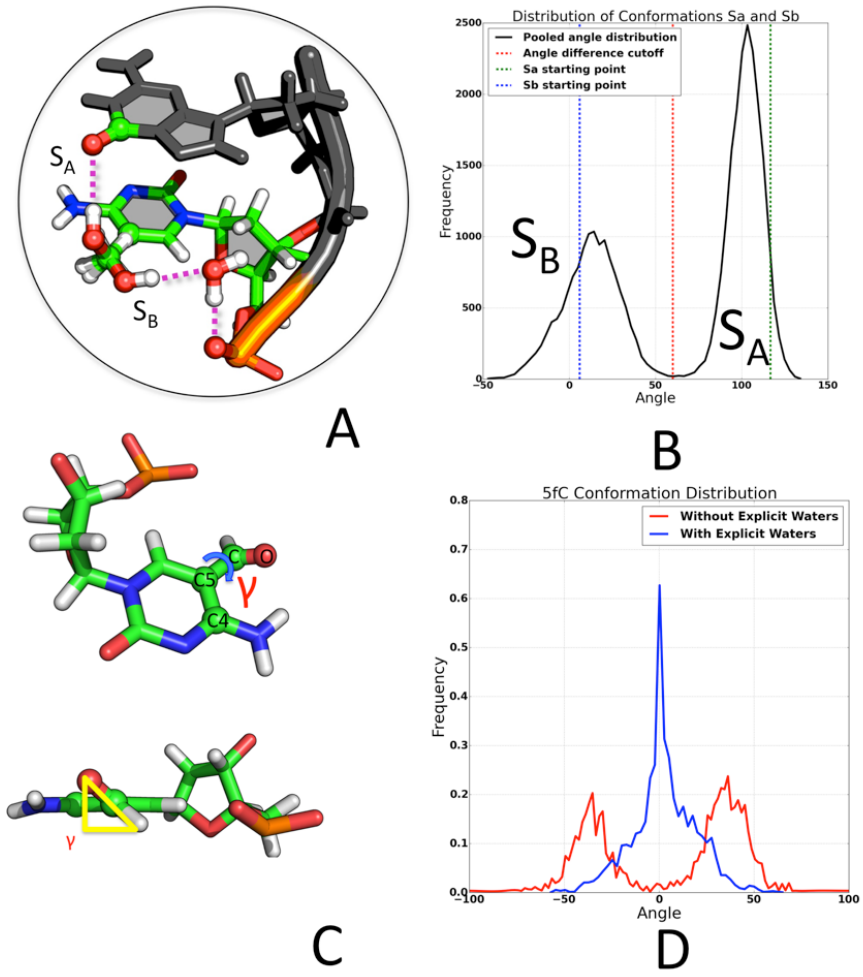

Figure 6.2: *5hmC and 5fC configurations during simulation.* **A** Two predominant configurations were observed for 5hmC in the 4C5X crystal structure: one where the hydroxy-group is oriented towards the adjacent guanine (configuration A, 70% occupancy), and the other where it is pointed towards the phosphate backbone (configuration B, 30% occupancy). **B** Distribution of 5hmC angles generated during simulation. The two modes of the distribution are centred closely to the angles of configuration A and B observed in the crystal structure. The proportion of configuration A observed during simulation is 60%, which is close to the 70% predicted from the experimental structure (PDB: 4C5X). **C** For the analysis of the 5fC orientation we defined a new dihedral angle along C4-C5-C-O, called gamma. **D** We performed simulations with explicit and implicit waters. The distribution of gamma angles for both simulations is shown. in explicit water the 5fC configuration remained in the position observed in the crystal structure, whereas in implicit water a bimodal distribution centred around -40 and 40 degrees was observed. Figure adopted from [265].

ions were added to neutralise the system.

B-DNA model generation For the comparative study of B-DNA vs F-DNA stability it was necessary to obtain a hydrated B-DNA model. The B-DNA models were created

using the Make-NA service [274]. Epigenetic marks were added by threading atoms to structure as described previously [30]. The orientation of the formyl in the model was identical to the orientation of formyl present in the structure of Raiber et al. (PDB: 4QKK) [88].

Simulation details Simulations were carried out at 300K temperature. We used the Amber bsc0 potential [194] as this energy function showed good agreement with experimental results in previous MOSAICS epigenetics applications [30, 32]. We used a distance-dependent dielectric for bulk solvent effects together with explicit waters to model water mediated hydrogen bond interactions. The explicit waters were used as crystal waters where available (F-DNA) or added using VMD [275] (B-DNA). The parameters together with tutorials are available on our website (Epigenetics page on http://www.cs.ox.ac.uk/mosaics/).

Epigenetic makeup The inverse epigenetic problem depends on the set of considered types of epigenetic modifications, $\hat{\mathrm{I}}\mathrm{e}$ and the set of possible locations, Ω_l that may host a modification. For example, assuming that F-DNA is the template one may choose two 'types' of epigenetic modifications, $\Omega_e = f, \oslash$ (formyl and lack of modification) at all 6 native locations in the CpG island, $(\mathrm{CpG})_3$. Therefore, $|\Omega_l| = 6$ and $|\Omega_e| = 2$ (refers to the size of the set), giving rise to 2^6 possibilities. Unfortunately, energies related to individual epigenetic makeups may not be directly comparable because each epigenetic makeup defines a new system. Instead, each epigenetic makeup can also be threaded to a reference structure (of the same system) that assumes a conformation distinct from the template. For F-DNA the reference structure can naturally be chosen as a straight B-DNA form. The energy difference between the template and reference structures accommodating the same epigenetic makeup is a good indicator of how much the epigenetic makeup stabilises the template. In our example we can calculate the energies for all 2^6 epigenetic makeups (out of which 24 are unique if we disregard symmetry) on both F-DNA (template) and B-DNA.

6.4 Results

We conducted two sets of simulations, one aimed at understanding the configuration preferences of the individual epigenetic marks 5mC (PDB: 4C63) and 5hmC (PDB: 4C5X) and the other at establishing the thermodynamic properties of a DNA dodecamer with up to six 5fC marks (PDB: 4QKK). Our simulations are in concordance with the experimental results as we correctly identified the energetically favourable structures for different sets of epigenetic marks. Based on these results we suggest epigenetic makeups for a dodecamer that may represent transition points at which the B-DNA form turns into the F-DNA form and vice versa.

5hmC and 5fC configurations The 5hmC crystal structure was resolved with the hydroxymethyl group of the epigenetic mark in two different configurations. Figure 6.2A shows the predominant configuration SA (70% occupancy), oriented towards the O6 oxygen of the adjacent guanine. Configuration SB (30% occupancy) formed a water-mediated hydrogen bond with the phosphate backbone. For validation purposes we attempted to reproduce these angle in our simulations.

We performed simulations starting in either the SA or SB configuration and recorded the 5hmC dihedral angle (C4-C5-OH-HO). We pooled the trajectories from both sets of simulations and the angle distributions are shown in 6.2. If there was no bias for configuration SA or SB in our simulations, or there was no switching between the two, we would have observed a bimodal distribution with equally populated modes. Instead we observed configuration A more frequently (60%) than configuration B (40%), as determined in the original study (Figure 6.2B). Thus, we did not generate any new configurations, but confirmed the experimentally determined configurations SA and SB (Figure 6.2). Simulations with explicit and implicit solvent led to the same result, yet with explicit solvent requiring a longer simulation time.

5hmC has local structural effects on DNA structure in simulation We simulated three structures of the Dickerson-Drew Dodecamer, with 5mC, 5hmc and C at po-

sition 9 in the CGCGAATT(C)GCG sequence (both on the forward and reverse strand), to study the effects of these epigenetic marks on the local DNA structure. We performed the simulations as described in the methods section below.

We analysed the results of our simulations using X3DNA [221]. The results showed that the helical parameters of the three sets of simulations differed only slightly, which is in line with experiments that show that individual modifications have little effect on the geometry of the DNA.

5hmC was shown to reverse the stabilising effect of 5mC. As this is likely due to subtle structural effects, we looked at possible mechanisms by which 5hmC may alter local base-pair geometry. For example, the O6 oxygen of the neighbouring guanine is amenable to a hydrogen bond interaction with 5hmC, which could affect the neighbouring base-pair.

Thus, we calculated the distances for the three hydrogen bonds of the 3'-adjacent G:C base-pair (Figure 6.3). If there were no differences between C, 5hmC and 5mC, the distributions of such distances should be similar.

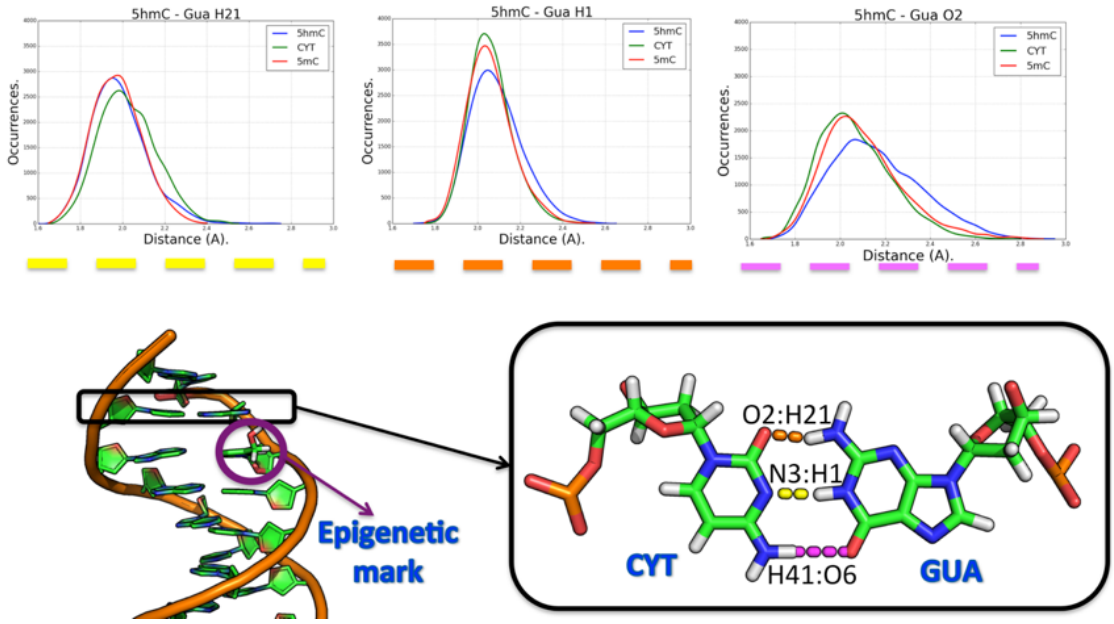

Figure 6.3: *Local base-pair distances for C, 5mC and 5hmC.* The base-pair distance distributions generated during simulation are similar between C, 5mC and 5hmC. The largest difference is found for the base pair adjacent to 5hmc, specificially the C/H41:G/O6 hydrogen bond. The O6 oxygen on guanine forms a second hydrogen bond with the adjacent 5hmC hydroxyl which could be the cause for the change in the distance. Figure adapted from [265].

We observed that the distance distributions for two of the three hydrogen bonds

(G/H1:C/N3 and G/H21:C/O2) were similar between the C, 5mC and 5hmC systems. However, a difference was seen between the 5hmC and 5mC systems for the G/O6:C/H41 distance (Figure 6.3). Interestingly, in our simulations 5hmC interacted with the 3'-adjacent G/O6 directly and via water-mediated hydrogen bonds. This interaction was also proposed by the authors of the crystal structure [271].

Furthermore, the orientation of the hydroxyl on 5hmC appeared to have an effect on the 3'-adjacent C/H21:G/O6 hydrogen bond. When the 5hmC hydroxyl was pointed towards G/O6, the longer C/H21:G/O6 distances were observed than when it was pointed in the opposite direction (Figure 6.4).

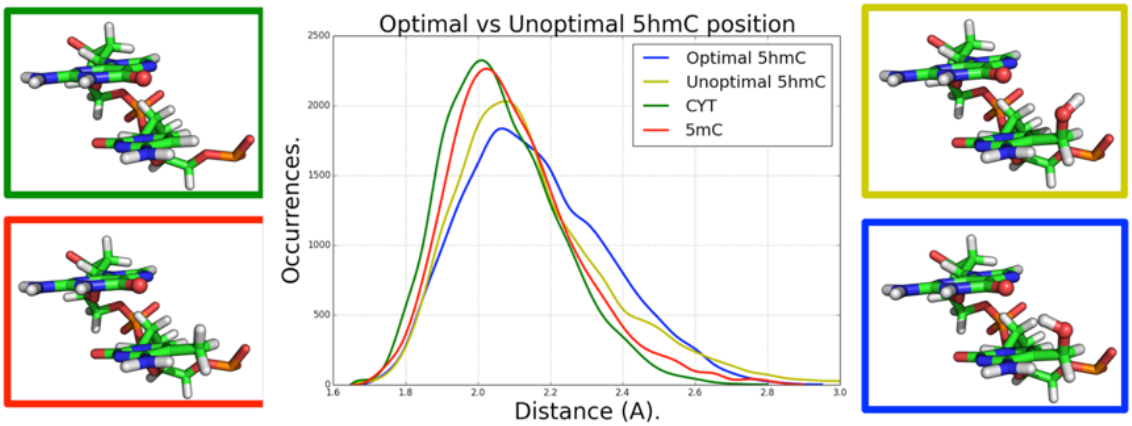

Figure 6.4: *The effect of 5hmC orientation on the neighbouring base pair.* The H21:O6 distance with regards to 5hmC âĂŞhydroxyl position. When simulations are started from an optimal hydrogen bond position for 5mCâĂŞOH:O6, the H21:O6 distribution shows an increase in larger distances than when the 5hmC-OH is placed in a suboptimal hydrogen position at the start of the simulation (yellow). This could be caused by a pulling effect by the 5hmC-hydroxyl on the O6 of the neighbouring guanine. Figure adopted from [265].

These results indicate that 5hmC might have a subtle effect on the local base-pairs, especially in certain sequence-specific contexts. Such minor effects by themselves are not sufficient to cause larger changes but the presence of several epigenetic marks could lead to a more substantial structural shift, which could affect the stability of the molecule.

After demonstrating that our approach could reproduce the experimentally known 5hmC configurations we analysed the orientational preference of a 5fC system (PDB: 4QKK). According to the study the formyl carbonyl oxygen lies in the same plane as the ring of the base (we calibrated the angle as being 0 in Figure 6.2). We started simulations

from structures with randomly selected 5fC angles between -30 and 30 degrees. For simulations in explicit solvent we observed angle distributions centred around $\gamma = 0$ (Figure 6.2). However, in implicit solvent a bimodal distribution was observed with the modes being centred around -40 and 40 degrees. Rare events of 180 degree carbonyl flips were also observed.

We concluded that for the purposes of accurately simulating the 5fC configuration explicit solvent was required. The original F-DNA paper noted that hydration was an important factor in facilitating the F-DNA form [88].

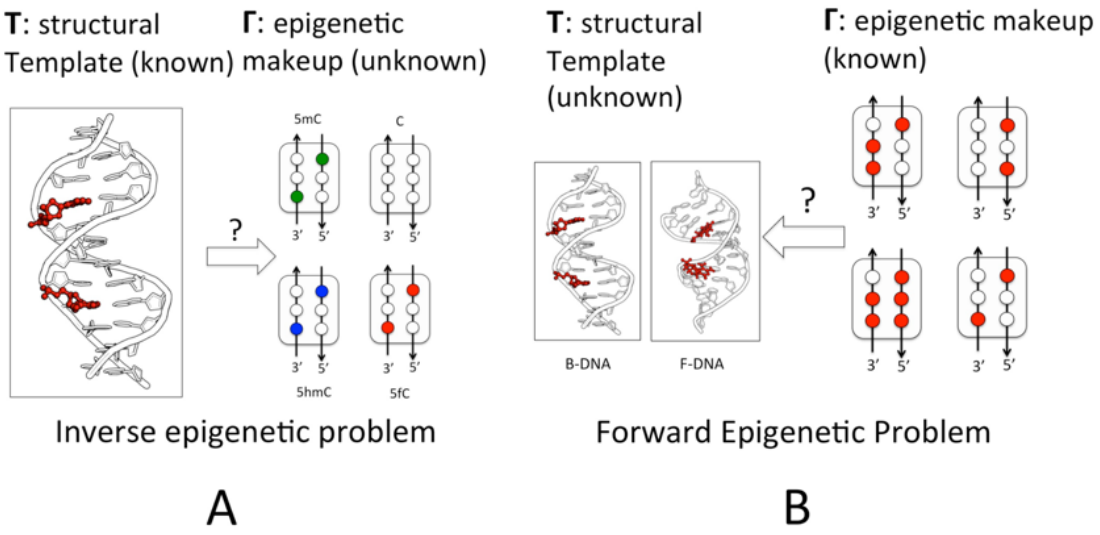

Figure 6.5: *The inverse and forward epigenetic problem.* **A** Inverse epigenetic problem: Given a structural template T, we attempt to identify the epigenetic makeup Γ, which is thermodynamically preferred. **B** Forward epigenetic problem: Given an epigenetic makeup Γ, we predict the structural changes with respect to a unmodified reference structure. Figure adapted from [265].

The structure of F-DNA energetically favours multiple 5fC epigenetic marks

The DNA dodecamer investigated here is not expected to take on the F-DNA form in the absence of 5fC modifications. Therefore, a simulation framework addressing the inverse epigenetic problem should be able to predict this.

After reproducing the experimental configurations of 5hmC and 5fC, we compared the effects that different epigenetic makeups have on a B-DNA and a F-DNA structure. We started by simulating the structures with individual epigenetic marks. As seen in the

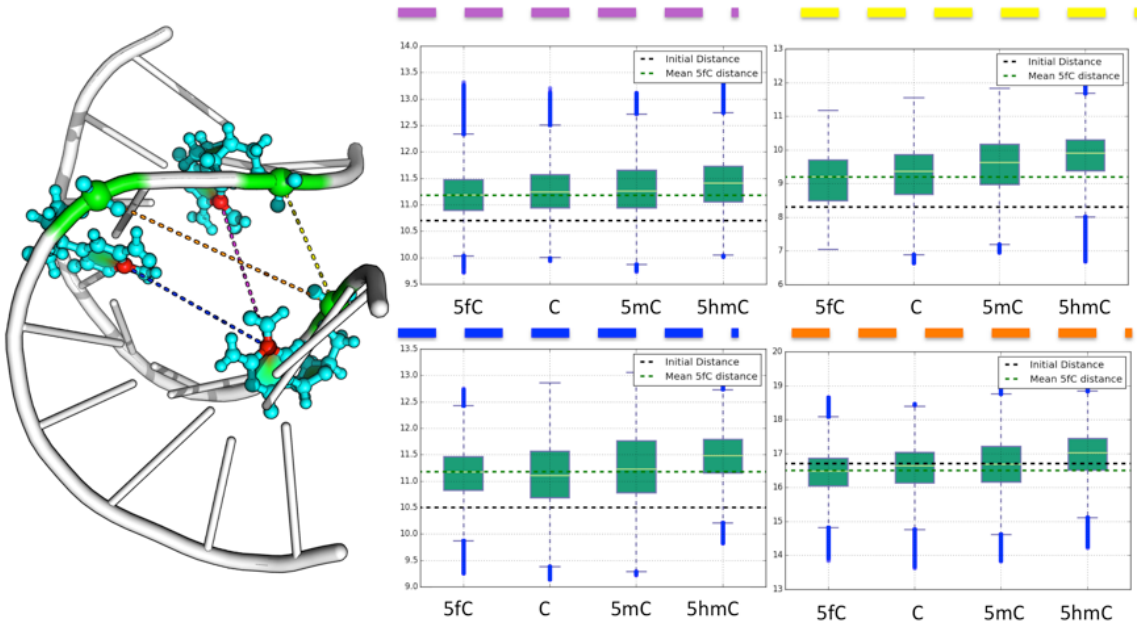

Figure 6.6: *The effect of different epigenetic makeups on F-DNA.* We attempt to solve the inverse epigenetics problem of an F-DNA dodecamer by quantifying its propensity to stay in the F-DNA form given a set of different epigenetic modifications (5fC, 5hmC, 5mC, C). We use the end-to-end distance as a proxy for distinguishing F-DNA from B-DNA (we include the C5 and phosphate atoms of the two pairs of terminal residues in the calculation). These distances are longer in B-DNA than in F-DNA, due to underwinding and bending of F-DNA. We show the distances generated during simulations of a dodecamer with six 5fC, six 5mC, six 5hmC marks and six unmodified cytosines. Box-plots of the end-to-end distances are sorted by upper quartile. The distances are shortest in the 5fC system (green line), indicating a higher preference for maintaining the folded F-DNA shape of F-DNA than the other epigenetic marks. Figure adopted from [265].

experimental studies [271, 276], such isolated epigenetic marks have little to no effect on the helical parameters of the B-DNA structure. Nevertheless, we observed small effects that may be enhanced in the presence of multiple epigenetic marks. Given that there are a number of combinations of epigenetic marks for a given stretch of CpG elements that may affect F-DNA and B-DNA differently we tested a range of different epigenetic makeups using the inverse and forward epigenetic problems (Figure 6.5).

In the forward epigenetic problem we would like to identify the structure with the lowest energy possible, given an epigenetic makeup. In the inverse epigenetic problem, we would like to identify the epigenetic makeup that is most energetically favourable for a given structure. In the first problem we sample the structural space and in the second we explore the space of all possible epigenetic makeups. Formalising the two problems

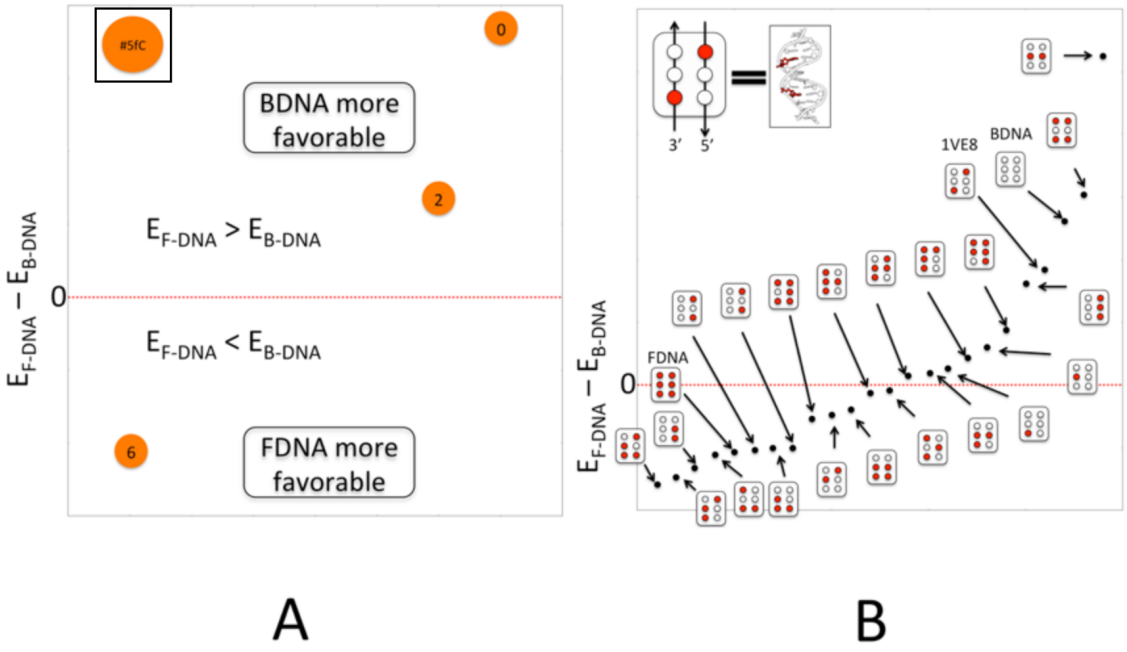

Figure 6.7: *Energy difference between B-DNA and F-DNA structural templates with different epigenetic makeups.* **A** Comparison of 5fC makeups for which at least one experimental structure has been solved. The cases with 0 and 2 5fC marks as well as the case with 6 5fC marks are correctly classified into B-DNA and F-DNA, respectively. The x-axis shows the number of formyl marks (from high to low). **B** The set of all possible 5fC makeups for the given dodecamer is presented (makeups are schematically shown next to each data point). The filled red dots correspond to positions of 5fC marks, the empty dots indicate unmodified C. Figure adapted from [265]. The x-axis shows the rank of the energies of the given 5fC makeups.

offers a systematic framework for investigating the interplay between epigenetic makeups and structure.

In order to test the propensity of the DNA dodecamer to maintain the F-DNA form, given different epigenetic marks (inverse epigenetic problem, Figure 6.5A), we defined four distances that act as a proxy for the curvature of the structure (distances between two pairs of C5 and two pairs of phosphate atoms, Figure 6.6). When the DNA is extended, as it is in B-DNA, these residues are far apart. In F-DNA these distances are smaller.

Using these distances we addressed the inverse epigenetic problem in order to test whether 5fC was the most preferable epigenetic mark for maintaining the F-DNA form. We used the original and generated mutations of the F-DNA structure solved by Raiber et al. [88] to yield C, 5mC, 5hmC and 5fC modified starting structures for our NMMC simulations. For each simulation, we recorded the four distances that we chose as bending

approximations. Our results show that 5fC maintained the closest distances and stayed most bent. This suggests that 5fC is the most suitable epigenetic mark for maintaining the F-DNA state in this dodecamer.

Next, we tested whether we could reproduce the expected energetic characteristics of formylated B-DNA and F-DNA, given a set of different epigenetic makeups. The F-DNA dodecamer should exhibit lower energies in the fully formylated state, whereas the B-DNA template should be most stable in the fully unmodified state. For a given number of formyl modifications, we set up simulations of both B-DNA and F-DNA forms, with the same number of atoms so as to be able to compare the energy of the systems in the two states. The energy difference ($E_{F-DNA} - E_{B-DNA}$) between F-DNA and B-DNA give an indication to which of the two structures is more favourable for a particular epigenetic makeup. Our simulations show that unmodified and two-5fC makeups were energetically more favourable in the B-DNA state. However, the six-5fC epigenetic makeup exhibited the lowest energies in the F-DNA state. Thus in each of the three extreme cases, our simulations confirm the experimental observations (Figure 6.7). Below we use the same approach to characterise epigenetic makeups in terms of the B-DNA/F-DNA transition.

Epigenetic design of DNA Given that our simulations managed to correctly classify experimentally known 5fC make ups into B-DNA and F-DNA, we decided to study in more detail the effect of different 5fC makeups on the B-DNA/F-DNA transition.

We enumerated all possible epigenetic makeups of a dodecamer with six modifiable cytosines. We assumed that each of the six available positions could either be formylated or unmodified (2^6 possible combinations, ignoring mirrored cases). We minimised the energy for each of the makeups in B-DNA and F-DNA and calculated the energy difference between the two states ($E_{F-DNA} - E_{B-DNA}$). We observed a high variation in the stabilising effects of different epigenetic makeups (Figure 6.7). This was to be expected as certain partially formylated makeups may not facilitate perfect B-DNA or F-DNA states. Rather, these makeups may represent transition states between B-DNA and F-DNA and could potentially give insights into how amenable the dodecamer is to structural change caused by epigenetic marks. These results may be considered as a starting point for

selecting interesting epigenetic makeups to be studied experimentally.

6.5 Discussion

There have been a number of recent publications regarding the structural effects of epigenetic marks on DNA that have changed our understanding of their biological roles [88, 270, 271, 276, 277]. Evidence is emerging that these modifications are not just demethylation intermediates but also have functional roles in themselves [278, 279]. Here, we presented a new approach for systematically studying epigenetic makeups in relevant sequence contexts, such as CpG repeats.

We have extended the molecular simulation software MOSAICS to include the sampling of 5hmC and 5fC modifications and we show that we are able to reproduce experimentally known 5hmC and 5fC configurations [88, 271]. Based on this result we studied the effects of isolated epigenetic marks. From our simulations it appears that single epigenetic marks only have minor structural effects on DNA, which has also been observed in experimental studies [88, 271, 276]. However, we noted that depending on the sequence context, there could be a larger effect if a critical threshold of epigenetic marks in the right positions was surpassed, such as in the structure of F-DNA. It has been suggested that these context-dependent interactions may arise from water-mediated hydrogen bonds that facilitate the maintenance of the F-DNA structure.

Our results seem to agree with this hypothesis. In our simulations explicit waters contributed significantly to the successful modelling of 5fC in particular. Because of its helical underwinding, the F-DNA structure traps water molecules at the centre of the major groove thereby contributing to a hydration network that leads to its stabilisation.

Nevertheless, the role of the F-DNA structure remains elusive. Understanding the behaviour of F-DNA with different epigenetic makeups may help elucidate its role in chromatin rearrangement and its ability to recruit transcription factors. A similar example is a bent structure of the 5hmC DNA-ngTET complex that seems to assist enzyme access to the 5hmC mark [280]. Investigating the molecular nature of such events requires the exhaustive testing of many epigenetic makeups. Our protocol presents an efficient

computational approach for evaluating the interplay between epigenetic makeups and DNA structure.

6.6 Conclusion

We defined and addressed the forward and inverse epigenetic problems by developing a robust computational simulation framework that is ready-to-use in the MOSAICS software package and available on our website (http://www.cs.ox.ac.uk/mosaics/). Our initial solutions to the inverse and forward epigenetic problems were in agreement with high-resolution crystal structures. Building on this result, we have tested all possible epigenetic makeups in a representative structure (PDB: 4QKK) and assessed their energetic feasibility. These predictions may hold valuable information on the transition between B-DNA and F-DNA and act as a starting point for choosing promising epigenetic makeups to be studied experimentally.

Chapter 7

Customised Natural Moves - Case study 1: Proteins

This work as published in: Demharter, S., Knapp, B., Deane, C. M., & Minary, P. (2016). Modeling Functional Motions of Biological Systems by Customized Natural Moves. *Biophysical Journal*, 111(4), 710-721. http://doi.org/10.1016/j.bpj.2016.06.028

7.1 Summary

Here, we demonstrate the application of the customised Natural Moves protocol on two protein case studies. As described in the previous chapter, the protocol allows the user to design and perform multiple simulation test cases to investigate the causal relationship between structural features and functional motions. I use this protocol to investigate the structural plasticity of the empty MHCII complex as well as the mechanism by which the peptide-loading chaperone HLA-DM stabilises the open form of the MHCII binding groove.

Peptide loading occurs within the antigen-presenting cell and is facilitated by HLA-DM. HLA-DM stabilises the open conformation of the MHCII binding groove when no peptide is bound and catalyses peptide exchange. While a structure of the MHCII/HLA-DM complex exists, the mechanism of stabilisation is still largely unknown.

7.2 The plasticity of the empty MHCII binding groove.

7.2.1 Introduction

The MHCII is a transmembrane protein that presents potentially harmful peptides to CD4+ T-cells [281]. The structure of the peptide loaded MHCII binding groove is well documented [282], however, to date no structure has been solved for the peptide-free MHCII [20] due to its dynamic nature. Several studies suggest that the absence of peptide destabilises the MHCII structure [159, 161]. Using our protocol we investigated the functional motions involved in the destabilisation of the peptide-free MHCII complex.

We designed multiple sets of customised Natural Moves and performed NMMC simulations to study the plasticity of the empty MHCII binding groove. Our simulations suggest that the $\beta 1$ helix can assume a number of transitory states that cause a narrowing of the binding groove in the absence of peptide.

7.2.2 Methods

NMMC simulations were initiated from an X-ray structure of the MHC II (HLA-DR) in complex with HLA-DM at a resolution of 2.6 Å (PDB:4GBX) [171]. The structure was coarse-grained using a 3-point per residue protein model [21]. We generated the MHC II model by removing the HLA-DM part of the structure file. In order to ensure extensive conformational sampling we performed Parallel Tempering using six replicas at temperatures 300K, 336K, 376K, 421K, 472K and 529K. We ran 15 independent repeats for each test case. Each repeat was run for 1,000,000 Monte Carlo iterations. These parameters were chosen so that the acceptance rate within each replica and the inter-replica exchange rate was at least 0.25 and 0.1, respectively. All data were collected from the replica with a canonical temperature of 300K. Distances were calculated with MDAnalysis [283] and binding-groove surface area was calculated using differential geometric analysis as described in [248].

7.2.3 Results

Step I: Define a hypothesis

The literature suggests that there are several structural features that may contribute to the plasticity of the empty binding groove. The C-terminal region of helix α1 has been shown to exhibit a distinct conformation in the absence of peptide by mass spectrometric mapping [165]. This region is also part of the binding site for the peptide-loading chaperone HLA-DM, and undergoes a structural change upon binding HLA-DM [171]. Therefore, we included this structural feature as an area of potential flexibility by introducing a molten zone at the C-terminal end of the α1 helix (α1-1 in Fig 7.1A).

Residues β53-68 on helix β1 are part of epitopes for conformation-sensitive antibodies that are selective for the empty binding groove [164, 284]. This region has been shown to undergo local structural changes by CD spectroscopy [164]. MD simulations and comparison of experimental MHCII structures revealed structural variability around a sharp kink in this region [158, 249, 285]. Given these observations we introduced a further molten zone at the N-terminal kink of the β1 helix (β1-1 in Fig 7.1A).

The second kink on the β1 helix has not been implicated in major structural changes. This is likely due to a disulphide bridge anchoring a conserved cysteine to the β-floor below. However, the segments on either side might still be influenced by flexibility in this kink, so a third molten zone was introduced at this point (β1-2 in Fig 7.1A).

Thus, our hypothesis states that conformational flexibility in the three unstructured regions in the two helices (α1 and β1) contributes to the variability of binding groove width and area in the empty MHCII complex.

Step II: Translate hypothesis into Natural Moves

Our hypothesis on binding groove flexibility provided us with a starting point for defining an initial set of segments, which can undergo 3-body rotations and translations. This resulted in an initial decomposition consisting of five segments (Fig 7.1B). We used secondary structure information to place molten zones between these segments. In this

coarse-grained protein case study we did not include any internal flexibility within the segments.

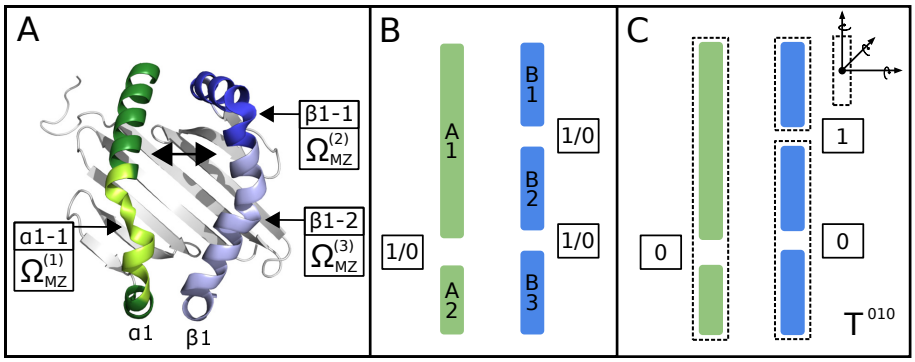

Figure 7.1: **Decomposing the MHCII binding groove into Natural Moves. A** Cartoon representation of the MHCII binding groove (peptide not shown). The three positions α1-1, β1-1 and β1-2 where we defined molten zones $\Omega^{(1)}_{MZ}$,$\Omega^{(2)}_{MZ}$ and $\Omega^{(3)}_{MZ}$ are highlighted by arrows. Helix α1 is shown in green; helix β1 in blue. The HLA-DM binding site is shown in light green (globular domain contacts not shown). The residues that form the epitope for antibodies specific for the empty binding groove are shown in dark blue. The two-headed arrow indicates where the binding groove width was measured for analysis (distance between centres of mass of residues α60–65 and β65–70). **B** The initial decomposition resulting from the choice of MZs is shown schematically. Helices α1 and β1 are shown as two green (A1,A2) and three blue rectangles (B1–B3). Each rectangle represents a helical segment that is linked to adjacent segments by molten zones. Each molten zone can be selectively switched on or off (1/0). **C** Example showing test case ${}^{010}T$. The resulting segments are outlined by dotted lines. The six degrees of freedom (three translations and three rotations) for each segment are shown on the top right.

Step III: Generate test cases

In this simple scenario, each of the three MZs may either be enabled or kept rigid, thereby splitting or grouping two neighbouring segments. As a result there are a total of $2^3 = 8$ different possible test cases that may be generated. For example, test case ${}^{010}T$ refers to a system in which MZs $\Omega^{(1)}_{MZ}$ and $\Omega^{(3)}_{MZ}$ are deactivated and $\Omega^{(2)}_{MZ}$ is activated. This creates 3 regions (1 in helix A and 2 in helix B) as shown in Fig 7.1C. Table 7.1 presents the remaining test cases.

Note that we also introduced permanently activated MZs at the end of the helices to allow for the free movement of all the segments (Fig 7.2).

Table 7.1: **All possible test cases that result from the initial decomposition. The set of segments is shown for each test case.**

Test case	Segments	No. of Segments
^{111}T	$\{S_{A1}, S_{A2}, S_{B1}, S_{B2}, S_{B3}\}$	5
^{110}T	$\{S_{A1}, S_{A2}, S_{B1}, S_{B2+B3}\}$	4
^{101}T	$\{S_{A1}, S_{A2}, S_{B1+B2}, S_{B3}\}$	4
^{011}T	$\{S_{A1+A2}, S_{B1}, S_{B2}, S_{B3}\}$	4
^{100}T	$\{S_{A1}, S_{A2}, S_{B1+B2+B3}\}$	3
^{010}T	$\{S_{A1+A2}, S_{B1}, S_{B2+B3}\}$	3
^{001}T	$\{S_{A1+A2}, S_{B1+B2}, S_{B3}\}$	3
^{000}T	$\{S_{A1+A2}, S_{B1+B2+B3}\}$	2

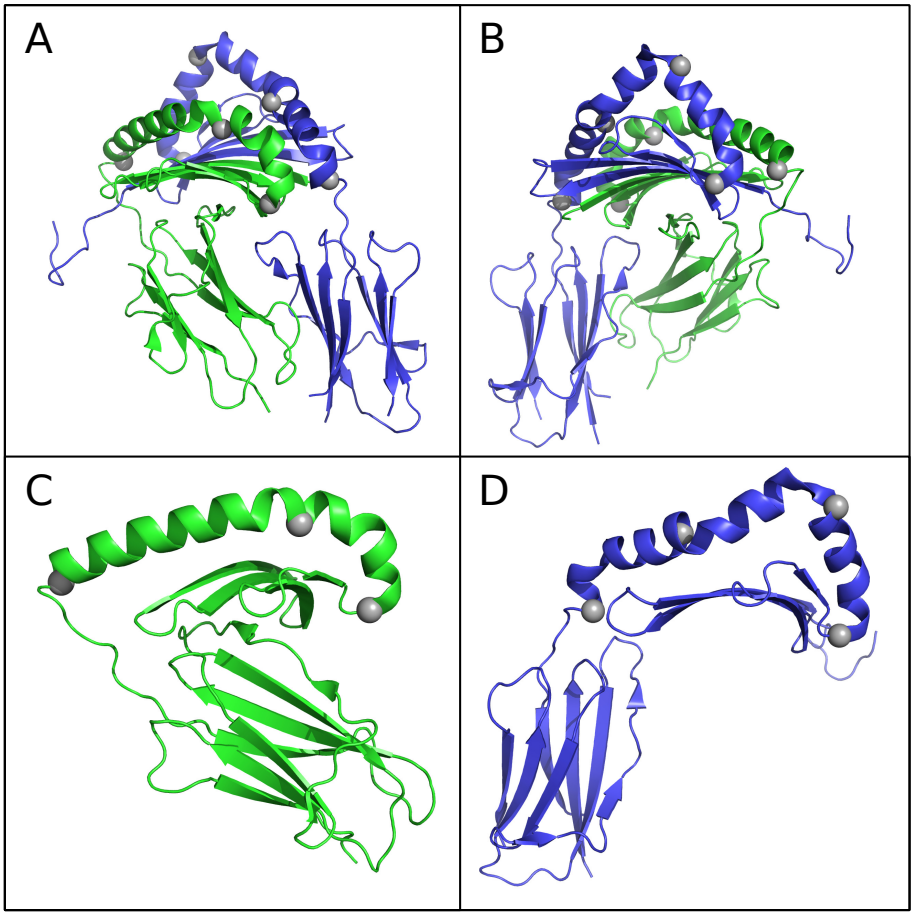

Figure 7.2: **Molten zones in MHCII. A B** MHCII is shown in cartoon representation. Chains A and B are coloured in green and blue, respectively. The molten zones are depicted as grey spheres. **C** Chain A and its three molten zones are shown. **D** Chain B and its four molten zones are shown.

Step IV: Conformational sampling and evaluation

Once the test cases were defined, we used Natural Move Monte Carlo (NMMC) [19] to generate the distributions seen in Fig 7.3.

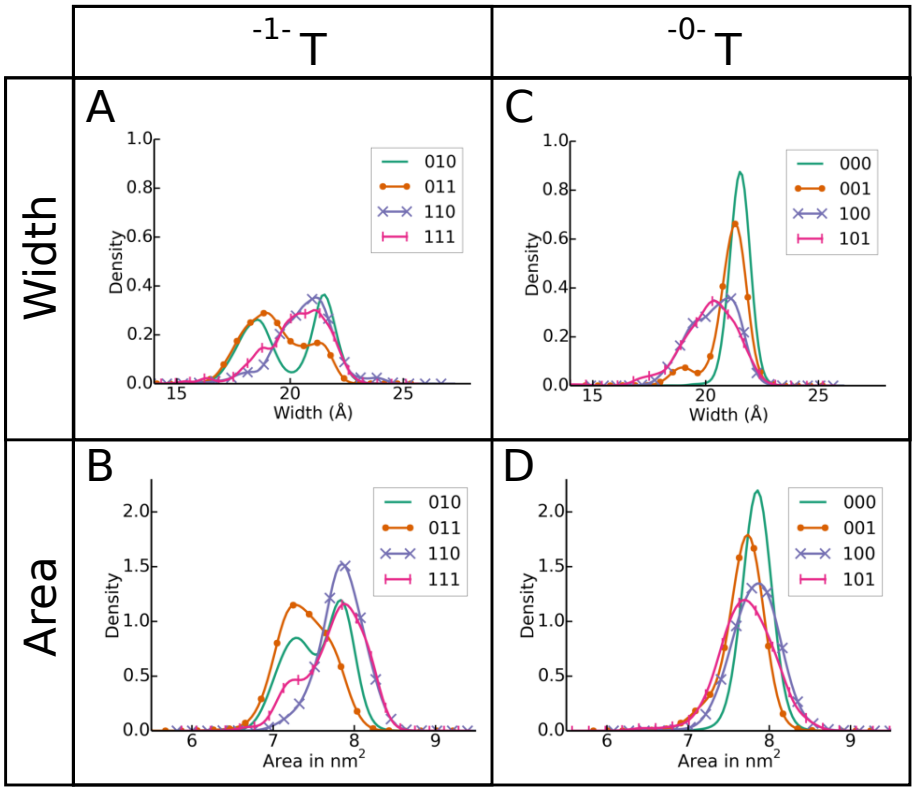

Figure 7.3: **Distributions of the binding-groove width and surface area generated during simulation. A,B** The left column shows test cases in which molten zone $\Omega_{MZ}^{(2)}$ was activated (^{-1-}T). Note the bimodal width and area distributions, which show that the MHCII binding groove takes on a wide and a narrow binding-groove conformation during simulation. **C,D** The right column shows test cases where the molten zone $\Omega_{MZ}^{(2)}$ in the β1-1 kink was deactivated (^{-0-}T). Note, that the binding-groove area remains stable for these test cases.

Fig 7.3 shows the binding-groove width as defined in Fig 7.1A and surface area distributions as calculated in [248] for all eight test cases. For clarity the test cases are shown in two groups. The first group includes the test cases in which $\Omega_{MZ}^{(2)}$ was activated (Fig 7.3A,B). The resulting bimodal width and surface area distributions show that the binding groove readily transitions between a wide and a narrow conformation. Depending on the test case the narrow population is more or less prominent. Test case ^{010}T for example exhibits a distribution with clearly defined wide and narrow populations. Note that the distribution was shifted towards the wide population in test cases ^{110}T and ^{111}T when $\Omega_{MZ}^{(1)}$ i.e. the α1-1 kink was activated. The second group shows test cases in which $\Omega_{MZ}^{(2)}$ was deactivated (Fig 7.3C,D). Some narrowing of the binding groove can be observed for test cases ^{100}T and ^{101}T, but the effect on the surface area is minimal. Generally the

binding groove remains in an open conformation when $\Omega^{(2)}_{MZ}$ i.e. the β1-1 kink is kept rigid (^{-0-}T).

Therefore, our customised Natural Move simulations suggest that the β1-1 kink plays a crucial role in facilitating a conformational change that results in the narrowing of the binding groove.

7.2.4 Discussion

All MHC class II structures with bound pepide (pMHCII) that have been solved to date are structurally highly similar. In the absence of peptide the MHCII is thought to undergo conformational changes [164, 165]. However, presumably due to its "floppy" nature in the absence of peptide [286], the structure of the empty MHCII has not yet been solved by X-ray crystallography. Other experimental techniques have been employed to show that the empty MHCII assumes at least two distinct forms, a peptide receptive and an averse form [159–163]. The receptive form mainly exists immediately after peptide dissociation and turns into the averse form within minutes. Given enough time, however, the averse form can isomerise back to the receptive form [160, 162].

The structural mechanisms underlying the conversion from receptive to averse are little understood. One simulation study suggested that partial unfolding of the α1 helix gives rise to a helical segment that binds the P1 pocket of the groove in a peptide-like fashion [285]. However, this effect was abrogated when the protonation state of the starting structure was adjusted [158, 249]. These studies suggested an involvement of the β1 rather than the α1 helix in the narrowing of the binding groove. In particular, they have shown that the region around the β1-1 kink is highly dynamic [158, 249]. Interestingly, the β1-1 kink is part of an epitope for two monoclonal antibodies that selectively bind the empty and not the peptide-loaded MHCII [164, 284]. Additionally, MD simulations on an empty MHC I complex have suggested that the helix which is the equivalent of the β1-helix in MHCII is responsible for the closing and opening of the binding groove [287].

In our simulations we have observed a similar role of the β1 helix in binding-groove plasticity. Only in test cases where $\Omega^{(2)}_{MZ}$ (the β1-1 kink) was active, was a significant

narrowing of the binding groove seen (Fig 7.3A,B). Previous observations in the literature regarding conformational heterogeneity of residues β53-68 around the β1-1 kink have been made [158, 164, 249, 284, 285], which are concordant with our own results suggesting that flexibility in the β1 helix provided by the β1-1 kink leads to a collapse of the binding groove.

7.3 Modulation of MHCII plasticity by the HLA-DM peptide loading chaperone.

The following work is in review: Demharter, S., Knapp, B., Deane, C. M., & Minary, P. (2017). Stabilisation of the empty MHCII binding groove by HLA-DM: A customised Natural Move Monte Carlo study. *Scientific Reports*.

7.3.1 Introduction

MHCII complexes rapidly take on an inactive peptide-averse state [160]. The peptide-exchange factor human leukocyte antigen DM (HLA-DM) is known to stabilise the empty MHCII complex [288] and promote a peptide-receptive state [289]. A structure of HLA-DM in complex with MHCII has been published [171], however the mechanism by which HLA-DM stabilises the MHCII binding groove is still a subject of debate [290].

Given that one of the key binding sites for HLA-DM on MHCII includes the $\beta 2$ globular subunit [171, 291] and that antibodies detect peptide-induced structural changes in the $\beta 2$ globular subunit [284], it was suggested that the HLA-DM chaperoning mechanism involves the long range transmission of structural changes from the globular subunits to the binding groove (Fig. 7.4)[158]. The hypothesis states that HLA-DM binding causes structural changes in the membrane-proximal $\beta 2$ subunit of MHCII that can be propagated to the binding groove to modulate peptide binding [284].

We evaluated this experimentally suggested hypothesis and investigated the underpinning mechanistic details of the HLA-DM/MHCII interaction with customised Natural Move Monte Carlo (cNMMC) [32] and Normal Mode Analysis. cNMMC is a protocol for hypothesis based modelling using Natural Move Monte Carlo (NMMC) [19], an established molecular simulation method that has been applied to investigate several biophysical research questions [22, 24, 31]. Our results suggest that HLA-DM stabilises the MHCII binding groove in an open conformation via a long range mechanism through the globular domains.

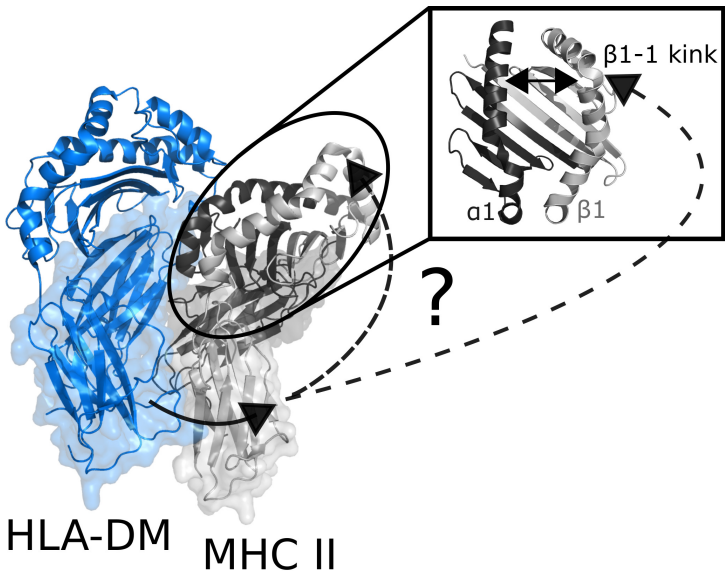

Figure 7.4: **Schematic showing hypothesis of HLA-DM assisted MHC class II binding-groove stabilisation.** HLA-DM (chain C and D in light and dark blue, respectively) binds laterally to the N-terminal side of the $\alpha 1$ helix and the two globular domains (in surface representation) of MHC class II (chain A and B in dark and light grey, respectively). Our hypothesis states that HLA-DM restrains the MHCII globular domains and thereby indirectly prevents the collapse of helix $\beta 1$ in the binding groove.

7.3.2 Methods

Natural Move Monte Carlo Simulations

All simulations were carried out with the MOSAICS software package [292]. All distributions were plotted with matplotlib [293] and pandas [294] using a bandwidth of 0.1. NMMC simulations were initiated from an X-ray structure of the MHCII (HLA-DR) in complex with HLA-DM at a resolution of 2.6 Å (PDB:4GBX) [171]. The structure was coarse-grained using a 3-point per residue protein model [21]. We generated the MHCII model by removing the HLA-DM part of the structure file. In order to ensure extensive conformational sampling we performed Parallel Tempering using six replicas at temperatures 300K, 336K, 376K, 421K, 472K and 529K. We ran 15 independent repeats for each test case. Each repeat was run for 6 x 1,000,000 Monte Carlo iterations each and the replica exchange rate was 0.1. The average acceptance rate within replicas was 0.5 and 0.3 for MHCII and MHCII/HLA-DM simulations, respectively. The average acceptance rate for jumps between replicas was 0.16. All data were collected at a canonical temperature

of 300K. Distances were calculated with MDAnalysis [283].

Normal Mode Analysis

Normal mode analysis was performed with NOMAD [134]. 30 modes were calculated using all-atom representation from PDB accession code 4GBX (without HLD-DM), a distance weight parameter for elastic constant (Å) of 5, a ENM Cutoff (Å) of 10, and an average rmsd (Å) of 1 in the output trajectories.

7.3.3 Results

HLA-DM is known to catalyse the peptide exchange in MHCII molecules. It binds laterally to the α1 region in the binding site and the α2 and β2 globular domains (*+DM* in Fig. 7.7A) and stabilises the peptide-receptive form of the MHCII. The structural mechanism by which this stabilisation occurs is largely unknown. Based on the hypothesis shown in Figure 7.4, we designed a set of customised Natural Moves to investigate a potential involvement of the MHCII globular domains in the stabilisation of the binding groove.

In the previous study we performed a set of customised NMMC simulations and identified that the collapse of the binding groove was largely driven by the plasticity of the β1-1 kink in the β1 helix [32]. Here, we simulated the same set of customised Natural Moves in the presence of HLA-DM using the same structure (PDB accession code: 4GBX - MHCII in complex with HLA-DM) as a starting point for investigating the mechanism underlying MHCII binding-groove stabilisation. In Figure 7.5 the binding-groove width data of Figure 7.3 is overlaid with the data generated during the simulations in the presence of HLA-DM. It clearly shows the stabilisation of the open binding groove in all eight test cases. Interestingly, the distance between the two MHCII globular domains during the simulations was also affected by HLA-DM as shown in Figure 7.6. While the globular domains drifted apart in the absence of HLA-DM, they remained in close proximity when HLA-DM was present.

As test case ^{010}T exhibited a well defined bimodal distribution of closed and open binding-groove states we chose to study the HLA-DM mechanism using this test case.

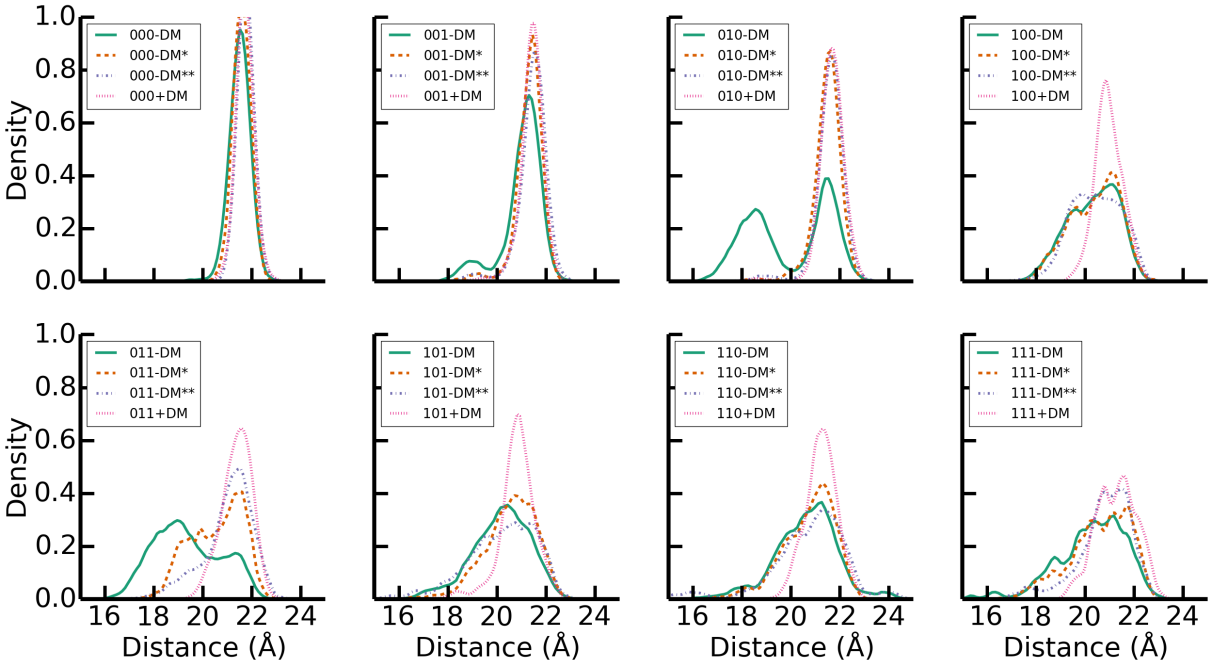

Figure 7.5: Binding groove width for the eight test cases. -DM: Original simulation with independent globular domains, -DM*: *-DM* dataset filtered for compact globular domains, -DM**: Simulations with fixed globular domains, +DM: Simulations in the presence of HLA-DM.

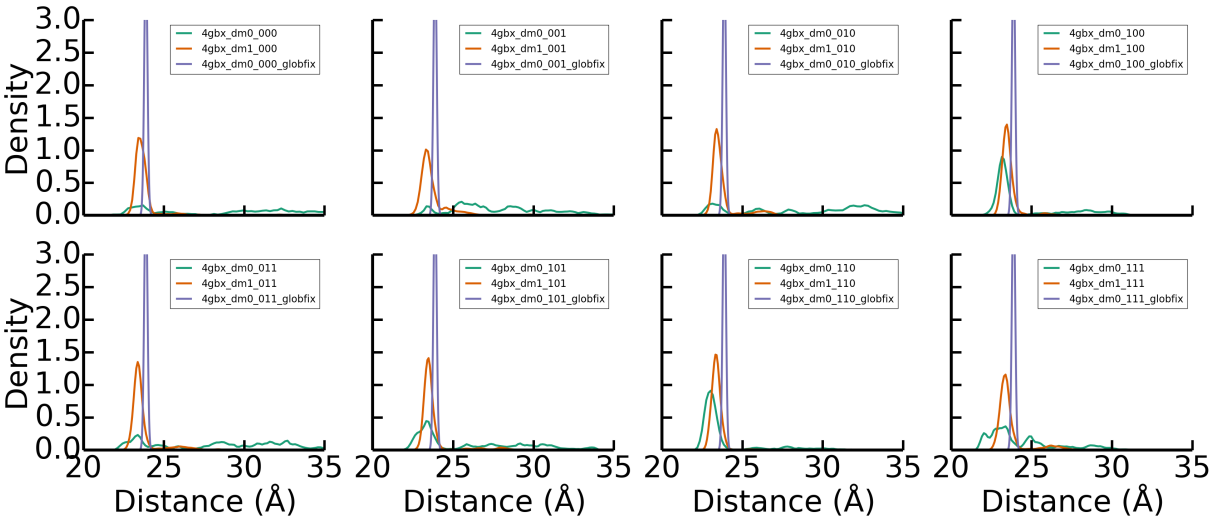

Figure 7.6: Distance between MHCII globular domains $\alpha 2$ and $\beta 2$. Default simulations are shown in green; simulations of MHCII in the presence of HLA-DM in orange; simulations with grouped MHCII globular domains shown in purple. All eight test cases are shown.

To further investigate whether the configuration of the globular domains was linked to the plasticity of the binding groove we generated a subset of the original simulation without HLA-DM *-DM* that excluded states with globular domains that had separated (distance between centres of mass of globular domains α_2 and β_2 <= 26Å and nearest

distance between binding-groove and the β_2 globular domain <= 6Å) from each other (*-DM**, schematically shown in Figure 7.7A). Interestingly, this distribution matched the binding-groove width distribution seen in the simulation with HLA-DM (*+DM* in Fig. 7.7C).

To confirm this result we designed a customised Natural Move simulation by grouping the two globular domains $\alpha 2$ and $\beta 2$ (*-DM***), thereby effectively ensuring that the two globular domains moved collectively throughout the simulation unlike in the *-DM* simulation where they moved independently. This customisation was sufficient to prevent the collapse of the binding groove in the absence of HLA-DM as shown by the *-DM*** distribution in Fig. 7.7C. In fact, the outcome was nearly identical to *+DM* and *-DM**.

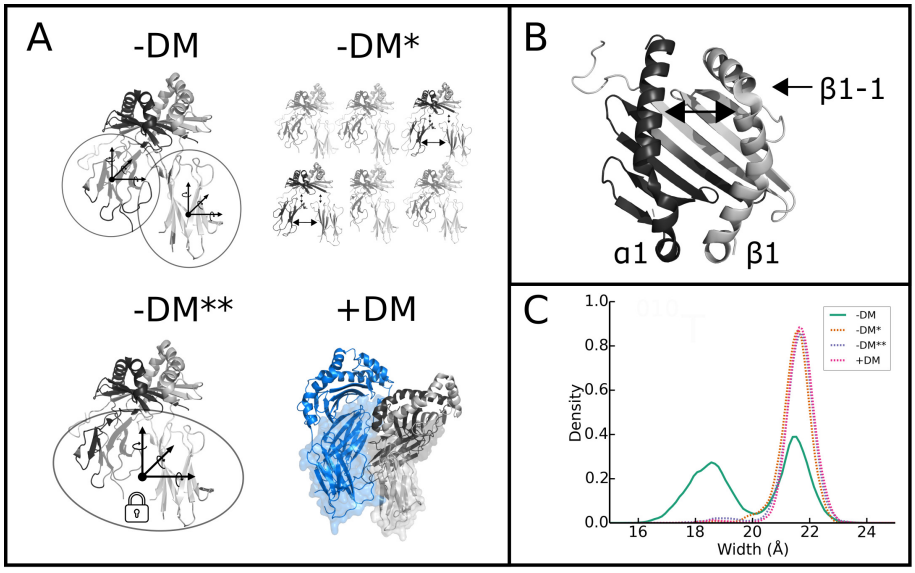

Figure 7.7: **HLA-DM stabilises the open MHCII binding-groove configuration indirectly through the globular domains. A** *-DM*: The two globular domains were allowed to move independently from each other; *-DM**: The *-DM* data set excluding structures with detached globular domains; *-DM***: The two globular domains were propagated as a rigid body unit; *+DM*: Simulation with independently moving globular domains in the presence of HLA-DM. The MHCII complex is depicted in grey (chain A and B in light and dark grey, respectively) and HLA-DM in blue. The membrane-distal $\alpha 1$ and $\beta 1$ domains are shown in cartoon and the globular domains $\alpha 2$ and $\beta 2$ in surface representation. **B** Top view of the MHCII binding groove. The two-headed arrow indicates the position at which the binding-groove width was measured and the single-headed arrow shows the $\beta 1$-1 kink. **C** Distributions of binding-groove widths. The solid green line shows data for simulations in the absence of HLA-DM (*-DM*), revealing a bimodal distribution. The dashed orange line shows a subset of the *-DM* data set filtered for states with compact globular domains (*-DM**). The dashed purple and pink lines show the results for simulations where the globular domains were moved as a rigid body unit (*-DM***) and where HLA-DM was present (*+DM*), respectively.

To further investigate the interplay between binding-groove and globular-domain motion, we performed normal mode analysis (NMA) on the MHCII complex (Fig. 7.8). Mode 18 captured the largest fluctuations in the binding-groove (Fig. 7.8A). It was also the lowest-frequency normal mode that showed a negative correlation between the binding-groove width and the separation of the globular domains as shown in Fig. 7.8B and C (for results of all modes see Fig. S1). Fig. 7.8D illustrates these measures.

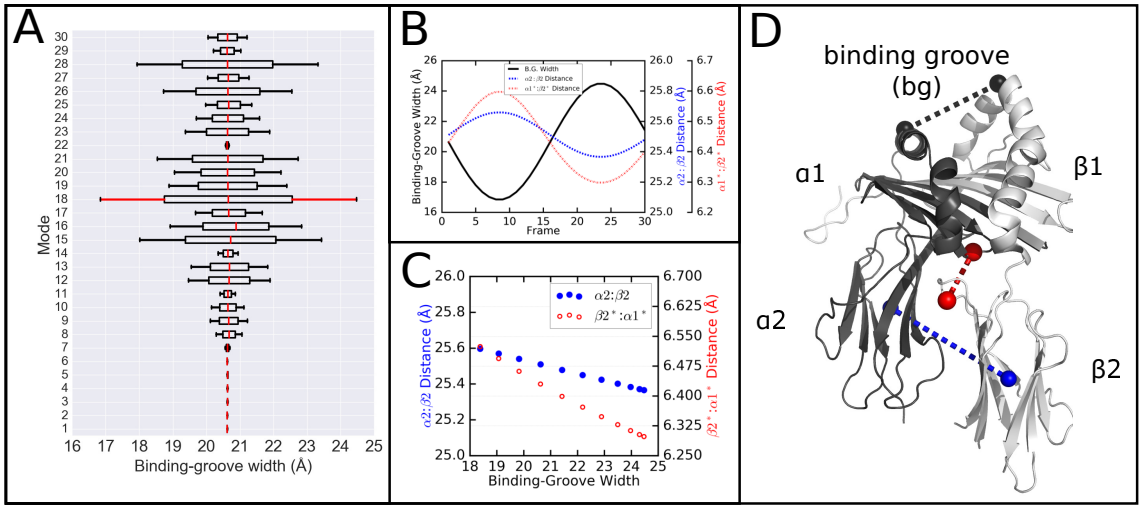

Figure 7.8: **Globular domain and β1 helix motion are linked in normal modes (results for mode 18 are shown). A** Minimum and maximum binding-groove widths for modes 1-30 are shown. Fluctuations were highest in mode 18 (highlighted in red). **B** The binding-groove width (black), the distance between the centres of mass of the two globular domains (blue) and the nearest distance between globular domain β2 and the α1 (red) are shown for mode 18. **C** Correlation plots of the α2:β2 (blue) and α1*:β2* (red) distances plotted against the binding-groove width. Results for mode 18 are shown. **D** A cartoon representation of the MHCII complex (chain A in grey, chain B in white). The following distances are highlighted: binding-groove width (black), distance between the globular domains α2:β2 (blue) and nearest distance between the binding groove and the β2 globular domain α1*:β2* (red).

Thus, our results suggest that the conformation of the MHCII globular domains play a crucial role for the stability of the binding-groove. In the presence of HLA-DM the MHCII globular domains are kept in a compact configurations beneath the binding-groove. In the absence of HLA-DM the globular domains separate, which allows the binding groove to collapse. By introducing customised Natural Moves that propagated the globular domains collectively rather than separately separation was prevented and the binding groove remained in its open state. When the simulation without HLA-DM and with

independent globular domains was filtered for states with compact globular domains, the proportion of open binding-groove states was significantly increased, thus showing a dependency between the globular domain separation and the binding groove. Normal mode analysis corroborated these findings. The mode that captured the largest binding-groove motions showed a similar dependency.

7.3.4 Discussion

Here, we applied customised Natural Move Monte Carlo and normal mode analysis to investigate the HLA-DM/MHCII interaction. We found a possible long range mechanism that implicates the membrane-proximal globular domains in stabilising the open state of the empty MHCII binding groove. Given our calculations, we propose a chaperoning mechanism where in the absence of HLA-DM, the MHCII globular domains change conformation, resulting in a long-range effect which causes the binding groove to collapse. Specifically, it seems that the β1-1 kink on β1 helix plays a major role in this plasticity. HLA-DM prevents this collapse indirectly by restraining the globular domains in a compact conformation and thereby stiffening the β1 helix and preventing its collapse.

One of the key binding sites for HLA-DM on MHCII includes the β2 globular subunit [171, 291] and antibodies detect peptide-induced structural changes in the β2 globular subunit [284]. Carven et al. suggested a HLA-DM chaperoning system involving the long range transmission of structural changes from the globular subunits to the binding groove. Specifically, they propose that HLA-DM binding causes structural changes in the membrane-proximal β2 subunit of MHCII that can be propagated to the binding groove to modulate peptide binding [284].

This led to the hypothesis that binding groove plasticity may be linked to the configuration of the globular domains. We tested this hypothesis in three sets of simulations: **1.** Simulation in the absence of HLA-DM with independently moving globular domains **2.** Simulation in the presence of HLA-DM with independently moving globular domains **3.** Simulation in the absence of HLA-DM with customised Natural Move simulations where the globular were propagated as a unified segment. Simulation 1 exhibited a bimodal

distribution of open and closed binding-groove states, whereas simulations 2 and 3 only generated open states. Thus the immobilisation of the globular domains by the interaction of HLA-DM and the customised Natural Moves have prevented the binding-groove from collapsing. To further strengthen this observation we filtered the conformation trajectories of the simulation 1 for states with compact globular domains. As expected, all closed binding-groove states were removed leading to a distribution resembling that of simulation 2 and 3.

In addition, the mode capturing the largest binding-groove fluctuations in our normal mode analysis showed a negative correlation between the separation of the globular domains and the binding-groove width.

Carven et al. have used chemical labelling and mass spectrometry to characterise residues that are involved in conformational changes in the globular domain [165]. In addition they identified a conformation-sensitive antibody that selectively bound to the globular domains of empty MHC class II proteins [284]. Furthermore, studies comparing MHCII crystal structures have shown conformational diversity in the globular domains, specifically rigid body motions of the β2 domain of up to 15 degrees [285, 295].

These experimental observations together with our results suggest an indirect stabilising mechanism by which HLA-DM modulates the conformational plasticity of the MHCII binding groove through its globular domains. This could explain how HLA-DM affects the stability of the MHCII β1-1 kink, one of the most flexible parts of the binding groove, while binding at the opposite end of the structure.

Thus, we used observations in the literature and our own results to define hypotheses regarding the plasticity of the MHCII binding groove and the mechanism by which HLA-DM may stabilise it. Our protocol allowed us to test and substantiate these hypotheses *in silico* with customised Natural Moves.

While the calculations shown here suggest a potential mechanism by which HLA-DM acts on MHCII, further experiments targeting the globular domains specifically are needed to confirm this hypothesis.

7.4 Files

The following Mosaics parameters were used for all test cases:

nmmc.input

```
~sim_gen_def[
   \simulation_typ{PT}  MIN PT EEMC SEQ_PT SEQ_EEMC NM DBFR
   \minimize_type{stsamc}
   \energy_report{2}
   \num_procs{1} # of processors to be used (default is 1) use replica_number+1
   \prop_type{tors} cart: cartesian, tors: torsional
   \prop_tors_sig{0} 1.e-5 proposal sig 0 < number < 2 Pi, usually 1.e-5
   \prop_rot_sig{0} 1.e-5 {0 <= radian < 2 Pi}
   \prop_trans_sig{0} 1.e-4 {Angstrom >= 0.0}
   \prop_clos_sig{1.e-4} 1.e-3 {Angstrom >= 0.0}
   \replica_number{5} :5 10 number >=0 replicas:0, 1, 2, 3, 4, ....
   \total_step_mc{1500000} 2000000 :10   number > 1
   \local_step_md{1} 10     number > 1
   \time_step_md{0.5} :0.5
   \statistics_freq{1000} 200
   \write_energy_unit{kcal} kcal  Ha: atomic unit, kcal: kcal/mol
   \prob_eemc_jump{0.10}   :0.15   number in 0,1
   \temperature{300} 300
   \stsamc_type{trigonom}
   \stsamc_period{2000} 4000 10000
   \stsamc_ampl{1000} 2500
   \stsamc_shift{0}
   \energy_gap{1.12} :1.25 number > 1.0 defined as E_i = n^(i): n = 1.2
   \eemc_disk_size{10}  1000
   \burn_in_B{0} 2
   \burn_in_N{0} 2
   \postprop_minimize{clos} clos
   \postprop_minimize_itmax{8} {integer >= 0}
   \postprop_minimize_energy{bond_bend} {bond,bond_bend,bond_bend_tors,bond_bend_tors_onfo,all}
   \extend_inter{3bond_conn} 3bond_conn, 4bond_conn, off:default
   \cancel_res_inter{off} local, neighbor (includes local), off
   \rinter_switch_length{0.0} 1.0 real inter switching length in A
   \inter_list{none} lnk_list none
   \EEMC_Emin{-0.3} -0.04 in Ha
   \EEMC_Emax{0.0}  0.0 in Ha
   \random_seed{-21516931}
]

~sim_mol_def[
  \system_def{residue} primitive
  \cgres_model{KB_3pt}  KB_3pt, off
```

```
42
43    \mol_parm_file{../top_3pt_prot_na.rtf}
44    \bond_database_file{../par_3pt_prot_na.prm}
45    \bend_database_file{../par_3pt_prot_na.prm}
46    \tors_database_file{../par_3pt_prot_na.prm}
47    \onfo_database_file{../par_3pt_prot_na.prm}
48    \inter_database_file{../par_3pt_prot_na.prm}
49
50    \pos_init_file{init.pdb}
51    \pos_out_file{last_frame.pdb}
52    \atom_pos_file{sampled.pos.pdb}
53    \tors_pos_file{sampled.tors_pos}
54    \epot_file{sampled.pot_energy}
55    \einter_file{sampled.inter_energy}
56    \region_database_file{region.data}
57    \energy_term{bond}
58    \energy_term{bend}
59    \energy_term{tors}
60    \energy_term{onfo}
61    \energy_term{inter}
62 ]
```

Example of a region file for test case ^{010}T:

region.data

```
1  ~region[\element_top_type{segment}
2
3      \dependency_type{independent}
4
5      \nseg{2}
6      \ncenter{2}
7      \segments_firstres{A:3,B:1}
8      \segments_lastres{A:44,B:50}
9
10     \segments_baseres{A:23,B:25}
11
12     \centers{A:23,B:25}
13
14     \prop_trans_sig{1.e-5}
15     \prop_rot_sig{1.e-6}
16     \prop_trans_sig_freeres{0}
17     \prop_rot_sig_freeres{0}
18 ]
19
20 ~region[\element_top_type{segment}
21
22     \dependency_type{independent}
```

```
23
24      \nseg{3}
25      \ncenter{3}
26      \segments_firstres{A:46,A:53,A:57}
27      \segments_lastres{A:51,A:55,A:76}
28
29      \segments_baseres{A:48,A:54,A:66}
30
31      \centers{A:48,A:54,A:66}
32
33      \prop_trans_sig{1.e-5}
34      \prop_rot_sig{1.e-6}
35      \prop_trans_sig_freeres{0}
36      \prop_rot_sig_freeres{0}
37  ]
38
39  ~region[\element_top_type{segment}
40
41      \dependency_type{independent}
42
43      \nseg{1}
44      \ncenter{1}
45      \segments_firstres{A:78}
46      \segments_lastres{A:182}
47
48      \segments_baseres{A:130}
49
50      \centers{A:130}
51
52      \prop_trans_sig{1.e-5}
53      \prop_rot_sig{1.e-6}
54      \prop_trans_sig_freeres{0}
55      \prop_rot_sig_freeres{0}
56  ]
57
58  ~region[\element_top_type{segment}
59
60      \dependency_type{independent}
61
62      \nseg{1}
63      \ncenter{1}
64      \segments_firstres{B:52}
65      \segments_lastres{B:62}
66
67      \segments_baseres{B:57}
68
69      \centers{B:57}
70
```

```
71      \prop_trans_sig{1.e-5}
72      \prop_rot_sig{1.e-6}
73      \prop_trans_sig_freeres{0}
74      \prop_rot_sig_freeres{0}
75  ]
76
77  ~region[\element_top_type{segment}
78
79      \dependency_type{independent}
80
81      \nseg{2}
82      \ncenter{2}
83      \segments_firstres{B:64,B:79}
84      \segments_lastres{B:77,B:90}
85
86      \segments_baseres{B:70,B:84}
87
88      \centers{B:70,B:84}
89
90      \prop_trans_sig{1.e-5}
91      \prop_rot_sig{1.e-6}
92      \prop_trans_sig_freeres{0}
93      \prop_rot_sig_freeres{0}
94  ]
95
96  ~region[\element_top_type{segment}
97
98      \dependency_type{independent}
99
100     \nseg{1}
101     \ncenter{1}
102     \segments_firstres{B:92}
103     \segments_lastres{B:190}
104
105     \segments_baseres{B:141}
106
107     \centers{B:141}
108
109     \prop_trans_sig{1.e-5}
110     \prop_rot_sig{1.e-6}
111     \prop_trans_sig_freeres{0}
112     \prop_rot_sig_freeres{0}
113 ]
```

The region file was generated with the `auto_region.py` script (it requires a file named params.py in the same folder, shown below the script):

`auto_region.py`

```
#! /usr/bin/python2.7

#from __future__ import print_function

import math
import os
import sys

my_dir = os.getcwd()
sys.path.append(my_dir)

import params

my_file = "%s/%s" % (my_dir,params.my_pdb)

my_chains = params.chains
moltenZones = params.moltenZones

#Get number of chains
no_chains = len(my_chains)

#Instantiate lists
my_inputs = []
in_first_res = []
in_last_res = []
segments = []

#Instantiate lists that will store first_res and last_res of all segments
all_first_res = []
all_last_res = []
all_centre = []
stride_list =[]

# Create header for input structure file (init.pdb)
header = "CBLC ~"
for i in params.chains:
    header += i
header += "\n"

continue_var = 'y'
cter = 0

# Create stride for input structure file (init.pdb)
for j in my_chains:

    f = open(my_file)

    i = 0
```

```
    prev_res = ""

    stride = "STRIDE >"

    for line in f:
        if "ATOM" in line and line[21] in j and line[23:26] != prev_res:

            current_res = line[23:26].strip()

            if i == 0:
                in_first_res.append(int(line[23:26]))

            if current_res in moltenZones[cter]:
                stride += "C"

            else:
                stride += "R"

            i += 1
            prev_res = line[23:26]

        else:
            pass

    in_last_res.append(int(in_first_res[cter]) + i-1)

    stride += "\n"
    stride_list.append(stride)
    f.close()
    cter += 1

# Strip init.pdb of header and stride
my_tmp = "temp.pdb"
f = open(my_file)
f_tmp = open(my_tmp,"w")

for line in f:
    if "STRIDE" in line or "CBLC" in line or line in ['\n', '\r\n',"END"]:
        pass
    else:
        f_tmp.write(line)

f.close()
f_tmp.close()
os.rename(my_tmp,my_file)

# Add new header and stride to init.pdb
for i in reversed(range(len(my_chains))):
```

```
97
98      with open(my_file, 'r') as original: data = original.read()
99      with open(my_file, 'w') as modified: modified.write(stride_list[i] + data + "\n")
100
101 with open(my_file, 'r') as original: data = original.read()
102 with open(my_file, 'w') as modified: modified.write(header + data)
103
104 # A function to calculate the centres of segments. Segments are stretches of residues between molt
105 def calc_centres(start,end,molten):
106
107     start = int(start)-1
108     end = int(end)+1
109
110     centre = []
111     seg_start = []
112     seg_end= []
113
114     len_seg = int()
115     temp_centre = int()
116
117     # Add start and end residues to moltenible residues list for calculation of segment lengths
118     print(molten)
119     molten.insert(0,start)
120     molten.append(end)
121
122     # Loop through moltenible residues list to calculate segment lengths and centres
123     for i in range (0,len(molten)-1):
124
125         # Make lists of last and first res
126         seg_start.append(int(molten[i]) + 1)
127         seg_end.append(int(molten[i + 1]) - 1)
128
129         # Calculate length of segments
130         len_seg = (seg_end[i] - seg_start[i]) + 1
131
132         # If length is even then divide by 2 and add to start residue
133         if len_seg % 2 == 0:
134
135             temp_centre = int((len_seg/2))
136             centre.append(temp_centre + int(seg_start[i])-1)
137
138         # If length is odd then divide by 2, round up (ceiling) and add to start residue
139         else:
140
141             temp_centre = int(math.ceil(len_seg/2.0))
142             centre.append(temp_centre + int(seg_start[i])-1)
143
144     return(seg_start,seg_end,centre)
```

```
145
146 # Store ouput from calc_centres in separate lists
147 for i in range(int(no_chains)):
148
149     segments.append(calc_centres(in_first_res[i],in_last_res[i],moltenZones[i][:]))
150
151     all_first_res.append(segments[i][0])
152     all_last_res.append(segments[i][1])
153     all_centre.append(segments[i][2])
154
155
156 # Write region file
157 def print_region(start,end,centre,no_segments):
158
159     in_nseg = no_segments
160     in_ncenter = no_segments
161     in_first = start
162     in_last = end
163     in_centre = centre
164     rot_setting = params.rot
165     trans_setting = params.trans
166     rot_free_setting = '0'
167     prop_free_setting = '0'
168
169     element = "{segment}"
170     dependency = "{independent}"
171     nseg = "{%s}" % in_nseg
172     ncenter = "{%s}" % in_ncenter
173     firstres = "{%s}" % in_first
174     lastres = "{%s}" % in_last
175     baseres = "{%s}" % centre
176     centers = "{%s}" % centre
177     prop_trans_sig = "{%s}" % trans_setting
178     prop_rot_sig = "{%s}" % rot_setting
179     prop_trans_sig_freeres = "{%s}" % rot_free_setting
180     prop_rot_sig_freeres = "{%s}" % prop_free_setting
181
182     template = """~region[\element_top_type{0}
183
184     \dependency_type{1}
185
186     \nseg{2}
187     \ncenter{3}
188     \segments_firstres{4}
189     \segments_lastres{5}
190
191     \segments_baseres{6}
192
```

```
193         \centers{7}
194 
195         \prop_trans_sig{8}
196         \prop_rot_sig{9}
197         \prop_trans_sig_freeres{10}
198         \prop_rot_sig_freeres{11}
199     ]
200         """
201 
202         print >>f, template.format(element, dependency, nseg, ncenter, firstres, lastres, baseres, cen
203 
204 
205     def print_options(no_chains,all_first_res,all_last_res,all_centre):
206         cter = int()
207         for i in range(int(no_chains)):
208             print("\nchain %d" % i)
209             cter = 0
210             for j in all_centre[i]:
211                 print("segment %d" % cter, all_first_res[i][cter],all_last_res[i][cter],all_centre[i][
212                 cter +=1
213 
214         return
215 
216 
217     # Take segment combination list and return lists containing start, end and centre residues accordi
218     def make_regions(all_first_res,all_last_res,all_centre,seg_comb):
219 
220         seg_comb_list = seg_comb.split()
221         seg_com_list_split = [i.split(':') for i in seg_comb_list]
222 
223         custom_start = []
224         custom_end = []
225         custom_centre = []
226 
227         no_segments = int()
228 
229         for i in range(len(seg_com_list_split)):
230 
231             chain = int(seg_com_list_split[i][0])
232             segment = int(seg_com_list_split[i][1])
233 
234             custom_start.append(my_chains[chain]+":"+str(all_first_res[chain][segment]))
235             custom_end.append(my_chains[chain]+":"+str(all_last_res[chain][segment]))
236             custom_centre.append(my_chains[chain]+":"+str(all_centre[chain][segment]))
237 
238         no_segments = len(custom_centre)
239 
240         return(custom_start,custom_end,custom_centre,no_segments)
```

```
241
242
243 seg_comb = params.segment_combination
244
245 f = open('region.data', 'w')
246
247 for i in range(len(seg_comb)):
248
249     #print_options(no_chains,all_first_res,all_last_res,all_centre)
250     first,last,centre,no_segments = make_regions(all_first_res,all_last_res,all_centre,seg_comb[i]
251     first = ",".join(map(str, first))
252     last = ",".join(map(str, last))
253     centre = ",".join(map(str, centre))
254     print_region(first,last,centre,no_segments)
255
256 f.close()
```

The file named `params.py` contains parameters that are loaded by `auto_region.py` to generate test cases. Below is an example for test case ^{010}T:

params.py

```
1 my_pdb = "init.pdb"
2 chains = ["A","B"]
3 moltenZones = [["45","52","56","77"],["51","63","78","91"]]
4 segment_combination = ["0:0 1:0","0:1","0:2","0:3","0:4","1:1","1:2 1:3","1:4"]
5 rot = "1.e-6"
6 trans = "1.e-5"
```

The following script was used to calculate the binding-groove distances.

`dist_mhcii.py`

```
1 #! /usr/bin/env python
2 # usage: dist_mhcii.py mosaics_trajectory.pdb
3
4 from MDAnalysis import *
5 import numpy
6 import math
7 import sys
8
9 my_traj = sys.argv[1]
10
11 u = Universe("../init.pdb",my_traj)
12 v = Universe("../init.pdb")
13
```

```
14 end = my_traj.find('.pdb')
15 fout_name = my_traj[0:end] + '_dist.dat'
16
17 helix_A = u.selectAtoms("segid A and resid 60:65")
18 helix_B = u.selectAtoms("segid B and resid 65:70")
19
20 f = open(fout_name,'w')
21
22 for ts in u.trajectory:
23
24     distance = numpy.linalg.norm(helix_A.centerOfMass() - helix_B.centerOfMass())
25
26     f.write('%7.3f\n' % distance)
27
28 f.close()
```

Chapter 8

Customised Natural Moves - Case study 2: Nucleic Acids

This work as published in: Demharter, S., Knapp, B., Deane, C. M., & Minary, P. (2016). Modeling Functional Motions of Biological Systems by Customized Natural Moves. *Biophysical Journal*, 111(4), 710-721. http://doi.org/10.1016/j.bpj.2016.06.028

8.1 Summary

Here, I present the use of our customised Natural Move Protocol to study the structural effects of epigenetic marks on nucleic acid structures. We use cNMs to modulate translations and rotations of segments as well as torsion and bend angles of bonds and compare different sets of cNMMC simulations to infer the effect of epigenetic marks on structural parameters.

8.2 Introduction

Here we investigated the effect of 5hmC on local base pair arrangement in the Dickerson-Drew dodecamer (DDD); a simple model system that has recently attracted interest due to a new crystal structure with added hydroxymethyl epigenetic marks on cytosines A9 and B21 [271]. Two hydroxymethyl configurations were found in this structure. One points

towards the backbone phosphate oxygen 5hmC/OP2, the other forms a weak hydrogen bond with the 3'-adjacent G/O6. For the purpose of this case study we focused on the latter as it was estimated to be the most prevalent configuration in the crystal [271]. A schematic of the system is shown in Figure 8.1A. Next we apply the four main steps of our protocol to investigate the effects of 5hmC on this structure.

8.3 Methods

The Dickerson-Drew dodecamer in configuration A (5-hydroxymethyl epigenetic marks point towards the O6 oxygen of the 3'-adjacent guanine (G/O6)) [271] at 1.3 Å resolution was used as the starting point. The missing hydroxyl hydrogens were added and oriented towards the 3'-adjacent G/O6. Hydrogens were added to the remaining atoms using pymol1.7's `h_add` command [273]. The 3'- and 5'-terminal base pairs were removed. An all-atom representation was used with the Amber99-bsc0 force field [194] and a dielectric dampening model [296]. Single temperature Natural Move Monte Carlo was performed at 300K. We ran 30 independent repeats of 5,000,000 Monte Carlo iterations for each test case. Helical parameters were analysed using x3DNA [221].

8.4 Results

Step I: Define a hypothesis

Lercher et al. observed that the 5hmC hydroxyl formed a non-canonical hydrogen bond with the 3'-G/O6 [271]. This oxygen is already part of a canonical (Watson-Crick) hydrogen bond with the C on the opposing strand. No structural differences between the DDD with and without the epigenetic mark were observed, suggesting that any effects that 5hmC might have on the surrounding base pairs cannot be seen in a static structure. We investigate the hypothesis that the hydroxyl-group on 5hmC subtly interferes with the 3'-adjacent G-C base pair.

Step II: Translate hypothesis into Natural Moves

Given our hypothesis we defined two sets of cNMs. The first set contained the two torsion angles around bonds C5-C5M ($\Omega_{\phi}^{(1)}$) and C5M-O5 ($\Omega_{\phi}^{(2)}$) in the 5hm epigenetic mark. This gave us control over the orientation of the hydroxyl group during simulation. The second set of cNMs described the collective movement of 5hmC and the 3'-adjacent G, when the MZ between them (${}^{*}\Omega_{MZ}^{(1)}$) was deactivated. This cNM was meant to simulate the stabilising effect caused by a non-canonical intra-strand hydrogen bond between the two neighbouring nucleotides. Figure 8.1B shows the cNMs. Note that the depiction of molten zone ${}^{*}\Omega_{MZ}^{(1)}$ is an abstraction as some of the detail was omitted for simplicity.

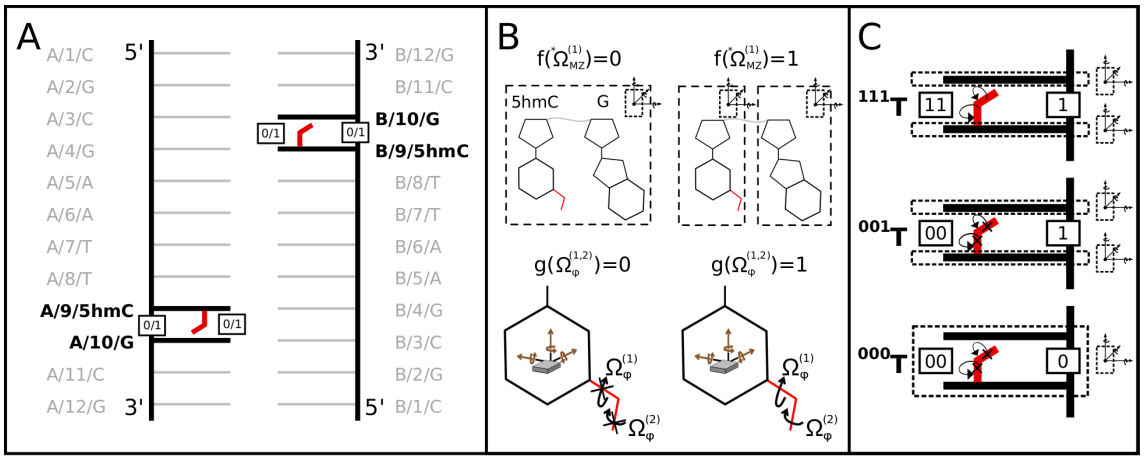

Figure 8.1: **Defining customised Natural Moves for 5-hydroxymethylcytosine in the Dickerson-Drew Dodecamer. A** Schematic showing the Dickerson-Drew dodecamer with two added 5-hydroxymethyl (5hm) epigenetic marks. The red lines represent the 5hm epigenetic marks and the thick black horizontal lines represent the bases that are directly affected by the customised Natural Moves. **B** The grey line connecting the two nucleotides represents an abstracted backbone chain that may undergo chain breaks during NMMC moves. The dotted rectangles show the collective motion of two neighbouring nucleotides when the interjacent melting zone ${}^{*}\Omega_{MZ}$ is deactivated or activated. The red lines show the epigenetic mark with the arrows highlighting the torsion angles around C5-C5M and C5M-O, the sampling of which may be deactivated or activated depending on the test case. **C** Test cases ${}^{111}T$, ${}^{001}T$ and ${}^{000}T$ are shown. The dotted lines show individual or collective degrees of freedom depending on the state of the interjacent molten zone (active/inactive). The arrows on the epigenetic marks represent rotations around the two torsion angles of 5hm which may be active or inactive. Note, that only one of the two epigenetic marks is shown. However both modifications are treated equivalently in each case.

Step III: Generate test cases

Given the cNMs that we defined above we get a decomposition vector $\mathbb{D}$ of length 3 (see

methods). The first two elements refer to rotational freedom along the two torsion angles $\Omega_{\phi}^{(1)}$ and $\Omega_{\phi}^{(2)}$ in the hydroxyl group of 5hmC and the third refers to ${}^{*}\Omega_{MZ}^{(1)}$ that consists of the backbone atoms between 5hmC and the 3'-adjacent G. Similar to the protein example in chapter 7, each element in $\mathbb{D}$ can either be on or off (1/0), i.e. the relative arrangement of G and 5hmC in the case of ${}^{*}\Omega_{MZ}^{(1)}$ and the sampling of torsion angles included in $\Omega_{\phi}^{(1)}$ and $\Omega_{\phi}^{(2)}$ can either be activated or deactivated. Thus, for a decomposition vector $\mathbb{D}_{DNA} : \{g(\Omega_{\phi}^{(1)}), g(\Omega_{\phi}^{(2)}), f({}^{*}\Omega_{MZ}^{(1)})\}$ of length three we get the following $2^3 = 8$ possible test cases: ${}^{000}T, {}^{001}T, {}^{010}T, {}^{100}T, {}^{011}T, {}^{101}T, {}^{110}T, {}^{111}T$.

Note, that we only considered test cases where both of the torsion angles were either active or inactive as we were only interested in a fully flexible or fixed epigenetic mark for this study. Therefore, we omitted test cases ${}^{010}T, {}^{100}T, {}^{011}T$ and ${}^{101}T$. The remaining test cases included ${}^{000}T, {}^{001}T, {}^{110}T$ and ${}^{111}T$. Test case ${}^{110}T$ was also ignored as it is very similar to test case ${}^{000}T$ due to the deactivated molten zone restraining the orientation of the two neighbouring bases. Thus, the set of test cases we included in our study were ${}^{000}T, {}^{001}T$ and ${}^{111}T$.

Step IV: Conformational sampling and evaluation

We ran four sets of simulations of the DDD: the three test cases ${}^{111}T$, ${}^{001}T$, ${}^{000}T$ and a simulation without the epigenetic mark that served as a control.

Figure 8.2 shows the effect of each of the test cases on the distance of the intra-strand hydrogen bonds between each of the two 5hmC hydroxyl groups and the 3'-adjacent guanine O6 atoms.

Figure 8.3A schematically depicts the base pairs and degrees of freedom of interest. Figure 8.3B shows the distributions of the parameters shear, stretch and propeller, that changed progressively as we applied the different test cases. Note that we only show the distributions for the base pairs around one of the epigenetic marks, but the effect was seen on both ends. Interestingly, the shear was most affected in the GC base pair 3'-adjacent to 5hmC, while the stretch and propeller were mostly changed in the 5'-adjacent base pair. No large differences between the modified ${}^{111}T$ system and the unmodified control were observed. However, once the orientation of epigenetic mark was fixed (test case ${}^{001}T$) a

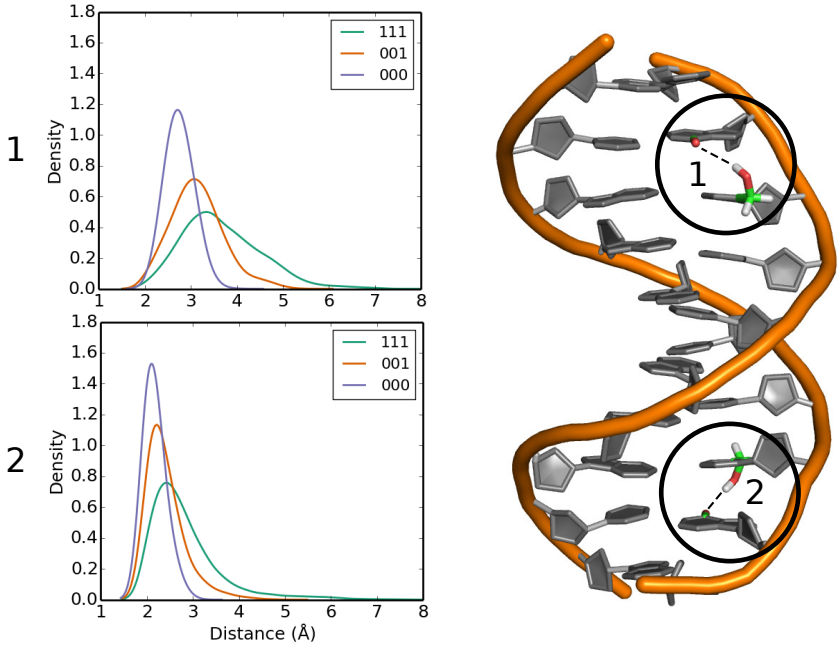

Figure 8.2: **The effect of customised Natural Moves on an intra-strand hydrogen bond.** Distance distributions of the two non-canonical hydrogen bonds between the hydroxyl hydrogens on 5hmC and the O6 oxygen of the 3'-adjacent guanine as highlighted on the right. All three test cases are shown. The X-ray structure, which we used as our starting structure, is not totally symmetric so we do not expect totally symmetrical effects as we move from ${}^{111}T$ to ${}^{000}T$.

subtle shift in the distribution was detected. The effect was further increased when the relative movement between 5hmC and the 3'-adjacent G was deactivated (test case ${}^{000}T$). Changes were also observed in base pair parameters stagger, buckle and opening, but the effects were less systematic and did not correspond to the increasing 'epigenetic signal' encoded in our test cases (all base pair parameter values are shown in Figure 8.4). We did not investigate changes in the base stack parameters (Figure 8.5), as we expected that the non-canonical epigenetic (intra-strand) hydrogen bond formation, which we enforce by customised Natural Moves, could directly impose particular base stacking. However, we were more interested to study distributions over DNA base pair parameters, which were less directly affected by hydrogen bond formation between adjacent (on the same strand) bases.

The base-pair and base-stack parameters of the remaining base pairs showed no significant changes as shown in Figures 8.4 and 8.5.

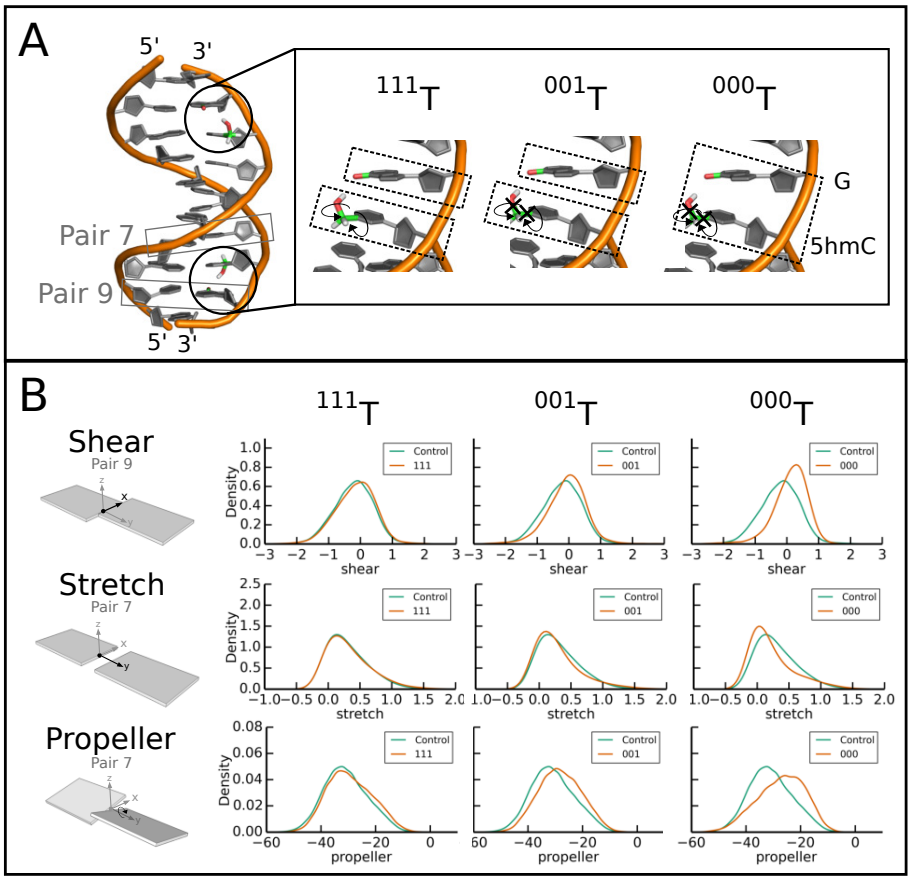

Figure 8.3: **The effect of 5-hydroxymethylcytosine on the Dickerson-Drew dodecamer is amplified by customised Natural Moves. A** The Dickerson-Drew dodecamer is depicted with the backbone in orange and bases in grey. The two 5hmC modifications are coloured based on atom type (O:red, C:green, H:white). The sets of degrees of freedom chosen for the 5hmC modification are shown on the right. Curved arrows indicate free torsional sampling, while the red crosses indicate fixed chi torsion on rotations around corresponding bonds (Chi1: C5-C5M, Chi2: C5M-OH). ${}^{111}T$: Full sampling of all torsion angles; ${}^{001}T$: Fixed torsion angles in 5hmC; ${}^{000}T$: Fixed torsion angles in 5hmC and relative orientation between 5hmC and G. **B** Distributions of the shear, stretch and propeller are shown for the three different test cases. Each column compares simulations without modification (control) in green against test cases ${}^{111}T$, ${}^{001}T$ and ${}^{000}T$ in orange. The shear is shown for base pair 7 and the stretch and propeller are shown for base pair 9.

8.5 Discussion

A sequence of enzymatic reactions drives a cycle of epigenetic cytosine modifications including 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) [297, 298]. 5mC has been shown to increase dsDNA stability, which is consistent with its role in gene expression at CpG islands [299].

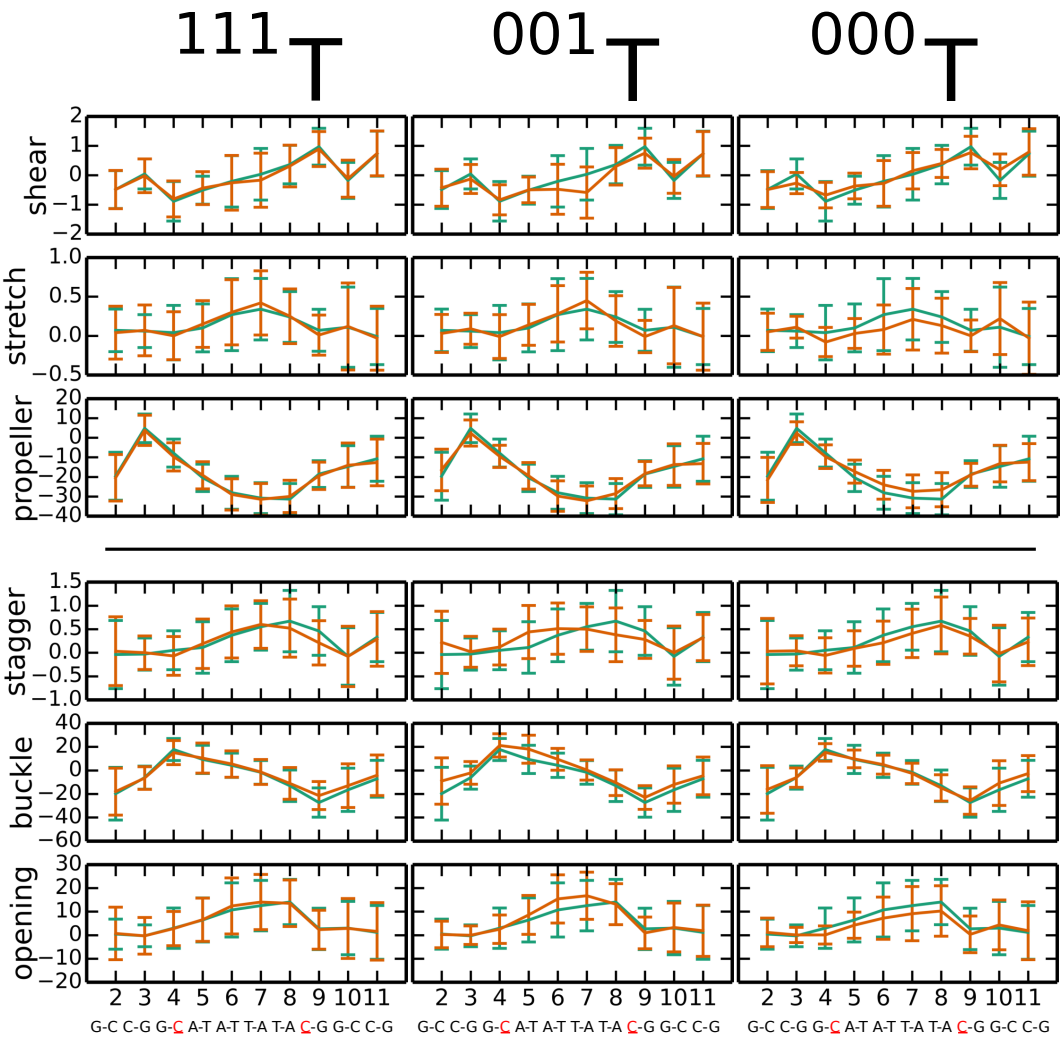

Figure 8.4: **The effect of three test cases on the base pair parameters.** The parameters for all base pairs in the three test cases ${}^{111}T$, ${}^{001}T$ and ${}^{000}T$ (orange) are compared against the control simulation without modification (green). The top half of the figure shows the three parameters shown in Figure 8.3. All other parameters do not show any systematic changes caused by the customised Natural Moves. Displacement parameters (shear, stretch, stagger) are shown in Ångstrom and angular parameters (buckle, propeller, opening) are shown in degrees. The vertical bars show the standard deviation. The red underlined characters show the positions of the epigenetic mark.

5hmC, sometimes referred to as the sixth base of the mammalian genome, can partly reverse the 5mC stabilising given the right sequence context [272] and a study investigating a 27-bp oligonucleotide has observed that 5hmC increases DNA flexibility in MD simulations [300]. Several DNA structures with 5hmC epigenetic marks have been solved to date but no significant structural effects on the DNA helical parameters have been found [271, 276, 277]. This is in contrast to a structure of a DNA dodecamer comprising three 5-formyl CpG sites that showed how 5fC causes large structural changes that lead

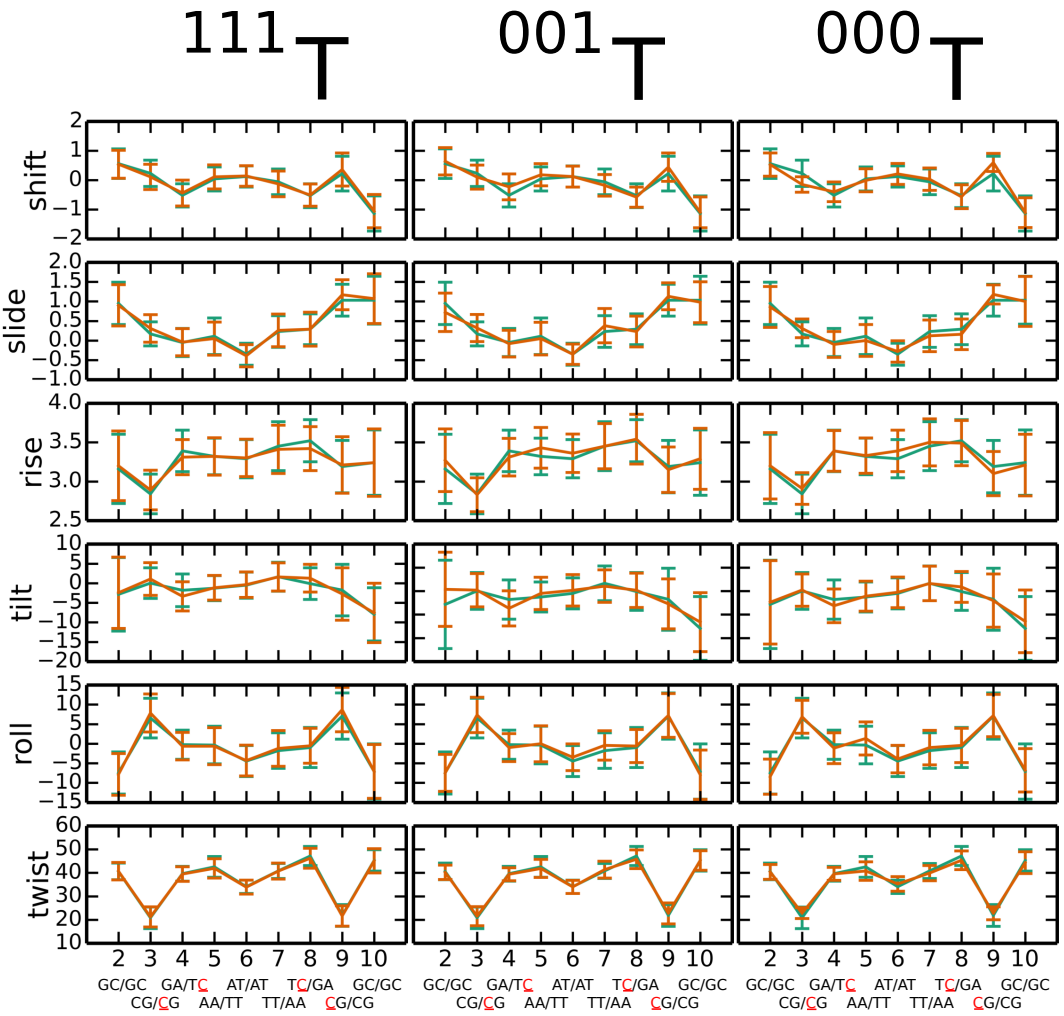

Figure 8.5: **The effect of three test cases on the base stack parameters.** The parameters for the three test cases ^{111}T, ^{001}T and ^{000}T (orange) are compared against the control simulation without modification (green). No systematic changes were observed in any of the test cases. Displacement parameters (shift, slide, rise) are shown in Ångstrom and angular parameters (tilt, roll, twist) are shown in degrees. The vertical bars show the standard deviation. The red underlined characters show the positions of the epigenetic mark.

to helical underwinding [88].

In order to demonstrate how customised Natural Moves can be used to study the effects of epigenetic marks we chose a recent high resolution structure of the Dickerson-Drew dodecamer comprising a 5hmC epigenetic modification. When performing traditional Natural Move Monte Carlo we found that the presence of a single 5hmC epigenetic mark in DDD causes only minimal change in some of the helical parameters of the 3'-adjacent adjacent base pair. These results agree with the general view that a single 5hmC epigenetic mark has a limited structural effect on the surrounding helical parameters, which makes

it difficult to identify experimentally [271, 276, 277]. The results are also in concurrence with Lercher et al. [271] who found that their crystal structures with and without 5hmC were nearly identical with an rmsd of 0.35 Å and a 0.8 Å widening of the major groove at the site of modification.

However, using customised Natural Moves we constrained the 5hmC hydroxyl group in the experimentally determined configuration and thereby increased non-canonical intra-strand hydrogen bonding during simulation and we were able to amplify some of the changes caused by the presence of 5hm. The effect was further increased when we deactivated the relative movement between 5hmC and the 3'-G, thereby effectively emulating the stabilising effect of an intra-strand hydrogen bond. Thus, using customised Natural Moves we were able to detect and amplify subtle structural effects on DNA helical parameters caused by a single epigenetic mark in the DDD.

8.6 Files

The following parameters were used for all simulations (parameters are defined in appendix A).

nmmc.input

```
~sim_gen_def[
   \simulation_typ{PT} DBFR NM NMEN PT EEMC SEQ_PT SEQ_EEMC DBFR DBFREN
   \minimize_tol{1.e-3}
   \minimize_type{bfgs} cg bfgs samc stsamc
   \minimize_report{2} 0  1  2
   \energy_report{2}
   \prop_type{tors} tors cart bend
   \prop_tors_sig{1.e-5} 1.e-5 0.0
   \prop_trans_sig{1.e-5} .5e-5 0.0
   \prop_rot_sig{5.e-6} 1.e-6 0.0
   \prop_tors_type{full} full side_chain
   \replica_number{0} 10            number > 1
   \prob_eemc_jump{0.15}
   \eemc_disk_size{10}
   \energy_gap{1.1} 1.1 E_i = a + b*(energy_gap)^i, (a,b) scaling
   \total_step_mc{100000}  2000 100            number > 1
   \local_step_md{1} 10           number > 1
   \time_step_md{0.4} 0.5
```

```
    \statistics_freq{1000} 10
    \burn_in_B{0} 2
    \burn_in_N{0} 2
    \write_energy_unit{kcal} kcal  Ha: atomic unit, kcal: kcal/mol
    \temperature{300} 300
    \inter_list{none} none lnk_list
    \random_seed{-7316982}
    \EEMC_Emin{-1.0} -0.04 in Ha
    \EEMC_Emax{0.0}  0.0 in Ha
    \prop_notors_sig_scale{20} {real >= 0.0}
]

~sim_mol_def[
  \system_def{residue} primitive
  \implicit_solvent{off} scp, off
  \ddd{DD0S} ce DD0S
  \ddd_D{80.0}
  \ddd_D0{4.0}
  \ddd_S{0.4}
  \ddd_c{0.5} 2.5
  \ddd_e{6.0} 2.0
  \neutralize{nucl} nucl, prot, all, off
  \mol_parm_file{../TOPPOT/top_database/amber/99-bs0/top_all99-bs0_prot_dna_chidef_eg_fixedCHM.rtf
  \bond_database_file{../TOPPOT/pot_database/amber/99-bs0/mosaics_amber99-bs0.bond}
  \bend_database_file{../TOPPOT/pot_database/amber/99-bs0/mosaics_dna_amber99-bs0_eg.bend}
  \tors_database_file{../TOPPOT/pot_database/amber/99-bs0/mosaics_amber99-bs0.tors_and_impr}
  \onfo_database_file{../TOPPOT/pot_database/amber/99-bs0/mosaics_amber99-bs0.onfo}
  \inter_database_file{../TOPPOT/pot_database/amber/99-bs0/mosaics_amber99-bs0.vdw}
  \region_database_file{region.data}
  \pos_init_file{init.pdb}
  \pos_out_file{sampled.pos_out.pdb}
  \atom_pos_file{sampled.pos.pdb}
  \tors_pos_file{sampled.tors_pos}
  \epot_file{sampled.pot_energy}
  \einter_file{sampled.inter_energy}
  \hessian_file{sampled.hessian}
  \eighess_file{sampled.eighess}
]
```

8.6.1 Enabling customised Natural Moves in Mosaics

The movement of 5hmC was disabled in order generate test cases ${}^{001}T$ and ${}^{000}T$. This was achieved by changing the chi groups of the CHM entry in the topology file `top_all99-bs0_prot_dna_` from:

```
CHIGROUP C2 O2 N3 C4 N4 C5 C5M C6 H6
CHIGROUP H42 H41
CHIGROUP OH H52 H53
CHIGROUP HO
```

to

```
CHIGROUP C2 O2 N3 C4 N4 C5 C5M C6 H6
CHIGROUP H42 H41
```

For test case ^{000}T a region file was used to group the two 5hmCs and their 3'-adjacent Gs:

```
~region[element_top_type{residue}
    \dependency_type{independent}

    \nres{16}
    \residues{A:1,A:2,A:3,A:4,A:5,A:6,A:7,A:10,B:11,B:12,B:13,B:14,B:15,B:16,B:17,B:20}

    \ncenter{1}
    \centers{A:5}

    \prop_trans_sig{0}
    \prop_rot_sig{0}
    \prop_trans_sig_freeres{1.e-5}
    \prop_rot_sig_freeres{5.e-6}
    \prop_trans_sig_respair{0}
    \prop_rot_sig_respair{0}
]

~region[element_top_type{residue}
    \dependency_type{independent}

    \nres{2}
    \residues{A:8,A:9}

    \ncenter{1}
    \centers{A:8}

    \prop_trans_sig{1.e-5}
    \prop_rot_sig{5.e-6}
    \prop_trans_sig_freeres{0}
    \prop_rot_sig_freeres{0}
    \prop_trans_sig_respair{0}
```

```
    \prop_rot_sig_respair{0}
]

~region[element_top_type{residue}
    \dependency_type{independent}

    \nres{2}
    \residues{B:18,B:19}

    \ncenter{1}
    \centers{B:18}

    \prop_trans_sig{1.e-5}
    \prop_rot_sig{5.e-6}
    \prop_trans_sig_freeres{0}
    \prop_rot_sig_freeres{0}
    \prop_trans_sig_respair{0}
    \prop_rot_sig_respair{0}

]
```

8.6.2 X3DNA analysis scripts

The following script was used to format the pdb trajectories for x3dna. `reorder_pymol.pdb`

```
#! /usr/bin/python
# usage: reorder_pymol.py mosaics_trajectory.pdb

import __main__
__main__.pymol_argv = ['pymol','-qc']
#__main__.pymol_argv = ['pymol','']
import sys,time,os
import pymol
pymol.finish_launching()

my_file = sys.argv[1]

end = my_file.find(".pdb")
fileName = my_file[0:end]
newFileName = fileName + "_reordered.pdb"

pymol.cmd.load(my_file,fileName)

pymol.cmd.save(newFileName,fileName,state=0)

pymol.cmd.quit()
```

The following script was used to analyse trajectories and extract parameters of interest.

`calc_x3dna_parameters.sh`

```
#! /bin/bash

my_params=( 'shift' "slide" "tilt" "roll" "rise" "twist" "propeller" "buckle" "opening" "shear" "s
#my_params=( 'major_gw_refined' 'major_gw_pp' 'minor_gw_refined' 'minor_gw_pp')

my_files=( "shift_cat.dat" "slide_cat.dat" "rise_cat.dat" "tilt_cat.dat" "roll_cat.dat" "twist_cat

my_folders=( "init_chm_full" "init_nochm_full" "init_chm_fullfixedCHM" "init_chm_fullfixedGC" )

for my_folder in "${my_folders[@]}"; do

    my_setups=$PWD"/"$my_folder"_rep"
    my_folder_cat=$my_folder"_cat"

    for rep in {1..32}; do

        setup=$my_setups$rep
        cd $setup
            echo $setup

            # reorder pdb file for compatibility with x3dna. The script requires pymol.
            ./reorder_pymol.py sampled.pos.pdb

            # x3dna - find base pairs in starting structure and save in my_bps.inp
            find_pair init.pdb my_bps.inp

            # run x3dna to calculate base pair/stack parameters
            x3dna_ensemble analyze -b my_bps.inp -e sampled.pos_reordered.pdb

            # extract parameter files from auxiliary file
            for param in "${my_params[@]}"; do
                x3dna_ensemble extract -p $param -o $param".dat"
            done

        cd ..

    done

done
```

Chapter 9

Conclusion

The work presented in this thesis can be divided into two parts. In the first part (chapters 4,5 and 6) I have presented a number of novel biological applications for traditional Natural Move Monte Carlo including the simulation of MHC/peptide detachment pathways, diabody hinge flexibility and epigenetic marks on DNA structures. In the second part (chapters 7 and 8) I presented a protocol, which enables the user to systematically choose and modify degrees of freedom according to their own intuition, experimental results or other data in order to test hypothesis regarding the functional motions of biological systems. We refer to this type of simulation as customised Natural Move Monte Carlo and applied the protocol to different research questions. We investigate the plasticity of the MHCII binding groove in the absence of peptide and the molecular mechanism by which the peptide-loading chaperone HLA-DM stabilises the empty MHCII binding groove. We also apply the protocol to a DNA system, where we use customised Natural Moves to enhance the structural effects that different epigenetic marks may have on DNA structure. These case studies provide a first indication of what may be possible with the targeted modulation of degrees of freedom for the purpose of probing molecular structures and understanding functional motions.

The thesis began with a brief introduction to the field of protein structure research chapter 1. In chapter 2 I highlighted the importance of computational methods as well as experimental techniques for the investigation of functional aspects of protein motion and function. I provided background to the main biological systems used in this thesis and

described the fundamentals of protein and DNA structure, gave an overview of techniques and discoveries regarding collective motion in biological molecules as well as introduced various aspects of molecular simulations including different types of models, degrees of freedom and algorithms.

Novel applications for traditional NMMC

In chapter 3 I listed detailed descriptions of the various methods used in this thesis, including the stochastic chain closure algorithm central to NMMC and the physics- and knowledge-based potentials used for nucleic acids and proteins, respectively. Furthermore, I outlined NMMC as well as the NMA method used in one of the chapters (chapter 7) and explained the working details of x3DNA, a popular structural analysis tool that I used to characterise conformational changes in epigenetically modified DNA molecules (chapter 8).

After the first two introductory chapters I presented the first NMMC research application in this thesis. Here, we applied NMMC to a commonly studied immunobiology system, the MHCI/peptide complex. The aim of this study was to prove that NMMC can be used as a training-free method to accurately classify MHC class I binding and non-binding peptides and provide insight into the dissociation process of non-binders. The computational efficiency of NMMC allowed us to explore the MHCI/peptide interaction for over 30 peptides. Previously, only a couple of pMHCI complexes had been simulated long enough to see some degree of dissociation. Even the longest MD simulation of a pMHCI complex to date (1000 ns) has not revealed more than partial detachment. We benchmarked our results against known binding affinities and our results were able to reliably distinguish binders from non-binders. While anchor residues are not considered to be the main contributing factor for pMHC binding, we observed good agreement between N- or C-terminal detachment and the presence/absence of preferred anchor residues.

We found that simulated detachment speed was correlated with binding data from experiments and experimental binding affinity data with an AROC fof 0.85; this is similar to sequence-based prediction methods and more accurate than structural docking tech-

niques. This suggests that coarse-grained NMMC applied to pMHC is informative at the biophysical scale and is sensitive to important features responsible for the MHCI/peptide interaction. We ran 100 independent replica simulations for each peptide to avoid simulations getting trapped in local minima or outlier trajectories skewing the results. Our boot strap analysis showed that at least 25 to 50 replicas were needed for a reliable result.

While more accurate methods exist, this was the first time that a training-free method could both predict the binding preference as well as generate detachment pathways. NMMC was orders of magnitude faster and as such allowed us to explore pMHC detachment processes on a scale that was not possible with all-atom molecular dynamics simulations previously. This study provided a proof of concept for studying small scale biophysical processes with a knowledge-based, coarse-grained Natural Move Monte Carlo method.

The second NMMC application was a study of the hinge flexibility in two diabodies developed by the Garcia lab. Diabodies are small bivalent and sometimes bi-specific antibody fragments that consist of a heavy-chain variable domain (VH) connected to a light-chain variable domain (VL) on the same polypeptide chain (VH-VL). By connecting them with a linker that is too short to allow pairing between the two domains on the same chain, the domains can only bind the complementary domains of another chain and thereby create two antibody fragments that assemble *in vivo* and create two antigen-binding sites. Moraga et al. developed a number of diabodies specific against the Erythropoietin receptor (EpoR), that dimerised the EpoR ectodomains. The dimerisation of receptors often plays a key role in the transduction of signals across the cell membrane. For example, cytokines, a class of secreted glycoproteins that participate in the regulation of cell fate and function bind to extracellular domains of their respective receptors, which triggers dimerisation and signalling. The diabodies elicited a range of different signalling amplitudes, from full to minimal agonism. Interestingly, the solved structures varied in EpoR dimer configuration and the proximity of the N-terminal ends. Thus, it appears that the signalling activity of GPCR receptors can be 'tuned' by modulating the geometry of their dimerisation.

Due to concerns that flexibility in the hinge connecting the two VH/VL domains in the diabody may have an important effect on this mechanism, we simulated two selected structures. I performed Natural Move Monte Carlo on the two diabody structures in order to investigate their dynamic range. Specifically, I assessed the flexibility in the two linkers connecting the VH/VL modules of the diabodies. I found that for both diabodies the relative arrangement of the two VH/VL domains was maintained throughout simulation with a only a few preferred configurations, each with high similarity to the native structure.

In chapter 6 we investigated the effects of different epigenetic marks (5mC, 5hmC, 5fC and 5caC) on DNA dodecamers. It is thought that the cumulative effect of several epigenetic marks can cause structural alterations in DNA. It was shown that six formylated CpG bases in close proximity are sufficient to convert the dodecamer from a B-DNA to a novel structural form referred to as F-DNA. Other structures with only two formylated CpG bases have been shown to have F-DNA-like base step parameter values, however the overall structure remained in the B-DNA form. This seemed to suggest that a critical number of 5fC marks is required to shift the structure from B-DNA to F-DNA.

The experimental testing of all possible epigenetic makeups is prohibitive. Computational approaches have been applied, however, these methods have traditionally been computationally expensive. Here, we used NMMC as an efficient computational framework to simulate a range of epigenetic makeups for different DNA structures.

We were able to reproduce experimental observations from two recent crystal structures that contain 5hmC and 5fC, respectively. We also show that compared to experiment our protocol correctly identifies the energetically favourable forms for structures with different epigenetic marks; 5fC is energetically more favourable in the F-DNA form and 5hmC in the B-DNA form. The computational efficiency and straight forward application of this protocol has the potential to facilitate comprehensive computational investigations of epigenetic systems.

Furthermore, we defined the forward and inverse epigenetic problems in an effort to develop a robust computational simulation framework for the testing of epigenetic

marks. The forward problem is defined as the search for a structural conformation given a known epigenetic makeup, while the inverse epigenetic problem aims to predict an epigenetic makeup given a known structure. Our initial solutions to the inverse and forward epigenetic problems were in agreement with high-resolution crystal structures. Building on this result, we have tested all possible epigenetic makeups in a representative structure and assessed their energies. These predictions may hold valuable information on the transition between B-DNA and F-DNA and may be of value as a starting point for experimental studies.

Customised Natural Moves and case studies

After having listed three novel applications for traditional NMMC in the first part of this thesis, I introduced a novel theoretical framework based around the customisation capabilities inherent to NMMC. NMMC simulations require as an input a decomposition of the structure into segments and regions that are connected by molten zones, which allow for the chain closure. The translational and rotational step size of each segment can be independently defined. This level of control over different parts of a molecular system was unprecedented when NMMC was first published. In the customised Natural Move protocol presented in this thesis we make use of this feature and lay down a formal description for the generation of test cases. Test cases may be regarded as NMMC simulations with a unique set of degrees of freedom and parameters that represent a specific research question or hypothesis with respect to the functional motions of a biological system.

We used cNMs to modulate translations and rotations of segments as well as torsion and bend angles of bonds and compare different sets of cNMMC simulations to infer causal relationships in functional motions. We used two case studies to demonstrate its application. First we investigated functional motions in the class II major histocompatibility complex (MHCII) and second we studied the structural effects of an epigenetic mark on a DNA model system.

In chapter 7 we demonstrated the application of the customised Natural Moves protocol on two protein case studies. We investigated the structural plasticity of the empty

MHCII complex as well as the mechanism by which the peptide-loading chaperone HLA-DM stabilises the open form of the MHCII binding groove.

We investigated the functional motions responsible for the collapse of the peptide-free MHCII binding groove. Using pre-existing information derived from experimental as well as simulation studies we identified potentially labile regions, which we were able to selectively include or exclude in the test cases (using molten zones). Our simulations showed that the test cases with active molten zones in the β1-helix were able to assume a number of transitory states that cause a narrowing of the binding groove in the absence of peptide.

In the second part of chapter 7 we investigated the mechanism by which HLA-DM stabilises the empty MHCII binding groove. While a structure of the MHCII/HLA-DM complex exists, the mechanism of stabilisation is still largely unknown. Again, we use existing knowledge from experiments and simulations to guide the design of our test cases. From this information we generated a hypothesis, which we were able to translate into customised Natural Moves. In addition to customised NMMC we also used normal mode analysis to investigate this interaction. We found that the suggested long range mechanism that implicates the membrane-proximal globular domains in stabilising the open state of the empty MHCII binding groove gave results that were in accordance with the hypothesis.

After having presented two protein case studies I also demonstrated the cNMMC protocol on a nucleic acid system, specifically the Dickerson-Drew Dodecamer in the presence and absence of an epigenetic mark (5hmC).

We investigated the effect of 5hmC on local base pair arrangement. The authors of the structure observed that the 5hmC hydroxyl formed a non-canonical hydrogen bond with the neighbouring G/O6. This oxygen is already part of a canonical (Watson-Crick) hydrogen bond with the C on the opposing strand. However, no structural differences between the DDD with and without the epigenetic mark were observed, suggesting that any effects that 5hmC might have had on the surrounding base pairs cannot be seen in a static structure. Using traditional NMMC as well as cNMMC we investigated the

hypothesis that the hydroxyl-group on 5hmC subtly interferes with the 3'-adjacent G-C base pair.

When performing traditional NMMC we found that the presence of a single 5hmC epigenetic mark in DDD only caused a minimal change in some of the helical parameters of the 3'-adjacent adjacent base pair. These results agree with the experimental data, which shows that a single 5hmC epigenetic mark has a limited structural effect on the surrounding helical parameters, which makes it difficult to detect. However, using customised Natural Moves we constrained the 5hmC hydroxyl group in the experimentally determined configuration and thereby increased non-canonical intra-strand hydrogen bonding during simulation. We were able to amplify some of the changes caused by the presence of 5hmC. The effect was even further increased when we deactivated the relative movement between 5hmC and the 3'-G, thereby effectively emulating the stabilising effect of an intra-strand hydrogen bond. Thus, using customised Natural Moves we were able to detect and amplify subtle structural effects on DNA helical parameters caused by a single epigenetic mark in the DDD.

Appendix A

Investigating structural effects of epigenetic marks on RNA models

A.1 MOSAICS Parameter Definitions

Header parameters

Header parameters for pdb file	
parameter	*meaning*
CBLC >AB	Constant Bond Length Closure use for chains A and B, moving in bond angle and torsional space.
CLOSCOMPL >AB	Propagation of fragments in chains A and B as complexes. Complexes are propagated in steps as defined by \prop_rot_sig{}, \prop_trans_sig{} and \prop_tors_sig{}.
STRIDE >FCCF	Indicates which residues are fixed (F) and allowed to move (C).

Simulation parameters

	~sim_gen_def[] parameters
\simulation_typ{ }	Defines the simulation type to be used. PT or PTMC: Parallel Tempering Monte Carlo MIN: energy minimization EEMC: Equi-energy Monte Carlo PTSTG or PTSTGMC: PTMC with staging transformation NM: Normal Modes NMEN: Elastic normal modes
\replica_number{0}	The highest order replica number in PTMC. Must be an integer ≥ 0. Using 0 defines PTMC with a single replica (conventional MC). Using n defines PTMC with $0, \ldots, n$, $n+1$ replicas.
\total_step_mc{10000}	Total number of steps taken in simulation. Must be an integer ≥ 1.
\local_step_md{1}	Number of molecular dynamics steps taken in each hybrid monte carlo step to generate moves. Only used in cartesian sampling. Must be an integer ≥ 1.
\time_step_md{0.4}	Time step for outer (inter-molecular energies) hybrid Monte Carlo local molecular dynamics move in fs, ≥ 0.
\intra_respa_step{1}	Time step for the inner loop of hybrid Monte Carlo is scaled by the inverse of this value. This is for the evaluation of bond, bond angle and torsional (intra-molecular) energies. Only used in cartesian sampling.
\statistics_freq{200}	Frequency of data printout. Must be an integer ≥ 1.
\prob_eemc_jump{0.15}	For PT simulations. The probability of attempting a PT jump, ≥ 0.

System parameters

~sim_mol_def[] parameters for input file	
parameter	*meaning*
\mol_parm_file{}	Input file with molecular topology.
\bond_database_file{}	Input file containing parameters of the bond interactions.
\bend_database_file{}	Input file containing bend angle potential parameters.
\tors_database_file{}	Input file containing torsional angle potential parameters.
\onfo_database_file{}	Input file containing onefour potential parameters.
\inter_database_file{}	Input file containing inter-molecular interaction parameters.
\pos_init_file{}	Input pdb file of initial conformation.
\pos_out_file{}	Output pdb file (only final conformation is saved).
\atom_pos_file{}	Output trajectory file.
\epot_file{}	Output potential energy file.
\einter_file{}	Output intermolecular energy file.
\cgres_model{}	Type of coarse grained model to use off (default): Full atomic details used. KB_1pt: Knowledge-based 1 center per residue model. KB_3pt: Knowledge-based 3 center per residue model. (Mostly used for proteins.) KB_5pt: Knowledge-based 3 center per residue model. (Mostly used for RNA.) KB_UA: Knowledge-based united atom model. Go: Go model BNL: BNL model
\tors_pos_file{}	Output file of torsional angles.

Bibliography

[1] H. M. Berman. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, jan 2000.

[2] Ron O Dror, Robert M Dirks, J P Grossman, Huafeng Xu, and David E Shaw. Biomolecular simulation: a computational microscope for molecular biology. *Annual review of biophysics*, 41:429–52, jan 2012.

[3] Jianhan Chen, Charles L Brooks, and Jana Khandogin. Recent advances in implicit solvent-based methods for biomolecular simulations. *Current opinion in structural biology*, 18(2):140–8, apr 2008.

[4] Alexey Onufriev. Chapter 7 Implicit Solvent Models in Molecular Dynamics Simulations: A Brief Overview. *Annual Reports in Computational Chemistry*, 4:125–137, 2008.

[5] J A McCammon, B R Gelin, and M Karplus. Dynamics of folded proteins. *Nature*, 267(5612):585–90, jun 1977.

[6] David E Shaw, Paul Maragakis, Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, Michael P Eastwood, Joseph A Bank, John M Jumper, John K Salmon, Yibing Shan, and Willy Wriggers. Atomic-level characterization of the structural dynamics of proteins. *Science (New York, N.Y.)*, 330(6002):341–6, oct 2010.

[7] Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, and David E Shaw. How fast-folding proteins fold. *Science (New York, N.Y.)*, 334(6055):517–20, oct 2011.

[8] Peter L Freddolino, Anton S Arkhipov, Steven B Larson, Alexander McPherson, and Klaus Schulten. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure (London, England : 1993)*, 14(3):437–49, mar 2006.

[9] Adam L. Beberg, Daniel L. Ensign, Guha Jayachandran, Siraj Khaliq, and Vijay S. Pande. Folding@home: Lessons from eight years of volunteer distributed computing. In *2009 IEEE International Symposium on Parallel & Distributed Processing*, pages 1–8. IEEE, may 2009.

[10] Modesto Orozco. A theoretical view of protein dynamics. *Chemical Society reviews*, 43(14):5051–66, jul 2014.

[11] Rafael C Bernardi, Marcelo C R Melo, and Klaus Schulten. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et biophysica acta*, 1850(5):872–877, may 2015.

[12] Mojie Duan, Jue Fan, Minghai Li, Li Han, and Shuanghong Huo. Evaluation of Dimensionality-Reduction Methods from Peptide FoldingâĂŞUnfolding Simulations. *Journal of Chemical Theory and Computation*, 9(5):2490–2497, may 2013.

[13] Marissa G Saunders and Gregory A Voth. Coarse-graining methods for computational biology. *Annual review of biophysics*, 42:73–93, jan 2013.

[14] Zhiyong Zhang, Lanyuan Lu, Will G Noid, Vinod Krishna, Jim Pfaendtner, and Gregory A Voth. A systematic methodology for defining coarse-grained sites in large biomolecules. *Biophysical journal*, 95(11):5073–83, dec 2008.

[15] Ivet Bahar, Timothy R Lezon, Lee-Wei Yang, and Eran Eyal. Global dynamics of proteins: bridging between structure and function. *Annual review of biophysics*, 39:23–42, jan 2010.

[16] Juan R Perilla and Thomas B Woolf. Towards the prediction of order parameters from molecular dynamics simulations in proteins. *The Journal of chemical physics*, 136(16):164101, apr 2012.

[17] Ivet Bahar and A J Rader. Coarse-grained normal mode analysis in structural

biology. *Current opinion in structural biology*, 15(5):586–92, oct 2005.
[18] Peter Minary and Michael Levitt. Conformational optimization with natural degrees of freedom: a novel stochastic chain closure algorithm. *Journal of computational biology : a journal of computational molecular cell biology*, 17(8):993–1010, aug 2010.
[19] Adelene Y L Sim, Michael Levitt, and Peter Minary. Modeling and design by hierarchical natural moves. *Proceedings of the National Academy of Sciences of the United States of America*, 109(8):2890–5, mar 2012.
[20] Bernhard Knapp, Samuel Demharter, Reyhaneh Esmaielbeiki, and Charlotte M Deane. Current status and future challenges in T-cell receptor / peptide / MHC molecular dynamics simulations. *Briefings in Bioinformatics*, (November 2014):1–10, 2015.
[21] Peter Minary and Michael Levitt. Probing protein fold space with a simplified model. *Journal of molecular biology*, 375(4):920–933, jan 2008.
[22] Junjie Zhang, Peter Minary, and Michael Levitt. Multiscale natural moves refine macromolecules using single-particle electron microscopy projection images. *Proceedings of the National Academy of Sciences of the United States of America*, 109(25):9845–50, jun 2012.
[23] Adelene Y L Sim, Peter Minary, and Michael Levitt. Modeling nucleic acids. *Current opinion in structural biology*, 22(3):273–8, jun 2012.
[24] Ignacio Moraga, Gerlinde Wernig, Stephan Wilmes, Vitalina Gryshkova, ChristianÂăP. Richter, Wan-Jen Hong, Rahul Sinha, Feng Guo, Hyna Fabionar, TomÂăS. Wehrman, Peter Krutzik, Samuel Demharter, Isabelle Plo, IrvingÂăL. Weissman, Peter Minary, Ravindra Majeti, StefanÂăN. Constantinescu, Jacob Piehler, and K.ÂăChristopher Garcia. Tuning Cytokine Receptor Signaling by Re-orienting Dimer Geometry with Surrogate Ligands. *Cell*, 160(6):1196–1208, 2015.
[25] L.R. Dodd, T.D. Boone, and D.N. Theodorou. A concerted rotation algorithm for atomistic Monte Carlo simulation of polymer melts and glasses. *Molecular Physics*, 78(4):961–996, mar 1993.
[26] P. V. Krishna Pant and Doros N. Theodorou. Variable Connectivity Method for the Atomistic Monte Carlo Simulation of Polydisperse Polymer Melts. *Macromolecules*, 28(21):7224–7234, oct 1995.
[27] Minghong G. Wu and Michael W. Deem. Analytical Rebridging Monte Carlo: Application to cis/trans Isomerization in Proline-Containing, Cyclic Peptides. apr 1999.
[28] Daniel Hoffmann and Ernst-Walter Knapp. Polypeptide folding with off-lattice Monte Carlo dynamics: the method. *European Biophysics Journal*, 24(6):387–403, 1996.
[29] Heinz Sklenar, Daniel Wüstner, and Remo Rohs. Using internal and collective variables in Monte Carlo simulations of nucleic acid structures: chain breakage/closure algorithm and associated Jacobians. *Journal of computational chemistry*, 27(3):309–15, feb 2006.
[30] Peter Minary and Michael Levitt. Training-free atomistic prediction of nucleosome occupancy. *Proceedings of the National Academy of Sciences of the United States of America*, 111(17):6293–8, apr 2014.
[31] Bernhard Knapp, Samuel Demharter, Charlotte M Deane, and Peter Minary. Exploring peptide/MHC detachment processes using hierarchical natural move Monte Carlo. *Bioinformatics (Oxford, England)*, 32(2):181–6, jan 2016.
[32] Samuel Demharter, Bernhard Knapp, C.M. Charlotte M. Deane, and Peter Minary. Modeling Functional Motions of Biological Systems by Customized Natural Moves. *Biophysical Journal*, 111(4):710–721, aug 2016.
[33] A G Ladurner, L S Itzhaki, V Daggett, and A R Fersht. Synergy between simulation and experiment in describing the energy landscape of protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, 95(15):8473–8, jul 1998.
[34] Ryan W Benz, Francisco Castro-Román, Douglas J Tobias, and Stephen H White.

Experimental validation of molecular dynamics simulations of lipid bilayers: a new approach. *Biophysical journal*, 88(2):805–17, feb 2005.

[35] Kresten Lindorff-Larsen, Nikola Trbovic, Paul Maragakis, Stefano Piana, and David E Shaw. Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. *Journal of the American Chemical Society*, 134(8):3787–91, feb 2012.

[36] Sebastian Kmiecik and Andrzej Kolinski. Characterization of protein-folding pathways by reduced-space modeling. *Proceedings of the National Academy of Sciences of the United States of America*, 104(30):12330–5, jul 2007.

[37] Michael Levitt and Arieh Warshel. Computer simulation of protein folding. *Nature*, 253(5494):694–698, feb 1975.

[38] Valentina Tozzini, Joanna Trylska, Chia-en Chang, and J Andrew McCammon. Flap opening dynamics in HIV-1 protease explored with a coarse-grained model. *Journal of structural biology*, 157(3):606–15, mar 2007.

[39] I Bahar and R L Jernigan. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *Journal of molecular biology*, 266(1):195–214, feb 1997.

[40] A Voegler Smith and C K Hall. alpha-helix formation: discontinuous molecular dynamics on an intermediate-resolution protein model. *Proteins*, 44(3):344–60, aug 2001.

[41] âĂă Siewert J. Marrink, *, âĂă H. Jelger Risselada, âĂą Serge Yefimov, , § D. Peter Tieleman, and Alex H. de VriesâĂă. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. 2007.

[42] Paul K. Weiner and Peter A. Kollman. AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions. *Journal of Computational Chemistry*, 2(3):287–303, 1981.

[43] Walter R. P. Scott, Philippe H. Hünenberger, Ilario G. Tironi, Alan E. Mark, Salomon R. Billeter, Jens Fennen, Andrew E. Torda, Thomas Huber, Peter Krüger, and Wilfred F. van Gunsteren. The GROMOS Biomolecular Simulation Program Package. *The Journal of Physical Chemistry A*, 103(19):3596–3607, may 1999.

[44] Bernard R. Brooks, Robert E. Bruccoleri, Barry D. Olafson, David J. States, S. Swaminathan, and Martin Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2):187–217, 1983.

[45] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926, jul 1983.

[46] B. E. Hingerty, R. H. Ritchie, T. L. Ferrell, and J. E. Turner. Dielectric effects in biopolymers: The theory of ionic saturation revisited. *Biopolymers*, 24(3):427–439, mar 1985.

[47] M J Sippl. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *Journal of molecular biology*, 213(4):859–83, jun 1990.

[48] Shoshana J. Wodak and Marianne J. Rooman. Generating and testing protein folds. *Current Opinion in Structural Biology*, 3(2):247–259, apr 1993.

[49] Christopher M Summa and Michael Levitt. Near-native structure refinement using in vacuo energy minimization. *Proceedings of the National Academy of Sciences of the United States of America*, 104(9):3177–82, feb 2007.

[50] R Samudrala and J Moult. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *Journal of molecular biology*, 275(5):895–916, feb 1998.

[51] H Taketomi, Y Ueda, and N GÅD. Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions. *International journal of peptide and protein research*, 7(6):445–59, jan 1975.

[52] D Hamada, S Segawa, and Y Goto. Non-native alpha-helical intermediate in the

refolding of beta-lactoglobulin, a predominantly beta-sheet protein. *Nat Struct Biol*, 3(10):868–73., 1996.

[53] Paul C Whitford, Karissa Y Sanbonmatsu, José N Onuchic, Cech T R, Guerriertakada C Mcclain W H, Altman S, Yonath A, Sengupta J Gao N Frank J, Gao H, Taylor D J, Baucom A Lieberman K Earnest T N Cate J H Yusupov M M, Yusupova G Z, Noller H F, Will C L Wahl M C, Lührmann R, Ghildiyal M D, Zamore P, Bartel D P, Sonenberg N Fabian M R, Filipowicz W, Simon I Fuxreiter M, Bondos S, Vallabhapurapu S M, Karin, Nieto M A, Baldwin A S, Neely K E Vignali M, Hassan A H, Workman J L, Reck-Peterson S L Olave I A, Crabtree G R, Levinthal C DeBrunner J T P Munck E, Dill K A, Thirumalai D C, Hyeon, Montal M Leopold P E, Onuchic J N, Bryngelson J D G, Wolynes P, Bryngelson J D G, Wolynes P, Socci N D Bryngelson J D, Onuchic J N, Wolynes P G, Schnapp B J Svoboda K, Schmidt C F, Block S M, Forde N R Bustamante C, Chemla Y R, Izhaky D, Hanson J Tan Y, Yang H, Vale R D A, Milligan R, Thai V Kerns S J Karplus M Henzler-Wildman K A, Lei M, Kern D, Sligar S G Frauenfelder H, Wolynes P G, Socci N D J, Onuchic, Luthey-Schulten Z Onuchic J N, Wolynes P G, Onuchic J N Shea J E, Brooks C L, Onuchic J N Shea J E, Brooks C L, Onuchic J N Cho S S, Levy Y, Wolynes P G, Komives E A Ferreiro D U, Hegler J A, Wolynes P G, Sherrington D S, Kirkpatrick, Edwards S F W, Anderson P, Binder K P, Young A, Onsager L, Ising E, Glotzer S C, Sastry S, Angell C A, Schmalian J Stevenson J D, Wolynes P G, Xia X Y G, Wolynes P, Thirumalai D Kirkpatrick T R, Wolynes P G, Garrahan J P D, Chandler, Angell C A Ediger M D, Nagel S R, Inoue A, Frauenfelder H, Iben I et Al, Bryngelson J D G, Wolynes P, Luthey-Schulten Z A Goldstein R A, Wolynes P G, GÅD N, Nymeyer H Clementi C, Onuchic J N, Chavez L L Cheung M S, Onuchic J N, Gosavi S Schug A Sanbonmatsu K Y Whitford P C, Noel J K, Onuchic J N, Yue K Z Fiebig K M Yee D P Thomas P D Dill K A, Bromberg S, Chan H S, Shakhnovich E Sali A, Karplus M, GÅD N H, Taketomi, Socci N D N, Onuchic J, Cieplak M Maritan A Shrivastava I, Vishveshwara S, Banavar J R, Onuchic J N Socci N D, Wolynes P G, Garcia A E Chahine J Onuchic J N, Nymeyer H, Socci N D, Camacho C J D, Thirumalai, Klimov D K D, Thirumalai, Sorenson J M T, Head-Gordon, Gutin A M Shakhnovich E, Farztdinov G, Karplus M, Karplus M Dinner A, Sali A, Shakhnovich E, Saven J G G, Wolynes P, Onuchic J N Yang S, Levine H, Garcia A E Yang S, Onuchic J N, Levine H, Onuchic J N Chavez L L, Clementi C, Leite V B P Chahine J, Oliveira R J, Wang J, Hummer G, Best R G, Hummer, Stamati H Kavraki L E Das P, Moll M, Clementi C, Chahine J Leite V B P Oliveira R J, Whitford P C, Wang J, Chahine J Wang J Onuchic J N Oliveira R J, Whitford P C, Leite V B P, Levy Y Cho S S, Wolynes P G, Guo Z L, Brooks C, Garcia A E Cheung M S, Onuchic J N, Levy Y N, Onuchic J, Leite V B P Oliveira L C, Silva R T H, Chahine J, Onuchic J N Hillson N, Garcia A E, Wales D J J, Dewsbury P E, Karanicolas J L, Brooks C, Hyeon C Thirumalai D Pincus D L, Cho S S, P'Brien O E Hu C-K Kouza M, Li M S, Thirumalai D, Matsunaga Y Rylance G J Johnston R L Komatsuzaki T, Hoshino K, Wales D J, Wallin S Chan H S, Zhang Z, Liu Z, Hills R D L, Brooks C, Hyeon C D, Thirumalai, Plotkin S S N, Onuchic J, Plotkin S S, Givaty O Y, Levy, Matysiak S Das P, Clementi C, Samiotakis A S, Cheung M, Sułkowska J I M, Cieplak, Sułkowska J I M, Cieplak, Pincus D L Hyeon C, Morrison G, Thirumalai D, Cieplak M I, Sułkowska J, Vendruscolo M O'Brien E P, Christodoulou J, Dobson C M, Bellesia G E, Shea J, Eastwood M P G, Wolynes P, Wang J Plotkin S S, Wolynes P G, Jennings P A Clementi C, Onuchic J N, Jennings P A Gosavi S, Chavez L L, Onuchic J N, Jennings P A Chavez L L, Gosavi S, Onuchic J N, Finke J M Onuchic J N Andrews B T, Gosavi S, Jennings P A, Jennings P A Gosavi S, Whitford P C, Onuchic J N, Norcross T S O, Yeates T, Wolynes P G McCammon J A, Karplus M, Garcia A E N, Onuchic J, Norcross T S Yeates T O, King N P, Sawaya M R Goldschmidt L King N P, Jacobitz A W, Yeates T O, Finke J M Heidary D K Onuchic J N Roy M, Chavez L L, Jennings P A, Daumy G O Cong Y Heidary D K, Roy M, Jennings P A, Szymczak P Sułkowska J I, Sulkowski P, Cieplak M, Sułkowska J I Noel J K, Onuchic J N, Onuchic J N Capraro D T, Roy M, Jennings P A, Jennings P A Baxter E L, Onuchic J N, Garcia A E Clementi C, Onuchic J N, Zhou Y A, Linhananta, Ding J Luo Z, Zhou Y, Saunders J Hennelly S P Onuchic J N Whit-

ford P C, Schug A, Sanbonmatsu K Y, Kussell E L Shimada J, Shakhnovich E I, Rhee Y M Jayachandran G Vishal V Sorin E J, Nakatani B J, Pande V S, Lai J Kim H Abeysirigunawardena S Mayerle M Woodson S A Ha T Chen K, Eargle J, Schulten Z, Shaw D E et Al, Powers G Lombardgillooly K Shuster D Mcintyre K W Ryan D E Levin W Madison V Greenfeder S A, Varnell T, Ju G, Caffes P Vigers G P A, Anderson L J, Brandhuber B J, Kemp G J L Koussounadis A I, Ritchie D W, Secombes C J, Onuchic J N Miyashita O, Wolynes P G, Onuchic J N Whitford P C, Wolynes P G, Itoh K M, Sasai, Takada S Portman J J, Wolynes P G, Sułkowska J I M, Cieplak, Schug A Lammert H, Onuchic J N, Thirumalai D H, Lorimer G, Huse M J, Kuriyan, Spahn C M T Munro J B, Sanbonmatsu K Y, Blanchard S C, Sakai H Chong K T Takeuchi S Nakagawa A Nada S Okada M Ogawa A, Takayama Y, Tsukihara T, Miyashita O Miller M Tasken K Onuchic J N Adams J A Woods V L Wong L, Lieser S A, Jennings P A, Henzler-Wildman K Hadjipavlou G Eisenmesser E Z Wolf-Watz M, Thai V, Kern D, Block S M, Ermolenko D N Korostelev A, Noller H F, Dorywalska M Marshall R A, Aitken C E, Puglisi J D, Blanchard S, Tsai C J Wolfson H Kumar S, Ma B, Nussinov R, Goldstein H, Chennubhotla C Bahar I, Tobi D, Lee-Wei Yang Chennubhotla C, Rader A J, Ivet Bahar, Bahar I Wang Y, Rader A J, Jernigan R L, Tama F H, Sanejouand Y, Tirion M M, Gosavi S Whitford P C, Onuchic J N, Sen T Z Kloczkowski A Kurkcuoglu O, Doruker P, Jernigan R L, Doruker P Kurkcuoglu O, Kurkcuoglu Z, Jernigan R L, Garcia J V Kloczkowski A Sen T Z, Feng Y, Jernigan R L, Watts S D Bahar I Jiang J, Shrivastava I H, Amara S G, Frank J Tama F, Mikel Valle, Brooks C L, Tozzini V Trylska J, McCammon J A, Geggier P Terry D Munro J B Onuchic J N Spahn C M T Sanbonmatsu K Y Blanchard S C Whitford P C, Altman R B, Eargle J Cornish P Ha T Luthey-Schulten Z Trabuco L G, Schreiner E, Schulten K, Maragakis P M, Karplus, Qasba P K Ramakrishnan B Schuyler A D, Jernigan R L, Chirikjian G S, Kubitzki M B L, de Groot B, Lu Q J, Wang, Phillips G N Daily M D, Cui Q, Korkut A W, Hendrickson, Biswas R Jana B, Adkar B V, Bagchi B, Haliloglu T Kantarci-Carsibasi N, Doruker P, Krumhansl J A García A E, Frauenfelder H, Korostelev A F, Noller H, Lu M Vyas N K Quiocho F A Wang Q Poon B K, Chen X, Jianpeng Ma, Best R B Chen Y-G Hummer G, Levy Y Whitford P C, Miyashita O, Onuchic J N, Okazaki K S, Takada, Darden T-Gohlke H Luo R Merz K M Onufriev A Simmerling C Wang B Case D A, Cheatham T E, Woods R J, Brooks B et Al, Chen J L, Brooks C, Karplus M A, McCammon J, Gorfe A A Grant B J, McCammon J A, Dror R O Klepeis J L, Lindorff-Larsen K, Shaw D E, Zuckerman D M, Brokaw J B J-W, Chu, Arora K L, Brooks C, Watkins L P Bhattacharyya S Brokaw J Chu J-W Hanson J A, Duderstadt K, Yang H, Henzler-Wildman K A et Al, Vendruscolo M M, Dobson C, Dror R O Lindorff-Larsen K, Piana S, Shaw D E, Petridis L Schulz R, Lindner B, Smith J C, Wüller M C E, Schulz G, Reinstein J Müller C W, Schlauderer G J, Schulz G E, Proba K Schlauderer G J, Schulz G E, Takada S Onuchic J N Okazaki K, Koga N, Wolynes P G, Yang S B, Roux, Zwanzig R, Kramers H, Hofrichter J Kubelka J, Eaton W A, Saven J G Tang J, Kang S-G, Gai F, Onuchic J N Whitford P C, Sanbonmatsu K Y, Okamoto Y, Gnanakaran S Nymeyer H, Garcia A E, Rodinger T R, Pomes, Portman J Sanbonmatsu K Y Gnanakaran S, Nymeyer H, Garcia A E, Roux B, Trygubenko S A Carr J M, Wales D J, Vanden-Eijnden E Maragliano L, Fischer A, Ciccotti G, Terrell R Sheppard D, Henkelman G, Koslover E F J, Wales D, Bell A T Peters B, Heyden A, Chakraborty A, Uberuaga B P Henkelman G, Jonsson H, Sugita Y Y, Okamoto, Garcia A E Y, Sanbonmatsu K, Berne B J Zhou R H, Germain R, Zhou R H, Rao F A, Caflisch, Baumketner A J-E, Shea, Khalili M Scheraga H A, Liwo A, Mathews D H A, Case D, Wales D J, Vanden-Eijnden E Miller T F, Chandler D, Haldane J S B, Pauling L, Horiuchi T N, GÅD, Bae E N, Phillips G, Wolynes P G Miyashita O, Onuchic J N, Sparrman T Wallgren M Olsson U Rundqvist L, Adén J, Wolf-Watz M, Li J Schrank T, Elam W, Hilser V J, Wolynes P G Lammert H, Onuchic J N, Hyeon C N, Onuchic J, Hyeon C N, Onuchic J, Tripathi S J, Portman J, Adams J A Hyeon C, Jennings P A, Onuchic J N, Olsson U M, Wolf-Watz, Zhang L Peng C, Head-Gordon T, Onuchic J N Walczak A M, Wolynes P G, Onuchic J N Schultz D, Ben Jacob E, Wolynes P G, Portman J J Shoemaker B A, Wolynes P G, Onuchic J N Levy Y, Cho S S, Wolynes P G, Onuchic J N Levy Y, Wolynes P G, Sarkar K Gruebele M Chen K, Eargle J, Luthey-Schulten Z, Vuzman D Y, Levy, Sturte-

vant J M Munson M, Obrien R, Regan L, Shen T Onuchic J N Levy Y, Cho S S, Wolynes P G, Levy Y Schug A, Whitford P C, Onuchic J N, Lemke E A Lavinder J J Ferreon A C M Magliery T J Onuchic J N Gambin Y, Schug A, Deniz A A, Breaker R R, Penedo J C Blouin S, Mulhbacher J, Lafontaine D A, Nahvi A Winkler W, Breaker R R, Hennelly S P Y, Sanbonmatsu K, Dussault A-M Mulhbacher J Ennifar E Penedo J C Heppell B, Blouin S, Lafontaine D A, Lin J-C D, Thirumalai, Walter N G Feng J, Brooks C L, Foster D A N Woodside M T Greenleaf W J, Frieda K L, Block S M, Green R F, Noller H, Hopfield J J, Rodnina M V W, Wintermeyer, Kim H D Chu S Blanchard S C, Gonzalez R L, Puglisi J D, Kelley A C Gao Y-G Murphy F V Weir J R Schmeing T M, Voorhees R M, Ramakrishnan V, Feldman M B Terry D S-Altman R B Munro J B Geggier P, Dave R, Blanchard S C, Gao H Li W Valle M-Zavialov A Frank J, Sengupta J, Ehrenberg M, Lovmar M Johansson M, Ehrenberg M, Mittelstaet J Konevega A L Wohlgemuth I, Pohl C, Rodnina M V, Frank J T, Spahn C M, Joseph S Sanbonmatsu K Y, Tung C-S, Altman R B Blanchard S C-Onuchic J N Whitford P C, Geggier P, Sanbonmatsu K Y, Rodnina M V Byrne R T, Konevega A L, Antson A A, Shi H B, Moore P, Wintermeyer W Rodnina M V Fischer N, Konevega A L, Stark H, Chladek S M, Sprinzl, Eargle J Rebecca W A, Luthey-Schulten Z, Sethi A Trabuco L G Eargle J, Black A A, Luthey-Schulten Z, Bashan A et Al, Sanbonmatsu K Y, Dalke A Humphrey W, and Schulten K. Biomolecular dynamics: orderâĂŞdisorder transitions and energy landscapes. *Reports on Progress in Physics*, 75(7):076601, jul 2012.

[54] Ronald D. Hills and Charles L. Brooks. Insights from Coarse-Grained GÅD Models for Protein Folding and Dynamics. *International Journal of Molecular Sciences*, 10(3):889–905, mar 2009.

[55] Hüseyin Kaya and Hue Sun Chan. Solvation Effects and Driving Forces for Protein Thermodynamic and Kinetic Cooperativity: How Adequate is Native-centric Topological Modeling? *Journal of Molecular Biology*, 326(3):911–931, 2003.

[56] MM Tirion. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Physical review letters*, 77(9):1905–1908, aug 1996.

[57] Yongmei Wang, A J Rader, Ivet Bahar, and Robert L Jernigan. Global ribosome motions revealed with elastic network model. *Journal of structural biology*, 147(3):302–14, sep 2004.

[58] I Bahar, A R Atilgan, and B Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding & design*, 2(3):173–81, jan 1997.

[59] B Brooks and M Karplus. Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme. *Proceedings of the National Academy of Sciences of the United States of America*, 82(15):4995–9, aug 1985.

[60] N Go, T Noguti, and T Nishikawa. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proceedings of the National Academy of Sciences of the United States of America*, 80(12):3696–700, jun 1983.

[61] M Levitt, C Sander, and P S Stern. Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *Journal of molecular biology*, 181(3):423–47, feb 1985.

[62] Lei Yang, Guang Song, and Robert L Jernigan. Protein elastic network models and the ranges of cooperativity. *Proceedings of the National Academy of Sciences of the United States of America*, 106(30):12347–52, jul 2009.

[63] Lei Yang, Guang Song, and Robert L. Jernigan. How Well Can We Understand Large-Scale Protein Motions Using Normal Modes of Elastic Network Models? *Biophysical Journal*, 93(3):920–929, 2007.

[64] Ivet Bahar, Burak Erman, Robert L. Jernigan, Ali Rana Atilgan, and David G. Covell. Collective Motions in HIV-1 Reverse Transcriptase: Examination of Flexibility and Enzyme Function. *Journal of Molecular Biology*, 285(3):1023–1037, jan 1999.

[65] Basak Isin, Pemra Doruker, and Ivet Bahar. Functional motions of influenza virus hemagglutinin: a structure-based analytical approach. *Biophysical journal*, 82(2):569–81, feb 2002.

[66] Qiang Cui, Guohui Li, Jianpeng Ma, and Martin Karplus. A Normal Mode Analysis of Structural Plasticity in the Biomolecular Motor F1-ATPase. *Journal of Molecular Biology*, 340(2):345–372, jul 2004.

[67] Adam Van Wynsberghe, Guohui Li, and Qiang Cui. Normal-Mode Analysis Suggests Protein Flexibility Modulation throughout RNA Polymerase's Functional Cycle. *Biochemistry*, 43(41):13083–13096, oct 2004.

[68] O Keskin, I Bahar, D Flatow, D G Covell, and R L Jernigan. Molecular mechanisms of chaperonin GroEL-GroES function. *Biochemistry*, 41(2):491–501, jan 2002.

[69] F. Tama, M. Valle, J. Frank, and C. L. Brooks. Dynamic reorganization of the functionally active ribosome explored by normal mode analysis and cryo-electron microscopy. *Proceedings of the National Academy of Sciences*, 100(16):9319–9323, aug 2003.

[70] Florence Tama and Charles L. Brooks. Diversity and Identity of Mechanical Properties of Icosahedral Viral Capsids Studied with Elastic Network Normal Mode Analysis. *Journal of Molecular Biology*, 345(2):299–314, 2005.

[71] Sibsankar Kundu, Julia S Melton, Dan C Sorensen, George N Phillips, and Jr. Dynamics of proteins in crystals: comparison of experiment with simple models. *Biophysical journal*, 83(2):723–32, aug 2002.

[72] Raffaella Burioni, Davide Cassi, Fabio Cecconi, and Angelo Vulpiani. Topological thermal instability and length of proteins. *Proteins: Structure, Function, and Bioinformatics*, 55(3):529–535, apr 2004.

[73] Sibsankar Kundu, Dan C. Sorensen, and George N. Phillips. Automatic domain decomposition of proteins by a Gaussian Network Model. *Proteins: Structure, Function, and Bioinformatics*, 57(4):725–733, dec 2004.

[74] K D Gibson and H A Scheraga. Minimization of polypeptide energy. I. Preliminary structures of bovine pancreatic ribonuclease S-peptide. *Proceedings of the National Academy of Sciences of the United States of America*, 58(2):420–7, aug 1967.

[75] S Kirkpatrick, C D Gelatt, and M P Vecchi. Optimization by simulated annealing. *Science (New York, N.Y.)*, 220(4598):671–80, may 1983.

[76] Robert H. Swendsen and Jian-Sheng Wang. Replica Monte Carlo Simulation of Spin-Glasses. *Physical Review Letters*, 57(21):2607–2609, nov 1986.

[77] Fritz M. Pohl and Thomas M. Jovin. Salt-induced co-operative conformational change of a synthetic DNA: Equilibrium and kinetic studies with poly(dG-dC). *Journal of Molecular Biology*, 67(3):375–396, 1972.

[78] V I Ivanov, L E Minchenkova, E E Minyat, M D Frank-Kamenetskii, and A K Schyolkina. The B to A transition of DNA in solution. *Journal of molecular biology*, 87(4):817–33, aug 1974.

[79] W K Olson, A A Gorin, X J Lu, L M Hock, and V B Zhurkin. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proceedings of the National Academy of Sciences of the United States of America*, 95(19):11163–8, sep 1998.

[80] Xiang-Jun Lu, Zippora Shakked, and Wilma K. Olson. A-form Conformational Motifs in Ligand-bound DNA Structures. *Journal of Molecular Biology*, 300(4):819–840, jul 2000.

[81] R Dickerson, M Bansal, C R Calladine, S Diekmann, W Hunter, O Kennard, E von Kitzing, R Lavery, H Nelson, W Olson, W Saenger, Z Shakked, H Sklenar, D Soumpasis, C-S Tung, A-J Wang, and V Zhurkin. Definitions and nomenclature of nucleic acid structure parameters. *J. Mol. Biol.*, 205(4):787–91, 1989.

[82] Xiang-Jun Lu and Wilma K Olson. Resolving the discrepancies among nucleic acid conformational analyses. *Journal of Molecular Biology*, 285(4):1563–1575, jan 1999.

[83] Wilma K Olson, Manju Bansal, Stephen K Burley, Richard E Dickerson, Mark Gerstein, Stephen C Harvey, Udo Heinemann, Xiang-Jun Lu, Stephen Neidle, Zippora Shakked, Heinz Sklenar, Masashi Suzuki, Chang-Shung Tung, Eric Westhof, Cynthia Wolberger, and Helen M Berman. A standard reference frame for the description of nucleic acid base-pair geometry. *Journal of Molecular Biology*, 313(1):229–237, 2001.

[84] D.J. Klein, T M Schmeing, P B Moore, and T A Steitz. The kink-turn: a new RNA secondary structure motif. *The EMBO Journal*, 20(15):4214–4221, aug 2001.
[85] ChulHee Kang, Xiaohua Zhang, Robert Ratliff, Robert Moyzis, and Alexander Rich. Crystal structure of four-stranded Oxytricha telomeric DNA. *Nature*, 356(6365):126–131, mar 1992.
[86] Stephen. Neidle. *Principles of nucleic acid structure*. Elsevier, 2008.
[87] Elene V. Hackl, Svetlana V. Kornilova, and Yurij P. Blagoi. DNA structural transitions induced by divalent metal ions in aqueous solutions. *International Journal of Biological Macromolecules*, 35(3-4):175–191, apr 2005.
[88] Eun-Ang Raiber, Pierre Murat, Dimitri Y Chirgadze, Dario Beraldi, Ben F Luisi, and Shankar Balasubramanian. 5-Formylcytosine alters the structure of the DNA double helix. *Nature Structural & Molecular Biology*, 22(1):44–49, dec 2014.
[89] Ken A Dill and Justin L MacCallum. The protein-folding problem, 50 years on. *Science (New York, N.Y.)*, 338(6110):1042–6, nov 2012.
[90] R. H. Austin, K. W. Beeson, L. Eisenstein, H. Frauenfelder, and I. C. Gunsalus. Dynamics of ligand binding to myoglobin. *Biochemistry*, 14(24):5355–5373, dec 1975.
[91] H Frauenfelder, S G Sligar, and P G Wolynes. The energy landscapes and motions of proteins. *Science (New York, N.Y.)*, 254(5038):1598–603, dec 1991.
[92] Hans Frauenfelder, Gregory A. Petsko, and Demetrius Tsernoglou. Temperature-dependent X-ray diffraction as a probe of protein structural dynamics. *Nature*, 280(5723):558–563, aug 1979.
[93] Katherine A. Henzler-Wildman, Ming Lei, Vu Thai, S. Jordan Kerns, Martin Karplus, and Dorothee Kern. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature*, 450(7171):913–916, dec 2007.
[94] Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, 450(7172):964–72, dec 2007.
[95] Toshiko Ichiye and Martin Karplus. Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins: Structure, Function, and Genetics*, 11(3):205–217, nov 1991.
[96] A M Lesk and C Chothia. Mechanisms of domain closure in proteins. *Journal of molecular biology*, 174(1):175–91, mar 1984.
[97] Ugur Emekli, Dina Schneidman-Duhovny, Haim J Wolfson, Ruth Nussinov, and Turkan Haliloglu. HingeProt: automated prediction of hinges in protein structures. *Proteins*, 70(4):1219–27, mar 2008.
[98] Mark Gerstein and Nathaniel Echols. Exploring the range of protein flexibility, from a structural proteomics perspective. *Current Opinion in Chemical Biology*, 8(1):14–19, 2004.
[99] Dominique Bourgeois and Antoine Royant. Advances in kinetic protein crystallography. *Current Opinion in Structural Biology*, 15(5):538–547, oct 2005.
[100] I Schlichting, J Berendzen, K Chu, A M Stock, S A Maves, D E Benson, R M Sweet, D Ringe, G A Petsko, and S G Sligar. The catalytic pathway of cytochrome p450cam at atomic resolution. *Science (New York, N.Y.)*, 287(5458):1615–22, mar 2000.
[101] S. Walter Englander. Hydrogen exchange and mass spectrometry: A historical perspective. *Journal of the American Society for Mass Spectrometry*, 17(11):1481–1489, nov 2006.
[102] Arthur G. Palmer. NMR Characterization of the Dynamics of Biomacromolecules. *Chemical Reviews*, 104(8):3623–3640, aug 2004.
[103] K Pervushin, R Riek, G Wider, and K Wüthrich. Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proceedings of the National Academy of Sciences of the United States of America*, 94(23):12366–71, nov 1997.
[104] Remco Sprangers and Lewis E Kay. Quantitative dynamics and binding studies of the 20S proteasome by NMR. *Nature*, 445(7128):618–22, feb 2007.

[105] Reto Horst, Eric B Bertelsen, Jocelyne Fiaux, Gerhard Wider, Arthur L Horwich, and Kurt Wüthrich. Direct NMR observation of a substrate protein bound to the chaperonin GroEL. *Proceedings of the National Academy of Sciences of the United States of America*, 102(36):12748–53, sep 2005.

[106] Shakeel Ahmad Shahid, Benjamin Bardiaux, W Trent Franks, Ludwig Krabben, Michael Habeck, Barth-Jan van Rossum, and Dirk Linke. Membrane-protein structure determination by solid-state NMR spectroscopy of microcrystals. *Nature Methods*, 9(12):1212–1217, nov 2012.

[107] Sua Myong, Benjamin C. Stevens, and Taekjip Ha. Bridging Conformational Dynamics and Function Using Single-Molecule Spectroscopy. *Structure*, 14(4):633–643, 2006.

[108] L Stryer and R P Haugland. Energy transfer: a spectroscopic ruler. *Proceedings of the National Academy of Sciences of the United States of America*, 58(2):719–26, aug 1967.

[109] Linda Columbus and Wayne L Hubbell. A new spin on protein dynamics. *Trends in Biochemical Sciences*, 27(6):288–295, 2002.

[110] Martin Karplus and Joseph N. Kushick. Method for estimating the configurational entropy of macromolecules. *Macromolecules*, 14(2):325–332, mar 1981.

[111] Turkan Haliloglu, Ivet Bahar, and Burak Erman. Gaussian Dynamics of Folded Proteins. *Physical Review Letters*, 79(16):3090–3093, oct 1997.

[112] D J Jacobs, A J Rader, L A Kuhn, and M F Thorpe. Protein flexibility predictions using graph theory. *Proteins*, 44(2):150–65, aug 2001.

[113] Stephen Wells, Scott Menor, Brandon Hespenheide, and M F Thorpe. Constrained geometric simulation of diffusive motion in proteins. *Physical Biology*, 2(4):S127–S136, nov 2005.

[114] Harold A. Scheraga, Mey Khalili, and Adam Liwo. Protein-Folding Dynamics: Overview of Molecular Simulation Techniques. *Annual Review of Physical Chemistry*, 58(1):57–83, may 2007.

[115] Emanuele Paci and Martin Karplus. Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulations. *Journal of Molecular Biology*, 288(3):441–459, may 1999.

[116] Donald Hamelberg, John Mongan, and J. Andrew McCammon. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *The Journal of Chemical Physics*, 120(24):11919, jun 2004.

[117] Stewart A. Adcock And and J. Andrew McCammon*. Molecular Dynamics:âĂĽ Survey of Methods for Simulating the Activity of Proteins. 2006.

[118] Benoît Roux. The calculation of the potential of mean force using computer simulations. *Computer Physics Communications*, 91(1):275–282, 1995.

[119] Peter G. Bolhuis, David Chandler, Christoph Dellago, and Phillip L. Geissler. T RANSITION P ATH S AMPLING : Throwing Ropes Over Rough Mountain Passes, in the Dark. *Annual Review of Physical Chemistry*, 53(1):291–318, oct 2002.

[120] Arthur F. Voter. A method for accelerating the molecular dynamics simulation of infrequent events. *http://oasc12039.247realmedia.com/RealMedia/ads/click_lx.ads/www.aip.org/pt/adcenter/pdf 37/20939943/x01/AIP-PT/JCP_ArticleDL_0117/PTBG_orange_1640x440.jpg/434f71374e3* jun 1998.

[121] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(1-2):141–151, nov 1999.

[122] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12562–6, oct 2002.

[123] S. Doerr, M. J. Harvey, Frank No?, and G. De Fabritiis. HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *Journal of Chemical Theory and Computation*, 12(4):1845–1852, apr 2016.

[124] Ulrich H.E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chemical Physics Letters*, 281(1-3):140–150, dec 1997.

[125] Benedict Leimkuhler and Sebastian Reich. A Metropolis adjusted Nosé-Hoover thermostat. *ESAIM: Mathematical Modelling and Numerical Analysis*, 43(4):743–755, jul 2009.

[126] Arthur F. Voter, Francesco Montalenti, and Timothy C. Germann. Extending the Time Scale in Atomistic Simulation of Materials. *Annual Review of Materials Research*, 32(1):321–346, aug 2002.

[127] S. Doerr and G. De Fabritiis. On-the-Fly Learning and Sampling of Ligand Binding by High-Throughput Molecular Simulations. *Journal of Chemical Theory and Computation*, 10(5):2064–2069, may 2014.

[128] A Amadei, A B Linssen, and H J Berendsen. Essential dynamics of proteins. *Proteins*, 17(4):412–25, dec 1993.

[129] W Wriggers and K Schulten. Protein domain movements: detection of rigid domains and visualization of hinges in comparisons of atomic coordinates. *Proteins*, 29(1):1–14, sep 1997.

[130] S Hayward and H J Berendsen. Systematic analysis of domain motions in proteins from conformational change: new results on citrate synthase and T4 lysozyme. *Proteins*, 30(2):144–54, feb 1998.

[131] H Berendsen. Collective protein dynamics in relation to function. *Current Opinion in Structural Biology*, 10(2):165–169, apr 2000.

[132] Vadim Alexandrov, Ursula Lehnert, Nathaniel Echols, Duncan Milburn, Donald Engelman, and Mark Gerstein. Normal modes for predicting protein motions: a comprehensive database assessment and associated Web tool. *Protein science : a publication of the Protein Society*, 14(3):633–43, mar 2005.

[133] E. Eyal, L.-W. Yang, and I. Bahar. Anisotropic network model: systematic evaluation and a new web interface. *Bioinformatics*, 22(21):2619–2627, nov 2006.

[134] Erik Lindahl, Cyril Azuara, Patrice Koehl, and Marc Delarue. NOMAD-Ref: visualization, deformation and refinement of macromolecular structures based on all-atom normal mode analysis. *Nucleic acids research*, 34(Web Server issue):W52–6, jul 2006.

[135] Karsten Suhre and Yves-Henri Sanejouand. ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic acids research*, 32(Web Server issue):W610–4, jul 2004.

[136] Yunfen He, J.-Y. Chen, J.R. Knab, Wenjun Zheng, and A.G. Markelz. Evidence of Protein Collective Motions on the Picosecond Timescale. *Biophysical Journal*, 100(4):1058–1065, feb 2011.

[137] B Brooks and M Karplus. Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme. *Proceedings of the National Academy of Sciences of the United States of America*, 82(15):4995–9, aug 1985.

[138] Liliane Mouawad and David Perahia. Motions in Hemoglobin Studied by Normal Mode Analysis and Energy Minimization: Evidence for the Existence of Tertiary T-like, Quaternary R-like Intermediate Structures. *Journal of Molecular Biology*, 258(2):393–410, 1996.

[139] Chunyan Xu, Dror Tobi, and I Bahar. Allosteric changes in protein structure computed by a simple mechanical model: hemoglobin TR2 transition. *Journal of molecular biology*, 333(1):153–68, oct 2003.

[140] Zheng Yang, Peter Májek, Ivet Bahar, PB Sigler, ZH Xu, HS Rye, SG Burston, WA Fenton, HR Saibil, NA Ranson, D Thirumalai, GH Lorimer, A Horovitz, KR Willison, W Zheng, BR Brooks, D Thirumalai, W Zheng, BR Brooks, D Thirumalai, I Kass, A Horovitz, Y Chen, K Reilly, Y Chang, BL de Groot, G Vriend, HJC Berendsen, J Ma, M Karplus, J Ma, PB Sigler, Z Xu, M Karplus, BT Sewell, RB Best, S Chen, AM Roseman, GW Farr, F Shewmaker, MJ Kerner, M Hayer-Hartl, G Klein, C Georgopoulos, G Stan, D Thirumalai, GH Lorimer, BR Brooks, NA Ranson, DK Clare, GW Farr, D Houldershaw, AL Horwich, C Chennubhotla, I Bahar, O Keskin, I Bahar, D Flatow, DG Covell, RL Jernigan, C Hyeon, GH Lorimer, D Thirumalai, A Horovitz, A Amir, O Danziger, G Kafri, Z Xu, AL Horwich, PB Sigler, PG Bolhuis, D Chandler, C Dellago, PL Geissler, R El-

ber, R Elber, M Karplus, R Czerminski, R Elber, R Olender, R Elber, S Huo, JE Straub, H Yang, H Wu, D Li, L Han, S Huo, H Jónsson, G Mills, KW Jacobsen, R Elber, D Shalloway, R Elber, A Ghosh, A Cárdenas, J Franklin, P Koehl, S Doniach, M Delarue, J Schlitter, M Engels, P Krueger, E Jacoby, A Wollmer, J Ma, M Karplus, J Ma, TC Flynn, Q Cui, AGW Leslie, JE Walker, TC Flynn, L Swint-Kruse, Y Kong, C Booth, KS Matthews, J Zhang, C Li, K Chen, W Zhu, X Shen, C Dellago, PG Bolhuis, D Chandler, DM Zuckerman, TB Woolf, A van der Vaart, M Karplus, D Branduardi, FL Gervasio, M Parrinello, I Bahar, AJ Rader, Q Cui, I Bahar, H Yu, L Ma, Y Yang, Q Cui, C Chennubhotla, I Bahar, I Bahar, C Chennubhotla, D Tobi, MK Kim, RL Jernigan, GS Chirikjian, O Miyashita, JN Onuchic, PG Wolynes, PC Whitford, O Miyashita, Y Levy, JN Onuchic, P Maragakis, M Karplus, J-W Chu, GA Voth, P Doruker, AR Atilgan, I Bahar, AR Atilgan, SR Durell, RL Jernigan, MC Demirel, O Keskin, J Ma, F Tama, CL Brooks, E Eyal, L-W Yang, I Bahar, C Xu, D Tobi, I Bahar, L Mouawad, D Perahia, P Petrone, VS Pande, L Yang, G Song, RL Jernigan, S Nicolay, YH Sanejouand, M Rueda, P Chacón, M Orozco, DG Teotico, ML Frazier, F Ding, NV Dokholyan, BRS Temple, P Májek, R Elber, H Weinstein, AJ Rader, DH Vlad, I Bahar, F Tama, CLI Brooks, C Chennubhotla, Z Yang, I Bahar, O Yifrach, A Horovitz, NA Ranson, GW Farr, AM Roseman, B Gowen, WA Fenton, O Yifrach, A Horovitz, J Monod, J Wyman, JP Changeux, A Horovitz, O Yifrach, F Glaser, T Pupko, I Paz, RE Bell, D Bechor-Shental, M Landau, I Mayrose, Y Rosenberg, F Glaser, E Martz, WA Fenton, Y Kashi, K Furtak, AL Norwich, HR Saibil, A van der Vaart, J Ma, M Karplus, Z Sun, DJ Scott, PA Lund, D Chandler, RB Best, G Hummer, OF Lange, N-A Lakomek, C Fares, GF Schroder, KFA Walter, HM Berman, J Westbrook, Z Feng, G Gilliland, TN Bhat, DK Clare, PJ Bakkes, H van Heerikhuizen, SM van der Vies, HR Saibil, and W Kabsch. Allosteric Transitions of Supramolecular Systems Explored by Network Models: Application to Chaperonin GroEL. *PLoS Computational Biology*, 5(4):e1000360, apr 2009.

[141] W. Zheng, B. R. Brooks, and D. Thirumalai. Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. *Proceedings of the National Academy of Sciences*, 103(20):7664–7669, may 2006.

[142] Nina M Goodey and Stephen J Benkovic. Allosteric regulation and catalysis emerge via a common route. *Nature Chemical Biology*, 4(8):474–482, aug 2008.

[143] B. F. Volkman. Two-State Allosteric Behavior in a Single-Domain Signaling Protein. *Science*, 291(5512):2429–2433, mar 2001.

[144] Rieko Ishima, Darón I Freedberg, Yun-Xing Wang, John M Louis, and Dennis A Torchia. Flap opening and dimer-interface flexibility in the free and inhibitor-bound HIV protease, and their implications for function. *Structure*, 7(9):1047–S12, 1999.

[145] Frans A.A. Mulder, Anthony Mittermaier, Bin Hon, Frederick W. Dahlquist, and Lewis E. Kay. Studying excited states of proteins by NMR spectroscopy. *Nature Structural Biology*, 8(11):932–935, nov 2001.

[146] Wayne L. Hubbell, David S. Cafiso, and Christian Altenbach. Identifying conformational changes with site-directed spin labeling. *Nature Structural Biology*, 7(9):735–739, sep 2000.

[147] David L. Farrens, Christian Altenbach, Ke Yang, Wayne L. Hubbell, and H. Gobind Khorana. Requirement of Rigid-Body Motion of Transmembrane Helices for Light Activation of Rhodopsin. *Science*, 274(5288), 1996.

[148] Yi-Shiuan Liu, Pornthep Sompornpisut, and Eduardo Perozo. Structure of the KcsA channel intracellular gate in the open state. *Nature Structural Biology*, 8(10):883–887, oct 2001.

[149] Robert J. Kadner, David S. Cafiso, Helen J. Merianos, Nathalie Cadieux, and Cindy H. Lin. Substrate-induced exposure of an energy-coupling motif of a membrane transporter. *Nature Structural Biology*, 7(3):205–209, mar 2000.

[150] Jianpeng Ma. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure (London, England : 1993)*, 13(3):373–80, mar 2005.

[151] Zhiyong Zhang, Jim Pfaendtner, Andrea Grafmüller, and Gregory A Voth. Defining

coarse-grained representations of large biomolecules and biomolecular complexes from elastic network models. *Biophysical journal*, 97(8):2327–37, oct 2009.
[152] Enrico Ravera, Loïc Salmon, Marco Fragai, Giacomo Parigi, Hashim Al-Hashimi, and Claudio Luchinat. Insights into domain-domain motions in proteins and RNA from solution NMR. *Accounts of chemical research*, 47(10):3118–26, oct 2014.
[153] Santiago Esteban-Martín, Robert Bryn Fenwick, and Xavier Salvatella. Synergistic use of NMR and MD simulations to study the structural heterogeneity of proteins. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2(3):466–478, may 2012.
[154] Jatin M Vyas, Annemarthe G Van der Veen, and Hidde L Ploegh. The known unknowns of antigen processing and presentation. *Nature reviews. Immunology*, 8(8):607–18, aug 2008.
[155] Eric W. Hewitt. The MHC class I antigen presentation pathway: strategies for viral immune evasion. *Immunology*, 110(2):163–169, oct 2003.
[156] M van de Rijn, C Bernabeu, B Royer-Pokora, J Weiss, J G Seidman, J de Vries, H Spits, and C Terhorst. Recognition of HLA-A2 by cytotoxic T lymphocytes after DNA transfer into human and murine cells. *Science (New York, N.Y.)*, 226(4678):1083–5, nov 1984.
[157] Max H. Schreier, N. N. Iscove, R. Tees, L. Aarden, and H. Vonboehmer. Clones of Killer and Helper T Cells: Growth Requirements, Specificity and Retention of Function in Long-Term Culture. *Immunological Reviews*, 51(1):315–336, aug 1980.
[158] Rakina Yaneva, Sebastian Springer, and Martin Zacharias. Flexibility of the MHC class II peptide binding cleft in the bound, partially filled, and empty states: a molecular dynamics simulation study. *Biopolymers*, 91(1):14–27, jan 2009.
[159] S Sadegh-Nasseri and H M McConnell. A kinetic intermediate in the reaction of an antigenic peptide and I-Ek. *Nature*, 337(6204):274–6, jan 1989.
[160] Joshua D. Rabinowitz, Marija Vrljic, Peter M. Kasson, Michael N. Liang, Robert Busch, J.Jay Boniface, Mark M. Davis, and Harden M. McConnell. Formation of a Highly Peptide-Receptive State of Class II MHC. *Immunity*, 9(5):699–709, 1998.
[161] S K Natarajan, M Assadi, and S Sadegh-Nasseri. Stable peptide binding to MHC class II molecule is rapid and is determined by a receptive conformation shaped by prior association with low affinity peptides. *Journal of immunology (Baltimore, Md. : 1950)*, 162(7):4030–6, apr 1999.
[162] Ravi V. Joshi, Jennifer A. Zarutskie, and Lawrence J. Stern. A Three-Step Kinetic Mechanism for Peptide Binding to MHC Class II Proteins âĂă. *Biochemistry*, 39(13):3751–3762, apr 2000.
[163] Peter M Kasson, Joshua D Rabinowitz, Lutz Schmitt, Mark M Davis, and Harden M. McConnell. Kinetics of Peptide Binding to the Class II MHC Protein I-E k. *Biochemistry*, 39(5):1048–1058, 2000.
[164] J A Zarutskie, A K Sato, M M Rushe, I C Chan, A Lomakin, G B Benedek, and L J Stern. A conformational change in the human major histocompatibility complex protein HLA-DR1 induced by peptide binding. *Biochemistry*, 38(18):5878–87, may 1999.
[165] Gregory J Carven and Lawrence J Stern. Probing the Ligand-Induced Conformational Change in HLA-DR1 by Selective Chemical Modification and Mass Spectrometric Mapping âĂă. *Biochemistry*, 44(42):13625–37, oct 2005.
[166] Janice S Blum, Pamela A Wearsch, and Peter Cresswell. Pathways of antigen processing. *Annual review of immunology*, 31:443–73, jan 2013.
[167] Phillip Morris, Jeffrey Shaman, Michelle Attaya, Miguel Amaya, Steven Goodman, Carolyn Bergman, John J. Monaco, and Elizabeth Mellins. An essential role for HLAâĂŞDM in antigen presentation by class II major histocompatibility molecules. *Nature*, 368(6471):551–554, apr 1994.
[168] P A Roche, M S Marks, and P Cresswell. Formation of a nine-subunit complex by HLA class II glycoproteins and the invariant chain. *Nature*, 354(6352):392–4, dec 1991.
[169] Jacques Neefjes, Marlieke L M Jongsma, Petra Paul, and Oddmund Bakke. Towards

a systems understanding of MHC class I and MHC class II antigen presentation. *Nature reviews. Immunology*, 11(12):823–36, dec 2011.

[170] Achal Pashine, Robert Busch, Michael P. Belmares, Jason N. Munning, Robert C. Doebele, Megan Buckingham, Gary P. Nolan, and Elizabeth D. Mellins. Interaction of HLA-DR with an Acidic Face of HLA-DM Disrupts Sequence-Dependent Interactions with Peptides. *Immunity*, 19(2):183–192, 2003.

[171] Wouter Pos, Dhruv K Sethi, Melissa J Call, Monika-Sarah E D Schulze, Anne-Kathrin Anders, Jason Pyrdol, and Kai W Wucherpfennig. Crystal structure of the HLA-DM-HLA-DR1 complex defines mechanisms for rapid peptide selection. *Cell*, 151(7):1557–68, dec 2012.

[172] Anne-Kathrin Anders, Melissa J Call, Monika-Sarah E D Schulze, Kevin D Fowler, David A Schubert, Nilufer P Seth, Eric J Sundberg, and Kai W Wucherpfennig. HLA-DM captures partially empty HLA-DR molecules for catalyzed removal of peptide. *Nature immunology*, 12(1):54–61, jan 2011.

[173] H. R. Drew, R. M. Wing, T. Takano, C. Broka, S. Tanaka, K. Itakura, and R. E. Dickerson. Structure of a B-DNA dodecamer: conformation and dynamics. *Proceedings of the National Academy of Sciences*, 78(4):2179–2183, apr 1981.

[174] R E Dickerson and H R Drew. Structure of a B-DNA dodecamer. II. Influence of base sequence on helix structure. *Journal of molecular biology*, 149(4):761–86, jul 1981.

[175] A V Fratini, M L Kopka, H R Drew, and R E Dickerson. Reversible bending and helix geometry in a B-DNA dodecamer: CGCGAATTBrCGCG. *The Journal of biological chemistry*, 257(24):14686–707, dec 1982.

[176] H R Drew and R E Dickerson. Structure of a B-DNA dodecamer. III. Geometry of hydration. *Journal of molecular biology*, 151(3):535–56, sep 1981.

[177] R. E. Dickerson, D. S. Goodsell, and S. Neidle. ...the tyranny of the lattice... *Proceedings of the National Academy of Sciences*, 91(9):3579–3583, apr 1994.

[178] B. Jayaram, D. Sprous, M. A. Young, , and D. L. Beveridge. Free Energy Analysis of the Conformational Preferences of A and B Forms of DNA in Solution. 1998.

[179] Jayashree Srinivasan, Thomas E. Cheatham, Piotr Cieplak, Peter A. Kollman, and David A. Case. Continuum Solvent Studies of the Stability of DNA, RNA, and PhosphoramidateâĹŠDNA Helices. 1998.

[180] Jonathan J Lee, George F Murphy, and Christine G Lian. Melanoma epigenetics: novel mechanisms, markers, and medicines. *Laboratory Investigation*, 94(8):822–838, aug 2014.

[181] Sylvain Guibert and Michael Weber. Functions of DNA methylation and hydroxymethylation in mammalian development. *Current topics in developmental biology*, 104:47–83, 2013.

[182] Hao Wu and Yi Zhang. Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation. *Genes & development*, 25(23):2436–52, dec 2011.

[183] N W Penn, R Suwalski, C O'Riley, K Bojanowski, and R Yura. The presence of 5-hydroxymethylcytosine in animal deoxyribonucleic acid. *The Biochemical journal*, 126(4):781–90, feb 1972.

[184] Skirmantas Kriaucionis and Nathaniel Heintz. The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain. *Science*, 324(5929), 2009.

[185] Minjia Tan, Hao Luo, Sangkyu Lee, Fulai Jin, Jeong Soo Yang, Emilie Montellier, Thierry Buchou, Zhongyi Cheng, Sophie Rousseaux, Nisha Rajagopal, Zhike Lu, Zhen Ye, Qin Zhu, Joanna Wysocka, Yang Ye, Saadi Khochbin, Bing Ren, and Yingming Zhao. Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell*, 146(6):1016–28, sep 2011.

[186] Eric S Lander. Initial impact of the sequencing of the human genome. *Nature*, 470(7333):187–97, feb 2011.

[187] Fabrício F. Costa. Non-coding RNAs: New players in eukaryotic biology. *Gene*, 357(2):83–94, 2005.

[188] Orly Wapinski, Howard Y Chang, F.H. Crick, et Al., C. Yanofsky, M. Guttman,

et Al., M. Guttman, et Al., A.C. Marques, C.P. Ponting, C.P. Ponting, et Al., J. Zhao, et Al., J.S. Mattick, M.E. Dinger, et Al., T.R. Mercer, et Al., J.S. Mattick, J.S. Mattick, I.V. Makunin, I.A. Qureshi, et Al., J.E. Wilusz, et Al., E. Pasmant, et Al., K.L. Yap, O. Harismendy, et Al., C.E. Burd, et Al., M.C. Tsai, et Al., R.A. Gupta, et Al., P. Ji, et Al., D. Bernard, et Al., V. Tripathi, et Al., J.C. Long, J.F. Caceres, M.A. Faghihi, et Al., T. Kino, et Al., C.M. Smith, J.A. Steitz, J.C. Webster, et Al., M. Huarte, et Al., M. Mourtada-Maarabouni, et Al., M. Halvorsen, et Al., G.A. Calin, et Al., J.K. Millar, et Al., R.S. Devon, et Al., J. Ekelund, et Al., M. Mutsuddi, I. Rebay, M.L. Moseley, et Al., R.S. Daughters, et Al., L.J. Scott, et Al., H.M. Broadbent, et Al., S.E. Wojcik, et Al., C. Lagier-Tourenne, D.W. Cleveland, T.J. Kwiatkowski, et Al., C. Vance, et Al., X. Wang, et Al., C. Bagni, W.T. Greenough, F. Zalfa, et Al., E.M. Johnson, et Al., T. Khanam, et Al., E. Mus, and et Al. Long noncoding RNAs and human disease. *Trends in cell biology*, 21(6):354–61, jun 2011.

[189] William a. Cantara, Pamela F. Crain, Jef Rozenski, James a. McCloskey, Kimberly a. Harris, Xiaonong Zhang, Franck a P Vendeix, Daniele Fabris, and Paul F. Agris. The RNA modification database, RNAMDB: 2011 update. *Nucleic Acids Research*, 39(SUPPL. 1):195–201, 2011.

[190] Yamei Niu, Xu Zhao, Yong-Sheng Wu, Ming-Ming Li, Xiu-Jie Wang, and Yun-Gui Yang. N6-methyl-adenosine (m6A) in RNA: An Old Modification with A Novel Epigenetic Function. *Genomics, Proteomics & Bioinformatics*, 11(1):8–17, 2013.

[191] Nian Liu, Qing Dai, Guanqun Zheng, Chuan He, Marc Parisien, and Tao Pan. N6-methyladenosine-dependent RNA structural switches regulate RNAâĂŞprotein interactions. *Nature*, 518(7540):560–564, 2015.

[192] Rui Wang, Zhipu Luo, Kaizhang He, Michael O Delaney, Doris Chen, and Jia Sheng. Base pairing and structural insights into the 5-formylcytosine in RNA duplex. *Nucleic acids research*, apr 2016.

[193] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087, dec 1953.

[194] Alberto Pérez, Iván Marchán, Daniel Svozil, Jiri Sponer, Thomas E Cheatham, Charles a Laughton, and Modesto Orozco. Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophysical journal*, 92(11):3817–3829, 2007.

[195] Junmei Wang, Piotr Cieplak, and Peter A. Kollman. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry*, 21(12):1049–1074, sep 2000.

[196] Wendy D. Cornell, Piotr Cieplak, Christopher I. Bayly, Ian R. Gould, Kenneth M. Merz, David M. Ferguson, David C. Spellmeyer, Thomas Fox, James W. Caldwell, and Peter A. Kollman. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society*, 118(9):2309–2309, jan 1996.

[197] Péter Várnai and Krystyna Zakrzewska. DNA and its counterions: a molecular dynamics study. *Nucleic acids research*, 32(14):4269–80, 2004.

[198] Simon C. Lovell, Ian W. Davis, W. Bryan Arendall, Paul I. W. de Bakker, J. Michael Word, Michael G. Prisant, Jane S. Richardson, and David C. Richardson. Structure validation by Cα geometry: Ï Ţ,ψ and Cβ deviation. *Proteins: Structure, Function, and Bioinformatics*, 50(3):437–450, jan 2003.

[199] H Lu and J Skolnick. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins*, 44(3):223–32, aug 2001.

[200] Jingtong Hou, Se-Ran Jun, Chao Zhang, and Sung-Hou Kim. Global mapping of the protein structure space and application in structure-based inference of protein function. *Proceedings of the National Academy of Sciences of the United States of America*, 102(10):3651–6, mar 2005.

[201] George Moraitakis, Andrew G Purkiss, and Julia M Goodfellow. Simulated dynam-

ics and biological macromolecules. *Reports on Progress in Physics*, 66(3):383–406, mar 2003.
[202] Wilfred F van Gunsteren, Jozica Dolenc, and Alan E Mark. Molecular simulation as an aid to experimentalists. *Current opinion in structural biology*, 18(2):149–53, apr 2008.
[203] Martin Karplus and Richard Lavery. Significance of Molecular Dynamics Simulations for Life Sciences. *Israel Journal of Chemistry*, 54(8-9):1042–1051, aug 2014.
[204] Bernhard Knapp, James Dunbar, and Charlotte M Deane. Large Scale Characterization of the LC13 TCR and HLA-B8 Structural Landscape in Reaction to 172 Altered Peptide Ligands: A Molecular Dynamics Simulation Study. *PLoS computational biology*, 10(8):e1003748, 2014.
[205] Kim L Morrison and Gregory A Weiss. Combinatorial alanine-scanning. *Current Opinion in Chemical Biology*, 5(3):302–307, jun 2001.
[206] M. A. Balsera, W. Wriggers, Y. Oono, and K. Schulten. Principal Component Analysis and Long Time Protein Dynamics. *The Journal of Physical Chemistry*, 100(7):2567–2572, jan 1996.
[207] A R Atilgan, S R Durell, R L Jernigan, M C Demirel, O Keskin, and I Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical journal*, 80(1):505–15, jan 2001.
[208] William Mattson and Betsy M. Rice. Near-neighbor calculations using a modified cell-linked list method. *Computer Physics Communications*, 119(2-3):135–148, jun 1999.
[209] Zhenhua Yao, Jian-Sheng Wang, Gui-Rong Liu, and Min Cheng. Improved neighbor list algorithm in molecular simulations using cell decomposition and data sorting method. *Computer Physics Communications*, 161(1-2):27–35, aug 2004.
[210] Lars Skjaerven, Siv M. Hollup, and Nathalie Reuter. Normal mode analysis for proteins. *Journal of Molecular Structure: THEOCHEM*, 898(1):42–48, 2009.
[211] Ivet Bahar, Ali Rana Atilgan, and Burak Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design*, 2(3):173–181, jun 1997.
[212] Konrad Hinsen. Analysis of domain motions by approximate normal mode calculations. *Proteins: Structure, Function, and Genetics*, 33(3):417–429, nov 1998.
[213] Michael Levitt, Christian Sander, and Peter S. Stern. Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme. *Journal of Molecular Biology*, 181(3):423–447, 1985.
[214] V. Zoete, O. Michielin, and M. Karplus. Relation between sequence and structure of HIV-1 protease inhibitor complexes: a model system for the analysis of protein flexibility11Edited by R. Huber. *Journal of Molecular Biology*, 315(1):21–52, jan 2002.
[215] Nathalie Reuter, Konrad Hinsen, and Jean-Jacques Lacapère. Transconformations of the SERCA1 Ca-ATPase: a normal mode study. *Biophysical journal*, 85(4):2186–97, oct 2003.
[216] Wenjun Zheng and Sebastian Doniach. A comparative study of motor-protein motions by using a simple elastic-network model. *Proceedings of the National Academy of Sciences of the United States of America*, 100(23):13253–8, nov 2003.
[217] Wenjun Zheng, Bernard R. Brooks, and D. Thirumalai. Allosteric Transitions in the Chaperonin GroEL are Captured by a Dominant Normal Mode that is Most Robust to Sequence Variations. *Biophysical Journal*, 93(7):2289–2299, oct 2007.
[218] Liliane Mouawad and David Perahia. Diagonalization in a mixed basis: A method to compute low-frequency normal modes for large macromolecules. *Biopolymers*, 33(4):599–611, apr 1993.
[219] D Perahia and L Mouawad. Computation of low-frequency normal modes in macromolecules: improvements to the method of diagonalization in a mixed basis and application to hemoglobin. *Computers & chemistry*, 19(3):241–6, sep 1995.
[220] Florence Tama, Florent Xavier Gadea, Osni Marques, and Yves-Henri Sanejouand. Building-block approach for determining low-frequency normal modes of macro-

molecules. *Proteins: Structure, Function, and Genetics*, 41(1):1–7, oct 2000.
[221] Xiang-Jun J. Lu and Wilma K. Olson. 3DNA: A software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Research*, 31(17):5108–5121, sep 2003.
[222] Marla S. Babcock, Edwin P.D. Pednault, and Wilma K. Olson. Nucleic Acid Structure Analysis: Mathematics for Local Cartesian and Helical Structure Parameters That Are Truly Comparable Between Structures. *Journal of Molecular Biology*, 237(1):125–156, 1994.
[223] R Lavery, K Zakrzewska, J S Sun, and S C Harvey. A comprehensive classification of nucleic acid structural families based on strand direction and base pairing. *Nucleic acids research*, 20(19):5011–6, oct 1992.
[224] Konstantin M. Kosikov, Andrey A. Gorin, Victor B. Zhurkin, and Wilma K. Olson. DNA stretching and compression: large-scale simulations of double helical structures. *Journal of Molecular Biology*, 289(5):1301–1326, 1999.
[225] A Sali and T L Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology*, 234(3):779–815, dec 1993.
[226] A Fiser, R K Do, and A Sali. Modeling of loops in protein structures. *Protein science : a publication of the Protein Society*, 9(9):1753–73, sep 2000.
[227] Chris Sander and Reinhard Schneider. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Structure, Function, and Genetics*, 9(1):56–68, jan 1991.
[228] Bruce I. Cohen, Scott R. Presnell, and Fred E. Cohen. Origins of structural diversity within sequentially identical hexapeptides. *Protein Science*, 2(12):2134–2145, dec 1993.
[229] Claire Marks, Jaroslaw Nowak, Stefan Klostermann, Guy Georges, James Dunbar, Jiye Shi, Sebastian Kelm, and Charlotte M. Deane. Sphinx: merging knowledge-based and ab initio approaches to improve protein loop prediction. *Bioinformatics*, 99(9):btw823, jan 2017.
[230] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins ^{âĂă}. *The Journal of Physical Chemistry B*, 102(18):3586–3616, apr 1998.
[231] Manfred J. Sippl. Recognition of errors in three-dimensional structures of proteins. *Proteins: Structure, Function, and Genetics*, 17(4):355–362, dec 1993.
[232] L J Stern, J H Brown, T S Jardetzky, J C Gorga, R G Urban, J L Strominger, and D C Wiley. Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature*, 368(6468):215–21, mar 1994.
[233] D Rognan, L Scapozza, G Folkers, and A Daser. Molecular dynamics simulation of MHC-peptide complexes as a tool for predicting potential T cell epitopes. *Biochemistry*, 33(38):11476–85, sep 1994.
[234] Bernhard Knapp, Ulrich Omasits, Barbara Bohle, Bernard Maillere, Christof Ebner, Wolfgang Schreiner, and Beatrice Jahn-Schmid. 3-Layer-based analysis of peptide-MHC interaction: in silico prediction, peptide binding affinity and T cell activation in a relevant allergen-specific model. *Molecular immunology*, 46(8-9):1839–44, may 2009.
[235] Y Altuvia, A Sette, J Sidney, S Southwood, and H Margalit. A structure-based algorithm to predict potential binding peptides to MHC molecules with hydrophobic binding pockets. *Human immunology*, 58(1):1–11, dec 1997.
[236] Arumugam Mohanapriya, Sajitha Lulu, Rajarathinam Kayathri, and Pandjassarame Kangueane. Class II HLA-peptide binding prediction using structural principles. *Human immunology*, 70(3):159–69, mar 2009.
[237] A Logean, A Sette, and D Rognan. Customized versus universal scoring functions: application to class I MHC-peptide binding free energy predictions. *Bioorganic &*

medicinal chemistry letters, 11(5):675–9, mar 2001.

[238] Mariyana Atanasova, Atanas Patronov, Ivan Dimitrov, Darren R Flower, and Irini Doytchinova. EpiDOCK: a molecular docking-based tool for MHC class II binding prediction. *Protein engineering, design & selection : PEDS*, 26(10):631–4, oct 2013.

[239] Atanas Patronov, Ivan Dimitrov, Darren R Flower, and Irini Doytchinova. Peptide binding prediction for the human class II MHC allele HLA-DP2: a molecular docking approach. *BMC structural biology*, 11(1):32, jan 2011.

[240] Didier Rognan. Molecular dynamics simulations: A tool for drug design. In *Three-Dimensional Quantitative Structure Activity Relationships -Volume 2*, pages 181–209. 1998.

[241] Irini Doytchinova, Peicho Petkov, Ivan Dimitrov, Mariyana Atanasova, and Darren R Flower. HLA-DP2 binding prediction by molecular dynamics simulations. *Protein science : a publication of the Protein Society*, 20(11):1918–28, dec 2011.

[242] Chen Yanover and Philip Bradley. Large-scale characterization of peptide-MHC binding landscapes with structural simulations. *Proceedings of the National Academy of Sciences of the United States of America*, 108(17):6981–6, apr 2011.

[243] Henry R Wilman, Jean-paul Ebejer, Jiye Shi, Charlotte M Deane, and Bernhard Knapp. Crowdsourcing Yields a New Standard for Kinks in Protein Helices. *Journal of Chemical Information and Modeling*, 54:2585–2593, 2014.

[244] Jeffrey Ishizuka, Kristie Grebe, Eugene Shenderov, Bjoern Peters, Qiongyu Chen, Yanchun Peng, Lili Wang, Tao Dong, Valerie Pasquetto, Carla Oseroff, John Sidney, Heather Hickman, Vincenzo Cerundolo, Alessandro Sette, Jack R Bennink, Andrew McMichael, and Jonathan W Yewdell. Quantitating T cell cross-reactivity for unrelated peptide antigens. *Journal of immunology (Baltimore, Md. : 1950)*, 183(7):4337–45, oct 2009.

[245] Georgii G Krivov, Maxim V Shapovalov, and Roland L Dunbrack. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, 77(4):778–95, dec 2009.

[246] Bernhard Knapp, Ulrich Omasits, and Wolfgang Schreiner. Side chain substitution benchmark for peptide/MHC interaction. *Protein science : a publication of the Protein Society*, 17(6):977–82, jun 2008.

[247] Daniele Narzi, Caroline M. Becker, Maria T. Fiorillo, Barbara Uchanska-Ziegler, Andreas Ziegler, and Rainer A. Böckmann. Dynamical Characterization of Two Differentially Disease Associated MHC Class I Proteins in Complex with Viral and Self-Peptides. *Journal of Molecular Biology*, 415(2):429–442, 2012.

[248] Birgit Hischenhuber, Hans Havlicek, Jelena Todoric, Sonja Höllrigl-Binder, Wolfgang Schreiner, and Bernhard Knapp. Differential geometric analysis of alterations in MH α-helices. *Journal of computational chemistry*, 34(21):1862–79, aug 2013.

[249] Bernd Rupp, Sebastian Günther, Talat Makhmoor, Andreas Schlundt, Katharina Dickhaut, Shashank Gupta, Iqbal Choudhary, Karl-Heinz Wiesmüller, Günther Jung, Christian Freund, Kirsten Falk, Olaf Rötzschke, and Ronald Kühne. Characterization of structural features controlling the receptiveness of empty class II MHC molecules. *PloS one*, 6(4):e18662, jan 2011.

[250] Qing Zhang, Peng Wang, Yohan Kim, Pernille Haste-Andersen, John Beaver, Philip E Bourne, Huynh-Hoa Bui, Soren Buus, Sune Frankild, Jason Greenbaum, Ole Lund, Claus Lundegaard, Morten Nielsen, Julia Ponomarenko, Alessandro Sette, Zhanyang Zhu, and Bjoern Peters. Immune epitope database analysis resource (IEDB-AR). *Nucleic acids research*, 36(Web Server issue):W513–8, jul 2008.

[251] Lianming Zhang, Keiko Udaka, Hiroshi Mamitsuka, and Shanfeng Zhu. Toward more accurate pan-specific MHC-peptide binding prediction: a review of current methods and tools. *Briefings in bioinformatics*, 13(3):350–64, may 2012.

[252] David W. Wright, Benjamin A. Hall, Owain A. Kenway, Shantenu Jha, and Peter V. Coveney. Computing Clinically Relevant Binding Free Energies of HIV-1 Protease Inhibitors. *Journal of Chemical Theory and Computation*, 10(3):1228–1241, mar 2014.

[253] Shunzhou Wan, Bernhard Knapp, David W. Wright, Charlotte M. Deane, and Peter V. Coveney. Rapid, Precise, and Reproducible Prediction of Peptide-MHC

Binding Affinities from Molecular Dynamics That Correlate Well with Experiment. *Journal of Chemical Theory and Computation*, 11(7):3346–3356, jul 2015.

[254] John J. O'Shea and William E. Paul. Mechanisms Underlying Lineage Commitment and Plasticity of Helper CD4+ T Cells. *Science*, 327(5969), 2010.

[255] Hongkai Zhang, Ian A Wilson, and Richard A Lerner. Selection of antibodies that regulate phenotype from intracellular combinatorial antibody libraries. *Proceedings of the National Academy of Sciences of the United States of America*, 109(39):15728–33, sep 2012.

[256] Johanna L. Barclay, Linda M. Kerr, Leela Arthur, Jennifer E. Rowland, Caroline N. Nelson, Mayumi Ishikawa, Elisabetta M. D'Aniello, Mary White, Peter G. Noakes, and Michael J. Waters. In Vivo Targeting of the Growth Hormone Receptor (GHR) Box1 Sequence Demonstrates that the GHR Does Not Signal Exclusively through JAK2. *Molecular Endocrinology*, 24(1):204–217, jan 2010.

[257] O Livnah, E A Stura, D L Johnson, S A Middleton, L S Mulcahy, N C Wrighton, W J Dower, L K Jolliffe, and I A Wilson. Functional mimicry of a protein hormone by a peptide agonist: the EPO receptor complex at 2.8 A. *Science (New York, N.Y.)*, 273(5274):464–71, jul 1996.

[258] O Livnah, D L Johnson, E A Stura, F X Farrell, F P Barbone, Y You, K D Liu, M A Goldsmith, W He, C D Krause, S Pestka, L K Jolliffe, and I A Wilson. An antagonist peptide-EPO receptor complex suggests that receptor dimerization is not sufficient for activation. *Nature structural biology*, 5(11):993–1004, nov 1998.

[259] Charles L. Brooks Iii Michael Feig, John Karanicolas. MMTSB Tool Set: Enhanced Sampling and Multiscale Modeling Methods for Applications in Structural Biology.

[260] Marcus D. Ballinger and James A. Wells. Will any dimer do? *Nature Structural Biology*, 5(11):938–940, nov 1998.

[261] R S Syed, S W Reid, C Li, J C Cheetham, K H Aoki, B Liu, H Zhan, T D Osslund, A J Chirino, J Zhang, J Finer-Moore, S Elliott, K Sitney, B A Katz, D J Matthews, J J Wendoloski, J Egrie, and R M Stroud. Efficiency of signalling through cytokine receptors depends critically on receptor orientation. *Nature*, 395(6701):511–6, oct 1998.

[262] J A Wells and A M de Vos. Structure and function of human growth hormone: implications for the hematopoietins. *Annual review of biophysics and biomolecular structure*, 22:329–51, 1993.

[263] Masayuki Kai, Kazuhiro Motoki, Hideaki Yoshida, Chie Emuta, Yukiko Chisaka, Kumi Tsuruhata, Chisato Endo, Mari Muto, Munetake Shimabe, Uichi Nishiyama, Tetsuya Hagiwara, Atsushi Matsumoto, Hiroshi Miyazaki, and Shiro Kataoka. Switching constant domains enhances agonist activities of antibodies to a thrombopoietin receptor. *Nature Biotechnology*, 26(2):209–211, feb 2008.

[264] Jamie B Spangler, Ignacio Moraga, Juan L Mendoza, and K Christopher Garcia. Insights into cytokine-receptor interactions from cytokine engineering. *Annual review of immunology*, 33:139–67, 2015.

[265] Konrad Krawczyk, Samuel Demharter, Bernhard Knapp, Deane, Charlotte M., and Peter Minary. Structural effects of epigenetic marks on DNA structure in silico. *Nucleic acids research (submitted)*, 2017.

[266] Carles Acosta-Silva, Vicenç Branchadell, Joan Bertran, and Antoni Oliva. Mutual relationship between stacking and hydrogen bonding in DNA. Theoretical study of guanine-cytosine, guanine-5-methylcytosine, and their dimers. *Journal of Physical Chemistry B*, 114(31):10217–10227, aug 2010.

[267] Tahir I. Yusufaly, Yun Li, and Wilma K. Olson. 5-Methylation of cytosine in CG:CG base-pair steps: A physicochemical mechanism for the epigenetic control of DNA nanomechanics. *Journal of Physical Chemistry B*, 117(51):16436–16442, dec 2013.

[268] Alexandra Teresa Pires Carvalho, Maria Leonor Gouveia, Charan Raju Kanna, Sebastian K T S Wärmländer, Jamie Platts, and Shina Caroline Lynn Kamerlin. Theoretical modelling of epigenetically modified DNA sequences. *F1000Research*, 4:52, jan 2015.

[269] Philip M D Severin, Xueqing Zou, Hermann E. Gaub, and Klaus Schulten. Cytosine methylation alters DNA mechanical properties. *Nucleic Acids Research*,

39(20):8740–8751, nov 2011.

[270] Alexandra T P Carvalho, Leonor Gouveia, Charan Raju Kanna, Sebastian K T S Wärmländer, Jamie A. Platts, and Shina Caroline Lynn Kamerlin. Understanding the structural and dynamic consequences of DNA epigenetic modifications: Computational insights into cytosine methylation and hydroxymethylation. *Epigenetics*, 9(12):1604–1612, jan 2014.

[271] Lukas Lercher, Michael a McDonough, Afaf H El-Sagheer, Armin Thalhammer, Skirmantas Kriaucionis, Tom Brown, and Christopher J Schofield. Structural insights into how 5-hydroxymethylation influences transcription factor binding. *Chemical communications (Cambridge, England)*, 50(15):1794–6, 2014.

[272] Armin Thalhammer, Anders S Hansen, Afaf H El-Sagheer, Tom Brown, and Christopher J Schofield. Hydroxylation of methylated CpG dinucleotides reverses stabilisation of DNA duplexes by cytosine 5-methylation. *Chemical communications (Cambridge, England)*, 47(18):5325–7, may 2011.

[273] Warren Lyford DeLano. The PyMOL Molecular Graphics System, Version 1.7.4 Schrödinger, LLC., 2014.

[274] Thomas J. Macke and David A. Case. Modeling Unusual Nucleic Acid Structures. In *Molecular Modeling of Nucleic Acids*, pages 379–393. nov 1997.

[275] William Humphrey, Dalke Andrew, and Klaus Schulten. {VMD} – {V}isual {M}olecular {D}ynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[276] Daniel Renciuk, Olivier Blacque, Michaela Vorlickova, and Bernhard Spingler. Crystal structures of B-DNA dodecamer containing the epigenetic modifications 5-hydroxymethylcytosine or 5-methylcytosine. *Nucleic Acids Research*, 41(21):9891–9900, 2013.

[277] Marta W. Szulik, Pradeep S. Pallan, Boguslaw Nocek, Markus Voehler, Surajit Banerjee, Sonja Brooks, Andrzej Joachimiak, Martin Egli, Brandt F. Eichman, and Michael P. Stone. Differential Stabilities and Sequence-Dependent Base Pair Opening Dynamics of WatsonâĂŞCrick Base Pairs with 5-Hydroxymethylcytosine, 5-Formylcytosine, or 5-Carboxylcytosine. *Biochemistry*, 54(5):1294–305, mar 2015.

[278] Achim Breiling and Frank Lyko. Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. *Epigenetics & Chromatin*, 8(1):24, jul 2015.

[279] Pijus Brazauskas and Skirmantas Kriaucionis. DNA modifications: Another stable base in DNA. *Nature chemistry*, 6(12):1031–3, dec 2014.

[280] Hideharu Hashimoto, June E Pais, Nan Dai, Ivan R Corrêa, Xing Zhang, Yu Zheng, and Xiaodong Cheng. Structure of Naegleria Tet-like dioxygenase (NgTet1) in complexes with a reaction intermediate 5-hydroxymethylcytosine DNA. *Nucleic acids research*, 43(22):10713–21, dec 2015.

[281] P Cresswell. Assembly, transport, and function of MHC class II molecules. *Annual review of immunology*, 12:259–93, jan 1994.

[282] E. Yvonne Jones, Lars Fugger, Jack L. Strominger, and Christian Siebold. MHC class II proteins and disease: a structural perspective. *Nature Reviews Immunology*, 6(4):271–282, apr 2006.

[283] Naveen Michaud-Agrawal, Elizabeth J Denning, Thomas B Woolf, and Oliver Beckstein. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *Journal of computational chemistry*, apr 2011.

[284] Gregory J Carven, Sriram Chitta, Ivan Hilgert, Mia M Rushe, Rick F Baggio, Michelle Palmer, Jaime E Arenas, Jack L Strominger, Vaclav Horejsi, Laura Santambrogio, and Lawrence J Stern. Monoclonal antibodies specific for the empty conformation of HLA-DR1 reveal aspects of the conformational change associated with peptide binding. *The Journal of biological chemistry*, 279(16):16561–70, apr 2004.

[285] Corrie A Painter, Anthony Cruz, Gustavo E López, Lawrence J Stern, and Zarixia Zavala-Ruiz. Model for the Peptide-Free Conformation of Class II MHC Proteins. *PLoS ONE*, 3(6):10, 2008.

[286] S Sadegh-Nasseri and RN N Germain. A role for peptide in determining MHC class II structure. *Nature*, 353(6340):167–70, sep 1991.

[287] Martin Zacharias and Sebastian Springer. Conformational flexibility of the MHC class I alpha1-alpha2 domain in peptide bound and free states: a molecular dynamics simulation study. *Biophysical journal*, 87(4):2203–14, oct 2004.

[288] Harald Kropshofer, Sven O Arndt, Gerhard Moldenhauer, Günter J Hämmerling, and Anne B Vogt. HLA-DM Acts as a Molecular Chaperone and Rescues Empty HLA-DR Molecules at Lysosomal pH. *Immunity*, 6(3):293–302, 1997.

[289] Robert Busch, Cornelia H Rinderknecht, Sujin Roh, Andrew W Lee, James J Harding, Timo Burster, Tara M C Hornell, and Elizabeth D Mellins. Achieving stability through editing and chaperoning: regulation of MHC class II peptide binding and expression. *Immunological reviews*, 207:242–60, oct 2005.

[290] Gijsbert M Grotenbreg, Melissa J Nicholson, Kevin D Fowler, Kathrin Wilbuer, Leah Octavio, Maxine Yang, Arup K Chakraborty, Hidde L Ploegh, and Kai W Wucherpfennig. Empty class II major histocompatibility complex created by peptide photolysis establishes the role of DM in peptide association. *The Journal of biological chemistry*, 282(29):21425–36, jul 2007.

[291] R C Doebele, R Busch, H M Scott, a Pashine, and E D Mellins. Determination of the HLA-DM interaction site on HLA-DR molecules. *Immunity*, 13(4):517–27, oct 2000.

[292] Peter Minary. Methodologies for Optimization and SAmpling In Computational Studies, (MOSAICS) version 3.9., 2007.

[293] J D Hunter. Matplotlib: A 2D graphics environment. *Computing In Science and Engineering*, 9(3):90–95, 2007.

[294] Wes McKinney. pandas: a Foundational Python Library for Data Analysis and Statistics. 2011.

[295] J H Brown, T Jardetzky, M A Saper, B Samraoui, P J Bjorkman, and D C Wiley. A hypothetical model of the foreign antigen binding site of class II histocompatibility molecules. *Nature*, 332(6167):845–50, apr 1988.

[296] R Rohs, C Etchebest, and R Lavery. Unraveling proteins: a molecular mechanics study. *Biophysical journal*, 76(5):2760–2768, 1999.

[297] MG Goll and TH Bestor. Eukaryotic cytosine methyltransferases. *Annu. Rev. Biochem.*, (74):481–514, 2005.

[298] Shinsuke Ito, Li Shen, Qing Dai, Susan C Wu, Leonard B Collins, James A Swenberg, Chuan He, and Yi Zhang. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science (New York, N.Y.)*, 333(6047):1300–3, sep 2011.

[299] Martin Münzel, Daniel Globisch, and Thomas Carell. 5-Hydroxymethylcytosine, the Sixth Base of the Genome. *Angewandte Chemie International Edition*, 50(29):6460–6468, jul 2011.

[300] Meni Wanunu, Devora Cohen-Karni, Robert R Johnson, Lauren Fields, Jack Benner, Neil Peterman, Yu Zheng, Michael L Klein, and Marija Drndic. Discrimination of methylcytosine from hydroxymethylcytosine in DNA molecules. *Journal of the American Chemical Society*, 133(3):486–92, jan 2011.