Interaction of the maturation protein of the bacteriophage MS2 and the F pilin of *Escherichia coli*

Timothy J. Spankie, MSci.

Thesis submitted to The University of Nottingham for partial fulfilment of the degree of Master of Research

September 2019

Abstract

Antimicrobial resistance is a ever increasing global concern, with the lack of new antibiotics and misuse of current available antibiotics exacerbating the issue. An alternative approach is to use bacteriophages to infect bacteria and select against the resistant genes, allowing for existing stock to be used again. The maturation protein at the centre of the bacteriophage MS2 infects F plasmid containing bacteria including the entire *Escherichia coli* (*E. coli*). The maturation protein in the centre of MS2 binds to the F pilus. This project investigates this interaction that delivers the viral DNA into bacteria. The model was ratified and extended through the use of protein-protein docking. A single subunit of the F pilus was compared to a trimer of subunits, and the trimer was found to be a more accurate model, as it included the steric hindrance of other monomer strands when approaching pilin attached to the maturation protein. Contacts between the maturation protein and F pilin were assesed during repeated Molecular Dynamics simulations, where the trimer model was shown to interact less, but within the same regions. Alanine scanning was performed before single-point mutations were made to explore making the maturation protein more versatile to other pili, without reducing the strength of binding to F pilin. After mutations, certain residues were found to be required for the maturation protein and F pilin to interact successfully. Other mutations had no effect on the interaction and some residues had a stronger interaction when mutated than before.

Acknowledgements

I would like to acknowledge the support of my supervisor Professor Jonathan Hirst, and the help of his research group; especially Ellen Guest for the introduction to CHARMM and NAMD, Steven Oatley for his help with python programming, and Alexe Haywood whose previous work led to the creation of this project. I would also like to thank the wider Computational And Theoretical Chemistry at Nottingham (CATC@N) group for always being there to provide support and stimulating conversation when needed.

The use of the University of Nottingham High Performance Computing (HPC) service Augusta, and the Tier 2 HPC Midlands Plus service Athena was appreciated for facilitating simulations. Without these services the same level of results would have been unobtainable.

Finally my thanks to my parents for their constant reassurance, proofreading and support.

Contents

List of Figures

List of Tables

Abbreviations

E. coli	*Escherichia coli*
AIRs	Ambiguous Interaction Restraints
CATC@N	Computational And Theoretical Chemistry at Nottingham
CHARMM	Chemistry at HARvard Macromolecular Mechanics
cryoEM	electron-counting cryo Electron Microscopy
HADDOCK	High Ambiguity Driven protein-protein DOCKing
HPC	High Performance Computing
HS	HADDOCK Score
MD	Molecular Dynamics
NAMD	NAnoscale Molecular Dynamics
NPT	Isobaric-Isothermal ensemble
NVT	Canonical ensemble
PBCs	Periodic boundary conditions
PDB	Protein Data Bank
PME	Particle Mesh Ewald
PPIs	Protein-protein Interactions
REU	Rosetta Energy Units
RMSD	Root Mean Square Deviation
TIP3P	Transferable Intermolecular Potential 3P
VMD	Visual Molecular Dynamics

1. Introduction

1.1 Antimicrobial resistance

Antimicrobial resistance, the ability for a microbe to survive when exposed to antimicrobials, is a global issue which threatens the treatment of infectious diseases in humans, animals and agriculture.[1] The natural phenomenon, due to evolution of conferring resistance, has been accelerated by inappropriate use of antimicrobials. This includes prescription of antimicrobials for viral infection which is ineffective, courses of antibiotics not being completed, allowing more resistant strains to survive, and the overuse of antimicrobials in farming for short-term benefits. Microbial infections have been treated and documented since ancient eras.[2] Modern antibiotics have been used since the discovery of penicillin in 1928 by Sir Alexander Fleming.[2,3] Recent developments in antibiotics have stalled due to challenging regulatory requirements and lowered economic incentives.[3,4] Even for the few antibiotics released recently, resistance was seen almost immediately.[5] Management of antimicrobial resistance is expensive, such as reversing the selective advantage of resistance that makes microbials susceptible to treatment again.[6]

Despite the simplicity of bacteria the cell structure is well developed, having unique biological structures and their well known pathogenicity.[7] Antimicrobial resistance can be transferred between bacteria through a process of genetic information sharing called conjugation.[8] A pilus, a hair-like appendage depolymerises from a donor cell, and plasmid DNA is transferred through the pilus tube to the recipient cell, as bacterial DNA resides inside the bacterial cytoplasm.[7] This is shown in Figure 1.1.

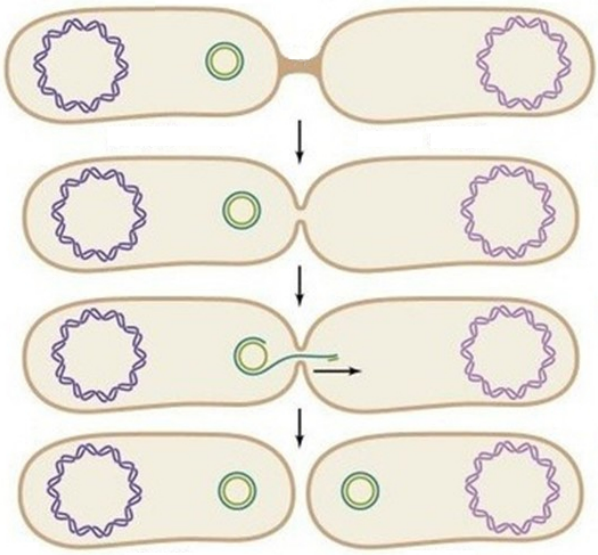


Figure 1.1: Conjugation of bacteria.

The pilus is essential for the passing of genetic information, and is a suitable target for bacteriophages.

1.2 Bacteriophages

A bacteriophage is a virus that infects a specific bacteria, whilst not affecting others such as safe microflora of a human. Bacteriophages consist of a protein shell that encapsulates either a DNA or RNA genome, dependent on the complexity of the bacteriophage. The bacterial cellular machinery is attacked, and prevented from producing bacterial components. Bacteriophages then use the resources of infected bacteria to replicate and so their quantity is dependent on the amount of bacteria.[9]

In this project the relationship between the bacteriophage MS2 and the F sex pilus of the *Escherichia coli* (*E. coli*) is examined, as shown in Figure 1.2. MS2 infects *E. coli* via the lytic cycle, beginning with attachment to the sex pilus of the bacterial host.[10] As the pilus retracts, the virion is forced from the capsid shell and the ssrNA genome is delivered into the host.[11] The bacteriophage consists of 89 coat protein dimers arranged in a T = 3 icosahedral lattice,[12] with a single copy of the maturation protein which binds to the pilus and protects the ssRNA genome inside.

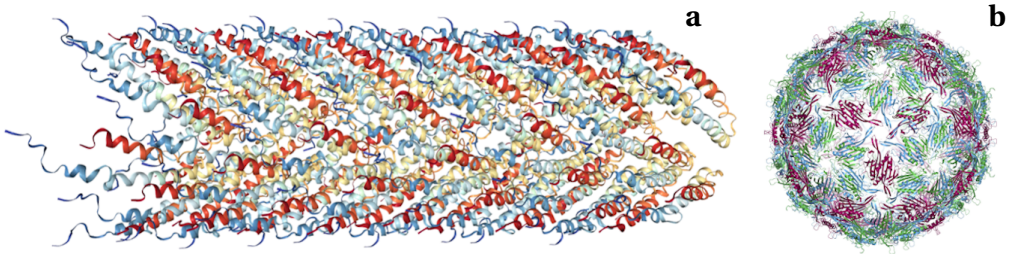


Figure 1.2: **a.** Structure of the bacterial sex F pilus (PDB ID 5LER).[10] **b.** Bacteriophage MS2, showing the coat protein arranged into an icosahedral shell with triangulation number T = 3 (PDB ID 2MS2).[13]

Bacteriophages multiply in the host bacterium, causing cell death through lysis, which results in further bacteriophages being released to infect further bacteria. The F plasmid as a self-transmissible conjugative plasmid, is able to conjugate the entire genome of the *E. coli*. If a bacteriophage is able to drive evolution of a bacteria towards loss of the antibiotic resistant plasmid, the life of existing antibiotics could be extended. Through manipulation of the structure the range of targets for a bacteriophage can be broadened, along with their application with potentials for use in medicine, agriculture and within the food industry.[14]

1.3 Molecular modelling of protein structures

To perform their biological functions, proteins must adopt a tertiary structure, known as the folded state.[15] Hydration and other hydrophobic interactions are the driving force for protein folding.[16–18] Hydrogen bonds formed throughout the protein stabilise the secondary and tertiary structures of the protein, which gives the proteins resultant form.[19,20] The three-dimensional conformations adopted are constrained by the underlying amino acid sequence based on Anfinsen's hypothesis and are stabilised by the balance between enthalpic and entropic contributions.[15]

Energy functions indirectly approximate the contributions to predict protein structure, predicting that realistic representations are unique, thermodynamically stable and low energy conformations. Early energy functions used force constants taken from vibrational spectra to parameterize harmomic torsional potentials which were combined with Lennard-Jones potentials to represent van der Waals forces.[21] These soon diversified into several commonly used energy functions such as AMBER,[22] OPLS,[23] GROMACS,[24] DREIDING[25]

and CHARMM.[26] The latter is used in this project, due to accessibility on available HPCs, ease of use and speed. Recent technological advancements have enhanced the capabilities of energy functions, with derivation of parameters from ab initio quantum calculations, development of X-ray crystallography methods and determination of NMR structure. Molecular modelling gives us the tools for discovering the specific contacts between the MS2 maturation protein and F pilus as the range of interactions can vary, from short-ranged amino acid to amino acid contacts, and longer ranged hydrophobic interactions, alongside van der Waals forces.[27]

Protein-protein Interactions (PPIs) vary dramatically dependent on composition, affinity and whether the association is temporary or not.[28] PPIs can be homoligomers between identical chains, such as the formation of the caspid shell of the MS2 bacteriophage, or heteroligomers such as between the non-identical maturation protein and pilin. Interactions can be instantaneous (transient), or act permanently, however these types of PPIs are not distinct, as in many biological processes an entire spectrum can be seen based on conditions imposed so are based on the timescale of the complex between two proteins. Protein stability is also affected by the surrounding solvent which is favourable as the energy gained from the protein forming a complex is compensated by the entropy gained as the accessible protein surface area is reduced.[29]

1.4 Project Aims

There were several key objectives for this project. Protein-protein docking was to be performed to study whether the best docking sites had been found, and for a control when a larger fraction of the F pilus was modelled. Expansion of the F pilus model allows a larger degree of motion in the simulation to be obtained, and a more realistic approach, at computational cost. Once the model has been improved, point mutations of the maturation protein can be made in an attempt to increase the selectivity of the MS2 bacteriophage, so other pili could be attached, and increase its efficacy. Overall the challenge is to find the residues that must be present to allow for the maturation protein to dock to the F pilus.

1.5 Thesis Outline

The literature discussed in the introduction raises the global issue of antimicrobial resistance. Methodology of this project is described in Chapter 2, and the various computational models are explained. Chapter 3 investigates the docking protocols of HADDOCK[30,31] and RosettaDock.[32] This leads to development of the model. Chapter 4 investigates using molecular dynamics (MD) simulations to access the differences between the monomer and trimer model. Chapter 5 uses point mutations to assess ways in which the maturation protein can be modified to allow for other pili to potentially dock, increasing the selectivity. Chapter 6 concludes the project and summarises potential future steps.

2. Theory

2.1 Model

A single subunit of the F pilus is shown below in Figure 2.1a. Consisting of three parallel α-helices, which was determined to a 5 Å resolution through electron-counting cryo Electron Microscopy (cryoEM) and helical reconstruction.[10] The subunits each with a phospholipid unit combine to form the tubular F pilus.

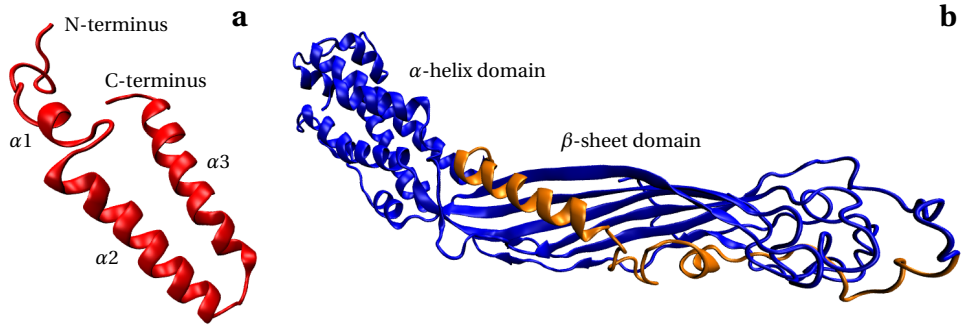


Figure 2.1: **a.** F Pilin structure. **b.** Maturation protein structure with the helix-loop-helix motif shown in orange.

The maturation protein shown in Figure 2.1b consists of an α-helix domain and a β-sheet domain and was constructed to a 3.6 Å resolution through a combination of cryoEM and asymmetric reconstruction.[11] The β-sheet domain contains a helix-loop-helix motif (residues 86 to 138) which interacts with the F pilus[33] as the β-sheet domain (residues 1-139, 226-268 and 314-378) is outside the MS2 caspid. The alpha-helix domain (residues 140-225, 269-313 and 375-393) remains inside the caspid structure. The F pilus is limited to where it can bind as the bacteriophage does not enter the pilus, so the interior face of the pilus is blocked.

2.2 Docking

Protein-protein docking was performed using HADDOCK (High Ambiguity Driven protein-protein DOCKing)[30,31] and RosettaDock[32] which is summarised in Figure 2.2

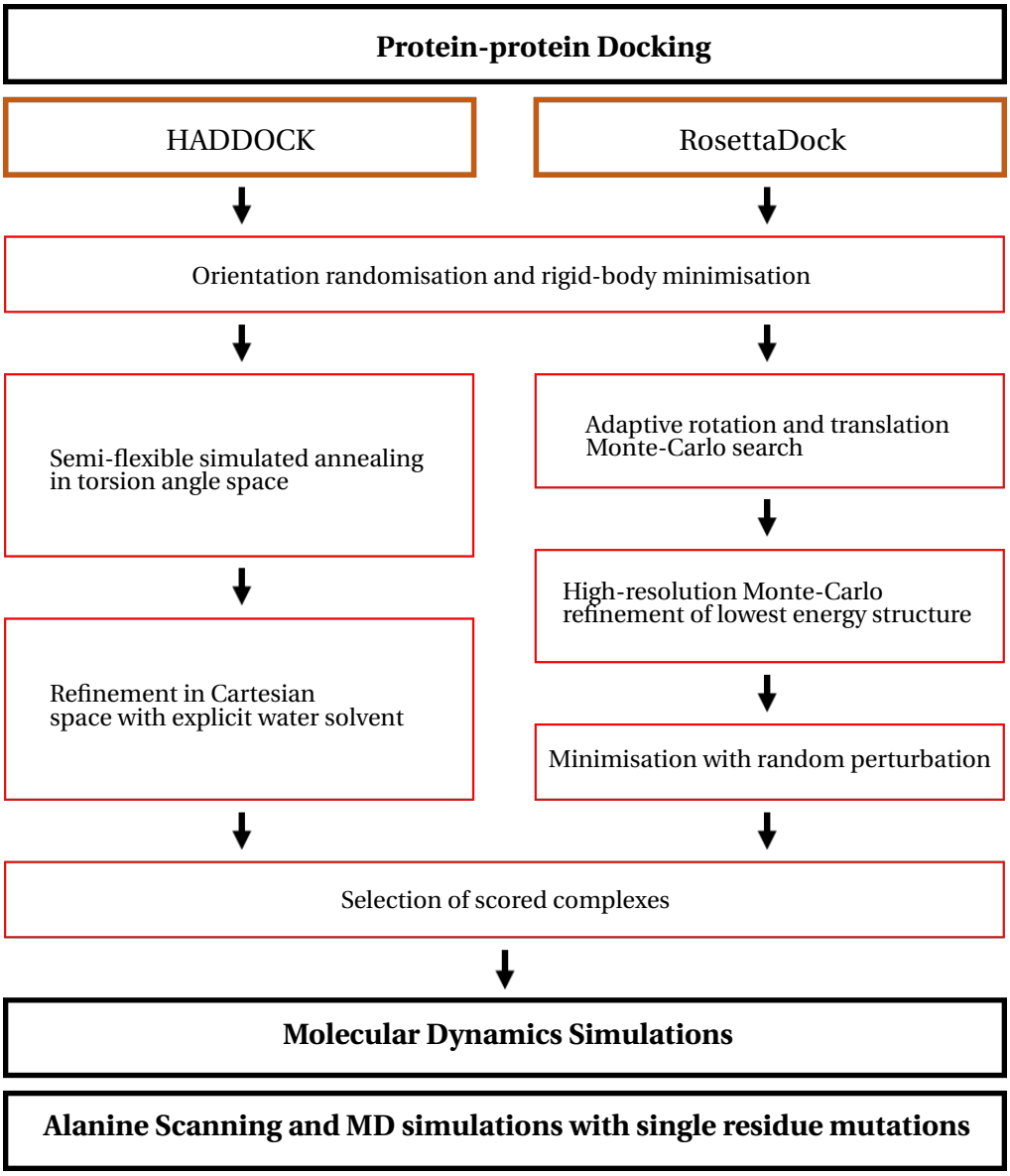


Figure 2.2: Summary of processes involved in protein-protein docking

HADDOCK

HADDOCK[30,31] follows a three stage docking protocol, with results clustered into similar structures, as follows:
Stage one - randomisation of orientations and rigid-body minimisation. In-

teracting partners are treated as rigid bodies, which are separated in space and rotated randomly about the centre of mass. Rigid bodies are energy minimised, where partners are allowed to rotate and translate to optimise interaction. Next the conformational space is sampled, depending on the Ambiguous Interaction Restraints (AIRs) that are defined (residues selected as active and passive).
Stage two - Semi-flexible simulated annealing in torsion angle space. A three step MD based refinement where first interacting partners are kept rigid and only the orientations are optimised. Flexibility is now introduced to the interface (analysis of intermolecular contacts within a 5 Å cutoff), allowing different binding poses from stage one to have different flexible regions defined. In the final step residues belonging to the flexible interface are allowed to move so the backbone and sidechains are granted freedom. AIRs are still important, as they might drive conformational changes.
Stage three - Refinement in Cartesian space with explicit solvent (water). Short explicit refinement models are subjected to a 300 K MD simulation, positional restraints on non-interface heavy atoms, later relaxed to allow all side chains to be optimised.

A set of 1 000 structures are generated, and clusters are created from structures that are similar. A maximum of 200 structures can be designated a cluster. Statistics of the top 10 clusters (if 10 are generated) are shown. The top cluster is the most reliable according to HADDOCK. For each cluster, the five best structures can be obtained. The scores given for the cluster do not match the individual scores, so the best overall score may not appear in clustering results.

HADDOCK score

The scores are returned as a HADDOCK Score (HS) which is only comparable to HADDOCK scores using the same scoring function. These vary between HADDOCK stages as shown in Table 2.1 For this reason, scores from the initial two stages of HADDOCK cannot be compared with the final explicit solvent stage. A lower or more negative score proposes a better structure.

HADDOCK scoring terms	Weighting		
	Stage 1	Stage 2	Stage 3
van der Waals intermolecular energy	0.01	1.0	1.0
Electrostatic intermolecular energy	1.0	1.0	0.2
Desolvation energy	1.0	1.0	1.0
Distance restraints energy	0.01	0.1	0.1
Buried Surface Area	- 0.01	- 0.01	0

Table 2.1: Weighting of HADDOCK scoring function for HADDOCK stages.

RosettaDock

RosettaDock[32] is a component of the RosettaCommons suite of programs, which follows the docking protocol as outlined below.
Docking begins with randomisation of orientations and minimisation of the rigid-body. Partner proteins are represented coarsely with the sidechains replaced with a single unified pseudo atom. A 500-step Monte-Carlo search is performed with adaptive rotation and translational steps adjusted dynamically to give an acceptance rate of 25 percent. The lowest energy structure is then selected for high-resolution refinement, where pseudo atoms are replaced with sidechain atoms from initial conformations. 50 Monte-Carlo with minimisation steps are made in which the rigid-body position is perturbed by a random direction and magnitude specified by a binomial-distribution around 0.1 Å and 3.0°. The rigid-body orientation is energy-minimised. If the score is acceptable sidechain conformations are optimised with Rotamer-Trials, an algorithm for packing sidechains, considering each residue only once, which is followed by a Metropolis criterion test. Every eight steps, an additional combinatorial sidechain optimization is carried out using the full sidechain packing algorithm, followed by an additional Metropolis criterion check.

Scores are returned in terms of Rosetta Energy Units (REU), which can only be compared to structures using the identical scoring function. A lower score suggests a better structure. The scoring weighting is shown in Table 2.2. The scoring is also weighted dependent on the balance of internal energies of each amino acid.

RosettaDock scoring terms	Weighting
Lennard-Jones attractive between atoms in different residues	1
Lennard-Jones repulsive between atoms in different residues	0.55
Lennard-Jones repulsive between atoms in the same residue	0.005
Solvation energy	0.9375
Coulombic electrostatic potential	0.875
Proline ring closure energy	1.25
Psi angle energy of residure preceding Proline	1.25
Hydrogen bonds backbone-backbone	1.17
Hydrogen bonds sidechain-backbone	1.17
Hydrogen bonds sidechain-sidechain	1.1
Disulfide geometry potential	1.25
Phi-psi angle preference (ramachandran)	0.25
Phi-psi angle preference (probability)	0.4
Omega backbone dihedral	0.625
Internal energy of sidechain rotamers	0.7
Torsional potential maintaining planar tyrosine hydroxyl	0.625

Table 2.2: Weighting of Rosetta scoring function for RosettaDock stages.

2.3 Molecular Dynamics

MD simulations were undertaken with NAnoscale Molecular Dynamics (NAMD) [34] using the CHARMM36 force field from Chemistry at HARvard Macromolecular Mechanics (CHARMM). [35,36] CHARMM was used for initial setup with help of the CHARMM-GUI [37,38] and for manipulating trajectories prior to analysis. The simulations were designed to replicate previous work as accurately as possible. [27]

MD simulations attempt to replicate the behaviour of biomolecular structures in a time-dependent nature through the use of a classical mechanics. An empirical force field describes the interactions between atoms, and integration of Newton's equations of motion are used to propagate the motion of particles.

Force Field

Interactions of atoms with other surrounding atoms exerts a force on the atom. This is stimulated by a force field. The CHARMM36 all-atom emperical

force field was used, and calculated from the potential energy where the force is equal to the negative of the energy gradient.

$$\boldsymbol{F} = -\nabla \boldsymbol{V} \tag{2.1}$$

The potential energy function consist of the following components;

$$V_{total} = V_{bond} + V_{angle} + V_{dihedral} + V_{improper} + V_{vdW} + V_{Coulomb} \tag{2.2}$$

where the total potential energy is dependent on the potential energy as a function of bond length, bond angle, dihedral angle, improper angle, van der Waals interactions and electrostatic interactions.

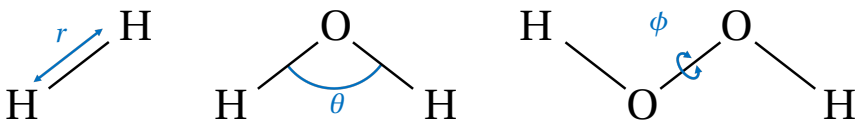


Figure 2.3: Bonding interactions of the force field, where r is the bond length, θ the bond angle and ϕ the torsion angle.

The first three bonding terms are shown schematically above. The potential for covalent bond stretching, the bonding angle between the bond and the dihedral angle between three bonds are shown in Equations 2.3, 2.4 and 2.5.

$$V_{bond} = \sum_{bond} k_r (r_i - r_{i,0})^2 \tag{2.3}$$

$$V_{angle} = \sum_{angle} k_\theta (\theta_i - \theta_{i,0})^2 \tag{2.4}$$

$$V_{dihedral} = \sum_{dihedral} k_\phi [1 + \cos(n_i \varphi - \delta_i)] \tag{2.5}$$

where k_r, k_θ, k_ϕ is the bond, angle and dihedral angle force constants respectively. r_i and $r_{i,0}$ are the bond length and equilibrium bond length. θ_i and $\theta_{i,0}$ are the bond angle and equilibrium bond angles. n_i is the multiplicity of the function, φ the torsion angle and δ_i the phase shift. The final bonding term is for the improper dihedral angle, which is used to select the correct geometry or chirality, and is the angle between planes of atoms ijk and jkl.

$$V_{improper} = \sum_{improper} k_{\varphi}(\varphi - \varphi_0)^2 \tag{2.6}$$

where k_{φ} is the improper angle force constant, and φ_0 is the equilibrium improper angle.

The final two terms are non-bonded interactions; van der Waals forces and electrostatic interactions, which are truncated to reduce computational cost as the interactions occur between all atoms that are not directly bonded. The Lennard-Jones potential is used to simulate weak non-polar interactions, and the Coulomb potential for electrostatics.

$$V_{vdW} = \sum_i \sum_{j>i} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{2.7}$$

$$V_{Coulomb} = \sum_i \sum_{j>i} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{2.8}$$

where ϵ_{ij} is the well depth, σ_{ij} is the distance where the Lennard-Jones potential is zero amd r_{ij} is the distance between atoms i and j. q_i and q_j are the charges on i and j respectively and ϵ_0 is the permittivity of free space. The two terms of the Lennard-Jones are a repulsive term proportional to r^{-12} and a attractive term proportional to r^{-6}. The potentials are both cutoff at 12 Å, with the Lennard-Jones beginning a smoothed truncation at 10 Å and the Columbic potential shifted to reach an effective zero potential at 12 Å.

Energy Minimisation

As a biomolecular structure changes conformation, so does the energy of the system. These energies form a multidimensional potential energy surface due to the large number of degrees of freedom in the structure. This energy landscape consists of a global minimum which is the state with the lowest energy, and many local minima which indicate states that are stable conformations. Stationary points within the potential energy surface occur, and these are detected by looking at the energy gradient.The energy gradient $\nabla \boldsymbol{V}$ is a vector consisting of all the possible first derivatives.

$$\nabla \boldsymbol{V} = \begin{bmatrix} \frac{\partial V}{\partial x_1} \\ \\ \frac{\partial V}{\partial x_2} \\ \\ \vdots \\ \\ \frac{\partial V}{\partial x_n} \end{bmatrix} \tag{2.9}$$

At a minimum, where the first derivatives $\boldsymbol{V}'(\boldsymbol{x})$ are zero as the force is equal to the negative gradient of energy as in Equation 2.1 and the second derivatives are all positive $\boldsymbol{V}''(\boldsymbol{x})$. The second derivatives form a n-by-n matrix called a Hessian $\boldsymbol{H}$ which determines the nature of the stationary point, from the force constants contained within.

$$\boldsymbol{H}_{ij} = \frac{\partial^2 \boldsymbol{V}}{\partial x_i \partial x_j} = \begin{bmatrix} \frac{\partial^2 V}{\partial x_1 \partial x_1} & \frac{\partial^2 V}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 V}{\partial x_1 \partial x_n} \\ \frac{\partial^2 V}{\partial x_2 \partial x_1} & \frac{\partial^2 V}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 V}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 V}{\partial x_n \partial x_1} & \frac{\partial^2 V}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 V}{\partial x_n \partial x_n} \end{bmatrix} \tag{2.10}$$

Energy minimisation, or geometry optimisation attempts to find the global minimum through a mathematical algorithm, which may get stuck on local minima. The geometry is altered in steps, with the energy being lowered slowly until it reaches the minimum. Various algorithms are used, some only using the first derivatives of energy such as steepest decent and conjugate gradient, and others such as the Newton-Raphson method are second order and use the second derivates as well.

Classical Molecular Dynamics

Numerical integration of Newton's equations of motion allows the movement of atoms to be calculated. The force on an atom i is related to its mass m_i and acceleration $\boldsymbol{a}_i$. The acceleration can also be determined from the potential energy through Equation 2.1. The energy gradient is calculated as a function of the atoms positions dependent of the interactions of the atoms, with each position being a function of time.

$$\boldsymbol{F}_i(t) = m_i \boldsymbol{a}_i(t) = -\boldsymbol{V}'(\boldsymbol{r}_i(t)) \tag{2.11}$$

for $i = 1,2,3,...n$. The velocity of a particle is the first derivative of the position, (Equation 2.12) and acceleration the second derivative and also the first derivative of the velocity (Equation 2.13).

$$\boldsymbol{v}_i(t) = \boldsymbol{r}_i'(t) \tag{2.12}$$

$$\boldsymbol{a}_i(t) = \boldsymbol{v}_i'(t) = \boldsymbol{r}_i''(t) \tag{2.13}$$

The use of numerical solutions allows approximated solutions of the differential equations 2.12 and 2.13.

$$\frac{\boldsymbol{F}_i(t)}{m_i} = \boldsymbol{r}_i''(t) \tag{2.14}$$

As the position and velocities at time t is known, positions and velocities before $t+\delta t$ and after $t-\delta t$ this point in time can be approximated where δt is the time step, as long as δt is small.

The timestep is selected to be shorter than the highest frequency of motion to prevent instabilities in integration. Translations have low frequencies, then rotation, torsions and vibrations the highest frequencies. The SHAKE[39] algorithm removes constrained hydrogen bond vibrations, meaning a 2 fs was used to lessen the overall effect on protein stability, but not shorter, as the shorter the time step the larger the computational cost due to the number of microscopic states explored.

The velocity Verlet integration is a second order algorithm used by NAMD. The Taylor expansion is used to approximate velocities of atoms at different time steps. Equation 2.15 calculates the position at $t+\delta t$ and Equation 2.16 at $t-\delta t$ from the positions and accelerations of the current step.

$$\boldsymbol{r}(t+\delta t) = \boldsymbol{r}(t) + \boldsymbol{r}'(t)\delta t + \frac{1}{2}\boldsymbol{r}''(t)\delta t^2 + O(\delta t^3) \tag{2.15}$$

$$\boldsymbol{r}(t-\delta t) = \boldsymbol{r}(t) + \boldsymbol{r}'(t)\delta t + \frac{1}{2}\boldsymbol{r}''(t)\delta t^2 - O(\delta t^3) \tag{2.16}$$

hence the position of the next step $\boldsymbol{r}(t+\delta t)$ can be calculated, through addition and rearrangement of Equations 2.15 and 2.16.

$$\boldsymbol{r}(t+\delta t) = 2\boldsymbol{r}(t) - \boldsymbol{r}(t-\delta t) + \boldsymbol{r}''(t)\delta t^2 \tag{2.17}$$

When substituted with Equation 2.13 gives:

$$\boldsymbol{r}(t+\delta t) = 2\boldsymbol{r}(t) - \boldsymbol{r}(t-\delta t) + \boldsymbol{a}(t)\delta t^2 \tag{2.18}$$

If the velocities $\boldsymbol{v}(t)$ of the current step and acceleration $\boldsymbol{a}(t)$ of the current and subsequent step $\boldsymbol{a}(t+\delta t)$ are known, the velocities of the atoms in the subsequent step $\boldsymbol{v}(t+\delta t)$ can be found.

$$\begin{aligned}\boldsymbol{v}(t+\delta t) &= \boldsymbol{v}(t) + \boldsymbol{v}'(t)\delta t + \frac{1}{2}\boldsymbol{v}''(t)\delta t^2 + O(\delta t^3) \\ &= \boldsymbol{v}(t) + \boldsymbol{a}(t)\delta t + \frac{1}{2}\boldsymbol{a}'(t)\delta t^2 + O(\delta t^3) \\ &= \boldsymbol{v}(t) + \frac{1}{2}\{\boldsymbol{a}(t) + \boldsymbol{a}(t+\delta t)\}\delta t + O(\delta t^3)\end{aligned} \tag{2.19}$$

Ensemble

An ensemble considers a large number of possibilities that a system could take, and acts as a probability distribution for the system. The ensemble used depends on the system involved and the conditions that are required to be maintained. This system uses the Canonical ensemble (NVT) and Isobaric-Isothermal ensemble (NPT) ensembles. The NVT ensemble can be used to identify thermodynamic phenomena at a constant temperature, as NVT keeps the number of particles, volume and temperature constant. The NPT ensemble is ideal for replicating experimental conditions for the system as it keeps particles, pressure and temperature constant. Other ensembles are not suitable in this case such as the Microcanocial ensemble which keeps the overall energy volume and particles constant, as temperature needs to be controlled.

The F pilin - maturation protein complex is heated in the NVT ensemble through velocity reassignment in simple Newtonian dynamics to a Maxwellian velocity distribution, shown in Equation 2.20.

$$f(v) = \left(\frac{m}{2\pi k_B T}\right)^{\frac{3}{2}} \exp\left(\frac{-mv^2}{2k_B T}\right) \tag{2.20}$$

where $f(v)$ is the distribution of velocities, m mass, and k_B is the Boltzmann constant, T the temperature and v being velocity.

This system also uses NVT during heating and initial equilibration and then uses NPT until the end of production, as it is closest to experimental conditions, which we are trying to replicate. The NPT ensemble requires a barostat to maintain pressure and a thermostat to maintain temperature. The pressure is controlled using the Nosé Hoover method[40] and Langevin dynamics[41] are use to control fluctuations of the barostat.

Langevin Dynamics is a stochastic method which represents the overall force on the system, including the effect of the explicit solvent on protein, and that the system is not in a vacuum. The Langevin equation[42] below (2.21) contains Newton's equations of motion from Equation 2.11 alongside a frictional drag force caused by the presence of solvent in the system, and a random force which replicates fluctuations caused by atoms colliding and interacting with the solvent. The Langevin equation also incorporates the Langevin thermostat to the system. The Newton component represents the overall force on an atom due to interaction with surrounding atoms.

$$\begin{aligned} m_i \boldsymbol{a}_i(t) &= \boldsymbol{F}(r_i(t)) - \gamma \boldsymbol{v}_i(t) m_i + \boldsymbol{R}(t) \\ &= -\Delta \boldsymbol{V} - \gamma \boldsymbol{v}_i(t) m_i + \boldsymbol{R}(t) \end{aligned} \tag{2.21}$$

where γ is the friction coefficient of the solvent and $\boldsymbol{R}(t)$ is the random force, which has a Gaussian distribution and a mean value of zero, independent of the velocity and position of an atom.

Temperature of the system is maintained through the frictional drag and random force terms.

Solvent Model

Previous work concluded that an explicit water model produces results that were more accurate than those from an implicit solvent model, and it was worth the extra computational cost.[27] An explicit solvent considers water molecules individually. The Transferable Intermolecular Potential 3P (TIP3P) water model[43] was chosen for its simplicity and wide use in the academic field.

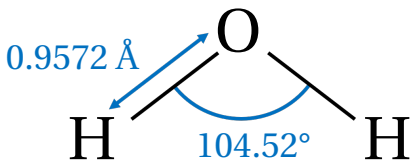


Figure 2.4: Geometry of the TIP3P water model.

The TIP3P water model is a rigid three-site structure defined with a hydrogen to oxygen distance of 0.9572 Å and a 104.52° hydrogen-oxygen-hydrogen bond angle. Oxygen has a partial charge of -0.834 and hydrogen +0.417.

Long-range electrostatics are the most computation expensive part of MD simulations along with other nonbonded potential energy calculations. The long-range electrostatics requirements can be reduced through approximation. This was done with the Ewald summation, using Particle Mesh Ewald (PME) [44] which approximates with a Fourier and a real space component. The long-range contribution uses a fast Fourier transform in Fourier space, and the short-range in the real space. A distance limitation for long-range electrostatic pair wise interactions is added to reduce the evaluated interactions.

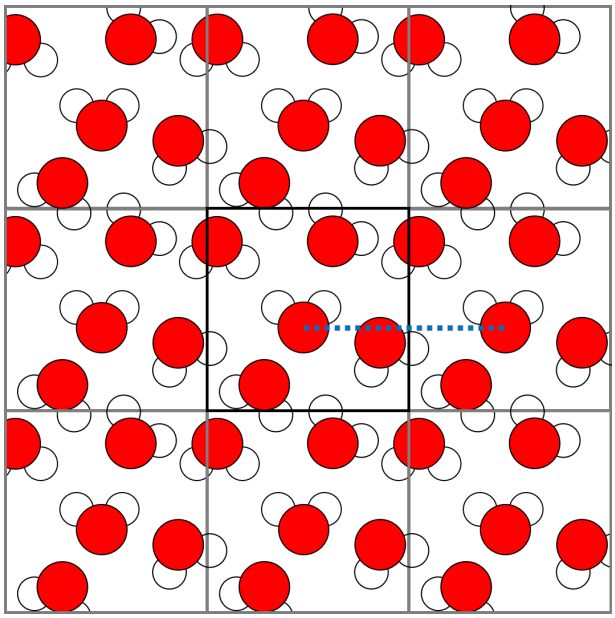


Figure 2.5: Model of a 2-dimensional square under PBCs. The blue dotted line shows the juxtaposition of a water molecule.

Thousands of water molecules are contained in a truncated octahedral unit cell, which results in the large amount of surface waters creating unwanted surface effects at the borders of the simulation. This is rectified th-

rough the use of Periodic boundary conditions (PBCs) which replicates bulk solvent. As shown in Figure 2.5 an infinite system is mimicked, as when a atom leaves the system on one side, it is replaced by an atom arriving on the opposite side, not unlike the video game Pacman. All the water molecules have long-ranged interactions calculated for a certain distance for all surrounding molecules, and hence when crossing the box, the atoms will still have an effect from neighbouring molecules even though they appear on the opposite side of the box.

Methodology

Initially created using CHARMM-GUI,[37,38] the docked complex of maturation protein and F pilin was solvated in a truncated octahedral TIP3P solvation box. The box was expanded to 10 Å from the edge of the complex, so the orientation and size of the pilin affected the size of the waterbox (and time of simulation due to extra water molecules. Bad contacts were removed through both the method of steepest descent and adopted basis Newton-Raphson algorithm for 50 steps, and PBCs were applied.

MD simulations were performed using the CHARMM36 forcefield in NAMD 2.12[34] with a 2 fs time step. A cutoff of 12 Å was used for van der Waals interactions, PME was used for long-range electrostatics and the SHAKE algorithm for the fixing of bond lengths including hydrogen.

Each simulation began with a 10 000 step energy minimisation using the conjugate gradient algorithm. Backbone and sidechain atoms were restrained with harmonic restraints with force constants of 10.0 and 5.0 kcal mol^{-1} Å^{-2} respectively. The system was then heated to 298 K by 3 K every 1 ps, giving 500 steps between each velocity reassignment to equilibrate. Then five 500 ps equilibration phases under the NPT ensemble were ran with the force constant restraints reduced to zero by 2.5 kcal mol^{-1} Å^{-2} per phase. A 20 ns production simulation was then run using the NPT ensemble. Langevin dynamics were applied with a friction coefficient of 5 ps^{-1} throughout minimisation, heating, equilibration and production.

Analysis

Molecular Dynamics (MD) trajectories were analysed through use of the python package MDTraj.[45] Trajectories were visualised using Visual Molecular

Dynamics (VMD)[46] alongside the toolkit VMD provides.

Root Mean Square Deviation

Root Mean Square Deviation (RMSD) is calculated using MDTraj[45] as in equation 2.22.

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\delta_j - \delta_i)^2} \tag{2.22}$$

where δ_i is the position of an atom i at the start of the preparation and δ_j is the position of atom i for the moment the RMSD is calculated for, with the complex being aligned with the original structure to remove translation effects.

PPIs were calculated by identifying residues with interactions with other residues within a certain distance. Residues were defined as being in contact with each other if their β-carbon atoms were within 8 Å of each other. This distance was used so interactions with long side chains would be included, including the residues arginine, lysine, tryptophan and tyrosine.

2.4 Point Mutations

Alanine scanning was calculated using the Robetta2.23[47] server, using a simple free energy function as in Equation .

$$\begin{aligned}\Delta G = {} & W_{attr}E_{LJattr} + W_{attr}E_{LJrep} + W_{HB(sc-bb)}E_{sc-bb} \\ & + W_{HB(sc-sc)}E_{sc-sc} + W_{sol}G_{sol} + W_{\phi/\psi}E_{\phi/\psi}(aa) + \sum_{aa=1}^{20} n_{aa}E_{aa}^{ref}\end{aligned} \tag{2.23}$$

where W is the relative weights of the different energy terms,[48] E_{LJattr} and E_{LJrep} are the attractive and repulsive Lennard-Jones terms, E_{sc-bb} and E_{sc-sc} orientation-dependent hydrogen bond potentials for sidechain-backbone and sidechain-sidechain respectively, G_{sol} the implicit solvation model, $E_{\phi/\psi}(aa)$ an amino acid type dependent backbone torsion angle propensity and E_{aa}^{ref} an amino acid type dependent reference energy, approximating interactions based on n_{aa} the number of amino acids of a certain type. The procedure is summarised in Figure 2.6.

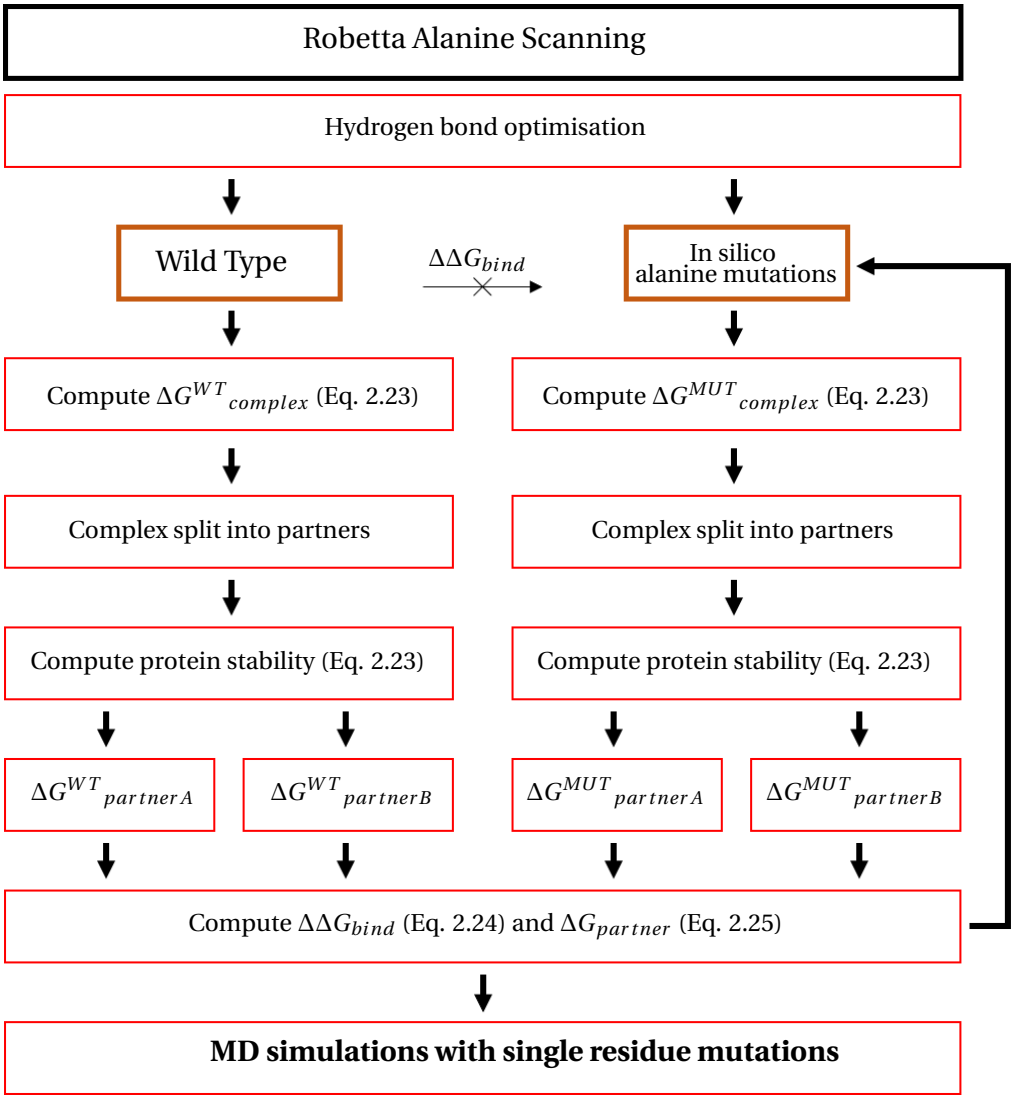


Figure 2.6: Summary of Robetta Alanine Scanning

Predicted change in binding free energy (ΔΔG)

Computational alanine scanning used two pathways, one for the original or wild type complex, and another for the mutated complex. Free energy is calculated first for the complex as $\Delta G^{WT}{}_{complex}$ and $\Delta G^{MUT}{}_{complex}$. Then the complex is split into partners and protein stability is computed, giving free energy for both partners, for four terms, $\Delta G^{WT}{}_{partnerA}$, $\Delta G^{WT}{}_{partnerB}$, $\Delta G^{MUT}{}_{partnerA}$ and $\Delta G^{MUT}{}_{partnerB}$ respectively. From this the predicted change in binding free energy ΔΔG can be calculated as in Equation 2.24 below.

$$\begin{aligned}\Delta\Delta G_{bind} = &(\Delta G^{WT}{}_{complex} - \Delta G^{WT}{}_{partnerA} - \Delta G^{WT}{}_{partnerB}) \\ &- (\Delta G^{MUT}{}_{complex} - \Delta G^{MUT}{}_{partnerA} - \Delta G^{MUT}{}_{partnerB})\end{aligned} \tag{2.24}$$

Predicted change in stability (ΔG)

For the stability prediction, fewer terms are required, as in equation 2.25 however this also means less variance between the mutated residues are seen due to the similarity in scores.

$$\Delta G_{partner} = \Delta G^{WT}{}_{partner} - \Delta G^{MUT}{}_{partner} \tag{2.25}$$

3. Comparison of docking models

Docking of the MS2 maturation protein and substructures from F pilus was performed using HADDOCK[30,31] and RosettaDock.[32] The relative merits of these programs were assessed.

3.1 Docking energy plots

To determine the respective merits of HADDOCK and RosettaDock, histograms were made for the energies reported by the two docking programs. For the HADDOCK structures, 200 structures were taken from the final stage of docking, of four independent runs for a total of 800 structures using selected residue for the F pilin. A65 used residues 1-16 and 65 of the F pilin as active residues. B60 used residues 1-16 and 60-65 of the F pilin as active residues. Both A65 and B60 were then run again with additional passively selected residues. Passive residues are those residues within 6.5 Å of active residues. All runs used residues 86-138 of MS2 as active restraints. These residues were selected as active through a combination of homology modelling and protein threading done previously by the School of Veterinary Medicine and Science. Crucially these portions of the proteins are exposed, which are not blocked from interaction by surrounding protein chains.

Meanwhile 1 000 structures were generated through 3 Å translations and rotations within 8° from RosettaDock for the monomer.

To test if the single monomer substructure of the F pilus was a good approximation of the whole F pilus structure, a larger portion of F pilin was used, a trimer rather than just a monomer. For the trimer the same strategy was followed, however the restraints were selected on only the middle substructure of the F pilin trimer in the case of HADDOCK, whilst RosettaDock is unable to be selective, and just rotates the trimer around the maturation

protein randomly.

HADDOCK

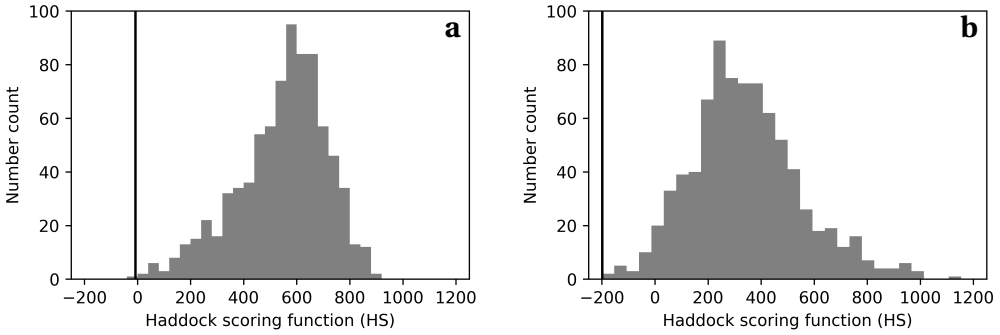


Figure 3.1: Histograms of the energies for 800 monomer structures **(a)** and 800 trimer **(b)** structures, created and scored by HADDOCK. The black line indicates the best scoring stucture (monomer = -7.8 HS, trimer = -198 HS).

The scores of the structures created by Haddock have a spread of scores between -200 to 1 200 HS. The best scoring monomer structure has a score of -7.77 HS, whilst the best scoring trimer structure scores -198 HS. When docking is done with the trimer, there is a tendency for lower, better scores. The most common scores are between 560 - 600 HS for the monomer and 200 - 240 HS for the trimer.

RosettaDock

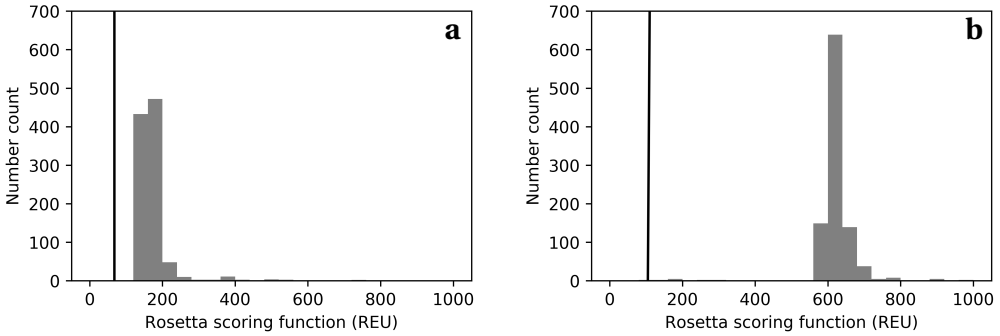


Figure 3.2: Histograms of the energies for 1 000 monomer structures **(a)** and 1 000 trimer structures (**b**) created and scored by RosettaDock. The black line indicates the best scoring stucture (monomer = 67.5 REU, trimer = 105 REU).

The RosettaDock results show a large number of structures with a similar energy, with significantly less variation than the HADDOCK scoring, with a range between 0 and 1 000 REU, which is not directly comparable to HS. The best scoring monomer structure scores 67.5 REU, and the best scoring trimer scores 105 REU. The majority of trimer scores are worse than the monomer structures. The most common scores are between 160 - 200 REU for the monomer, but 600 - 640 REU for the trimer, which is the reverse of what is seen for HADDOCK. RosettaDock cannot be given restraints to encourage certain residue to dock together, instead chains of protein move around (the monomer or trimer moves around the maturation protein). This means that RosettaDock is predisposed to bind the edge monomers of the trimer rather than the central monomer for steric reasons, which may explain the poorer performance of the monomer.

Rescoring

As the two scoring functions from HADDOCK and RosettaDock are not comparable, the structures were rescored in Rosetta after being processed through CHARMM which made the Protein Data Bank (PDB) files identical (PDBs from HADDOCK do not contain non-polar hydrogen atoms, whilst PDBs from RosettaDock do).

RosettaDock Monomer

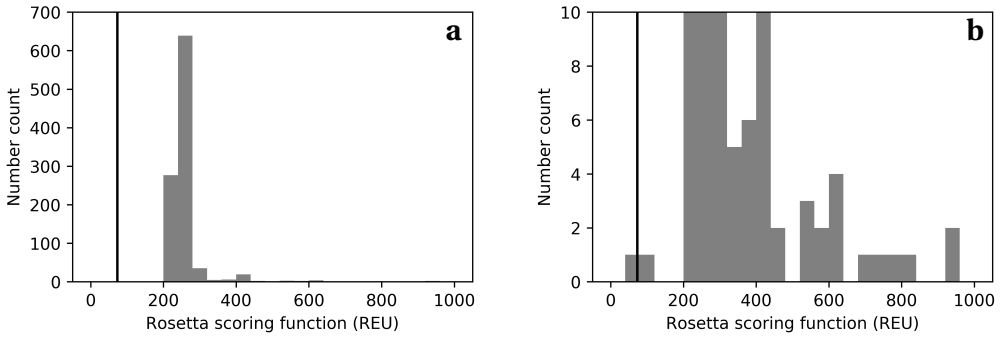


Figure 3.3: Histograms of the energies for 1 000 monomer structures created by RosettaDock and rescored by Rosetta **(a)**, rescaled to show scores of where only a few structures were found **(b)**. The black line indicates the best scoring stucture (73.2 REU).

The histogram varies slightly from Figure 3.2 due to the CHARMM processing where hydrogens are removed and later re-added, with the general trend re-

sulting in slightly worse scores, with the most common scores being between 240 - 280 REU . The best scoring structure scores 73.2 REU.

HADDOCK Monomer

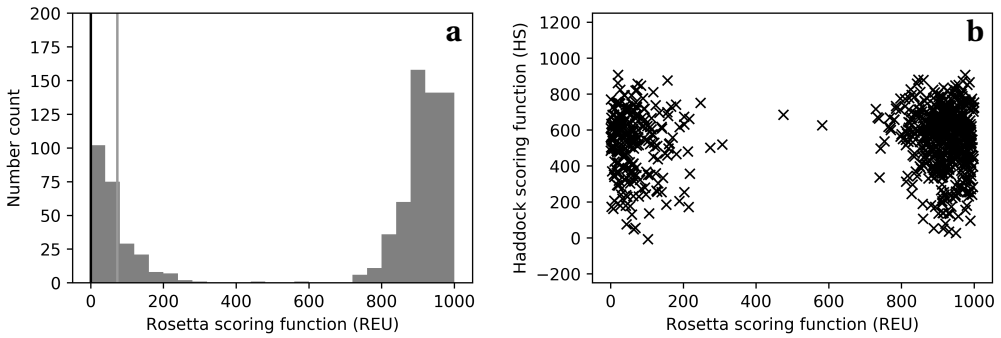


Figure 3.4: **a.** Histograms of the energies for 800 monomer structures created by HADDOCK and scored by Rosetta. The black line represents the lowest scoring structure (0.269 REU), and the grey line the comparative best score obtained from RosettaDock structures as in Figure 3.3 (73.2 REU). **b.** Plot of HADDOCK scoring function versus Rosetta scoring function for the 800 monomer structures.

When rescored by Rosetta, HADDOCK monomer structures tend to do better than the equivalent RosettaDock structures, with the best structure 72.9 REU lower at 0.269 REU, and the most common scores are between 0 - 40 REU. There is no direct correlation between HS and REU as shown in Figure 3.4 (a), but a discontinuity is evident, with a group of good scores and a group of bad scores. This was initially believed to two orientations of the F pilin being observed. However this is not the case, as histograms for individual orientations showed the same split of bad and good scores. From observation of poor scoring results, the poor scoring of HADDOCK structures in Rosetta is due to a poor interaction of the pilin with the edge of the β-sheet domain, as shown in Figure 3.5.

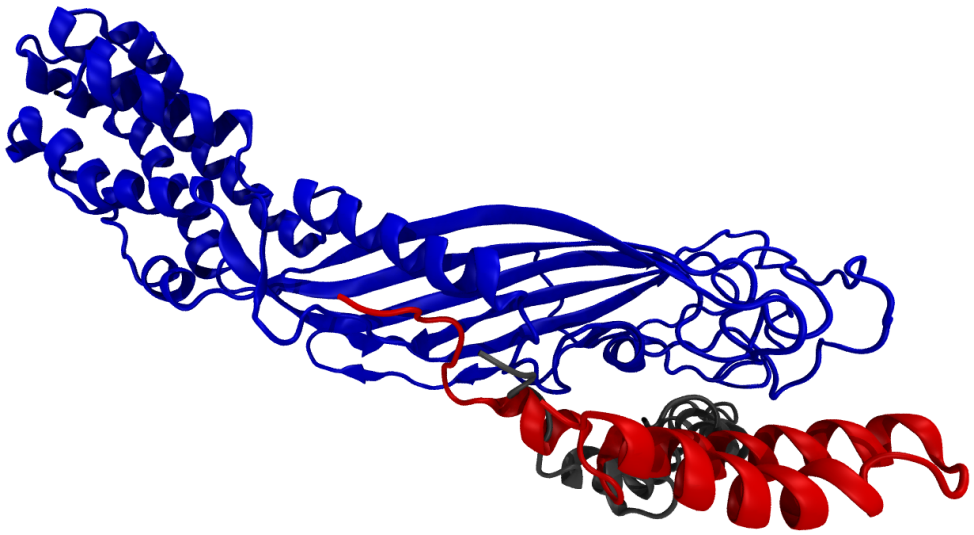


Figure 3.5: Docked structure of maturation protein and F pilin, good interaction with maturation protein (red), bad interaction (grey)

Through manual inspection of the structures, the poorest scoring structures arise when the pilin is too close to the MS2, where as only a small portion is connected in the better scoring cases. This may be due to how Rosetta deals with side chains, usually requiring prepacking which modifies steric clashes.

RosettaDock Trimer

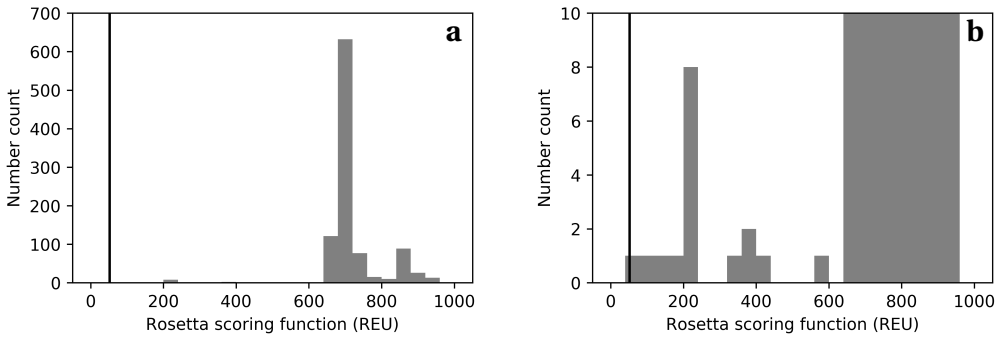


Figure 3.6: Histograms of the energies for 1 000 trimer structures created by RosettaDock and rescored by Rosetta **(a)**, rescaled to show energies of where only a few structures were found **(b)**. The black line indicates the best scoring stucture (51.6 REU).

Compared to the trimer originally in Figure 3.2 rescoring means the histogram varies slightly, with the general trend resulting in slightly worse scores, with

the most common scores being between 680 - 720 REU . The best scoring structure scores 51.6 REU, which is a better, lower score than the best rescored RosettaDock monomer structure.

HADDOCK Trimer

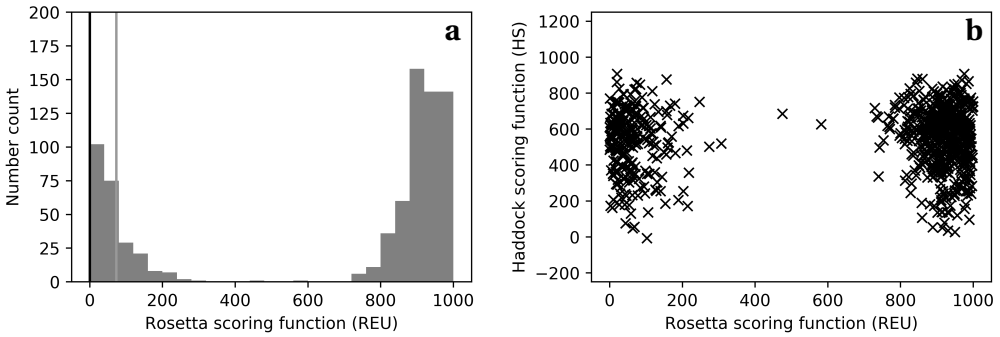


Figure 3.7: **a.** Histograms of the energies for 800 trimer structures created by HADDOCK and scored by Rosetta. The black line represents the lowest scoring structure (0.269 REU), and the grey line the comparative best score obtained from RosettaDock structures as in Figure 3.6 (73.2 REU). **b.** Plot of HADDOCK scoring function versus Rosetta scoring function for the 800 trimer structures.

Rescored HADDOCK trimer structures again did better in Rosetta than the equivalent RosettaDock structures, with the best structure 36.0 REU lower at 15.6 REU, less of a difference than with the monomer. The most common scores are also higher than the monomer at 240 - 280 REU. Despite this, 99% of HADDOCK trimer structures score lower than all but the best 2% of RosettaDock trimer structures. Figure 3.7 (b) shows no direct correlation between the Haddock and Rosetta scoring functions but no discontinuity is seen when the 800 HADDOCK structures are scored by Rosetta. This could be accounted for the steric hindrance of the two chains without active residues preventing the pilin from lying too close to the MS2 maturation protein.

From the above analysis, HADDOCK is a better model, as HADDOCK results score better in Rosetta than the majority of RosettaDock structures. However futher work is needed to assess how well it represents the docking of the maturation protein and the F pilus, in particular whether the trimer or monomer is a more accurate representation.

3.2 Clustering analysis

The best results of HADDOCK and RosettaDock were taken forward to a clustering algorithm MaxCluster, as seen in Table 3.1 for the monomer and Table 3.2 for the trimer structures. The structures were hierarchically clustered through maximum linkage into groups which had a variation in Root Mean Square Deviation (RMSD) for all backbone atoms of less than 1 Å.

Structure set	N^o structures	N^o clusters	Cluster size(s)
HADDOCK top 5%	40	5	14, 6, 7, 2, 2
HADDOCK best 10 [1]	10	3	3, 3, 2
RosettaDock top 5%	50	5	14, 13, 7, 2, 2
RosettaDock best 10 [2]	10	1	2
[1] + [2]	20	4	3, 3, 2(H), 2(R)

Table 3.1: Cluster analysis of monomer structures with a cutoff of 1 Å. (H) represents a cluster originating from only HADDOCK structures and (R) a cluster from only RosettaDock structures.

The clustering shows a number of similar structures of the top 5% of structures for both HADDOCK and RosettaDock. However, the similarities between the top 10 structures of each is less, and no clusters are shared between the top 10 structures of both when compared together.

Structure set	N^o structures	N^o clusters	Cluster size(s)
HADDOCK top 5%	40	7	8, 5, 4, 4, 3, 3, 2
HADDOCK best 10 [1]	10	2	4, 2
RosettaDock top 5%	50	1	25
RosettaDock best 10 [2]	10	0	
[1] + [2]	20	2	4, 2(H)

Table 3.2: Cluster analysis of trimer structures with a cutoff of 1 Å. (H) represents a cluster originating from only HADDOCK structures.

The trimer structures show similar trends to the monomer case. However RosettaDock structures now show little clustering. The lack of clusters containing a mixture of both HADDOCK and RosettaDock structures means that clustering isn't able to prove that the models generated are similar. Despite this, it does show the best HADDOCK structures were more closely related to each other than the RosettaDock structures were.

3.3 Analysis of top structures

Further to the clustering, RMSDs were calculated using MDTraj[45] between each structure for the best 10 models of HADDOCK and RosettaDock - so a 10 by 10 matrix for HADDOCK against RosettaDock results, the average and standard deviation are shown in Table 3.3 for the monomer and Table 3.4 for the trimer results. Also shown is the close interacting residues - the interface between MS2 and F pilin. If the interface differs between two structures, all atoms that are in the interface of either structure are included in the RMSD calculation.

Structure sets	Complex RMSD		Interface RMSD	
	Average (Å)	Closest (Å)	Average (Å)	Closest (Å)
H10, R10	9.58 ± 2.93	4.59	11.7 ± 4.82	4.16
H10, H10 [1]	7.88 ± 5.98	0.93	8.69 ± 6.33	1.20
R10, R10 [2]	5.83 ± 2.88	0.45	8.21 ± 5.64	0.312
Average of [1], [2]	6.85 ± 4.78	-	8.45 ± 5.96	-

Table 3.3: Monomer - Mean RMSDs and the closest RMSD for the complex and for the interface of the top 10 monomer structures from HADDOCK (H10) and RosettaDock (R10). Where the RMSD is calculated for all atoms but hydrogen, the complex is the docked protein - protein structure between the F pilin monomer and maturation protein of MS2 and the the interface defined as residues within an beta carbon distance of 8 Å to the other protein.

The results show that there is a high average RMSD difference between structures, even when only comparing within the same docking programs results, with the standard deviation also being high. Through inspection this is largely due to two orientations of the MS2 and F pilin interaction, which is discussed in more detail in Section 3.4. The best structures from RosettaDock show less variation than the best structures from HADDOCK, and the comparison of the two. The closest RMSD is significantly lower for both HADDOCK and RosettaDock on their own, supporting that they find similar structures internally but not compared to each other. The interface RMSDs are generally higher due to it involving much less of the MS2 maturation protein, which is a far larger protein, which does not move much, hence lowering a single RMSD when used in full.

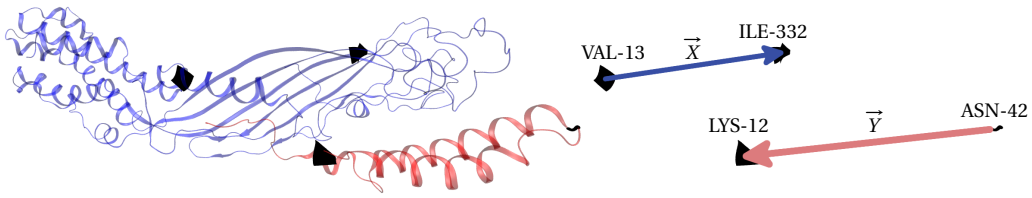
Structure sets	Complex RMSD		Interface RMSD	
	Average (Å)	Closest (Å)	Average (Å)	Closest (Å)
H10, R10	12.50 ± 6.85	5.41	12.3 ± 6.47	4.16
H10, H10 [1]	9.33 ± 8.83	0.99	8.63 ± 7.95	1.23
R10, R10 [2]	11.4 ± 7.46	3.47	10.4 ± 7.26	2.21
Average of [1], [2]	10.3 ± 8.19	-	9.51 ± 7.62	-

Table 3.4: Trimer - Mean RMSDs and the closest RMSD for the complex and for the interface of the top 10 trimer structures from HADDOCK (H10) and RosettaDock (R10). Where the RMSD is calculated for all atoms but hydrogen, the complex is the docked protein - protein structure between the F pilin trimer and maturation protein of MS2 and the the interface defined as residues within an C_β-C_β distance of 8 Å to the other protein.

The trimer shows the same pattern of results, with larger RMSDs and standard deviation due to the moving F pilin becoming larger. However, now HADDOCK outperforms RosettaDock, with the closest results being similar to the monomer, but for RosettaDock they are significantly increased. The comparison between the two docking methods is still larger than the individual sets.

3.4 Qualitative analysis of orientation

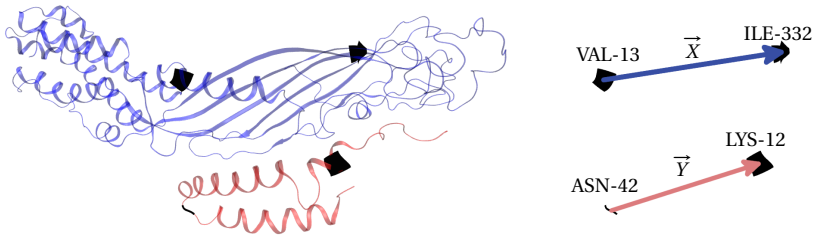
As mentioned previously two orientations were observed, and resulted in high RMSD values. The two orientations consist of one where the pilin is anti-parallel to the β-pleated sheets denoted as orientation A, shown in Figure 3.8 and another where the pilin is parallel to the β-pleated sheets denotated as orientation B, shown in Figure 3.9. These have also been seen in previous work which indicates that the A orientation has a stronger binding.



$$\vec{X} \cdot \vec{Y} = |X||Y|\cos\theta, \text{ where } 90° < \theta < 180°$$

Figure 3.8: Complex structure with orientation A and corresponding vectors.

To distinguish between the two orientations, vectors between different residues are taken for both MS2 and F pilin, to give the directions the proteins are facing. These are kept fixed, and the resulting dot product between the two vectors allows for distinction between orientations, with a negative $\cos\theta$ value for orientation A, and a positive $\cos\theta$ value for orientation B.



$$\vec{X} \cdot \vec{Y} = |X||Y|\cos\theta, \text{ where } 0° < \theta < 90°$$

Figure 3.9: Complex structure with orientation B and corresponding vectors.

Orientation ratios

Further investigation on the two orientations, and how likely they are to be the most accurate model was required. The ratio of orientation A and orientation B changes from the monomer to the trimer were calculated, as shown in Table 3.5.

Structure set	Monomer		Trimer	
	Ratio A:B	A/B (%)	Ratio A:B	A/B (%)
All HADDOCK	692:108	86/14	381:419	48/52
HADDOCK top 5%	22:18	55/45	24:16	60/40
HADDOCK best 10	7:3	70/30	9:1	90/10
HADDOCK best	1:0	100/0	1:0	100/0
All RosettaDock	977:23	98/2	977:23	98/2
RosettaDock top 5%	49:1	98/2	48:2	96/4
RosettaDock best 10	10:0	100/0	9:1	90/10
RosettaDock best	1:0	100/0	1:0	100/0

Table 3.5: Ratios of orientation A to orientation B for generated structures.

The ratios for RosettaDock are fairly consistent towards orientation A, but for HADDOCK, the trimer ratio shifts towards orientation B for all structures, but for the top structures orientation A is favoured more than it was in the monomer's case with 9 of the top 10 structures being orientation A.

3.5 Monomer and Trimer comparison

The RMSD between the docking of the maturation protein and either the monomer or the central subunit of the trimer is 1.78 Å. The difference is reduced due to the alignment of the maturation protein, but also increased by the edges of F pilin which is not interacting in either case with the maturation protein, as shown in Figure 3.10. This is evident in the buried surface area, which for the monomer is 2476 $Å^2$ and 2031 $Å^2$ for the trimer.

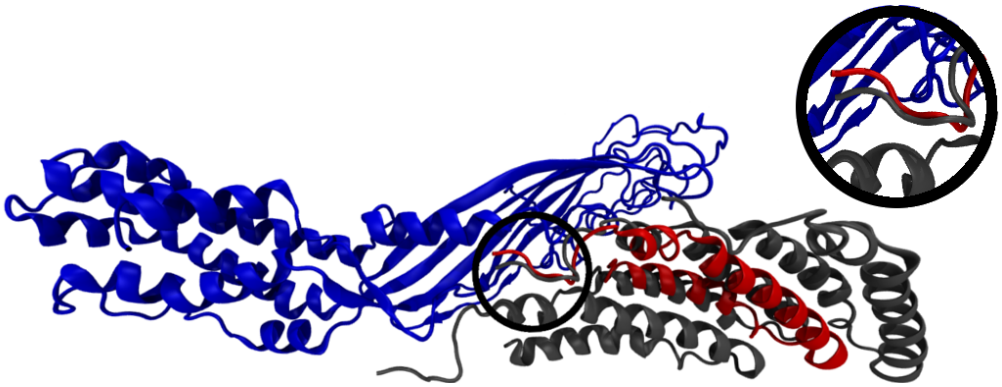


Figure 3.10: Docked structure of maturation protein and F pilin, monomer (red), and trimer (grey).

3.6 Conclusion

The best structures from RosettaDock share more similarity with each other than the best structures from HADDOCK in the monomer case, and the reverse for the trimer model. Despite this the trends that they show are broadly the same. The best HADDOCK structures score better in Rosetta than RosettaDock structures do, increasingly so for trimer structures. This, and the ability of HADDOCK to selectively bind to the middle monomer of the trimer, leads us to consider that HADDOCK docking is better. The trimer model is a more realistic overall model, despite less contacts made, due to including the steric blocking of other monomer strands when on approach and MD simulations show similar movement of the trimer and maturation protein as was seen for the bound monomer and maturation protein.

4. Molecular Dynamics Simulations

Molecular dynamics simulations were run after the complex was docked. MD simulations give an indication to what would happen to the complex in experimental conditions. We wanted to observe the contacts that were made in docking, and whether they remained throughout the duration of a 20 ns simulation, and did not dissociate. The preparation for production dynamics, which was improved is discussed in Appendix 7.A and benchmarking was performed to test the available computer clusters in Appendix 7.B.

4.1 Monomer Simulations

Analysis of contacts

The first use of MD simulations was for a monomer structure, which was used for a baseline compared to previous studies and is in the A orientation, which was found to be the best structure of four that were previously worked on.[27] This structure was also generated by HADDOCK.

The improved preparation protocol from Appendix 7.A was used, and five 20 ns production dynamic simulations were run for repeatable results. NAMD use a seeding based on the current clocktime of the system as a basis for velocity reassignment in the heating stage which creates independent results, and the seed numbers generated were kept.

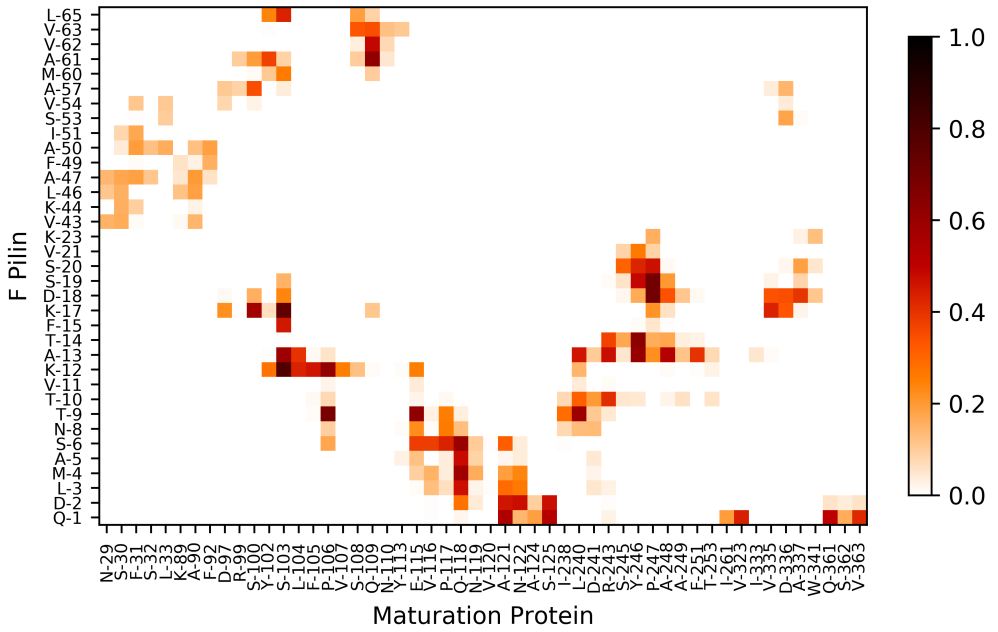


Figure 4.1: Map of contacts showing the fraction of time that residues between the monomer F pilin and the maturation protein stayed in contact for five 20 ns repetitions. A contact was defined as being in contact if the C_β-C_β of two residues were within 8 Å of each other.

The interactions between F pilin and the maturation protein were averaged and mapped, shown in Figure 4.1 which shows the fraction of time contacts were observed. There were several regions with a group of contacts such as between residues 240 - 261 of the maturation protein and 6 - 28 of the pilin, 115 - 125 of the maturation protein and 1 - 9 of the pilin and 29-99 of the maturation protein and 43 - 65 of the pilin. The abbreviations for residues can be found in Appendix 7.C. Five interactions are present in all repeats for more than half the simulation, with the highest being Ser103-Lys12 was present on average for 99% of the simulation. Ser103-Lys17 and Pro106-Thr9 were present for 88% of the time. Pro106-Lys12 is present for 80% of the simulation and Ser103-Ala13 for 78% of the simualtion. The maturation protein residues are from all from the helix-loop-helix motif in the beta sheet domain, and the residues from the pilin are between the N-terminus and end of the first alpha helix. This does mean that the model shows some promise, as these are parts of the pilin that are accessible when fully assembled as the tubular structure comprising of multiple subunits.

RMSD calculations

To assess the stability of the docked structures, RMSD calculations were made, using the maturation protein and the monomer of F pilin docked to the maturation protein (all of the monomer atoms, or atoms from the central substructure of the trimer).

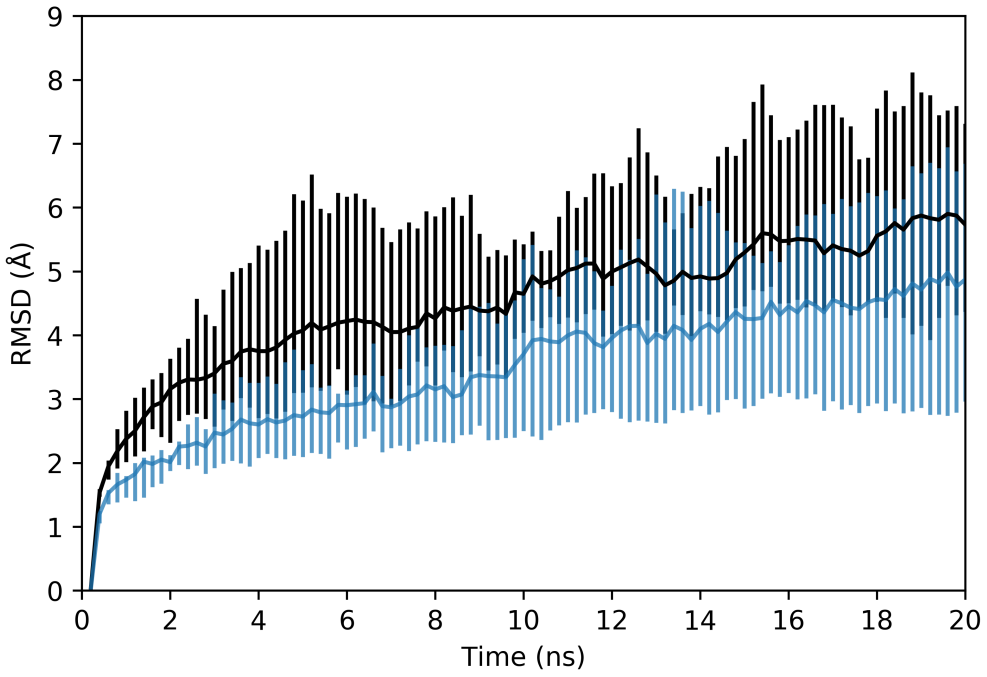


Figure 4.2: RMSDs for all non-hydrogen atoms, and the interface defined as residues within an C_β-C_β distance of 8 Å to the other protein at the start of production dynamics, averaged for the five repeats of the monomer.
Key: MS2 to F pilin complex average (black), Interface average (blue).

The average RMSD for the complex rises rapidly at the beginning, reaching 2.5 Å after 1 ns, then increases steadily throughout and reaches a final RMSD of 5.6 Å at 20 ns, whilst the maximum value occurring is 6.0 Å. During the second half of production dynamics, the RMSD rises by 0.7 Å, compared with the 4.9 Å for the first 10 ns. One repeat has significantly more deviance than the other four as shown in Figure 3.3. The standard deviation between values also is reduced for the second half of simulation from 0.78 Å to 0.34 Å.

The interface RMSD remains lower than that of the complex throughout the simulation. At 20 ns the RMSD reaches 4.5 Å, with a maximum throughout

of 4.7 Å. The first 10 ns shows an increase of 3.5 Å whilst the latter 10 ns has an increase of 1.0 Å which is higher than that of the complex as a whole during the final part of the production, so more more change is seen in the interface region than the rest of the protein-protein complex. The standard deviation is lower for the second half still at 0.36 Å compared to the first 10 ns standard deviation of 0.53 Å.

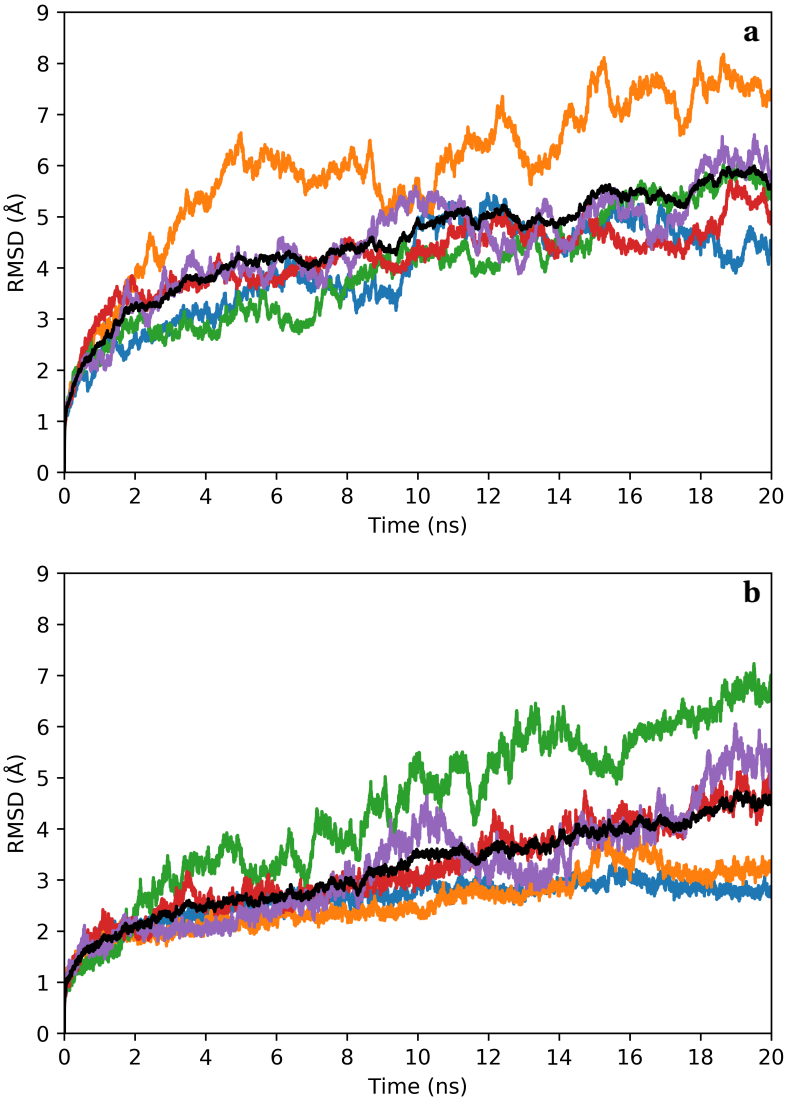


Figure 4.3: RMSDs for all non-hydrogen atoms, for each repeat and averaged for the five repeats of the monomer. **a.** Entire complex. **b.** The interface defined as residues within an C_β-C_β distance of 8 Å to the other protein at the start of production dynamics.
Key: Repeat 1 (blue), 2 (orange), 3 (green), 4 (red), lilac (5) and Average (black).

For repeats of the monomer both the complex and the interface shows a rapid initial increase, and then a more gradual continuous increase. Repeat 2 shows significant deviation from the rest in the complex RMSD, but the interface of repeats 3, 4 and 5 deviate significantly towards the end of the 20 ns.

Native contacts

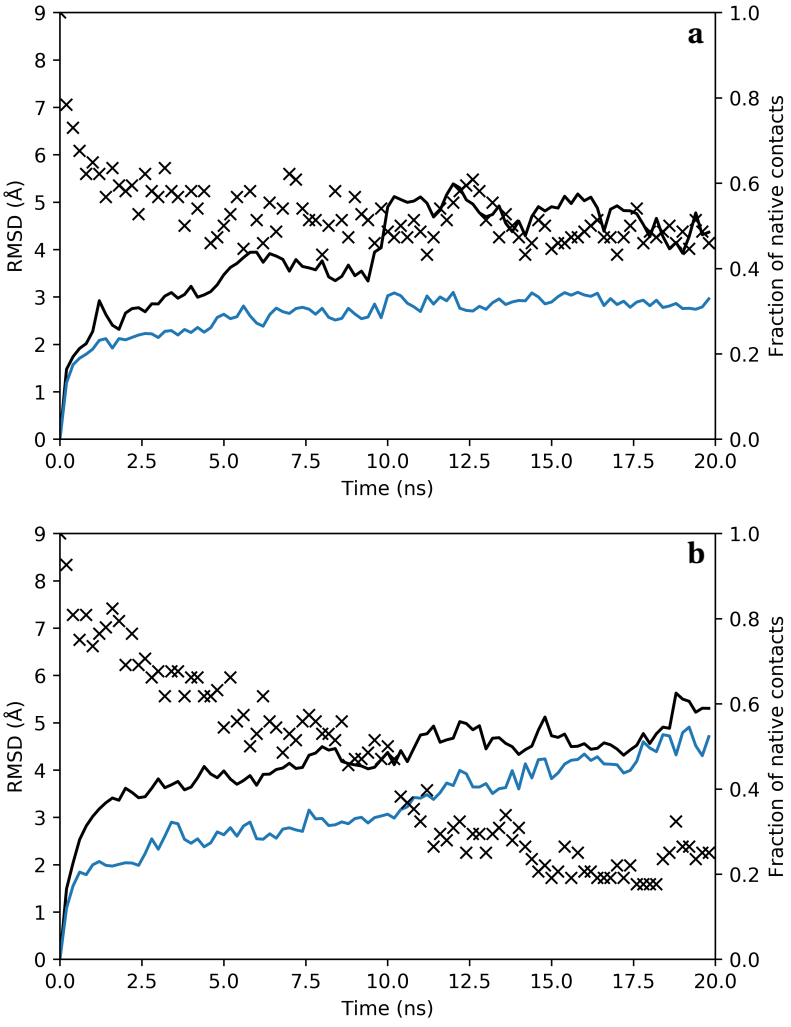


Figure 4.4: Fraction of native contacts remaining and RMSDs for all non-hydrogen atoms, for the monomer repeat with the highest textbf(a), **(b)** and lowest final percentage of native contacts, where native contacts and the interface are defined as residues within an C_β-C_β distance of 8 Å to the other protein at the start of production dynamics.
Key: Complex (black), Interface (blue) and Fraction of native contacts (crosses).

The monomer RMSDs showed that three repeats for the monomer and one repeat of the trimer had significantly higher RMSDs than the repeat with the lowest RMSD. In order to investigate whether the complex has disociated, the contacts that were in place at the start of the 20 ns simulation were measured to see if they remained and compared to the RMSD, again using MDTraj.[45] The strongest and weakest native contact fraction results for the monomer and trimer are shown in Figure 4.4 and Figure 4.9 respectively.

The lowest case for the monomer shows a steady drop in the fraction of native contacts remaining, however the best case for the monomer still shows a steady but slower drop, but the RMSD of the complex is substantially higher than that of the interface. There are three repeats with higher final native contact fractions of 0.46, 0.46 and 0.45 and then two lower scores of 0.30 and 0.27.

4.2 Trimer Simulations

Analysis of contacts

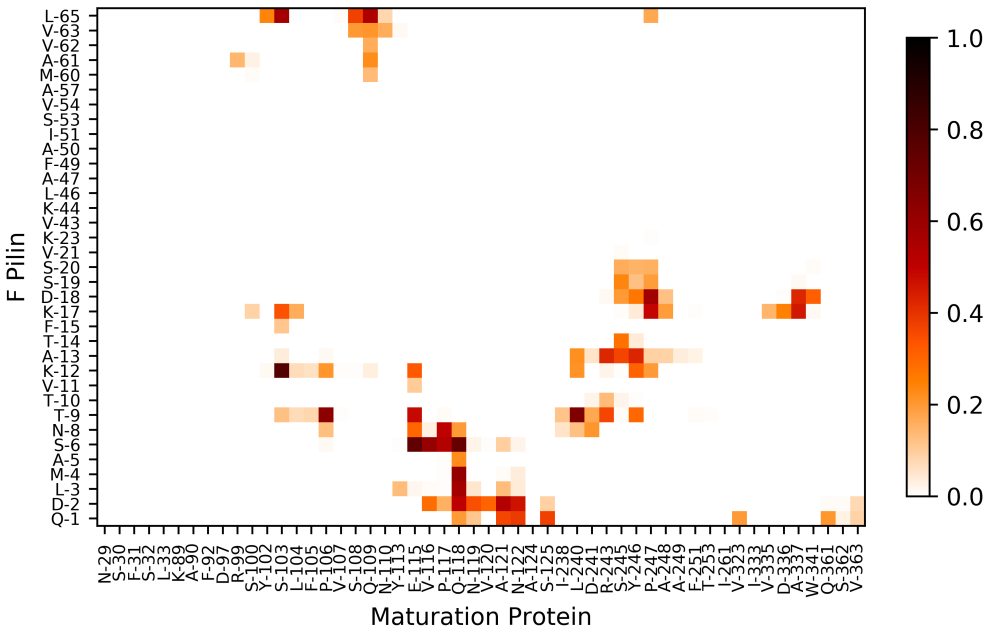


Figure 4.5: Map of contacts showing the fraction of time that residues between the trimer F pilin and the maturation protein stayed in contact for five 20 ns repetitions. A contact was defined as being in contact if the C_β-C_β of two residues were within 8 Å of each other.

The trimer structure scored best in HADDOCK in the previous chapter, and is of orientation A. Again five 20 ns repeats were averaged, shown in Figure 4.5. Fewer contacts are seen than in the monomer simulation previously. However, the regions where contacts are found are similar, yet the region of 29-99 of the maturation protein and 43 - 65 of the pilin minimal contacts were seen. Seven contacts are initially present after equilibration for all repeats of both the monomer and trimer F pilin models; Ser103-Lys12, Val116-Ser6, Ala121-Asp2, Asn122-Asp2, Arg243-Ala13, Pro247-Asp18 and Ala337-Asp18. Just three contacts were present more than 50% of the time for all trimer repeats; Ser103-Lys12, Glu115-Ser6 and Glu118-Ser6 for 97%, 94% and 93% of the time respectively. The residues are in the same region as the residues present in all monomer repeats, but only Ser103-Lys12 is seen for the majority of the time in both the monomer and trimer results. This could prove to be a useful benchmark for futher models.

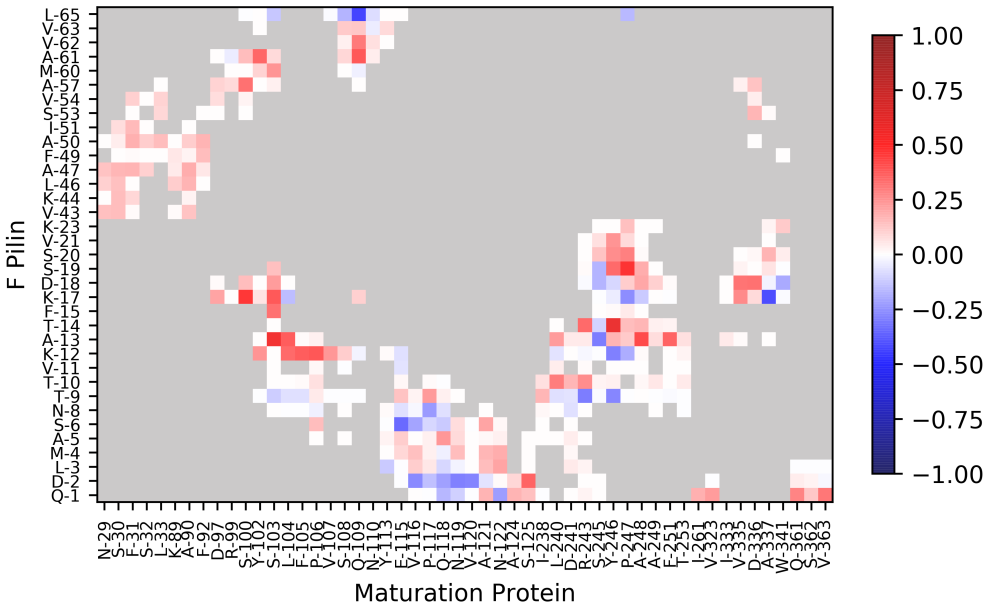


Figure 4.6: Map of contacts showing the difference in contacts made between the monomer and trimer. Red indicates a contact was present more in the monomer MD simulations and blue that a contact was present more in trimer MD simulations. Grey indicates contacts that were never seen.

The difference map Figure 4.6 shows that a significant number of contacts are shared between both the monomer and trimer models. The white boxes show where the contacts are present for a similar length of time, but does not discriminate for barely present contacts in both or ever present in both. There

are only a few contacts where the trimer has significantly more time present than the monomer, and even less where the contacts of the trimer is present much more consistently. This is consistent with the visualisation that less of the pilin can bind to the maturation protein with the trimer model, although within the region that early studies suggested.

RMSD calculations

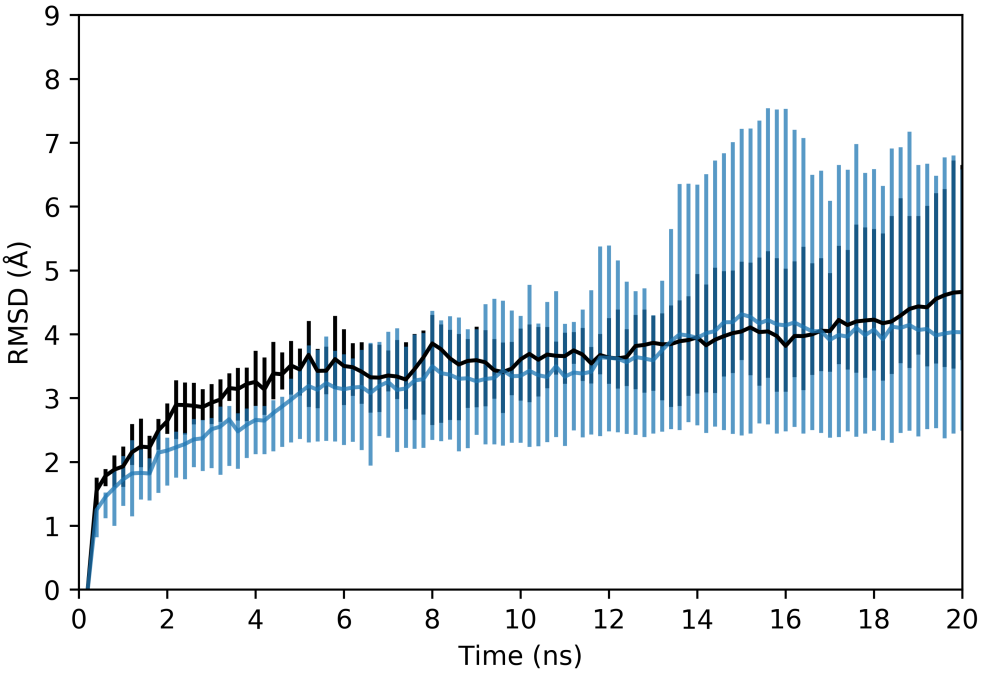


Figure 4.7: RMSDs for all non-hydrogen atoms, and the interface defined as residues within an C_β-C_β distance of 8 Å to the other protein at the start of production dynamics, averaged for the five repeats of the trimer.
Key: MS2 to F pilin complex average (black), Interface average (blue).

The average RMSD for the complex rises rapidly at the beginning, reaching 2.1 Å after 1 ns, then increasing steadily throughout and reaches a final RMSD of 4.7 Å at 20 ns, and a maximum value of 4.8 Å which is lower than seen for the monomer. The change in RMSD between 10 ns and 20 ns is larger than that seen for the monomer at 1.0 Å, but much smaller than the inital 10 ns of 3.6 Å like what was seen for the monomer. The standard deviation between values also falls to 0.29 Å for the second half from 0.59 Å from the first half of production dynamics.

RMSD calculations

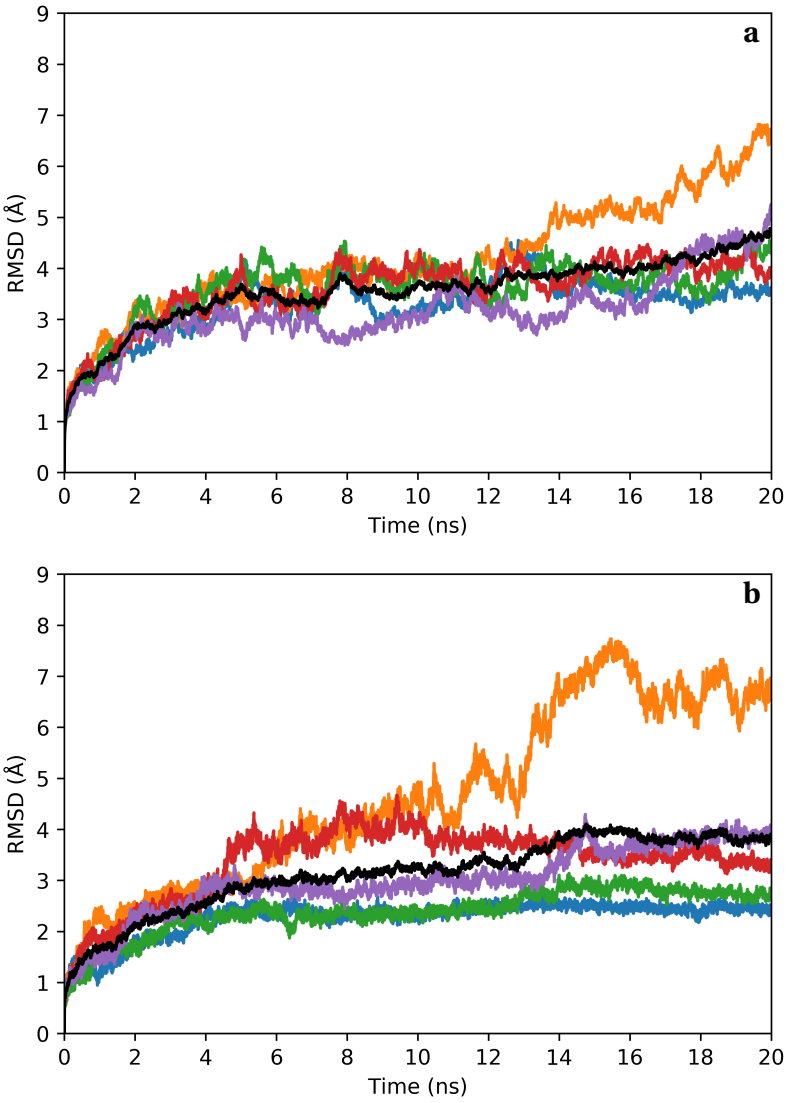


Figure 4.8: RMSDs for all non-hydrogen atoms, for each repeat and averaged for the five repeats of the trimer. **a.** Entire complex. **b.** The interface defined as residues within an C_β-C_β distance of 8 Å to the other protein at the start of production dynamics.
Key: Repeat 1 (blue), 2 (orange), 3 (green), 4 (red), lilac (5) and Average (black).

The interface RMSD remains lower than that of the complex for the majority of the simulation, but briefly rises above due to one repeat that is significantly above the rest as shown in Figure 4.8, which results in the large error

bars for the interface. The trimer interface average is still lower overall than the monomer interface at 20 ns with a RMSD of 3.8 Å, and a maximum value of 4.1 Å. The interface also performs better for the second half of production dynamics with a change of 0.6 Å, whilst increasing to 3.2 Å during the inital 10 ns. The standard deviation between values falls to 0.28 Å for the second half from 0.60 Å from the first half of production dynamics.

For repeats of the trimer both the complex and the interface shows a rapid initial increase. However relative to the monomer this flattens off, except for repeat 2 which shows significant variation for both the complex and the interface. Without the repeat with the exceptionly high RMSD, the stability and convergence of the trimer interface is greater than that of the monomer interface.

Native contacts

The repeat with the lowest final percentage of native contacts for the trimer shown in Figure 4.9 shows a steady drop in the fraction of native contacts remaining, with the final fraction being lower than the weakest case of the monomer, and higher RMSDs for both the complex and the interface. However the repeat with the highest final percentage of native contacts for the trimer drops at the start yet then remains steady and low. The RMSDs also remain fairly steady. There again are three repeats with a high final native contact fraction at 0.67, 0.55 and 0.51 and two repeats with lower score; 0.26 and 0.09. The lowest score is lower than the lowest score for the monomer repeats, though all three of the high scores are higher than the high scores of the monomer.

The average final native contact fraction is 0.37 for the monomer repeats and 0.41 for the trimer repeats. The minimum native contact fraction averaged for all repeats is lower at 0.27 for the monomer and 0.28 for the trimer. This again supports the trimer being a better model than the monomer, if only marginally. On the other hand there is more variation on the trimer with a standard deviation of 0.21 for the final native contact fraction, compared with the 0.11 of the monomer.

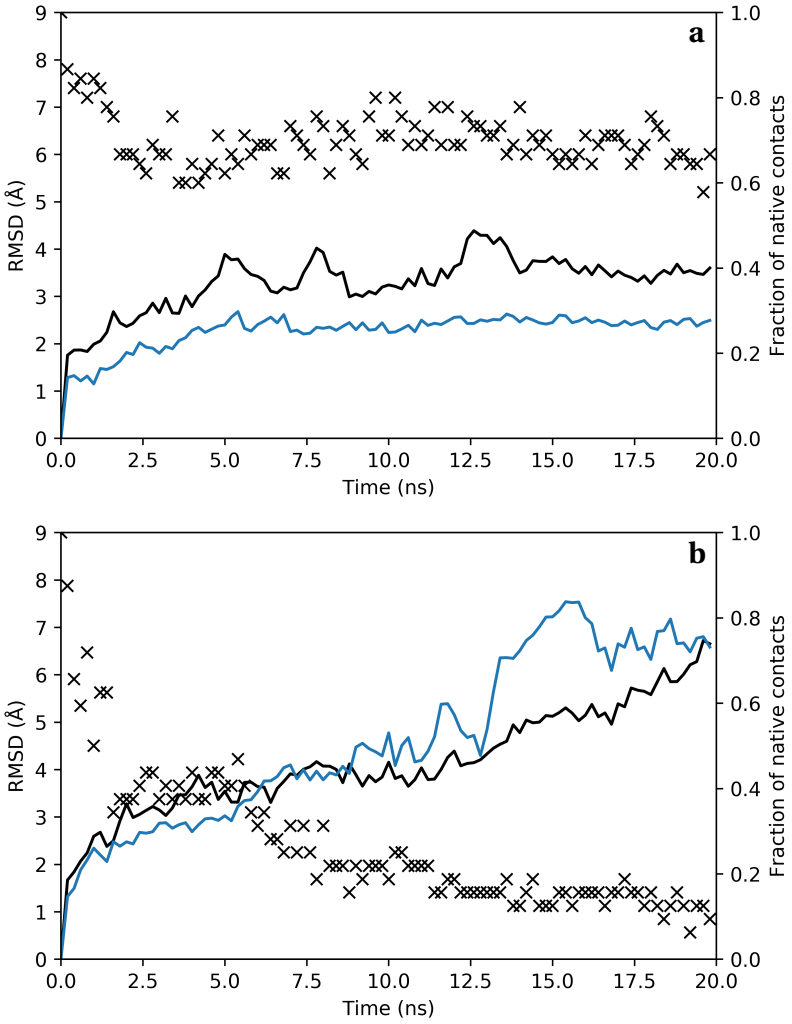


Figure 4.9: Fraction of native contacts remaining and RMSDs for all non-hydrogen atoms, for the trimer repeat with the highest textbf(a), **(b)** and lowest final percentage of native contacts, where native contacts and the interface are defined as residues within an C_β-C_β distance of 8 Å to the other protein at the start of production dynamics.
Key: Complex (black), Interface (blue) and Fraction of native contacts (crosses).

The production simulations have shown how the trimer model has less contacts than the monomer, but key contacts are shared. Analysis shows that the model has slight improvements on stability based on the fraction of native contacts remaing. Due to this improvements to the model by using the trimer F pilin can be taken foward towards mutation studies.

5. Point Mutations on MS2 Caspid Protein

5.1 Alanine scanning

For an indication of which residues should be mutated alanine scanning was calculated through use of the Robetta server for prediction of Gibbs free energy.[47] Table 5.1 below shows the residues that destabilised and stabilised the maturation protein of MS2 the most.

Effect	Residue	ΔG (kcal/mol)	Effect	Residue	ΔG (kcal/mol)
Stabilising	Gln118	-1.3	Destabilising	Arg259	7.2
	Trp341	-0.8		Trp364	3.6
	Asn122	-0.6		Ile114	2.6
	Gln361	-0.5		Val360	1.9
	Ser125	-0.5		Val323	1.7

Table 5.1: Change in predicted Gibbs free energy for the five most stabilising residues and five most destabilising residues.

These residues were taken forward to simulation dynamics with those residues replaced with alanine.

5.2 Initial Simulations

Stabilising

Ser125 was the most stabilising with a native contact fraction of 0.68 at the end and a minimum value of 0.53. Trp341 was the next most stabilising with 0.50 and a minimum of 0.36. The other residue mutations were lower than

that of the trimer average with final contact fraction of 0.28 for Gln118, 0.22 for Asn122 and 0.14 for Gln361. The overall average was 0.38 and an average minimum of 0.29 compared to the trimer of 0.41 at the end and an average 0.28 minimum which is not a significant difference, as the standard difference is high for the trimer at 0.20 for the final native contact value and 0.17 for the minimum value. This compares with standard deviations of 0.19 and 0.15 for the mutations end and minimum values respectively The number of residues defined as the interface at the start of production was significantly lower however with an average of 39 compared to the 50 of the trimer, which suggests a weaker binding, though the RMSD of the complex on average is similar at 4.4 Å compared to the 4.6 Å of the trimer, yet the interface RMSD is higher at 4.5 Å compared to 3.8 Å for the trimer, but the lower number of residues involves means each one contributes more.

Destabilising

For destabilising interactions, a lower score is desired than what was seen for the trimer and monomer models, to identify residues that must be kept for successful binding. Arg259 when mutated led to a contact fraction of just 0.16 at the end of production which was the desired result. Trp364 Val360 and Val323 were also moderately successful at 0.30, 0.31 and 0.39 respectively. Unfortunately Ile114 mutated to alanine had the opposite effect and stabilised at 0.68 so may have been misidentified by Robetta. This skewed the results, but still less residues in the interface were detected at 37 and a higher RMSD for the complex at 4.7 Å. The standard deviation is lower at 0.15 and 0.14 for the mutations end and minimum values respectively.

5.3 Further scanning

Given limited success with this method in which stabilising and destabilising residues are selected, further methods included within Robetta were tested. The first of these was a Hotspot method was used to identify residues which contribute the most to binding energy, for which both stabilising and destabilising residues were selected.[49] Both the Hotspot method and the original method were then selected based on the stability of the complex as a whole rather than just the maturation protein (described in Robetta as the partner protein). Each residue appearing in the top five of a method was then mu-

tated to alanine and a 20 ns simulation run (some residues appeared in several method lists). The results are then summarised for the methods, shown in Table 5.2.

Method	Complex RMSD		Interface RMSD		n^o residues	Native Contacts	
	max (Å)	end (Å)	max (Å)	end (Å)		min	end
DP(H)	5.4	4.7	4.4	3.8	37	0.22	0.35
SP(H)	5.8	4.4	5.1	4.5	39	0.29	0.38
DP	5.4	4.3	4.2	3.7	39	0.27	0.44
SP	5.4	4.7	5.6	4.8	39	0.22	0.29
DC(H)*	5.6	4.9	4.5	4.0	41	0.24	0.43
SC(H)	5.1	4.4	4.4	3.7	41	0.29	0.43
DC*	5.6	4.9	4.5	4.0	41	0.24	0.43
SC†	5.5	4.5	4.2	3.7	42	0.35	0.48
Trimer	6.5	5.7	5.6	4.7	50	0.31	0.42
Monomer	6.5	5.7	5.4	4.6	51	0.29	0.40

* Mutation residues in Destabilising Complex are identical to those in Destabilising Complex Hotspot.
† Six residue mutations were averaged due to identical ΔΔ G values.

Table 5.2: Average results of simulations for mutations selected by the different methods. Average from monomer and trimer simulations included for comparison. Where the complex is the docked protein - protein structure between the F pilin trimer and maturation protein of MS2 and the n^o residues is the amount of residues in the interface, defined as residues within an C_β-C_β distance of 8 Å to the other protein. Native Contacts is the fraction of native contacts present that were found at the start of production dynamics.
Key: Stabilising (S), Destabilising (D), Partner (P), Complex (C) and Hotspot (H).

These results are then plotted as an average in Figure 5.1 and as individual residues in Figure 5.2. The mutations used based on the stabilising partner method produced poor results, more than expected from destabilising residues. With exception to that result, the minimum fraction of the destabilising methods were lower than the trimer and monomer models, whilst the stabilising methods were similar or better than the trimer and monomer models, so would be suitable for replacement as no adverse affects were seen.

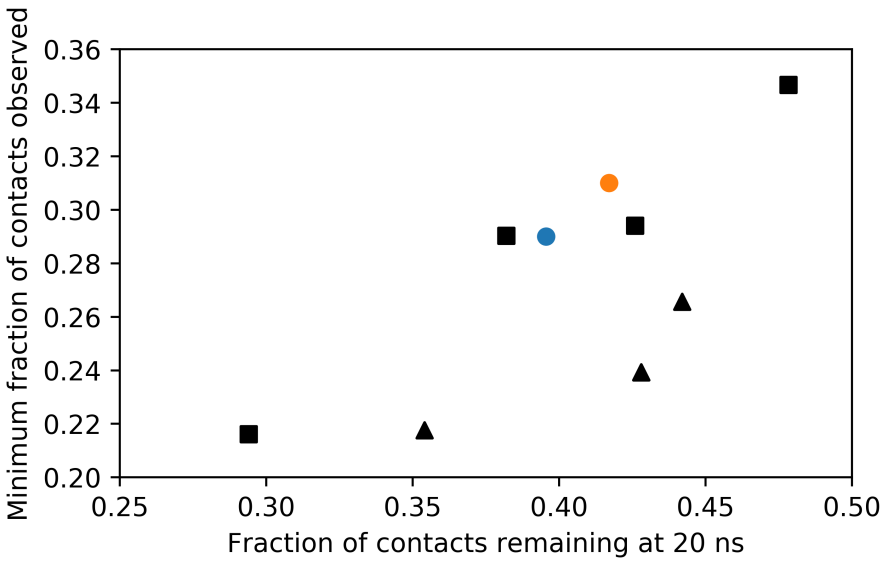


Figure 5.1: Average fraction of contacts remaining versus minimum fraction seen during production dynamics for each method using one 20 ns repeat for each mutation. Orange is the trimer average, blue the monomer average, and black the mutations. Stabilising methods are squares and destabilising methods are triangles.

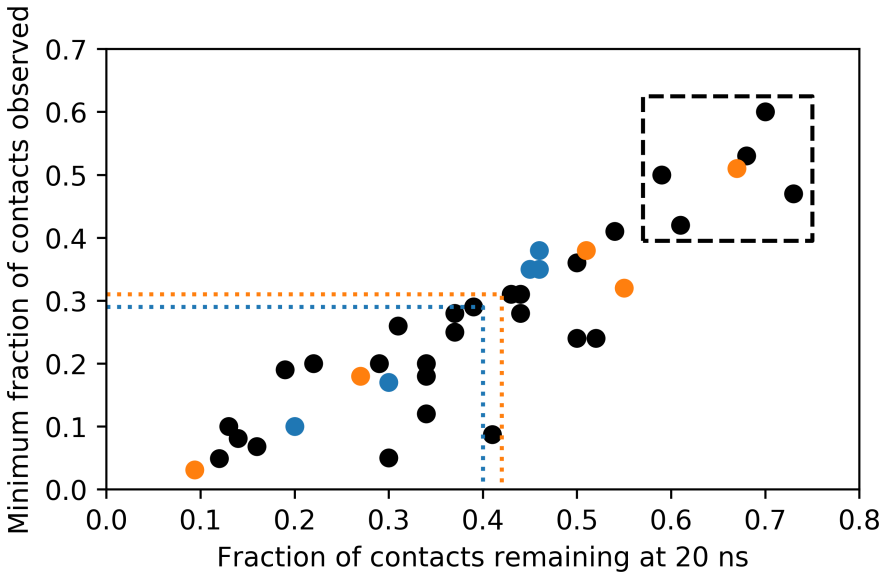


Figure 5.2: Fraction of contacts remaining versus minimum fraction seen during production dynamics for a 20 ns repeat for each mutation. Orange shows a trimer repeat, blue a monomer repeat, and black an average of two repeats for a individual mutation. Orange line shows average for trimer repeats, the blue line shows average of monomer repeats and the black box the region residues were taken forward.

Five results are clearly higher than the rest with a contact fraction above 0.6 after 20 ns, and are similar to the best run of the trimer. These are the mutations that show the most potential for allowing other pilins to bind. On the other end of the spectrum there are four that score very poorly at below 0.2 contact fraction remaining. These are more likely to be residues that are crucial for the maturation protein and F pilin to dock, and have to be conserved.

5.4 Promising Mutations

The five best mutations from the previous study were then repeated a further four times to see if the results were individual or would be consistent, as variance was seen between the different trimer and monomer results.

Ile114 scored the worst with a native contact final fraction of 0.42 and a minimum of 0.28 but this was not significantly different from the trimer average so mutating this residue should still be worthwhile. Val107 was next with an average final value of 0.45 and a minimum of 0.36, followed by Ser125 at 0.56 final value and 0.40 minimum and then Tyr339 at a final value of 0.55 and minimum of 0.42. Tyr102 performed the best after 5 repeats with an average final value of 0.60 and minimum of 0.43. It was the only mutation that even with the least number or residues in the interface at an average of 36, had the lowest RMSD for the interface at an average of 2.6 Å. It would be the first spot to consider to be changed to other residues to attempt to find other pilin docking sites. The repeats showed a large variance between results, however as not all of the repeats showed the docking failing, these mutations appear to be suitable for residue replacement.

5.5 Secondary Simulations

As it has been seen that results vary significantly between repeats, all mutations previously run for one 20 ns were run again to reduce the outliers. Individual mutations were averaged and the results are shown in 5.3

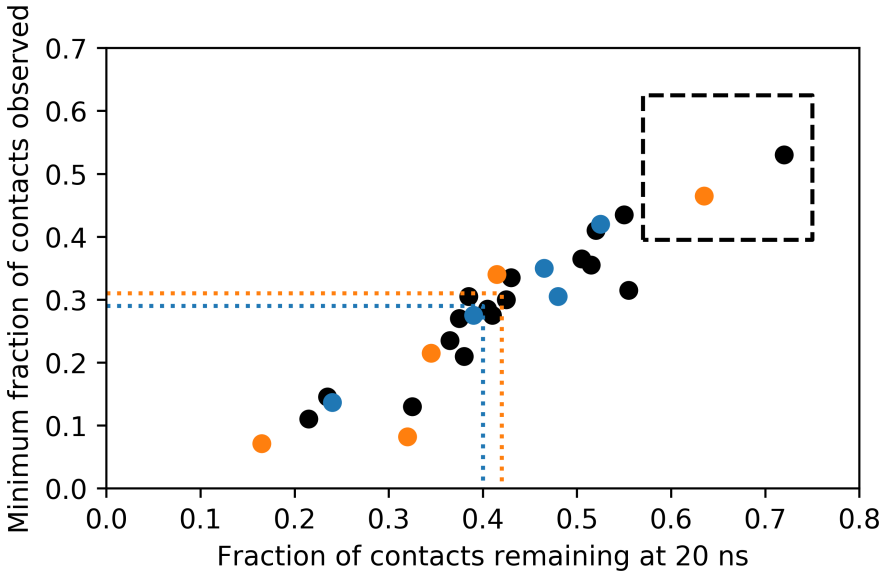


Figure 5.3: Fraction of contacts remaining versus minimum fraction seen during production dynamics for an average of two 20 ns repeats for each mutation. Orange shows a trimer repeat, blue a monomer repeat, and black an average of two repeats for a individual mutation. Orange line shows average for trimer repeats, the blue line shows average of monomer repeats and the black box the region residues were taken foward in Section 5.4

Only a single residue SER125 (0.72 fraction at end of 20 ns) remains within the region previously selected in Figure 5.2 after a second repeat. This mutation would be the first to change for attempting to dock other pilins. A further 11 residues have a native contact fraction minimum and end higher than that of the trimer average, so stabilise the binding when alanine. These could be mutated to other residues and hopefully increase the selectivity of the maturation protein. These are Tyr102 (0.64), Cys101 (0.56), Tyr339 (0.55), Val107 (0.53), Val323 (0.52), Ile114 (0.52), Trp341 (0.51), Ser362 (0.48), Val120 (0.47), His191 (0.43) and Val360 (0.43).

The averages for all of the methods shown in Figure 5.4 have now become more consistent, with the destabilising result being lower at 20 ns than the stabilising methods, which on average perform similarly or better than the monomer model. However not all the mutations as good as the trimer model due to the repeats of certain mutations having lower results than what was seen for the worst trimer repeat.

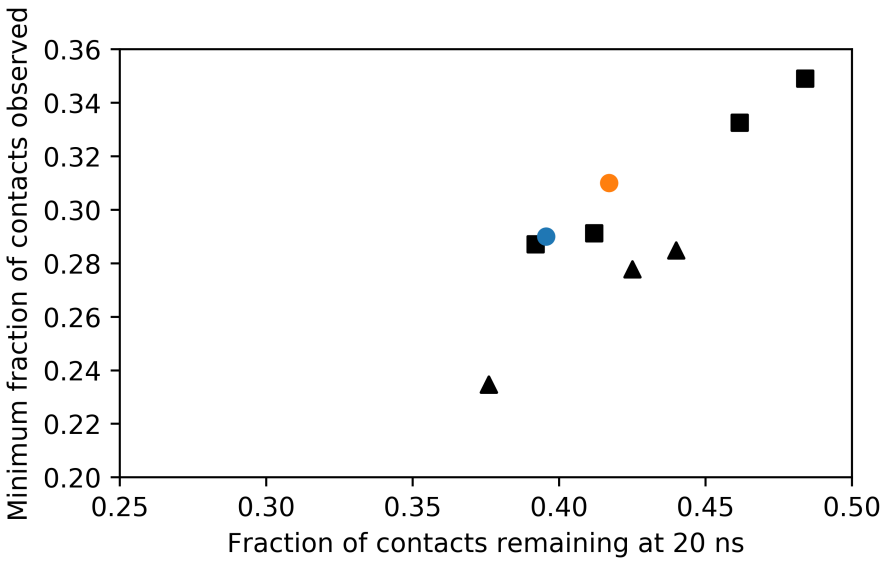


Figure 5.4: Average fraction of contacts remaining versus minimum fraction seen during production dynamics for each method using two 20 ns repeat for each mutation. Orange is the trimer average, blue the monomer average, and black the mutations. Stabilising methods are squares and destabilising methods are triangles.

Whilst trends are appearing for individual mutations on average little difference is seen for all stabilising and destabilising residues, with the average number of interfaces being 39 and 40 respectively, and average end fraction of native contacts being 0.41 and 0.42 respectively. Marginal difference is also seen when comparing hotspot mutations to non-hotspot mutations. The difference is favourable with stabilising hotspot mutations being higher at an average of 0.45 native contact fraction to 0.43 non-hotspot, and lower for destabilising hotspots at 0.40 end native contact fraction compared with 0.41 for the non-hotspot.

5.6 Crucial Residues

Of the seven residues that had contacts in all five trimer repeats in Section 4.2, two appeared in the alanine scanning runs, Ser103 and Asn122. Both were extremely low scoring, suggesting they needed to be present for the pilin and maturation protein to remain docked. Of the other five residues, two was alanine so a mutation was not required. The other three were simulated twice to see if they also were crucial to the binding. The three residues Val116, Arg243 and Pro247 all averaged at a higher end contact fraction than the av-

erage trimer result at 0.51, 0.47 and 0.49 respectively, which indicates that the residues are not required for successful binding of the pilin. This increases the validity of the alanine scanning, as even as the residues were seen in all repeats of both the monomer and trimer simulations. Those not selected by Robetta did not affect the stability of the complex.

In conclusion we have found that mutations to the maturation protein to alanine have been made in silico with both stabilising and destabilising affects. Certain residues are required for the interaction to remain such as Ser103 and Asn122, and others have little effect such as Val116, Arg243 and Pro247, whilst other mutations, in particular SER125 increased the stability of the complex when mutated and therefore make good targets for allowing other pilins to bind.

6. Conclusions

6.1 Summary

We have managed to replicate previous work, and successfully improved the model to include 3 subunits of the F pilus, instead of just one. Docking was successful with two programs HADDOCK and RosettaDock for both the monomer and trimer models, and the best orientation was found. HADDOCK docking was found to be produce more stable structures. Contacts made between the maturation protein and F pilin have been identified. Point mutations have been made and identified residues that could be changed to allow other pili to be attacked by the MS2 bacteriophage. Other residues have been found to be crucial to allow the F pilin to continue to bind successfully.

6.2 Future Work

This work has resulted in new pathways to be explored. The pilin model could be further extended for a more accurate model, at computational cost. Multiple mutations could be made at the same time to identify the combined effect and the detriment on the binding of F pilin, as just a single mutation reduces the average number of residues in the interface significantly. Other pili docking should be attempted such as the I plasmid-specific pili, to work out which residues the mutagen sites should be replaced with. This would extend the operational range of the MS2 bacteriophage, currently limited to binding with F plasmid-specific pili. Mutations would need to be optimised to allow docking with the I specific pili without compromising the action with F pilin. These mutations can then be taken forward towards experimental studies.

Bibliography

[1] World Health Organisation, *Global action plan on antimicrobial resistance.*, WHO Press, 2015, pp. 1–28.

[2] S. Sengupta, M. K. Chattopadhyay and H. P. Grossart, *Front. Microbiol.*, 2013, **4**, 1–13.

[3] L. J. V. Piddock, *Lancet Infect. Dis.*, 2012, **12**, 249–253.

[4] C. L. Ventola, *Pharm. Ther.*, 2015, **40**, 277–83.

[5] United States Centers for Disease Control, *Antibiotic resistance threats in the United States, 2013*, CDC, 2013, p. 28.

[6] M. Baym, L. K. Stone and R. Kishony, *Science*, 2016, **351**, aad3292.

[7] X. Wang, P. Zheng, T. Ma and T. Song, *Molecules*, 2018, **23**, 1307.

[8] R. E. Michod, H. Bernstein and A. M. Nedelcu, *Infect. Genet. Evol.*, 2008, **8**, 267–285.

[9] H. Brüssow and R. W. Hendrix, *Phage Genomics: Small is beautiful*, 2002.

[10] T. R. Costa, A. Ilangovan, M. Ukleja, A. Redzej, J. J. M. Santini, T. K. Smith, E. H. Egelman and G. Waksman, *Cell*, 2016, **166**, 1436–1444.e10.

[11] X. Dai, Z. Li, M. Lai, S. Shu, Y. Du, Z. H. Zhou and R. Sun, *Nature*, 2017, **541**, 112–116.

[12] K. Valegard, L. Liljas, K. Fridborg and T. Unge, *The three-dimensional structure of the bacterial virus MS2*, 1990.

[13] K. Toropova, G. Basnak, R. Twarock, P. G. Stockley and N. A. Ranson, *J. Mol. Biol.*, 2008, **375**, 824–836.

[14] A. E. Liana, C. P. Marquis, C. Gunawan, J. Justin Gooding and R. Amal, *J. Colloid Interface Sci.*, 2018, **514**, 227–233.

[15] C. B. Anfinsen, *Science*, 1973, **181**, 223–230.

[16] M. C. Bellissent-Funel, A. Hassanali, M. Havenith, R. Henchman, P. Pohl, F. Sterpone, D. Van Der Spoel, Y. Xu and A. E. Garcia, *Water Determines the Structure and Dynamics of Proteins*, 2016.

[17] W. Kauzmann, *Adv. Protein Chem.*, 1959, **14**, 1–63.

[18] R. L. Baldwin, *Proc. Natl. Acad. Sci. U. S. A.*, 2014, **111**, 13052–6.

[19] D. F. Stickle, L. G. Presta, K. A. Dill and G. D. Rose, *J. Mol. Biol.*, 1992, **226**, 1143–59.

[20] G. D. Rose and R. Wolfenden, *Annu. Rev. Biophys. Biomol. Struct.*, 1993, **22**, 381–415.

[21] R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, J. W. Labonte, M. S. Pacella, R. Bonneau, P. Bradley, R. L. Dunbrack, R. Das, D. Baker, B. Kuhlman, T. Kortemme and J. J. Gray, *J. Chem. Theory Comput.*, 2017, **13**, 3031–3048.

[22] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman, *J. Am. Chem. Soc.*, 1995, **117**, 5179–5197.

[23] W. L. Jorgensen and J. Tirado-Rives, *J. Am. Chem. Soc.*, 1988, **110**, 1657–1666.

[24] H. J. Berendsen, D. van der Spoel and R. van Drunen, *Comput. Phys. Commun.*, 1995, **91**, 43–56.

[25] S. L. Mayo, B. D. Olafson and W. A. Goddard, *J. Phys. Chem.*, 1990, **94**, 8897–8909.

[26] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, *J. Comput. Chem.*, 1983, **4**, 187–217.

[27] A. Haywood, *Masters thesis*, University of Nottingham, 2018.

[28] I. M. A. Nooren and J. M. Thornton, *Diversity of protein-protein interactions*, 2003.

[29] C. Chothia and J. Janin, *Nature*, 1975, **256**, 705–708.

[30] C. Dominguez, R. Boelens and A. M. Bonvin, *J. Am. Chem. Soc.*, 2003, **125**, 1731–1737.

[31] G. van Zundert, J. Rodrigues, M. Trellet, C. Schmitz, P. L. Kastritis, E. Karaca, A. Melquiond, M. van Dijk, S. de Vries and A. Bonvin, *J. Mol. Biol.*, 2016, **428**, 720–725.

[32] S. Chaudhury, M. Berrondo, B. D. Weitzner, P. Muthu, H. Bergman and J. J. Gray, *PLoS One*, 2011, **6**, e22477.

[33] K. C. Dent, R. Thompson, A. M. Barker, J. A. Hiscox, J. N. Barr, P. G. Stockley and N. A. Ranson, *Structure*, 2013, **21**, 1225–1234.

[34] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé and K. Schulten, *J. Comput. Chem.*, 2005, **26**, 1781–1802.

[35] B. R. Brooks, C. L. Brooks Iii, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York and M. Karplus, *J. Comput. Chem.*, 2009, **30**, 1545–1614.

[36] J. Huang and A. D. Mackerell, *J. Comput. Chem.*, 2013, **34**, 2135–2145.

[37] S. Jo, T. Kim, V. G. Iyer and W. Im, *J. Comput. Chem.*, 2008, **19**, 1859–1865.

[38] J. Lee, X. Cheng, J. M. Swails, M. S. Yeom, P. K. Eastman, J. A. Lemkul, S. Wei, J. Buckner, J. C. Jeong, Y. Qi, S. Jo, V. S. Pande, D. A. Case, C. L. Brooks, A. D. MacKerell, J. B. Klauda and W. Im, *J. Chem. Theory Comput.*, 2016, **12**, year.

[39] J. P. Ryckaert, G. Ciccotti and H. J. Berendsen, *J. Comput. Phys.*, 1977, **23**, 327–341.

[40] G. J. Martyna, D. J. Tobias and M. L. Klein, *J. Chem. Phys.*, 1994, **101**, 4177–4189.

[41] S. E. Feller, Y. Zhang, R. W. Pastor and B. R. Brooks, *J. Chem. Phys.*, 1995, **103**, 4613–4621.

[42] A. R. Leach, *Molecular Modelling - Principles and Applications*, Pearson Education, Harlow, 2nd edn., 2001, pp. 273–276.

[43] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem. Phys.*, 1983, **79**, 926–935.

[44] T. Darden, D. York and L. Pedersen, *J. Chem. Phys.*, 1993, **98**, 10089–10092.

[45] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane and V. S. Pande, *Biophys. J.*, 2015, **109**, 1528–1532.

[46] W. Humphrey, A. Dalke and K. Schulten, *J. Mol. Graph.*, 1996, **14**, 33–38.

[47] T. Kortemme, D. E. Kim and D. Baker, *Sci. Signal.*, 2004, **2004**, pl2.

[48] T. Kortemme and D. Baker, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 14116–14121.

[49] T. Clackson and J. A. Wells, *Science*, 1995, **267**, 383–386.

7. Appendices

7.A Adjusting preparation protocol

Initial preparation results

System preparation is a crucial part of running simulations, and ensuring that the model is as realistic as possible. On running the simulation for the first time, large increases in the RMSD of the complex were seen. The RMSD was calculated for all protein atoms that aren't hydrogen. The RMSD indicates how the complex is moving as an overall stucture, and sudden changes in RMSD could mean that initial contacts from the protein-protein docking are removed.

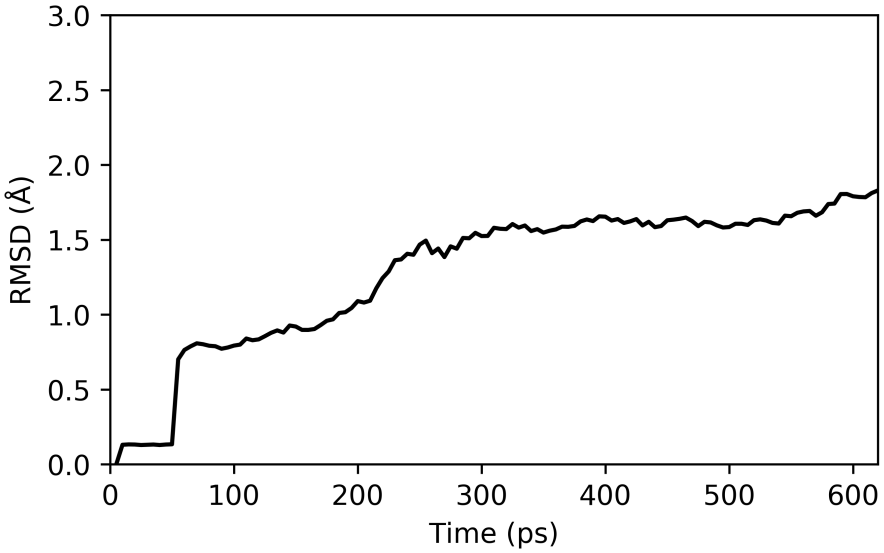


Figure 7.1: Initial RMSD of system preparation. Minimisation occurs before 20 ps, heating takes place between 20 ps and 120 ps, and equilibration between 120 ps and 620 ps. The majority of the perturbation occurs during the heating stage, with the RMSD increasing from 0.12 Å to 1.53 Å.

As a large quantity of the variation in RMSD was during the heating stage, with a large discontinuity between the minimisation and heating regimes, this was the area selected for improvement. The heating is done by slowly raising the temperature by an increment originally set as 3 K, with time spent at that temperature to equilibrate dependent on the total time allowed for heating, before increasing by a further 3 K and repeated until the final temperature is reached. Therefore the rate at which the heating occurs, and the time spent at each interval temperature can be controlled.

Improvements

Electrostatics

Initial efforts to improve the heating portion of system preparation were hampered by the temperature rapidly increasing between velocity resets. This resulted in the final temperature being greater than the wanted 298 K. This was isolated to being due to the electrostatics method employed; PME. By changing the interpolation order, in which the algorithm is calculated to 4 from 6 removes this issue. The value of 4 (cubic) is also the default for both CHARMM and NAMD.

Langevin dynamics

Keywords involving Langevin dynamics were removed from the minimisation regime due to the system being at 0 K. Both removal of keywords and setting the Langevin temperature to extremes resulted in no visual change to the trajectory of the complex.

Heating Increment

The heating stages were run for 50 000 and 100 000 steps to see how reducing the amount of heating, but with less time spent at each temperature would have an effect on the RMSD. It was found that the heating regime could be drastically improved upon by heating gradually, with the RMSD almost flatlining. However when the full setup was run, the RSMD dramatically increased after heating as the potential and kinetic energies equilibrated, even with heavy restraints, with the final RMSD being similar to runs with more rapid heating, as shown in Figure 7.2. Both 3 K and 0.3 K showed no discontinuity between heating and the next stage of equilibration, and similar final

RMSDs for several repeats, so 3K was chosen to allow for better comparison with work done by others previously.

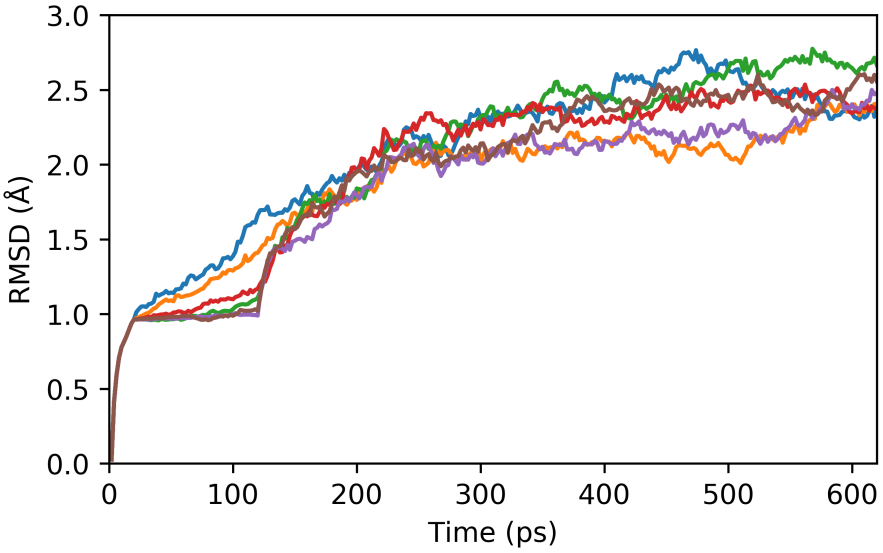


Figure 7.2: RMSDs for varying temperature increments between 0.006 K and 3 K during heating stage of preparation, after improvements have been made.
Key: 0.006 K (brown), 0.012 K (lilac), 0.03 K (red), 0.06 K K (green), 0.3 K (orange), 3 K (blue).

Summary

In conclusion, the modifications for preparation of the system, ready for production molecular dynamics has resulted in a smooth heating regime without artefacts and keeping a level of concordance with previous works to allow for direct comparison.

7.B Benchmarking

Due to the demanding nature of the simulations, benchmarking of the two available HPC clusters the local University of Nottingham High Performance Compute Service (Augusta) and the UK Tier 2 HPC Midlands Plus service (Athena). Augusta uses 40-core Intel Xeon Gold 6138 @ 2.0 GHz (dual-socket) nodes and Athena uses 28-core Intel Xeon E5-2680v4 @ 2.4 GHz (dual-socket) nodes. 5 ns of NAMD production simulation were run on 1 to 10 nodes on both services. The efficiency of NAMD means memory is not a consideration, with all runs using under 1 GB of RAM.

Augusta

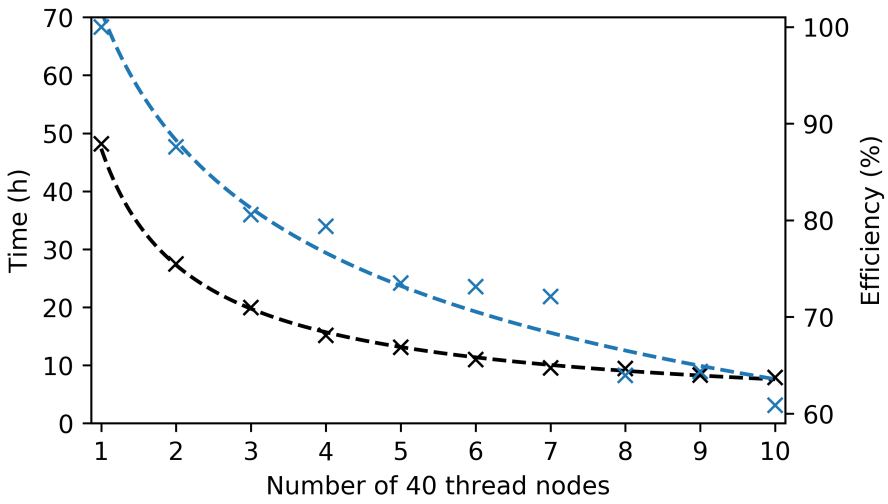


Figure 7.3: Time taken to run a 5 ns NAMD production simulation on the Augusta HPC (black). Efficiency of NAMD on multiple nodes of the Augusta HPC relative to the core hours used by a single node (blue).

A steady drop-off compared to ideal parallelisation is seen, with a speedup of 1.75 x for two nodes, 3.68 x for five nodes, and 6.09 x for ten nodes. There is a larger performance drop for use of more than seven nodes.

Athena

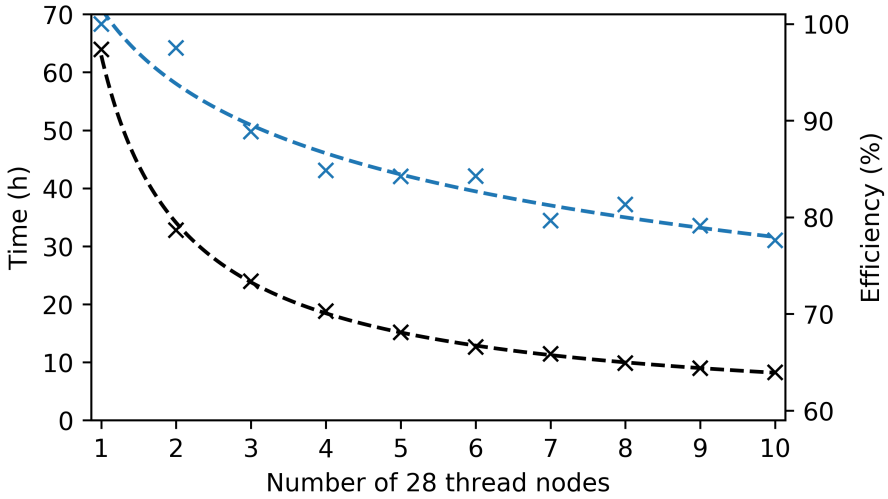


Figure 7.4: Time taken to run a 5 ns NAMD production simulation on the Athena HPC (black). Efficiency of NAMD on multiple nodes of the Athena HPC relative to the core hours used by a single node (blue).

A slower drop-off is seen for Athena, with a speedup of 1.95 x for two nodes, 4.21 x for five nodes, and 7.76 x for ten nodes. Use of two nodes is extremely efficient, with almost no penalty of parallelisation.

Speedup

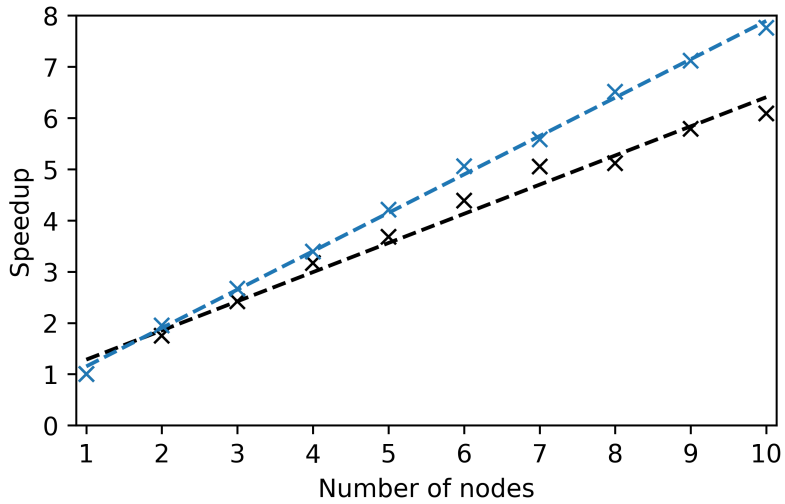


Figure 7.5: Speedup when using multiple nodes relative to a single node on the Augusta and Augusta HPCs.

Whilst similar for one and two nodes, the speedup of Augusta is substantially lower than that of Athena later on. This is due to the increased number of threads per node, hence a greater loss of speed due to communication between nodes, as for 10 nodes Augusta is using 400 cores to 280 cores used by Athena.

Trimer simulations

An expansion of the project is to look at simulating a larger proportion of the F pilin, by using multiple strands of the repeat unit rather than just one. which will increase the computational cost. Compared to the monomer single stranded simulation, a 5 ns trimer simulation took 1.17% longer on Augusta and 2.79% longer on Athena when using a single node. The increase is minute, as the bulk of the simulation cost is due to the explicit water box which is sized to 10 Å from the protein edge, which is not significantly changed by introduction of multiple strands of the pilin.

Summary

In conclusion, it appears that the Athena HPC is quicker, with greater speedup and higher efficiency when using multiple nodes, however as Augusta uses 40 core nodes and Athena 28 core nodes, the two are not comparable. Therefore it would be expected that communication between cores on Augusta takes more time. We cannot usually compare for the same number of cores as we would not be using full nodes on both HPCs, except for 7 nodes on Augusta and 10 nodes on Athena (280 cores). The comparative times suggest that Athena is 15.9% quicker, so when speed is of the essence, Athena should be used. This is to be expected with the higher clock speed of Athena Cores, but dosen't explain the greater efficiency, which is more likely due to the system architecture and compiling of NAMD.

7.C One-letter and three-letter codes for residues

Amino acid	Three-letter code	One-letter code
Glycine	GLY	G
Proline	PRO	P
Alanine	ALA	A
Valine	VAL	V
Leucine	LEU	L
Isoleucine	ILE	I
Methionine	MET	M
Cysteine	CYS	C
Phenylalanine	PHE	F
Tyrosine	TYR	Y
Tryptophan	TRP	W
Histidine	HIS	H
Lysine	LYS	K
Arginine	ARG	R
Glutamine	GLN	Q
Asparagine	ASN	N
Glutamic Acid	GLU	E
Aspartic Acid	ASP	D
Serine	SER	S
Threonine	THR	T

Table 7.1: one-letter and three-letter codes for each amino acid

7.D Activities undertaken towards Generic Training Program

- Diversity in Learning and Teaching
- Researcher skills and Endnote Training
- Research Integrity - Comprehensive
- Self directed learning in Python, CHARMM and NAMD.
- LaTeX for researchers; Introduction, and Further LaTeX for researchers
- Presentation at CATC@N seminar
- Attended 22 CATC@N seminars, 4 Theme seminars, 2 school colloquia, Dan Eley symposium and Postgraduate symposium.
- Attended Digital Research Compute Day, Nottingham and Midlands Computational Chemistry Conference, Loughborough.
- Attended two day workshop on Getting started with biomolecular simulations, Leeds.