

**Structural characterization of
flagellin from *E. coli* Nissle,
glycoprotein C from herpes simplex virus type 1,
and a nanobody-peptide tag system**

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Dipl.-Chem. Michael B. Braun

aus Lörrach

Tübingen

2021

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:

26.10.2021

Dekan:

Prof. Dr. Thilo Stehle

1. Berichterstatter:

Prof. Dr. Thilo Stehle

2. Berichterstatter:

Prof. Dr. Dirk Schwarzer

Abstract

The thesis presented here deals with four independent projects, which are briefly summarized in the following paragraphs.

Flagella enable bacteria to freely swim in solution granting them the motility to reach resourceful areas or to avoid an unfavorable environment. Therefore, the ability to produce a flagellum, is a major fitness factor for bacteria. Pathogens might rely on flagella to reach host surfaces and penetrating them. From a host point of view, flagella or their major protein FliC render an opportunity to detect bacteria and regulate a response appropriate to the bacterial threat. The regulation of the response can be dysfunctional, which can lead to inflammatory bowel disease. *E. coli* Nissle, a commensal showed positive effects in inflammatory bowel disease mouse models which partially depend on the major flagella protein FliC. We crystallized and solved the structure of the *E. coli* Nissle FliC. By comparison to known FliCs, we were able to identify two new structural features. First, the hypervariable region consists of three distinct domains and second, the hypervariable region is connected via a long linker to the constant region. Modeling the flagellum filament showed that the hypervariable regions might interact with each other forming an outer structure, which is connected via the linkers to the conserved flagellum core structure. Modeling the interaction with the immune receptor TLR5 that recognizes the conserved domain 1 of FliC showed that the domain should be well accessible and not sterically hindered by the three hypervariable domains. The dimerization of TLR5 with bound FliC to the active complex should also not be hindered by the hypervariable region.

Glycoprotein C (gC) is an attachment factor of Herpes simplex virus 1 that binds to heparan sulfate (HS) which is exposed on the cell surface of the host. This non-essential viral attachment protein allows lateral virus diffusion along cell surface and supports attachment of other viral proteins to their receptors on the cell surface, which ultimately results in membrane fusion or endocytotic uptake of the virus. The structure of gC was solved by experimental phasing and shows that gC

consists of three distinct domains and presumably forms a dimer. We identify the putative heparan sulfate binding area ranging from the middle to the N-terminal domain. Another function of gC, the protection from the complement system, needs further investigation. Here, the structure might help to design experiments to investigate how gC binds C3b and prevents further interaction of C3b with the complement system. The gC structure from HSV-1, solved in this work, can serve as model for the closely related HSV-2 gC (Identity 73%) and might help to explain differences in HS binding and C3b affinity.

We invented a nanobody (Nb)-based peptide tag system, which can be used for protein purification or to label proteins for super resolution microscopy. The nanobody binds a twelve amino acid long peptide with a low nanomolar affinity. We solved the structure Nb-peptide complex to explain the observed specificity for different amino acid sequences. We identified an essential tryptophan residue in the peptide sequence located in a deep hydrophobic pocket of the Nb, at another position basic residues are preferred. Some sidechains point towards the Nb and have to be rather small while other pointing away from the Nb with no preference for certain residues. The peptide is bound between the framework region and complementarity-determining region 3 (CDR3) of the Nb, mainly by backbone-backbone interactions.

In a side project, we contributed to the development of an adenovirus based vector system. The vector is based on the rare adenovirus serotype 43 with a low antibody prevalence in humans. The adenovirus was modified with affibodies changing the primary receptor to the ocotarget “epidermal growth factor receptor type 2” (Her2).

Zusammenfassung

Die hier vorgestellte Arbeit umfasst vier unabhängige Projekte, deren Ergebnisse in den folgenden Abschnitten zusammengefasst sind.

Flagellen ermöglichen Bakterien sich in Lösungen freischwimmend fortzubewegen, dadurch erhalten sie die Beweglichkeit um ressourcenreiche Bereiche zu erreichen oder eine unvorteilhafte Umgebung zu vermeiden. Aus diesem Grund ist die Fähigkeit eine Flagelle zu produzieren ein wichtiger Fitnessfaktor für Bakterien. Pathogene können auf die Flagelle angewiesen sein um Wirtsoberflächen zu erreichen und um sie zu durchdringen. Aus dem Blickwinkel des Wirtes, stellen Flagellen oder ihr Hauptproteinbestandteil FliC eine Möglichkeit dar, die Bakterien zu detektieren und eine angemessene Antwort auf die bakterielle Bedrohung zu initiieren. Die Regulierung dieser Reaktion kann dysfunktional sein, was zu chronischen, entzündlichen Darmerkrankungen führt. *E. coli* Nissle ein Kommensale zeigte in Mausmodellen positive Effekte bei chronisch entzündlichen Darmerkrankungen, welche teilweise auf dem Hauptflagellen Protein FliC beruhen. Wir kristallisierten und lösten die Struktur von *E. coli* Nissle FliC. Durch den Vergleich mit bekannten FliC Strukturen, konnten wir zwei neue strukturelle Eigenschaften identifizieren. Erstens, die hypervariable Region besteht aus drei unterscheidbaren Domänen und Zweitens, die hypervariable Region ist durch einen langen Linker mit der konstanten Region verbunden. Das Modellieren des Flagellen-Filaments zeigte, dass die hypervariablen Regionen mit einander interagieren und eine äußere Struktur formen könnten, welche über den Linker mit der konservierten Kernstruktur der Flagelle verbunden ist. Das Modellieren der Interaktion mit dem Immunrezeptor TLR5, welcher die konservierte Domäne 1 von FliC erkennt, zeigte, dass diese Domäne gut zugänglich sein sollte und nicht sterisch durch die drei hypervariablen Domänen gehindert werden sollte. Die Dimerisation von TLR5 mit gebundenem FliC zum aktiven Komplex sollte auch nicht durch die hypervariable Region behindert werden.

Glykoprotein C (gC) ist ein Bindungsfaktor von Herpes Simplex Virus 1, welcher Heparansulfat (HS) bindet, das an der Wirtsoberfläche exponiert ist. Dieses nicht essentielle virale Bindepotein erlaubt laterale Diffusion entlang der Zelloberfläche und unterstützt das Binden anderer viraler Proteine an ihre Rezeptoren, was letztlich zu einer Membranfusion oder der endozytotischen Aufnahme des Virus führt. Die Struktur von gC wurde durch experimentelles Phasieren gelöst und zeigte das gC aus drei abgetrennten Domänen besteht und vermutlich ein Dimer bildet. Wir identifizierten den Bereich der mutmaßlichen Heparansulfatbindung, welcher von der mittleren zur N-terminalen Domäne reicht. Eine weitere Funktion von gC, der Schutz vor dem Komplementsystem, benötigt weitere Untersuchungen. Hier könnte die Struktur helfen Experimente zu entwerfen, um herauszufinden wie gC C3b bindet und so weitere Interaktionen von C3b mit dem Komplementsystem verhindert. Die in dieser Arbeit gelöste gC Struktur von HSV-1 kann als Modell für den nahe verwandten HSV-2 (Identität 73%) dienen und könnte dabei helfen Unterschiede in HS Bindung und C3b Affinität zu erklären.

Wir haben einen Nanobody (Nb) basiertes Peptidetag System etabliert, welches zur Proteinreinigung oder zum Markieren von Proteinen in der Super-Resolution Mikroskopie verwendet werden kann. Der Nanobody bindet ein zwölf Aminosäuren langes Peptid mit niedriger nanomolarer Affinität. Wir haben die Struktur des Nb-Peptid-Komplexes gelöst, um die beobachtete Spezifität für verschiedene Aminosäuresequenzen zu erklären. Dabei konnten wir einen essentiellen Tryptophanrest in der Peptidsequenz identifizieren, welcher sich in einer tiefen hydrophoben Tasche des Nb befindet, an einer anderen Position sind basische Reste bevorzugt. Manche Seitenketten zeigen in Richtung des Nb und müssen eher klein sein, während andere vom Nb wegzeigen und damit keine Präferenz für einen speziellen Rest besitzen. Das Peptid ist zwischen der „Framework“ Region und der komplementaritätsbestimmenden Regionen 3 des Nb hauptsächlich über Backbone-Backbone Interaktionen gebunden.

In einem Nebenprojekt, waren wir beteiligt an der Entwicklung eines Adenovirus basierendem Vektorsystem. Der Vektor basiert auf dem seltenen Adenovirus Serotyp 43 mit einer niedrigen Prävalenz in Menschen. Der Adenovirus wurde mit Affibodies modifiziert, um den primären Rezeptor auf das Oncotarget „Epidermaler Wachstumsfaktor Rezeptor Typ 2“ (Her2) zu verändern.

Table of Contents

Abstract	I
Zusammenfassung	III
Table of Contents	VI
Abbreviations	XIII
I. Crystallography	1
1. Structural Biology.....	1
2. X-ray Crystallography	2
3. Crystallization	9
4. Crystal Symmetry	13
5. Data Quality	17
6. Radiation Damage	22
7. Electron Density Reconstruction	24
8. Solving the Phase Problem.....	26
8.1. Direct Methods	27
8.2. Molecular Replacement.....	30
8.3. Single Isomorphous Replacement (SIR).....	32
8.4. Single-Wavelength Anomalous Dispersion (SAD)	35
8.5. Marker Atom Substructure.....	38
9. Phase Improvement.....	40
9.1. Phase Combination.....	42
10. Structure Refinement	46

II. Flagellin from <i>E. coli</i> Nissle	51
1. Introduction	51
1.1. Intestinal Immune System	51
1.2. Inflammatory Bowel Disease	52
1.2.1. IBD Treatment	52
1.2.2. Flagellin Detection	53
1.3. Toll Like Receptors (TLRs).....	54
1.3.1. TLR-MAMP Interaction	54
1.3.2. TLR Signaling	57
1.3.3. TLR5 and IBD.....	59
1.4. Chemotaxis	60
1.5. Flagellum Components	63
1.6. Flagellum Growth	63
1.7. R-Type and L-Type Flagellum	66
2. Objectives	68
3. Materials and Methods	69
3.1. Materials and Buffers	69
3.1.1. Chemicals.....	69
3.1.2. Vector	69
3.1.3. Bacterial Strains	69
3.1.4. Buffers	70
3.1.5. Commercial Crystallization Screens	71
3.2. Molecular Biology.....	71
3.2.1. Preparation of Chemically Competent Cells.....	71
3.2.2. Transformation of Competent Cells	71

3.2.3.	Glycerol Stocks.....	72
3.2.4.	Plasmid DNA Isolation	72
3.2.5.	Site Directed Mutagenesis	73
3.2.6.	DNA Sequencing	74
3.3.	Protein-Biochemistry	75
3.3.1.	Protein Production	75
3.3.2.	Cell Lysis	75
3.3.3.	Immobilized Metal Ion Affinity Chromatography (IMAC)	76
3.3.4.	Dialysis and Proteolytic Cleavage	76
3.3.5.	SDS-Polyacrylamide Gel Electrophoresis (SDS-PAGE).....	77
3.3.6.	Size Exclusion Chromatography (SEC)	78
3.3.7.	Chemical Crosslinking	78
3.3.8.	Chemical Sidechain Modification	79
3.4.	Biophysical Methods	80
3.4.1.	Protein Concentration Determination	80
3.4.2.	Circular Dichroism (CD) Spectroscopy.....	80
3.4.3.	Thermal Shift Assay.....	80
3.5.	Structural Biology	81
3.5.1.	Precipitation Test.....	81
3.5.2.	Crystallization	81
3.5.3.	X-ray Diffraction Data Collection	81
3.5.4.	Soaking of Crystals with Heavy Atoms.....	82
3.6.	Software.....	83
4.	Results	84
4.1.	Purification of FliC Constructs	84

4.2.	Crystallization of FliC Δ D01	85
4.3.	FliC Δ D01 Phase Determination	86
4.4.	Crystallization of FliCD12	88
4.5.	FliCD12 Phase Determination	89
4.6.	Model of FliC Δ D0	90
4.7.	Structure of the Constant D1	91
4.8.	Transition of D1 to D2	92
4.9.	Structure and Transition of D2 and D3	93
4.10.	Structure and Transition of D3 and D4.....	95
4.11.	The β -Triangle	96
4.12.	Flexibility Between Domains	97
4.12.1.	FliC D12 Structure Alignment	97
4.12.2.	FliCD12 and FliCD234 Structure Alignment	98
4.13.	Modeling the Flagellum.....	99
4.13.1.	Monomer Interactions within the Flagellum	101
4.13.2.	The Flagellum Diameter	103
4.14.	FliC Stability	104
4.15.	FliC Δ D4 and FliC Δ D34 Purification.....	106
4.16.	Sequence Alignment of Flagellins.....	109
4.16.1.	The Conserved Region	111
4.16.2.	The Hypervariable Region.....	113
5.	Significance	115
6.	Discussion	116
6.1.	Sequence Comparison of Flagellins	116
6.1.1.	H-type Domain Organization.....	118

6.2.	D4 Deletion Mutant	119
6.3.	Flexibility Between Domains.....	120
6.4.	Model of the EcN Flagellum	121
6.4.1.	Flagellum Hypervariable Region Orientation.....	123
6.4.2.	The Wheel-Like Flagellum Model.....	124
6.5.	FliC Dimerization.....	126
6.6.	FliC TLR5 Interaction	128
III.	Glycoprotein C from HSV-1	131
1.	Introduction.....	131
1.1.	Herpesviridae	131
1.2.	HSV Pathogenesis	131
1.2.1.	The Lytic and Latent Phases	132
1.3.	HSV-1 Architecture	132
1.4.	HSV-1 Envelope Proteins and the Entry.....	133
1.5.	Glycoprotein C	135
1.5.1.	A Heparan Sulfate Receptor	135
1.5.2.	A C3b Receptor	137
1.5.3.	Shielding.....	138
2.	Objectives	139
3.	Materials and Methods	140
3.1.	Materials and Buffers	140
3.1.1.	Chemicals.....	140
3.1.2.	Cell line.....	140
3.1.3.	Buffers	140

3.2.	Protein-Biochemistry	141
3.2.1.	Protein Production	141
3.2.2.	Purification.....	141
3.3.	Structural Biology	142
3.3.1.	Crystallization	142
3.3.2.	Soaking of Crystals with Heavy Atom Derivatives	142
3.3.3.	X-ray Diffraction Data Collection	143
3.3.4.	Software	143
4.	Results	144
4.1.	Purification of gC.....	144
4.2.	gC Phase Determination	145
4.3.	gC Structure.....	146
4.3.1.	gC-Domain1	148
4.3.2.	Transition between D1 and D2	150
4.3.3.	gC-Domain2	151
4.3.4.	Transition between D2 and D3	152
4.3.5.	gC-Domain 3	153
4.4.	gC-Glycosylation	155
4.5.	gC-Dimer.....	157
5.	Significance	163
6.	Discussion	164
6.1.	The Heparan Sulfate Binding Site	164
6.1.1.	Important Residues for HS Binding.....	164
6.1.2.	Putative HS Binding Area	166
6.1.3.	The Dimerisation and HS Binding.....	168

6.2.	Shielding from immune recognition	170
6.3.	C3b Binding Regions.....	171
6.4.	Sequence Comparison with gC from HSV-2.....	173
IV.	Peptide Binding Nanobody BC2	176
1.	Introduction.....	176
1.1.	Antibodies and Antibodies Fragments	176
1.2.	Nanobody Production.....	178
1.3.	Unique Features of Nanobodies.....	178
1.4.	Super Resolution Microscopy as Application Field	180
2.	Results	181
2.1.	Structural Characterization	181
2.2.	Application	182
3.	Developments and Outlook.....	183
V.	Adenovirus Serotype 43 Vector	185
1.	Introduction.....	185
1.1.	Cancer	185
1.2.	Cancer treatment	186
2.	Results	187
VI.	References.....	189
VII.	Appendix.....	203
VIII.	Acknowledgements.....	252
IX.	Publications.....	253

Abbreviations

Amp	Ampicillin
APS	Ammonium peroxydisulfate
bp	Base pairs
DMAB	Dimethylaminoborane
<i>E. coli</i>	<i>Escherichia coli</i>
EDTA	Ethylenediaminetetraacetic acid
Gy	Gray (J kg ⁻¹); absorbed dose
IPTG	Isopropyl- β -D-thiogalactopyranosid
λ	Wavelength
o/n	Over night
OD ₆₀₀	Optical density or absorbance at a wavelength of 600 nm
PEG	Polyethylene glycol
RT	Room temperature
SDS	Sodium dodecyl sulfate
TEMED	N,N,N',N'- Tetramethylethylenediamine
TEV	Tobacco Etch Virus
Tris	2-Amino-2-(hydroxymethyl)propan-1,3-diol
wt	Wild type
% (v/v)	Percentage by volume
% (w/v)	"Percentage" by weight per volume (mass concentration)
k	Boltzmann constant 8.617×10^{-5} eVK ⁻¹

I. Crystallography

1. Structural Biology

The ambition of structural biology is to explain function and properties of biological matter on their molecular structure. There are three major techniques to gain structural information of biomolecules at atomic resolution: Nuclear magnetic resonance (NMR) spectroscopy, cryo-electron microscopy (cryo-EM), and X-ray crystallography. NMR investigates the energy differences in orientation of nonzero nuclear spin to a strong external magnetic field, thereby gaining information about the local magnetic surrounding and coupling of atomic spins in the near surrounding. Information about distance restraints between atoms are acquired and models fulfilling these restraints can be generated. The other two techniques have more in common with “classical” light microscopy where the used wavelength dictates the reachable resolution (Abbe limit). The wavelength of the radiation is in the range of interatomic distances and the major differences between the technics arise from the physical properties of the used radiation. Electromagnetic waves with such a small wavelength are X-rays (0.1-100 Å), where a wavelength of about 1 Å (12.3984 keV) is typically used in X-ray diffraction experiments. A suitable low wavelength can also be achieved with particles possessing a rest mass such as electrons or neutrons used in neutron-scattering, electron-diffraction, or cryo-EM. The difference between the methods arise through the strength of interaction with matter of the used radiation and how it is generated, manipulated and detected. The atomic cross-section for electrons (300 keV) is about 10^5 higher than for X-rays (12.3984 keV), therefore the probability of interaction is much higher. This enables the measurement of single particles but with a low radiation dose as ionizing inelastic events causing radiation damage and background noise have an even higher probability. Due to their charge, electron waves can be manipulated with magnets allowing to focus an electron beam on a sample and detector. Generation of an electron beam at a specific energy is comparatively easy using an electron emitter and an electric field to accelerate the produced electrons to the

desired energy. The model is built in the end in an electron density map that is reconstructed in 3D from a large number of 2D single particle images. The measured particles are randomly orientated, which allows the 3D reconstructions from different measured 2D projections.

Neutron and X-ray crystallography are very similar on the experimental setup. Both radiation types are not charged, and through the weak interaction with matter no useful lenses exist to refocus an image and the brilliance of existing sources is not sufficient to measure single molecules. Therefore, the signal needs to be amplified by using crystals of the molecules of interest. Neutron diffraction remains a niche method for visualizing hydrogens. In contrast to X-rays, neutrons are scattered at the nucleus and not at electrons. The atomic cross-section for neutron waves is small and for their generation a nuclear fission reactor (e.g. FRM II, MLZ, München) or a spallation neutron source (e.g. SINQ, PSI, Villigen) is needed, limiting the availability of this structure determination method. In contrast, laboratory or wavelength tunable synchrotron X-ray sources are widely distributed and allow to obtain high resolution data, which renders X-ray crystallography the most used method for structure determination today. In 2019, 9659 out of 11512 structures were determined by X-ray crystallography [1] (www.rcsb.org).

2. X-ray Crystallography

The atomic cross-section for elastic scattering of X-rays is small, this has the consequence that only 0.1-1% of all incoming photons are scattered. To address this, bright X-ray sources (e.g. 5×10^{11} ph/s PXIII, PSI, Villigen) and symmetrically ordered molecules forming a three dimensional (3D) lattice (crystals) are used to enhance the diffraction signal. How photons are scattered, by the electrons of an atom, depend on the wave functions (solutions of the Schrödinger equation) of every electron. The atomic scattering factor (f_S) can be calculated for every scattering vector ($\mathbf{S} = \mathbf{s}_{1(\text{scattered})} - \mathbf{s}_{0(\text{incoming})}$), if the electron density (ρ_r) of a given volume (V_{atom}) is known (1). In other words, the atomic scattering factor (f_S) describes how the electrons at every point of a volume contribute to the resulting scattering as a function of scattering direction.

$$(1) \quad f_S = \int_r^{V_{atom}} \rho_r \exp(2\pi i \mathbf{S} \mathbf{r}) dr$$

The scattering function of a molecule is the sum of all individual atomic scattering functions using a common origin. These functions are complex function with an amplitude and a phase, so phase differences have to be considered in the summation (2). In other words, the scattering functions of every atom are superimposed, resulting in a total scattering function of the molecule.

$$(2) \quad F_S = \sum_j^{atoms} f_{s,j} \exp(2\pi i \mathbf{S} \mathbf{r}_j)$$

In a crystal, protein molecules are packed in an ordered fashion. A crystal can be described as a representative set of molecules in a defined volume (unit cell) which is translated (u, v, w times in the corresponding direction) along its axes (a, b, c).

While the unit cell can be described with equation (2) summing over all atoms in the unit cell, the symmetric repetitive alignment of the unit cell leads to superposition of the unit cell scattering functions (3).

$$(3) \quad F_S^{cryst} = F_S^{cell} \sum_{u=0}^{u-1} \exp(2\pi i(uS\mathbf{a})) \sum_{v=0}^{v-1} \exp(2\pi i(vS\mathbf{b})) \sum_{w=0}^{w-1} \exp(2\pi i(wS\mathbf{c}))$$

For a large number of unit cells typical for a crystal (e.g. there is space for 10^{12} cubic unit cells with an edge length of 100 \AA in a 100 \mu m cubic crystal), the continuous scattering turns into a discrete diffraction pattern due to the sampling of the lattice. A diffraction image can be interpreted as the reciprocal space amplitude of the Fourier transform of the unit cell electron density sampled on the reciprocal space Fourier transform of the lattice. Consequently, the structure factors can be described in a lattice based coordinate system (4), where x_j, y_j, z_j are fractional coordinates (relative lengths of the basis vector system of the unit cell) of atom j to the origin.

$$(4) \quad F_{hkl} = \sum_j^{atoms} f_j \exp(2\pi i (hx_j + ky_j + lz_j))$$

The so-called Miller indices $[hkl]$ represent the direction of scattering. To understand how the discrete diffraction pattern emerges from the lattice, the diffraction process can be simplified as a reflection at a set of coplanar planes perpendicular to the scattering vector \mathbf{S} (Figure 1).

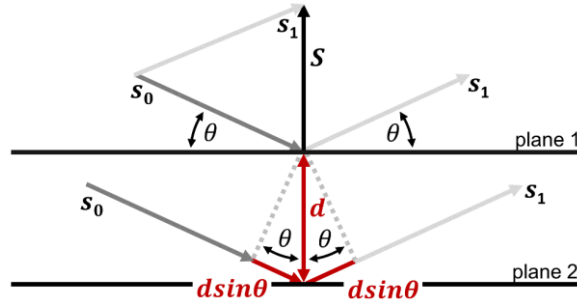


Figure 1: Diffraction simplified as reflection event. The scattering vector \mathbf{S} is perpendicular to a set of coplanar imaginary planes with a distance d . The incidence angle θ of reflected waves with direction \mathbf{s}_0 is equal to the emergent angle θ of the waves with direction \mathbf{s}_1 . A wave reflected at a parallel plane has an additional path length of $2d \sin \theta$.

The path length difference between waves reflected at a plane and a coplanar plane is $2d \sin \theta$. Where d is the distance between the planes and θ is the incidence and reflected angle with respect to the plane. This has two consequences. Firstly, waves reflected at coplanar planes with different distances d have different phase angles, which results in destructive interference. On the other hand, sets of collinear planes with the same distance d can interfere constructively. Secondly, the path difference $2d \sin \theta$ for the wave must be an integer multiple n of the wavelength λ of the incoming photon to result in constructive interference (5)(Bragg's law) [2].

$$(5) \quad n\lambda = 2d_{hkl} \sin \theta$$

The directions in which repetitive patterns in a crystal exist are defined by the unit cell. Every set of planes that intercepts the axis of a unit cell at rational numbers is linked to the translation of unit cell that builds up the crystal (interceptions with irrational numbers are not related to unit cell translation and therefore cannot lead to constructive interference as given by Bragg's Law). How often the unit cell is intercepted by a given set of collinear planes is related to the resolution of

information the corresponding diffraction condition contains. As an example the reflection condition with Miller index $[0,1,0]$ would intercept once with unit cell axis b . Consequently, the information that this diffraction condition contains, is a wave with the wavelength b . A reflection with two interceptions $[0,2,0]$ would result in a wavelength of $b/2$. The smaller the plane distance gets, the higher the resolution of the corresponding diffraction condition. Taken into account equation (5), the diffraction angle increases the smaller the plane distances are, or the higher the miller index is. The transformation to the reciprocal space with a common origin simplifies the calculations, as the rational interceptions of the unit cell can be expressed as lattice with integer numbers. The relationship of the real and reciprocal lattice is given by the equations (6) and (7).

$$(6) \quad a^* = \frac{b \times c}{V} \quad b^* = \frac{a \times c}{V} \quad c^* = \frac{a \times b}{V}$$

$$(7) \quad d_{hkl}^* = ha^* + kb^* + lc^*$$

The Miller indices $[hkl]$ represent the reciprocal lattice points and V the volume of the unit cell. The distances to the origin d_{hkl}^* represent plane distances in reciprocal space.

A diffraction experiment can be understood by the so-called Ewald construction (Figure 2), which can be briefly described as a different representation of Bragg's law. Here, a sphere with a radius of $1/\lambda$ (the so-called Ewald sphere) is drawn around the crystal. The reciprocal lattice originates where the beam intersects with the Ewald sphere. The reciprocal lattice points can be found at distances d_{hkl}^* from the index origin, where each distance represents a set of parallel real space planes. The origin of the real and reciprocal lattice is the same as the reciprocal lattice is nothing else but a different mathematical description of the same real space lattice. If a reciprocal lattice point intercepts with the Ewald sphere, the Bragg diffraction condition is fulfilled and a diffraction spot can be recordable on the detector. The scattered photon and its diffraction angle is described by the

vector connecting the real space origin with the reciprocal lattice point that intersects with the Ewald sphere. During a diffraction experiment the crystal and therefore the reciprocal lattice is rotated, which ultimately results in the intersection of different reciprocal lattice points with the Ewald sphere. Finally, this results in a set of recorded reflections described by the Miller indices, measured intensities and intensity errors, the so-called data set that is used for structure determination.

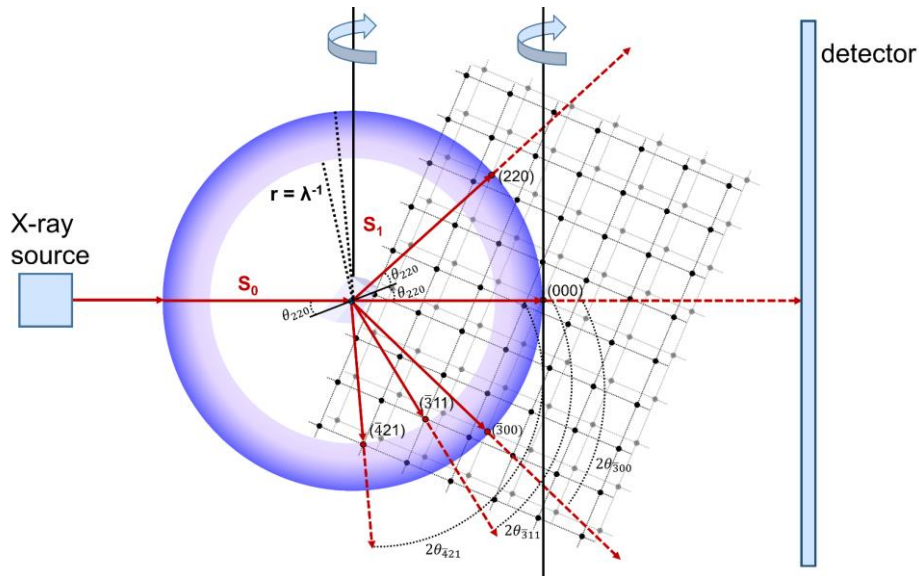


Figure 2: Ewald construction. An X-ray source (left) emits X-ray photons that hit a crystal and the detector (right). Around the crystal a spherical segment of the Ewald sphere with a radius of $1/\lambda$ is shown in blue. Two reciprocal lattice planes ($l = 0$ grey, $l = 1$ black) perpendicular to the beam are shown. When a reciprocal lattice point intercept with the Ewald sphere (red lattice points) the Braggs law is fulfilled resulting in a diffracted photon (dotted red lines) which can be recorded on the detector. The direction of each scattering vector s_1 can be drawn from the center of the Ewald sphere to the intersection point. During data collection the crystal and therefore the reciprocal lattice rotates and different lattice points intersect with the Ewald sphere. The diffraction angle θ for each reflection is half the angle to the beam center (000).

3. Crystallization

As discussed in chapter 2, structure determination by X-ray crystallography depends on the availability of crystals of the investigated molecules. Hence, crystallization of macromolecules is the limiting step to provide structure data. Due to protein stability, folding, flexibility and inhomogeneity crystallization can be challenging or even impossible. In general, crystals can form and grow in super-saturated solutions. In protein crystallography, different techniques (vapor diffusion, micro-batch, micro-dialyses, capillary counter diffusion and lipidic cubic phase) are used to achieve super-saturation (**Figure 3**).

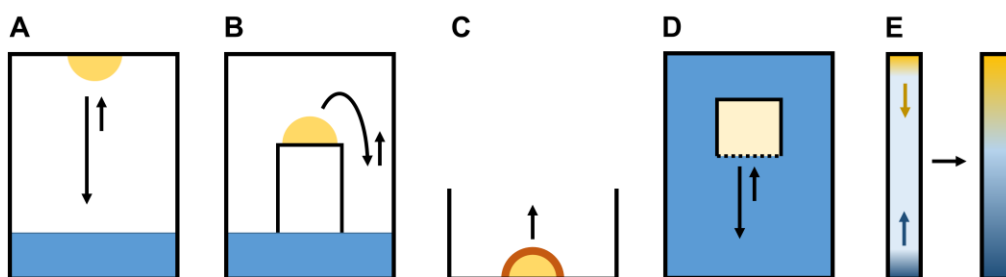


Figure 3: Crystallization techniques. **A:** Hanging drop vapor diffusion, a protein precipitation mixture (yellow) is placed in a sealed chamber “upside down” over a precipitate solution (blue). The net water transportation (indicated with black arrows) is from the drop to the precipitate solution. **B:** Sitting drop vapor diffusion, same as (A) but the protein precipitate mixture is placed on top of a surface. **C:** Micro-batch. A protein precipitation drop is covered with an oil (red) which slows down diffusion in an open system. **D:** micro-dialyses. A protein solution is placed separated by a dialyze membrane in a precipitation solution. **E:** capillary counter diffusion. A protein solution is placed on one side of a capillary and a precipitate solution on the other side. The diffusion towards each other can be slowed down by using a gel matrix (e.g. agarose).

The vapor diffusion technique is the most often used crystallization procedure applied for macromolecules. Here, a highly concentrated protein sample is mixed with a precipitate solution and placed in a small chamber above a reservoir containing a precipitate solution, separated by a gas phase. The precipitate solution usually contains chemicals which are highly soluble and bind water in their

hydration shell (e.g. PEG or salt), thereby competing with molecules of the solvation shell of protein molecules ("salting out"). In other words, free energy is gained when parts of the hydration shell solvate precipitate molecules. Also the energy loss, by partly release of solvating water in the crystallization process, is compensated by solvation of precipitate molecules. In this experimental setup, the protein precipitate mixture has a lower precipitate concentration compared to the reservoir. Over time, evaporated water from the drop will condense in the reservoir, increasing slowly the protein and precipitate concentration in the drop. Experimentally, different precipitation compounds, pH values, and additives are screened at different temperatures, as these parameters have a strong impact on protein solubility.

Supersaturated solutions thermodynamically prefer a phase transition, but the phase transition is kinetically hindered. Two soluble proteins might collide with each other in an orientation with enough energy to come close enough to bind, but the energy gained by binding might be lower than the energy lost through partly removal of the solvation. The lifetime of this dimer might be too short for a second collision which would increase the size. But at a certain point, local disturbance will not vanish and the one phase system will collapse. A "second" liquid phase with high protein concentration and protein-protein interactions, a gel-like structure with filamentous protein, amorphous precipitate or ordered crystal nuclei can appear. The amorphous precipitate is energetically less favored than ordered crystals, but it can be kinetically favored. The level of super-saturation when nucleation occurs is usually high. Any disturbance of the supersaturated solution can be the origin of a phase collapse, e.g. other protein contaminants, aggregate, impurities, solid impurities, phase boundaries (crystallization plate, air water interface). High supersaturation, to overcome the kinetic hindered nucleation, can be avoided by the introduction of external crystallization nuclei originating from previously grown crystals fragments. Lower temperatures energetically favor crystallization and slow down evaporation. The surface to volume ratio is higher for smaller crystals, this means that relatively more molecules will be on the surface, with less interaction, which is less favorable. Parameters that slow down evaporation can help to reduce

the number of nucleation (e.g. oil cover, larger drop size, dilution of reservoir). The growth of crystals depends on how supersaturated the solution is. Crystal growth depletes the solution from protein molecules, while evaporation increases the protein concentration. If vapor diffusion is too fast, too many crystallization nuclei can be generated, leading to small or intergrown crystals. The extreme high level of super saturation when nucleation occurs might not be optimal for crystal growth. Crystals that grow too fast can lead to misplaced molecules. The dislocation might then propagate over many molecules. Also, incorporation of impurities can lead to dislocation or growth arrest. Hence, formation of nuclei at lower super-saturation can improve the quality and size of crystals (**Figure 4**). The direct surroundings of a growing crystal are depleted of protein. The emerging concentration gradient leads to diffusion transport of proteins into the depleted zone. Diffusion is usually very slow for macromolecules, resulting in a zone with low level of super-saturation surrounding the crystal. Microgravity crystallization resulted in better ordered and larger crystals [3]. On earth, convective mixing driven by density differences dominates diffusion transport of proteins. A way to slow down this process is crystallization in gel-like matrices or highly viscous solutions, e.g. agarose gel, highly concentrated PEG solutions [4]. It must be assumed that not all proteins can be crystallized. Proteins can be intrinsically disordered or contain domains which are. They might also contain long flexible loops which hinder crystal formation. Their geometry in combination with their surface charge might not allow a stable symmetric orientation. Membrane proteins might contain a small hydrophilic interface which might be present only in one direction. Here, the protein surface can be changed by removal of unstructured domains, shortening of loops, mutation of surface residues, chemical modification of residues (e.g. methylation, acetylation), design of chimeric proteins with additional domains (e.g. lysozyme) or complexation with a ligand (e.g. nanobodies, antibody fragments, affibody, DARPin).

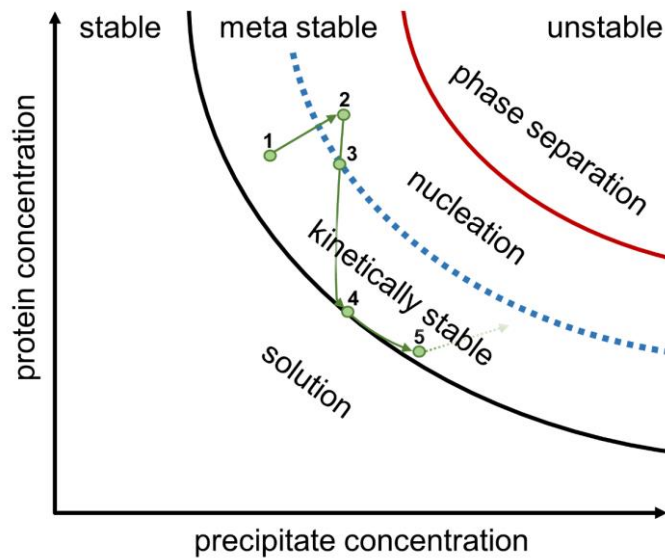


Figure 4: Protein crystallization diagram. A protein solution is stable below the solubility line (black). At higher concentration of protein or precipitate, phase separation is thermodynamically favored, but can be kinetically hindered. In this kinetically stable region crystals can grow if a nucleation core or crystal is available but cannot be formed. Above the dotted blue line attraction between molecules is high enough to form stable nucleation cores. Above the red line phase separation will occur (e.g. aggregation). A route of a successful crystallization experiment by vapor diffusion is shown in green. Starting at point 1, where no nucleation can occur. The concentration of protein and precipitate is increasing through water evaporation. At point 2 nucleation has occurred reducing the protein concentration to point 3, where no new nucleation cores are formed but existing nucleation cores grow further. At point 4 the solution is depleted from super saturated protein, the crystals can further grow as long as evaporation increases the concentration of protein and precipitants. At point 5 the crystal growths has stopped, concentration independent reasons such as impurity absorption at the surface, prevent further growth, the concentrations might increase further due to evaporation. This is a schematic diagram for the phase separation at high precipitate concentrations. In general, proteins are also not very soluble in solutions with a low ionic strength, allowing to reduce the precipitate concentration in order to generate crystals. Also different phases might be possible.

4. Crystal Symmetry

Crystals are patterns of objects that are regularly repeating in three dimensions. They can be described by a minimal volume, the unit cell, which is translated in three directions. The unit cell is defined by six parameters, the translation vectors (a , b , c) and the angles (α , β , γ), spanning the volume. Every crystal can be described by this six parameters as triclinic lattice (space group P1). A crystal might contain symmetry, which reduce the number of independent variables needed to describe the lattice. Two vectors can be orthogonal to each other. Then, the third vector can either be orthogonal (orthorhombic) or not (monoclinic). The orthorhombic lattice can have a higher symmetry when two (tetragonal) or three (cubic) vectors have the same length. Special cases are the trigonal or the hexagonal crystal systems, where a plane of the lattice is spun by three vectors which all have the same length and an angle of 120° between them, one of them is redundant ($a_1, a_2(b), a_3$). In this special cases every lattice point has six neighboring equidistant lattice points in the plane. The volume is spun by a vector orthogonal to the plane. Together, these different possibilities form the six primitive lattices of the seven crystal systems (**Figure 5**).

The symmetry of the unit cell content can allow a lattice where primitive lattices are stacked into each other. This so-called translational centered lattices can be described by additional vectors, in dimensions of the primitive lattice, pointing to additional lattice points. As there is no difference between the lattice points, these vectors link alternative origins to the unit cell.

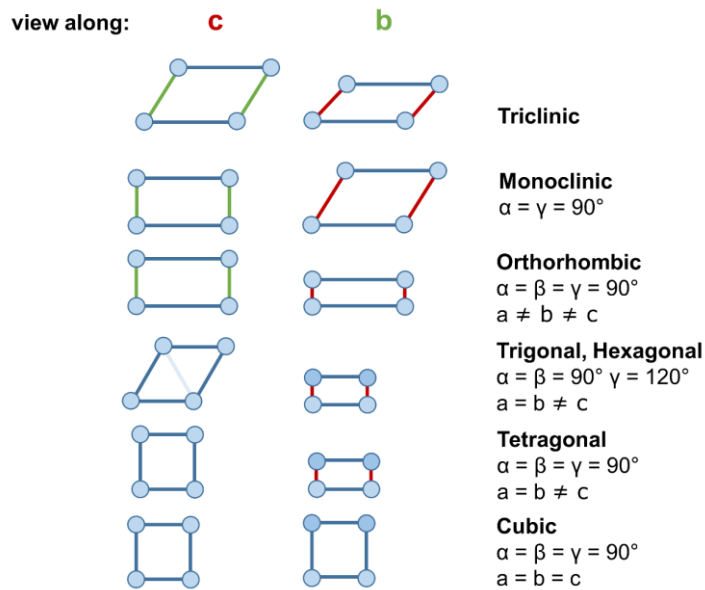


Figure 5: The six primitive lattices of the seven crystal systems. First and second row are views along the c-axis and b-axis, respectively. The b- and the c-axis is shown in green and red, respectively, if their length differs from the a-axis. The triclinic lattice has no symmetry restrictions, all other have one or more 90° angles. In the trigonal, hexagonal system the γ angle is 120° , the light blue line represents the fourth redundant direction to visualize the hexagonal and trigonal character of this primitive lattice.

Four different types of centering exist. In a base centered cell (C) the additional lattice point is in the middle of one plane of two orthogonal unit cell vectors (by convention $(\frac{1}{2} \frac{1}{2} 0)$ in the a, b plane). In a face centered cell (F) additional lattice points are in the middle of all three planes of the unit cell $(\frac{1}{2} \frac{1}{2} 0)$, $(\frac{1}{2} 0 \frac{1}{2})$, $(0 \frac{1}{2} \frac{1}{2})$. In a body centered cell (I) the additional lattice point is in the middle of the cell $(\frac{1}{2} \frac{1}{2} \frac{1}{2})$. There is a special case in the trigonal crystal system, the rhombohedral centering (R), with the centering $(\frac{2}{3} \frac{1}{3} \frac{1}{3})$ and $(\frac{1}{3} \frac{2}{3} \frac{2}{3})$. The combination of the seven crystal systems (**Figure 5**) with the centered lattices result in 14 Bravais lattices (**Table 1**). The asymmetric unit is the smallest cell which needs to be described together with the symmetry of the crystal to fully describe the crystal. The multiplicity is the number of symmetry equivalent “copies” of this asymmetric unit, which are generated by the symmetry of the lattice. Beside translational symmetry, rotational symmetry or mirror symmetry can exist.

Table 1: The seven crystal systems and the 14 Bravais types with their multiplicity. The trigonal and the hexagonal primitive cells are the same (hP).

Crystal system	Bravais types	Multiplicity	Lattice types
Triclinic	P	1	aP
Monoclinic	P, C	1, 2	mP, mC
Orthorhombic	P, C, I, F	1, 2, 2, 4	oP, ol, oC, oF
Tetragonal	P, I	1, 2	tP, tI
Trigonal	P, R	1, 3	hP, hR
Hexagonal	P	1	hP
Cubic	P, I, F	1, 2, 4	cP, cI, cF

Symmetry equivalent positions are linked by one or more symmetry operators. Therefore, it is enough to describe a single object and the symmetry operators (multiplicity of the symmetry operator) to fully construct the content. Point groups describe the complete symmetry of a molecule or the geometric arrangement of molecules or atoms. However, not all point groups are compatible with one of the seven crystal system (only point groups containing a “1”, 2, 3, 4, and 6-fold rotation axis). There are 32 crystallographic point groups and 11 of them are chiral (**Table 2**).

Table 2: The 11 chiral point groups in Hermann-Mauguin notation with the direction of view and multiplicity.

Crystal system	Point groups	Direction of view	Multiplicity
Triclinic	1		1
Monoclinic	2	[010]	2
Orthorhombic	222	[100], [010],[001]	4
Tetragonal	4, 422	[001], [100], [110]	4, 8
Trigonal	3, 32	[001], [100]	3, 6
Hexagonal	6, 622	[001], [100], [110]	6, 12
Cubic	23, 432	[100], [111], [110]	12, 24

In crystallography, the Hermann-Mauguin notation is used where a number n means a rotation axis of 360° divided by n . Point inversion (\bar{n}) and mirror symmetry (m) are not possible for chiral molecules such as proteins. Different combinations of rotation axes are possible and described by the point group of a crystal system. The multiplicity describes the number of objects generated from one object through the combination of symmetry operators. Rotation symmetry can exist in combination with translation by a fraction of the unit cell vector and in direction of the unit cell vector, a so-called screw axis (e.g. 3_1 describes a rotation of 120° and translation of $\frac{1}{3}$ of the unit cell, applied three times, the operators will shift the molecule to the same position in the neighboring unit cell) (**Figure 6**).

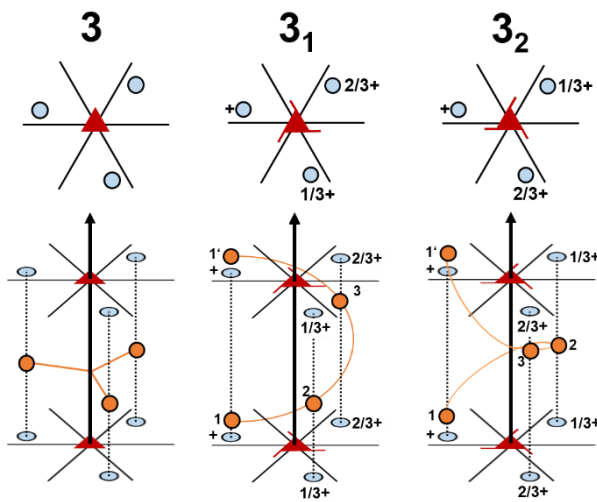


Figure 6: Rotation axis and screw axes. The 3-fold rotation axis and its two corresponding rotation axes are shown. The top row shows the view along the rotation axis with Hermann-Mauguin symbol for the rotation axis in red. Blue circles represent positions, the numbers next to circles emphasize the translation operation in unit cell units along the direction of the rotation axis. Bottom row shows a perspective view perpendicular to the rotation axis. The orange circles represent objects, where 3_1 results in a counterclockwise and 3_2 in a clockwise rotation.

A combination of the 11 chiral crystallographic point groups with the 14 Bravais lattices results in 65 possible space groups. A given point group samples the content of the asymmetric unit through its symmetry. For every symmetry related position in the Bravais lattice, the symmetry of the point group needs to be

applicable. Therefore, the multiplicity of the Bravais lattice and the point group are independent. To calculate the multiplicity of a specific space group the multiplicity of the Bravais lattice and the multiplicity of the point group are multiplied. Proteins crystallize in some space groups more frequently compared to others. The three most frequent space groups are $P2_12_12_1$, $P2_1$ and $C2$. Screw axes are more frequently observed than comparable space groups with plain rotation axes [5].

5. Data Quality

In the diffraction experiment, X-rays are manipulated by a crystal and measured usually on an area detector. Like every measurement, the intensity measurement has an accuracy and precision caused by systematic errors and random errors of the experimental setup. Three different sources of errors can be distinguished, the experimental setup, the crystal and errors caused by radiation damage introduced during the measurement. Radiation induced measurement errors are discussed in chapter 6. The experimental setup has a variety of different error sources. Beam quality aspects are beam divergence, non-perfect monochromaticity, fluctuations in intensity during the experiment, and the shape of the beam. Related to this, the beam may need to be switched on and off throughout the measurement with a beam shutter, introducing errors in the time the shutter is opening or closing. The crystal centering and rotation can be a source of errors. During the measurement, the crystal is rotated. A misalignment of the rotation axis or a miscentering of the crystal in the beam will result in differences of the diffraction intensities, as the crystal might move partly or completely out of the beam during the measurement. The rotation of the crystal needs to be precise and uniform throughout the measurement. Other movements such as vibrations of the crystal can lead to errors. The quality of the detector also influences the data quality. The distance of every detector pixel to the beam center and crystal needs to be known. Detectors are not perfectly planar and the beam center position has a measurement error.

Detectors have a certain area, "pixel", in which they read out the incoming radiation independent from other pixels. The pixel number defines the resolution of the detector. Important properties of the detector are the dynamic range, readout dead-time, blind regions, point spread function, detectable energy range, capability to detect the photon energy. The dynamic range is the minimum and maximum number of X-ray photons a detector can measure. The sensitivity of the detector may depend on the number of detected photons. The signal is amplified by the detection which might not be uniformly over the area and dynamic range. Detectors have a downtime, during which they are unresponsive to exposure when the signal is being read out. Modern counting detectors consist of smaller detector modules which are assembled to a larger area detector this is beneficial in the manufacturing process as bad performing modules [6] can be sort out or replaced when damaged. Also different designs can be easily realized, but with the downside of unresponsive area in between the modules. This blind regions also effect the detector surfaces directly adjacent, where reflections may only be partially recorded. A photon detected at a pixel may also influence adjacent pixels, spreading the signal. Some pixel of a detector may be damaged and unresponsive to X-rays. Photons within an energy range are detected. Here, X-ray fluorescents or high energy cosmic radiation contribute to the background noise. Direct counting detectors can discriminate between photon energies based on the signal caused by single photons and suppress unwanted background signals to some extent. Hybrid photon counting detectors are currently the best performing detector technology [7]. Other effects are caused by the geometry of the measurement. Reflections with higher diffraction angle hit the detector with a smaller angle than low resolution reflections which hit the detector close to 90° . This changes the spot shape and area from a round to a larger oval-shaped area. The diffraction experiment is usually carried out in air, which will absorb some of the radiation, effecting the higher diffraction angles more, as the distance between crystal and detector increases. This effect can be reduced or avoided by measuring in a helium atmosphere or vacuum [8].

Other experimental errors originate from variation of the crystal volume that is exposed to X-rays. This can change during the measurement as a consequence of the shape of the rotated crystal. The crystal volume effects the overall intensity of the measured spots. Other parameters leading to inaccurate measurements are crystal intrinsic. The molecules in a crystal are only ordered to some extent. A side chain or loop from a molecule that is at a crystallographic equivalent position can be present in two or more conformations throughout the crystal, whereas these conformations are not related to the crystallographic symmetry. Taking into account all unit cells in the exposed volume, the crystal then diffracts like two or more crystals, with differences only in the atomic scattering functions of the atoms at the different positions. In other words, two atomic scattering functions are needed for these atoms with a factor indicating to which percentage the actual conformation is present. The resulting structure factor can be calculated according to equation (4) but with a weighting factor n_j for every additional atom conformation j (8).

$$(8) \quad F_{hkl} = \sum_j^{atoms} n_j f_j \exp(2\pi i (hx_j + ky_j + lz_j))$$

The diffraction intensity is drastically reduced for these atoms, as is their electron density in the map. At a certain point, the contribution from different conformations will vanish in the background of the measurement. In other words, flexible parts of a protein will not contribute to the diffraction signal.

Thermal energy leads to a displacement of the atoms in the crystal around their center of mass. Equation (8) can be reformulated with the isotropic displacement factor B_j (9).

$$(9) \quad F_{hkl} = \sum_j^{atoms} n_j f_j^0 \exp(-B_j (\sin \theta / \lambda)^2) \exp(2\pi i (hx_j + ky_j + lz_j))$$

In addition to the displacement caused by thermal motion (classically called temperature factor) during the measurement, a general displacement of the atoms throughout the crystal decreases the diffraction signal. The B_{iso} is directly related to the mean square isotropic displacement $\langle u_{iso}^2 \rangle$ **(10)**.

$$(10) \quad B_{iso} = 8\pi^2 \langle u_{iso}^2 \rangle$$

The displacement will reduce the diffraction intensities as a function of the diffraction angle. If the atoms are displaced by 1 Å, no diffraction signal below 2 Å resolution will be observable while the diffraction around 20 Å will be less effected. The mean B-factor of all atoms can be determined from the measured intensities, applying Wilson statistics. The mean intensity of the diffraction spots is attenuated exponentially with the resolution $\sin^2\theta/\lambda^2$. This relation can be linearized in a logarithmic plot where the B-value correlates with the slope, the so called Wilson plot **(11)** [9].

$$(11) \quad \ln \frac{\overline{I_{hkl}}}{\sum (f_j^0)^2} = C - 2B \left(\frac{\sin\theta}{\lambda} \right)^2$$

The unit cell geometry is obtained via difference vector analysis of the assigned spots. During indexing and intensity integration, different measurement errors are acquired and minimized. The discrepancy between the observed and expected reflection positions are used to refine the unit cell geometry. During the integration process, a measured intensity is assigned to every spot, including an error estimate. Therefore, the background needs to be determined and the signal from different pixels and frames needs to be integrated. Profile fitting of the recorded signals at adjacent detector pixels enhances the data quality as intensities are corrected (upscaled or downweighted) on the basis of their expected profile shape.

Fluctuations in the intensity profile over the course of the experiment are reduced by scaling of the data. Reflections that are recorded multiple times, or are related through symmetry, are merged to one value with its standard deviation. This reduces the recorded data while the variance of the measurements can be determined. At the end of this process different types of errors can be calculated containing information about the quality of the measured data. One of the most important values is the $(I/\sigma)^{asymptotic}$. This value can be fitted from a I/σ versus intensity-counts plot and represents the highest I/σ which can be obtained in the experiment [10]. The plot contains the information, if beam intensity was too high and if a measurement with the same dose but more redundancy would have been beneficial. Other values are usually grouped in different resolution shells. Important values are the signal to noise ratio I/σ , the completeness as percentage of recorded hkl reflections, redundancy of the measured reflections, the R_{meas} -values giving the error between multiple times recorded reflections **(12)**, and the $CC_{1/2}$, a cross correlation between two randomly selected parts of the data. The $CC_{1/2}$ allows to determine to which resolution the data contain significant signals [11].

$$(12) \quad R_{meas} = \frac{\sum_{hkl} \sqrt{\frac{n}{(n-1)}} \sum_1^n |I_{i,hkl} - (I_{hkl})|}{\sum_{hkl} \sum_1^n |I_{i,hkl}|}$$

The degree of order in a crystal defines the resolution to which the crystal diffracts. The order in a crystals can depend on the unit cell directions, because of favored crystal contact in thus, distinct directions. The diffraction intensity and spot shape can vary according to which reflection is measured. If more than one crystal is measured at once, the recorded image will contain two diffraction patterns and some of the spots might overlap which increases the uncertainty of the measurement, or might not allow the interpretation. Cracks in the crystal may result in a similar observation, but the two diffraction patterns might almost overlap. Crystals are composed of small domains slightly misaligned towards each other, which lead to a blurring of the diffraction spots. The mosaicity of a crystal describes

how well the single domains are aligned. The blurring effect is stronger with increasing diffraction angle. Merohedral twinning describes a crystal disorder where different domains of the crystal are differently orientated, but the misalignment can be expressed with a defined symmetry operator that links both domains. As a result, the spots overlap perfectly, but they originate from structure factors which are not the same. The different intensities from particular hkl values are merge together through the diffraction experiment. Crystals can contain translational non-crystallographic symmetry. In this case two or more molecules can be present at different positions but with the same or nearly the same orientation in the unit cell. This will systematically modulate the diffraction pattern. The translation can be almost but not exactly a lattice translation with drastic effects on the diffraction pattern with a large number of systematically very weak and very strong reflections [12].

6. Radiation Damage

X-rays can interact in different ways with matter, elastic scattering (Rayleigh scattering), inelastic scattering (Compton scattering), or absorption. The cross section, a measure of the probability of a particle to interact with radiation, has about the same value for Compton and Rayleigh scattering for carbon at 10 keV [13]. The cross section for Compton scattering can be approximated as constant from 1 to 1000 keV, whereas the cross section for Rayleigh scattering decreases in this range more than three magnitudes [14]. In Compton scattering, the energy of the radiation is not constant, as a momentum transfer takes place and the scattered photons will contribute to the background in the diffraction experiment. The main source of radiation damage is due to the photoelectric effect, generating ions and strong ionizing electrons with kinetic energy. The energy of X-rays usually applied in an experiment is above all electron binding states of organic atoms (H, C, N, O, S, "Ca", "Cl"). The most inner electron shell (K-shell) has the highest cross section, as it is energetically closest to the used radiation. When an X-ray hits such an electron inelastically, the energy can be completely absorbed. A part of the energy then promotes the electron to continuum state while the exceeding rest is transferred in kinetic energy. This process has some consequences. The atom is ionized, has an unpaired electron and is in an excited state. An electron of a higher energy state will transit to the lower energy state, thereby releasing energy. This happens by either emission of an X-ray (characteristic X-ray radiation) or the emission of an electron (Auger electron). The emitted X-rays contribute to the recorded background or further induce radiation damage. The emitted electron (primary or Auger) has kinetic energy and bound electrons have a high cross section for inelastic scattering with these kinetic electrons, thereby more "free" electrons and positively charged atoms are generated. When the kinetic energy is low enough the electrons are captured by atoms producing negatively charged ions. Thereby, for every primary photoelectric event up to 500 low energy secondary electrons can be created [15]. This direct radiation damage will create ions and unpaired electrons and can only be reduced by lowering the absorbed X-ray dose. As a consequence, the solution surrounding a crystal should be

minimized, if possible the beam size should not exceed the crystal size (12 keV photoelectrons travel a few micrometers before they are absorbed [15] and the beam flux should be adjusted to the needs. Heavy atoms absorb X-rays better, therefore their concentration should be kept as low as necessary.

Beside an increased background signal, inelastic scattering has drastic chemical consequences. The generated radicals and ions will react with other atoms, resulting in bond cleavages and disorder in the crystal. There are more specific reactions which accumulate first, breakage of disulfide bridges, decarboxylation of aspartate and glutamate side chains, the loss of hydroxylgroups from tyrosines and carbon sulfur bond breakages [16]. At a certain dose, unspecific damage caused by X-rays accumulates to an extent that the diffraction pattern gradually vanishes, effecting the high resolution data first.

To reduce the impact of this indirect radiation damage, several procedures can be applied. Crystals can be measured at low temperatures, typically 100 K (cryo-crystallography), which slow down the chemical reactions. Problems which can arise from freezing are crystallization of surrounding water (anomaly of water), mechanical stress trough density changes, damage trough the transfer of the crystal in a solution suitable for freezing ("cryo" solution) by osmotic pressure. The addition of chemical radical scavengers seems not to improve data quality [17]. A larger crystal size increases the tolerated dose or allow data collection at different crystal positions, enhancing the signal to noise ration by merging equivalent reflections. If isotropic crystals can be produced, data sets from different crystals can be recorded and merged. X-ray Free-Electron Lasers (XFEL) allow to use tiny crystals in a short one shot exposure measurement. The extremely high photon flux directly destroys the sample, but radiation damage occurs after the diffraction process has happened (e.g. 10^{12} photons in 40 fs, >30 MGy [18]).

7. Electron Density Reconstruction

For the reconstruction of the electron density, the unit cell needs to be determined first. Structure determination by X-ray crystallography usually utilizes monochromatic light of a precisely determined wavelength. Therefore, the diffraction angle θ can be calculated for every recorded spot, as the crystal-detector distance and the distance from a recorded spot to the beam center is known. A vector system can be derived from the calculation of difference vectors of neighboring reflections, which ultimately yields unit cell parameters defining the crystal lattice. A lattice choice is made on the basis of data fitting, followed by the assignment of miller indices [hkl] for every measured spot (a process called indexing). If a crystal does contain additional symmetry elements not defined by the translational symmetry of the unit cell, the symmetry is also present in the reciprocal lattice, resulting in symmetry related reflection with equivalent intensity (in the range of measurement errors). In addition, translation symmetry (e.g. screw axes or lattice centering) yield reflections with intensity values equal to zero, so-called systematic absent reflections. This symmetry in the diffraction pattern with or without occurrence of systematic absent reflections is used to assign the likeliest space group of the crystal. Ultimately, only a successful model building with decent model parameters can validate the assignment of the space group, and can discriminate between two enantiomeric space groups (e.g. $P3_1$ and $P3_2$).

Chapter 2 summarized the diffraction theory and aimed to explain the appearance of diffraction patterns as well as the concept of the reciprocal lattice and the reciprocal space to describe it. This is helpful in calculating a theoretical diffraction pattern based on a model. Experimentally, the opposite is what needs to be achieved. The electron density map in real space needs to be calculated from a recorded diffraction pattern.

Equation (1) can be reformulated using discrete grid point coordinates of the unit cell. Due to the discrete diffraction pattern caused by the lattice, a Fourier summation is sufficient (13).

$$(13) \quad F_{hkl} = V \sum_{x=0}^a \sum_{y=0}^b \sum_{z=0}^c \rho_{xyz} \exp(2\pi i (hx + ky + lz))$$

Where the Miller indices hkl represent reciprocal lattice points and V the volume of the unit cell. The electron density ρ_{xyz} can be expressed as the inverse of the Fourier transform (14).

$$(14) \quad \rho_{xyz} = \frac{1}{V} \sum_{h=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} \sum_{l=-\infty}^{+\infty} F_{hkl} \exp(-2\pi i (hx + ky + lz))$$

The structure factors F_{hkl} are complex numbers with an amplitude F_{hkl} and a phase α_{hkl} (15).

$$(15) \quad \rho_{xyz} = \frac{1}{V} \sum_{h=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} \sum_{l=-\infty}^{+\infty} |F_{hkl}| \exp(i\alpha_{hkl} - 2\pi i (hx + ky + lz))$$

In other words, the electron density can be reconstructed by a summation of all waves of the lattice system. The phase relation of this different waves with a given lattice determines the geometry of their superposition, whereas the amplitude acts as weight of a given wave. As we can only measure the real part of this complex numbers and not their phases, the geometry of the wave superposition is unknown, which results in the so-called “phase problem”.

8. Solving the Phase Problem

X-rays hitting the detector can be measured. The contribution of a certain atom to a measured signal (which might be negative) can only be calculated when the position of the atom is known. That is exactly the aim of solving a structure. In other words, the real part of every diffracted wave can be measured but only simultaneously as a superposition of all waves of all atoms exposed to X-rays. The individual contribution of each atom to this superposition, the phase differences between each diffracted wave function is lost during the experiment. Experimentally this means that, for each structure factor amplitude gained from the intensity measurement of a diffraction spot, the corresponding phase is missing and needs to be generated.

8.1. Direct Methods

A way to calculate an electron density map is to use random phases for all structure factors, but this will result in a random electron density map. However, some facts are already known that are true for all atomic structures. First, the electron density needs to be positive, hence there are no areas possible with less than zero electrons. Second, electrons are spherically distributed around atomic nuclei. This assumption is generally true, but the measured data necessarily needs to contain atomicity, which is typically found at resolutions $\leq 1.2 \text{ \AA}$ [19]. At this atomic resolution the parameters which are needed to describe the crystal structure, are over-determined by the number of measured intensities. Third, especially in protein structures, the atoms are approximately equal in electron number and size. This assumption results in a phase relation of the structure factors (Sayre equation, triplet relation). With this relation and a set of random starting phases all other phases can be calculated (tangent formula). Different starting phases are chosen until an interpretable spherical electron density map is obtained. A more sophisticated approach such as the “shake and bake approach” uses dual space recycling [20]. Here, starting atoms are placed and phases are calculated from

these random “seed” atoms. Then phases are improved in reciprocal space by minimizing current and statistically expected values based on the triplet relation. In the next step, a real space map is calculated on the basis of the newly established phases, and atoms are placed at the highest peaks. This is done in a cyclic procedure where the next round of phase improvement starts with the calculation of new phases on the basis of the new model. This dual space cycling is repeated several times and with different starting seed atoms. A measure of the quality of phases, a figure of merit is calculated based on the fulfilment of the Sayre equation. The best solutions will also show a good correlation between calculated and measured reciprocal space data.

Patterson based methods utilize inter atomic distance information gained from the measured data. A map can be calculated using intensity values instead of structure factors, thereby ignoring the phases. The intensity is proportional to the squared absolute value of the structure factor **(16)**.

$$(16) \quad I_{hkl} = I_0 k \frac{N}{U} \lambda^3 LPA |F_{hkl}|^2$$

Where I_0 is the incoming beam intensity, k is a constant, N is the number of unit cells, U is the unit cell volume, λ is the wavelength of the incoming beam, L is the Lorentz factor correcting for different angular velocities of different reciprocal lattice points while the crystal is rotated, P is a polarization factor depending on the polarization of the incoming beam, while A is a correction factor for the absorption.

The Patterson map can be calculated in the Patterson space (uvw) directly from the intensities **(17)**.

$$(17) \quad P_{uvw} = \frac{1}{V} \sum_{h=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} \sum_{l=-\infty}^{+\infty} I_{hkl} \cos 2\pi (hu + kv + lw)$$

This map represents the autocorrelation of the density map **(18)**.

$$(18) \quad P_{uvw} = \int_R \rho(\mathbf{r})\rho(\mathbf{r} + \mathbf{t}_{uvw}) d\mathbf{r}$$

Were \mathbf{r} is a point in the real space R and \mathbf{t}_{uvw} is a point in the Patterson space P_{uvw} . Graphically, the electron density map is moved in u,v,w over the electron density map and the values of every point results in a Patterson map with the axis u,v,w .

This has some consequences. If the map is moved to $\langle 0 0 0 \rangle$, both electron density maps overlap perfectly, which results in the highest peak, the origin peak. If the map is moved that the electron density of two atoms overlap, a peak in the Patterson map will result. This leads to a very crowded map with N^2 peaks where N is the number of atoms in the unit cell. The distance of a peak to the origin peak represents the distance between two atoms. If the autocorrelation is moved in the opposite direction by the same distance, the correlation will be identical. Interpreted as distance, the distance between A and B $\langle uvw \rangle$ and B and A $\langle -u-v-w \rangle$ are identical. Hence, the Patterson map is centrosymmetric and every translational symmetry information (screw axis) of the unit cell is lost.

The Patterson map is very crowded with broad peaks for every interatomic distance present in the unit cell. The interpretation of Patterson maps of protein crystals is demanding and often impossible. But Patterson methods can be used in protein crystallography, if the crystal contains electron-rich atoms (heavy atoms) as marker atoms incorporated in the protein lattice. All other atoms are neglected by difference methods, reducing the problem to a few marker atoms with high peak intensities due to their high electron density. In the Patterson map, the distances and directions between the peaks are geometric restraints, which a model needs to fulfil. Once a suitable model is created, a translational search in real space needs to be performed to find the position of the marker atoms eventually resulting in a

plausible map. If the space group contains additional symmetry, additional information can be gained from Patterson maps. Additional symmetry also means, that there is more than one copy of each molecule in the unit cell. This will result in a distance vector that maps all symmetry related atoms to their symmetry equivalent position in the unit cell. This "self" distances peaks can be found at certain sections (or lines) in the Patterson map, the Harker sections (or lines). These Harker sections have to contain high peaks, if marker atoms are present. From the known symmetry of the unit cell, the influence of the symmetry elements on Patterson map distances is known. In other words, real space symmetry will be generated in certain areas of the Patterson map peaks. The symmetry element contains information about the transformation from Patterson to real space. For example, a 2 fold rotation axis along y has two equivalent positions at $\langle x, y, z \rangle$ and $\langle -x, y, -z \rangle$ this implies the distance between two equivalent atoms is $[2x, 0, 2z]$. Values for u and w from a peak in the corresponding Harker section $[u, 0, w]$ result from atoms at the real space lines $[u/2, y, w/2]$ and $[-u/2, y, -w/2]$. The problem here is underdetermined as the position in y is not known. For perpendicular symmetry elements, the problem can be solved and positions can be calculated from different Harker sections.

8.2. Molecular Replacement

In direct methods single atoms are placed in the unit cell. In molecular replacement a whole protein structure is placed. To gain meaningful phases, the challenge is to find a template which is structurally similar to the unknown structure, and to place it correctly in the unit cell. This method depends on how close the target structure is to the template. Previously solved structures with high sequence similarities are commonly used as search models, where the residues are shortened based on the sequence differences of the unknown structure. Potentially flexible parts, e.g. long loops and solvent accessible residues, might also be removed.

If the model is close enough, a good solution will show a high cross-correlation between calculated and observed structure factors. The map of the observed

structure factor amplitudes with phases from the correctly placed model will improve with the model similarity and coverage. The placement of the model in the unit cell is a 6-dimensional search with three rotational and three translational degrees of freedom. The calculation of the structure factors at every translational grid point, with all different orientations, from all atoms of the model, is computational demanding. To reduce the parameterization, the translational and rotational search can be performed separately. The rotation function search superposes the Patterson map of the measured data with the Patterson map of the search model. A value for the overlap of the two maps is calculated for every rotation angle of the search probe. A high overlap will result if the rotational orientation of the search model and measured structure are similar. The translational search places the molecule in the unit cell that the unit cell symmetry is fulfilled. In other words, the origin of the unit cell needs to be placed in a manner that the symmetry resembles the measured data. In P1 the origin is arbitrary and no translation search is needed. In polar unit cells, containing a rotation axis with no perpendicular symmetry element, molecule copies will be in a plane with a distance towards each other and the rotation axis in the center. The distance between the molecules can be determined by a translation that allow a correct placement of the rotation axis. In more symmetric space groups with perpendicular symmetry elements, the translation search is more difficult as the placement of the origin is restricted. The translation search can be done in Patterson space where again a calculated and measured Patterson map is compared. In the simplest case with a 2-fold axis, two molecules are placed in a P1 cell according to the two best solutions from the rotation search. One of the molecules is moved on a 2-dimensional grid and a Patterson map is calculated for every grid point and compared to the Patterson map of the measured data. The highest correlation of the overlapped Patterson maps can be expected when the intermolecular Patterson peaks of both maps will overlap. In this way the distance vector between the two molecules is determined and the 2-fold axis has to be located on half of this distance vector. In a last step, the solutions can be tested for meaningfulness by applying all symmetry operators and checking how well the molecules pack in the unit cell without major clashes.

8.3. Single Isomorphous Replacement (SIR)

In this method two data sets are recorded, a native data set of an unmodified crystal and one containing a heavy atom derivative. The method relies on the isomorphism between both crystals. Therefore, it is crucial that the derivatization with a heavy atom does not change the unit cell to large degree. If the native protein structure factor F_P is subtracted from the heavy atom derivative structure factor F_{PA} , a heavy atom only structure factor would result F_A (19).

$$(19) \quad F_A = F_{PA} - F_P$$

If a Patterson map is calculated, it will contain peaks for distances between heavy atoms and peaks for distances between heavy atoms and all other atoms (20).

$$(20) \quad F_{PA}^2 - F_P^2 = F_A^2 + F_A F_P^* + F_P F_A^*$$

An isomorphous difference Patterson map contains less noise as long as F_A is drastically smaller than F_P and F_{PA} (21).

$$(21) \quad \Delta F_{ISO}^2 = (F_{PA} - F_P)^2 = F_A^2 \cos^2(\varphi_{PA} - \varphi_A) \approx F_A^2$$

The squared isomorph difference structure factors ΔF_{ISO}^2 , can be used to calculate a difference Patterson map that subsequently can be utilized to find the marker atom substructure. The structure factors with phases of the solved marker atom substructure are then used to determine the phases of the protein structure factors. The vector summation, as emphasized by a Harker construction, results in two solutions with a closed triangle (Figure 7).

The “lack of closure” explained in chapter 9.1 can be calculated for every phase angle resulting in a probability distribution for the phase angle of every structure factor. The phase ambiguity leads to two phases with the highest probability. Uncertainties in the measurement of the structure factor amplitudes and the phases from the heavy atom solution will broaden the probability function. If the phases with the highest probability are chosen, 50% of them will be incorrect. A better choice for initial phases can be made by averaging the probability of all possible phase angle of a reflection. The result will be a complex number with an amplitude smaller than the structure factor amplitude. The phase of this complex number, the so-called “best phase” together with the amplitude as weighting term, the so-called “figure of merit” is a better choice for initial phases. With this method, all structure factors will have large phase errors, but they will be in average more close to the real phase, and the structure factors with large uncertainties will be weighted down.

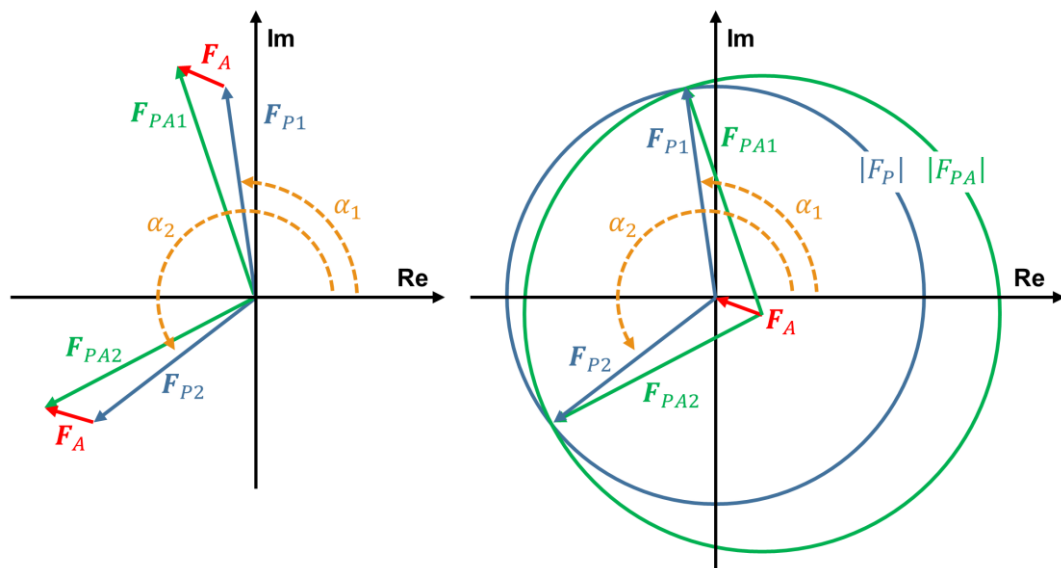


Figure 7: Vector summation of SIR data sets. The structure factor of the heavy atom substructure solution F_A is known (red). The structure factor amplitudes for the protein $|F_P|$ (blue) and the derivative $|F_{PA}|$ (green) are known and depicted as circles. The vector summation $F_{PA} = F_P + F_A$ has two intersections. The phase of the protein structure factor is either α_1 or α_2 (orange).

The phase ambiguity can be resolved by an additional derivative data set. One of the two solutions from the first SIR experiment will be similar to one of the two solutions from the additional derivative. The vector summation of the the so-called multiple isomorphous replacement (MIR), can be visualized as in SIR, but with an additional derivative (**Figure 8**).

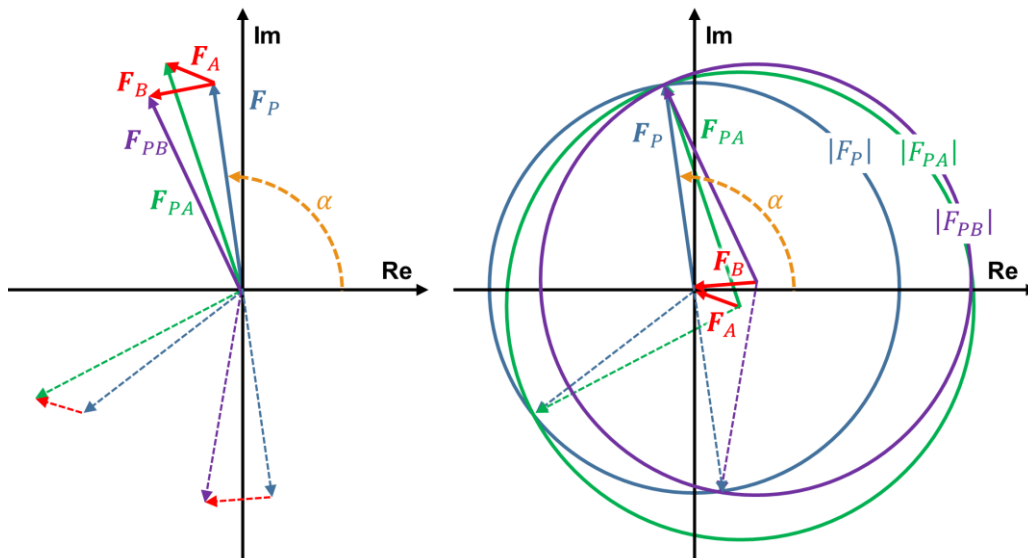


Figure 8: Vector summation of MIR data sets. The structure factor of the heavy atom substructure solution F_A and F_B are known (red). The structure factor amplitudes for the protein $|F_P|$ (blue), derivative A $|F_{PA}|$ (green) and derivative B $|F_{PB}|$ (purple) are known and depicted as circles. The two vector summations $F_P = F_{PA} - F_A = F_{PB} - F_B$ have one solution. Thin dotted lines indicate the incorrect solutions. The phase for this protein structure factor is α (orange).

8.4. Single-Wavelength Anomalous Dispersion (SAD)

Atomic scattering factors are dependent not only on the scattering angle but also the wavelength, which can be exploited to find the positions of marker atoms. The atomic scattering factor can be expressed as a wavelength independent part f_S^0 and two wavelength dependent parts f'_λ and f''_λ (22).

$$(22) \quad f_{S,\lambda} = f_S^0 + f'_\lambda + if''_\lambda$$

The dispersive part f'_λ effects the wave amplitude but not the phase. The anomalous scattering f''_λ arises through absorption where the phase is shifted $+90^\circ$, in other words, the imaginary part of the wavelength dependency. This phase shift leads to a breakdown of the centrosymmetry of the reciprocal space and therefore Friedel's law, which describes the equality of the intensities $I(hkl)$ and $I(-h-k-l)$, is not true anymore. The intensity difference of such reflections are used to determine the anomalous differences. The wavelength dependency declines with decreasing wavelength but not continuous. At certain frequencies (atom specific) the inner shell electrons resonate with the incoming X-rays and an energy transfer can occur. This leads to atom-specific absorption edges at given wavelengths. The presence of an atom in the sample and the exact wavelength of the absorption edge, which is to some degree dependent on the chemical state and environment of the atom, can be determined with a fluorescence scan. Here, the energy of the incoming wavelength is changed. If the energy exceeds the energy of an electron at an inner shell, a free electron will be produced, resulting in an empty state of an inner shell. This low energy state is then occupied by an electron from a higher energy state, which typically happens under the emission of a photon with specific energy. The heavier an elements is, the higher its absorption and therefore the anomalous scattering contribution. The anomalous scattering can be maximized if the diffraction data is recorded at the absorption edge or

slightly above. Not all elements have absorption edges at energies suitable for X-ray diffraction.

The anomalous differences between two Friedel pair F_{PA}^+ and F_{PA}^- can be exploited to calculate a difference Patterson map, which subsequently can be used to find the positions of the marker atoms in a similar procedure as described for the SIR case (23).

$$(23) \quad \Delta F_{ANO}^2 = (F_{PA}^+ - F_{PA}^-)^2 = F_A^2 \sin^2(\varphi_{PA} - \varphi_A) \approx F_A^2$$

If the substructure is solved, the marker atom positions are found, and initial protein phases can be calculated. As explained for the SIR experiment, a SAD experiment is underdetermined resulting in a phase ambiguity (Figure 9).

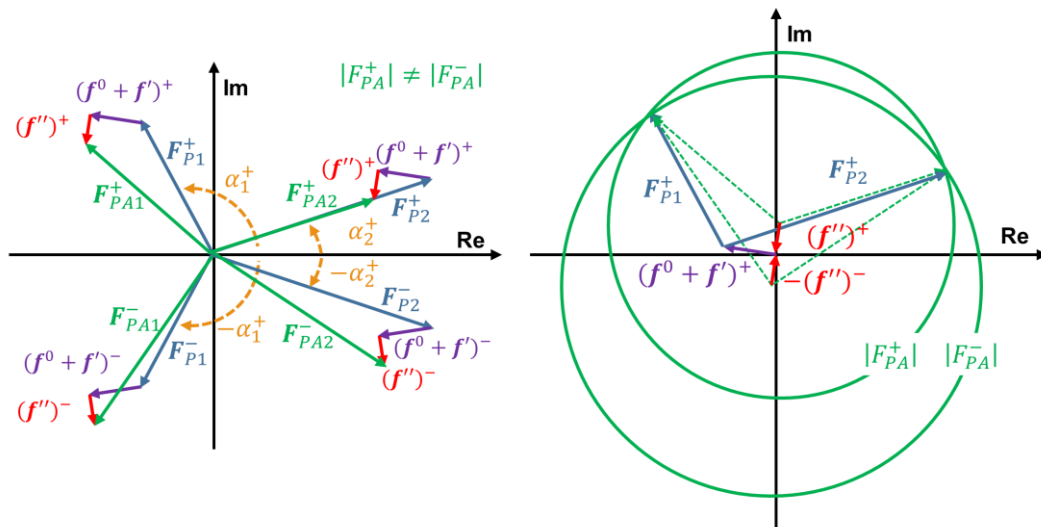


Figure 9: Vector summation of a SAD experiment. The reciprocal space is not centrosymmetric due to the anomalous scattering. Therefore, the measured structure factor amplitudes $|F_{PA}^+|$ and $|F_{PA}^-|$ (both green) are not equal. If the heavy atom substructure solution is known, $f^0 + f'$ (purple) and f'' (red) can be determined. The vector summation has two solutions (F_{P1}, F_{P2}) for each protein structure factor, which are centrosymmetric (F_P^+, F_P^-) (blue). The two solutions for the phase are shown in orange.

The “best phase” with the highest figure of merit can be calculated as described in chapter 9.1 by averaging the lack of closure for all phases. The phase ambiguity can be resolved by a combination of SAD with SIR or MIR, single or multiple isomorphous replacement with anomalous contribution (SIRAS or MIRAS). The combination of the two calculated difference Patterson maps (21)(23) theoretically also reduces the noise ($\cos^2 x + \sin^2 x = 1$)(24).

$$(24) \quad \Delta F_{ANO}^2 + \Delta F_{ISO}^2 = F_A^2$$

Multi-wavelength anomalous diffraction (MAD) uses the dispersive effect by measuring data sets of a protein crystal with marker atoms at different wavelengths. The dispersive difference between two data sets can be maximized by measurements at the wavelength of the peak and inflection point of the absorption edge. Additionally, a high and low remote data set can be measured above (high energy) and below (low energy) the absorption edge. The marker atom positions can then be identified using an equation system that is determined with two measured wavelengths. Where F_T includes f_0 but not the anomalous dispersive contributions of the marker atom (25)(26) [21].

$$(25) \quad F_+^2 = F_T^2 + \frac{f'^2 + f''^2}{(f^0)^2} F_A^2 + \frac{2f'}{f^0} F_T F_A \cos \alpha + \frac{2f''}{f^0} F_T F_A \sin(\varphi_T - \varphi_A)$$

$$(26) \quad F_-^2 = F_T^2 + \frac{f'^2 + f''^2}{(f^0)^2} F_A^2 + \frac{2f'}{f^0} F_T F_A \cos \alpha - \frac{2f''}{f^0} F_T F_A \sin(\varphi_T - \varphi_A)$$

8.5. Marker Atom Substructure

Experimental phasing methods such as SIR, MIR, SAD, MAD, SIRAS and MIRAS (and also radiation damage induced phasing RIP, UV-RIP [22, 23]) rely on the solution of a marker atom substructure. The marker atom substructure can be solved using Patterson and/or direct methods. Based on the calculated structure factors of the marker atom substructure, initial phases for the protein structure can be calculated. If the quality of the initial map based on the initial phases allows building of the protein structure, the phase problem can be solved. The quality of the substructure solution and its phasing power for the protein structure relies on the actual substructure in the crystal, the crystal quality, the crystal symmetry, the atom type and the quality of the measurement. The marker atoms may bind only partially or weak with a high degree of displacement around their mean position. The diffraction power of a crystal may suffer from the derivatization procedure. Heavy atoms increase the absorbed radiation dose and therefore enhance radiation damage. Higher crystal symmetry may lead to more centric reflections, which are related by the centrosymmetry of the reciprocal space and the point group symmetry. The phase difference of two centric reflections is restricted to 180° , which is not effected by anomalous scattering. This can be exploited for the error estimation, anomalous signal estimation and data scaling, which supports solving the phase problem. The used atom type and number of bound atoms dictates the change in measurable anomalous differences $\frac{\Delta F}{F}$ between Friedel pairs (27).

$$(27) \quad \frac{\Delta F}{F} = \frac{\Delta f_A}{f_P} \sqrt{\frac{n_A}{2n_P}}$$

anomalous difference: $\Delta f_A = 2f''$

dispersive difference: $\Delta f_A = |f'_{\lambda_1} - f'_{\lambda_2}|$

Where n_A is the number of heavy atoms, n_P is the number of protein atoms, and f_P is 6.7 electrons, the average “protein-atom” scattering factor of the origin reflection [0,0,0]. The change in intensities due to derivatization can be estimated with (28).

$$(28) \quad 2 \frac{\Delta F}{F} = k \frac{f_A}{f_P} \sqrt{\frac{n_A}{n_P}}$$

Where f_A is the number of electrons of the heavy atom and k is 2 for centric and $\sqrt{2}$ for acentric reflections. Measurements for the quality of substructure solution are R_{Cullis} (29) and the phasing power P_{iso} (30), and the mean figure of merit.

$$(29) \quad R_{Cullis} = \frac{\sum_{hkl} | |F_{PA}(obs) \pm F_P(obs)| - F_A(calc) |}{\sum_{hkl} |F_{PA}(obs) \pm F_P(obs)|}$$

$$(30) \quad P_{iso} = \frac{\sum_{hkl} F_A(calc)}{\sum_{hkl} |F_{PA}(obs) - F_{PA}(calc)|}$$

A marker atom substructure that consists of a single atom type is achiral, so the solution is centric, in contrast to the acentric chiral protein structure. As a consequence, there are two possibilities how to apply the handedness when the electron density map of the protein structure is calculated. The correct handedness can be determined after density modification calculations. Only in case of the correct handedness an interpretable map can be obtained, that shows protein features, e.g. α -helices and/or β -sheets. The wrong handedness will result in a noisy and flat electron density map.

9. Phase Improvement

Initial experimental phases are error-prone, but several methods are available to improve the initial phases. These methods incorporate knowledge about protein structures in general and information about the actual structure itself.

Crystals consist of regions containing protein and regions containing solvent. This observation is utilized in a procedure called solvent-flattening [24, 25]. Solvent flattening can be viewed as a low resolution solvent model building and refinement of the solvent borders. The solvent content of a crystal can be estimated from the size of the unit cell and the protein used for crystallization.

A statistical analysis of the mean solvent content of protein crystals resulted in the Matthews parameter of $2.58 \text{ \AA}^3 \text{ Da}^{-1}$, which represents the mean volume which is occupied per protein mass. This allows to estimate the protomer content in the asymmetric unit [5, 26]. Additional information can be obtained from the self-rotation function and the heavy atom substructure solution. The electron density in protein regions is about $0.43 \text{ e}\text{-}\text{\AA}^{-3}$ whereas the electron density of the solvent is about $0.33 \text{ e}\text{-}\text{\AA}^{-3}$. The asymmetric unit can be divided in a grid and subsequently the electron density can be determined in a sphere around every grid point. Grid points with low electron density are most likely solvent regions, whereas high density regions are most likely protein regions. Solvent regions are not ordered and should not contain map features. The protein envelope can be determined by analyzing how defined map features vary on the surface of spheres placed on grid points. Once the solvent and therefore the protein borders are determined, or vice versa, the solvent region can be flattened by setting it to an average value. A more bias free solvent modification method flips the electron density at the solvent region by setting it to average and subtracting the deviation from the average (solvent flipping)[27].

Based on the number of protein protomers in the asymmetric unit non-crystallographic symmetry (NCS) may be possible. NCS operators of symmetry related proteins within the asymmetric unit can be obtained from a self-rotation function, strong Patterson peaks or the marker atom substructure and validated by calculating a real space correlation. Once the symmetry operator and its position is known the electron density in this region can be averaged.

Depending on the resolution, the electron density distribution of proteins are similar. Histogram matching changes the electron density probability distribution to an ideal distribution.

As discussed in chapters 8.3 and 8.4, all experimental phasing techniques are based on error-prone measurements. As the diffraction signal decreases with increasing diffraction angle, higher resolution reflections have higher measurement errors. Therefore, the experimental phasing will result in poor quality high resolution phases. Another aspect is that the phasing techniques depend on difference methods which increases the uncertainty compared with one single data set. While the data at high resolution still contain information about the structure factor amplitudes their might be no phase information.

Long ranging real space information such as NCS or the solvent mask lead to a phase restrictions in the reciprocal space. This allows to some extent to generate phase information from other phases. This is done by increasing the resolution with measured structure factor, for which no phase information is available. In a real space map, areas are then averaged (NCS), flattened (solvent) and new phases are calculated. This allows to generate meaningful phase information by slowly increasing the resolution of a given data set, in a cyclic manner, while repeating density modification calculations (phase extension)[28].

9.1. Phase Combination

All density modifications are done in a cyclic procedure with different rounds of density modification and combination of the modified with the measured data. To calculate a map the observed amplitudes are combined with the initial phases. The density is then modified and new amplitudes and phases are calculated based on the modified density. This calculated phases are then combined with the initial phases and the observed amplitudes to calculate a new map. Then the cycle is repeated, until convergence is reached.

For the map calculation, Fourier coefficients with a weighting term and a phase are used (31).

$$(31) \quad F_{BEST} = m|F_P|\exp(i\phi_{BEST})$$

The quality of the used “best phase” is weighted with a figure of merit m , so that structure factors are down weighted based on the phase probability function. A basic phase probability function can be calculated by assuming that only the marker atom substructure contains errors. With the marker atom structure factors and the measured structure factor amplitudes, a vector summation forming a triangle can be calculated. For every phase angle a lack of closure of this triangle can be calculated (**Figure 10**).

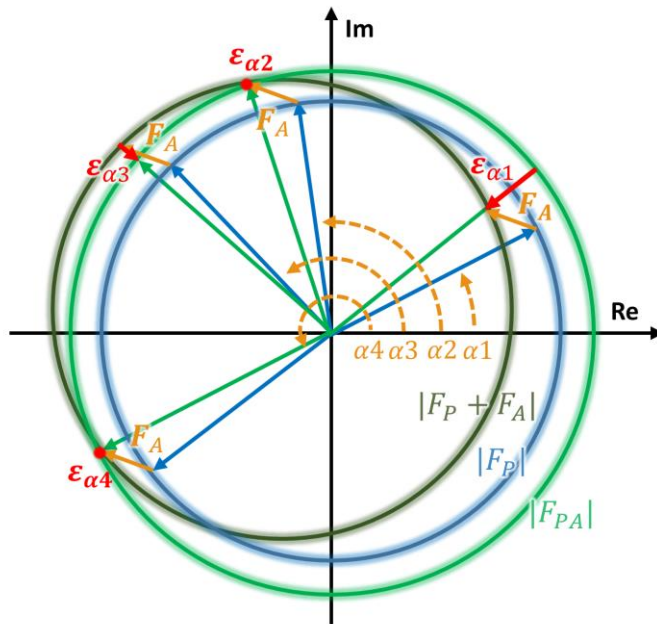


Figure 10: Vector diagram for the calculation of the lack of closure ε . The structure factor amplitudes of the protein (blue) and of its derivative (green) are shown as circles. These represent measured data, which contain errors, indicated by blurring of the circles. The structure factor including the phase angle of the derivative (orange) is known from the marker atom substructure solution. Four protein structure factor phases are shown as a vector summation. The “lack of closure” ε (red) is the difference vector to close the vector triangle. The phase angles α_2 and α_4 have a minimal lack of closure, indicated by a red dot.

The most probable phase will have the smallest lack of closure. If the distribution is not unimodal, a phase ambiguity is observed and more than one likely phase exists. The best choice for a phase is the weighted average of the probability distribution (**Figure 11**).

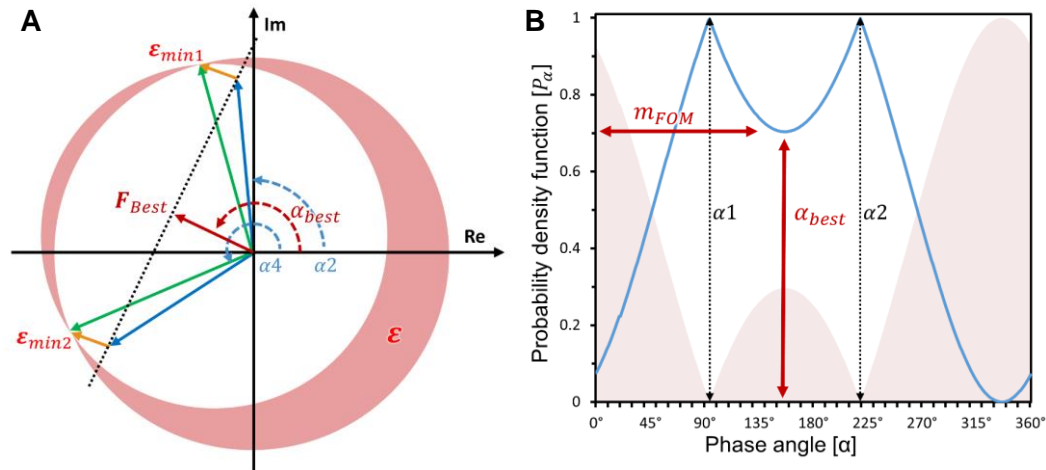


Figure 11: Lack of closure and best phase. **A:** The “lack of closure” vector diagram for the two most probable protein phases (blue). The lack of closure is shown as red area for all phases. The structure factor of the derivative (orange), the protein (blue), the protein including the derivative (green) and the best phase (red) are shown as vectors. **B:** Phase probability density function (blue) based on the “lack of closure” (red area). The best phase and the figure of merit are shown in red. The most probable phases are shown as dashed lines (black). The probability density function and the lack of closure was calculated with amplitudes and a derivative structure factor without errors, a more realistic probability density function would be more Gaussian shaped, due to measurement errors.

More sophisticated phase probability distributions account for the incompleteness of the model (Sim weights) and positional errors of the model (Luzzati D-factor). The anti-correlating Luzzati factor and the probability distribution variance can be combined to a single parameter σ_A^2 which is an estimation of the model error. The σ_A^2 parameter ranges from zero to one, where one would be a perfect model.

Phase information from different sources, e.g. density modification and initial experiment phasing need to be combined. Phase probability functions can be expressed with Hendrickson-Lattman coefficients **(32)**.

$$(32) \quad \text{prob}(\varphi|A, B, C, D) = K \exp(A \cos \varphi + B \sin \varphi + C \cos 2\varphi + D \sin 2\varphi)$$

Where A, B, C, D are the Hendrickson-Lattman coefficients and K is a constant. The joint probability can then be expressed by a summation of the respective Hendrickson-Lattman terms **(33)(34)**.

$$(33) \quad \text{prob}(\varphi)_{\text{joint}} = \prod_i \text{prob}_i(\varphi)_1 \prod_j \text{prob}_j(\varphi)_2$$

$$(34) \quad \text{prob}(\varphi)_{\text{joint}} = N \exp(\sum_k A_k \cos \varphi + \sum_k B_k \sin \varphi + \sum_k C_k \cos 2\varphi + \sum_k D_k \sin 2\varphi)$$

Where $(\varphi)_1$ and $(\varphi)_2$ are a sets of phases from different sources, k are the according coefficients (1,2) and N is a normalization constant.

10. Structure Refinement

When initial phasing and phase improvement is successful, the reconstructed electron density should be good enough to start model building of the protein structure. The structure refinement process aims to build the most accurate structure with the available data. The electron density map is calculated with measured amplitudes containing errors and initial phases with uncertainties from the phasing technique. The model structure is limited by the quality of the measurements and the available resolution, whereas the phases improve with every “correctly” placed atom. In addition to having a good correlation with the measured data, a structure must also fulfill chemical requirements regarding the primary amino acid sequence, bond distances, torsion angles, planarity, clashing etc. Structure refinement procedures use the experimental data and optimize the structure to fulfill both the electron density and chemically expected values.

A weighting can be applied to the refinement calculation which balances the model adjustments to fit the electron density and the restraints. The weighting is adjusted during the course of the model refinement and depends on the resolution and the quality of the data (35).

$$(35) \quad Q_{total} = w_{xray} Q_{xray} + Q_{geo} = w_{xray} \sum_{hkl} \frac{(F_{hkl}^{obs} - F_{hkl}^{calc})^2}{(\sigma_{hkl}^{obs})^2} + \sum_i \frac{(r_i^{obs} - i_i^{ideal})^2}{(\sigma_{r,i}^{obs})^2}$$

The residual Q_{total} represents the divergence between observed and calculated values, w_{xray} is the weighting term of the residual Q_{xray} of the experiment. The residual Q_{geo} of the restraint parameters r_i^{obs} , with their standard deviation σ and the ideal value i_i^{ideal} . For low-resolution data the X-ray term is down weighted as the data do not contain enough information to justify divergence from the ideal values.

To place an atom in the unit cell three, parameters for the position x , y , z need to be assigned and refined to optimize the structure. Symmetry equivalent atoms are not located at the exact same point in the unit cells throughout the crystal and in the course of the measurement. Different directional disorder of the crystallographic symmetry throughout the crystal can be addressed with anisotropy scaling. Vibration of atoms around their center of mass and displacement by disorder in the crystal, lead to a diffraction angle dependent attenuation, which can be modelled by a displacement factor (36).

$$(36) \quad f_s^B = f_s^0 T_s$$

Where the Debye-Waller factor T_s describes the attenuation of the atomic structure factor f_s^0 . The simplest model uses one parameter per atom describing an isotropic spherical displacement of every atom (37)(38).

$$(37) \quad T_s = e^{-B_{iso}(\sin \theta / \lambda)^2}$$

$$(38) \quad B_{iso} = 8\pi^2 \langle u_{iso}^2 \rangle$$

Where the B_{iso} -value is a measure for the squared displacement $\langle u_{iso}^2 \rangle$ around the mean position of an atom. Translation-libration-screw (TLS) parameterization applies an anisotropic displacement to groups of atoms using 20 additional parameters per group. An ellipsoid displacement of every atom needs six anisotropic displacement parameters per atom.

Atoms may be located at distinguishable points or only partially at one preferred location. This can be addressed with an additional occupancy parameter. For every additional location, additional x, y, z and displacement parameters are needed to described the model. The total number of parameters used in the refinement of the

structure are limited by the number of observations in the recorded data. This number has to be lower than the number of independent observations otherwise the model will be over-fitted. The independent or unique measured reflections can be counted from the indexed diffraction spots considering the given space group.

The R-value (39) compares the measured and calculated structure factor amplitudes and is used to inspect the progress of the refinement process.

$$(39) \quad R = \frac{\sum |F^{obs}| - k |F^{calc}|}{\sum |F^{obs}|}$$

A successful model building and refinement process is characterized by a decreasing R-value until a certain point is reached where the refinement does not further improve the model. To address over-parameterization, a small percentage of the measured structure factors are kept aside for validation purpose exclusively and are never used for model building and refinement. These structure factors are used to calculate the R_{free} -value (validation) while the rest is used to calculate the R_{work} (refinement) [29]. If the model is over refined the calculated structure factors will fit to the working set structure factors amplitudes and R_{work} will decrease, but the validation set (“free”) structure factor amplitudes will not fit with the model and the R_{free} will remain steady or might even increase. At an advanced stage of the model building and refinement, other parameters regarding the plausibility of the structure become important. These other parameters are the deviation from the ideal values of the bond length and angle, the planarity of planar groups, steric clashes of the structure, Ramachandran angles, cis-peptide bonds, coordination number of solvent atoms or ions and presence of electron density [30]. Outliers to the expected geometry need to be justifiable with the observed electron density.

A way to improve the data to parameter ratio are restraints. Constraints are “hard” restraints directly reducing the number of parameters. Here, parameter values are defined by other parameters or as constant. An example is planarity, where once

the plane is defined, all atoms of the planar group need to be on that plane. Another constraint is the sum of the occupancy of a residue with different conformations, which is usually set to one thereby reducing the occupancy parameters by one. Displacement factors can be defined as grouped displacement factors, e.g. applying a single displacement factor for all atoms of a residue. Rigid body refinement is an example, where a molecule is moved at once with constant atom positions within the molecule. NCS related molecules can be refined with a strict NCS (fixed rotational and translational symmetry) not allowing deviations between the NCS related molecules. "Soft" restraints on the other hand introduce stereochemical information serving as additional observations. Here, values are expected to be mean standard values with their variance commonly observed in protein structures. NCS for example, can also be refined using restraints allowing the NCS related molecules to deviate from each other by specified values. Moreover, displacement factors are depending on each other, where backbone atoms are commonly show lower displacement as sidechain and a value cannot be completely different the values of its linked atoms. If displacement factors are refined anisotropically, the movement is restricted mainly in the direction where bonds to other atoms exist. In summary, with higher resolution the number of measured observables increases, which allows to reduce the number of assumptions in the refinement process.

II. Flagellin from *E. coli* Nissle

1. Introduction

1.1. Intestinal Immune System

The gastrointestinal tract (GIT) has the important task to digest food and deliver nutrients through the epithelium into the body. The surface of the GIT acts as a barrier, separating inside and outside while allowing the transport of substances in one direction. Thereby, the epithelium surface is in close contact to a vast number of microbes inhabiting this resourceful niche. In a healthy individual, the microbiota lives in homeostasis with the host, where both influence and benefit from each other. The microbiome produces and supplies the host with vitamins K, B2, B9, B12 and metabolites such as short-chain fatty acids, while the host supplies nutrition, thereby influencing the composition of the microbiome [31, 32]. The commensal bacteria also protect from occasionally uptaken pathogens by occupying the biological niche in high number. Bacteria can secrete different antibacterial and immune stimulating substrates, protecting from other organisms and shaping the immune system [33]. The host senses permanently the microbiome and prevents uncontrolled growth by secretion of antibacterial peptides, bacteriolytic enzymes and IgA antibodies [34]. Goblet cells, the second most abundant cells in the GIT secret mucus, protecting the epithelial cell layer. This mucus is permanently secreted and augmented with antibacterial compounds keeping microbes away from the epithelial barrier [35, 36]. Pathogenic bacteria that breach the epithelial cell layer encounter immune cells resident in the lamina propria beneath the epithelial cell layer. This can lead to inflammation where the innate and adaptive immune system clear the pathogenic bacteria and support restoring the barrier integrity.

1.2. Inflammatory Bowel Disease

A dys-regulation of the immune response to the intestinal microbial flora can result in inflammatory bowel disease (IBD) [37]. More than 160 associated loci have been identified in genome wide associated studies with IBD patients [38, 39]. Beside a genetic predisposition, many other factors e.g. immune modulating drugs, surgery, transplantation, antibiotics, dietary, psychological stress, exposure to chemicals may be associated with an increased risk in developing IBD [40, 41]. IBD reduces the quality of life significantly, with impact on many aspects of usual life activities e.g. employability, education, social and interpersonal functioning [42], caused by symptoms such as abdominal pain, rectal bleeding, diarrhea, anemia, weight loss, fatigue and more [43]. IBD patients have a 0.5% per annum higher mortality rate [44], where IBD related colorectal cancer is responsible for 10-15% of the annual deaths [45]. The irritable bowel syndrome (IBS) associated with bacterial overgrowth and a gut motility disorder is described with similar but not as severe symptoms and an apparently normal mucosa. Both diseases, IBS and IBD share overlapping mechanisms such as increased gut permeability or dysbiosis [46]. IBD comprises a spectrum of diseases, where the main subtypes are Crohn's disease and ulcerative colitis (UC). The inflammation in Crohn's disease can affect any part of the GIT with a patch-like or segmented transmural inflammation pattern. In contrast UC is characterized as a mucosal inflammation affecting the rectum and varying lengths of the colon proximal to the rectum [43, 47]. Both subtypes are chronic diseases, cycling between inflammation flares and remission.

1.2.1. IBD Treatment

Beside genome wide association studies, mouse models are important for the understanding of IBD and the development of new treatments. Here, gene knockout mouse models and germ free mice have contributed to identify and validate key molecules, signaling pathways, effect of microorganism and cell types important for disease susceptibility, development and progression [48, 49].

Disease models such as the most commonly used dextran sodium sulfate (DSS) induced colitis are useful to investigate the effect of possible treatments [50, 51]. Different treatment strategies focus on anti-inflammatory drugs, immunomodulation, microbe modulation, symptom treatment or surgery. Besides, the well-established treatment with 5-aminosalicylic acid preparations (mesalazine) [52], antibodies targeting immune system modulators, e.g. tumor necrosis factor (TNF) [53] or interleukins 12/23 (IL-12/IL-23) [54], can be used. The probiotic bacteria *Escherichia coli* Nissle 1917 (EcN) seem to be equally effective in maintaining remission than mesalazine in UC [55, 56]. In the DSS induced colitis mouse model EcN prevented colonic damage [57-59] which could be linked to the main flagellum protein flagellin (FliC)[60].

1.2.2. Flagellin Detection

Monomeric FliC is detected by the host innate immune system by two different receptors located in the cytosol and at the cell surface. The cytosolic system is composed of two “nucleotide binding domain and leucine rich repeat containing proteins” (NLRs), namely the “NLR-family of apoptosis inhibitory protein” (NAIP) and the “NLR-family caspase recruitment domain containing protein 4” (NLRC4), which together form the NAIP-NLRC4 inflammasome [61, 62]. The system located at cell surface utilizes the Toll like receptor 5 (TLR5) [63] which detects monomeric FliC shed from the bacterial flagella. The flagellum is an important pathogenic factor for bacteria enabling them to move along gradients and penetrate the mucus covering the host surface. Beside the motility, depending on the pathogen, the flagellum mediate other functions, such as cell adhesion, mechano-sensing or biofilm formation [64]. The domains that build up the inner core of the flagellum are detected by TLR5. This domains exhibit a high degree of conservation between different bacteria, which allows the detection of a broad range of different bacteria. The genomic deletion of the TLR5 results in spontaneous colitis in mice [65] and hence might be an important factor in IBD.

1.3. Toll Like Receptors (TLRs)

Toll like receptors (TLRs) play a key role in the innate immunity and are linked to the adaptive immune system. These pattern recognition receptors (PRRs) detect microbe associated molecular patterns (MAMPs), where different TLRs recognize a broad range of different substrates, e.g. flagellum protein (TLR5), double-stranded RNA (TLR3), single-stranded RNA (TLR7, TLR8), lipopolysaccharides (LPS)(TLR4), lipoproteins (TLR2+TLR6 or TLR1) or non-methylated CpG DNA (TLR9). The overall structures and functions of different TLRs are comparable, they are single pass transmembrane receptors with a horseshoe-like shaped ectodomain consisting of leucine rich repeats (LRR) and a cytosolic Toll/interleukin receptor (TIR) domain. If the TLR specific MAMP is present a dimerized form is stabilized, where the C-terminal parts of the TLRs are in close proximity, resulting in a conformation change of the cytosolic TIR domains [66].

1.3.1. TLR-MAMP Interaction

The MAMPs, recognized by TLRs, are completely different molecular structures, with different TLR binding modes, but they all lead to a stabilization of either homo or hetero dimers. Double stranded RNA binds between two TLR3s at both sides of the horseshoe shape bridging the dimer together, lipopeptides are bound by a TLR1 and TLR2 heterodimer at the more C-terminal part of the interface. LPS on the other hand are bound by the co-receptor protein “myeloid differentiation factor 2” (MD-2), which are bound on each sides of a TLR4 dimer, with contacts to both monomers. TLR5 binds FliC monomers at the concave side of each horseshoe, while only few interactions are formed with the second TLR5. In the available crystal structure (3V47) only the D1 domains of the FliCs interact with the TLR5s while the hypervariable domains are pointing away from the TLR5-FliC heterodimer [67]. The TLR5 senses the highly conserved part responsible for the inner helical structure of the flagellum, which cannot be altered easily by the organism. Every major change in this region effects the stability of the flagellum

and therefore the motility of the organism. When compared with other TLRs, the binding interface that stabilizes the dimer, appears to be rather small (**Table 3**)(**Figure 12**), but the dimerization interface might be larger as the crystal structure does not comprise the full TLR5 (residues 22-181). Additionally, it was shown that the D0 domain of FliC is needed to activate the TLR5 receptor [68]. The D0-D1 domain transition cannot adopt the same structure as in the assembled flagellum as the D0 domain would clash with the cell membrane. The D0 might form additional interactions with the C-terminal part of the TLR5 or interact with the D0 domain bound by the opposed TLR5. Such a dimerization of the N-terminal with the C-terminal D0 plays a role in a different context, the export of FliC into the flagellum channel in the assembly process [69].

Table 3: Interface areas between TLRs and MAMPs calculated with the PISA server [70]. TLR-MAMP represents the main interface between a MAMP and a TLR monomer, while TLR*-MAMP represents the interface of the MAMP to the second TLR, forming the dimer. Dimer_{GAIN} is the total MAMP mediated interface gained through dimerization.

	TLR-MAMP	TLR*-MAMP	Dimer_{GAIN}
TLR3	1153 Å ²	1132 Å ²	(1132 Å ²)
TLR4	2×1016 Å ²	2×487 Å ²	974 Å ²
TLR12	1012 Å ²	475 Å ²	475 Å ²
TLR5	2×1637 Å ²	2×171 Å ²	342 Å ²

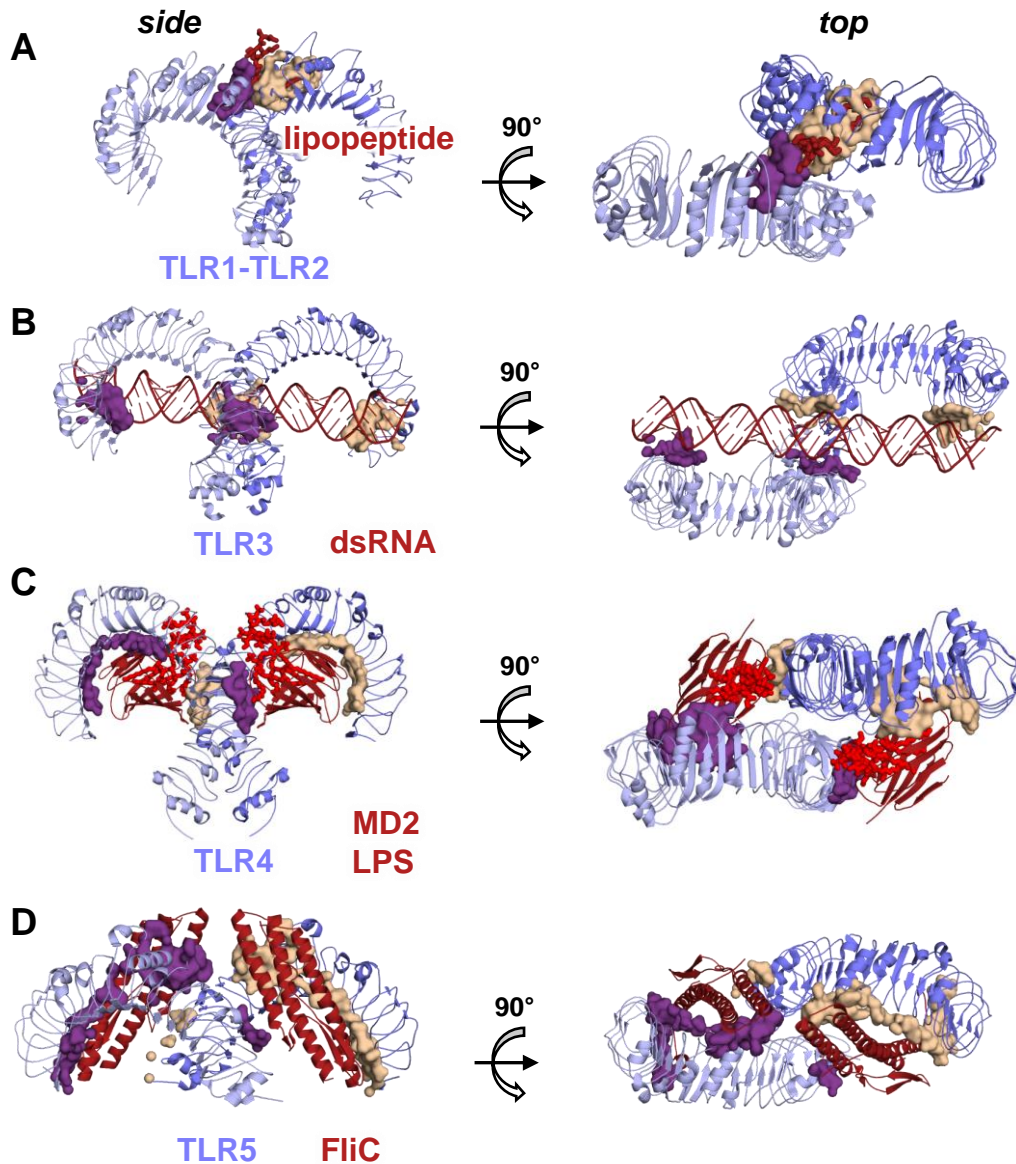


Figure 12: TLR dimers with their ligands. TLRs are shown in blue and ligands in red. The binding interface between ligands and TLRs is shown in surface representation. The interfaces are colored depending on the TLR monomers they originate from, either brown or purple. On the left, the TLRs are shown from the side with the C-terminal part in the middle pointing towards the cell surface. On the right, the view is rotated 90° as indicated. **A:** A lipopeptide interacting with a TLR1 and a TLR2 heterodimer (2Z7X). **B:** A dsRNA interacting with a TLR3 dimer (3CIY). **C:** The co-receptor MD2 with bound LPS interacting with TLR4 (3FXI). **D:** Two FliC monomers binding to a TLR5 dimer (3V47).

1.3.2. TLR Signaling

On the cytosolic side of the activated TLR, a TIR domain dimer acts as platform for other TIR-domain containing adaptor proteins. Myeloid differentiation primary response protein 88 (MyD88) is the most important adaptor protein interacting with most TLR-TIR domains. MyD88 contains besides a TIR-domain and an inter-domain region (~45 AAs) a death domain (DD) that allows interaction with other DD containing proteins (**Figure 13**). The “interleukin-1 receptor-associated kinase 4” (IRAK-4) binds to MyD88 utilizing its DD and recruits IRAK-1 or IRAK-2. The activation and association of the different IRAKs involves auto phosphorylation and cross phosphorylation. The complex formation on the platform of the TLR dimer consists of 7-8 MyD88 forming the first two layers, four IRAK-4 in the third layer and four IRAK-1 or IRAK-2 in the fourth layer [71]. The adapter proteins might allow higher order oligomeric structures linking different TLRs [72]. In the next step of the cascade the E3 ubiquitin ligase “TNF receptor associated factor 6” (TRAF6) binds to the complex which leads to K63 poly-ubiquitination [73, 74]. After dissociation of TRAF6/IRAK1, the “transforming-growth factor β activated kinase 1” (TAK1) [75], the “TAK1-binding protein 1” (TAB1) and TAB2 bind to TRAF6/IRAK1. IRAK1 is ubiquitinated and degraded while the complex of TRAF6/IRAK1/TAB1/TAB2/TAK1 translocates to the cytoplasm [76]. The “ubiquitin conjugating enzyme-13” (UBC-13) and “ubiquitin conjugating enzyme E2 variant-1” (UEV-1a) bind to the complex in the cytoplasm [77]. This poly-ubiquitinated complex activates “mitogen-activated protein kinases” (MAPKs), leading to the activation of the transcription factors p38 and the “Jun-N-terminal kinase” (JNK). Additionally, the I κ B kinase-complex (IKK) [78, 79] is activated, resulting in the degradation of the inhibitory I κ B α which then enables the release and translocation of the transcription factor (NF- κ B). This important transcription factors have a variety of effects depend on the cell-type, state of a cell, and other cell signals. Generally, the effects of this transcription factors allows the organism to cope with stress, playing an important role in inflammation, apoptosis, cell growth, and differentiation.

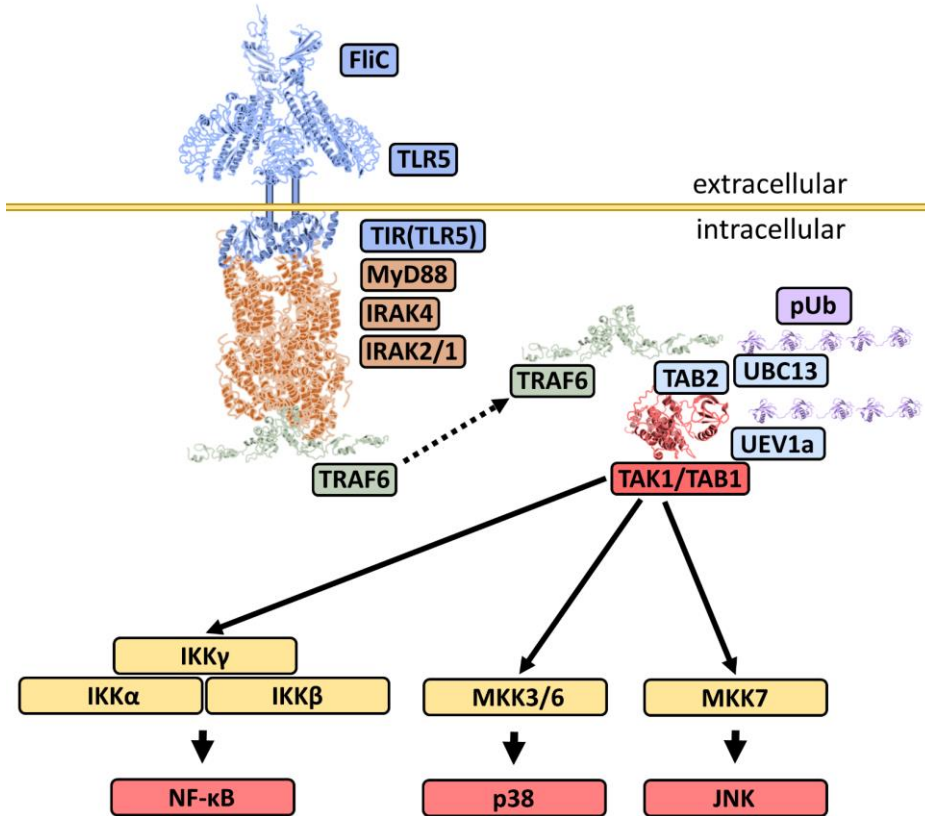


Figure 13: TLR5 signaling. FliC binds to TLR5, stabilizing the dimer, whereby the intracellular TIR-domains come in close proximity (3V47, 2J67; blue). This allows binding in the order MyD88, IRAK4, IRAK2 or IRAK1 (3MOP; orange). IRAK1 or IRAK2 activate TRAF6 (3HCS; green) which ubiquitinates itself and dissociates from the complex. The ubiquitin (2W9N; purple) allows binding of TAB2, UBC13, UEV1a (light blue), TAB1 and TAK1 (2EVA; red). This complex then activates MAPKs and IKK resulting in NF- κ B, P38 and JNK activation. This is a schematic figure of different available structures, neither are the shown interactions based on experiments, nor are the structures complete, from the same organisms or on a common scale.

1.3.3. TLR5 and IBD

The immune reaction caused through an external stimuli such as FliC is complex and depends on many aspects e.g. cell type, receptor localization, cell-cell signaling, presence of danger associated molecular patterns and presents of other pathogen associated molecular patterns. TLR5 is expressed by intestinal epithelial cells (IEC) and immune cells such as residual dendritic cells, macrophages and T-cells in the lamina propria. The TLR5 expression is also not homogenous throughout the GIT. In the small intestine TLR5 are not expressed in usual IEC, but in Paneth cells in the crypts of the small intestine, where they may help to protect the epithelial stem cells from pathogens. In the colon, TLR5 is expressed throughout the epithelial cells, coincident with the UC disease pattern. The localization of the TLR5 seems to be not restricted to the basolateral side [80] as previously reported [81, 82]. Upon TLR5 activation epithelial cells increase expression of chemokines e.g. “C-X-C motif chemokine 11” (Cxc11), Cxc12 and “C-C motif chemokine ligand 20” (Ccl20) which attract lymphocytes. Additionally, enzymes are produced that generate reactive oxygen species, which act antimicrobially or enhance inflammatory pathways [83, 84]. Other reactions in response to FliC detection, such as the production of antimicrobial peptides or epidermal growth factor, might depend on signals from immune cell of the lamina propria.

Special CD4⁺ T-cell subsets and dendritic cells seem to play a key role in the intestine immune homeostasis. T helper 17 cells (Th17) can produce interleukin 17 (IL-17), which can stimulate IEC to produce antimicrobial peptides [85]. Th17 cells also stimulate B cells to differentiate into IgA-secreting cells through IL-21 [86]. The translocation of the IgA through the epithelium is enhanced through IL-17 by increased Ig receptor production in IECs [86]. Th17 also show lower susceptibility than other T-cell subsets to the suppressive activity of T-regulatory cells (Treg) [87]. TLR5 is expressed in CD4⁺ T-cells and shows an effect upon flagellin stimulation. CD4⁺ CD25⁺ Treg cells express TLR5 and their suppressive capacity is increased by stimulation with FliC [88]. Treg-cells are suppressing effector T-cells, preventing inflammation and promoting tolerance. The Th17 on

the other hand rather promote inflammation and are important for pathogen clearance and mucosal barrier maintenance. A dysregulation of Th17 and Treg cells lead to inappropriate reaction of the immune system to the gut microbiota. Where low levels of Th17 cell signaling reduce antimicrobial peptide and IgA secretion of the IECs leading to an overgrowth. Also clearance on bacteria breaching the barrier is delayed, resulting in a more severe inflammation, affecting a larger area. When Treg cell suppression of Th1 and Th2 cells is not appropriate, barrier breaching bacteria result in strong inflammation damaging the gut more than needed. Dendritic cells permanently sense the microbiota in the intestine and present uptaken antigens to other immune cells. The dendritic cells influence the T-cell differentiation and thereby are important to shape the T-cell repertoire in order to gain homeostasis. Subsets of dendritic cells in the intestinal lamina propria can promote the development of Th17 cells through IL6 depending on TLR5 signaling [89]. Based on the bacterial stimuli and present regulatory cytokines DCs are activated and mature to different subsets which differ in their costimulatory molecule levels and signaling, which in turn effects the T-cells population [90].

1.4. Chemotaxis

Bacterial motility is achieved with different systems. On surfaces, some bacteria can move with protruding pili by binding and retracting the pili [91]. Other strategies to move on surfaces are thrust generation through extrusion of slime [92], or different mechanisms involving motor protein complexes [93].

In a liquid environment bacteria use flagella to swim freely. *E. coli* has 5 to 10 flagella randomly distributed on its surface. The thrust for movement is generated through the rotation of the long helical filamentous part of the flagellum [94]. The movement divides in two alternating modes, “run” and “tumble”. In the run mode all flagella rotate counter clockwise (CCW), resulting in the formation of a bundle of rotating flagella, generating thrust in one direction. In the tumble mode one or more flagella change their rotation direction to clockwise (CW) which distorts the bundle, resulting in an orientation change with no net movement. As a result, the

bacteria move in periods of straight movement interrupted by a random orientation change, mathematically describable as random walk. Combined with a sensor system they are able to adjust their locomotion to chemical gradients in their surroundings by changing the length of the straight movement periods, a phenomena called chemotaxis.

The switch in movement mode is regulated by a histidine-aspartate phosphorelay (HAP). Different chemotaxis proteins (Che) are involved in these processes (CheA, CheB, CheR, CheW, CheY, CheZ) (**Figure 14**). The tumble mode is initiated by the auto phosphorylation activity of CheA, which transfers the phosphoryl group to the response regulator (RR) CheY. The RR binds in the phosphorylated state to flagellin M (FliM), which is part of the flagella motor, causing a change in the direction of rotation. The RR is dephosphorylated to its inactive state by CheZ, resulting in a switch of the rotation direction back to CCW, ending the tumble mode. This auto phosphorylation activity of CheA is influenced by external signals and the elapsed time. A “clusters of methyl-accepting chemotaxis protein” (MCP) serve as sensors and are associated through CheW with CheA [95]. If a chemo attractant binds to a MCP the autophosphorylation rate of CheA is reduced, prolonging straight movement in the run mode. The sensitivity level of the sensors is adapted to activation levels in the near past. This is archived by the methylation dependent sensitivity of the MCPs. CheR methylates MCPs with a constant rate, thereby increasing CheA autophosphorylation rate. The phosphoryl group of CheA is transferred not only to the RR CheY, but also to CheB. If CheB is phosphorylated, it demethylates MCPs, reducing the auto phosphorylation rate of CheA in a negative feedback loop. An attractant increases the straight movement phase and decreases the sensitivity resulting in a gradient sensitive receptor over a wide concentration range of attractants [96].

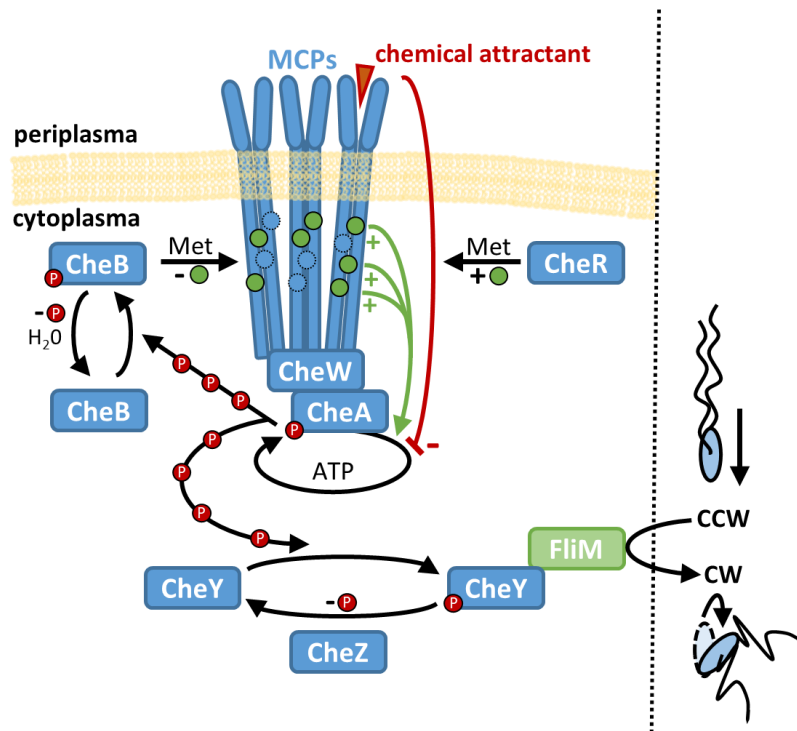


Figure 14: Proteins involved in chemotaxis. Different transmembrane MCPs sense chemicals in the periplasm. If an attractant binds to its MCP, the auto-phosphorylation activity of CheA is reduced. Auto phosphorylation also depends on the methylation state of the MCP. CheR methylates MCPs with a constant rate, increasing CheA activity, while phosphorylated CheB demethylates MCPs, decreasing CheA activity. If CheA is active, it transfers its phosphorylation to CheY or CheB. Phosphoryl-CheY can bind to FliM which changes the rotation direction of the flagellum to CW. CheZ dephosphorylates CheY and the rotation direction is changed back to CCW.

1.5. Flagellum Components

The rotating part of the bacterial flagellum called the rotor mainly consists of three parts: the rod, the hook and the filament. The rod is integrated in a stator (force generation), and reaches through the outer membrane and peptide glycan of the bacterium. The hook joints the rod with the filament. Its property of axial flexibility and twisting stiffness allows force transmission from the fixed rod to the filament in different orientations. The filament consists of only one protein, FliC, which makes up the most part of the flagellum.

Motor proteins as part of the stator generate torque at the rod, the hooks allows an axial reorientation, whereas the filament allows force transition to the medium resulting in a thrust along the filament axis [97]. The filament changes from a left handed supercoil to a right handed supercoil as rotation changes from CCW to CW. Due to the low Reynolds number, physical properties of bacterial movements are quite different then in a macroscopic world. The Reynolds number is the relation of the velocity of an object to the viscosity of the medium. For swimming bacteria, the viscosity of the media dominates inertia of the bacterium, leading to an “instant” stop if the accelerating force vanish.

1.6. Flagellum Growth

The flagellum consists of up to 20'000 subunits of FliC. When the flagellum growth, every subunit has to be exported trough a narrow channel with a diameter of about 20 Å. Subunits of the rod, hook, and filament are translocation and unfolded by a type III export machinery. The energy needed for this process is mainly delivered by the proton motive force, supported by ATP hydrolysis through the ATPase FliI. The narrow space inside the flagellum prevents folding of the subunits until they reach the tip of the polymer, where they rapidly fold and polymerize with the pre-existing polymer. The tip of the growing flagella filament is protected by a cap, which has a disc-like structure, with α -helical arms that interacts with the FliC. The

cap itself consists of FliD, which has been described to form a pentamer (*Salmonella* serovar Typhimurium) or hexamer (*Pseudomonas aeruginosa*)[98]. For both described oligomeric forms, there is a mismatch with the flagellum filament, which itself has an 11-fold helical symmetry with staggered monomers in a nearly 5.5-fold axial symmetry. It is thought that this mismatch of the α -helical arms with the filament generate an open space for the polymerization of the next monomer while other binding positions are blocked [99]. Through the folding and binding of a monomer at the free position, the whole cap is rotated and pushed to the next position so that the next position is free for another monomer. The cap acts as chaperon for both, the folding of monomers and the polymerization process, by guiding the monomers to their position, and protecting the growing tip of the polymer from misfolded or foreign proteins. It also protects from losing monomers to the surrounding, protecting, beside of the net loss of a building block, from the detection through the immune system. The folding and polymerization process of the monomers is thought to provide a drag force to the following monomers in the central channel allowing a fast and length independent growth [69] of the flagellum. This dragging momentum is passed through the whole flagellum channel through the parallel coiled coil association of the C-terminal to N-terminal parts of subsequent D0 domains. This linkage takes place at the export gate of the type III export machinery. FliC with its bound chaperon FliS docks at the ATPase ring complex (FliI, FliJ, FliH) and transits to the export cage (FliA)(**Figure 15**). FliJ transduces energy gained by ATP hydrolysis to the export cage ring, facilitating the release of the FliS chaperon. The terminal FliC part then binds to FliB of the export gate (FliA, FliB, FliO, FliP, FliQ, FliR) where a stronger binding with the terminal part of the previously exported FliC is established. When, at the other end of the flagellum, a monomer folds and polymerizes, the dragging force applies stress to the FliC-FliB bond resulting in the release of FliC into the channel.

The flagellum growth is even more complex as described here. The rod (FlgE, FlgB, FlgC, FlgF, FlgG), hook (FlgE) and hook-filament-junction (FlgK, FlgL) and filament (FliC) -proteins have to be exported in a controlled way. A substrate switch of the export gate protein FliB is expected when the hook reaches a certain length.

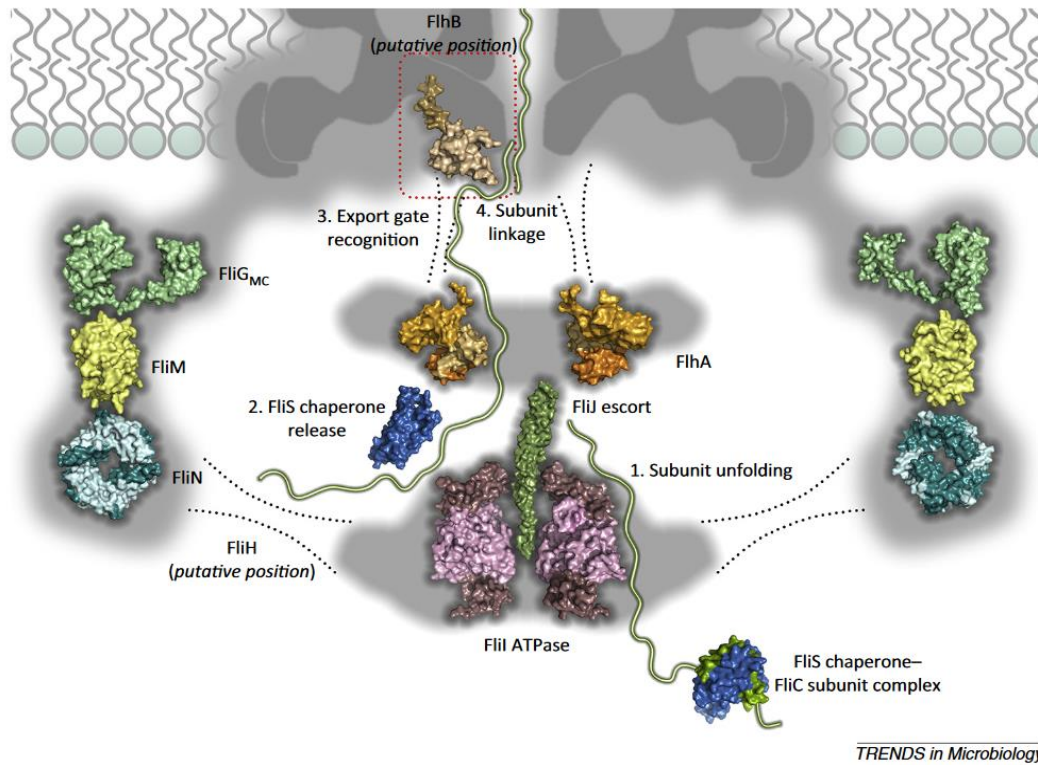


Figure 15: Interactions between FliC subunits with export machinery. The C ring (FliG:1LKV, FliM: 2HP7, FliN: 1YAB) is associated with the rotary ATPase complex through FliH (proposed position is indicated by dotted lines). The ATPase complex (FliI: 2DPY, FliJ: 3AJW) binds the FliC-chaperon complex (FliS-FliC: 1ORY) before entering the export cage (FliA: 3A5I) where FliS is released. The terminal part of FliC binds FliB at the export gate where then FliC subunits are linked. The Figure was adopted from Evans et al. 2014 [100].

The “ruler” protein FliK is assumed to probe the length of the hook and trigger at a given length a structural change in FliB switching substrate specificity from hook-like to filament-like proteins. Also, three different cap proteins are involved in the assembly stages, the rod cap FlgJ which has a muraminidase activity to penetrate the peptidoglycan, the hook cap FlgD, and the filament cap FliD. The different filament proteins also have different chaperons with different affinities for the export machinery [101].

1.7. R-Type and L-Type Flagellum

The FliC monomers of the flagellum change their conformation when the rotation direction of the rotor is changed. Two mutants have been designed, both forming straight flagellums, but with a different twist. These mutants are locked in different conformations forming either R-type or L-type filaments [97, 102, 103]. The inner core consisting of D0 is almost similar between the two types, but the orientation of D1 is different (**Figure 16**). The R-type and L-type conformations are associated with a CW or CCW rotation direction, respectively (**Figure 17**). Wild type flagellums consist of a mixture of R-type-like and L-type-like monomers which favor a different twist. The different twist prevalence is compensated by a curvature of the flagellum [104]. The ratio of the different types affects the strength and direction of the curvature. When the rotor changes the rotation direction, a number of flagellum monomers will change their conformation, this conformation change will then propagate over the whole flagellum, changing the overall curvature direction. The curvature of the flagellum enables the transformation of torque in linear force allowing to generate thrust.

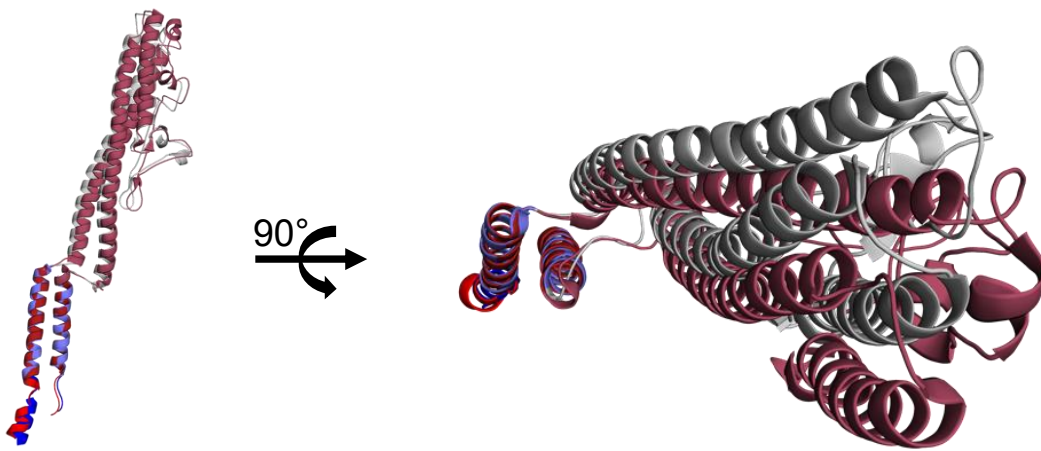


Figure 16: Alignment of FliC monomers originating from R-type or L-type flagellums. Cartoon representation of R-type FliC (5WJT) (blue, grey) and L-type FliC (5WJY) (red), viewed from the side and rotate 90° as indicated. Comparing R- and L-type flagellums shows that the D0 is similar in both types with an rmsd of 0.57 Å, where the first 9 amino acids and the last 11 amino acids show variation between the types (bright blue and bright red). Between the two D1s is a shift, where the R-type (grey) is shifted CCW compared with the L-type (light red), when viewed from tip towards rotor.

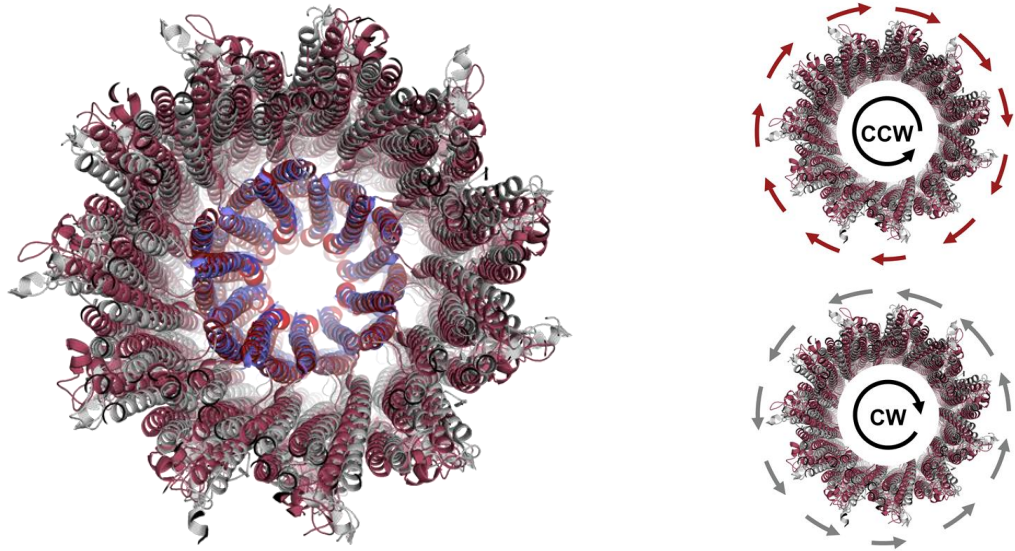


Figure 17: D0 alignment of R-type and L-type flagellums. D0s in the inner core of the flagellum do not change significantly when the flagellum switches from R-type (5WJT)(blue) to L-type (5WJY)(red), where the first and last residues of D0 do change (bright blue and red). The second layer of the core formed by D1 does shift significantly, the R-type D1s (grey) are orientated more CCW than the L-type D1s (light red), when viewed from the tip to the rotor of the flagellum. The D1 is oriented with the rotation direction CCW for the L-type and CW for the R-type, indicated with arrows (right side).

2. Objectives

A. Steimle and S. Menz from the group of Julia-Stefanie Frick and others found that there is a beneficial effect of *E. coli* Nissle (EcN) on the outcome of DSS induced colitis in mice, which they could attribute to the flagellin of this strain [60]. They hypothesized that the long hypervariable domain leads to stronger TLR5 signaling (**Figure 18**). They started a collaboration with us to gain structural information of the flagellin to identify the molecular mechanism influencing the TLR5 signaling. This investigation could lead to the development of a therapeutic for the treatment of IBD patients.

The hypervariable domains of flagellins are diverse between different *E. coli* strains whereas no structures are available so far. The structure of flagellin from EcN could be a first step in investigating structural and functional diversity within this species. Such investigations would help to associate structures and functions of unknown flagellins based on sequence similarity. In diagnosis, a broad understanding of flagellins might be of benefits in evaluating the intestinal microbiome.

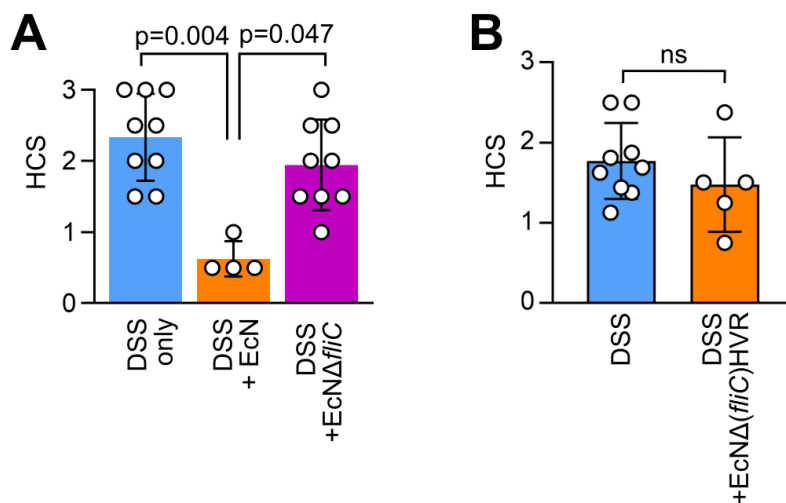


Figure 18: Adopted from Steimle *et al.* [60] (Fig 1d, Fig 4c) showing histological colitis scores (HCS) from colonic sections 7 days after DSS administration. **A:** Mice with no additional treatment, EcN or EcN:*fliC* deletion mutant or **B:** EcN:*fliC* hypervariable region deletion mutant administration.

3. Materials and Methods

3.1. Materials and Buffers

3.1.1. Chemicals

The chemicals used in this work were of analytical grade and obtained from: Sigma-Aldrich (Deisendorf, Germany), Roth (Karlsruhe, Germany), GE Healthcare (Uppsala, Sweden), Merck (Darmstadt, Germany) or Hampton Research (Aliso Viejo, USA).

3.1.2. Vector

For all constructs the pET-15b (Novagen, Merck, Darmstadt, Germany) vector was used. FliC_E, FliC_ETEV, FliC_ΔD0D1, FliC_ΔD4, FliC_ΔD4s and FliC_ΔD3D4 was provided by *T. Hagemann* (J. Frick, IMIT Tübingen, Germany). FliC_ΔD0 and FliC_ΔD0D3D4 was bought from Biocat (Heidelberg, Germany) which used restriction cloning with NcoI and BamHI and a synthetic insert. All other constructs were modified from existing constructs as described below (3.2.5).

3.1.3. Bacterial Strains

The *E. coli* BL21 strain was used for protein production, and the *E. coli* DH5α strain for the production of DNA (**Table 4**).

Table 4: *E. coli* strains used in this project. To amplify plasmid DNA the *E. coli* strain DH5α was used. Protein was production using the the *E. coli* BL21 (DE3) strain.

<i>E. coli</i> Strain	Genotype
DH5α	F ⁻ φ80 <i>lacZ</i> ΔM15 Δ(<i>lacZYA</i> -argF) U169 recA1 endA1 hsdR17 (r _k ⁻ , m _k ⁺) <i>gal phoA supE44 λ thi1 gyrA96 relA1</i>
BL21 (DE3)	F ⁻ <i>ompT hsdS_B</i> (r _B ⁻ m _B ⁻) <i>gal dcm</i> (DE3)

3.1.4. Buffers

Standard buffers were used for the immobilized metal ion affinity chromatography (IMAC) (**Chapter 3.3.3**) and size exclusion chromatography (SEC) (**Chapter 3.3.6**) for all FliC constructs (**Table 5**).

Table 5: Buffers used for the purification of FliC. Two different size exclusion buffers were used, for crystallization experiments the SEC-buffer, and for biological experiments DPBS-buffer. The pH of the Buffers were adjusted at 4°C.

Buffer	Components
Lysis Buffer / His _A Buffer	50 mM Tris pH 8.0 300 mM NaCl
His _B Buffer	50 mM Tris pH 8.0 300 mM NaCl 500 mM imidazole
SEC-Buffer	20 mM HEPES pH 7.5 150 mM NaCl
DPBS-Buffer	0.90 mM CaCl ₂ 0.49 mM MgCl ₂ 2.67 mM KCl 1.47 mM KH ₂ PO ₄ 137.93 mM NaCl 8.06 mM Na ₂ HPO ₄

3.1.5. Commercial Crystallization Screens

Initial crystallization screening experiments were done with commercially available screens (**Table 6**).

Table 6: Commercially available screens, used for initial crystallization experiments.

Screen	Company
JCSG	Molecular Dimensions, Suffolk, UK
Morpheus	Molecular Dimensions, Suffolk, UK
Wizard I- IV	Emerald BioSystems, Bainbridge Island, USA
Crystal Screens I, II	Hampton Research, Aliso Viejo, USA
PEG ION	Hampton Research, Aliso Viejo, USA

3.2. Molecular Biology

3.2.1. Preparation of Chemically Competent Cells

LB-medium was inoculated 1:100 with an o/n culture of *E. coli* and grown at 37°C while shaking until an OD₆₀₀ of 0.4-0.5 was reached. After 5 min incubation on ice, the cells were centrifuged at 3 000 × g for 10 min at 4 °C. The pellet was washed twice with 20 mL ice cold 0.1 M CaCl₂, resuspended in 4 mL 0.1 M CaCl₂, and incubated for 1 h on ice. Glycerol was added to a final concentration of 10 % (v/v). The cells were flash frozen in liquid nitrogen and stored at -80 °C.

3.2.2. Transformation of Competent Cells

Chemically competent cells (50 µL) were mixed with 1 µL (~100 ng) of plasmid DNA, or 9 µL of ligation mixture and incubated for 20-30 min on ice. After a heat shock at 42 °C for 30-45 s, followed by 2 min incubation on ice, 400 µL LB-medium (**Table 7**) was added and cells were grown for 1 h at 37 °C while shaking to develop

Amp resistance. Cells were then either used to inoculate a pre-culture, or plated on LB-agar containing 50 µg/mL Amp.

Table 7: LB-Media mixture.

LB-Media	LB-Agar Media
1% (w/v) tryptone	1% (w/v) tryptone
0.5% (w/v) yeast extract	0.5% (w/v) yeast extract
1% (w/v) NaCl	1% (w/v) NaCl
	6% (w/v) agar

3.2.3. Glycerol Stocks

Bacterial pre-cultures were mixed with 50% (v/v) glycerol, to a final concentration of 12.5% (v/v) glycerol, flash frozen in liquid nitrogen, and stored at -80°C.

3.2.4. Plasmid DNA Isolation

Plasmid DNA was isolated from o/n cultures using a Wizard Plus SV Miniprep kit (Promega, Mannheim, Germany), according to the manufacturer's protocol. The DNA concentration was measured using the UV-absorbance at $\lambda = 260$ nm, according to the Lambert-Beer law **(40)** (NanoDrop ND-1000, Thermo Fisher Scientific, Waltham, USA).

$$(40) \quad E_{\lambda} = \epsilon_{\lambda} \cdot c \cdot d$$

- E_{λ} : Absorbance
- ϵ_{λ} : decadic absorbance coefficient
- c: protein concentration
- d: path length

3.2.5. Site Directed Mutagenesis

Primers (Thermo Fisher Scientific, Waltham, USA) were designed to amplify the vector including the region of interest. Additional parts were introduced using primer with overhangs. The PCR program (Primus96plus, MWG-Biotech AG, Ebersberg, Germany) and reaction mixture are shown in **Table 8** and **Table 9**, respectively. The template DNA is methylated by *E. coli*, from which it was purified. This methylation was exploited by digestion of the template DNA for 1 h at 37 °C with 1 U/μL DpnI restriction enzyme (NEB, Frankfurt, Germany). The 5' ends were phosphorylated for 30 min at 37 °C using 2 μL reaction mixture, 0.5 μL T4 DNA PNK (NEB, Frankfurt, Germany) and 0.5 μL T4-buffer adjusted with water to a total volume of 5 μL. The phosphorylated DNA was ligated using a T4-ligase (NEB, Frankfurt, Germany) according to the manufacturer's protocol. The whole ligation mixture was used to transform chemically competent DH5α *E. coli* cells.

Table 8: PCR program. Cycle step 2-4 was repeated 30 times. Annealing temperatures (T_A) between 62 and 65 °C were used.

Cycle-Step	Time [s]	Temperature [°C]
1	300	98
2	45	98
3	45	T_A
4	420	72
5	600	72

Table 9: PCR reaction mixture. Q5 high fidelity polymerase (NEB, Frankfurt, Germany) was used.

Component	Volume [μL]	Concentration
H ₂ O	12	
5x Q5 Buffer	4	1x
10 mM dNTPs	0.8	400 μ M
10 μ M Primer for.	1	0.5 μ M
10 μ M Primer rev.	1	0.5 μ M
Template DNA	1	< 250 ng
Q5 Polymerase	0.2	0.2 U

3.2.6. DNA Sequencing

The DNA sequence was determined using Sanger sequencing [105] by Microsynth Seqlab (Göttingen, Germany) using T7 promoter and/or T7 terminator primers.

3.3. Protein-Biochemistry

3.3.1. Protein Production

For a pre-culture, the clone of interest was transferred in 10-60 mL LB-media containing 50 µg/mL Amp and incubated at 37 °C o/n while shaking. 1-6 L LB-media containing 50 mg/µL Amp was inoculated with the pre-culture 1:100. Cells were grown at 37 °C, while shaking, until an OD₆₀₀ of 0.6-0.7 was reached. The protein production was induced with 0.2-0.5 mM IPTG and the cells were shaken at 25 °C o/n or 37 °C for 4 h. Cells were harvested by centrifugation for 15 min at 9 200 × g (Sorval RC 6+, Thermo Fisher Scientific, Waltham, USA). The pellet was frozen in liquid nitrogen and stored at -80 °C. The SelenoMethionine medium complete kit (Molecular Dimensions, Suffolk, UK) was used to produce selenomethionine-containing protein, according to the manufacturer's protocol.

3.3.2. Cell Lysis

Cell pellets were resuspended in 4-6 mL Lysis buffer (**Table 10**) per gram pellet wet-weight. The suspension was supplemented with 1x “cOmplete” protease inhibitor (Roche, Basel, Switzerland) and 125-500 U Benzonase Nuclease (Merck KGaA, Darmstadt, Germany). Cells were lysed through sonication with an amplitude of 40%, for 2-3 min, with a cycle of 0.5 s pulse on followed by 0.5 s pulse off (Digital Sonifier 250, Branson Ultrasonics, Danbury, USA). The lysate was centrifuged for 45 min at 34 500 × g (Sorval RC 6+, Thermo Fisher Scientific, Waltham, USA).

Table 10: Lysis Buffer. Buffer used for protein purification, the pH was adjusted at 4°C.

Lysis Buffer
50 mM Tris pH 8.0
300 mM NaCl

3.3.3. Immobilized Metal Ion Affinity Chromatography (IMAC)

Filtered lysate was loaded with a flow rate of 0.5 mL/min on a 1 mL or 5 mL HisTrap FF crude column (GE Healthcare, Frankfurt, Germany) equilibrated with His_A buffer (Table 11) using an ÄKTAprime plus FPLC system (GE Healthcare, Frankfurt, Germany). After sample loading the column was washed with His_A and 2% His_B Buffer (10 mM imidazole) until the UV_{280nm} absorbance reached baseline. Bound proteins were eluted using a 10%, 20%, 30%, 60%, 100% His_B Buffer step elution (50 mM, 100 mM, 150 mM, 300 mM, 500 mM imidazole). Fractions were pooled based on SDS-PAGE analysis.

Table 11: IMAC Buffers. Buffers used for IMAC, the pH of the buffers was adjusted at 4°C.

His _A Buffer	His _B Buffer
50 mM Tris pH 8.0	50 mM Tris pH 8.0
300 mM NaCl	300 mM NaCl
	500 mM imidazole

3.3.4. Dialysis and Proteolytic Cleavage

Pooled protein fractions were dialyzed against SEC-buffer, or His_A buffer, using a Spectra/Por dialysis membrane at 4° C o/n (Spectrum Laboratories Inc, Rancho Dominguez, USA). If proteolytic cleavage of the His-tag was attempted, 1-2 mL TEV protease (0.5 mg/mL, produced in our laboratory) was added to the pooled fractions, before dialysis. The His-tagged TEV protease and undigested protein was removed by binding to a second IMAC. The flow through was pooled.

3.3.5. SDS-Polyacrylamide Gel Electrophoresis (SDS-PAGE)

To separate protein samples, SDS polyacrylamide gels with 12% or 15% acrylamide in the separation gel and 4% acrylamide in the stacking gel were used (**Table 12**). The samples were mixed with 4× SDS-sample buffer (**Table 13**) and heated for 5 min to 95°C. The electrophoresis was performed using a BIORAD Mini-PROTEAN Tetra System (BIORAD, Hercules, USA) at 200 V (PowerPac Basic, BIORAD, Hercules, USA) for 45-60 min, using 1× running buffer (Rotiphorese, Roth, Karlsruhe, Germany). The gels were stained with InstantBlue (Expedion Ltd, Over, UK).

Table 12: SDS polyacrylamide gel composition. The pH of the buffers used where adjusted at RT.

4× SDS Gels	Stacking Gel		Separation Gel	
	4% acrylamide	12% acrylamide	15% acrylamide	
10% (w/v) SDS	100 µL	150 µL	150 µL	
30% acrylamide	1.3 mL	6 mL	7.5 mL	
1.5 M Tris pH 6.8	2.5 mL			
1.5 M Tris pH 8.8		3.75 mL	3.75 mL	
H ₂ O	6.1 mL	5 mL	3.5 mL	
TEMED	10 µL	7.5 µL	7.5 µL	
10% (w/v) APS	100 µL	150 µL	150 µL	

Table 13: SDS-PAGE sample and running buffer composition. The pH of the buffers used where adjusted at RT.

4× SDS Sample Buffer	10× Running Buffer
20 ml 1 M Tris pH 6.8	0.25 M Tris
10 ml 10% SDS	1.92 M glycine
1.63 ml 0.5 M EDTA pH 8.0	1% (w/v) SDS
4 ml β-mercaptoethanol	
20 mg bromophenol blue	

3.3.6. Size Exclusion Chromatography (SEC)

Samples were filtered (0.22 μm) and loaded on an equilibrated SEC-column. For preparative SEC a HiLoad 16/60 Superdex 200 column (Pharmacia, Uppsala, Sweden) with a Pharmacia LKB GP250 plus system or a HiLoad 16/60 Superdex 75 column with a Pharmacia Biotech Amersham UV 900 P900 system were used with a flow rate of 1 mL/min. For analytical SEC a 3.2/30 Superdex 200 Increased column (GE Healthcare, Frankfurt, Germany) or a 3.2/30 Superdex 75 column (Pharmacia, Uppsala, Sweden) was used with an Ettan LC system (GE Healthcare, Frankfurt, Germany) with a flow rate of 0.05 mL/min. The SEC-buffer was used for SEC purification before crystallisation experiments were performed (Table 14). For stability experiments the DPBS-buffer was used for SEC. Absorbances at wavelengths of 280 nm, 254 nm and 215 nm were measured.

Table 14: Buffers composition used for SEC. The pH of the SEC-buffer was adjusted at 4°C.

SEC-Buffer	DPBS-Buffer
20 mM HEPES pH 7.5	0.90 mM CaCl ₂
150 mM NaCl	0.49 mM MgCl ₂
	2.67 mM KCl
	1.47 mM KH ₂ PO ₄
	137.93 mM NaCl
	8.06 mM Na ₂ HPO ₄

3.3.7. Chemical Crosslinking

The primary amine groups of lysine and/or of the N-terminus were cross-linked with 0.1-0.3% glutaraldehyde using 10 μg protein at RT, for 15-30 min. The reaction was quenched with Tris buffer, and the products of the reaction were analyzed with SDS-PAGE to validate intermolecular crosslinking.

3.3.8. Chemical Sidechain Modification

To change the chemical properties of the protein surfaces in order to influence crystal formation, sidechains were modified. Primary amines were methylated with formaldehyde and DMAB in SEC-Buffer (3.1.4). After 3 h of reaction at 4 °C additional DMAB and formaldehyde was added and incubated o/n (Table 15). The reaction was stopped with Tris and analyzed with liquid chromatography mass spectrometry using a Shimadzu LCMS 2020 (Duisburg, Germany) with a Phenomenex Kinetex (2.6 u C18 100 Å) (Aschaffenburg Germany) column by *J. Sindlinger* (Schwarzer lab, IFIB Tübingen, Germany). The basic property of primary amines (lysine, N-terminus), was changed through acetylation resulting in a non-basic sidechain (or N-terminus) or succinylation, resulting in an acidic sidechain (or N-terminus). For this reaction a protein concentration of 1 mg/mL was dialyzed in phosphate buffer pH 9, and over 2 h, acetic anhydride or succinic anhydride was added to a final concentration of 120 mM. The reaction was done on ice, while the pH was maintained at 9 by the addition 1 M NaOH. After additional 3 h of reaction time, the protein was dialyzed in 150 mM NaCl and 20 mM HEPES buffer at pH 8. The samples were analyzed with an analytical SEC or mass spectrometry.

Table 15: Compounds used for chemical modification of protein.

Methylation	Acetylation	Succinylation
per 1 mg protein	per 1 mg protein	per 1 mg protein
2×20 µL 1 M DMAB	120 µmol acetic	120 µmol succinic
2×40 µL 1 M formaldehyde	anhydride	anhydride

3.4. Biophysical Methods

3.4.1. Protein Concentration Determination

The protein concentration was determined according to the Lambert-Beer law (**40**) using the UV-absorbance at a wavelength of 280 nm (NanoDrop ND-1000, Thermo Fisher Scientific, Waltham, USA). The theoretical decadic absorbance coefficient for the different constructs (**Appendix 1**) was calculated with ProtParam [106].

3.4.2. Circular Dichroism (CD) Spectroscopy

To assess protein folding, CD-Spectroscopy was performed. Spectra from 195 nm to 250 nm wavelength were recorded with a protein concentration of about 0.3 mg/mL with a Jasco J-720 spectropolarimeter (Pfungstadt, Germany) at RT.

3.4.3. Thermal Shift Assay

To assess protein stability, protein melting curves were recorded. Therefore, a dye (protein thermal shift dye, Thermo Fisher Scientific, Waltham, USA) for which the fluorescence increases when it is bound to a protein was used. For excitation a wavelength of 580 nm was used, while the emission of 623 nm was measured. If a protein unfolds, hydrophobic parts become accessible to the dye, and the fluorescence increases. The fluorescence intensity of the protein and dye mixture (**Table 16**) was measured while slowly heating with a QuantStudio5 (Thermo Fisher Scientific, Waltham, USA).

Table 16: Thermal shift assay mixture.

Thermal Shift Assay Mixture
2 μ L 1-2 mg/mL protein
2.5 μ L dye (diluted 1:125)
15.5 μ L buffer

3.5. Structural Biology

3.5.1. Precipitation Test

To determine an appropriate starting concentration for initial crystallization experiments, a precipitation test was done. 0.5 μL of filtered (0.22 μm) protein solution was mixed on a glass slide, with different precipitate (30% (w/v) PEG4000 and 3 M ammonium sulfate) solutions. The protein concentration was stepwise increased until the protein precipitated within 1 min with at least one of the two precipitants.

3.5.2. Crystallization

Crystallization screens were done using 96-well sitting drop Intelli-plate (Art Robbins Instruments, Sunnyvale, USA) with either a Gryphon (Art Robbins Instruments, Sunnyvale, USA), or Freedom EVO (Tecan, Männedorf, Switzerland) crystallization robot. Drops were set up in a one to one ratio of protein solution and mother liquor, with a total drop size of 400 nL with the Gryphon, and 600 nL with the Freedom EVO. Crystals were optimized using 4 \times 6 (Hampton Research, Aliso Viejo, USA) hanging drop plates.

3.5.3. X-ray Diffraction Data Collection

To optimize the cryo-protectant or to test diffraction capacity of a crystal, the in house X-ray system was used. This system is composed of a rotating copper anode X-ray source (MicroMax-007HF (Rigaku, Sevenoaks, UK) producing CuK α radiation at $\lambda = 1.5418 \text{ \AA}$ and a mar345 image plate detector (marresearch, Norderstedt, Germany). Data sets were collected at the macromolecular crystallography beamline X06DA-PXIII of the Swiss Light Source (Paul Scherrer Institute, Villigen, Switzerland). Data sets for structure determination via

molecular replacement were collected with $\lambda = 1 \text{ \AA}$ (12.3984 keV). For anomalous data the wavelength was adjusted to specific values, according to the absorption edge of the element used.

3.5.4. Soaking of Crystals with Heavy Atoms

The HATODAS II [107] server was used to identify promising heavy atoms for initial soaking experiments. Therefore, crystals were soaked for 5 and 10 min in crystallization solutions containing different heavy atoms (**Table 17**), with a concentration of 10 mM or saturated if the solubility was lower. For Sm^{3+} and UO_2^{2+} compounds various crystals were soaked with concentrations ranging from 2.5 mM to 10 mM and soaking time between 1 min and 24 h. Additionally, crystals were soaked with 3 mM $(\text{NH}_4)_2\text{WS}_4$ or $\text{Lu}(\text{Ac})_3$ for 2 h, 1 mM $\text{Ta}_6\text{Br}_{14}$ for 1 week, 2.5 or 5 mM EuCl_3 for 24 h or 1 mM $\text{SeC}(\text{NH}_2)_2$ for 10 min.

Table 17: Heavy atom compounds used for soaking experiments.

HA-Compounds	
$\text{K}_2\text{Cl}_6\text{Pt}$	$\text{K}_2\text{Pt}(\text{NO}_2)_4$
SmCl_3	$\text{Sm}(\text{NO}_3)_3$
$\text{UO}_2(\text{NO}_3)_2$	UO_2Ac_2
$(\text{NH}_4)_2\text{Br}_6\text{Os}$	HoCl_3
Polyvalan [108]	

3.6. Software

The XDS program package was used to process (index, integrate, scale) the experimental diffraction image data sets, XSCALE to scale and merge datasets, XDSCONV to convert different crystallographic file formats [109]. From the PHENIX [110] program package XTRIAGE [111] was used to analyse data sets and PHENIX.REFINE [112] was used for simulated annealing and refinement. From the CCP4 [113] program package, POINTLESS [114] was used to identify possible space groups, MATTHEWS_COEF [26] was used to evaluate the possible unit cell content, CHAINSAW [115] to modify models for molecular replacement, PHASER [116] and MOLREP [117] for molecular replacement, REFMAC5 [118] for refinement and COOT [119] for model building and evaluation. PYMOL [120] was used for structure examination and figure generation. Experimental phasing was done using SHELX C/D/E [121] and AUTOSHARP [122] with SOLOMON [27] for density modification. MOLPROBITY [123] was used for model evaluation.

4. Results

4.1. Purification of FliC Constructs

The amino acid sequences of all used constructs are listed in **Appendix 2**. All constructs have been produced and purified as described in chapter 3.3 using a Ni-NTA column followed by size exclusion chromatography (**Figure 19**).

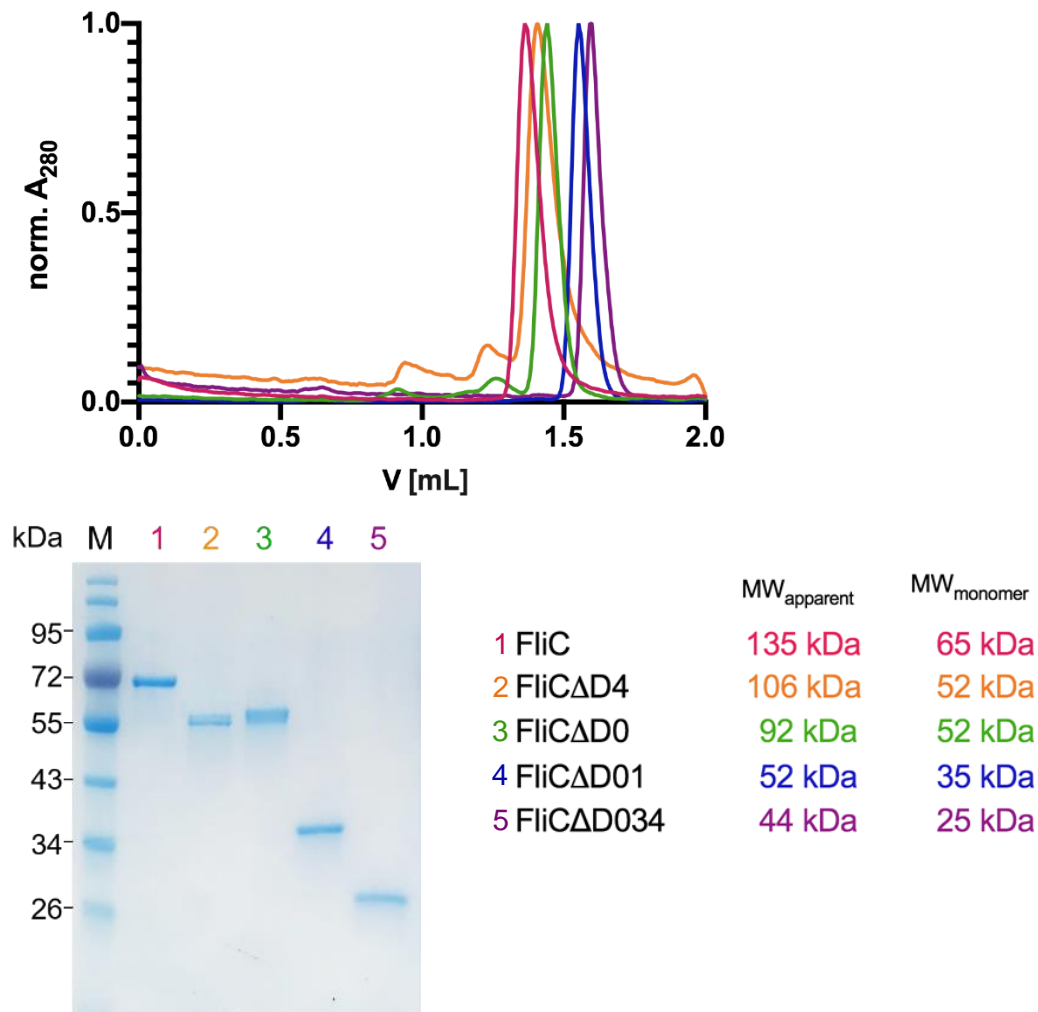


Figure 19: Analytical size exclusion chromatogram (SD200) and SDS-PAGE of representative purified FliC constructs. The normalized chromatograms of different constructs are shown in different colors, with the apparent molecular weight and monomer weight.

4.2. Crystallization of FliC Δ D01

Crystallization experiments using full length EcN FliC (**Figure 20 A**) with and without His-tag did not result in crystals. A shortened FliC, where the in solution unstructured D0 [124] which may hamper crystallization was removed FliC Δ D0 (Δ 1-47, Δ 555-595) (**Figure 20 B**), did not crystallize. Further shortening of the construct by removal of both constant domains, FliC Δ D01, (Δ 1-175, Δ 499-595) (**Figure 20 C**) resulted in protein which could be crystallized.

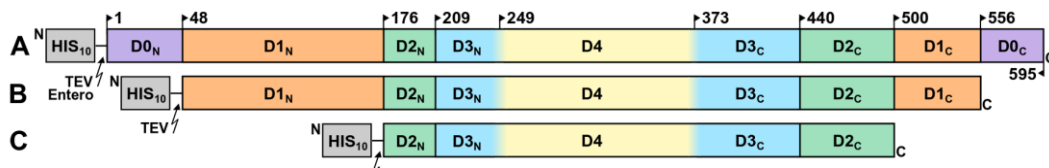
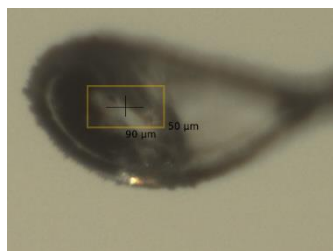


Figure 20: FliC constructs used for initial crystallization trails. The boxes are scaled and colored according to the five domains, residue numbers at domain boundaries are indicated. All constructs have a cleavable (TEV) N-terminal His₁₀-tag (scale magnified by three). **A:** Full length FliC a construct with an enterokinase cleavage site and/or TEV cleavage site was used. **B:** FliC Δ D0, where the putatively unstructured D0 was removed. **C:** FliC Δ D0D1 where both domains of the constant region were removed.

Initial screens with FliC Δ D01, resulted in crystals in more the 34 (**Appendix 3**) out of 384 screened (Hampton1-2, Wizard1-4, JCSG1-2) conditions. Crystals from 19 conditions were harvested and their diffraction qualities (resolution, anisotropy, mosaicity, number of lattices) were tested based on diffraction images generated at the in-house X-ray source. A data set for the best diffracting crystal was collected at the Swiss Light Source (X06DA - PXIII)(**Table 18**).

Table 18: Crystal used for data collection resulting in the highest resolution with the corresponding crystallization condition.

FliC Δ D01
FliC(Δ 1-176, Δ 500-595)
 19 mg/mL @ 20°C
 0.1 M HEPES 7.0
 0.2 M NaCl
 20% (w/v) PEG 6000

4.3. FliC Δ D01 Phase Determination

Molecular replacement attempts with the FliC structure from *Salmonella typhimurium* (1IO1) as well as domain fragments and poly-alanine models thereof, failed. Therefore, experimental phasing strategies were pursued. The anomalous signal from sulfur at 2.066 Å (6 keV) was not sufficient for native SAD phasing. The protein contains no cysteine and three methionine residues with a total molecular weight of 35.2 kDa. Crystals were grown with protein containing selenomethionine, but the anomalous signal was again not sufficient for the phase determination. Here, the problem was that only small crystals diffracting to a lower resolution could be produced. Further, native crystals were soaked with different heavy atom compounds (**Table 19**).

Table 19: Different heavy atom compounds tested for experimental phasing.

Heavy Atom Compounds			
K ₂ Cl ₆ Pt	K ₂ Pt(NO ₂) ₄	(NH ₄) ₂ WS ₄	(NH ₄) ₂ Br ₆ Os
SmAc ₃	Sm(NO ₃) ₃	Lu(Ac) ₃	HoCl ₃
EuCl ₃	UO ₂ Ac ₂	UO ₂ (NO ₃) ₂	
Ta ₆ Br ₁₄	Polyvalan [108]	SeUrea	

Soaks with Samarium compounds (SmAc_3 , $\text{Sm}(\text{NO}_3)_3$) showed anomalous signal at the appropriate wavelength (1.84 Å peak). Different soaking times varying from 30 seconds to days with different Samarium ion concentrations from 2 mM to 20 mM did not improve the measured anomalous signal to an extent that phases sufficient for further model building could be obtained. Neither SIRAS, SAD, nor MAD, using an inflect, peak and high remote data set, yielded sufficient phases. All data recorded had a high off origin Patterson peak (between 20-40%), indicating pseudo translational symmetry, weakening the phasing power. Different fine screening experiments could not reduce pseudo translational symmetry. From a crystal soaked for 10 min with 1 mM SeUrea, a SAD dataset was recorded at $\lambda = 0.978$ Å (12678 eV) (Table 20). The anomalous signal was above $1.33 d/\sigma_d$ up to a resolution of 3.44 Å. SHELX showed binding of 12 Selenium ions, where three bound with an occupancy above 50% (Figure 21).

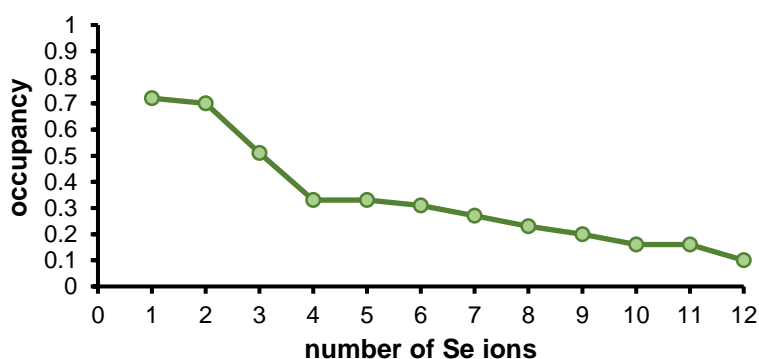
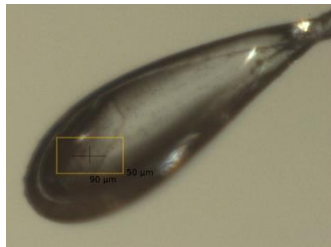


Figure 21: Occupancy of the 12 Se ions found with SHELX.

In the resolution shell between 3.63 Å and 3.44 Å, the phasing power was 1.01. After solvent flattening, a map could be calculated that allowed building of most of the protein peptide backbone. The preliminary R/R_{free} -values are 22.9%/27.1% at 2.30 Å resolution for the Se-derivative data set (Appendix 4). The preliminary model from the Se-derivative data set was used to solve the structure of the native

data set with molecular replacement. The current R/R_{free} -values for the structure of the native data set are 24.2%/27.4% at 1.65 Å resolution (**Appendix 4**).

Table 20: Crystal soaked for 10 min with 1 mM SeUrea used for SAD phase determination with the corresponding crystallization condition.



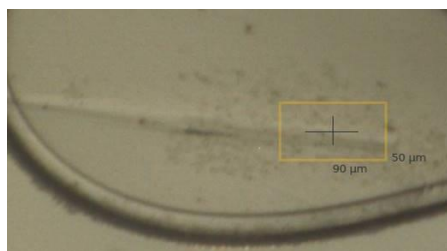
SeUrea derivative
FliC(Δ1-176, Δ500-595)
19 mg/mL @ 20°C
0.1 M citrate pH 4.5
10% (v/v) Propan-2-ol
20% (w/v) PEG 4000

4.4. Crystallization of FliCD12

With the crystal structure of the hypervariable domains, the transition between the hypervariable domains and conserved constant region as well as the constant regions was still unknown. Structures of FliC from *Bacillus subtilis* and *Salmonella dublin* in complex with *Zebrafish* TLR5 (5GY2; 3V47) as well as FliC from *Pseudomonas aeruginosa* (5WK5, 5WK6) or *Bacillus subtilis* (5WJT, 5WJU, 5WJV, 5WJW, 5WJX, 5WJY, 5WJZ) assembled to a flagellum, are available. These interactions between the TLR5 and FliC as well as the interactions in the flagellum mainly depend on the constant D1. Therefore, a structure of D1 and the first hypervariable domain, D2, would allow modeling of the TLR5 binding as well as the flagellum. Based on the D234 structure constructs were designed comprising D1 and D2. Crystallization trials with the first D12 construct (Δ1-47, Δ218-460<S, Δ555-595) did not result in crystals. Methylation, acetylation or succinylation of the lysines resulted in modifications of 12 out of 15 lysines, validated with ESI-LCMS (*J. Sindlinger*, Schwarzer lab), but not in crystals. In the size exclusion chromatography, the protein eluted at an apparent molecular weight of 44 kDa, whereas the monomer weight is 25 kDa. The incomplete chemical modification could be a result of a dimer formation, where three lysines are shielded from the solvent.

Based on an alignment of different FliC structures, shortened D1 constructs were designed (**Appendix 6**). Two screening conditions of this shortened constructs FliCD12(Δ 1-61, Δ 208-451<S, Δ 542-595) and FliC12(Δ 1-61, Δ 208-451<S, Δ 555-595), resulted in crystals (**Table 21**), from which data sets could be recorded at SLS (PXIII), which diffracted to a resolution of 2.03 Å and 1.75 Å, respectively.

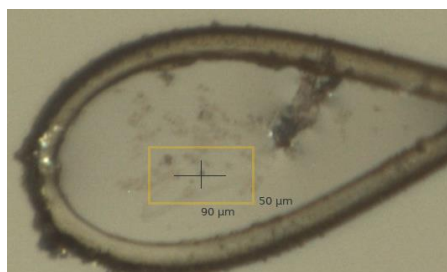
Table 21: Crystals used for data collection with the corresponding crystallization conditions.



Condition 1

FliCD12(Δ 1-61, Δ 208-451<S, Δ 542-595)

49 mg/mL @ 4°C
 0.1 M Tris 5.5
 0.2 M NH₄Ac
 25% (w/v) PEG 3350



Condition 2

FliC12(Δ 1-61, Δ 208-451<S, Δ 555-595)

45 mg/mL @ 4°C
 0.1 M Tris pH 8.5
 0.2 M MgCl₂
 30% (w/v) PEG 4000

4.5. FliCD12 Phase Determination

Initial phases were obtained by molecular replacement using Phaser. As search models the D2 from the previously solved FliC Δ D01 structure and D1 from *Salmonella* (1I01; 60-175, 403-450), modified that the model comprises only the sidechains to the last common atoms of the corresponding Nissle sidechain (CHAINSAW), were used. Initial phases were good enough to further improve the model. The preliminary R/R_{free}-values are 23.6/26.2% at 2.03 Å resolution for condition 1, FliCD12(Δ 1-61, Δ 208-451<S, Δ 542-595). The current R/R_{free}-values are 20.4/24.4% at 1.70 Å resolution for condition 2, FliCD12(Δ 1-61, Δ 208-451<S, Δ 555-595) (**Table 21**) (**Appendix 5**).

4.6. Model of FliC Δ D0

The structure of the hypervariable domains FliC Δ D01 and the FliCD12 structure both contain D2. This allows the generation of a FliCD1234 (FliC Δ D0) model by superposition of the D2 structures. The root mean square deviation (rmsd) of the superposition of the best fitting structures was 0.4 Å over 275 AAs, whereas the worst fitting superposition had an rmsd of 1.2 Å over 313 AAs (**Figure 22**).

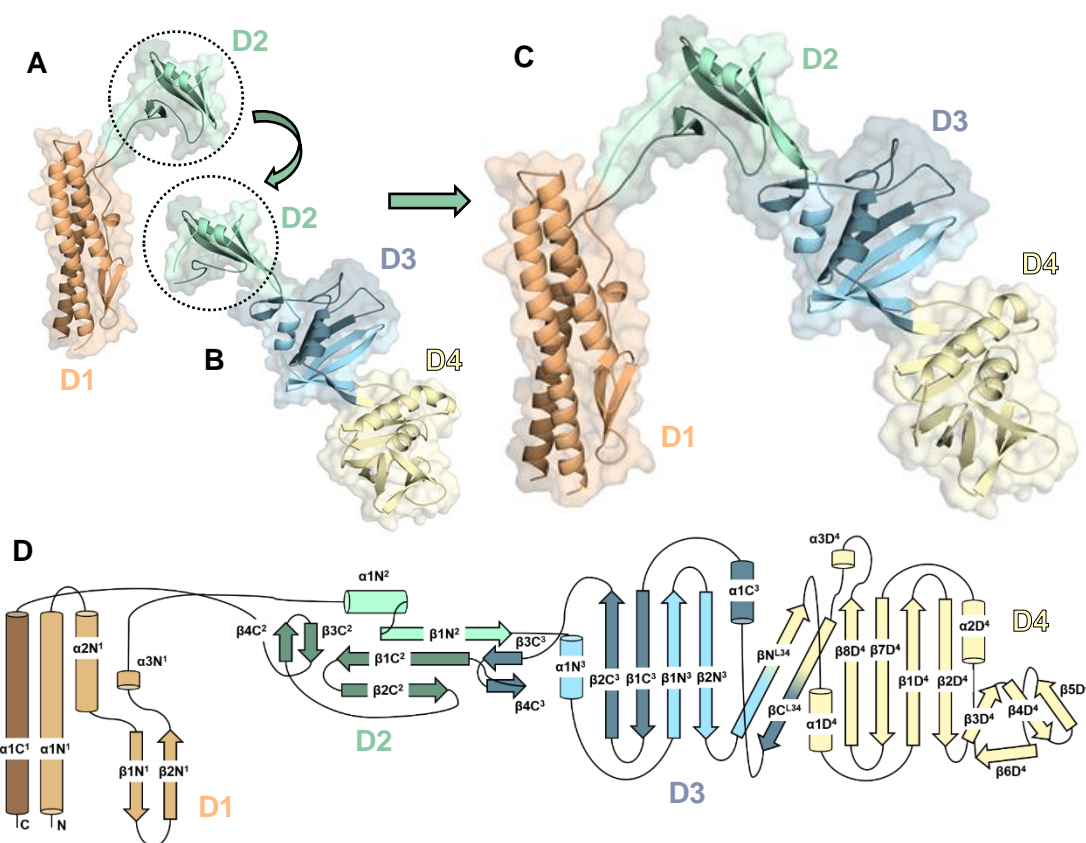


Figure 22: Cartoon and surface representations of crystal structures from two different FliC constructs and the model created from them. **A:** Structure of the D1 and D2. **B:** Structure of the D2, D3 and D4. Dotted circle indicate the D2 structures which were superimposed to generate **C:** model comprising the D1 (orange), D2 (green), D3 (blue) and D4 (yellow). **D:** Topology plot of the FliC D1234 model.

4.7. Structure of the Constant D1

The structure of the constant D1 is very similar to published structures from other organisms such as the *Salmonella Typhimurium* structure (1IO1), which had the lowest C α -rmsd of 0.57 Å between the D1s (**Figure 23**). The N-terminal and C-terminal α -helices form an anti-parallel coiled-coil structure ($\alpha 1N^1$; $\alpha 1C^1$). The N-terminal α -helix interact with an antiparallel α -helix ($\alpha 2N^1$), formed after a short loop. After a second loop region, a β -sheet of two antiparallel β -strands ($\beta 1N^1$; $\beta 2N^1$) interact with the C- and N-terminal coiled-coil. Next, a region with no secondary structure elements (except one α -helical turn ($\alpha 3N^1$)) follows, ending at a point close to the C-terminal strand. From here the N- and C-terminal strands interact with one another and form the transition to the first hypervariable domain, D2.

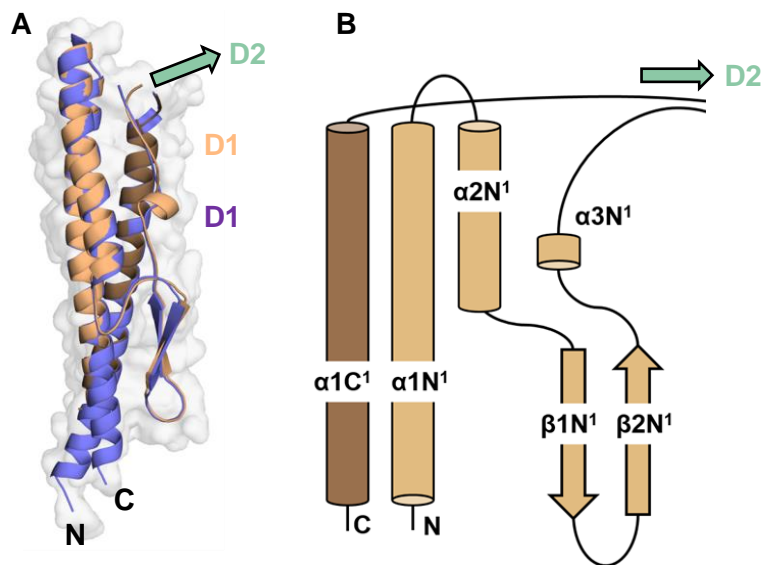


Figure 23: Superposition topology plot of D1. **A:** Superposition of the EcN (brown) and *S. Typhimurium* (1IO1) (purple) D1, with an rmsd of 0.57 Å. **B:** Topology plot of the EcN D1. The transition to D2 is indicated with green arrows.

4.8. Transition of D1 to D2

EcN FliC has a linker region between D1 and D2, this was unexpected as the transition in the *Salmonella* structure is seamless. The *Salmonella* structure shows a numerous interactions between the D1 and D2 domain, in EcN there are none. Since both termini of FliC are located in D0, the D1, D2 transition contains two strands which interact with each other. The length of the linker is about 30 Å (**Figure 24**). The first 21 Å out of the linker, from D1 to D2, are stabilized by ten hydrogen bonds and one ionic interaction between Glu500 and Lys505. In the 9 Å long second part of the linker, the N-terminal strain forms two hydrogen bonds with the C-terminal strain, which itself forms a β -hairpin. This C-terminal β -hairpin also interacts with the D2 core fold, which additionally stabilizes the linker region.

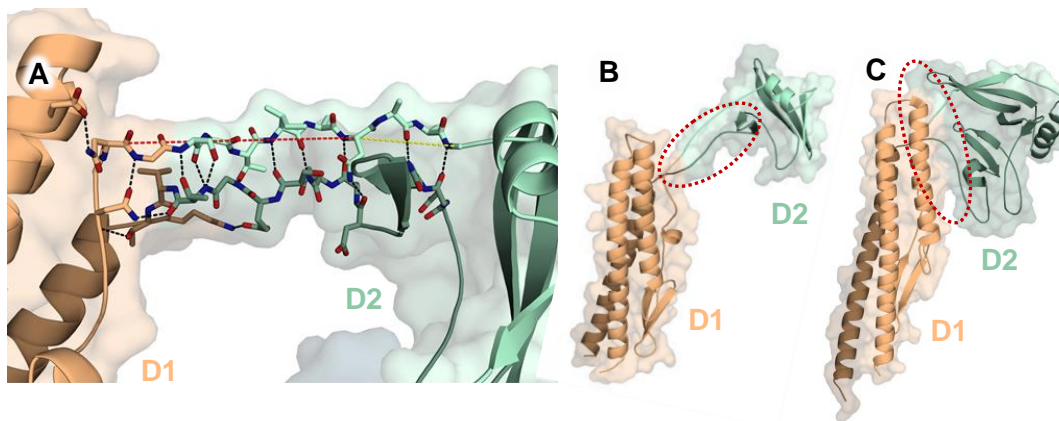


Figure 24: Transition between D1 of the constant region and the first hypervariable domain, D2. **A:** Interaction of the two linker strands. The hydrogen bonds and ionic interaction are depicted with black dotted lines. The distance between the D1 and D2 is depicted with a red dotted line (21 Å) and a yellow dotted line (9 Å). The region close to D2 is stabilized by a β -hairpin. **B:** FliC EcN structure of D1 (brown) and D2 (green) in cartoon and surface representation. The linker region is highlighted with a red dotted ellipse. **C:** *Salomonella* FliC structure, the region of the D1 and D2 interaction is highlighted with a red dotted ellipse.

4.9. Structure and Transition of D2 and D3

The D2 contains a three stranded antiparallel β -sheet and one α -helix forming a $\alpha\beta\beta$ -sandwich (**Figure 26**). The N-terminal strand forms the α -helix and one of the β -strands. This β -strand is also involved in the transition to D3, where it interacts with a β -hairpin structure of D3. This secondary structure element is part of the rather short transition between D2 and D3. Beside the two β -sheets of the $\alpha\beta\beta$ -sandwich D2 core fold, the C-terminal strand forms on both domain boundaries short β -hairpins, which interact with the D2 core fold (**Figure 25**). As motioned before (4.8) one of this β -hairpin stabilize the D1-D2 transition, the second one stabilizes the transition between D2 and D3.

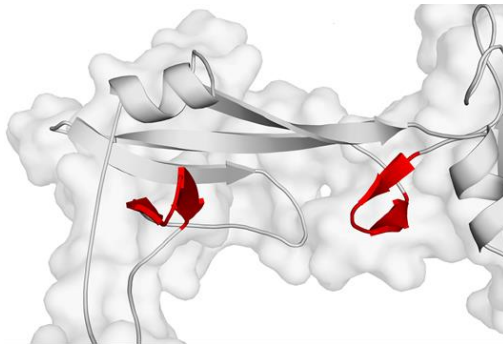


Figure 25: The transition of D2 is stabilized in both directions by β -hairpins (red).

The D3 domain consists of two $\alpha\beta$ -sandwiches where the four β -strands form a β -sheet with the two α -helices laying on the same side. One of the $\alpha\beta$ -sandwiches is formed by the N-terminal strand where the other is formed by the C-terminal strand. The D2 β -sheets from the $\alpha\beta\beta$ -sandwich and the β -sheets from the D3 $\alpha\beta$ -sandwiches lie in a plane, but their sheet orientations are orthogonal.

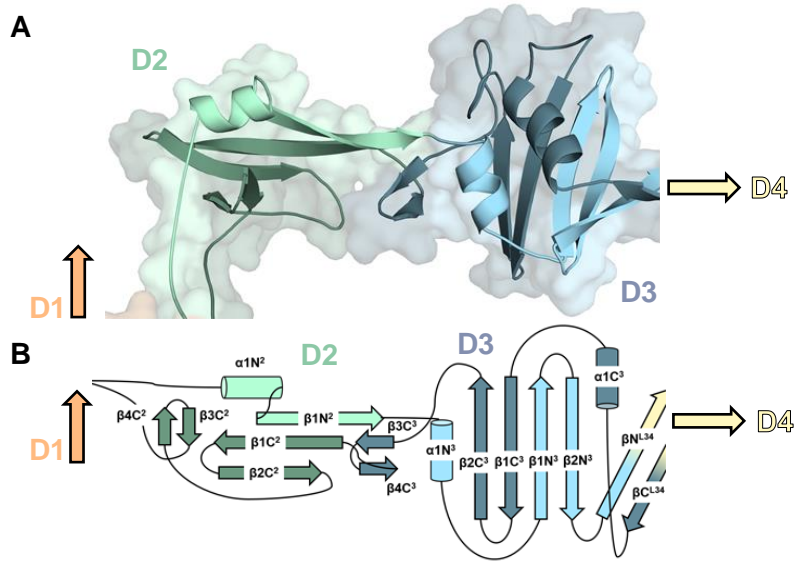


Figure 26: Structure and transition of D2 and D3. **A:** Cartoon and surface representation of D2 (green) and D3 (blue). **B:** Topology plot of D2 and D3. The N-terminal strand is shown in light the C-terminal strand in dark colors. The transition from D1 and to D4 are indicated with arrows.

4.10. Structure and Transition of D3 and D4

The core fold of D3 and D4 is very similar, both consisting of two $\alpha\beta$ -sandwiches, where the 4 β -strands form a β -sheet, with the two α -helices laying on the same side (**Figure 27**). The angle between the β -sheet planes of both domains is about 90° . The two domains are linked by a two stranded antiparallel β -sheet, where every β -strand also interacts with the β -strand of the $\alpha\beta$ -sandwich. Therefore, the core folds of both domains are connected.

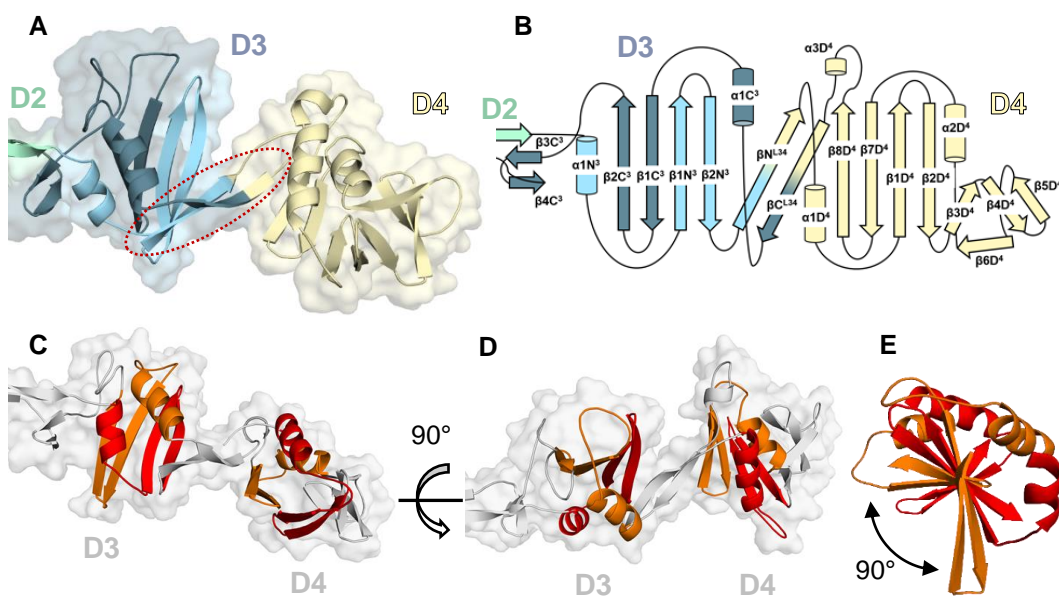


Figure 27: D3 and D4, cartoon and surface representation and topology plot. A: D3 (blue) and D4 (yellow), with the β -sheet transition between the two domains highlighted with a red dotted circle. **B:** Topology plot colored as **A**. **C:** Side view showing the $\alpha\beta$ -sandwiches of the N-terminal and C-terminal strand in red and orange, respectively. **D:** Side view as **C** but rotated 90° as indicated. **E:** View from D3 to D4 showing the $\alpha\beta$ -sandwiches to visualize the $\sim 90^\circ$ angle between the β -sheets.

4.11. The β -Triangle

The tip of D4 has an unusual fold, where three strands form an antiparallel triangle, with β -sheet like interactions, here termed β -triangle (**Figure 28**). There are three backbone interactions at two of the junctions, and two at the third junction. Two of the strands are additionally part of a β -sheet, where one interacts with a β -sheet of the $\alpha\beta$ -sandwich and the other is part of a two stranded β -sheet. Following this two stranded β -sheet, a loop reaches over this particular two stranded β -sheet and closes the triangle with last β -sheet like structure. Additionally the fold is stabilized by an ionic bond between Lys295 and Glu302.

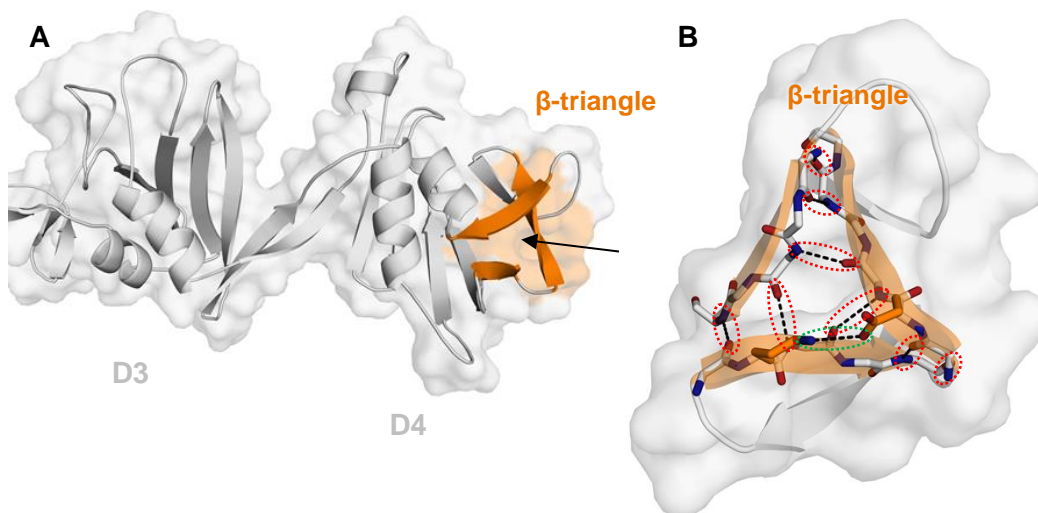


Figure 28: β -Triangle fold at the tip of D4. **A:** D3 and D4 shown in cartoon and surface representation, with the β -triangle fold highlighted (orange). **B:** Close-up view of the β -triangle fold with the β -sheet like backbone interactions, highlighted with red dotted ellipse, and the ionic interaction highlighted with a green dotted ellipse. View from D4 to D3 as indicated by the black arrow.

4.12. Flexibility Between Domains

4.12.1. FliC D12 Structure Alignment

Three structures of the FliC D12 were solved from two crystals (one asymmetric unit contained two protomers). This allows a structure alignment, whereby here, only D1 was aligned to visualize the differences in the relative position of D2 (**Figure 29**). The angle between the center of mass (COM) of D1, the C α of Pro507 at the tip of D1 and the COM of D2 varies between 130° and 136°. The angle between the C α of Pro507, the COM of D2 and the carbonyl C of Thr206, a point close to the transition to D3, varies between 92° and 107°. The structure shows that the first portion (21 Å) of the linker is stabilized by 10 hydrogens bonds and one ionic bond, where the last portion (9 Å) is stabilized by two hydrogens bonds and a β -hairpin structure.

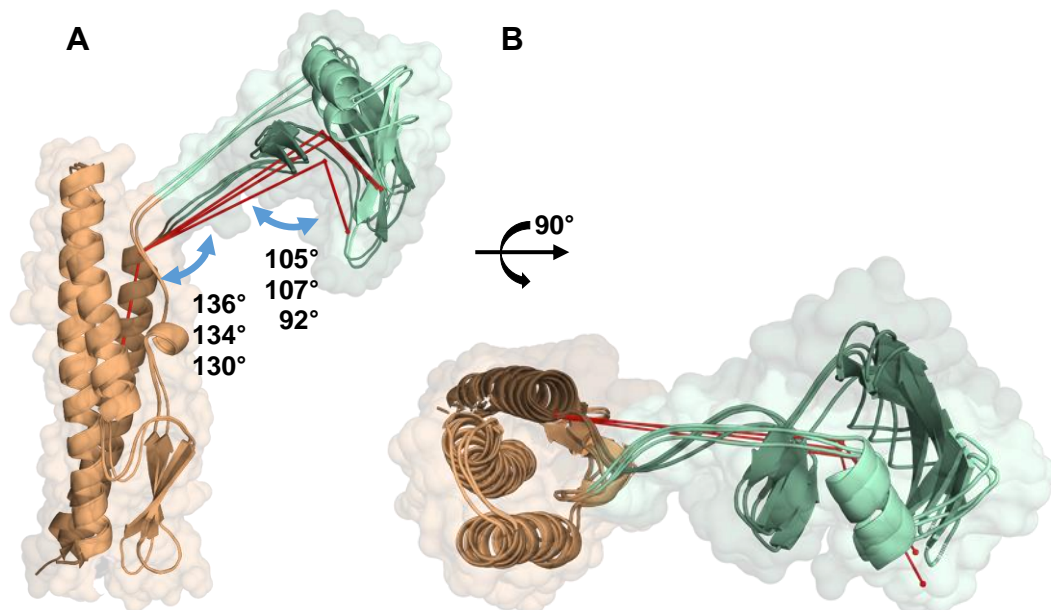


Figure 29: Alignment of three EcN D12 structures. **A:** Side view with angles between center of mass (COM) of D1, C α of Pro507, the COM of D2, and carbonyl-C of Thr206, connected with a red line. **B:** Top view of **A** rotated as indicated. The structures are shown in cartoon and surface representation with D1 (brown) and D2 (green).

4.12.2. FliCD12 and FliCD234 Structure Alignment

The alignment of D1 of the three D12 structures was superimposed at D2 with the three D234 structures, resulting in nine FliC D1234 models (**Figure 30**). When comparing the different angles between the C α of Pro507, the COM of D2 and the COM of D3, the angles vary between 82° and 100°. The variation in angles is only 3° higher than the angles driven from C α of Pro507 (tip of D1), the COM of D2 and the carbonyl C of Thr206, indicating only slight orientation divergences between D2 and D3. The angles between the COM of the D2, D3 and D4 is between 8° and 12°, indicating that there is not much flexibility deducible from the crystal structures among these domains. Nevertheless, due to four domain organization and angles differences between the domains, D4 seems to be able to cover a spectrum of positions in space with D1 as reference point.

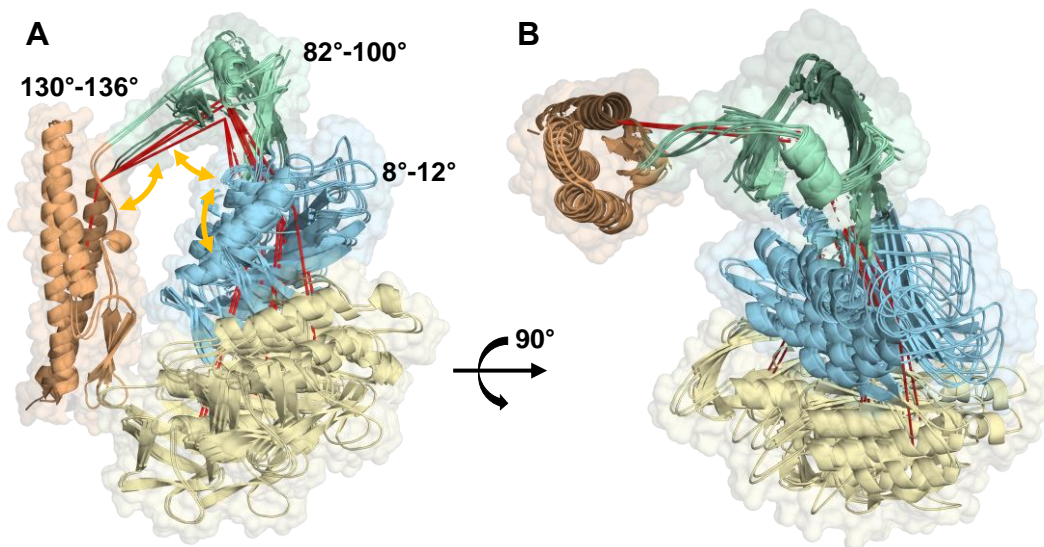


Figure 30: Model structure of FliC D1234. Structures are shown in cartoon and surface representation. The three FliCD12 structures were superimposed at D1, and for each structure, three FliC D234 structures were superimposed at the D2. **A:** Side view with the minimal and maximal angle between the domains. The COM of D1, C α of Pro507, the COM of the D2, D3 and D4 are connected with a red line. **B:** Top view rotated 90° as indicated.

4.13. Modeling the Flagellum

The two most distant structures from the structure assemble gained from alignment of the crystal structure (chapter 4.12.2) were selected to model the flagellum (Figure 31).

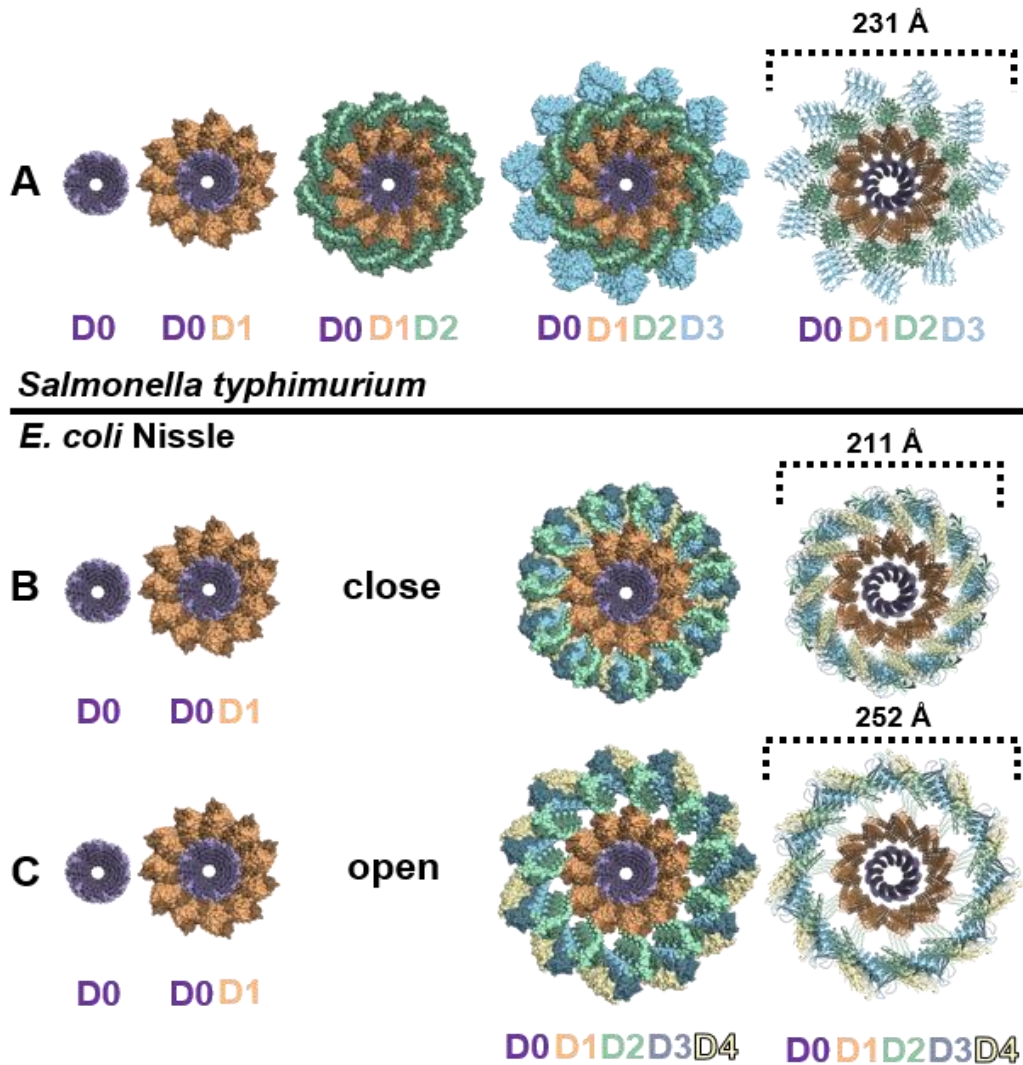


Figure 31: Three different flagellum models. Cartoon or surface representation axial viewed from the tip towards the stator of the flagellums, with D0 (purple), D1 (orange), D2 (green), D3 (blue) and D4 (yellow). **A:** Model of the *Salmonella typhimurium* flagellum. **B:** Closed model of the EcN flagellum. **C:** Open model of the EcN flagellum. The diameter of the models are indicated with dotted lines.

Therefore, D1 from these structures were aligned with each D1 from the flagellum cryo-EM structure of *Bacillus subtilis* (5WJT). The two resulting models differ in their shape as a result of the different hypervariable domain orientations. The model with the hypervariable regions closer to the D1 resulted in a closed flagellum, where the hypervariable region forms an outer ring-like structure. In the other model, D3 and D4 are moved outwards, opening the gap between the inner constant domain superstructure and the outer hypervariable domain superstructure (**Figure 31**).

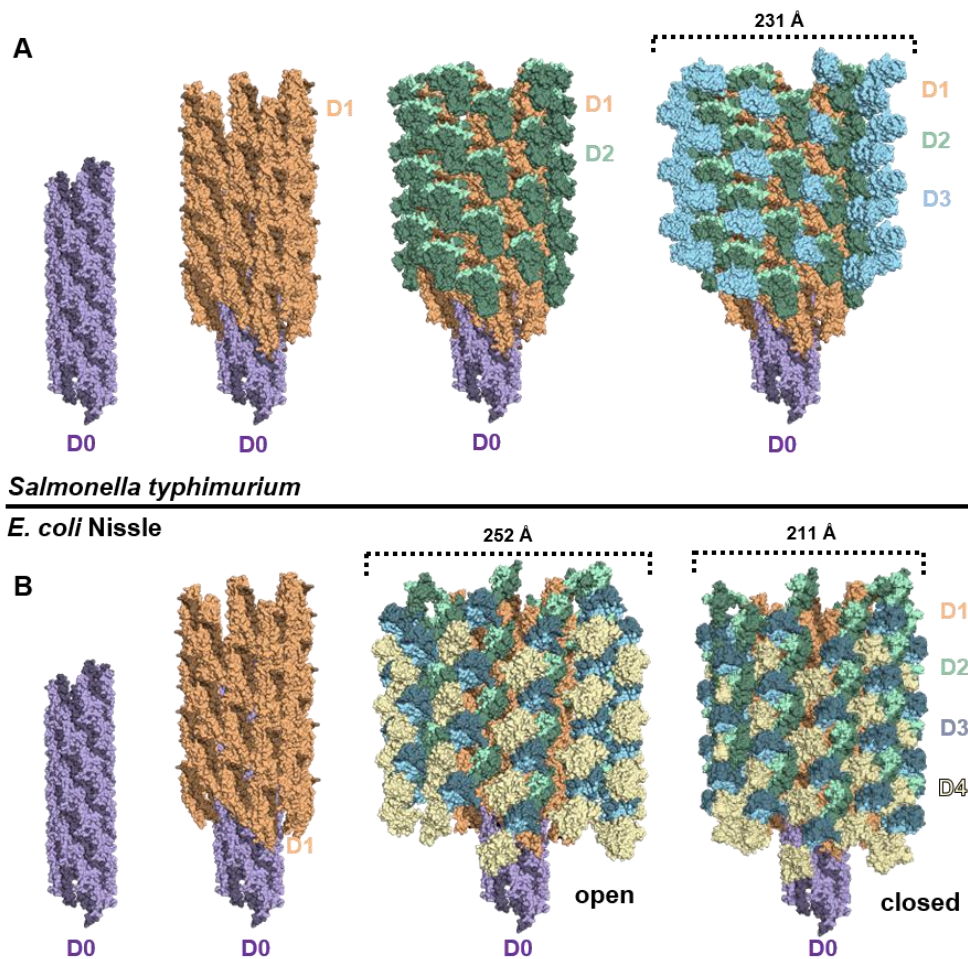


Figure 32: Three different flagellum models. Surface representation viewed from the side (radial) with D0 (purple), D1 (orange), D2 (green), D3 (blue), and D4 (yellow). **A:** Model of the *Salmonella typhimurium* flagellum. **B:** Model of the EcN flagellum in an opened and closed conformation. The diameter of the models are indicated with dotted lines.

4.13.1. Monomer Interactions within the Flagellum

Within the flagellum, the two constant domains, D0 and D1, of one monomer interact with six adjacent monomers, allowing the formation of a stable flagellum. The interaction patterns of D0 and D1 are different (**Figure 33**). Every D0 interacts with four adjacent D0s. Also, every D1 interacts with four adjacent D1s, but two of the interactions are not with the same monomer. So, every D01 monomer interacts with six adjacent monomers, two only via D0, two only via D1, and two via both D0 and D1. Between the two models of the flagellum, including the hypervariable region, is a difference in the connectivity. In one model the hypervariable domains are closer to the constant domains, with more putative interactions between different hypervariable domains. The other model has a more open conformation, with less contact between the domains. Both of the models have in common that there are interactions which connect the same two monomers in the axial flagellum direction. The open model shows additional interactions between two monomers which are also connected by D0. The closed model shows interactions of two monomers, which are neither connected through D0 nor D1. While there are no clashes in the open conformation model, there are clashing sidechains in loop regions of the closed model, which should be avoidable through a slight rearrangement of loops and sidechains.

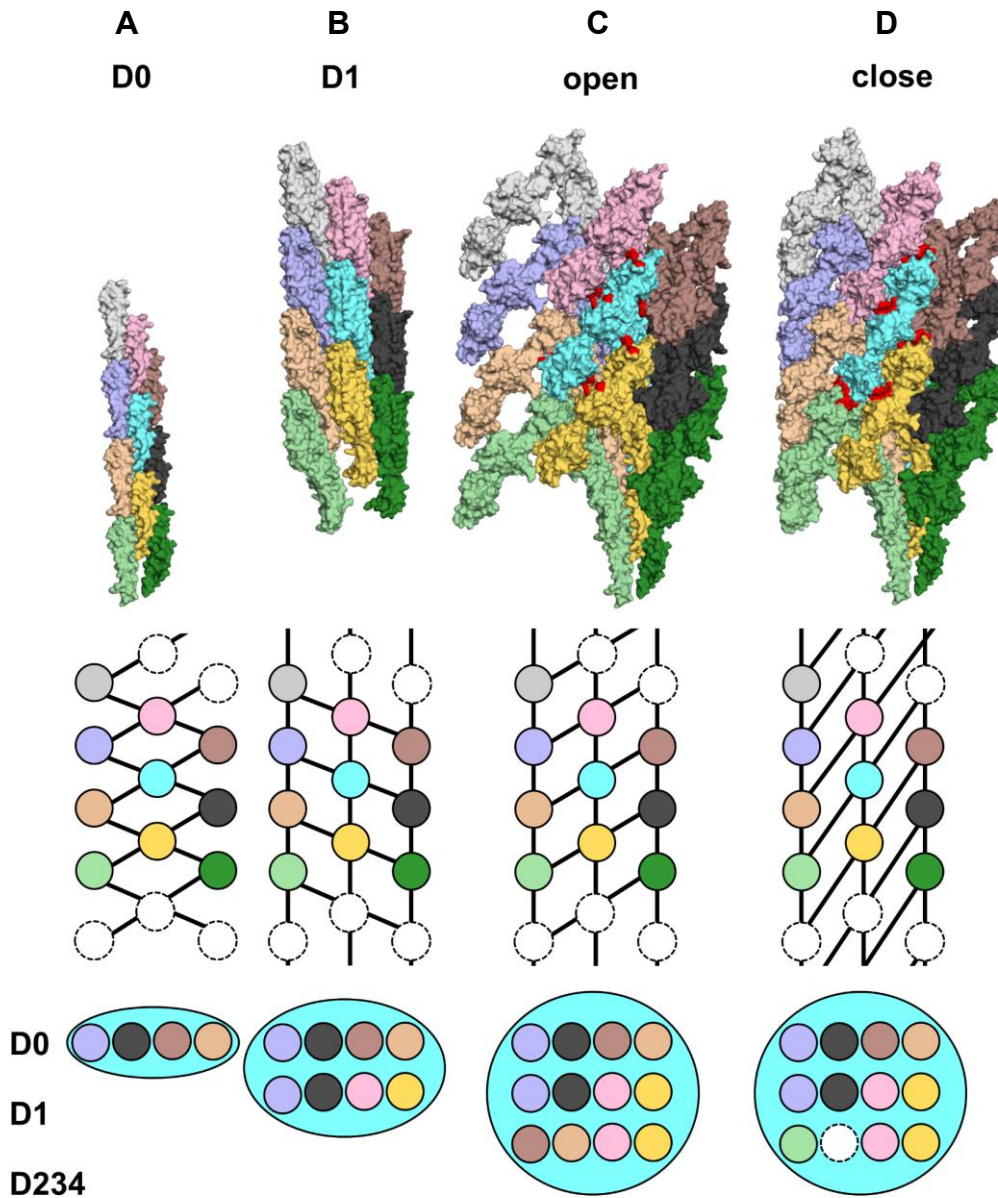


Figure 33: Schematic representation of the FliC interactions in the flagellum. The first row shows different domains, assembled to a flagellum in surface representation, whereas every monomer is colored differently. The second row indicates the interaction pattern between the monomers. The last row shows the interaction patterns of the blue monomer. **A:** Interactions between D0. **B:** Interactions between D1. **C:** Interactions between the D234 in the open conformation. **D:** Same as **C**, but in the close conformation. In surface representation of the open and closed conformation the interactions of the blue monomer are highlighted in red, whereas the most interaction are in the axial direction of the flagellum.

4.13.2. The Flagellum Diameter

In the *Salmonella typhimurium* flagellum model, the first hypervariable domain, D2 is directly associated with D1. Viewed along the axis, D2 seems to form a close ring directly adjacent to the D1 ring (**Figure 31**). The radial view shows that the hypervariable regions do not interact with each other (**Figure 32**). Each D2 has two neighbors in both tangential directions, but these are not close enough for interactions. The second hypervariable domain, D3 is located between two D2s when viewed along the axis, resulting in a gearwheel-like shape. The radial view shows that every D3 is located between two of the tangential neighboring D2 in a CW manner, without contacting them. In contrast to the *Salmonella* model, there are no interactions between the hypervariable and the constant regions in the EcN models, only a linker joins the regions. Here, the hypervariable domains form a ring-like structure, where in the closed model D3 and D4 are shifted towards the axis leading to a closed ring, with a regular ring shape. Between the constant domain ring and the hypervariable ring is a solvent filled gap. In the opened model D3 and D4 are shifted outward, leading to a more gear-like shape of the outer ring. The radial view shows that the hypervariable regions interact mostly with each other in the axial direction, forming structure with grooves. The solvent area between the regions is increased. The diameter of the flagellum varies 40 Å between the close (211 Å) and open (252 Å) models. The axial surface of revolution is 350 nm² for the closed and 500 nm² for the open model, therefore the open flagellum model has about 140% of the volume of the close model. Both models are driven from the two extremes of the FliC D1234 models, whereas the real structure of the flagellum could be an intermediate (**Figure 31**; **Figure 32**).

4.14. FliC Stability

Melting curves of five different FliC mutants showed that the deletion mutant FliC Δ D034, which lacks three domains was the most unstable FliC, with a melting point of 36.3 ± 0.5 °C (**Figure 34**). The most stable FliC, with a melting point of 53.3 ± 0.6 °C, was the variant without the constant domains (FliC Δ D01). These two variants could be crystallized. The other three FliCs, which did not form crystals in crystallization trails, have comparable melting points, with 42.9 ± 0.5 °C for the D4 deletion mutant, 45.7 ± 0.5 °C for the wt FliC and 46.4 ± 0.5 °C for the D0 deletion mutant.

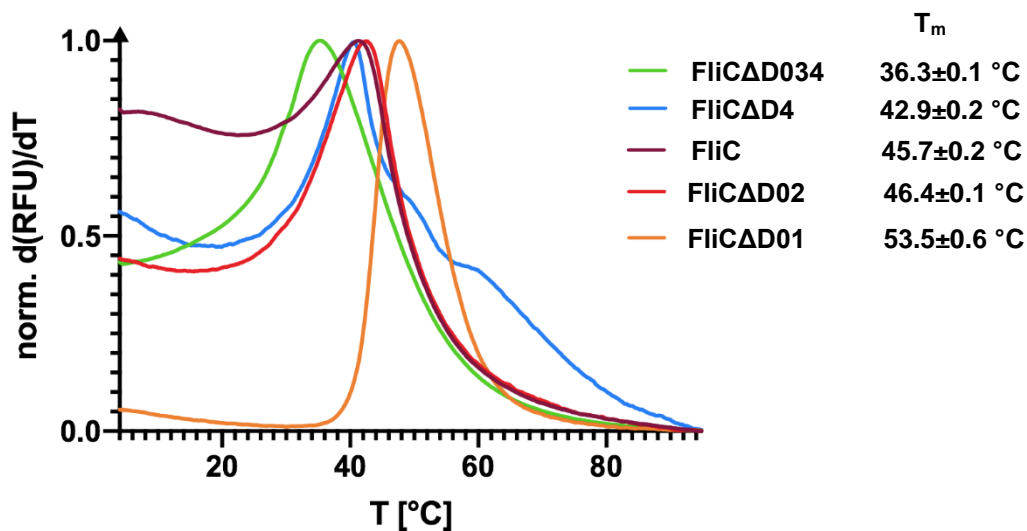


Figure 34: Temperature versus fluorescence signal change diagram. Melting curves have been determined for wt FliC and four different domain deletion mutants, using a thermos shift assay (3.4.3) in triplicate. Mean values are given with the standard deviation of the measurements, a more realistic error would be the highest standard deviation of 0.6 °C.

On the basis of the thermal stability experiments, the hypervariable region seems to be the most stable part of FliC. The melting temperature of the hypervariable domains is about 8 °C higher compared to FliC wt. The in solution unstructured D0

[124], does not influence the melting curve significantly. This suggests that unfolding of D1 reduces the stability of the hypervariable domains. Long-term stability experiments show degradation of D0 and D1, which is enhanced by higher temperatures (**Figure 35**).

The three hypervariable domains are structurally connected to each other by β -sheet structures. Therefore, the design and purification of a D4 deletion mutant was challenging. Only one out of five designed construct could be purified at all, but most of the sample was aggregated. This supports the importance of the structural connection for the folding of D3 and D4. The melting curve of this D4 deletion mutant construct (**Figure 37**; construct 6) was about 3 °C lower compared to wt FliC, whereas the deletion mutant D034 showed a reduction of the melting temperature of 9 °C.

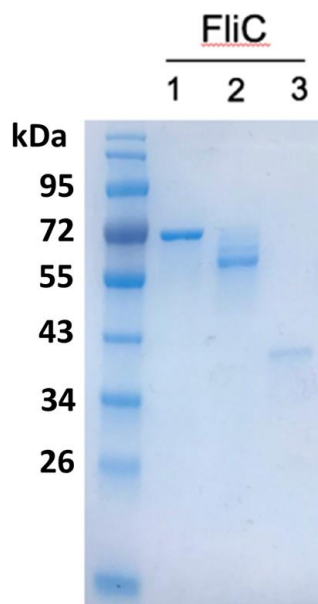


Figure 35: Degradation of wt EcN FliC. Equal amounts of purified FliC (1) was incubated at 4°C (2) or RT (3) for 20 days.

4.15. FliC Δ D4 and FliC Δ D34 Purification

Generation and purification of a D4 deletion mutant was challenging. Based on the structure of the hypervariable domain, five different constructs, with different linker lengths connecting the N-terminal and C-terminal strand of D3 were designed (**Figure 37**). All of the five constructs could be produced in *E. coli* (BL21), but only one did not show complete aggregation. This construct comprises most of the two stranded β -sheet and is connected via a five amino acid long linker. This construct suffers from severe formation of aggregation, but a relatively small fraction of the sample could be extracted that subjected to a second SEC seems stable against further aggregation (**Figure 38**). CD-spectroscopy showed that the protein was potentially folded (**Figure 36**), containing α -helices and β -sheets.

A deletion mutant without D3 and D4 (FliCD012) was designed with the same linkage of D2 as used for the FliCD12 construct that crystallized, but the addition D0 resulted in aggregation of the protein.

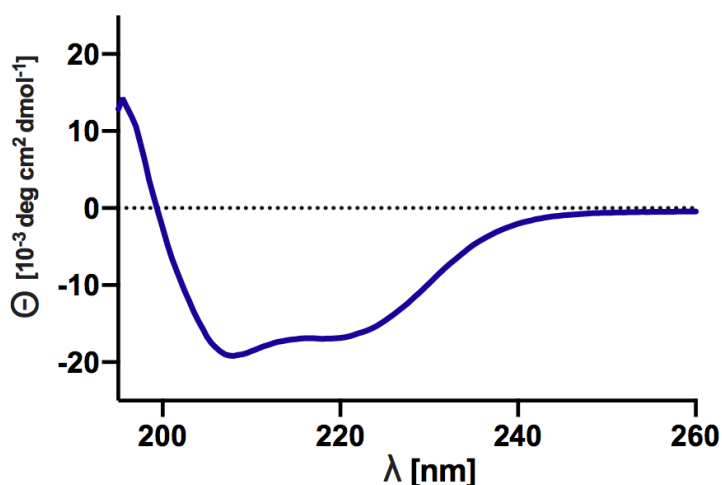


Figure 36: CD-spectroscopy of the D0123 mutant 6. The CD-spectroscopy showed that the FliC is folded and has α -helical content.

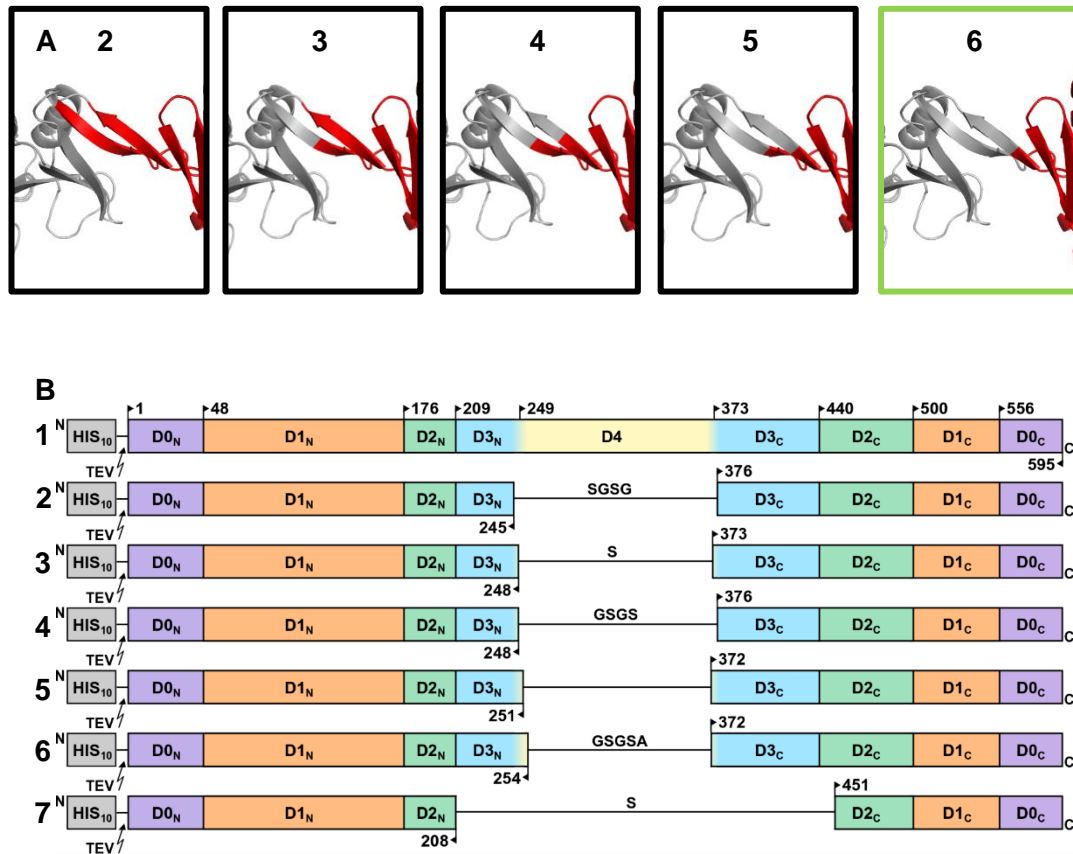


Figure 37: Different D4 deletion mutants. **A:** Cartoon representation of the transition between D3 and D4 domain. The part which has been removed is shown in red for the different mutants 2-6. Purification of mutant 2-5 resulted in aggregate, whereas only mutant 6 could be purified (green frame). **B:** Overview of the FliC constructs, where 1 is wt, 2-6 are Δ D4 mutants and 7 is a Δ D34 variant. The boxes are scaled and colored according to the five domains, domain boundaries are indicated by their first or last AA, respectively. All constructs contain a TEV-cleavable N-terminal His10-tag (scale magnified by a factor of 3).

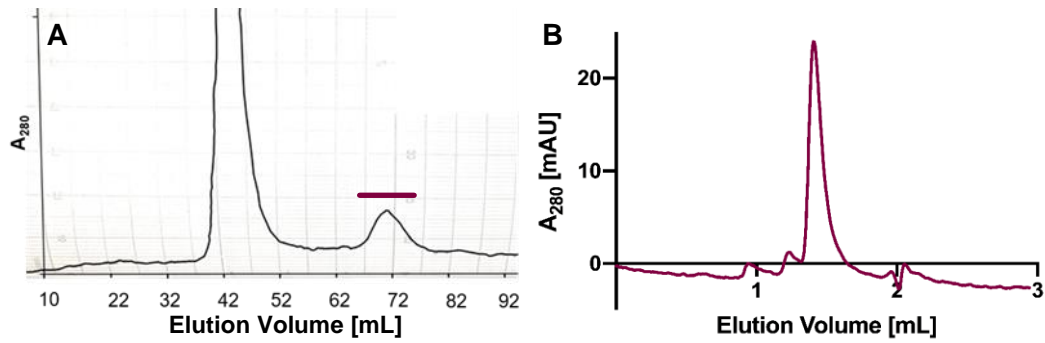


Figure 38: SEC of the D0123 mutant 6. **A:** The preparative SEC shows a dominate peak at the void volume (45 mL) of the column and a minor peak at about 70 mL indicated with a purple bar. **B:** Analytical SEC of the pooled peak at about 70 mL from **A** (purple bar). No aggregation peak was observed and the elution volume corresponds to the size of a monomer or dimer.

4.16. Sequence Alignment of Flagellins

Different *E. coli* strains vary in their flagellin structure, which is one criteria for their categorization (H-types). Different sequence alignments were performed to identify H-types with similar flagellins, and for which flagellins the EcN structure allow a structure prediction. A flagellin protein sequence alignment (**Appendix 7**) of representative H-type antigens from all H-types (**Appendix 8**) was generated with MUSCLE [125]. A phylogenetic tree (**Figure 39**) and a sequence identity matrix plot (**Figure 40**) show the relationship of the different FliCs. H12 for example has a sequence identity of 98.8% to EcN FliC (H1).

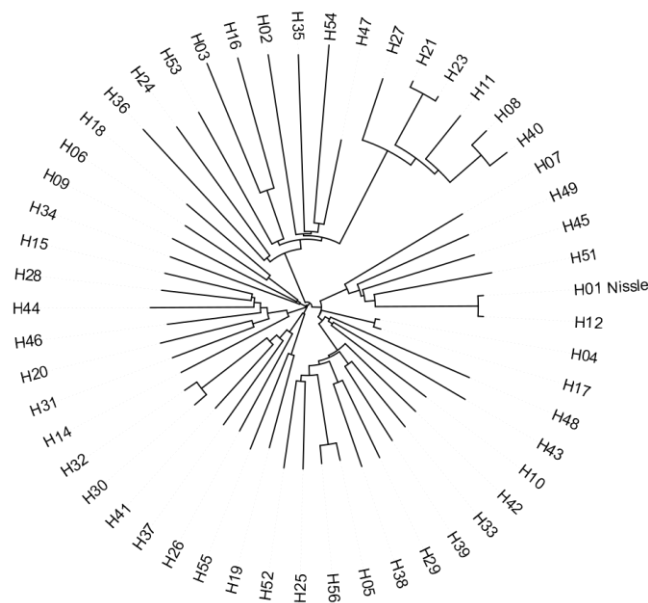


Figure 39: Phylogenetic tree of *E. coli* H-types. Figure based on a MUSCLE multiple protein sequence alignment of flagellin from representative H-types visualized with iTOL [126].

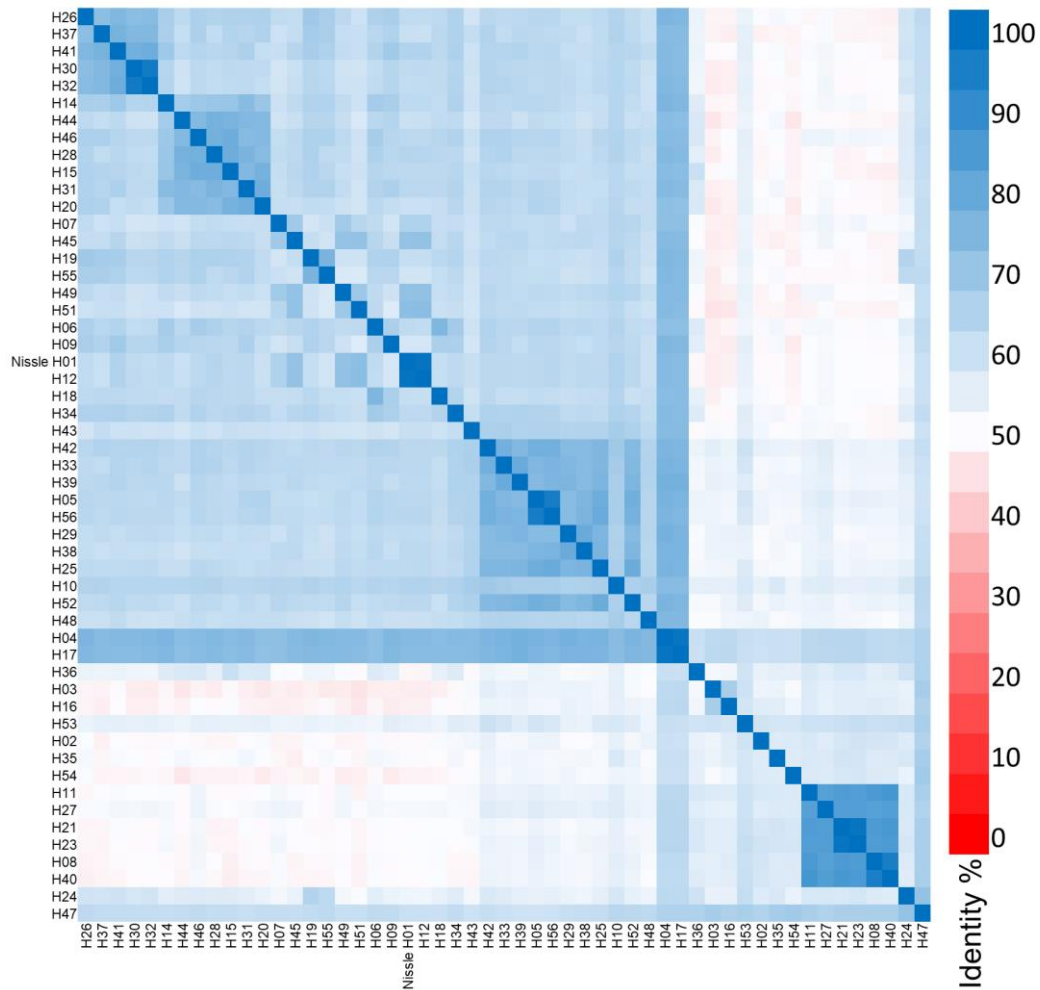


Figure 40: Sequence identity matrix plot of the *E. coli* H-types. Figure based on a MUSCLE multiple protein sequence alignment of flagellin. Blue color represent sequence identity above and red below 50%.

If the hypervariable and constant region are compared separately, differences within those regions can be identified. In this way small differences between the constant regions are not dominated by the huge differences between the hypervariable regions. When the constant region is neglected, the identity between the hypervariable regions can be assessed. Separate alignments (**Appendix 9**)(**Appendix 10**) and phylogenetic trees (**Figure 44**) were calculated for the

conserved region comprising the first 170 N-terminal plus the last 85 C-terminal AAs and the hypervariable region without those 170 N-terminal and 85 C-terminal AAs. This choice may not represent the exact same domain boundaries previously referred as constant and hypervariable regions by others [127, 128].

4.16.1. The Conserved Region

The phylogenetic tree based on the alignment of the conserved parts (**Figure 44** right) shows that a group of 38 H-types is more closely related to each other than a second group of 15 H-types (**Table 22**). Members of group 1 have a sequence identity above 91% within the group, whereas members of group 2 have a sequence identity above 84% within that group. The sequence identity of all conserved regions is above 74%. When the full sequences, including the hypervariable regions, are compared, the identity is above 45%. The consensus sequence (**Figure 41**) of the two groups differ in 27 AAs and at additional 34 positions are groups specific predominant AAs.

Table 22: H-types grouped based on their conserved domain.

Group 1	Group 2
H01, H04, H05, H06, H07, H09, H10, H12, H14, H15, H17, H18, H19, H20, H25, H26, H28, H29, H30, H31, H32, H33, H34, H37, H38, H39, H41, H42, H43, H44, H45, H46, H48, H49, H51, H52, H55, H56	H02, H03, H08, H11, H16, H21, H23, H24, H27, H35, H36, H40, H47, H53, H54

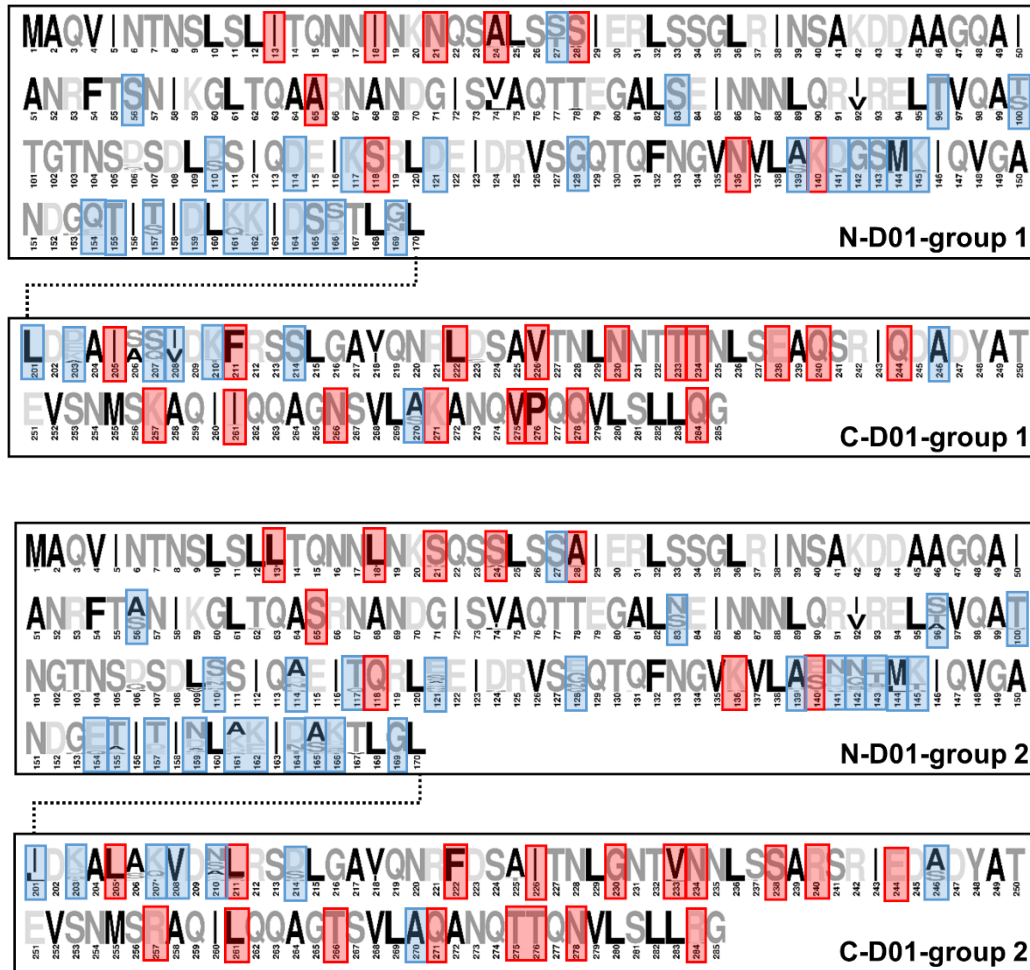


Figure 41: Consensus sequence of group 1 and group 2 of the conserved region. Positions with distinct differences are marked with red boxes, while positions with preference for certain residues are marked with a blue boxes.

4.16.2. The Hypervariable Region

The hypervariable region of different *E. coli* H-types varies substantially in sequence. An alignment of all H-types hypervariable regions resulted in an average sequence identity of 29% (**Figure 43**), whereas the average sequence identity of the full sequence was 44%. Based on the sequence alignment five groups A to E have been assigned, showing a sequence identity above 40% within their groups (**Figure 44**).

The EcN FliC has a sequence identity above 40% to H12, H45, H49, H51 (group C) but may also serve as model for most of the other H-types. The H-types H53, H35, H02, H54, H16, H03, H11, H23, H21, H27, H40, H08, H48, and H43 have a sequence identity to EcN FliC below 25% (**Figure 42**).

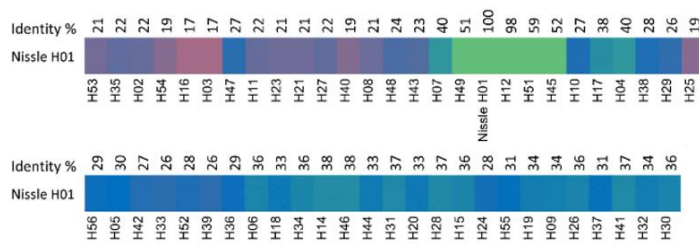


Figure 42: EcN sequence identity to other H-types. The sequence identity is color in a three color gradient from red (10%) over blue (25%) to green (50%). The corresponding sequence identity values to EcN are listed on top (EcN H01 row from **Figure 43**).

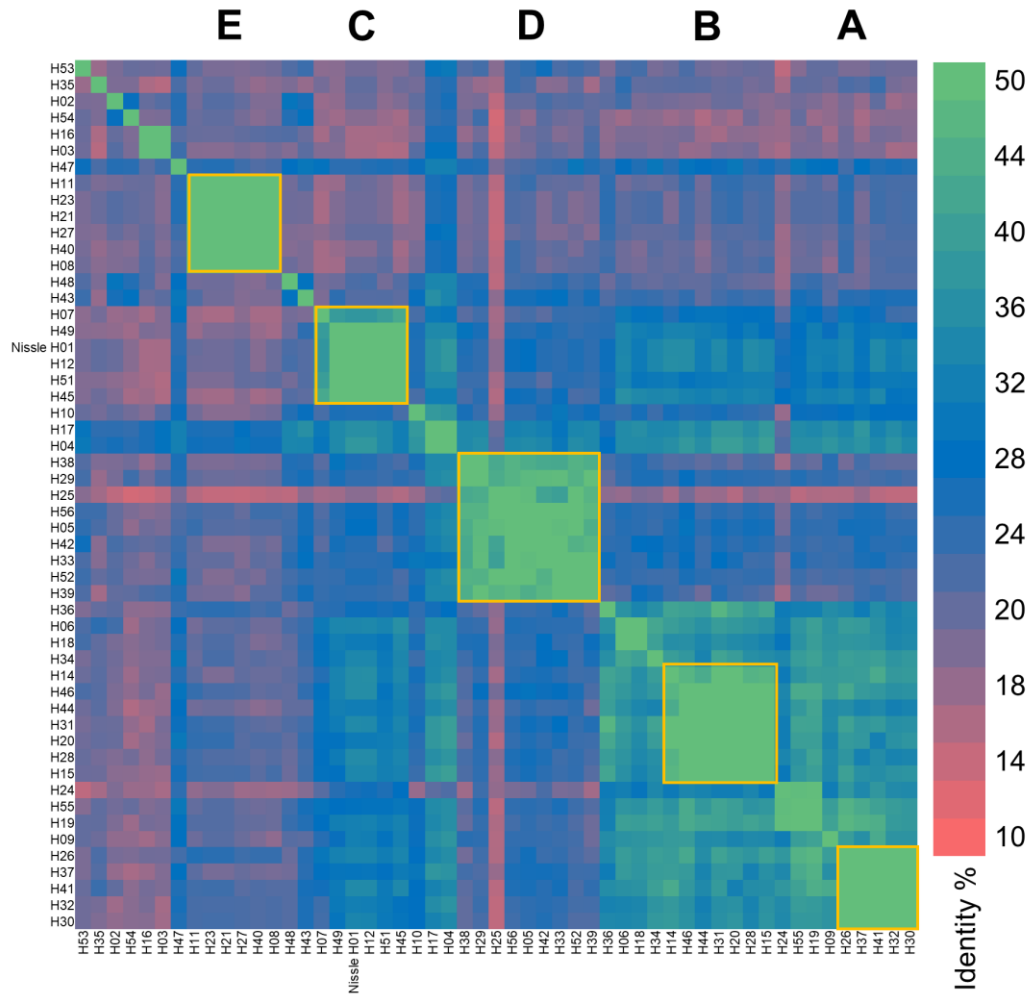


Figure 43: Hypervariable domain sequence identity matrix plot of *E. coli* H-types. The figure is based on a MUSCLE multiple protein sequence alignment. Five different groups from A to E are indicated above the plot and marked with orange boxes within the plot. The sequence identity is colored in a three color gradient from red (10%) over blue (25%) to green (50%).

5. Significance

We solved two crystal structures comprising different parts of flagellin (FliC) from the clinically relevant commensal EcN. This allows us to model the FliC structure from EcN, the first FliC structure from *E. coli*. The structures show two novel features, first the hypervariable region is joined with a long linker to the constant region, with no direct contacts between these regions, and second the hypervariable region has one additional domain compared to previously published FliC structures. Modeling of the flagellum showed a novel flagellum architecture, where the hypervariable domains form an outer ring-like structure connected with spoke-like linkers to the inner constant domain. These structures will help to model unknown FliCs and broaden the picture of structural diversity of flagellin and the flagellum. This structure from a commensal might contribute to the understanding, what distinguishes beneficial from virulent features. Sequence comparison of different *E. coli* FliC identifies two groups of constant regions and five groups of hypervariable regions that could be investigated further. Last, we modeled the FliC TLR5 interaction and postulated how the hypervariable domain could influence the TLR5 activation.

6. Discussion

6.1. Sequence Comparison of Flagellins

What can we learn from the EcN FliC structure about other *E. coli* strains? Classically, *E. coli* are categorized based on agglutination reactions with different polyclonal antibody sera against preparations of standard strains, which define a certain serotype [129]. Differences in the lipopolysaccharide (LPS) composition are referred to as O-antigens (186 O-types). The K-type (60 K-types) is based on capsular polysaccharides. H-typing differentiates based on flagellin antigens (53 H-types). The serologic classification of *E. coli* is time consuming and can be complicated due to cross reactivity of the used sera. Modern serotyping of *E. coli* relies on sequencing or PCR fragmentation [127, 130, 131]. EcN is monotrichously flagellated and has an O6:K5:H1 serotype [132].

All serotypes have conserved N- and C-terminal parts, which form the first and second domain (D0 and D1) with a sequence identity to EcN higher than 78%. In addition to this conserved “constant” domains, the middle part of the flagellum called “hypervariable region” can vary completely between different H-types. To correlate a function with the structure, an approach investigating the hypervariable and constant region separately could be a productive approach.

Based on sequence alignments, we categorized two groups of constant regions. These groups differ in 27 out of 254 AAs and show a prevalence for a certain AA in additional 34 positions. The grouping procedure might allow to investigate the effect on functions such, as TLR5 stimulation or flagellum stability, of representatives from each group.

Five groups (A to E) were identified for the hypervariable region (identity above 40%), which might have similar Flagellin structures within the groups. Not all H-types could be assigned to these groups, which illustrates the diversity of the hypervariable regions. To understand the structural diversity of the hypervariable region of *E. coli*, it would be reasonable to solve at least one structure from a member of each group (EcN FliC belongs to group C), and characterize the

structural differences. Noteworthy, a member of group E would be especially interesting as this group is most distantly related to the other groups. H-types so far not classified in the groups might also possess large differences in the 3D fold. As an example, the *Salmonella Typhimurium* (11O1) FliC, often referred as model structure, has a sequence identity above 30% to only three of the *E. coli* H-types hypervariable regions, which are not included in the groups (H54 43%, H4 35%, H17 33%).

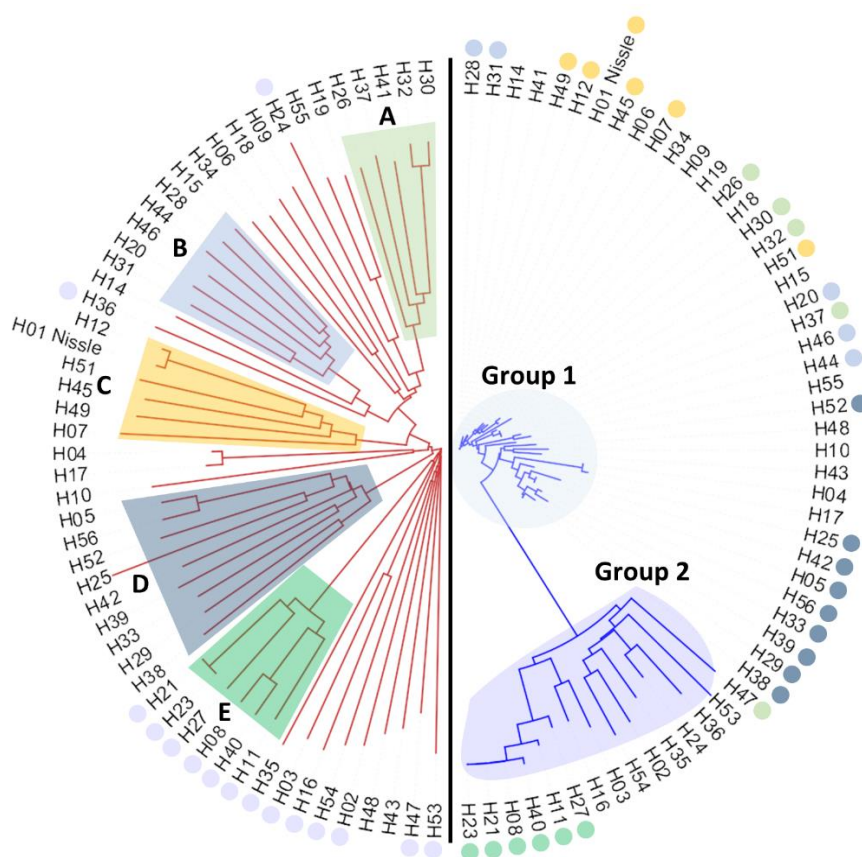


Figure 44: Phylogenetic tree of *E. coli* H-types. The phylogenetic tree of the variable (left, red) and conserved (right, blue) regions are vertically separated by a black line. The variable regions which share >40% similarity are boxed with different colors and letters from A to E. The conserved region (170 N-terminal and 85 C-terminal AAs) clusters in two groups. Group 1 comprises 38 closely related H-types, whereas group 2 is distant to group 1, with more variability within the group. Colored dots represent the grouping of the opposing phylogenetic tree.

6.1.1. H-type Domain Organization

The structure of the hypervariable domain showed three separated domains (D234). This was unexpected as in the literature EcN FliC have been compared to the *Salmonella* (11O1) FliC structure with two distinct hypervariable domains (D23). A comparison of the length of EcN with other H-types shows that EcN is one of the FliCs with the highest number of amino acids, whereby only five other H-types are larger (H51, H19, H55, H24, H9) (**Figure 45**). Based on the sequence alignment, they seem to have insertions mainly in D4, while H9 seem to have larger insertion (23 aa) in the D3. Three H-types (H4, H17, H47) probably only consist of one hypervariable domain (D2) due to their low number of AAs.

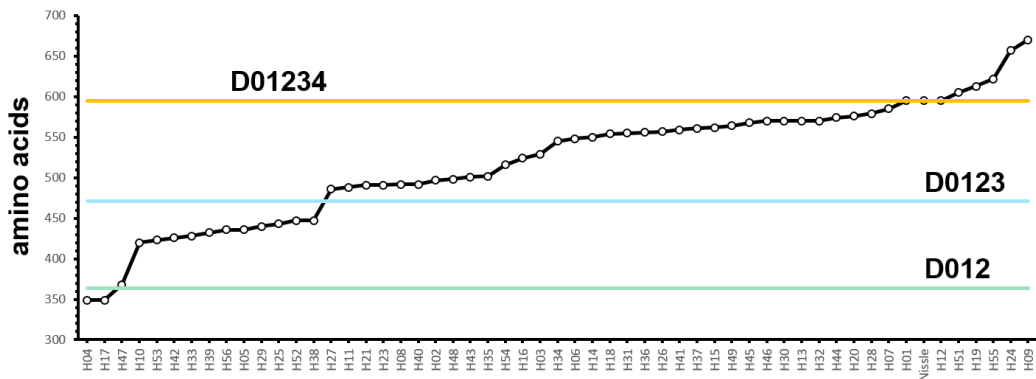


Figure 45: Size of the H-types as plotted by the number of amino acids for every H-type. The domain boundaries for EcN are indicated with colored lines.

The sequence alignment of EcN with H52 (group D) shows that large parts of all hypervariable domains are missing (**Figure 46**). The structures of group D may have some similarities to EcN, but might be folded differently and due to a much lower sequence length. This group might consist of two domains only. To model these H-types a structure of a member of this group would be beneficial.

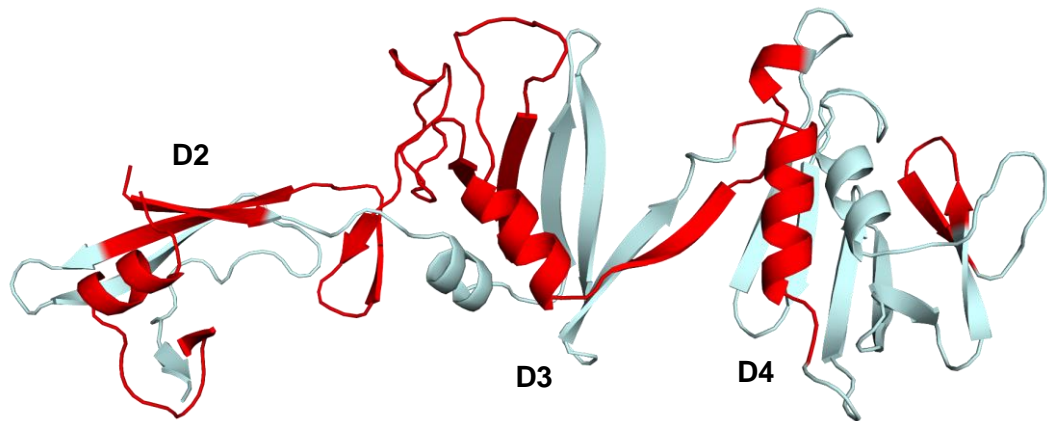


Figure 46: Sequence alignment of EcN and H52 visualized on the EcN structure. Larger parts missing in H52 are shown in red.

6.2. D4 Deletion Mutant

We designed a FliC D4 deletion mutant where the sequence after the domains connecting β -sheet was removed and showed that the protein is purifiable and folded. This mutant could be used to investigate the influence of this domain on TLR5 activation, flagellin stability or mucin binding.

Yang *et al.* [133] previously generated two EcN FliC chromosomal integrated deletion mutants (277-286 and 287-296) and compared them with EcN wt and a complete FliC deletion mutant. The structure presented here shows that these two deletion mutants are located within D4, partly removing β -strands, which are part of the $\alpha\beta$ -sandwich fold of D4 (**Figure 47**). This should impair the stability of D4. They showed that EcN harboring those mutations possess a flagellum (transmission electron microscopy images), but are unable to swim. This suggests that D4 integrity is not required for the formation of the flagellum, but affects the stability of the hypervariable domains, which effects the functionality of the flagellum. Interestingly this did not affect the H antigenicity and mucin-2 dependent

binding to IPEC-J2 (intestinal porcine enterocytes from jejunum of neonatal piglets) cells was only slightly reduced.

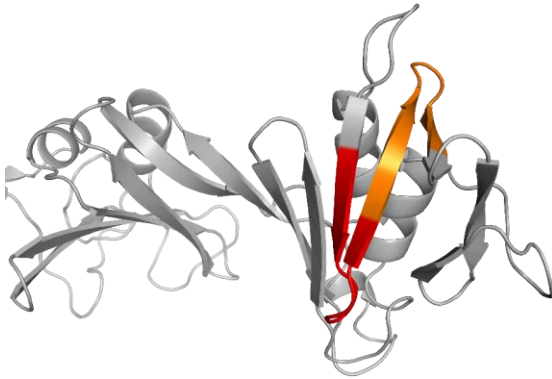


Figure 47: Cartoon representation of D4. The two deletion mutants investigated by *Yang et al.* are shown in red and orange. Both mutants comprising a loop and parts of two β -strands and should therefore influence the integrity of the $\alpha\beta\beta$ -sandwich fold of D4.

Due to the structural connection of the hypervariable domains, and the low sequence identity to most other H-types, mutations based on the EcN structure or models thereof need to be treated with caution. A proper mutant design would be enhanced substantially by a wider availability of structural data to better assign borders within the topology of the H-types.

6.3. Flexibility Between Domains

Crystal structures of FliC EcN show that there is a linker between the constant and hypervariable domains. The structure of the linker seems to be rather rigid, but slight changes in the angles between AAs in the linker region can change the orientation and position of the hypervariable and constant regions towards each other drastically. This was illustrated by the alignment of all crystal structures presented here, setting D1 as fixed reference point (**Figure 30**). In solution, with absence of crystal contacts, the degree of movement might probably be even higher.

6.4. Model of the EcN Flagellum

A cryo-EM structure of a FliC assembled to a flagellum allow the modelling of the flagellum with FliCs. Here, the cryo-EM structure of the flagellar filament (5WJT) from *Bacillus subtilis* comprising the D0 and D1 domain was aligned at D1 either with the *Salmonella typhimurium* (1IO1) FliC or FliC from EcN. As discussed in chapter 4.12.2 there are differences in the domain orientations in the crystal structures, allowing to generate an assemble of possible D1234 models (**Figure 30**). The two extremes, with the maximum displacement of the nine model assemblies, were used to create two models of the flagellum. One model is in an open conformation, while in the other model the hypervariable domains are more close to the constant domain helical structure.

The available data is ambiguous with regards to the interactions between the hypervariable domains. For the flagellum of EcN, both models show interactions between the hypervariable domains. In contrast to this, a *Salmonella typhimurium* flagellum model shows no interactions between the hypervariable regions. The cryo-EM reconstruction of the flagellar filament of *P. aeruginosa* (5WK5; EMD ID: 8855) shows electron density for D2 and D3, which interacts in axial direction with adjacent hypervariable regions. The resolution of the electron density for this domains did not allow model building [134]. This shows that the interactions in axial direction proposed based on both EcN models are not a unique feature of EcN. Additionally, tangential interactions seems possible as shown by the closed model, in this case the ridges formed by the protruding hypervariable domains, come in close contact forming a more regular rod-like structure.

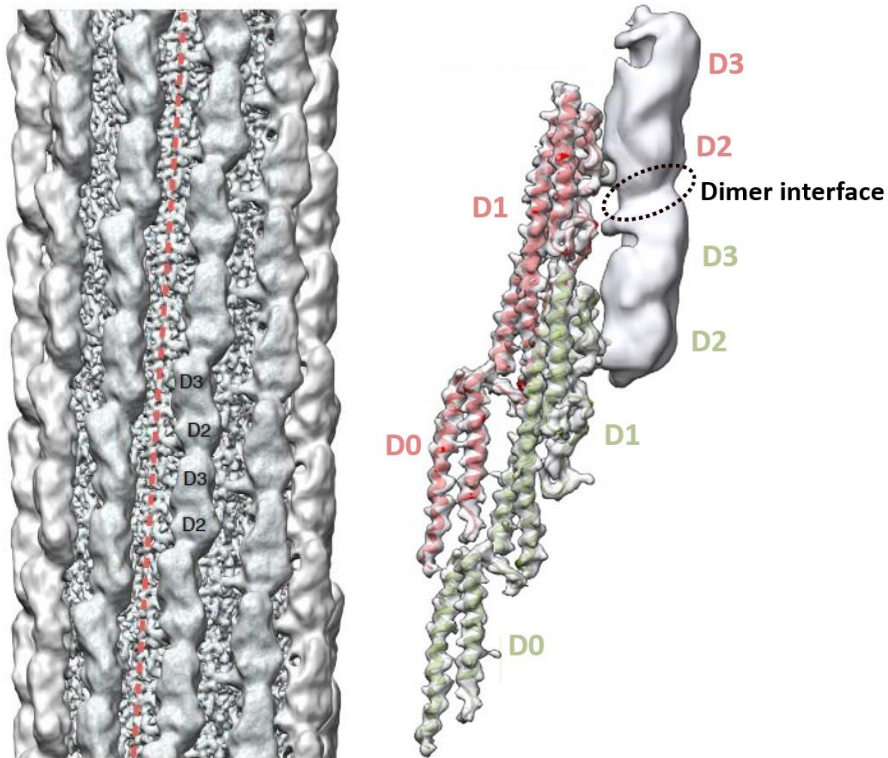


Figure 48: Cryo-EM reconstruction of the flagellum filament of *P. aeruginosa*. The electron density of the hypervariable domains have a lower resolution than the constant domains in the Cryo-EM reconstruction. The interactions are between D2 and D3 in an axial direction between adjacent monomers. In contrast to EcN FliC the hypervariable domains are not connected directly to the constant domain and not via a long linker, also in EcN the interactions seem to be between D2 and D3 as well as between D3 and D4. The figure was adapted from **Wang et al., 2017 [134]**.

6.4.1. Flagellum Hypervariable Region Orientation

The open and closed EcN flagellum models, as well as the *Salmonella typhimurium* model, show a comparable radial hypervariable region orientation. Viewed from the tip of the flagellum towards the stator, the hypervariable domains are protruding from the inner filament in a clockwise (CW) manner. In these two bacteria, the flagellum rotates most of the time counter clockwise (CCW) during the run mode (0.9-3.5 s), which produces thrust, leading to a trans-localization of the bacterium. The CW rotation in the tumble mode is shorter (0.2-0.4 s) and leads to reorientation of the bacteria [135]. The CW orientation of the hypervariable region is in line with the dominant CCW flagellum rotation direction. This architecture may increase stability against shear stress, where shear forces in a rotating flagellum would rather decrease the flagellum diameter, by pressing the monomer against each other, while a force in the opposite direction would bend the hypervariable domains outward, increasing the size.

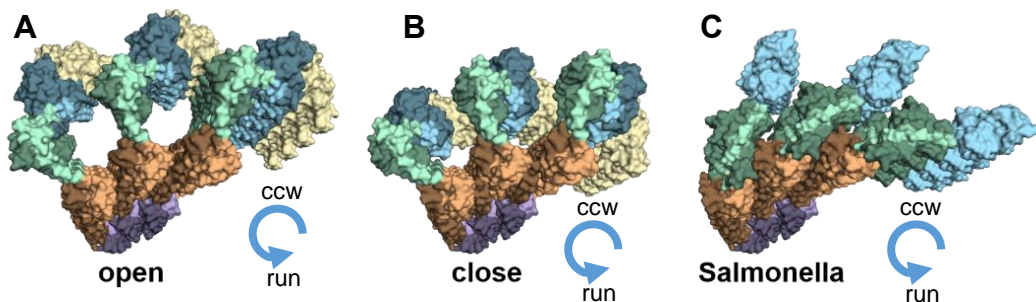


Figure 49: Three axial rows of monomers from different flagellum models. A: The opened model of EcN is shown, where D4 (yellow) points outwards with a large solvent accessible area between the hypervariable and constant region. **B:** The closed model of EcN shows a smaller solvent accessible area and D4 points inwards. **C:** The model of *Salmonella typhimurium* flagellum (1I01). Blue arrows indicate the counter clockwise rotation, leading to a run mode movement of the bacterium. Surface representation of FliCs with D0 (purple), D1 (orange), D2 (green), D3 (blue) and D4 (yellow). The N-terminal domain parts are shown in lighter colors.

6.4.2. The Wheel-Like Flagellum Model

The EcN flagellums presented here are models to gain experimental information about the flagellum. A more accurate model could be built with cryo-EM images that would allow to determine the diameter of the flagellum. This would allow to model a structure that resembles the measured diameter, or even fitting the x-ray structures into a cryo-EM reconstruction.

Nevertheless, the EcN models lead to some hypotheses that could be addressed experimentally. First, there are interactions between the hypervariable domains. There are some facts that support this statement: Due to the size of the three hypervariable domains, it would only be possible to generate a flagellum where the hypervariable domains do not interact, if the hypervariable domains are oriented in a strict radial direction. A domain orientation that would allow this was not observed in the crystal structures, where the domain orientation would lead to a more tangential orientation in the flagellum. Also, any shear forces generated through the flagellum rotation would need to be compensated only by the linker region, which is unlikely. As a consequence of interactions between hypervariable domains, some physical properties of the flagellum would change, which can be addressed with experiments. The stability of the flagellum should increase as well as the resistance against deformations. Here, the EcN flagellum could be compared with a flagellum missing hypervariable domains or with point mutations at possible interaction sites. Second, the flagellum hypervariable domains may adopt different orientations. The diameter of the two models vary by 40 Å, where the opened conformation has 140% the size of the closed model. This may represent a physiological adaptation to different properties of the media, such as viscosity, salt concentration, or pH. Here, size measurements of rotating flagellums in different solutions could provide some evidence. As this might not be easy to achieve, simulations might provide first clues.

The open model resembles a rod like structure with ridges of axial connected hypervariable domains separated by solvent accessible grooves. In the closed model, the ridges are moved to the side closing the gaps between them, resulting

in a smooth cylindrical shape. Both models have in common that the outer structure, which is stabilized by interactions, is connected with the linker region (**Figure 50**). What beneficial effects could such an additional polymerization provide? First, with increasing size (molecular weight) and surface area, sheer forces will increase, which will at a certain point exceed the binding forces provided by the constant regions. This limit needs to be either compensated by a highly stable constant region or by interactions of the hypervariable domains. Second, the linker region and an outer structured region might be a realization of a lightweight construction, where the linker is like a spoke in a wheel. This would save energy by reducing the number of needed AAs building blocks and the mass which needs to be accelerated when the flagellum changes the rotation direction. Another reason of this architecture might be that the linker allows bending flexibility of the flagellum. Third, the hypervariable domains multimeric structure might provide protection for the constant region multimer. On one hand, the outer structure with the linker region could extenuate exogenous radial forces. On the other hand, the hypervariable domain could shield the flagellum from antibodies or proteases. Lastly, the multimeric hypervariable domain structure might be crucial as a binding interface for other molecules, e.g. mucin type 2 [132]. If the flagellum could change its overall hypervariable domain orientation (open, closed) this might also be a possibility to modulate the binding affinities e.g. the groove in the open conformation may serve as a huge interface, whereas the binding might be different in a closed conformation.

The model based predictions are somewhat speculative but may help to design experiments to correlate structure with function, e.g. size measurements in different environments, thermal stability, protease stability, antibody mapping or binding experiments.

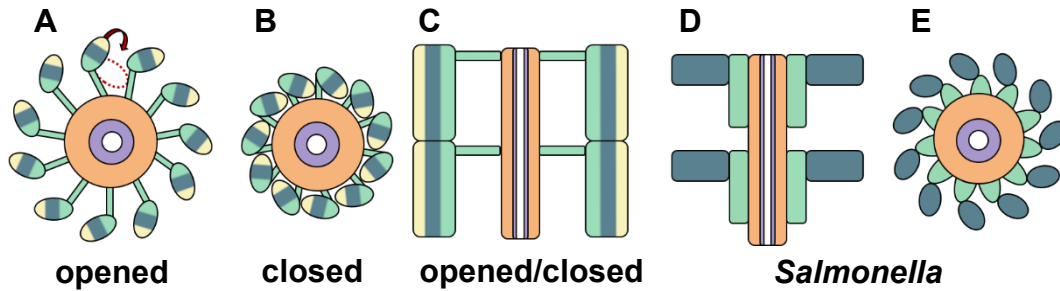


Figure 50: Schematic flagellum models. The opened (A) and closed (B) flagellum models from EcN viewed along the flagellum axis. The hypervariable region shift from opened to closed model is indicated with a red arrow and a dotted red circle. For comparison the *Salmonella* model is shown in axial orientation (E). The radial view of the opened/closed EcN (C) and *Salmonella* (D) model shows the main different between the flagellum structures. The hypervariable regions of EcN does not interact with the constant region but with other hypervariable regions in axial direction or in the closed model additionally in tangential direction. In the *Salmonella* model each hypervariable region only interacts with its constant region. D0 is shown in purple, D1 in orange, D2 in light green, D3 in dark green and D4 in yellow.

6.5. FliC Dimerization

SEC experiments report an apparent molecular weight for all constructs above the size of a monomer (Table 23). This is either an effect of dimerization or the non-globular shape of the molecule. The homo dimerization should prevent further oligomerization, as no additional peaks or shoulders were observed in the SEC. Crystal contact analysis suggest one potential interaction interface, which is also symmetric, and therefore would prevent further oligomerization, in agreement with the observed data. This interface is located at the tip of D4 and involves the β -triangle (Figure 51). In addition, this interface buries a solvent accessible area of 825 \AA^2 and was rated as biologic by the EEPIC-server [136].

Contradicting to this, the D4 deletion mutant did not reduce the apparent molecular weight in the SEC. This would suggest that more than a single dimerization interface exists. The D4 tip dimerization might also allow a dimerization of the D0, which was not included in the crystal structures. To validate if the possible D4

dimerization interface is of physiologic relevance, mutation studies, combined with functional data or affinity measurements are needed.

Table 23: Molecular weight of different constructs and their apparent weight on the SEC.

Construct	MW _{monomer}	MW _{apparent}
FliC	65 kDa	208%
FliCΔD4	52 kDa	204%
FliCΔD0	52 kDa	177%
FliCΔD01	35 kDa	149%
FliCΔD034	25 kDa	177%

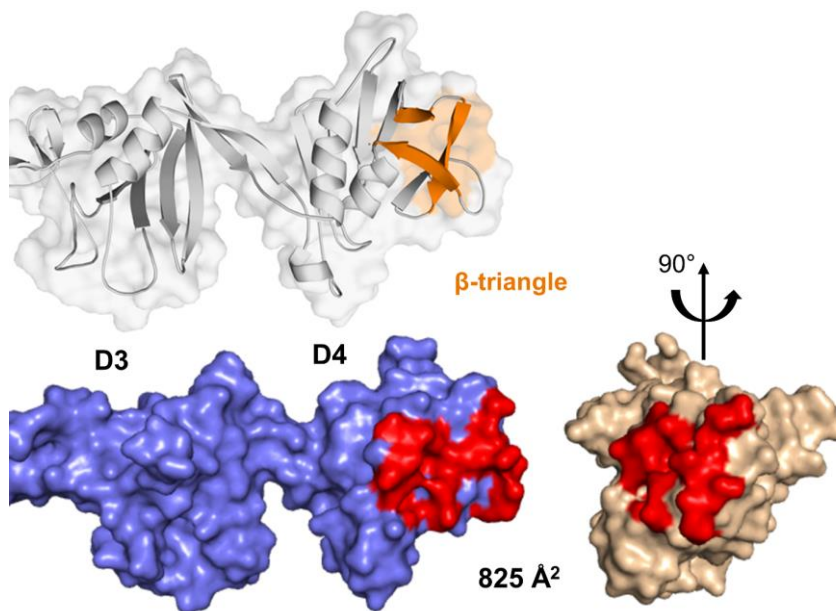


Figure 51: Possible D4 dimerization interface. Surface and cartoon representation of D3 and D4 including the β -triangle (orange). The possible dimerization interface comprises an area of 825 \AA^2 (red). For better visualization the interface is opened up and one monomer (brown) is rotated 90° as indicated.

6.6. FliC TLR5 Interaction

There are two structures of TLR5 (*Danio rerio* chimera) in complex with FliC (*Bacillus subtilis* or *Salmonella dublin*) available (5GY2, 3V47). These structures show that TLR5 interacts with D1 of FliC. The structure from *Bacillus subtilis* (5GY2) contains only D1, whereas the structure of *Salmonella dublin* includes a hypervariable region, which points away from the FliC-TLR5 interface. TLR5 signalling is induced when two TLR5s bound to FliCs form a dimer, whereas D1 from FliC contributes with some interactions to the TLR5 dimerization. In the *Salmonella dublin* structure, the two FliC hypervariable regions are in proximity, but do not interact (**Figure 52**).

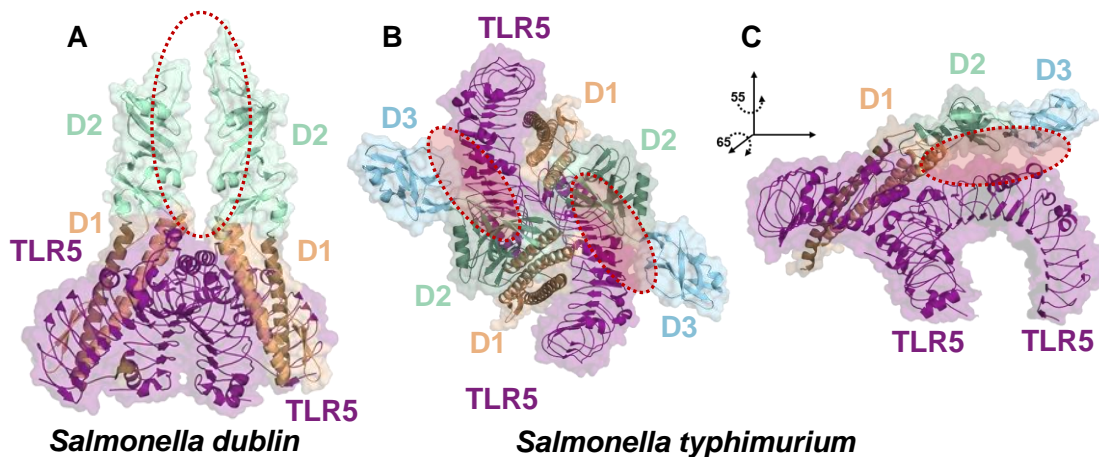


Figure 52: TLR5-FliC interaction. **A:** Side view of FliC from *Salmonella dublin* in complex with TLR5. The hypervariable domain (green) points upward away from TLR5. The two hypervariable domains come close to each other, but do not interact, highlighted with a dotted ellipse (red). **B:** Top view of the TLR5 interaction with FliC from *Salmonella typhimurium* (1I01). The model was generated through structure superposition of D1. Areas of possible interactions are highlighted with dotted ellipses (red). **C:** Side view of **B** rotated as indicated ($y = 55^\circ$; $z = 65^\circ$), with only one FliC to visualize the area of possible interaction between the FliC hypervariable region and TLR5. Protein chains are shown in cartoon and surface representation for TLR5 (purple), D1 (brown), D2 (green) and D3 (blue).

D1 of *Salmonella typhimurium* (1IO1) can be structurally aligned with D1 of *Salmonella dublin* (3V47) to generate a model of the TLR5 interaction. In this case, the hypervariable domain points towards the opposing TLR5. This could be a mechanism in which the hypervariable regions of FliC influence the TLR5 activation. An interaction would either gain energy through additional binding and dimer stabilization, or hamper the dimer formation through steric hindrance. This interaction could be either between the hypervariable domains of two FliC bound to TLR5, or between one hypervariable domain and the opposing TLR5.

When the TLR5 interaction of EcN FliC is modelled (superposed using D1 as reference), the hypervariable region points away from the TLR5 dimer and is far away from the TLR5 and the second FliC. Therefore, neither steric hindrance nor binding could be expected (**Figure 53**). The here proposed mechanisms how the hypervariable region could influence the TLR5 dimer interaction could not be observed for EcN FliC. The hypervariable domains are not close enough for binding or steric effects.

It was described that EcN FliC leads to a strong TLR5 activation and that the hypervariable region influences this interaction [60]. The model based on the crystal structure shows complete absence of interaction possibilities of the hypervariable domains caused by the linker. This renders D1 well accessible, which could be an explanation for the strong TLR5 activation. Further investigations with other FliCs containing a linker region could be of interest in this regard. Other aspects of the hypervariable domains might also influence the TLR5 activation. A dimerization of FliC, which does not need to be compatible with TLR5 dimerization, may increase local FliC concentrations. The hypervariable region may influence the stability of the FliC or may also facilitate additional binding to the cell surface or mucus [132].

Another aspect of FliC-TLR5 activation and the influence of the hypervariable region could be the *in vivo* environment, where multiple different bacteria are present at once. A well accessible FliC, such as EcN, could influence the TLR5 activation levels if the proposed mechanisms is valid. Two FliCs from different

organisms may bind two TLR5s and form an active TLR5 dimer. In this case, a well accessible FliC which also points away from the second bound FliC may increase TLR5 activation levels. A sterically hampered FliC might prefer to form an active TLR5 complex with a well accessible FliC such as EcN FliC, thereby increasing the concentration of active complexes. To validate the hypotheses, carefully designed TLR5 activation experiments may provide some evidence. Those studies would be rather challenging, as mutant with the same constant regions needs to be designed, whereas D0 also stabilizing the dimer, shows to be degradation sensitive.

The degradation sensitivity of D0 might be a bacterial mechanism to reduce detection by TLR5. Reducing the affinity of D0 towards each other is not an option, as the interaction of D0 is important for the bacteria for linking consecutive monomers in the export and transport mechanism of FliC through the flagellum channel. From a host point of view, the instability of the D0 could allow detection of living (or recently died) bacteria, shedding intact FliC, while partially degraded FliC would lead to a reduced TLR5 signals.

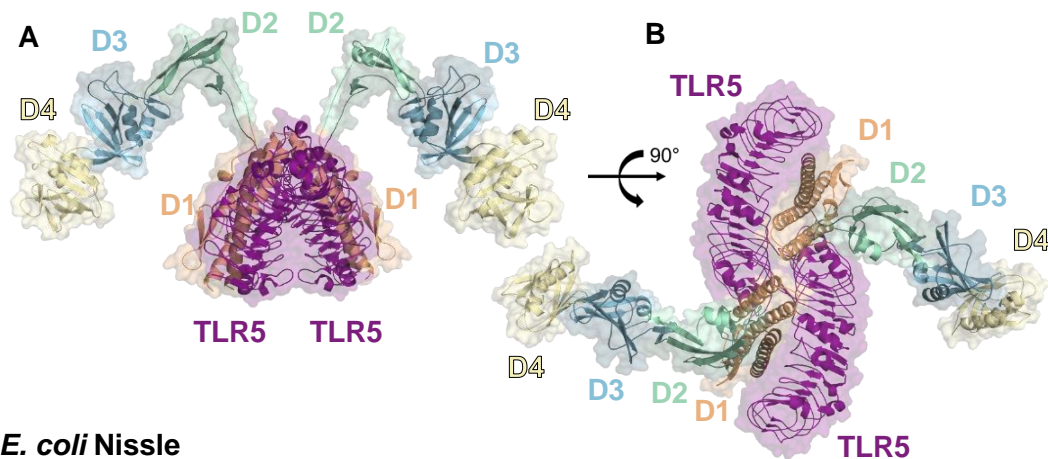


Figure 53: EcN FliC TLR5 interaction model. D1 of EcN FliC and the *Salmonella dublin* FliC were superposed to generate this model. **A:** Side view of the TLR5-FliC homodimer. **B:** As **A** but rotated 90° as indicated. The TLR5 is shown in purple, FliC in brown (D1), green (D2), blue (D3) and yellow (D4), respectively. The model shows that the hypervariable domains of FliC point away from the TLR5 dimer and also from each other.

III. Glycoprotein C from HSV-1

1. Introduction

1.1. Herpesviridae

Herpesviridae have a broad host and cell range. For humans, eight pathogenic types of herpes viruses are known: Herpes simplex virus type 1 (HSV-1; HHV-1), herpes simplex virus type 2 (HSV-2; HHV-2), Varicella zoster virus (VZV; HHV-3), Epstein-Barr virus (EBV; HHV-4), Cytomegalovirus (CMV; HHV-5), human herpesvirus 6 (HHV-6A/B), human herpesvirus 7 (HHV-7) and Kaposi's sarcoma associated herpesvirus (KSHV; HHV-8) [137].

1.2. HSV Pathogenesis

In 2016, about 491 million people were living with a HSV-2 infection and 3'583 million people with HSV-1 infections in the age cohort of 49 or younger [138]. The two viruses differ in their primary transmission. HSV-1 is orally and HSV-2 is sexually transmitted, but both viruses can infect any part of the skin or mucosa [137]. Infected cells lose their intact plasma membrane and form multinucleated large cells with degenerated nuclei. Vesicle-like fluid accumulations emerge between the epidermis and dermal layer, containing high loads of virus. Symptoms are skin lesions, ulcers and inflammation that are associated with pain and itchiness, but the infection can also be asymptomatic while shedding the virus [139]. An infection, especially with HSV-2, increases the risk of acquiring and transmitting other sexually transmitted pathogens, especially human immunodeficiency virus (HIV) [140, 141]. Immunocompromised individuals have a higher risk for a severe course of HSV associated diseases [142, 143]. In addition to the common infection of epithelial cells and sensory neurons, HSV cause life-threatening herpes simplex encephalitis (HSE) in one out of 250'000 to 500'000 individuals per year [144]. Here, the virus reaches the brain, where the lytic

replication mechanism and the subsequent immune reaction causes tissue damage with a broad range of symptoms such as fever, headache, behavioral changes, hemiparesis, reduced level of consciousness, seizures and amnesia [144]. Untreated, the mortality rate is above 70%, with survivors suffering from severe sequelae [145].

1.2.1. The Lytic and Latent Phases

After primary infection with HSV, the virus persists lifelong, interrupted by reoccurring infectious phases [146]. HSV infections are generally acquired by contact of the virus with mucosa. After infection of the epithelial cells, the virus replicates in these cells and enters termini of sensory neurons at the site of the primary infection. Next, the virus travels through the axon to the cell body of the neuron [147]. Here, the virus either replicates in a lytic process or establishes a latent infection. In the latent phase, a small subset of viral genes are expressed while most others are epigenetically silenced [148]. When this low level gene expression reaches a threshold, in combination with other stimuli, the cell switches to the lytic phase [149]. In the lytic phase, genes are activated that are needed for virus progeny production. Stress induced stimuli can induce transcription of lytic genes resulting in a switch from a latent to a lytic phase. The latent infected sensory neurons act as a virus reservoir, where in reoccurring lytic phases epithelial cells are infected, which shed the virus to the environment, ultimately resulting in the infection of new hosts [148].

1.3. HSV-1 Architecture

The diameter of HSV-1 is about 2'000 Å, and the virus is composed of four distinct layers: core, capsid, tegument and envelope. The core consists of a double stranded 152'000 bp DNA genome, which is condensed and protected in the capsid. The capsid has a pseudo icosahedral symmetry and a diameter of about 1250 Å. The triangulation number of the capsid is 16 formed by 150 hexons, eleven

pentons and one dodecameric portal vertex at a penton position, which is involved in genome packing [150, 151]. The tegument is a proteinaceous layer with a variable thickness containing proteins with key roles in the early events of the infection, such as capsid transportation through the cell or DNA insertion into the nucleus [152, 153]. The outer layer of HSV is the envelope, a lipid bilayer containing 15 viral proteins, which mediate viral entry and tropism [154].

1.4. HSV-1 Envelope Proteins and the Entry

HSV-1 typically infect epithelial cells and neurons but can also infect other cells such as fibroblasts and lymphocytes [155]. The virus enters the cells either by endocytosis or fusion of the viral and host membranes. If the virus is up taken by endocytosis, the endosomal membrane fuses with the viral membrane in a pH dependent step, releasing the capsid in the cytoplasm [156]. Once the capsid reaches the cytoplasm, the capsid facilitates the delivery to the nucleus by interaction with the microtubular network [157]. When the capsid has reached the nucleus, it binds to the nuclear pore complex and releases the viral DNA into the nucleus [158]. The early steps of infection, attachment and entry are mediated by the proteins on the outer side of the enveloped virus. The envelope contains 15 different proteins, where twelve of those are glycosylated [154]. In the first step, the virus attaches to the cell surface to bring the host and viral surface in close contact to facilitate binding of other viral receptors. The attachment step depends primarily on glycoprotein C (gC), which binds to heparan sulfate (HS) [159]. In absence of gC, glycoprotein B (gB), which also binds HS, can partially substitute this function of gC [160]. Therefore, gC is not essential for HSV entry, but the infectivity is reduced about tenfold in the absence of gC [159]. For virus entry, the glycoproteins gB, gD, gH and gL are essential [161]. Thereby, gD binds to the host entry receptors nectin-1, herpesvirus entry mediator (HVEM) or a specifically modified HS, which contains a sulfatation at the hydroxy-group of C3 [162]. The binding of gD to one of those entry receptor triggers a conformational change, which allows the interaction of gD with a gH/gL heterodimer. This activates the

gH/gL heterodimer, which can trigger a structural rearrangement of gB to a fusogenic state [163]. Activated gB inserts its membrane-penetrating N-terminus into the host membrane and pulls the viral and host membranes together to enable membrane fusion (**Figure 54**) [164, 165].

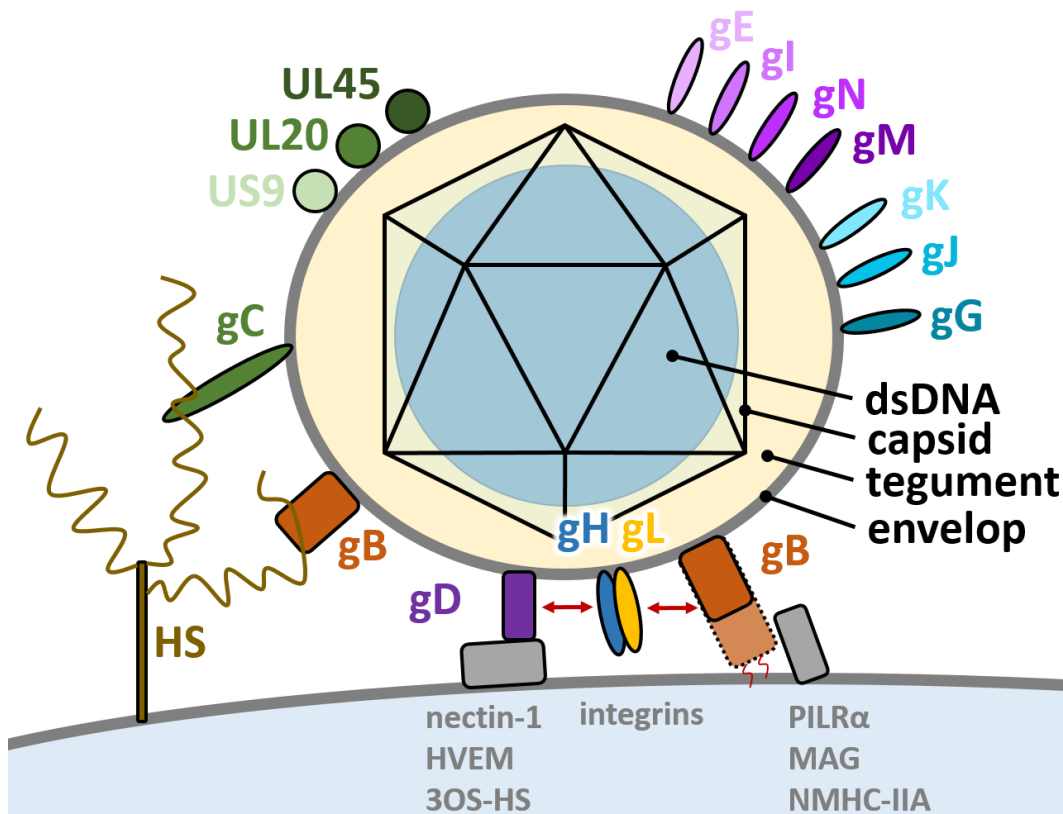


Figure 54: Schematic representation of the HSV envelope proteins. Attachment to HS is mediated by gC or gB. After the attachment, gD binds to a HSV entry receptor (nectin-1, HVEM, 3OS-HS). This binding event is transmitted as fusion signal via the gH/gL heterodimer to gB (red arrows). After a structural rearrangement, gB inserts into the host membrane and promotes fusion. gB can also interact with PILR α , MAG or NMHC-IIA and gH/gL with integrin resulting in fusion or endocytosis, respectively. The other envelope proteins which have other not essential functions are indicated.

gB can also engage directly to entry receptors, such as the “paired immunoglobulin-like type 2 receptor alpha” (PILR α), the “nonmuscle myosin heavy chain IIA” (NMHC-IIA), or the “myelin-associated glycoprotein” (MAG), which can all trigger, with the help of gH/gL, fusion of the viral and host membranes [162]. The heterodimer gH/gL can bind certain integrins, which triggers endocytosis. Additionally to these four glycoproteins, which are essential for entry, and gC, which enhances the viral attachment, ten other envelope protein exist, which influence the pathway that HSV uses to enter a cell. The glycoproteins gE and gI promote cell to cell spread across junctions [166], while gN, gM and the non-glycosylated envelope protein UL45 are involved in syncytia formation [167, 168]. The non-glycosylated protein US9 is involved in the anterograde transport through axons [169]. The non-glycosylated protein UL20 and the gK can interact with gB and are involved in the fusion with neuronal membranes [170, 171]. gJ protects the cells from apoptosis and seems to actively promotes ROS production [172]. gG is needed for an efficient apical infection of polarized epithelial cells [173].

1.5. Glycoprotein C

Glycoprotein C is not essential for the HSV attachment, entry, or steps in the viral live cycle, but it increases infectivity and fitness of the virus in the host environment. So far, three functions of gC have been described: First, the attachment to HS, second, the interference with complement component 3 (C3) of the innate immune system, and third, shielding of other envelope proteins from the recognition by the adaptive immune system.

1.5.1. A Heparan Sulfate Receptor

Heparan sulfate (HS), a glycosaminoglycan (GAG), serves as an attachment receptor for HSV-1 and HSV-2. In the absence of HS, the GAG chondroitin sulfate can also serve as a receptor [174]. GAGs are typically linked to a serine residue, starting with a β -glycosidic linked Xylose, which is β -4 linked to a galactose, followed

consists of 10 to 12 monosaccharides, while optimal binding is achieved with a HS length of 16 to 18 monosaccharides [177]. Preparations with HS containing 1.5 sulfate groups per disaccharide unit showed binding while preparations with 0.5 or 0.7 sulfate groups per disaccharide showed no significant binding. Each kind of sulfatation (NS, 2O, 6O) is needed for an optimal HS binding [178]. The rare sulfatation at the HS hydroxyl-group at C3 serves as receptor for gD, promoting membrane fusion [179]. N-sulfatation has been reported to be less important for gC binding than 2O- or 6O-sulfatation. Binding studies showed that the preferred disaccharide in 12-mer HS fragment might be IdoA(2S)-GlcNS(6S) [178]. On the surface of gC charged arginine residues are involved in the binding of HS. The crucial residues for HS binding, six arginine residues (Arg129, Arg130, Arg143, Arg147, Arg151, Arg160) and one isoleucine residue (Ile142), have been identified by mutational studies [180]. The binding of HS is probably driven by the negatively charged sulfate or acid groups of HS and the positively charged arginine residues at the surface of gC. Some charge patterns of the HS chain in combination with the “tertiary” structure of the HS chain resemble better the positions of charge residues at the gC surface than others. Therefore, distinct patterns bind stronger to gC generating some degree of specificity. Additional, interactions with nonpolar groups could increase the binding affinity and may contribute to specificity.

1.5.2. A C3b Receptor

HSV have several mechanisms to evade the host immune response. The antibody immune response is reduced by binding of the antibody Fc domain by HSV gE-gI [181, 182]. The viral peptide presentation by the “major histocompatibility complex” (MHC) class 1 of infected cells is hampered by the HSV protein ICP47, which prevents transportation of peptides to the MHC by the “transporter associated with antigen processing” (TAP) [183]. The complement system activity against the virus and infected cell is reduced by gC [184]. The viral gC can bind C3 and fragments thereof, such as C3b, iC3b and C3c. When C3 is bound by gC, the interaction with C5 of the complement system is blocked and therefore the formation of a C5

convertase. Without a C5 convertase the membrane attack complex leading to cell death will not form. The interaction of bound C3b with properdin, a positive complement regulator is also blocked [185]. The gC C3b binding site was reported to consist of four regions (region 1: 124-137, region 2: 276-292, region 3: 339-366 and region 4: 223-246) [186]. In a mouse model, 50-fold more virus that contains no gC than wt viruses was needed to cause a comparable disease [185].

1.5.3. Shielding

An additional function of gC is to protect the virus from adaptive immune recognition. gC has a mucin-like region at the N-terminus, while the C-terminus is anchored in the viral or cell membrane. This mucin-like domain reduces the accessibility of gC and other glycoproteins. For viruses that do not have gC, a lower concentrations of antibodies against the essential glycoproteins gB, gH and gD were needed to neutralize the infection with those viruses [187]. The mucin-like region protects from the complement system by sterically blocking properdin and C5 to access C3b once C3b is bound by gC [185]. The mucinlike region has been shown to modulate the gC GAG interaction, where the mucin region seems to increase the dissociation of the gC-GAG complex, which modulates virus mobility and allow releases of recently produced virus [180, 188].

2. Objectives

So far, no structure of gC is available. gC was previously purified and crystallized by M.H.C. Buch [189]. Data sets of the crystallized gC were recorded, but there is no phase information available. The phase information needs to be recovered to be able to determine the structure. Once phase information is obtained the structure needs to be build, refined and analyzed.

3. Materials and Methods

3.1. Materials and Buffers

3.1.1. Chemicals

The chemicals used in this work were of analytical grade and obtained from: Sigma-Aldrich (Deisendorf, Germany), Roth (Karlsruhe, Germany), GE Healthcare (Uppsala, Sweden), Merck (Darmstadt, Germany) or Hampton Research (Aliso Viejo, USA).

3.1.2. Cell line

Chinese hamster ovary (CHO) Lec 3.2.8.1 cells were used for the protein production. These cells contain four mutations in the N- and O-glycosylation pathway, resulting in a high mannose N-glycosylation with five to nine mannoses per glycosylation site and homogenous N-acetylgalactosamine O-glycosylation [190]. CHO Lec 3.2.8.1. cells stably transfected with a a Fc-tagged HPV-1 gC construct have been produced by M.H.C. Buch, 2016 [189].

3.1.3. Buffers

Buffers, used for the purification and cleavage of gC were adopted from M.H.C. Buch, 2016 [189].

Table 24: Buffers used for the purification. The pH of the buffers were adjusted at 4°C.

Buffer	Components
ProteinA-Binding Buffer	173 mM glycine pH 9.0 500 mM NaCl
SEC-Buffer	20 mM HEPES pH 7.5 150 mM NaCl
ProteinA-Elution Buffer	100 mM glycine pH 3.0

3.2. Protein-Biochemistry

3.2.1. Protein Production

Protein production and purification was adapted from M.H.C. Buch, 2016 [189]. The CHO cells were grown in multilayer flasks (Millicell HY Flasks T-600, Millipore) to 70-90 % confluence in MEM alpha medium (Gibco) containing 10% FBS (Gibco) supplemented with 100 mM sodium pyruvate (Sigma-Aldrich) and 200 mM L-glutamine (Gibco). For protein production, the medium was exchanged to EX-CELL 325 PF CHO serum free medium (SAFC, Sigma-Aldrich) supplemented with 100 mM sodium pyruvate (Sigma-Aldrich) and 200 mM L-glutamine (Gibco). Contrary to the protocol from M.H.C. Buch, Penicillin and Streptomycin were not used. The medium containing the secreted gC was stored at -20°C.

3.2.2. Purification

After thawing of 1 L cell medium, the medium was cleared from remaining cells and cell debris by centrifugation at 9180 \times g for 10 min and filtration with a 0.22 μ m filter (MF Mixed Cellulose Ester, Millipore). The filtered medium was applied with a flow of <1.5 mL/min to a 5 mL Protein A column (HiTrap Protein A HP, GE-Healthcare) at 4°C. The column was washed with ProteinA-Binding buffer (**Table 25**), until the UV_{280nm} absorption reached baseline. Afterwards, gC was either eluted with a 100 mM glycine buffer pH 3 or the Fc-tag was proteolytic cleaved with 0.5 mg TEV-protease per 1 L media on the column. If uncleaved gC was eluted, the pH was adjusted to pH 9 with ProteinA-Binding buffer, followed by proteolytic cleavage with 0.5 mg TEV-protease per 1 mg gC-Fc. The TEV-protease, which has about the same molecular weight as the cleaved gC, was removed with Ni-IMAC (HisTrap HP, GE-Healthcare). In order to remove aggregate, a size exclusion chromatography (SEC) was performed as final purification step using a Superdex 200i 10/300 (GE-Healthcare) column with a buffer containing 20 mM HEPES at pH 7.5 and 150 mM NaCl (SEC-buffer).

Table 25: Buffers used for protein purification. The pH of the buffers was adjusted at 4°C.

ProteinA-Buffer	SEC-Buffer
173 mM glycine pH 9.0	20 mM HEPES pH 7.5
500 mM NaCl	150 mM NaCl

3.3. Structural Biology

3.3.1. Crystallization

Crystallization experiments were performed using 4 × 6 (Hampton Research, Aliso Viejo, USA) hanging drop plates with a total drop size of 1 µL in a one to one ratio of protein solution and mother liquor. The crystallization solution contained 100 mM ammonium acetate, 100 mM Bis-Tris at pH 5.5 and 17% (w/v) PEG 10'000. Protein concentration varying from 2.2 to 4.4 mg/mL were used.

3.3.2. Soaking of Crystals with Heavy Atom Derivatives

The HATODAS II [107] server was used to identify promising heavy atom derivatives for initial soaking experiments. Therefore, crystals were soaked for 1 and 10 min in crystallization solutions containing different heavy atom derivatives (**Table 26**), with a concentration of 10 mM or saturated if the solubility was lower. For Uranyl-compounds various crystals were soaked with concentrations ranging from 2.5 mM to 20 mM and soaking times between 1 min and 24 h.

Table 26: Heavy atom compounds used for gC soaking experiments.

HA-Compounds	
K_2PtCl_6	$K_2Pt(NO_2)_4$
$SmAc_3$	$Sm(NO_3)_3$
$GdCl_3$	$PbCl_2$
$HgCl_2$	K_2HgI_4
$KAu(CN)_2$	UO_2Ac_2
$UO_2(NO_3)_2$	

3.3.3. X-ray Diffraction Data Collection

All crystals were pretested at the in house X-ray system. Data sets were collected at the macromolecular crystallography beamline X06DA-PXIII of the Swiss Light Source (Paul Scherrer Institute, Villigen, Switzerland), as described in chapter **II.3.5.3**.

3.3.4. Software

For data processing, scaling, analysis, refinement, phasing, model building and figure generation the software, as described in chapter **II.3.6** was used.

4. Results

4.1. Purification of gC

The purification of gC yielded about 1 mg gC-Fc protein per liter media, with variations between batch and passage number of the CHO cells. The harsh elution conditions of the protein A column (pH 3) probably reduced the yield, as aggregated gC was accumulating on the protein A column. But also proteolytic Fc-tag cleavage on the column did not significantly increase the yield, probably due to inefficient cleavage. Nevertheless, several purifications yielded enough pure protein for crystallization experiments, including experimental phasing (**Figure 56**).

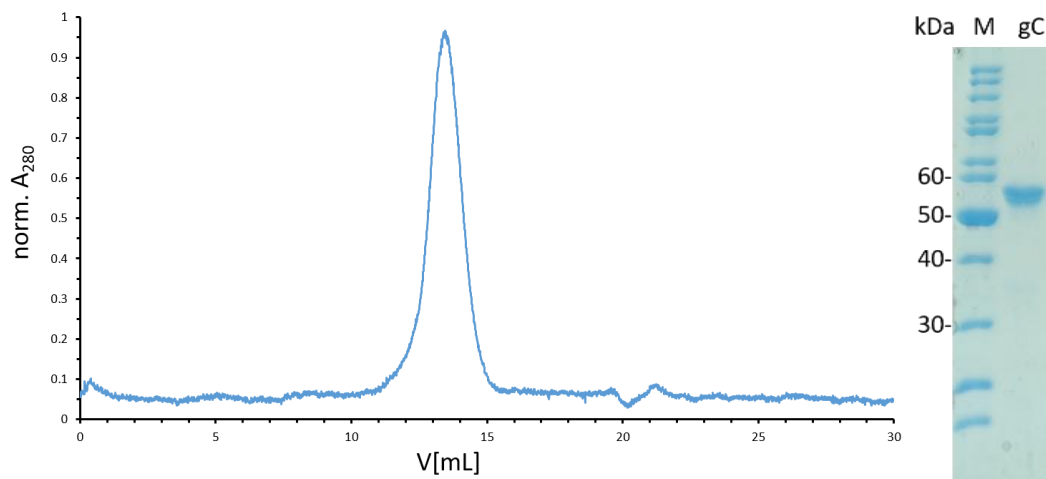


Figure 56: SEC and SDS-PAGE of cleaved gC. The SEC (Superdex 200i 10/300) shows a single peak at about 13.4 mL corresponding to a MW of 84 kDa. The SDS-PAGE showed a thicker band at about 55 kDa corresponding to differently glycosylated gC.

4.2. gC Phase Determination

Diffraction data from initial crystal soaking experiments with different heavy atoms containing compounds (**Table 26**), showed anomalous signal for uranyl soaked crystals. Due to the fragility of the crystals, optimization of the soaking procedure was needed. The highest anomalous signal could be measured from a crystal soaked for about 1 h in 5 mM uranyl acetate dissolved in the crystallization condition (**Table 27**). A SAD dataset was collected at a wavelength of 2.075 Å (5.975 keV). The anomalous signal was above $1.78 d/\sigma_d$ up to a resolution of 4.90 Å. By analysis of the data with SHELX 11 Uranyl ions were found, of which six bound with an occupancy above 50% (**Figure 57**). In the resolution shell between 5.24 Å and 4.90 Å, the phasing power was 1.0.

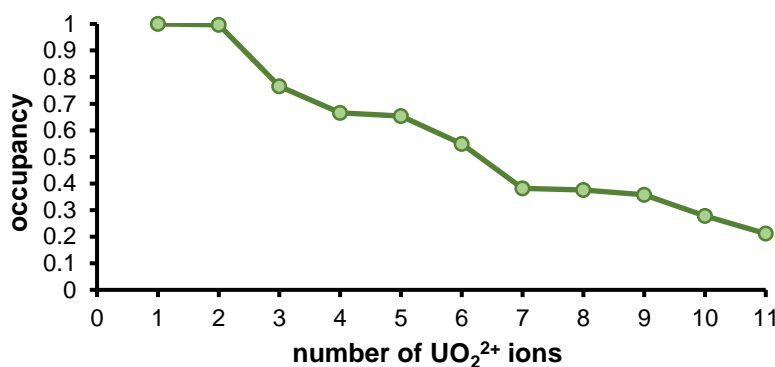
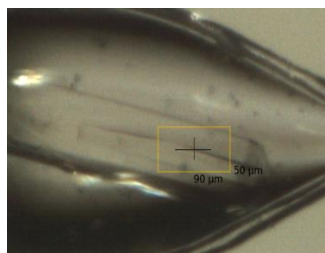


Figure 57: Occupancy of the 11 Uranyl ions found with SHELX.

After solvent flattening, an electron density map could be calculated, which allowed to build most of the protein peptide backbone. The AA sequence could be assigned and build based on the electron density of aromatic and di-cysteine side chains. At a certain point where, based on the electron density, no major model improvement were possible, the partially build model was used to solve the phases of the native dataset via molecular replacement. The model was further improved with the native dataset to current R/Rfree-values of 24.2% / 27.4% at 2.67 Å resolution.

Table 27: The crystal and the corresponding crystallization condition soaked with 5 mM uranyl acetate for 1 h used for SAD phase determination.



gC uranyl acetate derivative

4.2 mg/mL @ 20°C
0.1 M Bis-Tris pH 5.5
0.1 M NH₄Ac
17% (w/v) PEG 10'000

4.3. gC Structure

Glycoprotein C contains a highly glycosylated mucin-like N-terminal region (24-122) with five putative N-glycosylation and numerous clustered O-glycosylation sites [191]. This region, as well as the C-terminal part, which forms a transmembrane helix (478-511), was not included in the protein used for the crystallization (123-477). In the electron density map, the first residue 123 is well defined, whereas residue 466 was the last residue that could be built. The twelve missing residues at the C-terminal part of the construct are not ordered in the crystal and therefore probably are flexible. The structure shows that gC consists of three distinct domains each comprising an immunoglobulin (Ig)-like fold (**Figure 60**). The N-terminal domain 1 (D1) has an Ig-intermediate (Igl)-like fold, whereas domain 2 (D2) and domain 3 (D3) each possess an Ig-constant2 (IgC2)-like fold. The orientation of the β -sandwiches of D1 and D2 form a straight line, whereby a kink is visible between D2 and D3. The kink has an angle of about 110° when measured with the three points: C α of Gly247 at the tip of D1, C α of Arg343 at the kink between D2 and D3, and C α of Ser459 at the C-terminal part of D3.

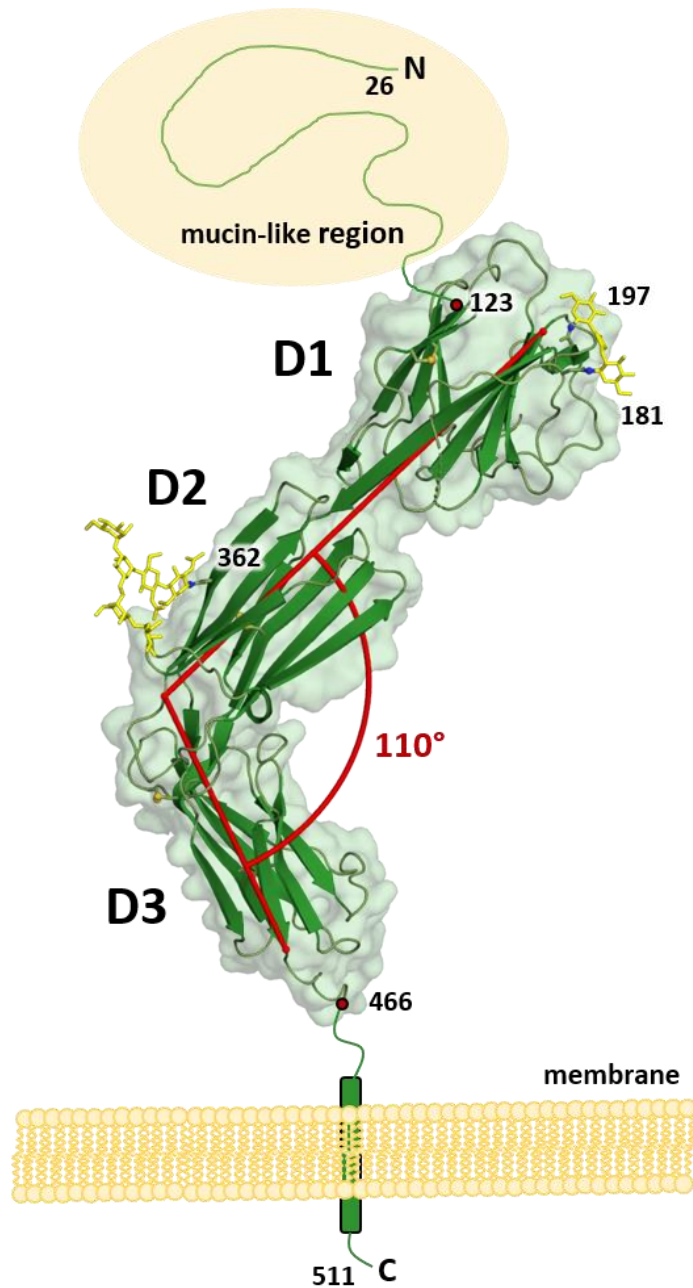


Figure 58: Schematic representation of gC from HSV-1 including the crystal structure. gC contains a mucin-like region with N- and O-glycosylation sites in the first 122 AAs depicted as yellow sphere. The crystal structure including AA 123 to 466 indicated with red dots, showed that gC consists of three IgG-like domains D1, D2 and D3, here shown in surface and cartoon representation. Three N-glycans attached to Asn residues (number indicated) could be observed and are shown as yellow stick representation. The C-terminal part of the protein anchors gC into the virus or host cell membrane via a trans-membrane helix and was not included in the used construct. The angle between D3 and D1/D2 is indicated with red lines.

4.3.1. gC-Domain1

Domain 1 of gC ranging from residue 123 to 264 consists of an Igl-like fold (**Figure 59**). This fold consists of a two layered β -sandwich, where one layer consists of four β -strands: A (125-127), B (140-146), E (226-231) and D (218-221), whereas the second layer consist of five β -strands: A' (133-135), G (254-264), F (240-247), C (155-161), and C' (189-193). The CC'-loop and the C'D-loop of gC are long, with 27 and 24 residues, respectively (**Figure 60**). Residues 164-172 of the CC'-loop were not resolved in the electron density, probably due to a high degree of flexibility. The CC'-loop covers the surface of the five stranded β -sheet, whereby the C'D-loop covers a side between the two sheets of the β -sandwich. The CC'-loop and C'D-loop contain one glycosylation each, at Asn181 and Asn197, respectively, whereas only the first GlcNAc moiety could be built in the electron density map. A structure similarity search with D1 using DALI and a representative set of the protein data bank (pdb) showed that the ten most similar structures differ in these two loops.

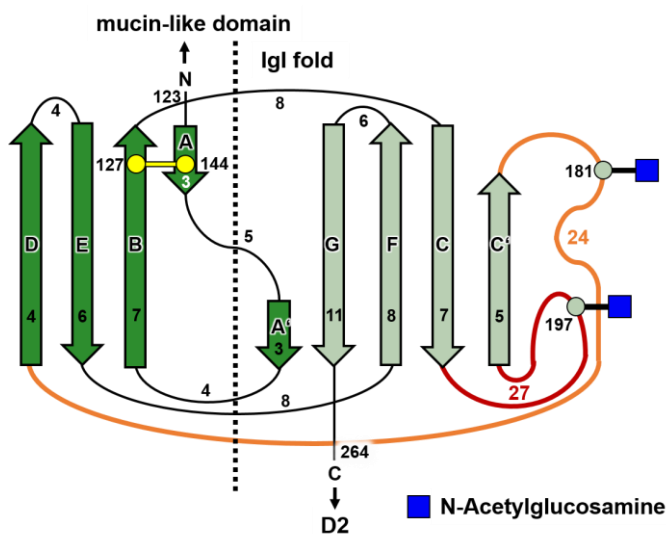


Figure 59: Topologic representation of the Igl-like fold of D1. The β -strands of the two-layered β -sandwich are labeled from A to G, here opened up at the dotted line with the length of each β -strand and loop indicated. The disulfide bond is shown in yellow with the residue numbers indicated. The two long loops, the CC'-loop (27 AAs) and the C'D-loop (24 AAs) are shown in red and orange, respectively. The N-glycosylations at Asn 181 and Asn 197 are shown as blue squares.

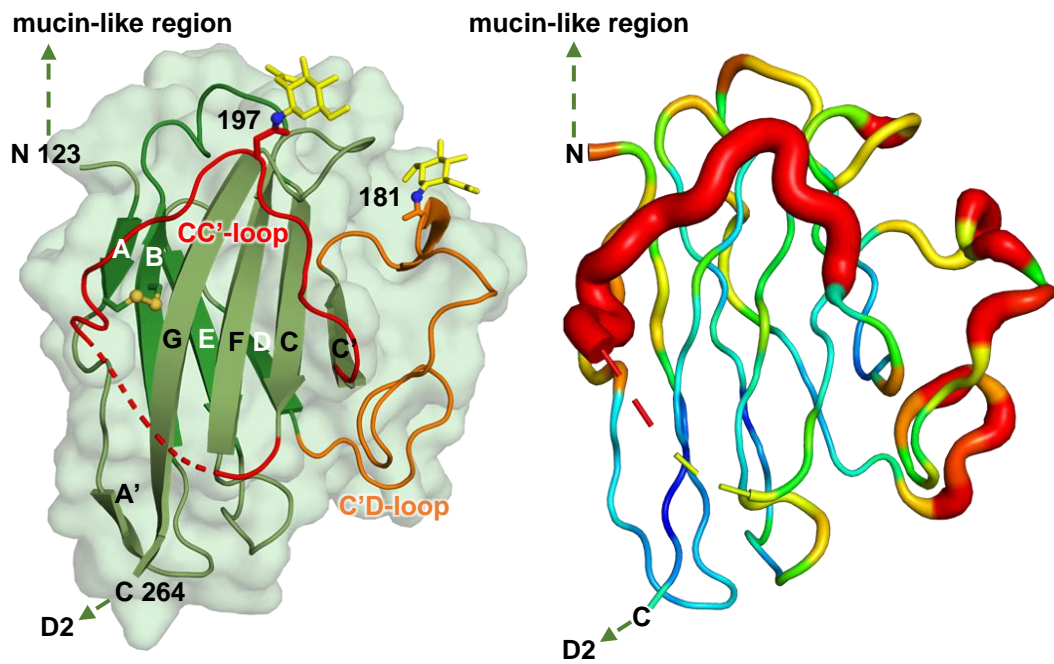


Figure 60: Structure of gC D1. The Igl-like D1 of gC is shown as cartoon and surface representation (**left side**). The β -strands are labeled from A to G and the first and last AA number of the domain is indicated. The N-glycosylations at Asn181 and Asn197 are shown in yellow stick representation. The disulfide bond between Cys127 and Cys144 is shown in dark yellow. The average C α -rmsd between gC and the ten best (**Appendix 14**) hits from a DALI [192] structure search is visualized in rainbow colors ranging from blue (≤ 0.5 Å) to red (≥ 3 Å) (**right side**). The highest rmsd was observed for the CC'-loop and the C'D-loop.

4.3.2. Transition between D1 and D2

The transition between D1 and D2 is seamless. The β -strand G of D1 forms two backbone interactions with the β -strand B of D2 in an antiparallel β -sheet manner (**Figure 61**). These are the only hydrogen bonds that stabilize the domain transition. Additional water mediated interactions between the two domains might be possible, but due to the low resolution of the crystal structure, determination of coordinated water molecules was not possible. The side chain of Arg265 of D2 is stabilized in its position by two hydrogen bonds to backbone carbonyls (Ala290 and Asp323) of D2. Additionally, this arginine forms a cation- π interaction to Tyr136 of D1 with a distance of 3.3 Å. Furthermore, Met263 and Tyr292 as well as Phe264 and Arg353 are forming van der Waals contacts between D1 and D2.

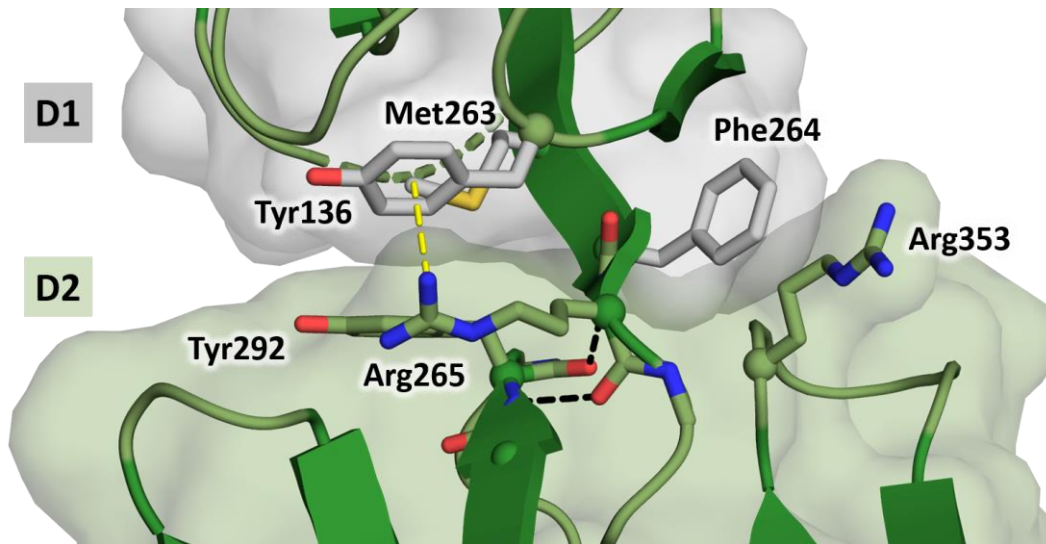


Figure 61: Transition between D1 and D2. Residues in the interface area between D1 (grey) and D2 (green) are shown in stick representation. The two antiparallel β -sheet backbone-backbone interactions of β -strand G of D1 and the β -strand B of D2, are shown with black dotted lines. The interface consists of two arginine residues, three aromatic residues and one methionine in van der Waals distance. Whereby, Arg265 forms a cation- π interaction with Tyr136 (yellow dotted line).

4.3.3. gC-Domain2

Domain 2, ranging from residue 265 to 371, has an IgC2-like fold (**Figure 62**). The structure consists of a two-layered β -sandwich, where one β -sheet is formed by four β -strands: A (268-273), B (283-291), E (324-332) and D (314-321), and the other β -sheet is formed by three β -strands: C (298-303), F (344-352) and G (358-366). The two β -sheets are linked by a disulfide bond between β -strand B (Cys286) and β -strand F (Cys347). The loops of the β -sandwich that face the side towards D3 (AB-9, CD-10, EF-11, GG'-13) are longer than the loops pointing towards D1 (BC-6, ED-2, FG-5) (**Figure 63**). Additionally, the residues connecting the A and B β -sheets form a three residue long parallel β -sheet with the C-terminus of the domain. Six sugars rings of the high mannose N-glycosylation could be observed at Asn362. A structure comparison using DALI [192] revealed that the ten most homologue structures (**Appendix 14**) varying primarily in the loop regions.

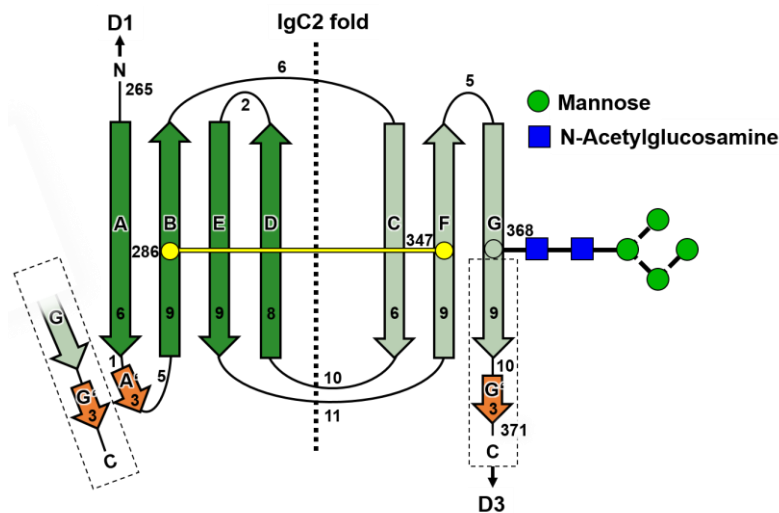


Figure 62: Topologic representation of the IgC2-like fold of D2. The β -strands of the two-layered β -sandwich are labeled from A to G', here opened up at the dotted line. The disulfide bond is shown in yellow with the residue numbers indicated. The length of the β -strands and loops are indicated. The N-glycosylation composition at Asn368 is indicated with squares and circles. The short two stranded β -sheet that links the AB-loop to the C-terminus is shown in orange.

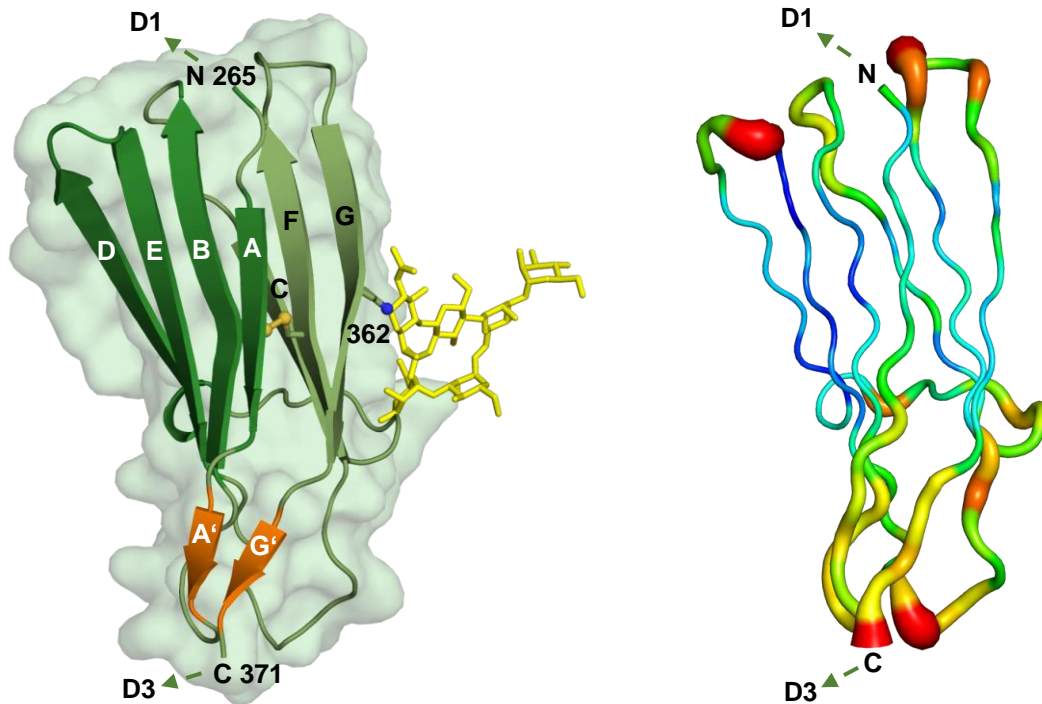


Figure 63: Structure of gC D2. The IgC2-like D2 of gC is shown in cartoon and surface representation (**left side**). The β -strands are labeled from A to G' and the first and last residue number of the domain is indicated. The N-glycosylation at Asn362 is shown in yellow stick representation. The parallel β -sheet, comprising three residues, that stabilizes the AB-loop and the C-terminal domain region is shown in orange. Average C α -rmsd between gC and the ten best (**Appendix 14**) matching structures from a DALI [192] structure based homology search, are visualized in rainbow colors ranging from blue (≤ 0.5 Å) to red (≥ 3 Å) (**right side**).

4.3.4. Transition between D2 and D3

The transition between D2 and D3 is seamless as the D1, D2 transition (**Figure 64**). The strand that connects the two domains is stabilized by a short parallel β -sheet (A'B') formed within the AB-loop. A hydrogen bond is formed between the side chain of Gln280 and the backbone carbonyl of Gly450. Met277 of D2 points into a hydrophobic pocket formed by D3. The bottom of this pocket is formed by Pro373, Tyr447 and Leu454, where the upper part is formed mainly by Leu370, Pro371, Arg372, Ile451 and Pro452.

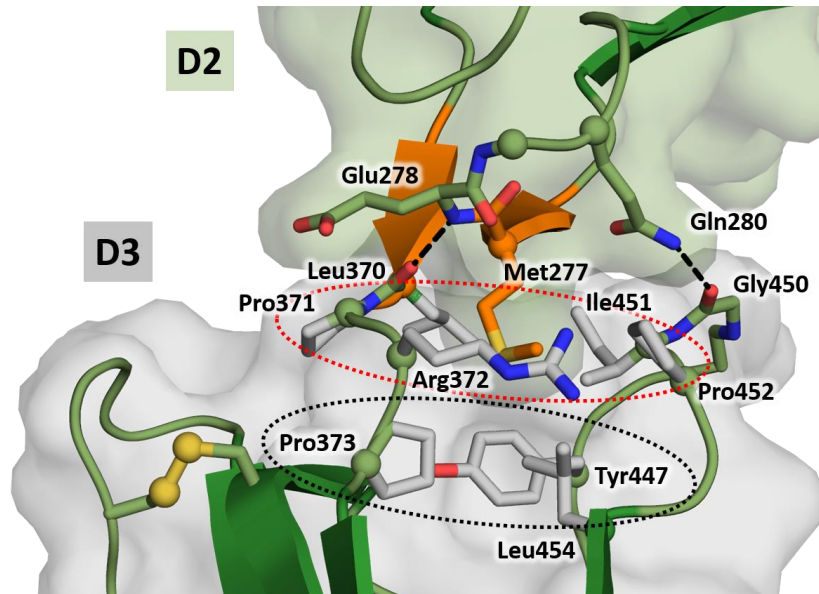


Figure 64: Interface between D2 and D3. Residues in the interface are shown in stick representation. The two hydrogen bonds between the domains are shown with black dotted lines. The short β -sheet at the interface area is shown in orange. Met277 of D2, shown in orange, points in a mainly hydrophobic pocket formed by D3. The three residues forming the bottom of the pocket are highlighted with a black dotted ellipse, whereas the five residues forming the upper part of the pocket are highlighted with a red dotted ellipse.

4.3.5. gC-Domain 3

Domain 3 ranging from AA 372 to 466 has an IgC2-like fold (**Figure 65**). The structure consists of a two-layered β -sandwich, where one β -sheet is formed by four β -strands: A (374-379), B (383-392), E (424-432) and D (413-416), and the other β -sheet is formed by three β -strands: C (396-401), F (439-445) and G (454-459) (**Figure 66**). The two β -sheets are linked by a disulfide bond between β -strand B (Cys386) and β -strand F (Cys442). Additionally, the ED-loop (417-423) is stabilized by an untypical disulfide bond between ED-loop Cys419 and β -strand B Cys390. A structure comparison with the pdb using DALI [192] showed that the ten best matching structures (**Appendix 15**) varying primarily in the loop regions (**Figure 66**).

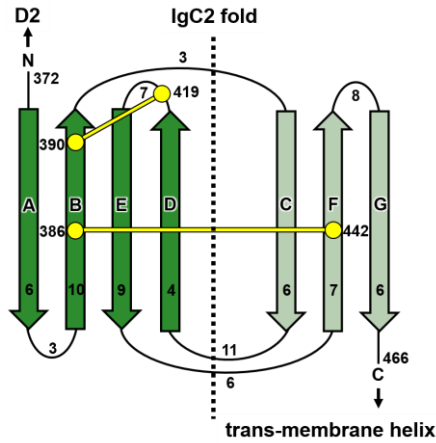


Figure 65: Topological representation of the IgC2-like fold of D3. The β -strands of the two-layered β -sandwich are labeled from A to G, here opened up at the dotted line. Disulfide bonds are shown in yellow with the residue numbers indicated. The length of the β -strands and loops are indicated.

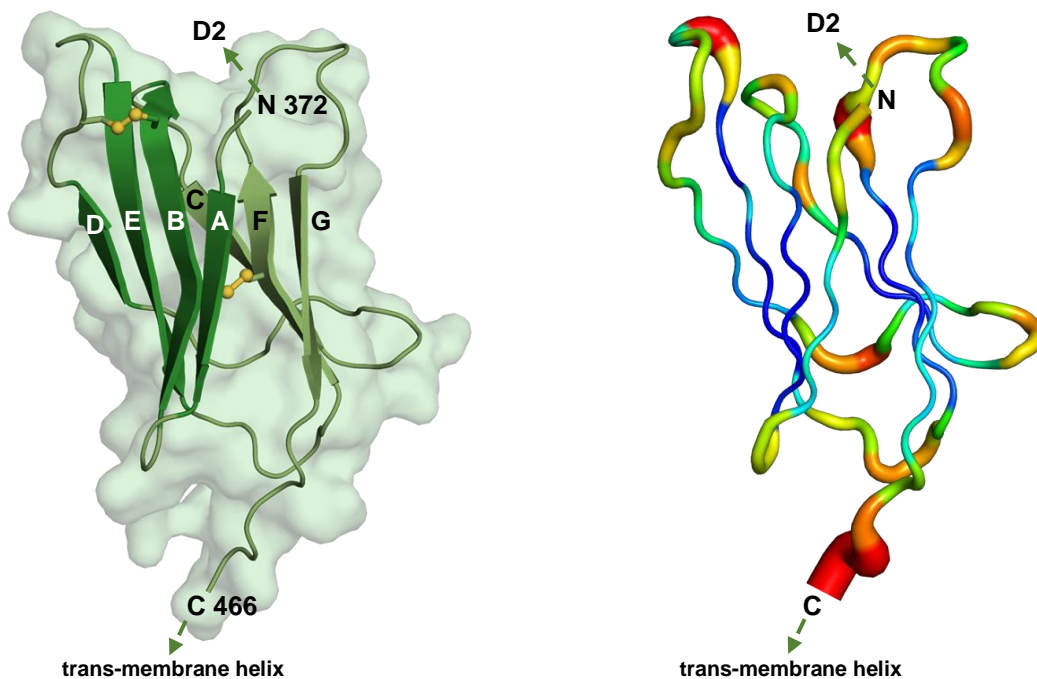


Figure 66: Structure of gC D3. The IgC2-like D3 of gC is shown in cartoon and surface representation (**left side**). The β -strands are labeled from A to G and the first and last residue number of the domain is indicated. Average $C\alpha$ -rmsd between gC and the ten best hits (**Appendix 15**) from a DALI [192] structure search are visualized in a rainbow color gradient ranging from blue (≤ 0.5 Å) to red (≥ 3 Å) (**right side**).

4.4. gC-Glycosylation

The gC protein, used for crystallization, was produced in glycosylation-deficient CHO cells, resulting in a high mannose type (5 to 9 mannoses) glycosylation pattern. The electron density map at Asn368 of D2 allowed to build six sugar rings. The first N-linked GlcNAc residue interacts, with a bond distance of 2.89 Å, via the hydroxyl group at C6 with the carboxyl group of Asp305. Additionally, there is an interaction of a terminal mannose. Here, the hydroxyl group at C2 interacts with the backbone amid of Gln307, forming a bond with a distance of 2.96 Å. This interaction might be only possible for some glycosylation types as usually the hydroxyl group of C2 is linked to additional carbohydrates. Nevertheless, these interactions probably decreased the flexibility of the high mannose type glycosylation at this position, allowing its structure elucidation.

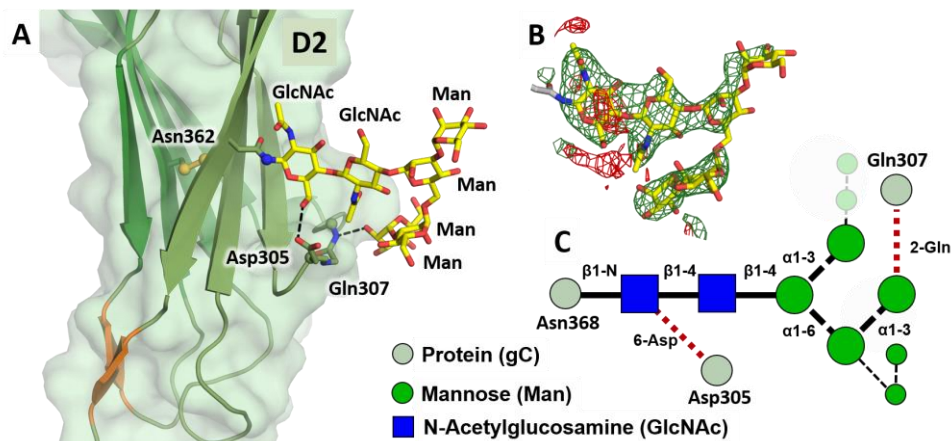


Figure 67: Glycosylation of D2. The glycosylation of Asn368 of D2 is shown in stick representation (yellow)(A). Simulated annealing difference omit map contoured at 3 sigma with the six build carbohydrates of the high mannose glycosylation (B). In the schematic figure (C), additional possible mannose residues are shown transparent and smaller. Two hydrogen bonds are formed between the protein and the carbohydrates, indicated with black (A) or red (C) dotted lines. The first hydrogen bond is formed between the hydroxyl group at C6 of the first GlcNAc moiety and the carboxyl group of Asp305. The second hydrogen bond is formed between the hydroxyl group at C2 of a terminal mannose and the backbone amid of Gln307.

The other three glycosylation sites are in D1, predicted based on the NXT/S consensus sequence for N-glycosylation (**Figure 68**). Here, the electron density map shows the first GlcNAc moiety at Asn181 and Asn197. The remaining sugar tree of the high mannose glycosylation is probably too flexible to be resolved in the electron density. For Asn148 no glycosylation was observed, here the glycosylation is either too flexible to be resolved in the electron density map or the residue was not glycosylated. All three putative glycosylation sites of D1 are at the N-terminal side of the β -sandwich, distant to the transition to D2. Asn148 is in the BC-loop, while Asn181 and Asn197 are in the two long loops, the CC'-loop and C'D-loop, respectively. Here, the C'D-loop connects β -strands at the opposed side of the β -sandwich, but the loop is orientated in a way that the glycosylation is at the N-terminal side of the β -sandwich.

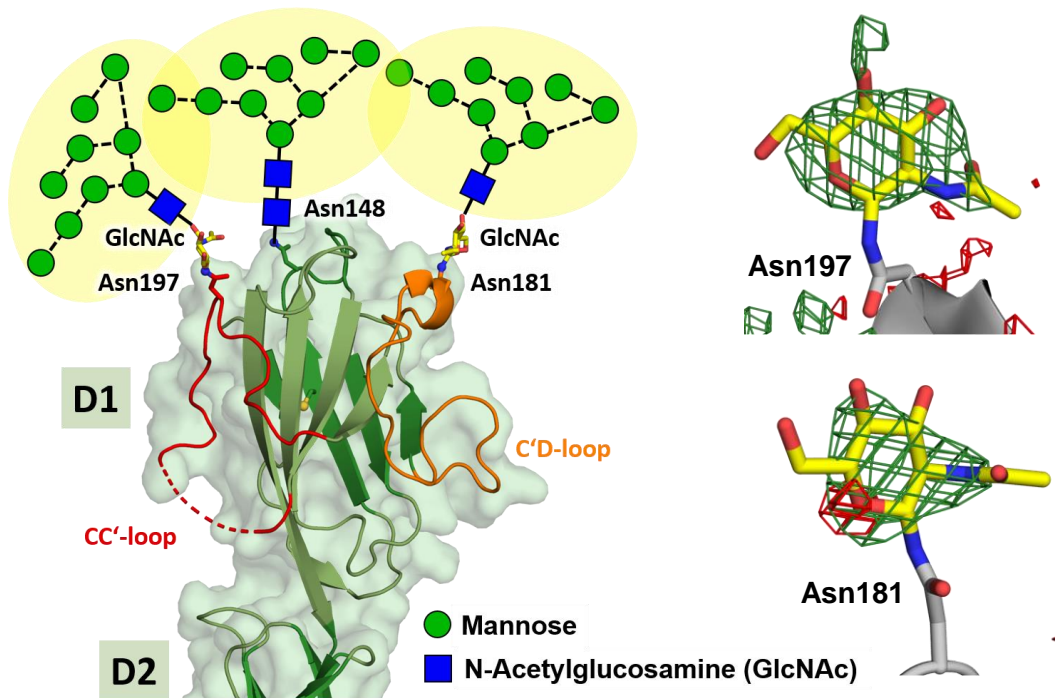


Figure 68: Glycosylation of D1. The CC'-loop and C'D-loop of D1 are glycosylated at Asn197 and Asn181, respectively. The first GlcNAc of the glycosylation could be built in the electron density map, whereby the rest of the high mannose glycosylation is indicated. No glycosylation was visible at Asn148, which might still be glycosylated. Simulated annealing difference omit map contoured at 3 sigma of the two GlcNAc residues (right side).

4.5. gC-Dimer

Glycoprotein C crystallized in space group $P6_522$ with one protomer in the asymmetric unit. From this protomer, four possible dimers are generated through crystallographic symmetry with a protein-protein interface. One of them ($x-y, x+1, z-1/6$) has a small interface area of 431 \AA^2 (PISA) and is physiologically unlikely, since the N-terminal domain interacts with the C-terminal domain. Two of the other possible dimers ($x-y, -y, -z$) ($-x+y-1, y, -z+1/2$) are orientated in a way that would allow membrane anchoring on the C-terminal side and a mucin region on the N-terminal side, but their interface area buries only 246 \AA^2 and 137 \AA^2 (PISA) [70], respectively, which is probably physiologically not significant. The fourth dimer ($-y, -x, -z+1/6$) is most probably a physiologic dimer with an interface area of 894 \AA^2 (PISA), where the two protomers are twisted around each other (**Figure 69; Figure 70**).

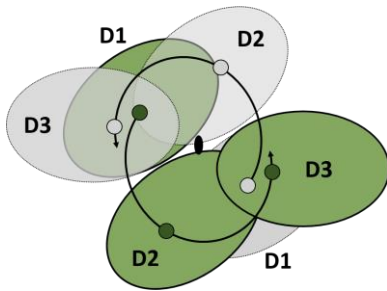


Figure 69: Dimer domain positions. The domains of the two protomers are shown along the two-fold symmetry (black ellipse) as ellipse either in green or grey. The domains are orientated CCW when viewed from D1 to D3 (outside towards membrane), indicated with arrows.

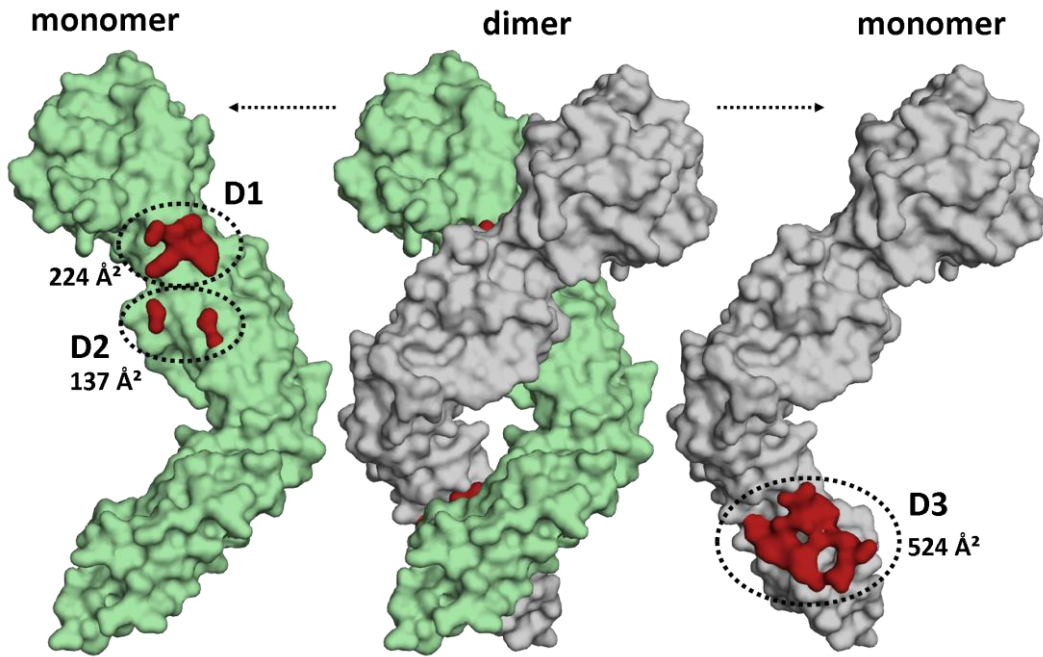


Figure 70: Putative dimer of gC. The surface representation of the gC dimer is shown in green and grey. To visualize the interface area, the monomers are translated to the sides as indicated with dotted arrows. The surface areas within a distance of 4.5 Å of the two protomers are shown in red. The involved domains are indicated with dotted black circles.

For the dimerization, more than one domain is in contact with each other, thereby the main interactions are formed between the two monomers of D3 (524 Å²). In comparison, the interface area between the two monomers of D2 and between the two monomers of D1 is 137 Å² and 224 Å², respectively. The D1 and D2 interface areas are in close proximity. The interactions have a rather hydrophilic character, which could be the reason why the PISA server, comparing free energies of solvated monomers with solvated multimers plus the energy gain through multimerisation, predicts this interaction as not physiological. On the other hand, the EPPIC server [193], which additionally takes conservation of interface residues into account, predicts this interaction as biological. The total interface area is split in two distinct sites, which are in a distance of about 42 Å to each other (center of mass of the interface residues).

The main interface area, between the monomers of D3 (**Figure 71**), is mainly hydrophilic, with only two hydrophobic residues Ile441 and Val453 involved in the interaction. Arg443 interacts with Glu455 of the same monomer, whereby this glutamate residue interacts also with the same glutamate residue of the other protomer. His456 interacts with Asp404. Glu439 interacts with His457 of the same monomer, whereby this histidine residue interacts also with the same histidine residue of the other protomer and Asp403.

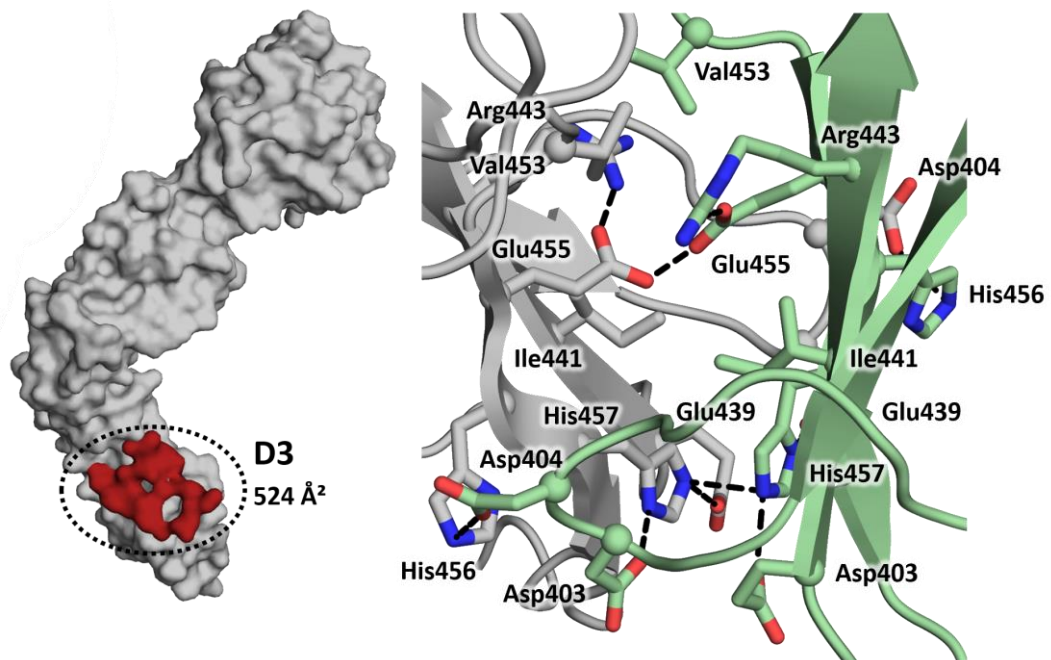


Figure 71: Dimer interface of D3. The interface area (red) of D3 is shown in surface representation (**left side**). Close up view of the D3 interactions, shown in cartoon and stick representation, with hydrogen bonds shown with black dotted lines (**right side**).

The other interface area is at the transition of D1 and D2 (**Figure 72**). At the D2 side of this interface His321 interacts with Thr270 and Ser268 interacts with Asp323. In the interface of D1 a water molecule or ion is complexed by four arginine residues (2×Arg135 and 2×Arg265). Arg135 forms a salt bridge to Glu233, whereas Arg265 forms also a cation- π interaction with Tyr136 of the same monomer. Additionally, water mediated interactions may be involved in this dimer

formation, which could not be resolved with the available data set. The interface areas between the monomers in D1 and D2 are not completely well defined by the electron density map probably due to the crystallographic two-fold axes along the dimer interface, averaging the differences in the orientation of the residues between the two protomers.

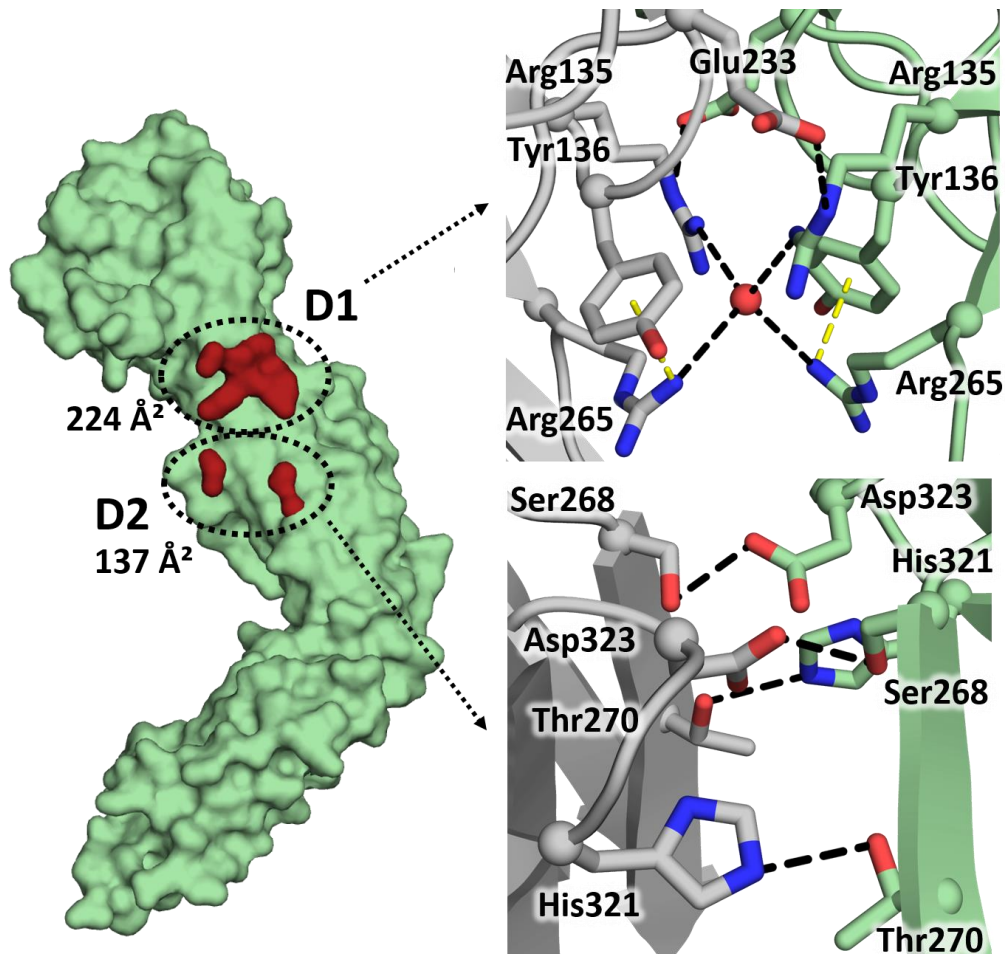


Figure 72: Dimer interfaces of D1 and D2. The interface area of the dimer is shown as surface representation in red (**left side**). Close up views of D1 and D2 shown in cartoon and stick representation, with hydrogen bonds shown with black dotted lines (**right side**). The cation- π interactions are shown with yellow dotted lines.

Based on the crystal structure, it is likely that gC forms a dimer. The full-length gC is additionally anchored in the membrane limiting the freedom of movement and pre-orientating the gC monomers suitable for dimerization. The mucin-like region at the N-terminal domain might additionally contribute to the dimerization. In solution, it is hard to distinguish between gC monomeric and gC dimeric species. The dimerization as observed in the crystal structure does not significantly alter the radius of gyration R_G when compared to the monomer **(41)**(**Figure 73**).

$$(41) \quad R_G = \sqrt{\frac{1}{m_{total}} \sum_i^N m_i r_i^2}$$

The radius of gyration R_G is calculated, with the distance r_i to the center of mass, m_i the mass of the corresponding atom and m_{total} the total mass of the object. Therefore, a monomer and dimer would be indistinguishable by methods that depend on gyration radii such as SEC, small angle X-ray scattering (SAXS) or dynamic light scattering (DLS). In the SEC, gC appears as a single peak (**Figure 56**), which could be a monomer, dimer or a mixture thereof.

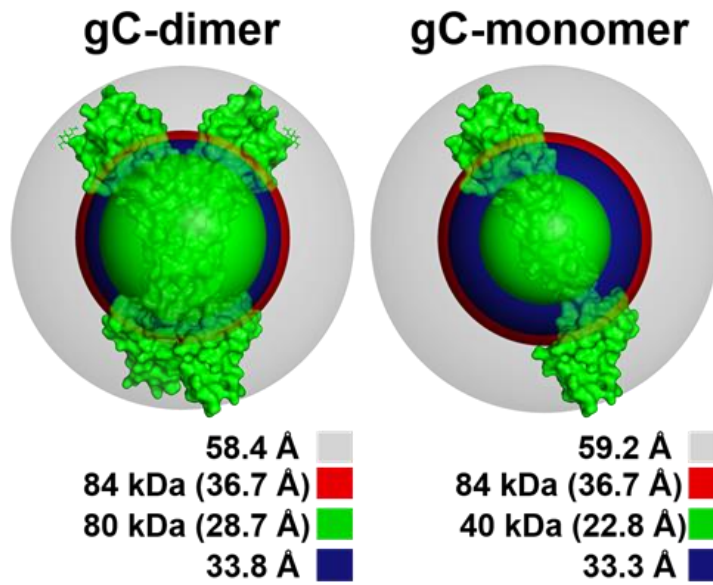


Figure 73: Different rotational radii of the gC monomer and dimer. The sphere, representing the distance from the centre of mass to the most distant C α -atom, is shown in grey. In radius of gyration of the apparent weight of gC, based on SEC, is shown in red. The radius of gyration of a perfect spherical protein, with the corresponding MW, is shown in green. The radius of gyration of the monomeric and dimeric structure is shown in blue.

5. Significance

We solved the structure of gC of HSV-1, the first gC structure of all herpes-virus. The structure revealed that gC consists of three distinct domains with an Ig-like fold. Based on the structure, we could show that gC might form a dimer, whereas D3 contributes most interactions. This could allow the design of dimer interface mutations, to validate the dimerization of gC and investigate the influence of the dimerization *in vivo*. Based on sequence conservation, the HSV-2 gC most likely has a similar domain composition and also forms a dimer. We could identify the area where HS binding most probably occurs and identify residues, which might contribute to this interaction. The C3b binding area is rather unclear, but based on the structure, mutants can now be designed to identify the C3b binding epitope. Additionally, we identified two long loops of D1 with an unknown function. These loops might bind yet unknown proteins or somehow protect the gC from antibody recognition. The structure could encourage further research to analyze the three functions of gC in HSV infections: HS-binding, C3b-binding and shielding from the recognition by the adaptive immune system. This structure might also facilitate HSV vaccine development where gC is a potential target [194-196].

6. Discussion

6.1. The Heparan Sulfate Binding Site

6.1.1. Important Residues for HS Binding

Heparan sulfate, a poly-sulfated GAG, serves as initial receptor for HSV [197], with gC or gB mediating binding [160, 198]. This GAG contains approximately 0.6-1.8 sulfate groups per disaccharide, with varying amount of sulfatation among the disaccharides [199, 200]. Together with the carboxyl group of the uronic acid, HS is negatively charged, therefore positively charged residues, such as arginines and lysines, are typically involved in GAG binding [200]. Residues that are important for HS binding have been identified for gC with mutational studies (**Figure 74**) [180]. In these studies, seven arginine residues and Ile142 were determined to have a strong effect on the HS binding. In the crystal structure, five of these arginine residues (Arg151, Arg147, Arg143, Arg129) are arranged in a line along the β -sheets of D1. The distances between a pair of successive C α s is less than 8.8 Å, whereas the sum along the line connecting the five arginine residues is 29.7 Å. The distance between two 1,4-linked glycosidic oxygens in HS is about 5.4 Å, so most likely several monosaccharides are bound by these five arginine residues. Two additional arginine residues (Arg160, Arg130) reported to be crucial for HS binding might not be involved in direct HS binding based on the structure. Arg160 lies in a pocket on the opposite side of the domain. Mutation of this arginine to alanine might have had a destabilizing effect on the fold, which probably affects the close C'D-loop. The same is true for Arg130, which is also located in a pocket with limited accessibility. Here, a mutation to alanine might also influence the stability of the domain or the close CC'-loop. This arginine residue is closer to the other arginine residues involved in binding, therefore charge effects may play contribute to binding or a complete rearrangement of the arginine side chain might be possible, but unlikely. The only hydrophobic residue Ile142 that was found to be important for GAG binding points in the hydrophobic core of D1, therefore involvement of this residue in binding is very unlikely. The structure revealed a

tryptophan residue (Trp126) at the surface of the domain next to two successive arginine residues with a C α distance of 5.3 Å to Arg145 and 4.0 Å to the closest side chain atom of Arg143. Aromatic residues, especially tryptophan residues, have been described to often be involved in CH- π binding of carbohydrates [201]. This solvent accessible tryptophan residue is also directly adjacent to the disulfide bonded Cys127 (unusual disulfide bond for an IgI-like fold), which stabilizes the tryptophan in its position. Therefore, this tryptophan residue might be involved in HS binding.

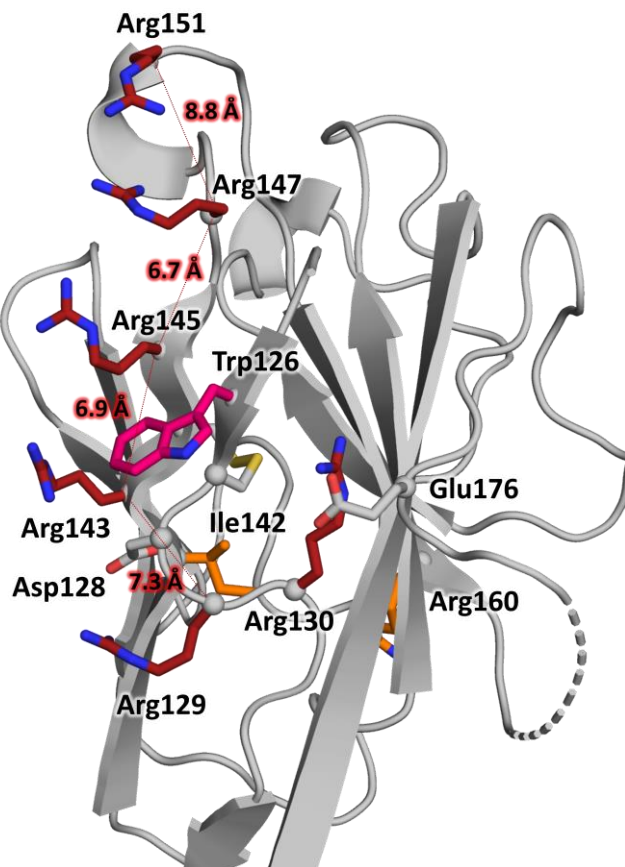


Figure 74: Residues important for GAG binding. Cartoon and stick representation of D1 with residues, which have been reported [180] to be important for GAG binding colored red or orange. Distances between C α s are indicated. Arg130 is in a pocket and might be not directly involved in GAG binding. Arg160 is located in a pocket on the opposite side of the domain and therefore most probably is not involved in direct GAG binding. Ile142 points towards the domain core in a hydrophobic pocket and probably is not involved in GAG binding. Trp126 shown in pink might be part of the GAG binding interface as the side chain point in the solvent region and is in close proximity to the important arginine residues.

6.1.2. Putative HS Binding Area

The minimum length of HS chain necessary for gC binding consists of 10 to 12 monosaccharides [177, 178]. A comparable 10 monosaccharide heparin structure likely (1e00) forms a wave-like straight structure with a chain length of about 39 Å (C1 first residue to hydroxyl at C4 of the tenth residue) (**Figure 75**).

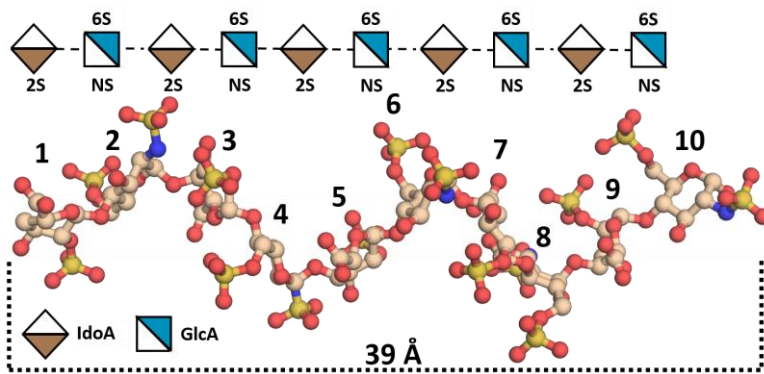


Figure 75: Heparin structure. The decameric heparin (1E00) is represented as sticks and spheres. The end to end distance of the decameric structure is about 39 Å, whereas the glycan adopts a wave-like but overall straight structure.

The distance between the five arginine residues, which were determined to be crucial for HS binding, is about 30 Å, a decameric GAG is about 10 Å longer whereas a linear 12 residue GAG would be even longer. D1 contains several additional arginine residues in near proximity (**Figure 76**). These residues could be involved in HS binding, which would better resemble the reported HS length for binding. On the other hand, the mucin-like domain contains also arginine residues, which might play a role in HS binding.

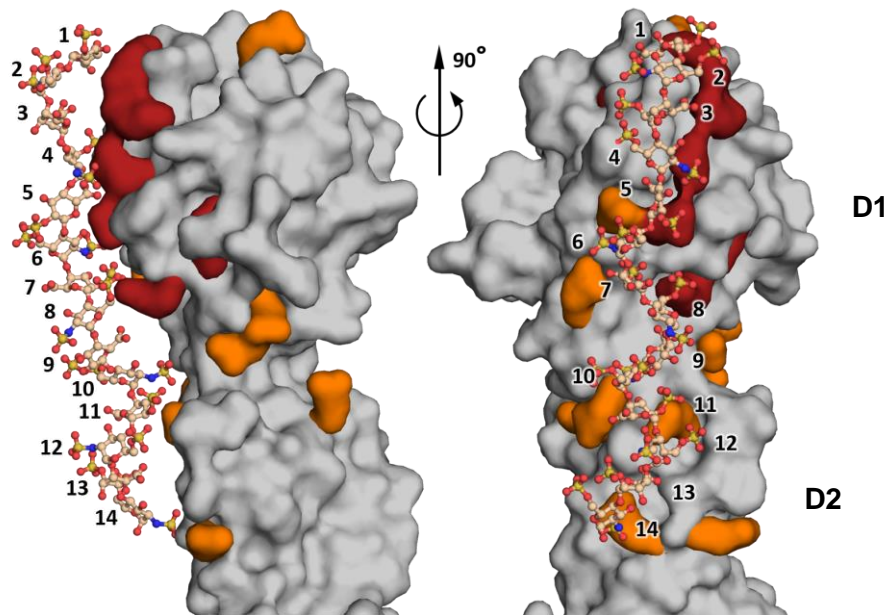


Figure 76: Schematic model of the gC GAG interaction. Surface representation of D1 from gC with the arginine residues colored orange, arginine residues, which are known to be important for HS binding are highlighted in red. A model comprising 14 residues based on a heparin structure (1E0O) is shown as sticks and spheres. The model aims to visualize the surface distribution of arginine residues with the actual size of a comparable GAG and not represent actual HS binding. This model was generated by manual placement of the heparin along the arginine containing surface area after visual inspection.

Along one side of gC, a positive patch is visible in the electrostatic surface potential ranging from the distal end of D1 to the D1-D2 transition (**Figure 77**). This patch might mediate binding of the negatively charged HS, but binding studies are needed to investigate if the arginine residues on the D1-D2 transition are also involved in the HS binding. At some positions of this positive patch are two arginine residues in close contact to each other, which might be responsible for a specificity in preferred binding of distinct sulfated HS (Arg143 and Arg228; Arg 135 and 353; Arg360 and Arg361). These residue might preferably bind di-sulfated or carboxyl- and sulfate group at the same or in directly adjacent monosaccharide subunit. To validate these residues as specificity determining residues, binding studies with mutated gC and differently modified HS are needed.

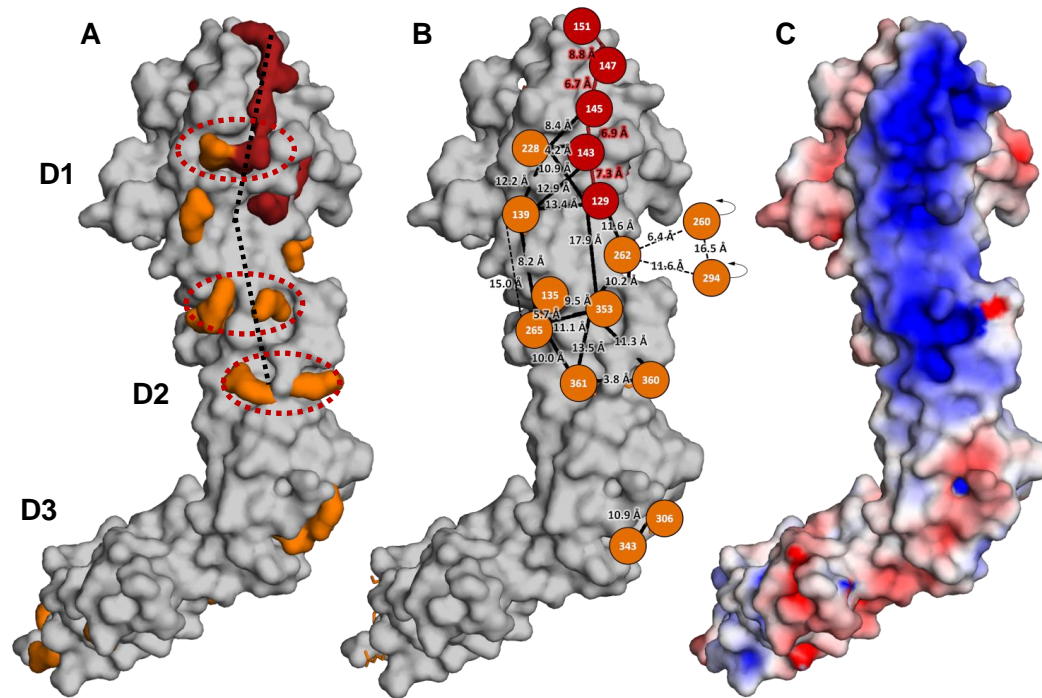


Figure 77: Putative GAG binding site of gC. Different surface representations of gC are shown. Arginine residues are colored in orange, whereas the arginine residues, which have been reported [180] to be essential for GAG binding are highlighted red (A). A possible location of the oligo glycan chain is indicated with black dotted lines, close residue pairs are marked with a red dotted circle. The residue numbers and the distances between C α s are indicated (B). Arg260 and Arg294, although not visible in this view, are indicated with arrows. The electrostatic surface potential of gC is shown as color gradient from +5 to -5 kTe⁻¹ (C). The negative charged GAG probably binds to the large patch with a positive surface potential.

6.1.3. The Dimerisation and HS Binding

The HS binding site is located between the two protomers of the gC dimer (Figure 78). The two protomers interact at the C-terminal part of D1 with an angle between the domains of about 92°, measured between the C α s of Phe146 at the C-terminal and the center between the C α s of Ser138 at the N-terminal ends of the β -strand B (the two close strands). The HS binding site is located in the cleft between the

two protomers, with a large area of positive electrostatic surface potential spreading over both protomers. The increased positive potential area of the dimer might increase the charge mediated attraction of the negatively charged HS. The positive potential area of the dimer might help to pre-orientate the HS in a favored way for binding. The dimerization interface in the D1-D2 transition contains arginine residues, which might allow HS binding within the dimerization interface. This would increase the distance between the protomers in the interface, which could change the overall orientation of these domains. One possibility might be that the cleft closes upon binding of both protomers to the same HS chain. This could be in concert with the opening of the dimer interface. Another possibility could be that the dimer simply binds one or two HS chains at each site next to the dimerization interface.

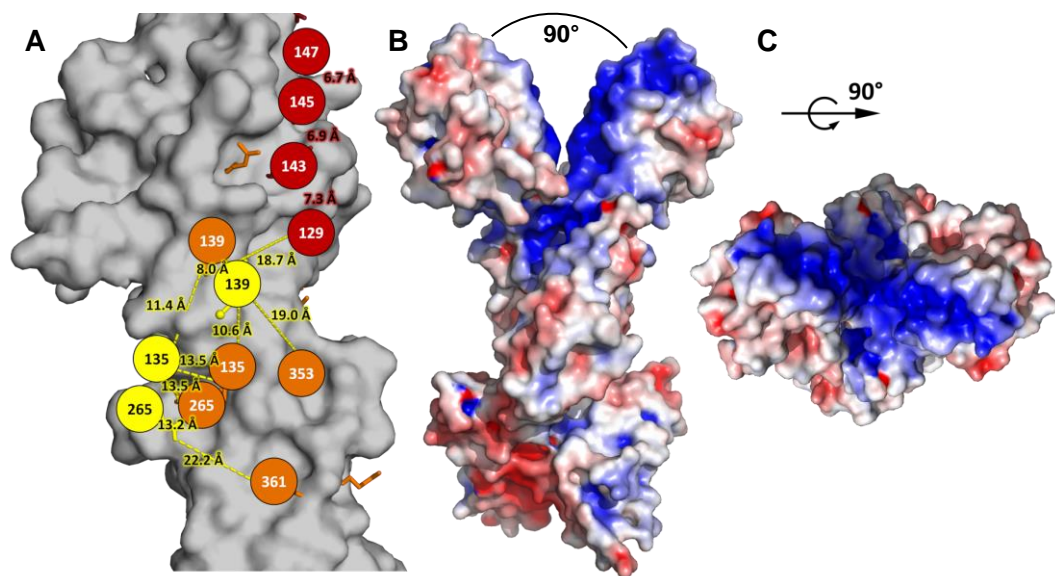


Figure 78: Putative dimer GAG binding site. Surface representation of a monomer with the arginine residues reported to be important for GAG binding indicated in red, others in orange and arginine residues from the second protomer in yellow (A). The electrostatic surface potential of the gC dimer shown as color gradient from +5 to -5 kTe^{-1} (B). 90° rotated view of B as indicated (C).

6.2. Shielding from immune recognition

Glycoprotein C is targeted by the antibody based immune response [202] and it was reported that gC shields gB from the recognition of neutralizing antibodies [187]. The gC dimer has a rod like shape with D3 close to the surface (**Figure 79**). D2 has a glycosylation site opposed to D2 from the second protomer. D1 has two to three glycosylation sites at the tip of D1, most distant to the surface. The high glycosylated mucin like region, not included in the structure, is also at the tip of D1. Overall, this results in a tree-like structure with a glycosylation shield opposed to the cell/virus surface with a distance between the mucin shield and the surface. This mucin-like region sterically blocks antibodies from reaching their epitopes at proteins between the shield and the virus surface. The dimerization of gC should increase the density of the glycan shield above the gC dimer. The putative HS binding site is in the middle between the two protomers. This allows the access of the HS, while the binding site is shielded from the larger antibodies through the glycosylation. As symmetric homodimer, the dimerization itself covers a part of the monomer by the dimerization interface, presenting two times the same surface area. The two long loops, CC'-loop and the C'D-loop of D1 are on the solvent accessible sides of the dimer. Their function is unclear, but they might shield the surface from antibody recognition by allowing immune escape mutations that might not alter the functionality of gC. These loops might be more flexible in solution, as in the crystal structure, where both loops form most of the crystal contacts. Both loops contain glycosylation sites, and if they are flexible in solution, the accessibility of gC by the immune system might be reduced. Another possibility is that those loops interact with a yet unknown interaction partner, such as other glycoproteins e.g. gB.

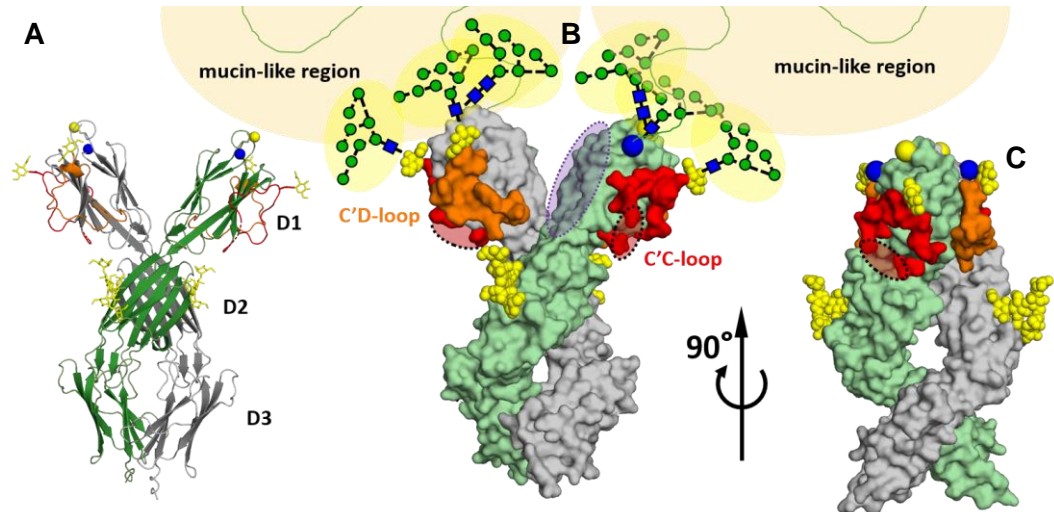


Figure 79: Schematic gC dimer representation. The gC dimer is shown in cartoon or surface representation (**A**; **B**; **C**). The C'C-loop and C'D-loop that covers the surface of D1 are colored red and orange, respectively. The missing residues of the C'C-loop are indicated with a black dotted ellipse. The glycosylation sites covering the tip of D1 are indicated in yellow. The amino-group at the N-terminal residue of the structure is indicated with a blue sphere while the connected mucin-like region is indicated in yellow. The HS-binding site is indicated with a purple dotted ellipse. Rotated view of **B** as indicated (**C**).

6.3. C3b Binding Regions

The gC C3b binding site was reported to consist of four regions (**Figure 80**) [186]. The region 1 (124-137) and the region 4 (223-246) are part of D1, whereas region 2 (276-292) and region 3 (339-366) are part of D2. The regions are on both sides of the domains and it is rather unclear where C3b actually binds. Region 1 and region 3 form a continuous surface contacting the glycosylation site (Asn364), which might be important for C3b binding [186]. The regions partially overlap with the HS binding site, which is consistent with the reported inhibition of C3b binding in presence of HS [203]. C3b contains large areas with negative surface potential, therefore the positive surface potential areas of gC, which are involved in the HS binding, might also be important for the C3b binding (**Figure 77**). Based on the gC structure, mutational studies can be designed targeting residues at the surface to

reliably define the epitope. A C3b-gC complex structure could elucidate the epitope, where D3 might be neglected as it seems to be not involved in direct C3b binding. C3b is a multi-domain protein with a molecular weight of about 180 kDa, which is more than triple the weight of gC (55 kDa). D3 might be needed to allow D1 and D2 to reach the binding epitope of the much larger C3b *in vivo*.

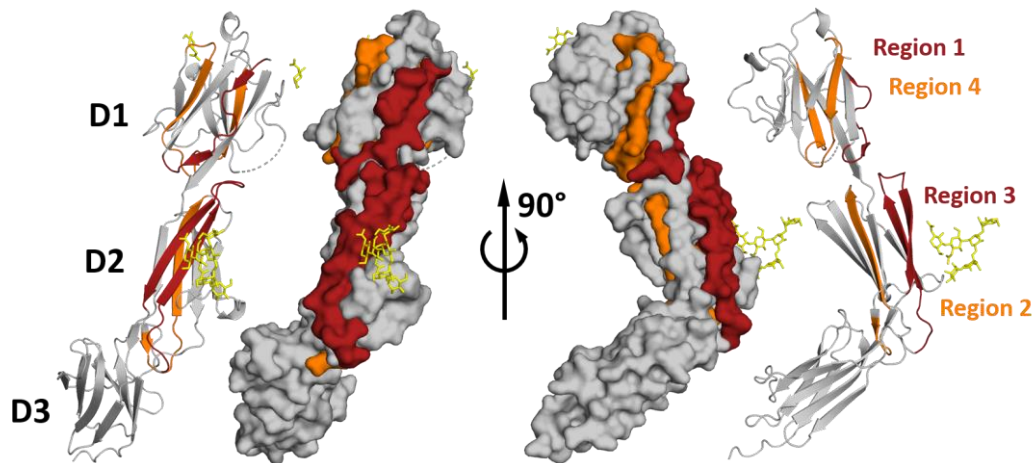


Figure 80: C3b binding epitope mapped on gC. Cartoon and surface representation of gC with the previously reported C3b binding regions 1-4 colored in red or orange. View from two directions rotated as indicated.

6.4. Sequence Comparison with gC from HSV-2

HSV-1 as well as HSV-2 contain gC (gC1 and gC2, respectively) and both bind to HS, but the binding seems to be less important for gC2 and is compensated by gB [204, 205]. The gC2-HS complex was reported to be three magnitudes less stable than the gC1-HS complex [206]. Differences in the affinities to C3b have also been reported, whereas gC2 seems to have a higher affinity for C3b.

The gC1 and gC2 have a protein sequence identity of 73.0% comparing the sequence of the gC1 structure (123-466). The sequence identity is higher for the β -sheets (85.4%) than for the loop regions (62.6%) with no deletions or insertions in the sequences. Therefore, gC2 has most probably a similar domain organization and fold. The dimer interface is conserved and differs only in one residue (321), with a histidine residue and asparagine residue for gC1 and gC2, respectively. Also the D1-D2 transitions differ, where gC1 contains a methionine and gC2 a valine residue at position 263. The D2-D3 transition differs in one residue 277, where gC1 contains a methionine and gC2 a leucine residue. The arginine residues that are important for gC1 HS binding are conserved in gC2 with one exception for position 147 that is replaced by a proline residue in gC2. This proline residue might be primarily responsible for the different HS binding properties of the two HSV strains (**Figure 81**).

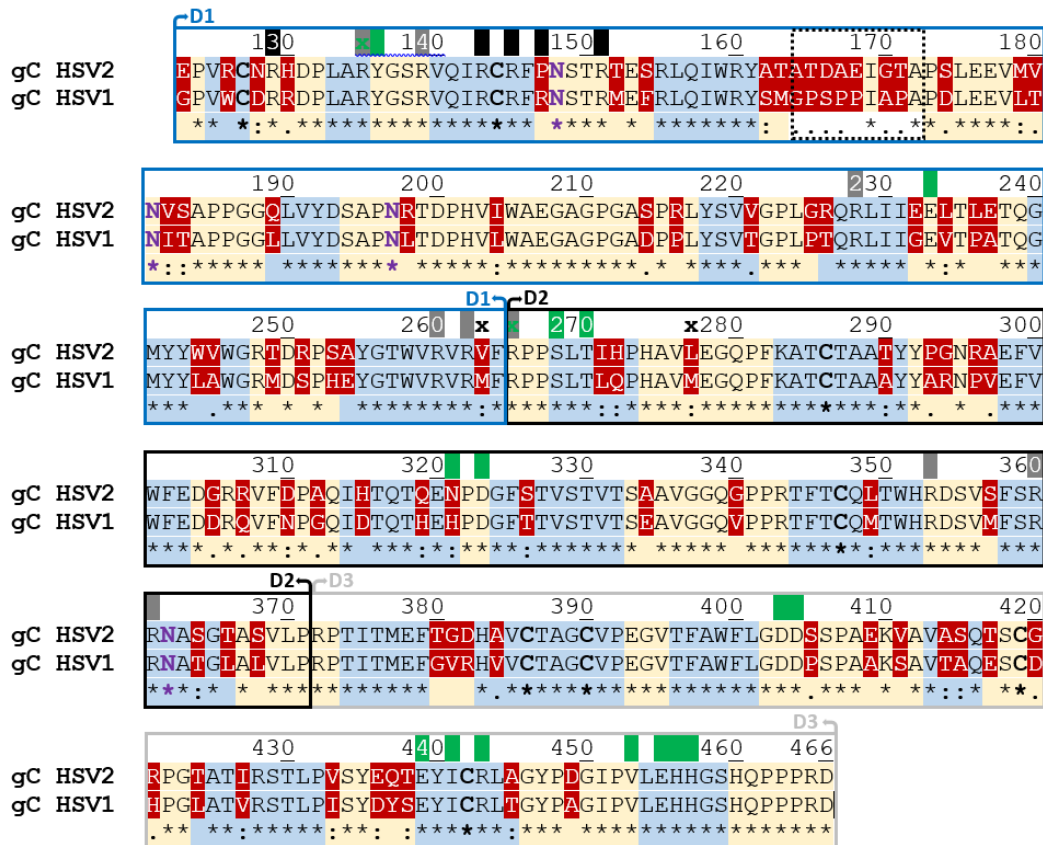


Figure 81: Sequence comparison between HSV-1 and HSV-2. Differences between the sequences are colored red. Linker regions and β -sheets are colored yellow and blue, respectively. Conserved disulfide bonds and glycosylation sites (purple) are shown in bold letters. Residues involved in the domain transition that differ are marked (x). Arginine residues that are involved in HS binding are indicated in black, while residues that might be involve are shown in grey (top row). The loop-region not included in the structure is highlighted with dotted lines. Residues that are involved in the dimer formation are shown in green.

Most sequence differences are distributed over the gC sequence (**Figure 82**) and a HS or a C3b complex-structures would help to determine their contribution in regards to the complex formation. One particular sequence that might be of interest in this regard is the CC'-loop, which contains a ten residues segment with nine sequence differences. A part of this loop was not resolved in the gC1 structure and contains four additional proline residues in gC2.

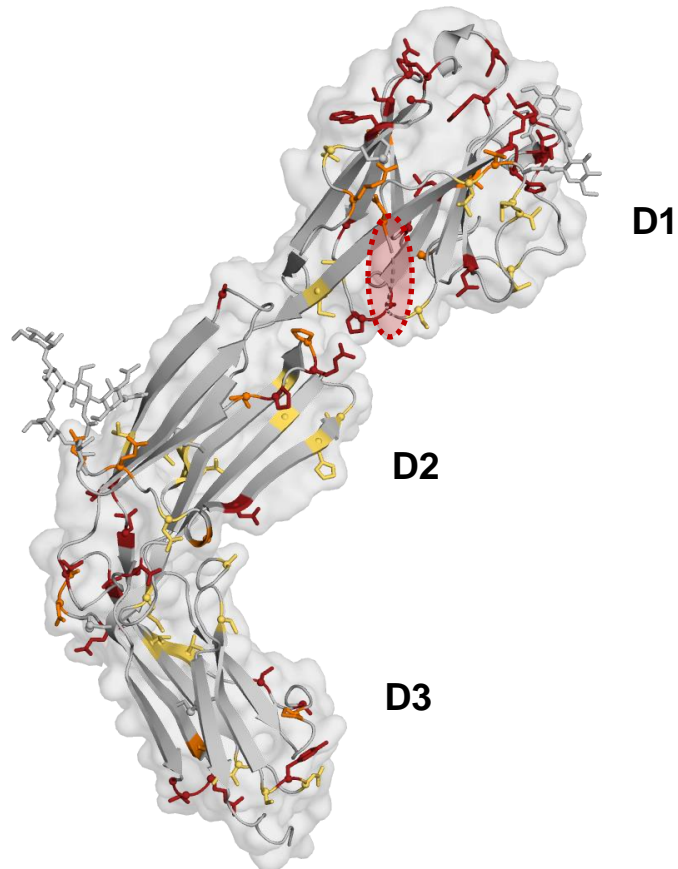


Figure 82: HSV-2 gC sequence differences mapped on the HSV-1 gC structure. The gC monomer is shown in grey surface and cartoon representation. The gC1 residues, which differ in gC2, are shown as sticks, colored according to their similarity, where red represents no similarity, orange low similarity and yellow high similarity.

IV. Peptide Binding Nanobody BC2

1. Introduction

1.1. Antibodies and Antibodies Fragments

Antibodies are important tools in modern biochemistry, due to their specific and strong binding capacity. The commonly used antibodies are soluble immunoglobulins of type G (IgG). These IgGs, secreted from B cells, consist of heavy chains and light chains, whereby antigen binding takes place at the tips of the Y-shaped protein (**Figure 84**). The antigen binding is achieved by six complementarity-determining regions (CDRs), three from each of the two variable domains, heavy and light. Comparing different antibody sequences, the CDRs are highly variable loops, with the capacity to binding a vast variety of antigens. The immunoglobulin variable (IgV) domains as part of the immunoglobulin superfamily (IgSF), consist of a β -sandwich, build by two antiparallel β -sheets, which are stacked onto each other stabilized by a conserved disulfide bond between the two β -sheets (**Figure 83**).

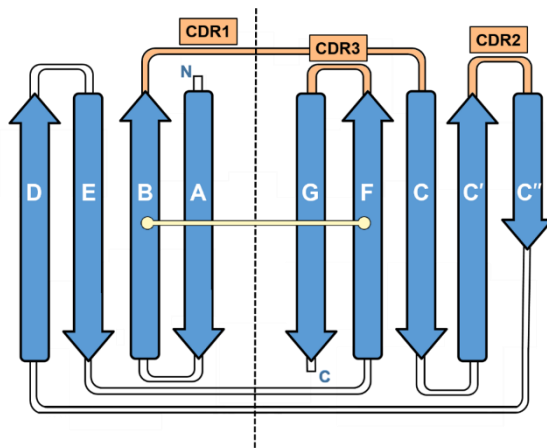


Figure 83: Topologic representation of an IgV domain. The IgV fold consists of two antiparallel β -sheets, which are stacked on each other, here opened up at the dotted line. The framework region is shown in blue, while the three CDRs are shown in orange. The disulfide bond is shown in yellow.

For various applications, the binding is of interest and other aspects of the antibody, such as the capacity to crosslink antigens, or the interaction with the immune system through the constant domains, are not intended. Proteolytic cleavage with papain produces “fragments antigen binding” (Fabs), consisting of the variable and first constant region of the heavy and light chain [207]. Engineered single chain variable fragments (scFv) consist only of the variable regions of the heavy and light chains, where one C-terminus is linked to the N-terminus of the other chain [208]. Beside this conventional antibody fragments [209], there is the possibility to generate from heavy chain antibodies (hclgG) single domain fragments, called nanobodies (Nb), with the capacity of high affinity binding [210](**Figure 84**).

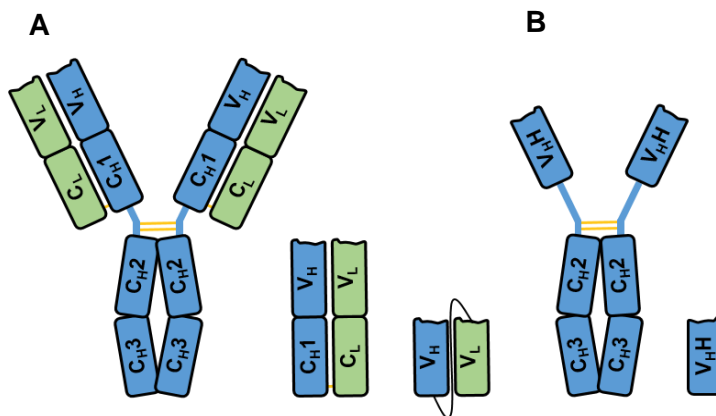


Figure 84: Schematic overview of different antibody types and fragments thereof. C stands for constant, V for variable region, H for heavy chains (blue), and L for light chains (green). Disulfide bonds are indicated in yellow. **A:** IgG antibody, Fab fragment and scFv. **B:** hclgG heavy chain antibody from Camelids and Nb thereof.

1.2. Nanobody Production

As part of the adaptive immune response, Camelids have beside conventional antibodies hclgGs, protecting them from pathogens or foreign substances. Production of Nbs against a specific protein can be induced by immunization of Camelids. The RNA from peripheral blood lymphocytes is then isolated and a cDNA library from the Nb open reading frames is generated. To select and identify binding proteins molecular displaying methods are used. In a phage display the cDNA from these libraries is ligated with phage DNA that, when the gene is expressed, a Nb fused with a phage coat protein is produced. As a result, phages are generated that present a Nb on their surface, while harboring the genetic information of this Nb in the inside. Immobilized target protein is then used to select for binding phages. The DNA of these phage can then be amplified, and sequenced, in order to characterize the presented Nb [211]. With this method, it is possible to do *ex vivo* affinity maturation by using an error prone PCR for the amplification introducing random mutations. The Nbs can then be produced for example in bacteria or cell culture.

1.3. Unique Features of Nanobodies

Nbs compared with conventional antibodies or fragments thereof have two unique characteristics, they are small in size and differ in the binding mode. First containing only one Ig-like domain, their size is between 12 and 15 kDa, which is half the size of the smallest antibody fragment, the scFv. The small size leads to a high tissue permeability and a fast blood clearance. In super-resolution microscopy the size has the advantage of a close spatial localization of the dye, conjugated to the Nb, and the bound antigen. Second, the differences in binding mode is caused by the fact that Nbs have only three CDRs instead of the six CDRs present in conventional antibodies. To achieve comparable affinities, the interactions of the Nbs have to be more efficiently concentrated in the compact paratope [212]. Conventional antibodies have a larger area where their CDR are distributed, due

to their two domain organization, which allows binding of flat, and convex surfaces. Nbs partially compensate the loss of three CDRs through an elongated CDR3 loop, which is often stabilized through a non-canonical disulfide bond. Nbs bind preferred to areas with a roughness, e.g. grooves, cavities and hinge regions, where their CDR3 loop can protrude in concave areas (**Figure 85**). This allows a more rigid binding, with interactions in multiple directions. It can be expected that binding small molecules such as peptides is more problematic for Nbs, as they cannot enclose an antigen between two CDR3 loops, which is the usual binding mode of antibodies for small molecules.

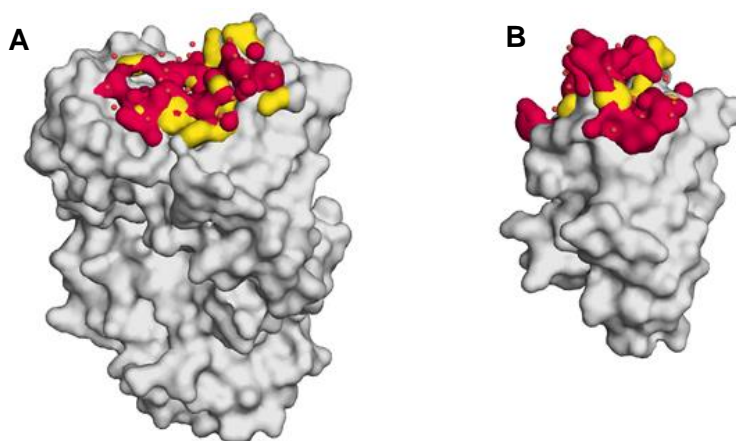


Figure 85: Binding interfaces of an antibody and a Nb, both binding lysozyme. The antibody (**A**)(1P2C) and the Nb (**B**)(1JTT) are shown in surface representation. The binding interface is colored red for distances ≤ 3.5 Å to the antigen, or interface water molecules, which are shown as red spheres, having distances of ≤ 3.5 Å to both, antigen and antibody or Nb. Distances ≤ 4.5 Å to the antigen are colored yellow. The Antibody shows a flat interface, while the Nb binds at a different location of lysozyme, with a concave interface.

1.4. Super Resolution Microscopy as Application Field

In fluorescence based super resolution microscopy, the small size of Nbs can have beneficial effects, resulting in images with high quality. For these techniques, the proteins of interest need to be labeled *in vivo* with fluorophores. As fluorophores, organic dyes are preferable in contrast to protein-based fluorophores such as GFP (green fluorescent protein), as they have a higher quantum yield and photo stability. Two problems arise, however. First, the dye needs to be delivered inside the cell, and second, the labeling needs to be specific. The first problem, transfecting cells can be solved using lipid protein mixtures, which fuse with the cell membrane. Another strategy is electroporation, or peptide carriers such as “trans-activator of transcription” (TAT) from human immunodeficiency virus (HIV) or virus protein 22 (VP22) from herpes simplex virus 1 (HSV-1), which promote the delivery of a fused protein across the cell membrane. For the second problem, there are different approaches. The introduction of an unnatural amino acid via a stop-codon is possible but challenging and time consuming. The fusion of enzymes (SNAP-, CLIP-, TMP-, Halo-tag), which catalyze the coupling of a fluorophore substrate have the disadvantage that the fusion with this enzymes (18-33 kDa) may alter the natural behavior of a protein. Specific antibodies against the protein may also alter its behavior. A compromise could be the usage of a small tag, with a specific antibody, where the changes to the natural proteins are kept minimal and later binding of an antibody does not directly interfere with the protein. Here, the small size of Nbs would be beneficial by limiting the size change and distance between dye and the protein. Also, the one domain organization, and therefore a low inherent flexibility, would lead to a good temporal localization of the dye. There are other labeling strategies (peptide [213], Affibodies [214]) and for the best results, different strategies need to be established and assessed in the individual case. To distinguish different target proteins, the label strategies for the different proteins needs to be orthogonal.

2. Results

2.1. Structural Characterization

Our cooperation partner produced Nbs against β -catenin by immunizing alpaca, followed with several rounds of affinity maturation using phage display [215]. One of these Nb had the property to bind a 12 amino acids (AAs) linear epitope with low nanomolar affinity. Such interactions are rather atypical for Nbs, which usually bind concave, rigid and structured epitopes [216]. To elucidate the binding mode, we solved the crystal structures of this Nb at high resolution, with and without the 12 AAs fragment. We observed that the peptide was elongated, and that binding to the Nb involved mainly backbone-backbone interactions, thereby forming a β -sheet by insertion of the fragment between the CDR3-loop and framework regions of the Nb. Further, the specificity of the Nb-antigen interaction was investigated, by mutating every AA of the peptide to every possible, and quantifying the amount of bound fragments in a competitive assay, using mass spectroscopy. Thereby, we identified that tryptophan at position 10 is indispensable, where at position 6 and 8 small AAs and at position 3 basic AAs are obligated. These findings are consistent with the crystal structure, where tryptophan at position 10 is in a hydrophobic pocket, and residues 6 and 8 are pointing toward the Nb, with limited space available. Residue 3 is involved in a charge-mediated interaction. The other side chains of the peptide are pointing away from the Nb, and contribute to specificity only by excluding a few AAs, mainly proline. At the Nb side, there is a charge-mediated interaction that we termed “headlock interaction” that reaches from CDR3 over the bound fragment to the framework region of the Nb. Comparing peptide binding to the wt Nb and headlock mutants, using SPR, showed that the headlock increases the binding affinity by lowering the off rate 10-fold.

2.2. Application

But what could be done with this nanobody? In our opinion, it has ideal properties needed as protein tag. First, the affinity is high, $K_D = 1.4 \pm 0.1$ nM, second, the binding is specific, third, it is elutable with the soluble peptide fragment. Moreover, the Nb is stable for months at 4°C, the C-terminus or N-terminus of the fragment can be fused to a protein, elution can also be achieved at pH >11, and binding is still possible under harsh conditions (2 %SDS; 4 M Urea or 1.5 M GdmCl). We have investigated these different aspects, and further showed that the Nb can be used to label tagged proteins for confocal microscopy.

Based on this work a commercially available tag system (Spot-Tag®, Spot-Trap®, Spot-Label) was developed (ChromoTek GmbH), which is based on the Nb peptide interaction, with a change of four AAs in the peptide sequence.

3. Developments and Outlook

Our cooperation partner meanwhile showed that the specific labeled bivalent Nb can be used to perform high-quality “direct stochastic optical reconstruction microscopy” (dSTORM) imaging in mammalian and yeast cells [217]. Their bivalent Nb had a lower off-rate than the monovalent Nb, but did not improved the staining specificity. The staining specificity was improved by changing the Nb labeling strategy of the bivalent Nb to be specific. The downside of a bivalent Nb is the capability of crosslinking different tagged proteins, thereby potentially affecting their native behavior.

A different approach to increase the affinity could be a duplication of the tag. A close second sequence, may lead to a rebinding instead of dissociation. The property of this Nb, which does not directly contact all side chains of the peptide, and therefore only requires an AA subset for specificity, could be a chance to design a tag with two overlapping consensus binding sequences. This would be possible for shifts of 1, 3 or 5-11 (**Figure 86**). The color coded is based on the supplementary table 2 from our publication [218]. The lighter the blue color the better the expected binding, red stands for no expected binding with values above 0.4. For a duplicated tag (2x12 AAs) it would be possible to have a third nested tag. In this case, for every additional 6 AAs an additional binding site would be created. If experimental data from such a tag sequences show higher affinity than the original β -catenin sequence that could also reduce problems which may arise from β -catenin binding.

Also noteworthy is that the 12 AA tag has been successfully used as purification tag for nucleolar pre-60S particles by Sanghai et al., finally resulting in a cryo-EM structure [219].

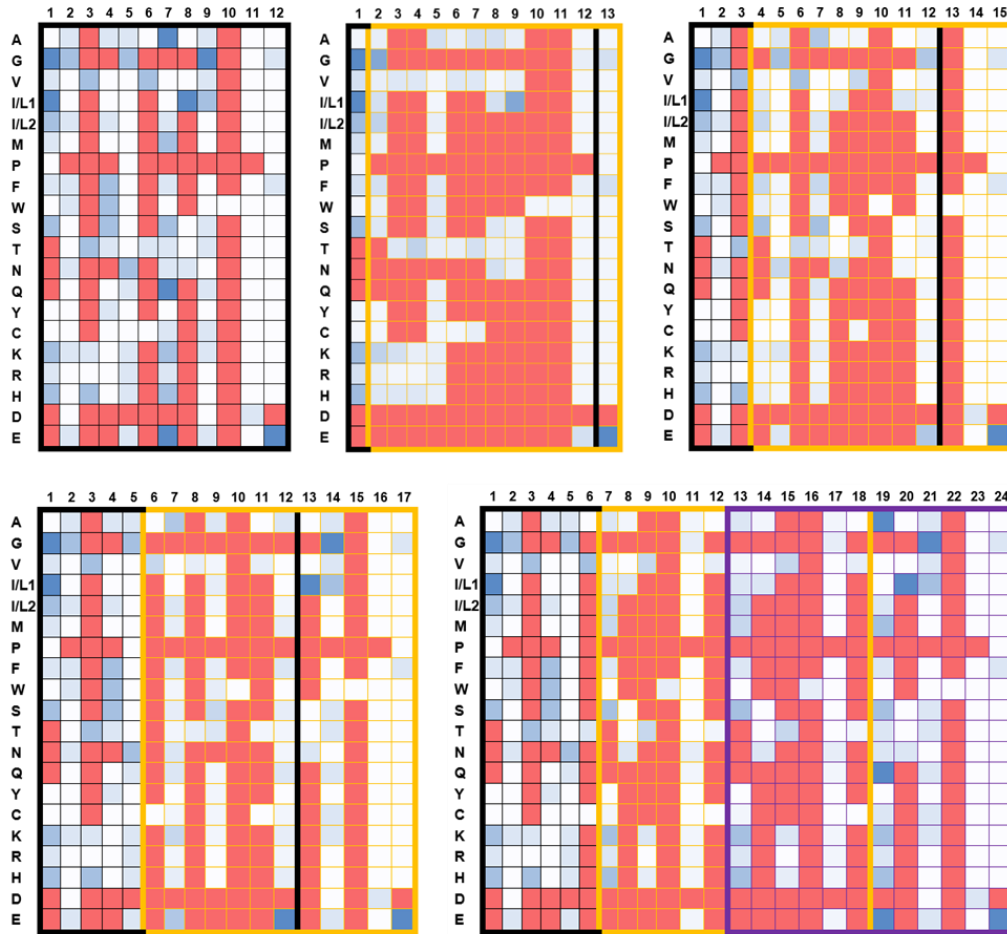


Figure 86: Consensus sequences of the BC2-tags. Top left: color coded figure based on supplementary table 2 from our publication, showing the consensus sequence of the tag. Values above 0.4 are shown in red, expected not to bind, whereas the lighter the blue color, the better the expected binding. In the other tables the overlap of two consensus sequences shifted by 1, 3, 5 and 6 is shown in yellow. For a shift of 6 AAs it would be possible to design a tag containing 1.5 consensus sequences, thereby a binding site would be added for every additional 6 AAs (third tag shown in purple).

V. Adenovirus Serotype 43 Vector

1. Introduction

1.1. Cancer

Cancer is a group of diseases, where cells of an organism divide uncontrolled and invade other tissues. This genetic disease is caused by mutations accumulating in the genome of a cell, or through several rearrangements in a process called chromothripsis [220], resulting in a breakdown of several key cell functions.

There are several mechanisms to keep the genome of cells and the organism stable. If a mutation occurs cellular repair processes correct those “errors”. When repairing does not work the cell will stop to proliferate, or die in a controlled fashion. The repair mechanism occasionally fails, whereby the damage, such as a single strand break, is repaired but with a permanent change in the genome of a cell [221]. Genetic preposition, where repair mechanisms or other effects that contribute to genome stability are impaired, can increase the risk of developing cancer, but anyhow cells with mutated genomes accumulate over time in an organism [222].

Environmental factors such as ionizing radiation, UV-radiation, chemical carcinogens and some fibrous materials can cause mutations. Also, viruses such as human papilloma virus (HPV16; HPV18) [223], hepatitis B/C virus (HBV; HCV) [224], Kaposi's sarcoma-associated herpesvirus (KSHV/HHV8) [225], Epstein-Barr-Virus (EBV) [226], human T-lymphotropic virus 1 (HTLV-1) [227] and Merkel cell polyomavirus (MCPyV) [228] can cause cancer in humans. In addition, some bacteria have been associated with cancer, e.g. *Helicobacter pylori* [229]. Two reactions of the cellular metabolism, oxidative respiration and lipid peroxidation, produce reactive

oxygen species (ROS) that can cause DNA damage. It has been estimated that about 70.000 lesions occur in a human cell per day [230, 231].

1.2. Cancer treatment

Classical treatments of cancer are the surgical excision of the cancer cells, the killing of cancer cells with drugs in chemotherapy or ionizing radiation in radiation therapy. In surgery, the tumor and surrounding healthy tissue is removed in order eliminate all cancerous cells but with the risk of triggering spreading when not achieved. In radiation therapy, the cancer is exposed to ionizing radiation to induce damage to the genome thereby inducing apoptosis. The ionizing radiation generates ROS, which then cause most of the DNA damage [232]. Chemotherapy focuses on rapid dividing cells, which distinguish cancer cell from most other cells. Platinum based drugs for example form adducts with DNA, resulting DNA damage triggers apoptosis [233]. Other strategies are the inhibition of nucleic acid synthesis, topoisomerases, microtubule assembly or disassembly [234-236].

Beside those classical treatments there are many therapeutic strategies focusing on different properties of cancer cells. Epidermal growth factor signaling includes kinase cascades e.g. rat sarcoma (RAS), rapidly accelerated fibrosarcoma (RAF), mitogen activated protein kinase (MAPK), that often show mutations in cancer. Cyclin-dependent kinases or other cell cycle kinases regulate cell division, which are often altered in cancer. Tumour suppressors such as the transcription factor p53, can arrest or senesce the cell cycle, induce apoptosis and inhibit angiogenesis. Viruses such as adenoviruses (Ad) and papillomaviruses inactivate p53 by facilitating its degradation, thereby preventing apoptosis [237, 238]. Vaccines against oncogenic viruses have been developed and approved

(Europe, North and South America, Australia and Japan) against human papilloma virus (HPV) causing cervical cancer [239]. Vaccines against cancer cells comprise the risk of development of an autoimmune disease, and so vaccination has to be directed against antigens that differ between cancer cells and healthy cells. To achieve this, there is the possibility to extract dendritic cells or T-cells from a patient and immunize or genetically modify them *ex vivo*. Those immune cells, presenting either cancer related antigens or recognizing cancer cells, are then reinjected. Gene therapy focuses on the delivery of wild type genes to facilitate the production of wild type protein or editing mutated genes in cancerous cells. Here, viruses can be used as vectors for delivery of DNA. Gendicine, an adenovirus vector containing a recombinant human p53 gene was approved in China (2003) for gene therapy [240]. Some viruses replicate efficient in cancer cells thereby killing them in a lytic process. Oncorine (H101) an oncolytic adenovirus was approved in China (2005), T-VEC (Imlygic), a HSV-1-based oncolytic virus, was approved in the USA and EU (2015), other oncolytic viruses are in clinical studies [241].

2. Results

We contributed to the development of serotype 43 driven Ad-vectors [242]. This serotype belonging to species D has a low seroprevalence, unlike the adenovirus serotype 5 (Ad5) of species C from which the available adenovirus vectors originate. Besides the high seroprevalance [243], Ad5 binds factor X, which was associated with off-target liver transduction after intravascular delivery [244]. In contrast, Ad43 does not bind factor X. Here, the off-target transduction was investigated, CD46 was identified as receptor and chimeric Ad5 fiber knobs with affibodys against human

epidermal growth factor receptor type 2 (HER2) were used to direct the virus against this onco-target. Integrin binding was assumed based on reduced infectivity when Chinese hamster ovary cells (CHO) were pretreated with GRGDSP-peptide or EDTA. This suggests an α_v -integrin as receptor, also described for other Adenoviruses [245]. The function of the receptor for Ad43 as well as the relevant β -subunit of the integrin need to be validated with human integrin in further experiments. In addition, we showed that prior immunization against Ad5 had no effect on the Ad43 based vector.

This new vector could be the basis for the development of a new oncolytic virus-based agent. Ad5 based drugs are injected intra humoral to reduce off target liver transduction. Here, with an Ad43 based vector off target delivery is lower, but as described for other Ad-vectors high levels of virus are found in Kupffer cells [246], a problem which needs to be addressed. To overcome clearance through the immune system several doses of Ad5 have to be administered, although Ad43 has a lower seroprevalence, immunity is gained rapidly. Combination therapy of several viruses, one after another, could increase effectivity thereby minimize virus doses and reduce side effects.

VI. References

1. Rupp, B., *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology*. Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology, 2010: p. 1-809.
2. Bragg, W.H. and W.L. Bragg, *The reflection of X-rays by crystals*. The Royal Society, 1913. **88**(605).
3. Owens, G.E., et al., *Comparative analysis of anti-polyglutamine Fab crystals grown on Earth and in microgravity*. Acta Crystallogr F Struct Biol Commun, 2016. **72**(Pt 10): p. 762-771.
4. McPherson, A. and L.J. DeLucas, *Microgravity protein crystallization*. NPJ Microgravity, 2015. **1**: p. 15010.
5. Chruszcz, M., et al., *Analysis of solvent content and oligomeric states in protein crystals - does symmetry matter?* Protein Science, 2008. **17**(4): p. 623-632.
6. Broennimann, C., et al., *The PILATUS 1M detector*. J Synchrotron Radiat, 2006. **13**(Pt 2): p. 120-30.
7. Forster, A., S. Brandstetter, and C. Schulze-Briese, *Transforming X-ray detection with hybrid photon counting detectors*. Philos Trans A Math Phys Eng Sci, 2019. **377**(2147): p. 20180241.
8. Panjikar, S., et al., *A step towards long-wavelength protein crystallography: subjecting protein crystals to a vacuum*. J Appl Crystallogr, 2015. **48**(Pt 3): p. 913-916.
9. Wilson, A.J.C., *Determination of absolute from relative x-ray intensity data*. Nature, 1942. **150**(151-2).
10. Diederichs, K., *Quantifying instrument errors in macromolecular X-ray data sets*. Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 6): p. 733-40.
11. Karplus, P.A. and K. Diederichs, *Linking crystallographic model and data quality*. Science, 2012. **336**(6084): p. 1030-3.
12. Read, R.J., P.D. Adams, and A.J. McCoy, *Intensity statistics in the presence of translational noncrystallographic symmetry*. Acta Crystallographica Section D-Biological Crystallography, 2013. **69**: p. 176-183.
13. Nave, C., *Radiation-Damage in Protein Crystallography*. Radiation Physics and Chemistry, 1995. **45**(3): p. 483-490.
14. Hubbell, J.H., et al., *Atomic form factors, incoherent scattering functions, and photon scattering cross sections*. Journal of Physical and Chemical Reference Data, 1975. **4**(471).
15. O'Neill, P., D.L. Stevens, and E.F. Garman, *Physical and chemical considerations of damage induced in protein crystals by synchrotron radiation: a radiation chemical perspective*. J Synchrotron Radiat, 2002. **9**(Pt 6): p. 329-32.
16. Close, D.M. and W.A. Bernhard, *Comprehensive model for X-ray-induced damage in protein crystallography*. J Synchrotron Radiat, 2019. **26**(Pt 4): p. 945-957.

17. Kmetko, J., et al., *Can radiation damage to protein crystals be reduced using small-molecule compounds?* Acta Crystallographica Section D-Biological Crystallography, 2011. **67**: p. 881-893.
18. Zhou, Q., et al., *Architecture of the synaptotagmin-SNARE machinery for neuronal exocytosis.* Nature, 2015. **525**(7567): p. 62-7.
19. Morris, R.J. and G. Bricogne, *Sheldrick's 1.2 angstrom rule and beyond.* Acta Crystallographica Section D-Structural Biology, 2003. **59**: p. 615-617.
20. Hauptman, H.A., *Shake-and-bake: An algorithm for automatic solution ab initio of crystal structures.* Macromolecular Crystallography, Pt B, 1997. **277**: p. 3-13.
21. Weis, W.I., et al., *Structure of the calcium-dependent lectin domain from a rat mannose-binding protein determined by MAD phasing.* Science, 1991. **254**(5038): p. 1608-15.
22. Ravelli, R.B.G., et al., *Specific radiation damage can be used to solve macromolecular crystal structures.* Structure, 2003. **11**(2): p. 217-224.
23. Nanao, M.H. and R.B. Ravelli, *Phasing macromolecular structures with UV-induced structural changes.* Structure, 2006. **14**(4): p. 791-800.
24. Wang, B.C., *Resolution of Phase Ambiguity in Macromolecular Crystallography.* Methods in Enzymology, 1985. **115**: p. 90-112.
25. Terwilliger, T.C., *Reciprocal-space solvent flattening.* Acta Crystallographica Section D-Biological Crystallography, 1999. **55**: p. 1863-1871.
26. Matthews, B.W., *Solvent content of protein crystals.* J Mol Biol, 1968. **33**(2): p. 491-7.
27. Abrahams, J.P. and A.G. Leslie, *Methods used in the structure determination of bovine mitochondrial F1 ATPase.* Acta Crystallogr D Biol Crystallogr, 1996. **52**(Pt 1): p. 30-42.
28. Reddy, V.S., *Application of the phase extension method in virus crystallography.* Crystallogr Rev, 2016. **22**(2): p. 128-140.
29. Brunger, A.T., *Free R-Value - a Novel Statistical Quantity for Assessing the Accuracy of Crystal-Structures.* Nature, 1992. **355**(6359): p. 472-475.
30. Engh, R.A. and R. Huber, *Accurate Bond and Angle Parameters for X-Ray Protein-Structure Refinement.* Acta Crystallographica Section A, 1991. **47**: p. 392-400.
31. Derrien, M. and P. Veiga, *Rethinking Diet to Aid Human-Microbe Symbiosis.* Trends Microbiol, 2017. **25**(2): p. 100-112.
32. LeBlanc, J.G., et al., *Bacteria as vitamin suppliers to their host: a gut microbiota perspective.* Current Opinion in Biotechnology, 2013. **24**(2): p. 160-168.
33. Garcia-Gutierrez, E., et al., *Gut microbiota as a source of novel antimicrobials.* Gut Microbes, 2019. **10**(1): p. 1-21.
34. Muller, C.A., I.B. Autenrieth, and A. Peschel, *Innate defenses of the intestinal epithelial barrier.* Cell Mol Life Sci, 2005. **62**(12): p. 1297-307.
35. Pelaseyed, T., et al., *The mucus and mucins of the goblet cells and enterocytes provide the first defense line of the gastrointestinal tract and interact with the immune system.* Immunol Rev, 2014. **260**(1): p. 8-20.
36. Kim, Y.S. and S.B. Ho, *Intestinal goblet cells and mucins in health and disease: recent insights and progress.* Curr Gastroenterol Rep, 2010. **12**(5): p. 319-30.

37. Peloquin, J.M. and D.D. Nguyen, *The microbiota and inflammatory bowel disease: Insights from animal models*. Anaerobe, 2013. **24**: p. 102-106.
38. Liu, T.C. and T.S. Stappenbeck, *Genetics and Pathogenesis of Inflammatory Bowel Disease*. Annual Review of Pathology: Mechanisms of Disease, Vol 11, 2016. **11**: p. 127-148.
39. Jostins, L., et al., *Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease*. Nature, 2012. **491**(7422): p. 119-124.
40. Ho, S.M., et al., *Challenges in IBD Research: Environmental Triggers*. Inflammatory Bowel Diseases, 2019. **25**: p. S13-S23.
41. Ghouri, Y.A., V. Tahan, and B. Shen, *Secondary causes of inflammatory bowel diseases*. World Journal of Gastroenterology, 2020. **26**(28): p. 3998-4017.
42. Knowles, S.R., et al., *Quality of Life in Inflammatory Bowel Disease: A Systematic Review and Meta-analyses-Part I*. Inflamm Bowel Dis, 2018. **24**(4): p. 742-751.
43. Yu, Y.Y.R. and J.R. Rodriguez, *Clinical presentation of Crohn's, ulcerative colitis, and indeterminate colitis: Symptoms, extraintestinal manifestations, and disease phenotypes*. Seminars in Pediatric Surgery, 2017. **26**(6): p. 349-355.
44. Card, T., R. Hubbard, and R.F.A. Logan, *Mortality in inflammatory bowel disease: A population-based cohort study*. Gastroenterology, 2003. **125**(6): p. 1583-1590.
45. Keller, D.S., et al., *Colorectal cancer in inflammatory bowel disease: review of the evidence*. Techniques in Coloproctology, 2019. **23**(1): p. 3-13.
46. Spiller, R. and G. Major, *IBS and IBD - separate entities or on a spectrum?* Nature Reviews Gastroenterology & Hepatology, 2016. **13**(10): p. 613-621.
47. Xavier, R.J. and D.K. Podolsky, *Unravelling the pathogenesis of inflammatory bowel disease*. Nature, 2007. **448**(7152): p. 427-434.
48. Mizoguchi, A., et al., *Genetically engineered mouse models for studying inflammatory bowel disease*. Journal of Pathology, 2016. **238**(2): p. 205-219.
49. Britton, G.J., et al., *Microbiotas from Humans with Inflammatory Bowel Disease Alter the Balance of Gut Th17 and RORgammat(+) Regulatory T Cells and Exacerbate Colitis in Mice*. Immunity, 2019. **50**(1): p. 212-224 e4.
50. Okayasu, I., et al., *A Novel Method in the Induction of Reliable Experimental Acute and Chronic Ulcerative-Colitis in Mice*. Gastroenterology, 1990. **98**(3): p. 694-702.
51. Eichele, D.D. and K.K. Kharbanda, *Dextran sodium sulfate colitis murine model: An indispensable tool for advancing our understanding of inflammatory bowel diseases pathogenesis*. World Journal of Gastroenterology, 2017. **23**(33): p. 6016-6029.
52. Iacucci, M., S. de Silva, and S. Ghosh, *Mesalazine in inflammatory bowel disease: A trendy topic once again?* Canadian Journal of Gastroenterology and Hepatology, 2010. **24**(2): p. 127-133.
53. Peyrin-Biroulet, L., et al., *Efficacy and safety of tumor necrosis factor antagonists in Crohn's disease: Meta-analysis of placebo-controlled trials*. Clinical Gastroenterology and Hepatology, 2008. **6**(6): p. 644-653.
54. Simon, E.G., et al., *Ustekinumab for the treatment of Crohn's disease: can it find its niche?* Therapeutic Advances in Gastroenterology, 2016. **9**(1): p. 26-36.

55. Bernstein, C.N., *Treatment of IBD: Where We Are and Where We Are Going*. American Journal of Gastroenterology, 2015. **110**(1): p. 114-126.
56. Jonkers, D., et al., *Probiotics in the Management of Inflammatory Bowel Disease A Systematic Review of Intervention Studies in Adult Patients*. Drugs, 2012. **72**(6): p. 803-823.
57. Schultz, M., et al., *Preventive effects of Escherichia coli strain Nissle 1917 on acute and chronic intestinal inflammation in two different murine models of colitis*. Clinical and Diagnostic Laboratory Immunology, 2004. **11**(2): p. 372-378.
58. Wang, Y.R., et al., *The administration of Escherichia coli Nissle 1917 ameliorates irinotecan-induced intestinal barrier dysfunction and gut microbial dysbiosis in mice*. Life Sciences, 2019. **231**.
59. Rodriguez-Nogales, A., et al., *The Administration of Escherichia coli Nissle 1917 Ameliorates Development of DSS-Induced Colitis in Mice*. Frontiers in Pharmacology, 2018. **9**.
60. Steimle, A., et al., *Flagellin hypervariable region determines symbiotic properties of commensal Escherichia coli strains*. PLoS Biol, 2019. **17**(6): p. e3000334.
61. Kay, C., et al., *Molecular mechanisms activating the NAIP-NLRC4 inflammasome: Implications in infectious disease, autoinflammation, and cancer*. Immunological Reviews, 2020. **297**(1): p. 67-82.
62. Lage, S.L., et al., *Emerging concepts about NAIP/NLRC4 inflammasomes*. Frontiers in Immunology, 2014. **5**.
63. Hayashi, F., et al., *The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5*. Nature, 2001. **410**(6832): p. 1099-1103.
64. Chaban, B., H.V. Hughes, and M. Beeby, *The flagellum in bacterial pathogens: For motility and a whole lot more*. Seminars in Cell & Developmental Biology, 2015. **46**: p. 91-103.
65. Vijay-Kumar, M., et al., *Deletion of TLR5 results in spontaneous colitis in mice*. Journal of Clinical Investigation, 2007. **117**(12): p. 3909-3921.
66. O'Neill, L.A. and A.G. Bowie, *The family of five: TIR-domain-containing adaptors in Toll-like receptor signalling*. Nat Rev Immunol, 2007. **7**(5): p. 353-64.
67. Yoon, S.I., et al., *Structural basis of TLR5-flagellin recognition and signaling*. Science, 2012. **335**(6070): p. 859-64.
68. Forstneric, V., et al., *The role of the C-terminal D0 domain of flagellin in activation of Toll like receptor 5*. PLoS Pathog, 2017. **13**(8): p. e1006574.
69. Turner, L., A.S. Stern, and H.C. Berg, *Growth of flagellar filaments of Escherichia coli is independent of filament length*. J Bacteriol, 2012. **194**(10): p. 2437-42.
70. Krissinel, E. and K. Henrick, *Inference of macromolecular assemblies from crystalline state*. Journal of Molecular Biology, 2007. **372**(3): p. 774-797.
71. Gay, N.J., M. Gangloff, and L.A. O'Neill, *What the Myddosome structure tells us about the initiation of innate immunity*. Trends Immunol, 2011. **32**(3): p. 104-9.
72. Motshwene, P.G., et al., *An oligomeric signaling platform formed by the Toll-like receptor signal transducers MyD88 and IRAK-4*. J Biol Chem, 2009. **284**(37): p. 25404-11.

73. Yin, Q., et al., *E2 interaction and dimerization in the crystal structure of TRAF6*. *Nat Struct Mol Biol*, 2009. **16**(6): p. 658-66.
74. Cohen, P. and S. Strickson, *The role of hybrid ubiquitin chains in the MyD88 and other innate immune signalling pathways*. *Cell Death Differ*, 2017. **24**(7): p. 1153-1159.
75. Mihaly, S.R., J. Ninomiya-Tsuji, and S. Morioka, *TAK1 control of cell death*. *Cell Death Differ*, 2014. **21**(11): p. 1667-76.
76. Conze, D.B., et al., *Lys63-linked polyubiquitination of IRAK-1 is required for interleukin-1 receptor- and toll-like receptor-mediated NF-kappaB activation*. *Mol Cell Biol*, 2008. **28**(10): p. 3538-47.
77. Jain, A., S. Kaczanowska, and E. Davila, *IL-1 Receptor-Associated Kinase Signaling and Its Role in Inflammation, Cancer Progression, and Therapy Resistance*. *Front Immunol*, 2014. **5**: p. 553.
78. Israel, A., *The IKK complex, a central regulator of NF-kappaB activation*. *Cold Spring Harb Perspect Biol*, 2010. **2**(3): p. a000158.
79. Polley, S., et al., *Structural Basis for the Activation of IKK1/alpha*. *Cell Reports*, 2016. **17**(8): p. 1907-1914.
80. Price, A.E., et al., *A Map of Toll-like Receptor Expression in the Intestinal Epithelium Reveals Distinct Spatial, Cell Type-Specific, and Temporal Patterns*. *Immunity*, 2018. **49**(3): p. 560-575 e6.
81. Gewirtz, A.T., et al., *Cutting edge: Bacterial flagellin activates basolaterally expressed TLR5 to induce epithelial proinflammatory gene expression*. *Journal of Immunology*, 2001. **167**(4): p. 1882-1885.
82. Rhee, S.H., et al., *Pathophysiological role of Toll-like receptor 5 engagement by bacterial flagellin in colonic inflammation*. *Proceedings of the National Academy of Sciences of the United States of America*, 2005. **102**(38): p. 13610-13615.
83. Ha, E.M., et al., *A direct role for dual oxidase in Drosophila gut immunity*. *Science*, 2005. **310**(5749): p. 847-50.
84. Ogier-Denis, E., S.B. Mkaddem, and A. Vandewalle, *NOX enzymes and Toll-like receptor signaling*. *Semin Immunopathol*, 2008. **30**(3): p. 291-300.
85. Kolls, J.K., P.B. McCray, Jr., and Y.R. Chan, *Cytokine-mediated regulation of antimicrobial proteins*. *Nat Rev Immunol*, 2008. **8**(11): p. 829-35.
86. Cao, A.T., et al., *Th17 cells upregulate polymeric Ig receptor and intestinal IgA and contribute to intestinal homeostasis*. *J Immunol*, 2012. **189**(9): p. 4666-73.
87. Annunziato, F., et al., *Phenotypic and functional features of human Th17 cells*. *J Exp Med*, 2007. **204**(8): p. 1849-61.
88. Crellin, N.K., et al., *Human CD4(+) T cells express TLR5 and its ligand flagellin enhances the suppressive capacity and expression of FOXP3 in CD4(+)CD25(+) T regulatory cells*. *Journal of Immunology*, 2005. **175**(12): p. 8051-8059.
89. Liu, H., et al., *TLR5 mediates CD172 alpha(+) intestinal lamina propria dendritic cell induction of Th17 cells*. *Scientific Reports*, 2016. **6**.
90. Frick, J.S., F. Grunebach, and I.B. Autenrieth, *Immunomodulation by semi-mature dendritic cells: A novel role of Toll-like receptors and interleukin-6*. *International Journal of Medical Microbiology*, 2010. **300**(1): p. 19-24.

91. Maier, B. and G.C.L. Wong, *How Bacteria Use Type IV Pili Machinery on Surfaces*. Trends Microbiol, 2015. **23**(12): p. 775-788.
92. Wolgemuth, C., et al., *How myxobacteria glide*. Curr Biol, 2002. **12**(5): p. 369-77.
93. Nan, B. and D.R. Zusman, *Novel mechanisms power bacterial gliding motility*. Mol Microbiol, 2016. **101**(2): p. 186-93.
94. Lauga, E., *Bacterial Hydrodynamics*. Annual Review of Fluid Mechanics, Vol 48, 2016. **48**: p. 105-130.
95. Sourjik, V., *Receptor clustering and signal processing in E. coli chemotaxis*. Trends Microbiol, 2004. **12**(12): p. 569-76.
96. Wadhams, G.H. and J.P. Armitage, *Making sense of it all: bacterial chemotaxis*. Nat Rev Mol Cell Biol, 2004. **5**(12): p. 1024-37.
97. Arkhipov, A., et al., *Coarse-grained molecular dynamics simulations of a rotating bacterial flagellum*. Biophys J, 2006. **91**(12): p. 4589-97.
98. Postel, S., et al., *Bacterial flagellar capping proteins adopt diverse oligomeric states*. Elife, 2016. **5**.
99. Al-Otaibi, N.S., et al., *The cryo-EM structure of the bacterial flagellum cap complex suggests a molecular mechanism for filament elongation*. Nat Commun, 2020. **11**(1): p. 3210.
100. Evans, L.D., C. Hughes, and G.M. Fraser, *Building a flagellum outside the bacterial cell*. Trends Microbiol, 2014. **22**(10): p. 566-72.
101. Vonderviszt, F. and K. Namba, *Function and Assembly of Flagellar Axial Proteins*. Austin (TX) Landes Bioscience, 2000-2013. **Madame Curie Bioscience Database**.
102. Asakura, S. and T. Iino, *Polymorphism of Salmonella Flagella as Investigated by Means of in-Vitro Copolymerization of Flagellins Derived from Various Strains*. Journal of Molecular Biology, 1972. **64**(1): p. 251-&.
103. Stadler, A.M., et al., *Correlation between Supercoiling and Conformational Motions of the Bacterial Flagellar Filament*. Biophysical Journal, 2013. **105**(9): p. 2157-2165.
104. Hasegawa, K., I. Yamashita, and K. Namba, *Quasi- and nonequivalence in the structure of bacterial flagellar filament*. Biophysical Journal, 1998. **74**(1): p. 569-575.
105. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors*. Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.
106. Gasteiger, E., et al., *Protein Identification and Analysis Tools on the ExpASY Server*. John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press, 2005: p. 571-607
107. Sugahara, M., et al., *Heavy-atom Database System: a tool for the preparation of heavy-atom derivatives of protein crystals based on amino-acid sequence and crystallization conditions*. Acta Crystallogr D Biol Crystallogr, 2005. **61**(Pt 9): p. 1302-5.
108. Engilberge, S., et al., *Crystallophore: a versatile lanthanide complex for protein crystallography combining nucleating effects, phasing properties, and luminescence*. Chem Sci, 2017. **8**(9): p. 5909-5917.
109. Kabsch, W., *Xds*. Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 2): p. 125-32.

110. Adams, P.D., et al., *PHENIX: a comprehensive Python-based system for macromolecular structure solution*. Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 2): p. 213-21.
111. Zwart, P.H., R.W. Grosse-Kunstleve, and P.D. Adams, *Xtriage and Fest: automatic assessment of X-ray data and substructure structure factor estimation*. CCP4 Newsletter Winter, 2005. **Contribution 7**.
112. Afonine, P.V., et al., *Towards automated crystallographic structure refinement with phenix.refine*. Acta Crystallogr D Biol Crystallogr, 2012. **68**(Pt 4): p. 352-67.
113. Winn, M.D., et al., *Overview of the CCP4 suite and current developments*. Acta Crystallogr D Biol Crystallogr, 2011. **67**(Pt 4): p. 235-42.
114. Evans, P.R., *An introduction to data reduction: space-group determination, scaling and intensity statistics*. Acta Crystallogr D Biol Crystallogr, 2011. **67**(Pt 4): p. 282-92.
115. Stein, N., *CHAINSAW: a program for mutating pdb files used as templates in molecular replacement*. Journal of Applied Crystallography, 2008. **41**: p. 641-643.
116. McCoy, A.J., et al., *Phaser crystallographic software*. J Appl Crystallogr, 2007. **40**(Pt 4): p. 658-674.
117. Vagin, A. and A. Teplyakov, *Molecular replacement with MOLREP*. Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 1): p. 22-5.
118. Murshudov, G.N., A.A. Vagin, and E.J. Dodson, *Refinement of macromolecular structures by the maximum-likelihood method*. Acta Crystallogr D Biol Crystallogr, 1997. **53**(Pt 3): p. 240-55.
119. Emsley, P., et al., *Features and development of Coot*. Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 4): p. 486-501.
120. Schrodinger, LLC, *The PyMOL Molecular Graphics System, Version 1.8*. 2015.
121. Sheldrick, G.M., *Experimental phasing with SHELXC/D/E: combining chain tracing with density modification*. Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 4): p. 479-85.
122. Vonrhein, C., et al., *Automated structure solution with autoSHARP*. Methods Mol Biol, 2007. **364**: p. 215-30.
123. Chen, V.B., et al., *MolProbity: all-atom structure validation for macromolecular crystallography*. Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 1): p. 12-21.
124. Vonderviszt, F., et al., *Terminal regions of flagellin are disordered in solution*. J Mol Biol, 1989. **209**(1): p. 127-33.
125. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput*. Nucleic Acids Res, 2004. **32**(5): p. 1792-7.
126. Letunic, I. and P. Bork, *Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees*. Nucleic Acids Res, 2016. **44**(W1): p. W242-5.
127. Joensen, K.G., et al., *Rapid and Easy In Silico Serotyping of Escherichia coli Isolates by Use of Whole-Genome Sequencing Data*. J Clin Microbiol, 2015. **53**(8): p. 2410-26.

128. Schoenhals, G. and C. Whitfield, *Comparative analysis of flagellin sequences from Escherichia coli strains possessing serologically distinct flagellar filaments with a shared complex surface pattern*. J Bacteriol, 1993. **175**(17): p. 5395-402.
129. Orskov, I., et al., *Serology, chemistry, and genetics of O and K antigens of Escherichia coli*. Bacteriol Rev, 1977. **41**(3): p. 667-710.
130. Banjo, M., et al., *Escherichia coli H-Genotyping PCR: a Complete and Practical Platform for Molecular H Typing*. J Clin Microbiol, 2018. **56**(6).
131. Fratamico, P.M., et al., *Advances in Molecular Serotyping and Subtyping of Escherichia coli*. Front Microbiol, 2016. **7**: p. 644.
132. Troge, A., et al., *More than a marine propeller--the flagellum of the probiotic Escherichia coli strain Nissle 1917 is the major adhesin mediating binding to human mucus*. Int J Med Microbiol, 2012. **302**(7-8): p. 304-14.
133. Yang, Y., et al., *The flagellin hypervariable region is a potential flagella display domain in probiotic Escherichia coli strain Nissle 1917*. Arch Microbiol, 2016. **198**(7): p. 603-10.
134. Wang, F., et al., *A structural model of flagellar filament switching across multiple bacterial species*. Nat Commun, 2017. **8**(1): p. 960.
135. Patteson, A.E., et al., *Running and tumbling with E. coli in polymeric solutions*. Sci Rep, 2015. **5**: p. 15761.
136. Duarte, J.M., et al., *Protein interface classification by evolutionary analysis*. BMC Bioinformatics, 2012. **13**: p. 334.
137. Fatahzadeh, M. and R.A. Schwartz, *Human herpes simplex virus infections: epidemiology, pathogenesis, symptomatology, diagnosis, and management*. J Am Acad Dermatol, 2007. **57**(5): p. 737-63; quiz 764-6.
138. James, C., et al., *Herpes simplex virus: global infection prevalence and incidence estimates, 2016*. Bull World Health Organ, 2020. **98**(5): p. 315-329.
139. Whitley, R., D.W. Kimberlin, and C.G. Prober, *Pathogenesis and disease*, in *Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis*, A. Arvin, et al., Editors. 2007: Cambridge.
140. Corey, L., et al., *The effects of herpes simplex virus-2 on HIV-1 acquisition and transmission: a review of two overlapping epidemics*. J Acquir Immune Defic Syndr, 2004. **35**(5): p. 435-45.
141. Tobian, A.A. and T.C. Quinn, *Herpes simplex virus type 2 and syphilis infections with HIV: an evolving synergy in transmission and prevention*. Curr Opin HIV AIDS, 2009. **4**(4): p. 294-9.
142. Rand, K.H., et al., *Cellular immunity and herpesvirus infections in cardiac-transplant patients*. N Engl J Med, 1977. **296**(24): p. 1372-7.
143. Whitley, R.J., et al., *Infections caused by herpes simplex virus in the immunocompromised host: natural history and topical acyclovir therapy*. J Infect Dis, 1984. **150**(3): p. 323-9.
144. Gnann, J.W., Jr. and R.J. Whitley, *Herpes Simplex Encephalitis: an Update*. Curr Infect Dis Rep, 2017. **19**(3): p. 13.
145. Whitley, R.J. and B. Roizman, *Herpes simplex virus infections*. Lancet, 2001. **357**(9267): p. 1513-8.

146. Knipe, D.M. and A. Cliffe, *Chromatin control of herpes simplex virus lytic and latent infection*. Nat Rev Microbiol, 2008. **6**(3): p. 211-21.
147. Buch, A., et al., *Inner tegument proteins of Herpes Simplex Virus are sufficient for intracellular capsid motility in neurons but not for axonal targeting*. PLoS Pathog, 2017. **13**(12): p. e1006813.
148. Bloom, D.C., N.V. Giordani, and D.L. Kwiatkowski, *Epigenetic regulation of latent HSV-1 gene expression*. Biochim Biophys Acta, 2010. **1799**(3-4): p. 246-56.
149. Collins-McMillen, D. and F.D. Goodrum, *The loss of binary: Pushing the herpesvirus latency paradigm*. Curr Clin Microbiol Rep, 2017. **4**(3): p. 124-131.
150. Liu, Y.T., et al., *Cryo-EM structures of herpes simplex virus type 1 portal vertex and packaged genome*. Nature, 2019. **570**(7760): p. 257-261.
151. Zhou, Z.H., et al., *Seeing the herpesvirus capsid at 8.5 Å*. Science, 2000. **288**(5467): p. 877-80.
152. Bowman, B.R., et al., *Structure of the herpesvirus major capsid protein*. EMBO J, 2003. **22**(4): p. 757-65.
153. Zhou, Z.H., et al., *Visualization of tegument-capsid interactions and DNA in intact herpes simplex virus type 1 virions*. J Virol, 1999. **73**(4): p. 3210-8.
154. Hilterbrand, A.T. and E.E. Heldwein, *Go go gadget glycoprotein!: HSV-1 draws on its sizeable glycoprotein tool kit to customize its diverse entry routes*. PLoS Pathog, 2019. **15**(5): p. e1007660.
155. Madavaraju, K., et al., *Herpes Simplex Virus Cell Entry Mechanisms: An Update*. Front Cell Infect Microbiol, 2020. **10**: p. 617578.
156. Dollery, S.J., M.G. Delboy, and A.V. Nicola, *Low pH-induced conformational change in herpes simplex virus glycoprotein B*. J Virol, 2010. **84**(8): p. 3759-66.
157. Sodeik, B., M.W. Ebersold, and A. Helenius, *Microtubule-mediated transport of incoming herpes simplex virus 1 capsids to the nucleus*. J Cell Biol, 1997. **136**(5): p. 1007-21.
158. Padeloup, D., et al., *Herpesvirus capsid association with the nuclear pore complex and viral DNA release involve the nucleoporin CAN/Nup214 and the capsid protein pUL25*. J Virol, 2009. **83**(13): p. 6610-23.
159. Herold, B.C., et al., *Glycoprotein C of herpes simplex virus type 1 plays a principal role in the adsorption of virus to cells and in infectivity*. J Virol, 1991. **65**(3): p. 1090-8.
160. Herold, B.C., et al., *Glycoprotein-C-Independent Binding of Herpes-Simplex Virus to Cells Requires Cell-Surface Heparan-Sulfate and Glycoprotein-B*. Journal of General Virology, 1994. **75**: p. 1211-1222.
161. Turner, A., et al., *Glycoproteins gB, gD, and gH/gL of herpes simplex virus type 1 are necessary and sufficient to mediate membrane fusion in a Cos cell transfection system*. J Virol, 1998. **72**(1): p. 873-5.
162. Agelidis, A.M. and D. Shukla, *Cell entry mechanisms of HSV: what we have learned in recent years*. Future Virology, 2015. **10**(10): p. 1145-1154.
163. Atanasiu, D., et al., *Regulation of herpes simplex virus gB-induced cell-cell fusion by mutant forms of gH/gL in the absence of gD and cellular receptors*. mBio, 2013. **4**(2).

-
164. Heldwein, E.E., et al., *Crystal structure of glycoprotein B from herpes simplex virus 1*. *Science*, 2006. **313**(5784): p. 217-20.
 165. Cooper, R.S. and E.E. Heldwein, *Herpesvirus gB: A Finely Tuned Fusion Machine*. *Viruses*, 2015. **7**(12): p. 6552-69.
 166. Dingwell, K.S. and D.C. Johnson, *The herpes simplex virus gE-gI complex facilitates cell-to-cell spread and binds to components of cell junctions*. *J Virol*, 1998. **72**(11): p. 8933-42.
 167. El Kasmi, I. and R. Lippe, *Herpes simplex virus 1 gN partners with gM to modulate the viral fusion machinery*. *J Virol*, 2015. **89**(4): p. 2313-23.
 168. Haanes, E.J., et al., *The UL45 Gene-Product Is Required for Herpes-Simplex Virus Type-1 Glycoprotein B-Induced Fusion*. *Journal of Virology*, 1994. **68**(9): p. 5825-5834.
 169. Snyder, A., K. Polcicova, and D.C. Johnson, *Herpes simplex virus gE/gI and US9 proteins promote transport of both capsids and virion glycoproteins in neuronal axons*. *J Virol*, 2008. **82**(21): p. 10613-24.
 170. Chouljenko, V.N., et al., *The herpes simplex virus type 1 UL20 protein and the amino terminus of glycoprotein K (gK) physically interact with gB*. *J Virol*, 2010. **84**(17): p. 8596-606.
 171. Musarrat, F., et al., *The Amino Terminus of Herpes Simplex Virus 1 Glycoprotein K (gK) Is Required for gB Binding to Akt, Release of Intracellular Calcium, and Fusion of the Viral Envelope with Plasma Membranes*. *J Virol*, 2018. **92**(6).
 172. Aubert, M., et al., *The antiapoptotic herpes simplex virus glycoprotein J localizes to multiple cellular organelles and induces reactive oxygen species formation*. *J Virol*, 2008. **82**(2): p. 617-29.
 173. Tran, L.C., et al., *A herpes simplex virus 1 recombinant lacking the glycoprotein G coding sequences is defective in entry through apical surfaces of polarized epithelial cells in culture and in vivo*. *Proc Natl Acad Sci U S A*, 2000. **97**(4): p. 1818-22.
 174. Mardberg, K., et al., *Herpes simplex virus type 1 glycoprotein C is necessary for efficient infection of chondroitin sulfate-expressing gro2C cells*. *Journal of General Virology*, 2002. **83**: p. 291-300.
 175. Oldberg, A., L. Kjellen, and M. Hook, *Cell-surface heparan sulfate. Isolation and characterization of a proteoglycan from rat liver membranes*. *J Biol Chem*, 1979. **254**(17): p. 8505-10.
 176. Peerboom, N., et al., *Binding Kinetics and Lateral Mobility of HSV-1 on End-Grafted Sulfated Glycosaminoglycans*. *Biophysical Journal*, 2017. **113**(6): p. 1223-1234.
 177. Lycke, E., et al., *Binding of herpes simplex virus to cellular heparan sulphate, an initial step in the adsorption process*. *J Gen Virol*, 1991. **72 (Pt 5)**: p. 1131-7.
 178. Feyzi, E., et al., *Structural requirement of heparan sulfate for interaction with herpes simplex virus type 1 virions and isolated glycoprotein C*. *J Biol Chem*, 1997. **272**(40): p. 24850-7.
 179. O'Donnell, C.D. and D. Shukla, *The Importance of Heparan Sulfate in Herpesvirus Infection*. *Virol Sin*, 2008. **23**(6): p. 383-393.

180. Mardberg, K., et al., *Mutational analysis of the major heparan sulfate-binding domain of herpes simplex virus type 1 glycoprotein C*. J Gen Virol, 2001. **82**(Pt 8): p. 1941-50.
181. Frank, I. and H.M. Friedman, *A novel function of the herpes simplex virus type 1 Fc receptor: participation in bipolar bridging of antiviral immunoglobulin G*. J Virol, 1989. **63**(11): p. 4479-88.
182. Ndjamen, B., et al., *The herpes virus Fc receptor gE-gI mediates antibody bipolar bridging to clear viral antigens from the cell surface*. PLoS Pathog, 2014. **10**(3): p. e1003961.
183. Fruh, K., et al., *A viral inhibitor of peptide transporters for antigen presentation*. Nature, 1995. **375**(6530): p. 415-8.
184. Fries, L.F., et al., *Glycoprotein-C of Herpes-Simplex Virus-1 Is an Inhibitor of the Complement Cascade*. Journal of Immunology, 1986. **137**(5): p. 1636-1641.
185. Lubinski, J., et al., *In vivo role of complement-interacting domains of herpes simplex virus type 1 glycoprotein gC*. J Exp Med, 1999. **190**(11): p. 1637-46.
186. Hung, S.L., et al., *Structural Basis of C3b Binding by Glycoprotein-C of Herpes-Simplex Virus*. Journal of Virology, 1992. **66**(7): p. 4013-4027.
187. Komala Sari, T., K.A. Gianopoulos, and A.V. Nicola, *Glycoprotein C of Herpes Simplex Virus 1 Shields Glycoprotein B from Antibody Neutralization*. J Virol, 2020. **94**(5).
188. Delguste, M., et al., *Regulatory Mechanisms of the Mucin-Like Region on Herpes Simplex Virus during Cellular Attachment*. ACS Chem Biol, 2019. **14**(3): p. 534-542.
189. Buch, M.H.C., *PhD thesis: Structural Studies of Polyomavirus and Herpesvirus*, University Tuebingen. 2016.
190. Stanley, P., *Chinese-Hamster Ovary Cell Mutants with Multiple Glycosylation Defects for Production of Glycoproteins with Minimal Carbohydrate Heterogeneity*. Molecular and Cellular Biology, 1989. **9**(2): p. 377-383.
191. Norden, R., et al., *O-linked glycosylation of the mucin domain of the herpes simplex virus type 1-specific glycoprotein gC-1 is temporally regulated in a seed-and-spread manner*. J Biol Chem, 2015. **290**(8): p. 5078-91.
192. Holm, L., *DALI and the persistence of protein shape*. Protein Sci, 2020. **29**(1): p. 128-140.
193. Bliven, S., et al., *Automated evaluation of quaternary structures from protein crystals*. Plos Computational Biology, 2018. **14**(4).
194. Awasthi, S., J.M. Lubinski, and H.M. Friedman, *Immunization with HSV-1 glycoprotein C prevents immune evasion from complement and enhances the efficacy of an HSV-1 glycoprotein D subunit vaccine*. Vaccine, 2009. **27**(49): p. 6845-53.
195. Egan, K.P., et al., *An HSV-2 nucleoside-modified mRNA genital herpes vaccine containing glycoproteins gC, gD, and gE protects mice against HSV-1 genital lesions and latent infection*. PLoS Pathog, 2020. **16**(7): p. e1008795.
196. Egan, K., et al., *Vaccines to prevent genital herpes*. Transl Res, 2020. **220**: p. 138-152.
197. WuDunn, D. and P.G. Spear, *Initial interaction of herpes simplex virus with cells is binding to heparan sulfate*. J Virol, 1989. **63**(1): p. 52-8.

198. Laquerre, S., et al., *Heparan sulfate proteoglycan binding by herpes simplex virus type 1 glycoproteins B and C, which differ in their contributions to virus attachment, penetration, and cell-to-cell spread*. J Virol, 1998. **72**(7): p. 6119-30.
199. Sarrazin, S., W.C. Lamanna, and J.D. Esko, *Heparan sulfate proteoglycans*. Cold Spring Harb Perspect Biol, 2011. **3**(7).
200. Xu, D. and J.D. Esko, *Demystifying heparan sulfate-protein interactions*. Annu Rev Biochem, 2014. **83**: p. 129-57.
201. Hudson, K.L., et al., *Carbohydrate-Aromatic Interactions in Proteins*. Journal of the American Chemical Society, 2015. **137**(48): p. 15152-15160.
202. Cairns, T.M., et al., *Dissection of the antibody response against herpes simplex virus glycoproteins in naturally infected humans*. J Virol, 2014. **88**(21): p. 12612-22.
203. Huemer, H.P., et al., *Factors Influencing the Interaction of Herpes-Simplex Virus Glycoprotein-C with the 3rd Component of Complement*. Archives of Virology, 1992. **127**(1-4): p. 291-303.
204. Herold, B.C., et al., *Differences in the susceptibility of herpes simplex virus types 1 and 2 to modified heparin compounds suggest serotype differences in viral entry*. J Virol, 1996. **70**(6): p. 3461-9.
205. Gerber, S.I., B.J. Belval, and B.C. Herold, *Differences in the role of glycoprotein C of HSV-1 and HSV-2 in viral binding may contribute to serotype differences in cell tropism*. Virology, 1995. **214**(1): p. 29-39.
206. Rux, A.H., et al., *Kinetic analysis of glycoprotein C of herpes simplex virus types 1 and 2 binding to heparin, heparan sulfate, and complement component C3b*. Virology, 2002. **294**(2): p. 324-32.
207. Zhao, Y.H., et al., *Two routes for production and purification of Fab fragments in biopharmaceutical discovery research: Papain digestion of mAb and transient expression in mammalian cells*. Protein Expression and Purification, 2009. **67**(2): p. 182-189.
208. Ahmad, Z.A., et al., *scFv Antibody: Principles and Clinical Application*. Clinical & Developmental Immunology, 2012.
209. Nelson, A.L., *Antibody fragments Hope and hype*. Mabs, 2010. **2**(1): p. 77-83.
210. Eyer, L. and K. Hruska, *Single-domain antibody fragments derived from heavy-chain antibodies: a review*. Veterinarni Medicina, 2012. **57**(9): p. 439-513.
211. Pardon, E., et al., *A general protocol for the generation of Nanobodies for structural biology*. Nature Protocols, 2014. **9**(3): p. 674-693.
212. Henry, K.A. and C.R. MacKenzie, *Antigen recognition by single-domain antibodies: structural latitudes and constraints*. Mabs, 2018. **10**(6): p. 815-826.
213. Lotze, J., et al., *Peptide-tags for site-specific protein labelling in vitro and in vivo*. Mol Biosyst, 2016. **12**(6): p. 1731-45.
214. Lofblom, J., et al., *Affibody molecules: engineered proteins for therapeutic, diagnostic and biotechnological applications*. FEBS Lett, 2010. **584**(12): p. 2670-80.

215. Traenkle, B., et al., *Monitoring interactions and dynamics of endogenous beta-catenin with intracellular nanobodies in living cells*. Mol Cell Proteomics, 2015. **14**(3): p. 707-23.
216. Zavrtnik, U., et al., *Structural Basis of Epitope Recognition by Heavy-Chain Camelid Antibodies*. J Mol Biol, 2018. **430**(21): p. 4369-4386.
217. Virant, D., et al., *A peptide tag-specific nanobody enables high-quality labeling for dSTORM imaging*. Nat Commun, 2018. **9**(1): p. 930.
218. Braun, M.B., et al., *Peptides in headlock--a novel high-affinity and versatile peptide-binding nanobody for proteomics and microscopy*. Sci Rep, 2016. **6**: p. 19211.
219. Sanghai, Z.A., et al., *Modular assembly of the nucleolar pre-60S ribosomal subunit*. Nature, 2018. **556**(7699): p. 126-129.
220. Rode, A., et al., *Chromothripsis in cancer cells: An update*. Int J Cancer, 2016. **138**(10): p. 2322-33.
221. Hakem, R., *DNA-damage repair; the good, the bad, and the ugly*. EMBO J, 2008. **27**(4): p. 589-605.
222. Hao, D.P., L. Wang, and L.J. Di, *Distinct mutation accumulation rates among tissues determine the variation in cancer risk*. Scientific Reports, 2016. **6**.
223. Li, N., et al., *Human papillomavirus type distribution in 30,848 invasive cervical cancers worldwide: variation by geographical region, histological type and year of publication*. International Journal of Cancer, 2011. **128**(4): p. 927-935.
224. El-Serag, H.B., *Epidemiology of Viral Hepatitis and Hepatocellular Carcinoma*. Gastroenterology, 2012. **142**(6): p. 1264-+.
225. Mesri, E.A., E. Cesarman, and C. Boshoff, *Kaposi's sarcoma and its associated herpesvirus*. Nature Reviews Cancer, 2010. **10**(10): p. 707-719.
226. Khan, G. and M.J. Hashim, *Global burden of deaths from Epstein-Barr virus attributable malignancies 1990-2010*. Infectious Agents and Cancer, 2014. **9**.
227. Goncalves, D.U., et al., *Epidemiology, Treatment, and Prevention of Human T-Cell Leukemia Virus Type 1-Associated Diseases*. Clinical Microbiology Reviews, 2010. **23**(3): p. 577-+.
228. Schadendorf, D., et al., *Merkel cell carcinoma: Epidemiology, prognosis, therapy and unmet medical needs*. European Journal of Cancer, 2017. **71**: p. 53-69.
229. Polk, D.B. and R.M. Peek, *Helicobacter pylori: gastric cancer and beyond*. Nature Reviews Cancer, 2010. **10**(6): p. 403-414.
230. Tubbs, A. and A. Nussenzweig, *Endogenous DNA Damage as a Source of Genomic Instability in Cancer*. Cell, 2017. **168**(4): p. 644-656.
231. Lindahl, T. and D.E. Barnes, *Repair of endogenous DNA damage*. Cold Spring Harbor Symposia on Quantitative Biology, 2000. **65**: p. 127-133.
232. Zou, Z., et al., *Induction of reactive oxygen species: an emerging approach for cancer therapy*. Apoptosis, 2017. **22**(11): p. 1321-1335.
233. Wang, D. and S.J. Lippard, *Cellular processing of platinum anticancer drugs*. Nat Rev Drug Discov, 2005. **4**(4): p. 307-20.
234. Li, J., et al., *Recent advances in delivery of drug-nucleic acid combinations for cancer treatment*. Journal of Controlled Release, 2013. **172**(2): p. 589-600.

-
235. Delgado, J.L., et al., *Topoisomerases as anticancer targets*. *Biochemical Journal*, 2018. **475**: p. 373-398.
 236. Jordan, M.A. and L. Wilson, *Microtubules as a target for anticancer drugs*. *Nature Reviews Cancer*, 2004. **4**(4): p. 253-265.
 237. Querido, E., et al., *Identification of three functions of the adenovirus E4orf6 protein that mediate p53 degradation by the E4orf6-E1B55K complex*. *Journal of Virology*, 2001. **75**(2): p. 699-709.
 238. Lechner, M.S. and L.A. Laimins, *Inhibition of P53 DNA-Binding by Human Papillomavirus E6 Proteins*. *Journal of Virology*, 1994. **68**(7): p. 4262-4273.
 239. Drolet, M., et al., *Population-level impact and herd effects following the introduction of human papillomavirus vaccination programmes: updated systematic review and meta-analysis*. *Lancet*, 2019. **394**(10197): p. 497-509.
 240. Chen, G.X., et al., *Clinical utility of recombinant adenoviral human p53 gene therapy: current perspectives*. *Onco Targets Ther*, 2014. **7**: p. 1901-9.
 241. Russell, L. and K.W. Peng, *The emerging role of oncolytic virus therapy against cancer*. *Chin Clin Oncol*, 2018. **7**(2): p. 16.
 242. Belousova, N., et al., *Native and engineered tropism of vectors derived from a rare species D adenovirus serotype 43*. *Oncotarget*, 2016. **7**(33): p. 53414-53429.
 243. Nwanegbo, E., et al., *Prevalence of neutralizing antibodies to adenoviral serotypes 5 and 35 in the adult populations of The Gambia, South Africa, and the United States*. *Clinical and Diagnostic Laboratory Immunology*, 2004. **11**(2): p. 351-357.
 244. Waddington, S.N., et al., *Adenovirus serotype 5 hexon mediates liver gene transfer*. *Cell*, 2008. **132**(3): p. 397-409.
 245. Mathias, P., M. Galleno, and G.R. Nemerow, *Interactions of soluble recombinant integrin alpha v beta 5 with human adenoviruses*. *Journal of Virology*, 1998. **72**(11): p. 8669-8675.
 246. Manickan, E., et al., *Rapid Kupffer cell death after intravenous injection of adenovirus vectors*. *Molecular Therapy*, 2006. **13**(1): p. 108-117.

VII. Appendix

Appendix 1: Molecular weights and extinction coefficients of the different FliC constructs.

Construct	Mutations	M _w [Da]	E ₂₈₀ [M ⁻¹ cm ⁻¹]
FliC		61024	20860
FliCΔD0	(Δ1-47, Δ555-595)	51899	19370
FliCΔD01	(Δ1-175, Δ499-595)	32206	19370
FliCD12	(Δ1-47, Δ208-451>S, Δ555-595)	28021	5960
FliCD12	(Δ1-61, Δ208-451>S, Δ555-595)	26506	5960
FliCD12	(Δ1-61, Δ208-451>S, Δ542-595)	25264	5960
FliCΔD4	(Δ255-371)>GSGSA	53124	17880
FliCΔD4	(Δ249-375)>GSGS	52017	17880
FliCΔD4	(Δ246-375)>SGSG	51668	17880
FliCΔD4	(Δ249-372)>S	52103	17880
FliCΔD4	(Δ251-372)	52465	17880
FliCΔD34	(Δ208-451)>S	40185	8940

Appendix 2: EcN FliC constructs used in this thesis. Highlighted are the His-tag and liker sequence is (orange), the Enterokinase cleavage site (blue), the TEV cleavage site (green) and additional amino acids used to links the strands after a domain deletion (red).

FliCe

MGHHHHHHHH HHSSGHI DDD DKHMAQVINT NSLSLITQNN INKNQSALSS SIERLSSGLR
 INSAKDDAAG QAIANRFTSN IKGLTQAARN ANDGISVAQT TEGALSEINN NLQRIRELTV
 QASTGTNSDS DLDSIQDEIK SRLDEIDRVS GQTQFNGVNV LAKDGSMKIQ VGANDGQTIT
 IDLKKIDSST LGLNGFNVNG SGTIANKAAT ISDLTAAKMD AATNTITTTN NALTASKALD
 QLKDGDTVTI KADAAQTATV YTYNASAGNF SFSNVSNNNTS AKAGDVAASL LPPAGQTASG
 VYKAASGEVN FDVDANGKIT IGGQEAYLTS DGNLTTNDAG GATAATLDGL FKKAGDGQSI
 GFNKTASVTM GGTTYNFKTG ADAGAATANA GVSFTDTASK ETVLNKVATA KQGTAVAANG
 DTSATITYKS GVQTYQAVFA AGDGTASAKY ADNTDVSNTAT ATYTDADGEM TTIGSYTTYKY
 SIDANNGKVT VDSGTGTGKY APKVGAEVYV SANGTLTTDA TSEGTVTKDP LKALDEAISS
 IDKFRSSLGA IQNRLDSAVT NLNNTTTNLS EAQSRIQDAD YATEVSNMSK AQIIQQAGNS
 VLAKANQVPQ QVLSLLQG

FliCeTEV

MGHHHHHHHH HHSSGHI DDD DKHTTENLYF Q SMAQVINTN SLSLITQNNI NKNQSALSSS
 IERLSSGLRI NSAKDDAAGQ AIANRFTSNI KGLTQAARNA NDGISVAQTT EGALSEINNN
 LQRIRELTVQ ASTGTNSDSD LDSIQDEIKS RLDEIDRVSG QTQFNGVNVL AKDGSMKIQV
 GANDGQTITI DLKKIDSIDL GLNGFNVNGS GTIANKAATI SDLTAAKMDA ATNTITTTNN
 ALTASKALDQ LKDGDTVTIK ADAAQATVY TYNASAGNFS FSNVSNNTSA KAGDVAASLL
 PPAGQTASGV YKAASGEVNF DVDANGKITI GGQEAYLTS DGNLTTNDAGG ATAATLDGLF
 KKAGDGQSIG FNKTASVTMG GTTYNFKTGA DAGAATANAG VSFTDTASKE TVLNKVATAK
 QGTAVAANGD TSATITYKSG VQTYQAVFAA GDGTASAKYA DNTDVSNATA TYTDADGEMT
 TIGSYTTKYS IDANNGKVTV DSGTGTGKYA PKVGAEVYVS ANGLTTTDDAT SEGTVTKDPL
 KALDEAISSI DKFRSSLGAI QNRLDSAVTN LNNTTTNLSE AQSRIQDADY ATEVSNMSKA
 QIIQQAGNSV LAKANQVPQQ VLSLLQG

FliCΔD0 (Δ1-47, Δ555-595)

MGHHHHHHHH HHSSGHITTE NLYFQSQAIA NRFTSNIKGL TQAARNANDG ISVAQTTEGA
 LSEINNNLQR IRELTVQAST GTNSDSDLDS IQDEIKSRLD EIDRVSGQTQ FNGVNVLAKD
 GSMKIQVGAN DGQTTIDDLK KIDSIDLGLN GFNVNGSGTI ANKAATISDL TAAKMDAATN
 TITTTNNALT ASKALDQLKD GDTVTIKADA AQTATVYTYN ASAGNFSFSN VSNNTSAKAG
 DVAASLLPPA GQTASGVYKA ASGEVNFVDV ANGKITIGGQ EAYLTS DGNL TTNDAGGATA
 ATLDGLFKKA GDGQSIGFNK TASVTMGGTT YNFKTGADAG AATANAGVSF TDTASKETVL
 NKVATAKQGT AVAANGD TSA TITYKSGVQT YQAVFAAGD TASAKYADNT DVSNATATYT
 DADGEMTTIG SYTTKYSIDA NNGKVTVDSD TGTGKYAPKV GAEVYVSANG TLTTDDATSEG
 TVTKDPLKAL DEAISSIDKF RSSLGAIQNR LDSAVTNLNN TTTNLSEAQS RIQD

FliCΔD01 (Δ1-175, Δ499-595)

MGHHHHHHHH HHSSGHITTE NLYFQSNVSG TIANKAATIS DLTAAKMDAA TNTITTTNNA
 LTASKALDQL KDGDTVTIKA DAAQTATVYT YNASAGNFSF SNVSNNTSAK AGDVAASLLP
 PAGQTASGVY KAASGEVNF VDANGKITIG GQEAYLTS DGNL TTNDAGGATA TAATLDGLFK
 KAGDGQSIGF NKTASVTMGG TTYNFKTGAD AGAATANAGV SFTDTASKET VLNKVATAKQ
 GTAVAANGDT SATITYKSGV QTYQAVFAAG DGTASAKYAD NTDVSNATAT YTDADGEMTT
 IGSYTTKYSI DANNGKVTVD SGTGTGKYAP KVGAEVYVSA NGTLTTDDATS

FliCΔ12 (Δ1-47, Δ208-451>S, Δ555-595)

MGHHHHHHHH HHSSGHITTE NLYFQSQAIA NRFTSNIKGL TQAARNANDG ISVAQTTEGA
 LSEINNNLQR IRELTVQAST GTNSDSDLDS IQDEIKSRLD EIDRVSGQTQ FNGVNVLAKD
 GSMKIQVGAN DGQTTIDDLK KIDSIDLGLN GFNVNGSGTI ANKAATISDL TAAKMDAATN
 TITTTNNSGS YTTKYSIDAN NGKVTVDSDT GTGKYAPKVG AEVYVSANGT LTDDATSEGT
 VTKDPLKALD EAISSIDKFR SSSLGAIQNR LDSAVTNLNT TTNLSEAQSR IQD

FliCD12 ($\Delta 1-61$, $\Delta 208-451$ >S, $\Delta 555-595$)

MGHHHHHHHH HHSSGHITTE NLYFQSTQAA RNANDGISVA QTTEGALSEI NNNLQRIREL
 TVQASTGTNS DSDLDSIQDE IKSRLDEIDR VSGQTQFNGV NVLAKDGSMK IQVGANDGQT
 ITIDLKIDS DTLGLNGFNV NSGTIANKA ATISDLTAAK MDAATNTITT TNN^SGSYTTK
 YSIDANNGKV TVDSGTGTGK YAPKVGAEVY VSANGTLTTD ATSEGTVTKD PLKALDEAIS
 SIDKFRSSLG AIQNRLLDSAV TNLNNTTTL SEAQSRIQD

FliCD12 ($\Delta 1-61$, $\Delta 208-451$ >S, $\Delta 542-595$)

MGHHHHHHHH HHSSGHITTE NLYFQSTQAA RNANDGISVA QTTEGALSEI NNNLQRIREL
 TVQASTGTNS DSDLDSIQDE IKSRLDEIDR VSGQTQFNGV NVLAKDGSMK IQVGANDGQT
 ITIDLKIDS DTLGLNGFNV NSGTIANKA ATISDLTAAK MDAATNTITT TNN^SGSYTTK
 YSIDANNGKV TVDSGTGTGK YAPKVGAEVY VSANGTLTTD ATSEGTVTKD PLKALDEAIS
 SIDKFRSSLG AIQNRLLDSAV TNLNNTTTL

FliC Δ 4 ($\Delta 255-371$ >GSGSA)

MGHHHHHHHH HHSSGHI^{DDD} DKHTTENLYF Q^SMAQVINTN SLSLITQNNI NKNQSALSSS
 IERLSSGLRI NSAKDDAAGQ AIANRFTSNI KGLTQAARNA NDGISVAQTT EGALSEINNN
 LQRIRELTVQ ASTGTNSDSD LDSIQDEIKS RLDEIDRVSG QTQFNGVNVL AKDGSMKIQV
 GANDGQTITI DLKKIDSDDL GLNGFNVNGS GTIANKAATI SDLTAAKMDA ATNTITTTNN
 ALTASKALDQ LKDGDTVTIK ADAAQATVY TYNASAGNFS FSNVSN^{GSGS} A^TDTASKETV
 LNKVATAKQG TAVAANGDTS ATITYKSGVQ TYQAVFAAGD GTASAKYADN TDVSNATATY
 TDADGEMTTI GSYTTKYSID ANNGKVTVDS GTGTGKYAPK VGAEVYVSAN GTLTTDATSE
 GTVTKDPLKA LDEAISSIDK FRSSLGAIQN RLDSAVTNLN NTTNLSEAQ SRIQDADYAT
 EVSNMSKAQI IQQAGNSVLA KANQVPQQVL SLLQG

FliC Δ 4 ($\Delta 249-375$ >GSGS)

MGHHHHHHHH HHSSGHI^{DDD} DKHTTENLYF Q^SMAQVINTN SLSLITQNNI NKNQSALSSS
 IERLSSGLRI NSAKDDAAGQ AIANRFTSNI KGLTQAARNA NDGISVAQTT EGALSEINNN
 LQRIRELTVQ ASTGTNSDSD LDSIQDEIKS RLDEIDRVSG QTQFNGVNVL AKDGSMKIQV
 GANDGQTITI DLKKIDSDDL GLNGFNVNGS GTIANKAATI SDLTAAKMDA ATNTITTTNN
 ALTASKALDQ LKDGDTVTIK ADAAQATVY TYNASAGNFS ^{GSGS}SKETVL NKVATAKQGT
 AVAANGDTS TITYKSGVQT YQAVFAAGDG TASAKYADNT DVSNATATYT DADGEMTTIG
 SYTTKYSIDA NNGKVTVDSG TGTGKYAPKV GAEVYVSANG TLTDDATSEG TVTKDPLKAL
 DEAISSIDKF RSSLGAIQNR LDSAVTNLNN TTTNLSEAQS RIQDADYATE VSNMSKAQII
 QQAGNSVLAK ANQVPQQVLS LLLQG

FliCAD4 ($\Delta 246-375>SGSG$)

MGHHHHHHHH HHSSGHI DDD DKHTTENLYF QSMAQVINTN SLSLITQNNI NKNQSA LSSS
 IERLSSGLRI NSAKDDAAGQ AIANRF TSN I KGLTQAARNA NDGISVAQTT EGALSE INNN
 LQRI RELTVQ ASTGTNSDSD LDSIQDEIKS RLDEIDRVSG QTQFNGVNV L AKDGS MKIQV
 GANDGQTITI DLKKIDS DTL GLNGFN VNGS GTIANKAATI SDLTAAKMDA ATNTITTTNN
 ALTASKALDQ LKDGDTVTIK ADAAQ TATVY TYNASAGSGS GSKETVLNKV ATAKQGTAVA
 ANGDTSATIT YKSGVQTYQA VFAAGDGTAS AKYADNTDVS NATATYTDAD GEMTTIGSYT
 TKYSIDANNG KVTVD SGTGT GKYAPKVGAE VYVSANGTLT TDATSEGTVT KDPLKALDEA
 ISSIDKFRSS LGAIQNR LDS AVTNLNN TTT NLSEAQSRIQ DADYATEVSN MSKAQIIQQA
 GNSVLAKANQ VPQQVLSLLQ G

FliCAD4 ($\Delta 249-372>S$)

MGHHHHHHHH HHSSGHITTE NLYFQ SMAQV INTNSLSLIT QNNINKNQSA LSSSIERLSS
 GLRINSAKDD AAGQAIANRF TSN I KGLTQA ARNANDGISV AQTTEGALSE INNNLQRI RE
 LTVQASTGTN SDSDLDSIQD EIKSRLDEID RVSGQTQFNG VNV LAKDGS M KIQVGANDGQ
 TITIDLKKID SDTLGLNGFN VNGSGTIANK AATISDLTAA KMDAATNTIT TTNNALTASK
 ALDQLKDGDT VTIKADAAQT ATVYTYNASA GNFS DTASK ETVLNKVATA KQGTAVAANG
 DTSATITYKS GVQTYQAVFA AGDGTASAKY ADNTDVS NAT ATYTDADGEM TTIGSYTTKY
 SIDANNGKVT VDSGTGTGKY APKVGAEVYV SANGTLT TDA TSEGTVTKDP LKALDEAISS
 IDKFRSSLGA IQNR LDSAVT NLNNTTTNLS EAQSRIQDAD YATEVSNMSK AQIIQAGNS
 VLAKANQVPQ QVLSLLQ G

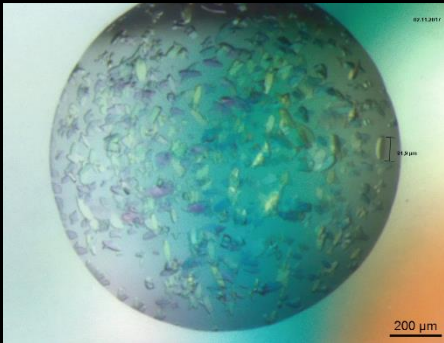
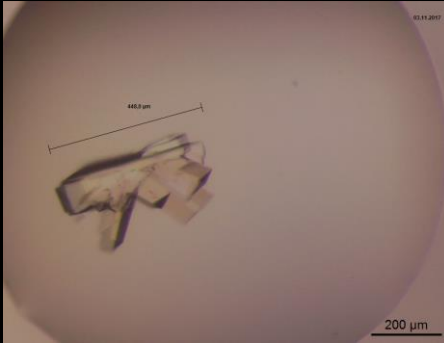
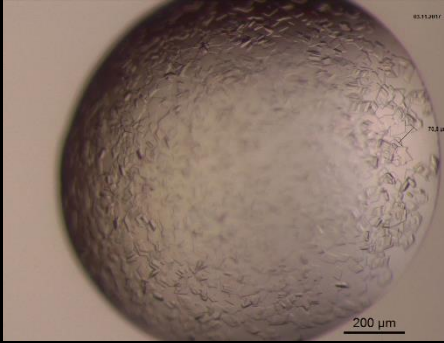

FliCAD4 ($\Delta 252-371$)

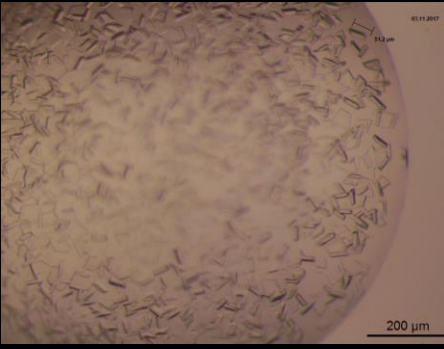


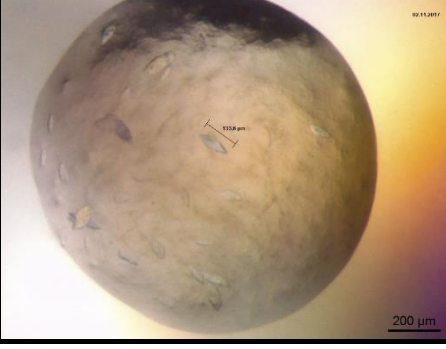
MGHHHHHHHH HHSSGHITTE NLYFQ SMAQV INTNSLSLIT QNNINKNQSA LSSSIERLSS
 GLRINSAKDD AAGQAIANRF TSN I KGLTQA ARNANDGISV AQTTEGALSE INNNLQRI RE
 LTVQASTGTN SDSDLDSIQD EIKSRLDEID RVSGQTQFNG VNV LAKDGS M KIQVGANDGQ
 TITIDLKKID SDTLGLNGFN VNGSGTIANK AATISDLTAA KMDAATNTIT TTNNSGSYTT
 KYSIDANNGK VTVDSGTGTG KYAPKVGAEV YVSANGTLT DATSEGTVTK DPLKALDEAI
 SSIDKFRSSL GAIQNR LDSA VTNLNN TTTN LSEAQSRIQ ADYATEVSNM SKAQIIQQA
 NSVLAKANQV PQQVLSLLQ G

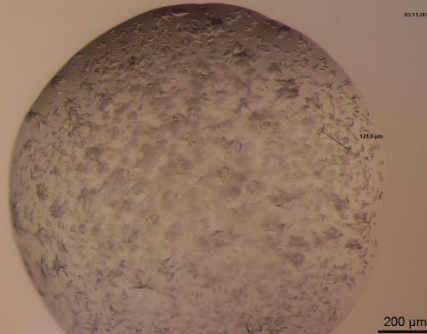

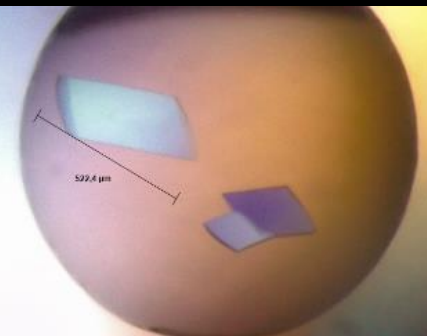

FliCAD34 ($\Delta 209-450>S$)


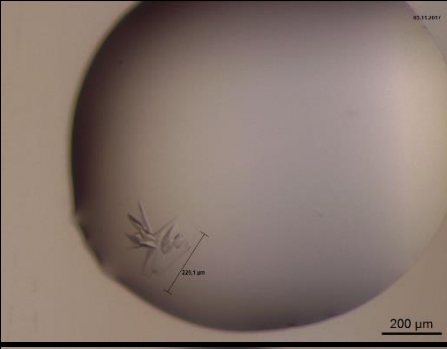


MGHHHHHHHH HHSSGHIHTT ENLYFQ SMAQV VINTNSLSLI TQNNINKNQS ALSSSIERLS
 SGLRINSAKD DAAGQAIANR FTSNIKGLTQ AARNANDGIS VAQTTEGALS EINNNLQRI R
 ELTVQASTGT NSDSDLDSIQ DEIKSRLDEI DRVSGQTQFN GNV LAKDGS M MKIQVGANDG
 QTITIDLKKI DSDTLGLNGF NVNGSGTIAN KAATISDLTA AKMDAATNTI TTTNNSGSYT
 TKYSIDANNG KVTVD SGTGT GKYAPKVGAE VYVSANGTLT TDATSEGTVT KDPLKALDEA
 ISSIDKFRSS LGAIQNR LDS AVTNLNN TTT NLSEAQSRIQ DADYATEVSN MSKAQIIQQA
 GNSVLAKANQ VPQQVLSLLQ G

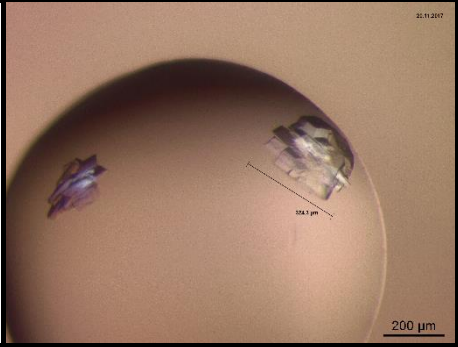
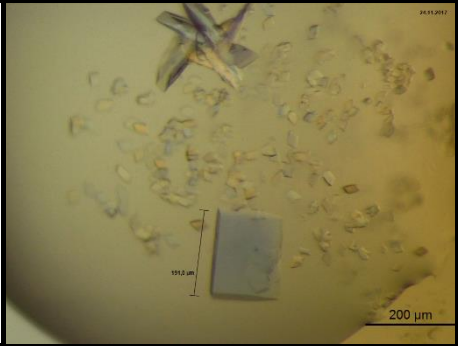


Appendix 3: FliCD234 crystallization conditions. Initial crystallization conditions with images of the crystals from commercially available crystallization screens (Hampton Crystal Screen1-2, JCSG 1-2, Wizard1-4).

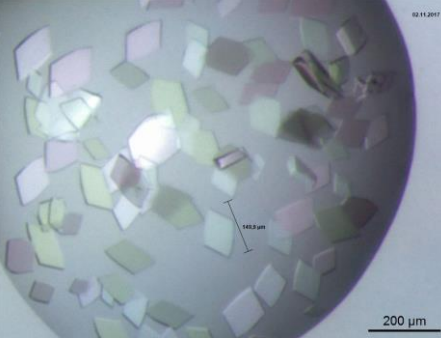

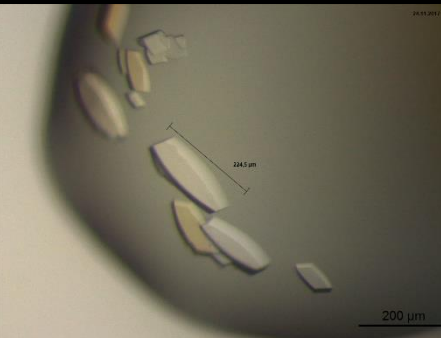

Hampton-CS 1-18	0.1 M Sodium cacodylate pH 6.5 0.2 M Magnesium acetate 20% (w/v) PEG 8000	 Micrograph showing a dense field of small, multi-colored (green, blue, purple) crystals. A scale bar in the bottom right corner indicates 200 μm. A small inset in the top right corner shows a magnified view of a crystal with a dimension of 11.9 μm.
Hampton-CS 1-20	0.1 M Sodium acetate pH 4.6 0.1 M Ammonium sulfate 25% (w/v) PEG 4000	 Micrograph showing a few large, flat, multi-colored crystals. A scale bar in the bottom right corner indicates 200 μm. A horizontal dimension line above one of the crystals indicates a length of 46.0 μm.
Hampton-CS 1-32	2.0 M Ammonium sulfate	 Micrograph showing a dense field of small, multi-colored crystals. A scale bar in the bottom right corner indicates 200 μm. A small inset in the top right corner shows a magnified view of a crystal with a dimension of 73.3 μm.
Hampton-CS 1-46	0.1 M Sodium cacodylate pH 6.5 0.1 M Calcium acetate hydrate 18% (w/v) PEG 8000	 Micrograph showing a dense field of small, multi-colored crystals. A scale bar in the bottom right corner indicates 200 μm. A small inset in the top right corner shows a magnified view of a crystal with a dimension of 11.2 μm.

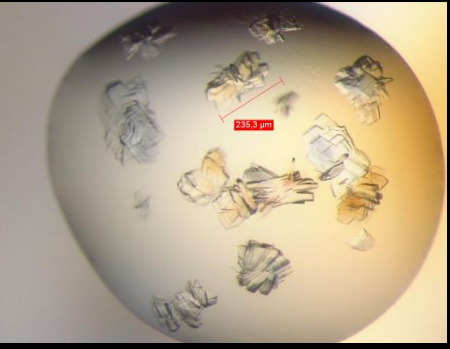


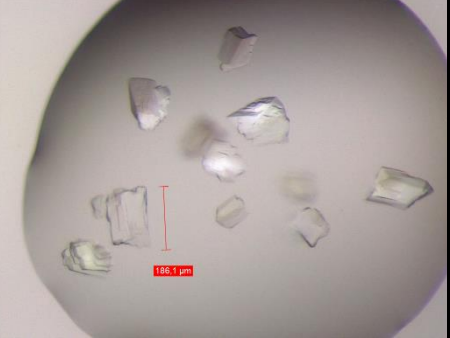
Hampton-CS 1-47	0.1 M Sodium acetate pH 4.6 2.0 M Ammonium sulfate	
Hampton-CS 2-13	0.1 M Sodium acetate pH 4.6 0.1 M Ammonium sulfate 30% (w/v) PEG 2000 MME	
Hampton-CS 2-21	0.1 M MES pH 6.5 0.1 M Sodium phosphate monobasic 0.1 M Potassium phosphate monobasic 2.0 M Sodium chloride	
Hampton-CS 2-23	0.1 M MES pH 6.5 1.6 M Ammonium sulfate 10% (v/v) 1,4-Dioxane	



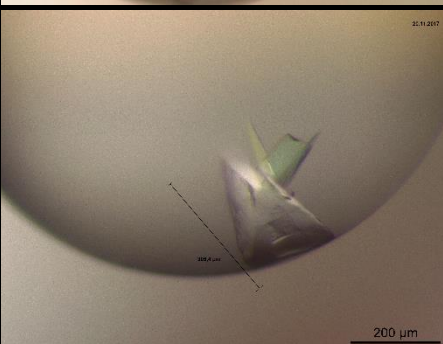

<p>Hampton-CS 2-24</p>	<p>0.1 M MES pH 6.5 0.05 M Cesium chloride 30% (v/v) Jeffamine M-600</p>	
<p>Hampton-CS 2-30</p>	<p>0.1 M HEPES pH 7.5 5% (v/v) MPD 10% (w/v) PEG 6000</p>	
<p>JCSG-plus 1-25</p>	<p>0.1 M Phosphate/citrate pH 4.2 0.2 M Sodium chloride 20% (w/v) PEG 8000</p>	
<p>JCSG-plus 1-35</p>	<p>0.1 M Sodium acetate pH 4.6 2.0 M Ammonium sulfate</p>	

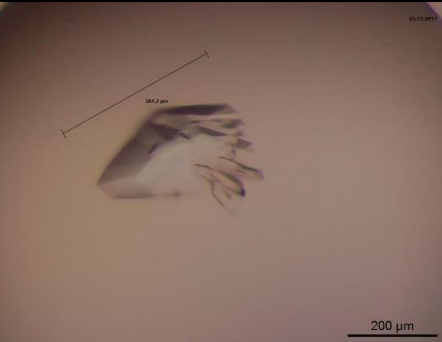
JCSG-plus 1-40	0.1 M Sodium acetate pH 4.5 0.2 M Lithium sulfate 30% (w/v) PEG 8000	
JCSG-plus 1-42	0.1 M Tris pH 8.5 0.1 M Magnesium chloride 20% (w/v) PEG 8000	
JCSG-plus 1-45	0.17 M Ammonium sulfate 15% (v/v) Glycerol 25.5% (w/v) PEG 4000	
JCSG-plus 2-05	0.1 M CAPS pH 10.5 40% (v/v) MPD	

<p>JCSG-plus 2-39</p>	<p>0.1 M BIS-Tris pH 5.5 25% (w/v) PEG 3350</p>	
<p>JCSG-plus 2-43</p>	<p>0.1 M BIS-Tris pH 5.5 0.2 M Ammonium sulfate 25% (w/v) PEG 3350</p>	
<p>JCSG-plus 2-44</p>	<p>0.1 M BIS-Tris pH 5.5 0.2 M Sodium chloride 25% (w/v) PEG 3350</p>	
<p>Wizard 1-17</p>	<p>0.1 M Sodium acetate pH 4.5 0.2 M Lithium sulfate 30% (w/v) PEG 8000</p>	

<p>Wizard 1-31</p>	<p>0.1 M Phosphate/citrate pH 4.2 0.2 M Sodium chloride 20% (w/v) PEG 8000</p>	 <p>Micrograph showing numerous small, multi-faceted, translucent crystals. A scale bar indicates 200 μm. A measurement of 165.3 μm is shown for one crystal.</p>
<p>Wizard 2-03</p>	<p>0.1 M Tris pH 8.5 0.2 M Magnesium chloride 20% (w/v) PEG 8000</p>	 <p>Micrograph showing a dense population of small, irregular, translucent crystals. A scale bar indicates 500 μm. A measurement of 198.7 μm is shown for one crystal.</p>
<p>Wizard 2-19</p>	<p>0.1 M Phosphate/citrate pH 4.2 1.6 M Sodium phosphate monobasic 0.1 M Potassium phosphate dibasic</p>	 <p>Micrograph showing several larger, elongated, translucent crystals. A scale bar indicates 200 μm. A measurement of 245 μm is shown for one crystal.</p>
<p>Wizard 2-28</p>	<p>0.1 M MES pH 6.0 0.2 M Calcium acetate 20% (w/v) PEG 8000</p>	 <p>Micrograph showing a dense population of small, irregular, translucent crystals. A scale bar indicates 200 μm. A measurement of 195.9 μm is shown for one crystal.</p>

Wizard 3-18	0.1 M Citrate pH 4.0 1 M Lithium chloride 20% (w/v) PEG 6000	 Micrograph showing several irregular, light-colored crystals. A red double-headed arrow indicates a length of 235.9 μm. The background is a light yellowish-grey.
Wizard 3-19	0.2 M Ammonium nitrate 20% (w/v) PEG 3350	 Micrograph showing several small, colorful crystals (blue, green, yellow) on a light background. A black double-headed arrow indicates a length of 102.3 μm. A scale bar in the bottom right corner indicates 200 μm.
Wizard 3-30	0.17 M Ammonium sulfate 15% (v/v) Glycerol 25.5% (w/v) PEG 4000	 Micrograph showing several light-colored, irregular crystals. A black double-headed arrow indicates a length of 205.4 μm. A scale bar in the bottom right corner indicates 200 μm.
Wizard 3-32	16% (w/v) PEG 8000 0.04 M Potassium phosphate monobasic 20% (v/v) Glycerol	 Micrograph showing several light-colored, irregular crystals. A red double-headed arrow indicates a length of 186.1 μm. The background is a light greyish-blue.

Wizard 3-47	0.1 M MES pH 6.5 0.2 M Ammonium sulfate 30% (w/v) PEG 5000 MME	
Wizard 4-05	0.1 M Acetate pH 5.5 2 M Ammonium sulfate 2% (v/v) PEG 400	
Wizard 4-31	0.1 M HEPES pH 7.0 0.02 M Magnesium chloride 20% (w/v) Polyacrylic acid 5100	
Wizard 4-33	0.1 M HEPES pH 7.5 0.8 M Potassium phosphate dibasic 0.8 M Sodium phosphate monobasic	

<p>Wizard 4-34</p>	<p>0.1 M MES pH 6.0 0.2 M Ammonium chloride 20% (w/v) PEG 6000</p>	 <p>204.2 μm</p> <p>200 μm</p>
------------------------	--	--

Appendix 4: Crystallographic data and refinement statistic for the FliCD234 data sets.

	FliCD234	FliCD234-SeUrea
Data collection		
Wavelength (Å)	1.0	0.978
Resolution (Å)	50.00 (1.75-1.65)	46.63 (2.36-2.30)
Space Group	4 (P2 ₁)	4 (P2 ₁)
Cell dimension		
<i>a, b, c</i> (Å)	56.87, 61.31, 90.44	57.85, 72.87, 93.65
α, β, γ (°)	97.808	95.187
Reflections	283494 (35707)	4462058 (319004)
Unique	132670 (18996)	67587 (4988)
Redundancy	2.1 (1.9)	66.0 (13.5)
Completeness (%)	91.1 (80.8)	100 (100)
<i>I</i> / σ <i>I</i>	9.36 (0.80)	27.87 (4.99)
<i>R</i> _{meas}	6.3 (113.1)	26.5 (215.1)
CC _{1/2}	99.8 (51.6)	100 (97.7)
Wilson B (Å ²)	33.0	35.6
SAD-phasing		
Phasing power		0.64* (1.01)**
R-Cullis		0.92* (0.81)**
FOM acentric		0.23* (0.32)**
<i>D</i> / σ <i>D</i>		1.23 (1.33)**
Refinement		
<i>R</i> _{work} / <i>R</i> _{free}	24.2/27.4	22.9/27.1
No. atoms / B-value (Å ²)		
Protein A	1999/28.8	2082/33.7
Protein B	1761/27.2	2091/32.0
Water	447/33.8	268/34.6
RMSD angels (°)	0.721	0.587
RMSD bonds (Å)	0.006	0.003
Ramachandran favoured (%)	96.8	92.3
Ramachandran outlier (%)	0	1.5

* For SAD-phasing the resolution range 47.15 - 2.36 Å was used.

** The anomalous signal was used to a resolution shell of 3.63 - 3.44 Å.

Appendix 5: Crystallographic data and refinement statistic for the FliCD12 data sets.

	FliCD12 ($\Delta 1-61, \Delta 208-451 < S, \Delta 555-595$)	FliCD12 ($\Delta 1-61, \Delta 208-451 < S, \Delta 542-595$)
Data collection		
Wavelength (Å)	1.0	1.0
Resolution (Å)	30.00 (1.74-1.70)	30.00 (2.08-2.03)
Space Group	5 (C2)	5 (C2)
Cell dimension		
<i>a, b, c</i> (Å)	115.42, 27.88, 88.24	169.90, 27.45, 99.55
α, β, γ (°)	128.583	106.138
Reflections	89928 (3164)	235267 (16381)
Unique	23968(1396)	29436(2159)
Redundancy	3.8 (2.3)	8.0 (7.6)
Completeness (%)	96.5 (75.9)	99.8 (100)
<i>I</i> / σ <i>I</i>	11.73 (1.28)	5.75 (1.22)
<i>R</i> _{meas}	7.5 (88.1)	33.7 (200.2)
CC _{1/2}	99.8 (51.5)	99.3 (49.1)
Wilson B (Å ²)	30.0	29.3
Refinement		
<i>R</i> _{work} / <i>R</i> _{free}	20.4/24.4	23.6/26.2
No. atoms / B-value (Å ²)		
Protein A	1692/26.7	1695/31.8
Protein B	-	1685/30.5
Water	190/34.5	244/33.8
RMSD angels (°)	0.604	0.412
RMSD bonds (Å)	0.005	0.002
Ramachandran favoured (%)	99.6	98.9
Ramachandran outlier (%)	0	0.3

Appendix 6: Sequence alignment of different FliC constructs from different organisms. In the corresponding crystal structures (2ZBI, 1IO1, 3V47) the last build amino acids are marked (green). Different D0 deletion mutants were designed by shortening the existing mutant in two steps (red and orange) N-terminal and C-terminal. AAs are marked based on their similarity, with an asterisk for a perfect aligned, with a colon mark for a strong similarity or with a dot for a weak similarity.

N-term

```

2ZBI          -----TAQIKGLTQAQRNANDGISLAQTAEALGEISNNLQRIRELAVQASN
1IO1          -----FTSNIKGLTQASRNANDGISIAQTTEGALNEINNNLQRVRELAVQSAN
3V47          GSAKDPQAIANRFTSNIKGLTQASRNANDGISIAQTTEGALNEINNNLQRVRELSVQATN
E. coli Nissle -----QAIAANRFTSNIKGLTQAARNANDGISVAQTTEGALSEINNNLQRIRELTVQAST
                *****:***** *****:***:***.*.*****:***:*.:.

```

C-term

```

2ZBI          VSSFTNAQQTITQIDNALKDINTARADLGAVQNRFTSTVANLQSM-----
1IO1          EAAATTENPLQKIDAALAQVDTLRSDLAAVQNRFNSAITNLGNTVNLNLSAR-----
3V47          AAAKKSTANPLASIDSALSKVDAVRSSLGAIQNRFDSAITNLGNTVTNLNSARSRIED
E. coli Nissle ----TGTKDPLKALDEAISSIDKFRSSLGAIQNRRLDSAVTNLNNTTTNLSEASRIQD
                :. . : : : * * : . : : * : . * : * * : * : : * * . . * * . * : * * : *

```

Appendix 7: FliC protein sequence alignment of *E. coli* H-types (MUSCLE). AAs are marked based on their similarity, with an asterisk for a perfect aligned, with a colon mark for a strong similarity or with a dot for a weak similarity.

H06	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H18	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H10	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H07	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H19	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H55	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H45	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H34	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H44	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H46	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H15	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H28	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H49	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H20	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H31	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H37	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H41	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H30	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H32	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H42	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H05	MAQVINTNSLSLITQNNINKNQSALSTSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H56	MAQVINTNSLSLITQNNINKNQSALSTSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H48	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H14	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H12	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H01	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
Nissle	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H09	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H52	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H33	MAQVINTNSLSLITQNNINKNQSALSTSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H39	MAQVINTNSLSLITQNNINKNQSALSTSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H51	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H43	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H26	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H04	MAQVINTNSLSLITQNNINKNQSALSTSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H17	MAQVINTNSLSLITQNNINKNQSALSTSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H29	MAQVINTNSLSLITQNNINKNQSALSTSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H38	MAQVINTNSLSLITQNNINKNQSALSTSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H36	MAQVINTNSLSLLTQNNLNKSQSSLSAIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H53	MAQVINTNSLSLLTQNNLNKSQSSLSAIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H27	MAQVINTNSLSLITQNNINKNQSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H11	MAQVINTNSLSLLTQNNLNKSQSSLSAIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H21	MAQVINTNSLSLLTQNNLNKSQSSLSAIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H23	MAQVINTNSLSLLTQNNLNKSQSSLSAIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H08	MAQVINTNSLSLLTQNNLNKSQSSLSAIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H40	MAQVINTNSLSLLTQNNLNKSQSSLSAIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H02	MAQVINTNSLSLLTQNNLNKSQSSLSAIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H35	MAQVINTNSLSLLTQNNLNKSQSSLSAIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H03	MAQVINTNSLSLLTQNNLNKSQSSLSAIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H16	MAQVINTNSLSLLTQNNLNKSQSSLSAIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H47	MAQVINTNSLSLLTQNNLNKSQSSLSAIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
H54	MAQVINTNSLSLLTQNNLNKSQSSLSAIERLSSGLRINSKDDAAGQAIANRFTSNIKG	60
	***** ***:**:*:*****:*****	

H06	DEIDRVSGQTQFNGVNVLA	KDGS	MKI	QV	GAND	QTI	TID	LK	IDS	DTL	GL	NG	FNV	NG	KGE	180
H18	DEIDRVSGQTQFNGVNVLS	KDGS	MKI	QV	GAND	GETI	TID	LK	IDS	DTL	LN	LAG	FNV	NG	E	180
H10	EEIDRVSSQTQFNGVNVLA	KDGM	MNI	QV	GAND	QTI	TID	LK	IDS	STL	LN	LS	F	D	N	180
H07	DEIDRVSGQTQFNGVNVLA	KDGS	MKI	QV	GAND	GETI	TID	LK	IDS	DTL	GL	NG	FNV	NG	KGT	180
H19	DEIDRVSGQTQFNGVNVLS	KDGS	MKI	QV	GAND	GETI	TID	LK	IDS	DTL	LN	LAG	FNV	NG	KGS	180
H55	DEIDRVSGQTQFNGVNVLA	KDGS	MKI	QV	GAND	GETI	TID	LK	IDS	STL	LN	LT	G	FNV	NG	180
H45	DEIDRVSGQTQFNGVNVLA	KDGS	MKI	QV	GAND	QTI	TID	LK	IDS	DTL	GL	NG	FNV	NG	KGT	180
H34	DEIDRVSGQTQFNGVNVLA	KDGS	MKI	QV	GAND	QTI	TID	LK	IDS	DTL	GL	S	G	FNV	NG	180
H44	DEIDRVSGQTQFNGVNVLA	KDGS	MKI	QV	GAND	QTI	TID	LK	IDS	STL	LN	LT	G	FNV	NG	180
H46	DEIDRVSGQTQFNGVNVLA	KDGS	MKI	QV	GAND	QTI	TID	LK	IDS	STL	KL	T	G	FNV	NG	180
H15	DEIDRVSGQTQFNGVNVLA	KDGS	MKI	QV	GAND	QTI	TID	LK	IDS	STL	LN	LT	G	FNV	NG	180
H28	DEIDRVSGQTQFNGVNVLA	KDGS	MKI	QV	GAND	QTI	TID	LK	IDS	DTL	GL	S	G	FNV	NG	180
H49	DEIDRVSGQTQFNGVNVLA	KDGS	MKI	QV	GAND	QTI	TID	LK	IDS	DTL	GL	NG	FNV	NG	KGT	180
H20	DEIDRVSGQTQFNGVNVLA	KDGS	MKI	QV	GAND	QTI	TID	LK	IDS	STL	KL	T	G	FNV	NG	180
H31	DEIDRVSGQTQFNGVNVLA	KDGS	MKI	QV	GAND	QTI	TID	LK	IDS	DTL	GL	S	G	FNV	NG	180
H37	DEIDRVSGQTQFNGVNVLA	KDGS	MKI	QV	GAND	QTI	TID	LK	IDS	STL	KL	T	G	FNV	NG	180
H41	DEIDRVSGQTQFNGVNVLA	KDGS	MKI	QV	GAND	QTI	TID	LK	IDS	DTL	GL	S	G	FNV	NG	180
H30	DEIDRVSGQTQFNGVNVLS	KDGS	MKI	QV	GAND	GETI	TID	LK	IDS	STL	KL	T	S	FNV	NG	180
H32	DEIDRVSGQTQFNGVNVLS	KDGS	MKI	QV	GAND	GETI	TID	LK	IDS	STL	KL	T	S	FNV	NG	180
H42	AEIDRVSGQTQFNGVNVLA	KNGS	LNI	QV	GAND	QTI	SID	LQ	IDS	S	AL	GL	S	G	F	180
H05	DEIDRVSGQTQFNGVNVLA	KDGS	MKI	QV	GAND	QTI	SID	LQ	IDS	STL	GL	NG	F	S	V	180
H56	DEIDRVSGQTQFNGVNVLA	KDGS	MKI	QV	GAND	QTI	SID	LQ	IDS	STL	GL	NG	F	S	V	180
H48	DEIDRVSGQTQFNGVNVLA	KNGS	MKI	QV	GAND	NQTI	SID	LQ	IDS	AKT	GL	D	G	F	S	180
H14	DEIDRVSGQTQFNGVNVLA	KDGS	MKI	QV	GAND	QTI	TID	LK	IDS	DTL	GL	S	G	FNV	NG	180
H12	DEIDRVSGQTQFNGVNVLA	KDGS	MKI	QV	GAND	QTI	TID	LK	IDS	DTL	GL	NG	FNV	NG	S	180
H01	DEIDRVSGQTQFNGVNVLA	KDGS	MKI	QV	GAND	QTI	TID	LK	IDS	DTL	GL	NG	FNV	NG	S	180
Nissle	DEIDRVSGQTQFNGVNVLA	KDGS	MKI	QV	GAND	QTI	TID	LK	IDS	DTL	GL	NG	FNV	NG	S	180
H09	DEIDRVSGQTQFNGVNVLS	KDGS	MKI	QV	GAND	GETI	TID	LK	IDS	DTL	LN	LAG	FNV	NG	KGS	180
H52	DEIDRVSGQTQFNGVNVLA	KDGS	MKI	QV	GAND	QTI	SID	LQ	IDS	STL	GL	K	G	F	S	180
H33	DEIDRVSGQTQFNGVNVLA	KNGS	MAI	QV	GAND	QTI	IN	D	LQ	IDS	STL	GL	G	G	F	180
H39	DEIDRVSGQTQFNGVNVLA	KNGS	MAI	QV	GAND	QTI	SID	LQ	IDS	STL	GL	S	G	F	S	180
H51	SEIDRVSGQTQFNGVNVLS	KDGS	LKI	QV	GAND	QTI	SID	LK	IDS	DTL	GL	NG	FNV	NG	S	180
H43	EEIDRVSGQTQFNGVNVLA	KDGM	TKI	QV	GAND	QTI	SID	LK	IDS	STL	GL	T	G	F	V	180
H26	DEIDRVSGQTQFNGVNVLS	KDGS	MKI	QV	GAND	GETI	TID	LK	IDS	DTL	LN	LAG	FNV	NG	AGS	180
H04	DEIDRVSGQTQFNGVNVLS	KNDS	MKI	QI	GAND	NQTI	SIG	LQ	IDS	STL	LN	L	K	G	F	180
H17	DEIDRVSGQTQFNGVNVLS	KNDS	MKI	QI	GAND	NQTI	SIG	LQ	IDS	STL	LN	L	K	G	F	180
H29	DEIDRVSGQTQFNGVNVLA	KDNT	MKI	QV	GAND	QTI	SID	LQ	IDS	STL	GL	NG	F	S	V	180
H38	DEIDRVSGQTQFNGVNVLA	KDGS	MKI	QV	GAND	QTI	SID	LQ	IDS	STL	GL	NG	F	S	V	180
H36	EEINRVSGQTQFNGVNVLA	SNS	MTI	QV	GAND	GEAIT	IDL	KEI	TAET	L	GL	T	G	FNV	NG	180
H53	AEIDRVSGQTQFNGVNVLA	SNS	LSI	QV	GAND	GEKID	IDL	KEI	DT	G	T	L	GL	AN	F	180
H27	EEIDRVSEQTQFNGVNVLA	ENNE	MKI	QV	GAND	GETI	TIN	LAKI	DAKT	L	GL	D	G	F	N	180
H11	EEIDRVSEQTQFNGVNVLA	ENNE	MKI	QV	GAND	GETI	TIN	LAKI	DAKT	L	GL	D	G	F	N	180
H21	EEIDRVSEQTQFNGVNVLA	ENNE	MKI	QV	GAND	GETI	TIN	LAKI	DAKT	L	GL	D	G	F	N	180
H23	EEIDRVSEQTQFNGVNVLA	ENNE	MKI	QV	GAND	GETI	TIN	LAKI	DAKT	L	GL	D	G	F	N	180
H08	EEIDRVSEQTQFNGVNVLA	ENNE	MKI	QV	GAND	GETI	TIN	LAKI	DAKT	L	GL	D	G	F	N	180
H40	EEIDRVSEQTQFNGVNVLA	ENNE	MKI	QV	GAND	GEAIT	IN	LAKI	DAKT	L	GL	D	G	F	N	180
H02	SEIDRVSGQTQFNGVNVLA	SQDM	TI	QV	GAND	GETI	TIK	LQE	INS	DTL	GL	S	G	F	I	180
H35	QEIDRVSNQTQFNGVNVLA	SQDM	TKI	QV	GAND	GETI	TID	LKE	INS	KT	L	GL	D	K	L	180
H03	SEIDRVSQQTQFNGVNVLA	TNQ	TKI	QV	GAND	QTI	EI	GL	KI	D	AD	T	L	GL	K	180
H16	SEIDRVSNQTQFNGVNVLA	SQDM	TKI	QV	GAND	GETI	EI	AL	D	KI	DAKT	L	GL	D	N	180
H47	DEIDRVSQQTEFNGVNVLA	GDK	TLKI	QV	GAND	NQSI	DIN	LK	IDS	TV	L	K	R	D	L	180
H54	NEIDRVSGQTQFNGVNVLA	SKNT	LTI	QV	GAND	GETI	DIN	LKE	INS	Q	T	L	GL	D	K	180

H06	TANTAATLKDMSGFTAAAAP-----GGTVGVYQYTDKS----AVASSVDILN	223
H18	TANTAATLKDMMVGLKLDNTG-----VTTAGVNRVIADK----AVASSTDILN	223
H10	SVKDGATINKQVAVGAGDFK-----DKASGSLG	208
H07	ITNKAATVSDLTSAGAKLNTTGT-----LYD-----LKTENTLLTTDAAF	221
H19	VANTAATDNLTLAGFTAGT-----KAAD--GTVT----YS--KNVQFAAATASNVL	225
H55	VSNTAATDNLTLAGFTAGT-----TPAAD--GTVT----YS--KDVDNAKAAAASNVL	226
H45	IANKAATVSDDLTAAGATGTG-----P-----YA--VTTNNTALSASDALS	218
H34	KASVAATADGMVKDGYIKGL-----TSSDG----STAYT--KTTANTAAKGSDDILA	225
H44	VANKAATKADLTAAQLTTAAGGTIAAPAADANGVTK----YT--VSAGLNESTVADVFA	234
H46	VANTAATKDELAAAAAAGT-----TPAVGTDGVTK----YT--VDAGLNKATAANVFA	226
H15	VANTAATKADLTAAQLSA-----PGAADANGTVT----YT--VSAGYKESTAADVIA	226
H28	VANTAATKSDLAQAQLLA-----PGTADANGTVT----YT--VSAGLKTSTAADVIA	226
H49	IANKAATISDLAATGANVTN-----S-----SNIVVTTKFNALDAATAFS	220
H20	VANTAATKADLAAAAIGTPG-----AA---DSTGAIAYT--VSAGLTKTTAADVLS	226
H31	VANTAATKDDDLVAASVSAA-----VGNEYT--VSAGLSKSTAADVIA	220
H37	VDNAKATDANLTTAGFTQGV-----VDSNGNST---WTKSTTTNFDAATAVNVLA	227
H41	VANAKATEADLTGAGFSQGA-----VDTNGNST---WTKSTTTNYSAAATADLLS	227
H30	VDNAKATEADLTGAGFSQGA-----VV-SGNST---WTKSTVTFNAATATDVLA	226
H32	VANAKATEADLTGAGFSQSA-----VV-SGNST---WTKSTVTFNAATATDVLA	226
H42	KLSDTVTQVG-DGSA-APVK-----VDLDAAT-----DI---GTALG	213
H05	NVSDSITQITGAAGT-KPVG-----VDFTAVAK-----DL---TTATG	214
H56	NVSDSITQITGAAGT-KPVG-----VDFTAVAK-----DL---TTATG	214
H48	VTSAPVT---AFGATTN-----NIKLT-----	201
H14	VANTAASKADLVAAANATVVG-----N---KYT-----VSAGYDAAKASDLLA	219
H12	IANKAATISDLTAAKMDAAT-----NTIT-----TTNNALTASKALD	217
H01	IANKAATISDLTAAKMDAAT-----NTIT-----TTNNALTASKALD	217
Nissle	IANKAATISDLTAAKMDAAT-----NTIT-----TTNNALTASKALD	217
H09	VANTAATSDLLKLAGFTKGT-----TDTNGVTA---YT--NTISNDKAKASDLLA	225
H52	KVSDAITTVPGANAGDAPVT-----VKFGAN-----DTA-----AAAMAKTLGI	219
H33	KLSDSITQVVGASGS-LADVK-----L-----SSVASALGV	209
H39	KLSDSITQVVGASGS-LADVK-----L-----SAVATKLG	210
H51	IANKAATISDLTAQKAVD-----NNGTYKVTTSNAA-----LTASQALS	220
H43	ISTAVTGAATTTYADSAVA-----IDIGTDSGIAADAALGTINFDNNT-----	225
H26	VDNAKATGKDLTDAGFTASA-----ADANGKIT---YTKDVTTKFDKATAADVLG	227
H04	FSAAKLTAA-----	189
H17	FSAAKLTAA-----	189
H29	ETSEAITQLPN-----G-----ENAP-----IAVKMDASVLTDLNI	211
H38	SVGDAITQLPG-----ETA-----ADAP-----VTIKFDDSVKTDLKL	213
H36	VNNTVATAKD-----LTDKG-----FISTDNGKTYT--GSAGLANAKAGDVFG	221
H53	STSTKTVANG-----GDIVLS	196
H27	ATGSDLISKFKATGT--D-----NYQIN-----GTDNYT	207
H11	ATGSDLISKFKATGT--D-----NYDVG-----G-DAYT	206
H21	ATGSDLISKFKATGT--D-----NYQIN-----GTDNYT	207
H23	ATGSDLISKFKATGT--D-----NYQIN-----GTDNYT	207
H08	ATGSDLISKFKATGT--D-----NYDVG-----G-KTYT	206
H40	ATGSDLISKFKATGT--D-----NYDVG-----G-DAYT	206
H02	LKAA-----T-----AETTYFG	192
H35	TSDLAATATTELAPAKTDVK-----NLFTTDGAT-----APTFLK-	215
H03	PTSGAVALKSEMSPTLTSV-----NATTGKN-----GTNYAFG	213
H16	PMSSAVALKSEAPDLTKV-----NATDGSV-----GGAKAFG	213
H47	-----Q-----YKDGAT-----T-----	187
H54	PSGENVKVDTTTTTT-----YTDGTA-----IKNKA--	207

H06	AVAGADGNK-----V-TTSADVFGT--PAAAVTYTYNKDINSYSAASDD-	265
H18	AVAGVDGSK-----V-STEADVFGAAAPGTPVEYTYHKDNTNTYTAS-AS-	266
H10	TLK-----LVEKDGKYYVNDTKSSKYYDAEVD	235
H07	KLG--NGDK-----VTVGGVDYTYNAKSGDFTTTKSTA	252
H19	AAK--DGDE-----ITF-AGNNGTG--IAATGGTYTYHKDSNSYSFSATAA	266
H55	AAK--NGDT-----ISF-AGNNGTG--ITATAGTYTYNKASDSYSFSATAA	267
H45	RLK--TGDT-----VTT-----TGSSAAIYTYDAAKGNFTTQATVA	252
H34	ALK--TGDK-----ITATGA-NSLA--DNATSTTYTYNATSNTFSYTADGV	266
H44	GLG--DTA-----VVNANITSGFNAV---TGNNYTYHKDNTNDFTFNATIA	274
H46	NLA--DGA-----VVDASISNGFGAA---AATDYTYNKATNDFTFNASIA	266
H15	SIK--DGSAPT-----SAITATINNGFGDSSALTSNDYTYDPAKGDFTYDVASS	273
H28	SLA--NNA-----KVNATIANGFGSP---TATDYTYNSATGDFTYSATIA	266
H49	KLK--DGDS-----VAV-----AAQKYTYNASTNDFTFTEN--T	249
H20	SLA--DGTT-----ITATGVKNGFA--AGATSNAKLNKDNNTFTYDT--T	266
H31	SLT--DGAT-----VTAAGVSNFGA--AGATGNAYKFNQANNTFTYNT--T	260
H37	AVK--DGST-----I-NYTGTTNGL--GIAATSAYTYH--DSTKSYTFDST	266
H41	TIK--DGST-----V-TYAGTDTGL--GVAAAGNYTYD--ANSKYSFNAN	266
H30	SVS--GGST-----ISGYTGTNGL--GVAASTAYTYN--ATSKYSFSDAT	266
H32	SVS--GGST-----ISGYTGTNGL--GVAASTAYTYN--ATSKYSFSDAT	266
H42	QKV--NASS-----LTLHNLDKDG---AATENYVV--SYGSDNYAA---	248
H05	KTV--DVSS-----LTLHNTLDAKG---AATSQFVV--QSGNDFYSA---	249
H56	KTV--DVSS-----LTLHNTLDAKG---AATAQFVV--QSGSDFYSA---	249
H48	-----GIT-----LSTEAATDTGGTNPASIEGYVT---DNGNDYYAKITG	238
H14	GVS--DGDT-----VQA-TINNGFG--TAASATNYKYD--SASKYSFDTT	258
H12	QLK--DGDT-----VTI-----KAD--AAQTATVYTYNASAGNFSFSN--V	252
H01	QLK--DGDT-----VTI-----KAD--AAQTATVYTYNASAGNFSFSN--V	252
Nissle	QLK--DGDT-----VTI-----KAD--AAQTATVYTYNASAGNFSFSN--V	252
H09	NIT--DGSV-----ITGGGAN--AF--GVAANKGYTYD--AASKYSFAAD	263
H52	SDT--SGLS-----LH--NVQSADG--K--ATGTYVV--QSGNDFYSA---	252
H33	-DA--STLT-----LH--NVQTPAG--A--ATANYVV--SSGSDNYSV---	241
H39	-NA--STLS-----LH--EVQDSAG--D--GTGTFVV--SSGSDNYAV---	242
H51	KLS--DGDT-----VD--IATYAG--GTSSTVSYKYDADAGNFSYNN--T	257
H43	-----GK-----YYAQITS-AA-NPGLDGAYEIHVNDADGSFTVAAS	260
H26	KAA--AGDS-----ITYAGTDTGLG--VAADASTYTY--NAANKSYTFDAT	267
H04	-DG--TAIA-----AA--DVKDAGG--K--QVNL-----	209
H17	-DG--TAIA-----AA--DVKDAGG--K--QVNL-----	209
H29	TDA--SAVS-----LH--NVTK--GG--V--ATSTYVV--QYGDKSYA----	242
H38	TDA--SGLS-----LH--NLKDENG--N--LTNQYVV--QNGGKSYA----	245
H36	KMVD-TGTVTT-----TIDNGF--GTAQSNYKYDKASNSFSFDIDDA	261
H53	SKTI-KAEIQI----DSHSADPK---AADGLY--ALKDGTGYAVKDKDGAYHAAVKNA	244
H27	VNVD-SGAVQN-----E-----DGDAIFVSAADGSLTTKSDTK	239
H11	VNVD-SGAVKD-----T-----TGNDIFVSAADGSLTTKSDTK	238
H21	VNVD-SGVVQD-----K-----DGKQVYVSAADGSLTTSSDTQ	239
H23	VNVD-SGVVQD-----K-----DGKQVYVSAADGSLTTSSDTQ	239
H08	VNVE-SGAVKN-----D-----ANKDVVSAADGSLTTSSDTK	238
H40	VNVD-SGAVKD-----K-----DNKDVVSAADGSLTTSSDTK	238
H02	STVK-LADANTLDA-----DITATVKGTT---TPGQRDGNIMSDANGKLYVVKVAGS	239
H35	-YVD-NTD-----P-----DKLFVSNNGTNYAASYDK	241
H03	ATFR-TDDVNT--YFGKAVNTGD-KTNVLG-T---EA---EVFQIQVEGQTYFVAQNGD	262
H16	SNYK-NADVET--YFGTGNVQDTKDTTDDATGTA---GT---KVYQVQVEGQTYFVQDNN	264
H47	--LK--GDIKSTSVQKFDTT-----E-AAISPKDKGKYYAVVSKS	224
H54	-ALA--TDVNNASSIGVSDAIPG-----D-IKFNETSGKYYVAITST	245

H06	-----ISSANLAAFLNPQAGDTTKATVTIGG---KDQDV-NIDKSGNLTA	306
H18	-----VDATQLAAFLNPEAGGTTAATVSIENGTTAQEQKV-IIAKDGLSLTA	311
H10	-----TSKGGKINF	243
H07	GTGVDAQAQATDSAKKRDLAATLHADVGVKSVNGSYTTKD---GTVSF-ETDSAGNITI	307
H19	S-----KDSLLSTLAPNAGDFTAKVTIGS---KSQEV-NVSKDGTITS	306
H55	S-----KDSLLSMLAPNAGDSFTASVSIIGG---KAQDV-NVSKDGTITT	307
H45	-----DGDVVFANFLKPAAGTTASGVYTRST---GDVKF-DVDANGDVTI	294
H34	N-----QTNAANLIPAAKTTAASVTIGG---TAQNV-NIDDSGNITS	306
H44	A---GDA---TTPSNSAKLQSLTPKAGDTAHLNVKIGA---TSVDV-VLSSDGKITA	322
H46	A---GAA---AGDSNSAALQSFLLTPKAGDTANLSVKIGT---TSVNV-VLASDGKITA	314
H15	-----ANNTAAQVQSFLTPKAGDTANLKVTVGT---TSVDV-VLASDGKITA	316
H28	A---GTN---SGDSNSAALQSFLLTPKAGDTANLNVKIGS---TSIDV-VLASDGKITA	314
H49	V-----ATGTATDLGATLKAAGQSQSGTYTFAN---GKVNFDVDASGNITI	294
H20	-----ATTAELQSYLTPKAGDTATFSVEIGG---TTQDV-VLSSDGKITA	307
H31	-----STAAELQSYLTPKAGDTATFSVEIGS---TKQDV-VLASDGKITA	301
H37	GAAVAG-----AASSLQGTG---GTDNTAKITIDG---SAQEV-NIAKDGKITD	309
H41	GLTGAN-----TATALKGYL---GTGANTAKISIGG---TEQEV-NIAKDGKITD	309
H30	ALTNGDGT---GATTKVADVVKAYA---ANGDNQAQISIGG---SAQDV-KIASDGLTLD	316
H32	ALTNGDGT---AGSTKVADVVKAYA---ANGDNQAQISIGG---SAQEV-KIASDGLTLD	316
H42	-----S-VA-DDGTVTL	258
H05	-----S-INHTDGKVTL	260
H56	-----S-IDHASGEVTL	260
H48	-----GDNDGKY-----YAV-TVANDGTVTM	258
H14	---TASA-----ADVQKYLTPGVGD TAKGTITIDG---SAQDV-QISSDGKITS	300
H12	SNNTSAKA-----GDVAASLLPPAQQTASGVYKAAS---GEVNF-DVDANGKITI	298
H01	SNNTSAKA-----GDVAASLLPPAQQTASGVYKAAS---GEVNF-DVDANGKITI	298
Nissle	SNNTSAKA-----GDVAASLLPPAQQTASGVYKAAS---GEVNF-DVDANGKITI	298
H09	G---ADS-----AKTLSIINPNTGDSSQATVTIGG---KEQKV-NISQDGKITA	305
H52	-----S-VN-AGGVVTL	262
H33	-----S-VEDSSGTVTL	252
H39	-----S-VDAASGAVNL	253
H51	ANKTSAAA-----GTLADTLLPAAGQTKTGTYKAAT---GDVNF-NVDATGNLTI	303
H43	DKQAGAAP-----GTALT-----SGKVQ-----TA-TTTPGTAVDV	290
H26	GV-AKADA-----GTALKGYLG---ASNTGKINIGG---TEQEV-NIAKDGKITD	309
H04	-----	209
H17	-----	209
H29	-----A-SVDAGGTVKL	253
H38	-----A-TVAANGVTL	256
H36	AGTTA-----PQVATY---LNPTANDKTKASVEING---SSQAVI-IDHNGKMTA	304
H53	DGK-----VTWDS---SDTSVT-----AT	260
H27	VGGTGIDA-----TGLAKAAVSLAKDASIKYQG-ITFTN---KGTGAFDGSNGTLIA	288
H11	IAGTGIDA-----TALAAAAKNKAQNDKFTFNG-VEFTT---TT--AADGNGNGVISA	285
H21	FK---IDA-----TKLAVAAKDLAQGNKIVYEG-IEFTN---TGTGAIPATGNGKLT	285
H23	FK---IDA-----TKLAVAAKDLAQGNKIVYEG-IEFTN---TGTGAIPATGNGKLT	285
H08	VSGESIDA-----TELAKLAIKLADKGSIEYKG-ITFTN---NTGAELDANGKGVLT	287
H40	VDGKSIDA-----TELAKLAINLADQKSIDYKG-ITFTN---KSGTAFDANGKGVLT	287
H02	D-----KPAENGYE-VTVED---DPTSPD---AGKLLGL	267
H35	D-----TNSIKFNSTGTGTGTAGTDFIDA---TGKDVH---VGGGRVMA	280
H03	T-----ATNAYSLKQSGSGYEK-VTVDS---KAVQIA---NFGGRVTA	299
H16	T-----NTNGFTLLKQNSTGYEK-VQVGG---KDVQLA---NFGGRVTA	301
H47	T-----TTD-----NNGIYA-ASVDS---DGNVTI---DASKKVTV	253
H54	E-----HND-----KNGDYE-ITVDK---DGSAAAL---VAGQSSPK	274

H06	ADD-----GAVLYMDATGNLTKNNAG---GDTQATLAKLATATGAKAATIQT	351
H18	ADD-----GAALYLDGDTGNLSKTNAG---TDTQAKLSDLMANNANAKTVITTD	356
H10	NST-----NESGT---TPTAA-----	256
H07	GGG-----QAYVDDAGNLTNNAGS---AAKADMKALLKAASEGS-----D	345
H19	SDG-----KALYLDEKGNLTQTGS---GTTKAATWDLNLMANTDITGKDAY-G	349
H55	TDG-----KSLYLDQKGNLTQTGS---GTTIAATWDLNLMANTDITGVDATSG	351
H45	GGK-----AAYLDATGNLSTNNPGI---ASSAKLSDLFASGSTLA-----T	332
H34	SDG-----DQLYLDSTGNLTKNQ---GNPKKATVSGLLGNTDAKGT---V	347
H44	KDG-----SKLFIGIDGNLTQNSAGA---GIKPATLDALNTQNT---TTPAAAV	364
H46	KDG-----SALYIDSTGNLTQNSAGT---V-TAATLDGLTKNHDATGAVG---	355
H15	KDG-----SALYIDSTGNLTQNSAGL---T-SAKLA-TLT-----GLQSGV	353
H28	KDG-----SELFIDVDGNLTQNNAGT---V-KAATLDALTKNWHHTGTP-GAV	357
H49	GGG-----KAFLV-GGALTTNDP---TGSTPATMSSLFKAADKDA-----	332
H20	KDG-----SKLYIDTTGNLTQNGGNGVGTLAEATLSGLALNKNGLT-A---V	351
H31	KDG-----SKLYIDTTGNLTQNGG---GTLEEATLNGLAFNHSGPAAA---V	342
H37	TDG-----KALYIDSTGNLTKNG---S-DTLTQATLNDVLTGANS--VDD---	348
H41	TNG-----DALYLDITGNLTKNY---A-GSPPAATLDNLVLA---TVN---	346
H30	VNG-----DALYIGSDGNLTKNQ---A-GGPAATLDGIFNGANGNAAVD---	357
H32	TNG-----DALYIGADGNLTKNQ---A-GGPAATLDGIFNGANGHDAVD---	357
H42	NKT-----DITYSGG---DITGATKD--DTLTKV-----	282
H05	NKA-----DVEYTDNDGLTAAATQK--DQLIKV-----	287
H56	NKA-----DVEYKTDNDGLTAAATQK--DQLIKV-----	287
H48	ATG-----ATANATVTDANTTKA-----	276
H14	SNG-----DKLYIDTTGRLTKNGFS---ASLTEASLSTLAANNTKA-----	338
H12	GG-----QKAYLTSNLDGNLTTND--A-GGATAATLDGLFKKAGDQSIGFKK	341
H01	GG-----QEAYLTSNLDGNLTTND--A-GGATAATLDGLFKKAGDQSIGFKN	341
Nissle	GG-----QEAYLTSNLDGNLTTND--A-GGATAATLDGLFKKAGDQSIGFKN	341
H09	ADD-----NATLYLDKQGNLTKTNA---GNDTAATWDGLISNSDSTGAVPVGV	350
H52	NTT-----NVTFTDPANGVTTAT-QT--GQPIKV-----	288
H33	NTT-----DIGYTDTPANGVTTGS-MT--GKYVKV-----	278
H39	NTT-----DVTYDDATNGVTGAT-QN--GQLIKV-----	279
H51	G-G-----QQAYLTTDGNLTTNN--S-GGAATATLKEFLTLAGDGKSLGNGG	346
H43	TAA-----KTAL--A-----AAGADT-----	304
H26	TNG-----DALYLDSTGNLTKNTANL--GAADKATVDKLFAGAQD-----	347
H04	-----LSYTDTASNS-----	219
H17	-----LSYTDTASNS-----	219
H29	NKA-----DVTYNDAAANGVKN-ATQI--G-----SLV-----	277
H38	NKA-----NVTYSVAVANGIDT-ATQS--G-----QLV-----	280
H36	ADD-----NAELFIDNSGNLTKNNKTG---GKAATLENLALNKTGTNTA----	345
H53	DVDQTT--KVTE-----FQEVYKK-----VAANTSGLAA-----	287
H27	NIDGK-----DVTFTID---A-----TGKDAT-----	307
H11	EIDGK-----SVTFTVT---DA--D-----KKAS-----	304
H21	NVDGK-----AVEFTIS---GS--AD-----TSGTSAT-----	308
H23	NVDGK-----AVEFTIS---GS--AD-----TSGTSAT-----	308
H08	NIDGQ-----DVQFTID---SN--AP-----TGAGAT-----	309
H40	NIDGQ-----DVKFTIN---ST--AA-----TGADAT-----	309
H02	AL-----AGTQPQAGN---LKEVT-TVK---GK--GAIDVQLGTDATA-----	302
H35	ANDIKGRITDDNVAAKPQLFLDQSGKYHI-----SKDNGAT-----	315
H03	FVDDGTAAHNALS-----V-----DLQKGTVGKALS-----	325
H16	FVEDNGS---ATS-----V-----DLAAGKMGKALA-----	324
H47	DL-----	255
H54	SLEDVGATKNVTAY--QVANTQSNTQSVDATVS---AG--AISELKTGGVDNTA-----	321

H06	KGTFTSDGTAFDGASMS-----	368
H18	KGTFTANTTKFDGVDIS-----	373
H10	-TEVTTVGRDVK-----	267
H07	GASLTFNGTEYTIKATPATTSP-----VAPLIPGGITYQATV-----	383
H19	NSAAAAVGTVIEAKGMTITSAGGNA-----QVLKDAAYNAAYATSITGTPGDAGA--	400
H55	HSAATAVKTIKVKDMTITSAGGNA-----QVATDKAYNDKYAARIVAGDTAAQAD--	402
H45	TGSIQLSGTTYN-FGAAATS-----GVTYTKTV-----	359
H34	KTTIKTE-----AGVTVTAE-----	362
H44	PVTI-----	368
H46	-VDI-----	358
H15	ASTI-----	357
H28	STVI-----	361
H49	QSSIDFGGKYE-FAGGNSTNGG-----GVKFKDT-----	361
H20	KST-----	354
H31	QST-----	345
H37	-TRIDFD-----SGMSVTLKVNSTVD-----ITGASISAA-----AMTN-	382
H41	-ATIKFD-----SGMTVDYTAG-TGAN-----ITGASISAD-----DMAA-	379
H30	-AKITFG-----SGMTVDFTQASKKVD-----IKGATVSAE-----DMDT-	391
H32	-AKITFG-----SGMTVDFTQVSNVND-----IKGATVSAE-----DMNT-	391
H42	-----	282
H05	-----	287
H56	-----	287
H48	-TTITSGGTPVQ-----IDNTAGSATA-----NLGAVSLVKLQ--DSKGNSTD-----	316
H14	-TTIDIGGTSIS-FT-----GNSTT-----	356
H12	TASVTMGGTTYN-FKGTADADAATA-----NAGVSFTDTAS-----KETVLNK	383
H01	TASVTMGGTTYN-FKGTADAGAATA-----NAGVSFTDTAS-----KETVLNK	383
Nissle	TASVTMGGTTYN-FKGTADAGAATA-----NAGVSFTDTAS-----KETVLNK	383
H09	ATTITITSGTA--SGMSV-QSAGAGIQTSTNSQILAGGAFAAKVS--IEGGAATDILVAS	405
H52	-----	288
H33	-----	278
H39	-----	279
H51	TATVTLDNNTYN-FKAAANVTDGAGVIA-----AAGVYATATV-----SKDVILAQ	391
H43	-SGLKLV-----QLSNTDSAGKVT-----NVGY-----	326
H26	-ATITFDS-----GMTAKFDQTAGTVD-----FKGASIS-----ADAMAS-	381
H04	-----	219
H17	-----	219
H29	-----QVGADAN-----	284
H38	-----QVGADST-----	287
H36	---V-----	346
H53	-----	287
H27	---L-----	308
H11	---L-----	305
H21	---V-----	309
H23	---V-----	309
H08	---I-----	310
H40	---I-----	310
H02	-----	302
H35	---Y-----	316
H03	---F-----	326
H16	---Y-----	325
H47	-----	255
H54	-----	321

H06	-----IDTNTFANAVK-----	379
H18	-----VDASTFANAVK-----	384
H10	-----LDASALKANQ-----S	278
H07	-----SKDVVLSETKAAA-----ATSS	400
H19	AGAAATAGNAAVG-----ALGATAVDNNTADVADISISASQMASILQD-----KD	445
H55	-----TAADTAEA-----TATTTAVANTTADVSGVTISASQMASILKD-----KD	442
H45	-----SADTVLSTVQSAATANTAVTGAT	382
H34	-----GNTGTVKIEGATVSASAF-----	380
H44	-----TTEDKTEIKLAGATVA-----GQSGAIVVTGARISAEAMQSATKT-----	408
H46	-----TTADGATISLAGSANAA-TGTQSGAITLKNVRI SADALQSAAKG-----	401
H15	-----TTEDGTNI-----DI-----AANGNIGLTGVRI SADSLQSATKS-----	391
H28	-----TTEDETTFTLAGGTD-----TTSGAITVANARMSAESLQSATKS-----	401
H49	-----VSSDALLAQVKADSTAN-----NVK	381
H20	-----ITTADNTSIVLNGSSDGTGNAGTEGTIAVTGAVISSAALQSASK-----T	399
H31	-----ITTADGTSIVLAGSGD-FGTTKTAGAINVTGAVISADALLSASK-----A	389
H37	-----E	383
H41	-----K	380
H30	-----A	392
H32	-----A	392
H42	-----	282
H05	-----	287
H56	-----	287
H48	-----	316
H14	-----	356
H12	-----VATAKQGGKAVAADGDTSA-----T	402
H01	-----VATAKQGTAVAANGDTSA-----T	402
Nissle	-----VATAKQGTAVAANGDTSA-----T	402
H09	NGNITAADGGSALYLDATTGGFTTTAGGNTAAS-----LDNLIANSK-----DAT	449
H52	-----	288
H33	-----	278
H39	-----	279
H51	-----LQSASQAAATATDGDIVA-----T	410
H43	-----	326
H26	-----T	382
H04	-----	219
H17	-----	219
H29	-----	284
H38	-----	287
H36	KSSITTESGKIAIAGSTDGTT-----AGKISATNVKISAEDL-----	384
H53	-----NQTFKSY-----	294
H27	-----KT-SDPVY-----	315
H11	-----IT-SETVY-----	312
H21	-----AP-TTALY-----	316
H23	-----AP-TTALY-----	316
H08	-----TT-DTAVY-----	317
H40	-----TT-DIDVY-----	317
H02	-----S-----ITGAKL-----FKLEDA-----	315
H35	-----ANATVNAT-----TGEVTY-----DSGT-AAA-----	337
H03	-----NDSQMSVY-----VDGKNL-----EIKQVLDA-----	348
H16	-----NDAPMSVY-----FGGKNL-----DVHQVQDT-----	347
H47	-----ADA-----	258
H54	-----AGNAKL-----VKMTYTDS-----	335

H06	-----NDTYTATVGA-	389
H18	-----NETYTATVGV-	394
H10	L-----VVY-----K-----	283
H07	ITFNSGVLSKTI-----GF-----	414
H19	FT-----LS--DGSDTYNVTSNAV	462
H55	FA-----LN--SGTTQYAVTGSTG	459
H45	IKYNTGIQ-----SATASFGG-	398
H34	-----TGIAY-----SAN--TGGNTYAVAANN-	400
H44	TGFTTGTTVAA-----NTGKVT-----	426
H46	TVINVDNGADDI-----SVSKTG--VV-----TT---	423
H15	TGFTVGTGATGL-----TVGTDG--K-----	410
H28	TGFTVDVGATGN-----SAGDIK--VD-----SKGI-	425
H49	ITFNNGPL-----SFTASFQ--	396
H20	TGFTVGTVDI-----AGYIS--VGTDGSVQA----	423
H31	TGFTSGAY-----T--VGTGTVK-----	406
H37	-----LTGKAY-----TVV--NGAESYAVAT-N-	403
H41	-----LSGKAY-----TVA--NGAESYDVAAVT-	401
H30	-----LTGQAY-----TVA--NGAQSFVAAA-G-	412
H32	-----LTGQAY-----TVA--NGAQSYDAAA-D-	412
H42	---A-----ANSDGEAVGFAT--VQGKNEYITDGV-	307
H05	---A-----ADSDGSAAGYVT--FQKKNYATTVST-	312
H56	---A-----ADSDGAAAGYVT--FQKKNYATTAPA-	312
H48	-----TY-----ALKD--TNGNLYAADVNE-	334
H14	-----PNTITYSVTGAKVDQAAFDKAVST-----SGNDVD--FTTAGYSVDG---	396
H12	ITYKSGVQTYQA-----VFAA--GDGTASAKYAD--	429
H01	ITYKSGVQTYQA-----VFAA--GDGTASAKYAD--	429
Nissle	ITYKSGVQTYQA-----VFAA--GDGTASAKYAD--	429
H09	LTVTSGTQNTVYSTTGSQAQFTSLAKVDTVNVTNAHVSAEGMAN--LTKSNFTIDMGG-	506
H52	---TTNSAGA-----AVGYVT--IQGKDYLAGADG-	313
H33	---GADALGA-----AVGYVT--VQGNFKADAGA-	303
H39	---TSDANGA-----AVGYVT--IQGKNYQAGATG-	304
H51	INYKSGVMIG-----SATFTN--GKGTADGMTSGT-	438
H43	-----GLQN--DSGTIFATDYDG-	342
H26	LN-----N-----GSYTAN--VGGKAYAVTAG--	402
H04	-----T-----KYAVVDS--	227
H17	-----T-----KYAVVDS--	227
H29	-N-----D-----AVGFVT--VQKKNYVANDS--	303
H38	-G-----T-----PKAFVS--VQKKSFGIDDA--	306
H36	-----KAKADTAGFTTS-----TGFTVAAGG-	405
H53	-----KKDNGELGYVVENADGTFNRANV-	317
H27	-----KNSAGQFTT-----TKV-	327
H11	-----KNSAGLYTT-----TKV-	324
H21	-----KNSAGQLTA-----TKV-	328
H23	-----KNSAGQLTA-----TKV-	328
H08	-----KNSAGQFTT-----TKV-	329
H40	-----KNSAGQFTT-----TKV-	329
H02	-----NGK-DTGSFAL--IGDDGKQYAANV-	337
H35	-----LGAGIEVGSALKEIAKPDAGVAVDL-	362
H03	-----DGKPKAGAFAAQT--ADGKSLAVNI-	371
H16	-----QGNPVPNSFAAKT--SDGTYIAVNV-	370
H47	-----	258
H54	-----NGKKVEGGYALKV--GDDYYAADY--	357

H06	-----KT---YSVTTGSAAA-----D-TAYMSNGVL-----SDTPPTY	418
H18	-----TLPATYTVNN--GTA-----A-SAYLVDGKV-----SKTPAEY	424
H10	-----DKSGNDAYIIQTKDVTNQST-----F	305
H07	-----TAGESSDAAKSYVDDKGGITNVADYT-VSY-SVNKDNGSVTVA--GYA---S	459
H19	TI-----NGKAANID-----DSGAI TDQ--TSKVVNY	487
H55	AVTYDPDTPAATGDIVSAYVD-----DAGTLT TD--ANKTVKY	496
H45	-----VNTNGAGNSNDTYTDADKELTTASYT-INY-NVDKDTGTVTVA--S-----N	442
H34	-----TTNGFLA-----GDDLTDQ--AQTVSTY	421
H44	-----IGGNQAYTQTDGTLA-----AKNETEI	448
H46	-----GGAPTYTDADGKLT-----TTNTVDY	444
H15	-----VTIGGTTAQSYTSKDGSLT-----TDNTTKL	436
H28	-----VQYTGTVFEDAYTKADGSLT-----TDNTTNL	453
H49	-----NGVSGSAASNAAYIDSEGLTTTESY-NTNY-SVDKDTGAVSV-----T	438
H20	-----YDAATSGNKASYTNTD-GTL-----T--TDNTTKL	450
H31	-----SGGNDVYNKADGTL-----T--TDNTTKY	429
H37	-----NTVKTADAKNVYVDASGKLTDDKATV-----TETY	435
H41	-----GAVTTAGNSPVYADADGKLTTSASNTV-----TQTY	433
H30	-----GAVTATTGGATVNI GADGELTTATNKTV-----TETY	444
H32	-----GAVTATTGGATVNI GAEGELTTAANKTV-----TETY	444
H42	-----KNQS--TAA--PTDIA-----	319
H05	-----ALDDNTAAK--ATDNK-----	326
H56	-----ALNDDTTAT--ATANK-----	326
H48	-----TT-GAVSVKTIITYTDS SGAASS-----PTAV-KLGGDDGKTEVV--DIDGTY	378
H14	-----ATGAVTKGVAPVYIDNNGAL TTS-----DTVDF	424
H12	-----KADVSNATATYTDADGEMTTIGSY-TTKY-SIDANNGKVTV-----D	469
H01	-----NTDVSNATATYTDADGEMTTIGSY-TTKY-SIDANNGKVTV-----D	469
Nissle	-----NTDVSNATATYTDADGEMTTIGSY-TTKY-SIDANNGKVTV-----D	469
H09	-----T-----GTVTYTVSNGDVKAAA-----NA-DVYVEDGALSAN--ATKDVTY	544
H52	-----KDAIENGGDAATNEDTKIQLTDEL-----	337
H33	-----LVNSK--NAAGSQNV T SAI-----	320
H39	-----VDVLANSVGAAPT TAVDTGT-----	324
H51	-----TPVVATGAKA--VYVDGNNELTSTASY-DTTY-SVNADTGAVKV-----V	479
H43	-----TTVTPGAETV TYKDASGN-----STTAAV-TLGGSDGKTNL-----V	379
H26	-----AVQTGGAD--VYKDTTGAL TTEDET V TAT--YGFADGKVS-----	440
H04	-----ATGK--YMEATV-V--ITG-----	241
H17	-----ATGK--YMAATV-V--ITS-----	241
H29	-----LVNANGAA--GAEATRVT--IDGDGTNQA-KIEL-----	332
H38	-----ALKNNTGD--A-TATQPG--TSGTTVVA A-SIHLSTGK-----	338
H36	-----D-----QKATLNGSEAYVKDGGFTIDNT-----AKY	431
H53	-----DSKTG-----	322
H27	-----ENKAA-----TASDLLNNAKKGSSLVVNGADYEVSADGKT V T-----GLGRTM	372
H11	-----DNKAA-----TLDL DLNAAKKTGSTLVVNGATYDVSADGKTITETASGN NKVM	373
H21	-----ENKAA-----TLDL DLNAAKKTGSTLVVNGATYDVSADGKTITETASGN NKVM	377
H23	-----ENKAA-----TLDL DLNAAKKTGSTLVVNGATYDVSADGKTITETASGN NKVM	377
H08	-----ENKAA-----TLDL DLNAAKKTGSTLVVNGATY NVSADGKT V TDTPGAPKVM	378
H40	-----ENKAA-----TLDL DLNAAKKTGSTLVVNGATY NVSADGKT V TDTPGAPKVM	378
H02	-----DQKTGAVSVK TMSYTDADGVKHDNVKV-----ELGSDGKTEVV T--ATDGKT	383
H35	-----TGVSG-----VTGAQ-----LFAKKDGS GYVIK--GTAD--	389
H03	-----DG-NG-----NTSVVKDADGNV E W V-----VDKDGA AKTVV--RKDDKI	408
H16	-----DAATG-----NTSVITDPNGKAV E W A-----VKNDGSAQAIM--REDDKV	408
H47	-----AGD-----	261
H54	-----ESTSKTVTVRTTSYKDV DGV PQKGLN-----KIGGADGKTETV--TIGEKT	401

H06	YAQADGSITTTEDAAAGKLVYKGS DGKLTDTT SKAESTSDPLAALDDAISQIDKFRSSL	478
H18	FAQADGTTTSGENAATSKAIYVSANGNLTTNTTSESEATTNPLAALDDAIASIDKFRSSL	484
H10	NA-----ANISDAGVLSIGASTTAPS NLTANPLKALDDAIASVDKFRSSL	350
H07	ATDTNKDYAPA----IGTAVNVNSAGKITTE TETSAGSATNPLAALDDAISSIDKFRSSL	515
H19	FAHTNGSVTND----TGSTIYATEDGSLTTDAATKAETTADPLKALDEAISSIDKFRSSL	543
H55	YAHTNGSVTND----SGSAIYATEAGKLTTEASTAAETTANPLKALDDAISQIDKFRSSL	552
H45	GAGATGKFAAT----VGAQAYVNSTGKLTTE TTSAGTATKDPLAALDEAISSIDKFRSSL	498
H34	YSQADGTVTNS----AGKEIYKDADGVYSTEN--KTSKTS DPLAALDDAISSIDKFRSSL	475
H44	FLQKDGSI TNN----SGKAVYVQEDGKFTTDAATKAATTADPLKALDDAISSIDKFRSSL	504
H46	FLQTDGGSVTNG----SGKGVYTDAAKGFTTDAATKAATT DPLKALDDAISQIDKFRSSL	500
H15	YLQKDGGSVTNG----SGKAVYVEADGDFTTDAATKAATT DPLKALDEAISQIDKFRSSL	492
H28	FLQKDGTVTNG----SGKAVYVSADGNFTTDAETKAATTADPLKALDEAISSIDKFRSSL	509
H49	GGSGTGKYAAN----VGAQAYVGADGKLTNTTSTG SATKDPLNALDEAIASIDKFRSSL	494
H20	YLQKDGGSVTNG----SGKAVYVEADGDFTTDAATKAATT DPLAALDDAISQIDKFRSSL	506
H31	YLQDDGGSVTNG----SGKAVYVDATGKLTTEATKAATTADPLKALDEAISSIDKFRSSL	485
H37	HEFANGNIYDD----KGAAVYAAADGSLTTE TTSKSEATANPLAALDDAISQIDKFRSSL	491
H41	HEFANGNIYDD----KGS SLYKAADGSLTSEAKGKSEATADPLKALDEAISSIDKFRSSL	489
H30	HEFANGNILDD----DGAALYKAADGSLTTEATGKSEVTTDPLKALDDAIASVDKFRSSL	500
H32	HEFANGNILDD----DGAALYKAADGSLTTEATGKSEAT DPLKALDDAIASVDKFRSSL	500
H42	-----QTI---DLDTADEFTGASTADPLALLDKAIAQVDTFRSSL	356
H05	-----VVVELSTAKPTAQFSGASSADPLALLDKAIAQVDTFRSSL	366
H56	-----VVVELSTATPTAQFSGASSADPLALLDKAIAQVDTFRSSL	366
H48	DSA-----DL----NGNLQTLTAGGEALTAVANGKTTDPLKALDDAIASVDKFRSSL	428
H14	YLQDDGGSVTNG----SGKAVYKDADGKLT TDAETKAATTADPLKALDEAISSIDKFRSSL	480
H12	SGTGTGKYAPK----VGAEVYVSANGTLT TTDATSEGTVTKDPLKALDEAISSIDKFRSSL	525
H01	SGTGSGKYAPK----VGAEVYVSANGTLT TTDATSEGTVTKDPLKALDEAISSIDKFRSSL	525
Nissle	SGTGTGKYAPK----VGAEVYVSANGTLT TTDATSEGTVTKDPLKALDEAISSIDKFRSSL	525
H09	FEQKNGAITNS----TGGTIYETADGKLTTEAT TASSSTADPLKALDEAISSIDKFRSSL	600
H52	-----DVDGSVKTAATATFSGTATNDPLALLDKAIASQVDTFRSSL	377
H33	-----GDIANKANANIYGTSSADPLALLDKAIASVDKFRSSL	358
H39	-----LQLSGTGATTELGKTATQNP LALLDKAIASVDKFRSSL	362
H51	SGTGTGKFEAV----AGADAYVSKDGKLT TETTSAGTATKDPLAALDAAISSIDKFRSSL	535
H43	TAA----DGKT----YGATALNGADLSDPNNTVK SVADNAKPLAALDDAIAMVDKFRSSL	431
H26	-----DG----EGSTVYKAADGSITKDATT KSEATDPLKALDDAISQIDKFRSSL	487
H04	-----TAAAVTVGAAEVAGAATADPLKALDAAIAKV DFRSSL	279
H17	-----TAAAVTVGATEVAGAATAEPLKALDAAIAKV DFRSSL	279
H29	-----SQNGDTAATSEFAGASTNDPLTLLDKAIASVDKFRSSL	370
H38	-----NSVDADVTA STEFTGASTNDPLTLLDKAIASVDKFRSSL	377
H36	YVQEDGAI TNG----SGKVAYKDADGKLT TDAKTETAKTTDPMAKLDKALAKVDALRSDL	487
H53	-----VVS-VGTKISTSPDVLATIDNALKIVDSQRSSL	354
H27	YLS----K--S----EGGSP----ILVKEDA AKSLQSTTNPLETIDKALAKVDNLRSDL	417
H11	YLS----K--S----EGGSP----ILVNEDA AKSLQSTTNPLETIDKALAKVDNLRSDL	418
H21	YLS----K--S----EGGSP----ILVNEDA AKSLQSTTNPLETIDKALAKVDNLRSDL	422
H23	YLS----K--S----EGGSP----ILVNEDA AKSLQSTTNPLETIDKALAKVDNLRSDL	422
H08	YLS----K--S----EGGSP----ILVNEDA AKSLQSTTNPLETIDKALAKVDNLRSDL	423
H40	YLS----K--S----EGGSP----ILVNEDA AKSLQSTTNPLETIDKALAKVDNLRSDL	423
H02	YSVS---DLQG---KSLK-----TDSIAAISTQKTEDPLA AIDKALSQVDSLRSNL	428
H35	-----NKE----VLF EAKVAADGKV----TKGDQLTADPLKSID DALSQVDQFRSSL	433
H03	YGAS---VT-G---FGGTPTVNVDTTAIDASELKGMTTAKPLEKLDTALAKVDKLRSSL	460
H16	YTAN---IT-N---KTATK-----GAELSASDLKALAT TNPLSALDEALAKVDKLRSSL	455
H47	-----LTK-----TKVDEDATAATKTSNPLSKIDDAISDVDSLRS DL	299
H54	YAAD---KLKD---HDF-----SKQATLGE EATTTTVNPLDAIDKALAQVDSLRS DL	447

. . . : * * : * * ** *

H06	GAVQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	538
H18	GAIQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	544
H10	GAVQNRLD SAIANLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	410
H07	GAIQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	575
H19	GAVQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	603
H55	GAVQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	612
H45	GAIQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	558
H34	GAIQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	535
H44	GAVQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	564
H46	GAIQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	560
H15	GAIQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	552
H28	GAVQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	569
H49	GAIQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	554
H20	GAIQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	566
H31	GAVQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	545
H37	GAIQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	551
H41	GAVQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	549
H30	GAVQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	560
H32	GAVQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	560
H42	GAVQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	416
H05	GAVQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLSKANQV	426
H56	GAVQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLSKANQV	426
H48	GAVQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	488
H14	GAVQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	540
H12	GAIQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	585
H01	GAIQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	585
Nissle	GAIQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	585
H09	GAVQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	660
H52	GAVQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLSKANQV	437
H33	GAVQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLSKANQV	418
H39	GAVQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	422
H51	GAIQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	595
H43	GAVQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLSKANQV	491
H26	GAVQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLAKANQV	547
H04	GAVQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLSKANQV	339
H17	GAVQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLSKANQV	339
H29	GAVQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLSKANQV	430
H38	GAVQNRLD SAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLSKANQV	437
H36	GAIQNRFDSTITNLGNTVNNLTSARSRIEDADYATEVSNMSRAQIILQQAGTSVLAQANQT	547
H53	GAIQNRFD SAI TNLGNTVNNLSSARSRIEDADYATEVSNMSRAQIILQQAGTSVLAQANQT	414
H27	GAVQNRFD SAI TNLGNTVNNLSSARSRIEDADYATEVSNMSRAQIILQQAGTSVLAQANQT	477
H11	GAVQNRFD SAI TNLGNTVNNLSSARSRIEDADYATEVSNMSRAQIILQQAGTSVLAQANQT	478
H21	GAVQNRFD SAI TNLGNTVNNLSSARSRIEDADYATEVSNMSRAQIILQQAGTSVLAQANQT	482
H23	GAVQNRFD SAI TNLGNTVNNLSSARSRIEDADYATEVSNMSRAQIILQQAGTSVLAQANQT	482
H08	GAVQNRFD SAI TNLGNTVNNLSSARSRIEDADYATEVSNMSRAQIILQQAGTSVLAQANQT	483
H40	GAVQNRFD SAI TNLGNTVNNLSSARSRIEDADYATEVSNMSRAQIILQQAGTSVLAQANQT	483
H02	GAIQNRFD SAI TNLGNTVNNLSSARSRIEDADYATEVSNMSRAQIILQQAGTSVLAQANQT	488
H35	GAVQNRFD SAI TNLGNTVNNLSSARSRIEDADYATEVSNMSRAQIILQQAGTSVLAQANQT	493
H03	GAVQNRFD SAI TNLGNTVNNLSSARSRIEDADYATEVSNMSRAQIILQQAGTSVLAQANQT	520
H16	GAVQNRFD SAI TNLGNTVNNLSSARSRIEDADYATEVSNMSRAQIILQQAGTSVLAQANQT	515
H47	GAVQNRFD SAI TNLGNTVNNLSSARSRIEDADYATEVSNMSRAQIILQQAGTSVLAQANQT	359
H54	GAVQNRFDSTITNLGNTLNNLSSARSRIEDADYATEVSNMSRAQIILQQAGTSVLAQANQT	507
	** : *** . : * . : * * . * * . * * . : * : * * * * * * . * * : * * * . * * * . : * * .	

H06	PQQVLSLLQG	548
H18	PQQVLSLLQG	554
H10	PQQVLSLLQG	420
H07	PQQVLSLLQG	585
H19	PQQVLSLLQG	613
H55	PQQVLSLLQG	622
H45	PQQVLSLLQG	568
H34	PQQVLSLLQG	545
H44	PQQVLSLLQG	574
H46	PQQVLSLLQG	570
H15	PQQVLSLLQG	562
H28	PQQVLSLLQG	579
H49	PQQVLSLLQG	564
H20	PQQVLSLLQG	576
H31	PQQVLSLLQG	555
H37	PQQVLSLLQG	561
H41	PQQVLSLLQG	559
H30	PQQVLSLLQG	570
H32	PQQVLSLLQG	570
H42	PQQVLSLLQG	426
H05	PQQVLSLLQG	436
H56	PQQVLSLLQG	436
H48	PQQVLSLLQG	498
H14	PQQVLSLLQG	550
H12	PQQVLSLLQG	595
H01	PQQVLSLLQG	595
Nissle	PQQVLSLLQG	595
H09	PQQVLSLLQG	670
H52	PQQVLSLLQG	447
H33	PQQVLSLLQG	428
H39	PQQVLSLLQG	432
H51	PQQVLSLLQG	605
H43	PQQVLSLLQG	501
H26	PQQVLSLLQG	557
H04	PQQVLSLLQG	349
H17	PQQVLSLLQG	349
H29	PQQVLSLLQG	440
H38	PQQVLSLLQG	447
H36	TQNVLSLLR-	556
H53	TQNVLSLLR-	423
H27	TQNVLSLLR-	486
H11	TQNVLSLLQG	488
H21	TQNVLSLLR-	491
H23	TQNVLSLLR-	491
H08	TQNVLSLLR-	492
H40	TQNVLSLLR-	492
H02	TQNVLSLLR-	497
H35	TQNVLSLLR-	502
H03	TQNVLSLLR-	529
H16	TQNVLSLLR-	524
H47	TQNVLSLLR-	368
H54	TQNVLSLLR-	516
	*:*****:	

Appendix 8: Different H-types with NCBI-entry (National Center for Biotechnology Information) used for sequence alignments.

H-types					
H01	L07387	H21	AIHL01000060	H41	AY250020
H02	AIHA01000023			H42	AY250021
H03	AB128916	H23	AB028476	H43	AIGA01000038
H04	AJ605764			H44	AB269770
H05	AY249990			H45	AY250023
H06	AY249991	H26	AY250008	H46	AB028478
H07	AB028474	H27	AM231154	H47	EF392694
H08	AJ865465	H28	AAJT02000052	H48	AY250025
H09	AY249994	H29	JH965342	H49	AY250026
H10	AY249995	H30	AY250011		
H11	AY337465	H31	CP000247	H51	AY250027
H12	AB028475	H32	AY250014	H52	AY250028
		H33	AY250015	H53	AB128917
H14	AY249998	H34	AY250016*	H54	AB128918
H15	AY249999	H35	EF392692	H55	AB269771
H16	AB128919	H36	EF392693	H56	AY250029
H17	AJ515904	H37	AY250017	Nissle	CP022686.1
H18	AY250001	H38	AY250018		
H19	AY250002	H39	AY250019		
H20	AY250003	H40	AJ884568		

Appendix 9: FliC protein sequence alignment of the constant region of different *E. coli* H-types (MUSCLE). AAs are marked based on their similarity, with an asterisk for a perfect aligned, with a colon mark for a strong similarity or with a dot for a weak similarity.

```

H04      MAQVINTNSLSLITQNNINKNQSSALSTSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H17      MAQVINTNSLSLNTQNNINKNQSSALSTSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H48      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H10      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H44      MAQVINTNSLSLITQNNINKNQSSMSTAIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H55      MAQVINTNSLSLITQNNINKNQSSMSTAIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H51      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H43      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H52      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H42      MAQVINTNSLSLITQNNINKNQSSALSTSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H33      MAQVINTNSLSLITQNNINKNQSSALSTSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H39      MAQVINTNSLSLITQNNINKNQSSALSTSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H05      MAQVINTNSLSLITQNNINKNQSSALSTSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H56      MAQVINTNSLSLITQNNINKNQSSALSTSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H29      MAQVINTNSLSLITQNNINKNQSSALSTSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H38      MAQVINTNSLSLITQNNINKNQSSALSTSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H15      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H37      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H46      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H20      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H30      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H32      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H06      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H07      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H34      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H49      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H01      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H45      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H12      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H41      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H14      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H28      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H31      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H18      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H19      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H09      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H26      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H47      MAQVINTNSLSLLTQNNLNKSQSSLSSAIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H53      MAQVINTNSLSLLTQNNLNKSQSSLSSAIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H36      MAQVINTNSLSLLTQNNLNKSQSSLSSAIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H02      MAQVINTNSLSLLTQNNLNKSQSSLSSAIERLSSGLRINSKDDAAGQAIANRFTANIKG
H54      MAQVINTNSLSLLTQNNLNKSQSSLSSAIERLSSGLRINSKDDAAGQAIANRFTANIKG
H35      MAQVINTNSLSLLTQNNLNKSQSSLSSAIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H27      MAQVINTNSLSLITQNNINKNQSSALSSSIERLSSGLRINSKDDAAGQAIANRFTANIKG
H11      MAQVINTNSLSLLTQNNLNKSQSSLSSAIERLSSGLRINSKDDAAGQAIANRFTANIKG
H21      MAQVINTNSLSLLTQNNLNKSQSSLSSAIERLSSGLRINSKDDAAGQAIANRFTANIKG
H23      MAQVINTNSLSLLTQNNLNKSQSSLSSAIERLSSGLRINSKDDAAGQAIANRFTANIKG
H08      MAQVINTNSLSLLTQNNLNKSQSSLSSAIERLSSGLRINSKDDAAGQAIANRFTANIKG
H40      MAQVINTNSLSLLTQNNLNKSQSSLSSAIERLSSGLRINSKDDAAGQAIANRFTANIKG
H03      MAQVINTNSLSLLTQNNLNKSQSSLSSAIERLSSGLRINSKDDAAGQAIANRFTSNIKG
H16      MAQVINTNSLSLLTQNNLNKSQSSLSSAIERLSSGLRINSKDDAAGQAIANRFTANIKG
***** **.* **.*:.*:*****.***

```

H04 LTQAARNANDGISLAQTAEGALSEINNNLQRIRELTVQASTGTNSDSDLSSIQDEIKSRL
H17 LTQAARNANDGISLAQTAEGALSEINNNLQRIRELTVQASTGTNSDSDLSSIQDEIKSRL
H48 LTQAARNANDGISVAQTTEGALSEINNNLQRVRELTVQATTGTNSESDLSSIQDEIKSRL
H10 LTQAARNANDGISVAQTTEGALSEINNNLQRVRELTVQATTGTNSESDLSSIQDEIKSRL
H44 LTQAARNANDGISVAQTTEGALSEINNNLQRIRELTVQSSTGTNSKSDLSSIQDEIKSRL
H55 LTQAARNANDGISVAQTTEGALSEINNNLQRIRELTVQSSTGTNSESDLSSIQDEIKSRL
H51 LTQAARNANDGISLAQTEGALSEINNNLQRVRELTVQATTGTNSDSDLSSIQDEIKSRL
H43 LTQAARNANDGISLAQTEGALSEINNNLQRVRELTVQATTGTNSESDLSSIQDEIKSRL
H52 LTQAARNANDGISVAQTTEGALSEINNNLQRIRELTVQASTGTNSDSDLSSIQDEIKSRL
H42 LTQAARNANDGISLAQTEGALSEINNNLQRVRELTVQATTGTNSDSDLSSIQDEIKSRL
H33 LTQAARNANDGISLAQTEGALSEINNNLQRVRELTVQATTGTNSDSDLSSIQDEIKSRL
H39 LTQAARNANDGISLAQTAEGALSEINNNLQRVRELTVQATTGTNSDSDLSSIQDEIKSRL
H05 LTQAARNANDGISLAQTEGALSEINNNLQRVRELTVQATTGTNSDSDLSSIQDEIKSRL
H56 LTQAARNANDGISLAQTEGALSEINNNLQRVRELTVQATTGTNSDSDLSSIQDEIKSRL
H29 LTQAARNANDGISLAQTEGALSEINNNLQRVRELTVQATTGTNSDSDLSSIQDEIKSRL
H38 LTQAARNANDGISLAQTEGALSEINNNLQRVRELTVQATTGTNSDSDLSSIQDEIKSRL
H15 LTQAARNANDGISVAQTTEGALSEINNNLQRVRELTVQATTGTNSQSDLSSIQDEIKSRL
H37 LTQAARNANDGISVAQTTEGALSEINNNLQRVRELTVQATTGTNSQSDLSSIQDEIKSRL
H46 LTQAARNANDGISVAQTTEGALSEINNNLQRVRELTVQATTGTNSQSDLSSIQDEIKSRL
H20 LTQAARNANDGISVAQTTEGALSEINNNLQRVRELTVQATTGTNSQSDLSSIQDEIKSRL
H30 LTQAARNANDGISVAQTTEGALSEINNNLQRIRELTVQATTGTNSTSDLSSIQDEIKSRL
H32 LTQAARNANDGISVAQTTEGALSEINNNLQRIRELTVQATTGTNSDSDLSSIQDEIKSRL
H06 LTQAARNANDGISVAQTTEGALSEINNNLQRIRELTVQASTGTNSDSDLSSIQDEIKSRL
H07 LTQAARNANDGISVAQTTEGALSEINNNLQRIRELTVQASTGTNSDSDLSSIQDEIKSRL
H34 LTQAARNANDGISVAQTTEGALSEINNNLQRIRELTVQASTGTNSDSDLSSIQDEIKSRL
H49 LTQAARNANDGISVAQTTEGALSEINNNLQRIRELTVQASTGTNSDSDLSSIQDEIKSRL
H01 LTQAARNANDGISVAQTTEGALSEINNNLQRIRELTVQASTGTNSDSDLSSIQDEIKSRL
H45 LTQAARNANDGISVAQTTEGALSEINNNLQRIRELTVQASTGTNSDSDLSSIQDEIKSRL
H12 LTQAARNANDGISVAQTTEGALSEINNNLQRIRELTVQASTGTNSDSDLSSIQDEIKSRL
H41 LTQAARNANDGISVAQTTEGALSEINNNLQRIRELTVQASTGTNSDSDLSSIQDEIKSRL
H14 LTQAARNANDGISVAQTTEGALSEINNNLQRIRELTVQASTGTNSDSDLSSIQDEIKSRL
H28 LTQAARNANDGISVAQTTEGALSEINNNLQRIRELTVQASTGTNSDSDLSSIQDEIKSRL
H31 LTQAARNANDGISVAQTTEGALSEINNNLQRIRELTVQASTGTNSDSDLSSIQDEIKSRL
H18 LTQAARNANDGISVAQTTEGALSEINNNLQRIRELTVQATTGTNSDSDLSSIQDEIKSRL
H19 LTQAARNANDGISVAQTTEGALSEINNNLQRIRELTVQATTGTNSTSDLSSIQDEIKSRL
H09 LTQAARNANDGISVAQTTEGALSEINNNLQRIRELTVQASTGTNSDSDLSSIQDEIKSRL
H26 LTQAARNANDGISVAQTTEGALSEINNNLQRIRELTVQATTGTNSDSDLSSIQDEIKSRL
H47 LTQASRNANDGISVAQTTEGALNEINNNLQRIRELTVQATNGTNSQSDMESIQAEITQRL
H53 LTQASRNANDGISIAQTTEGALNEINNNLQRVRELVQSQNGTNSDSDVQSIQEEIQRL
H36 LTQASRNANDGISIAQTTEGALNEINNNLQRVRELVQATNGTNSPDLSSIQNEITQRL
H02 LTQASRNANDGISVAQTTEGALNEINNNLQRIRELSVQATNGTNSDSDLSSIQAEITQRL
H54 LTQASRNANDGISVAQTTEGALNEINNNLQRVRELTVQATNGTNSDSDLSSIQAEITQRL
H35 LTQASRNANDGISVAQTTEGALNEINNNLQRIRELSVQATNGTNSDSDLSSIQAEITQRL
H27 LTQASRNANDGISVAQTTEGALNEINNNLQRVRELTVQATNGTNSDSDLSSIQAEITQRL
H11 LTQASRNANDGISVAQTTEGALNEINNNLQRVRELTVQATNGTNSDSDLSSIQAEITQRL
H21 LTQASRNANDGISVAQTTEGALNEINNNLQRIRELSVQATNGTNSDSDLSSIQAEITQRL
H23 LTQASRNANDGISVAQTTEGALNEINNNLQRIRELSVQATNGTNSDSDLSSIQAEITQRL
H08 LTQASRNANDGISVAQTTEGALNEINNNLQRIRELSVQATNGTNSDSDLSSIQAEITQRL
H40 LTQASRNANDGISVAQTTEGALNEINNNLQRIRELSVQATNGTNSDSDLSSIQAEITQRL
H03 LQASRNANDGISLAQTEGALSEINNNLQRVRELVQATNGTNSQSDLSSIQDEITQRL
H16 LTQASRNANDGISVAQTTEGALSEINNNLQRIRELSVQATNGTNSDSDLSSIQDEITQRL
:

H04 DEIDRVSGQTQFNGVNVLSKNDMSMKIQIGANDNQTISIGLQQIDSTTLNLLDAAIAKVDK
H17 DEIDRVSGQTQFNGVNVLSKNDMSMKIQIGANDNQTISIGLQQIDSTTLNLLDAAIAKVDK
H48 DEIDRVSGQTQFNGVNVLAKNGSMKIQVIGANDNQTITIDLKQIDAKTLGLLDDAIASVDK
H10 EEIDRVSSQTQFNGVNVLAKDGKMNQVIGANDGQTITIDLKIDSSTLNLLDDAIASVDK
H44 DEIDRVSGQTQFNGVNVLAKDGSMKIQVIGANDGQTITIDLKIDSSTLNLLDDAIAISSDK
H55 DEIDRVSGQTQFNGVNVLAKDGSMKIQVIGANDGETITIDLKIDSSTLNLLDDAISQIDK
H51 SEIDRVSGQTQFNGVNVLSKDGSLKIQVIGANDGQTISIDLKIDSSTLGLLDDAIAISSDK
H43 EEIDRVSGQTQFNGVNVLAKDGTMKIQVIGANDGQTISIDLKIDSSTLGLLDDAIAMVDK
H52 DEIDRVSGQTQFNGVNVLAKDGSMKIQVIGANDGQTISIDLKIDSSTLGLLDDKAISQVDT
H42 AEIDRVSGQTQFNGVNVLAKNGSLNIQVIGANDGQTISIDLKIDSSTLGLLDDKAIAQVDT
H33 DEIDRVSGQTQFNGVNVLAKNGSMKIQVIGANDGQTINIDLKIDSSTLGLLDDKAIASVDK
H39 DEIDRVSGQTQFNGVNVLAKNGSMKIQVIGANDGQTISIDLKIDSSTLGLLDDKAIASVDK
H05 DEIDRVSGQTQFNGVNVLAKDGSMKIQVIGANDGQTISIDLKIDSSTLGLLDDKAIAQVDT
H56 DEIDRVSGQTQFNGVNVLAKDGSMKIQVIGANDGQTISIDLKIDSSTLGLLDDKAIAQVDT
H29 DEIDRVSGQTQFNGVNVLAKDNTMKIQVIGANDGQTISIDLKIDSSTLGLLDDKAIASVDK
H38 DEIDRVSGQTQFNGVNVLAKDGSMKIQVIGANDGQTISIDLKIDSSTLGLLDDKAIASVDK
H15 DEIDRVSGQTQFNGVNVLAKDGSMKIQVIGANDGQTITIDLKIDSSTLNLLDDEAISQIDK
H37 DEIDRVSGQTQFNGVNVLAKDGSMKIQVIGANDGQTITIDLKIDSSTLKLDDAISQIDK
H46 DEIDRVSGQTQFNGVNVLAKDGSMKIQVIGANDGQTITIDLKIDSSTLKLDDAISQIDK
H20 DEIDRVSGQTQFNGVNVLAKDGSMKIQVIGANDGQTITIDLKIDSSTLKLDDAISQIDK
H30 DEIDRVSGQTQFNGVNVLSKDGSMKIQVIGANDGETITIDLKIDSSTLKLDDAIASVDK
H32 DEIDRVSGQTQFNGVNVLSKDGSMKIQVIGANDGETITIDLKIDSSTLKLDDAIASVDK
H06 DEIDRVSGQTQFNGVNVLAKDGSMKIQVIGANDGQTITIDLKIDSSTLGLLDDAISQIDK
H07 DEIDRVSGQTQFNGVNVLAKDGSMKIQVIGANDGETITIDLKIDSSTLGLLDDAIAISSDK
H34 DEIDRVSGQTQFNGVNVLAKDGSMKIQVIGANDGQTITIDLKIDSSTLGLLDDAIAISSDK
H49 DEIDRVSGQTQFNGVNVLAKDGSMKIQVIGANDGQTITIDLKIDSSTLGLLDEAIAISDK
H01 DEIDRVSGQTQFNGVNVLAKDGSMKIQVIGANDGQTITIDLKIDSSTLGLLDEAIAISSDK
H45 DEIDRVSGQTQFNGVNVLAKDGSMKIQVIGANDGQTITIDLKIDSSTLGLLDEAIAISSDK
H12 DEIDRVSGQTQFNGVNVLAKDGSMKIQVIGANDGQTITIDLKIDSSTLGLLDEAIAISSDK
H41 DEIDRVSGQTQFNGVNVLAKDGSMKIQVIGANDGQTITIDLKIDSSTLGLLDEAIAISSDK
H14 DEIDRVSGQTQFNGVNVLAKDGSMKIQVIGANDGQTITIDLKIDSSTLGLLDEAIAISSDK
H28 DEIDRVSGQTQFNGVNVLAKDGSMKIQVIGANDGQTITIDLKIDSSTLGLLDEAIAISSDK
H31 DEIDRVSGQTQFNGVNVLAKDGSMKIQVIGANDGQTITIDLKIDSSTLGLLDEAIAISSDK
H18 DEIDRVSGQTQFNGVNVLSKDGSMKIQVIGANDGETITIDLKIDSSTLNLLDDAIAISDK
H19 DEIDRVSGQTQFNGVNVLSKDGSMKIQVIGANDGETITIDLKIDSSTLNLLDDEAIAISSDK
H09 DEIDRVSGQTQFNGVNVLSKDGSMKIQVIGANDGETITIDLKIDSSTLNLLDDEAIAISSDK
H26 DEIDRVSGQTQFNGVNVLSKDGSMKIQVIGANDGETITIDLKIDSSTLNLLDDAISQIDK
H47 DEIDRVSGQTQFNGVSVLGEDKTLKIQVIGANDNQSIDINLKKIDSTVLKIDDAISDVDS
H53 AEIDRVSGQTQFNGVVKVLTSDSKLSIQVIGANDGEKIDIDLKIDTGTGLGIDNALKIVDS
H36 EEINRVSGQTQFNGVVKVLASDNSMTIQVIGANDGEAITIDLKEITAETLGLLDDKALAKVDA
H02 SEIDRVSGQTQFNGVVKVLASDQDMTIQVIGANDGETITIKLQEINSDTLGLIDKALSQVDS
H54 NEIDRVSGQTQFNGVVKVLASKNTLTIQVIGANDGETIDINLKEINSQTLGLIDKALAQVDS
H35 QEIDRVSNQTQFNGVVKVLASQQTMKIQVIGANDGETITIDLKEINSKTLGLIDDALSQVDQ
H27 EEIDRVSEQTQFNGVVKVLAENNEMKIQVIGANDGETITINLAKIDAKTLGLIDKALAKVDN
H11 EEIDRVSEQTQFNGVVKVLAENNEMKIQVIGANDGETITINLAKIDAKTLGLIDKALAKVDN
H21 EEIDRVSEQTQFNGVVKVLAENNEMKIQVIGANDGETITINLAKIDAKTLGLIDKALAKVDN
H23 EEIDRVSEQTQFNGVVKVLAENNEMKIQVIGANDGETITINLAKIDAKTLGLIDKALAKVDN
H08 EEIDRVSEQTQFNGVVKVLAENNEMKIQVIGANDGETITINLAKIDAKTLGLIDKALAKVDN
H40 EEIDRVSEQTQFNGVVKVLAENNEMKIQVIGANDGEAITINLAKIDAKTLGLIDKALAKVDN
H03 SEIDRVSGQTQFNGVVKVLATNQTMKIQVIGANDGQTEIGLKDIDATLGLLDDTALAKVDK
H16 SEIDRVSNQTQFNGVVKVLASDQTMKIQVIGANDGETIEIALDKIDAKTLGLLDEALAKVDK
**:* **:* **:* **:* **:* **:* **:* **:* **:* **:* **:* **:* **:* **:* **:* **:*

H04 FRSSLGAVQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLS
H17 FRSSLGAVQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLS
H48 FRSSLGAVQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H10 FRSSLGAVQNRLLDSAIANLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H44 FRSSLGAVQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H55 FRSSLGAVQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H51 FRSSLGAIQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H43 FRSSLGAVQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLS
H52 FRSSLGAVQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLS
H42 FRSSLGAVQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H33 FRSSLGAVQNRLLSSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLS
H39 FRSSLGAVQNRLLSSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIVQQAGNSVLS
H05 FRSSLGAVQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLS
H56 FRSSLGAVQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLS
H29 FRSSLGAVQNRLLSSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLS
H38 FRSSLGAVQNRLLSSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLS
H15 FRSSLGAIQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H37 FRSSLGAIQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H46 FRSSLGAIQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H20 FRSSLGAIQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H30 FRSSLGAVQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H32 FRSSLGAVQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H06 FRSSLGAVQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H07 FRSSLGAIQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H34 FRSSLGAIQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H49 FRSSLGAIQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H01 FRSSLGAIQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H45 FRSSLGAIQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H12 FRSSLGAIQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H41 FRSSLGAVQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H14 FRSSLGAVQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H28 FRSSLGAVQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H31 FRSSLGAVQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H18 FRSSLGAIQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H19 FRSSLGAVQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H09 FRSSLGAVQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H26 FRSSLGAVQNRLLDSAVTNLNNTTTNLSEAQSRIQDADYATEVSNMSKAQIIQQAGNSVLA
H47 LRSDLGAVQNRFDSDAITNLGNTVNNLSSARSRIEDSDYATEVSNMSRAQILQQAGTSVLA
H53 QRSSLGAIQNRFDSDAITNLGNTVNNLSSARSRIEDSDYATEVSNMSRAQILQQAGTSVLA
H36 LRSDLGAIQNRFDSTITNLGNTVNNLSSARSRIEDADYATEVSNMSRAQILQQAGTSVLA
H02 LRSDNLGAIQNRFDSDAITNLGNTVNNLSSARSRIEDADYATEVSNMSRAQILQQAGTSVLA
H54 LRSDLGAVQNRFDSTITNLGNTLNNLSSARSRIEDADYATEVSNMSRAQILQQAGTSVLA
H35 FRSSLGAVQNRFDSDAITNLGNTVNNLSSARSRIEDSDYATEVSNMSRAQILQQAGTSVLA
H27 LRSDLGAVQNRFDSDAITNLGNTVNNLSSARSRIEDADYATEVSNMSRAQILQQAGTSVLA
H11 LRSDLGAVQNRFDSDAITNLGNTVNNLSSARSRIEDADYATEVSNMSRAQILQQAGTSVLA
H21 LRSDLGAVQNRFDSDAITNLGNTVNNLSSARSRIEDADYATEVSNMSRAQILQQAGTSVLA
H23 LRSDLGAVQNRFDSDAITNLGNTVNNLSSARSRIEDADYATEVSNMSRAQILQQAGTSVLA
H08 LRSDLGAVQNRFDSDAITNLGNTVNNLSSARSRIEDADYATEVSNMSRAQILQQAGTSVLA
H40 LRSDLGAVQNRFDSDAITNLGNTVNNLSSARSRIEDADYATEVSNMSRAQILQQAGTSVLA
H03 LRSSLGAVQNRFDSDAITNLGNTVNNLSSARSRIEDSDYATEVSNMSRAQILQQAGTSVLA
H16 LRSSLGAVQNRFDSDAITNLGNTVNNLSSARSRIEDADYATEVSNMSRAQILQQAGTSVLA
*.***:***:.*::**.* .**:.*.***:*.*****.***:***.***:

H04	KANQVPQQVLSLLQG
H17	KANQVPQQVLSLLQG
H48	KANQVPQQVLSLLQG
H10	KANQVPQQVLSLLQG
H44	KANQVPQQVLSLLQG
H55	KANQVPQQVLSLLQG
H51	KANQVPQQVLSLLQG
H43	KANQVPQQVLSLLQG
H52	KANQVPQQVLSLLQG
H42	KANQVPQQVLSLLQG
H33	KANQVPQQVLSLLQG
H39	KANQVPQQVLSLLQG
H05	KANQVPQQVLSLLQG
H56	KANQVPQQVLSLLQG
H29	KANQVPQQVLSLLQG
H38	KANQVPQQVLSLLQG
H15	KANQVPQQVLSLLQG
H37	KANQVPQQVLSLLQG
H46	KANQVPQQVLSLLQG
H20	KANQVPQQVLSLLQG
H30	KANQVPQQVLSLLQG
H32	KANQVPQQVLSLLQG
H06	KANQVPQQVLSLLQG
H07	KANQVPQQVLSLLQG
H34	KANQVPQQVLSLLQG
H49	KANQVPQQVLSLLQG
H01	KANQVPQQVLSLLQG
H45	KANQVPQQVLSLLQG
H12	KANQVPQQVLSLLQG
H41	KANQVPQQVLSLLQG
H14	KANQVPQQVLSLLQG
H28	KANQVPQQVLSLLQG
H31	KANQVPQQVLSLLQG
H18	KANQVPQQVLSLLQG
H19	KANQVPQQVLSLLQG
H09	KANQVPQQVLSLLQG
H26	KANQVPQQVLSLLQG
H47	QANQTTQNVLSLLR-
H53	QANQTTQNVLSLLR-
H36	QANQTTQNVLSLLR-
H02	QANQTTQNVLSLLR-
H54	QANQTTQNVLSLLR-
H35	QANQTTQNVLSLLR-
H27	QANQTTQNVLSLLR-
H11	QANQTTQNVLSLLQG
H21	QANQTTQNVLSLLR-
H23	QANQTTQNVLSLLR-
H08	QANQTTQNVLSLLR-
H40	QANQTTQNVLSLLR-
H03	QANQTTQNVLSLLR-
H16	QANQTTQNVLSLLR- :***.!.*:*****.

Appendix 10: FliC protein sequence alignment of the hypervariable region of different *E. coli* H-types (MUSCLE). AAs are marked based on their similarity, with an asterisk for a perfect aligned, with a colon mark for a strong similarity or with a dot for a weak similarity.

H06	LKKIDSDTLGLNGFNVGKGE	180
H18	LKKIDSDTLNLAGFNNGEGE	180
H10	LKKIDSSTLNLS SFDATNLGT	180
H07	LKKIDSDTLGLNGFNVGKGT	180
H19	LKKIDSDTLNLAGFNVGKGS	180
H55	LKKIDSSTLNLTGFNVNGKGS	180
H45	LKKIDSDTLGLNGFNVGKGT	180
H34	LKKIDSDTLGLSGFNVNGSAD	180
H44	LKKIDSSTLNLTGFNVNGEGS	180
H46	LKKIDSSTLKLTFGNVNGS--	178
H15	LKKIDSSTLNLTGFNVNGSGS	180
H28	LKKIDSDTLGLSGFNVNGGGA	180
H49	LKKIDSDTLGLNGFNVGKGT	180
H20	LKKIDSSTLKLTFGNVNGSGS	180
H31	LKKIDSDTLGLSGFNVGKGA	180
H37	LKKIDSSTLKLTFGNVNGKAA	180
H41	LKKIDSDTLGLSGFNVGKGA	180
H30	LKKIDSSTLKLTSFNVGKGA	180
H32	LKKIDSSTLKLTSFNVGKGA	180
H42	LQKIDSSALGLSGFSVAGGAL	180
H05	LQKIDSSTLGLNGFSVSGQSL	180
H56	LQKIDSSTLGLNGFSVSAQSL	180
H48	LKQIDAKTLGLDGFVKNNDT	180
H14	LKKIDSDTLGLSGFNVNGGGA	180
H12	LKKIDSDTLGLNGFNVNGSGT	180
H01	LKKIDSDTLGLNGFNVNGSGT	180
Nissle	LKKIDSDTLGLNGFNVNGSGT	180
H09	LKKIDSDTLNLAGFNVGKGS	180
H52	LQKIDSSTLGLKGFVSIGNAL	180
H33	LQKIDSSTLGLGGFVSNNAL	180
H39	LQKIDSSTLGLSGFSVSQNSL	180
H51	LKKIDSDTLGLNGFNVNGSGT	180
H43	LKKIDSSTLGLTGFDVSTKAN	180
H26	LKKIDSDTLNLAGFNVNGAGS	180
H04	LQQIDSTTLNLKGFTVSGMAD	180
H17	LQQIDSTTLNLKGFTVSGMAD	180
H29	LQKIDSSTLGLNGFSVSKNAL	180
H38	LQKIDSSTLGLNGFSVSKNAV	180
H36	LKEITAE TLGLTFGNVGKGS	180
H53	LKKIDTGTGLANFVDSKFD	180
H27	LAKIDAKTLGLDGFNIDGAQK	180
H11	LAKIDAKTLGLDGFNIDGAQK	180
H21	LAKIDAKTLGLDGFNIDGAQK	180
H23	LAKIDAKTLGLDGFNIDGAQK	180
H08	LAKIDAKTLGLDGFNIDGAQK	180
H40	LAKIDAKTLGLDGFNIDGAQK	180
H02	LQEINSDTLGLSGFGIKDPTK	180
H35	LKEINSKTLGLDKLDVRNTFS	180
H03	LDKIDADTLGLKDFSVASAKV	180
H16	LDKIDAKTLGLDNF SVAPGKV	180
H47	LKKIDSTV LKLRDL DVVSET-	179
H54	LKEINSQTLGLDKLNVQRAYT	180

H06	TANTAATLKDMSGFTAAAAP-----GGTVGVYQYTDKS----AVASSVDILN	223
H18	TANTAATLKDMMVGLKLDNTG-----VTTAGVNRVIADK----AVASSTDILN	223
H10	SVKDGATINKQVAVGAGDFK-----DKASGSLG	208
H07	ITNKAATVSDLTSAGAKLNTTTG-----LYD-----LKTENTLLTTDAAF	221
H19	VANTAATDNLTLAGFTAGT-----KAAD--GTVT----YS--KNVQFAAATASNVL	225
H55	VSNTAATDNLTLKLAGFTAGA-----TPAAD--GTVT----YS--KDVDNAKAAAASNVL	226
H45	IANKAATVSDDLTAAGATGTG-----P-----YA--VTTNNTALSASDALS	218
H34	KASVAATADGMVKDGYIKGL-----TSSDG----STAYT--KTTANTAAKGSDILA	225
H44	VANKAATKADLTAAQLTTAAGGTIAAPAADANGVTK----YT--VSAGLNESTVADVFA	234
H46	VANTAATKDELAAAAAAGT-----TPAVGTDGVTK----YT--VDAGLNKATAANVFA	226
H15	VANTAATKADLTAAQLSA-----PGAADANGTVT----YT--VSAGYKESTAADVIA	226
H28	VANTAATKSDLAQAQLLA-----PGTADANGTVT----YT--VSAGLKTSTAADVIA	226
H49	IANKAATISDLAATGANVTN-----S-----SNIVVTTKFNALDAATAFS	220
H20	VANTAATKADLAAAAIGTPG-----AA---DSTGAIAYT--VSAGLTKTTAADVLS	226
H31	VANTAATKDDDLVAASVSAA-----VGNEYT--VSAGLSKSTAADVIA	220
H37	VDNAKATDANLTTAGFTQGV-----VDSNGNST---WTKSTTTNFDAATAVNVLA	227
H41	VANAKATEADLTGAGFSQGA-----VDTNGNST---WTKSTTTNYSAAATADLLS	227
H30	VDNAKATEADLTGAGFSQGA-----VV-SGNST---WTKSTVTFNAATATDVLA	226
H32	VANAKATEADLTGAGFSQSA-----VV-SGNST---WTKSTVTFNAATATDVLA	226
H42	KLSDTVTQVG-DGSA-APVK-----VDLDAAT-----DI---GTALG	213
H05	NVSDSITQITGAAGT-KPVG-----VDFTAVAK-----DL---TTATG	214
H56	NVSDSITQITGAAGT-KPVG-----VDFTAVAK-----DL---TTATG	214
H48	VTSAPVT---AFGATTN-----NIKLT-----	201
H14	VANTAASKADLVAAANATVVG-----N---KYT-----VSAGYDAKASDLLA	219
H12	IANKAATISDLTAAKMDAAT-----NTIT-----TTNNALTASKALD	217
H01	IANKAATISDLTAAKMDAAT-----NTIT-----TTNNALTASKALD	217
Nissle	IANKAATISDLTAAKMDAAT-----NTIT-----TTNNALTASKALD	217
H09	VANTAATSDLLKLAGFTKGT-----TDTNGVTA---YT--NTISNDKAKASDLLA	225
H52	KVSDAITTVPGANAGDAPVT-----VKFGAN-----DTA-----AAAMAKTLGI	219
H33	KLSDSITQVVGASGS-LADVK-----L-----SSVASALGV	209
H39	KLSDSITTIIGNTTAASKNVD-----L-----SAVATKLG	210
H51	IANKAATISDLTAQKAVD-----NNGTYKVTTSNAA-----LTASQALS	220
H43	ISTAVTGAATTTYADSAVA-----IDIGTDISGIAADAALGTINFDNNT-----	225
H26	VDNAKATGKDLTDAGFTASA-----ADANGKIT---YTKDVTTKFDKATAADVLG	227
H04	FSAAKLTAA-----	189
H17	FSAAKLTAA-----	189
H29	ETSEAITQLPN-----G-----ENAP-----IAVKMDASVLTDLNI	211
H38	SVGDAITQLPG-----ETA-----ADAP-----VTIKFDDSVKTDLKL	213
H36	VNNTVATAKD-----LTDKG-----FISTDNGKTYT--GSAGLANAKAGDVFG	221
H53	STSTKTVANG-----GDIVLS	196
H27	ATGSDLISKFKATGT--D-----NYQIN-----GTDNYT	207
H11	ATGSDLISKFKATGT--D-----NYDVG-----G-DAYT	206
H21	ATGSDLISKFKATGT--D-----NYQIN-----GTDNYT	207
H23	ATGSDLISKFKATGT--D-----NYQIN-----GTDNYT	207
H08	ATGSDLISKFKATGT--D-----NYDVG-----G-KTYT	206
H40	ATGSDLISKFKATGT--D-----NYDVG-----G-DAYT	206
H02	LKAA-----T-----AETTYFG	192
H35	TSDLAATATTELAPAKTDVK-----NLFTTDGAT-----APTFLK-	215
H03	PTSGAVALKSEMSPTLTSV-----NATTGKN-----GTNYAFG	213
H16	PMSSAVALKSEAAPDLTKV-----NATDGSV-----GGAKAFG	213
H47	-----Q-----YKDGAT-----T-----	187
H54	PSGENVKVDTTTTTT-----YTDGTA-----IKNKA--	207

H06	AVAGADGNK-----V-TTSADVFGT--PAAAVTYTYNKDINSYSAASDD-	265
H18	AVAGVDGSK-----V-STEADVFGAAAPGTPVEYTYHKDNTNTYTAS-AS-	266
H10	TLK-----LVEKDGKYYVNDTKSSKYYDAEVD	235
H07	KLG--NGDK-----VTVGGVDYTYNAKSGDFTTTKSTA	252
H19	AAK--DGDE-----ITF-AGNNGTG--IAATGGTYTYHKDSNSYSFSATAA	266
H55	AAK--NGDT-----ISF-AGNNGTG--ITATAGTYTYNKASDSYSFSATAA	267
H45	RLK--TGDT-----VTT-----TGSSAAIYTYDAAKGNFTTQATVA	252
H34	ALK--TGDK-----ITATGA-NSLA--DNATSTTYTYNATSNTFSYTADGV	266
H44	GLG--DTA-----VVNANITSGFNAV---TGNNYTYHKDNTNDFTFNATIA	274
H46	NLA--DGA-----VVDASISNGFGAA---AATDYTYNKATNDFTFNASIA	266
H15	SIK--DGSAPT-----SAITATINNGFGDSSALTSNDYTYDPAKGDFTYDVASS	273
H28	SLA--NNA-----KVNATIANGFGSP---TATDYTYNSATGDFTYSATIA	266
H49	KLK--DGDS-----VAV-----AAQKYTYNASTNDFTFTEN--T	249
H20	SLA--DGTT-----ITATGVKNGFA--AGATSNAKLNKDNNTFTYDT--T	266
H31	SLT--DGAT-----VTAAGVSNFGA--AGATGNAYKFNQANNTFTYNT--T	260
H37	AVK--DGST-----I-NYTGTTNGL--GIAATSAYTYH--DSTKSYTFDST	266
H41	TIK--DGST-----V-TYAGTDTGL--GVAAAGNYTYD--ANSKYSFNAN	266
H30	SVS--GGST-----ISGYTGTNGL--GVAASTAYTYN--ATSKYSFSDAT	266
H32	SVS--GGST-----ISGYTGTNGL--GVAASTAYTYN--ATSKYSFSDAT	266
H42	QKV--NASS-----LTLHNLDKDG---AATENYVV--SYGSDNYAA---	248
H05	KTV--DVSS-----LTLHNTLDAKG---AATSQFVV--QSGNDFYSA---	249
H56	KTV--DVSS-----LTLHNTLDAKG---AATAQFVV--QSGSDFYSA---	249
H48	-----GIT-----LSTEAATDTGGTNPASIEGYVT---DNGNDYAKITG	238
H14	GVS--DGDT-----VQA-TINNGFG--TAASATNYKYD--SASKYSFDTT	258
H12	QLK--DGDT-----VTI-----KAD--AAQTATVYTYNASAGNFSFSN--V	252
H1	QLK--DGDT-----VTI-----KAD--AAQTATVYTYNASAGNFSFSN--V	252
Nissle	QLK--DGDT-----VTI-----KAD--AAQTATVYTYNASAGNFSFSN--V	252
H09	NIT--DGSV-----ITGGGAN--AF--GVAANKGYTYD--AASKYSFAAD	263
H52	SDT--SGLS-----LH--NVQSADG--K--ATGTYVV--QSGNDFYSA---	252
H33	-DA--STLT-----LH--NVQTPAG--A--ATANYVV--SSGSDNYSV---	241
H39	-NA--STLS-----LH--EVQDSAG--D--GTGTFVV--SSGSDNYAV---	242
H51	KLS--DGDT-----VD--IATYAG--GTSSTVSYKYDADAGNFSYNN--T	257
H43	-----GK-----YYAQITS-AA-NPGLDGAYEIHVNDADGSFTVAAS	260
H26	KAA--AGDS-----ITYAGTDTGLG--VAADASTYTY--NAANKSYTFDAT	267
H04	-DG--TAIA-----AA--DVKDAGG--K--QVNL-----	209
H17	-DG--TAIA-----AA--DVKDAGG--K--QVNL-----	209
H29	TDA--SAVS-----LH--NVTK--GG--V--ATSTYVV--QYGDKSYA----	242
H38	TDA--SGLS-----LH--NLKDENG--N--LTNQYVV--QNGGKSYA----	245
H36	KMVD-TGTVTT-----TIDNGF--GTAQSNYKYDKASNSFSFDIDDA	261
H53	SKTI-KAEIQI----DSHSADPK---AADGLY--ALKDGTGYAVKDKDGAYHAAVKNA	244
H27	VNVD-SGAVQN-----E-----DGDAIFVSAATDGSLLTKSDTK	239
H11	VNVD-SGAVKD-----T-----TGNDIFVSAADGSLLTKSDTK	238
H21	VNVD-SGVVQD-----K-----DGKQVYVSAADGSLLTSSDTQ	239
H23	VNVD-SGVVQD-----K-----DGKQVYVSAADGSLLTSSDTQ	239
H08	VNVE-SGAVKN-----D-----ANKDVVSAADGSLLTSSDTK	238
H40	VNVD-SGAVKD-----K-----DNKDVVSAADGSLLTSSDTK	238
H02	STVK-LADANTLDA-----DITATVKGTT---TPGQRDGNIMSDANGKLYVKVAGS	239
H35	-YVD-NTD-----P-----DKLFVSNNGTNYAASYDK	241
H03	ATFR-TDDVNT--YFGKAVNTGD-KTNVLG-T---EA---EVFQIQVEGQTYFVAQNGD	262
H16	SNYK-NADVET--YFGTGNVQDTKDTTDDATGTA---GT---KVYQVQVEGQTYFVQDNN	264
H47	--LK--GDIKSTSVQKFDTT-----E-AAISPKDKGKYYAVVSKS	224
H54	-ALA--TDVNNASSIGVSDAIPG-----D-IKFNETSGKYYVAITST	245

H06	-----ISSANLAAFLNPQAGDTTKATVTIGG----	KDQDV-NIDKSGNLTA	306
H18	-----VDATQLAAFLNPEAGGTTAATVSIENGTTAQEQKV-I	IAKDGSLTA	311
H10	-----	TSKKGINF	243
H07	GTGVDAQAATDSAKKRDALAAFLHADVGVKSVNGSYTTKD---	GTVSF-ETDSAGNITI	307
H19	S-----KDSLLSTLAPNAGDFTTAKVTIGS----	KSQEV-NVSKDGTITS	306
H55	S-----KDSLLSMLAPNAGDSFTASVSIIGG----	KAQDV-NVSKDGTITT	307
H45	-----DGDVNFANTLKPAAAGTTASGVYTRST----	GDVKF-DVDANGDVTI	294
H34	N-----QTNAANLIPAAAGTTAASVTIGG----	TAQNV-NIDDSGNITS	306
H44	A----GDA---TTPSNSAKLQSI LTPKAGDTAHLNVKIGA----	TSVDV-VLSSDGKITA	322
H46	A----GAA---AGDSNSAALQSFLTPKAGDTANLSVKIGT----	TSVNV-VLASDGKITA	314
H15	-----ANNTAAQVQSFLTPKAGDTANLKVTVGT----	TSVDV-VLASDGKITA	316
H28	A----GTN---SGDSNSAQLQSFLTPKAGDTANLNVKIGS----	TSIDV-VLASDGKITA	314
H49	V-----ATGTATTDLGATLKAAGQSQSGLTYTFAN----	GKVNFDVDASGNITI	294
H20	-----ATTAELQSYLTPKAGDTATFSVEIGG----	TTQDV-VLSSDGKITA	307
H31	-----STAAELQSYLTPKAGDTATFSVEIGS----	TKQDV-VLASDGKITA	301
H37	GAAVAG-----AASSLQGTG--GTDNTAKITIDG----	SAQEV-NIAKDGKITD	309
H41	GLTGAN-----TATALKGYL--GTGANTAKISIGG----	TEQEV-NIAKDGKITD	309
H30	ALTNGDGT--GATTKVADV LKAYA--ANGDNQAQISIGG----	SAQDV-KIASDGT LTD	316
H32	ALTNGDGT--AGSTKVADV LKAYA--ANGDNQAQISIGG----	SAQEV-KIASDGT LTD	316
H42	-----	S-VA-DDGTVTL	258
H05	-----	S-INHTDGKVTL	260
H56	-----	S-IDHASGEVTL	260
H48	-----	GDNDGKY-----YAV-TVANDGTVTM	258
H14	----TASA-----ADVQYLT PPGVDTAKGTITIDG----	SAQDV-QISSDGKITS	300
H12	SNNTSAKA-----GDVAASLLPPAQGTASGVYKAAS----	GEVNF-DVDANGKITI	298
H01	SNNTSAKA-----GDVAASLLPPAQGTASGVYKAAS----	GEVNF-DVDANGKITI	298
Nissle	SNNTSAKA-----GDVAASLLPPAQGTASGVYKAAS----	GEVNF-DVDANGKITI	298
H09	G----ADS-----AKTLSIINPNTGDSSQATVTIGG----	KEQKV-NISQDGKITA	305
H52	-----	S-VN-AGGVVTL	262
H33	-----	S-VEDSSGTVTL	252
H39	-----	S-VDAASGAVNL	253
H51	ANKTSAAA-----GTLADTLLPAAGQTKTGTYKAAT----	GDVNF-NVDATGNLTI	303
H43	DKQAGAAP-----GTALT-----SGKVQ-----	TA-TTTPGTAVDV	290
H26	GV-AKADA-----GTALKGYLG---ASNTGKINIGG----	TEQEV-NIAKDG SITD	309
H04	-----	-----	209
H17	-----	-----	209
H29	-----	A-SVDAGGTVKL	253
H38	-----	A-TVAANGNVTL	256
H36	AGTTA-----PQVATY---LNPTANDKTKASVEING----	SSQAVI-IDHNGKMTA	304
H53	DGK-----	VTWDS-----SDTSVT-----AT	260
H27	VGGTGIDA-----TGLAKAAVSLAKDASIKYQG-ITFTN----	KGTGAFDGSNGTLIA	288
H11	IAGTGIDA-----TALAAAANKAQNDFTFNG-VEFTT----	TT--AADGNNGVYSA	285
H21	FK---IDA-----TKLAVA AKDLAQGNKIVYEG-IEFTN----	TGTGAIPATGNGK LTA	285
H23	FK---IDA-----TKLAVA AKDLAQGNKIVYEG-IEFTN----	TGTGAIPATGNGK LTA	285
H08	VSGESIDA-----TELAKLAIKLADKGSIEYKG-ITFTN----	NTGAELDANGKGV LTA	287
H40	VDGKSIDA-----TELAKLAINLADQKSIDYKG-ITFTN----	KSGTAFDANGKGV LTA	287
H02	D-----	KPAENGYE-VTVED-----DPTSPD---AGK LKL G	267
H35	D-----TNSIKFNSTGTGTGTAGTDFIDA-----	TGKDVH--VGGGRVMA	280
H03	T-----ATNAYSLLKQSGSGYEK-VTVDS----	KAVQIA--NFGGRVTA	299
H16	T-----NTNGFTLLKQNSTGYEK-VQVGG----	KDVQLA--NFGGRVTA	301
H47	T-----TTD-----	NNGIYA-ASVDS-----DGNVTI--DASKKVTV	253
H54	E-----HND-----	KNGDYE-ITVDK-----DGS AAL--VAGQSSPK	274

H06	ADD-----GAVLYMDATGNLTKNAG----GDTQATLAKLATATGAKAATIQT	351
H18	ADD-----GAALYLDGTGNLSKTNAG----TDTQAKLSDLMANNANAKTVITTD	356
H10	NST-----NESGT---TPTAA-----	256
H07	GGG-----QAYVDDAGNLTNNAGS----AAKADMKALLKAASEGS-----D	345
H19	SDG-----KALYLDEKGNLTQTGS---GTTKAATWDLNMANTDITGKDAY-G	349
H55	TDG-----KSLYLQKGNLTQTGS---GTTIAATWDLNMANTDITGVDATSG	351
H45	GGK-----AAYLDATGNLSTNNPGI----ASSAKLSDLFASGSTLA-----T	332
H34	SDG-----DQLYLDSTGNLTKNQA---GNPKKATVSGLLGNTDARGTA---V	347
H44	KDG-----SKLFIGIDGNLTQNSAGA---GIKPATLDALTQNT---TTPAAAV	364
H46	KDG-----SALYIDSTGNLTQNSAGT---V-TAATLDGLTKNHDTGAVG---	355
H15	KDG-----SALYIDSTGNLTQNSAGL---T-SAKLA-TLT-----GLQSGSV	353
H28	KDG-----SELFIDVDGNLTQNNAGT---V-KAATLDALTKNWHTTGTP-GAV	357
H49	GGE-----KAFLV-GGALTTNDP---TGSTPATMSSSLFKAADDKDA-----	332
H20	KDG-----SKLYIDTTGNLTQNGGNGVGLAEATLSGLALNKNGLT-A---V	351
H31	KDG-----SKLYIDTTGNLTQNGG---GTLEEATLNGLAFNHSGPAAA---V	342
H37	TDG-----KALYIDSTGNLTKNNG---S-DTLTQATLNDVLTGANS---VDD---	348
H41	TNG-----DALYLDITGNLTKNY--A-GSPPAATLDNVLASA---TVN---	346
H30	VNG-----DALYIGSDGNLTKNQ--A-GGPAATLDGIFNGANGNAAVD---	357
H32	TNG-----DALYIGADGNLTKNQ--A-GGPAAATLDGIFNGANGHDAVD---	357
H42	NKT-----DITYSGG---DITGATKD--DTLIKV-----	282
H05	NKA-----DVEYTDTDNGLTTAATQK--DQLIKV-----	287
H56	NKA-----DVEYKDTDNGLTTAATQK--DQLIKV-----	287
H48	ATG-----ATANATVTDANTTKA-----	276
H14	SNG-----DKLYIDTTGRLTKNGFS---ASLTEASLSTLAANNTKA-----	338
H12	GG-----QKAYLTSDBGNLTND--A-GGATAATLDGLFKKAGDQSIGFKK	341
H01	GG-----QEAYLTSDBGNLTND--A-GGATAATLDGLFKKAGDQSIGFKN	341
Nissle	GG-----QEAYLTSDBGNLTND--A-GGATAATLDGLFKKAGDQSIGFKN	341
H09	ADD-----NATLYLDKQGNLTKTNA---GNDTAATWDGLISNSDSTGAVPVGV	350
H52	NTT-----NVTFTDPANGVTTAT-QT--GQPIKV-----	288
H33	NTT-----DIGYTDTANGVTTGS-MT--GKYVKV-----	278
H39	NTT-----DVTYDDATNGVTTGAT-QN--GQLIKV-----	279
H51	G-G-----QQAYLTTDGNLTNN--S--GGAATATLKEFLTLAGDGKSLGNGG	346
H43	TAA-----KTAL--A-----AAGADT-----	304
H26	TNG-----DALYLDSTGNLTKNANTANL--GAADKATVVKLFFAGAQD-----	347
H04	-----LSYTDTASNS-----	219
H17	-----LSYTDTASNS-----	219
H29	NKA-----DVTYNDAANGVKN-ATQI--G-----SLV-----	277
H38	NKA-----NVTYSVDVANGIDT-ATQS--G-----QLV-----	280
H36	ADD-----NAELFIDNSGNLTKNKNTG---GKAATLENLALNKTGTNTA---	345
H53	DVDQTT---KVTE-----FQEVYKK-----VAANTSGLAA-----	287
H27	NIDGK-----DVTFTID---A-----TGKDAT---	307
H11	EIDGK-----SVTFTVT---DA--D-----KKAS---	304
H21	NVDGK-----AVEFTIS---GS--AD-----TSGTSAT---	308
H23	NVDGK-----AVEFTIS---GS--AD-----TSGTSAT---	308
H08	NIDGQ-----DVQFTID---SN--AP-----TGAGAT---	309
H40	NIDGQ-----DVKFTIN---ST--AA-----TGADAT---	309
H02	AL-----AGTQPQAGN---LKEVT-TVK---GK--GAIDVQLGTDATATA---	302
H35	ANDIKGRTDDNVAAKPQLFLDQSGKYHI-----SKDNGAT---	315
H03	FVDDGTAAHNALS-----V-----DLQKGTVGKALS---	325
H16	FVEDNGS---ATS-----V-----DLAAGKMGKALA---	324
H47	DL-----	255
H54	SLEDVGATKNVTAY--QVANTQSNTQSVDATVS---AG--AISELKTGGVDNTA-----	321

H06	KGTFTSDGTAFDGASMS-----	368
H18	KGTFTANTTKFDGVDIS-----	373
H10	-TEVTTVGRDVK-----	267
H07	GASLTFNGTEYTI AKATPATTSP-----VAPLIPGGITYQATV-----	383
H19	NSAAA AVGTVIEAKGMTITSAGGNA-----QVLKDAAYNAAAYATSITTGTPGDAGA--	400
H55	HSAATAVKTI IKVKDMTITSAGGNA-----QVATDKAYNDKYAARIVAGDTAAQAD--	402
H45	TGSIQLSGTTYN-FGAAATS-----GVTYTKTV-----	359
H34	KTTIKTE-----AGVTVTAE-----	362
H44	PVTI-----	368
H46	-VDI-----	358
H15	ASTI-----	357
H28	STVI-----	361
H49	QSSIDFGGKKYE-FAGGNSTNGG-----GVKFKDT-----	361
H20	KST-----	354
H31	QST-----	345
H37	-TRIDFD-----SGMSVFLDKVNSTVD-----ITGASISAA-----AMTN-	382
H41	-ATIKFD-----SGMTVDYTAG-TGAN-----ITGASISAD-----DMAA-	379
H30	-AKITFG-----SGMTVDFTQASKKVD-----IKGATVSAE-----DMDT-	391
H32	-AKITFG-----SGMTVDFTQVSNNVD-----IKGATVSAE-----DMNT-	391
H42	-----	282
H05	-----	287
H56	-----	287
H48	-TTITSGGTPVQ-----IDNTAGSATA-----NLGAVSLVKLQ--DSKGNDDT-----	316
H14	-TTIDIGGTSIS-FT-----GNSTT-----	356
H12	TASVTMGGTTYN-FKGTGADADAATA-----NAGVSFTDTAS-----KETVLNK	383
H01	TASVTMGGTTYN-FKGTGADAGAATA-----NAGVSFTDTAS-----KETVLNK	383
Nissle	TASVTMGGTTYN-FKGTGADAGAATA-----NAGVSFTDTAS-----KETVLNK	383
H09	ATTTITSGTA--SGMSV-QSAGAGIQTSTNSQILAGGAFAAKVS--IEGGAATDILVAS	405
H52	-----	288
H33	-----	278
H39	-----	279
H51	TATVTLDNNTTYN-FKAAANVTDGAGVIA-----AAGVYTYTATV-----SKDVILAQ	391
H43	-SGLKLV-----QLSNTDSAGKVT-----NVGY-----	326
H26	-ATITFDS-----GMTAKFDQTAGTVD-----FKGASIS-----ADAMAS--	381
H04	-----	219
H17	-----	219
H29	-----QVGADAN-----	284
H38	-----QVGADST-----	287
H36	---V-----	346
H53	-----	287
H27	---L-----	308
H11	---L-----	305
H21	---V-----	309
H23	---V-----	309
H08	---I-----	310
H40	---I-----	310
H02	-----	302
H35	---Y-----	316
H03	---F-----	326
H16	---Y-----	325
H47	-----	255
H54	-----	321

H06	-----IDTNTFANAVK-----	379
H18	-----VDASTFANAVK-----	384
H10	-----LDASALKANQ-----S	278
H07	-----SKDVVLSETKAAA-----ATSS	400
H19	AGAAATAGNAAVG----ALGATAVDNTTADVADISISASQMASILQD-----KD	445
H55	-----TAADTAEA-----TATTTAVANTTADVSGVTISASQMASILKD-----KD	442
H45	-----SADTVLSTVQSAATANTAVTGAT	382
H34	-----GNTGTVKIEGATVSASAF-----	380
H44	----TTEDKTEIKLAGATVA---GQSGAIVVTGARI SAEAMQSATKT-----	408
H46	----TTADGATISLAGSANAA-TGTQSGAITLKNVRI SADALQSAAKG-----	401
H15	----TTEDGTNI-----DI---AANGNIGLTGVRI SADSLSATKS-----	391
H28	----TTEDETTFTLAGGTDA---TTSGAITVANARMSAESLSATKS-----	401
H49	-----VSSDALLAQVKADSTAN---NVK	381
H20	---ITTADNTSIVLNGSSDGTGNAGTEGTIAVTGAVIS SAALQSASK-----T	399
H31	---ITTADGTSIVLAGSGD-FGTTKTAGAINVTGAVISADALLSASK-----A	389
H37	-----E	383
H41	-----K	380
H30	-----A	392
H32	-----A	392
H42	-----	282
H05	-----	287
H56	-----	287
H48	-----	316
H14	-----	356
H12	-----VATAKQ GKAVAADGDTSA-----T	402
H01	-----VATAKQGTAVAANGDTSA-----T	402
Nissle	-----VATAKQGTAVAANGDTSA-----T	402
H09	NGNITAADGSALYLDATTTGGFTTTAGGN TAAS-----LDNLIANSK-----DAT	449
H52	-----	288
H33	-----	278
H39	-----	279
H51	-----LQSASQAAATATDGD TVA-----T	410
H43	-----	326
H26	-----T	382
H04	-----	219
H17	-----	219
H29	-----	284
H38	-----	287
H36	KSSITTESGTKIAIAGSTDGTT---AGKISATNVKISAE DL-----	384
H53	-----NQTFKSY-----	294
H27	-----KT-SDPVY-----	315
H11	-----IT-SETVY-----	312
H21	-----AP-TTALY-----	316
H23	-----AP-TTALY-----	316
H08	-----TT-DTAVY-----	317
H40	-----TT-DIDVY-----	317
H02	-----S---ITGAKL---FKLEDA-----	315
H35	-----ANATVNAT---TGEV TY---DSGT-AAA-----	337
H03	-----NDSQMSVY---VDGKNL---EIKQVLDA-----	348
H16	-----NDAPMSVY---FGGKNL---DVHQVQDT-----	347
H47	-----ADA-----	258
H54	-----AGNAKL---VKMTYTDS-----	335

H06	-----NDTYTATVGA-	389
H18	-----NETYTATVGV-	394
H10	L-----VVY-----K-----	283
H07	ITFNSGVLSKTI-----GF-----	414
H19	FT-----LS--DGSDTYNVTSSNAV	462
H55	FA-----LN--SGTTQYAVTGSTG	459
H45	IKYNTGIQ-----SATASFGG-	398
H34	-----TGIAY-----SAN--TGGNTYAVAANN-	400
H44	TGFTTGTTVAA-----NTGKVT-----	426
H46	TVINVDNGADDI-----SVSKTG--VV-----TT---	423
H15	TGFTVGTGATGL-----TVGTDG--K-----	410
H28	TGFTVDVGATGN-----SAGDIK--VD-----SKGI-	425
H49	ITFNNGPL-----SFTASFQ--	396
H20	TGFTVGTVDVT-----AGYIS--VGTDGSVQA----	423
H31	TGFTSGAY-----T--VGTDGVVK----	406
H37	-----LTGKAY-----TVV--NGAESYAVAT-N-	403
H41	-----LSGKAY-----TVA--NGAESYDVAAVT-	401
H30	-----LTGQAY-----TVA--NGAQSFVAAA-G-	412
H32	-----LTGQAY-----TVA--NGAQSYDAAA-D-	412
H42	---A-----ANSDGEAVGFAT--VQGKNEYITDGV-	307
H05	---A-----ADSDGSAAGYVT--FQKKNYATTVST-	312
H56	---A-----ADSDGAAAGYVT--FQKKNYATTAPA-	312
H48	-----TY-----ALKD--TNGNLYAADVNE-	334
H14	-----PNTITYSVTGAKVDQAAFDKAVST-----SGNDVD--FTTAGYSVDG---	396
H12	ITYKSGVQTYQA-----VFAA--GDGTASAKYAD--	429
H01	ITYKSGVQTYQA-----VFAA--GDGTASAKYAD--	429
Nissle	ITYKSGVQTYQA-----VFAA--GDGTASAKYAD--	429
H09	LTVTSGTQNTVYSTTGSQAQFTSLAKVDTVNVTNAHVSAEGMAN--LTKSNFTIDMGG-	506
H52	---TTNSAGA-----AVGYVT--IQGKDYLAGADG-	313
H33	---GADALGA-----AVGYVT--VQGNFKADAGA-	303
H39	---TSDANGA-----AVGYVT--IQGKNYQAGATG-	304
H51	INYKSGVMIG-----SATFTN--GKGTADGMTSGT-	438
H43	-----GLQN--DSGTIFATDYDG-	342
H26	LN-----N-----GSYTAN--VGGKAYAVTAG--	402
H04	-----T-----KYAVVDS--	227
H17	-----T-----KYAVVDS--	227
H29	-N-----D-----AVGFVT--VQKKNYVANDS--	303
H38	-G-----T-----PKAFVS--VQKKSFGIDDA--	306
H36	-----KAKADTAGFTTS----TGFTVAAGG-	405
H53	-----KKDNGELGYVVENADGTFNRANV-	317
H27	-----KNSAGQFTT-----TKV-	327
H11	-----KNSAGLYTT-----TKV-	324
H21	-----KNSAGQLTA-----TKV-	328
H23	-----KNSAGQLTA-----TKV-	328
H08	-----KNSAGQFTT-----TKV-	329
H40	-----KNSAGQFTT-----TKV-	329
H02	-----NGK-DTGSFAL--IGDDGKQYAANV-	337
H35	-----LGAGIEVGSALKEIAKPDAGVAVDL-	362
H03	-----DGKPKAGAFAAQT--ADGKSLAVNI-	371
H16	-----QGNPVPNSFAAKT--SDGTYIAVNV-	370
H47	-----	258
H54	-----NGKKVEGGYALKV--GDDYYAADY--	357

H06	-----KT---YSVTTGSAAA-----D-TAYMSNGVL-----SDTPPTY	418
H18	-----TLPATYTVNN--GTA-----A-SAYLVDGKV-----SKTPAEY	424
H10	-----DKSGNDAYIIQTKDVTNQST-----F	305
H07	-----TAGESSDAAKSYVDDKGGITNVADYT-VSY-SVNKDNGSVTVA--GYA---S	459
H19	TI-----NGKAANID-----DSGAITDQ--TSKVVNY	487
H55	AVTYDPDTPAATGDIVSAYVD-----DAGTLTDD--ANKTVKY	496
H45	-----VNTNGAGNSNDTYTDADKELTTASYSY-INY-NVDKDTGTVTVA--S-----N	442
H34	-----TTNGFLA-----GDDLTDQ--AQTVSTY	421
H44	-----IGGNQAYTQTDGTLA-----AKNETEI	448
H46	-----GGAPTYTDADGKLT-----TTNTVDY	444
H15	-----VTIGGTTAQSYTSKDGSLT-----TDNTTKL	436
H28	-----VQYTGTVFEDAYTKADGSLT-----TDNTTNL	453
H49	-----NGVSGSAASNAAYIDSEGLTSTESY-NTNY-SVDKDTGAVSV-----T	438
H20	-----YDAATSGNKASYTNTD-GTL-----T--TDNTTKL	450
H31	-----SGGNDVYNKADGTL-----T--TDNTTKY	429
H37	-----NTVKTADAKNVYVDASGKLTDDKATV-----TETY	435
H41	-----GAVTTAGNSFVYADADGKLTTSASNTV-----TQTY	433
H30	-----GAVTATTGGATVNIAGDELTTATNKTV-----TETY	444
H32	-----GAVTATTGGATVNIAGDELTTAANKTV-----TETY	444
H42	-----KNQS--TAA--PTDIA-----	319
H05	-----ALDDNTAAK--ATDNK-----	326
H56	-----ALNDDTTAT--ATANK-----	326
H48	-----TT-GAVSVKTIITYTDS SGAASS-----PTAV-KLGGDDGKTEVV--DIDGTY	378
H14	-----ATGAVTKGVAPVYIDNNGALTS-----DTVDF	424
H12	-----KADVSNATATYTDADGEMTTIGSY-TTKY-SIDANNGKVTV-----D	469
H01	-----NTDVSNAATATYTDADGEMTTIGSY-TTKY-SIDANNGKVTV-----D	469
Nissle	-----NTDVSNAATATYTDADGEMTTIGSY-TTKY-SIDANNGKVTV-----D	469
H09	-----T-----GTVTYTVSNGDVKAAA-----NA-DVYVEDGALSAN--ATKDVTY	544
H52	-----KDAIENGGDAATNEDTKIQLTDEL-----	337
H33	-----LVNSK--NAAGSQNVTSAI-----	320
H39	-----VDVLANSGVAAPTTAVDTGT-----	324
H51	-----TPVVATGAKA--VYVDGNNELTSTASY-DTTY-SVNADTGAVKV-----V	479
H43	-----TTVTPGAETVYKDasGN-----STTAAV-TLGGSDGKTNL-----V	379
H26	-----AVQTGGAD--VYKDTTGALTEDETVTAT--YGFADGKVS-----	440
H04	-----ATGK--YMEATV-V--ITG-----	241
H17	-----ATGK--YMAATV-V--ITS-----	241
H29	-----LVNANGAA--GAEATRVT--IDGDGTNQA-KIEL-----	332
H38	-----ALKNNTGD--A-TATQPG--TSGTTVVA--SIHLSTGK-----	338
H36	-----D-----QKATLNGSEAYVKDGGFTIDNT-----AKY	431
H53	-----DSKTG-----	322
H27	-----ENKAA-----TASDLLNNAKKGSSLVVNGADYEVVSADGKTVT---GLGRTM	372
H11	-----DNKAA-----TLDLNLNAAKKTGSTLVVNGATYDVSADGKTITETASGNKVM	373
H21	-----ENKAA-----TLDLNLNAAKKTGSTLVVNGATYDVSADGKTITETASGNKVM	377
H23	-----ENKAA-----TLDLNLNAAKKTGSTLVVNGATYDVSADGKTITETASGNKVM	377
H08	-----ENKAA-----TLDLNLNAAKKTGSTLVVNGATYDVSADGKTITETASGNKVM	378
H40	-----ENKAA-----TLDLNLNAAKKTGSTLVVNGATYDVSADGKTITETASGNKVM	378
H02	-----DQKTGAVSVKTMSTYTDADGVKHDNVKV-----ELGSDGKTEVV--ATDGKT	383
H35	-----TGVSG-----VTGAQ-----LFAKKDGSYVIK--GTAD--	389
H03	-----DG-NG-----NTSVVKDADGNVEWV-----VDKDGAAKTVV--RKDDKI	408
H16	-----DAATG-----NTSVITDPNGKAVEWA-----VKNDGSAQAIM--REDDKV	408
H47	-----AGD-----	261
H54	-----ESTSKTVTVRTTSYKDVDPQKGLN-----KIGGADGKTETV---TIGEKT	401

H06	YAQADGSIITTTEDAAAGKLVYKGS DGKLTDTT SKAESTSDPLAALDDAISQIDKFRS	478
H18	FAQADGTTITSGENAATSKAIYVSANGNLTTNTTSESEATTNPLAALDDAIASIDKFRS	484
H10	NA-----ANISDAGVLSIGASTTAPSNLTANPLKALDDAIASVDKFRS	350
H07	ATDTNKDYAPA----IGTAVNVNSAGKITTEETTSAGSATNPLAALDDAISSIDKFRS	515
H19	FAHTNGSVTND----TGSTIYATEDGSLTTDAATKAETTADPLKALDEAISSIDKFRS	543
H55	YAHTNGSVTND----SGSAIYATEAGKLTTEASTAAETTANPLKALDDAISQIDKFRS	552
H45	GAGATGKFAAT----VGAQAYVNSTGKLTTEETTSAGTATKDPLAALDEAISSIDKFRS	498
H34	YSQADGTVTNS----AGKEIYKDADGVYSTEN--KTSKTS DPLAALDDAISSIDKFRS	475
H44	FLQKDGSI TNN----SGKAVYVQEDGKFTTDAATKAATTADPLKALDDAISSIDKFRS	504
H46	FLQTDGGSVTNG----SGKGVYTDAAKGFTTDAATKAATT DPLKALDDAISQIDKFRS	500
H15	YLQKDGGSVTNG----SGKAVYVEADGDFTTDAATKAATT DPLKALDEAISQIDKFRS	492
H28	FLQKDGTVTNG----SGKAVYVSADGNFTTDAETKAATTADPLKALDEAISSIDKFRS	509
H49	GGSGTGKYAAN----VGAQAYVGADGKLTNTTSTG SATKDPLNALDEAIASIDKFRS	494
H20	YLQKDGGSVTNG----SGKAVYVEADGDFTTDAATKAATT DPLAALDDAISQIDKFRS	506
H31	YLQDDGGSVTNG----SGKAVYVDATGKLTTEATKAATTADPLKALDEAISSIDKFRS	485
H37	HEFANGNIYDD----KGAAVYAAADGSLTTEETTSKSEATANPLAALDDAISQIDKFRS	491
H41	HEFANGNIYDD----KGS SLYKAADGSLTSEAKGKSEATADPLKALDEAISSIDKFRS	489
H30	HEFANGNILDD----DGAALYKAADGSLTTEATGKSEVTTDPLKALDDAIASVDKFRS	500
H32	HEFANGNILDD----DGAALYKAADGSLTTEATGKSEAT DPLKALDDAIASVDKFRS	500
H42	-----QTI---DLDTADEFTGASTADPLALLDKAIAQVDTFRS	356
H05	-----VVVELSTAKPTAQFSGASSADPLALLDKAIAQVDTFRS	366
H56	-----VVVELSTATPTAQFSGASSADPLALLDKAIAQVDTFRS	366
H48	DSA-----DL---NGGNLQTLTAGG EALTAVANGKTTDPLKALDDAIASVDKFRS	428
H14	YLQDDGGSVTNG----SGKAVYKDADGKLTDAETKAATTADPLKALDEAISSIDKFRS	480
H12	SGTGTGKYAPK----VGAEVYVSANGTLTDDATSEGT VTKDPLKALDEAISSIDKFRS	525
H01	SGTGSGKYAPK----VGAEVYVSANGTLTDDATSEGT VTKDPLKALDEAISSIDKFRS	525
Nissle	SGTGTGKYAPK----VGAEVYVSANGTLTDDATSEGT VTKDPLKALDEAISSIDKFRS	525
H09	FEQKNGAITNS----TG GTIYETADGKLTTEATTASSSTADPLKALDEAISSIDKFRS	600
H52	-----DVDG SVKTAATATFSGTATNDPLALLDKAIASQVDTFRS	377
H33	-----GDIANKANANIYGTSSADPLALLDKAIASVDKFRS	358
H39	-----LQLSGTGTATELKG TATQNPPLALLDKAIASVDKFRS	362
H51	SGTGTGKFEAV----AGADAYVSKDGKLTTEETTSAGTATKDPLAALDAAISSIDKFRS	535
H43	TAA----DGKT---YGATALNGADLSDPNNTVKS VADNAKPLAALDDAIAMVDKFRS	431
H26	-----DG---EGSTVYKAADGSITKDATTKSEAT DPLKALDDAISQIDKFRS	487
H04	-----TAAAVTVGA AEVAGAATADPLKALDAAIAKVDKFRS	279
H17	-----TAAAVTVGATEVAGAATAEPLKALDAAIAKVDKFRS	279
H29	-----SQNGDTAATSEFAGASTNDPLTLLDKAIASVDKFRS	370
H38	-----NSVDADVTA STEFTGASTNDPLTLLDKAIASVDKFRS	377
H36	YVQEDGAI TNG----SGKVAYKDADGKLTDAKTETAKTTDPM AKLDKALAKVDALRS	487
H53	-----VVS-VG TKISTSPDVLATIDNALKIVDSQRS	354
H27	YLS----K--S----EGGSP-----ILVKEDA AKSLQSTTNPLETIDKALAKVDNLR	417
H11	YLS----K--S----EGGSP-----ILVNEDA AKSLQSTTNPLETIDKALAKVDNLR	418
H21	YLS----K--S----EGGSP-----ILVNEDA AKSLQSTTNPLETIDKALAKVDNLR	422
H23	YLS----K--S----EGGSP-----ILVNEDA AKSLQSTTNPLETIDKALAKVDNLR	422
H08	YLS----K--S----EGGSP-----ILVNEDA AKSLQSTTNPLETIDKALAKVDNLR	423
H40	YLS----K--S----EGGSP-----ILVNEDA AKSLQSTTNPLETIDKALAKVDNLR	423
H02	YSVS---DLQG---KSLK-----TDSIAAISTQKTEDPLA AIDKALSQVDSLRS	428
H35	-----NKE---VLF EAKVAADGKV----TKGDQLTADPLKS IDDAALSQVDQFRS	433
H03	YGAS---VT-G---FGGTPTVNVDTT AIDASELKGMTTAKPLEKLDALAKVDKLR	460
H16	YTAN---IT-N---KTATK-----GAELSASDLKALAT TNPLSALDEALAKVDKLR	455
H47	-----LTK-----TKVDEDATAATKTSNPLSKIDDAIS DVDSLRS	299
H54	YAAD---KLKD---HDF-----SKQATLGE EATTTVNPLDAIDKALAQVDSLRS	447

. . . : * * : * * **

Appendix 11: Crystallographic data and refinement statistic for the gC data sets.

	gC-nativ	gC-UO₂²⁺
Data collection		
Wavelength (Å)	1.0	2.075
Resolution (Å)	48.20 (2.83-2.67)	47.94 (3.20-3.12)
Space Group	179 (P6 ₅ 22)	179 (P6 ₅ 22)
Cell dimension		
a, c (Å)	134.77, 170.941	133.65, 171.00
Reflections	1036183 (155641)	9381339 (599176)
Unique	49334 (7899)	30424 (2234)
Redundancy	21.0 (19.7)	308.4 (268.2)
Completeness (%)	99.8 (98.9)	99.6 (98.5)
<i>I</i> / σ <i>I</i>	18.23 (1.01)	26.56 (0.27)
<i>R</i> _{meas}	16.8 (340.5)	55.3 (1035.0)
CC _{1/2}	99.9 (48.0)	99.9 (70.0)
Wilson B (Å ²)	77.2	89.6
SAD-phasing		
Phasing power		1.25 (1.00)*
R-Cullis		0.77 (0.81)*
FOM acentric		0.17 (0.33)*
<i>D</i> / σ <i>D</i>		1.86 (1.78)*
Refinement		
<i>R</i> _{work} / <i>R</i> _{free}	24.5/26.9	
No. atoms / B-value (Å ²)		
Protein	2573/79.1	
Carbohydrates	100/103.8	
RMSD angels (°)	1.029	
RMSD bonds (Å)	0.009	
Ramachandran favoured (%)	94.0	
Ramachandran outlier (%)	2.1	

* The anomalous signal was used to a resolution shell of 5.24-4.90 Å

Appendix 12: HSV-1 gC construct used in this thesis. Highlighted are the export signaling sequence (blue), the TEV cleavage site (green) and the Fc-tag (orange).

```

MAPGRVGLAV VLWSSLWLGA GVAGGETAS TGPVWCDRRD PLARYGSRVQ IRCRFRNSTR
MEFRLQIWRY SMGSPPIAP APDLEEVLTN ITAPPGLLV YDSAPNLDP HVLWÆGAGP
GADPPLYSVT GPLPTQRLII GEVTPATQGM YYLAWGRMDS PHEYGTVWRV RMFRPPSLTL
QPHAVMEGQP FKATCTAAAY YPRNPVEFVW FEDDRQVFNQ GQIDTQTHEH PDGFTTVSTV
TSEAVGGQVP PRTFTCQMTW HRDSVMFSRR NATGLALVLP RPTITMEFGV RHVVCTAGCV
PEGVTFAWFL GDDPSPAAS AVTAQESCDH PGLATVRSTL PISYDYSEYI CRLTGYPAGI
PVLEHHGSHQ PPPRPDTERQ VIEAIEGTEN LYFQGTHTCP PCPAPELLGG PSVFLFPPKP
KDTLMISRTP EVTCVVVDVS HEDPEVKFNW YVDGVEVHNA KTKPREEQYN STYRVVSVLT
VLHQDWLNGK EYKCKVSNKA LPAPIEKTIS KAKGQPREPQ VYTLPPSREE MTKNQVSLTC
LVKGFYPSDI AVEWESNGQP ENNYKTTTPV LSDGSFFLY SKLTVDKSRW QQGNVFCSV
MHEALHNYHT QKSLSLSPGK

```

Appendix 13: DALI server 3D structure comparison results. The ten best matching structures from comparison of gC_D1 with the PDB25, a not redundant representative subset of the PDB, used for a Ca-rmsd plot.

Pdb-chain	Rmsd [Å]	Description
5lfu-A	2.3	MYELIN-ASSOCIATED GLYCOPROTEIN
6p11-A	2.1	ZWEI IG DOMAIN PROTEIN ZIG-8
6o89-H	2.3	ANTI-CD28XCD3 CODV-FAB HEAVY CHAIN
4i2x-E	2.0	FABOX117 LIGHT CHAIN
6eg1-A	2.0	DEFECTIVE PROBOSCIS EXTENSION RESPONSE 2, ISOFORM
1igf-M	2.0	IGG1-KAPPA B13I2 FAB (LIGHT CHAIN)
4f91-A	3.1	BUTYROPHILIN SUBFAMILY 3 MEMBER A1
3fn3-A	2.8	PROGRAMMED CELL DEATH 1 LIGAND 1
1hxm-B	2.4	GAMMA-DELTA T-CELL RECEPTOR
6opa-S	2.4	ENVELOPE GLYCOPROTEIN GP160

Appendix 14: DALI server 3D structure comparison results. The ten best matching structures from comparison of gC_D2 with the PDB25, a not redundant representative subset of the PDB, used for a Ca-rmsd plot.

Pdb-chain	Rmsd [Å]	Description
2pet-A	2.3	LUTHERAN BLOOD GROUP GLYCOPROTEIN
6lx3-C	2.2	INTERLEUKIN-2, IMMUNOGLOBULIN HEAVY CONSTANT ALPHA
4i2x-E	2.9	FABOX117 LIGHT CHAIN
4of8-A	2.5	IRREGULAR CHIASM C-ROUGHEST PROTEIN
1i3r-B	2.5	H-2 CLASS II HISTOCOMPATIBILITY ANTIGEN, E-K ALPH
6cr1-H	2.3	LIGHT CHAIN OF ADALIMUMAB EFAB (VL-IGE CH2)
2wqr-B	2.2	IG EPSILON CHAIN C REGION
3so5-B	2.9	LEUCINE-RICH REPEATS AND IMMUNOGLOBULIN-LIKE DOMA
6x4g-C	2.3	INDUCIBLE T-CELL COSTIMULATOR
3oq3-B	2.3	INTERFERON ALPHA-5

Appendix 15: DALI server 3D structure comparison results. The ten best matching structures from comparison of gC_D3 with the PDB25, a not redundant representative subset of the PDB, used for a Ca-rmsd plot.

Pdb-chain	Rmsd [Å]	Description
6x4g-C	2.0	INDUCIBLE T-CELL COSTIMULATOR
5opi-C	2.5	H-2 CLASS I HISTOCOMPATIBILITY ANTIGEN, D-B ALPHA
3fn3-A	3.3	PROGRAMMED CELL DEATH 1 LIGAND 1
1i3r-B	2.3	H-2 CLASS II HISTOCOMPATIBILITY ANTIGEN, E-K ALPH
4i2x-E	2.1	FABOX117 LIGHT CHAIN
4of8-A	2.5	IRREGULAR CHIASM C-ROUGHEST PROTEIN
1igf-M	2.5	IGG1-KAPPA B13I2 FAB (LIGHT CHAIN)
4f9l-A	2.5	BUTYROPHILIN SUBFAMILY 3 MEMBER A1
5zo2-A	2.8	CELL ADHESION MOLECULE 4
6cr1-H	2.3	LIGHT CHAIN OF ADALIMUMAB EFAB (VL-IGE CH2)

VIII. Acknowledgements

Thilo, vielen Dank für die Unterstützung und die Freiheit bei der Gestaltung der eigenen Projekte. Deine positive und offene Sichtweise ist Inspiration.

Dirk vielen Dank das du dich bereit erklärt hast Gutachter dieser Arbeit zu sein.

Für wissenschaftliche Diskussionen und Unterstützung möchte ich Melanie Dietrich, Bärbel Blaum, Luisa Ströh, Elena Ostertag und Georg Zocher danken.

Für nicht wissenschaftliche Diskussionen und Unterstützung möchte ich ebenfalls Melanie Dietrich, Bärbel Blaum, Luisa Ströh, Elena Ostertag und Georg Zocher, sowie Christina Harprecht, Michael Buch und Stefanie Ott danken.

Außerdem möchte ich allen Studenten danken die ich betreuen durfte:

Kalmurat Zinur, Maria Wahle, Philip Rössler, Raika Sieger, Jasmin Kuhn, Malena Frick, Andriko von Kugelgen, Katarzyna Glowacz, Raed Shalaby, Melanie Maier, Ruoshi Zhang.

Für die Arbeit an gemeinsamen Projekten möchte ich Johannes Heidrich, Michael Buch und Thomas Hagemann sowie Julia-Stefanie Frick, Tomas Bergström und Frank Böckler danken.

Julia Sindlinger danke ich für die MS-Versuche.

Besonderer Dank gilt Melanie Dietrich, Elena Ostertag, Georg Zocher, Melanie Maier und Stefanie Ott.

IX. Publications

Michael B. Braun, Bjoern Traenkle, Philipp A. Koch, Felix Emele, Frederik Weiss, Oliver Poetz, Thilo Stehle & Ulrich Rothbauer *Peptides in headlock--a novel high-affinity and versatile peptide-binding nanobody for proteomics and microscopy*. Scientific Reports, 2016. **6**: p. 19211.

Natalya Belousova, Galina Mikheeva, Chiyi Xiong, Loren J. Stagg, Mihai Gagea, Patricia S. Fox, Roland L. Bassett, John E. Ladbury, Michael B. Braun, Thilo Stehle, Chun Li, Victor Krasnykh *Native and engineered tropism of vectors derived from a rare species D adenovirus serotype 43*. Oncotarget, 2016. **7**(33): p. 53414-53429.