Structure and Function of Essential Components of Contractile Injection Systems

PAR

Sergii BUTH

EPFL

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2014

To the barbell at the university sports center…

ACKNOWLEDGEMENTS

Firstly I would like to thank my supervisor, Petr Leiman for giving me an opportunity to do my Ph.D thesis in his laboratory and contribute to very interesting projects. After 4.5 years of doctoral research, contractile injection systems became dear to my heart. And not only because they are of a high scientific interest from a structural, evolutional and microbiological point of view but also of their putative medical application as an antibacterial treatment. I am very grateful to Petr for his continuous guidance, advices and support. The experience received, competences and skills acquired under his supervision are priceless. Also I appreciate very much the amount of freedom he allowed me to have in all aspects of work. I would also like to thank him for an opportunity to attend a number of local and international conferences, especially the Regio meetings on biocrystallograpgy, Phage Virus Assembly and FASEB Virus structure and Assembly meetings. These events allowed communicating with the experts from oversees and get feedback on my results thus were very motivating and productive.

I am grateful to my thesis jury members who came back with a positive answer to my call – it was a challenge to assemble a jury for the end of August. Prof. Cristian Heinis, Dr. Florence Pojer, Dr. Jochen Klumpp and Dr. Vincent Olieric thank you very much.

Now I would like to appreciate a person without whom I am not sure I would have succeeded with the doctorate. Chris Browning, who was not only a very good colleague with such important traits for a colleague as patience, responsibility, ability to hear and understand a problem, willing to come and help anytime anywhere and, finally, good sense of humor and music taste (hail to the HardRadio, keep it loud!), but became a great friend of mine with a brilliant set of hobbies and interests, some of which we shared. Thanks a lot for all the friendship and support I received during all this time. Petr, thanks a lot for hiring us at the same time!

The results reported here would not be possible without an extensive contribution of Mikhail Shneider. I want to express my gratitude for all the work with construct engineering, cloning, protein expression and purification he has done for our common projects. I also appreciate all those scientific discussions and insights into the research field he provided. Thanks to his hobbies such as photography, history and the curiosity of discovering the surroundings I have dramatically broaden my horizons about Switzerland and its closest neighbors.

I am thankful to all other lab/floor/institute/EPLF members with whom I have being working all this time. It was a wonderful time thanks to you, guys!

ABSTRACT

My Doctoral thesis is dedicated to the structural and functional characterization of contractile injection systems, a recently defined class of multicomponent biological assemblies that is based on common ancestry of its members: bacteriophages with contractile tails, R-type pyocins, type VI secretion system (T6SS), *Photorhabdus* virulence cassette and the *Serratia entomophila* antifeeding prophage. The accumulating body of evidence suggests that the structure of the main components of all contractile tail-like systems is conserved, thus resulting in a similar morphology and mechanism of action. The aim of this work was to characterize features common to all the class members by means of modern bioinformatic tools and to determine the structure of the key proteins/protein complexes for some contractile systems.

This structural information can be of a great importance for theoretic applications: to prove bioinformatic findings and to use the resolved structures as models to characterize other members of the class. On the other hand, structural findings about the membrane-puncturing needles and the tail fibers employed by all contractile systems and extensively studied in this work, are crucial for understanding of such fundamental biological processes as virus attachment to the host cell, cell envelope penetration, and translocation of large protein complexes across lipid membranes. This knowledge is key to the development of new methods for antibacterial therapy that is required for the treatment of diseases caused by antibiotic resistant bacteria.

R-type pyocins that are used by Pseudomonas aeruginosa in intra-species competition were of great interest in this research work for two reasons: 1) their potential medical application for antibacterial treatment; 2) being the simplest members of the contractile ejection system family they perfectly match the role of a model system. R2 pyocin structure was characterized by means of CryoEM and X-ray crystallography. A CryoEM reconstruction of the pyocin baseplate and a tail fragment was calculated. The structures of the R2 and R1 pyocin receptor binding fibers (which are disordered in CryoEM reconstructions) were determined by X-ray crystallography thus completing structural characterization of the host-binding part of the pyocin particle.

The structure of the T4 cell-puncturing device fragment was characterized in great detail, revealing interesting and previously unknown features such as, for example, the presence of fatty acid molecules inside the beta-helical domain of gp5. The identity of the mysterious cell-puncturing device capping protein, whose existence was predicted more than 10 years ago, was determined, and its crystal structure has been solved.

The knowledge obtained from studying phage and pyocin structure and function was applied to the type VI secretion system leading to the discovery and characterization of the exceptional role of PAAR-repeat proteins which function to sharpen and diversify the central spike of the T6SS organelle. These findings allowed to propose new, improved model of T6SS machinery and shed light on previously unknown mechanisms of the T6SS effectors' translocation the into the prey cells.

Keywords: type VI secretion system, T6SS, R-type pyocin, bacteriophage, T4, cell-puncturing device, contractile injection systems, CryoEM, X-ray crystallography.

RESUME

Mon travail de thèse de doctorat a été dédié à la caractérisation structurale et fonctionnelle des systèmes d'injection contractiles, une classe récemment définie, basée sur l'ascendance commune de ses membres: les bactériophages à queue contractiles, pyocins de type R, le système de sécrétion de type VI (SST6), cassette de virulence de *Photorhabdus* et le *Serratia entomophila* antifeeding prophage. La structure des composantes principales de tous les systèmes avec une queue contractile est conservée, ce qui se traduit par une morphologie et un mécanisme d'action similaire. Il était important de caractériser tous les traits communs pour tous les membres de ce groupe, employant des outils bioinformatiques modernes ainsi que de déterminer la structure des composantes clés pour certains des membres.

Cette information structurale peut avoir une grande importance du côté théorique : pour prouver les résultats bioinformatiques et pour utiliser les structures déterminées en tant que modèles pour caractériser d'autres membres de la classe. D'autre part, les découvertes concernant la structure des aiguilles de perforation de la membrane et des fibres de queue, présentées dans cette thèse, sont cruciales pour la compréhension du processus vital et hautement complexe **de l'**infection de bactéries par des bactériophage. Il en est de même pour l'attachement à la cellule hôte et la pénétration de l'enveloppe cellulaire. Cette information est la clé pour le développement des nouvelles méthodes de thérapie antibactérienne comme le traitement par thérapie bactériophage des maladies causées par des bactéries résistantes aux antibiotiques lesquelles sont actuellement en émergence. En outre, cette connaissance peut être étendue sur la relation bactérie-pyocin qui permet de développer les pyocins comme une nouvelle approche dans la conception de médicaments antibactériens.

Les pyocins de type R ont été d'un grand intérêt pendant cette recherche, non seulement pour leur application médicale potentielle pour le traitement antibactérien, mais aussi parce qu'ils sont les membres les plus simples de la classe et répondent parfaitement au rôle du système de modélisation. La structure de la pyocin R2 a été caractérisée par des moyens de microscopie électronique et cristallographie aux rayons X. Une reconstruction CryoEM de la semelle de pyocin et un fragment de la queue ont été calculé. Les structures des fibres de pyocin R1 et R2 (qui sont mal définis dans la carte CryoEM) ont été déterminées par cristallographie aux rayons X et caractérisées.

La structure du fragment du dispositif cellulaire de ponction du bactériophage T4 a été caractérisée en détail, révélant des caractéristiques intéressantes jusqu'alors inconnues. La

structure et l'identité de la protéine qui couvre le dispositif cellulaire de perforation du bactériophage T4, dont l'existence a été prédite il y a plus de 10 ans mais identifiée seulement maintenant, a finalement été dévoilée.

Les connaissances accumulées dans le domaine du phage T4 ont été utilisé pour comprendre le système de sécrétion de type VI et ces notions ont conduits à la découverte et à la caractérisation du rôle exceptionnel des protéines PAAR dans le SST6 qui aiguisent et diversifient sa pointe. Ces découvertes permettent de proposer maintenant un modèle amélioré de la machinerie du SST6 et éclaire d'un jour nouveau les mécanismes jusqu'ici inconnus de la translocation d'effecteurs de SST6 dans les cellules scibles.

Mots-clés: système de sécrétion de type VI, SST6, pyocin de type R, bactériophage, dispositif cellulaire de ponction, T4, systèmes d'injection contractiles, CryoEM, cristallographie aux rayons X.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

μl	microliter
3D	three-dimensional
Å	Angstrom
SLS	Swiss Light Source
bp	base pair
CryoEM	electron cryo-microscopy
CTF	contrast transfer function
Da	Dalton
DNA	deoxyribonucleic acid
E.coli	*Escherichia coli*
EDTA	ethylene diamine tetraacetic acid
EM	electron microscope/microscopy
ESI	electrospray ionization
EC	collision energy
eV	electron volt
gp	gene product
His-tag, His6	histidine reach tag
IPTG	isopropyl-β-D-thiogalactopyranoside
kb	thousand of nucleic acid bases
kbp	thousand of nucleic acid base pairs
kDa	thousand of Daltons
LPS	lipopolysaccharide
LB	Luria-Bertani broth
MAD	multiwavelength anomalous diffraction
mM	milimole
MR	molecular replacement
MS	mass spectrometry/mass spectrometer
MW	molecular weight
NCS	non-crystallographic symmetry
nm	nanometer

OB-fold   oligonucleotide/oligosaccharide binding fold

PAGE   polyacrylamide gel electrophoresis

PDB   Protein Data Bank

phage   bacteriophage

RMSD   root mean square deviation

RB/RBP   receptor binding/receptor binding protein

SAD   single wavelength anomalous diffraction

SDS   sodium dodecyl sulfate

Se   selenium

SeMet   selenomethionine

TEM   transmission electron microscope

Tris   tris(hydrohymethyl)-aminomethane

LIST OF TABLES

LIST OF FIGURES

1   INTRODUCTION. OVERVIEW OF CONTRACTILE INJECTION SYSTEMS

R-type pyocins, bacteriophages with contractile tails, type VI secretion system (T6SS) of Gram-negative bacteria, the *Photorhabdus* virulence cassette and the *Serratia entomophila* antifeeding prophage belong to the recently defined class of contractile injection systems based on common ancestry (Bönemann, Pietrosiuk and Mogk 2010), (Leiman and Shneider, Contractile tail machines of bacteriophages 2012), (Pukatzki, et al. 2007) (Figure 1.1).
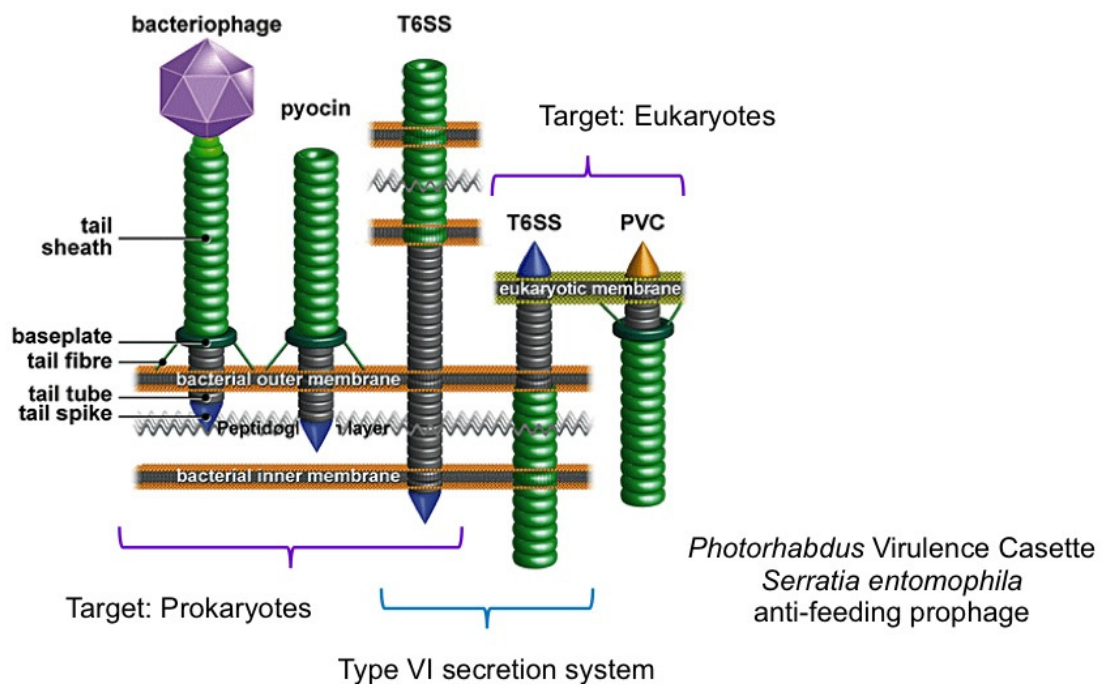
Figure 1.1 Organization of the contractile injection systems class.
Key members of the contractile injection systems class with identification of their targets and main components. The figure adapted from (Bönemann, Pietrosiuk and Mogk 2010).

A contractile injection system contains (at a minimum) a baseplate, a tail tube, an external contractile sheath and a central tail spike.

The baseplate contains proteins responsible for host cell recognition and attachment. During phage/pyocin attachment to the host cell baseplate changes its conformation (Kostyuchenko, et al. 2003), which triggers contraction of the sheath resulting in disruption of the cell envelope by the central spike protein and insertion of the tail tube into the prey cell. Similarities between phages with contractile tails and pyocins will be discussed in detail in chapter **2** of this thesis.

Analysis done in (Leiman, Basler, et al. 2009) revealed structural and functional similarity of the Hcp (the *h*emolysin-*c*oregulated-*p*rotein) protein from T6SS to the bacteriophage T4 gp19 tail

tube protein. Bioinformatic analysis showed homology between the T6SS VgrG (*valine-glycine repeat protein G*) proteins and T4 gp5-gp27 protein complex, also known as T4 cell-puncturing device (Pukatzki, et al. 2007). One of the smallest T6SS proteins *E.coli* c3402 was found to be homologous to gp25 in T4 – one of the most conserved proteins in all phages with contractile tails. These bioinformatic findings were further confirmed by *Pseudomonas aeruginosa* VgrG – T4 gp5-gp27 and *Pseudomonas aeruginosa* Hcp1 – T4 gp19 structure comparison (Leiman Lab, unpublished data), (Leiman, Basler, et al. 2009).

Based on all these findings and on extensive bioinformatic analysis (Leiman and Shneider, Contractile tail machines of bacteriophages 2012) proposed the model of the structure of the simplest contractile tail (Figure 1.2). The model represents a minimal set of the proteins, required to construct a contractile injection system that functions as a bacteriophage tail.
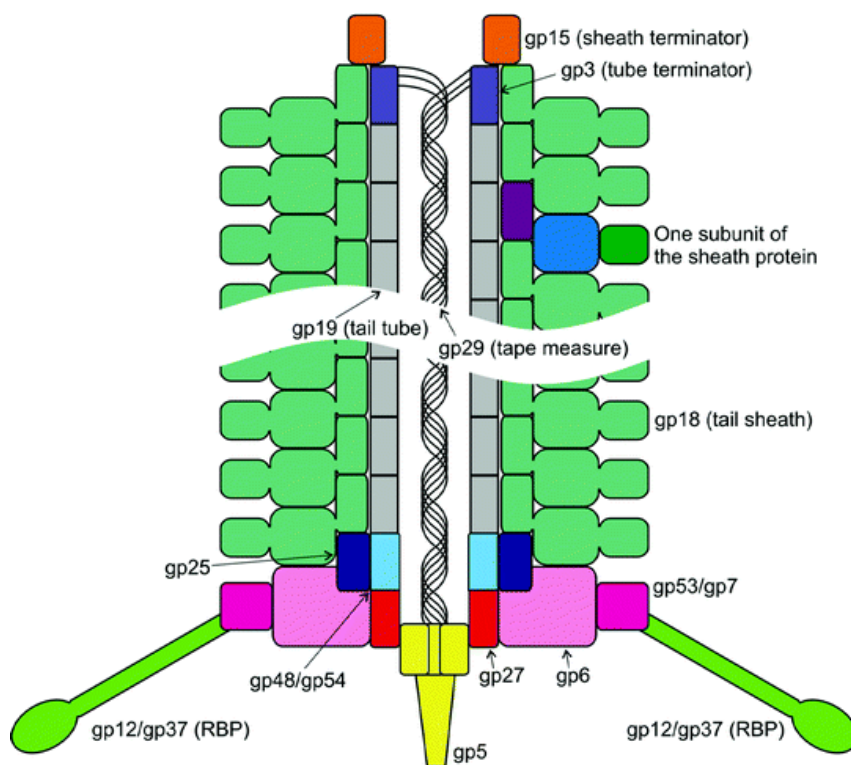
Figure 1.2 Structure of the simplest contractile tail.
The proteins are labeled with the names of their T4 phage orthologs and colored in distinctive colors. Figure adapted from (Leiman and Shneider, Contractile tail machines of bacteriophages 2012).

According to Leiman and Shneider (Leiman and Shneider, Contractile tail machines of bacteriophages 2012) a minimal baseplate is composed of gp6, gp25, and gp53. Proteins that attach the baseplate to the host cell (also called receptor binding proteins, RPB) emanate from the baseplate. A contractile ejection system also contains a set of proteins that required for the

assembly of its middle part: tail tube initiator, tape measure protein, tail terminator complex, tail tube and the sheath.

The findings summarized above show that the structure of main components of all contractile tail-like systems is conserved, resulting in a conserved morphology and a common mechanism of action. Moreover, Leiman and Shneider (Leiman and Shneider, Contractile tail machines of bacteriophages 2012) came to the conclusion that the presence of just two proteins – orthologs of T4 gp25 and gp27 – in a phage-like cluster of genes is a required and sufficient condition to define this cluster as a contractile tail relative.

2 STRUCTURAL STUDIES OF R-TYPE PYOCINS

2.1 Introduction and objectives

At least 500 million acute infections per year require antibiotic therapy. Under this selective pressure, genes responsible for antibiotic resistance are spreading through all types of common bacterial pathogens at an ever increasing rate. Recent CDC and WHO reports show that more than 60% of *Staphylococcus aureus* that cause infections are methicillin resistant (MRSA), more than 30% of *Enterococci* are vancomycin resistant (VRE), and more than 30% of *Pseudomonas aeruginosa* are fluoroquinolone resistant (FQRP). The current high rate of antibiotic resistance acquisition by bacterial pathogens makes development of new antibiotics financially infeasible for private companies. Instead, several natural agents that can lyse bacteria are being investigated as antimicrobials. Many of them are derived from bacterial viruses (bacteriophages). In particular, endolysins (phage lytic enzymes) are extremely effective against a broad range of Gram-positive bacteria (e.g. *S. pneumoniae*) (Fischetti 2006), (Hermoso, Garcia and Garcia 2007), whereas phages and their virion components are being tested as antimicrobials against Gram-negatives (Watanabe, et al. 2007), (Scholl and Martin 2008), (Williams, et al. 2008) and (Scholl, Cooley, et al. 2009).

Many aspects of bacteriophage-bacteria-eukaryotic organism interactions have to be addressed before phage becomes a reliable therapeutic tool. Many phages carry toxin genes (Herold, Karch and Schmidt 2004), (Hayashi, Baba, et al. 1990) and, due to the diversity of phage genomes, about 50% of genes in well-studied phages have unknown functions and potentially can be toxic (Miller, et al. 2003). Therefore DNA-free phage components such as tails and enzymes involved in host cell recognition and attachment (RBPs) represent good candidates as new antimicrobials. It is essential to understand the function and structure of these components at the atomic level of detail, similar to the level of knowledge attained today for small molecule antibiotics.

One of the promising DNA-free bactericidal agents is the pyocin, a natural killer of the important human pathogen *Pseudomonas aeruginosa.* The pyocin DNA resides silently in the bacterial genome until it is activated with a UV or mitomycin C treatment (Nakayama, et al. 2000). Assembled particles of R- and F-type pyocins consist of about 10 different protein species in multiple copies and resemble contractile and non-contractile phage tails, respectively. A pyocin-lysogenized bacterium can produce up to 200 pyocin particles that leave the host by causing its lysis. Each of the newly released particles can bind to a bacterium and form a pore in its envelope causing depolarization of the cytoplasmic membrane and resulting in host death.

Like phages, pyocins are very potent bacteriocins and one pyocin particle is enough to kill a bacterial cell. However, pyocins have a very narrow host spectrum, which is limited to several *P. aeruginosa* species or even strains. Therefore, tools for manipulating the pyocin's killing spectrum are required in order to make it into a tunable antibacterial tool. This is impossible to achieve without detailed bioinformatic and functional characterization of the pyocin particle, and phage is instrument in this analysis.

The relationships between bacteriocins and phages have been considered since 1952 (Jacob and Wollman 1952, Kabsch 2010). The physiological similarities, such as the mutagenic agents induction, bactericidal action and the necessity for a receptor were observed. But, at the same time it was noticed that the pyocins do not replicate themselves in the sensitive cells. Later, the electron microscopy investigations showed that that R-type pyocins resemble the contractile tail of bacteriophages (Ishii, Nishi and Egami 1965) whereas the F-type pyocins are similar to the non-contractile bacteriophage tails (Takeya, et al. 1967).

These observations were taken as a starting point in the search for possible phage ancestors of these pyocins. It was found that some phages are neutralized by antipyocin sera: phages PS3 (Ito and Makoto 1970), PS17 (Kageyama, et al. 1980), UCTX (Hayashi, Matsumoto, et al. 1994) are neutralized by anti-R sera, and phage KF1 by anti-F sera (Kuroda, Kagiyama and Kageyama 1983), (Nakayama, et al. 2000). Also were found sequence homologies between pyocins and phages.

The analysis of the DNA sequence of R2/F2 pyocins gene cluster of the PAO1 strain of *Pseudomonas aeruginosa* (the region between *trpE* and *trpGCD* genes) containing R2 (16 ORFs) and F2 (16 ORFs) genes with those of phages P2 and λ, respectively was reported (Figure 2.1). Among the 16 ORFs from the R2 pyocin-specific region, 12 ORFs show significant similarities with the tail genes of P2 phage. And 8 ORFs of the F2 pyocin region show similarities with the tail genes of λ phage, moreover, the order of these genes was conserved.

These observations demonstrate that R2-type pyocins are relatives of the P2 phage and F2-type of λ phage. The G + C content of the sequences of R- and F-types pyocin genes are very similar to that of the P. aeruginosa chromosome (66.6%). However some ORFs corresponding to the tail fibers are AT rich, which could mean that they may come from other genomes. This observation suggests that R- and F-type pyocins became evolutionarily specialized as bacteriocins, rather than being simple remnants of phages (Nakayama, et al. 2000).
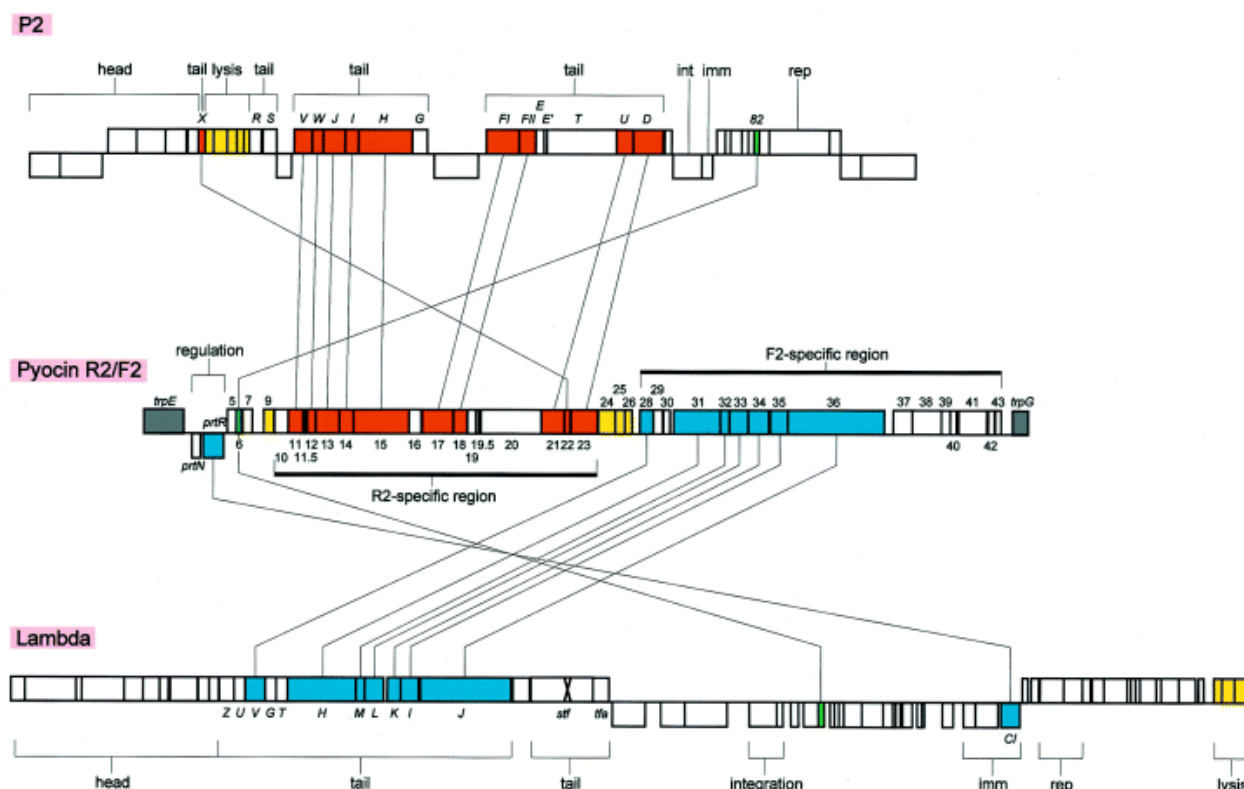
Figure 2.1 Comparison of genetic organization between the R2/F2 pyocin gene locus and the P2 and λ phage genomes.
Figure adapted from (Nakayama, et al. 2000).

Host specificity of phages (and, therefore, the pyocins) is determined by the tail fibers and/or tailspikes, which appear to have a set of conserved, N-terminal, tail-binding domains and a rather wide pool of C-terminal, host cell receptor-binding domains. Phages expand their host range by exchanging these modules, although these events are rare because they result from non-homologous recombination. Therefore, non-homologous recombination can be exploited in designing new pyocin fibers provided the structure or at least the borders of the N- and C-terminal domains of the fiber are known precisely. It should be possible to engineer tail fibers and tailspikes with any desired host specificity that can be used to complement fiberless pyocins. This can be achieved by fusing the N-terminal domain of the pyocin fiber with the C-terminal receptor-binding domain of the fiber from the phage with the desired host specificity.

Pioneering practical reports of these ideas were presented relatively recently by (Williams, et al. 2008), (Scholl, Cooley, et al. 2009). In the first work, using the above described techniques of fusing the C terminal part of the P2 tail fiber to the N terminus of R2 pyocin tail fiber (gene *PA0620* or *prf15*), they have changed the spectrum of R2 pyocin to kill *Escherichia coli* strain C and several uropathogenic *E. coli* (UPEC) strains, which are targeted by P2. By a similar approach they have created pyocins that kill *Yersinia pestis*.

It is now clear that phage fibers require a chaperone 'module', which is encoded at or near the C terminus of the fiber. The module might be part of the fiber and form its C-terminal domain, or can comprise a separate gene product, which is located next to the 3' end of the fiber gene in the genome. Scholl *et al*. found that pyocin and phage tail fibers often require the cognate chaperone for proper folding, i. e. the folding chaperone has some specificity towards its folding partner. Scholl *et al.* also found that for some fusion constructs, the cognate chaperone alone was not sufficient for fiber attachment to the particle, and the original pyocin fiber chaperone gp16 (PA0621) is also required (unpublished data). These results show an importance of the inclusion of the pyocin and/or phage tail fiber/spike chaperone into the expression construction.

Another important fact about the fusion of the N-terminal tail-binding domains with the C-terminal host cell receptor-binding domains is that the initial attempt of replacing the R2 tail fiber with the P2 tail fiber done by Scholl and co-authors was unsuccessful. This construct did not result in a functional pyocin particle even when the P2 chaperone gene was coexpressed. The problem was in the impossibility to predict any particular junction that might be suitable for forming properly folded chimeric tail fibers using secondary or tertiary structural analysis programs. However, the solution was found by fusing a large number of different N-terminal parts of R2 *prf15* to different portions of gene H encoding the C-terminal part of the P2 tail fiber. The analysis of a large series of fusions between R2 *prf15* and gene H of P2 led to the identification of a tentative 'universal' acceptor site on the prf15 gene, located near residue 140.

In the second work (Scholl, Cooley, et al. 2009), the authors focused on changing of the pyocin specificity by using the tail spike of phage φV10, which belongs to the *Podoviridae* family and infects the pathogenic *E. coli* strain O157. The reason behind this was to expand the range of target bacteria, particularly to strains that produce thick surface polysaccharide structures, such as, for example, *E. coli* O157. The AVR2-V10 chimeric pyocin was created, which retains the catalytic O-antigen-degrading activity of the donor phage tail spike, a peculiarity not known in any natural phage tail-like bacteriocin. The chimeric fiber was created using the fusion sites close to that found in the previous work. Different C-terminal portions of the φV10 tail spike were tried, and it was noted that not all N-terminal domains, which worked well before, resulted functional particles.

Fusions that gave active pyocin particles were tested on bacterial lipopolysaccharides (LPS). Incubation of LPS with AVR2-V10 resulted in the degradation of LPS. Another important fact was discovered during this work. The Shiga toxin production, which is often induced by the antibacterial antibiotic treatment, was not increased during the pyocin treatment at any

concentration of the AVR2-V10 particles, suggesting that pyocins might have therapeutic applications in the future. A schematic showing different steps in the engineering and retargeting procedure of the pyocins is shown in Figure 2.2.
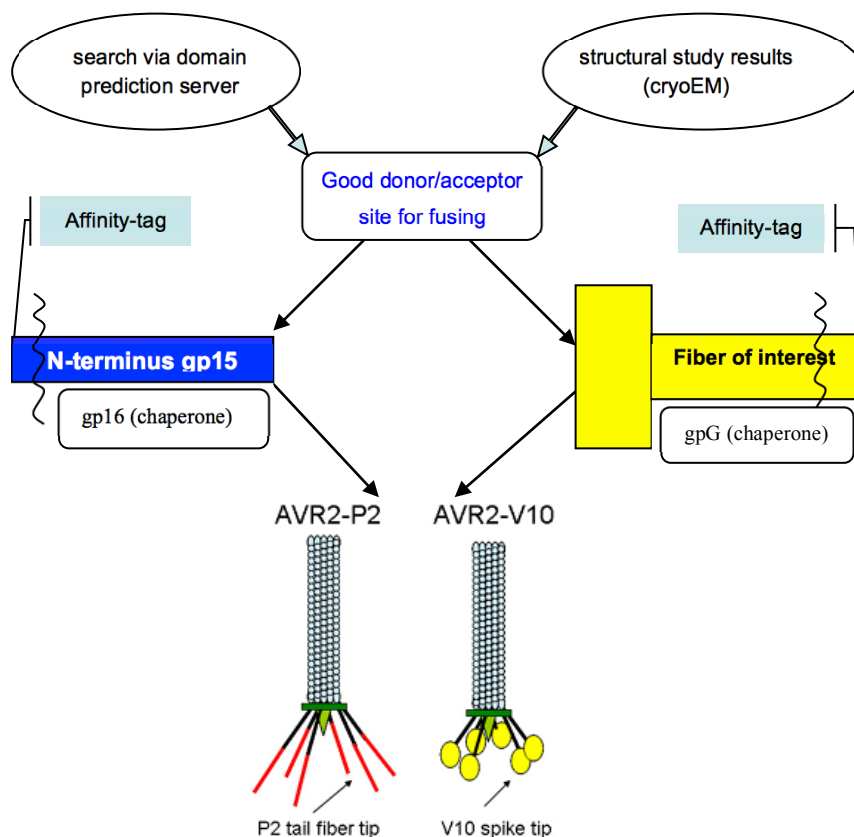
Figure 2.2 Schematic representation of the pyocin structural studies plan, leading to engineering of the pyocins with desired host specificity.
AVR2-P2 and AVR2-V10 successfully created chimeric pyocin particles, active in a killing way (Scholl, Cooley, et al. 2009).

One of the most important observations characterizing all pyocins carrying foreign fibers is that these chimeras are unstable and lose their killing capacity upon storage much faster than the wild type pyocin. Thus, the trial-and-error methods for creating new fibers that is derived from bioinformatic studies is a good proof-of-the-concept technique, but its applicability is limited when a technologically and medically relevant product is required.

One of the main goals of this thesis was to obtain structural data about the overall organization of the pyocin particle and detailed information on the structure of the pyocin fiber that could facilitate design of stable chimeras. The overall structure of the pyocin was studied by cryo-electron microscopy. The atomic structure of the receptor-binding part of the fiber was determined by X-ray crystallography.

2.2   CryoEM reconstruction of the R2 pyocin

Pyocin sample for data collection was prepared by Dr. Dean Scholl (Avidbiotics Corp., S. San Francisco, USA), data collection and micrographs scanning were done in the laboratory of Prof. Hong Zhou (UCLA). My task was to process raw micrograph data, box the particles and to calculate a 3D model.

2.2.1   Materials and methods

The micrographs of R2 pyocins were recorded at the Electron Imaging Center for NanoMachines (EICN) directed by Prof. Hong Zhou on a Titan Krios 300 kV FEG transmission electron microscope (TEM). Data was acquired with following parameters: aberration 2.86, magnification x59000, pixel size 2.14 Å/pix. Scanned digitized micrographs were binned 2x2 resulting into 4.28 Å/pix pixel size to reduce the impact of read noise on the signal to noise ratio (SNR). For binned micrographs contrast transfer function (CTF) parameters were found using CTFFIND (Mindell and Grigorieff 2003). Binned micrographs were Wiener filtered for noise suppression and better visibility of the pyocin particles with simultaneous application of earlier found CTF parameters. This step together with binning was crucial as CryoEM images are known for producing low specimen contrast (Figure 2.3). Particle picking was done manually due to the very low contrast produced by pyocin particles. BOXER from Eman program package with 192 voxel size was used (Ludtke, Baldwin and Chiu 1999). The data was further processed with the SPIDER software package (Frank, et al. 1996). A total of 2017 particles were selected for reconstruction. A geometric object consisting of solid cylinders was used as the initial model (Figure 2.4 **a)**). The structure resulted after the first reconstruction round (Figure 2.4 **b)**) was used for reference projection creation followed by particle class averaging. Additional noise reduction was applied at the latest reconstruction rounds. The final contrast transfer function corrected reconstruction has a resolution of 28 Å, based on the Fourier Shell Correlation (FSC) with a cutoff value of 0.5. Alternatively used FSC cutoff of 0.143 gave the resolution of 23 Å (Figure 2.5). FSC was calculated by Sergey Nazarov after 60 iterations in FREALIGN (Grigorieff 2007).
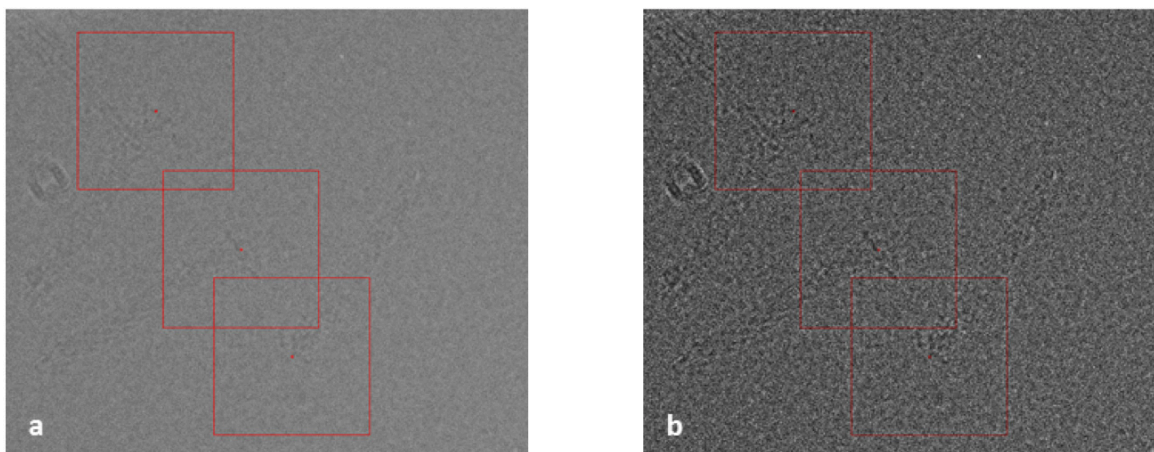
Figure 2.3 Fragment of the pyocin microphotograph binned 2x2 and Wiener filtered.
a) original "working" micrograph fragment; **b)** fragment **a)** modified in graphic editor. Red boxes show the reconstructed part of the particle.
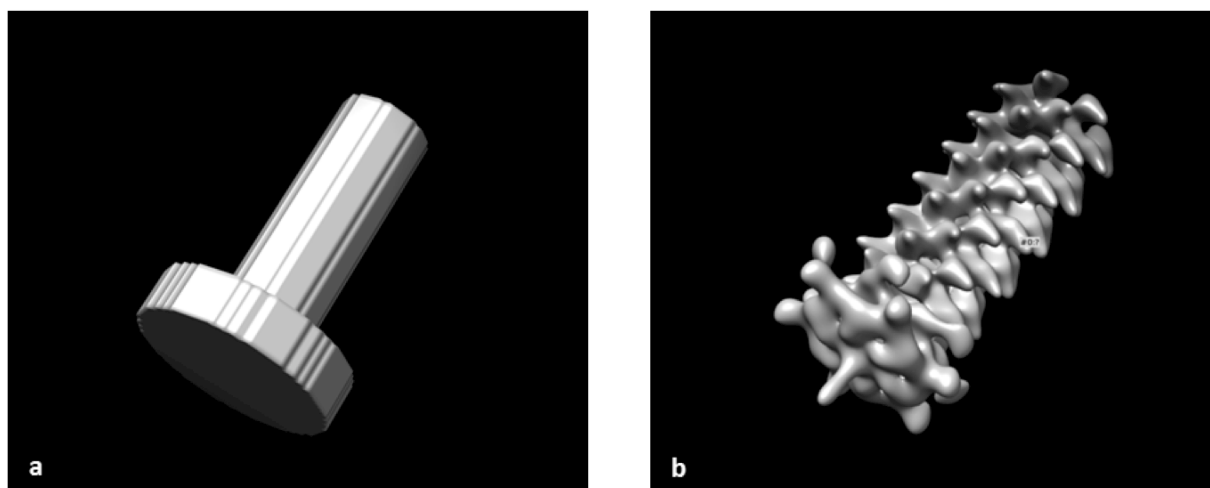
Figure 2.4 R2 pyocin initial model a) and the model after the first refinement cycle b).
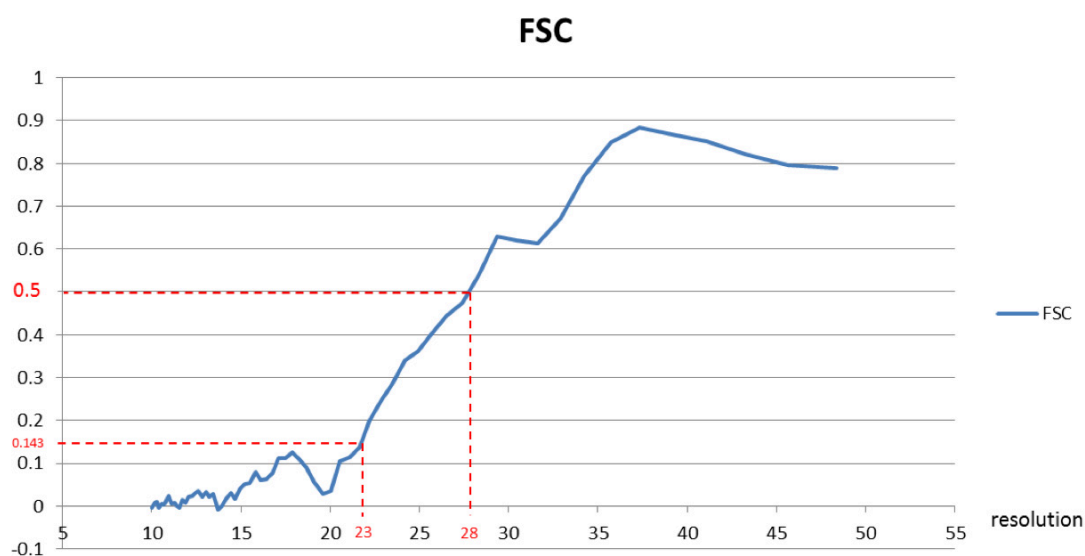
Figure 2.5 Resolution versus FSC plot for R2 pyocin reconstruction.

2.2.2 Results and discussion

A single particle 3D CryoEM reconstruction of R2 pyocin in the extended conformation at 28 Å resolution (FSC cutoff 0.5) was calculated (Figure 2.6).
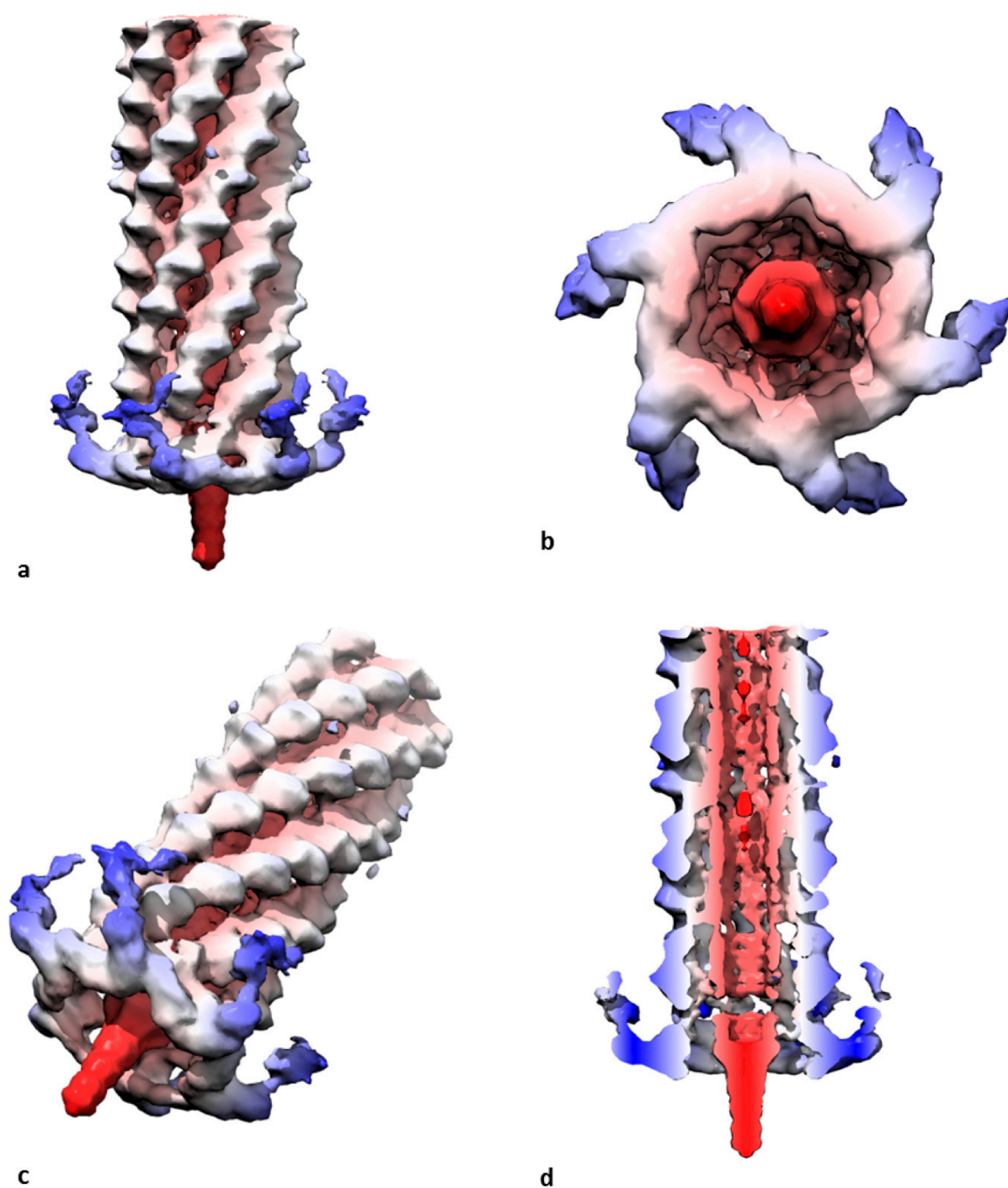
Figure 2.6 Projections of 3D reconstructed R2 pyocin particle.
a) particle side view; **b)** end-on view (bacteria binding end); **c)** tilted side view exposing membrane piercing spike; **d)** cut-away side view. The map is colored in a red-white-blue pattern starting from the center (red) to the periphery (blue).

Resolution of the current R2 pyocin reconstruction allows characterizing main features of the particle. The R2 pyocin baseplate is much simpler than T4 baseplate (Kostyuchenko, et al. 2003). The baseplate contains six tail fibers of a single type (PAO620). Only the N-terminal part of the fiber (about ¼ of the total length) can be seen in the final reconstruction. The C-terminal receptor-binding part is disordered because it is likely to have different orientations in different particles (Figure 2.6, colored in blue, Figure 2.7, segment colored in green). The X-ray structure of the C-terminal fragment of PA0620 covers this lack complementing the CryoEM model Figure 2.7 (the structure of PA0620 will be described in great details in **2.2**).

The central part of the baseplate in T4 phage particle is composed of 3 proteins in multiple copies: $(gp25)_6$-$(gp6)_{12}$-$(gp53)_6$ (starting from the particle center and going to the periphery). Comparison of the T4 and R2 baseplate structures along with bioinformatic analysis suggests that the R2 baseplate is composed of the following proteins $(PA0617)_6$-$(PA018)_{12}$-$(PA0619)_6$ (Figure 2.7, colored in magenta). The pyocin proteins are smaller then their T4 orthologs (108 residues in PA0617 against 132 in gp25; 295 residues in PA0618 against 660 in gp6; 177 residues in PA0619 against 196 in gp53). Unfortunately, the resolution of the reconstruction is not high enough to precisely determine and segment separate proteins from the assembly.

PA0616 forms the central membrane-puncturing needle of the pyocin (colored in coral red in (Figure 2.6, Figure 2.7). Its N-terminal OB-fold domain is plugged into a donut-shaped trimer of PA0628 (Figure 2.7, in blue). The pointy C-terminal domain of PA0616 is protruding from the plane of the baseplate outwards. PA0628 connects the 3-fold symmetrical cell-puncturing device with the 6-fold symmetrical baseplate. The PA0628-PA0616 complex is analogous to the T4 cell-puncturing device composed of gp27-gp5-gp5.4.

PA0622 (386 residues) is an ortholog of gp18 (659 residues) from T4 that forms the contractile sheath around the tail tube gp19. Individual subunits of the sheath (PA0622) can be distinguished in the pyocin reconstruction (Figure 2.6, Figure 2.7). T4 gp19 is orthologous to PA0623, which comprises the central tube that represents an empty channel and could be observed in the cut-away views of Figure 2.6 **d)** and Figure 2.7, colored in pale red and gray, respectively.

PA0626 (329 amino acids) is the tail tube initiator protein (Figure 2.7, colored in gold). It is situated between PA0628 and PAO623. Its T4 ortholog is gp54 (Kostyuchenko, et al. 2003) (Leiman and Shneider, Contractile tail machines of bacteriophages 2012).
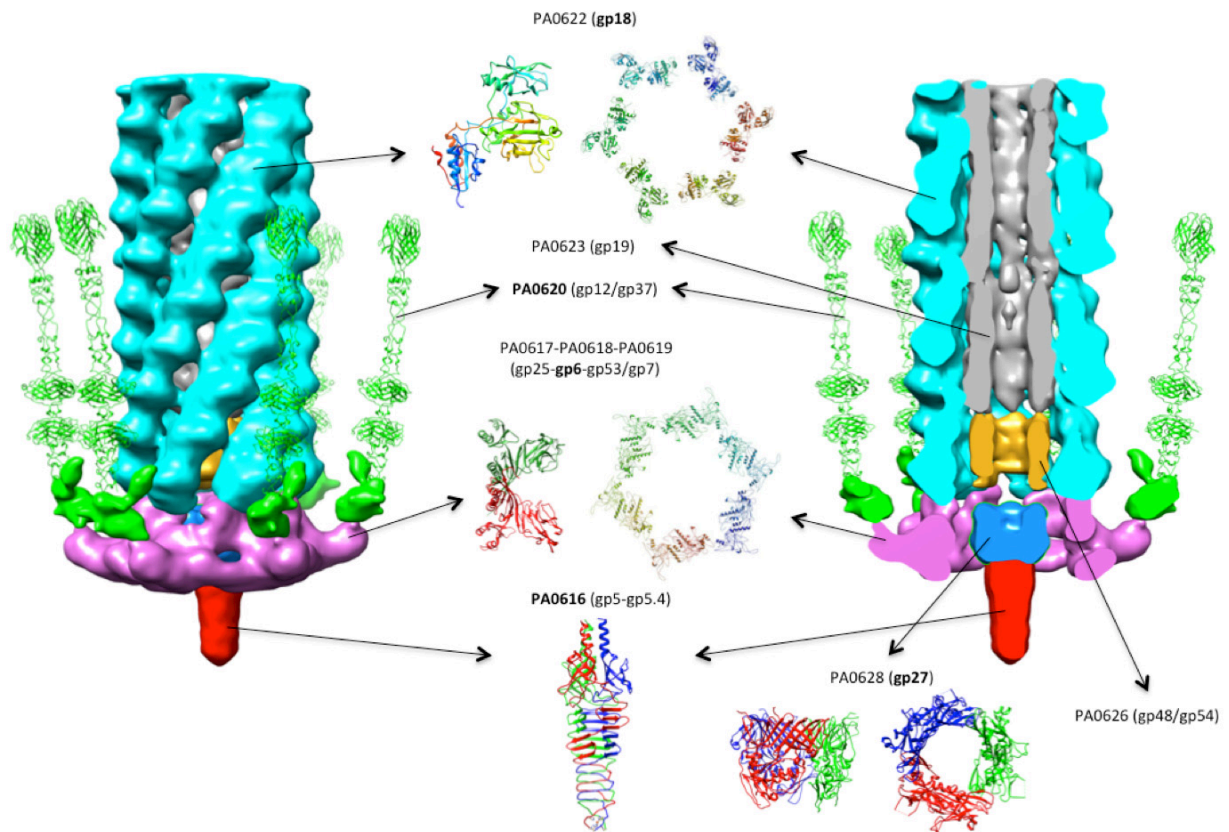
Figure 2.7 Segmentation and main components assignment on the 3D reconstruction of the R2 pyocin particle (side view and cut-away side view).
Different proteins/protein assemblies are colored in distinct colors and labeled. Known pyocin structures or their T4 orthologs are shown as ribbon diagrams and assigned to their corresponding densities. The first label (PA0XXX, where XXX is a number) stands for gene number from the *P. aeruginosa* PAO1 genome (that contains R2 pyocin gene cluster). The label in parentheses (gpXX, where XX is a number) represents gene number for T4 orthologs. Labels in bold indicate that the structure is available. (Aksyuk, Leiman and Shneider, et al. 2009), (Aksyuk, Leiman and Kurochkina, et al. 2009), (Kanamaru, et al. 2002), Leiman lab, unpublished data.

2.3   Crystallographic studies of the R1 and R2 receptor binding fibers

2.3.1   Introduction

Biochemical work with tail fibers whose function is to anchor a pyocin particle to the cell surface is not a trivial task. These proteins have evolved to be sticky: their N terminus is attached to the pyocin baseplate whereas their C terminus binds to a receptor on the host cell surface. These properties predispose pyocin fibers to non-specific aggregation. Nevertheless, we set an ambitious goal to determine the structure of the fiber by X-ray crystallography, which required producing a large amount of pure protein.

Experiments performed by Dean Scholl and colleagues (Williams, et al. 2008) were extremely useful in understanding the domain organization and functional requirements of the pyocin fiber. Scholl and colleagues (Williams, et al. 2008) noticed that phage P2 and pyocin fibers (gpH and PA0620, respectively) have 21% sequence identity overall and more than 40% in the N-terminal region (Figure 2.8). They proceeded to fusing several N-terminal fragments of the pyocin fiber to C-terminal fragments of the P2 fiber that had different lengths. Fiberless pyocin particles were complemented with these chimeric fibers and their killing activity was tested on *E. Coli C*, which is infected by P2, but completely resistant to the pyocin (Figure 2.9). Two fusion regions resulted in active pyocin particles (Figure 2.9). We used this information to create a series of deletion mutants of the pyocin fiber having different lengths.

Similar to the P2 fiber, the pyocin fiber requires a chaperone for assembly (Williams, et al. 2008). The chaperone gene for both fibers (gene *H* in P2 and gene *PA0621* in the pyocin) is situated immediately downstream from the fiber gene (Figure 2.10) (Köhler, Donner and van Delden 2010).

There are 5 types of pyocins (R1 through R5) each of which has a narrow spectrum limited to the strains of *P. aeruginosa* strains that have similar surface LPS (Figure 2.11). The five types of the pyocins are virtually identical even down to the DNA level. The only significant difference in these pyocins is the fiber and its chaperone, which is clearly related to their different hosts. The fiber sequences can be divided into two clades with the first one containing the R1 and R5 fibers and the second – R2, R3 and R4. The clades are nevertheless close to each other and, for example, the R1 and R2 sequences show 85.9% sequence identity and 92.8% similarity. Therefore, the domain organization of the R1 fiber is similar to that of the R2 fiber.

Figure 2.8 Sequence alignment of the P2 tail fiber (gpH) with the R2-type pyocin fiber (PA0620).
Colors and symbols definition according to ClustalW version 2 (ClustalW2) (Goujon, et al. 2010). Fragments yielding bactericidal particles are mapped on the alignment. Numbers on top correspond to P2 gpH, bottom numbers correspond to R2 pyocin PA0620.
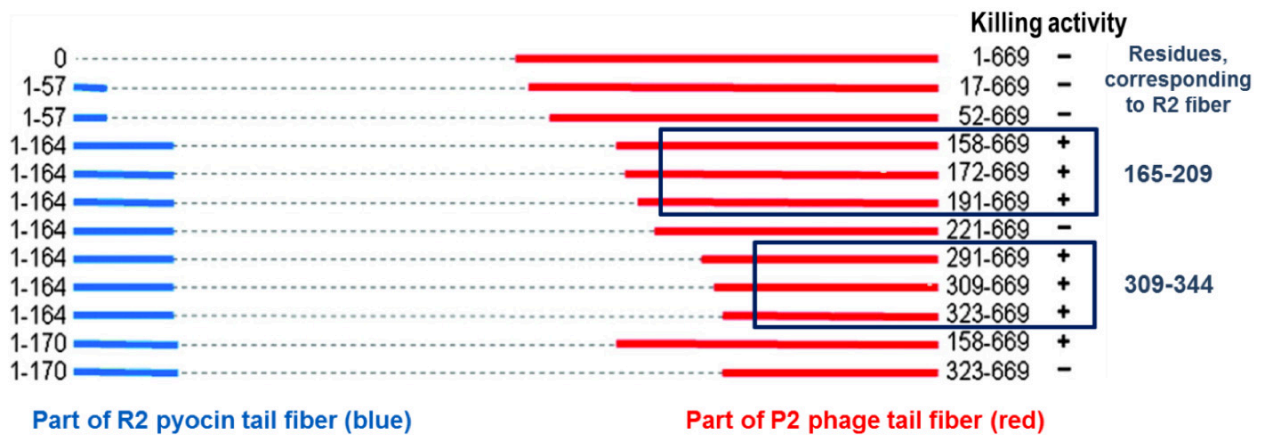
Figure 2.9 Killing activity of various fusions of the R2 pyocin fiber to the P2 phage tail.
The boxes highlight the fragments that lead to bactericidal activity. The figure was adapted from (Williams, et al. 2008).
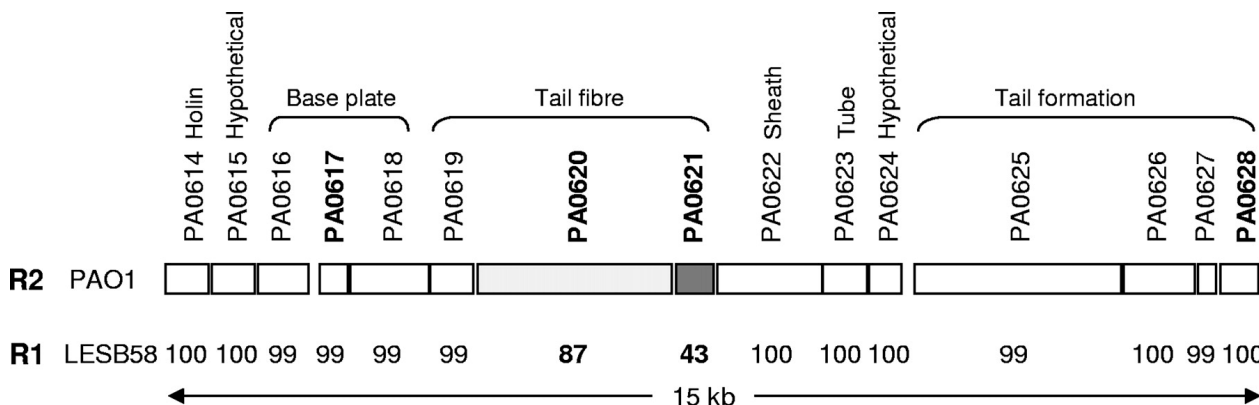
Figure 2.10 Amino acid sequence identity of selected proteins comprising R-pyocins encoded by PAO1 (type R2) and LESB58 (type R1).
Genes corresponding to the pyocin tail fiber and its chaperone are in bold. Figure was adapted from (Köhler, Donner and van Delden 2010).
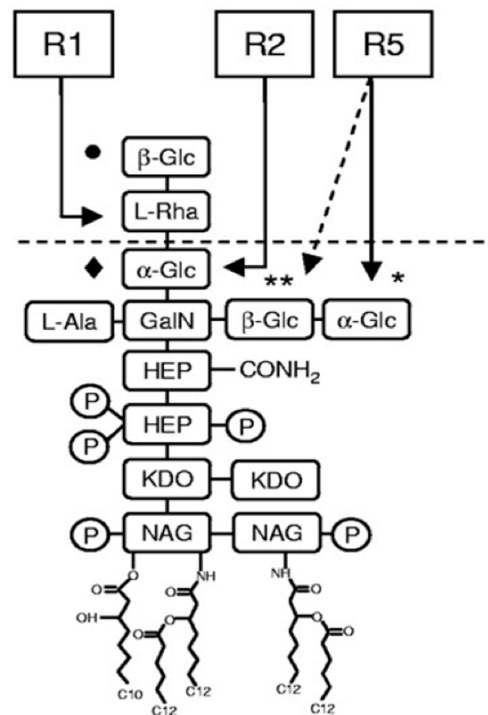
Figure 2.11 LPS structure of *P. aeruginosa* PAO1 (serotype O5).
Specificity of different pyocins from R family to LPS constituting monosaccharides is shown. Figure adapted from (Köhler, Donner and van Delden 2010)

2.3.2 Materials and methods

2.3.2.1 Construct engineering and cloning

All constructs used in this study were designed and created by Dr. Mikhail Shneider.

We first expressed the full length PA0620 fiber and two of its deletion mutants called d1 and d2 in the presence of its chaperone PA0621. The size of the mutants was chosen to correspond to the length of the P2 fiber fragments that could be fused with the pyocin fiber successfully (Figure 2.9). The PA0620d1 mutant contained residues 173 – 691, and PA0620d2 – residues 124 – 691. Note that both mutants had intact C-terminus.

The PA0620d2 mutant was insoluble. The full-length protein was soluble but showed a tendency to aggregation and therefore was unsuitable for a large-scale purification. The PA0620d1 mutant was soluble and behaved well enough to be eventually subjected to crystallization, although the crystals diffracted poorly (see below). We treated the full-length protein with trypsin and analyzed the resulting products by mass spectrometry. A fragment containing residues 322 – 691 was identified and mapped onto the trypsin cleavage map calculated by PeptideCutter (Artimo, et al. 2012). We called this deletion mutant PA0620d3. It was successfully crystallized and its structure determined. We used a fragment of similar size (residues 323-701) to determine the structure of the R1-type fiber (UNP: B7V4K6, strain LESB58, gene PLES_06171).

2.3.2.2 Protein expression and purification

2.3.2.2.1 R2 pyocin fiber full length protein, d1 and d2 deletion mutants

The full length PA00620 protein, PA0620d1 and PA0620d2 deletion mutants were cloned into pESL plasmid (kanamycin selection). The vector was designed to express fusion constructs containing an N-terminal His-tag, a SlyD protein (FKBP-type peptidyl-prolyl cis-trans isomerase and folding chaperone), a linker, a TEV cleavage site and the fragment of interest. PA0621 was expressed in pACUC plasmid (chloramphenicol selection). Both plasmids were transformed into B834 (DE3) strain of *E. coli* competent cells. The transformed cells were grown at 37 °C in the LB medium, complemented with kanamycin and chloramphenicol at the concentration of 100 and 34 μg/ml, respectively, until the optical density reached the value of 0.6 at 600 nm. The medium was cooled on ice to the temperature of 18-20 °C followed by expression induction by addition of IPTG to a final concentration of 1 mM. After further incubation at 18 °C overnight (approximately 16 hours), the cells were harvested by centrifugation at 5180 g, 4 °C. The cell pellet was resuspended in 1/50th of the original cell volume in a 20 mM Tris-HCl pH 8.0 buffer

complemented with 300 mM NaCl, 5 mM imidazole (Im) and 0.02% NaN3. The cells were lysed by sonication. The cell debris was removed by centrifugation at 35000 g for 15 minutes at 4 °C. Cells, pellet and supernatant were analyzed with the SDS-PAGE. Analysis revealed that full length protein expressed insoluble; deletion mutant d2 showed problems with folding and aggregated shortly. Further procedure was continued with d1 mutant. The supernatant was loaded onto the Ni^{2+}-precharged 5 ml GE HisTrap FF Crude column (GE Healthcare Life Sciences) that was equilibrated with 20 mM Tris-HCl pH 8.0, 300 mM NaCl. Protein was eluted in two steps: 1) with buffer containing 20 mM Tris-HCl pH 8.0, 300 mM NaCl, 50mM Im – to remove nonspecific bounded proteins and 2) with 20 mM Tris-HCl pH 8.0, 300 mM NaCl, 250 mM Im buffer – actual protein elution. SDS-PAGE analysis of the eluted fractions and the flow through showed poor binding of the fusion construct to the metal affinity column. Nevertheless, most of the protein came out in the flow through, thus the purification till certain degree was achieved due to the non-specific binding of expression side products to the column resin. Purification was continued with the flow through material. It was set up for overnight His-tag and SlyD digestion with TEV protease against 10 mM Tris-HCl pH 8.0, complemented with 3 mM DTT, 1.5 mM EDTA. After His-tag and SlyD were removed part of the target protein aggregated. Digested protein was further purified with ion-exchange chromatography (GE MonoQ 10/100 GL column connected to an AKTApurifier 100 system) in 20 mM Tris-HCl pH 8.0 buffer using 0 to 1 M NaCl linear gradient. Relevant fractions were further purified by size exclusion chromatography using a GE HiLoad 16/60 Superdex 200 PG (GE Healthcare Life Sciences) column connected to the AKTApurifier 100 system (GE Healthcare Life Sciences). A 10 mM Tris-HCl pH 8.0, 200 mM NaCl buffer was used.

2.3.2.2.2 Native and SeMet derivative of R2 pyocin PA0620d3 deletion mutant

The PA0620d3 deletion mutant was cloned together with PA0621 (downstream of PA0620d3) into pTSL plasmid (ampicillin selection). The expression construct had same composition as pESL described in **2.2.2.2.1**. The same purification protocol was used as for PA0620d1 mutant. The only difference was the type of growth medium used (2xTY for the native protein and SelenoMet Medium (Molecular Dimensions) for the SeMet derivative). The growth medium contained ampicillin at a concentration 200 μg/ml and chloramphenicol at 34 μg/ml.

2.3.2.2.3 R1 pyocin fiber fragment

The fragment of the R1 pyocin fiber (residues 323-701) was expressed and purified following the same procedure as for PA0620d3.

2.3.2.3 Crystallization, data collection and structure determination

2.3.2.3.1 R2 pyocin fiber

PA0620d1 and d3 were brought to concentrations of 9 and 11 mg/ml, respectively. The initial crystallization screening was carried out by the sitting drop method in 96 well SWISSCI 2-lens MRC plates using Jena Bioscience crystallization screens. Optimization of crystallization conditions was performed in 24 well-plates (Jena Bioscience) by hanging drop vapor diffusion. Crystallization drops of the 24-well plate setup contained 1.5 μl of the protein solution in 10 mM TrisCl pH 8.0, 150 mM NaCl (200 mM for PA0620d1) mixed with an equal volume of the well solution. Best crystals of the PA0620d1 were obtained with the protein having the initial concentration of 5 mg/ml and equilibrated against 750 μl of the well solution containing 8-9% PEG 4000, 100 mM Hepes pH 7.0, 200-300 mM KH_2PO4 at +18∘C. Best PA0620d3 crystals were obtained with the protein at 7 mg/ml and equilibrated against 750μl of the well solution containing 3-5% PEG 6000, 100 mM Tris·HCl pH 8.5, 160-210 mM KCl at +18°C Figure 2.12.

a)   b)
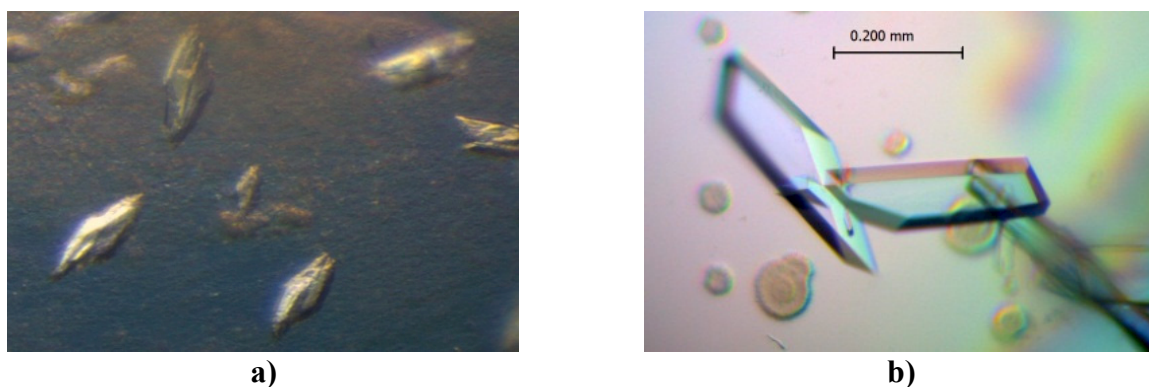
Figure 2.12 Crystals of the PA0620d1 a) and PA0620d3 b).

For data collection, the crystals were dipped for 20-45 seconds into cryo solutions containing either 25% of PEG 4000 for PA0620d1 or 25% of ethylene glycol for PA0620d3 in addition to the well solution components and flash frozen in a vaporized nitrogen stream at 100 K. Crystals of PA0620d1 showed anisotropic diffraction up to 3.5 Å in the best direction (Figure 2.13 **a)**). Attempts to improve the quality of PA0620d1 crystals failed and the work with PA0620d1 was stopped.

Best crystals of PA0620d3 diffracted to better than 1.7 Å resolution. Unfortunately, it was impossible to collect such high resolution data even when using the largest available PILATUS 6M detector (424 x 435 mm^2, 2463 x 2527 pixels) because of excessive spot overlaps due to the crystal having a long axis and a relatively high mosaicity (Figure 2.13 **b)**).

The structure was solved by single-wavelength anomalous diffraction (SAD) technique using a SeMet derivative (Hendrickson and Ogata 1997). The SAD data were collected at the wavelength of the maximum Se adsorption (K-line) thus maximizing the anomalous signal.

PA0620d3 crystals belonged to $P2_12_12_1$, #19 space group with a = 56.12, b = 126.35 and c = 431.25 Å unit cell parameters. Data collection was carried out at the X06SA PXI Pilatus beam line of the Swiss Light Source (SLS) at the Paul Scherrer Institute (SLS, Villigen, Switzerland) at the wavelength of 1 Å. The diffraction data was indexed, integrated, and scaled with XDS (Kabsch 2010). Assuming that the protein is trimeric the Matthews coefficient (Matthews 1968) was calculated. The program did not give a definitive answer: two possibilities, having 2 (3.22 $Å^3$/Da with total probability of 0.21) or 3 (2.15 $Å^3$/Da with total probability of 0.78) trimers were further considered. The Se sites were located with SHELX_CDE program from CCP4 program package (Winn, et al. 2011). SHELXC anomalous signal strength statistics is given in Table 2.1. Table 2.2 shows SHELXD phasing statistics for typical correct and incorrect solutions. There are 6 methionine residues per polypeptide chain, giving in total 36 methionones in case of 2 molecules per asymmetric unit and 54 in case of 3. A total of 50 sites were found with a sharp occupancy drop off after 30 (from 0.49 for site #30 to 0.21 for site #31, Figure 2.14). Visual inspection of Se site positions showed that there the asymmetric unit contained 2 molecules. The first 30 sites were then given to the program SOLVE (Terwilliger, Automated structure solution, density modification and model building 2002) for further refinement and phasing. The phases were improved by NCS averaging and solvent flattening with the program RESOLVE (Terwilliger, Automated structure solution, density modification and model building 2002). ARP/wARP (Lamzin, Perrakis and Wilson 2001) was used for automated model building. Further refinement of the atomic model was performed with Coot (Emsley and Cowtan 2004), Refmac5 (Winn, et al. 2011) and PHENIX with TLS (Adams, et al. 2010). Details of data reduction and refinement are given in the Table 2.3.
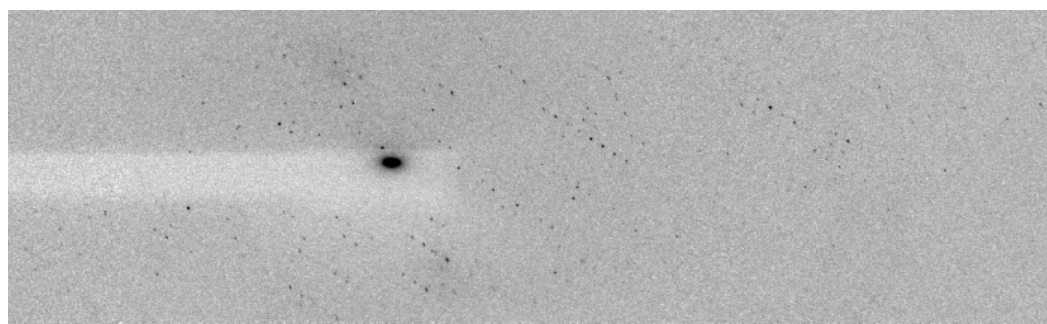
a)
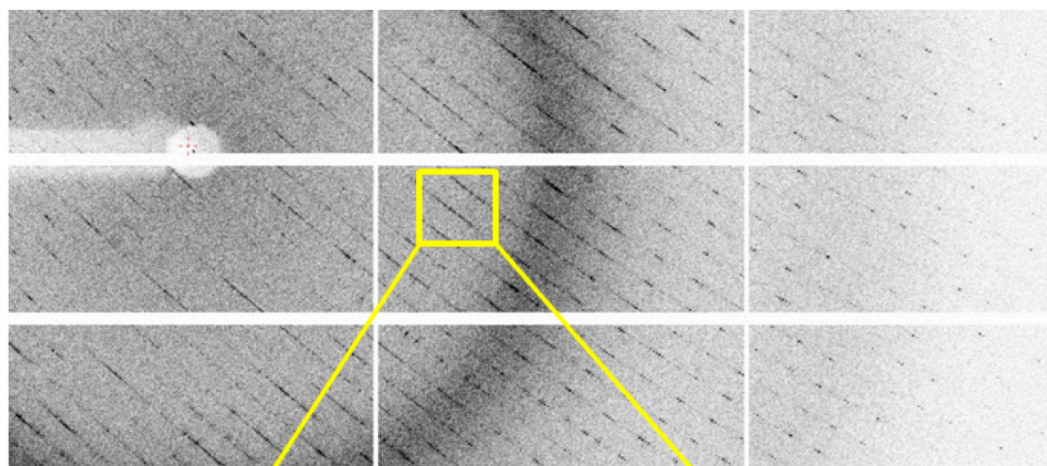
Zoom
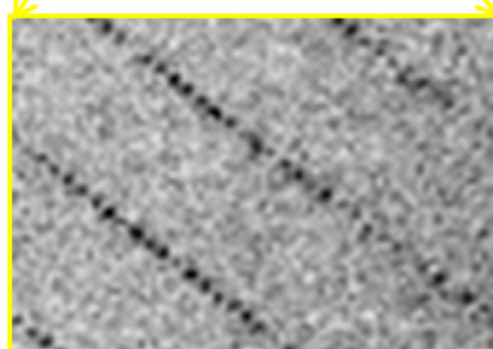
b)

Figure 2.13 Diffraction pattern of PA0620d1 and and PA0620d3 crystals.
a) Anisotropic diffraction up to 3.5 Å in the best direction of PA0620d1 crystal (0.5-degree oscillation). **b)** Diffraction pattern of PA0620d1 crystal and its zoomed region (0.25-degree oscialltion).
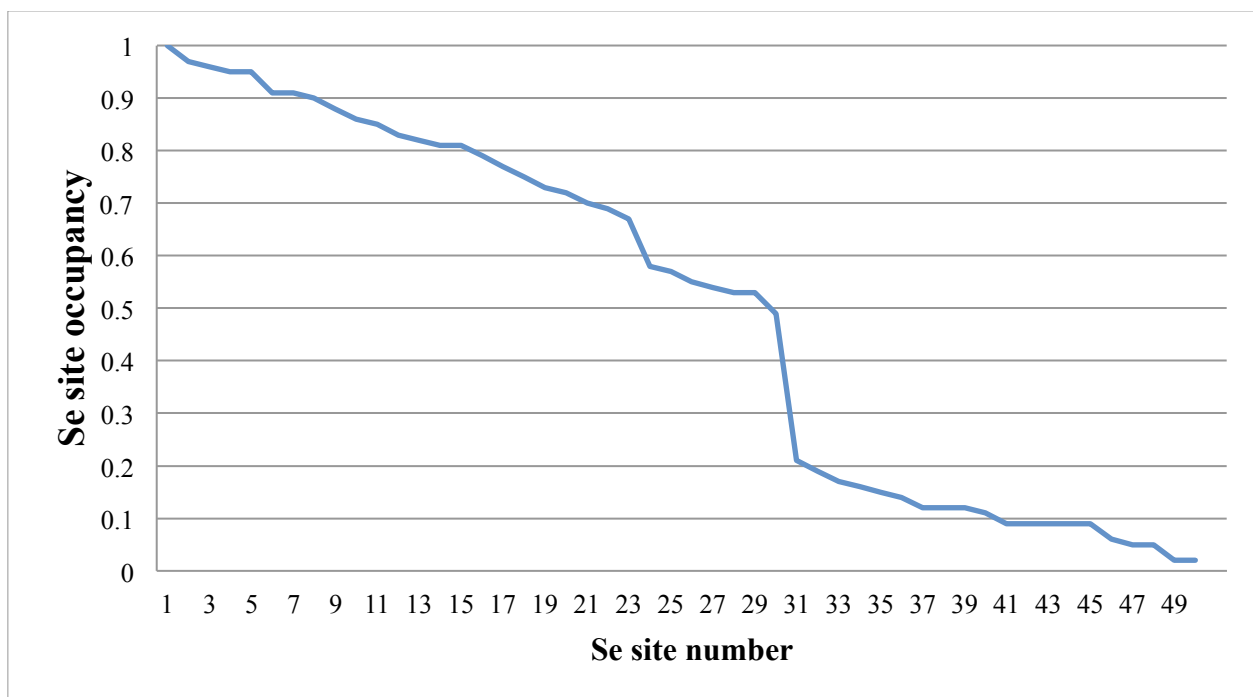
Figure 2.14 Refined occupancy against peak number for SHELXD location of the 36 Se atoms from R2 pyocin fiber SAD data.

Table 2.1 SHELXC anomalous signal strength statistics on R2 pyocin fiber SAD data.

N(data) – number of reflection; $<I/sigI>$ – mean intensity over intensity standard deviation within a resolution shell; %Compleate – reflections completeness for a resolution shell; $<d''/sig>$ – mean anomalous dispersion over anomalous dispersion standard deviation within a resolution shell.

* For a non-random anomalous dispersion signal $<d''/sig>$ must be greater then 0.80.

	Inf-8.0	8.0-6.0	6.0-5.0	5.0-4.0	4.0-3.5	3.5-3.1	3.1-2.9	2.9-2.7
N(data)	1922	2333	3666	10600	11720	15648	10535	13107
$<I/sig>$	63.3	52.6	53.8	51.1	43.1	31.7	21.5	15.2
%Complete	52.0	49.0	62.5	81.1	89.5	90.3	83.7	79.8
$<d''/sig>$*	3.08	2.27	1.96	1.70	1.56	1.32	1.03	0.88

Table continuation

	2.7-2.5	2.5-2.3	2.3-2.12
N(data)	17109	25488	23122
$<I/sig>$	10.6	7.1	5.2
%Complete	77.2	83.9	61.8
$<d''/sig>$	0.74	0.66	0.69

Table 2.2 SHELXD phasing statistics on R2 pyocin fiber SAD data. Examples of typical correct and incorrect solutions.

R – R factor, Rcryst=(Σ||Eo|-|Ec||)/Σ|Eo|; Min. function – minimal function, a measure of the mean-square difference between the values of the triplets calculated using a particular set of phased and the expected probabilistic values of the same triplets as given by the ratio of modified Bessel functions; <cos> – cos ($\Phi_{H,K}$), a parameter of a minimal function, amplitudes of known heavy atoms. Ra – R factor based on anomalous data; CC all – correlation coefficient calculated with all data; CC weak – correlation coefficient with only the weaker *E*-values that were not used in the dual-space recycling; PATFOM – Patterson figure of merit.

	Typical correct solution	**Typical incorrect solution**
R	0.314	0.439
Min. function	0.428	0.557
<cos>	0.361	0.193
Ra	0.336	0.530
CC all	48.56/48.59	19.36/48.59
CC weak	28.77/28.90	7.31/28.90
PATFOM	5.31/5.31	0.89/5.31

2.3.2.3.2 R1 pyocin fiber

For crystallization, the R1 fiber fragment was concentrated to 9 mg/ml. The initial crystallization screening was carried out by the sitting drop method in 96 well SWISSCI 2-lens MRC plates using Jena Bioscience crystallization screens. Optimization of crystallization conditions was performed in 24 well-plates (Jena Bioscience) by hanging drop vapor diffusion. Crystallization drops of the 24-well plate setup contained 1.25 μl of the protein solution in 10 mM TrisCl pH 8.0, 150 mM NaCl mixed with an equal volume of the well solution. Best crystals were obtained with the protein having the initial concentration of 7.4 mg/ml and equilibrated against 500 μl of the well solution containing 4% PEG 4000, 100 mM PIPES pH 6.1, 140 mM Na_2(mal) at +18∘C. These crystals diffracted to 2.3 Å resolution and belonged to $P2_12_12_1$, #19 space group with a = 144.08, b= 154.46 and c = 198.83 Å unit cell parameters. Data collection was carried out at the X06SA PXI Pilatus beam line of the Swiss Light Source (SLS) at the Paul Scherrer Institute (SLS, Villigen, Switzerland) at the wavelength of 1 Å. The diffraction data was indexed, integrated, and scaled with XDS (Kabsch 2010). A Matthews coefficient of 2.33 $Å^3$/Da (Matthews 1968) was calculated with total probability of 0.98 for 2 molecules in asymmetric unit.

The structure of the R1 pyocin fiber was solved by molecular replacement with PHASER (McCoy, et al. 2007) using the PA0620d3 structure as a search model. The PA0620d3 structure had to be split into 2 fragments of about equal lengths because the R1 fiber showed a significant bent in the middle. This was caused by the crystal packing that was different to that of the R2 fiber Figure 2.16. Coot (Emsley and Cowtan 2004) and PHENIX with TLS (Adams, et al. 2010) were used for the refinement of the atomic model. The details of data reduction and refinement are given in the Table 2.3.

Table 2.3 X-ray data collection and refinement statictics for R2 and R1 pyocin fiber structures.
Data in parenthesis represent statistics for the highest resolution shell.

Data collection	R2 fiber SeMet derivative	R2 fiber native	R1 fiber native
Wavelength	0.97965 Å	1.0 Å	1.0 Å
Number of frames	1440	780	720
Frame width (°)	0.25	0.25	0.25
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
Cell dimensions (Å)	$a = 56.20$, $b = 126.82$, $c = 433.43$	$a = 56.12$, $b = 126.35$, $c = 431.25$	$a = 144.08$, $b = 154.46$, $c = 198.83$
Number of trimers per asymmetric unit	2	2	2
Resolution (Å)	49.9 – 2.4	50.0 – 1.9	49.0 – 2.2
R_{meas}	0.071 (0.18)	0.13 (0.55)	0.11 (0.61)
$<I / \sigma_I>$	20.84 (8.2)	8.77 (2.6)	8.8 (2.1)
Completeness (%)	99.3 (96.1)	99.6 (98.5)	99.2 (96.2)
Redundancy	6.71 (6.17)	3.8 (3.8)	3.4 (3.2)

Refinement			
Number of reflections			
Working		246067	192452
Test		3691	9610
R_{work} / R_{free}		0.171 / 0.214	0.190 / 0.226
B-factor ($Å^2$)		23.5	40.0
R.m.s. deviations			
Bond lengths (Å)		0.020	0.006
Bond angles (°)		1.105	0.698
Number of atoms			
Protein		16894	16836
Non-protein (solvent and ligands)		4087	2021
Ramachandran plot (%)			
Most favored		97.2	96.7
Additionally allowed		2.8	3.2
Outliers		0.0	0.1

2.3.3 Results and discussion

2.3.3.1 Morphology overview and crystal packing of the R2 and R1 pyocin fiber structures

The R2 and R1 fibers have a similar domain organization and overall structure. Each protein is a 240 Å-long fiber formed by three intertwined polypeptide chains and that contains five domains. The N-terminal domain is a short rod that is built by a repeating alpha-helix-turn motif. It is followed by two tandem knob domains that have similar structures. The two knobs are followed by a shaft domain that is ~ 97 Å long and contains a buried Fe ion approximately in the middle. The C-terminal domain is a globular structure with a lectin-like fold (Figure 2.15).
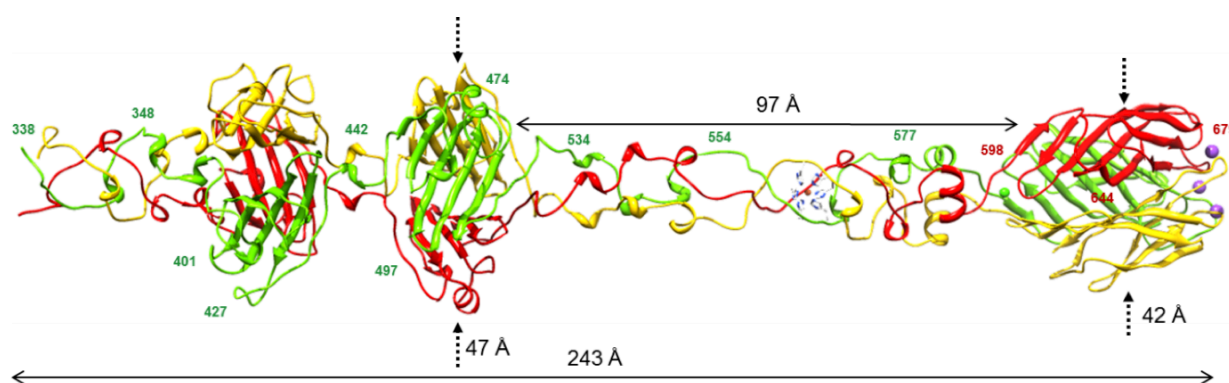
Figure 2.15 Domain organization of the R2 and R1 fibers.
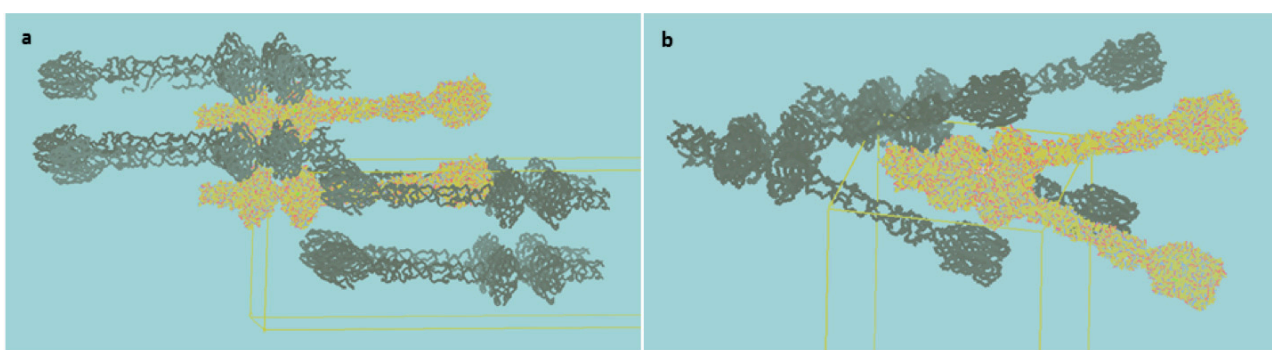
Figure 2.16 Crystal packing of the R2 a) and R1 b) pyocin fibers.

2.3.3.2 Analysis of the N terminal and shaft domains

The fold of the N terminal rod domain is somewhat similar to that of the shaft domain. Three polypeptide chains extensively intertwine with each other to form a structure with complex topology Figure 2.17 **a), b)**. The fold of the polypeptide chain could be characterized by a helix-plus-linker motif: short fragments of α-helices mainly composed of 3 residues alternate with linkers of different length (from 2 to 7 residues) Figure 2.17 **c), d)**. After each α-helix the polypeptide chain non-regularly changes its run direction that eventually results in clockwise twisting of the chain and makes the intertwining of chains possible.

This motif is topologically similar to the helix-turn-helix motif of the stem domain of the phage phi29 head fiber (Xiang and Rossmann 2011). However, the phi29 stem domain is highly regular: each repeat unit corresponds to a 45° turn and a translation of 19.5 Å along the helical axis Figure 2.17 **e)**.
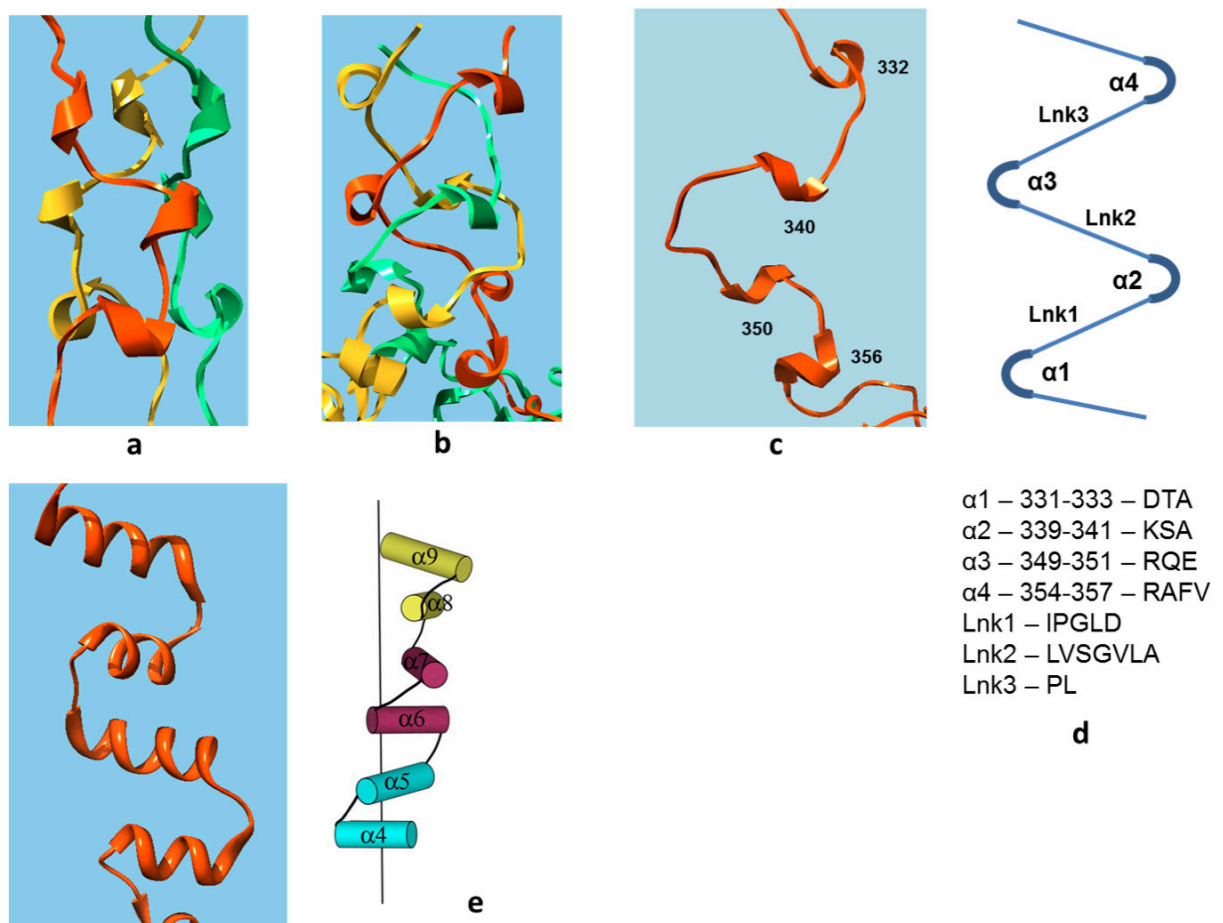
Figure 2.17 Helix-plus-linker motif of the fiber N-terminal domain and the shaft.
a) Ribbon diagrams of the fiber shaft; **b)**, the N-terminal domain; **c)** single chain of the N-terminal domain; key residues are numbered. For **a)** and **b)** three chains of the trimer are colored in yellow, green and orange; **d)** schematic representation and detailed topology of the helix-plus-linker motif for the fragment, depicted in **c)**; **e)** helix-turn-helix motif of the stem domain fragment of phage phi29 head fiber. R2 fiber model and numbering was used for illustration.

2.3.3.3 Analysis of the fiber knob domains

The tandem knob domains (residues 359-440 and 444-530, Figure 2.18) have a similar fold that is formed by an antiparallel β-sheet containing 5 strands that are connected by loops containing as many as 17 residues. Six residues out of these 17 form a short α-helix. The two domains show only 19.7 % sequence identity on superposition with 71 aligned C-alpha atoms (out of 82) and RMSD of 2.1 Å. These domains have clearly evolved from a single ancestor as a result of an ancient gene duplication event. The inter-strand loops form the outer part of the knobs whereas the β-sheets constitute their core.
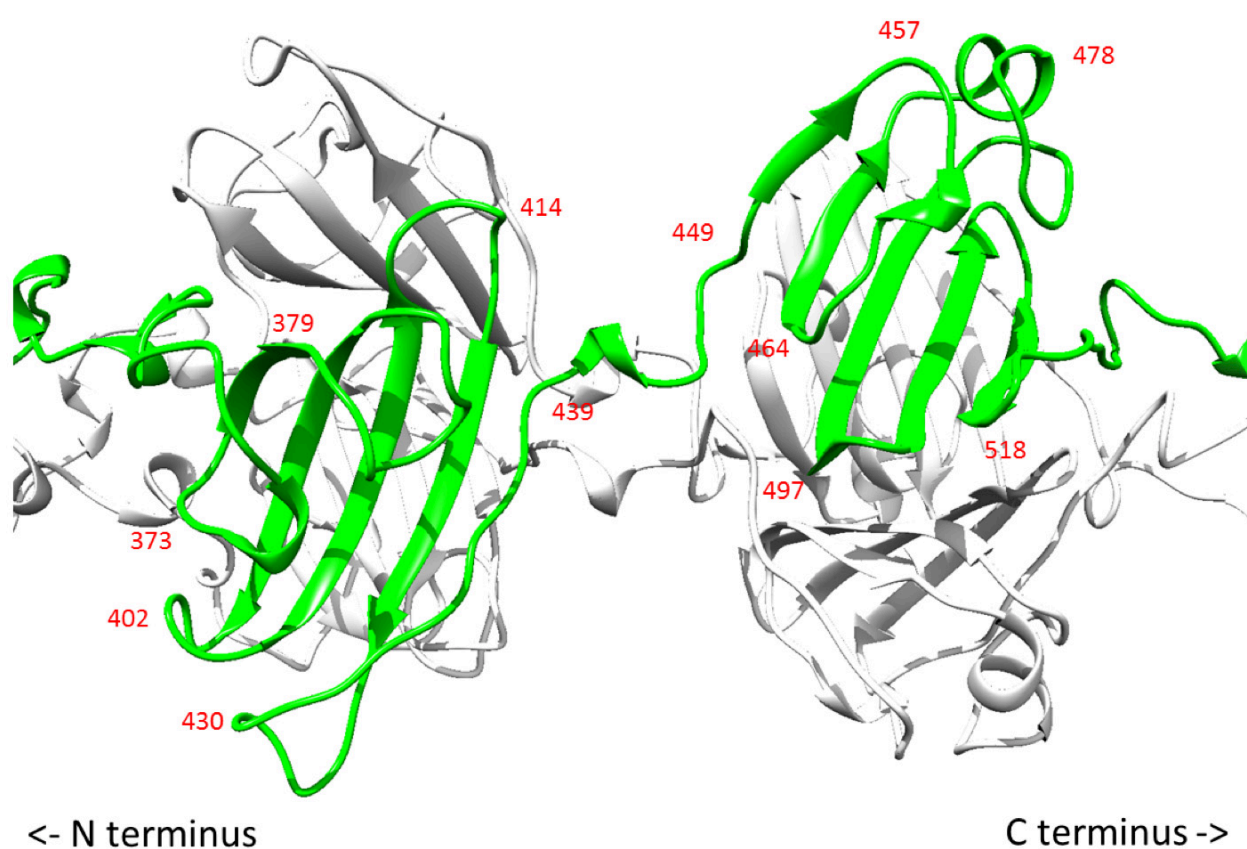
Figure 2.18 Ribbon representation of two mirroring knob domains.
Key residues are numbered (R1 fiber model and numbering were used in this figure).

2.3.3.4 Localization and analysis of metal ions buried inside the fiber's core

There are two metal coordination sites within the fiber's core. One of them is responsible for coordination of an iron ion and the other – for coordination of a hydrated calcium ion. The iron binding site is situated approximately in the middle of the shaft domain. $[Ca(H_2O)_6]^{2+}$ is situated in the very beginning of the C-terminal domain (Figure 2.19 **a)**). The iron binding site is characterized by a double histidine motif (HxH) in the amino acid sequence (residues H361 and H363 for R1; H362 and H364 for R2 fiber). Each of the three polypeptide chains contributes both histidines of its HxH motif resulting in the octahedral coordination of the centrally positioned iron atom (Figure 2.19 **b)**). The average Fe-His ($N^{\varepsilon 2}$) bond distance is 2.16 ± 0.08 Å in the R1 fiber structure and 2.22 ± 0.05 Å in the R2 structure. The average value of all iron binding sites with a coordination number of 6 found in all structures contained in the PDB and determined with a resolution of 2.2 Å or better is 2.09 ± 0.12 Å (MESPEUS_10, (Hsin, et al. 2008)).

The hydrated calcium ion has an octahedral geometry as well – $Ca(H_2O)_6^{2+}$. It is coordinated by three symmetry-related D613 residues (Figure 2.19 **c)**). The average Ca-H_2O bond distance is 2.27 ± 0.05 Å in the R2 fiber and 2.38 ± 0.03 Å in the R1 fiber. This is again in agreement with a database value of 2.46 ± 0.22 Å calculated for a coordination number of 6, a resolution limit of 2.2 Å by MESPEUS_10 (Hsin, et al. 2008).

Consistency of bond lengths and geometries with the database values and temperature factors with those of surrounding waters and amino acid atoms suggests that metal identities were determined correctly.

The functions of the buried metal ions is most likely related to the folding of the proteins. They might form reference points for three nascent protein chains that are about to fold into a trimer. A similar iron and calcium binding sites are found in other fibrous proteins, for which a similar function was proposed. There is no experimental data to support this hypothesis at this point. (Browning, et al. 2012), (van Raaij, Mitraki, et al. 1999). The metal-binding sites also likely to contribute to the elastic properties of the fiber because the long iron-containing shaft domain clearly displays certain amount of flexibility, as the R1 and R2 fibers can bend in the middle due to crystal packing (see **2.2.2.3.2**, Figure 2.16 **b)**).
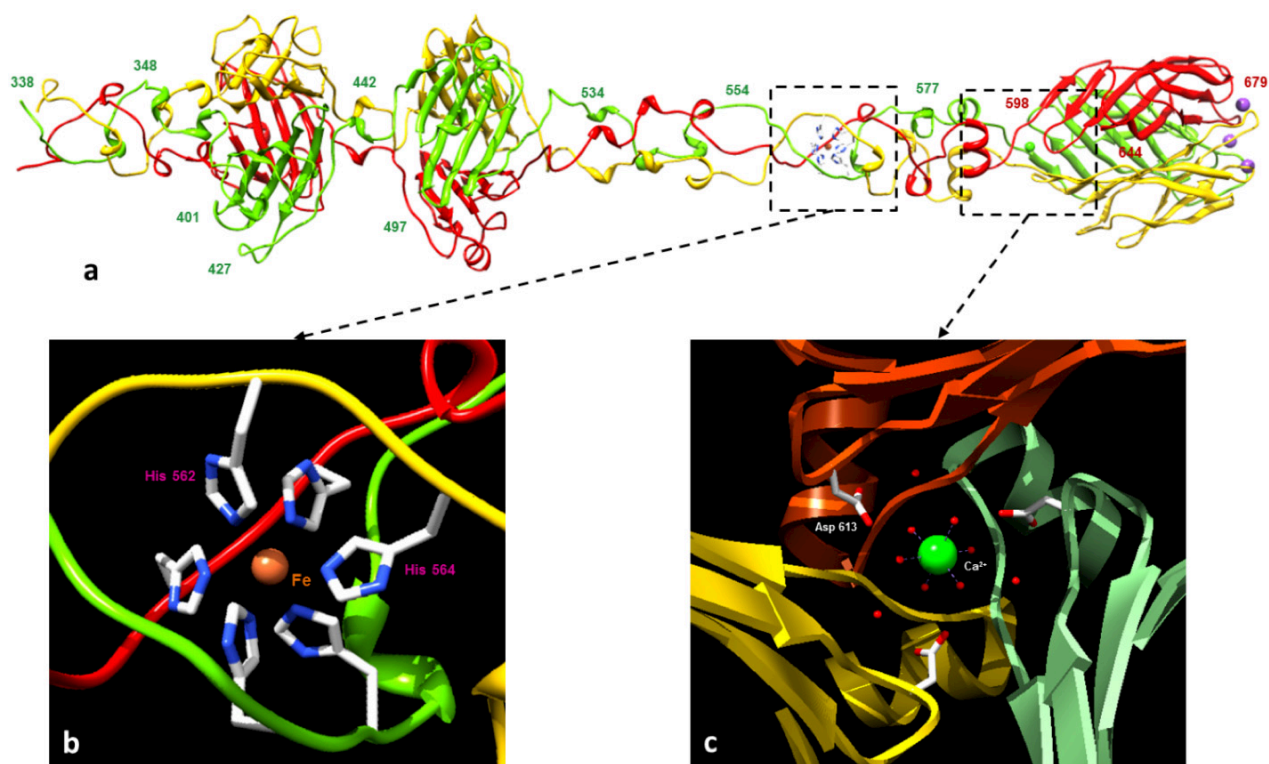
Figure 2.19 Location of the metal ion coordination sites.
a) Position of the metal ion sites in the overall structure; details of the iron **b)** and calcium **c)** coordination site in pyocin fibers; Numbers for the key residues are given (R2 fiber model and numbering were used in this figure).

2.3.3.5 Analysis of the C-terminal receptor binding (RB) domain and putative RB sites

Residues 596-691 and 595-701 comprise the C-terminal receptor-binding (RB) domain of the R2 and R1 fibers, respectively. The polypeptide chain forms two β-sheets stacked on top of each other. Both β-sheets are antiparallel and contain 3 and 5 strands of different lengths (Figure 2.20). The folds of the R1 and R2 fiber domains are very similar and the two structures can be superimposed with a RMSD of 1.4 Å between 87 equivalent atoms out of 96 used in the alignment. The sequence identity of this superposition is 42% (Figure 2.18). Nevertheless, the surfaces of the R1 and R2 C-terminal RB domains are markedly different.

The loops connecting the β-strands in both structures have different geometry and create surface cavities that coordinate metal ions. Each of the three symmetry-related cavities of the R2 fiber contains a potassium ion, whereas the R1 cavities coordinate sodium ions. The potassium cavity of R2 is located on the interface of two subunits and is much deeper than the sodium cavity of R1, which is formed by residues belonging to the same polypeptide chain. The cavities are located at different places on the RB surface and the distance between their centers, which is found upon superimposing the two domains onto each other, is about 8 Å (Figure 2.22).

The identification of metal ion was based on the analysis of the coordination geometry and temperature factors (Hsin, et al. 2008). The findings are consistent with the composition of crystallization solutions: the R2 fiber was crystallized in the presence of K^+, whereas the R1 fiber – in the presence of Na^+ (see **2.2.2.3**).

The symmetry-related coordination spheres were heterogeneous because some of the water molecules comprising the coordination polyhedron were missing in several sites (especially in the R1 structure) thus complicating the analysis of the sites' geometry. The metal ion coordination polyhedron in both structures can be described as a twisted trigonal (or triangular) prism containing 6 ligands on average (Figure 2.21, Table 2.5). The potassium ion of the R2 fiber is significantly more solvent exposed than the sodium ion of R1. There are at least 3 water ligands in the coordination sphere of the potassium ion, whereas the sodium sphere of the R1 fiber contained only 1 water molecule.

A search for structures with folds resembling that of the R2 fiber using the Dali server (Holm and Rosenström 2010) identified hits only for the C-terminal domain (Table 2.4). The most similar structures were those of Discoidin-I and Discoidin-II, which are galactose- and N-acetylgalactosamine-binding lectins, respectively; *Helix Pomatia* agglutinin – oligosaccharides lectin-binding protein, and phage receptor binding protein – the tail spike of phage SF6. All

identified proteins (besides being trimeric), which are structurally similar to the RB domain of the pyocin fiber, are trimeric and function to bind saccharides. It is therefore likely that the C-terminal domain of the pyocin fiber is responsible for binding to the sugar portion of LPS receptor.

Discoidin-I binds a sugar molecule in one of its surface groves that is somewhat negatively charged. The negatively charged metal-binding cavities on the pyocin surface are part of prominent groves. In the crystal structures reported here, a major part of the groves is occupied by hydrated metal ions that are likely to have originated from the crystallization solution. (Figure 2.22). In the R2 C-terminal head domain, the groves or cavities are grouped together thus creating a "supercavity" at the domain extremity around the 3-fols axis. In R1 fiber, the groves start at 3-fold axis and extend towards the N terminus of the domain similar to the discoidin-1 lectin-binding domain, which was crystallized with a ligand molecule bound to it (Figure 2.22 **c)**). Summarizing the observations reported above, we propose that residues that coordinate the metal ions in R1 and R2 RB domains are responsible for recognition and attachment to the sugar moiety of the bacterial LPS receptor.

Table 2.4 Results of Dali database structural similarity search for R2 and R1 pyocin fibers.
First number corresponds to the hits for the structure of R2 pyocin fiber, second number – to the structure of R1 fiber. Z-score – amount of structural similarity; RMSD – RMSD of structure fragments superposition; lali - number of residues was used in superposition; #res – number of residues in the compared structure from PDB; %id – percent of sequence identity in compared structures.

Z score	RMSD	lali	# res	% id	Hit description
7.9/7.9	3.2/7.7	94/100	254/253	14/14	Discoidin-1, Subunit A; GalNAc-binding lectin
7.8/7.8	3.2/4.8	92/96	255/256	13/16	Discoidin-2; GalNAc-binding lectin
6.7/8.1	2.4/2.5	72/80	100/99	15/18	Helix pomatia agglutinin; oligosaccharides binding lectin
6.5/6.1	5.3/5.4	111/114	511/509	7/13	Tail spike of bacteriophage SF6, RBP
4.2/-	6.6/-	114/-	263/-	4/-	p2 lactococal phage RBP
-/5.1	-/3.4	-/92	-/205	-/8	C-terminal head domain of the adenovirus short fiber
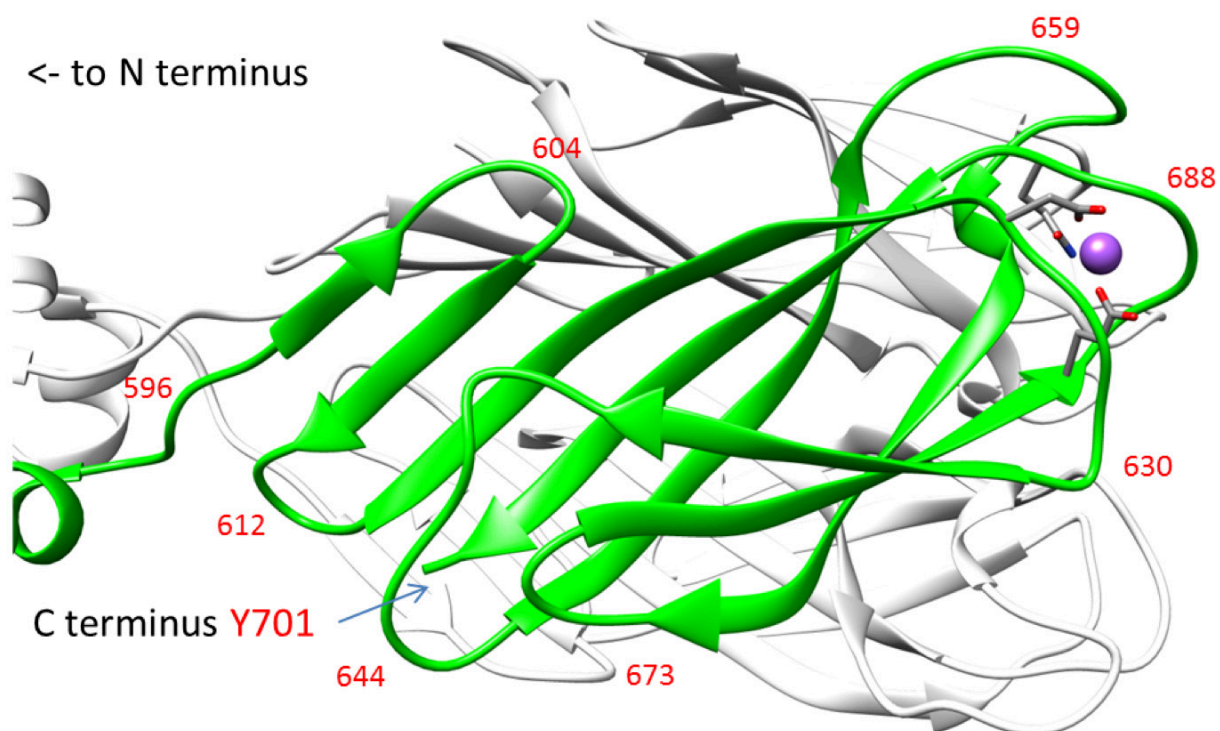
Figure 2.20 Structural details of the fiber C-terminal domain.
For the fold tracing simplicity one of the chains is colored in green. Key residues are numbered (R1 fiber model and numbering was used for illustration).
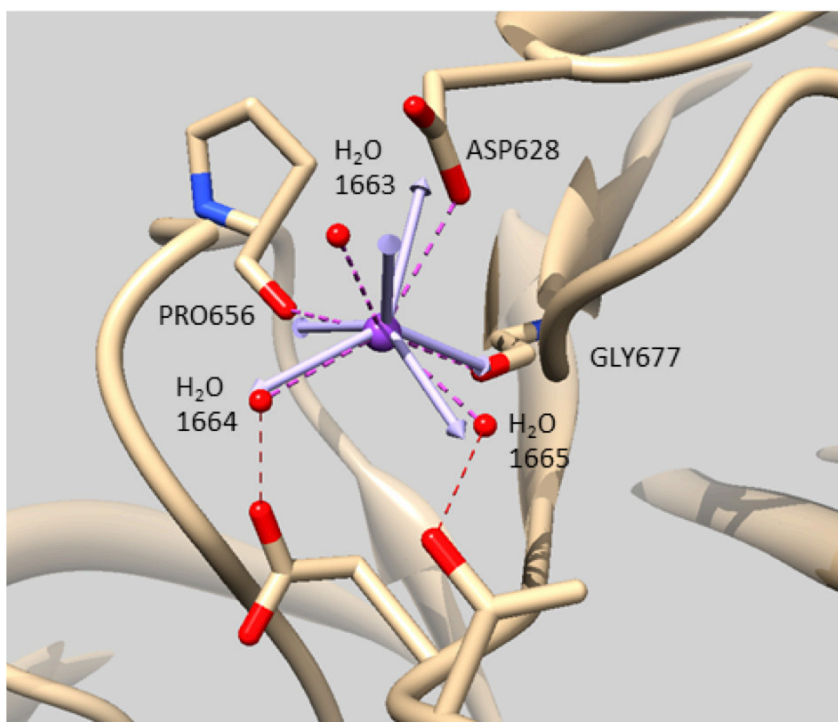
Figure 2.21 Analysis of the potassium coordination sphere in R2 RB domain.
Protein polypeptide chain showed with ribbon representation is colored in gold. Side chains are showed for residues coordinating potassium ion (in violet) or water molecules from potassium ion's inner coordination sphere. Ideal trigonal prism configuration of potassium ion coordination bonds is showed with violet-blue arrows. Short contacts (H-bonds and coordination bonds) are showed with dash lines.

Table 2.5 Bond length and probable geometry analysis within potassium coordination polyhedron.
Output is generated by UCSF Chimera (Pettersen, et al. 2004).

Coordinator	Distance (Å)	Distance RMSD	Best Geometry
HOH 1663.W O	2.82	0.000	N/A
GLY 677.C O	2.85		N/A
HOH 1664.W O	2.92		N/A
PRO 656.B O	2.92	0.389	square planar
HOH 1665.W O	2.96	0.822	square pyramid
ASP 628.C OD1	2.98	0.535	trigonal prism
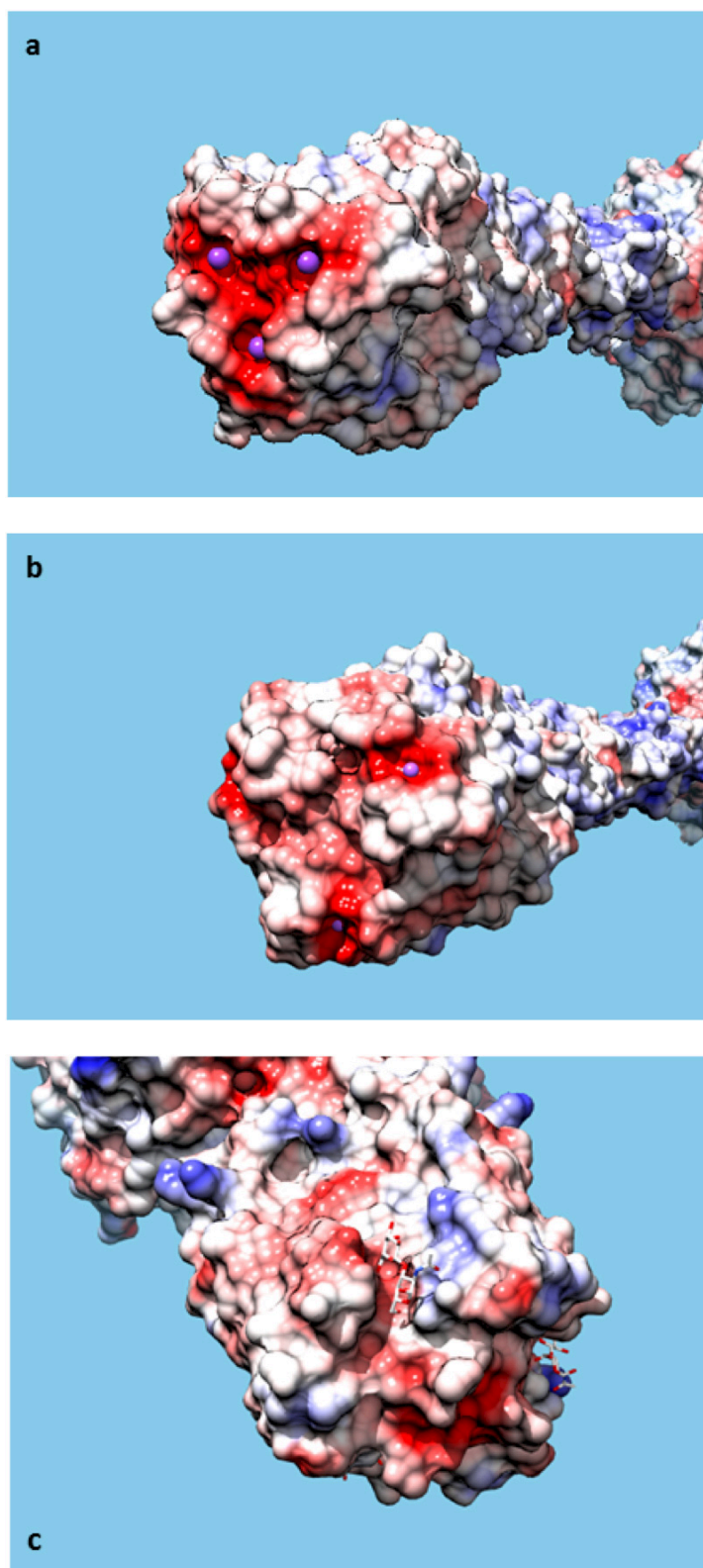
Figure 2.22 Surface representation of R2 a), R1 b) pyocin RB domains and discoidin-1 lectin-binding domain c) with mapped coulombic coloring.
K^+ and Na^+ ions are colored in magenta. Molecular surface is colored according to coulombic surface charge with blue, white and red corresponding to the most positive, neutral, and negatively charged patches, respectively.
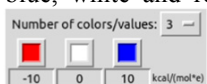
Köhler et al. (Köhler, Donner and van Delden 2010) described sugar moiety-binding specificity of different pyocins allowing us to derive saccharide-specificity of pyocin fibers (Figure 2.11). We purified bacterial LPS from *Pseudomonas* strains that are sensitive to R1 and R2 pyocins and set up co-crystallization and soaking experiments with these LPS and with monosaccharides that mimic sugar moieties of LPS from R1 and R2 sensitive strains. The L-rhamnose residue if LPS appeared to be critical for the R1 pyocin binding and the α-glucose – for the R2 pyocin. Therefore, we set up co-crystallization experiments for both types of fibers with the following sugars: L-rhamnose, D-glucose, D-galactose (that represented a sensitive monosaccharide and two controls).

Unfortunately, neither R1 fiber crystals that grew in the presence of L-rhamnose (3, 5 and 10 mM) nor R2 fiber crystals produced in the presence of glucose and galactose (3, 5 and 10 mM) contained any monosaccharide bound to the fiber molecules. Co-crystallization experiments of R2 pyocin fiber with bacterial LPS purified from a sensitive *Pseudomonas aeruginosa* strain 13s gave negative results as well (this LPS was purified by Dr. Dean Scholl, Avid biotics Corp., South San Francisco, USA). To overcome the problem of competition between the K^+/Na^+ ions and LPS/monosaccharides for receptor binding sites a series of experiments including 24 hours preliminary soaking of LPS/monosaccharides with the protein material was done, but this did not result in binding of the substrate to the fiber in the crystals. Unfortunately, no other crystallization condition lacking K^+/Na^+ ions was found.

(Figure 2.23 **a)**). Note, that this pyocins bind to different P. aeruginosa strains. Analysis of all 13 available at the GenBank pyocin fiber sequences (types R1 through R5) (Figure 2.23 **b)**) showed two divergent regions also located on the C-terminal and knob domains.

We propose that during infection the pyocin fiber and the baseplate functions similar to those of T4, in which the baseplate is activated by reorientation of the fiber. Clearly, two sites that are located at different part of a long fiber are needed for a reorientation, assuming that the first binding event occurs stochastically. In this model, both binding sites (either on the C-terminal domain or on the knob domain) can serve as a primary recognition site.
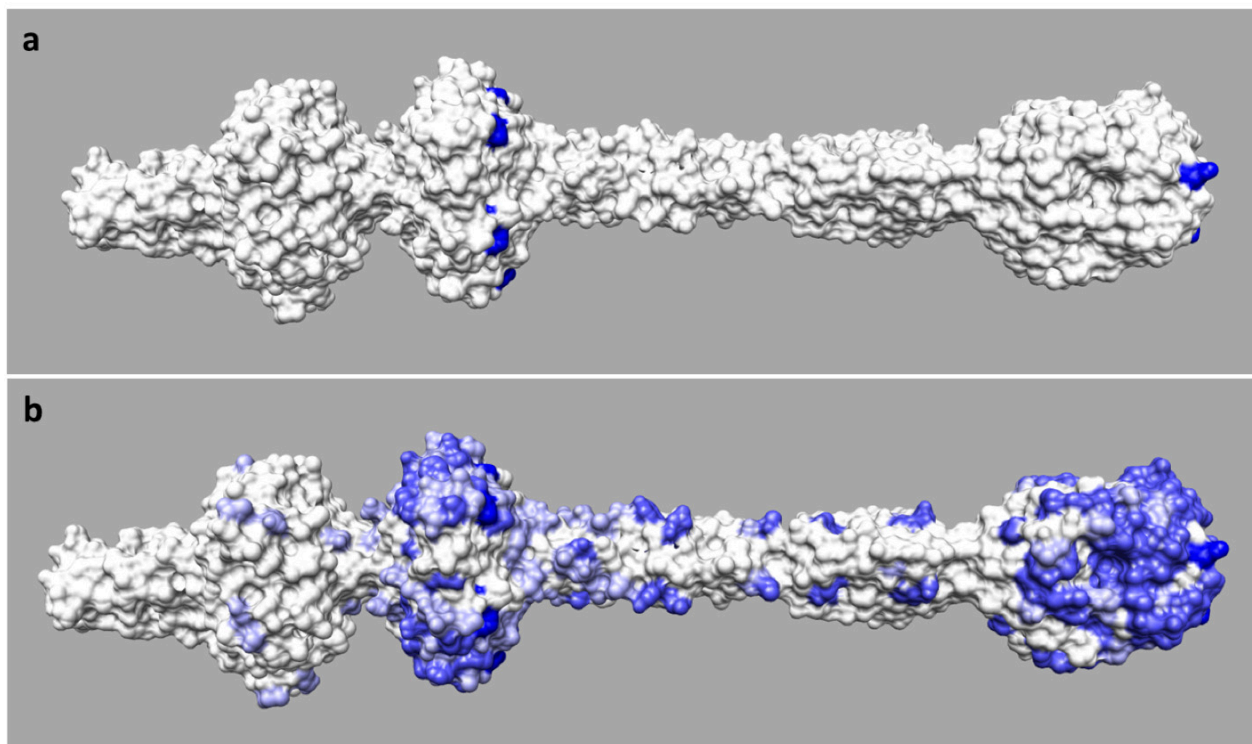
Figure 2.23 Structure conservation in pyocins and closest homologues.
a) map of structure preservations within R-type pyocin family: R2-R3-R4. Conservation threshold is 0.66; **b)** map of structure preservations within pyocins R1-R2-R3-R4-R5 and 13 closest homologues. Conservation threshold is 0.95. White color corresponds to fully conserved residues, blue to non-conserved. Blue-white gradient corresponds to a given residue conservation. Structure conservations are mapped onto surface representation of the R2 pyocin fiber.

3 CRYSTALLOGRAPHIC STADIES OF THE CELL-PUNCTURING DEVICES IN CONTRACTILE INJECTION SYSTEMS

3.1 Refined structure of bacteriophage T4 gp5 β-helix

3.1.1 Introduction

Bacteriophage T4 (Figure 3.1 **a)**) is a double-stranded DNA virus, which infects *Escherichia coli*. It belongs to the *Myoviridae* family of phages with long contractile tails. A 172 kbp genome, more than 40 structural proteins in multiple copies makes T4 one of the most complex viruses (Kutter, et al. 1990). It is also one of the best genetically and biochemically studied phages. The T4 virus particle consists of a 5-fold-symmetric head containing the genomic DNA, a 6-fold-symmetric bilayered contractile tail, and six long tail fibers attached to the baseplate (Eiserling and Black 1994). Bacteriophage T4 has a very efficient mechanism for infecting cells (Goldberg, Grinius and Letellier 1994). The key component of this process is the baseplate, located at the end of the phage tail, which regulates the interaction of the tail fibers and the DNA ejection machinery. The central part of the baseplate is a complex of gene product (gp) 5 (63kDa), gp27 (44kDa) and gp5.4 (10kDa) – a protein, whose existence was predicted more than 10 years ago but the identity has been established only recently by us (Kostyuchenko, et al. 2003) (Figure 3.1 **b)**). This complex is required to penetrate the outer cell membrane of *E. coli* and to disrupt the intermembrane peptidoglycan layer, promoting the subsequent entry of the phage DNA into the host. The structure of the gp5-gp27 complex has previously been determined using X-ray crystallography to 2.9 Å resolution (Kanamaru, et al. 2002), presenting 4 functional domains: 1) gp27; 2) gp5 N-terminal OB-fold domain; 3) gp5 lysozyme domain and 4) gp5 C-terminal β-helical domain (Figure 3.1 **c)**). By far the most remarkable of the above is the β-helical domain (Figure 3.1 **d)**), which was proposed (but not yet proven) to puncture the cell membrane during tail contraction (Kanamaru, et al. 2002).

Gp5 belongs to the class of trimeric fibrous proteins. There are only four known motifs among such kind of proteins, namely: the collagen triple-helix (Traub and Piez 1971), (Bella, et al. 1994), the α-helical coiled coil (other oligomers are possible, dimers, tetramers, and pentamers) (Lupas 1996), the triple β-spiral/shaft (van Raaij, Mitraki, et al. 1999) and the intertwined triple β-helix (van Raaij, Schoehn, et al. 2001), (Kanamaru, et al. 2002). The unique properties of oligomeric fibrous proteins, such as varying length, stability and oligomerization make them suitable scaffolding units for controllable production of one, two- or three-dimensional biological macroassembles. Collagens and alpha-helical coiled coils are well studied structures

and widely found in living organisms, unlike triple β-helices, which were discovered recently and require detailed studies on their structural organization, folding properties and stability.
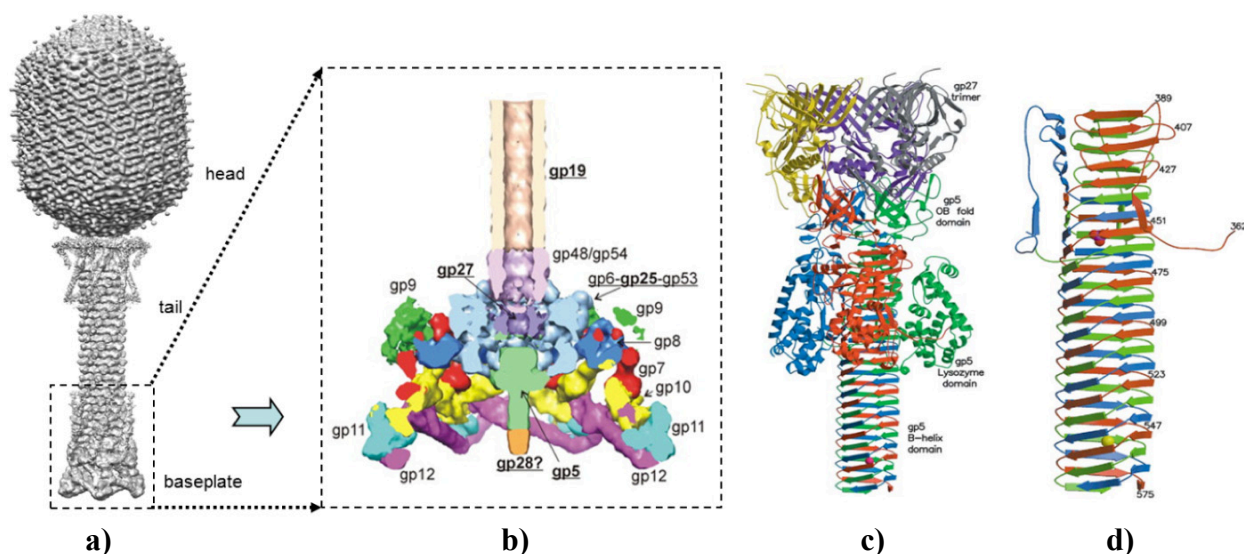
Figure 3.1 T4 cell-puncturing device placement in the phage particle.
a) general view of T4 phage; **b)** baseplate cut-away view; **c)** gp5-gp27 complex; **d)** gp5 β-helical domain. Figure was adapted from (Kostyuchenko, et al. 2003) and (Kanamaru, et al. 2002).

Recent attempts to express soluble and correctly folded isolated β-helical fragments failed indicating the necessity of an extra-oligomerization domain (chaperone), either native or artificial (Papanikolopoulou, et al. 2004). Also, no data existed on folding of isolated triple β-helix.

At least ten crystal structures containing an extended triple β-helix are known, all of them have been discovered in phages (Bernstein, et al. 1977), (Browning, et al. 2012) and Leiman lab, unpublished data. Out of all them, gp5 is the most perfect, longest and regular structure of trimeric β-helix.

The previously resolved structure of the gp5-gp27 complex highlighted many structural details, especially for gp5, however many details remained uncharacterized or characterized incorrectly. To improve on this, Dr. Sergei Budko in Andreas Engel's laboratory created several deletion mutants of gp5, comprising different parts of the C-terminal β-helical domain. These mutants were very stable SDS-resistant trimers. Sergei then moved to Michael Rossmann's lab, where he found crystallization conditions for one of these mutants, called gp5R483 – a C-terminal fragment, consisting of residues R483-G575 (Figure 3.2). This mutant represents the whole solvent exposed part of the β-helix domain. Gp5R483 crystals, produced by Dr. Budko diffracted to better than 2 Å resolution. I continued this project – refined initial crystallization conditions

which resulted in crystals diffracting up to 1.3 Å resolution and allowed to obtain a much more precise description of the remarkable intertwined β-helical structure.
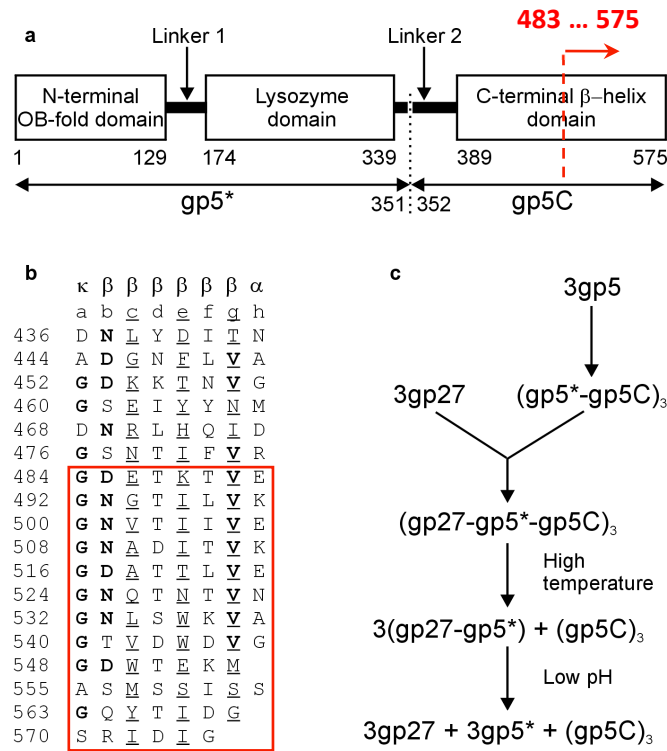
Figure 3.2 Map of the gp5R483 fragment.
In red are shown fragment borders. Figure adapted from (Kanamaru, et al. 2002).

3.1.2 Materials and methods

3.1.2.1 Construct engineering

Gene encoding gp5R483 deletion mutant was PCR amplified and cloned into the pHisTrx2 (Kammerer, et al. 1998) vector using restriction sites *BamHI* and *EcoRI*. The vector was designed to express fusion constructs containing an N-terminal His-tag, a thioredoxin A, a linker with a thrombin cleavage site and the fragment of interest. The DNA inserts were verified by Sanger dideoxy DNA sequencing.

3.1.2.2 Protein expression and purification

The recombinant protein was expressed as fusion construct at 37 °C in *E. coli* BL21 (DE3) host strain (Novagen) after IPTG induction at final concentration of 1 mM. The Selenomethionine (SeMet) mutant of R483 was expressed in modified M9 medium in the presence of SeMet using the B834 (DE3) strain of *E. coli*. Purification of the fusion proteins was carried out by immobilized metal affinity chromatography on a HisTrap HP column (Amersham Biosciences) and the separation of gp5 mutants after thrombin cleavage was carried out as described in the manufacturer's instructions. Before thrombin cleavage, the fusion proteins were additionally purified on an anion exchange HiTrap Q HP column (Amersham Biosciences). The cleavage was performed at 20 °C for 16 hours with thrombin (Novagen).

3.1.2.3 Crystallization, data collection and structure determination

Crystallization, data collection and structure determination for SeMet derivative of gp5R483 were done by Dr. Sergei Boudko.

The crystallization conditions for both, native and SeMet derivative proteins, contained 0.02M Calcium chloride, 0.1M Sodium acetate (pH 5.2), 15% 2-methyl-2,4-pentanediol (MPD). Despite the small amount, the presence of calcium was crucial for crystallization. The SeMet derivative crystals grew to a size of 0.6 mm x 0.6 mm x 0.6 mm in hanging drops after about one to two weeks at 20 °C. The crystals were briefly dipped into a cryoprotectant solution containing 0.02M Calcium chloride, 0.1M Sodium acetate (pH 5.2), 30% MPD and then flash frozen in a vaporized nitrogen stream at 100 K.

SeMet crystals were used for a three-wavelength, MAD data collection procedure (Hendrickson and Ogata 1997) at APS synchrotron (Argonne, Chicago). The programs DENZO and SCALEPACK were used to process the diffraction data sets (SeMet and native) (Otwinowski

and Minor 1997). The program SOLVE (Terwilliger and Berendzen, Automated MAD and MIR structure solution 1999) was used to compute the phases based on the 6 Se atoms per 285 residues. The phases were then improved by solvent flattening, and an atomic model (about 80% of all residues) was built with the program RESOLVE (Terwilliger, Automated structure solution, density modification and model building 2002), ARP/wARP 6.1 (Lamzin, Perrakis and Wilson 2001), and manually using the program XtalView (McRee 1999). The program Refmac5 (Winn, et al. 2011) was used for the refinement.

The native protein crystallization conditions were further fine-tuned by varying the pH of the buffer (between 4.6 and 5.4), the concentration of $CaCl_2$ (0.05 and 0.06 M) and the concentration of MPD (between 9 and 11 %) using the hanging drop method (2 µl drop size) in 24 deep-well plates. For crystallization, the protein was brought to the concentration of 20 mg/ml. The best conditions (0.05M calcium Chloride, 0.1M Sodium acetate (pH 5.2), 11% MPD) yielded native crystals of 0.3 mm x 0.2 mm x 0.5 mm after 4 days of incubating at +18 °C (Figure 3.3 **a)**). The mother liquor served as a cryoprotector for data collection in a vaporized liquid nitrogen stream at 100 K. Crystals belong to $C222_1$, #20 space group with $a = 57.61$, $b = 72.76$, $c = 130.49$ Å unit cell parameters with one gp5R483 trimer in the asymmetric unit. Diffraction data were collected at the SLS (Villigen, Switzerland) PXIII beamline using a wavelength of 0.9 Å. The diffraction extended up to 1.3 Å resolution with a very clean diffraction pattern and exhibited X-cross, which is a characteristic of a helical structure (Figure 3.3 **b)**). The first layer line could be seen at approximately 4.7 Å resolution, corresponding to the spacing between β-strands. The structure was solved by molecular replacement using a MOLREP program from CCP4 program suit (Winn, et al. 2011). The structure of gp5R483 SeMet served as a search model. The structure was initially refined by Refmac5. The final refinement was done with SHELXL (Sheldrick 2008). Details of data reduction and refinement are given in the Table 3.1.
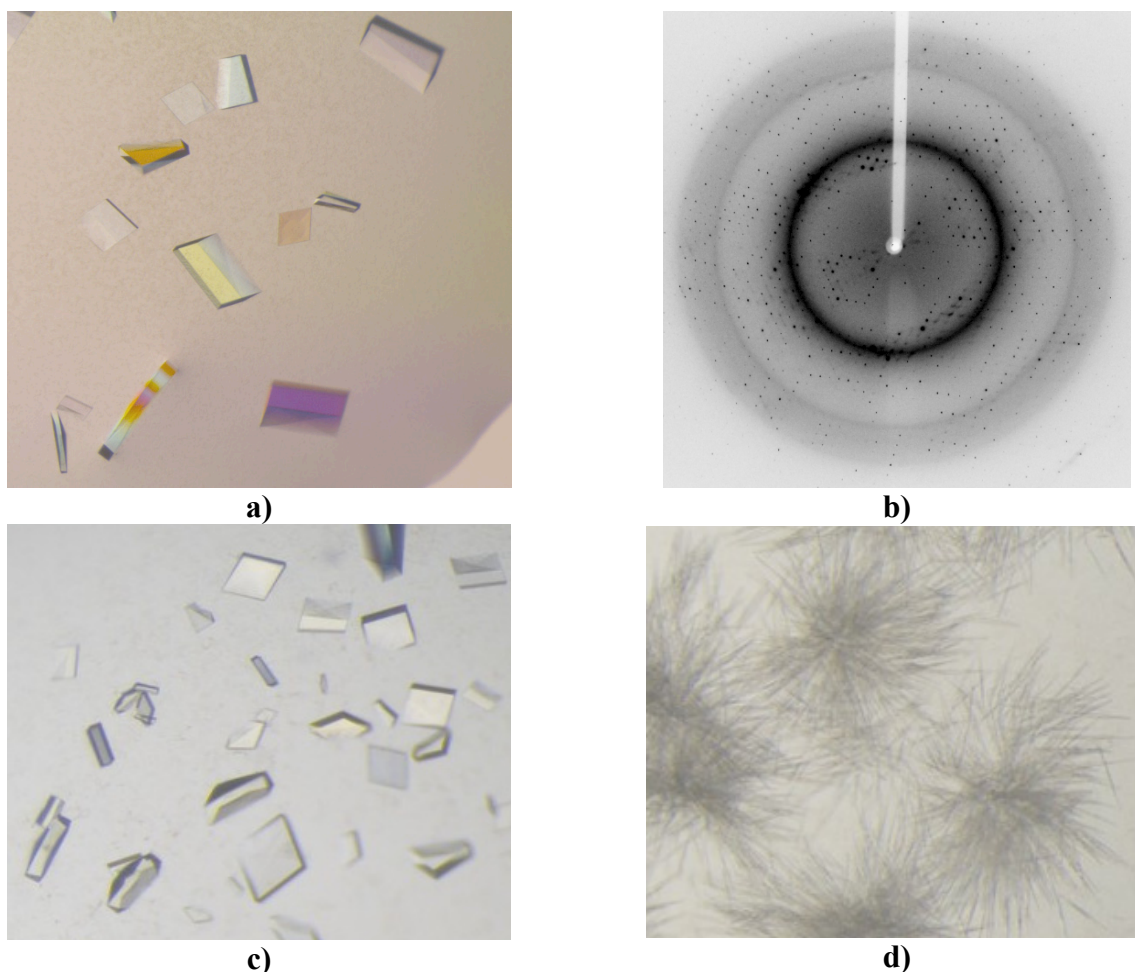
Figure 3.3 Gp5R483 crystallization and data collection.
a) crystals of gp5R483 grown in the presence of Ca^{2+}; **b)** diffraction pattern of the gp5R483 crystals **a)**, diffracting up to 1.3 Å resolution; crystals of gp5R483 grown in the presence of Sr^{2+} **c)** and Zn^{2+} **d)**.

To determine the origin of the metal cation incorporated into the hydrophobic core of the protein several other bivalent cations were tried for crystallization (Mg^{2+}, Sr^{2+}, Zn^{2+}, Ni^{2+} and Mn^{2+}). Their initial concentration was derived from that of $CaCl_2$ (10, 30, 50 and 70 mM) multiplied by a coefficient of difference in solubility of the hexahydrates (or, the most stable hydrates) according to the formula:

(New metal concentration) = ($CaCl_2$ concentration) * (solubility of $MeCl_2$) / (solubility of $CaCl_2$)

Only conditions containing Sr^{2+} (0.06-0.1M $SrCl_2$, 11% MPD, 0.1M NaAc, pH 5,4) resulted in crystals suitable for X-ray data collection (Figure 3.3 **c)**). Zn^{2+} containing conditions (0.011-0.015M $ZnCl_2$, 11% MPD, 0.1M NaAc, pH 5,4) resulted in very thin needles that were not suitable for crystallographic studies) (Figure 3.3 **d)**). Unlike the earlier experiments with $CaCl_2$ grown crystals, the mother liquor was a poor cryoprotector for $SrCl_2$ grown crystals. A good cryoprotector was obtained by increasing the concentration of MPD to 25%. The crystals were briefly washed in this solution before flash freezing in a vaporized nitrogen stream at 100 K. X-

ray excitation spectrum was recorded for all types of crystals. The $SrCl_2$ grown crystals showed a clear presence of the Sr in the X-ray fluorescent emission spectrum. Diffraction data was collected at the wavelength corresponding to the Sr absorption edge ($\lambda = 0.769$ Å).

Three data sets using identical inverse beam data collection parameters were collected for: 1) a $CaCl_2$ grown crystal and $CaCl_2$ containing cryoprotector; 2) $SrCl_2$ grown crystal, $SrCl_2$ cryoprotector; 3) $SrCl_2$ grown crystal, $CaCl_2$ cryoprotector. The latter combination was used to remove the unbound Sr cations. All gp5R483 derivatives crystallized in $P2_12_12_1$ or $C222_1$ space groups with very close unit cell parameters to those of the 1.3 Å resolution structure collected earlier (Table 3.1). Indexing, integrating and scaling were done using HKL2000 (Otwinowski and Minor 1997). The refined model of gp5R483 was used to solve the phase problem of the Sr derivatives by molecular replacement employing PHASER (McCoy, et al. 2007). Anomalous map calculations were done with FFT program from CCP4 program package (Winn, et al. 2011). Data collection and data processing statistics are given in Table 3.1. The structure of gp5R483 mutant was deposited in the Protein Data Bank under accession number 4JJ2.

3.1.2.4 Mass spectrometry analysis

In order to determine the molecular weight of the inclusion compounds, few sets of ESI-Q-TOF MS experiments were performed by Dr. Laure Menin at the mass spectroscopy service at EPFL, BCH 1525 on Q-TOF Ultima (Waters) instrument. The first set of experiments was made using the gp5R483 protein in the conditions used for crystallization with a concentration of 20mg/ml. To reduce the low molecular weight impurities, for the rest MS experiments, gp5R483 was first crystallized, the crystals were then collected, washed in the crystallization solution free from protein, and dissolved in distilled water. Afterwards, the crystals were dialyzed against distilled water by 3 repeated buffer exchanges using a Millipore microconcentration device (Merck Millipore) with the cutoff of 10 kDa. The final concentration of thus purified protein was 3.9 mg/ml and 6.7 mg/ml for the 2nd and 3rd MS experiments respectively. In the first and second experiments the increasing of the sample cone voltage was chosen as the denaturation-in-source method. The heating was used for the third experiments as the denaturation-in-source method because in the previous two experiments a large amount of the trimer could still be observed.

3.1.2.5 Gp5R483 inclusion compounds extraction

To extract the inclusion compounds from the gp5R483 trimer a modified Folsch procedure was applied (Folsch, Lees and Sloane Stanley 1957). Aqueous solution of gp5R483 crystals, approximate volume 300 μl, was mixed in a 1.5 ml low-bind reaction tube (Eppendorf) with 750μl of the extraction mixture composed of dichloromethane and methanol, 2:1 v/v ratio. The tube was vortexed at 16,000 RPM for about 3 minutes. Then organic phase was removed using a syringe (Hampton), transferred to another tube and dried under vacuum until a resulting volume of approximately 150 μl. This sample was then used for GC-IE-MS and GC-CI-MS analysis using EI/CI-1200L (Varian) GC-MS, equipped with Capillary column VF-5ms (5% phenyl-methyl 95% dimethylpolysiloxane column, 0.25mm x 30 m).

Table 3.1 X-ray data collection and refinement statistics of gp5R483.
* Two acquisitions, for high and low resolution; same crystal was used.
** Main task was BDF synthesis.
*** Number of peaks above the noise threshold.
Data in parenthesis represent statistics for the highest resolution shell.

Data collection	a) CaCl$_2$/CaCl$_2$ **	b) SrCl$_2$/SrCl$_2$**	c) SrCl$_2$/CaCl$_2$**	d) CaCl$_2$/CaCl$_2$
Beamline	PXII (SLS)	PXII (SLS)	PXII (SLS)	PXIII (SLS)
Wavelength (Å)	0.769	0.769	0.769	0.975
Oscillation angle, step (°)	0-100, 180-280, 1	0-100, 180-280, 1	0-100, 180-280, 1	0-360, 0-360, 1*
Number of frames	200	200	200	720 (360 + 360)*
Space group	P212121, #19	C2221, #20	C2221, #20	C2221, #20
Cell dimensions:	57.40, 72.14,	57.43, 72.34,	57.42, 72.19,	57.61, 72.76,
a, b, c (Å)	129.59	129.66	129.16	130.49
Number of trimers per asymmetric unit	1	1	1	1
Resolution (max) (Å)	1.7	1.5	1.7	1.3
Rmerge (%)	0.066 (0.34)	0.075 (0.415)	0.064 (0.328)	0.069 (0.299)
$<I / \sigma I>$	7.6	9.1	12.3	11.3
Completeness (%)	97.24 (98.7)	97.79 (85.5)	94.18 (96.3)	99.59 (87.0)
Redundancy	8.5 (8.4)	7.7 (5.5)	8.8 (8.6)	10.2 (4.3)

Refinement				
Number of reflections				
Working				69991
Test				4689
R$_{work}$ / R$_{free}$				0.1527/0.1777
B-factor (Å^2)				19.4
R.m.s. deviations				
Bond lengths (Å)				0.0169
Bond angles (°)				0.0413
Number of atoms				
Protein				2259
Solvent and ligands				388
Ramachandran plot (%)				
Most favored				97.4
Additionally allowed				2.6
Outliers				0.0
Number of anomalous scatterers (asymmetric unit)	0***	13***	13***	–

3.1.3   Results and discussion

3.1.3.1   Crystal structure of gp5R483

The centered orthorhombic crystals of the gp5R483 mutant contain one molecule of the trimer (chains A, B and C) per asymmetric unit. Each chain resembles a corkscrew consisting of 12 beta-strands. Three such corkscrews intertwine and form a triangular prism, whose faces are represented by 3 parallel β-sheets (Figure 3.4). Based on the inner core packing the structure could be visually separated into two approximately equal halves, the amino- and carboxy-terminal halves. The amino-terminal half has a hydrophobic inner channel about 10 Å in diameter in which three elongated density moieties were found, whereas the carboxy-terminal half is tightly packed with large hydrophobic residues such as tryptophans and methionines. There is a metal ion, sandwiched between these large residues. It was initially identified as potassium (Kanamaru, et al. 2002).

Two gp5 trimers interact with each other with their C-terminal parts to form a dimer. In the area of close contact, the distances between the closest Cα atoms of the neighboring trimers within the dimer are in the range of 4.6-4.9 Å (Figure 3.8). The high resolution structure of the C-terminal fragment of the gp5 β-helix showed a great deal of flexibility with about 15% of residues showing double conformations, especially near the C terminus. The main chain of three C-terminal terminal residues for one of the chains had a double conformation in addition to that of the side chains.
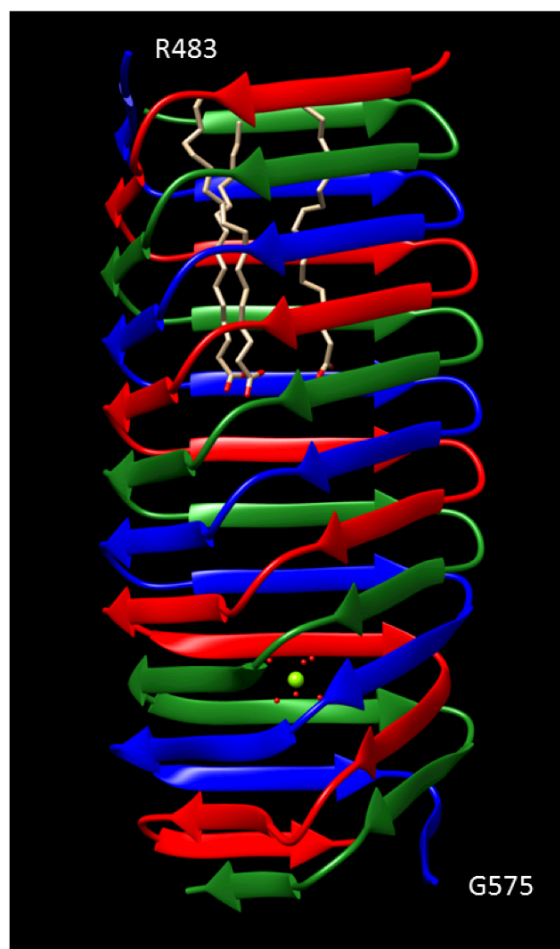
Figure 3.4 Ribbon diagram of the gp5R483 structure.
Three polypeptide chains are colored in red, green and blue. N- and C-termini are indicated. Inclusion compounds such as fatty acid molecules and Mg^{2+} cation are also shown.

3.1.3.2   Analysis and identification of the buried metal ion

The C-terminal half of the protein contains a buried ligand that is sandwiched between large hydrophobic residues (Figure 3.5 **a)**) and coordinated by three symmetry-related glutamates Glu552 (Figure 3.5 **b)**). The 1.3 Å resolution structure of gp5R483 allowed a clear analysis of the coordination sphere of this ligand. The ligand consists of a centrally positioned atom – most probably a metal ion – that is hydrated by 6 water molecules (Figure 3.5 **b)**, Figure 3.6). Metal geometry analysis in Chimera (Pettersen, et al. 2004) (Table 3.2) showed that the coordination polyhedron could be described as a slightly distorted octahedron with a RMDS of 0.24 Å (Table 3.3) (Figure 3.6).

In order to establish the identity of the metal cation, the X-ray excitation spectra of several crystals were recorded at the SLS and APS synchrotrons. None showed any signature lines for metals. An attempt was made to establish the identity of the metal by placing a light metal ion into the putative metal site position and performing the crystallographic refinement. Na^+, K^+, Ca^{2+} and Mg^{2+} ions were tried. Taking into account the coordination bond distances, the geometry of the coordination polyhedron and the coordination number, the choice was narrowed down to Ca or Mg – two very close analogues. With all other equal parameters, coordination bond lengths were closer to the value for Mg: the experimental value is 2.16 ± 0.06 Å; database $Mg^{2+}–H_2O = 2.17 \pm 0.15$ Å comparing to $Ca^{2+}–H_2O = 2.46 \pm 0.22$ Å (MESPEUS_10 was used for this analysis (Hsin, et al. 2008)). Temperature factor refinement was less conclusive but it appears to favor Ca (Table 3.4). Unfortunately, the absorption edges for either Mg or Ca are not easily accessible at synchrotron beamlines (Ca K-edge = 4.0381 keV, Mg K-edge = 1.3050 keV). Therefore there are no decisive means of proving that the metal is Mg using spectroscopic techniques. Notably, there is only one Mg atom per trimer comprised of 2010 non-hydrogen atoms, further complicating the analysis.

a)   b)
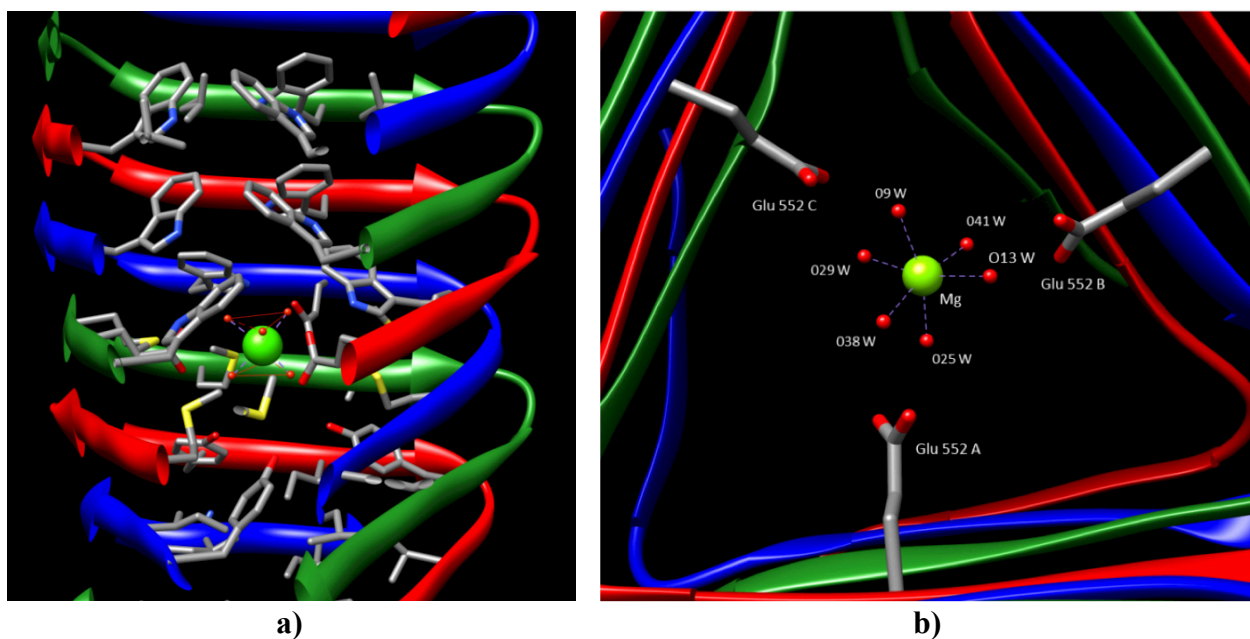
Figure 3.5 Metal ion buried inside the gp5R483 core.
a) water coordination polyhedron, buried deep into the hydrophobic core; b) hydrated Mg^{2+} ion, coordinated by three glutamates (Glu552).
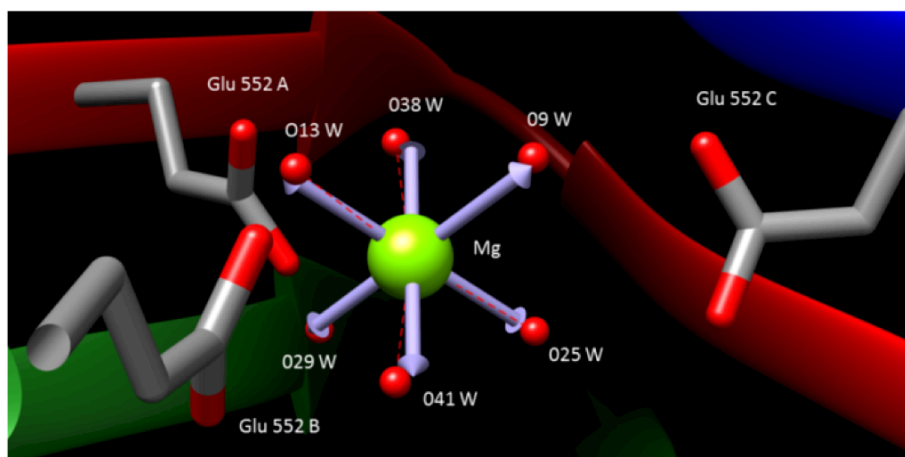
Figure 3.6 Coordination polyhedron geometry analysis of buried magnesium ion.
Outer coordination sphere (Glu residues) and inner (water molecules) are shown. In purple is shown idealized octahedral bond configuration for Mg^{2+} ion.

Table 3.2 Bond length and probable geometry analysis within magnesium coordination polyhedron.

Coordinator	Distance (Å)	Distance RMSD	Best Geometry
HOH 41.S O	2.09	0.000	N/A
HOH 29.S O	2.11		N/A
HOH 25.S O	2.14		N/A
HOH 38.S O	2.17	0.602	tetrahedral
HOH 9.S O	2.21	0.465	trigonal bipyramid
HOH 13.S O	2.25	0.240	octahedron

Table 3.3 Magnesium coordination polyhedron bond angles analysis.

Angle	Value	Ideal
O41 Mg O38	169.39º	value for the
O9 Mg O29	165.04º	octahedron:
O13 Mg O25	165.34º	180º

Table 3.4 B-factors ($Å^2$) of Mg^{2+} and Ca^{2+} ions, surrounding waters and the Glu552 O1 and O2 atoms.

Atoms GLU	Value	Atoms W	Value
O1A	21.74	O29W	29.57
O2A	19.12	O38W	27.05
O1B	22.15	O41W	25.52
O2B	20.47	O13W	22.32
O1C	20.33	O9W	22.05
O2C	18.74	O25W	23.81
Mg^{2+}	19.36		
Ca^{2+}	25.53		

The role of the buried metal ion is to neutralize the charge of the three buried Glu residues suggesting that this metal ion is present in the structure from the moment the protein folds into a trimer. However, $CaCl_2$ was used in the crystallization conditions at the concentration of about 50 mM and therefore it might have substitute the original ion during crystallization. We therefore decided to set up a set of crystallization experiments in the presence of different bivalent metal salts (see **3.1.2.3**). Two additional sets of crystallization conditions were obtained which gave crystals, the first one containing Sr^{2+} (a close analogue of Ca and Mg) and the other one – Zn^{2+} (Figure 3.3). The crystals grown in the presence of Zn^{2+} were too thin to study by X-ray single crystal diffraction.

We then collected diffraction data at the wavelength corresponding to the Sr absorption peak (λ = 0.769 Å) using the Ca-grown crystals and Sr-grown crystals and calculated Bijvoet difference Fourier (BDF) maps. The Sr-grown crystals were washed in the cryoprotector solution, containing Ca^{2+}, instead of Sr^{2+} (to remove unbound/non-specifically bound Sr^{2+} and reduce the noise level of anomalous map).

The Bijvoet difference Fourier (BDF) synthesis showed very prominent peaks for the Sr cations (Figure 3.7 **a)**, **b))**, which bind to the exterior of the molecule. We found that clusters of strong Sr anomalous sites are formed at the C-terminus–C-terminus contact area of two neighboring gp5R483 trimers that form a dimer of trimers (Figure 3.8). Not surprisingly, no significant peak was found in the Ca grown/Ca cryo dataset. In addition to the strong peaks, many weaker peaks were found in the BDF maps of the Sr-grown/Sr-cryo and the Sr-grown/Ca-cryo datasets. Sr binds to protein molecules non-covalently, and the distribution of Sr binding sites occupancies does not have a sharp falloff (Figure 3.9). The Sr site was considered to be above the noise level if it satisfied the following criteria:

1. Strong peak in BDF synthesis;
2. Strong peak in 2Fo-Fc map;
3. Proper chemical environment (appropriate distance to the protein molecule).

We found that both, Sr-grown/Sr-cryo and the Sr-grown/Ca-cryo datasets contained a peak close, but not precisely overlapping with the buried Mg cation. This peak was 17th and 12th in height in the Sr-grown/Sr-cryo and the Sr-grown/Ca-cryo datasets, respectively. In both of these datasets, the peak is at or below the noise level of the map.

This proves that the buried metal cation does not diffuse into or out of the protein during crystallization, but instead is present from the moment the protein folds into a trimer in the cell cytoplasm.
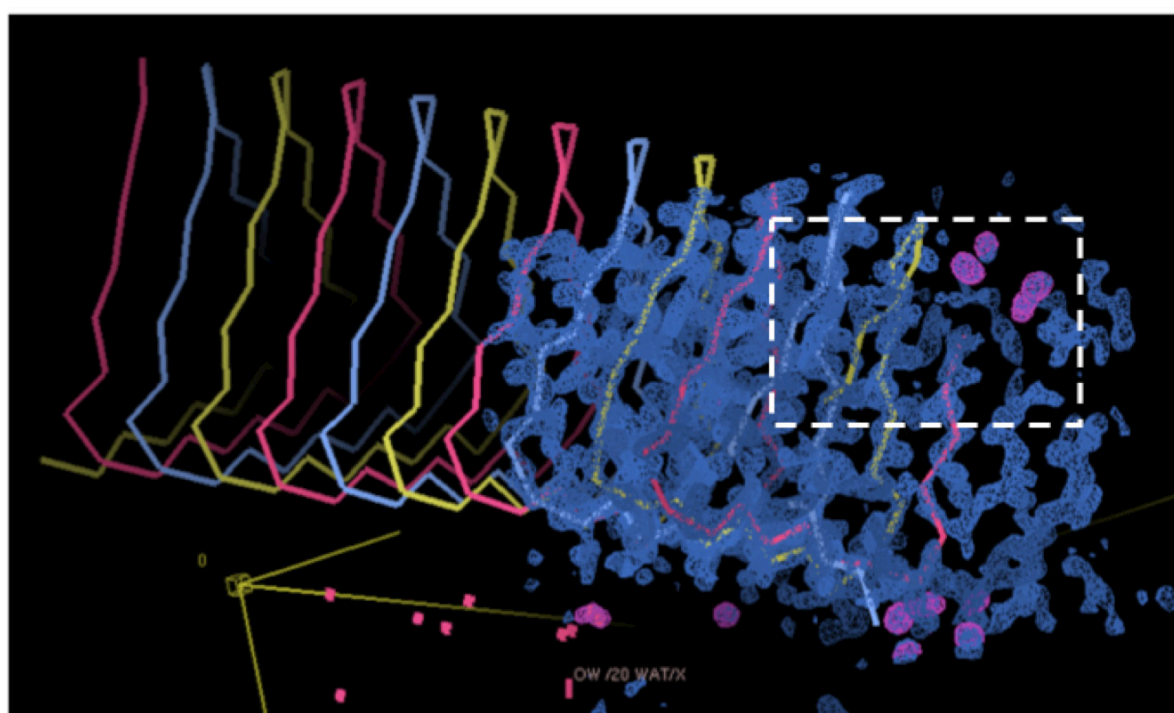
a)
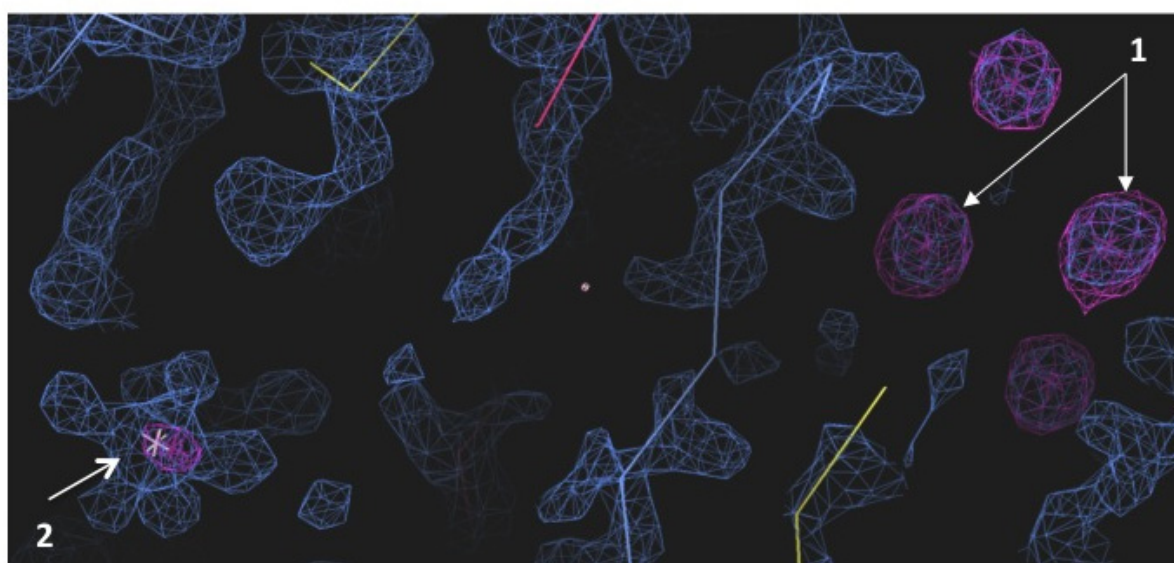
b)

Figure 3.7 Sr anomalous peaks in the Bijvoet difference Fourier (BDF) synthesis.
a) Cα trace diagram of the gp5R483 with 2Fo-Fc map (blue, contoured at 2.3 σ) for the protein and associated ligands and BFD map (purple, contoured at 6.0 σ) of the Sr cations bound to the exterior of the protein in the Sr-grown/Ca-cryo dataset; **b)** zoomed region showed in **a)** with a dotted line. 1 – exterior Sr anomalous peaks; 2 – Mg^{2+} polyhedron density).
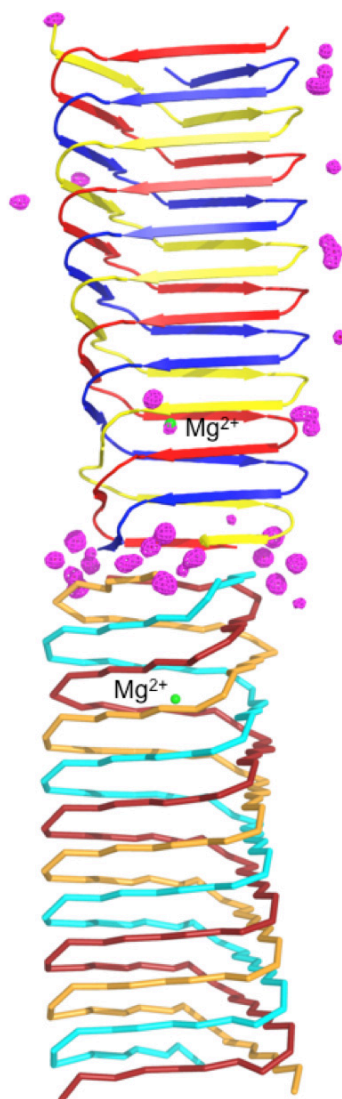
Figure 3.8 Gp5R483 dimer and Sr^{2+} sites.
Ribbon diagram (top trimer) and Cα trace (bottom trimer) representation of gp5R483 dimer. Exterior Sr anomalous peaks are colored in magenta. Buried Mg ions are colored in green and labeled.
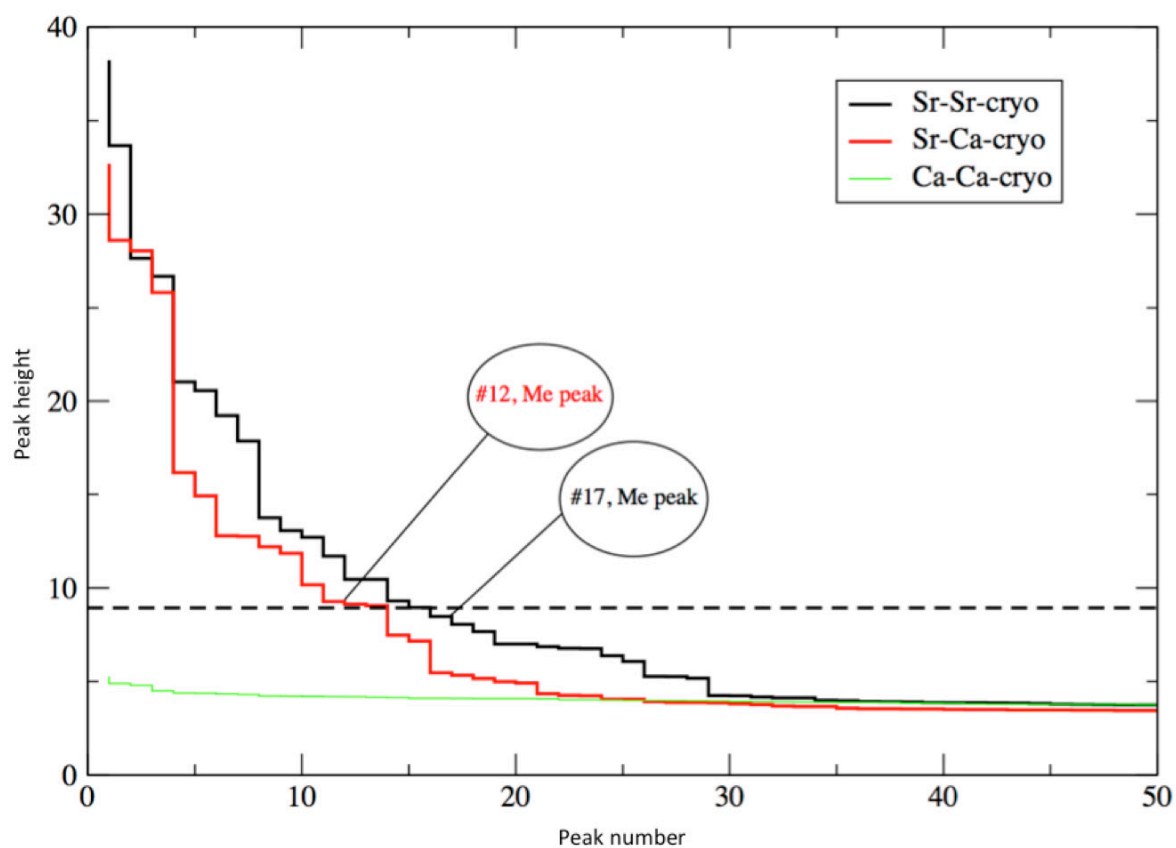
Figure 3.9 The occupancy distributions for Sr^{2+} anomalous peaks.
Black, red and green curves correspond to peaks from Sr-grown/Sr-cryo, Sr-grown/Ca-cryo and Ca-grown/Ca-cryo datasets, respectively. The dashed line is the noise level.

3.1.3.3 Identification of the internal extended compounds

The N-terminal part of the β-helical prism (residues 484 to 521) is hollow and forms a large cage with hydrophobic walls. There are three symmetry-related, ~20 Å long density moieties inside this compartment. The electron density and Uppsala natural small molecule database (Kleywegt 2007) search suggest that this density might correspond to a fatty acid molecule with a tail of 15-18 carbon atoms and the MW between 240 and 290 Daltons (Da) (Figure 3.10 **a)**). An attempt was made to establish the identity of these molecules using various mass-spectrometry (MS) techniques. The MS experiments show that the difference between the mass of the gp5R483 trimer, containing these inclusion compounds (family of peaks K), and the mass of the monomer times three ("empty" trimer, family of peaks D, that was predominating) is 757 Da**,** which might correspond to the mass of 3 inclusion fatty acids – 252 Da per each molecule, provided they are the same (Figure 3.11). Considering the existence of these two forms of gp5R483 trimer (with and without inclusion compounds) and nonconcluding analysis of the low MW zone of the spectra, was decided to conduct MS experiments in denaturation conditions. The screen for these conditions was done and included an increase of the sample cone voltage (SC) and the collision energy (CE), sample heating and usage of a denaturation buffer (formaldehyde, water, and acetonitrile in proportion 1:49:50 v/v). Each of these methods resulted into a trimer dissociation: intensity of the peak, corresponding to the gp5R483 trimer decreases when intensity of the peak, corresponding to the monomer dramatically increases (Figure 3.12 shows such dissociation achieved by increasing of the SC and CE), but, most probably, due to the dissociation of the peak D, corresponding to an empty trimer, as no significant peak stands out in the low MW zone. Some MS experiments were also performed in the negative mode (results were not informative, data not shown).

The low MW zone of the spectra was analyzed for each MS experiment as well (Figure 3.13). Initially, MS analysis was conducted on the protein sample previously used to set up crystallization trials. The spectra showed peak at 235 Da (Figure 3.13), which is near to the area of the expected mass value and a second peak at 339 Da, which was not a satisfactory value for the expected mass range, calculated from the electron density. On the other hand were found peaks (245 Da, 272 Da, 282 Da, and 301 Da) whose intensities are dependent on the applied cone voltage and the collision energy, suggesting that they might result from denaturation of the protein, which occurs as the cone voltage and collision energy are increased. The peaks stand out dramatically in comparison to 235 and 339 Da peaks as seen in Figure 3.13 bottom spectra, where the highest sample cone voltage and collision energy were used (SC = 100 V, EC = 15V)).

However, the spectra in this region are highly contaminated with impurities. Therefore a precise determination of either of the above listed signals corresponding to one of the inclusion compounds was impossible.

To reduce the low molecular weight contamination, the MS experiment was run on the protein purified by crystallization. Was crystallized a large amount of the protein, crystals were collected, washed in the protein-free crystallization solution and dialyzed against pure water (see **2.1.2.4**). This sample was then used for the MS experiments. The resultant low MW spectrum is shown in Figure 3.14. This time noticeable peaks were observed for the fragments with the MW of 269 and 338 Da. The height of these peaks dramatically increased after incubation of a sample during one hour at 95 °C. The 269 Da peak was also observed during another round of analysis along with 295 Da peak Figure 3.15. However both of them were lost in the high level of the background noise that stood out after an increase of the SC and CE.

Poor reproducibility of the peaks in the low MW zone, heterogeneity of the peak corresponding to the trimer + inclusion compounds (McCoy, et al. 2007) (McCoy, et al. 2007) suggest that the composition of the inclusion compounds is heterogeneous. Different dissociating conditions were tried were insufficient to unfold the trimer completely and liberate enough of the inclusion compounds for further MS/MS experiments. The same problem was encounted by the authors in (Potier, et al. 2003). They found the possibility to overcome these difficulties by carrying an organic extraction of the inclusion compounds followed by GC-MS analysis of the extract. This allowed them to determine the composition of the heterogeneous mixture. A modified Folsch procedure (see Materials and methods, **2.1.2.5**) was applied for the dissolved gp5 crystals sample. The chromatogram obtained during the experiment had 5 significant peaks, which were in the area of fatty acids elution. The mass spectra of these peaks contained the fingerprints corresponding to the 16:0, hexadecanoic acid (palmitic acid) with probability of 72.4 %, peak 1; 18:0 octadecanoic acid (stearic acid) with probability of 70.0 %, peak 2_1; 18:1 octadecenoic acid and isomers (oleic acid) with probability (sum) of 32.64 %, peak 2_2; esters of the hexadecanoic acid, probability (sum of the main ones) > 90.0%, peak 3 and esters of the octadecenoic/octadecenoic acid, peak is broaden, and most probably contains 2 non-resolved peaks of each type of esters.

These findings prove that the inclusion compounds are a mixture of a few fatty acids with 26-28 Da mass difference, which leads to a heterogeneous mass distribution of the peak corresponding to the gp5R483 trimer with inclusion compounds.

As a final step, the atomic models of the stearic, oleic and palmitic acids were built, placed into the corresponding electron density inside the hydrophobic channel of gp5 and refined. Taking into account that the electron density corresponding to the inclusion compounds was weaker than for the protein residues, the refinement may be considered as successful – no peaks above 3.5 RMSD in Fo-Fc map were detected, the geometry and density fit for the ligands are correct (Figure 3.10 **b)**, **c))**. Mapping the refined fatty acid models into the cavity of the gp5R483 (surface representation) shows that 3 fatty acids fill the cavity space very well (Figure 3.10 **d))**.

It could be hypothesized, that the function of these inclusion compounds is to stabilize the hydrophobic core of the protein, to constitute some sort of hydrophobic rod of rigidity, especially considering how difficult is to denature the trimer containing these molecules.
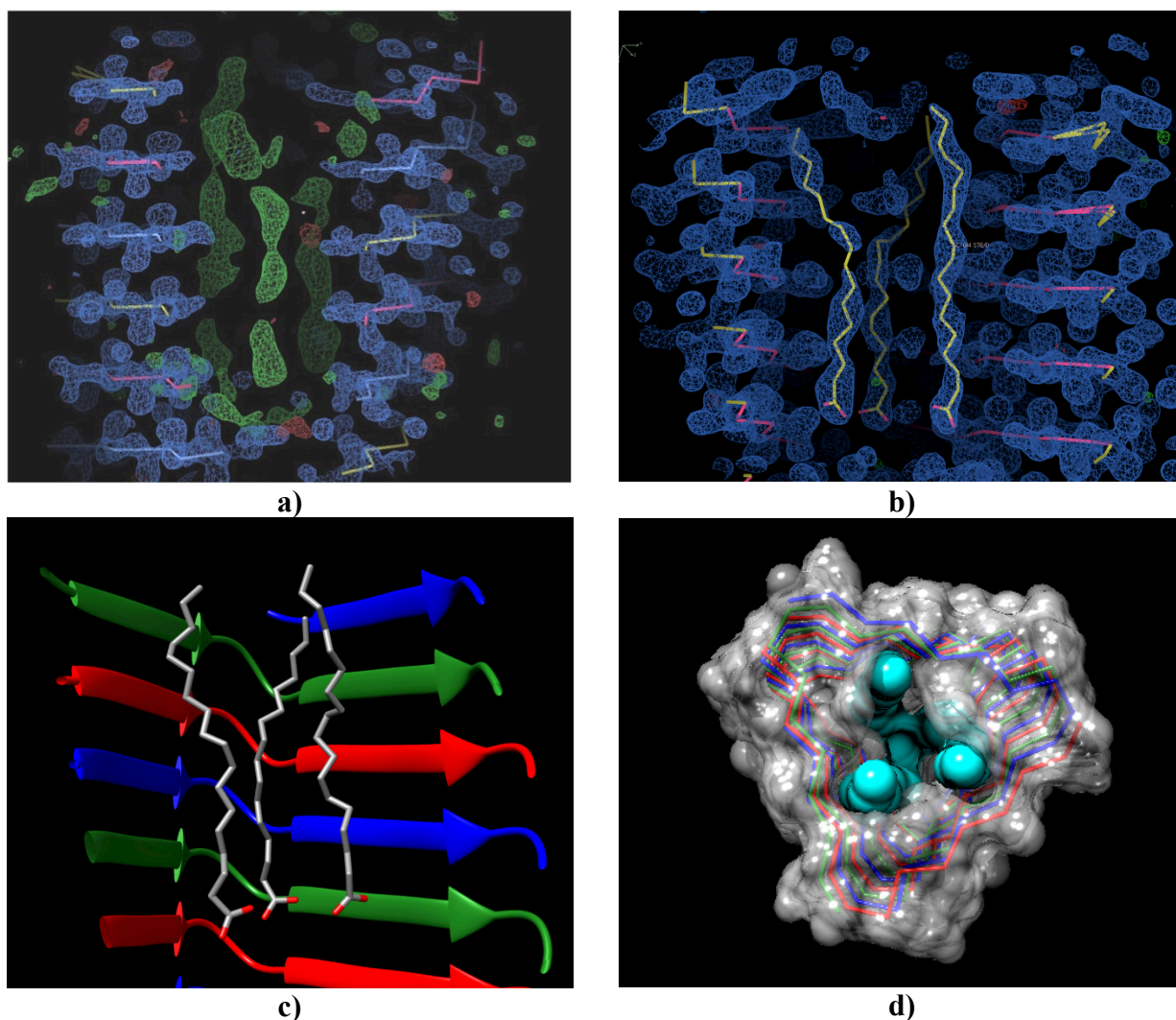
Figure 3.10 Identification of the internal extended compounds.
a) initial, model-free Fo-Fc map (green), which shows the inclusion compounds; **b)** three fatty acids, fitted into the density of 2Fo-Fc map; **c)** the model of the three fatty acids inside of the internal cavity of the β-helix (cut away view, ribbon diagram representation for protein polypeptide chains); **d)** top view of the inclusion compounds inside the β-helix cavity; protein polypeptide chains are represented with grey semi-transparent molecular surface and Cα trace (red, green and glue), inclusion compounds are represented with cyan molecular surface.
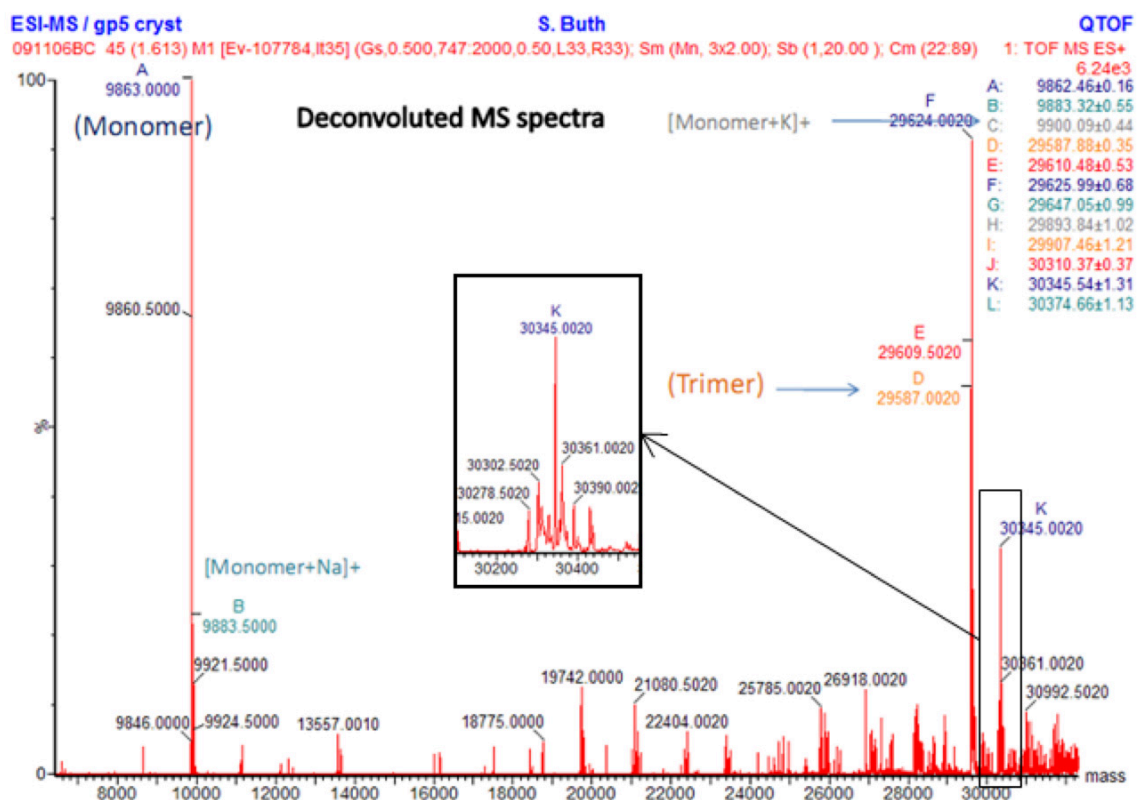
Figure 3.11 The mass-spectra, containing the information about different forms of gp5R483.
Peak A corresponds to the mass of the monomer, peak D (Ax3) to the mass of the trimer and peak K to the mass of the trimer with inclusion compounds.
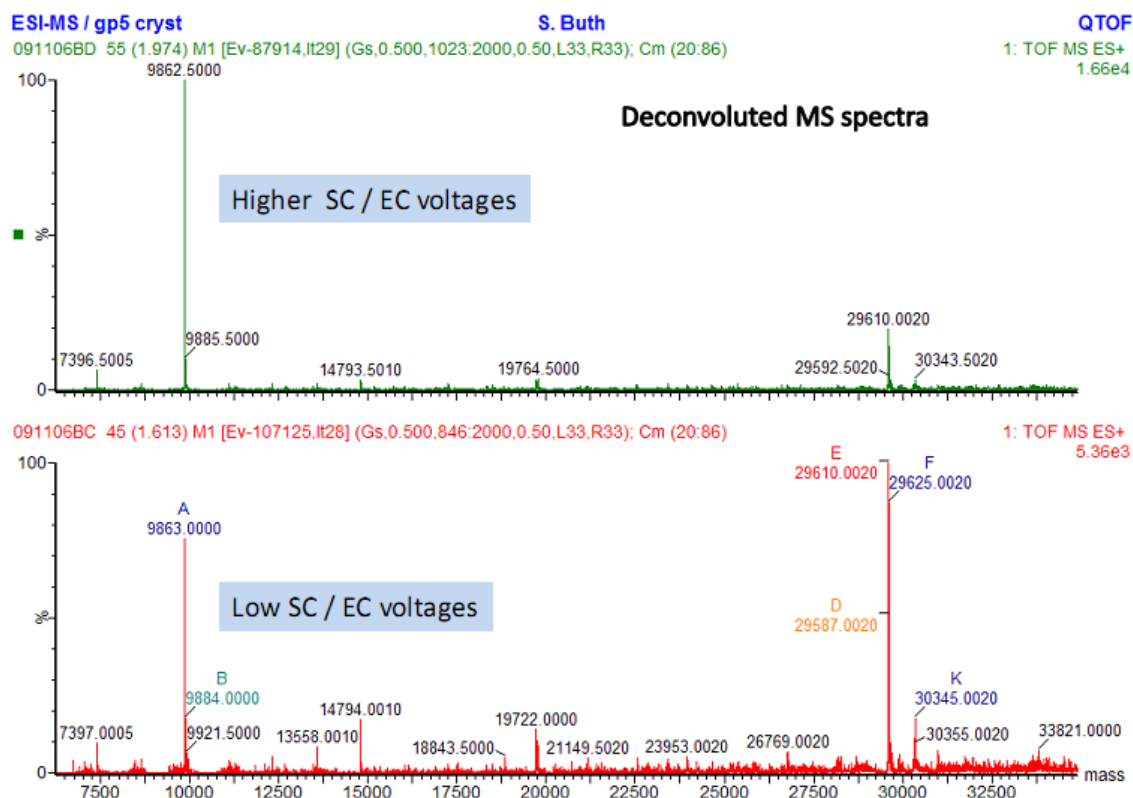
Figure 3.12 Mass spectra of gp5R483 trimer dissociated in-source.
The sample cone voltage (SC) and collision energy (EC) were used to promote gp5R483 trimer dissociation.
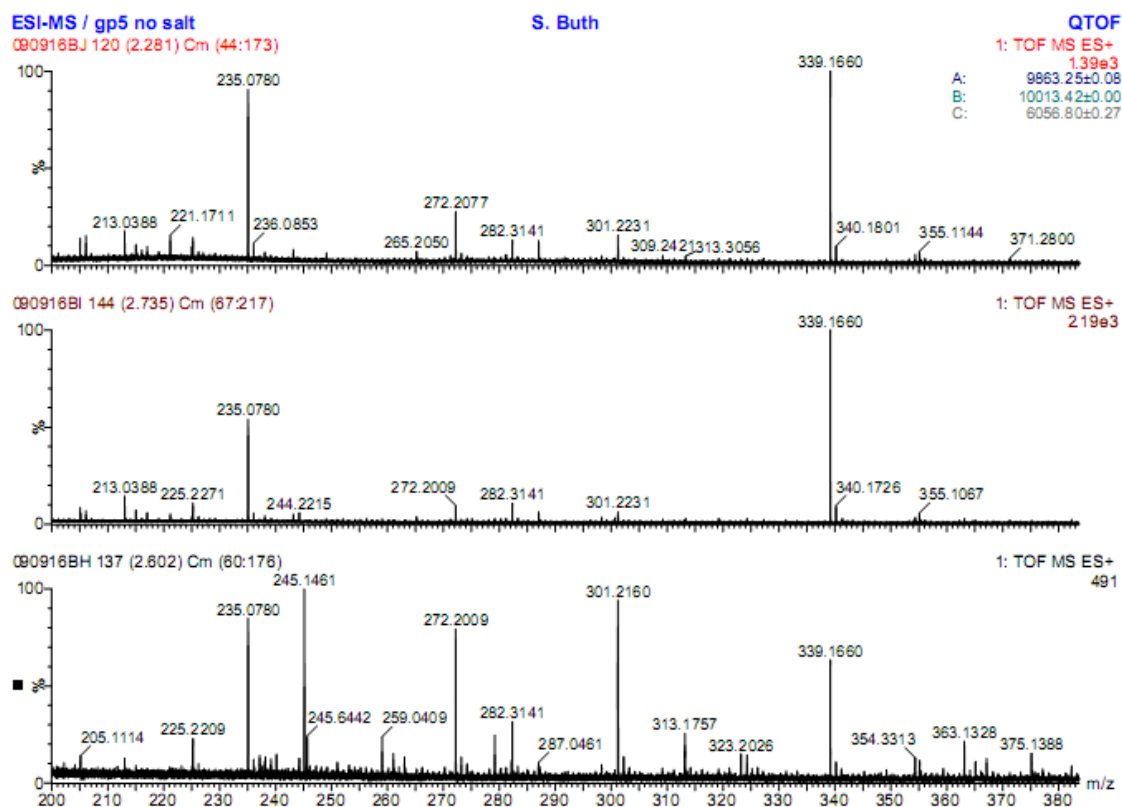
Figure 3.13 Gp5R483 mass spectra low MW zone analysis.
The sample applied cone voltage (SC) and collision energy (EC) were step increased from top to bottom: 1) SC = 35 V, EC = 4 V, 2) SC = 50 V, EC = 10 V, 3) SC = 100 V, EC = 15V.
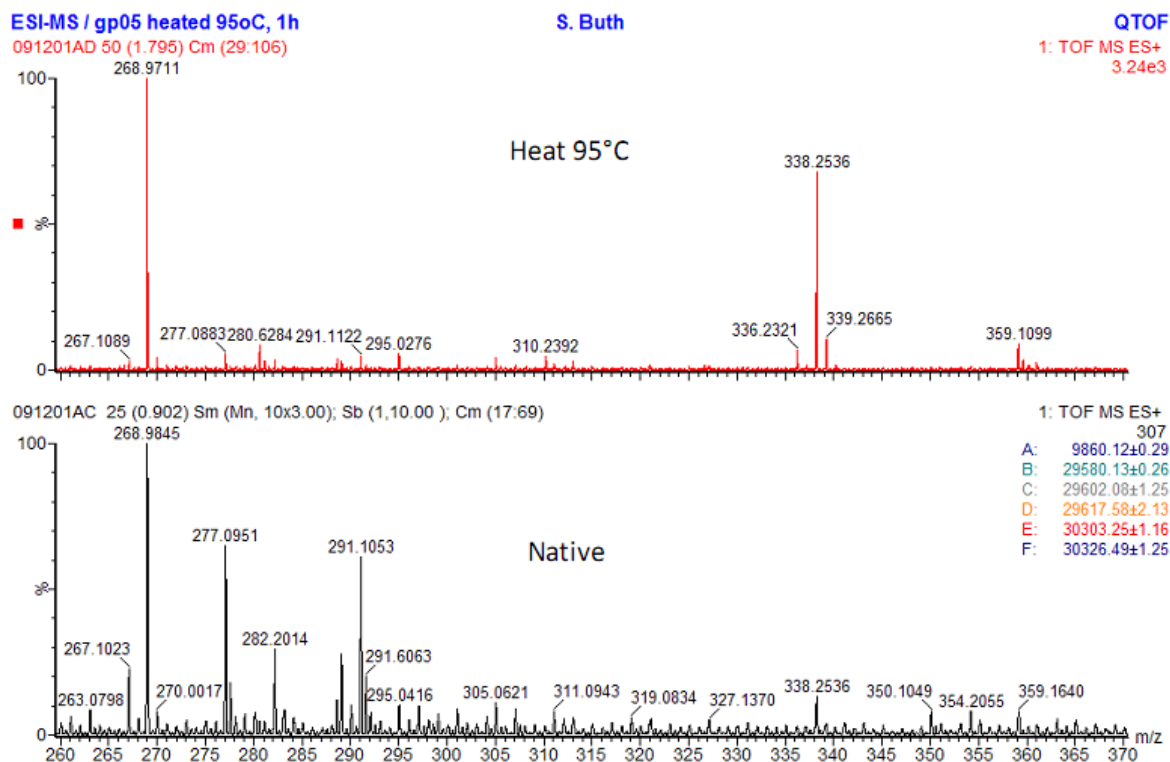
Figure 3.14 Mass spectra of gp5R483 denatured by heating.
Sample was prepared from the protein purified by crystallization. Low MW zone of the spectra obtained by heat dissociation of the sample.
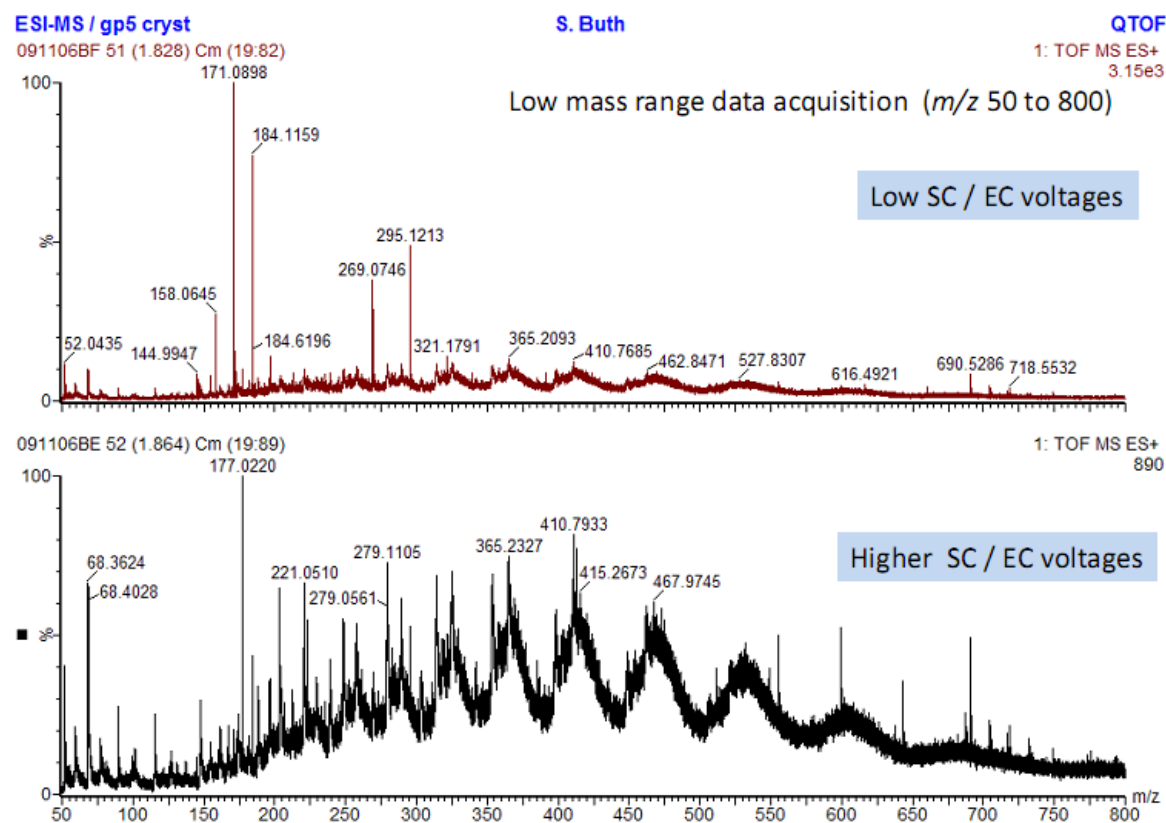
Figure 3.15 Mass spectra low MW zone analysis of gp5R483 dissociated in-source.
Sample was prepared from the protein purified by crystallization. Low MW zone of the spectra obtained by increasing of the SC/EC voltages applied to the sample.
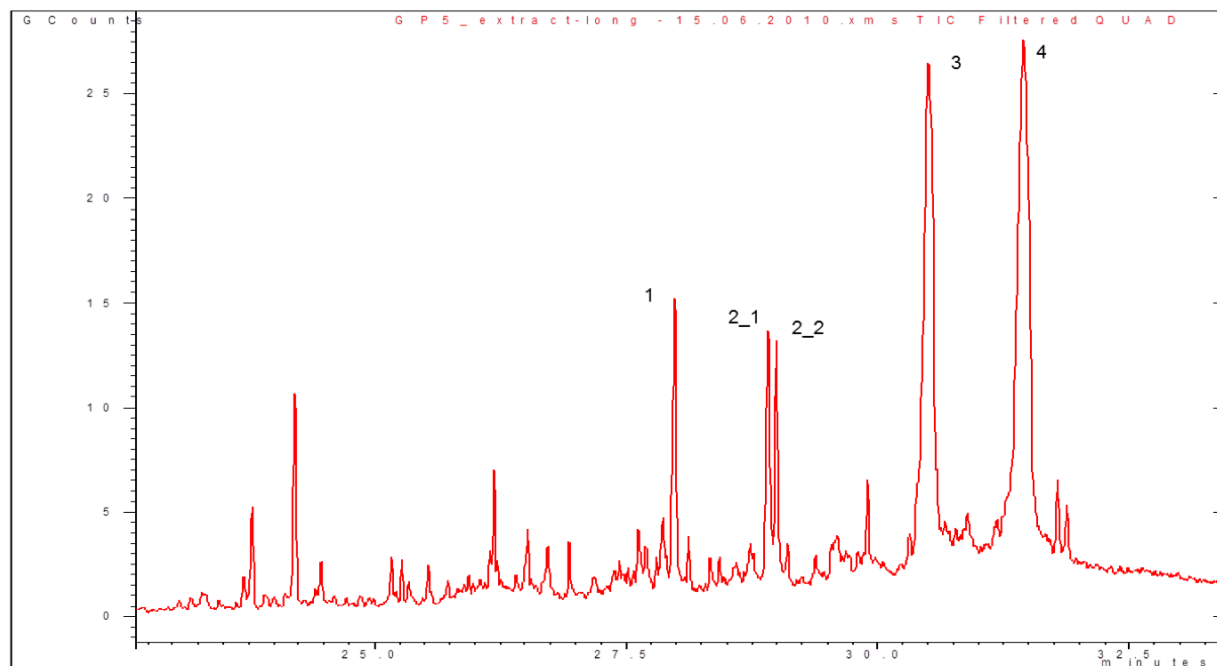
Figure 3.16 Gas chromatography of gp5R483 organic extract.
Peak 1: 16:0 Hexadecanoic acid (palmitic acid); probability = 72.4%.
Peak 2_1: 18:1 Octadecenoic acid (oleic acid); probability (sum) = 32.64%.
Peak 2_2: 18:0 Octadecanoic acid (stearic acid); probability = 70.0%.
Peak 3: Esters of the hexadecanoic acid; probability (sum, main ones) > 90.0%.
Peak 4: Esters of the octadecenoic/octadecenoic acid; probability (sum, main ones) > 50.0%.

3.2   Structure of gp5G484-gp5.4 complex

3.2.1   Introduction

Various phages with contractile tails solve the cell membrane piercing problem differently. P2 and other P2-like phages carry out cell membrane piercing using a sharp, spike-like trimeric protein, which is armored with an Fe ion coordinated by six histidine residues (double histidine motif (2xHis) per each chain of the trimer) (Browning, et al. 2012). For other phages, including T4 phage, one of the best studied bacteriophages, cell-puncturing proteins/protein complexes are known as well, but they do not have spike-like shape of the tip (Kanamaru, et al. 2002).

Analyzing the cryoEM reconstruction of the T4 phage baseplate (Kostyuchenko, et al. 2003) one observes that below T4 cell-puncturing device (expanding from the C terminus of the gp5 β-helix – the outward facing domain) is an extra density that might accommodate another protein or protein complex (Figure 3.17). Also, the analysis of the high resolution crystal structure of the C-terminal gp5 β-helix fragment (see **3.1.3.1**) shows that the distal end of the T4 cell-puncturing device ends with the blunt end and possesses a surface that can accommodate this protein/protein complex. This structure also revealed a high level of disorder of the residues at the gp5 C terminus (up to the model main chain split for alternative conformations), which suggests that cell wall piercing is unlikely to be carried out by only the gp5-gp27 complex, but requires another protein to stabilize and sharpen its tip. This critical piece of information regarding this and other similar spikes has been missing for years.

Analysis of genomes of a large number of bacteriophages (even evolutionary distant) had revealed a group of proteins that were mainly found downstream of the gp5 protein (T4 genome terminology) and seemed to be genetically linked. Considering a large family of T-even phages, ORFs homologous to the gp5.4 protein from T4 were always found near gp5-like proteins. This allowed hypothesizing that gp5.4 caps the gp5-gp27 complex. Gp5.4 is a 97 amino acid long protein (GenBank ID AAD42485.1).

a)   **b)**   **c)**
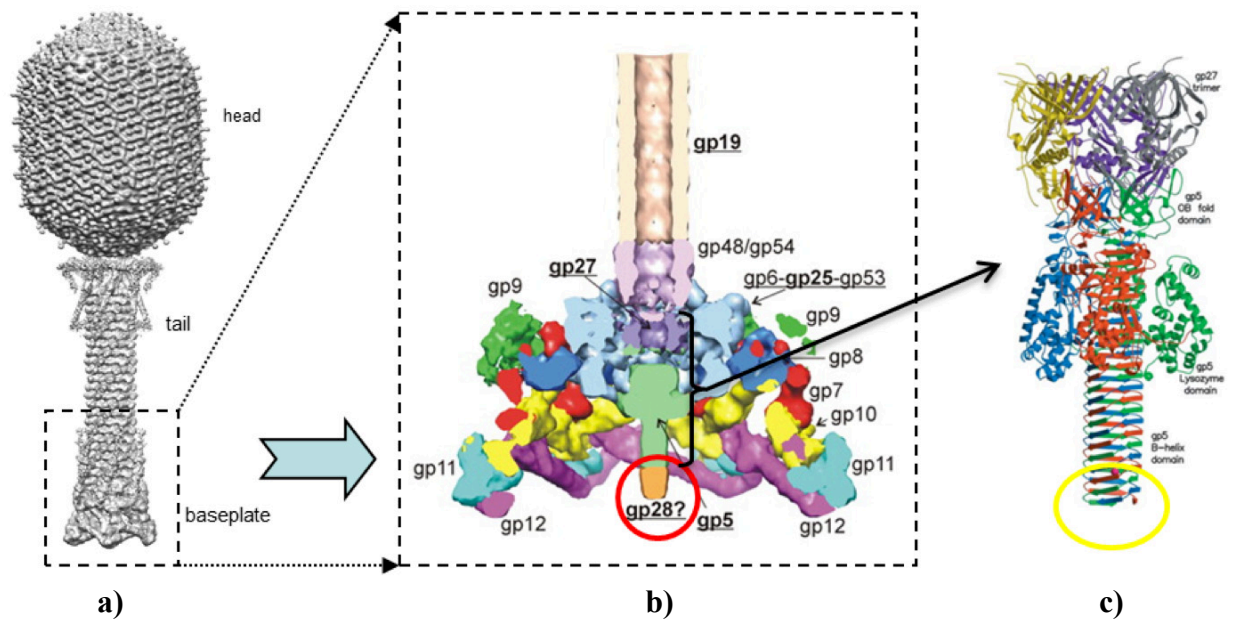
Figure 3.17 T4 cell-puncturing device caping protein placement.
a) general view of the T4 phage; **b)** baseplate cut-away view; in red is circled an unannotated density; **c)** gp5-gp27 complex a.k.a the T4 cell-puncturing device; in yellow is circled its blunt end. Figure adapted from (Kostyuchenko, et al. 2003) and (Kanamaru, et al. 2002).

3.2.2 Materials and Methods

3.2.2.1 Construct engineering

The knowledge about structure and properties of the C-terminal deletion mutant of gp5 β-helix containing residues R483-G575 was of great help in working with gp5.4. Gp5R483 behaved and crystallized well and thus it was chosen as a co-expression partner for gp5.4, as its carrier. Also the gp5R483 structure could be used as a search model for molecular replacement. The gp5 β-helix fragment was modified by removing R483 to avoid having a bulky residue at the N terminus as it was disordered in the 1.3 Å structure and could potentially have interfered during crystallization of the gp5–gp5.4 complex. This gave rise to a new mutant, gp5G484. It was cloned together with gp5.4 into a pEEva2 plasmid (a derivative of pET23a). Gp5G484 was genetically modified to contain a histidine-rich (His6, His-tag) extension via TEV-cleavage site (ENLYFQG) and SGS linker at the N-terminus. Dr. Mikhail Shneider did construct engineering and cloning.

3.2.2.2 Protein expression and purification

The pEEva2 plasmid carrying gp5-gp5.4 DNA complex was transformed into the BL21 (DE3) strain of *E. coli*. The transformed cells were grown at 37 °C in the LB medium, complemented with ampicillin at the concentration of 200 μg/ml until the optical density reached the value of 0.6-0.8 at 600 nm. The medium was cooled on ice to the temperature of 18-20 °C followed by induction the gp5-gp5.4 complex expression by addition of IPTG to a final concentration of 1 mM. After further incubation at 18 °C overnight (approximately 16 hours), the cells were harvested by centrifugation at 5180 g, 4 °C. The cell pellet was resuspended in $1/50^{th}$ of the original cell volume in a 20 mM Tris-HCl pH 8.0 buffer complemented with 300 mM NaCl, 5mM imidazole (Im) and 0.02% NaN_3. The cells were lysed by sonication. The cell debris was removed by centrifugation at 35000 g for 15 minutes at 4 °C. The supernatant was loaded onto the Ni^{2+}-precharged 5ml GE HisTrap FF Crude column (GE Healthcare Life Sciences), which was equilibrated with 20 mM Tris-HCl pH 8.0, 300 mM NaCl. Protein was eluted with 20 mM Tris-HCl pH 8.0, 300 mM NaCl, 250mM Im buffer using two-step gradients on an AKTApurifier 100 system (GE Healthcare Life Sciences): 1) 15% (37.5 mM) Im – to remove nonspecific bounded proteins and 2) 100% (250 mM) Im – actual protein elution. Upon completion of the chromatography run, the fractions of the elution peak were pulled together and dialyzed overnight with simultaneous TEV His-tag cleavage against 10 mM Tris-HCl pH 8.0, complemented with 1ml of TEV-protease (at the concentration of 1mg/ml), 3 mM DTT, 1.5 mM

EDTA for His-tag removal. Digested protein was further purified with ion-exchange chromatography (GE Mono Q 10/100 GL column connected to an AKTApurifier 100 system) in 20 mM Tris-HCl pH 8.0 buffer using 0 to 1 M NaCl linear gradient Figure 3.18 **a)**. Selected fractions of the ion-exchange chromatography were analyzed on SDS-PAGE gel Figure 3.18 **b)**. **Components 1** and **2** (Figure 3.18 **b)**) were further purified by size exclusion chromatography using a GE HiLoad 16/60 Superdex 200 PG (GE Healthcare Life Sciences) column connected to the AKTApurifier 100 system (GE Healthcare Life Sciences). A 10 mM Tris-HCl pH 8.0, 150 mM NaCl buffer was used.
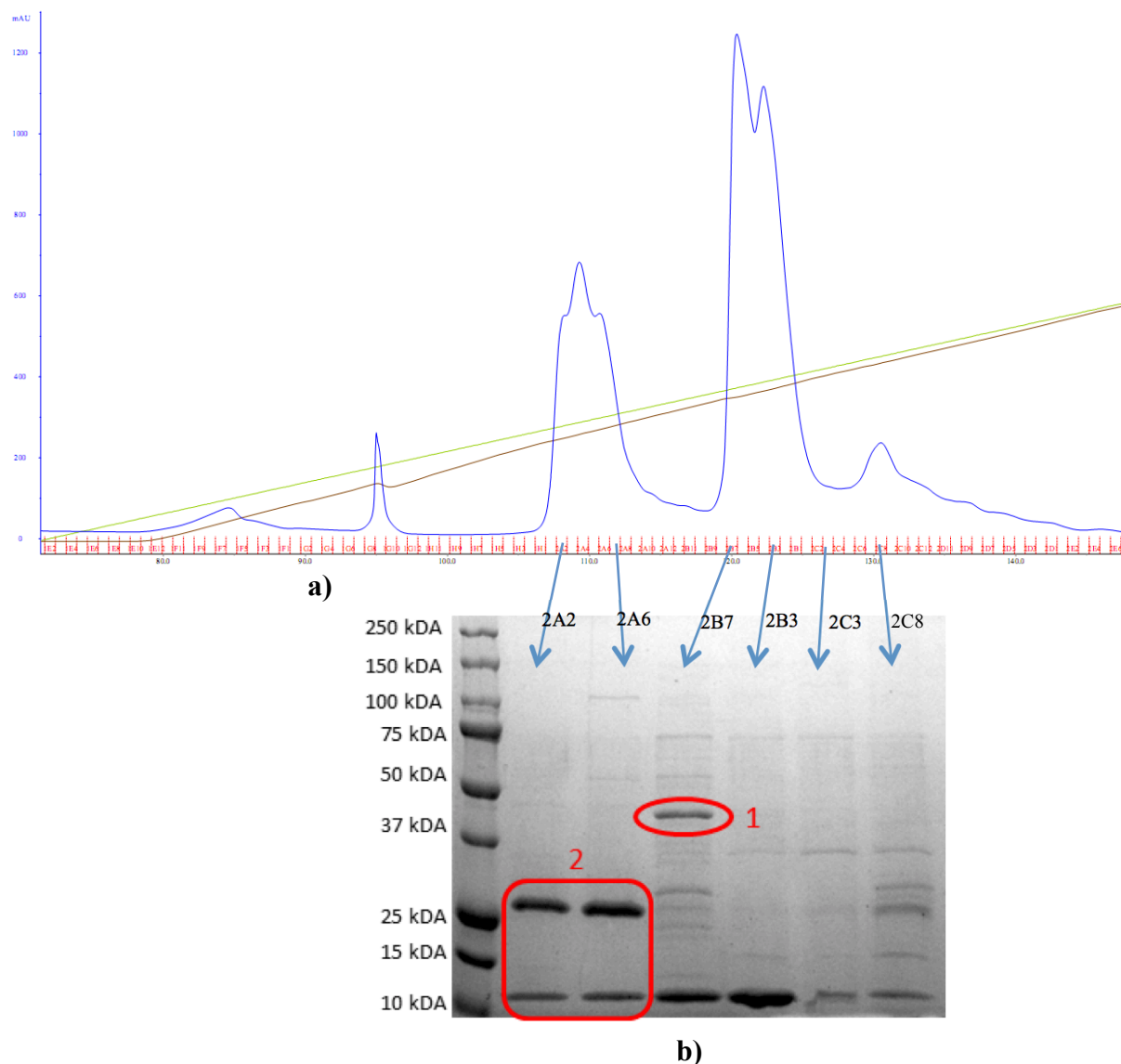
Figure 3.18 Purification and fraction analysis of the gp5-gp5.4 complex.
a) ion-exchange purification of the complex; **b)** purification analysis with SDS-PAGE. In red are frames bands, corresponding to the **Components 1** and **2**.

3.2.2.3 Crystallization, data collection and structure determination

3.2.2.3.1 **Component 1**

Judging from the unannotated EM density at the very tip of gp5, the MW of gp5.4 was expected to be around 30 kDa (3x10kDa) and have a structure somewhat similar to gp5 – an intertwined triple β-helix or coiled coil or gpV/gp138 apex tip-like structure (Browning, et al. 2012). Gp5G484-gp5.4 complex was expected to run in the range of 58-60 kDa on the SDS-PAGE gel (29kDa – trimer of gp5G484 together with 30 kDa for gp5.4 trimer). The highest MW band, present on the gel corresponds to 45-46 kDa Figure 3.18 **b)**, **Component 1**. Previously it was observed that the tight trimer of gp5R483 runs on a SDS gel in a lower MW range then was expected (data not shown). Considering this fact it was believed that the observed band of 45-46 kDa may correspond to the complex of interest and proceeded with crystallization.

The protein complex was brought to the concentration of 22 mg/ml in 10 mM Tris-HCl pH 8.0, 150 mM NaCl buffer. Initial crystallization screening was carried out employing the method of sitting drop in 96 well SWISSSCI MRC 2 plates using Jena Bioscience crystallization screens. Optimization of initial hits from the successful crystallization conditions was carried out in 24 well plates employing the method of hanging drop vapor diffusion. Crystals were obtained by mixing 1.25 μl of purified protein complex with 1.25 μl of reservoir solution and allowed to equilibrate against 500μl of 22% PEG 4000, 200 mM Li_2SO_4, 100 mM Tris-HCl pH 8.5 at 18 °C. Prism-like crystals appeared in about 5 days and continued to grow for another week reaching dimensions of 0.2 mm x 0.15 mm x 0.1 mm (Figure 3.20 **a)**). Crystals appeared to be exceptionally sensitive to the temperatures above 20 °C as they did not survive a 3 hours synchrotron trip and handling at the beamline, thus dissolved while being handled. Crystals re-appeared within a week or two in the same drops, were flash-frozen in the liquid nitrogen and transported frozen to the beamline. They belong to the $P2_1$, #4 space group with a = 110.29, b = 73.79, c = 111.60 Å, β = 113.39° unit cell parameters. Data collection was carried out at the X06SA PXI Pilatus beam line of the Swiss Light Source (SLS) at the Paul Scherrer Institute (SLS, Villigen, Switzerland) at the wavelength of 1 Å. Best crystals diffracted to better than 2 Å resolution. The diffraction data was indexed, integrated and scaled with XDS (Kabsch 2010), details are summarized in Table 3.5. The Matthews coefficient (Matthews 1968) at 2.40 $Å^3$/Da suggested with total probability of 0.86 that the asymmetric unit accommodates 3 molecules of gp5-gp5.4 complex. The structure of the **Component 1** was determined by molecular replacement using the program PHASER (McCoy, et al. 2007), with the 1.3 Å structure of gp5R483 deletion mutant as a search model. Structure solution revealed that the crystallized

Component 1 is a dimer of two solo gp5G484 trimers (totaling of 58 kDa). The model was built manually with Coot (Emsley and Cowtan 2004) and refined with REFMAC5 (Winn, et al. 2011) using NCS.

3.2.2.3.2 Component 2

Despite the negative outcome with **Component 1** there was another peak from the ion-exchange purification step. A mass spectrometry analysis on the material from this peak revealed a mass, precisely corresponding to the MW of gp5.4 without first Met residue (Figure 3.19). There was no evidence of gp5G484. Nevertheless gp5-gp5.4 complex formation was not confirmed by mass spectrometry (or by any other means) purification of the **Component 2** continued, followed by crystallization.
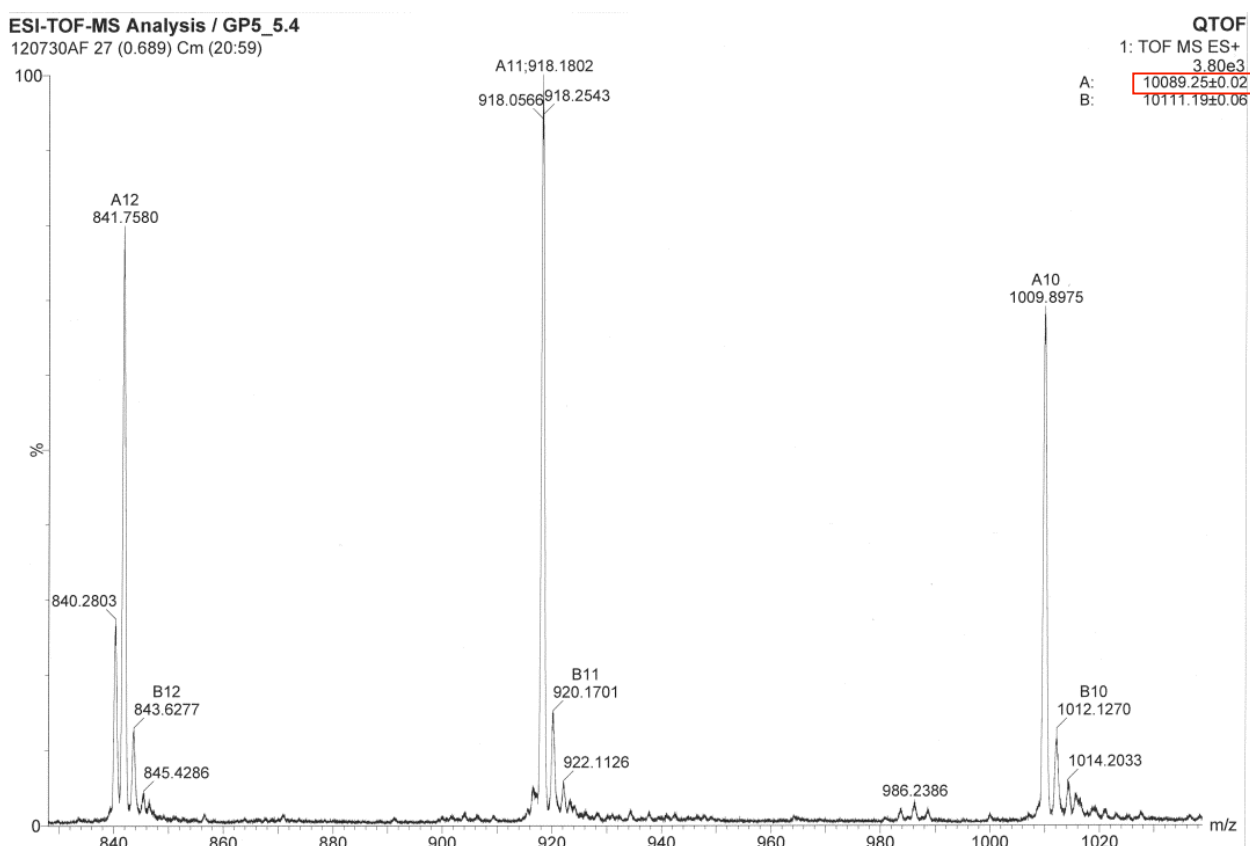
Figure 3.19 ESI-TOF mass spectrometry analysis of the gp5-gp5.4 protein complex.
Deconvoluted MW of the family of peaks A (framed in red) corresponds to the mass of the gp5.4 protein without first methionine residue.

The protein complex was brought to a concentration of 20 mg/ml in 10 mM Tris-HCl pH 8.0, 150 mM NaCl buffer. Initial crystallization screening was carried out employing the method of sitting drop in 96 well SWISSSCI MRC 2 plates using Jena Bioscience crystallization screens. Optimization of crystallization conditions was carried out in 24 well plates employing the method of hanging drop vapor diffusion. Crystals were obtained by mixing 1.25 μl of purified

protein complex with 1.25 μl of reservoir solution and allowed to equilibrate against 500μl of 26% PEG 2000, 80 mM MgCl, 100 mM Tris-HCl pH 8.5 at 18 °C. Crystals appeared in about 3 days quickly reaching their maximum dimensions of 0.6 mm x 0.1 mm x 0.05 mm forming quartz-like clusters Figure 3.20 **b)**. For data collection, the crystals were dipped for 20-30 seconds into the cryoprotectant solution containing 25% v/v of ethylene glycol in additional to the crystallization solution components and flash frozen in a vaporized nitrogen stream at 100 K. Crystals belonged to P2$_1$, #4 space group with a = 46.30, b = 49.33, c = 84.06 Å, β = 96.19^o unit cell parameters. Data collection and fluorescence scan were carried out at the X06SA PXI Pilatus beam line of the Swiss Light Source (SLS) at the Paul Scherrer Institute (SLS, Villigen, Switzerland) at the wavelength of 1 Å. Best crystals diffracted to better than 1.2 Å resolution. The diffraction data was indexed, integrated and scaled with XDS (Kabsch 2010), details are summarized in Table 3.5. As oligomeric state of the gp5.4 was unknown the Matthews coefficient was not of a great use. The structure of the **Component 2** was determined by molecular replacement using the program PHASER (McCoy, et al. 2007). The 1.3 Å structure of the gp5R483 deletion mutant was used as a search model. Based on the unit cell parameters was made an assumption that the asymmetric unit accommodates 1 molecule of gp5-gp5.4 complex, thus a search request was specified to find one copy of the gp5G484 trimer. This assumption was further confirmed by the Matthews coefficient of 2.40 Å^3/Da with total probability of 1.0. Structure solution revealed that crystallized **Component 2** is the complex of interest and consists of a gp5G484 trimer and a monomer of gp5.4 (totaling of 40.4 kDa). The model was built manually with Coot (Emsley and Cowtan 2004) and refined with SHELXL (Sheldrick 2008).

| a) | b) |

Figure 3.20 Crystals of Components 1 and 2 resulted from gp5-gp5.4 complex purification.
a) Component 1 (gp5G484 dimer) and **b) Component 2** (gp5G484-gp5.4 complex) crystals (the scale bar length is 200μm).

Table 3.5 X-ray data collection and refinement statistics of Component 1 (gp5G484 dimer) and Component 2 (gp5G484-gp5.4 complex).
Data in parenthesis represent statistics for the highest resolution shell.

Data collection	Component 1	Component 2
Wavelength	1.0 Å	1.0 Å
Number of frames	990	1440
Frame width (°)	0.25	0.25
Space group	P1 2(1) 1 (#4)	P1 2(1) 1 (#4)
Unit cell parameters:		
a, b, c (Å)	110.29, 73.79, 111.60	46.30, 49.33, 84.06
α, β, γ (°)	90, 113.39, 90	90, 96.19, 90
Biologic assembly	Dimer of trimers	Tetramer (trimer + monomer)
Number of biological assemblies per asymmetric unit	3	1
Resolution (Å)	60.0 – 2.0	46.0 – 1.15
R_{meas}	0.095 (0.363)	0.057 (0.436)
$<I / \sigma_I>$	6.7 (2.4)	12.1 (2.3)
Completeness (%)	96.6 (95.9)	95.2 (81.3)
Redundancy	2.4 (2.5)	3.2 (2.5)

Refinement		
Number of reflections		
Working	110537	124490
Test	5818	4516
R_{work} / R_{free}	0.219/0.279	0.127/0.171
B-factor (Å^2)	35.3	22.9
R.m.s. deviations		
Bond lengths (Å)	0.019	0.013
Bond angles (°)	1.849	0.030
Number of atoms		
Protein	12217	2893
Solvent and ligands	992	638
Ramachandran plot (%)		
Most favored	98.9	95.2
Additionally allowed	1.1	4.5
Outliers	0.0	0.3

3.2.3 Results and discussion

3.2.3.1 Morphology overview and dimer formation peculiarities of the gp5G484 dimer

The structure of the gp5 β-helix was initially analyzed by (Kanamaru, et al. 2002) and its C-terminal fragment, gp5R483, was analyzed in greater detail in paragraph **3.1** of this manuscript. This paragraph will focus on the analysis of the surface contacts of two gp5G484 trimers, forming a dimer of trimers. Similar oligomer formation was already observed for gp5R483 – it formed pseudodimers via Sr^{2+} sites (see **3.1.3.2**).

In the crystallographic unit cell three dimers of trimers are present, totaling 6 gp5G484 molecules. Dimerization occurs via the C termini of the trimers. Four out of six trimers have same fold of the gp5 C terminus as previously reported by (Kanamaru, et al. 2002) and in paragraph **3.1.** Two of these "usual" gp5 fragments form one type of dimers of trimers (Figure 3.21 **a)**). Two other "usual" β-helix fragments form two pairs of dimers with gp5G484 trimers, whose last β-strands (residues 571-575) fold backwards, into antiparallel manner, stacking on top of the preceding β-strand (Figure 3.21 **b)**). This fold became available due to reorientation of the R571, which has become a part of the tight turn along with conventional S570. For both forms of dimer of trimers the side chain of the R571, which was oriented towards the outside of the protein molecule and disordered in free gp5R483, is oriented towards the negatively charged site of the opposite gp5G484 molecule, formed by G569, D568 and G575. In this configuration the R571 guanidinium group is fixed by H-bonds formation with backbone O from G569, $O^{\varepsilon 1}$ from D568 and G575 OXT. R571 side chain is well ordered and can be easily modeled (Figure 3.22). No other amino acid side chain pairs, forming H-bonds were detected. However, in the close contact area present an H-bond network between the opposite dimer main chains, typical to β-sheet structures.

a)   **b)**
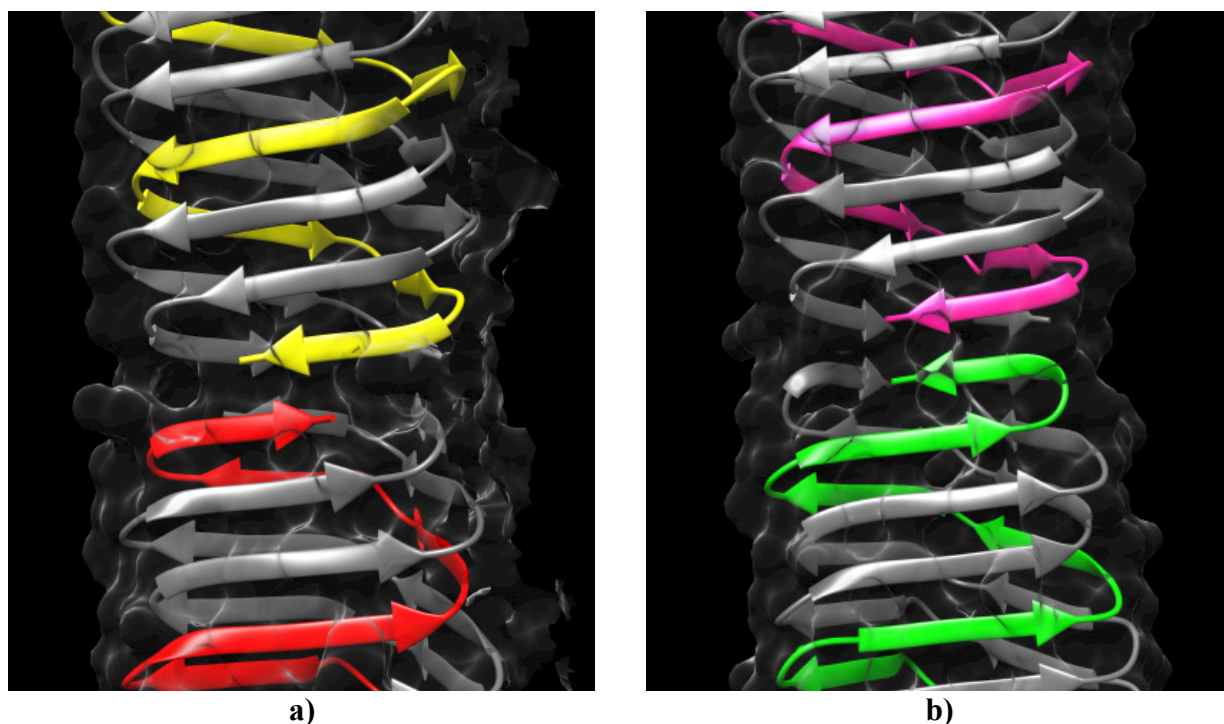
Figure 3.21 Two forms of the gp5G484 dimer of trimers.
Semi-transparent molecular surface with ribbon diagram representation of the polypeptide chains for each type of dimers. For clarity a single pair of polypeptide chains forming intertrimer contacts are colored in yellow and red for **a)** and magenta and acid green for **b)**.
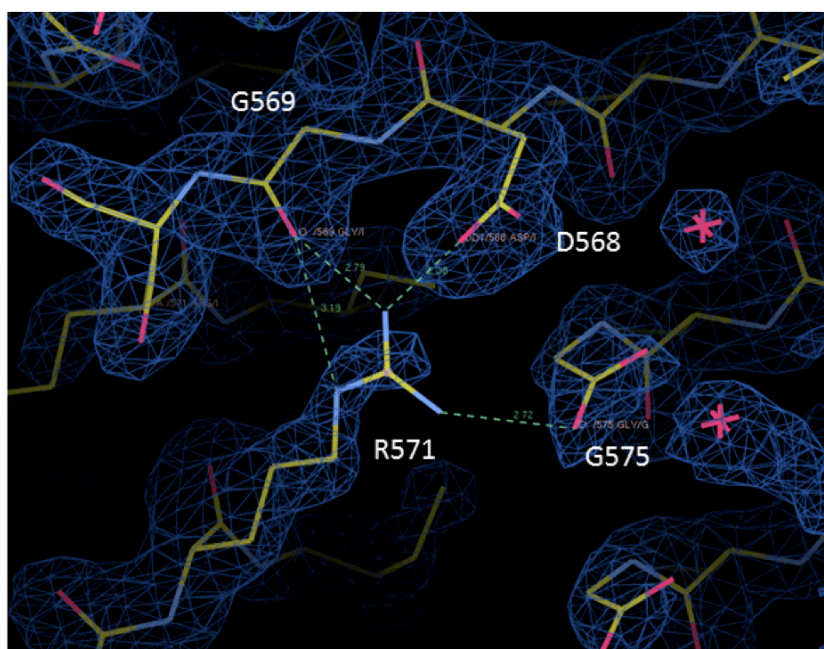
Figure 3.22 "New" orientation of the R571 in gp5G484 dimer of trimers.
H-bond network stabilizing R571 side chain is showed with dash line. Residues, forming H-bond network are numbered. 2Fo-Fc map is contoured at 1.5σ.

3.2.3.2 Morphology overview of the gp5G484-gp5.4 structure

The structure of the gp5G484-gp5.4 complex was determined by molecular replacement and refined to 1.15 Å resolution. Three polypeptide chains of gp5 fragment constitute an intertwined triple β-helix, however remarkably gp5.4 turned out to be a monomer that is folded into a triangular pyramid that decorates the blunt end of the gp5 β-helix. The overall shape of the gp5R484-gp5.4 complex reminds of a sharpened pencil of 102 Å long and 28 Å wide. The gp5G484 forms the "pencil's" body and gp5.4 – its sharpened tip.

Spike proteins from P2 and φ92 consist of 3 domains: OB-fold, β-helix and apex domains (Figure 3.23 **b), c)**) whereas the T4 cell-puncturing device has additional two domains: gp27 and gp5 lysozyme domain (Figure 3.1 **c)**). Gp27 is situated above an OB-fold domain that crowns full-length β-helix, and suits as an interface linker between the hexameric baseplate and trimeric gp5 (Kostyuchenko, et al. 2003). Gp5 lysozyme domain spans along the faces of β-helix and meant to digest bacterial peptidoglycan. The gp5G484-gp5.4 fragment of T4 injection machinery comprised only from a β-helical domain and an apex. The structure was deposited in the Protein Data Bank under accession number 4KU0.

3.2.3.3 β-helix comparison in gp5, gpV and gp138

The full-length T4 β-helix is much longer compared to those of P2 and φ92: 110 Å against 60 and 41 Å in gpV and gp138 respectively; the gp5G484 fragment is 52 Å long. Each face of the gp5 β-prism starts with a 5-strended antiparallel β-sheet. Then the structure becomes fully intertwined with each chain of the trimer wrapped between the other two (Figure 3.24 **a)**). The gpV completely follows the trend of gp5 – starts with an antiparallel 3-stranded β-sheet followed by 12-stranded intertwined structure (Figure 3.24 **b)**) whereas in gp138 β-helix is composed of three antiparallel β-sheets, two of which lay along the same face of imaginary trigonal prism, and the third one, C-terminal β-sheet, is clockwise shifted around the threefold axis of the protein (Figure 3.24 **c)**).
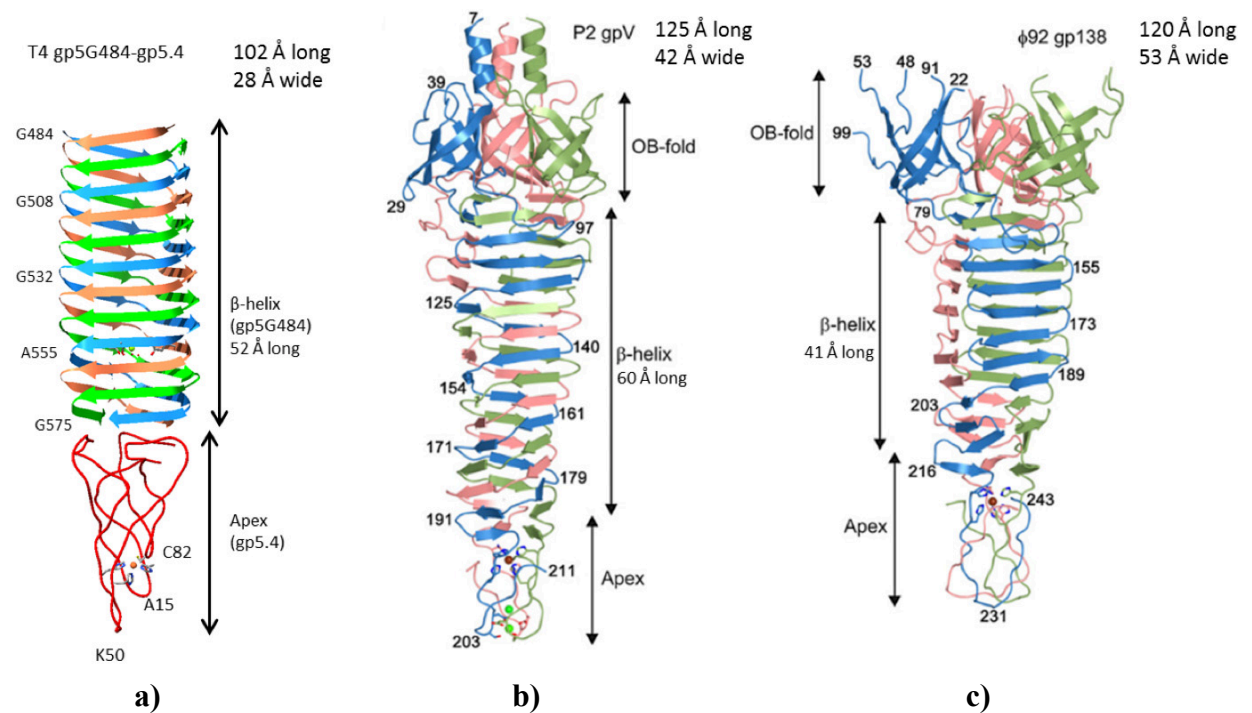
Figure 3.23 Morphology overview and comparison of T4 gp5G484-gp5.4 a), P2 gpV b) and φ92 gp138 c).

Ribbon representation is used. Different polypeptide chains are colored in different colors.
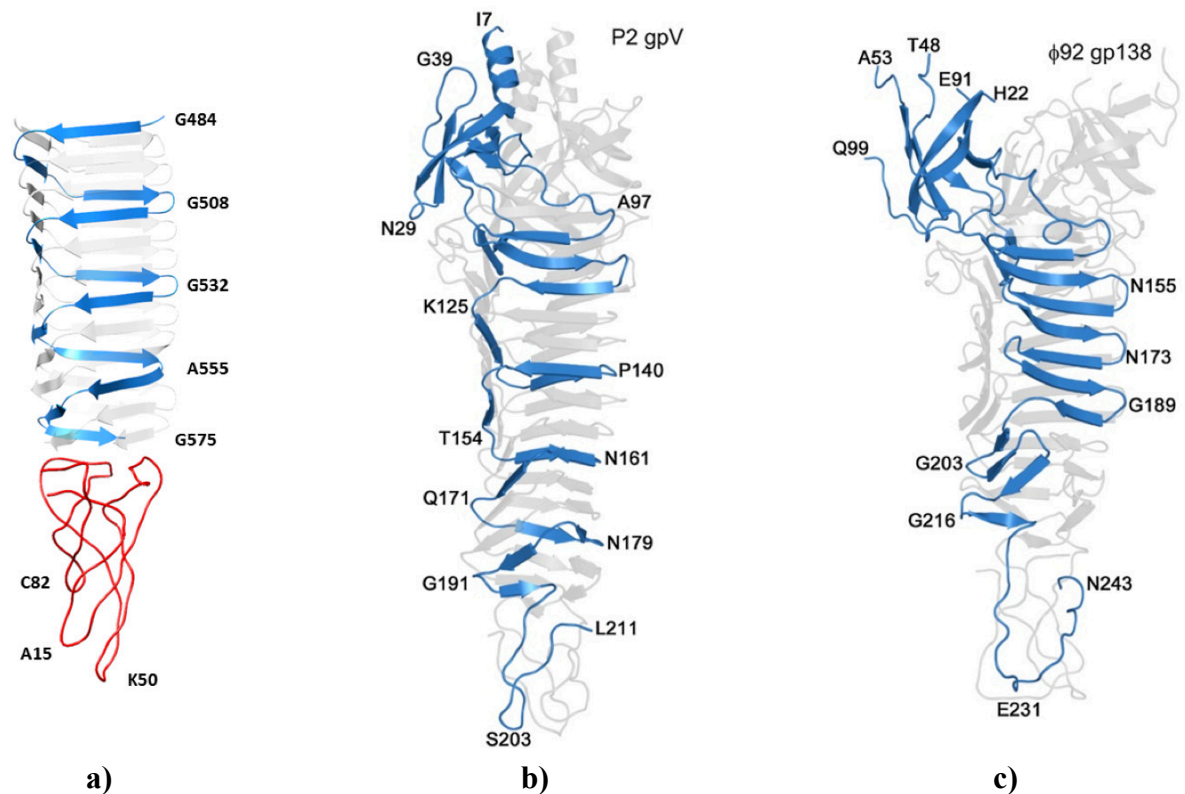
Figure 3.24 Topology of the β-helix in a) T4 gp5G484-gp5.4 complex b) P2 gpV and c) φ92 gp138.

Ribbon representation is used. For clarity single polypeptide chain comprising β-helix is colored in sky blue. Gp5.4 is colored in red.

3.2.3.4   Analysis of the gp5.4 protein structure and its comparison to the apex domains in gpV and gp138

Gp5.4 protein is a monomeric protein. It belongs to the PAAR (Proline-Alanine-Alanine-aRginine) repeat superfamily (the superfamily will be characterized in details in **3.3.3**) and form a sharp conical extension of the gp5-gp27 protein complex, aka T4 cell puncturing device. Its polypeptide chain is folded into three loops of different length that slightly wrap around each other (Figure 3.26 **a)**). The first loop is composed of residues 2-29 and has an intermediate length, the second loop is the longest – residues 30-66 and the third loop is the shortest – residues 67-96 (Figure 3.26 **b)**). The interloop regions (residues 31-33 and 67-69) together with 4 terminal residues (94-97) are 3 β-strands that form the "base" – a surface where gp5.4 attaches to gp5. Each β-strand of the "base" is parallel to the corresponding last β-strand from each chain of gp5G484. Close contact surfaces are composed of hydrophobic residues: Ile, Lys, Phe, Val, Arg. Figure 3.25 **a)** – a surface representation and coulombic surface coloring of gp5G484-gp5.4 complex perfectly demonstrates a hydrophobic patch that lies on the border between gp5G484 and gp5 and corresponds to the contact area between negatively charged gp5.4 and gp5G484. In the close contact area there are no pairs of charged residues, which would contribute to complex formation via side chain coulomb interactions. However, there are three peripheral plug – jackplug pairs (similar to the situation, described in paragraph **3.2.3.1** for gp5G484 dimer of trimers), resulting into 9 hydrogen bond formation (three per pair). These pairs are formed by backbone O from G569, $O^{\varepsilon 1}$ from D568 and OXT that belongs to G575 on one side and Lys residues (28, 66 and 93) on another. Also, there is an intermolecular main chain hydrogen bond network present, which is typical for β-sheet interaction. The network is composed of 11 H-bonds. This analysis allows us to conclude that hydrophobic interactions along with a network of H-bonds are the forces that are required for the gp5G484-gp5.4 complex to be formed and stabilized.

The apex domains in gpV and gp138 are trimeric and constituted of residues 191–211 and 216–245 respectively (20 and 29 residues per monomer). In both structures, the apex domain tapers from about 23 Å to 9 Å in diameter. Residues D203-S203-G204 comprise the sharp turn of the gpV apex loops. In gp138 these residues are G231-T232-G233-G234-S235. Nevertheless the gp138 tip is composed of 5 residues per chain against 3 in gpV both tips have about the same width. The reason is that the tip of gp138 contains a well-defined apex residue E231, which points along the axis of the trimer and form the actual extremity. The gp5.4 tapers from 27Å to 6Å against 9 Å in gpV and gp138 what makes it the sharpest amongst discussed spikes (Figure

3.25). This characteristic of the gp5.4 is achieved due to the morphology of the longest loop sharp turn and the fact, that the only this loop makes the protein extremity. The loop sharp turn is constituted only of 4 residues: K50-K51-P52-Y53.

a)   b)   c)
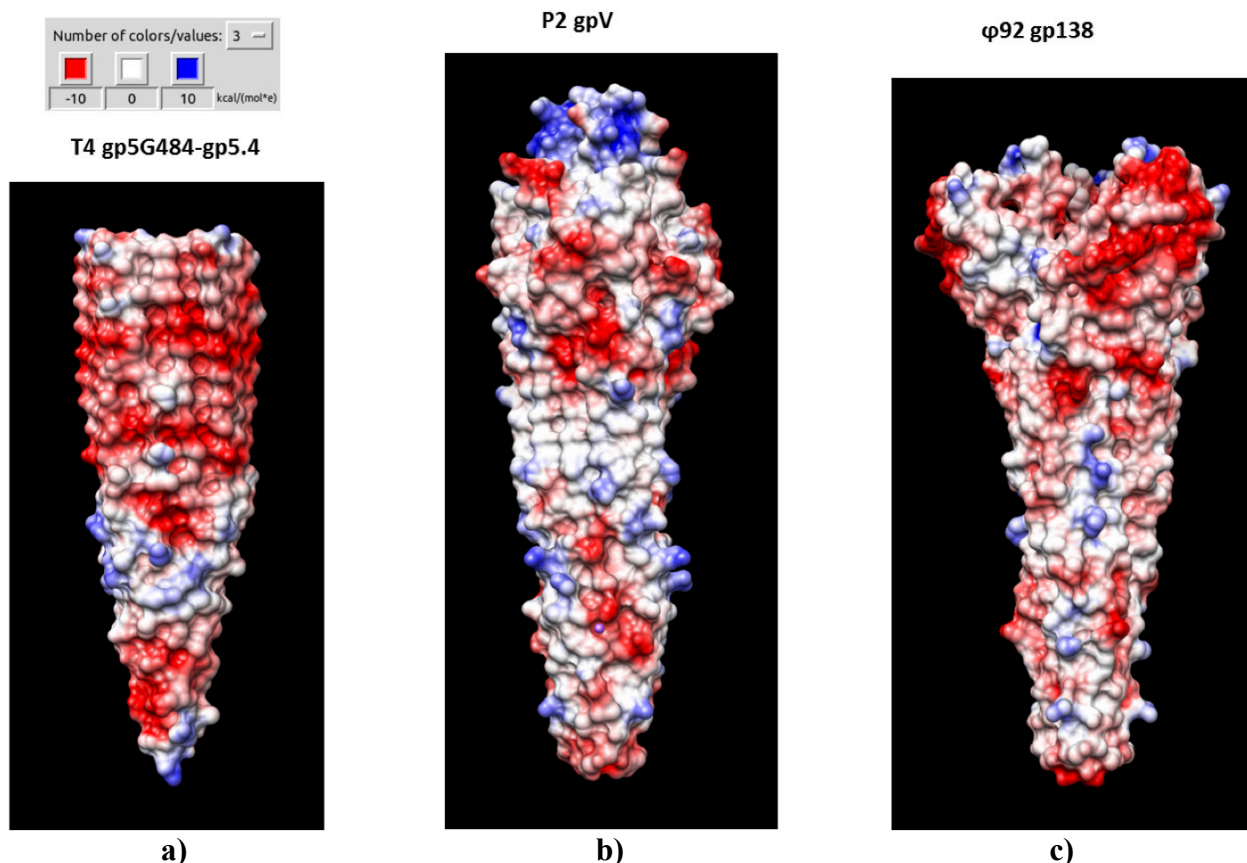
Figure 3.25 Surface representation and columbic surface coloring for gp5G484- gp5.4 complex, gpV and gp138.
Molecular surface is colored according to coulombic surface charge with blue, white and red corresponding to the most positive, neutral, and negatively charged patches, respectively.
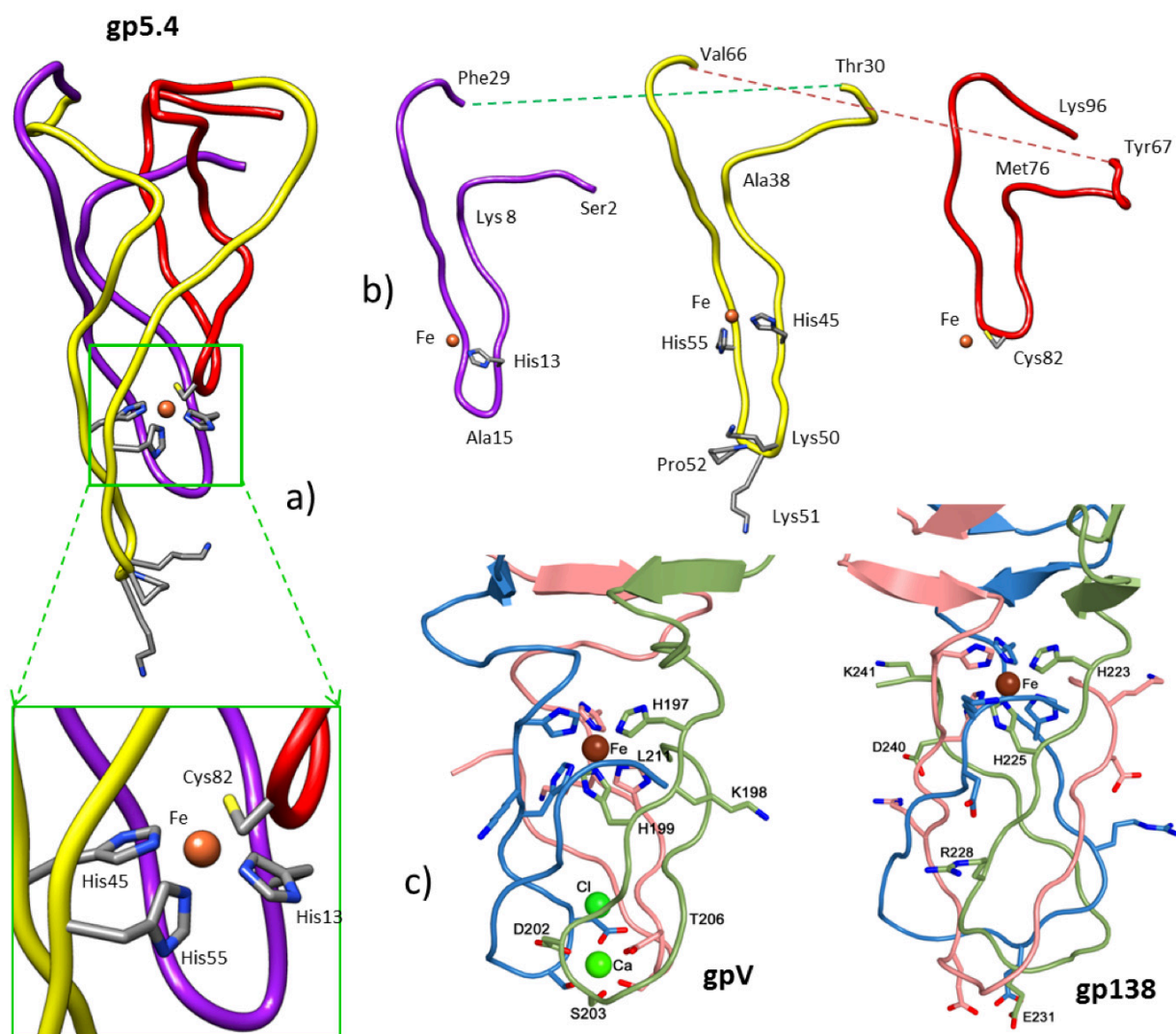
Figure 3.26 Detailed analysis of the gp5.4 architecture and its comparison to gpV and gp138 apex domains.

a) ribbon diagram of gp5.4 and zoomed region, containing Fe coordination site; b) analysis of the loops composing gp5.4; c) ribbon diagram of gpV and gp138 apex domains.

The most remarkable feature about all three structures is the organization of the apex interiors. There is an iron ion present in all three structures. In gpV and gp138 each of the three chains constituting the apex domain contains a double-histidine motif – two histidine residues that point into the structure interior with the spacing of one amino acid in between (H197 and H199, K198 is the spacer in gpV; H223 and H225 with T224 as a spacer in gp138). Three pairs of these histidines create an octahedral coordination site for the iron ion positioned on the three-fold axis of the protein (Figure 3.26 **c)**). In gp5.4 iron ion coordination site is tetrahedral and constituted of an ensemble of three His (H13, H45 and H55) and a single Cys (C82) residue (Figure 3.26 **a)**, **b)**). The average Fe-His ($N^{\varepsilon 2}$) bond distance is 2.01 ± 0.01 Å and Fe-Cys (S^{γ}) is 2.31 Å. These values agree with an average value of 2.16 ± 0.15 Å for Fe-His ($N^{\varepsilon 2}$) bond distance and 2.30 ± 0.05 Å for Fe-Cys (S^{γ}) (for the average bond length values calculation were chosen the structures

with coordination number CN = 4 and tetrahedral geometry (max bond angle deviation from the ideal in tetrahedron was 20°), resolution limit 2.2 Å; MESPEUS_10 was used for analysis (Hsin, et al. 2008)). Each loop of gp5.4 (like each chain in gpV and gp138) donates a residue or two for coordination site constitution thus making out of the iron ion a hardness element that adds extra rigidity to the protein during bacterial cell membrane piercing. A possible other function of the iron ion is to be a protein folding reference point – to keep the loops in register to each other during the folding. (Browning, et al. 2012) Also propose these functions for the iron ion in gpV and gp138.

Analyzing the gp5.4 protein family, it was established, that the only four amino acids, three histidines and a cysteine, are absolutely conserved in all gp5.4 orthologs (Figure 3.32 **a)**). The same applies for the family of proteins, where belong P2 gpV and ϕ92 gp138. Alignment of their amino acid sequences to the protein database shows strong conservation of the double-histidine motif – the residues that form metal coordination sites, but low sequence identity (Browning, et al. 2012). Considering the central spike structures of phages P2 and phi92, T4 and T6SS VgrG, as well as extensive bioinformatical study (Leiman and Shneider, Contractile tail machines of bacteriophages 2012) and (Browning, et al. 2012) it becomes evident that the architecture of the central spike proteins in all contractile systems is conserved. Nevertheless central spikes of phages, infecting the same host, are very diverse in terms of their sequences, they show a very high thermal stability even in the presence of denaturing agents (Leiman Lab, unpublished data, (Browning, et al. 2012)), arguing that they interact with the host cell membrane non-specifically and most likely act as membrane-piercing drills.

In the **Figures 3.20**, **3.21**, **3.22**, **3.23** parts **b)** and **c)**, depicting gpV and gp138, were adapted from (Browning, et al. 2012).

3.3 Type 6 secretion system (T6SS) warhead proteins

Most of the following subchapter was adapted from the paper (Shneider, et al. 2013).

3.3.1 Introduction

The bacterial type VI secretion system (T6SS) is a large multi-component macromolecular machine that plays an important role in the ecology of many Gram negative bacteria. T6SS is responsible for translocation of a wide range of toxic effector molecules allowing predatory cells to kill both prokaryotic as well as eukaryotic prey cells. The T6SS organelle is a member of contractile injection systems and functionally analogous to contractile tails of bacteriophages and is thought to attack cells by initially disrupting cell membrane with a trimeric protein complex called the VgrG spike (Pukatzki, et al. 2007), (Leiman, Basler, et al. 2009). Neither the exact protein composition of the T6SS organelle nor the mechanisms of effector selection and delivery are known. Here below is reported that proteins from the PAAR (Proline-Alanine-Alanine-aRginine) repeat superfamily form a sharp conical extension on the VgrG spike, which is further involved in attaching effector domains to the spike.

Careful examination of VgrG sequences showed that a β-structural repeat, which is presumed to be responsible for β-helix formation (Pukatzki, et al. 2007), either extends to the very C terminus of the protein or terminates with a glycine/serine-rich stretch. The X-ray crystal structure of one full-length VgrG trimer revealed the glycine/serine-rich stretch bends the polypeptide chain away from the β-helix without disturbing its tip, and that all VgrG β-helices have blunt ends resembling that of T4 gp5 β-helix (Leiman Lab, unpublished data). Due to solubility problems of most tested VgrG and PAAR proteins, was used a soluble fragment of T4 gp5 β-helix (residues 484-575), which is known to fold into a stable native-like trimeric structure, as a platform for creating β-helices mimicking the blunt end of various VgrG spikes. The last two β-strands at the tip of gp5 β-helix were replaced to mimic complementary VgrGs from *Escherichia coli* CFT073 and *Vibrio cholera* V52. These gp5-VgrG chimeras were then co-expressed with PAAR proteins.

3.3.2 Materials and methods

Dr. Mikhail Shneider designed the gp5G484-PAAR expression vectors and purified the proteins.

3.3.2.1 Construct engineering and cloning

The gp5 fragment containing residues 484-575 was cloned into the in-house designed expression vector pEEva2 (a derivative of the pET-23a plasmid (Novagen)). The T7-tag was replaced with a MGSSH$_6$SSG His-tag followed by a TEV protease cleavage site (ENLYFQG) and SGS linker. Upon expression and after TEV cleavage, the gp5 fragment contained four residues (GSGS) upstream of residue 484. The C-terminal residues of gp5 were modified by performing PCR with long primers containing the required mutations to mimic corresponding VgrG tips.

For expression in *cis*, PAAR genes were cloned downstream from the gp5 fragment. For expression in *trans*, PAAR genes were cloned into the pATE vector (chloramphenicol selection), a derivative of the pACYCDuet-1 expression plasmid (Novagen). The dual cloning site of the pACYCDuet-1 vector was replaced by the multiple cloning site from the pEEva2 vector.

3.3.2.2 Expression and purification of gp5-PAAR complexes

Gp5-PAAR complexes were expressed in E. coli B834 (DE3) cells grown in the 2xTY medium containing ampicillin at 100 μg/ml for the *cis* constructs and, additionally, chloramphenicol at 34 μg/ml for the *trans* constructs. The cultures with a total volume of 2 liters were incubated at 37° C with shaking at 200 rpm until the culture optical density reached 0.6 (at 600 nm wavelength). The culture was cooled down to 18 °C, and the protein expression was induced by an addition of IPTG to a final concentration of 1 mM. The expression continued overnight. Cells were harvested by centrifugation at 5180 g at 4 °C for 15 min. The cell pellet was resuspended in a lysis buffer that contained 20 mM Tris-HCl pH 8.0, 300 mM NaCl, 5 mM Imidazole. The cells were lysed by ultrasonication on ice. The lysate was centrifuged at 35000 g, 4 °C for 15 min. The supernatant was loaded onto a Ni^{2+} precharged column (5ml GE HisTrap FF Crude), connected to the AKTApurifier 100 system (GE Healthcare Life Sciences). The non-specifically bound material was removed by washing the column with a washing buffer (50 mM Tris-HCl pH 8.0, 300 mM NaCl, 20 mM Imidazole). The affinity bound material was eluted with an elution buffer (20mM Tris-HCl pH 8.0, 300 mM NaCl, 250 mM Imidazole). The fractions containing the target protein were pulled together and set up for the His-tag overnight digestion with TEV-protease against 10 mM Tris-HCl pH 8.0, 3 mM DTT and 1.5 mM EDTA buffer. The digested protein was further purified by ion-exchange chromatography performed with a GE

Mono Q 10/100 GL column connected to an AKTApurifier 100 system. The sample was loaded onto the column, pre-equilibrated with buffer A (20 mM Tris-HCl pH 8.0) and eluted with a linear gradient against buffer B (20 mM Tris-HCl pH 8.0, 1M NaCl) using a linear gradient from 0 to 65% of the buffer B concentration. Relevant fractions were combined and concentrated using Sartorius ultrafiltration devices with a molecular weight cutoff of 10,000 to a volume of ~5 ml. This sample was then loaded onto a GE HiLoad 16/60 Superdex 200 size-exclusion column pre-equilibrated with 10 mM Tris-HCl pH 8.0, 150 mM NaCl. The fractions containing pure gp5G484-PAAR complexes were combined and concentrated to 25 mg/ml with the help of similar Sartorius ultrafiltration unit. The protein was stored in the same buffer at +4 °C until it was used for crystallization. All purification buffers and the final protein solution contained NaN_3 at a concentration of 0.02% (w/v).

3.3.2.2.1 Identification of gp5-PAAR complexes

The experience with gp4G484-gp5.4 showed that two protein species – gp5G484-PAAR complex and gp5G484 (dimer of trimers) were accumulating during protein purification. The β-helical fragment of gp5 does not fully denature in sodium dodecyl sulfate, and runs during SDS-PAGE experiment as two species of ~27 and ~10 kDa (Figure 3.18 **b)**, sample bands 1 and 2, 2A2 and 2A6 fraction, respectively. PAAR proteins (MW ~9.7 kDa) have the same SDS-PAGE mobility as the fastest migrating band of gp5G484 (MW ~10 kDa). However, these species are separable by high resolution anion exchange chromatography (monoQ resin). The gp5 fragment is a highly negatively charged β-helix ((Kanamaru, et al. 2002), Figure 3.25 **a)**) that binds to the monoQ anion exchange resin stronger than gp5G484-PAAR complex. Thus, proteins can be separated: gp5G484-PAAR complexes are eluted earlier than gp5 fragment (all oligomeric forms) Figure 3.18 **a)**.

3.3.2.3 Crystallization and structure determination of gp5-PAAR complexes

The initial crystallization screening was carried out by the sitting drop method in 96 well SWISSCI 2-lens MRC plates using Jena Bioscience crystallization screens. Optimization of crystallization conditions was performed in 24 well-plates (Jena Bioscience) by hanging drop vapor diffusion. Crystallization drops of the 24 well-plate setup contained 1.25 μl of the protein solution in 10 mM Tris-HCl pH 8.0, 150 mM NaCl mixed with an equal volume of the well solution. Best crystals of the gp5G484-c1882 complex were obtained with the protein having the initial concentration of 13.5 mg/ml and equilibrated against 500 μl of the well solution containing 100-150 mM $CaCl_2$, 13-15% PEG 3350, 100 mM MES pH 6.5 (Figure 3.27 **a)**). Best

crystals of the gp5-VCA0105 complex were obtained with the protein at 15 mg/ml and equilibrated against 500µl of the well solution containing 13-14% PEG 2000, 100 mM NaAc pH 5.0 (Figure 3.27 **b)**).

a)   b)
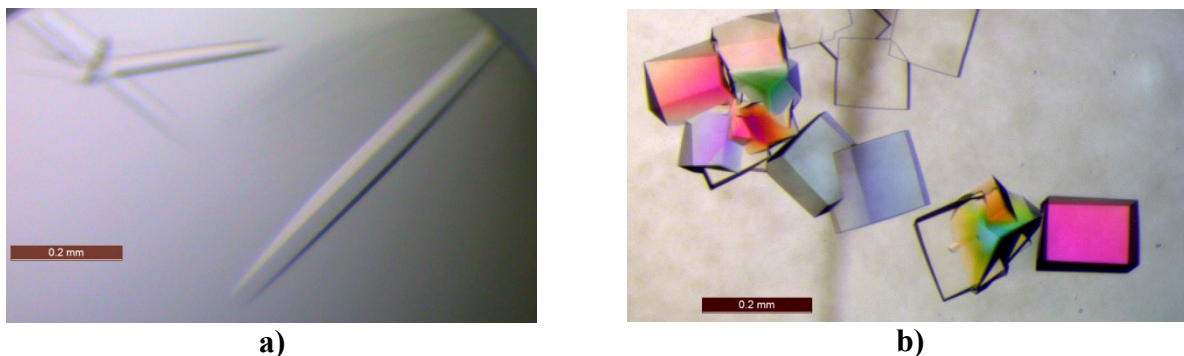
Figure 3.27 Crystals of the gp5_c1883-c1882 a) and gp5_VCA0018-VCA0105 b).

For data collection, the crystals were dipped for 20-45 seconds into cryo solutions containing either 30% of glycerol for gp5G484-c1882 or 25% of 2-methyl-2,4-pentanediol for gp5G484-VCA0105 in addition to the well solution components and flash frozen in a vaporized nitrogen stream at 100 K. Collection of diffraction data and fluorescent scans was carried out at the PXI and PXIII beam lines of the Swiss Light Source (SLS) at the Paul Scherrer Institute (Villigen, Switzerland), respectively, using X-rays with a wavelength of 1.000 Å for both crystals. Best gp5-c1882 and gp5-VCA0105 crystals diffracted to 3.4 Å and 1.9 Å resolution limits, respectively. The diffraction data was indexed, integrated, and scaled with XDS (Kabsch 2010, Kammerer, et al. 1998). The structure of the gp5-c1882 complex was solved by molecular replacement with PHASER (McCoy, et al. 2007) using the gp5R483 structure (paragraph **2.1**) as a search model. As the asymmetric unit contained four gp5-c1882 complexes, non-crystallographic symmetry (NCS) averaging was used to improve the electron density prior to model building. The model was built manually with Coot (Emsley and Cowtan 2004) and refined with REFMAC5 (Winn, et al. 2011) and PHENIX (Adams, et al. 2010) using NCS for torsion angles restraints. The structure of gp5-c1882 was subsequently used as a search model to solve the structure of the gp5-VCA0105 complex by molecular replacement with PHASER (McCoy, et al. 2007). There was only one complex per asymmetric unit. The structure was refined with PHENIX (Adams, et al. 2010, Ackermann 2003, Ackermann 2003, Ackermann 2003) and Coot (Emsley and Cowtan 2004, Fischetti 2006). Details of data reduction and refinement are given in the Table 3.6.

Table 3.6 X-ray data collection and refinement statistics of gp5_VCA0018-VCA0105 and gp5_c1883-c1882 complexes.
Data in parenthesis represent statistics for the highest resolution shell.

Data collection	gp5_VCA0018-VCA0105	gp5_c1883-c1882
Wavelength	1.0 Å	1.0 Å
Number of frames	720	480
Frame width (°)	0.25	0.25
Space group	$P4_222$	$P6_122$
Cell dimensions (Å)	a = 113.70, c = 76.90	a = 187.16, c = 238.89
Number of gp5-PAAR protein complexes per asymmetric unit	1	4
Resolution (Å)	80.4 – 1.9	48.6 – 3.4
R_{meas}	0.06 (0.37)	0.15 (1.01)
$<I / \sigma_I>$	20.0 (4.9)	12.6 (2.1)
Completeness (%)	99.9 (99.3)	99.8 (98.9)
Redundancy	6.9 (7.0)	7.0 (6.7)

Refinement		
Number of reflections		
Working	40054	33366
Test	2007	1692
R_{work} / R_{free}	0.16 / 0.21	0.184 / 0.243
B-factor ($Å^2$)	41.4	162.9
R.m.s. deviations		
Bond lengths (Å)	0.007	0.002
Bond angles (°)	0.995	1.039
Number of atoms		
Protein	2880	10848
Solvent and ligands	513	60
Ramachandran plot (%)		
Most favored	100.0	97.7
Additionally allowed	0.0	2.3
Outliers	0.0	0.0

3.3.3   Results and discussion

3.3.3.1   Morphology overview and structure analysis of the gp5_VgrG-PAAR complexes

The gp5-c1882 and gp5-VCA0105 complexes were purified and crystallized. The structures were solved by molecular replacement (Rossmann and Blow 1962) using the gp5R483 β-helix fragment as a search model. Structures of both complexes from T6SS are very similar to the gp5G484-gp5.4 complex structure from T4 phage, with minor differences. In all three complexes a single chain of the PAAR protein folds into a symmetrical cone-shaped structure with a sharp tip and a triangular base fully occupying the blunt end of the β-helix (Figure 3.28). The cone contains 9 short β-strands, three of which create its base and participate in binding to gp5 and six others form three β-hairpins that point toward the vertex of the cone, but have different lengths. The PAAR proteins interact with the gp5 β-helix via a virtually flat hydrophobic patch and 14 or 16 hydrogen bonds for c1882 or VCA0105, respectively (Figure 3.29). In both proteins, 12 hydrogen bonds (11 in gp5-gp5.4 structure) between the main chain atoms of the tip of gp5 and those of the PAAR domain form a perfect triangle surrounding the central hydrophobic patch creating a unique binding platform (Figure 3.30). PISA software (Krissinel and Henrick 2007) shows that ~16.5% of the PAAR protein surface is buried in this interface and the free energy of interaction between the VgrG tip and c1882 or VCA0105 PAAR proteins is -5.5 kcal/mol or -3.4 kcal/mol, respectively.

The three loops of the PAAR proteins from T6SS are stabilized by a Zn^{2+} ion positioned close to the cone's vertex, when in T4 gp5.4 the stabilizing is done by an Fe^{3+} ion (Figure 3.28 **d**). For gp5-VCA0105 was done a crystallographic refinement with all putative metal ions, which were fitted into the corresponding density and refined (Mn^{2+}, Fe^{3+}, Co^{2+}, Ni^{2+}, Cu^{2+} and Zn^{2+}). Bond length, bond angles, B-factor and the height of the peak in difference map suggest that the metal ion is whether Zn^{2+} or Fe^{3+}. To answer definitively on the coordinated metal ion identity question were done X-ray fluorescent experiments on the crystals of gp5-VCA0105 and gp5-c1882. These experiments proved that the metal ion in both T6SS PAAR proteins is Zn^{2+} (Figure 3.31). What made to rethink the conclusions for gp5.4. Following refinement of the gp5-gp5.4 structure with a Zn^{2+} ion revealed a strong peak up to 4.7 RMSD in the Fo-Fc map (the noise level is at ~2.7 RMDS), when for the Fe^{3+} there is no such peak. This proves the accuracy of the earlier conclusions on gp5.4.
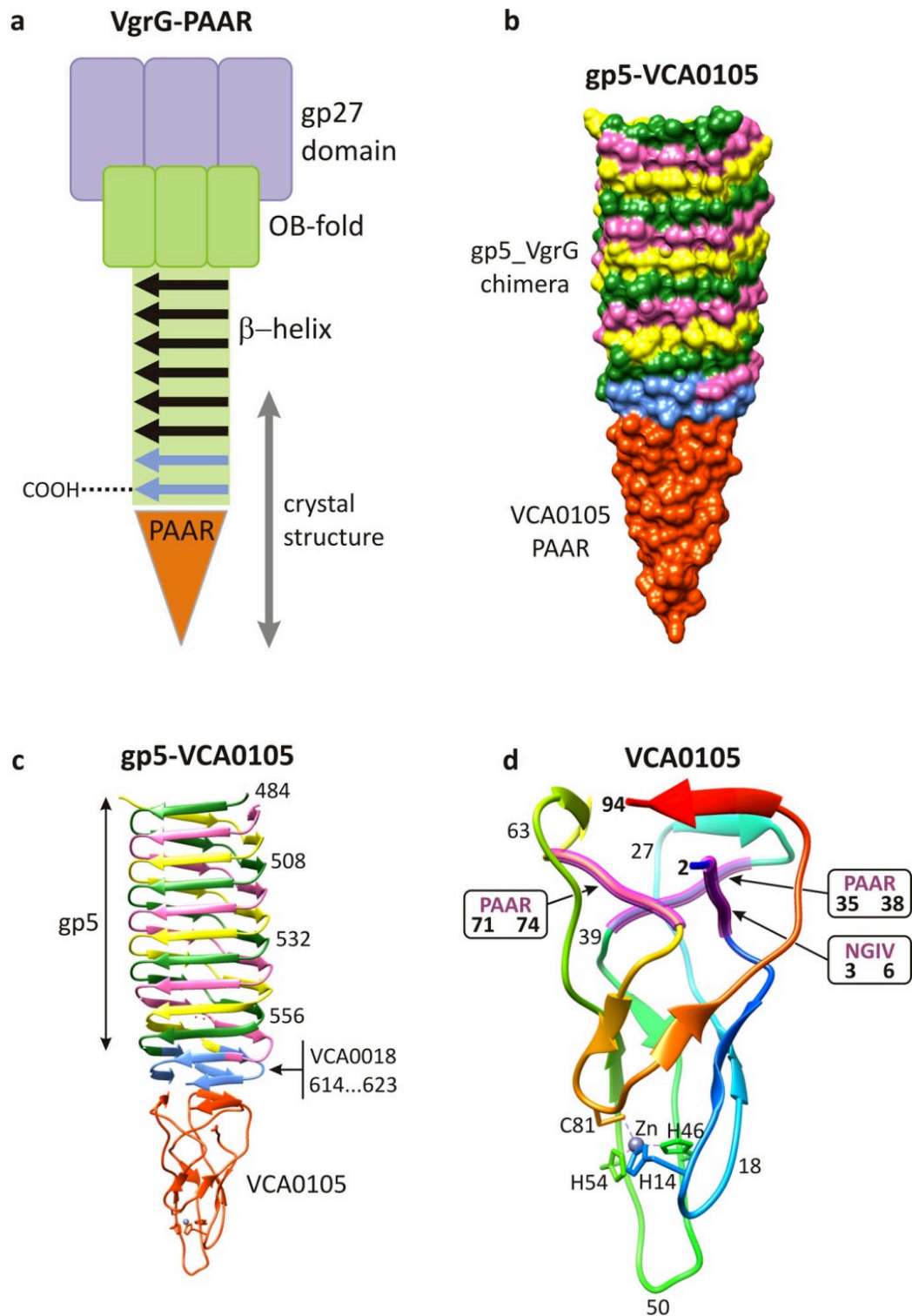
Figure 3.28 Crystal structure of the VCA0105 PAAR-repeat protein bound to its VgrG-like partner.
a) schematic representation of the conserved domains comprising the VgrG-PAAR complex. The last strands of the β-helix that form the PAAR binding site are in light blue. Gray arrow shows the fragment roughly corresponding to the crystal structure; **b)** molecular surface representation of the gp5_VCA0018-VCA0105 complex crystal structure; **c)** ribbon diagram of the gp5_VCA0018-VCA0105 complex; d) the polypeptide chain of the VCA0105 PAAR protein is colored in rainbow colors with N terminus in blue and C terminus in red. Residues responsible for Zn binding are labeled as well as PAAR motifs are highlighted in magenta.
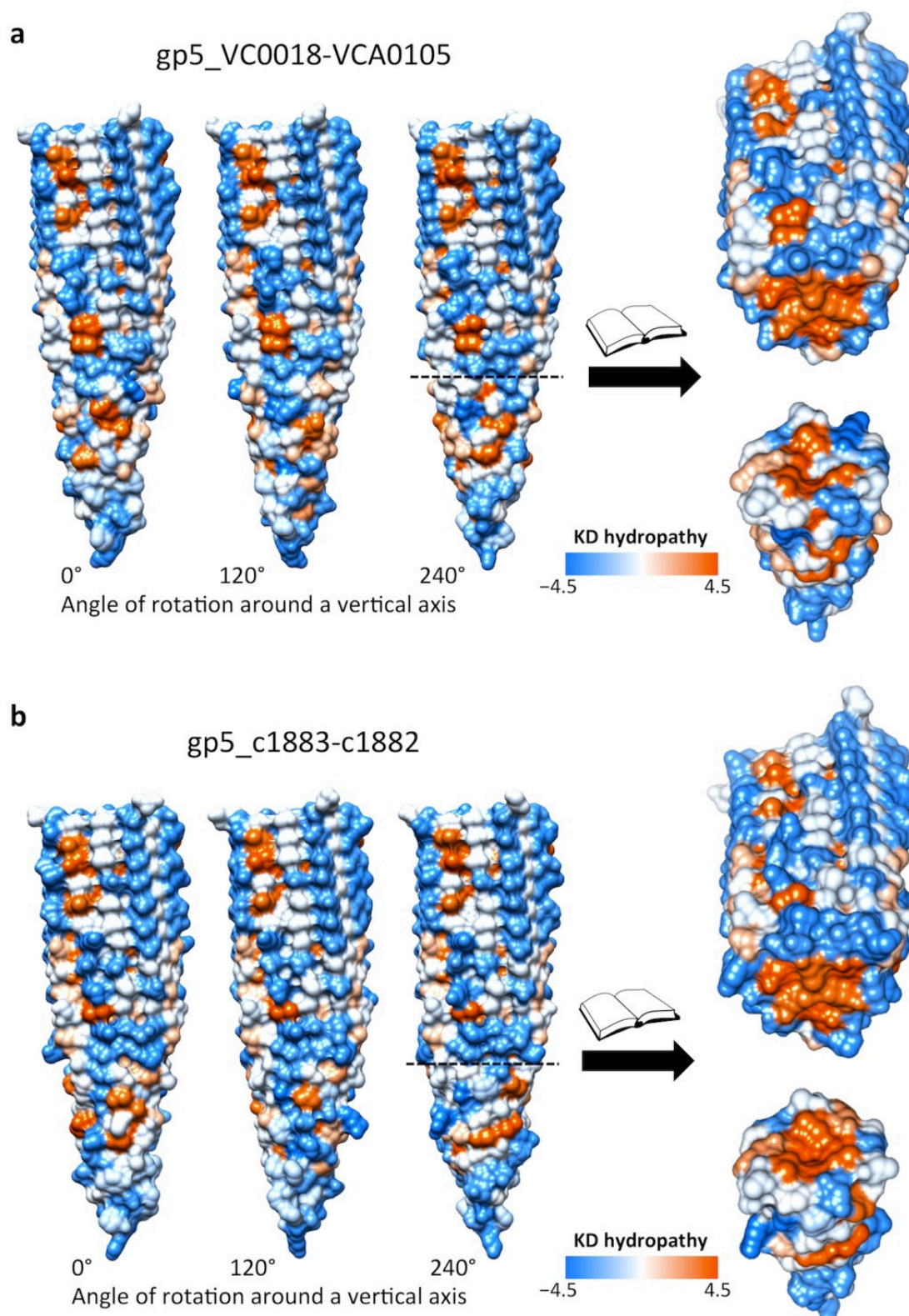
Figure 3.29 Surface representation of gp5-PAAR complexes and VgrG-PAAR interface.
Molecular surfaces are colored according to their hydrophobicity with sky blue, white, and orange corresponding to the most hydrophilic, neutral, and hydrophobic patches, respectively.
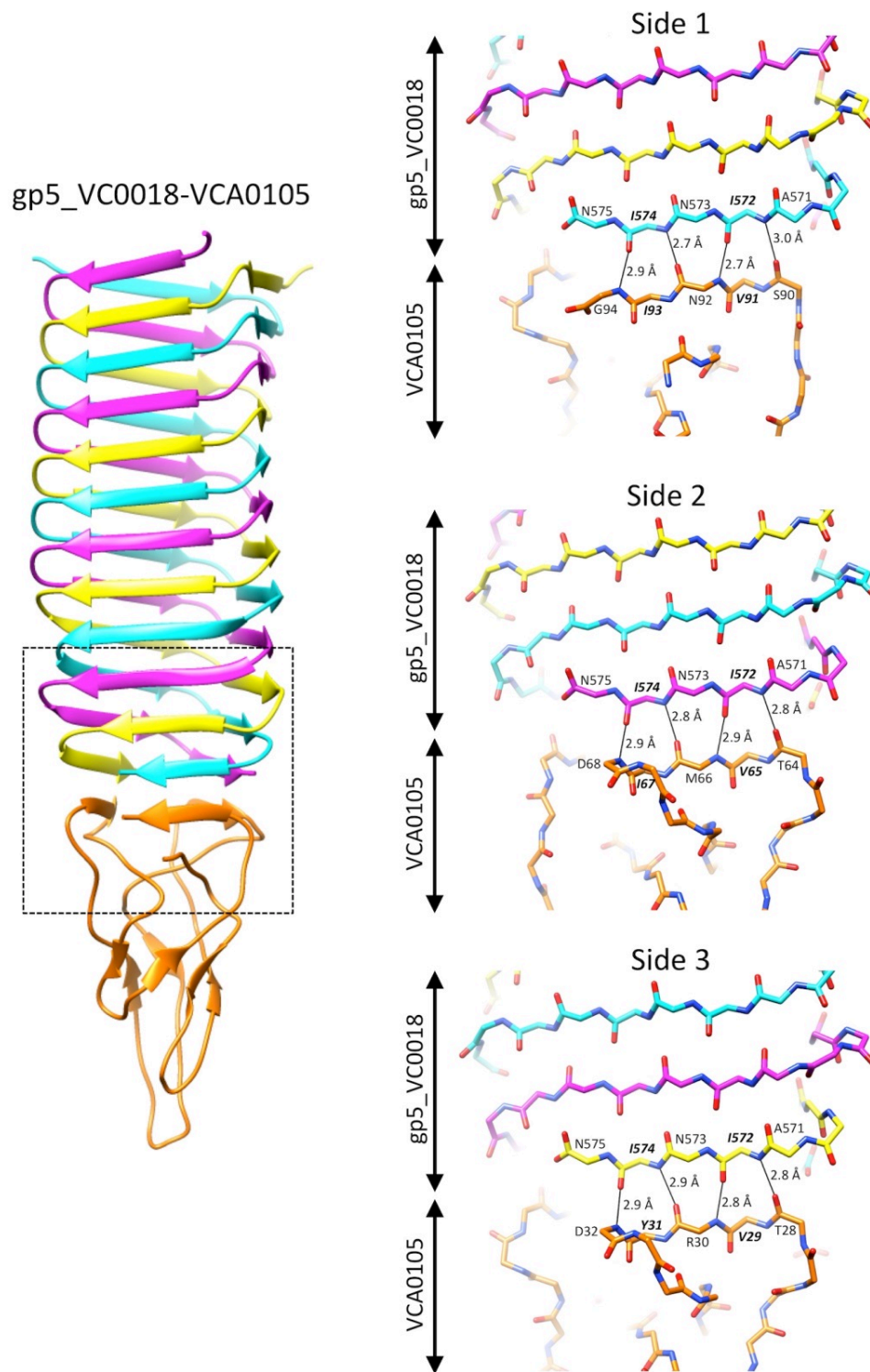
Figure 3.30 Main chain hydrogen bonding network of VgrG-PAAR interface.
The three panels on the right show the main chain hydrogen within each face of gp5 β-helix. Side chains are not shown for clarity.
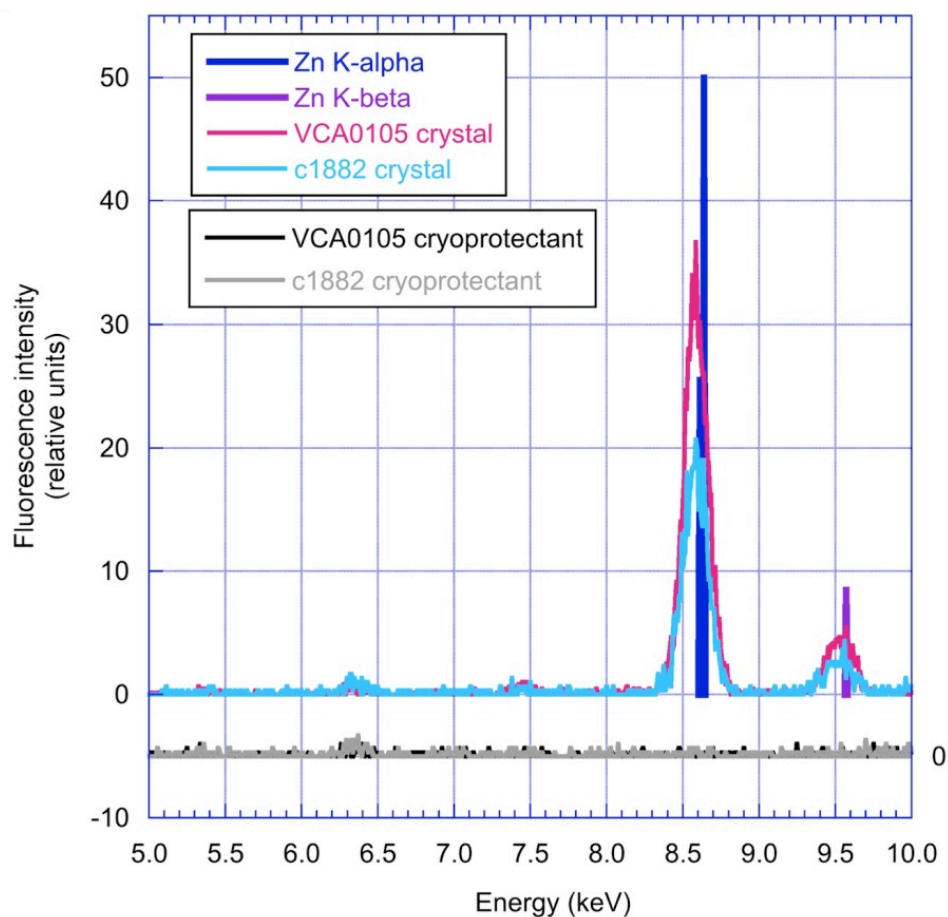
Figure 3.31 X-ray fluorescence spectra of gp5_VC0018-VCA0105 and gp5_c1883-c1882 crystals and their cryoprotectant solutions.
The excitation wavelength is 1.0 Å (~12.4 keV).

3.3.3.2 Conserved features of PAAR proteins

A metal binding site in T6SS PAARs has the same composition as in T4 PAAR – it consists of three histidines and one cysteine. These four residues are very well conserved in close homologs (Figure 3.32 **a)**). In more distant homologs these residues are replaced with similar or complementary metal-binding residues (arginines, lysines and glutamines) suggesting that they also coordinate a metal ion roughly at the same position.

The metal ion, being a natural ligand for this site, stabilizes the pointed tip of the PAAR domain and is likely to be important for its integrity during penetration of the target cell envelope.

The PAAR motif sequence is also conserved, but with a lower degree of conservation than the metal (Zn) binding site (Figure 3.32 **a)**). The PAAR motif function is to stabilize the fold by forming the central scaffold of the structure where the three parts of the polypeptide chain meet and intertwine (Figure 3.32 **b), c)**). The modules of the three-PAAR-motif-scaffold are braced against each other through the main chain hydrogen bonds that are shielded by hydrophobic residues from all sides (Figure 3.32 **d)**). c1882 and VCA0105 show 25.8 and 29.0 % sequence identity with gp5.4 and 1.08 and 1.27 Å RMSD of C_α atoms positions of the backbone after models superposition, respectively. However, sequence identity between c1882 and VCA0105 is 61.3 % and structure superposition RMSD (C_α atoms) is only 0.53 Å (Figure 3.33).
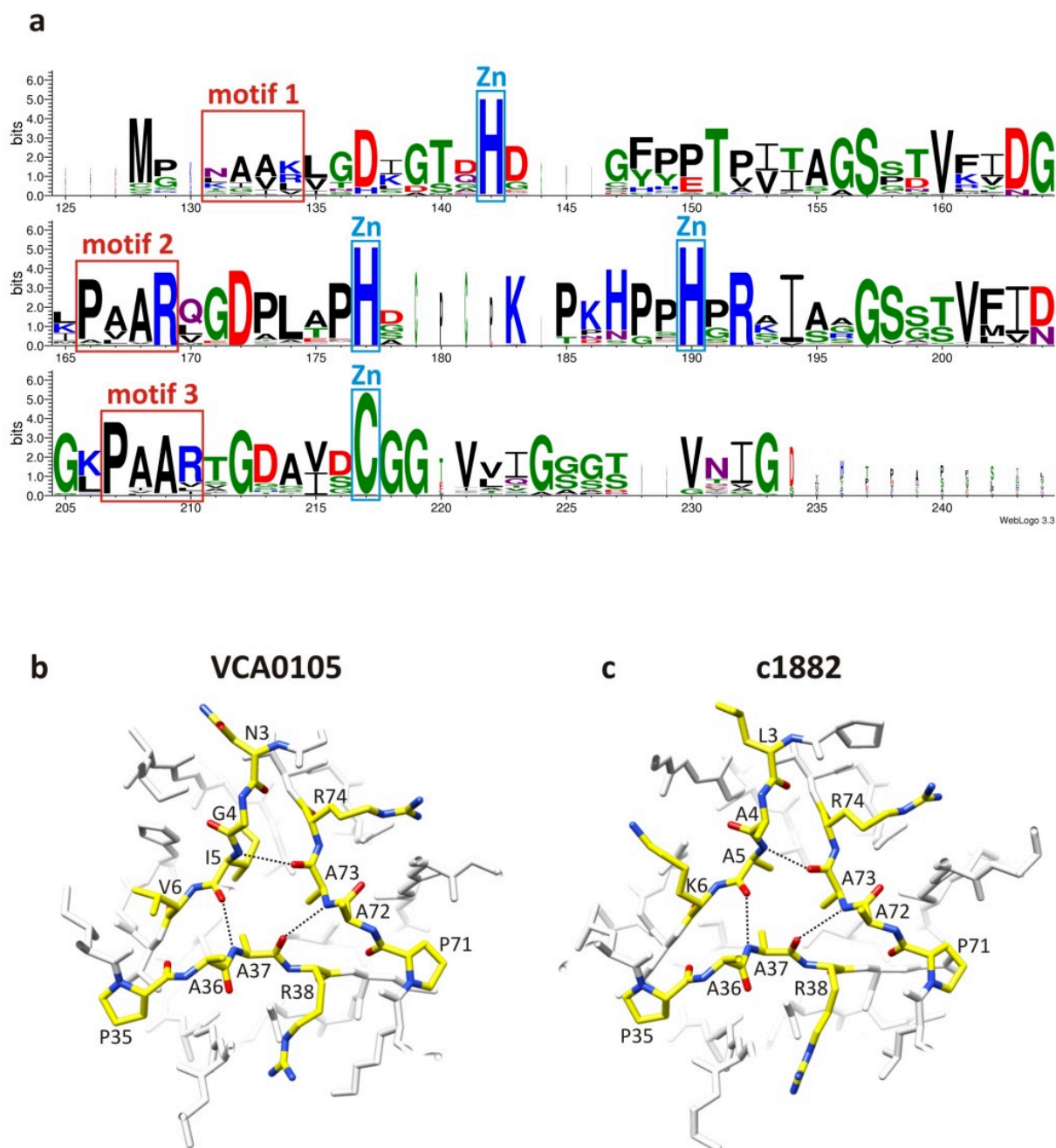
Figure 3.32 Conserved features of PAAR proteins.
a) WebLogo35 sequence alignment of VCA0105 homologs identified with BLAST36. The conserved PAAR motif and residues forming the metal (Zn) binding site are labeled. Pseudotrimeric organization of the three interacting PAAR motifs in VCA0105 **b)** and c1882 **c)**.
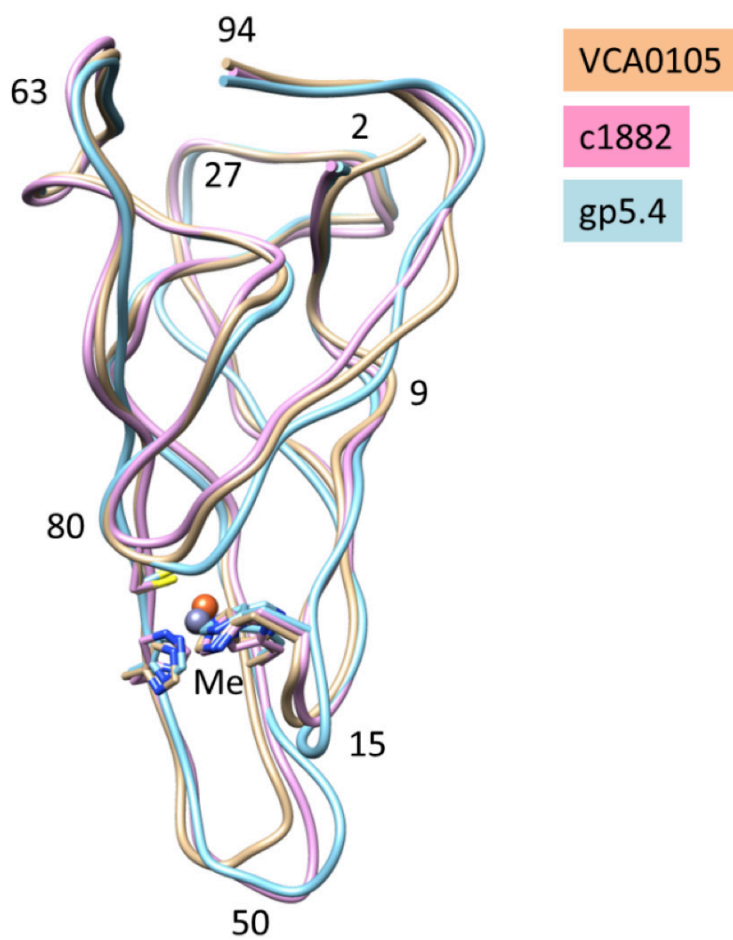
Figure 3.33 Superposition of VCA0105, c1882 PAAR structures onto T4 gp5.4.
Residue numbers (T6SS PAAR) are given at strategic positions.

3.3.3.3 Bioinformatic analysis of PAAR proteins

The PAAR-repeat proteins compose a diverse superfamily called CL15808 in the CDD database (Marchler-Bauer, Zheng and Bryant 2013) that contains three families PF05488, COG4104, and PF13665 (or DUF4150). The first two families are similar and describe PAAR domains that are on average ~95 residues long whereas the PF13665 family is somewhat more distant and its typical members contain ~130 amino acids. c1882, VCA0105 and gp5.4 are the representatives of the PF05488 family. Hundreds of hypothetical proteins in the database contain PAAR domains that are extended both N- and C-terminally. These extension domains are predicted to have various functions (Figure 3.34). The crystal structure shows that the termini of the PAAR domain are open to solution and thus can be extended without distorting the VgrG binding site (Figure 3.33). Very similar putative effector domains can be found fused to the C termini of VgrG proteins. Binding of these larger PAAR proteins to the tip of VgrG spikes would decorate the T6SS spike with a great variety of effector domains.

VCA0284, the larger of the two *V. cholerae* V52 PAAR proteins, carries a transthyretin domain (TTR) at its C terminus, which is a very common architecture of PAAR proteins (Figure 3.34 **c)**). TTR is an immunoglobulin-like domain that is known to form oligomeric structures in which these domains interact with each other or with other partners (Hamburger, et al. 1999). Thus, PAAR-associated TTR domains may act as adapters to further decorate the VgrG tip with effectors displaying TTR domains or serve to bind the spike to other TTR domain-containing proteins.
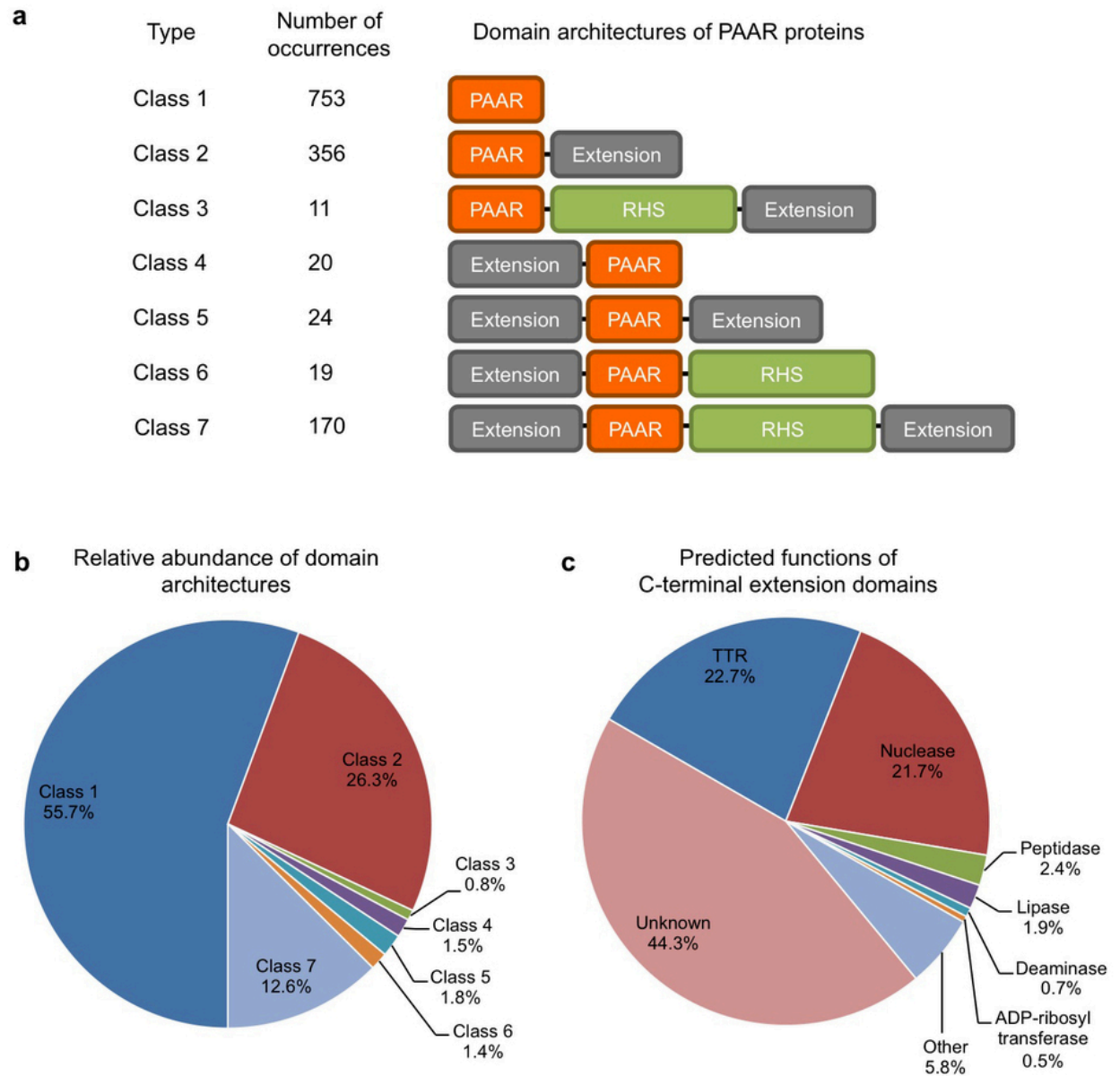
Figure 3.34 Bioinformatic analysis of PAAR proteins.
a) Domain organization of all known bacterial (non-phage) PAAR proteins; **b**) Relative abundance of the seven distinct domain organizations; **c**) Predicted functions of the C-terminal domains. TTR stands for transthyretin domain.

3.3.3.4   T6SS machine multiple effector translocation VgrG (MERV) model

c1882 and VCA0105 proteins as well as gp5.4, T4 PAAR protein cap the spike proteins of their contractile injection system. These proteins are VgrG in T6SS and gp5-gp27 protein complex in T4 phage. In addition to already proposed functions for T4 gp5.4 as sharpening and stabilizing the end of the spike β-helix during the initial piercing event of the target cell envelope, the structural and bioinformatic findings summarized above allow hypothesizing on additional PAAR-repeat protein functions. These functions are in close relation with T6SS feature of translocating the effector domains into the target cells. Because the crystal structures demonstrated that the canonical PAAR-repeat domain (~95 residues) can be extended N- or C-terminally without distorting its structure or its VgrG β-helix binding site was made a prediction that large PAAR proteins carrying effector domains will probably also bind to VgrG spikes and be translocated into target cells by the T6SS organelle. Considering the findings reported here and other published data was proposed that there are five mechanisms by which effectors can be incorporated into the T6SS spike complex (Figure 3.35). Three of them: 1) C-terminal extensions of the VgrG spike (Pukatzki, et al. 2007); 2) binding surface features on the VgrG protein (Dong, et al. 2013), (Hachani, et al. 2011); and 3) N- or C-terminal extensions of the PAAR protein (Koskiniemi, et al. 2013); are supported by direct or indirect experimental evidence. Two others: 4) binding surface features or additional domains (for example, the TTR domain) on PAAR proteins and 5) incorporation into the cavity formed by the gp27 domain of VgrG, remain speculative. Thus, the T6SS machine may be capable of delivering a multifunctional 'cargo' or multiple effector translocation VgrG (MERV) spike into the target prey cell in a single molecular translocation event driven by T6SS sheath contraction.
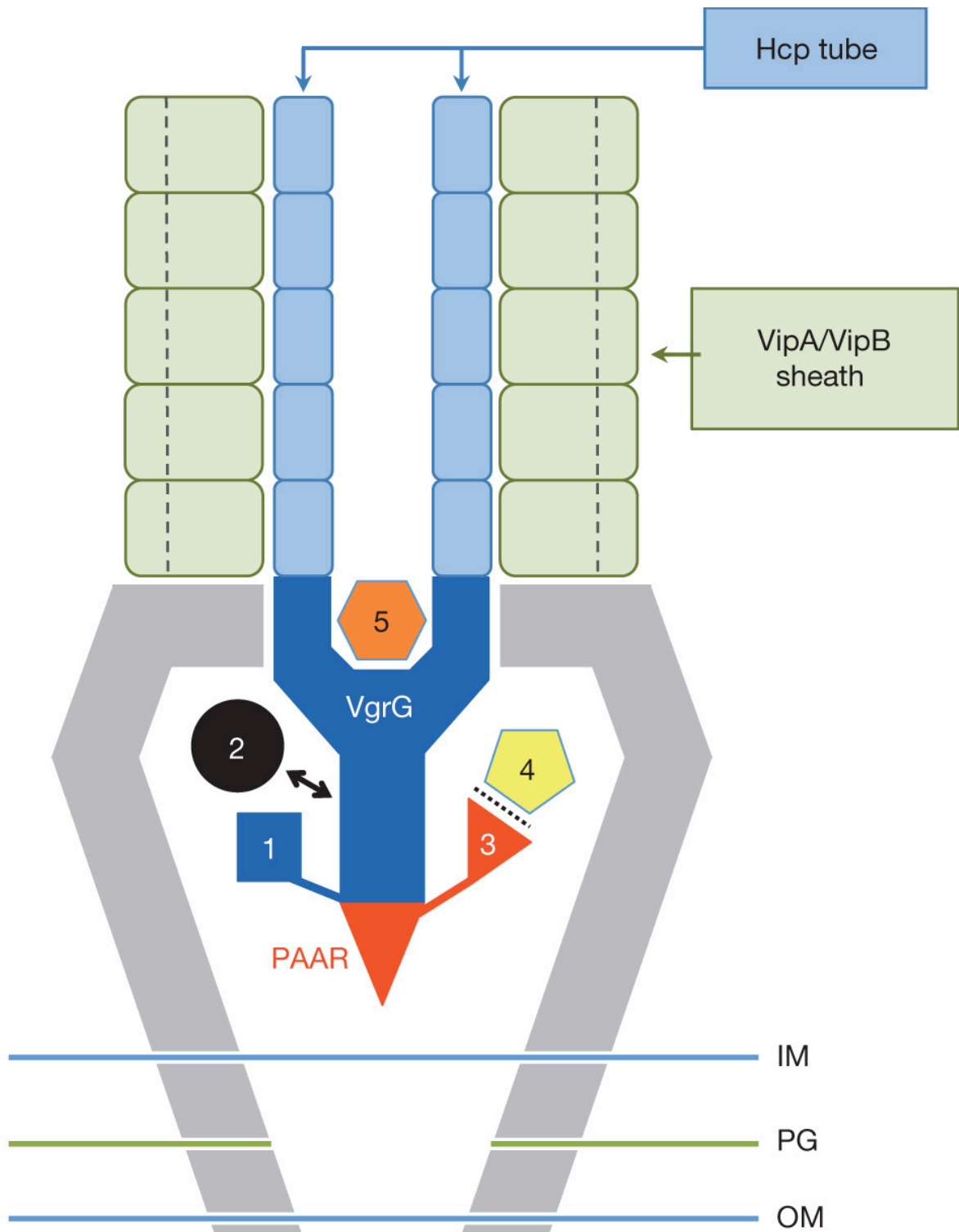
Figure 3.35 T6SS machine multiple effector translocation VgrG (MERV) model.
Effectors are predicted to be loaded onto the spike complex by five distinct mechanisms: 1) C-terminal extensions of the VgrG spike; 2) non-covalent binding to the VgrG spike; 3) N- or C-terminal extensions of the PAAR protein; 4) non-covalent binding to the PAAR protein or its extension domains; 5) incorporation into the cavity formed by the gp27 domain of VgrG. A single T6SS sheath contraction event translocates the VgrG spike with all of its cargo proteins into a nearby target cell. Other proteins making up the T6SS 'baseplate' (grey color) are not labeled but presumably reside within or attached to the inner and outer membranes and peptidoglycan layer (IM, OM, and PG, respectively).

3.4 Conclusions and perspectives

This Ph.D. thesis was focused on studies of three contractile ejection systems: R-type pyocins, bacteriophage T4, and the type VI secretion system. Proteins, responsible for attaching these large macromolecular machines to the host cell surface or involved in disruption of the lipid envelope were studied.

R-type pyocins are being developed into a tunable antimicrobial agent by our collaborators in AvidBiotics Corp. in S. San Francisco, CA, USA. Host specificity of a pyocin particle is determined by the fibers that emanate from the baseplate. The structure of the fiber is of great benefit for designing new chimerical pyocin particles that carry foreign fibers, which allow the pyocins to bind and kill specific pathogens. We have isolated the receptor-binding part of the fiber, and studied it using X-ray crystallography. Crystal structures of two distinct types of fibers that characterize all other fibers have been determined. Unfortunately, experiments aimed at finding the receptor-binding site on the fiber have been unsuccessful. Nevertheless, structural characterization and bioinformatic analysis allowed us to propose a function for the pyocin fibers that describes a mechanism by which the orientation of the fiber might trigger contraction of the pyocin particle.

In addition to the crystallographic studies of the fiber, the structure of the R-type pyocin particle was studied by cryo-electron microscopy (cryoEM). The cryoEM structure showed that similar to phage T4 the pyocin particle carried a centrally position spike protein, which is responsible for membrane piercing during infection. The receptor-binding part of the fiber was disordered in the cryoEM map. Crystallographic studies of the receptor-binding part of the fiber complement the cryoEM studies thus giving a complete structure of the pyocin particle. This structural information makes it possible to design stable fusions of pyocin fibers with other phage tail fibers in other places besides the "sweet spot" around residue 140 found by the trial-and-error method by our AvidBiotics colleagues.

Future aims in pyocin structural and functional characterization are the following:

1. Determining the structure of the N-terminal domain of the fiber. This study is complicated because this domain shows a high propensity to aggregation and it does not fold when expressed on its own. It might be possible to create a fusion of this domain with a stable, well behaving (in terms of solubility and folding) fragment of a trimeric fiber or tailspike from other phage.

2. Complete atomic-resolution structure of the baseplate. Such a structure, when combined with the fiber N-terminal domain structure, will explain how the receptor-binding signal

is transferred to the baseplate and to the sheath triggering its contraction. Three proteins $(PA0617)_6$-$(PA018)_{12}$-$(PA0619)_6$ form the major part of the baseplate, and it might possible to produce this complex in a form suitable for crystallization.

The other focus area of this thesis work was the central spike proteins found in all contractile tail-like systems. We characterized the structure of the β-helical domain of the T4 spike protein gp5 in great detail. Perhaps, the most remarkable finding here was that long fatty acid molecules are buried inside the gp5 β-helix.

We have been able to establish the composition and the structure of the complete T4 cell-puncturing spike. The identity of the protein decorating the tip of the spike has remained unknown for many years. We have solved this puzzle and found that gp5.4 forms a sharp conical extension to the spike. Gp5.4 belongs to the family of Proline-Alanine-Alanine-aRginine (PAAR) repeat proteins, which are ubiquitous in phages and the type VI secretion system. The polypeptide chain of PAAR-repeat proteins has a complex topology and is organized around a buried iron or zinc ion. The structure of the gp5-gp5.4 complex was determined to a resolution of 1.15 Å.

One of the important aspects of this thesis was work related to the bacterial type VI secretion system. The central spike protein in T6SS is called VgrG and it is a T4 gp5 ortholog. Employing a fragment of gp5 β-helix as a bait, and modifying its gp5.4 binding interface to mimic the tip of various VgrG spikes we were able to purify and solve crystal structures of several complexes that contained a chimerical gp5 with different PAAR-repeat proteins bound to it (VCA0105 from *Vibrio cholerae* and c1882 from *Escherichia coli* CFT). Structural information along with bioinformatic analysis of the T6SS clusters revealed putative, previously unknown, means of effector delivering into pray cells involving PAAR domains.

The topology of PAAR-repeat proteins resembles that of the apex domain of spikes containing the HxH double histidine motive found in P2 and φ92 phages (and their relatives). The latter also contain a centrally positioned iron ion. Combining all the data describing membrane-piecing spikes of contractile injection systems, it is evident that the architecture of the central spike proteins in all contractile systems is conserved. Furthermore, structural conservation with low sequence identity suggests that the central spikes of contractile systems interact with the host cell membrane non-specifically and most likely act as membrane-piercing drills.

To the contrary, bioinformatics suggests that the central spike proteins from contractile tail-like systems that target Gram-positive bacteria have enzymatic domains associated with them, and

thus interact with the host cell surface in a specific way. This is an interesting direction of future research.

Another interesting but challenging project is to verify whether multiple effectors can be attached to a single VgrG spike by means of their PAAR domains (or in any other way). Tools for expressing complexes containing several proteins – some of which are toxic – is a requirement in this project. Some of the proteins that bind to VgrG spikes contain more than 2000 amino acids, making such a project extremely difficult but exciting.

3.5   Molecular graphics

The following molecular graphics programs were used for preparation of the figures presented in this work. I would like to thank the people involved in the development and support of these programs. I am very grateful to the staff of UCSF Chimera project for supporting a very useful bulleting board and for creating excellent video tutorials and guides on program's special features.

Without such programs as Coot, PyMOL and UCSF Chimera it would be impossible to present the results of this thesis in a clear manner.

Figure	Program name
2.4, 2.6, 2.7, 2.15, 2.17-22, 2.24, 3.5-7, 3.11, 3.22, 3.24-27, 3.29-31, 3.33, 3.34	UCSF Chimera (Pettersen, et al. 2004)
2.16, 3.8, 3.11	Coot (Emsley and Cowtan 2004)
3.9, 3.23	PyMOL (Schrödinger n.d.)

3.6 Description of author contribution to various projects during his Ph.D. work at the Laboratory of Structural Biology and Biophysics

While a Ph.D. student at LBBS, the author of this thesis contributed to many projects some of which are discussed in this thesis and some have been or will be described elsewhere.

Projects described in this thesis:

Paragraph 2.2. CryoEM reconstruction of the R2 pyocin.

Processing of scanned micrographs: CTF correction, filtering, particles boxing. Initial model creation, iterative refinement, intermediate results analysis and removal of bad particles. Reconstruction analysis and its graphical representation (segmentation of reconstruction, preparation of figures).

Paragraph 2.3. Crystallographic studies of the R1 and R2 pyocin fibers.

Expression and purification of the d3 deletion mutant of R1 and R2 pyocin fibers, initial screening and optimization of crystallization conditions, co-crystallization experiments, data collection and processing, structure solution and refinement, characterization of the structures, molecular graphics.

Paragraph 3.1. Structure of the bacteriophage T4 gp5 β-helix.

Optimization of initial crystallization conditions, crystallization experiments with different bivalent metal ions, data collection and processing of gp5 Sr derivative data, structure solution, anomalous maps calculations, samples preparation for mass spectroscopy (including organic extraction), mass spectrometry results interpretation (supervised by Dr. Laure Menin), structure characterization, molecular graphics. Structure deposition to PDB.

Paragraph 3.1. Structure of gp5G484 dimer and gp5G484-gp5.4 complex.

Protein expression and purification, initial screening and optimization of crystallization conditions, data collection and processing, structures solution and refinement, structures characterization, molecular graphics, deposition of gp5G484-gp5.4 structure to PDB.

Paragraph 3.3.Type 6 secretion system (T6SS) warhead proteins (PAAR-repeat proteins).

Initial screening and optimization of crystallization conditions, data collection and processing, structures solution and refinement, structures analysis and deposition to PDB.

Projects that have been or will be described elsewhere:

Structural studies of bicyclic peptide ligands that bind to the human urokinase-type plasminogen activator protein.

Optimization of crystallization conditions, data collection and processing, structure solution and refinement, structures characterization, deposition of structure to PDB.

Structure and substrate specificity of the *Acinetobacter baumannii* phage AP22 gp54 tailspike.

Protein expression and purification, initial screening and optimization of crystallization conditions, co-crystallization with *A. baumannii* polysaccharide, data collection and processing, structures solution and refinement, structure characterization, molecular graphics.

Structural studies of PA0618 – the largest protein of the pyocin baseplate.

Protein expression and purification, initial screening and optimization of crystallization conditions, data collection and processing.

BIBLIOGRAPHY

1. Ackermann, H W. "Bacteriophage observations and evolution." *Res Microbiol.*, 2003: 245-251.

2. Adams, P D, et al. "PHENIX: a comprehensive Python-based system for macromolecular structure solution." *Acta Crystallogr D Biol Crystallogr.*, 2010: 213-221.

3. Aksyuk, A A, et al. "The tail sheath structure of bacteriophage T4: a molecular machine for infecting bacteria." *EMBO J.*, 2009: 821-829.

4. Aksyuk, A A, P G Leiman, M M Shneider, V V Mesyanzhinov, and M G. Rossmann. "The structure of gene product 6 of bacteriophage T4, the hinge-pin of the baseplate." *Structure*, 2009: 800-808.

5. Artimo, P, et al. "ExPASy: SIB bioinformatics resource portal." *Nucleic Acids Res.*, 2012: 597-603.

6. Bönemann, G, A Pietrosiuk, and A Mogk. "Tubules and donuts: a type VI secretion story." *Molecular Microbiology* 76, no. 4 (2010): 815–821.

7. Bella, J, M Eaton, B Brodsky, and H M Berman. "Crystal and molecular structure of a collagen-like peptide at 1.9 resolution." *Science*, 1994: 75-81.

8. Bernstein, F C, et al. "The Protein Data Bank: A Computer-based Archival File For Macromolecular Structures." *J Mol Biol.* 112, no. 3 (1977): 535-542.

9. Browning, C, M M Shneider, V D Bowman, D Schwarzer, and P G Leiman. "Phage pierces the host cell membrane with the iron-loaded spike." *Structure,* 2012: 326-339.

10. Dong, T G, B T Ho, D R Yoder-Himes, and J J Mekalanos. "Identification of T6SS-dependent effector and immunity proteins by Tn-seq in Vibrio cholerae." *Proc Natl Acad Sci USA*, 2013: 2623-2628.

11. Eiserling, F A, and L W Black. In *Molecular Biology of Bacteriophage T4*, edited by J D Karam, K N Kreuzer and D H Hall, 209-212. Washington, DC: American Society for Microbiology, 1994.

12. Emsley, P, and K Cowtan. "Coot: model-building tools for molecular graphics." *Acta Crystallogr D Biol Crystallogr.*, 2004: 2126-2132.

13. Fischetti, V A. "Using phage lytic enzymes to control pathogenic bacteria." *BMC Oral Health*, 2006: S16.

14. Folsch, J, M Lees, and G H Sloane Stanley. "A simple method for the isolation and purification of total lipides from animal tissues." *J Biol Chem.*, 1957: 497-509.

15. Frank, J, et al. "SPIDER and WEB: Processing and visualization of images in 3D electron microscopy and related fields." *J Struct Biol.*, 1996: 190–199.

16. Goldberg, E, L Grinius, and L Letellier. In *Molecular Biology of Bacteriophage T4*, by J D Karam, 347-356. Washington, DC: American Society for Microbiology, 1994.

17. Goujon, M, et al. "A new bioinformatics analysis tools framework at EMBL-EBI." *Nucleic acids res.*, 2010: W695-699.

18. Grigorieff, N. "FREALIGN: high-resolution refinement of single particle structures." *J Struct Biol.*, 2007: 117–125.

19. Hachani, A, et al. "Type VI secretion system in Pseudomonas aeruginosa: secretion and multimerization of VgrG proteins." *J Biol Chem.*, 2011: 12317-12327.

20. Hamburger, Z A, M S Brown, R R Isberg, and P J Bjorkman. "Crystal structure of invasin: a bacterial integrin-binding protein." *Science*, 1999: 291-295.

21. Hayashi, T, et al. "Cytotoxin-converting phages, phi CTX and PS21, are R pyocin-related phages." *FEMS Microbiol Lett.*, 1994: 239-244.

22. Hayashi, T, T Baba, H Matsumoto, and Y Terawaki. "Phage-conversion of cytotoxin production in Pseudomonas aeruginosa." *Mol Microbiol.*, 1990: 1703-1709.

23. Hendrickson, W A, and C M Ogata. "Phase determination from multiwavelength anomalous diffraction measurements." *Methods Enzymol.*, 1997: 494-523.

24. Hermoso, J A, J L Garcia, and P Garcia. "Taking aim on bacterial pathogens: from phage therapy to enzybiotics." *Curr Opin Microbiol.*, 2007: 461-472.

25. Herold, S, H Karch, and H Schmidt. "Shiga toxin-encoding bacteriophages--genomes in motion." *Int J Med Microbiol.*, 2004: 115-121.

26. Holm, L, and P Rosenström. "Dali server: conservation mapping in 3D." *Nucl Acids Res.*, 2010: W545-549.

27. Hsin, K, Y Sheng, M M Harding, P Taylor, and M D Walkinshaw. "MESPEUS: a database of the geometry of metal sites in proteins." *J Appl. Cryst.* 41, no. 5 (2008): 963-968.

28. Ishii, S I, Y Nishi, and F Egami. "The fine structure of a pyocin." *J Mol Biol.*, 1965: 428-431.

29. Ito, S, and K Makoto. "Relation between pyocins and bacteriophage in Pseudomonas aeruginosa." *J Gen Appl Microb.*, 1970: 10.

30. Jacob, F L S, and E Wollman. "Sur la biosynthèse d'une colicine et sur son mode d'action." *Ann Inst Pasteur*, 1952: 21.

31. Köhler, T, V Donner, and C van Delden. "Lipopolysaccharide as Shield and Receptor for R-Pyocin-Mediated Killing in Pseudomonas aeruginosa." *J Bacteriol.*, 2010: 1921-1928.

32. Kabsch, W. "XDS." *Acta Cryst D Biol Crystallogr.*, 2010: 125-32.

33. Kageyama, M, T Shinomiya, Y Aihara, and M Kobayashi. "Characterization of a bacteriophage related to R-type pyocin." 1980: 951-957.

34. Kammerer, R A, T Schulthess, R Landwehr, A Lustig, D Fischer, and J Engel. "Tenascin-C hexabrachion assembly is a sequential two-step process initiated by coiled-coil alpha-helices." *J Biol Chem.*, 1998: 10602-10608.

35. Kanamaru, S, et al. "Structure of the cell-puncturing device of bacteriophage T4." *Nature*, 2002: 553-557.

36. Kleywegt, G J. "Crystallographic refinement of ligand complexes." *Acta Cryst D Biol Crystallogr.*, 2007: 94-100.

37. Koskiniemi, S, et al. "Rhs proteins from diverse bacteria mediate intercellular competition." *Proc Natl Acad Sci U S A*, 2013: 7032-7037.

38. Kostyuchenko, V A, et al. "Three-dimensional structure of bacteriophage T4 baseplate." *Nature Structural Biology*, 2003: 688-693.

39. Krissinel, E, and K Henrick. "Inference of macromolecular assemblies from crystalline state." *J Mol Biol.*, 2007: 774-797.

40. Kuroda, K, R Kagiyama, and M Kageyama. "Isolation and characterization of a new bacteriophage, KF1, immunologically cross-reactive with F-type pyocins." *J Biochem.*, 1983: 61-71.

41. Kutter, E M, B Guttmann, G Mosig, and W Rüger. "T4 genomic map." In *Genomic Maps*, edited by S J O'Brien, 1–27. Cold Spring Harbor, N. Y.: Cold Spring Harbor Laboratory Press, 1990.

42. Lamzin, V S, A Perrakis, and K S Wilson. "The ARP/wARP suite for automated construction and refinement of protein models." In *International Tables for Crystallography. Volume F: Crystallography of biological macromolecules*, edited by M G Rossmann and E Arnold, 720-722. Dordrecht: Kluwer Academic Publishers, 2001.

43. Leiman, P G, and M M Shneider. "Contractile tail machines of bacteriophages." In *Viral Molecular Machines*, edited by V.B. Rao M.G. Rossmann, 93-114. New York: Springer, 2012.

44. Leiman, P G, et al. "Type VI secretion apparatus and phage tail-associated protein complexes share a common evolutionary origin." *Proc Natl Acad Sci USA*, 2009: 4154-4159.

45. Ludtke, S J, P R Baldwin, and W Chiu. "EMAN: semiautomated software for high-resolution single-particle reconstructions." *J Struct Biol.*, 1999: 82-97.

46. Lupas, A. "Coiled coils: new structures and new functions." *Trends in Biochem Sci.*, 1996: 375-382.

47. Marchler-Bauer, A, C Zheng, and S H Bryant. "CDD: conserved domains and protein three-dimensional structure." *Nucleic Acids Res.*, 2013: 348-352.

48. Matthews, B W. "Solvent content of protein crystalls." *J Mol Biol.*, 1968: 491-497.

49. McCoy, A J, R W Grosse-Kunstleve, P D Adams, M D Winn, L C Storoni, and R J Read. "Phaser crystallographic software." *J Appl Cryst.*, 2007: 658-674.

50. McRee, D E. "XtalView/Xfit - a versatile program for manipulating atomic coordinates and electron density." *J Struct Biol.*, 1999: 156-165.

51. Miller, E S, E Kutter, G Mosig, F Arisaka, T Kunisawa, and W Rüger. "Bacteriophage T4 genome." *Microbiol Mol Biol Rev.*, 2003: 86-156.

52. Mindell, J A, and N Grigorieff. "Accurate determination of local defocus and specimen tilt in electron microscopy." *J Struct Biol.*, 2003: 334–347.

53. Nakayama, K, et al. "The R-type pyocin of Pseudomonas aeruginosa is related to P2 phage, and the F-type is related to lambda phage." *Mol Microbiol.*, 2000: 213-231.

54. Otwinowski, Z, and W Minor. "Processing of X-ray Diffraction Data Collected in Oscillation Mode." *Methods in Enzymol.*, 1997: 307-326.

55. Papanikolopoulou, K, S Teixeira, H Belrhali, V T Forsyth, A Mitraki, and M J van Raaij. "Adenovirus fibre shaft sequences fold into the native triple beta-spiral fold when N-terminally fused to the bacteriophage T4 fibritin foldon trimerisation motif." *J Mol Biol.*, 2004: 219-227.

56. Pettersen, E F, et al. "UCSF Chimera--a visualization system for exploratory research and analysis." *J Comput Chem.*, 2004: 1605-1612.

57. Potier, N, et al. "Using nondenaturing mass spectrometry to detect fortuitous ligands in orphan nuclear receptors." *Protein Sci.*, 2003: 725-733.

58. Pukatzki, S, A T Ma, A T Revel, D Sturtevant, and J J Mekalanos. "Type VI secretion system translocates a phage tail spike-like protein into target cells where it cross-links actin." *Proc Natl Acad Sci USA*, 2007: 15508-15513.

59. Rossmann, M G, and D M Blow. "The detection of sub-units within the crystallographic asymmetric unit." *Acta Cryst.*, 1962: 24-31.

60. Scholl, D, and D W Jr Martin. "Antibacterial efficacy of R-type pyocins towards Pseudomonas aeruginosa in a murine peritonitis model." *Antimicrob Agents Chemother.*, 2008: 1647-1652.

61. Scholl, D, et al. "An engineered R-type pyocin is a highly specific and sensitive bactericidal agent for the food-borne pathogen Escherichia coli O157:H7." *Antimicrob Agents Chemother.*, 2009: 3074-3080.

62. Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.5.0.4.

63. Sheldrick, G M. "A short history of SHELX." *Acta Cryst. A*, 2008: 112-122.

64. Shneider, M M, Buth, S A, B T Ho, M Basler, J J Mekalanos, and P G Leiman. "PAAR-repeat proteins sharpen and diversify the Type VI secretion system spike." *Nature*, 2013: 350-353.

65. Takeya, K, Y Minamishima, K Amako, and Y Ohnishi. "A small rodshaped pyocin." *Virology*, 1967: 166-8.

66. Terwilliger, T C. "Automated structure solution, density modification and model building." *Acta Cryst D Biol Cryltallogr.*, 2002: 1937-1940.

67. Terwilliger, T C, and J Berendzen. "Automated MAD and MIR structure solution." *Acta Crystallogr D Biol Crystallogr.*, 1999: 849-861.

68. Traub, W, and K A Piez. "The chemistry and structure of collagen." *Adv Protein Chem.*, 1971: 243-252.

69. van Raaij, M J, A Mitraki, G Lavigne, and S Cusack. "A triple beta-spiral in the adenovirus fibre shaft reveals a new structural motif for a fibrous protein." *Nature*, 1999: 935-938.

70. van Raaij, M J, G Schoehn, M R Burda, and S Miller. "Crystal structure of a heat and protease-stable part of the bacteriophage T4 short tail fibre." *J Mol Biol.*, 2001: 1137-1146.

71. Watanabe, R, et al. "Efficacy of bacteriophage therapy against gut-derived sepsis caused by Pseudomonas aeruginosa in mice." *Antimicrob Agents Chemother.*, 2007: 446-52.

72. Williams, S R, D Gebhart, D W Martin, and D Scholl. "Retargeting R-Type Pyocins To Generate Novel Bactericidal Protein Complexes." *Appl Environ Microbiol.*, 2008: 3868-3876.

73. Winn, M D, et al. "Overview of the CCP4 suite and current developments." *Acta Crystallogr. D Biol. Crystallogr.*, 2011: 235–242.

74. Xiang, Y, and M G Rossmann. "The structure of bacteriophage phi29 head fibers has a supercoiled triple repeating helix-turn-helix motif." *Proc Natl Acad Sci USA*, 2011: 4806-4810.

CURRICULUM VITAE

PERSONAL DATA:

Name, Surname:	Sergii (Sergey) Buth
Date and place of birth:	26 July 1986, Sumy, Ukraine
Address:	Avenue de Tivoli 24, 1007 Lausanne, Switzerland
Phone:	+41789441661 (cell)
Email:	sergii.buth@epfl.ch
Languages:	English (fluent), French (intermediate), Ukrainian and Russian (mother tongues)

EDUCATION:

01.2009 – 11.2013   **Ph.D.** Ecole Polytecnique Federale de Lausanne, (SwissFederal Institute of Technology, Lausanne), Switzerland
Major: structural biology
Thesis director: Prof. Petr G. Leiman

09.2007 – 06.2008   **M.Sc.** Department of Chemistry, Kiev National Taras Shevchenko University, Ukraine

09.2003 – 06.2007   **B.Sc.** Department of Chemistry, Kiev National Taras Shevchenko University, Ukraine

PROFESSIONAL EXPERIENCE:

12.2013 – onwards   Postdoctoral research assisatant, Laboratory of Structural biology and Biophysics, EPFL, Lausanne, Switzerland

01.2009 – 11.2013   Graduate research assistant, Laboratory of Structural biology and Biophysics, EPFL, Lausanne, Switzerland

09.2009 – 01.2012   Graduate teaching assistant at EPFL undergraduate courses, EPFL, Lausanne, Switzerland

04.2008   Research assistant (visiting scholar) at the Laboratory of the Organic Coordination Chemistry, University of Strasbourg, France

09.2005 – 06.2008	Undergraduate/graduate research assistant at the Institute of Organic Chemistry NAS of Ukraine, X-ray single crystal diffraction center, Kiev, Ukraine

AWARDS AND HONORS:

2011	"Outstanding presentation" award at XXIInd Biennial Conference on Phage/Virus Assembly, Port Aransas, TX, USA
2009	FEMS Young Scientist Meeting Grant for XXIst Biennial Conference on Phage/Virus Assembly attendance, Annecy, France
2008	French government scholarship for join France-Ukraine doctorate (withdrawn by the applicant favoring position at the EPFL), Kiev, Ukraine
2008	Struchkov's prize (second diploma) of the X-ray Diffraction Center, Institute of Organoelement compounds, Moscow, Russia
2008	National academy of science of Ukraine award for outstanding students, Kiev, Ukraine
2007	Fellowship of the Pinchyk foundation, Kiev, Ukraine

LIST OF PUBLICATIONS

1. M. M. Shneider, **S. A. Buth**, B. T. Ho, M. Basler, J. J. Mekalanos, P. G. Leiman. "PAAR-repeat proteins sharpen and diversify the Type VI secretion system spike." *Nature*, 2013: 350-353.

2. S. Chen, I. Rentero Rebollo, **S. A. Buth**, J. Morales-Sanfrutos, J. Touati, P. G. Leiman, C. Heinis. "Bicyclic peptide ligands pulled out of cysteine-rich peptide libraries." *J. Am. Chem. Soc.*, 2013: 6562-6569.

3. S. Chen, D. Gfeller, **S. A. Buth**, O. Michielin, P. G. Leiman, C. Heinis. "Improving binding affinity and stability of peptide ligands by substituting glycines with D-amino acids." *ChemBioChem*, 2013: 1316-1322.

4. L.M. Yagupolskii, N.V. Kondratenko, O.I. Chernega, A.N. Chernega, **S.A. Buth**, Y.L. Yagupolskii. "A novel synthetic route to thiacyanine dyes with perfluorinated polymethine chain." *Dyes and Pigments*, 2008: 242–246.

5. L. M. Yagupolskii, I. I. Maletina, L. V. Sokolenko, Yu. G. Vlasenko, **S. A. Buth**. "N-Perfluoroalkylsulfonylimido derivatives of arenecarboxylic acid amides and their oxidative aza Hofmann rearrangement." *J. Fluorine Chem.*, 2008: 486-492.

6. A. S. Kanischev, V. M. Timoshenko, **S. A. Buth**, A. N. Chernega, Yu. G. Shermolovich. "About regioselectivity of alkylation reactions of polyfluor-alkylsubstituted NH-pyrazoles and NH-1,2,3-triazoles by olefines." *Zh. Org. Farm. Khim. (Russ.) (J. Org. Pharm. Chem.)*, 2008: 65-70.

7. K. G. Nazarenko, N. A. Shtil, **S. A. Buth**, A. N. Chernega, M. O. Lozinskii, A. A. Tolmachev. "The reactions of 2-chloro-1,4-benzothiazine-3-one derivatives with 'push-pull' enamines. Synthesis of 3-oxo-3,4-dihydro-2H-1,4-benzothiazine-2-yl derivatives of 1,3-dicarbonyl compounds." *Tetrahedron*, 2008: 4478-4485.

8. A. G. Balia, A. N. Chernega, **S. A. Buth**, A. N. Vasylenko, V. S. Brovarets, B. S. Drach. "Synthesis of the new derivatives of 2,5-diamino-1,3-thiazole and 2-thiohydantoin on the base of amidophenacylating reagents." *Zh. Obshch. Khim. (Russ.) (J. Gen. Chem.)*, 2008: 1427-1435.

9. P. V. Savitskii, R. I. Vas'kevich, Yu. L. Zborovskii, V.I. Staninets, **S. A. Buth**, A. N. Chernega. "Reactions of heterocyclization of 1-(benzothiazol-2-yl)-4-phenylthiosemicarbazide." *Zh. Org. Khim. (Russ.) (J. Org. Chem.)*, 2008: 407-410.

10. A. I. Vas'kevich, R. I. Vas'kevich, V. I. Staninets, **S. A. Buth**, A. N. Chernega. "Reaction of 2-(propin-2-ylthio)-5,6,7,8-tetrahydrobenzo[4,5]thieno[2,3-d]pyrimidin-4(3H)-one with arylsulfenyl chlorides." *Zh. Org. Khim. (Russ.) (J. Org. Chem.)*, 2007: 1530-1535.

11. A. S. Kanishchev, Yu. P. Bandera, V. M. Timoshenko, E. B. Rusanov, **S. A. Buth**, Yu. G. Shermolovich. "Synthesis of 5-polyfluoroalkyl-4-(p-tolylsulfonyl)-pyrazoles and 4-polyfluoroalkyl-5-(p-tolylsulfonyl)-pyrimidines from 1-dimethylamino-2-(p-tolylsulfonyl)-polyfluoroalk-1-en-3-ones." *Khim. Get. Soedin., SSSR (Russ.) (Chem. Heterocycl. Compd.)*, 2007: 1052-1058.

12. A. S. Kanishchev, Yu. P. Bandera, V. M. Timoshenko, **S. A. Buth**, A. M. Nesternko, Yu. G. Shermolovich. "4-Aryl(benzyl)sulfonyl-5-polyfluoroalkyl-2-chloro-1,2,3-triazoles – the first example of monocyclic N–Cl 1,2,3-triazoles." Khim. Get. Soedin., SSSR (Russ.) (Chem. Heterocycl. Compd.) 2007, 9, 1342-1352.

13. N. V. Briukhovetska, N. P. Kolesnyk, A. N. Chernega, **S. A. Buth**, Yu. G. Shermolovich. "The synthesis of chiral 1,2,4-benzothiadiazine derivatives." *Tetrahedron: Asymmetry*, 2007: 271–275.

14. V. I. Lubenets, S. V. Vasyiyuk, O. V. Goy, **S. A. Buth**, A. N. Chernega, V. P. Novikov. "Synthesis and properties 2,3-dioxo-1,2,3,4-tetrahydroquinoxaune-6-thiosulfonic acid S-esters." *Zh. Org. Farm. Khim. (Russ.) (J. Org. Pharm. Chem.)*, 2007: 56-63.