KU LEUVEN

FACULTY OF SCIENCE

Computational design of repeat proteins with putative antifreeze activity

Jeroen VRANCKEN

Supervisor: Prof. Dr. Marc De Maeyer

Mentor: Dr. Arnout Voet

Thesis presented in
fulfillment of the requirements
for the degree of Master of Science
in Biochemistry & Biotechnology

Academic year 2015-2016

Dankwoord

Eerlijk gezegd dacht ik om het slim te spelen en het dankwoord al te schrijven voor het eerste semester gedaan was. Bij nader inzien was het niet zo handig. Je kan wel iets schrijven, maar er kan nog zoveel veranderen of misgaan in de tijd die volgt.

Daarom zou ik graag als eerste Marc De Maeyer, Arnout Voet en Els Deridder willen bedanken. Ze hielpen me op weg in het labo en met mijn thesis, en bovendien gaven ze me ook tal van advies hoe ik de dingen beter kon aanpakken en doen. Ook wil ik ze graag bedanken voor het vertrouwen en de kans die ze me gaven om mijn thesis hier in dit labo te mogen doen, bedankt.

Verder wil ik graag mijn ouders, broer en zus bedanken. Ze kennen me ten slotte al het langste en kunnen gelukkig tegen mijn stoten en stunten. Daarmee wil ik ze graag bedanken dat ze er steeds voor me zijn en dat ze steeds zorgen zullen blijven maken, zelfs als het niet nodig is. En een klein extra dank je voor de broer, voor het meermaals herstellen van mijn laptop. Want ja, soms gebeurt dat wel eens.

Of course I would also like to thank the Biomol team and the students in the lab (Boyin, Ellen, Joren, Ovia, Thijs, Tom, Wim, and Xiaoyu) for making the lab an enjoyable, fun, and interesting place to be.

Furthermore, I would like to thank BioMacS, and especially professor Waelkens, for giving the opportunity to analyse the purified samples via a MALDI TOF MS.

And last, but not least, my gems. Thank you Alessandra, Charlotte, Elfri, Kwinten, Lieselot, Margaux, Paulien, and Sarah, or just simply a big thanks to the 'midwives', as most of them are a part of it. For the corrections, answers to my many annoying questions, the enjoyable and sometimes late or crazy talks, the postcards, the adventures, the support and worries, and the bad jokes which only few people would get.

So yes, thank you all.

Table of Contents

List of Figures

List of Tables

Glossary

2D	Two-dimensional
3D	Three-dimensional
AA	Acrylamide
AFGP	Antifreeze glycoprotein
AFP	Antifreeze protein
Ala, A	Alanine, an amino acid
APS	Ammonium persulphate
ARS	Ancestral sequence reconstruction
Asn, N	Asparagine, an amino acid
Asp, D	Aspartic acid, an amino acid
BAA	Bis-acrylamide
bp	Base pairs
BSA	Bovine serum albumin
CD	Circular dichroism
CfAFP	*Choristoneura fumiferana* antifreeze protein
CHCA	α-cyano-4- hydroxycinnamic acid
CLIPS	Crystal Lattice Interacting Protein Scaffolds
CV	Column volume
Cys, C	Cysteine, an amino acid
DNA	Deoxyribonucleic acid
dNTPs	Deoxynucleoside triphosphates
EDTA	Ethylenediaminetetraacetic acid
ESI	Electrospray ionization
Gly, G	Glycine, an amino acid
GST	Glutathione S-transferase
His$_6$	Hexahistidine
His, H	Histidine, an amino acid
IBP	Ice-binding protein
IBS	Ice-binding site
Indel	Deletions and insertions
IPTG	Isopropyl β-D-1-thiogalactopyranoside
IRI	Ice recrystallization inhibition
LB	Lysogeny broth
LpIBP	*Lolium perenne* antifreeze protein
m/z	Mass to charge ratio
MALDI	Matrix-assisted laser desorption/ionization
MBP	Maltose-binding protein
MD	Molecular dynamics
mdeg	Millidegrees
MpAFP	*Marinomonas primoryensis* antifreeze protein

MW	Molecular weight
n.d.	No date
Ni-NTA	Nickel-nitrilotriacetic acid
NMR	Nuclear magnetic resonance
OD	Optical density
Ori	Origin of replication
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
PDB	Protein data bank
RE$_3$Volutionary	REverse Engineering Evolutionary
RiAFP	*Rhagium inquisitor* antifreeze protein
RMSD	Root-mean-square deviation
RNA	Ribonucleic acid
RPM	Rounds per minute
SbwAFP	Spruce budworm antifreeze protein
SDS	Sodium dodecyl sulphate
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
Ser, S	Serine, an amino acid
TAE	Tris-acetate-EDTA buffer
TEMED	Tetramethylethylenediamine
TFA	Trifluoroacetic acid
TH	Thermal hysteresis
THP	Thermal hysteresis protein
Thr, T	Threonine, an amino acid
TiAFP	*Typhula ishikariensis* antifreeze protein
TmAFP	*Tenebrio molitor* antifreeze protein
TOF MS	Time-of-flight mass spectrometry
Usp	Universal Stress Protein
Val, V	Valine, an amino acid
WT	Wild type
Xxx	Any amino acid

Summary

During this project, an attempt was made to create Crystal Lattice Interacting Protein Scaffolds (CLIPS) by redesigning antifreeze proteins (AFPs). The chosen AFPs contain linear repetitive structures and can thus be used as building blocks by varying the amount of repeats. This unique feature gives them potential applications, such as binding or coordinating crystal lattices, as it is derived from AFPs.

The project was split into two parts. During the first part, a procedure was developed, based on RE$_3$Volutionary protein design, which is a form of reverse engineering evolution, to obtain CLIPS. It allowed us to obtain reconstructed ancestral sequences and sort them based on their properties. Based on this ranking, the most promising sequence was chosen and this construct was ordered and used for the second part, the expression and purification of CLIPS.

However, during multiple attempts only the wild type AFPs were successfully expressed and purified. This prevented the additional goal of analysing whether the redesigned proteins retained antifreeze activity and if so, a comparison with the wild type AFPs. Explanations and improvements were sought, as there were no purified redesigned proteins present. This lead to a new protein design strategy based on the RE$_3$Volution design principle, this time two instead of one single repeat were chosen as a template. The newly found sequences showed significant differences with the original results, making them very promising candidates for *in vitro* validation, which will hopefully result in stable proteins and insights to improve computational protein design.

Samenvatting

Tijdens dit project werd een poging ondernomen om ‘Crystal Lattice Interacting Protein Scaffolds’ (CLIPS) te ontwerpen vertrekkende van antivries proteïnen (AFPs). De gekozen AFPs bevatten lineair herhaalde structuren, waardoor ze ideaal zijn voor het gebruik als bouwstenen door het aantal herhalingen te laten variëren. Deze unieke eigenschap maakt de AFPs zeer geschikt voor een aantal doeleinden zoals het coördineren van metaal- en ijsroosters, doordat de CLIPS afgeleid zijn van AFPs.

Het project werd gesplitst in twee delen. Tijdens het eerste gedeelte werd een procedure opgesteld voor het bekomen van CLIPS. Dit werd gedaan via RE$_3$Volutionary methoden, een vorm van reverse engineering evolutie. Hierdoor worden verscheidene potentieel evolutionaire sequenties verkregen die gerangschikt werden op basis van hun fysicochemische eigenschappen. Via deze resultaten kan de beste sequentie gekozen worden voor het tweede deel: de expressie en purificatie van CLIPS.

Tijdens verscheidene pogingen werden enkel de wild-types AFPs tot expressie gebracht. Dit voorkwam het controleren of de nieuwe CLIPS antivries activiteit hebben, zoals het bijkomende doel beschreef. Als dit zo was dan konden ze vergeleken worden met de wild-types. Verklaringen en verbetering werden gezocht omdat er geen CLIPS opgezuiverd waren. Dit leidde tot een nieuwe aanpakmethode gebaseerd of de RE$_3$Volutionary methode. Deze keer werden er echter twee lussen in plaats van één lus beschouwd als een herhalingsstructuur. De nieuw bekomen sequenties vertoonden significante verschillen met de eerste resultaten. Hierdoor zijn ze veelbelovend voor een verdere *in vitro* validatie, wat hopelijk zal resulteren in stabiele proteïnen en een verbeterd inzicht in computationeel proteïne ontwerp.

1

General introduction

1.1 Antifreeze proteins

Antifreeze proteins (AFPs) have been studied for almost 50 years (DeVries and Wohlschlag, 1969). New AFPs are still being discovered. These proteins are named after their activity: they have an affinity for ice which lowers the freezing point of body fluids without a significant effect on the melting point, thereby generating a thermal hysteresis (TH).

The AFPs make it possible for organisms to survive sub-zero environments, giving rise to new advantages such as a safe place against predators and new possibilities when a population gets too large. The freezing point of most body fluids is already slightly depressed by colligative properties. These can be ions or small organic molecules that are dissolved in the solution. AFPs, which are only present in a few organisms, will depress the freezing point even more by using a different mechanism (DeVries, 1984).

The AFPs were first discovered in some Antarctic fish such as *Trematomus borchgrevinki* and *Dissostichus mawsoni*, but are present in several other organisms as well. The first discovered AFP was in fact an antifreeze glycoprotein (AFGP) which is only found in fish. It allows marine fish to survive in seawater by lowering the freezing point of the blood by ~ 1 °C, thus reaching a freezing point around -2.0 °C, which is just below the freezing point of the seawater (-1.9 °C) (DeVries et al., 1970; Davies and Hew, 1990).

After the discovery of the AFGP, more fish were examined for antifreeze activity, which led to the discovery of several more AFPs (Davies et al., 1988).

1.1.1 Fish antifreeze proteins

The newly discovered fish AFPs were arranged into two categories. The first category contained the AFGPs, while the second category contained the AFPs. The latter category was, based on their molecular architecture, subdivided into three groups. The groups were named type I to type III, after the order in which they were first discovered (Davies et al., 1988). Nearly ten years later, another AFP type was discovered and is referred to as type IV (Deng et al., 1997). Although type IV AFPs are only found in low concentrations in a few species, it is considered a different AFP. It is thought that the extremely low antifreeze activity of type IV AFP arose by coincidence or that selection never occurred on this type because a better AFP, namely type I, was already present in the same species. This might explain the extreme low activity, low concentrations, and the aggregation and precipitation when higher concentrations were used in *in vitro* experiments. Nevertheless, due to their antifreeze potential, they are categorized as AFPs (Gauthier et al., 2008).

The differences between the four fish AFP types and the AFGPs is mostly found in the conservation and type of the ice-binding surface (IBS). The structural differences can be seen in figure 1.1 A-D on page 3, while information about the Protein Data Bank (PDB) code, molecular weight (MW) and TH activity can be found in table 1.1 (Wu et al., 2001; Nishimiya et al., 2008; Hobbs et al., 2011; Middleton et al., 2012; Hakim et al., 2013).

Table 1.1: Properties and identification of AFPs

	AFGP	**Type I**	**Type II**	**Type III**
Organism	Antarctic notothenioids[1]	*Myoxocephalus scorpius*[1]	*Hemitripterus americanus*[1]	Zoarcidae family
PDB code	/	4KE2	2PY2	4UR4
MW (kDa)	2.6-33	3.3-4.5	14-24	6
TH (°C)	0, 2-0, 8	0.82	0.45	0.7
Concentration[2]	20	10	0.15 mM	0.4 mM

	Type IV	**TisAFP**	**SbwAFP**	**MpAFP**
Organism	*Myoxocephalus octodecemspinosus*	*Lolium perenne*	*Choristoneura fumiferana*	*Marinomonas primoryensis*
PDB code	/	3ULT	1M8N	3P4G
MW (kDa)	12	13.5	9	34
TH (°C)	0.5	0.25	1.08	2
Concentration[2]	2 mM	2	20 μM	0.5

	TmAFP	**RiAFP**	**TisAFP**
Organism	*Tenebrio molitor*	*Rhagium inquisitor*	*Typhula ishikariensis*
PDB code	1EZG	4DT5	3VN3
MW (kDa)	8.4	13	23
TH (°C)	3.4	6	0.42
Concentration[2]	2	0.5	48 μM

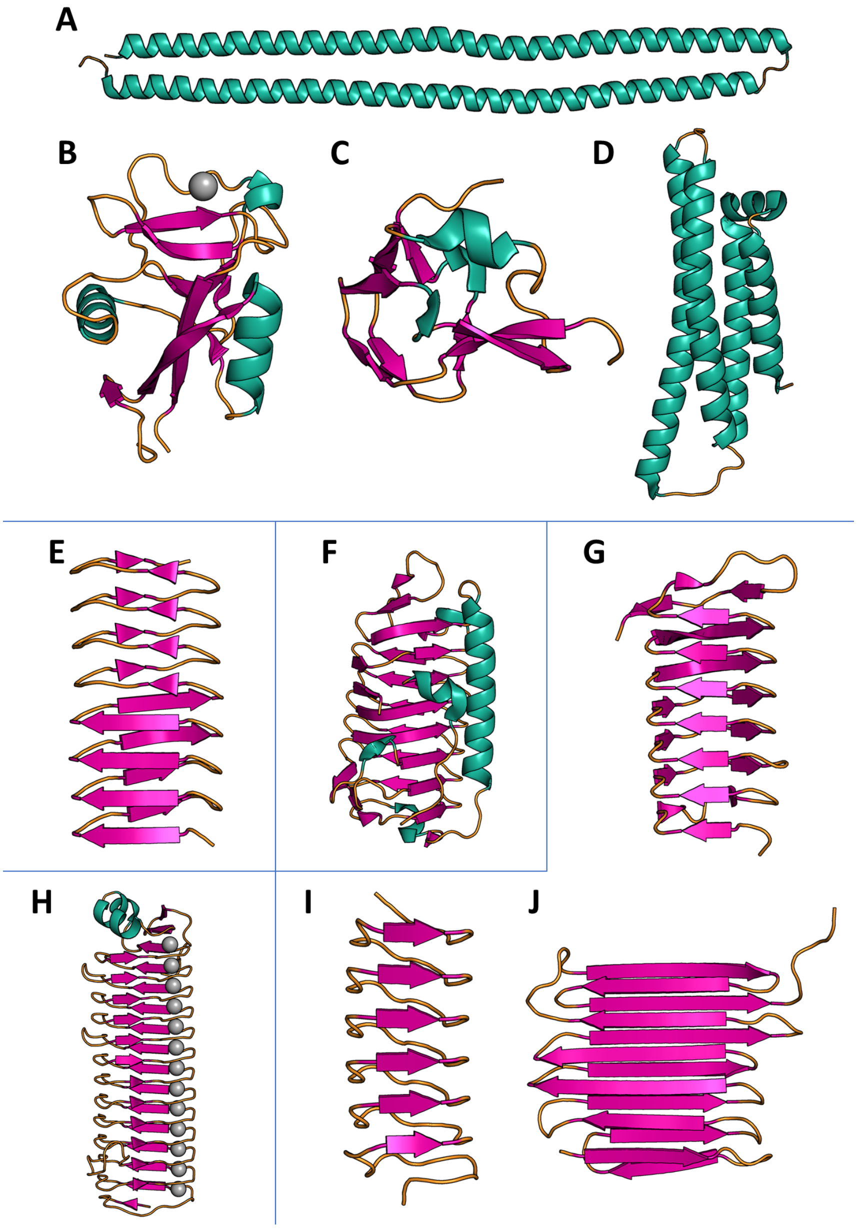


Figure 1.1: **X-ray crystal structures from different antifreeze and ice-binding proteins.** α-helical structures are shown in blue, β-sheets in purple, coils and loops in orange, and calcium ions are shown as grey spheres. Cartoon **A** shows a fish type I AFP from *Pseudopleuronectes americanus* (PDB: 4KE2). Cartoons **B** and **C** show a fish type II and III AFP from *Clupea harengus* (PDB: 2PY2) and *Zoarces viviparus* (PDB: 4UR4), respectively. Fish type IV AFP is very similar to the human apolipoprotein E3 (PDB: 1EA8), shown in cartoon **D**. A vegetal AFP from *Lolium perenne* (PDB: 3ULT) is shown in **E**. Cartoon **F** shows a fungal AFP from *Typhula ishikariensis* (PDB: 3VN3). Cartoon **G** shows an AFP from the insect *Choristoneura fumiferana* (PDB: 1M8N) and **H** shows a bacterial AFP from *Marinomonas primoryensis* (PDB: 3P4G). Cartoons **I** and **J** show insect AFPs from *Tenebrio molitor* (PDB: 1EZG) and *Rhagium inquisitor* (PDB: 4DT5), respectively.

AFGP

The genes for the AFGPs arose approximately 5 to 15 million years ago, the same time as when the freezing of the Antarctic Ocean occurred (Chen et al., 1997b). For the Antarctic icefish, belonging to the suborder Notothenioidei, the gene probably diverged from a trypsinogen encoding gene through alternative splicing and tandem duplication. The AFGP from Arctic cod, belonging to the family *Gadidae*, are very similar to the previous described AFGPs, yet there is evidence that they probably are a rare example of a very similar convergent evolution. There is quite some evidence to confirm this assumption, such as the codon choice for the typical alanine-alanine-threonine repeat and that the AFGP gene loci is different for Antarctic icefish and Arctic cod (Chen et al., 1997b).

The most common AFGP, for both types of fish, consists of an alanine-threonine repeat with a disaccharide (β-D-galactosyl-α-N-acetyl-D-galactosamine, where the two sugar groups are linked $1 \rightarrow 3$) attached to the threonine. The ratio of alanine and threonine in these repeats is 2:1, although some species alternate between alanine and proline, and the Arctic cod alternates between threonine and arginine as well (Chen et al., 1997a).

Like the AFPs, the AFGPs have different isoforms too. These isoforms are present as polyproteins in the AFGP gene instead of processing existing proteins through cleaving or splicing. These different isoforms not only allow small structural changes such as a proline residue instead of an alanine residue, but allows variation in the protein's size as well, giving additional ice-binding residues. The observed correlation between the length and activity of different isoforms is interesting as well and will be explained further in section 1.1.2 (Yeh and Feeney, 1996; Harding et al., 2003; Li and Jin, 2004).

Type I

The fish type I AFP, as shown in figure 1.1 A, has an α-helix content of more than 95% and an alanine content of about 65%. Each winding consists of 11 amino acids with the following consensus sequence: Thr-$(\text{Xxx})_3$-Ala-$(\text{Xxx})_3$-Ala-Xxx-Xxx, with Thr as threonine, Ala as alanine and Xxx as any amino acid (Sun et al., 2014).

Type II

In contrast with the α-helical type I, type II has a globular structure with a higher β-sheet content and an unusual high amount of cysteine residues. Type II has two subcategories, namely the calcium dependent and calcium independent AFPs. The calcium dependent protein is shown in figure 1.1 B. The structure between both groups is mostly the same, both contain two twisted β-sheets with an anti-parallel orientation and an α-helix on both sides. The biggest difference can be found in the IBS. For the calcium dependent proteins,

[1] Also present in different fish species

[2] The unit of concentration is mg/mL unless stated differently

the IBS is mostly made up of four residues (Thr96, Leu97, Thr98, and Thr115, according to the herring AFP (2PY2) deposited in the PDB). These residues are located on a flat side of the protein and are able to form a hydrogen bond with the calcium binding site, suggesting that the divalent ion is part of the IBS as well. The calcium independent proteins do not have a calcium ion and probably evolved from a common ancestor during which they lost the calcium binding site in order to develop a potentially larger IBS (Liu et al., 2007; Davies and Hew, 1990).

Type III

Type III, as seen in figure 1.1 C, is a globular protein with an even greater β-sheet content than type I and II. It contains eight short β-strands, forming three anti-parallel β-sheets, a loop, several turns, and a big hydrophobic core. In contrast to type I, it contains no repeat sequences and it is thought to adsorb to ice through hydrogen bonds that are present on the β-sheets (Sönnichsen et al., 1993).

Type IV

Type IV is glutamine-rich and, like type I, has an amphipathic α-helical structure. The hypothetical model as proposed in Deng and Laursen, 1998 consist out of four helix regions forming a helix bundle and is very similar to the apolipoprotein E3 which is shown in figure 1.1 D. It is thought that type IV AFPs diverged from these apolipoproteins, but their antifreeze activity did no increase much due to the lack of evolutionary selection. Both the fish type I AFP and type IV AFP are found in *Myoxocephalus octodecimspinosis*. Because of the former having a more efficient antifreeze activity, type IV never evolved as well as the other fish AFPs. Still, they are classified as a distinct subgroup due to their amphipathic helix-bundle and unusual high glutamine content (Deng and Laursen, 1998; Gauthier et al., 2008).

1.1.2 Arthropods antifreeze proteins

Around the same time it became clear that there should be a similar mechanism in other organisms, such as the cold enduring insects. However, initially only saccharides and alcohols were found to lower the freezing point in these insects. Later different molecules were found as well. These molecules were named “thermal hysteresis proteins” (THPs) due to the difference they cause between the freezing point and melting point of a solution (Davies, 2014). It was only in 1997 that the first AFP from *Tenebrio molitor* (TmAFP, figure 1.1 I) was successfully purified while retaining its activity, even though the function of these proteins was known for about twenty years (Graham et al., 1997). AFPs were found in different insects and arthropods as well, e.g., *Choristoneura fumiferana* (CfAFP, also known as spruce budworm antifreeze protein or SbwAFP, figure 1.1 G), spiders (Duman, 1979), and centipedes (Tursman et al., 1994).

After comparing the insect AFPs with the fish AFPs, it became clear that the activity of the former is about an order of magnitude larger than the activity of the latter (Tyshenko et al., 1997). Another difference is the composition of the amino acids and the orientation of homologous side chains. While the fish AFPs have a large content of alanine, the insect AFPs have a less significant alanine content. However, some insect AFPs have a more significant cysteine content, e.g., the AFP from *T. molitor* has a cysteine content around 12% while this content lays around 22% for the AFP from *Dendroides canadensis* (Duman and Horwath, 1983).

Also remarkable is the orientation of the amino acids' side chains present in the IBS. For insect AFPs this orientation is mostly parallel, creating a two-dimensional (2D) roster or ice-binding array with almost equal distances. This ice-binding array generates a better match with the structured ice lattice, allowing a more efficient binding with the ice.

Choristoneura fumiferana

C. fumiferana, commonly known as spruce budworm, faces temperatures up to -30 °. In order to survive these harsh circumstances, they trap themselves in a cocoon and use both colligative and non-colligative properties to avoid freezing. The non-colligative properties are based on producing AFPs and cryoprotective agents, and neutralising or removing potential ice nucleators. While for the colligative properties, the organisms will dehydrate in order to have a higher concentration of salts, saccharides, and alcohols to lower the freezing point.

In this organism, one AFP is used, but it has different isoforms. The use of different isoforms is more beneficial than providing different AFPs and in this case it might explain the dual character of the antifreeze activity and the neutralizing of potential ice nucleators. All the protein isoforms fold as highly regular left-handed β-helical structures with 15 amino acid per tandem repeat or turn. The cross-section has a triangular shape from which one side is the IBS consisting out of a 2D Thr-Xxx-Thr array.

The difference between the most common isoform, CfAFP-337, and the largest isoform, CfAFP-501 shown in figure 1.1 G, is the insertion of 31 amino acids which is equal to two repeats. Due to the two extra repeats in CfAFP-501 the IBS is extended and possesses an antifreeze activity that is three times as good as the CfAFP-337. This difference in activity, together with the difference in activity of the different isoforms from AFGPs, gave rise to the hypothesis that the antifreeze activity is correlated with not only the composition of the IBS, but with the length as well (Leinala et al., 2002a,b).

Rhagium inquisitor

Another insect AFP with a unique fold is the AFP from *Rhagium inquisitor*, shown in figure 1.1 J. It is probably the largest determined insect AFP and the largest AFP without incorporated ions. It folds as a compressed β-superhelix and consists out of two parallel β-sheets. The central part of the superhelix has a left-handed orientation and the two sheets consist out of six and seven strands placed on top of each other and kept together by interlacing amino acids. In contrast to most of the other insect AFPs, the cysteine content is relatively low and only one disulphide bridge and three hydrogen bonds are present. The IBS is located on the most flat side of the protein with a 2D threonine array. The consensus sequence for this is (Thr-Xxx)$_3$-Thr (Hakim et al., 2013).

Tenebrio molitor

In contrast with RiAFP, the AFP from *T. molitor* probably contains the smallest β-helical turn identified. The protein is shown in figure 1.1 I and is made up of seven repeats, each consisting of 12 amino acids, resulting in a very regular right-handed β-helix. The consensus sequence for this protein is Thr-Cys-Thr-Xxx-Ser-(Xxx)$_2$-Cys-(Xxx)$_2$-Ala-Xxx, with Cys being a cysteine and Ser a serine. Like most insect AFPs the IBS consists of a threonine array, while both cysteine residues point inwards and form eight disulphide bridges and numerous of inter-loop hydrogen bonds. The cross-section is a pseudo-rectangle that is divided into two parts, or channels, due to the disulphide bridges. The two channels are occupied by the conserved alanine and serine residues, creating more hydrogen bonds and space for internal bound water molecules, thus leaving no space for a hydrophobic core (Liou et al., 2000).

1.1.3 Other antifreeze proteins

As more AFPs were found in addition to the preceding classes, such as bacteria, fungi and plants (Duman and Olsen, 1993), it became clear that the overall fold and IBS sometimes show similarities, but that this is not always the case. Some of the AFPs found in protists, e.g., the sea-ice diatoms, are similar to some fish AFPs. However, the former possess a weaker antifreeze activity (Raymond, 2000). On the other hand, the AFP found in the bacteria *Marinomonas primoryensis* (MpAFP) differs in both structure and activity (Gilbert et al., 2005).

Lolium perenne

Antifreeze activity is found in some plants and vegetables as well. *Lolium perenne*, e.g., is a very common rye-grass and possesses antifreeze activity. Due to its low TH and higher than usual ice recrystallization inhibition (IRI) the protein is called an ice-binding protein

(IBP) and is shown in figure 1.1 E. The protein has eight tandem repeats, each containing 14 or 15 amino acids, and does not contain a cap structure. The consensus sequence is present twice in every repeat, giving rise to a two-fold rotation axis. The consensus sequence, Xxx-Xxx-Asn-Xxx-Val-Xxx-Gly, contains a non-conserved IBS with the sequence Xxx-Val-Xxx. Asn stands for asparagine, while Val and Gly stands for valine and glycine, respectively. The valine residue is projecting inwards while the two neighbouring residues are projecting outwards and form the IBS. The inward projecting asparagine residue in the consensus sequence gives rise to two internal asparagine ladders due to the fact that in each turn the consensus sequences occurs twice. These internal asparagine ladders give the protein an increased stability through the formed hydrogen bond network between the asparagine residues and the peptide backbone.

For a long time it was not sure whether the protein had one or two IBSs, due to the fact that it had two equally flat surfaces. Recently it has been confirmed that only one side possesses antifreeze activity and that the presence of the other flat surface may explain the unusual elevated IRI activity (Middleton et al., 2009).

The IBS of the LpIBP contains 16 residues (eight turns with two contributing residues), but in contrast with the insect AFPs only five out of the 16 residues are threonine residues. The remaining 11 residues are mostly serine or valine. For this reason, the IBS is not conserved and the outward projecting residues are marked as Xxx in the consensus sequence. Another big difference with the insect AFPs is the lack of the typical regular side chain orientation, which might result in an additional decrease of the TH (Middleton et al., 2009, 2012).

Marinomonas primoryensis

MpAFP is a bacterial AFP from *M. primoryensis* and is probably the largest AFP structure discovered yet. The full protein's mass is about 1.5 MDa, consists out of five distinct domains and is involved in cell adhesion. From these domains, only the second and fourth domain are highly repetitive and the fourth domain is the only domain containing antifreeze activity. Therefore, the MpAFP that is mentioned and shown in figure 1.1 H is only the fourth domain of the larger protein. This domain folds as a right-handed β-helix with 19 amino acids and a calcium ion per tandem repeat, with a few exceptions, such as an insertion of eight amino acids in one of the repeats. Each repeat contains three short β-strands and is made up of the following consensus sequence: Xxx-Gly-Thr-Gly-Asn-Asp-$(\text{Uuu})_4$-Gly-Gly-$(\text{Uuu})_3$-Gly-$(\text{Uuu})_3$ with Asp as aspartic acid and Uuu alternating between any amino acid and any hydrophobic amino acid. The IBS consists out of the first six residues of the consensus sequence and these residues are responsible for the calcium binding site as well. The aspartic acid residues will interconnect the different repeats while threonine and asparagine or aspartic acid residues are responsible for the long and flat IBS (Gilbert et al., 2005; Garnham et al., 2011).

Typhula ishikariensis

In order to use its pathogenic activity, the snow mold fungus *Typhula ishikariensis* uses AFPs to get to dormant plants that are covered under snow. The AFP from *T. ishikariensis* (TiAFP), shown in figure 1.1 F, is one of the irregular AFPs. Although it produces an isoform with a high TH activity, TiAFP is still classified as a moderate active AFP. Regardless of that one hyperactive isoform, the AFP does not show any similarity with insect AFPs, nor other hyperactive AFPs. TiAFP folds as a right-handed β-helix with six loops that each contain 18 to 27 amino acids and an adjacent α-helix next to the β-helix. In contrast with the insect AFPs, there are no cysteine residues present, nor does the IBS contain Thr-Xxx-Thr repetitions. The IBS is located on the flattest surface of the protein and shows a very high degree of irregularity when compared to the IBS from hyperactive AFPs. As mentioned before, the AFP does not contain any disulphide bridges, but it contains aromatic residues that might increase the stability of the hydrophobic core (Kondo et al., 2012).

1.2 History of the antifreeze hypothesis

To gain a better understanding about the functionality of AFPs, it is important to know the atomistic mechanism of lowering the freezing point of solutions and what the responsible key amino acids are. This improved understanding might help to give insight in possible ways to improve AFPs for commercial applications. However, it is difficult to define a universal antifreeze mechanism as the diversity of the discovered AFPs keeps increasing. Initially it was thought that all the AFPs possess a common mechanism, but since the discovery of the AFPs from the suborder Zoarcoidei, the existence of a universal antifreeze mechanism has been questioned. The doubt arose mainly because these AFPs are classified as type III AFPs, which lack sequence repeats and have an unusual high hydrophobic content as well (Sönnichsen et al., 1993).

In this section, the evolution of the first hypothesis to the current model will be explained.

1.2.1 The hindered growth of ice crystals

After observing scanned electron micrographs in the late seventies, it became clear that AFPs become a part of the freezing solvent due to an irreversible binding with ice (Raymond and DeVries, 1977). If the binding would not be irreversible, the ice crystals would continue to grow even in the presence of AFPs. As this is not the case, it was stated that the binding is irreversible (Knight et al., 1993).

It has been known as well that the growth rate of crystal surfaces depends on the presence of impurities and adsorbents. Thus, it is thought that large polymers with several repeating units have a larger influence on the growth rate than smaller polymers or polymers without repeating units. As such, the overall shape of crystals will most likely change in the presence of AFPs and will also depend on the AFP. This is why growing ice crystals with AFPs can be observed and allows differentiation (Buckley, 1952).

Raymond and DeVries observed that for the fish AFPs type I to III, the crystals grow in hexagonal bipyramids. They explained this phenomena by the Gibbs-Thomson effect or the Kelvin effect (Raymond and DeVries, 1977). Ice typically grows in hexagonal crystals, as can be seen in figure 1.2 on page 12. Once nucleates are formed, water molecules can be epitaxially incorporated into the basal planes of the hexagonal ice crystals. These ice crystals will grow primarily along the a-axis and with small steps along the c-axis (figure 1.2 D) (Knight, 1967; Franks, 1982).

The AFPs will bind to the basal plane of ice, where it forces ice to grow in between the other bound AFPs. The width of these openings depends on both the concentration of AFPs in the liquid phase and the size of an AFP. The latter has an important role in the propagation of ice as well. The steps along the c-axis are half the width of a bound AFP and it is only possible for ice to grow between two bound AFPs (Raymond, 1976).

Ice displays a radial growth and when it grows between the bound AFPs, an increase in the surface area will occur. As more AFPs are being adsorbed, the local curvature will increase even more, halting the ice growth because it is thermodynamically unfavourable to bind more free water molecules to the ice lattice. This is called the Gibbs-Thomson effect, where the ice growth will stop because the energy difference between the water molecules in the bulk and in the surface is too big. This arises from surface tension, due to the force that the bulk exerts on other water molecules to keep them at the surface. For this reason water is naturally found back in spherical volumes, to reduce the surface over volume ratio (Isgro et al., 2014).

The Gibbs-Thomson formula:

$$\Delta T = T_0 - T = \frac{2\,\Omega\,\gamma\,T_0}{\rho_{min}\,\Delta\,H_0} \tag{1.1}$$

With T_0 being the freezing temperature and T the melting temperature of water and ice. Ω is the molar volume of ice and ΔH_0 the latent heat of fusion necessary for a phase transition (Yeh and Feeney, 1996).

According to the Gibbs-Thomson formula 1.1, there are two ways to lower the local freezing point of a solution.

The first method to alter the freezing point is by changing the interfacial energy or surface tension (γ). In most cases a minimal surface is preferred such that the interfacial energy can be kept low. An increase in surface would therefore result in more molecules being part of the outer layer, which is unfavourable as they have a higher energy than the bulk molecules.

A second method is by halting the growth of ice crystals. Normally, when ice crystals start growing, their radius (ρ_{min}) can be interpreted as infinite. This infinite radius does not lead to any change in the freezing temperature (ΔT), as can be observed in normal water bodies. When AFPs are introduced to this solution, they will obstruct the growth of ice crystals. The ice that grows in between the different bound AFPs will thus have a smaller radius than the rest of the crystal. Their radius cannot be considered infinite and thus the change in freezing temperature does not equal zero, creating a thermal hysteresis.

The different types of AFPs have their own structures and sequences, as depicted in figure 1.1. This influences the binding to the ice crystals and, consequently, the change in freezing temperature as well, giving every AFP a specific temperature range in which they are active. When the temperature is above the melting point, there will be no ice formation and hence the AFP cannot have an effect. When the temperature is lowered below the AFP's temperature range, the formed ice crystals will grow primarily along either the c-axis or the a-axis, as mentioned briefly before in section 1.2.1. This growth will occur exceptionally fast and can be explained by the inhibition mechanisms and preferences of AFPs.

1.2.2 The first hypothesis

One year after the discovery of the first AFP, a mechanism by which the newly discovered AFP might function was proposed (DeVries et al., 1970). According to this hypothesis, AFPs could bring structure to water molecules in the liquid phase or at least be able to immobilize the water molecules. This would lead to the limitation of free water molecules that are able to be incorporated into the ice lattice. The hypothesis rose in popularity after it was discovered that the side chains of AFPs contain a higher than usual amount of hydroxyl and polar groups.

However, seven years later NMR experiments and isopiestic determinations revealed that AFPs can bind less water molecules than expected initially. When this was compared to other proteins in solution, it was noted that for similar sized proteins, AFPs bind only slightly more water molecules (Haschemeyer et al., 1977; Duman et al., 1980).

1.2.3 The ice lattice matching model

In the seventies, several site directed mutation experiments were performed on AFPs to observe the influence on the ice growth. During these experiments, regular spaced aspartic and glutamic acid residues were targeted, leading to a strong decrease in both the potential hydrogen bonding groups and the antifreeze activity. Because of this correlation and due to the similar distance between carboxyl groups of aspartic and glutamic acid residues, and the oxygen atoms from the basal plane of crystallized water, it was thought that these

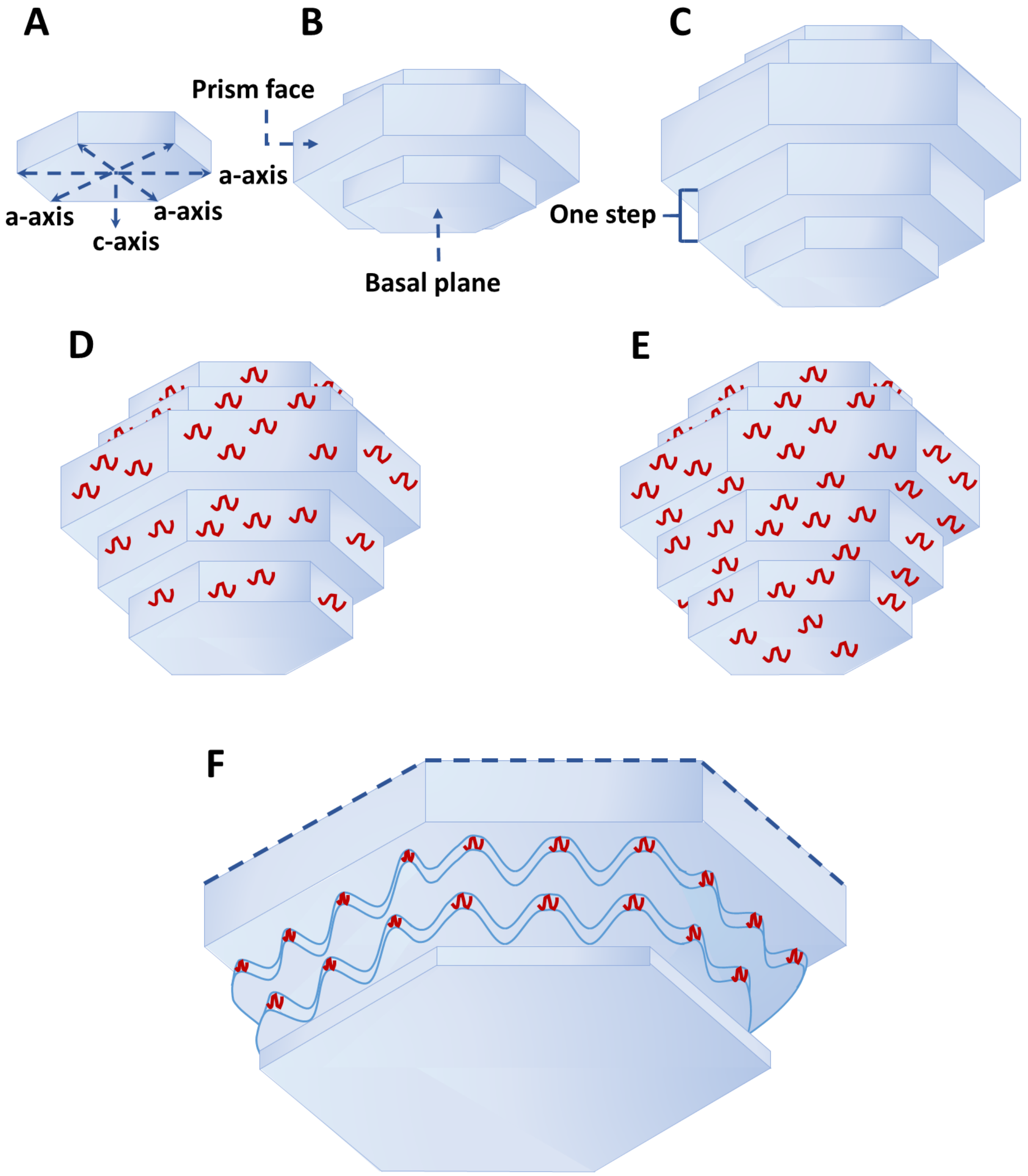


Figure 1.2: **Schematic figure depicting the growth of an ice crystal, the adsorption of AFPs and the inhibition of ice growth.** Panels **A-C** show the growth of an ice crystal. Ice crystals will primarily grow along the a-axis, while the growth along the c-axis will occur in small steps. Panel **D** and **E** show the adsorption preference of moderate active AFPs and hyperactive AFPs, respectively. The hyperactive AFPs will adsorb to the basal plane, as well as the prism faces, while the moderate active AFPs only adsorb to the prism faces. Panel **F** shows only a part of the ice crystal. Here the adsorption of AFPs limit the ice growth of the crystal. Ice can only grow between the different bound AFPs, leading to an increase in the local curvature. This energetically unfavourable process will lead to the halt of the ice front.

hydrogen bonds were responsible for the antifreeze activity (Shier et al., 1972; Duman and DeVries, 1976).

It was then suggested that AFPs would interact via hydrogen bonds at the water-ice interface due to the similar positions of the hydrogen bond donors and acceptors, hence the name ice lattice matching model. The model is shown in figure 1.3 A and 1.3 B (DeVries and Lin, 1977; DeVries, 1984).

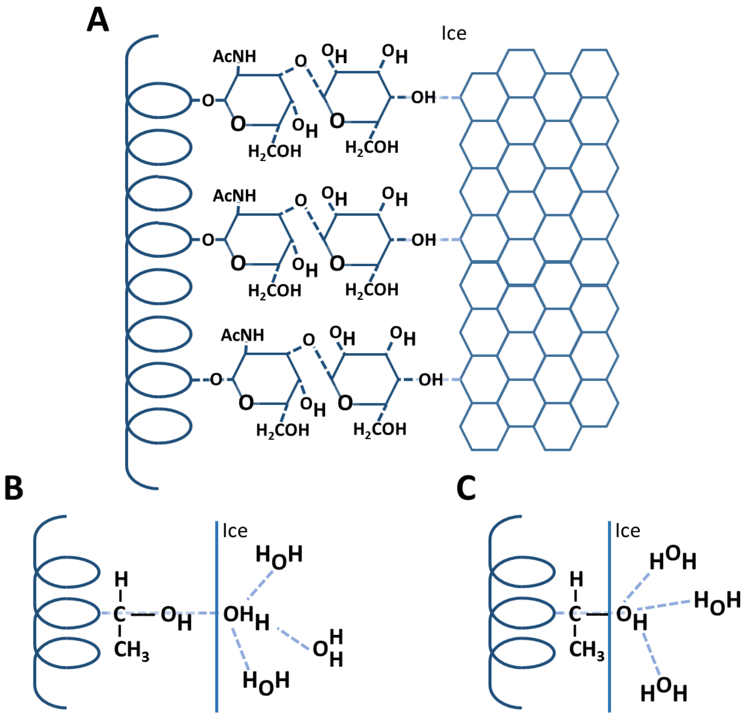


Figure 1.3: **Simplified representations of the ice lattice matching model (panels A and B) and the ice lattice occupying model (panel C).** Panel **(A)** is based on the fish AFGPs and shows the ice lattice matching model, where hydrogen bonds occur between the attached disaccharide and the ice. The disaccharides are attached to the threonines, which occurs as every third residue. The repeat distance between the hydroxyl groups matches with the repeat distance between the oxygen atoms in the ice lattice. Panels **B** and **C** are based on the fish type I AFPs, where adsorption occurs through the threonine residues. According to the ice lattice matching model, the hydroxyl group of threonine, is not incorporated into the ice lattice and thus can only form one hydrogen bond with the ice (panel **B**). The two remaining hydrogen bonds with the surrounding solvent are not shown here. For the ice lattice occupying model (panel **C**) the hydroxyl group is incorporated into the ice lattice and is able to form three hydrogen bonds with the ice, resulting in a practically irreversible binding (DeVries and Lin, 1977; Knight et al., 1993).

1.2.4 The ice lattice occupying model

In the early nineties a new model for the α-helical fish AFPs was postulated (Knight et al., 1993). This model described how the AFPs' polar groups interact with the ice, but it had three shortcomings.

The first drawback was the dynamic ice-water interface. It is more difficult to incorporate variables in a dynamic ice-water interface than in a static one. The transition from water to ice, or vice versa, does not occur suddenly in the dynamic interface, but will occur gradually. During this transition, three factors play a role, namely (i) the density profile, (ii) the molecular orientational and translational order, and (iii) the rate of molecular diffusion (Karim and Haymet, 1988). Because of these variables and difficulties, the dynamic interface model was converted to a static interface model. Two corrections were further applied to the static interface to make up for the errors that could otherwise occur. The first correction was that the AFP adsorption occurs quasi irreversible. The second correction was that the AFPs adsorb with a specific orientation to the crystals, namely with the IBS. If one or both corrections are not fulfilled, then ice crystals would be able to continue to grow, even in the presence of AFPs.

The second problem deals with the interaction between the AFPs and the ice. Originally, it was thought that these were connected by hydrogen bonds between the ice surface and the polar groups of the amino acids facing the ice. Since these AFPs are not incorporated in the ice, only one hydrogen bond would be possible between a polar group of the AFP and the ice, while two more hydrogen bonds are possible with the surrounding water. This method of bonding can be seen in figure 1.3 B and does not explain the irreversible adsorption of AFPs to ice. The one hydrogen bond between ice and the polar group can easily be broken and replaced with a hydrogen bond between the polar group and the surrounding water. Because of this manifestation, a different approach was used in which the polar groups from the AFPs are incorporated in the ice surface, as can be seen in figure 1.3 C. According to the new model, all three hydrogen bonds of each polar group need to be broken at the same time for an AFP to be released from the ice while the reformation of these hydrogen bonds with the ice should not occur. This three-fold increase of hydrogen bonds results in the irreversible adsorption of the AFPs to the ice, as the chance that a complete detachment of an AFP is as good as zero. The specificity of this approach will increase as well, since the polar groups need to be small enough to fit in the cavities of the ice crystal lattice surface and they must be able to bind tetrahedrally.

The final problem encountered by the lattice occupying model was the orientation of the AFPs. To solve this, molecular models were used to suggest specific bonding arrangements. Through these models it was noted that the bonding should take place via an adsorption plane with a 2D array of bonds instead of a linear array. This array allows the AFPs to bind to specific planes of the ice crystal, which contributes to the specific ice crystal shapes as briefly mentioned before (Knight et al., 1993).

1.2.5 Modification of the ice lattice occupying model

In the beginning of the 21^{st} century the lattice occupying model, which was based on the fish type I AFPs, was modified to allow a better fit for the AFP from the spruce budworm (SbwAFP). This resulted in three different mechanisms, which are shown in figure 1.4 (Leinala et al., 2002b). The improved model is still applicable for the α-helical type I AFPs, as it is based on the hydroxyl and methyl groups of threonine and type I AFPs contains both the alanine and threonine residues.

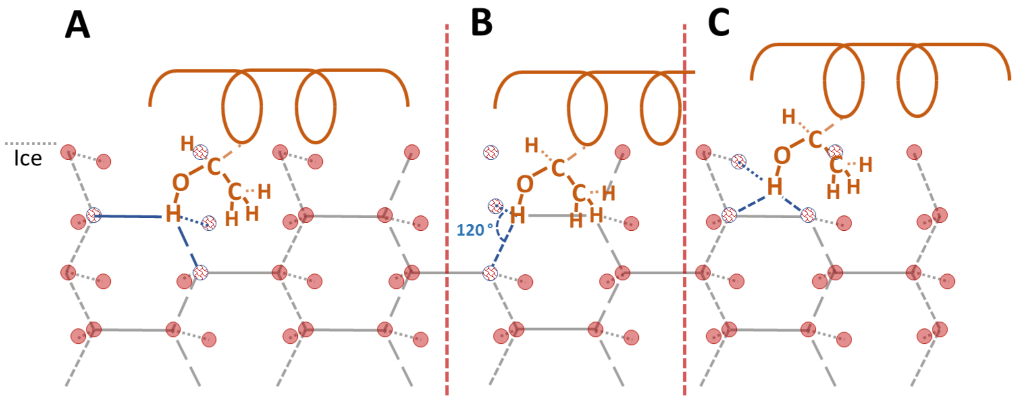


Figure 1.4: **The modified representations of the ice lattice occupying model, showing all three mechanisms.** The prism face of ice is shown, with red spheres representing the oxygen positions in the ice lattice. The grey lines represent the hydrogen bonds between the ice while the blue lines represent the newly formed hydrogen bonds with the threonine residues. The uninterrupted line represent this plane, while the dotted lines are either in front or behind this plane. Panel **(A)** shows the first mechanism, in which the methyl group fits in the neighbouring cavity and does not affect the three hydrogen bonds of the hydroxyl group. Panel **(B)** shows the second mechanism, where both the hydroxyl and methyl group occupy an oxygen position in the ice lattice. This results in the loss of one hydrogen bond for the hydroxyl group. The remaining two hydrogen bonds are altered and form a bond angle around 120 °. In the third mechanism, panel **(C)**, both hydroxyl and methyl groups are placed in a cavity. The methyl group does not disrupt any hydrogen bonds and the hydroxyl group is able to make three hydrogen bonds (Leinala et al., 2002b).

The first mechanism (figure 1.4 A) postulates that the hydroxyl group could be incorporated into the ice prism face, while the methyl group fits into a neighbouring cavity. The methyl group is only space filling in this case, while the hydroxyl group is able to form the three hydrogen bonds.

The second mechanism assumes that both groups will fit in a cavity on the ice prism face. Although by doing so the methyl group would take up the place of an oxygen position, resulting in one hydrogen bond less available for the hydroxyl group. The two remaining hydrogen bonds would be altered and have a bond angle around 120°, as shown in figure 1.4 B.

For the last mechanism (figure 1.4 C), both the hydroxyl and the methyl group would be incorporated indirectly. Here it is the methyl group that fits in a cavity, while the hydroxyl group fits in a small surface cavity located on the prism plane of the ice. Neither the hydroxyl group, nor the methyl group would occupy an oxygen position in the ice lattice. In contrast with the previous mechanism, the methyl group would not disrupt any hydrogen bonds and will take on a clathrate arrangement. The hydroxyl group is able to make the three hydrogen bonds, resulting in a very high surface complementarity.

1.2.6 Focus on hydrogen bonding

All the previous theories were based on hydrogen bonds between AFPs and ice or water, yet hydrogen bonds cannot explain the different affinity between the AFPs and ice or water, nor can it explain the tightness of the adsorption.

The reason for a different affinity for the crystalline ice state is not yet clear. As there is no significant difference between the liquid and the crystalline phase, it would seem more advantageous for the AFPs to remain in the liquid phase. For AFPs in the liquid phase, the rearrangement of water molecules and the orientating of the IBS can occur more easily and, in contrast to the crystalline phase, the AFPs do not have to overlap with a correct crystalline lattice position for hydrogen bonds to occur (Haymet et al., 1999).

For a while, there has been a disagreement about the hydrogen bonds. These disagreements were based on the amount of hydrogen bonds that occur with the solvent and their effect. It was thought that the type I AFP lacked sufficient hydrogen bonds for a tight binding to occur, hence the generation of the ice lattice occupying model where the hydroxyl groups would be incorporated in the ice lattice. This threefold increase in hydrogen bonds could explain the tightness and irreversibility of the binding, yet there is no agreement for this model as the threonine's hydroxyl groups do not protrude sufficiently enough from the IBS to be incorporated into the ice and form the three hydrogen bonds (Chao et al., 1997).

In the late nineties it became clear that hydrogen bonds might not be the main binding mechanism between AFPs and ice. During multiple experiments, the regularly spaced threonine residues of short fish type I AFPs were replaced with other amino acids, such as valine, serine, and alanine residues. As stated in section 1.2.4, it was thought that these threonine residues would be incorporated in the ice where multiple hydrogen bonds would occur. The replacement with serine residues barely showed any activity, even though the length of the side chain is about the same for both residues and has the potential to form hydrogen bonds. Whereas the replacement with valine residues, which contains a methyl group instead of a hydroxyl group, lost only about 15% of the antifreeze activity when it was compared to the wild type AFP (Chao et al., 1997; Zhang and Laursen, 1998; Haymet et al., 1998).

These experiments suggest that there are more forces contributing to the adsorption of AFPs besides the hydrogen bonds. The other forces should have at least the same magnitude as the hydrogen bonds, as the mutation of the IBS to amino acids with hydroxyl groups barely displayed any activity. Haymet, *et al.* discovered six different kinds of interactions that might contribute during the adsorption. These include the hydrogen bonding due to the nature of the solvent, hydrophobicity after comparing the wild type with the valine mutations and the relative size of the side chains when these were compared to the alanine and valine mutations (Haymet et al., 1999).

These different interactions will be incorporated in the following models.

1.2.7 Shape complimentary

As the contribution of other factors was proven, the IBS was observed for different characteristics. It became clear that the IBS is located at a more hydrophobic surface of an AFP. These surfaces are relatively flat and conserved when they were compared to the AFPs' hydrophilic surfaces and other AFPs (Davies et al., 2002). For these reasons it was thought that there must be a complimentary surface between the AFP and the ice lattice for attractive van der Waal forces.

The experiment of Gagne, *et al.* in 2003 illustrated that both the hydroxyl and methyl groups are important for the AFPs. In this experiment the threonine residues where modified to similar non-natural residues, which have a hydrogen instead of a hydroxyl group. When the antifreeze activity was compared to the wild type, it only showed 9% of the activity and thus suggesting that both the hydrogen bonds and van der Waal forces are important for the binding of AFPs with water molecules (Haymet et al., 2001; Gagne et al., 2003).

1.2.8 The main stabilization forces

In the beginning of the 21^{st} century it was thought that there are a number of interactions that play a role in the binding of AFPs with ice. The surface complementarity will lead to van der Waal interactions, while the contact between the IBS and the ice will generate a number of hydrogen bonds. Both interactions are needed for the antifreeze activity.

Once an AFP is dissolved in water, water molecules will associate with it. This is done by forming hydrogen bonds with the protein backbone and side chains. If the AFP interacts with the ice lattice, the IBS with its surrounding hydrophobic residues will be oriented towards the ice lattice. This results in the release of bound water molecules and less hydrophobic residues that are exposed to the solvent. The AFP's first solvation layer will be less constrained and more water molecules will be present as free water molecules, increasing the entropy of the solution which is also known as the hydrophobic effect (Davies et al., 2002; Scotter et al., 2006).

Some of the previous described mechanisms, such as the mechanism depicted in figure 1.4 B, does not show a perfect fit between the AFP and the ice lattice. However, there is a partial compensation for the loss of hydrogen bonds by hydrophobic interactions and van der Waal forces (Davies et al., 2002).

1.2.9 Clathrate water molecules

More recently Molecular Dynamics simulations (MD) gave a better insight in the hydrophobic surface of the IBS and how it might be able to arrange nearby water molecules, resulting in the bound water molecules as described in section 1.2.8. These bound water molecules are the water molecules that associate with the protein backbone and side chains via hydrogen bonds. By doing this, they form a lattice around the AFP and are called clathrate water molecules. In contrast to what was thought earlier, the clathrate water molecules would mimic the ice planes and thus facilitate the adsorption of the AFP to ice. Releasing the clathrate water molecules that are close to the IBS into the solvent would reduce the amount of clathrate water molecules being able to recognize the ice, and would thus be an unfavourable process. However, the idea of releasing water molecules for an entropic gain can still be applied to the excess of water molecules surrounding the protein and the IBS, but a description of how the hydrophobic surface is able to arrange the clathrate water molecules to mimic the ice planes was not possible through MD simulations. Nonetheless, a hypothesis concerning the formation of clathrate water molecules was already formed fifty years ago (Garnham et al., 2011). This hypothesis postulates that pentagonal rings of water molecules are present between the hydrophobic side chains of α-helices or β-sheets, only if these are separated by one water layer (Scheraga et al., 1962). The pentagonal water rings were later observed in crystal structures as well (PDB: 1CRN) (Teeter, 1984).

Clathrate water molecules can be found in all of the AFPs, although a difference can be observed between hyperactive and moderately active AFPs. The difference in activity is described further in section 1.3.2 on page 21. Both kind of AFPs are able to arrange water molecules to mimic the prism faces of ice and thus allowing irreversible adsorption to these faces. However, only the hyperactive AFPs are capable in adsorbing to the basal plane in addition to the prism faces, via a small difference in the orientation of the arranged water molecules.

The mechanism of clathrate water molecules explains the correlation between the antifreeze activity and the concentration of AFPs as well. Only a small fraction of AFPs would be able to adsorb to the ice lattice; the other fraction of AFPs would have an improper orientation of clathrate water molecules and side chains or does not have enough clathrate water molecules due to an exchange with the solvent, and both would prevent an immediate adsorption. It is possible to increase the fraction that is able to adsorb to the ice lattice, and thus the TH, by increasing the AFP concentration or the annealing time.

In contrast to earlier beliefs, the IBS will use hydrogen bonds to anchor the clathrate water molecules. These water molecules will mimic the ice planes and facilitate the adsorption, while the hydrophobic effect is needed to arrange the water molecules for a correct mimicry. Instead of being driven by entropic gains, the adsorption process is now primarily driven by enthalpic gains (Garnham et al., 2011).

1.2.10 Conclusion

By connecting the dots of what was already known and what was observed after studying the newly discovered AFPs, a first logical hypothesis was made. Conclusions were made about the adsorption of AFPs by observing the growth of ice crystals, leading to the hypothesis that adsorption had to occur as good as irreversible and that the ice growth was halted due to the Gibbs-Thomson effect, which is still agreed on today.

As the first hypothesis was merely based on what was observed, and also included the fact that AFPs could bind more water molecules than other proteins, it became clear rather soon that they were mistaken. Nonetheless, less than ten years later the most common ice-binding residues were discovered, leading to new hypothesises.

Again, the new hypothesises were based on what was discovered and it was thought, through logical reasoning, that the main mechanisms for AFPs were the hydrogen bonds. Yet, this hypothesis could not explain everything, such as the preference for water and the tightness of the bond, so the next step was trying to improve the current hypothesis.

By focusing on the hydrogen bonds scientists were led to a wrong track and it was only in the nineties that they discovered that there are more forces contributing to the adsorption of AFPs. This led to the current hypothesis that water molecules will associate with the protein, by forming hydrogen bonds with the protein backbone and side chains, and that the hydrophobic protruding residues are responsible for arranging the water molecules. This will result in clathrate water molecules that mimic the crystal lattice of ice, facilitating the adsorption of the AFP to the ice lattice.

1.3 Function of an antifreeze protein

1.3.1 Acquired qualities

One of the possible ways to distinguish between AFPs, is to categorize them based on the acquired qualities of the host organism. These can either be to avoid extracellular ice formation or to tolerate it while avoiding the intracellular ice formation (Atici and Nalbantoglu, 2003). They are then called freeze tolerant organisms or freeze avoiding organisms, with the former being able to survive the freezing of body fluids while the latter tries to prevent this (Duman, 2001). The AFPs that occur in freeze tolerant organisms are sometimes called ice-binding proteins (IBP).

Freeze avoidance

The freeze avoiding organisms need to be able to withstand the coldest temperatures they encounter. Naturally, these temperatures will depend on the habitat where the organisms reside, but most AFP-producing organisms may easily face temperatures of -20 °C. Without any mechanisms these organisms would freeze to death, but by preventing the occurrence of intracellular ice formation, they would be able to withstand these temperatures. One of the possible ways is to remove all the ice nucleators from the body, or counter effect them with AFPs. Another possibility is by supercooling the body fluids, this can be done by either non-colligative properties like antifreeze proteins, colligative properties like small molecules, or by a combination of both.

Freeze tolerant

Once the temperature exceeds the depressed freezing point, the ice crystals will grow very fast and recrystallization can occur. Recrystallization is the rejoining of smaller ice crystals to form a bigger ice crystal. It mostly occurs during freeze-thaw cycles when the temperature is fluctuating around the freezing temperature. The growth and recrystallization of ice will inflict extensive damage to the tissue because of the sudden increase in volume. It is said that the ice crystals ‘burst’ because the crystals grow in an explosive manner (Davies, 2014).

For organisms it is more beneficial to be freeze tolerant if they are exposed to less extreme sub-freezing conditions, for smaller periods of time or if the temperature is continuously fluctuating around the freezing point, as the latter results in freeze-thaw cycles and these cycles are beneficial for ice recrystallization. For these reasons it is thought that freeze tolerant organisms evolved to yield low antifreeze activity to prevent extensive bursting from happening. Low antifreeze activity is characterized by a low TH, which causes a small depression of the freezing point with less intensive burst patterns. Yet these AFPs will have a higher than usual IRI, which prevents the recrystallization and thus is able to prevent more damage in these circumstances. Because of this low antifreeze activity, these proteins are referred to as ice-binding proteins (IBPs). At the moment it is not certain how the mechanisms for IRI works, but it is thought that the IBPs will display a similar mechanism as the AFPs for the depression of the freezing point (Yu, 2010). However, experiments showed that for the inhibition of ice recrystallization, a lower IBP concentration is required (Duman, 2001).

Another possibility for the freeze tolerant organisms is to induce the extracellular ice formation by using ice nucleators. This induction will cause water to be withdrawn from the cytoplasm to the growing extracellular ice crystals. The cytoplasm will become more concentrated, resulting in a depression of the freezing point by colligative properties. Although cell death can occur when the cytoplasm will become too hydrated (Middleton et al., 2012). Probably all plants, and some arthropods as well, are freeze tolerant organisms (Duman and Horwath, 1983; Atici and Nalbantoglu, 2003).

1.3.2 Activity

As mentioned before in section 1.2.9, it is possible to distinguish between different categories of AFPs based on their activity, in particular the hyperactive AFPs and the moderate active AFPs: the former possess a TH that is about one or two orders of magnitude higher than the TH of fish AFPs. The other AFPs with a TH similar to the fish AFPs are categorised as moderate active AFPs (Scotter et al., 2006; Drori et al., 2014). The difference in activity probably arose due to the different environments organisms had to adapt to. Fish, e.g., only have to depress the freezing point of their body fluids by one degree to survive, while insects sometimes have to face temperatures around −20 °C and thus need a different strategy (DeVries et al., 1970; Sanders, 1991; Cheng, 1998). The difference is not only found in the activity, but in the crystal adsorption and burst pattern as well.

An explanation for the difference in TH might be found in the adsorption model. As mentioned briefly in section 1.2.9, the hyperactive AFPs are capable to adsorb to the basal plane in addition to the different prism faces of an ice crystal, as can be seen in figure 1.2 E on page 12. The slightly different arrangement of their clathrate water molecules allows them to adopt a lattice structure similar to the ice lattice from the basal planes and the prism faces.

The difference in adsorption preferences lead to different burst patterns when the temperature exceeds the TH (Scotter et al., 2006). The burst patterns for both the hyperactive AFPs and the moderate active AFPs can be seen in figure 1.5.

Hyperactive AFPs

The best known hyperactive AFPs originate from insects. They consist of a conserved 2D array of exposed threonine amino acids separated by an inward pointing amino acid

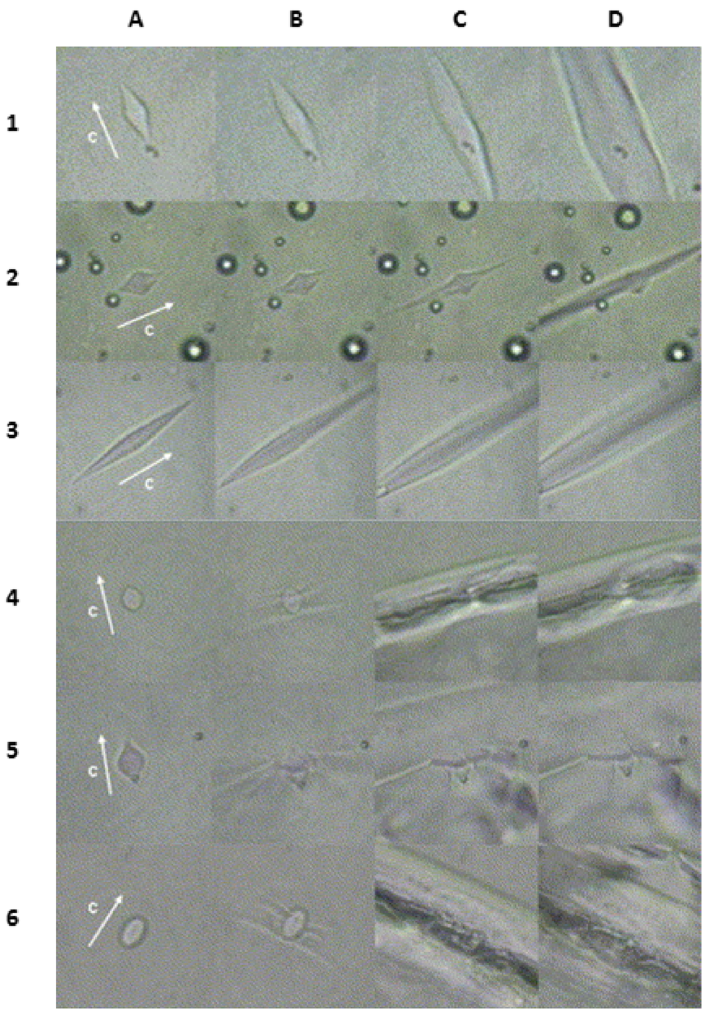


Figure 1.5: **Burst patterns of ice crystals in the presence of hyperactive AFPs and moderate active AFPs from Scotter et al., 2006, used with permission.** Series **1-3** show the burst pattern of moderate active AFPs, whereas series **4-6** show the burst pattern of hyperactive AFPs. Multiple images were taken at a 0.4 s interval during the burst, shown as frames **A-D**. The c-axis direction is indicated with a white arrow. The c-axis is considered as the weak axis of an ice crystal for moderate active AFPs, as these AFPs are unable to adsorb to the basal planes of ice, resulting in a rod-like burst pattern once the temperature is decreased past the temperature range of an AFP. For the hyperactive AFP, the weak axis is the a-axis. So once the temperature is decreased, an elongation will occur, conform to the prism face. The moderate active AFP series were obtained from, respectively, a type I AFP from *P. americanus* and *Myoxocephalus scorpius*, and a recombinant type II AFP from *C. harengus*. The hyperactive AFP series were obtained from an insect AFP from *T. molitor*, a hyperactive type I AFP from *P. americanus*, and a bacterial AFP from *M. primoryensis*, respectively. Experiments were conducted in the presence of 100 mM ammonium bicarbonate with a pH of 7.9, except for MpAFP which was conducted in 20 mM Tris-HCl with a pH of 7.5 and 10 mM $CaCl_2$.

(Thr-Xxx-Thr). This conserved array is found back in other hyperactive AFPs as well, although it is not required in order to be a hyperactive AFP. The MpAFP, e.g., has a conserved array consisting out of threonine and asparagine amino acids (Thr-Xxx-Asn), whereas the AFP from a snow flea has a glycine content of about 45% and will probably use a different approach (Garnham et al., 2011; Mok et al., 2010). Yet all of these will be able to adsorb to the basal planes, as well as the prism faces of ice.

Another interesting characteristic of hyperactive AFPs is the exceptional regular side chain orientation of the IBS when they are compared to other side chains or to the side chains from the IBS of moderate active AFPs. It allows an arrangement of clathrate water molecules with a higher similarity to the ice lattice.

It is thought that the a-axis of ice crystals are the vulnerable points for hyperactive AFPs. Thus once the temperature is lowered past the AFP's temperature range, a fast elongation conform to the prism face will occur, resulting in an increase of the hexagonal shape as a typical burst pattern and can be seen in figure 1.5 4 - 6 (Scotter et al., 2006).

Moderate active AFPs

The other AFPs are categorised as moderate active AFPs. These do not only include most of the fish AFPs, but also the AFPs derived from plants and other organisms. The moderate active AFPs are unable to adsorb to the basal planes of ice crystals and will only prevent the growth of ice at the prism faces of ice crystals. Growth of the ice crystal can be observed along the c-axis when the temperature exceeds the TH, resulting in a rod-like burst pattern as can be seen in figure 1.5 1 - 3.

If there is no AFP present, growth will primarily occur at the prism faces of the ice crystals, generating a relatively flat ice sheet (Scotter et al., 2006).

1.4 Applications for antifreeze proteins

1.4.1 Current situation

A lot of questions were asked after the discovery of the first AFP and the rising interest in its putative applications. Not only was there an academic interest in the structure and mechanism, but a need to explain how AFPs evolved as well. During the years following the discovery, multiple hypotheses were formed and evaluated, leading to a better understanding of AFPs for putative applications for a wide variety of areas.

1.4.2 Commercial applications

The unique quality of repressing ice crystal formation gives AFPs a great potential value. Not only has it applications in the food industry or modification of crop plants, but as well in cryopreservation, cryosurgery, and more.

Nowadays, the use of AFPs as additions for the preservation of food, organs or cells is still a complicated procedure. A change in concentration or in the type of AFP that is used may conflict damage, rather than protecting them against ice induced damage (Griffith and Ewart, 1995). Yet if applied correctly the use of AFPs may facilitate many procedures.

In this section, a few applications will be described further.

Cryopreservation

The cryopreservation of cells is not only dependent on the concentration and type of AFPs, but on the cell type that is used as well. The survival rate and functions of cells would increase after rapid freezing, although for some cells the increase of the survival rate has a limited effect.

To prove the impact on the organs' function, rat hearts were frozen and thawed before they were flushed with cardioplegic solution and assessed on cardiac output. During this experiment different concentrations of AFGPs were used and the results were compared with the untreated hearts. The different concentrations had an influence on the shape of the ice crystals and on both the TH and IRI. Nevertheless, the storage of rat hearts was not enhanced with a change in concentration and even though less ice formation was observed when AFGPs were used, damage would still occur (Wang et al., 1994).

Not all the different AFPs are fit as cryopreservatives, some will cause interference or damage to the cells or membranes. However, via AFGPs it is possible to store human platelets for 21 days at 4 °C. Before this addition it was not possible to store them at temperatures lower than 18 °C due to the cold activation of platelets. This activation would introduce a change in the platelets' shape and cause them to release their content, resulting in a loss of function (Tablin et al., 1996).

In contrast to the human platelets, red blood cells will become hemolyzed during freeze-thaw cycles in the presence of AFGPs or relatively big AFPs. However, type I to III AFPs were able to reduce hemolysis when they were applied in a micromolar quantity. When the concentrations were elevated to a millimolar quantity, the hemolysis would be enhanced. A correlation between the TH and the ability to protect the red blood cells was seen after observing the crystal growth of the type III AFP mutant through cryo-electron microscopy. Although the surviving fraction of the red blood cells is more dependent on the IRI than on the TH (Griffith and Ewart, 1995; Chao et al., 1996).

Cryosurgery

During cryosurgery, skin lesions with abnormal tissue are subjected to extremely cold temperatures via a cryoprobe. This cryoprobe will cause cell death to the surrounding tissue via either a direct change in chemical and physical environment or due to the change in biological response. Depending on the assessment technique, the immediate cell death, the delayed cell death or both can be measured, but possible interference might occur as well. This interference might be due to cell death prior to freezing and is dependent on the cell tissue and previous conducted treatments, which may result in different temperature thresholds.

The objective of cryosurgery is to treat the abnormal tissue, while trying to inflict as little damage as possible to the normal, healthy, tissue. In some way this is the reverse of the evolutionary goal of AFPs, as one of the mechanisms of cryosurgery is the formation of intracellular ice, which is avoided in organisms. The actual contribution of AFPs is not yet known in detail and most knowledge is obtained from using animal models. Although it is known that AFPs will cause an increased amount of damage due to the spicular needle-shaped ice crystals and that this will elevate the amount of cell deaths.

The damage that occurs is primarily related to the freezing and is mainly dependent on two factors, namely which tissue is used and the freezing rate. Depending on the tissue, the threshold temperatures at which the cells start to die will vary. The freezing rate is important as well, as slow freezing rates favour the ice formation outside the cells, resulting in the transport of water molecules to the extracellular space. The cell will become more concentrated and will, eventually, become hydrated and thus through this way more cells will die at a delayed time. However, when the freezing rates are increased, the formation of ice will occur intracellular, causing more damage immediately. Other factors that contribute to the cryo-damage are the time and temperature the cells are kept at.

The AFPs will then be applied in the bloodstream, near the tumour. By doing this, the necessary cooling rate for the formation of intracellular ice can be lowered while enhancing the cryoinjury in the direct surroundings of the probe (Muldrew et al., 2001; Koushafar et al., 1997).

Food industry

During the food preservation, food will be frozen, stored, transported, and stored again before its consumption. During this time it is subjected to several freeze-thaw cycles, encouraging the recrystallization of ice crystals and having an effect on both the quality and the texture of frozen food. The recrystallization of ice crystals can cause cellular damage, resulting in a loss of nutrients and a reduced water content. The addition of carefully chosen AFPs can prevent or decrease the reduced quality and texture by preventing the recrystallization (Griffith and Ewart, 1995).

Unilever

Unilever introduced a genetically modified type III AFP from *Macrozoarces americanus* into their ice cream. They claim that by doing this, it is possible to reduce the content of fat and sugars while increasing the amount of fruit. In addition, it would not change the texture of ice cream, even though the fat content is reduced, and it would improve the quality of the ice after freeze-thaw cycles. Approval for the use of the modified AFP in Europe as an additive for ice cream was given in 2009, while for other countries, such as the USA, it was already on the market since 2003 [1,2,3].

1.5 Repeat structure

Many of the discovered AFPs have a β-helical architecture. This architecture gives rise to a periodicity of a specific number of amino acids, which also contains the IBS, that is translationally repeated throughout the protein. The periodicity generates repeats, loops or turns, and translational symmetry. Not all the AFPs with a β-helical structure contain repeats and in some cases there can be multiple repeats present in one turn, hence the distinction between repeats and turns.

As mentioned before, AFPs tend to have isoforms. Most of the times, the isoforms will differ in the amino acid sequence. However, it occurs as well that an isoform contains a loop insertion or deletion, altering the translational symmetry.

1.5.1 The evolutionary origin

A common accepted hypothesis postulates that multiple domains, or in this case, repeats, arose from one domain, or repeat, when they share a high similarity in both the sequence, structure, and function (Emmert-Streib, 2012). Due to genetic drift and evolutionary pressure, the multiple domains and repeats with the same sequence underwent alterations while retaining the structural and functional similarity, as can be seen in figure 1.6.

It is possible that over the course of years the original sequence has diversified too much to talk about repeats, even though the repetitive ice-binding residues are still present. This can be seen in the AFP from spruce budworm, where there is no consensus sequence present due to the high variation between isoforms (Doucet et al., 2000). However, these isoforms were interesting for another reason as well, namely, the activity. When these

[1] Unilever (n.d.), “Cool ice cream innovations.” https://www.unileverusa.com/about/innovation/product-innovations/cool-ice-cream-innovations. Accessed: 2015-11-21.

[2] Cummins, J., M.-W. Ho, and M. Hooper (2006), “GM protein in ice cream.” http://www.i-sis.org.uk/GMPIIC.php. Accessed: 2015-11-21.

[3] Sander Voormolen (2009), “In stores next summer: ice cream that doesn’t melt.” http://vorige.nrc.nl//international/Features/article2339979.ece. Accessed: 2016-05-15.

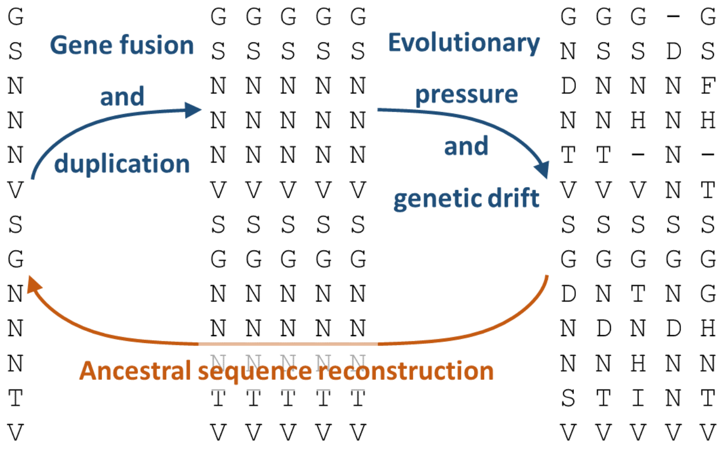

Gene fusion
and
duplication
Evolutionary
pressure
and
genetic drift
Ancestral sequence reconstruction
G S N N N V S G N N N T V
G S N N N V S G N N N T V
G S N N N V S G N N N T V
G S N N N V S G N N N T V
G S N N N V S G N N N T V
G S N N N V S G N N N T V
G N D N T V S G D N N S V
G S N N T V S G N D N T V
G S N H - V S G T N H I V
- D N N N N S G N D N T V
G S F H - T S G G H N T V

Figure 1.6: **A representation of both the evolution of a sequence (shown in blue) and the ancestral sequence reconstruction (shown in orange).** A tandem repeat of sequences probably arose due to gene fusion and duplication. Over the following years this sequence was prone to evolutionary pressure and genetic drift, causing changes in the sequence. Through ancestral sequence reconstruction, also known as ARS, a repeat sequence is derived, which might possibly be the sequence that multiple repeats once had in common.

isoforms were observed, a correlation between the amount of repeats and the antifreeze activity was found, resulting in a new hypothesis (Li and Jin, 2004).

Via ancestral sequence reconstruction, part of the REverse Engineering Evolutionary (RE$_3$Volutionary) protein design, it is possible to reconstruct putative ancestral sequences. It is possible that most domains or repeats once shared these ancestral sequences before genetic drift occurred.

Via RE$_3$Volutionary protein design it is possible to superimpose the putative ancestral sequences on the different segments to obtain identical domains and repeats, as if gene fusion and duplication had just occurred. This method has been experimentally validated with the pseudosymmetric sensor domain of a *Mycobacterium tuberculosis* protein kinase (PDB: 1RWL), which gave rise to the Pizza protein family (Voet et al., 2014).

1.6 Pizza proteins

The Pizza protein family is a successful result of reverse engineering evolution according to the duplication and functionary. Via RE$_3$Volutionary protein design it was possible to obtain proteins that are highly thermostable and symmetrical. The original protein, the pseudosymmetric sensor domain of a *M. tuberculosis* protein kinase, is part of the β-propeller protein family, which contains strong suggestions that gene duplication and fusion occurred during their evolution due to the high similarity between the different blades. This protein family, and more specifically the sensor domain of a *M. tuberculosis* protein kinase, was then chosen as a template for the Pizza proteins.

The Pizza protein family contains multiple proteins with a different amount of identical blades per protein. Pizza6, e.g., contains six identical blades and is found back in a monomeric form while Pizza2 and Pizza3 only have two or three blades, respectively, but will form trimers and dimers to regain a stable six-fold structure (Voet et al., 2014).

The highly symmetrical Pizza proteins were modified further to give these proteins a specific function. This was done by rationally introducing metal ion binding sites through point mutations. Via the incorporation of these binding sites it is feasible to grow metallic clusters or quantum dots, which are known as biomineralisation, or adsorb to specific metal or inorganic surfaces (Voet et al., 2015).

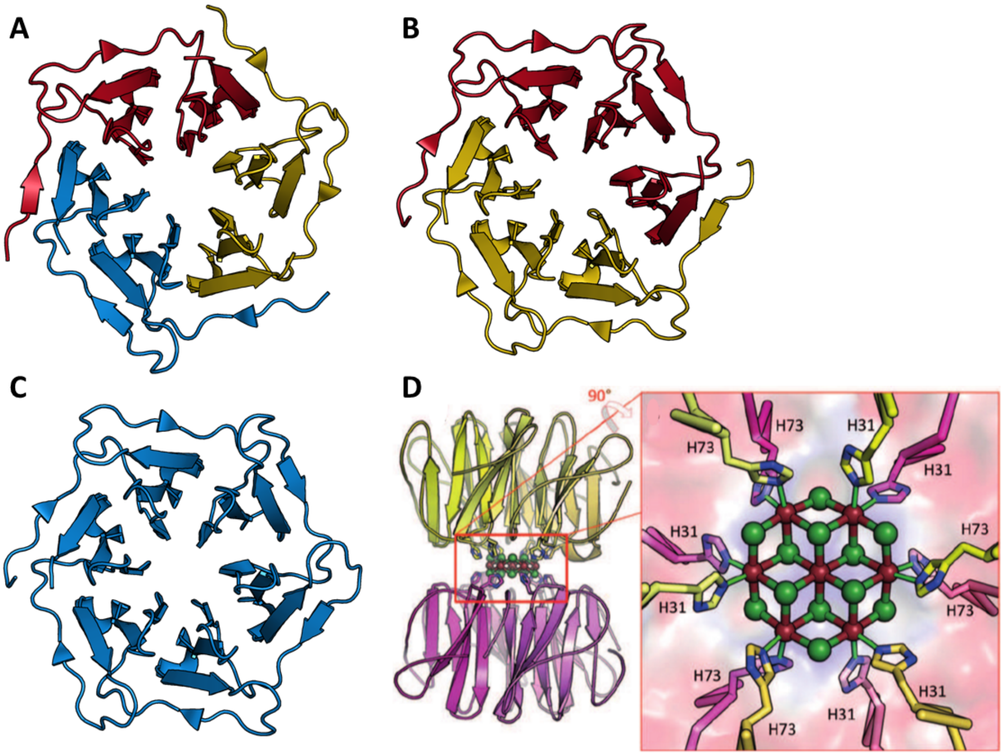


Figure 1.7: **Cartoon representation of the Pizza proteins**. Pizza2, Pizza3, and Pizza6 are shown in cartoon **A**, **B**, and **C**, respectively. Every colour indicates a different monomer, as Pizza2 needs three monomers to generate a stable six-fold structure. Panel **D**, from Voet et al. (2015), used with permission, shows the successful biomineralisation of a $CdCl_2$ nanocrystal templated by Pizza proteins. Two Pizza2 trimers are placed on top of each other, allowing them to template the nanocrystal. The cadmium and chloride ions are shown as brown and green spheres, respectively.

1.6.1 Biomineralisation

The crystallization process of inorganic material is called biomineralisation. This process contains two main steps, namely the nucleation and the crystal growth. The latter tends to be autocatalytically, which makes nucleation the crucial step.

Many organisms will use biomineralisation as a way to strengthen existing tissue. Via biomineralisation they can gain mechanical strength or additional protection. E.g., the production of bone by osteoblasts, and the calcification process resulting in the building of shells. In some cases organic material can be used as well, e.g., cellulose and all its derivatives (Ochlal, 1991).

In more recent years modified proteins were used successfully to study the results and effects of biomineralisation, as well as the possible applications that biomineralisation can have. An example of this is the modified Pizza protein family that has been used successfully to generate a cadmium chloride nanocrystal (Voet et al., 2015).

1.7 Redesign of AFPs

It should then be possible to redesign certain AFPs to create perfect repeating proteins or Crystal Lattice Interacting Protein Scaffolds (CLIPS), by following the principles of RE$_3$Volutionary protein design. As has been done before with the Pizza proteins, which resulted in proteins from which the number of repeats can be varied easily. The translational repeating proteins could be interesting as scaffold proteins for crystal binding applications, to induce the biomineralisation like the modified Pizza proteins or to coordinate metallic arrays and ice lattice structures as they are based on AFPs.

The fact that AFPs contain translational symmetry should be an advantage, as this gives the opportunity to tune the length to crystallographic lattices by varying the number of repeats.

2

Aims

The development of Crystal Lattice Interacting Protein Scaffolds, CLIPS, would be very interesting for the synthetic biotechnology and nano-biotechnology, such as binding or coordinating crystal lattices, as the length of these proteins can easily be modulated. Therefore, during this thesis the main goal was to develop a procedure to computationally design length-variable CLIPS by using antifreeze proteins (AFPs) as templates. The AFPs are ideal to serve as template proteins for this project, as some of these have a linear repetitive structure and are able to coordinate an ice crystal lattice.

The development of CLIPS can be divided into two parts. The first goal is to develop a protein design procedure which may be applied on multiple template proteins to obtain CLIPS. During this procedure, ancestral sequences will be derived via a so-called $\mathrm{RE_3}$Volutionary protein design method. The procedure will also create perfect translational symmetric backbone models on which the ancestral sequences will be mapped. This allows us to compute the properties of each sequence and from the alignment with the backbone. These properties, such as the energy value, RMSD, and the amount of differences present between multiple sequences, can be used to obtain a ranking to find the most favourable sequence.

Once the method is developed and has yielded a promising repeat protein sequence, we can try to experimentally validate the procedure. If unsuccessful, an improved design strategy may be attempted. During this part, the gene sequence of the full protein is ordered from a vendor and will be brought to expression as multiple proteins varying in length. Thus, the goal of the second part is to obtain CLIPS.

An additional goal could be set by analysing the thermal hysteresis (TH) of the redesigned proteins in relation to the number of repeats and the original AFP, as the CLIPS are based on AFPs. However, this additional goal is only possible if the CLIPS are successfully brought to expression as the current success-rates of computational protein design are low, as the availably algorithms are not yet fully mature. Still, the main goal of this thesis is to develop a functioning procedure.

3

Materials and Methods

3.1 Computational methods

3.1.1 PyMOL

PyMOL was initially written by Warren L. DeLano, using the Python programming language hence the ´Py' part in the name. It was developed as an open-source and user sponsored system for the visualization of molecular systems. Currently it is commercially available via Schrödinger, inc, but is free for academic usage. With PyMOL, it is possible to display three-dimensional (3D) molecular structures and it allows the user to apply changes to the structures, as well as to produce high quality images. The changes can be made both though the graphical interface and the command line (Schrödinger, LLC, 2015).

3.1.2 Strap

Strap stands for 'interactive structure based sequences alignment program'. The 'free-for-use' computer program is based on a Drag-and-Drop system where protein alignments can be made based on the sequence, 2D structure or 3D structure. The input can come directly from a database or it can be in ASCII text format, while the output can be exported for further use with text processors (Gille and Frömmel, 2001).

3.1.3 FastML server

The FastML server is a bioinformatics tool that can be used for the reconstruction of ancestral sequences. The input required for the reconstruction is a phylogenetic tree and a multiple sequence alignment from either codons, nucleotides or protein sequences. The phylogenetic tree should have a Newick format while the multiple sequence alignment should have a Fasta format. The tool will use the relation between the different sequences according to the positions in the phylogenetic tree and as a result of several algorithms, the most probable ancestral sequence per node will be reconstructed. At the same time possible deletions and insertions, also known as indels, are taken into account.

The first part is the reconstruction of characters and can be done with two different methods. These are the joint and the marginal reconstruction methods (Pupko et al., 2000, 2002). It is not necessary that both types of reconstruction give the same sequence per node, since the former method looks for all possible node sequences while the latter only looks for the most probable sequence.

During the second part possible indels are sought by scanning the multiple sequence alignments. This is done by giving every character a binary code for the absence (0) or presence (1) of indels. Afterwards, the indels need to be reconstructed for every node present in the phylogenetic tree. The idea behind this is that deletions and insertions occur over time and are present in the future sequences, which are susceptible for alterations as well.

Both parts are brought together to generate an improved ancestral sequence (Ashkenazy et al., 2012).

3.1.4 Brugel

Brugel is a multi-task system and allows the usage of procedures and single commands to apply manipulations and calculations on proteins and DNA structures. The applied changes can easily be assessed via visualisation tools, such as PyMOL (Delhaise et al., 1988).

When using Brugel, one general script was written for all the manipulations, but it was altered for each specific protein model. This script is listed in the appendix, see appendix B.3.1.

The first part of the script contains the used library, models, and procedure to assemble repeats containing the first and last few residues. In order to proceed with the script, the start and end residue number of the repeats and cap structures must be known. For this reason, the ensembles ‘repeatx’, ‘start’, and ‘end’ is used in the script to indicate which repeat is mentioned and what the start and end residue number of the referred repeat is. The procedure ‘create_rep’ is used to make segments that can be aligned, as not all the repeats contain an equal amount of residues due to indels.

With the help of two preceding repeats, two types of matrices can be made. One to displace the template repeat in the direction of the C-terminal, while the other matrix will displace the template repeat to the N-terminal. After displacing the template repeat to either terminal, the newly obtained template repeat is saved and used to continue the process. Via these matrices a backbone model can be obtained which contains the cap structures and the superimposed template repeats. The repeats need to be identified anew, as only the cap structures and template repeats are present, removing the previous indels. The matrices are generated again and can then be applied to the new model to generate more, or less, repeats and displace the cap structures along with the repeats. The displaced repeats and cap structures are saved every time and when finished, they are reassembled to create structures with a varying amount of repeats.

The script does not contain the part in which the template repeat is superimposed on the other repeats, as this is dependent on the used model and is similar to generating more, or less, repeats.

Furthermore, the ensemble 'last-repeat' stands for the last superimposed template repeat which becomes 'last-repeat+1' if the matrix for increasing the amount of repeats is applied once. The same is applicable for the cap structure, which becomes 'capb+1' after applying the same matrix. 'first' stands for the number of the first residue of the protein structure, and 'final-structure' is the name for the complete model.

3.1.5 PyRosetta

PyRosetta is a combination of the Python language with the Rosetta Protein Modelling Suite and can be used for biochemical and biomedical research. It is possible for users to easily write little programs that utilise the Rosetta sampling and scoring functions in their algorithms, such as the prediction of the protein structure, docking simulations, and the design or redesign of functional proteins. Over the years Rosetta has increased its structure prediction quality to an atomic-detail accuracy and is now, by its more simple use, a time-saving method when compared to non-computational methods (Kaufmann et al., 2010).

During this research, PyRosetta was used for protein redesign. To be able to do this a protein backbone, in PDB format, is necessary, as well as different sequences in FASTA format and a Python script.

As a result, the Python script will map the sequences on the backbone model and generate a new model (PDB file) per sequence, accompanied with a file containing the corresponding Rosetta scores.

Furthermore, the RosettaDesign algorithm can be used to energetically optimize the structure and sequence, while the Rosetta FastRelax protocol can be used to relax the protein after the mapping. For the latter, the 'Talaris2014' scoring function is used.

Both the commands and script are listed in the appendix, see appendix B.3.2.

3.1.6 Computational design of proteins

During this master research, the computational protein design can be divided into two parts. One is where every turn is treated as a single turn, while in the second method two turns will form a repeating unit and these units are used for the computational work.

Repetitive repeat structure

Figure 3.1 shows a flowchart that describes the computational process of redesigning AFPs. The process is inspired by the RE$_3$Volutionary protein design method. The process starts with the selection of an appropriate protein and retrieving the 3D structure from the PDB. During this research, the IBP from *L. perenne*, as seen in figure 3.1, and the AFP from *M. primoryensis* were chosen as templates for the CLIPS.

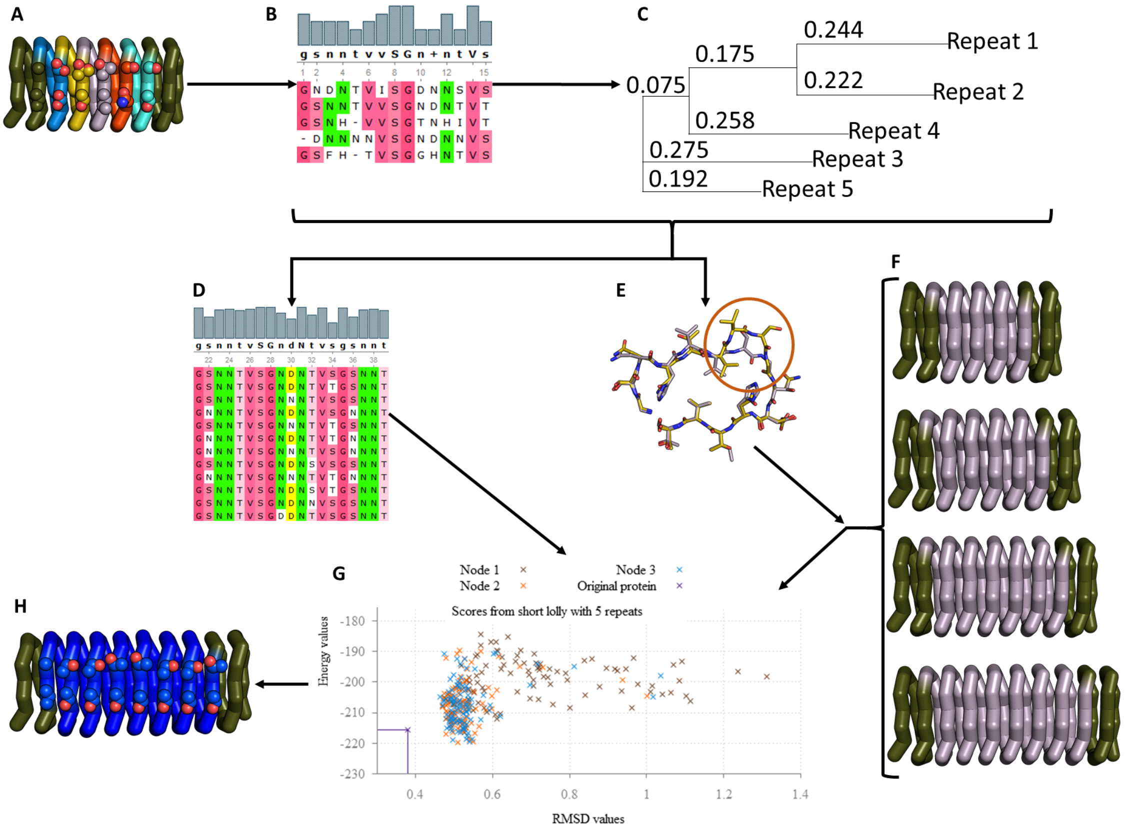


Figure 3.1: **A flow-chart of LpIBP and the applied computational alterations.** The process starts from the WT protein **(A)** from which the repeats are identified, isolated and aligned **(B)**. An unrooted phylogenetic tree is established via webtools **(C)**. Together with the aligned sequences they are used to identify suitable template repeats **(E)** and to create putative ancestral consensus sequences **(D)**. The template repeat is used to create a backbone model with different lengths **(F)** on which the putative ancestral consensus sequences are imposed and scored using a computational protein design tool based on PyRosetta **(G)**. The top sequences are manually validated and the best sequence is selected **(H)**.

From these crystal structures, the sequences were inspected and the repeats were identified. The multiple repeats contain a high sequence and structural similarity. Once the repeat sequences were identified, they were aligned with Strap to represent the sequential conservation (figure 3.1 B). From these alignments a repeat was chosen that could serve as a template structure for future structural modelling (figure 3.1 E). This template should have a highly conserved sequence without any indels when it is compared to the other repeat sequences. The inward pointing residues and the residues that are a part of the IBS are also taken into consideration, as the former are a part of the hydrophobic core and stabilize the protein, while the latter are needed for specific interactions.

The repeat sequences were then used to generate an unrooted phylogenetic tree with Strap (figure 3.1 C). The phylogenetic tree shows different branches connected with a branch point or node. The branches stand for the most conserved sequences and every new branch symbolizes a diverged sequence. Meaning that the most conserved sequence, the ancestral consensus sequence, should be present in the first node and that all the other sequences diverged from this. A trustworthy ancestral consensus sequence can only be found when a great number of sequences are compared. In this case this is less important, as the method is merely used to create multiple consensus sequences.

The phylogenetic tree can then be used to validate the template repeat by looking at the moment when it diverged from the ancestral consensus sequence and the template can then be used for the modelling part.

The chosen template repeat is superimposed on all the repeats of the original wild type (WT) protein via the Brugel software. As such, the redesigned protein backbone contains a specific number of repeats without the presence of insertions or deletions. From the redesigned backbone model, different backbone models that vary in length can be obtained by either adding or removing template repeats and displacing the cap structures. The results from this modification are shown in figure 3.1 F. These obtained protein backbone models serve as backbone models on which the putative ancestral consensus sequences can be mapped.

The possible ancestral consensus sequences are computed through the FastML server (figure 3.1 D) and are then mapped on the backbone models via PyRosetta. Scores for both the energy optimisation and the RMSD values are then collected via a Python script and used to generate a graph (figure 3.1 G).

The top ten or top twenty sequences with the lowest scores and RMSD denotion are evaluated and the best sequence is chosen manually by checking for consensus sequences, significant deviations and the amount of repeating residues within a single repeat. The best sequence is then selected for further modifying.

Repetitive unit structure

A second method to redesign proteins containing repeats, is to take two turns instead of one to create a template repeating unit, and repeat this unit along a linear axis. The use of an extra turn will generate some diversity between the linear repeats, as the immediate neighbour will be a different turn. This may result in an increase of the protein's stability and a divergence of the DNA sequence. By generating more diversity, as the WT proteins all have different repeats, these redesigned proteins will probably have different characteristics and properties when compared to the previous redesigned proteins and the WT proteins.

Three possibilities for the creation of units are further explained.

1. Consecutive repeats

   During this method, a repeat together with the consecutive repeat is considered as a single unit, resulting in unit 1 being repeat 1 and repeat 2; unit 2 being repeat 2 and repeat 3; and so on. Via this method, the possible influences of consecutive repeats are preserved, as consecutive repeats might have had an influence on one another during the evolution.

   After the generation of a phylogenetic tree with the different units, the possible ancestral sequences are obtained and are incorporated between the cap structures of the WT protein. These sequences are then superimposed on backbone models via PyRosetta, generating energy values and RMSD values from which promising sequences can be chosen for further optimization.

2. Overall unit formation

   In contrast with the previous consecutive repeat method, every repeat is combined with the remaining repeats to sample all the possibilities, and thus, it is not solemnly based on possible influences during the evolution like the previous method.

3. Sampling the ancestral reconstructed sequences

   This method combines the obtained repeats from a previous conducted ancestral sequence reconstruction (see ‘Repetitive repeat structure’ from section 3.1.6), while the two previous methods (1. Consecutive repeats and 2. Overall unit formation) both focus on making units from repeats occurring in the WT protein.

   During the ancestral sequence reconstruction of repetitive repeat structures, a phylogenetic tree and putative ancestral sequences were generated by comparing single repeats to each other. Now, these putative ancestral sequences are collected anew and used to identify the newly obtained repeats. These repeats are then combined to each other to create double repeats or units.

   During the unit formation, the use of identical repeats in a single unit is prevented, since this is similar to the previous conducted ancestral reconstruction of repetitive repeat structures.

   In contrast to the previous two methods, the units will now be directly incorporated between the cap structures, as template repeats, and superimposed on backbone models via PyRosetta, instead of first generating an unrooted phylogenetic tree and possible ancestral sequences.

The code to create double repeating units from the sequences of previously conducted ancestral sequence reconstruction can be found in the appendix (see appendix B.3.3).

3.1.7 Sequence editing

The most optimal sequence from the computational protein design is run through different websites and tools for possible sequence optimisations. The base sequence is back translated into amino acids via the entelechon webtool[1]. This tool utilises a library of DNA codons encoding the same base and uses this library to predict the preferable DNA codons for a specific expression organism, while trying to avoid codon bias by alternating between different codons encoding the same amino acid where possible. Another advantage of this webtool is that it takes into account which restriction sites should be avoided. This is especially useful when restriction sites are introduced into the gene. In order to introduce silent restriction sites, the gene code of these restriction sites should differ from the already present restriction sites in the plasmid. If not, the plasmid would be cleaved open and transcription would fail.

Different silent restriction sites are manually inserted in the gene sequence. Silent restriction sites are alterations of the DNA sequence without altering the amino acid sequence. Through this insertion it is possible to restrict at two places with the same enzyme to remove one or multiple repeats of the AFP. The restriction enzymes were chosen with the WatCut tool[2]. The chosen restriction enzymes are not present in the pET28 plasmid.

Theoretical parameters, such as the molecular weight, extinction coefficient, and theoretical pI, were computed through the ProtParam tool[3].

3.1.8 Gnuplot

Gnuplot originated in 1986 as a private software for a better understanding of mathematical functions. These functions are inserted in a command-line, while the output shows the visualisation of these functions. The founders of the software, Thomas Williams, Colin Kelley, Russell Lang and many others, decided to improve the applications of the software, such as including the use for data viewing and web scripting. Nowadays gnuplot is still copyrighted, but it is freely distributed under the GNU license and thus available for everyone[4].

[1] http://www.entelechon.com/bttool/bttool.html

[2] http://watcut.uwaterloo.ca/template.php?act=silent_new

[3] http://web.expasy.org/protparam/

[4] http://www.gnuplot.info/

3.2 Experimental methods

3.2.1 Competent cells

Competent cells are commonly used for the transformation of DNA plasmids into *Escherichia coli*. Different strains are suitable for different applications as well, resulting in higher efficiencies. Two different competent cell strains were used in the following lab work.

For the laboratory plasmid production, the non-pathogenic DH5α strain was used. It is a common choice for the production of plasmids, partly due to the extra features this strain contains, e.g., the endA1 and hsdR17 mutations (Taylor et al., 1993).

The BL21 (DE3) was used to bring plasmids to expression to produce the desired protein. This strain is deficient in the Lon and OmpT proteases, resistant to phage T1 (fhuA2), and contains a chromosomal copy of T7 RNA polymerase gene. The DE3 notations refers to the presence of the T7 RNA polymerase gene which makes them very convenient for the expression of proteins. To induce the expression of proteins isopropyl β-D-1-thiogalactopyranoside is used, known as IPTG (Carl Roth GmbH + Co.KG, Karlsruhe, Germany). IPTG mimics lactose, which results in the transcription of the lac operon.

3.2.2 Plasmids

Vectors are able to carry genes into host cells where they can be replicated and expressed if necessary. The most common types of vectors are artificial chromosomes, cosmids, plasmids, and viral vectors. They all consist of the same basic structures: an origin of replication, a multiple cloning site, and a selectable marker.

Plasmids are a commonly chosen vector for *E. coli* and consist out of a double stranded circular DNA sequence. The transformed plasmids are dependent on the replication mechanisms of the organism, but do not rely directly on the chromosomal DNA for replication. The plasmids contain an origin of replication, commonly abbreviated as ori, which like the name suggests contains the sequence where the replication will initiate. The ori of plasmids may differ with the bacterial genome's ori. The multiple cloning site or polylinker site, is a region that contains a distinct set of restriction sites where the gene of interest can be inserted. The selectable marker makes it possible to distinguish between organisms that contain the plasmid and those that do not. Usually the marker in *E. coli* is an antibiotic resistance gene.

In the here presented labwork, pET28 plasmids (shown in figure 3.2) are used for the cloning and expression of recombinant AFPs. The pET plasmids in general are developed especially for recombinant proteins in *E. coli*, making them very powerful vectors for cloning and expression. The lower case letter next to the number of a pET plasmid

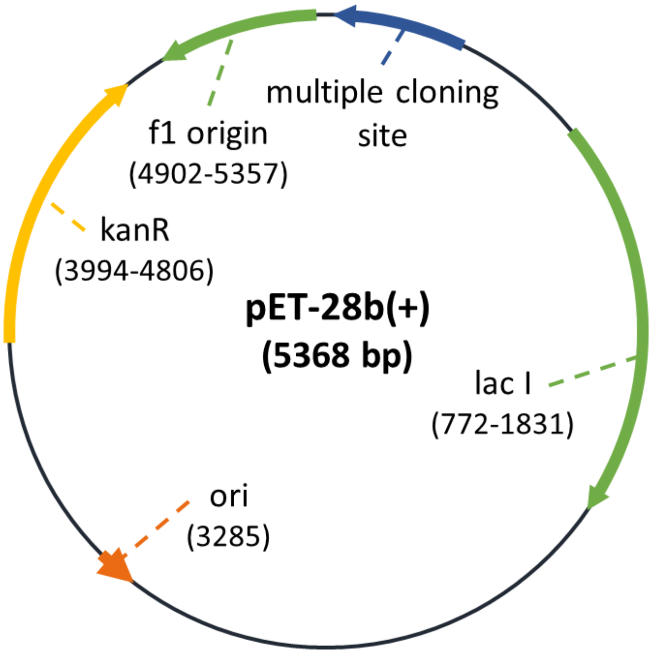


Figure 3.2: **A vector map of the plasmid pET-28b.** The plasmid contains a ColE1 origin of replication (shown in orange), a lacI gene (shown in green), a multiple cloning site (shown in blue), a f1 origin of replication (shown in green as well), and the kanamycin resistance gene (shown in yellow).

indicates that it is a translation vector and stands for the position of the reading frame relative to the BamHI cloning site recognition sequence. During this research there will be no difference between the different pET28 plasmids as the NdeI and XhoI restriction sites are used and the BamHI is located between these two restriction sites. Translation vectors are used for target genes without a ribosomal binding site, they then contain a T7 ribosome binding site, which is highly efficient. Expression of the pET plasmids does not need to occur all the time, i.e., in some cases the recombinant proteins can be damaging for the cell. For this reason transcription of the target genes happen under T7 transcription and the expression of these genes can only occur in strains with a T7 RNA polymerase source, e.g., the BL21 (DE3) *E. coli* strain (Novagen, 2003).

3.2.3 PCR

Polymerase Chain Reaction (PCR) is a common and relatively easy method to amplify small amounts of DNA exponentially. The technique is commonly used and there are only four main components needed. These are the template DNA, primers, a polymerase enzyme, and deoxynucleoside triphosphates (dNTPs). The four components are mixed in a buffer that contains ions and so on for an optimal pH concentration and interaction between the DNA polymerase and the DNA itself.

The fraction of DNA that needs to be amplified is present in the template DNA. The primers, single stranded DNA oligonucleotides, are complementary to the beginning and ending of either the DNA fraction or the DNA template, and may contain required mutations as well. A few guidelines about primers were made to obtain better results when performing a PCR. These guidelines include that the optimal length of a primer is between 18 and 22 base pairs (bp), the amount of guanine and cytosine should make up 40 to 60% of the primer sequence, and the melting temperature of the primer should be between 52 and 58 °C. The primer melting temperature can be calculated with equation 3.1. The dNTPs are the building blocks that are used to make the amplified DNA, while a heat stable polymerase enzyme is required for the replication activity.

$$T_m(°\text{C}) = 2(A_{content} + T_{content}) + 4(G_{content} + C_{content}) - 5 \tag{3.1}$$

Table 3.1: Standard PCR program. Based on the compounds the program can slightly vary.

Step	Temperature (°C)	Time (s)	Number of cycles
Initial denaturation	95	180	1
Denaturation	95	30	25 - 35
Annealing	57	30	
Extension	72	60	
Final extension	72	480	1
Pause	18	Pause	1

Table 3.1 shows a standard PCR program. It contains five steps and a pause phase, in which the DNA is kept at a low temperature in the appliance. The first step is pursued at 95 °C, during this phase the DNA undergoes denaturation, hence the name of the step, and the double strand will separate into two single strands. The temperature is then lowered to 50 - 60 °C in the annealing phase. During this phase the primers are able to hybridize with their complementary sequence at the beginning and ending of the wanted DNA fraction. The heat stable DNA polymerase is able to bind on this double stranded sequence and during the subsequent phase, the elongation phase which occurs at 72 °C, the polymerase can extend the primers in a 5' – 3' fashion. During this phase, the dNTPs are incorporated based on the original DNA template sequence. The last three steps are typically repeated for 35 cycles so that the wanted DNA sequence is amplified exponentially. After the last cycle, the mixture is typically kept at 72 °C for 8 to 10 minutes so that there are no unfinished sequences present.

During the here presented experiments, PCR is conducted in a thermocycler (Westburg, Leusden, The Netherlands), allowing fast and precise control of the temperature.

The PCR products are typically subjected to agarose gel electrophoresis (with a PCR purification afterwards), digestion, and ligation in order to obtain the desired final constructs, free from primers and enzymes (see section 3.2.10, 3.2.11, and 3.2.3, respectively).

Gene amplification

For the amplification of a gene the Pfu DNA polymerase was used. Pfu DNA polymerase is a thermostable polymerase and possesses 'proofreading' activity. This is a 3' to 5' exonuclease activity and checks whether the incorporated nucleotide is correct or not. If it is not correct it will remove the nucleotide at the 3' end. Compared to Taq DNA polymerase, Pfu is slower, but contains a lower error rate, namely $2.6 \, \mathrm{x} \, 10^{-6}$ errors per nucleotide per cycle. Primers were ordered from Integrated DNA Technologies (IDT, Redwood City, California, USA) and both the buffers and enzymes were ordered from Thermo Fisher Scientific Baltics (UAB, Vilnius, Lithuania).

To perform a PCR for gene amplification, the PCR program from table 3.1 was followed and each PCR microcentrifuge tube contained the following:

- 1 µL template DNA (plasmids were ordered from IDT, Redwood City, California, USA or GenScript, Taito, Taito-ku, Tokyo, Japan)
- 1 µL 200 µM dNTP
- 5 µL Pfu buffer with $MgSO_4$
- 2.5 µL forward primer
- 2.5 µL reverse primer
- 0.7 µL Pfu DNA polymerase
- 37.3 µL Milli-Q

Digestion

Both the desired DNA fragment and the plasmid need to be cleaved by the same restriction enzymes in order to be able to insert the desired DNA fragment into the plasmid. Two different restriction enzymes, namely the NdeI and XhoI restriction enzymes, are used to cleave the ends of the DNA fragment and the plasmid to avoid the ligation of plasmids without the DNA fragment or two DNA fragment with each other.

The microcentrifuge tubes for digestion were incubated at 37 °C for two hours and contained the following:

- DNA fragment (varying concentration, at least 500 ng)
- 2.5 µL NdeI
- 2.5 µL XhoI

- 7.5 µL Fast digest buffer
- Milli-Q till a final volume of 75 µL is achieved

The digested products are either used immediately or stored at -20 °C. When they are used, they are first subjected to an agarose gel electrophoresis and PCR purification afterwards to dismiss the enzymes and unwanted DNA fragments (see section 3.2.10 and 3.2.11, respectively).

Ligation

After the digestion, the DNA fragment needs to be inserted in a plasmid. For this to happen, both the DNA fragment and the plasmid are incubated together at room temperature in the presence of a ligation enzyme.

For the ligation reaction, 15 ng of both the DNA fragment and plasmid are required. The final volume should be 5 µL. In the case that the DNA fragment, the plasmid or both have a concentration that is too low to achieve 5 µL, then 10 or 15 µL is used. The amount of the other compounds are doubled, or tripled, respectively. The compounds for a final volume of 5 µL are listed below.

- 0.5 µL T4 ligase buffer
- DNA fragment (varying concentration)
- Preferred plasmid, cleaved with the same restriction enzymes (varying concentration)
- 0.5 µL T4 enzyme
- Milli-Q till a final volume of 5 µL is achieved

The ligation products, plasmids with an incorporated DNA fragment, can be transformed immediately into *E. coli*. For this protocol see section 3.2.4 followed by section 3.2.5, only now there are a few modifications: use 92 µL competent cells (DH5α) with 8 µL ligation products and 550 µL LB instead of 100 µ LB.

Colony PCR analysis

One way to check for false positives is through colony PCR analysis. In this method a PCR is conducted on the lysed *E. coli* instead of only the plasmid. Two primers are chosen in such a way that the insert is amplified, making it relatively easy to detect false positives. In the case of pET vectors, a T7 terminator primer and a T7 promoter primer were selected.

PCR colony analysis uses DreamTaq™ DNA polymerase and (10x) DreamTaq™ buffer, both ordered from Thermo Fisher Scientific Baltics (UAB, Vilnius, Lithuania). Both already include a DreamTaq™ Green buffer with two tracking dyes, making it possible to directly use the PCR products for an agarose gel electrophoresis, as described in section 3.2.10.

To perform a PCR colony analysis, the PCR program from table 3.1 was followed for 25 cycles and with minor adjustments. The adjustments consist of a 120 seconds initial denaturation, an annealing phase at 56 °C and a final extension for 540 seconds. Each PCR microcentrifuge tube contained a 20 µL mixture composed of the following:

- 2 µL 10x DreamTaq™ buffer
- 0.4 µL 10 mM dNTP
- 0.5 µL 10 µM T7 terminator primer
- 0.5 µL 10 µM T7 promoter primer
- 0.2 µL DreamTaq™ DNA polymerase
- 16.4 µL Milli-Q

3.2.4 Transformation of *E. coli*

Competent cells are able to take up extracellular genetic material. This process is called transformation and can occur either by a direct uptake of the genetic material or by incorporation. It is possible to force the direct uptake of genetic material by either electrical or thermal stimulation. During this master thesis only the latter method, heat shock, is used.

Once the transformation is complete, the *E. coli* DH5α strain is able to replicate the amount of genetic material. In most cases this is under the form of plasmids that not only contain the gene of interest, but a specific antibiotic resistance as well. The antibiotic resistance allows the selection of successfully transformed bacteria by using lysogeny broth plates (LB, Sigma-Aldrich Chemie, Steinheim, Germany) with the antibiotic which resistance is integrated in the plasmid.

The colonies growing on these plates can be used for cultivation, in which a larger quantity of the plasmids can be obtained.

Protocol for the transformation with the heat shock technique:

1. Thaw the competent cells on ice. Normally they are held at -80 °C to preserve their competency.
2. Mix 50 µL competent cells with 1 µL plasmid DNA and incubate the cells on ice for 30 minutes.
3. Place the cells at 43 °C for 45 seconds, before placing them on ice for three to five minutes.
4. Add 100 µL LB to the microcentrifuge tubes with competent *E. coli* cells before placing them at 37 °C for one hour.

5. Plate the cells on LB agar plates with antibiotic for which resistance is integrated in the plasmid.
6. Incubate the plates overnight at 37 °C.

3.2.5 Plasmid miniprep

After the transformation of plasmids that contain the desired DNA fragment in DH5α, a colony is grown overnight at 37 °C in a 5 mL culture of LB with antibiotics. The following morning the plasmid DNA is extracted from the *E. coli* by using the GeneJet™ Plasmid Miniprep Kit (Thermo Fisher Scientific Baltics, UAB, Vilnius, Lithuania).

Protocol for the plasmid miniprep:

1. Centrifugate the cells at 12,000 rpm for one minute, after which the supernatant is discarded.
2. The cells are resuspended in a 250 µL resuspension solution. The resuspension solution, together with the lysis solution, elution, neutralisation, and wash buffer, are a part of the GeneJet™ Plasmid Miniprep Kit.
3. Add 250 µL lysis solution and invert the microcentrifuge tubes at least six times. Wait three minutes before going to the next step.
4. Add 350 µL neutralisation buffer and invert the microcentrifuge tubes at least six times.
5. Centrifugate the microcentrifuge tubes for five minutes and transfer the supernatant to the included spin columns, centrifugate the spin columns again for one minute. The DNA now binds to the column's silica membrane. The continuation of this protocol is similar to the final part of the protocol in section 3.2.11.
6. Discard the flow-through and wash the column twice with 500 µL wash buffer. After every addition centrifugate the column.
7. Centrifugate the column an extra time and elute the DNA in a new microcentrifuge tube with 40 µL elution buffer.
8. Measure the concentration of the recovered DNA by UV-Vis spectroscopy (Nano-Drop 2000, Thermo Fisher Scientific, Bleiswijk, The Netherlands).

3.2.6 Plasmid sequencing

After the plasmid miniprep, small samples of the colonies that look promising for future work were send for sequence determination. The company responsible for sequencing is LGC genomics (Berlin, Germany) and determines the DNA sequence through an automated-Sanger method. The results are aligned with the designed structures through NCBI's nucleotide BLAST (Zhang et al., 2000). The nucleotide Blast query and example can be found in appendix B.4.

3.2.7 Cultivation and expression

Protocol for the cultivation and expression of transformed colonies from plasmids with kanamycin resistance:

1. Incubate one colony of BL21 (DE3) from section 3.2.4 overnight at 37 °C in 100 mL LB with 0.1 mL kanamycin (Carl Roth GmbH + Co.KG, Karlsruhe, Germany). The progress can be observed visually, but for a more accurate description the optical density (OD) can be measured at 600 nm by using LB as a reference. The following morning the OD from a ten-fold diluted sample should at least be 0.4.
2. Add 15 mL of the starter culture (previous step) to 1 L LB with 1 mL kanamycin and incubate at 37 °C.
3. Observe the progress by measuring the OD of the culture. Before continuing the procedure, the OD should be between 0.6 and 0.7. Normally it takes 2 - 2.5 hours to reach the desired OD.
4. Place the culture at room temperature for ten minutes before adding 1.177 mL IPTG to induce the protein expression.
5. Incubate the culture with IPTG overnight at 23 °C.
6. Centrifugate the culture at 6, 100 rpm for twelve minutes, discard the supernatant and store the pellet at -20 °C.

3.2.8 Cell lysis

At first, an attempt was made to lyse the cells with a sonicator (Branson Sonifier 450, VWR international S.A.S, Fontenay-Sous-Bois, France). For this, the obtained pellet from section 3.2.7 was resuspended in 10 - 20 mL Dulbecco's Phosphate-Buffered Saline (DPBS, Gibco™, Fisher Scientific UK, Leicestershire, UK) for LpIBP, while the pellet for MpAFP was resuspended in 10 - 20 mL resuspension buffer (see appendix B.1). One cOmplete Protease Inhibitor Cocktail™ tablet (Roche Diagnostics, Mannheim, Germany) was then added before using a sonicator to break the cells open. This was done in five cycles of 45 seconds with a duty cycle equal to 90% and for the output control an intensity of 2.3. In between cycles, the mixture was stored on ice. Afterwards, the mixture was centrifuged at 21, 000 rpm and 4 °C for one hour. The pellet was discarded and the supernatant is used for purification.

Due to the relative low yield of the sonicator, a freeze-thaw method with liquid nitrogen was used to improve the yield.

Protocol for the freeze-thaw cycles with liquid nitrogen:

1. Resuspend the pellet in 10 mL DPBS.
2. Add 1.5 mL protease inhibitor stock solution and 300 µL lysozyme stock solution to the resuspended pellet. The composition of both stock solutions can be found

back in appendix B.1.

3. Immerse the falcon tubes with the resuspended pellet in the liquid nitrogen. Make sure that the whole solution is frozen before continuing the protocol.
4. Thaw the solution by placing it in circulation water.
5. Repeat the last two steps three times. The solution should now have a slime-like structure.
6. Add 150 µL DNase stock solution and 1 mL $MgCl_2$ stock solution. Both stock solutions and compositions can be found back in appendix B.1.
7. Place the solution on ice until it becomes non-sticky. Normally it takes 30 to 60 minutes.
8. Centrifugate the solution at 6, 000 rpm and 4 °C for ten minutes.
9. Transfer the supernatant to a new falcon tube and centrifugate for ten minutes at 7, 250 rpm and 4 °C. The supernatant is used for the purifications process, while the pellet is discarded.

3.2.9 Purification

Thermal denaturation

Some proteins, e.g., the LpIBP, possess the unique ability to refold after thermal denaturation. This ability is useful and makes it possible to increase the purity of the LpIBP by either using thermal denaturation before or after the His-tag affinity chromatography. During this lab work, the thermal denaturation of LpIBP was applied before the purification with Ni-NTA resin.

To perform the thermal denaturation, incubate the recombinant protein at 70 °C for five minutes. Afterwards, let the sample cool down slowly by placing it on the bench top for about ten minutes and fifteen minutes on ice afterwards to assure a slow cooling process. To separate the denatured proteins from the refolded proteins, centrifugate the sample at 12, 000 rpm for one minute. The supernatant is then used to continue the purification process with.

His-tag affinity chromatography

The purification of AFPs is mainly done using Ni-NTA resin (QIAGEN GmbH, Hilden, Germany). It is an affinity chromatography method in which nickel ions are immobilized on the resin due to four chelation sites in the NTA. This will result in a greater and stronger binding capacity of hexahistidine (His_6) tagged recombinant proteins, and a high degree of purity. One mL resin has an approximate binding capacity of 50 mg His_6-tagged recombinant protein. While the recombinant protein is bound to the chelating agent, the impurities are washed away with a buffer containing low concentrations of imidazole. Imidazole is able to bind to the chelating agent as well and acts as a competitor. When low concentrations of imidazole are used, the chelating agent still has a higher affinity for the His$_6$-tagged recombinant proteins, but when the concentration of imidazole is elevated to $100 - 250$ mM, the recombinant protein is eluted. Elution can occur by a reduction in the pH as well. During this master thesis the pH is kept at 8.0 while the concentration of imidazole is increased.

Protocol for the use of Ni-NTA resin columns:

1. The Ni-NTA resin is stored as a 50% aqueous suspension with 20% (v/v) ethanol. Before the Ni-NTA resin can be used, remove the ethanol and wash the resin first with Milli-Q and then with the same buffer as in which the recombinant protein is present. The amount of Ni-NTA resin present in the column is called 'bed volume', during this master thesis a bed volume of 4 mL is typically used. The continuation of this protocol is based on a bed volume of 4 mL.
2. Add the recombinant protein to the Ni-NTA resin and let it mix overnight in a end-over-end shaker for an optimal interaction.
3. Place a polyethylene disc at the bottom of the 20 mL column (Bio-Rad, Hercules, CA, USA) and load the column with the mixture.
4. Remove the safety cap and collect the flow-through.
5. Once the flow-through is collected, add 10 mL of the 5 mM imidazole buffer (wash buffer 1, see appendix B.1).
6. Once the first wash fraction is collected, add 20 mL of the 20 mM imidazole buffer (wash buffer 2, see appendix B.1) and collect it in fractions of 10 mL.
7. To elute the recombinant protein, add 30 mL of the 250 mM imidazole buffer (elution buffer, see appendix B.1). The first 15 mL is collected in fractions of 1 mL, the last 15 mL as fractions of 5 mL.

The recombinant protein was identified in the collected fractions using SDS-PAGE analysis (see section 3.2.10).

Superdex purification

The susequent step of the purification is via a prepacked size exclusion chromatography column. The matrix consists of dextran, different branched glucan molecules varying in length, and is covalently attached to cross-linked agarose beads. The smaller molecules are able to interact with the beads, while the larger molecules will just pass through, and thus, elute first. The Superdex columns are designed in such a way that they have the same separation range as the Sephadex columns. The range of the high resolution depends on the size of the column, e.g., $3 - 70$ kDa for the Superdex 75 and $10 - 600$ kDa for the Superdex 200(Drevin and Johansson, 1991).

Before the sample can be injected, the Superdex column is washed with two column volumes (CV) of 3 M NaCl and, 2 CV of 1 M NaOH, 2 CV of Milli-Q water, and finally 2 CV of the final buffer. Fractions will be taken after the injection, but only the fractions corresponding to the peaks will be analysed via SDS-PAGE.

3.2.10 Gel electrophoresis

One of the methods to separate DNA, RNA, and proteins is through gel electrophoresis. Hereby a separation based on the size and charges of the molecules is obtained. The two most frequently used techniques are either a polyacrylamide gel for Sodium Dodecyl Sulphate PolyAcrylamide Gel Electrophoresis (SDS-PAGE) or an agarose gel electrophoresis. The latter is used for DNA and RNA fragments, since it has a greater separation range while the SDS-PAGE is more used for protein separation or DNA fragments up to 500 bp.

Polyacrylamide gel

Different gels for the SDS-PAGE can be made to obtain a better separation of a specific molecular weight (MW) range, from which the target protein is a part. The compositions of the used gels are shown in table 3.2 and the gels are run on a Mini-PROTEAN™ Tetra Cell from Bio-Rad (Hercules, CA, USA).[5] During the SDS-PAGE the protein fractions are denatured first. This is done with the ionic detergent sodium dodecyl sulphate (SDS), giving the protein fractions a negative charge as well. The samples are applied at the top of the gel and migrate through the gel matrix due to the electrical current that is applied on the mini tank. Within a given time, bigger proteins will migrate less far than smaller proteins due to the size of proteins and the composition of the gel matrix. To obtain the best separation degrees, the 12% gel is used for MWs between 10 and 200 kDa, while the 16% gel is better for MWs between 3 and 100 kDa. Due to the arbitrary chosen amount of running and stacking gel, and the possibility that the degree of separation can vary from run to run, a prestained sample containing ten proteins with known MW is applied next to the samples for easy comparison (PageRuler™ Prestained Protein Ladder, Thermo Fisher Scientific, Bleiswijk, The Netherlands).

The SDS-PAGE gel consists out of two parts. The running gel, which is the main part of the SDS-PAGE gel, and the stacking gel, which contains wells in which the samples and the prestained protein ladder are applied. Both the running and stacking gel need about 30 minutes to polymerise.

Protocol for preparing the polyacrylamide gels:

1. Mix all the compounds, except for APS and TEMED, from the running gel carefully before degassing the solution for five minutes.

[5] AA stands for acrylamide, BAA for bis-acrylamide and TEMED for tetramethylethylenediamine

Table 3.2: Composition of SDS-PAGE gels

	Running gel		Stacking gel
Compounds	**12%**	**16%**	**4%**
1.5 M Tris-HCl, pH = 8.8	5 mL	5 mL	0 mL
0.5 M Tris-HCl, pH = 6.8	0 mL	0 mL	3.125 mL
30% AA / BAA	8 mL	10.6 mL	1.625 mL
10% SDS	200 µL	200 µL	125 µL
Milli-Q	6.8 mL	4.2 mL	9.15 mL
10% APS	175 µL	175 µL	150 µL
TEMED	17 µL	17 µL	17 µL

2. When the mixture is degassed, add and mix the APS and TEMED carefully. (APS: Fisher Scientific UK, Leicestershire, UK and TEMED: Merck Group, Darmstadt, Germany).
3. Pour the mixture in the casting frames, while keeping in mind that the stacking gel with wells comes on top. These wells should not touch the running gel, since the applied samples can then distribute over the width of the running gel. Add a layer of isopropyl alcohol (Acros Organics, Geel, Belgium) on top of the running gel to prevent dehydration.
4. When the running gel is almost polymerized, prepare the stacking gel like the running gel. When the running gel is polymerized, remove the isopropyl alcohol layer and add the stacking gel on top.
5. Carefully place the comb on top of the stacking gel to form the wells.
6. When the stacking gel is polymerized as well, the gel can either be used immediately or stored at 4 °C for future use.

When the gels are ready for use, place them in the electrophoresis tank and fill the tank with running buffer (see appendix B.1). The wells are rinsed with the running buffer before the samples or the prestained protein ladder are loaded.

Protocol for preparing the samples:

1. Mix the samples containing the protein fractions with premade laemmli buffer (1:1) containing β-mercaptoethanol (see the appendix B.1).
2. Boil the samples for five minutes and centrifugate at $7,000$ rpm for one minute before dispensing 15 µL of the samples into the wells. Usually the first well is used to load 3.5 µL of the prestained protein ladder.
3. Close the electrophoresis tank and run the SDS-PAGE gel at a constant current of 20 mA and a voltage of 200 V. When the desired separation is achieved, switch off the current and put the gel in a staining solution for 30 minutes.

4. Destain the gel in a destaining solution, until the desired degree of destaining has been achieved. Every ten minutes the destaining solution is refreshed by filtering it with activated carbon. Both the staining and destaining solutions can be found back in appendix B.1.

Agarose gel

As mentioned before with the polyacrylamide gel, it is possible to obtain a better separation of a specific MW range by altering the composition of the gel. With agarose gels, the concentration of agarose is altered. For the following experiments, agarose gels of 0.9% and 1.5% agarose in TAE buffer (1x, see appendix B.1) are used. Dissolve the agarose in the heated TAE buffer and pour it into the casting frame once the temperature is tolerable for the material. The width of the wells depend on the type and amount of samples that are run and the comb is then placed on the casting frame before the solution is poured into the frame. Once the gel is ready for use, it is placed inside the gel electrophoresis tank (Mupid™-One Submarine Mini Electrophoresis System, Tokyo, Japan) and the gel is run at 100 V until the desired separation is achieved.

If there is no loading dye present in the samples, a 5x loading dye (Fisher Scientific UK, Leicestershire, UK) is used.

If a PCR purification is performed afterwards, then the samples are split in a small fraction (for one well) and a main fraction (for about three wells). This brings the opportunity of prolonged exposure of the smaller fractions to ethidium bromide solution and UV light without causing harm to the main fractions. The smaller fractions are then placed next to the O'GeneRuler™ Plus DNA Ladder of 100 bp or 1 kb (Thermo Fisher Scientific, Bleiswijk, The Netherlands) to function as a reference. Once the gel has run, the small fractions and DNA ladder are separated from the main fractions and put in an ethidium bromide solution. Ethidiumbromide is able to intercalate with the DNA helix and its fluorescent qualities makes it possible to visualise the binding with DNA under UV light. During the intercalation the fluorescence intensity increases almost 20-fold. The staining takes about 15 minutes for an adequate detection. After the staining, the fractions are observed under UV light. From the smaller reference fractions, the desired DNA band can be cut and discarded, while the rest of this gel (the DNA ladder and the small fractions with a hole where the desired DNA band was) is used as a reference to quickly find the desired DNA bands in the main fractions. The main fractions are put for a minimum amount of time in ethidium bromide solution and UV light is only used to check whether the DNA bands are on the same position as they were for the separated fractions. The DNA bands from the main fractions are then cut and collected in an microcentrifuge tube for further treatment (see section 3.2.11).

3.2.11 PCR purification

The acquired DNA from agarose gels from section 3.2.10 is purified using GeneJET™ PCR Purification Kit (Thermo Fisher Scientific Baltics, UAB, Vilnius, Lithuania).

Protocol for the PCR purification:

1. Add binding solution (part of the GeneJET™ PCR Purification Kit) to the microcentrifuge tubes containing the cut DNA bands until the gel fragment is immersed in binding solution.
2. Heat the microcentrifuge tubes at 55 °C until the gel is completely dissolved.
3. If the desired DNA fragment contains less than 500 bp, then add an equal amount of isopropyl alcohol to the mixture as was done with the binding solution.
4. Decant the mixture in the purification column and centrifugate at $12,000$ rpm for one minute. The binding buffer contains agents that denature proteins and promotes the binding of the DNA to the column's silica membrane.
5. Wash the column twice with 700 µL wash buffer, after every addition centrifugate the column.
6. Centrifugate the column an extra time and elute the DNA in a new microcentrifuge tube with 40 µL Milli-Q.
7. Measure the concentration of the recovered DNA by UV-Vis spectroscopy (NanoDrop 2000, Thermo Fisher Scientific, Bleiswijk, The Netherlands).

3.2.12 Bradford protein assay

The Bradford protein assay is based on the absorption shift from 465 to 595 nm due to the binding of Coomassie Brilliant Blue G-250 to the protein and can thus be used as a method to determine the protein concentration from purified samples. Coomassie Brilliant Blue G-250 exists in two colours, i.e., red and blue. When Coomassie Brilliant Blue G-250 is added to a purified protein sample, it will bind to the protein and convert to a blue colour. This dye binding process takes approximately two minutes to complete and results in a complex with a high sensitivity and stability that lasts up to an hour, after this time it starts to precipitate. The Bradford protein assay is a robust method with little to no interferences. The use of proper buffer controls can eliminate alkaline buffering agents and small amounts of detergents such as SDS, while cations and carbohydrates give little to no interference (Bradford, 1976).

Before the Bradford protein assay is performed, the samples with the desired proteins are first collected and then prepared for dialysis. For the dialysis, the samples were poured inside a SnakeSkin™ dialysis membrane (Thermo Fisher Scientific, Bleiswijk, The Netherlands), which is sealed at the bottom and at the top. Samples were dialysed overnight at 4 °C in a 1 L volume filled with a target buffer (see appendix B.1) and a magnetic stirring bar before the Bradford protein assay can be resumed.

The composition of the Bradford dye reagent, GFP$_{\text{UV}}$ storage buffer, and the reference protein solution can be found in appendix B.1. The first column of a 96-well plate is used for the calibration curve with the reference protein solution. The reference protein solution contains BSA, which is typically used to determine unknown protein quantities. BSA has the advantage that it is stable, relatively cheap, and has no effect in many biochemical reactions. The first well does not contain any reference protein solution, while the second well contains 10 µL. Every following well contains 10 µL more than the previous one, so that the last well of the first column (well H1) contains 70 µL of the reference protein solution. GFP$_{\text{UV}}$ storage buffer is added to the wells so that the final volume of these wells is 100 µL. The second column is a duplicate from the first column, and thus contains the same reference protein solution and dilutions. Dilutions of the purified protein sample are made, usually these are a 25-fold and a 12.5-fold dilution. From these dilutions 20, 40, 60, and 80 µL are added to four consecutive wells and again GFP$_{\text{UV}}$ storage buffer is added to obtain a final volume of 100 µL. The following column is again a duplicate of the purified protein dilutions series. To each used well 150 µL of the Bradford dye reagent is added and the OD was measured at both 595 nm and 405 nm with the Tecan Safire2™ (Tecan Group Ltd., Männedorf, Switzerland). The acquired results were used to calculate the calibration curve and from this the concentration from the protein samples could be measured.

3.2.13 Circular Dichroism spectroscopy

Circular Dichroism (CD) spectroscopy is a popular method to analyse secondary structures. It is based on the absorbency of polarized light by chiral molecules, and proteins, due to the asymmetric nature of the peptide bond. When a beam of equally amounts of left and right circularly polarized light is radiated through a chiral sample, one polarization state will be absorbed to a greater extent, generating a positive or negative CD signal.

The spectra is measured as a function of the wavelength and expressed in millidegrees (mdeg). As mentioned, CD spectroscopy is mostly used to analyse or monitor secondary structure of proteins, due to the fact that different CD activity is measured for different secondary structure elements. The α-helix, e.g., has two characteristic minima at 208 and 222 nm, while the β-sheet has only one characteristic minima at 215 nm.

The secondary structures can be monitored in function of temperature as well, and secondary structures can easily change with a change in pH or upon interaction with other molecules (Woody, 1995).

The estimation of α-helical and β-sheet structures can be made through different web tools, such as K2D2 and K2D3 (Perez-Iratxeta and Andrade-Navarro, 2008; Louis-Jeune et al., 2011).

3.2.14 Matrix-assisted laser desorption/ionization mass spectrometry

The Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry (MALDI MS) is a soft ionization mass spectrometry method that can be used to determine the molecular weight of biomolecules and organic molecules up to 1 MDa. In contrast to electrospray ionization (ESI), the generated ions will only carry one charge, facilitating the processing of the results. The term soft ionization indicates that fragmentation will only occur in very low quantities. MALDI can be coupled to mass analysers as well to obtain high resolution and very high quality data.

Before utilizing MALDI, the sample must be mixed with matrix material and applied to a target where solvent evaporation will occur. This evaporation leads to a concentration distribution of the sample on the target. It is possible that the concentration differs per region on the target, meaning that the results are not reproducible. This technique is called a solvent-based technique. Other techniques, such as the solvent-free sample preparation, allow a more homogeneous sample surface on the target, resulting in reproducible measurements. These techniques are more time consuming, which is why the relative easily solvent-based technique is more performed. When the matrix is exposed to the laser beam, it will absorb the emitted energy and release it again as heat. The sample will use this heat to vaporize and form ions that can be analysed through a Time-Of-Flight Mass Spectrometry (TOF MS) (Constans, 2005; Hyzak et al., 2011).

Time-of-flight mass spectrometry

During the mass spectrometry a selection on the ions will occur. This is based on the mass to charge ratio (m/z), but after a MALDI the charge is typically equal to one. During a TOF MS, both the electric field and the distance to the detector will be held constant, while the time it takes for ions to be detected will be measured. The measured time is then dependent on the velocity and thus the m/z, and is easily converted, with the known parameters, to the m/z, resulting in spectra showing the abundance of ions versus the m/z (Guilhaus, 1995).

4

Experiments and results

The aim of this research is to develop a procedure to generate CLIPS. These proteins contain a linear repetitive structure, with a relatively flat IBS, that can be easily modulated in length. As the template proteins are AFPs, it is possible to check whether the redesigned proteins retain the antifreeze activity and how this activity is influenced by the used procedure and by the amount of repeats.

4.1 Choice of template AFPs

All AFPs' coordinates were obtained from the PDB and evaluated based on their structure and the protein purification process. A suitable AFP satisfies the following requirements; (i) the structure of the AFP resembles a linear rod and (ii) contains a multiple repeat structure. If not, twisting and bending may occur when more repeats are introduced.

Structural analysis revealed that most of the hyperactive AFPs have a rod-like structure while the fish AFPs tend to vary in shape. Yet, in the following research the moderate active AFP from *L. perenne* and the hyperactive AFP from *M. primoryensis* were chosen as templates, due to the lack of disulphide bridges and inclusion body formation. Both are briefly described in section 1.1.3 and shown in figure 4.1 and 4.2.

4.2 Calippo

During the first part of this master thesis, the focus was placed on the AFP from *M. primoryensis*, MpAFP. The name of this project was inspired by the shape of the protein, the calcium ions that are present inside the protein, and the antifreeze activity. The number following 'Calippo' describes the number of repeats present in the protein.

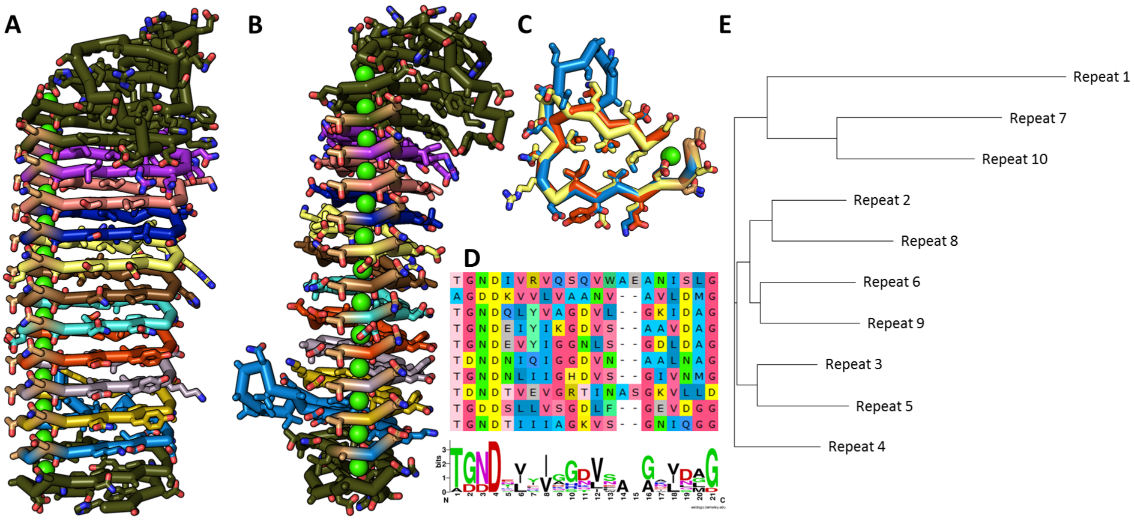


Figure 4.1: **Structural analysis of the AFP from *M. primoryensis***. Panels **A** - **B** show the front and side view of the MpAFP. All the repeats are shown in a different colour and the identified cap structures are shown in olive green. Panel **C** shows the structural alignment of the short template, the long template and the first repeat, as this repeat contains an insertion of eight amino acids. Panel **D** shows the alignment of the different identified repeats with a sequence logo underneath. The conserved calcium binding site (GTGND) is clearly visible. The repeats were used to generate an unrooted phylogenetic tree as well, which is shown in panel **E**.

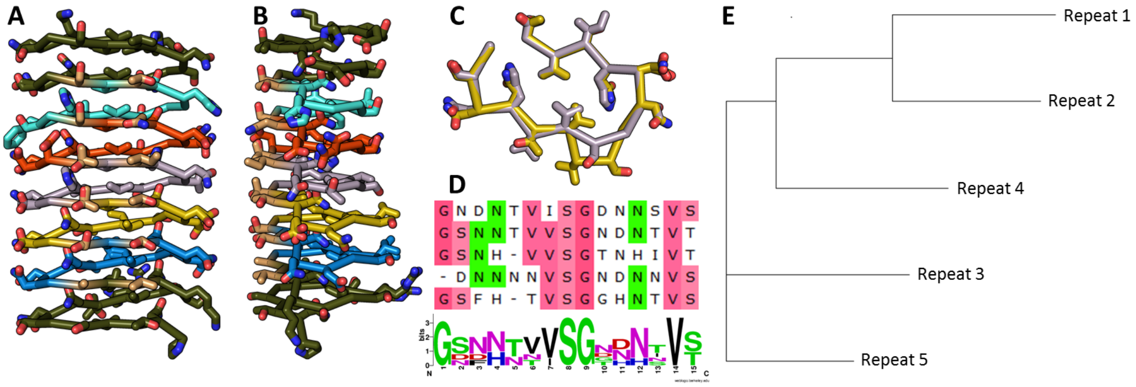


Figure 4.2: **Structural analysis of the IBP from *L. perenne***. The front and side view of the LpIBP are shown in panel **A** and **B**. The cap structures are shown in olive green, while every repeat is shown in a different colour. Both the template repeats are structurally aligned in panel **C**, while the identified repeats are aligned in panel **D** with a generated sequence logo underneath. Panel **E** shows the unrooted phylogenetic tree, which was generated from the different repeats.

4.2.1 Computational protein design

The first step of each project was to redesign existing AFPs or IBPs. The main focus of the newly designed proteins is to vary in length when compared to the abundant WT isoform. Later, they can be analysed to see if they still contain a TH or IRI. If they do, they can be compared and ranked relatively to the WT protein and each other.

After retrieving the MpAFP structure from the PDB (3P4G), the structure was split into two cap structures and ten enclosed repeats. A template repeat was chosen from the structurally aligned repeats, as described in section 3.1.6. This repeat was then utilized to construct the backbone model by superimposing the template repeat on the other repeats via Brugel, resulting in a protein with two cap structures and ten identical repeat structures in between. Via a Brugel procedure, similar to the procedure described in section 3.1.4, five different backbone models were generated, containing eight to twelve repeats as seen in figure 4.3.

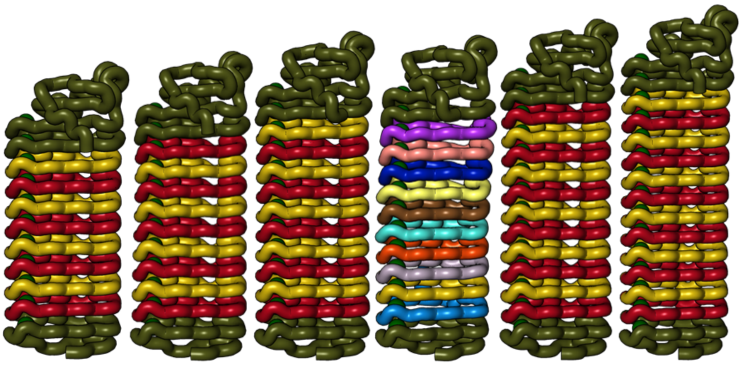


Figure 4.3: **Backbone model of the redesigned Calippo proteins**. A comparison of the different used backbone models, including both the redesigned models as the WT MpAFP model. The caps are coloured in olive green, while the repeats of the redesigned backbone models are coloured red and yellow in alternating style. Each repeat from the WT MpAFP has a different colour. The backbones range from eight to twelve repeats.

Putative ancestral sequences were generated and were then mapped via PyRosetta on these backbone models. The obtained results are shown in figure 4.4 and were further processed, until one promising sequence, with a low energy, remained. This sequence was translated into a gene sequence, further optimized and at least three different restriction enzymes, AatII, BshTI, and KpnI, were chosen for the different silent restriction sites, allowing to remove up to six repeats when all three restriction enzymes are combined. The combination and results of the restriction enzymes can be seen in figure 4.5.

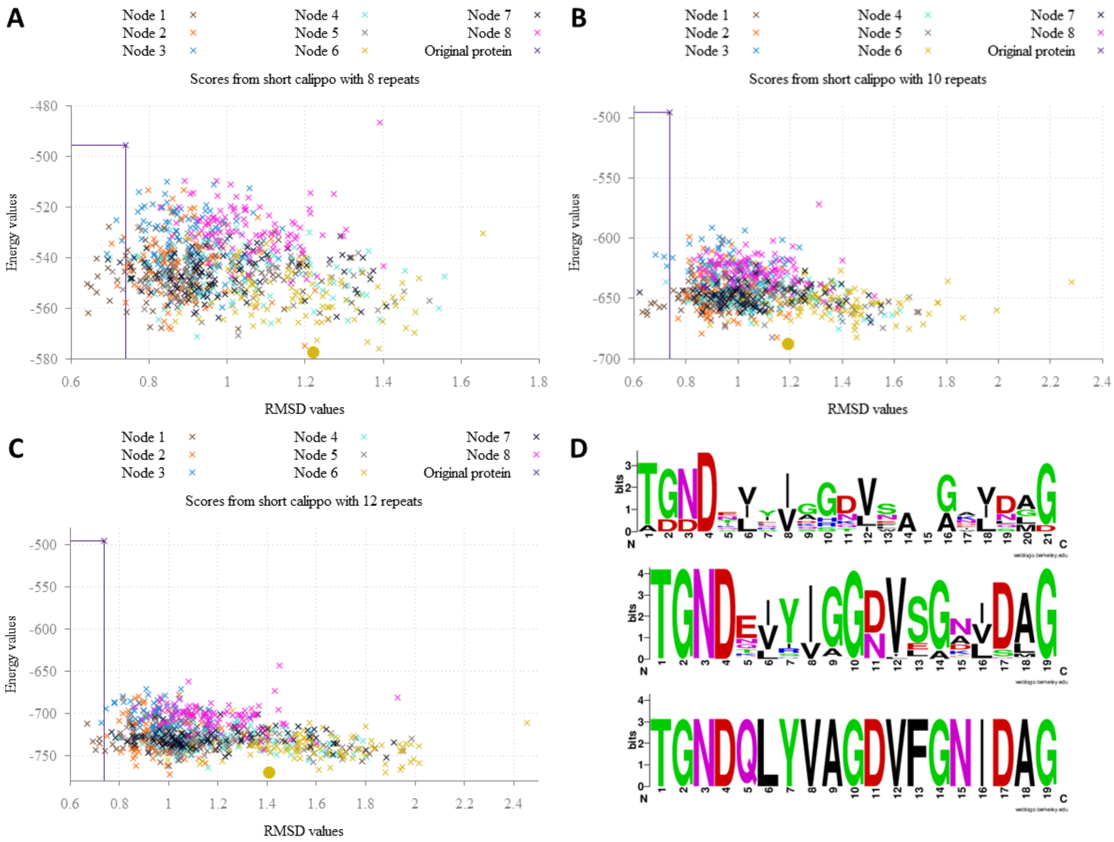


Figure 4.4: **Ancestral reconstruction dot plots of redesigned Calippo proteins, with energy values as function of RMSD.** Graph **A - C** represent all the possible sequences for Calippo8, Calippo10, and Calippo12. The different colours indicate the different nodes, while the yellow coloured circle indicates the sequence which was chosen for the following experiments. Through PyRosetta the different sequences were energetically minimized and compared with the original backbone, and thus generating energy values and RMSD values per superimposed sequence. The values of all the sequences per node, from the phylogenetic tree, are shown in the graphs. A reference point, the purple x, was made by imposing the original WT MpAFP sequence on the original backbone. All the sequences that have an energy value and RMSD value smaller than the reference point are, theoretically, more favourable sequences. Panel **D** shows a comparison of different sequence logos. The first logo shows the alignment of the different repeats, the second logo shows the alignment of all the possible ancestral sequences, while the third logo is the chosen sequence.

4.2.2 *In vitro* experiments

Sequence delivery and processing

The constructs were ordered from a vendor, namely GenScript (Piscataway, NJ, USA). The sequence for Calippo12, with three different silent restriction sites, was delivered in a pET28 vector. Because the gene of interest is already in the correct plasmid, only the restrictions, as depicted in figure 4.5, need to be applied in order to obtain the smaller desired protein sequences.

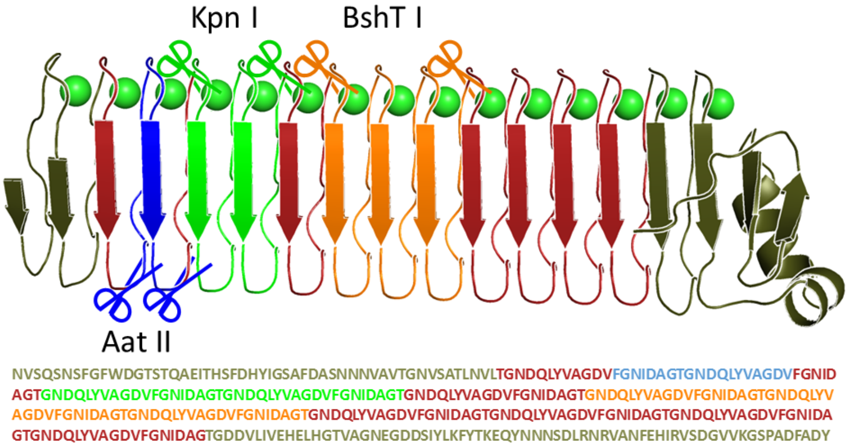

Kpn I
BshT I
Aat II
NVSQSNSFGFWDGTSTQAEITHSFDHYIGSAFDASNNNVAVTGNVSATLNVLTGNDQLYVAGDVFGNIDAGTGNDQLYVAGDVFGNID
AGTGNDQLYVAGDVFGNIDAGTGNDQLYVAGDVFGNIDAGTGNDQLYVAGDVFGNIDAGTGNDQLYVAGDVFGNIDAGTGNDQLYV
AGDVFGNIDAGTGNDQLYVAGDVFGNIDAGTGNDQLYVAGDVFGNIDAGTGNDQLYVAGDVFGNIDAGTGNDQLYVAGDVFGNIDA
GTGNDQLYVAGDVFGNIDAGTGDDVLIVEHELHGTVAGNEGDDSIYLKFYTKEQYNNNSDLRNRVANFEHIRVSDGVVKGSPADFADY

Figure 4.5: **Calippo12's cleavage pattern.** A cartoon representation of Calippo12 is shown, with underneath the corresponding structure's sequence with conserved colour patterns. Each restriction site is placed twice, as a silent mutation, to span an integer number of repeats. By combining the different restriction patterns shorter Calippo proteins can be obtained. The scissor symbols in the figure correspond to the gene code of the enzyme restriction sites and have the same colour as the repeats that will be removed. Restriction enzyme AatII is shown in blue, while restriction enzymes KpnI and BshTI are shown in green and orange, respectively.

For the restriction reactions, three small samples were taken, one sample for each restriction enzyme. As such, three new protein sequences were obtained, namely Calippo11, Calippo10, and Calippo9. As a control, all three plasmids were sequenced before continuing. Calippo9, which was cleaved with the BshTI restriction enzyme, was used to continue the process and was combined with the two other restriction enzymes separately, resulting in Calippo7 and Calippo8. Again, as a control, both new plasmids were sequenced so that a final combination could be made by combining the three restriction enzymes to obtain Calippo6, which was once again sequenced.

Transformation and expression

The plasmids were transformed into *E. coli* BL21 strains for protein expression, which were induced with 1.177 mL IPTG when an OD_{600} of 0.6 was reached.

All the plasmids were successfully transformed in *E. coli* cells and brought to expression. However, only the WT showed a clear expression pattern. During the purification via affinity chromatography it became clear that the Calippo proteins had failed to express. The expression pattern and purification of the WT protein and a Calippo protein are shown in figure 4.6 and figure 4.7, respectively.

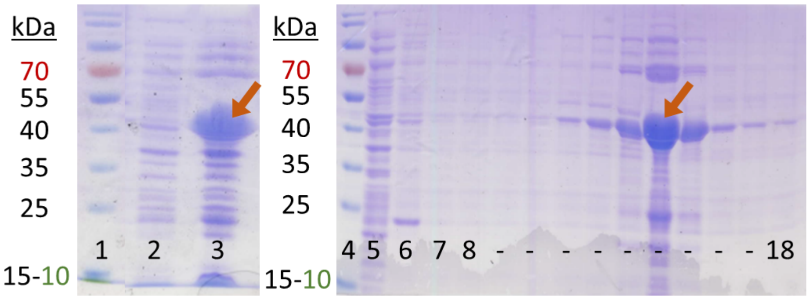


Figure 4.6: **Expression and purification of the WT MpAFP on a 12% SDS-PAGE-gel.** The first and fourth lane show the prestained protein marker, while lane 2 and 3 are samples from before and after the induction with IPTG, respectively. The orange arrow indicates the production of the desired protein. Lane 5 shows the flow-through, after applying the resuspended pellet on the Ni-NTA resin column. Lane 6 and 7 show the fractions of the washing procedure, while lane 8 to 18 are different fractions from the elution procedure. The orange arrow on the second gel indicates lane 14 with fraction 12 and contains the highest concentration of MpAFP.

4.3 Lolly

The second attempt at redesigning a protein was based on the IBP from *L. perenne*, LpIBP. The name of this project is inspired by the the genus of the organism in which the WT protein occurs, namely *Lolium*, and the fact that the WT protein has an antifreeze activity. The number following 'Lolly' describes the amount of repeats that are present in that protein.

4.3.1 Computational protein design

As with the MpAFP, the Lolly part started with the redesign of the existing WT protein, in this case the LpIBP. A few things are clearly different when the LpIBP is compared with the MpAFP, such as, (i) the size of the protein, (ii) the average size and amount

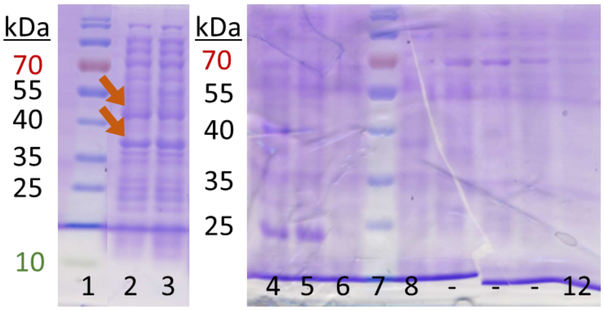


Figure 4.7: **Expression and purification of Calippo10 on a 12% SDS-PAGE-gel.** Calippo10 was chosen to represent the different Calippo proteins, as it is about the same size and weight as the WT MpAFP. Lanes 1 and 7 show the prestained protein marker. The first gel shows the samples from before and after the induction with IPTG, lanes 2 and 3, respectively. The orange arrow indicates two possible positions in the lane that might be due to the protein expression of the desired protein. However, for both positions a band of the same intensity is observed before induction, possible indicating that there is a very low induction or no induction at all. Lanes 4 to 6 show fractions of the washing procedures, while the remaining lanes are fractions of the elution procedure. As visible in both gels, no Calippo proteins were purified, nor brought to expression. The results of the other Calippo proteins were similar to the results shown here.

of repeats, (iii) the absence of calcium ions, and (iv) the difference in both IRI and TH activity. Despite these differences, most of the design processes will be identical, but during *in vitro* experiments, different buffers will be used because of the calcium ions that are present in the MpAFP.

The LpIBP was retrieved from the PDB (3ULT). After observing the structure, two cap structures and five enclosed repeats were identified. However, in the literature the protein is described as having eight tandem repeats without any cap structures. After observing the sequence, some deviations were found in the first and last repeats and to avoid putative folding frustrations these terminal repeats were identified as cap structures. The five remaining repeats were structurally aligned and compared as described before. The template repeat was superimposed on the remaining repeats via Brugel, as described earlier, but this time backbone models were obtained for proteins with three to seven repeats, as can be seen in figure 4.8.

The ancestral sequences generated from the repeats were mapped on the protein backbone via the PyRosetta procedure and the generated data is collected, processed, and shown in figure 4.9. The chosen sequence was further processed until one promising sequence remained. This sequence was then further modified, such as the translation in bases and the insertion of the silent restriction sites for the restriction enzymes BamHI and BmtI.

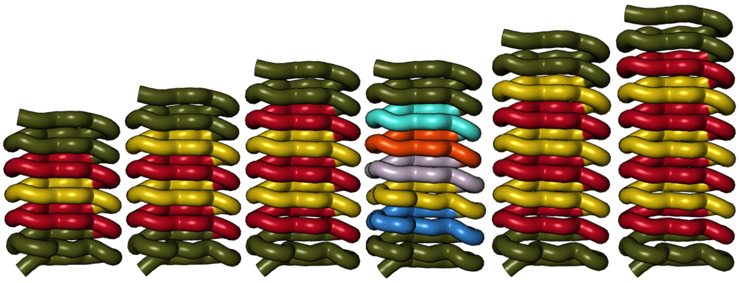


Figure 4.8: **Backbone model of the redesigned Lolly proteins with single repeats**. A comparison is shown of the different used backbones during the second part of the thesis with the WT LpIBP, from which every repeat has a different colour. The cap structures are coloured in olive green, while the repeats from the redesigned models are coloured red and yellow, in alternating style.

4.3.2 *In vitro* experiments

Sequence delivery and processing

In order to produce the Lolly proteins, the gene sequence was ordered in a geneblock vector from Integrated DNA Technologies (IDT, Redwood City, California, USA). After arrival, the gene of interest was placed in a pET28 plasmid, cleaved with NdeI and XhoI restriction enzymes. The plasmids with inserts were transformed into *E. coli* BL21 (DE3) cells for both the production of plasmids and the expression of proteins. Once the protein is expressed, the protein needs to be purified before it can be checked and assessed on antifreeze activity.

Due to the short repeat length and the high similarity between every repeat it was impossible to synthesize the full Lolly7 construct. The different Lolly gene sequences were then ordered separately and only Lolly3 and Lolly4 were successfully synthesized and cloned in a pUC vector.

Transformation and expression

Before expression could be induced, the plasmids containing the redesigned Lolly sequences and the WT sequence were successfully transformed into *E. coli* BL21 cells. The expression was then induced using 1.177 mL IPTG once the OD_{600} reached 0.6, as described in section 3.2.7. The expression pattern of the redesigned Lolly proteins appeared successful when they were compared with the expression pattern of the WT protein. Both expression patterns are shown in figure 4.10.

Purification

Following protein expression, the pellet was resuspended in DPBS. Two lysis techniques, one with a sonicator and one freeze-thaw method with liquid nitrogen, were applied on

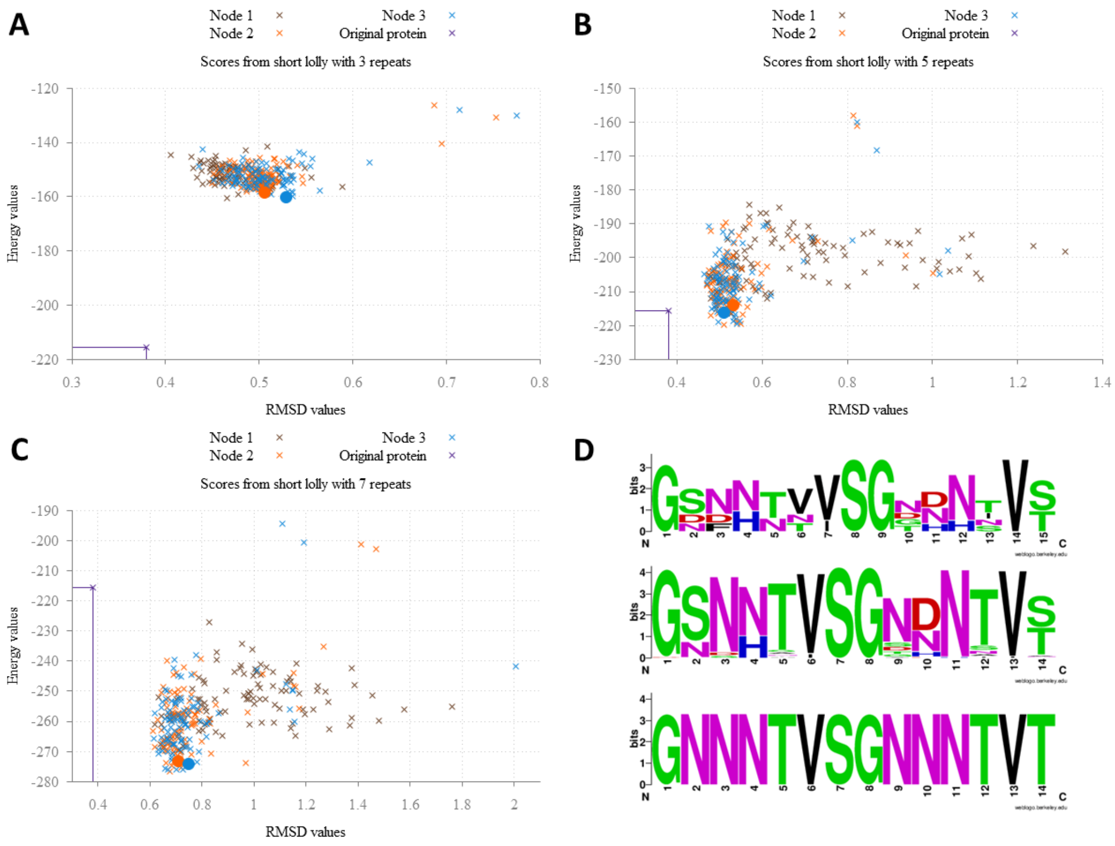


Figure 4.9: **Ancestral reconstruction dot plots of the redesigned Lolly proteins, with energy values in function of RMSD values.** Graph **A** - **C** represent all the possible sequences for Lolly3, Lolly5, and Lolly7, respectively. The different colours indicate the different nodes, while the two coloured circles share the same sequence and were chosen for the following experiments. Via PyRosetta the different sequences were energetically minimised and compared with the original backbone, and thus generating energy values and the RMSD per superimposed sequence. The values of all the sequences per node, from the phylogenetic tree, are shown in the graphs. A reference point, the purple x, was made by imposing the original WT LpIBP sequence on the original backbone. Panel **D** shows three sequence logos. The first one is the comparison between the different repeats from the WT LpIBP, the second one is the alignment of all the possible ancestral sequences, while the last logo is the sequence of the chosen repeat.

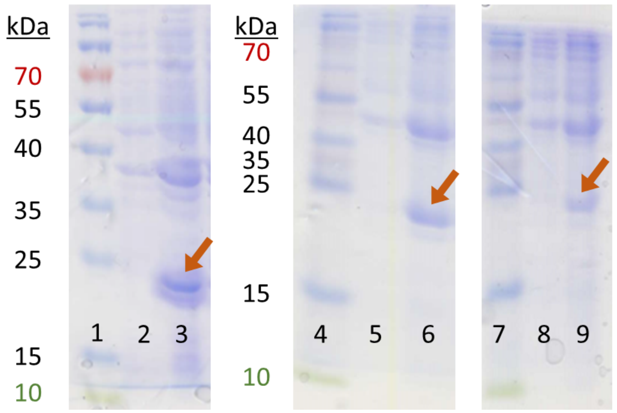


Figure 4.10: **The induction of Lolly proteins shown on a SDS-PAGE-gel.** The induction of the WT LpIBP is shown on a 12% gel, while both redesigned Lolly proteins are shown on 16% gels. Lane 1, 4, and 7 show the prestained protein marker while the lanes following this marker are, respectively, samples from before and after the induction with IPTG. The orange arrow indicates the production of the desired protein.

the WT protein to see which one held a bigger yield of the protein. The freeze-thaw method, as described in section 3.2.8, had a yield that was almost ten times higher, so for the subsequent work only the freeze-thaw method was utilized.

After cell lysis, the desired protein needs to be purified. For this, all three proteins were subjected to thermal denaturation and affinity chromatography for purification. Lolly3 and Lolly4 still revealed low concentrations of unwanted proteins and Lolly3 was subjected to the Superdex 75 to see if a third purification resulted in a more pure sample. The estimated protein concentrations for the WT, Lolly3, and Lolly4 were 0.037, 0.052, and 0.024 M, respectively, and were placed on a SDS-PAGE-gel for comparison, as can be seen in figure 4.11. Lolly3 appeared slightly smaller on the gel as expected.

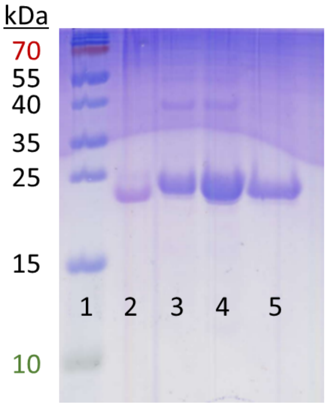


Figure 4.11: **16% SDS-PAGE-gel of the comparison of Lolly proteins.** The first lane shows the prestained protein marker, followed by the WT LpIBP, Lolly4, Lolly3, and Lolly3 after purification with the Superdex75.

Protein validation

The next step when proteins are obtained, is to verify the proteins by checking the estimated mass of the protein and the fold. To get an idea of the proportion between α-helices and β-sheets the proteins were analysed via CD spectroscopy and different tools were used to compute the content of α-helices and β-sheets, as mentioned in section 3.2.13. A rough estimation of the protein's mass can be obtained through a SDS-PAGE gel, shown in figure 4.11. The CD spectra of all three proteins can be seen in figure 4.12. As visible, there is a clear difference between the spectra of the WT protein and the redesigned Lolly proteins. This difference was confirmed via MALDI TOF MS. The MALDI TOF MS spectra can be found in appendix B.5.

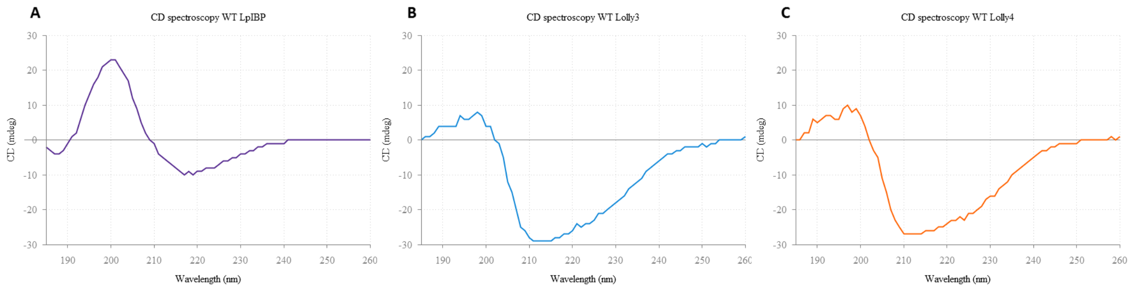


Figure 4.12: **CD spectra of LpIBP and the different Lolly proteins.** The three graphs show the CD spectra of the WT LpIBP, Lolly3, and Lolly4, respectively. As visible, both redesigned Lolly proteins indicate on the presence of α-helical structures (absoption band around 190 nm), while the WT LpIBP has a very high β-sheet content. Remarkable are the low values at 200 nm for the redesigned Lolly proteins. The computed α-helices content for the WT LpIBP, Lolly3, and Lolly4 was 2.69, 86.82, and 57.63%, while the computed β-sheet content was 43.46, 11.76, and 23.55%, respectively.

4.3.3 Double repeats

To allow greater diversity at the genetic level and remove putative folding frustration, a novel design was developed. The difference with the previous design method is located in the template repeats. This time two repeats were selected to function as a single repeating template unit instead of only one repeat per template repeat. A Brugel procedure was followed to obtain different backbones, as shown in figure 4.13 and afterwards, putative ancestral sequences were generated in three different ways before superimposing the sequences on the backbones via PyRosetta, as previously described in section 3.1.6 on page 37. The mapping of the sequences on the backbones takes a long time to process, especially because of the large amount of sequences that are being sampled for the multiple backbones. For this reason only the data of the Lolly proteins is available.

The generated data from the three different methods was collected and processed, and shown in figure 4.14 where they are compared with the initial approach, where one repeat was used per template repeat. The results will be further discussed in the next chapter.

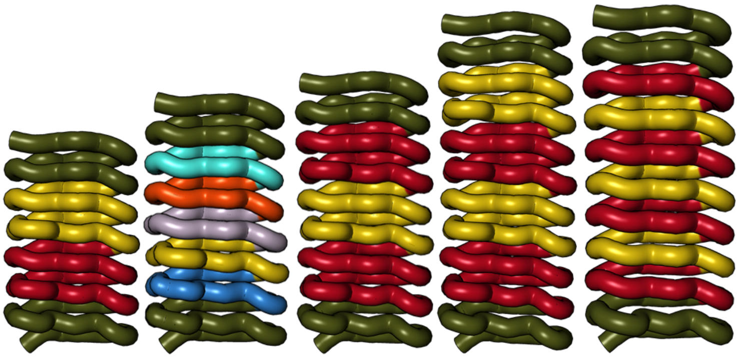


Figure 4.13: **Backbone model of the redesigned Lolly proteins with two repeats as one unit**. A comparison is made between the backbones of the redesigned Lolly proteins, containing double repeats, with the WT LpIBP and Lolly7 with single repeats, as the last backbone model. The cap structures are coloured in olive green, while every repeat of the WT LpIBP is shown in a different colour. The repeats of the redesigned backbone models are shown in an alternating style of red and yellow.

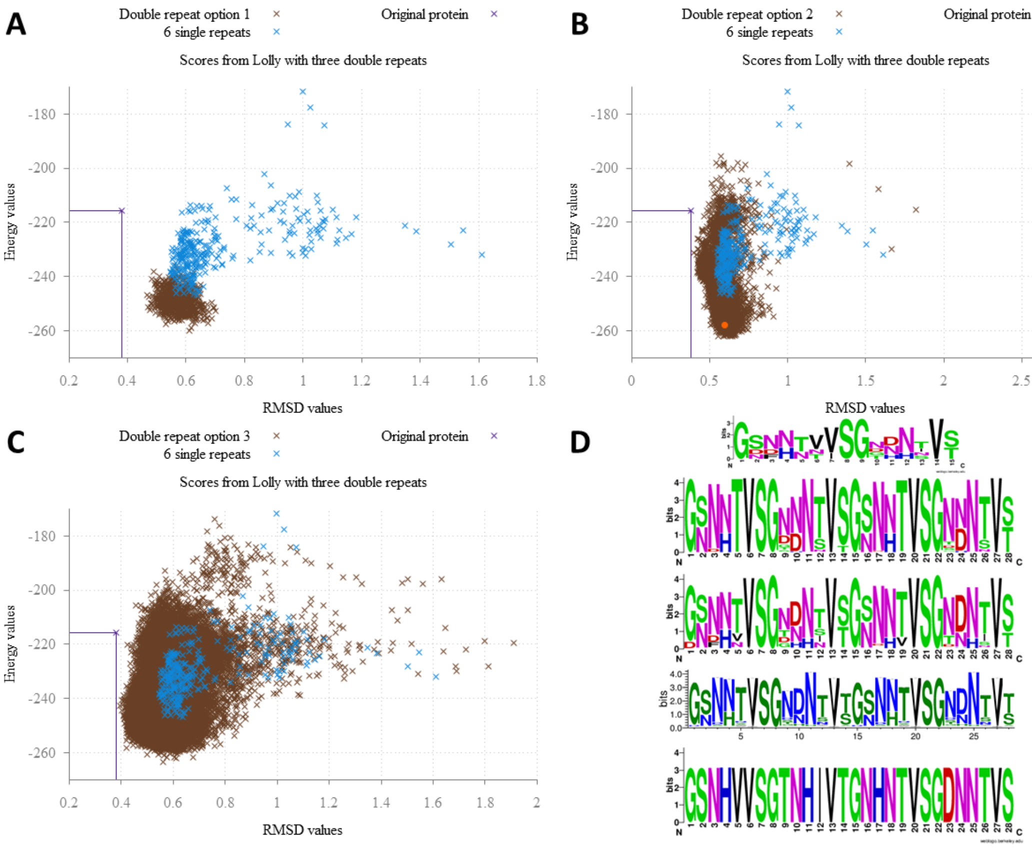


Figure 4.14: **Ancestral reconstruction dot plots of the redesigned Lolly proteins with double repeats**. Graphs **A - C** display all the repeat sequences from method 1, 2, and 3, respectively. The blue spots represent sequences based on the initial single repeat per template approach, while the brown spots represent the sequences based on the new approach where two repeats were chosen per template repeat. The purple x, in all three graphs, is the reference point, which was made by imposing the original WT LpIBP sequence on the original backbone. The orange dot present in graph B presents the sequence which looked the most promising. Panel **D** shows the sequence logo plots, starting from the comparison between the different repeats from the WT LpIBP, a comparison from all the sequences from method 1, 2, and 3, and the last one shows the sequence which looked the most promising.

5

Discussion

The main goal of this Master thesis is to develop a procedure to generate CLIPS, Crystal Lattice Interacting Protein Scaffolds, for possible applications that require interaction with a crystal-lattice or for biomineralization. For these purposes AFPs were chosen as the most promising template proteins, as they do not exhibit the typical bending and twisting like other repeating proteins and contain a variety of different structures with a repeating architecture.

We focus on the AFPs that contain a flat IBS, which is able to interact with the ice crystal lattice, and has repeating units along a linear axis, resulting in a β-helical architecture. Theoretically it would be possible to modify these AFPs to produce proteins with an adjustable length. Unlike other natural proteins with tandem repeats, AFPs are not subjected to twisting and bending. Examples of twisting and bending are shown in figure 5.1.

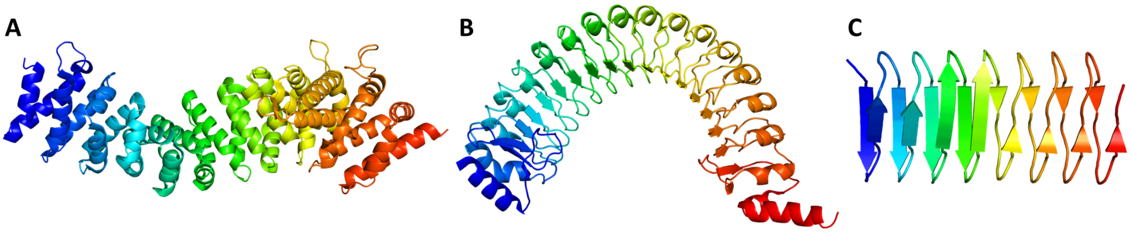


Figure 5.1: **Cartoon representations of the crystal structures of translational repeating proteins.** Panel **A** and **B** show an armadillo repeat domain from *Caenorhabditis elegans* (PDB: 4R10) and a leucine rich repeat domain from *Chlorobaculum tepidum* (PDB: 5IL7), respectively. Both possess repeat structures, but the repetitive structure will twist and bend while the ice-binding protein from *L. perenne* (PDB: 3ULT) shows a repetitive structure along a linear axis, without the twisting and bending.

There are quite some AFPs that fit the desired criteria for this project. For this reason, further restrictions have been applied, such as no extra stabilisation due to disulphide bridges and the absence of inclusion body formation during protein expression. As a result, the two most promising candidates were antifreeze proteins from *L. perenne* and *M. primoryensis*.

During the following steps, both proteins were used separately as two different projects. Both proteins were used as models for the computational protein redesign, where backbone models were created with a varying number of repeats. At the same time the repeats were compared, aligned, and modified to create repeats with the same amount of amino acids. These repeats could then be used to create an unrooted phylogenetic tree and possible ancestral sequences. The ancestral sequences were mapped on the designed backbones, energetically optimized and sorted, so that the most promising sequence can be obtained. The corresponding genetic sequence was further modified, so that it contained the silent restriction sites, before ordering the expression construct.

5.1 Calippo

Only the largest Calippo construct, Calippo12, was ordered from GenScript. By means of minor adjustments, three different silent restriction sites were introduced in the sequence, allowing to derive five smaller protein sequences from Calippo12. The base sequence for the derived Calippo proteins was verified via sequence determination, but only the WT MpAFP was expressed successfully, as can be seen in figures 4.6 and 4.7.

The sequence logo from figure 4.4 D shows that only the calcium-binding site is conserved in the WT MpAFP, as the rest of the sequence clearly shows multiple amino acids per residue number (the first logo). The final chosen sequence (the third logo) shows similarities with both the WT logo and the logo from the ancestral sequences, but lacks any clear consensus sequence, except for the calcium-binding site, which also is the IBS. Similar observations were made during other protein design, e.g., the Pizza family, where a divergence from the consensus sequence is necessary for stable proteins (Voet et al., 2014).

Initially it was thought that the Calippo proteins would have a higher chance to succeed, as they have relatively large repeat sequences compared to other AFPs, resulting in a putative higher variety between successive residues and repeats at the codon level.

A possible explanation for the lack of expression can be found in the codons, namely a codon bias due to the repetitive structures. A shortage of low-abundance tRNA might occur during translation, which results in either point mutations or incomplete proteins (Rosano and Ceccarelli, 2014). There may be an increased chance for codon bias to occur, as the redesigned Calippo proteins contain up to twelve identical repeats. However, codon optimization tools were used to redesign the sequence, and to validate the sequence again afterwards, to limit the chance of codon bias. Still, the chance that codon bias occurs cannot be excluded, but it can easily be checked by using low plasmid copy numbers to reduce the metabolic pathways. This will reduce the use of tRNAs and thus generate a bigger chance of success, if codon bias is the cause. Another method to improve the odds, is by translocating the produced proteins to the periplasm or the medium by using signal peptides, or start from redesigned proteins with low amounts of repeats, as the lack of expression might be due to the higher amount of identical repeats.

Other reasons, such as protein toxicity and the influence of the His$_6$-tag, have a smaller chance of being responsible for the lack of protein after purification, as the WT AFP was expressed correctly. For instance, protein toxicity has a relatively smaller chance as the only changes applied to the natural proteins were localised in the repetitive structure and were inspired by ancestral reconstruction, meaning that no new residues were introduced in the redesigned sequences. The influence of the His$_6$-tag might depend on the terminal it is attached to and can disrupt the folding or state of the protein, causing aggregation or possible degradation.

However, aggregation and degradation can be caused by other reasons as well, such as small changes in the backbone or in the side chain conformations and due to the repetitive character of small β-turns along a linear axis. The small changes in conformation and orientation may cause changes in the interaction with not only the solvent, but between differently charged residues as well, resulting in a potential loss of entropy (Baker, 2010). Although no new amino acids were introduced in the sequence, the protein's core was conserved as well to potentially increase the stabilisation. However, during protein translation, the protein folding will commence, resulting in interactions between newly formed β-sheets, the surrounding residues, and consecutive β-sheets. Thus, the temporal folds will probably be different when they are compared with the final achieved helical architecture. These interactions may be unfavourable, causing putative folding frustrations.

5.2 Lolly

At the beginning of the Lolly part, the obtained sequences after ancestral reconstruction were sorted on energy. The sequences with the lowest energy were compared and the most promising sequence was chosen. This choice was based on the position and repetition of certain amino acids, e.g., repeating aspartic acid residues, protruding phenylalanine or isoleucine residues, and led to the sequence of Lolly7. The differences between the chosen sequence, the ancestral sequences and the WT sequence can be seen in the sequence logos from figure 4.9 D. As visible, the repeat sequence is highly conserved and almost found back in all three sequences.

Lolly7, with two different silent restriction sites, to derive Lolly6, Lolly5, and Lolly4 from Lolly7, was initially ordered from IDT. However, due to codon bias and repetitive residues and repeats, the construct failed to synthesise. Subsequently, the Lolly constructs were submitted separately after a new optimisation round and only Lolly3 and Lolly4 were synthesized successfully.

First, the WT LpIBP was brought to expression, which can be seen in the second lane of figure 4.11. It shows that the LpIBP migrated at a different position than expected, which is called gel shifting. The difference between the predicted and apparent mass has been confirmed before by Lauersen *et al.* (Lauersen et al., 2011). Later, when Lolly3 and Lolly4 were brought to expression a similar band appeared on the SDS-PAGE-gel. As visible in figure 4.11, the mass of the redesigned Lolly proteins are more or less equivalent. Although, at first Lolly3 appeared to be smaller than Lolly4 and both Lolly proteins seem to have a bigger mass than the WT protein with five repeats, which might be due to gel shifting as well. This phenomena occurs when there is a change in the amount of SDS that is linked with the protein and mostly occurs in proteins with elevated amounts of accessible hydrophobic and charged residues, as SDS has a preference to accumulate at these residues (Rath et al., 2009).

For further validation a CD spectra was measured for both the WT LpIBP and the redesigned Lolly proteins. The three spectra can be seen in figure 4.12. The WT LpIBP clearly shows signs of a protein that is very rich in β-sheets, as expected, while the samples from the redesigned proteins have a lower signal over noise ratio and dominantly display α-helical structures. For the redesigned Lolly proteins, the signal over noise ratio starts to decrease remotely fast when the wavelength gets close to 200 nm and it reaches its limit at 195 nm. The increasing noise may indicate the presence of salt-like structures, since no chiral-activity was measured, but the increase of α-helical structures might indicate that a different protein is expressed.

To verify whether the correct protein was purified or a different protein, a MALDI TOF MS was assessed, which confirmed that the isolated protein did not correspond to the designed proteins. Results of the MALDI TOF MS can be found in appendix B.5. Again, the WT protein showed expected results with very low contamination, while both redesigned Lolly proteins did not show any of the corresponding peaks, indicating that the desired protein was not present in these samples. In both samples, the purified protein had the same MW, which is 2 kDa bigger than the MW of Lolly3 and 1 kDA bigger than Lolly4.

The nearly 25 kDa bands present in the SDS-PAGE-gel of both redesigned Lolly proteins (figure 4.11) are also present in the wash fractions of the redesigned Calippo proteins, indicating that the protein is probably a household protein from *E. coli* that is expressed under stress conditions, such as the universal stress proteins (Usp) (Farewell et al., 1998).

Like the Calippo proteins, a possible explanation can be found in the codon bias and the repetitive structures, as each turn consist out of two repeats with seven amino acids. These seven residues contain three consecutive aspartic acid residues, a glycine residue and the repetitive IBS, which might not generate enough difference to avoid codon bias.

Again, folding problems leading to aggregation and degradation can be caused by the

His_6-tag, changes in the entropy or conformation, or misfolding during protein translation.

5.2.1 Double repeats

After the results were known for both projects, an improved approach was developed to design the CLIPS, which may have a higher chance of success. Instead of focusing on one turn and using this turn as a template repeat over the whole protein, two turns would now be considered as a single repeating template unit. However, these two turns should be non-identical, as the initial method uses identical repeating turns. The new repeating units may avoid structural incompatibilities, such as steric hindrance and repelling charges, and may have a stabilizing effect on each other. Another advantage of the double repeats is that there is a greater diversity present in the gene sequence.

Three different methods were used, as described in section 3.1.6, and gave significant varying results as can be seen in figure 4.14.

The first method was utilised via consecutive repeats. During this method, two consecutive repeats were treated as a single unit, conserving the link between multiple repeats as they might have had an influence on each other during the evolution of the repeats. This can be easily seen in figure 5.2 A, as it shows the least deviation and has a low RMSD value. The repeats were used to generate an unrooted phylogenetic tree and possible ancestral sequences, which were mapped on the backbone model.

In the second method, the potential influence of consecutive repeats was discarded, by combining all five repeats with each other. The increase in combinations will result in a higher sampling and a more complex, but evolutionary meaningless, unrooted phylogenetic tree. Although a rise in both the energy and RMSD value can be observed, the deviation will be limited as this method is only based on the five original repeats.

During the last method, all the possible ancestral sequences of single repeats were combined with each other to form a single repeating unit existing of two non-identical repeats. This method allows coincidental combination, as a huge amount of sequences are sampled without any evolutionary relation. The deviation during this method increases largely, as can be seen in figure 5.2, as the RMSD of the sequences increase over the three different methods. However, due to the large coincidental sampling this deviation will generate sequences with a lower energy value as well when it is compared to the first method.

Another interesting observation made is that there is a significant difference between the population of the single template repeat backbone sequences, consisting of six repeats, and the sequences of double template repeat backbone, consisting of three double repeats (figure 4.14 A - C). The difference in energy can be observed in figure 5.2 E as well. However, via the third method a huge amount of sequences were sampled that held no evolutionary relation. This method will contain sequences that scored equally well to

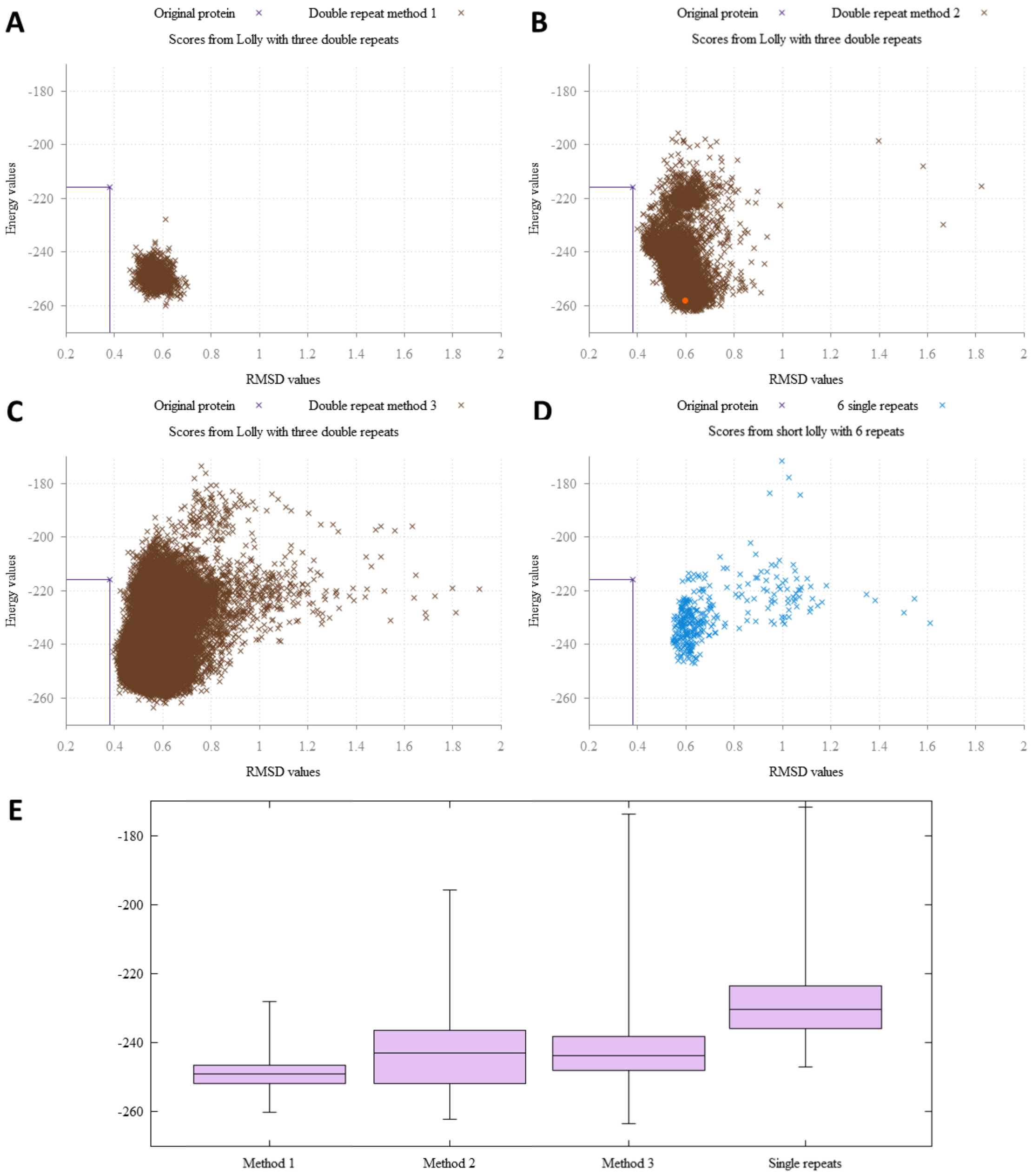


Figure 5.2: **Comparison of the energy values of the backbones with three double repeats**. Graphs **A - C** show the three different methods used for generating the double repeating template unit, while graph **D** shows a backbone with a single repeating template unit, equal in length, for comparison. Panel **E** shows a boxplot variant where it is clearly visible that the third method has a higher sampling and that all three methods score better than the initial attempt with single repeating units.

the initial method, but also sequences that are more favourable and less favourable than the sequences of the single repeating template unit with six single repeats.

In total, 185 sequences with the lowest energy were chosen from the three methods. These sequences were compared with each other and the most promising sequence was chosen based on (i) the amount of differences between two repeats from the same repeating unit, (ii) the amount of consecutive aspartic acid residues and (iii) the amount of aspartic acid residues present in the repeating unit, and (iv) the amount of β-sheet breaking residues present in the β-sheets. Due to the lack of time the sequence could not be ordered and expressed in time. However, the increase of diversity in consecutive turns and the extra care in choosing the desired sequence, may have a favourable effect on the β-helical structure, protein expression and purification.

When the sequence logos are compared (figure 4.14 D), the highly conserved repeat sequence can be found back in all three methods. However, this time the final chosen sequence displays more variation when it is compared to the consensus sequence. During the initial design of the single template repeat a sequence that was very similar to the consensus sequence was experimentally tested. These experiments did not yield any purified protein, suggesting that the consensus sequence is not always a stable sequence for further protein design. The increase in variation may then avoid putative folding frustration and may have a stabilising effect on consecutive repeats.

5.3 Recommendations

5.3.1 Multiple sequences

One way of improving the odds in computational protein design is by using multiple sequences that vary in only a few amino acids. This may not only result in a successful sequence, but it may also help to understand the specificity of amino acids in the case of β-helical structures and AFPs with an IBS.

However, even by using more sequences, and thus a higher degree of sampling, there is no guarantee there will be a successful sequence, as the current success-rates of protein design are relatively low and the mechanisms are not yet fully understood.

For instance, the lab of Baker tested about 87 proteins from which only two proteins were a success (Fleishman et al., 2011). Similar experiments were repeated in other papers from the Baker lab and showed similar results.

Baker and coworkers made it clear that for future computational protein design it is important to know what went wrong. Because one of the reasons that there is so little known about protein design is because there is a lack of time and funding to focus on the unsuccessful designs. However, even if not everything is known, the newly discovered information can be incorporated in existing algorithms, which can then be refined to generate better predictions about the structure and energy of proteins.

5.3.2 Recombinant protein tags

His_6-tag

The location of the His_6-tag can have an influence on the protein folding and function as well. However, a relatively easy method to test the influence of the His_6-tag is by either switching the tag from the C-terminal to the N-terminal, or vice versa.

Different tag systems

Different tag systems, such as the GST tag and the MBP tag, are possible to use as well.

The glutathione S-transferase tag (GST tag) results in fusion proteins with a high affinity for glutathion coated beads. Although the native state of the protein is altered, the tag can easily be removed with thrombin.

The maltose-binding protein (MBP) can be used in fusion proteins too. Although the mechanism is not fully understood, MPB tags would increase the solubility of the fusion protein and can be purified via amylose columns. After cleaving with specific proteases, the protein can be separated via affinity chromatography.

Both switching the position of the His_6-tag and utilizing different tags could be explored in the future. Therefore, the goal is to identify a highly expressing, easy to purify and handle protein.

5.3.3 Cell strains

Another influence on the expression pattern of proteins is the cell strain that is used. Different BL21 cell strains vary in the active genes and incorporated plasmids and may have effects on the protein expression that are not always fully understood. For instance, BL21 (DE3) pLysS contains an additional plasmid with a T7 lysozyme encoding gene. Target genes that are normally under the control of the T7 promoter will maintain a low background expression, but it will not interfere with IPTG induced expression via IPTG.[1]

Although codon optimisation tools were used in the beginning of this project to avoid the use of non-encoded tRNAs, it is possible to use special strains such as the BL21-CodonPlus (DE3)-RIPL strains and Rosetta™ strains that encode for extra tRNAs that usually limit the translation. These strains allow the use of rare or new tRNAs, as the production of these tRNAs is increased or introduced, and allows high-level expression of recombinant genes that were otherwise not possible (Novagen, n.d.).

[1] Promega Corporation (2016), “BL21(DE3)pLysS Competent Cells” https://be.promega.com/products/cloning-and-dna-markers/cloning-tools-and-competent-cells/bacterial-strains-and-competent-cells/bl21_de3_plyss-competent-cells. Accessed: 2016-05-11.

6

Conclusion and future perspectives

A computational redesign algorithm was developed successfully, even though the purification and expression of the desired proteins did not have any yield. The lack of correctly expressed proteins led us to evaluate and redo the computational part with a new approach. The new approach of redesigning and selecting proteins showed more promising and improved results when it was compared to the initial approach. Previous experiments showed that the repeat sequence should not be completely similar to the consensus sequence, as the initial attempt was very similar to the consensus sequence and had no yield (figure 4.9). The new method clearly showed more deviations (figure 4.14), which results in a higher diversity in the gene sequence and may also have a stabilising effect on consecutive repeats. However, there was not enough time to test this sequence via *in vitro* experiments due to the time consuming PyRosetta simulations.

If future experiments are to be conducted, it is possible to test whether the new approach is indeed better, as there are more differences present between two consecutive repeats and both repeats contain more diverse residues when compared to the first approach, suggesting that this method may result in more favourable sequences. Future experiments would not only allow us to test the correlation between the amount of repeats present in the redesigned protein and the TH, but it would allow us as well to see if there is a maximum amount of repeats to obtain an optimal, or maximum, activity. The redesigned proteins could serve as CLIPS as well, and it might help improve the understanding of AFP mechanisms, resulting in an improvement of the commercially used AFPs.

However, no redesigned proteins were expressed during this master thesis, which still is a common occurrence in computational protein design. Although most reasons of why this happens are unknown, multiple labs are investing time and resources to obtain a better knowledge on how computational protein design can be improved to yield better success-rates.

References

Ashkenazy, H., O. Penn, A. Doron-Faigenboim, O. Cohen, G. Cannarozzi, O. Zomer, and T. Pupko (2012), “FastML: A web server for probabilistic reconstruction of ancestral sequences.” *Nucleic Acids Research*, 40, 580–584.

Atici, Ö. and B. Nalbantoglu (2003), “Antifreeze proteins in higher plants.” *Phytochemistry*, 64, 1187–1196.

Baker, D. (2010), “An exciting but challenging road ahead for computational enzyme design.” *Protein Science*, 19, 1817–1819.

Bradford, M. M. (1976), “A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding.” *Analytical Biochemistry*, 72, 248–254.

Buckley, H. E. (1952), *Crystal growth.* Wiley, New York.

Chao, H., P. L. Davies, and J. F. Carpenter (1996), “Effects of antifreeze proteins on red blood cell survival during cryopreservation.” *The Journal of Experimental Biology*, 199, 2071–2076.

Chao, H., M. E. Housten, R. S. Hodges, C. M. Kay, B. D. Sykes, M. C. Loewen, P. L. Davies, and F. D. Sönnichsen (1997), “A diminished role for hydrogen bonds in antifreeze protein binding to ice.” *Biochemistry*, 36, 14652–14660.

Chen, L., A. L. DeVries, and C.-H. C. Cheng (1997a), “Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod.” *Proceedings of the National Academy of Sciences of the United States of America*, 94, 3817–3822.

Chen, L., A. L. DeVries, and C.-H. C. Cheng (1997b), “Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish.” *Proceedings of the National Academy of Sciences of the United Stated of America*, 94, 3811–3816.

Cheng, C.-H. C. (1998), “Evolution of the diverse antifreeze proteins.” *Current Opinion in Genetics and Development*, 8, 715–720.

Constans, A. (2005), “Pioneering ionization technique paved the way for proteomics.” *The Scientist*, 19, 37–41.

Davies, P. L. (2014), “Ice-binding proteins: A remarkable diversity of structures for stopping and starting ice growth.” *Trends in Biochemical Sciences*, 39, 548–555.

Davies, P. L., J. Baardsnes, M. J. Kuiper, and V. K. Walker (2002), “Structure and function of antifreeze proteins.” *Philosophical Transactions of the Royal Society B: Biological Sciences*, 357, 927–935.

Davies, P. L. and C. L. Hew (1990), “Biochemistry of fish antifreeze proteins.” *The FASEB Journal*, 4, 2460–2468.

Davies, P. L., C. L. Hew, and F. L. Fletcher (1988), “Fish antifreeze proteins: Physiology and evolutionary biology.” *Canadian Journal of Zoology*, 66, 2611–2617.

Delhaise, P., M. Bardiaux, M. De Maeyer, M. Prevost, D. Vanbelle, J. Donneux, I. Lasters, E. Vancustem, P. Alard, and S. Wodak (1988), “The brugel package - toward computer-aided-design of macromolecules.” *Journal of molecular graphics*, 6, 219.

Deng, G., D. W. Andrews, and R. A. Laursen (1997), “Amino acid sequence of a new type of antifreeze protein , from the longhorn sculpin *Myoxocephalus octodecimspinosis*.” *Federation of European Biochemical Societies Letters*, 402, 17–20.

Deng, G. and R. A.” Laursen (1998), “Isolation and characterization of an antifreeze protein from the longhorn sculpin, *Myoxocephalus octodecimspinosis*." *Biochimica et Biophysica Acta*, 1388, 305–314.

DeVries, A. L. (1984), "Role of glycopeptides and peptides in inhibition of crystallization of water in polar fishes." *Philosophical Transactions of the Royal Society of London*, 304, 575–588.

DeVries, A. L., S. K. Komatsu, and R. E. Feeney (1970), "Chemical and physical properties of freezing point-depressing glycoproteins from Antarctic fishes." *Journal of Biological Chemistry*, 245, 2901–2908.

DeVries, A. L. and Y. Lin (1977), "Structure of a peptide antifreeze and mechanism of adsorption to ice." *Biochimica et Biophysica Acta*, 495, 388–392.

DeVries, A. L. and D. Wohlschlag (1969), "Freezing resistance in some Antarctic fishes." *Science*, 163, 1073–1075.

Doucet, D., M. G. Tyshenko, M. J. Kuiper, S. P. Graether, B. D. Sykes, A. J. Daugulis, P. L. Davies, and V. K. Walker (2000), "Structure-function relationships in spruce budworm antifreeze protein revealed by isoform diversity." *European Journal of Biochemistry*, 267, 6082–6088.

Drevin, I. and B.-L. Johansson (1991), "Stability of Superdex 75 prep grade and Superdex 200 prep grade under different chromatographic conditions." *Journal of Chromatography*, 547, 21–30.

Drori, R., Y. Celik, P. L. Davies, and I. Braslavsky (2014), "Ice-binding proteins that accumulate on different ice crystal planes produce distinct thermal hysteresis dynamics." *Journal of The Royal Society Interface*, 11, 10.

Duman, J. and K. Horwath (1983), "The role of hemolymph proteins in the cold tolerance of insects." *Annual review of physiology*, 45, 261–270.

Duman, J. G. (1979), "Subzero temperature tolerance in spiders: The role of thermal hysteresis factors." *Journal of Comparative Physiology*, 131, 347–352.

Duman, J. G. (2001), "Antifreeze and ice nucleator proteins in terrestrial arthropods." *Annual Review of Physiology*, 63, 327–357.

Duman, J. G. and A. L. DeVries (1976), "Isolation, characterization and physical properties of protien antifreezes from the winter flounder, *Pseudopleuronectes americanus*." *Comparative Biochemistry and Physiology*, 533, 375–380.

Duman, J. G. and T. M. Olsen (1993), "Thermal hysteresis protein activity in bacteria, fungi, and phylogenetically diverse plants." *Cryobiology*, 30, 322–328.

Duman, J. H., J. L. Patterson, J. J. Kozak, and A. L. DeVries (1980), "Isopiestic determination of water binding by fish antifreeze glycoproteins." *Biochimica et Biophysica Acta*, 626, 332–336.

Emmert-Streib, F. (2012), "Limitation of gene duplication models: Evolution of modules in protein interaction networks." *Public Library of Science One*, 7, 1–13.

Farewell, A., K. Kvint, and T. Nyström (1998), "*uspB*, a new σ^{S}-regulated gene in *Escherichia coli* which is required for stationary-phase resistance to ethanol." *Journal of Bacteriology*, 180, 6140–6147.

Fleishman, S. J., J. E. Corn, E.-M. Strauch, T. A. Whitehead, J. Karanicolas, and D. Baker (2011), "Hotspot-centric De Novo design of protein binders." *Journal of Molecular Biology*, 413, 1047–1062.

Franks, F. (1982), *The properties of aqueous solutions at subzero temperatures.* Plenum Press, New York. In Water, volume 7: Water and Aqueous Solutions at Subzero Temperatures.

Gagne, M., L. Spyracopoulos, S. P. Graether, Z. Jia, P. L. Davies, and B. D. Sykes (2003), "Spruce budworm antifreeze protein: Changes in structure and dynamics at low temperature." *Journal of Molecular Biology*, 327, 1155–1168.

Garnham, C. P., R. L. Campbell, and P. L. Davies (2011), "Anchored clathrate waters bind antifreeze proteins to ice." *Proceedings of the National Academy of Sciences*, 108, 7363–7367.

Gauthier, S. Y., A. J. Scotter, F.-H. Lin, J. Baardsnes, G. L. Fletcher, and P. L. Davies (2008), "A re-evaluation of the role of type IV antifreeze protein." *Cryobiology*, 57, 292–296.

Gilbert, J. A., P. L. Davies, and J. Laybourn-parry (2005), "A hyperactive , Ca^{2+}-dependent antifreeze protein in an Antarctic bacterium." *FEMS Microbiology Letters*, 245, 67–72.

Gille, C. and C. Frömmel (2001), "STRAP: editor for STRuctural Alignments of Proteins." *Bioinformatics*, 17, 377–378.

Graham, L. A., Y-C. Liou, V. K. Walker, and P. L. Davies (1997), "Hyperactive antifreeze protein from beetles." *Nature*, 388, 727–728.

Griffith, M. and K. V. Ewart (1995), "Antifreeze proteins and their potential use in frozen foods." *Biotechnology advances*, 13, 375–402.

Guilhaus, M. (1995), "Principles and instrumentation in time-of-flight mass spectrometry." *Journal of Mass Spectrometry*, 30, 1519–1532.

Hakim, A., J. B. Nguyen, K. Basu, D. F. Zhu, D. Thakral, P. L. Davies, F. J. Isaacs, Y. Modis, and W. Meng (2013), "Crystal structure of an insect antifreeze protein and its implications for ice binding." *The Journal of Biological Chemistry*, 288, 12295–12304.

Harding, M. M., P. I. Anderberg, and A. D. J. Haymet (2003), "'Antifreeze' glycoproteins from polar fish." *European Journal of Biochemistry*, 270, 1381–1392.

Haschemeyer, A. E. V., W. Guschlbaur, and A. L. DeVries (1977), "Water binding by antifreeze glycoproteins from Antarctic fish." *Nature*, 269, 87–88.

Haymet, A. D. J., L. G. Ward, and M. M. Harding (1999), "Winter flounder "antifreeze" proteins: Synthesis and ice grotwh inhibition of analogues that probe the relative importance of hydrophobic and hydrogen-bonding interactions." *Journal of the American Chemical Society*, 121, 941–948.

Haymet, A. D. J., L. G. Ward, and M. M. Harding (2001), "Hydrophobic analogues of the winter flounder 'antifreeze' protein." *Federation of European Biochemical Societies letters*, 491, 285–288.

Haymet, A. D. J., L. G. Ward, M. M. Harding, and C. A. Knight (1998), "Valine substituted winter flounder 'antifreeze': Preservation of ice growth hysteresis." *Federation of European Biochemical Societies letters*, 430, 301–306.

Hobbs, R. S., M. A. Shears, L. A. Graham, P. L. Davies, and G. L. Fletcher (2011), "Isolation and characterization of type I antifreeze proteins from cunner, *Tautogolabrus adspersus*, order Perciformes." *Federation of European Biochemical Societies*, 278, 3699–3710.

Hyzak, L., R. Moos, F. von Rath, V. Wulf, M. Wirtz, D. Melchior, H.-W. Kling, M. Köhler, S. Gäb, and O. J. Schmitz (2011), "Quantitative matrix-assisted laser desorption ionization-time-of-flight mass spectrometry analysis of synthetic polymers and peptides." *Analytical Chemistry*, 83, 9467–9471.

Isgro, T. A., M. Sotomayor, and E. Cruz-Chu (2014), *Case study: Water and ice.* University of Illinois.

Karim, O. A. and A. D. J. Haymet (1988), "The ice/water interface: A molecular dynamics simulation study." *The Journal of Chemical Physics*, 89, 6889–6896.

Kaufmann, K. W., G. H. Lemmon, S. L. DeLuca, J. H. Sheehan, and J. Meiler (2010), "Practically useful: What the Rosetta Protein Modelling Suite can do for you." *Biochemistry*, 49, 2987–2998.

Knight, C. A. (1967), *The freezing of supercooled liquids.* D. Van Nostrand Company, Inc., New Jersey.

Knight, C. A., E. Driggers, and A. L. DeVries (1993), “Adsorption to ice of fish antifreeze glycopeptides 7 and 8.” *Biophysical Journal*, 64, 252–259.

Kondo, H., Y. Hanada, H. Sugimoto, T. Hoshino, C. P. Garnham, P. L. Davies, and S. Tsuda (2012), “Ice-binding site of snow mold fungus antifreeze protein deviates from structural regularity and high conservation.” *Proceedings of the National Academy of Sciences*, 109, 9360–9365.

Koushafar, H., L. Pham, C. Lee, and B. Rubinsky (1997), “Chemical adjuvant cryosurgery with antifreeze proteins.” *Journal of Surgical Oncology*, 66, 114–121.

Lauersen, K. J., A. Brown, A. Middleton, P. L. Davies, and V. K. Walker (2011), “Expression and characterization of an antifreeze protein from the perennial rye grass, *Lolium perenne*.” *Crybiology*, 62, 194–201.

Leinala, E. K., P. L. Davies, D. Doucet, M. G. Tyshenko, V. K. Walker, and Z. Jia (2002a), “A β-helical antifreeze protein isoform with increased activity.” *The Journal of Biological Chemistry*, 277, 33349–33352.

Leinala, E. K., P. L. Davies, and Z. Jia (2002b), “Crystal structure of β-helical antifreeze protein points to a general ice binding model.” *Structure*, 10, 619–627.

Li, C. and C. Jin (2004), “Letters to the Editor: ^{1}H, ^{13}C and ^{15}N resonance assignments of the antifreeze protein cfAFP-501 from spruce budworm at different temperatures.” *Journal of Biomolecular NMR*, 30, 101–102.

Liou, Y.-C., A. Tocilj, P. L. Davies, and Z. Jia (2000), “Mimicry of ice structure by surface hydroxyls and water of a β-helix antifreeze protein.” *Nature*, 406, 322–324.

Liu, Y., Z. Li, Q. Lin, J. Kosinski, J. Seetharaman, J. M. Bujnicki, Sivaraman J., and C.-L. Hew (2007), “Structure and evolutionary origin of calcium dependent herring type II antifreeze protein.” *Public Library of Science ONE*, 2, 1–11.

Louis-Jeune, C., M. A. Andrade-Navarro, and C. Perez-Iratxeta (2011), “Prediction of protein secondary structure from circular dichroism using theoretically derived spectra.” *Proteins: Structure, Function, and Bioinformatics*, 80, 374–381.

Middleton, A. J., A. M. Brown, P. L. Davies, and V. K. Walker (2009), “Identification of the ice-binding face of a plant antifreeze protein.” *Federation of European Biochemical Societies Letters*, 583, 815–819.

Middleton, A. J., C. B. Marshall, F. Faucher, M. Bardolev, I. Braslavsky, R. L. Campbell, V. K. Walker, and P. L. Davies (2012), “Antifreeze protein from freeze-tolerant grass has a beta-roll fold with an irregularly structured ice-binding site.” *Journal of Molecular Biology*, 416, 713–724.

Mok, Y.-F., F.-H. Lin, L. A. Graham, Y. Celik, I. Braslavsky, and P. L. Davies (2010), “Structural basis for the superior activity of the large isoform of snow flea antifreeze protein.” *Biochemistry*, 49, 2593–2603.

Muldrew, K., J. Rewcastle, B. J. Donnelly, J. C. Saliken, S. Liang, S. Goldie, M. Olson, R. Baissalov, and G. Sandison (2001), “Flounder antifreeze peptides increase the efficacy of cryosurgery.” *Cryobiology*, 42, 182–189.

Nishimiya, Y., H. Kondo, M. Takamichi, H. Sugimoto, M. Suzuki, A. Miura, and S. Tsuda (2008), “Crystal structure and mutational analysis of calcium-independent type II antifreeze protein from longsnout poacher, *Brachyopsis rostratus*.” *Journal of Molecular Biology*, 382, 734–746.

Novagen (2003), *pET system manual*, 10 edition.

Novagen (n.d.), “Competent cells: What a difference a strain makes.” Merck KGaA, Darmstadt, Germany.

Ochlal, E.-L. (1991), “Biomineralization principle.” *Principles and Applications in Bioinorganic Chemistry*, 68, 627–630.

Perez-Iratxeta, C. and M. A. Andrade-Navarro (2008), "K2D2: Estimation of protein secondary structure from circular dichroism spectra." *BioMed Central Structural Biology*, 8, 1–5.

Pupko, T., I. Pe'er, M. Hasegawa, D. Graur, and N. Friedman (2002), "A branch-and-bound algorithm for the for the inference of ancestral amino-acids sequences when the replacement rate varies among sites: Application to the evolution of five gene families." *Bioinformatics*, 18, 1116–1123.

Pupko, T., I. Pe'er, R. Shamir, and D. Graur (2000), "A fast algorithm for join reconstruction of ancestral amino acids sequences." *Molecular Biology and Evolution*, 17, 890–896.

Rath, A., M. Glibowicka, V. G. Nadeau, G. Chen, and C. M. Deber (2009), "Detergent binding explains anomalous SDS-PAGE migration of membrane proteins." *Proceedings of the Natoinal Academy of Sciences*, 106, 1760–1765.

Raymond, J. A. (1976), *Adsorption inhibition as a mechanism of freezing resistance in polar fishes.* Ph.D. thesis, University of California.

Raymond, J. A. (2000), "Distribution and partial characterization of ice-active molecules associated with sea-ice diatoms." *Polar Biology*, 23, 721–729.

Raymond, J. A. and A. L. DeVries (1977), "Adsorption inhibition as a mechanism of freezing resistance in polar fishes." *Proceedings of the National Academy of Sciences of the United States of America*, 74, 2589–93.

Rosano, G. L. and E. A. Ceccarelli (2014), "Recombinant protein expression in Escherichia coli: Advances and challenges." *Frontiers in Microbiology*, 5, 1–17.

Sanders, C. J. (1991), *Biology of North American spruce budworms.* Elsevlier, Amsterdam. In tortricid pests, Their biology, natural enemies and control, volume 7: Tortricids in forestry.

Scheraga, G. A., G. Nemethy, and I. Z. Steinberg (1962), "The contribution of hydrophobic bonds to the thermal stability of protein conformations." *Journal of Biological Chemistry*, 237, 2506–2508.

Schrödinger, LLC (2015), "The PyMOL molecular graphics system, version 1.8."

KEY: PyMOL
ANNOTATION: PyMOL The PyMOL Molecular Graphics System, Version 1.8, Schrödinger, LLC.

Scotter, A. J., C. B. Marshall, L. A. Graham, J. A. Gilbert, C. P. Garnham, and P. L. Davies (2006), "The basis for hyperactivity of antifreeze proteins." *Cryobiology*, 53, 229–239.

Shier, W. T., Y. Lin, and A. L. DeVries (1972), "Structure and mode of action of glycoproteins from an Antarctic fish." *Biochimica et Biophysica Acta*, 263, 406–413.

Sönnichsen, F., B. Sykes, H. Chao, and P. L. Davies (1993), "The nonhelical structure of antifreeze protein type III." *Science*, 259, 1154–1157.

Sun, T., F.-H. Lin, R. L. Campbell, J. S. Allingham, and P. L. Davies (2014), "An antifreeze proteins folds with an interior network of more than 400 semi-clathrate waters." *Science*, 343, 795–798.

Tablin, F., A. E. Oliver, N. J. Walker, L. M. Crowe, and J. H. Crowe (1996), "Membrane phase transition of intact human platelets: Correlation with cold-induced activation." *Journal of Cellular Physiology*, 168, 305–313.

Taylor, R. G., D. C. Walker, and R. Mclnnes (1993), "*E.coli* host strains significantly affect the quality of small scale plasmid DNA preparations used for sequencing." *Nucleic Acids Research*, 21, 1677–1678.

Teeter, M. M. (1984), "Water structure of a hydrophobic protein at atomic resolution: Pentagon rings of water molecules in crystals of crambin." *Proceedings of the National Academy of Sciences*, 81, 6014–6018.

Tursman, D., J. G. Duman, and C. A. Knight (1994), "Freeze tolerance adaptations in the centipede *Lithobius forficatus*." *Journal of Experimental Zoology*, 268, 347–353.

Tyshenko, M. G., D. Doucet, P. L. Davies, and V. K. Walker (1997), "The antifreeze potential of the spruce budworm thermal hysteresis protein." *Nature Biotechnology*, 15, 887–890.

Voet, A. R. D., H. Noguchi, C. Addy, D. Simoncini, D. Terada, S. Unzai, S.-Y. Park, K. Y. J. Zhang, and J. R. H. Tame (2014), "Computational design of a self-assembling symmetrical β-propeller protein." *Proceedings of the National Academy of Sciences*, 111, 15102–15107.

Voet, A. R. D., H. Noguchi, C. Addy, K. Y. J. Zhang, and J. R. H. Tame (2015), "Biomineralization of a cadmium chloride nanocrystal by a designed symmetrical protein." *Angewandte Chemie International Edition*, 54, 9857–9860.

Wang, T., Q. Zhu, X. Yang, J. R. Layne, and A. L. DeVries (1994), "Antifreeze glycoproteins from Antarctic Notothenioid fishes fail to protect the rat cardiac explant during hypothermic and freezing preservation." *Cryobiology*, 31., 185–192.

Woody, R. W. (1995), "Circular dichroism." *Methods in Enzymology*, 246, 34–71.

Wu, Y., J. Banoub, S. V. Goddard, M. H. Kao, and G. L. Fletcher (2001), "Antifreeze glycoproteins: relationship between molecular weight, thermal hysteresis and the inhibition of leakage from liposomes during thermotropic phase transition." *Comparative Biochemistry and Physiology Part B*, 128, 265–273.

Yeh, Y. and R. E. Feeney (1996), "Antifreeze proteins : Structures and mechanisms of function." *Chemical Reviews*, 96.

Yu, S. O. W. (2010), *Antifreeze proteins: Activity comparisons and de novo design of an ice-binding protein*. Master's thesis, Queen's University, Kingstone, Canada.

Zhang, W. and R. A. Laursen (1998), "Structure-function relationships in a type I antifreeze polypeptide. The role of threonine methyl and hydroxyl groups in antifreeze activity." *The Journal of Biological Chemistry*, 273, 34806–34812.

Zhang, z., S. Schwartz, L. Wagner, and W. Miller (2000), "A greedy algorithm for aligning DNA sequences." *Journal of Computation Biology*, 7, 203–214.

Appendices

Appendix A

Risk analysis

Firstly the rules of the laboratory, in which the work is conducted, will be followed. These rules are dependent on which work is mainly conducted in these labs and the presence of special lab appliances. When working with a sonicator, e.g., safety glasses and ear protection must be worn. Anyone else in the same room should be wearing safety glasses and ear protection as well.

Most of the frequently handled products during this master thesis, such as Trizma™ base, SDS, and ethanol, can be irritating for the eyes, skin and/or the respiratory tract. For this reasons, gloves and safety glasses are highly recommended and mandatory in most of the used labs.

Although some products are used without immediate risks, such as NaCl and glycine, gloves and safety glasses are still mandatory as these products are usually components of a buffer, or are used in labs where it is mandatory.

Before experiments are started, the potential risks and the safety measurements of each step should be validated. This allows the extra care for more dangerous products or extra attention to inflammable products, but by doing so, the risks and treatments are known for when something does happen. Ethidium bromide, e.g., is a potential carcinogen from which it is thought to be genotoxic, a frame-shift mutagen and teratogen. One of the labs has a special zone for ethidium bromide and equipment that can be used for ethidium bromide stained gels. When leaving this zone, gloves should be thrown away and the hands need to be washed before taking new gloves.

During this master thesis, lab work will be conducted with genetically modified *E. coli* strains DH5α and BL21 (DE3). The strains are considered non-pathogenic, as they are modified in such a way that they do not contain the pathogenic mechanisms responsible for most of the enteric infections. The introduced pET plasmids contain genes coding for AFPs, which do not provide any fitness advantage to the bacterium, as the cells are unable to survive out of the laboratory and expressed AFPs only have an advantage during freezing periods. The cells are either subjected to lysis or are disposed after autoclaving the culture with cells.

Appendix B

Products, appliances, and additional figures

B.1 Composition of buffers and products

Bradford dye reagent: Dissolve 25 mg of Coomassie Brilliant Blue G-250 in 12.5 ml 95% ethanol. Add 25 ml of 85% w/v H_3PO_4 and dilute to a final volume of 250 ml with Milli-Q. Pure ethanol is from Fisher Scientific UK, Leicestershire, UK, other products are from Sigma-Aldrich Chemie, Steinheim, Germany.
Destaining solution: 50% methanol (Fisher Scientific UK, Leicestershire, UK), 10% acetic acid (VWR international S.A.S, Fontenay-Sous-Bois, France).
Dialysis solution for LpIBP: 5 mM Tris (Trizma™ base, Sigma-Aldrich Chemie, Steinheim, Germany), pH 7.8.
DNase stock solution: A stock solution of 2 M DNase I (Roche Diagnostics, Mannheim, Germany).
Elution buffer for LpIBP: 0.5 M NaCl (Fisher Scientific UK, Leicestershire, UK), 50 mM Tris, 250 mM imidazole (Alfa Aesar GmbH and Co KG, Karlsruhe, Germany), pH 8.0.
Elution buffer for MpAFP: 0.5 M NaCl, 50 mM Tris, 2% (v/v) glycerol, 2 mM $CaCl_2$, 250 mM imidazole, pH 7.5.
Laemmli buffer containing β-mercaptoethanol: 3.55 mL deionized water, 1.25 mL 0.5 M Tris, 2.5 mL glycerol (Acros Organics, Geel, Belgium), 2.0 mL 10% (w/v) SDS (Acros Organics, Geel, Belgium), 0.2 mL 0.5% (w/v) bromophenol blue (Sigma-Aldrich Chemie, Steinheim, Germany), 0.5 mL β-mercaptoethanol (Sigma-Aldrich Chemie, Steinheim, Germany), pH 6.8.
Lysozyme stock solution: A stock solution of 3.33 M lysozyme (Sigma-Aldrich Chemie, Steinheim, Germany).
$MgCl_2$ stock solution: A stock solution of 12.95 mM $MgCl_2$ (VWR international S.A.S,

Fontenay-Sous-Bois, France) in DPBS (Gibco™, Fisher Scientific UK, Leicestershire, UK).

Protease inhibitor stock solution: Dissolve one cOmplete Protease Inhibitor Cocktail™ tablet (Roche Diagnostics, Mannheim, Germany) in 10 mL DPBS.

Reference protein solution: 10 mM Tris, 0.2 M NaCl, 0.075 mM bovine serum albumin (BSA), pH 8.0. NaCl is from Fisher Scientific UK, Leicestershire, UK, the other products from Sigma-Aldrich Chemie, Steinheim, Germany.

Resuspension buffer for MpAFP: 50 mM tris, 150 mM NaCl, 2 mM $CaCl_2$, pH 7.5.

SDS running buffer (10x): 0.25 M Tris, 1.9 M glycine (Sigma-Aldrich Chemie, Steinheim, Germany), 0.1% SDS, pH 8.0.

Staining solution: 0.05% Coomassie Brilliant Blue G-250, 50% methanol (Fisher Scientific UK, Leicestershire, UK), 10% acetic acid (VWR international S.A.S, Fontenay-Sous-Bois, France).

GFP_{UV} storage buffer: 10 mM Tris, 0.2 M NaCl, pH 8.0.

TAE buffer (1x): 40 mM Tris, 20 mM acetate (Acros Organics, Geel, Belgium), 1 mM EDTA (Acros Organics, Geel, Belgium), pH around 8.6. Typically a 50x stock solution is prepared.

Wash buffer 1 for LpIBP: 0.5 M NaCl, 50 mM Tris, 2% (v/v) glycerol, 5 mM imidazole (Alfa Aesar GmbH and Co KG, Karlsruhe, Germany), pH 8.0.

Wash buffer 2 for LpIBP: 0.5 M NaCl, 50 mM Tris, 2% (v/v) glycerol, 20 mM imidazole, pH 8.0.

Wash buffer 1 for MpAFP: 0.5 M NaCl, 50 mM Tris, 2% (v/v) glycerol, 2 mM $CaCl_2$, 5 mM imidazole, pH 7.5.

Wash buffer 2 for MpAFP: 0.5 M NaCl, 50 mM Tris, 2% (v/v) glycerol, 2 mM $CaCl_2$, 12.5 mM imidazole, pH 7.5.

Wash buffer 3 for MpAFP: 0.5 M NaCl, 50 mM Tris, 2% (v/v) glycerol, 2 mM $CaCl_2$, 20 mM imidazole, pH 7.5.

B.2 Remaining appliances

Autoclave: Astell scientific (Kent, United Kingdom)

MALDI TOF/TOF MS: Applied Biosystems 4800 MALDI TOF/TOF Mass Spectrometer (Applied Biosystems, Foster City, CA, USA)

pH meter: 713 pH Meter (Metrohm, Antwerp, Belgium)

Table centrifuge: Eppendorf™ MiniSpin™ (Thermo Fisher Scientific, Bleiswijk, The Netherlands) and Galaxy 14D centrifuge (VWR international S.A.S, Fontenay-Sous-Bois, France)

B.3 Scripts

B.3.1 Brugel

The following script was used for the backbone manipulations, as is described in section 3.1.4

```
lib_read(pdbnoth2001)
pdb_read(template.pdb)
pdb_read(23.pdb)

proc_define(create_rep/o,rep:mask:out, start1:residue:in,end1:residue:in
    , start2:residue:in,end2:residue:in, start3:residue:in,end3:residue:
    in)
    m_segment(@@rep_start, @start1, @end1)
    m_segment(@@rep_middle, @start2, @end2)
    m_segment(@@rep_end, @start3, @end3)
    m_or(@@rept, @@rep_start, @@rep_middle)
    m_or(@rep, @@rept, @@rep_end)
proc_end

proc_define(define_reps/o)
    m_segment(capa,a‘start’,a‘end’)
    m_segment(capb,a‘start’,a‘end’)
    m_segment(‘repeatx’,a‘start’,a‘end’)
    create_rep(‘repeatx’,a‘start’,a‘end’,a‘start’,a‘end’,a‘start’,a‘end
    ’)
    create_rep(‘repeatx’,a‘start’,a‘end’,a‘start’,a‘end’,a‘start’,a‘end
    ’)
proc_end
define_reps

fit(mat_back/o,{‘repeatx’,main,%m_and},1,{‘repeatx-1’,main,%m_and},1)
fit(mat_front/o,{‘repeatx-1’,main,%m_and},1,{‘repeatx’,main,%m_and},1)

mat_apply(mat_front,all,1)
pdb_write(‘last-repeat+1’.pdb,‘last-repeat’,{%set_last})
pdb_write(‘capb+1’.pdb,capb,{%set_last})

modfil_clear
pdb_read(capa.pdb)
pdb_read(‘all-repeats’.pdb)
pdb_read(‘last-repeat+1’.pdb)
pdb_read(‘capb+1’.pdb)

udc_renum(a‘first’,{all,%ml_nares},‘first’,a,off)
chain_rename(all,a,0a)
chain_rename(all,a,0b)
chain_rename(all,a,0c)
pdb_write(‘final-structure’.pdb,all,1)
modfil_clear

pdb_read(‘final-structure’.pdb)
m_describe(all)
modfil_clear
```

B.3.2 PyRosetta

To run the PyRosetta script with default options, the following command can be used:

```
python sequence_mapping.py --backbone input_backbone.pdb --sequences
    input_sequences.fasta
```

While the following command can be used to receive more information about the available options:

```
python sequence_mapping.py -help
```

The following script is used to superimpose and optimize the sequences on the protein backbone using PyRosetta:

```
#!/usr/bin/env python

import os
import optparse
import re
#from rosetta import *
from rosetta import Pose
from rosetta import pose_from_pdb
from toolbox import *

one_to_three = {’A’: ’ALA’,
    ’R’: ’ARG’,
    ’N’: ’ASN’,
    ’D’: ’ASP’,
    ’C’: ’CYS’,
    ’E’: ’GLU’,
    ’Q’: ’GLN’,
    ’G’: ’GLY’,
    ’H’: ’HIS’,
    ’I’: ’ILE’,
    ’L’: ’LEU’,
    ’K’: ’LYS’,
    ’M’: ’MET’,
    ’F’: ’PHE’,
    ’P’: ’PRO’,
    ’S’: ’SER’,
    ’T’: ’THR’,
   'W': 'TRP',
   'Y': 'TYR',
   'V': 'VAL',
   }

def sequence_mapping(pdb_file, sequence_file, score_file, relax, jobs):
   if os.path.exists( os.getcwd() + '/' + pdb_file ) and pdb_file:
      init()
      pose = Pose()
      score_fxn = create_score_function('talaris2013')
   if (relax):
      refinement = FastRelax(score_fxn)
      pose_from_pdb(pose, pdb_file)
   if os.path.exists( os.getcwd() + '/' + sequence_file ) and
       sequence_file:
      fid = open(sequence_file,'r')
      fod = open(score_file,'w')
      data = fid.readlines()
      fid.close()
      sequences = []
      read_seq = False
      for i in data:
   if not len(i):
      continue
   elif i[0] == '>':
      read_seq = True
      fasta_line = re.split(':|\s+|\||\\n',i[1:])
      name_cpt=0
      while (name_cpt<len(fasta_line) and not fasta_line[name_cpt]):
         name_cpt+=1
   if name_cpt<len(fasta_line):
      job_output = fasta_line[name_cpt]
   else:
      print 'Error: Please enter an identifier for sequences in your
         fasta file'
      exit(1)
   elif read_seq:
      seq=list(i)
      resn=1
      for j in i:
         if j!='\n' and resn<=pose.total_residue():
         mutator = MutateResidue( resn , one_to_three[j] )
         mutator.apply( pose )
         resn+=1
      elif resn>pose.total_residue():
         print 'WARNING: couldn\'t mutate residue number '+str(resn)+',
              sequence too long for backbone...'
      resn+=1
  if (relax):
    jd = PyJobDistributor(job_output, jobs, score_fxn)
    jd.native_pose = pose
    scores = [0]*(jobs)
    counter = 0
    decoy=Pose()
    while not jd.job_complete:
      decoy.assign(pose)
      resn=1
      refinement.apply(decoy)
      jd.output_decoy(decoy)
      scores[counter]=score_fxn(decoy)
      counter+=1
    for i in range(0, len(scores)):
      fod.writelines(job_output + '_' + str(i+1) + ' : '+str(scores[
        i])+'\n')
    else:
      pose_packer = standard_packer_task(pose)
      pose_packer.restrict_to_repacking()
      packmover = PackRotamersMover(score_fxn, pose_packer)
      packmover.apply(pose)
      fod.writelines(job_output+' : '+str(score_fxn(pose))+'\n')
      pose.dump_pdb(job_output+'_1.pdb')
    else:
      print 'Bad fasta format'
      exit(1)
      fod.close()
    else:
      print 'Please provide a valid sequence file, '+sequence_file+'
          doesn\'t exist'
    else:
      print 'Please provide a valid backbone file, '+pdb_file+'
        doesn\'t exist'

parser=optparse.OptionParser()
parser.add_option('--backbone', dest = 'pdb_file',
default = '',
help = 'the backbone in PDB format' )

parser.add_option('--sequences', dest = 'seq_file',
default = '',
help = 'the sequences to map' )

parser.add_option('--out', dest = 'score_out',
default = 'scores.sc',
help = 'the score file to output' )
```

```
parser.add_option(’−−clean’, action=”store_true”, dest = ’clean_pdb’,
default = False,
help = ’makes the pdb Rosetta friendly’ )

parser.add_option(’−−no_relax’,action=”store_false”, dest = ’relax’,
default = True,
help = ’no relaxation after sequence mapping’ )

parser.add_option(’−−nstruct’, dest = ’jobs’,
default = ’1’,
help = ’number of relaxations per sequence’ )

(options,args) = parser.parse_args()

pdb_file=options.pdb_file
sequence_file = options.seq_file
score_file=options.score_out
clean_pdb=options.clean_pdb
relax=options.relax
jobs=int(options.jobs)

if clean_pdb:
    cleanATOM( pdb_file )
    sequence_mapping(pdb_file[:−4]+’.clean.pdb’, sequence_file,
        score_file, relax, jobs)
else:
    sequence_mapping(pdb_file, sequence_file, score_file, relax, jobs)
```

B.3.3 Formation of double repeats

The following code creates double repeating units from sequences of previously conducted ancestral sequence reconstruction and prevents the formation of identical double repeats in one unit.

```
#!/usr/bin/python
# −*− coding: utf−8 −*−

import sys

try:
    f1 = sys.argv[1]
    f2 = sys.argv[2]
except:
    print ”Sequence files required”
    sys.exit(1)
```

```
seqs1 = []
seqs2 = []

for i in open(f1):
    seqs1.append(i.strip())

for i in open(f2):
    seqs2.append(i.strip())

for a, i in enumerate(seqs1):
    for b, j in enumerate(seqs2):
        if i == j:
            pass
        print ”>”, a+1, ”_” , b+1
        print i+j
```

B.4 Blast results

When the LGC genomics finishes sequencing, a DNA sequence is send back. This sequence can be compared with the original designed sequence through NCBI's nucleotide BLAST (Zhang et al., 2000). During this nucleotide BLAST two sequences are aligned and different scores are shown. Normally the sequenced DNA should be a complete match with the designed sequence, although mutations can happen.

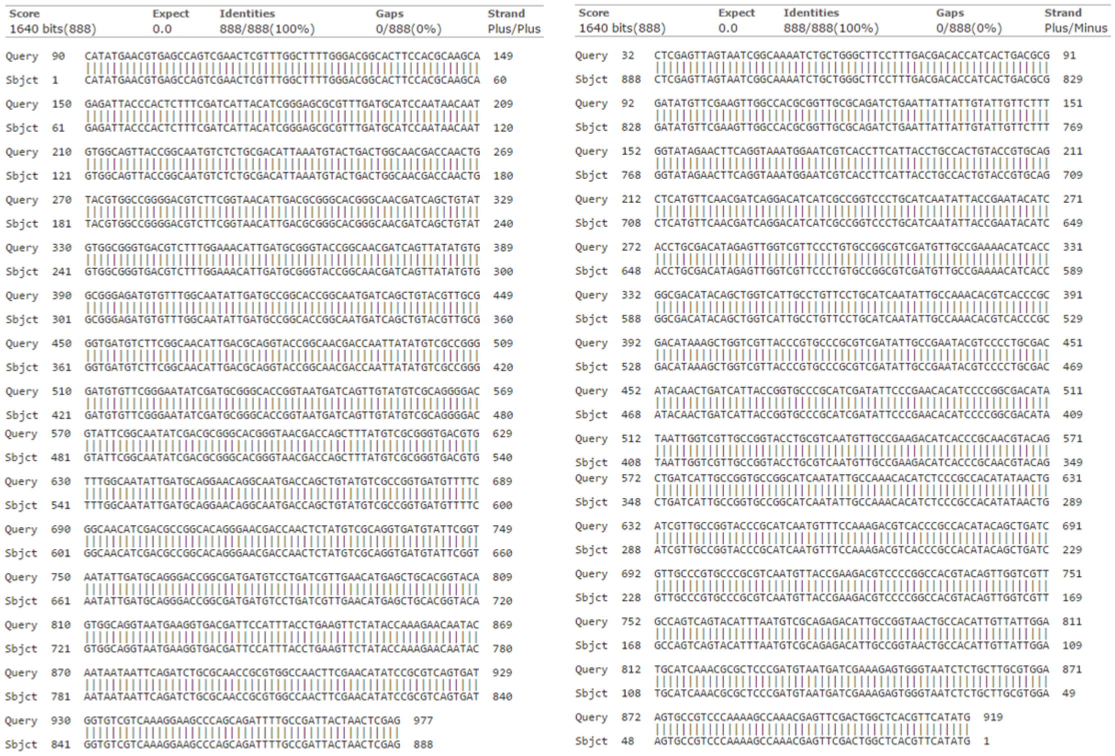
The query ID from every BLAST can be found back in table B.1 and B.2, while figure B.1 shows the alignment results of a BLAST, in this case the alignment of Calippo9.

Table B.1: ID of the query blasts from the redesigned proteins part I

Lolly3	**Lolly4**	**Calippo6**	**Calippo7**
Query_97051	Query_10389	Query_110645	Query_137053
Query_21401	Query_38481	Query_142809	Query_43043
Query_67775	Query_171797		Query_135829
Query_93535			

Table B.2: ID of the query blasts from the redesigned proteins part II

Calippo8	**Calippo9**	**Calippo10**	**Calippo11**
Query_245161	Query_22781	Query_159055	Query_79153
Query_26157	Query_180055	Query_60897	Query_4153
Query_232349			



```
Score            Expect  Identities     Gaps       Strand
1640 bits(888)   0.0     888/888(100%)  0/888(0%)  Plus/Plus

Query  90   CATATGAACGTGAGCCAGTCGAACTCGTTTGGCTTTTGGGACGGCACTTCCACGCAAGCA  149
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    CATATGAACGTGAGCCAGTCGAACTCGTTTGGCTTTTGGGACGGCACTTCCACGCAAGCA  60

Query  150  GAGATTACCCACTCTTTCGATCATTACATCGGGAGCGCGTTTGATGCATCCAATAACAAT  209
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   GAGATTACCCACTCTTTCGATCATTACATCGGGAGCGCGTTTGATGCATCCAATAACAAT  120

Query  210  GTGGCAGTTACCGGCAATGTCTCTGCGACATTAAATGTACTGACTGGCAACGACCAACTG  269
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  GTGGCAGTTACCGGCAATGTCTCTGCGACATTAAATGTACTGACTGGCAACGACCAACTG  180

Query  270  TACGTGGCCGGGGACGTCTTCGGTAACATTGACGCGGGCACGGGCAACGATCAGCTGTAT  329
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181  TACGTGGCCGGGGACGTCTTCGGTAACATTGACGCGGGCACGGGCAACGATCAGCTGTAT  240

Query  330  GTGGCGGGTGACGTCTTTGGAAACATTGATGCGGGTACCGGCAACGATCAGTTATATGTG  389
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241  GTGGCGGGTGACGTCTTTGGAAACATTGATGCGGGTACCGGCAACGATCAGTTATATGTG  300

Query  390  GCGGGAGATGTGTTTGGCAATATTGATGCCGGCACCGGCAATGATCAGCTGTACGTTGCG  449
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301  GCGGGAGATGTGTTTGGCAATATTGATGCCGGCACCGGCAATGATCAGCTGTACGTTGCG  360

Query  450  GGTGATGTCTTCGGCAACATTGACGCAGGTACCGGCAACGACCAATTATATGTCGCCGGG  509
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361  GGTGATGTCTTCGGCAACATTGACGCAGGTACCGGCAACGACCAATTATATGTCGCCGGG  420

Query  510  GATGTGTTCGGGAATATCGATGCGGGCACCGGTAATGATCAGTTGTATGTCGCAGGGGAC  569
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421  GATGTGTTCGGGAATATCGATGCGGGCACCGGTAATGATCAGTTGTATGTCGCAGGGGAC  480
Query  570  GTATTCGGCAATATCGACGCGGGCACGGGTAACGACCAGCTTTATGTCGCGGGTGACGTG  629
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  481  GTATTCGGCAATATCGACGCGGGCACGGGTAACGACCAGCTTTATGTCGCGGGTGACGTG  540

Query  630  TTTGGCAATATTGATGCAGGAACAGGCAATGACCAGCTGTATGTCGCCGGTGATGTTTTC  689
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  541  TTTGGCAATATTGATGCAGGAACAGGCAATGACCAGCTGTATGTCGCCGGTGATGTTTTC  600

Query  690  GGCAACATCGACGCCGGCACAGGGAACGACCAACTCTATGTCGCAGGTGATGTATTCGGT  749
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  601  GGCAACATCGACGCCGGCACAGGGAACGACCAACTCTATGTCGCAGGTGATGTATTCGGT  660

Query  750  AATATTGATGCAGGGACCGGCGATGATGTCCTGATCGTTGAACATGAGCTGCACGGTACA  809
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  661  AATATTGATGCAGGGACCGGCGATGATGTCCTGATCGTTGAACATGAGCTGCACGGTACA  720

Query  810  GTGGCAGGTAATGAAGGTGACGATTCCATTTACCTGAAGTTCTATACCAAAGAACAATAC  869
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  721  GTGGCAGGTAATGAAGGTGACGATTCCATTTACCTGAAGTTCTATACCAAAGAACAATAC  780

Query  870  AATAATAATTCAGATCTGCGCAACCGCGTGGCCAACTTCGAACATATCCGCGTCAGTGAT  929
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  781  AATAATAATTCAGATCTGCGCAACCGCGTGGCCAACTTCGAACATATCCGCGTCAGTGAT  840

Query  930  GGTGTCGTCAAAGGAAGCCCAGCAGATTTTGCCGATTACTAACTCGAG  977
            ||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  841  GGTGTCGTCAAAGGAAGCCCAGCAGATTTTGCCGATTACTAACTCGAG  888
```

```
Score            Expect  Identities     Gaps       Strand
1640 bits(888)   0.0     888/888(100%)  0/888(0%)  Plus/Minus

Query  32   CTCGAGTTAGTAATCGGCAAAATCTGCTGGGCTTCCTTTGACGACACCATCACTGACGCG  91
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  888  CTCGAGTTAGTAATCGGCAAAATCTGCTGGGCTTCCTTTGACGACACCATCACTGACGCG  829

Query  92   GATATGTTCGAAGTTGGCCACGCGGTTGCGCAGATCTGAATTATTATTGTATTGTTCTTT  151
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  828  GATATGTTCGAAGTTGGCCACGCGGTTGCGCAGATCTGAATTATTATTGTATTGTTCTTT  769

Query  152  GGTATAGAACTTCAGGTAAATGGAATCGTCACCTTCATTACCTGCCACTGTACCGTGCAG  211
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  768  GGTATAGAACTTCAGGTAAATGGAATCGTCACCTTCATTACCTGCCACTGTACCGTGCAG  709

Query  212  CTCATGTTCAACGATCAGGACATCATCGCCGGTCCCTGCATCAATATTACCGAATACATC  271
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  708  CTCATGTTCAACGATCAGGACATCATCGCCGGTCCCTGCATCAATATTACCGAATACATC  649

Query  272  ACCTGCGACATAGAGTTGGTCGTTCCCTGTGCCGGCGTCGATGTTGCCGAAAACATCACC  331
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  648  ACCTGCGACATAGAGTTGGTCGTTCCCTGTGCCGGCGTCGATGTTGCCGAAAACATCACC  589

Query  332  GGCGACATACAGCTGGTCATTGCCTGTTCCTGCATCAATATTGCCAAACACGTCACCCGC  391
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  588  GGCGACATACAGCTGGTCATTGCCTGTTCCTGCATCAATATTGCCAAACACGTCACCCGC  529

Query  392  GACATAAAGCTGGTCGTTACCCGTGCCCGCGTCGATATTGCCGAATACGTCCCCTGCGAC  451
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  528  GACATAAAGCTGGTCGTTACCCGTGCCCGCGTCGATATTGCCGAATACGTCCCCTGCGAC  469

Query  452  ATACAACTGATCATTACCGGTGCCCGCATCGATATTCCCGAACACATCCCCGGCGACATA  511
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  468  ATACAACTGATCATTACCGGTGCCCGCATCGATATTCCCGAACACATCCCCGGCGACATA  409

Query  512  TAATTGGTCGTTGCCGGTACCTGCGTCAATGTTGCCGAAGACATCACCCGCAACGTACAG  571
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  408  TAATTGGTCGTTGCCGGTACCTGCGTCAATGTTGCCGAAGACATCACCCGCAACGTACAG  349
Query  572  CTGATCATTGCCGGTGCCGGCATCAATATTGCCAAACACATCTCCCGCCACATATAACTG  631
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  348  CTGATCATTGCCGGTGCCGGCATCAATATTGCCAAACACATCTCCCGCCACATATAACTG  289

Query  632  ATCGTTGCCGGTACCCGCATCAATGTTTCCAAAGACGTCACCCGCCACATACAGCTGATC  691
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  288  ATCGTTGCCGGTACCCGCATCAATGTTTCCAAAGACGTCACCCGCCACATACAGCTGATC  229

Query  692  GTTGCCCGTGCCCGCGTCAATGTTACCGAAGACGTCCCCGGCCACGTACAGTTGGTCGTT  751
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  228  GTTGCCCGTGCCCGCGTCAATGTTACCGAAGACGTCCCCGGCCACGTACAGTTGGTCGTT  169

Query  752  GCCAGTCAGTACATTTAATGTCGCAGAGACATTGCCGGTAACTGCCACATTGTTATTGGA  811
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  168  GCCAGTCAGTACATTTAATGTCGCAGAGACATTGCCGGTAACTGCCACATTGTTATTGGA  109

Query  812  TGCATCAAACGCGCTCCCGATGTAATGATCGAAAGAGTGGGTAATCTCTGCTTGCGTGGA  871
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  108  TGCATCAAACGCGCTCCCGATGTAATGATCGAAAGAGTGGGTAATCTCTGCTTGCGTGGA  49

Query  872  AGTGCCGTCCCAAAAGCCAAACGAGTTCGACTGGCTCACGTTCATATG  919
            ||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  48   AGTGCCGTCCCAAAAGCCAAACGAGTTCGACTGGCTCACGTTCATATG  1
```

Figure B.1: **NCBI's nucleotide BLAST of the ordered Calippo9 sequence compared to the results from sequencing.** The blast shows the alignment between both sequences. For the first alignment the promoter was used as primer, while for the second alignment the terminator primer was used. The first row shows the characteristics of the alignment, such as the amount of identities and gaps. For further use, it is important that both sequences are identical and that no gaps occur.

B.5 MALDI TOF MS Spectra

The mass of the proteins, present in the purified samples, was analysed via MALDI TOF MS. The following figures show the MS spectra for the WT LpIBP, Lolly3, and Lolly3 after further purification via the Superdex 75, respectively. Figure B.2 shows the WT LpIBP, where one clear peak is present and corresponds to the mass of the LpIBP. Some minor peaks are still present, indicating that it is not fully purified. However, both figure B.3 and figure B.4 show a set of peaks that do not correspond to the expected mass of Lolly3. They also contain broad peaks and a bad resolution. Lolly4, not shown here, showed very similar results to Lolly3.

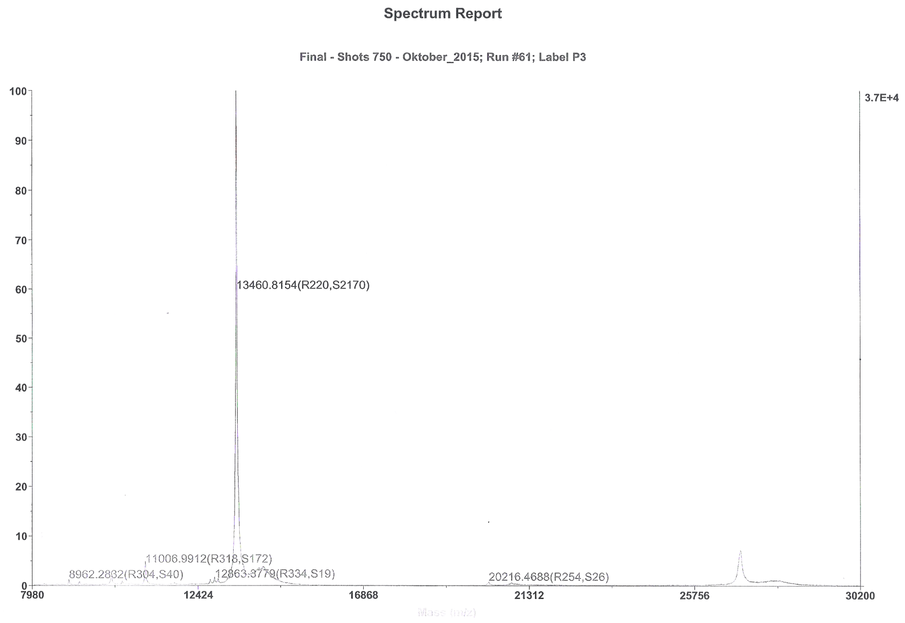


Figure B.2: **MALDI TOF spectrum of the WT LpIBP.** Analysis of the WT LpIBP was performed on an Applied Biosystems 4800 MALDI TOF/TOF Mass Spectrometer (Applied Biosystems, Foster City, CA, USA). 5 mg/mL α-cyano-4- hydroxycinnamic acid (CHCA) was used as a matrix in 50/50 acetonitrile/water with 0.1% trifluoroacetic acid (TFA).

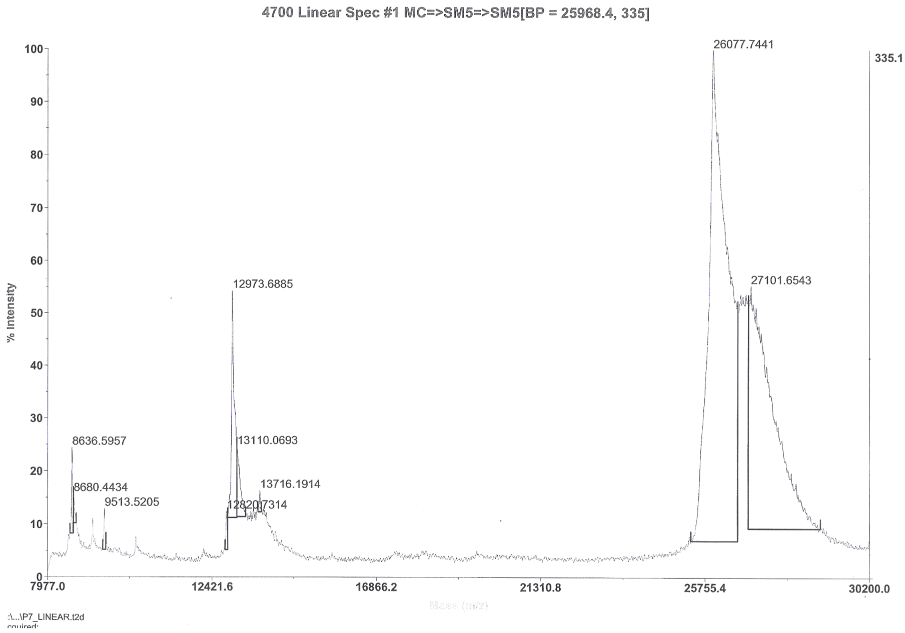


Figure B.3: **MALDI TOF spectrum of the redesigned Lolly**3**.** Analysis of the redesigned Lolly3 was performed on an Applied Biosystems 4800 MALDI TOF/TOF Mass Spectrometer (Applied Biosystems, Foster City, CA, USA). 5 mg/mL α-cyano-4- hydroxycinnamic acid (CHCA) was used as a matrix in 50/50 acetonitrile/water with 0.1% trifluoroacetic acid (TFA).

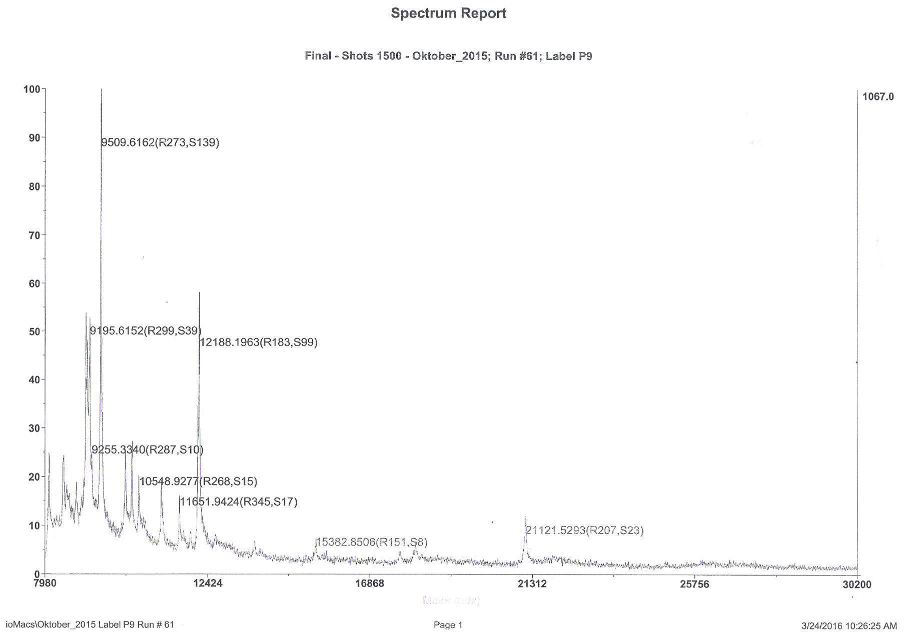


Figure B.4: **MALDI TOF spectrum of the redesigned Lolly**3 **after extra purification.** Analysis of the redesigned Lolly3 after extra purification was performed on an Applied Biosystems 4800 MALDI TOF/TOF Mass Spectrometer (Applied Biosystems, Foster City, CA, USA). 5 mg/mL α-cyano-4- hydroxycinnamic acid (CHCA) was used as a matrix in 50/50 acetonitrile/water with 0.1% trifluoroacetic acid (TFA).

Laboratorium Biomoleculaire Modelling
Celestijnenlaan 200 G
3001 LEUVEN, BELGIË
tel. + 32 16 32 75 21
fax + 32 16 32 79 74
www.kuleuven.be