Biochemical, biophysical, and structural studies of seed proteins from *Moringa oleifera* and implications for traditional water purification

Martine Moulin

Faculty of Natural Sciences

Doctor of Philosophy

March 2019

A mon père,

Abstract

Moringa oleifera is a tropical plant that belongs to the Moringaceae family and is native to northern India. Moringa *oleifera* is called *The Miracle Tree* because many of its parts have valuable applications. In particular, seed extracts from the plant have been used in traditional water treatment throughout Africa. The application of this extract to untreated water causes a 95% reduction of turbidity, and a decrease in particle and bacterial content. Numerous laboratories have demonstrated these effects but the precise nature and properties of the active components in the extract has been unclear.

A purified fraction of cationic coagulant proteins from seed extract has been characterised at a molecular level using biochemical and biophysical methods including chromatography, mass spectrometry and tandem mass spectrometry. These studies have allowed the identification of two main isoforms (*Mo*-CBP3-3 and *Mo*-CBP3-4) of the *Moringa oleifera* chitin binding protein *Mo*-CBP3, (a 2S albumin protein) present in this fraction of interest. The X-ray crystallographic structure of one isoform, *Mo*-CBP3-4, was determined to a resolution of 1.68 Å. This structure provides detailed information relating to this diverse family of albumins as well as insights for the observed flocculating properties of the protein. Neutron reflection studies of the structure and composition of interfacial layers at the solid/solution interface show that the crude seed extract behaves differently from the purified fraction. These results have been related to the anti-microbial, coagulation and flocculation properties of both the seed extract and the purified fraction and are placed in the context of water purification. Future directions could include the development of a recombinant expression system for large scale production of perdeuterated protein allowing contrast matching studies by neutron reflection and a neutron crystallographic structure to understand the nature of hydration interactions.

Acknowledgments

I would like to extend my acknowledgments to former and current members of the laboratory for their support and encouragement

Most of all, I am forever grateful to my family for their unlimited encouragement, especially my mother, but also Denis, Lysandre and Ambroise for their love and constant support over these 6 years !

Table of contents

List of Figures

5.13: Model representation of the crude extract (CE) versus *Mo*-CBP3 isoforms adsorbed layer on Silica (SiO_2) Different seed proteins are represented as symbol (● red spot could be *Mo*-CBP3 isoforms,◉ green spot,🌙 yellow moon, and ⬠ grey octagon as non-identified seed proteins).
a) For CE, an increase of adsorption (up to 0.5mg/ml) was observed as a diffuse layer having a thickness of 60 Å. The dense layer of the CE was about 5.5 mg m^{-2}
b) For *Mo*-CBP3 isoforms a clear adsorption was obtained as a well-defined layer having a thickness of 15 Å and a density of layer about 1.3 ± 0.2 mg m^{-2}. This adsorption was observed with a low concentration at 0.025 mg/ml.P.160.

5.14: Schematic showing the crystal structure of *Mo*-CPB3-4 (helix) **a)** and surface representation **b))** showing the polar versus non-polar areas of the protein. Red represents the most hydrophobic and white the most hydrophilic regions according to the Eisenberg hydrophobicity scale (Eisenberg *et al*., 1984).P.161.

5.15: Model representing flocculation and coagulation mechanisms of impurities that could occur in dirty water **a)** in presence of CE or **b)** with purified fraction (*Mo*-CBP3 isoforms). According to neutron reflectivity (NR) data obtained from collaborators, CE material was able to interact with both negative and neutral surfaces (*ie* particles) but the NR data obtained for the purified fraction demonstrated that 20 times more CE was needed to cover the surface at the silica surface (a). These *Mo*-CBP3 isoforms were highly positively charged and interacted more specifically with the silica surface with low concentration of materials.P.163.

List of Tables

xxi

List of Abbreviations

AD	*Anno Domini*
CE	Crude extract
CEA	Commisariat à l'Energie Atomique et aux Energies alternatives
CHU	Centre hospitalier universitaire
CV	Column volume
CM	Carboxymethyl
DTT	1,4-Dithiothreitol
ε	Molar absorption coefficient
EDTA	Ethylenediaminetetraacetic acid
ESI	Electrospray ionisation
ESRF	European synchrotron radiation facility
HPLC	High-performance liquid chromatography
HTX	High troughput crystallisation
IBS	Institut de biologie structurale
ILL	Institut Laue-Langevin
IEP	Isoelectric point
kDa	Kilodalton
mAU	Milli absorption unit
Mo	*Moringa oleifera*
***Mo*-CBP**	*Moringa oleifera* Chitin binding protein
MR	Molecular replacement

mS	milliSiemens
MS	Mass spectrometry
mU	Milliunit
MW	Molecular Weight
MWCO	Molecular weight cut-off
NTU	Nephelometric turbidity units
Pcps	Pyrrolidone carboxyl peptidases
Pfu	Polymerase
PSB	Partnership for structural biology
PVDF	polyvinylidene fluoride
RMSD	Root mean squre deviation
RPC	Reverse phase chromatography
SAD	Single-wavelength anomalous diffraction
SDS	Sodium dodecyl sulfate
SEC	Size exclusion chromatography
TCEP	Tris (2-carboxyethyl) phosphine

1-Introduction

1-1 The use of natural coagulants in water purification history

The first reference to the use of the *Moringa* plant for water purification is believed to have been made in the book of Exodus. This refers to the period after which Moses had parted the waters of the Red Sea and had led the Israelites to a place in the wilderness called Marah (in the Sinai Peninsula). The Israelites, being desperately short of water, were faced with the prospect of drinking water that was too "bitter". Moses is said to have asked God, who advised Moses to take the seeds from a particular tree and cast them into the water. The exact quotation from Exodus, chapter 25 verse 15, is given below:

Then Moses made Israel set out from the Red Sea, and they went into the wilderness of Shur. They went three days in the wilderness and found no water. When they came to Marah, they could not drink the water of Marah because it was bitter; therefore it was named Marah. And the people grumbled against Moses, saying, "What shall we drink?" And he cried to the Lord, and the Lord showed him a log, and he threw it into the water, and the water became sweet. (Bible Exodus 15:22 new internal version(NIV))

An internet search (http://sinai-moringa.com/moringa-in-ancient-egypt/) relating to the history of *Moringa* provides references to its use in a variety of ceremonial and medicinal contexts. The indigenous knowledge and use of *Moringa* is referenced in more than 80 countries and known in over 200 local languages. It has been used by various societies (Romans, Greek, Egyptian and Indian) for thousands of years dating back to 150 A.D. Historical texts reveal that ancient kings and queens used *Moringa* leaves and fruit in their diet to maintain mental alertness and healthy skin. Ancient Maurian warriors of India were fed with *Moringa* leaf extract in the warfront; the elixir drink was believed to add the extra energy and relieve them of the stress and pain incurred during war.

The historical record between these early references and later more recent research on *Moringa* and its uses is very poor. In 1850, Pereira published a general survey on the purification of drinking water, and mentioned the use of seeds of *Strycknos potatorum* in India and beans, haricots and castor seeds in Nubia. He described their use by rural communities along the Nile valley in removing turbidity from water by rubbing the inside of earthenware vessels with various kinds of seeds. This caused the sedimentation of a number of impurities, yielding clear water. A number of natural flocculating or coagulating agents of plant or soil origin exist. For example, in a text edited by Al Azharia Jahn (1960), detailed descriptions are given of the use of natural coagulants in the treatment of rural water supplies. In 1977, the same author described traditional methods of water purification in Sudan and related these to geographic and socioeconomic conditions (Jahn, 1977). Interestingly this author provides further context from the Muslim culture, mentioning that "a Sharif descended from the prophet prayed to Allah for aid with the water available from the Nile" (legend related from pious people of the Nile valley). Mention is made of the use of soil in water improvements. Jahn also extensively reviewed the distribution and the local use of various plants in water purification, including seeds of the Nirmali tree in India, mucilage from cactus leaves of *Opuntia tuna*, *Opuntia ficus indica,* and related species in Mexico and Peru. The natural polymers responsible for the water coagulating effect of plants differ in their chemical nature, but are mainly polysaccharides and proteins. One of the most promising traditional agents for turbidity removal seems to be the crushed seeds of *Moringa oleifera* tree.

1-2 The *Moringa* Tree

1-2-1 *Moringa* family

These are 13 different types of Moringa tree in the genus *Moringa* of family *Moringaceae* that have been identified throughout the world. Its name can be *Moringa*, drumstick tree (from the appearance of the long, slender, triangular seed-pods), horseradish tree (from the taste of the roots, which resembles horseradish), bean oil tree, or benzoil tree (from the oil which is derived from the seeds). Only a few of these species have been seen of interest in the context of water treatement: *Moringa oleifera, Moringa stenopetala, Moringa peregrine, Moringa longituba, Moringa drouhardii* and*Moringa ovalifolia* (Jahn, 1986).

For the purposes of this thesis work we have focused specifically on *Moringa oleifera* (*Mo*) largely because of its prevalence throughout sub-Saharan Africa, India, and Latin America, and because of the focus that has been placed on using it in the context of developing more modern purification approaches.

Figure 1.1: Distribution of Moringa Trees throughout the world (taken from https://www.treesforlife.org/our-work/our-initiatives/moringa

1-2-2 *Moringa oleifera* ("The miracle tree")

Moringa oleifera (*Mo*) is indigenous to south Asia, where it grows in the Himalayan foothills from north-eastern Pakistan to northern West Bengal, India. It has been introduced and become naturalised in other parts of India, Pakistan, Afghanistan, Bangladesh, Sri Lanka, Asia, the Arabian peninsula, Africa, southern Florida, throughout the west Indies and from Mexico to Peru, Paraguay and Brazil.

Mo is a small fast growing shrub or tree that is approximately 12m (36 ft) in height when mature, and can live for up to 20 years. This tree is also called, the miracle tree, because many of its parts have valuable applications. It has been used in villages in developing countries for centuries for traditional medicine, food and cooking oil, wood and water purification.

Figure 1.2: Photographs of the *Moringa oleifera* tree, its flowers (kindly provided by A.Rennie) and its leaves (taken from http://www.treesthatfeed.org/moringa-tree)

1-3 Moringa seeds

1-3-1 Morphology and appearance

Mo seeds are globular, about 1 cm in diameter (Figure1.3). They are three-angled, with an average weight of about 0.3 g, 3-winged with wings produced at the base of the seed to the apex, 2–2.5 cm long, 0.4–0.7 cm wide; the kernel is responsible for 70%–75% of the weight. Mo seeds have a brownish semi-permeable seed hull. Each tree can produce 15000 to 25000 seeds per year.

Figure 1.3: Seeds (A), Kernels (B), fruits (C) and oil extract (D) of *Moringa oleifera* (Leone *et al.,* 2016)

1-3-2 Seed development phase

Seed development is complex, with the first stage consisting of extensive cell division and differentiation/morphogenesis/patterning (Wobus *et al*, 1999). This is followed by the arrest of growth and subsequently a maturation stage in which there is accumulation of storage compounds, the development of tolerance to desiccation and the termination of growth. Having completed its development, the seed then enters a dormancy period that is ultimately ended by germination (Harada, 1997). During germination the seedling draws on these reserves until the plant establishes photosynthetic capability. (Baud *et al*, 2002).
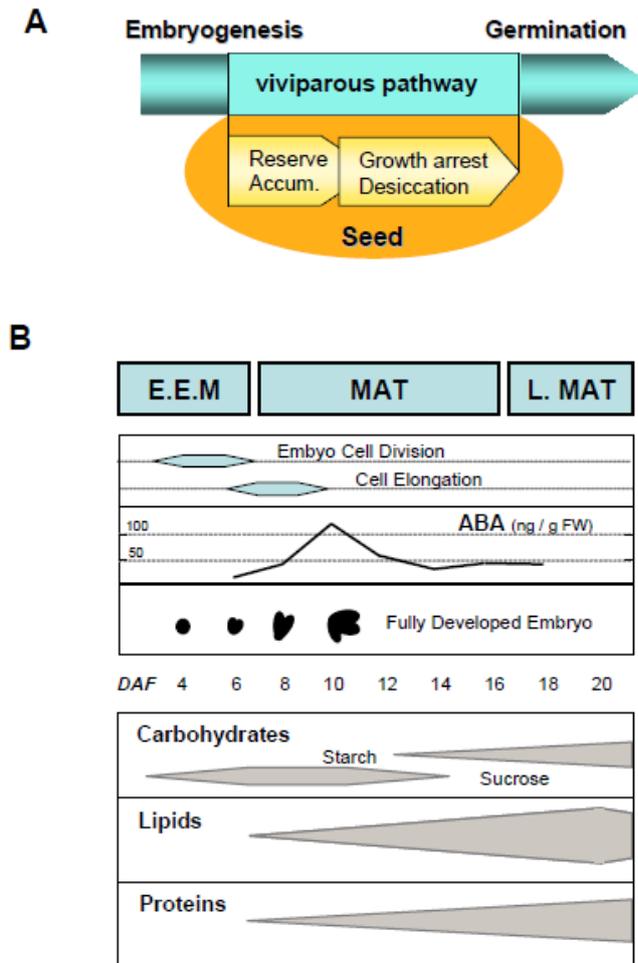
Figure 1.4: Development within the seeds. Phases and Features – taken from Vicente-Carbajosa and Carbonero (2005). (A) Representation of embryo development within the seed, followed by maturation (accumulation of storage compounds, development of resistance to desiccation) and dormancy. The end point of the process is germination. (B) Time course of Arabidopsis seed development (as an example), indicating events in embryo morphogenesis (E.E.M) and maturation (MAT) and late maturation (L.M.A.T).

The early and mid phases of maturation are dominated by the action of the plant hormone

abscisic acid (ABA). The transcription of major seed storage proteins occurs mainly during this

period. Subsequently, ABA levels decline and late maturation follows, characterised by the

synthesis of LEA (Late Embryogenesis Abundant) proteins that are associated with the

dehydration process and the acquisition of desiccation tolerance.
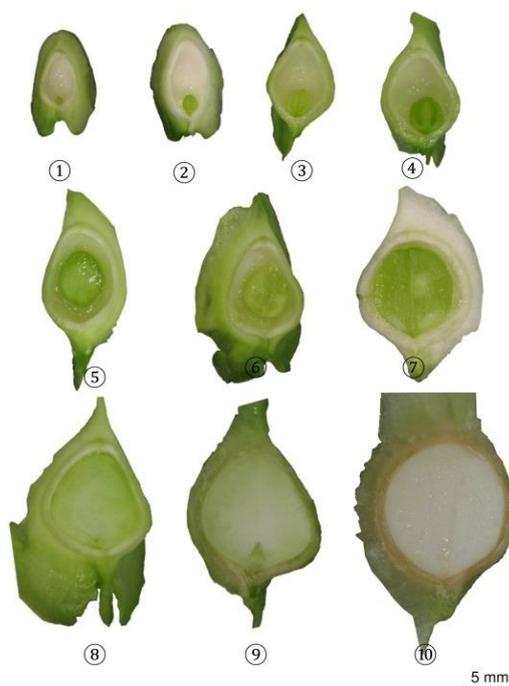
Figure 1.5: Cross-sections of a *Moringa oleifera* seed at different stages of development based on the fruit diameter (mm) (1) 8 mm;(2) 10 mm; (3) 12 mm;(4) 14mm;(5) 16 mm; (6) 18 mm; (7) 20 mm; (8) 22 mm (9) 24 mm; (10) 26 mm (taken from Muhl *et al.*, 2016).

During seed development, numerous physical and biochemical changes are observed. The seed volume changes through cell division, enlargement and differentiation (Figure 1.5), involving the deposition of storage compounds (oil, starch and protein). Starch is the first compound to be synthesised, and its initial accumulation acts as a temporary reserve enabling the regulation and facilitation of growth and biosynthesis of additional storage compounds such as oil and protein.

1-3-3 Nutritional composition of seeds

The main compounds found in mature *Mo* seed are oil, protein, and starch and have been analysed by Duke *et al.,* (1984), Makkar *et al.,* (1997), Oliveira *et al.,* (1999), Abdulkarim *et al.,* (2005), Anwar *et al.,* (2005) as summarised in Table 1.1. The percentage of proteins in *Mo* seeds is greater than found for many other leguminous crops, with an average content between 28 to 38 %.

Mo seed composition	Abdulkarim *et al* (2005)	Makkar *et al* (1997)	Duke *et al* (1984)	Oliveira *et al* (1999)	Anwar *et al* (2005)
Oil	30.8 %±2	41.7 %	34.7 %	41.2%±2.2	33.2 to 40.9 %
Protein	38.3 %±1.03	36.7 %	38.4 %	33.3 ±1.2	28.5 to 34 %
Starch	16.5 %	17.8 %	17.1 %	21.1 %	Not given

Table 1.1: Composition of oil, protein and starch in seeds of *Moringa oleifera* as given by Duke *et al* .,(1984), Makkar *et al.,* (1997), Oliveira *et al.,* (1999), Abdulkarim *et al.,* (2005), Anwar *et al.,* (2005).

This variation amongst the reported values (Table 1.1) may be attributed to varying agro-climatic conditions as well as the time of harvest (Signh *et al.*,1992). Muhl *et al.* (2016) have shown that protein accumulation remained fairly constant throughout seed development, decreasing slightly from 29.1% at a fruit diameter of 10 mm to 24.8 % at maturity (Figure 1.6). The protein content is less affected by the irrigation amount than starch and oil.
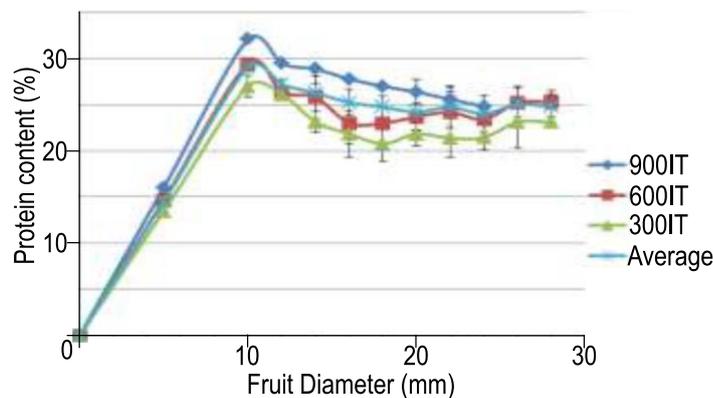
Figure 1.6: Changes in the protein content (%) of developing *Moringa oleifera* seeds at various developmental stages, and the effect of three irrigation treatments (IT). 900IT – 900 mm of water/annum, 600IT – 600 mm/annum, 300IT – 300 mm/annum. Vertical bars (±) indicate standard errors in measurement. (Adapted from Mulh *et al.*,2016)

1-3-4 Exploitation in traditional water purification

The exploitation of Moringa seeds in traditional water purification is still used in certain rural areas (*eg* in Sudan). Moreover some charity associations teach families this simple process of combining crushed Moringa seeds with water, helping with the provision of a sustainable supply of clean water

Traditional water purification procedure

The procedure used in traditional water purification consists of crushing the white seed kernels to a powder and mixing it with clean water (Figure 1.7). The mixture obtained is then transferred to the container of the turbid water. The water is stirred rapidly for two minutes and then slowly thereafter for 10-15 minutes. After one hour, impurities sediment. Only one seed is necessary to purify one litre of slightly contaminated water and two seeds for very dirty water.

Figure 1.7: Traditional water purification procedure using *Moringa* seeds – as taught to families in developing countries by the charity association "strong harvest international" (taken from https://www.strongharvest.org/moringa-basics/).

Lea (2010) published an application of this low-cost *Mo* protocol for water treatment and

provided an overall general guideline for dosage rates (Table 1.2). It can be noted that *Mo*

seeds are not an effective coagulant at low turbidity water (<50 NTU) water.

Raw water turbidity (NTU)	Dose range (seeds/liter)	Dose range (mg/ml)
< 50 NTU (low)	1 seed/4 liters	50mg/liter
50-150 NTU (medium)	1 seed/2 liters	100mg/liter
150-250 NTU (high)	1 seed/ liter	200mg/liter
>250 NTU (extreme)	2 seeds/ liter	400mg/liter

Table 1.2: *Mo* dosage rates. One shelled seed (~200 mg) is used to treat 1 litre of very turbid surface water. (Doerr, 2005). NTU (nephelometric turbidity units)

Advantages

Dirty water contains chemicals and biological impurities *i.e.* suspended and dissolved

inorganic and organic compounds and micro organisms. These compounds may come from

natural sources and leaching of waste deposits. Inorganic compounds, in general, originate

from weathering and leaching of rocks, soils and sediments (calcium, magnesium, sodium

sulphate , nitrate..) whereas organic compounds originate from decaying plants and animal

matters and from agricultural runoffs (detergents, pesticides, herbicides and solvent). It has

been shown by Madsen *et al.*,(1987) that the use of *Mo* seeds can reduce 90-99 % of bacterial

content and decrease the turbidity of 80 to 99.5%. This study was performed using fresh

samples from the Blue and White Nile.

It is very effective for high turbidity water and shows similar effects to that of alum, which is

commonly used in water treatment. However, there are concerns over the use of alum, which

is known to cause gastrointestinal problems (Drinking Water and Health, 1982) and may be

implicated in Alzheimer's disease (Vijayaraghavan *et al.*, 2011; Matilainen *et al.*, 2011). It also

produces greater sludge volumes. It has been demonstrated that *Mo* seed extract possesses

anti-microbial properties for both Gram-negative and Gram-positive bacterial cells (Broin *et al.*, 2002). Microorganisms can be removed by settling in the same manner as the removal of colloids in coagulated and flocculated water. The improvement and optimisation of the flocculation and studies as to how the seed protein could be used in practical conditions has been reported in a number of studies (see *e.g.* McConnachie *et al.*, 1999; Beltrán-Heredia &. Sánchez-Martín, 2009; Pritchard *et al.*, 2010).

1-4 Seed storage proteins

The seed storage proteins play an important natural role because they provide a source of nutrients (carbon nitrogen and sulphur resources) for subsequent growth and development of the plant. Seed storage proteins were among the earliest of all proteins to be characterized. Wheat gluten proteins were first isolated in 1745 (Beccari, 1745) and Brazil nut globulin was crystallised in 1859 by Maschke. Despite wide variation in their detailed structures, all seed storage proteins have a number of common properties. Firstly, they are synthesized at high levels in specific tissues and at certain stages of development. Their synthesis is regulated by nutrition, and they act as a sink for surplus nitrogen. However, most also contain cysteine and methionine, and sufficient supplies of sulphur are therefore required for their biosynthesis. Many seeds contain separate groups of storage proteins, some rich in sulphurous amino acids and others less so. The presence of these groups may allow the plant to maintain high levels of storage protein synthesis despite variations in sulphur availability. A second common property is their presence in the mature seed in discrete deposits called protein bodies. Finally, all storage protein fractions are mixtures of components that exhibit polymorphism both within single genotypes and among genotypes of the same species. This polymorphism

arises from the presence of multigene families and in some cases, proteolytic processing and glycosylation.
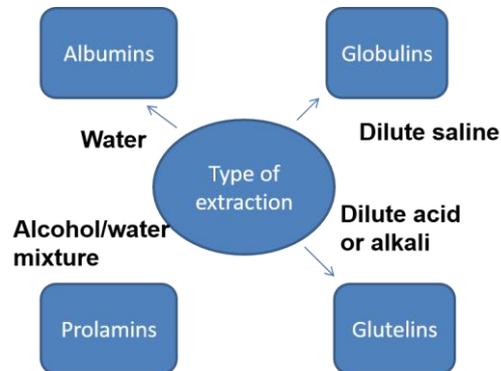
Figure 1.8: Seed proteins are classified into four groups according to their solubility in specific solvents used successively to extract the proteins according to Osborne (1924).

The major seed proteins include albumins, globulins, glutelins and prolamins. Osborne (1924) classified the seed storage proteins on the basis of their extraction and solubility in water (albumins), dilute saline (globulins), alcohol/water mixtures (prolamins), and dilute acid or alkali (glutelins) (Figure 1.8). A recent study by Baptista and co-workers in 2017, described the fractionation of seed proteins according to the Osborne classification and showed that *Mo* seeds are mainly composed of globulins and albumins which represent 53% and 44% of total proteins of the seed, respectively. From turbidity measurements, they assessed both protein families as having high coagulant potential in water treatment.

1-4-1 The 2S albumin storage protein

The 2S albumin proteins were defined as a group based on their sedimentation coefficients (S_{20w}) of ~ 2 (Youle and Huang, 1981). They are widely distributed in both mono- and di-cotyledonous plants and are a major group of seed storage proteins. They represent 70 % of seed storage protein present in the seed. They are widely studied in the Cruciferae - notably

oil seed rape (napins) and Arabidopsis. Storage proteins accumulate primarily in the protein storage vacuoles (PSVs) and are assembled and folded in the endoplasmic reticulum (ER) which is also the site of disulphide bond formation. The entry of 2S albumin into the ER occurs as a result of a signal peptide that is cleaved during translocation into the lumen of the ER. They are synthetized as single precursor protein (preproalbumin) that is proteolytically cleaved with the loss of linker peptide and short peptides from both the N and C termini to obtain the mature form (Figure 1.9).
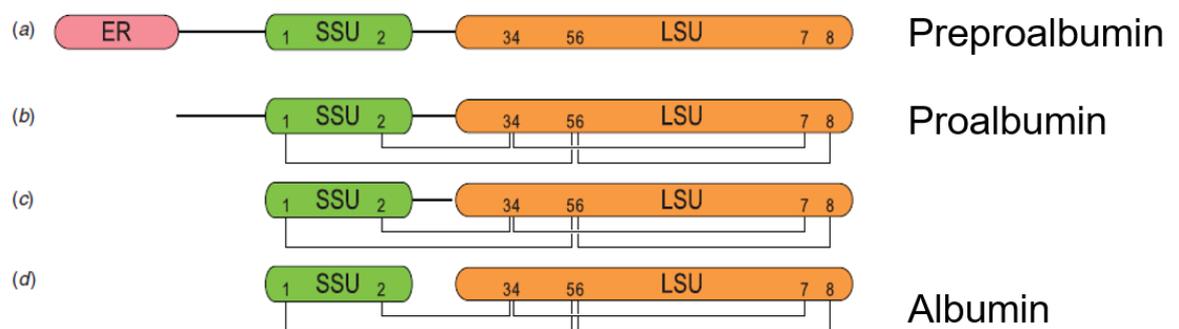
Figure 1.9: Schematic representation of the proteolytic cleavage steps leading to the mature albumin form (taken from Mylne *et al.*, 2014). (a) Preproalbumin is composed of an endoplasmic reticulum signal sequence (ER, pink), a small (SSU, green) and large (LSU, orange) albumin subunits. (b) The ER signal is removed upon entry into the ER, resulting in disulphide bonds formation. (c) The endo-protease AEP removes the N-terminal pro region (solid black line) and cleaves after the residue preceding the large albumin subunit. (d) An ASP exo-protease matures the small albumin subunit by trimming back the exposed tail (solid black line) resulting in mature hetero-dimeric albumin.

It has been shown that some ER luminal chaperones and enzymes may assist in these processes. Despite differences in their subunit structure and synthesis, all the 2S albumins are compact globular proteins with conserved cysteine residues.
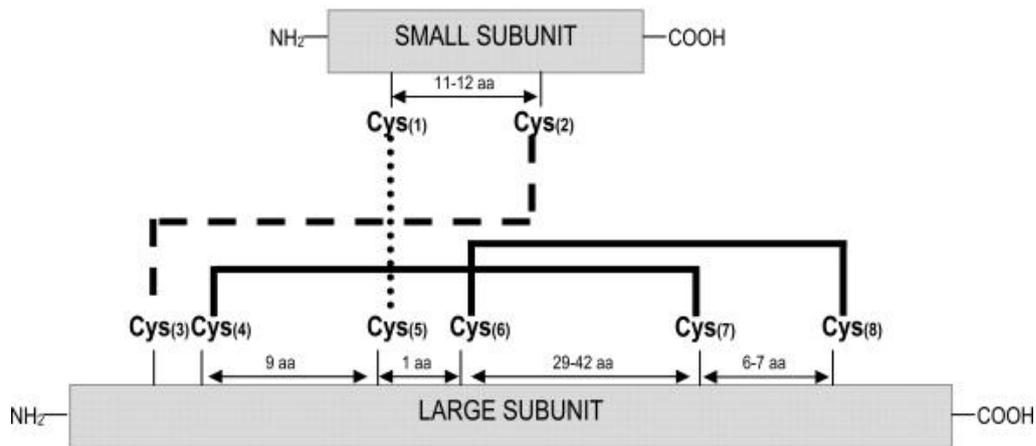
Figure 1.10: Schematic representation of the disulphide bond patterns formed between the eight conserved cysteine residues in the 2S albumin family (from Moreno *et al.*, 2008).

These eight cysteins follow a conserved pattern (...C...C.../...C...CXC...C...C...), which is called the eight-cysteine motif (8CM) (Figure 1.10). This sequence motif is a characteristic structural feature of all 2S albumins and is also found in other members of the prolamin superfamily (Shewry *et al.*, 1995; José-Estanyol *et al.*, 2004). As storage proteins, they are utilized by the plant as a source of nutrients during subsequent germination and seedling growth, and studies have demonstrated that 2S albumins can play a protective role in plants against fungal attack (Agizzio *et al.*, 2006).

1-5 Scientific studies of Moringa seed proteins and their properties

In a contemporary scientific context, the literature relevant to this thesis project seems to commence in the 1970s. Specific emphasis is placed on *Moringa* varieties that are referred to in Arabic as *"shagarra al rauwaq"*. Jahn, in 1979, published a further paper on *Mo* in which the coagulation principles are described, noting that these properties were comparable to those of alum. Jahn also notes a bacteriostatic effect of *Moringa* but only in a transient mode. In 1982, Barth et *al.* described a study of material, assumed to be *Mo* seed proteins, that were

fractionable by chromatography techniques. Their experiment on flocculation with a colloidal silica suspension showed that the active substances are cationic polymers. They demonstrated that the coagulation efficiency significantly depends on an optimal dosage, suitable conditions of stirring, and a pH between 5 and 9. The acidic hydrolysis of the fractionable material has revealed the presence of amino acids - mainly glutamine and argnine - with a calculated isoelectric point (IEP) higher than 8.6.

Many studies have reported that for most of the exploitable processes suggested so far, the active material from the crude extract (CE) from the seeds is a mixture of various components. The details of the extraction of protein from the seeds have been identified as important: the most effective flocculating agent was obtained using a 1 M saline solution although the specific type of cation appears to be less significant than the concentration (Madrona *et al.*, 2010 and 2011). The presence of a mixture of proteins with a composition that might depend in part on growth conditions and the possibility to alter the composition of the extracted material may partially explain the varying molecular masses of *Moringa* seed proteins.

1-5-1 Different Moringa proteins identified so far

Some laboratories have demonstrated a principal component of this seed extract to consist of dimeric proteins with molecular weight (MW) in the region of 12-14 KDa and having a isoelectric point (IEP) between 10 and 11 (Ndabigengesere *et al.,* 1995). Other reports describe a single polypeptide chain have 60 amino acids, an apparent molecular mass of 6.5 Kda and a pI of greater than 10, called MO2.1 and MO2.2; these were sequenced by Gassenschmidt *et al*. in 1995. MO2.1 and MO2.2 are natural variants that differ by a single residue of this polypeptide. The cDNA was cloned by Brouin *et al.* (2002) who demonstrated the flocculent activities and the antimicrobial properties. In 2008, MoL was identified as a *Mo*

seed lectin that agglutinates human as well as rabbit erythrocytes and has a binding specificity for glycoproteins. MoL is a homodimer (14KDa) in which the monomers (~7KDa) are linked by disulfide bonds (Katre *et al*, 2008). More recently, a new coagulant lectin named cMoL that agglutinates human and rabbit red blood cells was purified from Mo seeds. cMoL is composed of a polypeptide chain of about 11.9 kDa that forms homotrimers of approximately 30 kDa (Luz *et al.*, 2013).

In 2012, a novel chitin-binding protein (CBP) was purified from the seeds of *Mo* and named *Mo*-CBP3 by Gifoni *et al*. *Mo*-CBP3 is a 14-kDa (estimated by gel filtration) thermostable antifungal protein that inhibits the spore germination and mycelial growth of the ascomycete *Fusarium solani* and other fungi (Batista *et al.*, 2014). Freire and co-workers demonstrated that *Mo*-CBP3 belongs to the 2 S albumin family due the presence of an eight-cysteine motif; these are synthetized as precursors, which are then proteolytically cleaved to form the mature form. Four isoforms of this precursor have been isolated namely *Mo*-CBP3-1, *Mo*-CBP3-2, *Mo*-CBP3-3 and *Mo*-CBP3-4, which differ only from each other by a few amino acid residues. The *Mo*-CBP3 precursors consist of 160 to 163 (Figure 1.11) amino acids including the N-terminal signal peptide and the linker peptides. Based on comparative sequence analysis, it was speculated that *Mo*-CBP3 precursors are likely to be processed similarly to seed storage albumins like pro-Mabinlin-II (a seed storage protein from a tree *Capparis masaikai*).

Figure 1.11: Multiple alignment of the amino acid sequences of the *Mo*-CBP3 precursors with proMabinlin-II (Freire *et al*., 2015).

In 2015, Ullah *et al.* published the model structure of the *Mo*-CBP3-4 isoform, confirming the compact fold due to the presence of a small and large chain linked together by four disulfide bonds. This arrangement is key to the extreme stability of the protein as well as it heat resistance and proteolysis.

1-5-2 Flocculation activities of *Moringa* seed proteins

Mo protein from the seeds is the active flocculating agent and some studies have focussed on establishing an understanding at a microscopic level as to how it works. The usual means to achieve flocculation of colloids in water purification is either by screening and neutralisation of charge using polyvalent salts, or by the addition of polymers that act as 'bridging' flocculants that bind to multiple particles. However *Moringa* proteins cannot act in either of these ways. The seed proteins are small, cationic and are difficult to denature in solution. They show a marked tendency to self-associate in aqueous solution (Kwaambwa & Rennie, 2012) unless the concentration is very low. The model that emerges for flocculation is one that involves adsorption of protein to a wide range of different particles and this will then favour their association (Hellsing *et al*., 2014). Essentially, the strong attractive interactions between the molecules are transferred to the particles that are covered with protein. This allows effective

coagulation and hetero-coagulation of a broad range of particulate impurities. The binding of seed protein to interfaces of silica (Kwaambwa *et al.*, 2010) and alumina (Kwaambwa *et al.*, 2015) to aqueous solutions has been investigated by neutron reflection and the results give a picture of a capability to adsorb a broad range of different materials as multi-layers of molecules.

1-5-3 Antibacterial activities of Moringa seed proteins

Beside their flocculation activity, additional experiments on toxicology and microbiology have also shown a transitory inhibition of different bacterial strains like *Staphylococcus aureus* and *Pseudomonas*. Berger *et al.*,(1984) have published a toxicological assessment of seeds from *Mo* and *Moringa stenopetala*, and emphasise these seeds as a source of coagulants for the treatment of domestic water. Madsen *et al.,* (1987) have published an estimation of the turbidity reduction to 80-99.5%, paralleled by a primary bacterial reduction of 90-99% in the first 1 to 2 hours of treatment. These authors also noted a secondary bacterial increase due to regrowth in the supernatant for several strains like *Salmonella* and *Shigella*, in somes cases for *Esherichia coli* but not for *Vibrio, Streptococus faecilis* and *Clostridium*. This activity has been ascribed to *Moringa* synthesized derivatives of benzyl isothiocyanates (a known aromatic antibacterial compound obtained by methylene dichloride extraction) as well as to *Moringa* seed derived peptides (Suarez *et al.,* 2005). Water-soluble peptides released from the crushed seed kernels belong to a group of cationic peptides often displaying antimicrobial activity by its interaction with negatively charged microbial surfaces, with its amphiphilic structure allowing their incorporation into cellular membranes. The antibacterial activity of such a *Moringa* seed peptide (MO2.1 also called Flo) against human pathogens including *Pseudomonas aeruginosa* and *Streptococcus pyogenes* has been studied in detail (Suarez et

al., 2005). The authors suggest partly distinct molecular mechanisms for sedimentation and bactericidal activity involving membrane destabilization by a hydrophobic loop. The behaviour of the proteins as an antibacterial agent has also been investigated by cryo-EM on *E. coli* cells showing a fusion of inner and outer membranes (Shebek *et al.*, 2015). The strategy used to assess antibacterial activity of Mo cationic peptide (MOCP) involves their immobilization on sand (Jerri *et al.*, 2012).

1-6 Scientific rationale for the PhD work

At the inception of this project, following discussions with Prof. A. Rennie (Uppsala University), seeds from *Moringa oleifera* were made available courtesy of Dr. Kwaambwa of Namibia (University of Science and Technology, Windhoek). Early compositional analysis demonstrated the presence of small protein components within the water soluble CE. One particular fraction of purified protein within this extract was found to be the most abundant species. This particular fraction was of interest because of its properties in relation to those of the bulk extract. Of special focus was the aggregation-inducing properties and how these could be understood using state-of-the-art biochemical and biophysical techniques.

The ultimate goal of the thesis work was therefore to establish a rational understanding of these properties in terms of the protein structure and its surface interaction behaviour.

Quite apart from the fundamental scientific interest in this rather remarkable system, there are several reasons why this sort of information may be of central importance for a fuller exploitation of *Moringa* both (i) at a local level where refinements of procedure could further optimise traditional water purification methods in rural areas (ii) at a national/regional level where there is governmental interest (eg in southern African countries such as Botswana and

Namibia) in how this type of system could be developed on a large scale at water purification plants. Key examples of this are summarised below.

- In Botswana and Namibia, the exploitation of traditional water purification approaches is seen as a priority to develop effective water purification for remote areas that may lack resources, power, and trained personnel. The development of exploitation of *Moringa* products for this use has been seen as of higher priority than its use in larger water treatment plants for towns and cities.

- The Ministry of Agriculture, Water and Forestry in Namibia has established a project for a plantation to grow various *Moringa* trees in the Rundu district. The aim is to investigate both optimal growing conditions and to provide seeds for water purification trials.

- There is interest in developing the growth of *Moringa* as a 'cash crop'. Apart from the use of seeds in water purification, seed oil can be sold or used as a fuel. Leaves are exploited as a salad crop and dried to make teas. Encouraging wider growth could therefore bring economic development to rural areas.

- In Botswana, the Botswana Water Utilities Corporation has undertaken a number of tests such as those reported at a Workshop in 2015 on water and sustainable development in Perugia where *Moringa* protein was seen to be as effective as other flocculating agents but where work remained to be performed to verify the residual protein content and the consequent oxygen demand that should be maintained at a low level. The water situation (droughts) and economic resources available have slowed this work, although the corporation has identified that the rising cost of chemicals purchased abroad pose major challenges.

1-7 Aims and objectives

Given the interest in these seeds in terms of their potential for modern applications in water purification, this thesis project set out to try to isolate the various components of the crude see extract and to purify and characterise them using advanced biochemical and biophysical methods. The main emphasis thereafter was (i) to establish information in terms of their bacteriocidal/bacteriostatic and flocculation properties (ii) to determine structural information at a molecular level that could be related to the properties of interest. The specific objectives are set out below :

(a) To perform a detailed biochemical and biophysical characterisation and activity assays (antibacterial and flocculation activities) of the fractionated CE from seeds of *Moringa oleifera*, and to identify and study the main component protein. This is described in Chapter 3.

(b) To carry out a detailed crystallographic structure analysis of the main protein component (*Mo*-CBP3-4). This structure determination and its function related to the water purification properties and is described in Chapter 4.

(c) To undertake an extensive analysis of the way in which *Mo*-CBP3-4 interacts with different types of surface using neutron reflectometry studies and compare these data to those obtained for the crude extract by our collaborators This is of interest for an understanding of the mechanism of its coagulation-flocculation properties. The findings are shown and discussed in Chapter 5.

2-Material and methods

This thesis work required the extraction, purification and characterisation of various proteins from Moringa *oleifera* (*Mo*), including the *Mo*-CBP3 protein and its different isoforms *Mo*-CBP3-1, *Mo*-CBP3-2, *Mo*-CBP3-3 and *Mo*-CBP3-4. The purification approaches required detailed chromatographic techniques based on charge separation and on size exclusion methods, making extensive use of polyacrylamide gel electrophoresis, as well as exploiting mass spectrometry for quality control and fragment analysis. Pure protein was crystallised using advanced robotic techniques and a detailed crystallographic study carried out using facilities at the European Synchrotron Radiation Facility (ESRF). The structural analysis did not allow the structure to be sequenced completely, and further proteolytic and N-terminal sequencing methods were used. A complementary study using neutron reflectivity was performed to further understand the mechanism of coagulation. This chapter summarises the various techniques used throughout the project.

2-1 Protein purification

2-1-1 Origin of materials: extraction from seeds

The proteins present in the seed are classified in four groups according to their solubility in specific solvents used successively to extract the proteins. Osborne in 1924 classified the seed storage proteins on the basis of their extraction and solubility - in water (albumins) dilute saline (globulins), alcohol/water mixtures (prolamins) and dilute acid or alkali (glutelins). The protein seed extract also called crude extract (CE) provided by Prof A. Rennie (Uppsala University, Sweden) and Dr. M. Kwaambwa (Polytechnic of Namibia, Windhoek) was prepared from seeds of Moringa *oleifera* collected in Namibia.

Figure 2.1: On the left,the *Moringa oleifera* (Mo) seeds are circular with a brownish semi-permeable seed hull (from Olagbemide *et al.,* 2014). On the right, Mo kernels are white after the husk removal. The kernel is responsible of 70-75% of the weight.

The procedure used for the extraction of *Moringa* protein involves treatment with petroleum ether to remove oil, followed by the extraction of water soluble proteins. This means that the major proteins present in the seed are albumin proteins. This procedure has been described by Kwaambwa *et al.,* (2012).

2-1-2 Protein Purification

Ion exchange chromatography: CM sepharose

The CE is not pure and *Mo*-CBP3 isoforms need to be separated from contaminants. For the purification of proteins, polypeptides, nucleic acids polynucleotides and other charged biomolecules the most common technique employed is that of ion exchange chromatography (Bonnerjera *et al.,* 1986). The separation using this method is based upon the reversible adsorption of the charged molecules to immobilised ion exchange groups of opposite charge. The first stage in the purification is the equilibration of the ion exchanger where the pH and ionic strength is adjusted using an appropriate buffer. At this stage the exchange groups are associated with exchangeable counter-ions. The next step is adsorption of the sample where molecules of appropriate charge displace the counter-ions, binding reversibly to the gel.

Unbound contaminants are washed using starting buffer. The following stage involves elution using an elution gradient of buffer by either increasing the ionic strength or adjusting the pH. This elution will cause the molecules to elute according to their strength of binding so that the more weakly binding ones will elute first (Figure 2.2). The final step is removal of the remaining molecules - and subsequently the regeneration of the column prior to the next purification.

The type of ion exchanger depends on the insoluble matrix to which charged groups are covalently bound. The ion exchanger used to purify *Mo*-CBP3 isoforms was a carboxymethyl (CM) anion exchanger column where its functional group is $-O-CH_2-COO^-$ and charged at pH 7 (Ghebremichael *et al.*,2005). The protein elution process involved increasing ionic strength (sodium chloride).
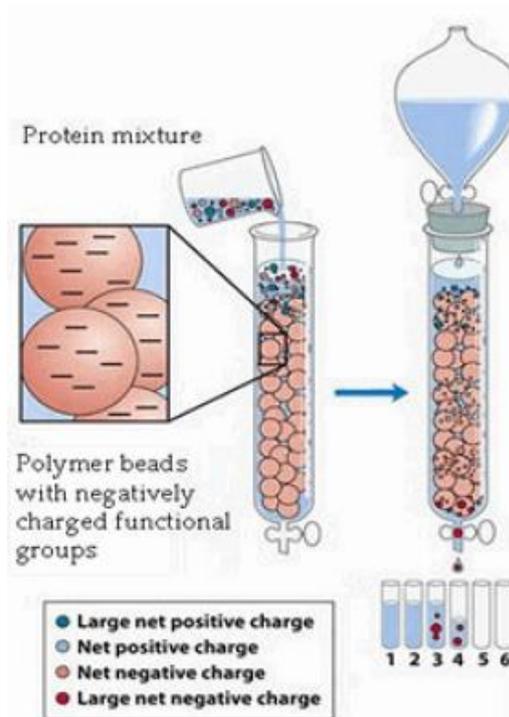
Figure 2.2: Cation exchange chromatography illustrated. Positively charged groups on the protein bind to negatively charged groups on the cation exchange resin. Increasing salt concentration produce cations that displace the proteins (image taken http://www.biochemden.com/ion-exchange-chromatography/).

Size exclusion chromatography (SEC)

Other types of chromatography include gel filtration or gel permeation chromatography. This technique separates proteins on the basis of molecular size and is used as the last step of purification. The stationary phase in size exclusion chromatography (SEC) contains porous hydrophilic gel beads made of cross-linked dextran, polyacrylamide and polystyrene. The principle of the technique is the diffusion of molecules into the porous cavities of the beads. The molecules larger than the pores cannot enter inside the beads whereas the smaller ones can.

Large molecule that cannot enter the pores of chromatography beads

Target protein that can use a fraction of the pore volume of the beads

Salt or other low-molecular weight substances that can use the entire pore volume of the beads
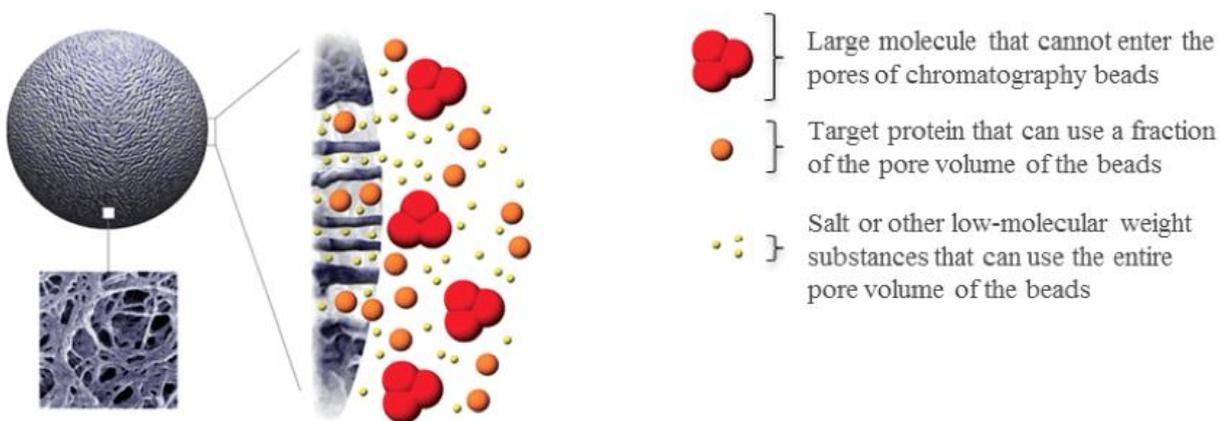
Figure 2.3: Separation of two proteins by using size-exclusion chromatography (taken from GE healthcare). The protein mixture is loaded on the top of the gel. Then the large molecules pass through the column faster than the small molecules

Reduction and alkylation methods and reverse phase chromatography (RPC)

The technique of reversed-phase chromatography (RPC) was used to separate chains of the isoforms after a reduction and alkylation treatment for further characterisation.

Reduction and alkylation of cysteine residues was performed to separate both chains of the protein linked by disulphide bridges. The reduction to the thiol is achieved by the reaction with sulfhydryl or phosphine groups such as dithiotreitol (DTT) or tris-2-

carboxyethylphosphine hydrochloride (TCEP). The irreversible alkylation of the SH groups was carried out with iodoacetamide which transformed the cysteine residues to the stable S-carboxyamidomethylcysteine (Figure 2.4). These bonds are irreversibly broken up for *Mo-CBP3* protein characterisation and peptide mapping. This covalent addition of a carbamidomethyl group (57.07 Da) prevents the formation of disulfide bonds. The iodoacetamide is unstable and light-sensitive and the alkylation was performed in the dark.
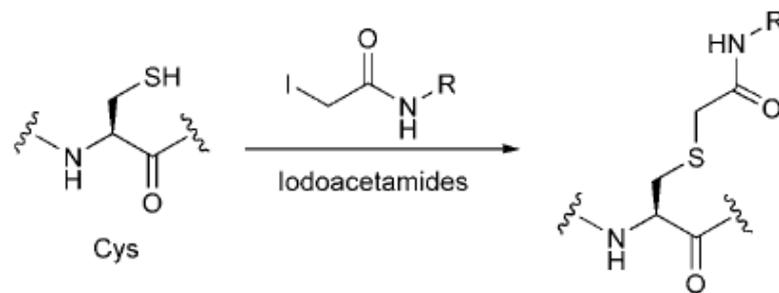
Figure 2.4: Modification of cysteine residue using iodoacetamides (image taken from Chalker *et al.,* 2009).

Reverse phase chromatography

Reverse phase chromatography (RPC) consists of the adsorption of hydrophobic molecules onto a hydrophobic solid support in a polar mobile phase. Decreasing the mobile phase polarity by adding more organic solvent reduces the hydrophobic interaction between the protein and the solid support, resulting in de-sorption. The more hydrophobic the protein the more time it will spend on the solid support and the higher the concentration of organic solvent that is required to produce de-sorption.
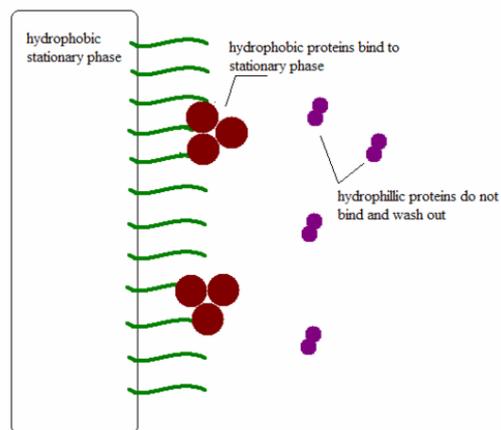
Figure 2.5: Principle of reverse phase chromatography (RPC) (scheme taken from hhttps://en.wikibooks.org).

The stationary phase is packed with silica modified with silyl ethers containing non-polar alkyl groups C8 (octylsilane) or C18 (octadecylsilane) (Figure 2.6). Both refer to the alkyl chain length of the bonded phase of the column. This creates a hydrophobic stationary phase. As the C18 column has the highest degree of hydrophobicity due to the length of the carbon chain, it was chosen to separate both chains of *Mo*-CBP3 protein. The polar phase contains polar organic solvents such as methanol, butanol or acetonitrile.
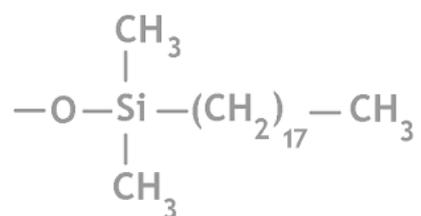
Figure 2.6: The C18 column is an octadecyl carbon chain (C18)-bonded silica.The C8 column exhibts the same formula but contains 7 CH_2 groups instead of 17 for the C18

2-1-3 Gel and western blot analysis

Gel electrophoresis

To assess if the purification of *Mo*-CBP3 proteins was successful and the level of purity is sufficient, sodium dodecyl sulphate (SDS) polyacrylamide gel electrophoresis (PAGE) analysis was used. This method is based on the use of electric field to separate different polypeptides according to their molecular weight. To separate proteins in an electrical field in a manner that is based on their molecular weight only, the tertiary structure has to be destroyed by reducing the protein to a linear molecule and masking the intrinsic net charge of the protein. SDS (a detergent) is used to do this and is normally present in the SDS-PAGE sample buffer. The use of SDS, along with boiling, and the presence of a reducing agent (normally DTT or β-mercaptoethanol) breaks down protein-protein disulphide bonds and disrupts the tertiary structure of proteins. This denatures the folded proteins to linear molecules. SDS also coats the protein with a uniform negative charge, which masks the intrinsic charges on the R-groups. SDS binds fairly uniformly to the linear proteins (around 1.4g SDS/ 1g protein), meaning that the charge of the protein is becomes approximately proportional to its molecular weight. SDS is also present in the gel to make sure that once the proteins are linearised and their charges masked, they stay that way throughout an electrophoresis run.
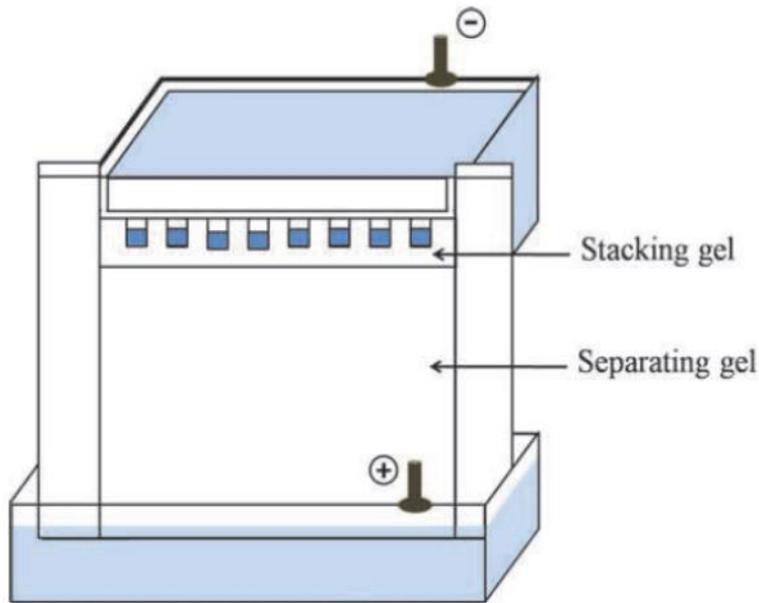
Figure 2.7: Polyacrylamide gel electrophoresis (PAGE).The equipment composes of the upper and lower chambers which are filled with an electrode buffer. The gel is polymerized in the space between two glass plates and then connected between the two chambers. Protein samples, suspended in a bromophenol blue/SDS loading buffer are loaded into wells at the top of the system. Molecules migrate into the gel in response to the applied electric field. For SDS-PAGE, the protein migrate from cathode to anode.

Following migration, PAGE gels are typically stained using a mixture of Coomassie Brilliant Blue

R_250 mixed with methanol and acetic acid (the latter helps fix the protein while staining).

The final *destai*n step removes excess dye using a solution of methanol and acetic acid was.

Tris-Tricine gel analysis

The resolution of smaller proteins (<10 kDa) using PAGE is hindered by the continuous

accumulation of free dodecyl sulfate (DS) ions (from the SDS sample and running buffers) in

the stacking gel. This zone of stacked DS micelles causes mixing of the DS ions with the smaller

proteins, resulting in fuzzy bands and decreased resolution. The mixing also interferes with

the fixing and staining of smaller proteins. To help solve this problem, Schagger and von Jagow

(1987) developd a Tris-Tricine protocol based on the Tris-Glycine system This modified system

uses a low pH in the gel buffer and substitutes tricine for glycine in the running buffer. The

smaller proteins and peptides that migrate with the stacked DS micelles in the Tris-Glycine protein gel system are well-separated from DS ions in Tris-Tricine gel system, resulting in sharper bands and higher resolution. Tricine–SDS-PAGE is the preferred electrophoretic system for the resolution of proteins smaller than 30 kDa.

Western blot

Western blotting is an important technique used in molecular biology and useful for protein detection. The technique consists of three steps: the separation of proteins based on size, the transfer to a solid support and the detection/visualisation of the target protein with the help of an appropriate primary and secondary antibody.

Production of a Moringa polyclonal antibody

Antibody production involves the preparation of antigen samples *ie Mo*-CBP3 purified protein and their safe injection into an animal in a manner that evokes high expression levels of antigen-specific antibodies, which can be recovered from the serum of the animal (Figure 2.8). The production of a *Moringa* polyclonal antibody was carried out by Eurogentec and was used as an antigen-specific probe to detect specifically the *Mo*-CBP3 protein.
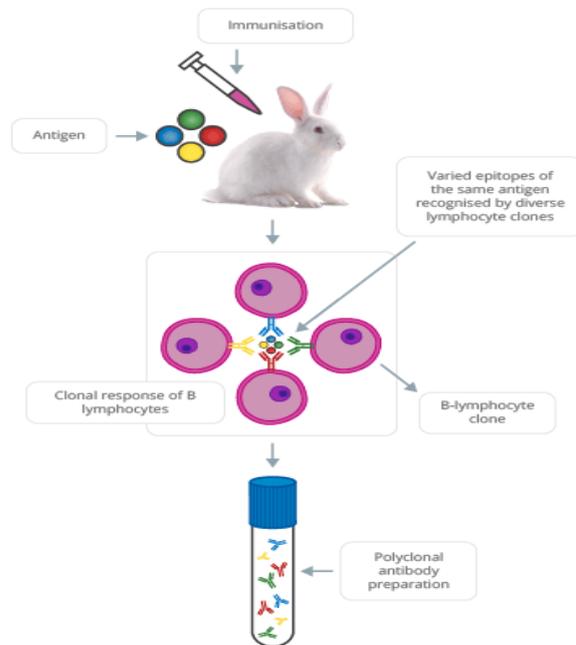
Figure 2.8: Schematic diagram summarising the production of a polyclonal antibody (image taken from immunostep.com website).

Transfer and colorimetric detection on a membrane

The protein mixture is separated based on molecular weight by gel electrophoresis. The separated bands are then transferred to a membrane – resulting in a band for each protein. The membrane is then blocked to prevent for non-specific binding by placing it in a dilute solution of protein (typically Bovine serum Albumin (BSA) or non-fat dry milk). After blocking, the membrane is incubated with an unlabelled primary antibody specific to the protein of interest. The unbound antibody is washed off leaving only the antibody that is bound to the protein of interest. After washing, a labelled secondary antibody is used to detect the presence of the first antibody, and thus the target protein. The most common detection methods use secondary antibodies conjugated to alkaline phosphatase (AP) or horseradish peroxidase (HRP). When the enzyme is exposed to a substrate solution, a coloured precipitate is deposited on the blot and the colorimetric detection reaction proceeds until stopped. The

thickness of the band corresponds to the amount of protein. Detection limits for colorimetric substrates are in the low nanogram range.
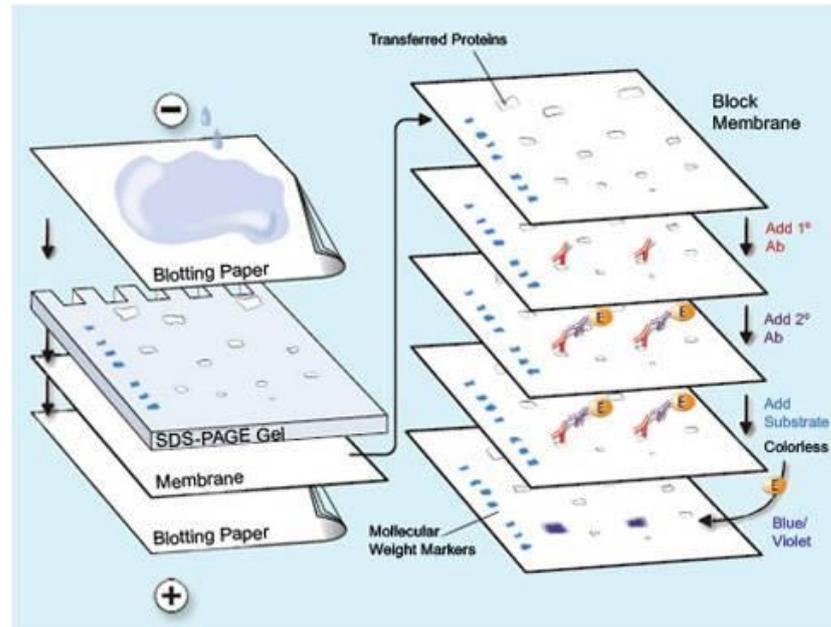
Figure 2.9: Principle of the western blot method (taken from www.Komabiotech.co.kr).

Isoelectric focusing gel (IEF)

Isoelectric focusing (IEF) is an electrophoretic technique for the separation of proteins based on their isolelectric point (IEP). The IEP is the pH at which a protein has no net charge and thus, does not migrate further in an electric field. IEF gels are used to determine the IEP of a protein and to detect minor changes in the protein due to post-translational modifications such as phosphorylation and glycosylation. Gels are cast with amphoteric molecules to set up pH gradients. Proteins migrate in an electric field until a stable pH gradient is formed and they settle into their IEP. A high finishing voltage is applied to focus the proteins into narrow zones. High voltage cannot be used during the initial stages of IEF as movement of carrier ampholytes generates excessive heat.

2-1-4 Concentration and determination of protein concentration

To measure protein concentrations, two different techniques were used - a direct UV measurement at 280 nm, or a *Bradford assay*, which is a colorimetric method.

UV measurement

To determine the protein concentration an equation derived from Beer-Lambert's law was used (Equation 2.1).

$$\left(\frac{Absorbance\ 280 \times Dilution\ Factor}{Extinction\ coefficient\ (M^{-1}cm^{-1})}\right) \times Relative\ Molecular\ mass\ (Da)$$

$$= Protein\ concentration\ (mg/ml)$$

Equation 2.1: The Beer-Lambert equation.

Absorbance is measured at 280 nm using a UV visible spectrophometer (GeneQuant 1300), based mostly on tyrosine and tryptophan side chains in the protein absorbing at this wavelength. It should be also noted that there is a small contribution from phenylalanine and from disulphide bonds. The method is simple and direct, and the sample can be recovered. Moreover the specific absorption value must be calculated theoretically for each protein. In the case of *Moringa* protein, the determination of the extinction coefficient was performed using the amino acid analysis (AAA) platform at the Institut de Biologie Structurale (IBS), since the amino acid sequence was unknown.

Bradford assay

The Bradford protein assay is used for the rapid determination of protein concentration. Among the various methods available, it has been commonly used because it is considered fast and sensitive (Bradford,1976). The procedure is a sample one for the determination of

total protein concentration in solutions that depends upon the change in absorbance based on the proportional binding of the dye Coomassie Blue G-250 to proteins. A set of standard was prepared from a stock solution of proteins having known concentration. The Bradford values obtained from the standard is used to construct a standard curve. This standard curve was then used to determine the protein concentration for the samples studied in this thesis work.

2-2-Methods for biochemical characterisation

Amino acid composition and N-terminal sequencing analyses were carried out at Institut de Biologie Structurale (IBS), Grenoble using the amino acid analysis (AAA) platform with the help of J.-P. Andrieu.

2-2-1 Amino acid composition

Amino acid compositional analysis is a classical protein analysis technique and is essential in the identification of native proteins. The method is also used in the quantification of peptides and proteins as it is the only method to determine absolute protein quantities in solution. The determination of the amino acid composition of a polypeptide is a complex analytical process, consisting of two steps, complete hydrolysis of the substrate to liberate the residues followed by chromatographic analysis and quantification of the liberated amino acids (Figure 2.10).
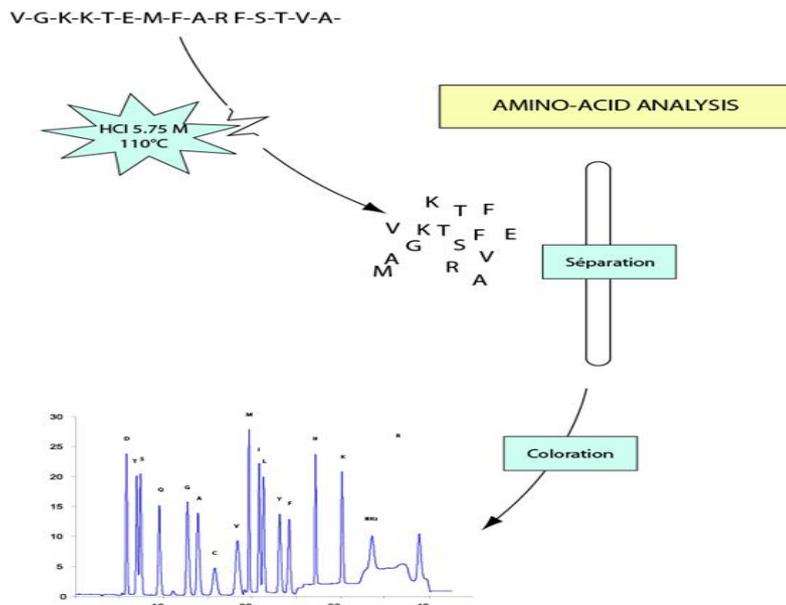
Figure 2.10: The scheme represents the different steps of amino acid composition analysis. The sample is hydrolysed with 5.75 M H-Cl for 20 hours at 110°C. The liberated amino acids are then separated on a cation exchange resin, and after staining, the absorbance of the compound formed is measured at two wavelengths (scheme kindly provided by J.-P. Andrieu).

2-2-2 N-terminal sequencing

The potential presence of the cyclisation of N-terminal glutamine to pyroglutamate (pyro-Glu) for *Mo*-CBP3 protein, as described by Freire *et al.,* (2015), leads to a blocked chain and in this case, the automated Edman degradation fails. This is why the *Mo*-CBP3 protein was pre-treated with a pyroglutamate aminopeptidase of *Pyrococcus furiosus or Pfu* pyroglutamate aminopeptidase.

Aminopeptidase digestion

The *Pfu* pyroglutamate aminopeptidase was used for the removal of the N-terminal pyroglutamic acids before Edman degradation. The hyperthermophilic Archeaeon *Pyrococcus furiosus (Pfu)* enzyme was used because it is superior in both thermostability and specific activity to any other Pyrrolidone carboxyl peptidases (Pcps) found so far.
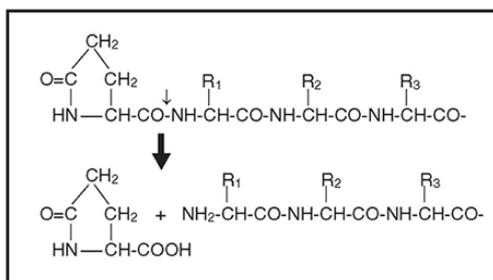
Figure 2.11: *Pyrococcus furiosus* (*Pfu*) Pyroglutamate Aminopeptidase Activity (image taken from www.clontech.com)

N-terminal sequencing using the Edman degradation method

N-terminal sequencing is a process to identify the order of amino acid arrangements in a peptide or protein. This method was also applied to the full length protein as a cleavage product obtained after limited proteolysis. The N-terminal amino acid was pre-digested by *Pfu* pyroglutamate aminopeptidase (Figure 2.11) and the *Mo*-CBP3 proteins were blotted on polyvinylidene difluoride (PVDF) membrane after a Tris-Tricine gel. The protein band was cut from the membrane and analysed on a sequencer performing Edman degradation.

The Edman degradation, developed by Pehr Edman in the 1950's, consists of the labelling of the amino-terminal residue and its cleavage from the peptide without disrupting the peptide bonds between other amino acid residues. The reagent used is phenyl isothiocyanate (PITC), which reacts with the amino nitrogen of the N-terminal amino acid in order to obtain a phenylthiocarbamoyl (PTC) derivative. This PTC derivative is then treated with HCl in an anhydrous solvent. The N-terminal amino acid is cleaved from the remainder of the peptide and leads to a thiazolone which rearranges to a phenylthiohydantoin (PTH) derivative (Figure 2.12)
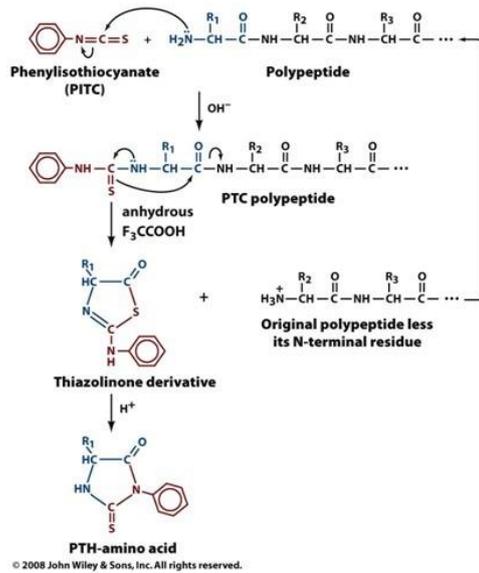
Figure 2.12: The Edman degradation pathway is based on the reaction of phenylisothiocyanate (PITC) with the free amino group of the *N*-terminal residue. The phenylthiohydantoin (PTH) derivative obtained are removed one at a time and identified by chromatography (taken from Handbook of proteins: structure functions and methods 2008).

This last product is isolated and identified using chromatography (Figure 2.13). The remainder

of the peptide is subjected to a second Edman degradation.
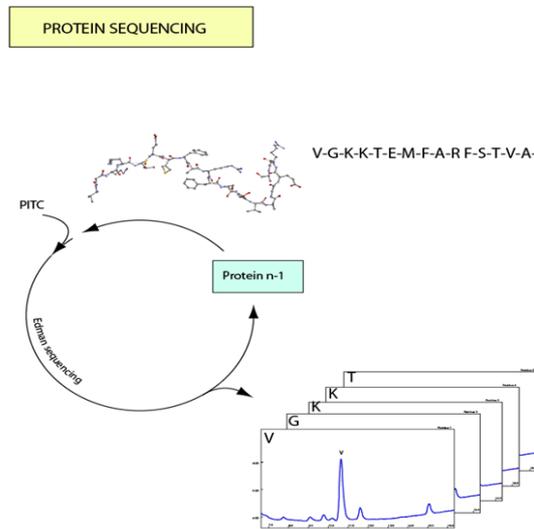
Figure 2.13: The scheme represents the different steps of the N-terminal sequencing method. Edman degradation involves a series of chemical steps that remove the amino acid from the amino terminal end of polypeptide. The released amino acid derivative is identified and the process is repeated through several rounds of amino acid removal and identification (scheme kindly provided by J.-P.Andrieu).

The sequence obtained was submitted to automatic alignment software, performed using the NCBI_BLAST search system (Coordinators NR. Database Resources of the National Center for Biotechnology,2017) Generally, a sequencing of 5 to 8 residues is sufficient to identify a protein present in this database. Larger sequences are possible if necessary - up to a maximum of 20.

2-2-3 Proteolysis

Limited proteolysis

Limited proteolysis is a simple biochemical method which was performed to identify *Moringa* seed proteins. During limited proteolysis, a protein is incubated with a relatively low concentration of different proteases such as trypsin, which cuts at recognition sites throughout the protein, normally at exposed regions such as loops and other flexible regions. Trypsin predominantly cleaves peptide chains at the carboxyl side of the amino acids lysine and arginine, except when either is followed by proline. Following digestion with a protease, samples are analysed using Tris-Tricine gels to identify cleavage products. The appearance of lower molecular weight bands denotes digestion of the full length protein. Variables that can be adjusted to optimise results are: type of protease, protease dilution, temperature or time of incubation. The cleavage products are transferred on a PVDF membrane and sent for N-terminal sequencing analysis.

Trypsin in gel digestion

In gel digestion is a part of sample preparation for the mass spectrometric identification of proteins. The technique consists of 4 steps: destaining, reduction and alkylation (R&A) of the cysteines in the protein, proteolytic cleavage of the protein, and peptide extraction. The step

of reduction and alkylation of cysteines residues may be optional but has the advantage that it allows the protease to better access the protein. This method was introduced by Rosenfeld (1992) and many modifications and improvements have since been made. After staining, the gel was extensively washed with pure water to remove all SDS prior to gel band excision. The gel band was cut with a razor blade and placed in a 0.5 ml Eppendorf tube. The gel piece was first washed with 50 µl of 50 mM NH_4HCO_3 for 30 minutes in a thermomixer, then the buffer was discarded and a second wash was performed with 50 µl of 50% 50mM NH_4HCO_3/ 50% acetonitrile (ACN) for 30 minutes. This second step was repeated twice and a last wash of 5 minutes was carried out with 50 µl of ACN 100% and the solvent discarded. 10µl of 19 ng/µl Trypsin (Promega) were added to the gel pieces and they were covered with 25 mM NH_4HCO_3. Digestion was carried out in a shaker overnight at 37 °C. The reaction was stopped by addition of 2 µl of 50 % of trifluoroacetic acid (TFA); the supernatant was kept. The peptide extraction was performed by washing successively with 50 µl of 5% ACN / 0.1% TFA, then 50 µl of 50% acetonitrile / 0.1% TFA was added to the gel slice and agitated for 20 minutes. Finally 50 µl of ACN was added and shaking carried out for 5 minutes at RT. The eluates were lyophilized to a volume of 10 µl volume and samples analyzed using MALDI-TOF MS. MS data were processed automatically using Mascot Distiller software (v. 2.5.1, Matrix Science; Perkins *et al*., 1999). The searches were carried out using a sequence database containing the *Moringa* sequences derived.

2-2-4 Glycosylation detection of the state of *Mo*-CBP3 isoform

Glycoproteins can be readily located in polyacrylamide gels and distinguished from other non-glycosylated proteins. The detection methods include the periodic acid-Schiff (PAS) stain. The presence of carbohydrate was evaluated by specific staining of the *Mo*-CBP3 band after Tris-Tricine gel electrophoresis using the Glycoprotein Detection Kit (Sigma). This detection system is a modification of Periodic Acid-Schiff (PAS) methods and yields magenta bands with a light pink or colourless background. The detection limit has been found to be in the range of 25-100 ng for carbohydrates depending on the nature and the degree of glycosylation of the protein. Horseradish peroxidase with a carbohydrate content of 16% is used as a positive control in the kit.

2-3-Methods for activity characterisation

Several tests were conducted to characterise *Mo*-CBP3 isoforms including minimum inhibitory concentration (MIC) determination, antifungal activity, flocculation on latex beads, and coagulation assays on living cells.

2-3-1 Determination of the Minimal Inhibitory Concentration (MIC)

MICs are defined as the lowest concentration of an antimicrobial substance that will inhibit the visible growth of a microorganism after overnight incubation. MICs are used by diagnostic laboratories mainly to confirm resistance, but also as a research tool to determine the *in vitro* activity of new antimicrobials, and data from such studies have been used to determine MIC breakpoints. The tube dilution test is the standard method for determining levels of resistance to an antibiotic. Serial dilutions of the antibiotic are made in a liquid method which is inoculated with a standardized number of organisms for a prescribed time The results are

measured using agar dilution or broth microdilution, usually following the guidelines of a

reference body such as the CLSI (Clinical and Laboratory Standards Institute), BSAC (British

Society for Antimicrobial Chemotherapy) or EUCAST (European Committee on Antimicrobial

Susceptibility Testing) (Figure 2.14). The lowest concentration of antibiotic preventing

appearance of turbidity is considered to be the MIC. The standard strains used for the MIC of

Moringa protein are *S. aureus* ATCC29213 and *E. coli* ATCC25922, following CLSI guidelines.
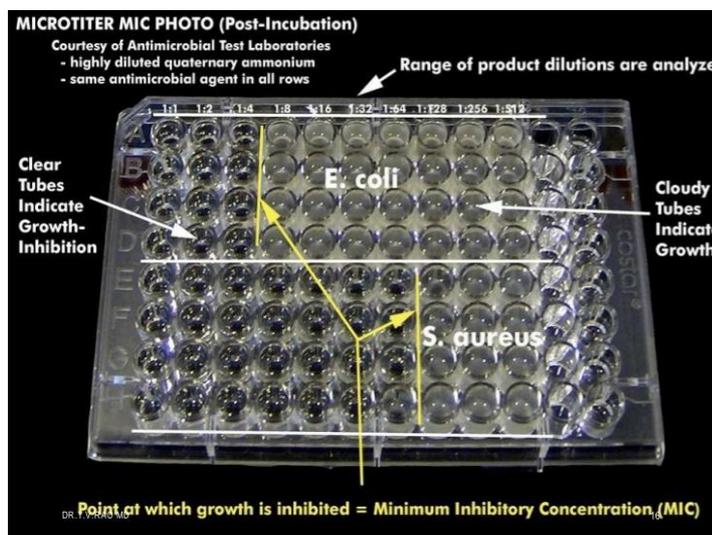
Figure 2.14: Example of the minimum inhibitory concentration (MIC) determination using a microplate system. The clear wells indicate the growth-inhibition whereas the cloudy wells indicate the growth. The point at which growth is inhibited is called the MIC (http://www.slideshare.net/doctorrao/minimum-inhibitory-concentration).

The broth medium was Mueller-Hinton 2 broth (MH2B, bioMérieux, Marcy L'Etoile, France).

4.8 mg of freeze-dried Moringa protein were dissolved in MH2B to reach a concentration of

20 g/L and 20 mg of seed extract were dissolved in MH2B to reach a concentration of 80 g/L.

One row of a 96-well microtiter plate was filled with 90 µL of two-fold serial dilutions of

Moringa protein, seed extract or gentamicin (Panpharma) taken as control, in MH2B medium,

so as to obtain final concentrations ranging from 0.02 to 10 mg/ml for Moringa protein ; 0.08

to 40 mg/ml for seed extract and 0.03 to 16 mg/l for gentamicin. A bacterial inoculum (10 µL per well, 5×10^5 CFU/mL of final inoculum) was then added to each well. Antibiotic free cultures were used as positive controls and MH2B served as a negative control. Microplates were incubated at 37°C in ambient atmosphere and the MICs were read after 20 h culture incubation.

2-3-2 Antifungal and synergistic activities

The antifungal and synergistic tests were carried out at the Centre Hospitalier Universitaire de la Tronche-Grenoble (CHU) in the Parasitologie–Mycologie group supervised by Professor M. Cornet. The antifungal test is based on the growth inhibition of various human pathogenic fungi in presence of *Mo*-CBP3 isoforms. After 24 and 48 hours, the turbidity of the suspension is visually observed and the test is repeated three times. The synergy tests were performed by mixing protein with antifungals at different concentrations and the MIC was determined using the same human pathogens. The reference method used to determine the MIC of an antifungal is called EUCAST (Rodriguez –Tudela et *al.,* 2008).

2-3-3 Gel diffusion assay for chitinase activity

A gel-diffusion assay was performed to evaluate chitinase activity. The assay is based upon diffusion of the enzyme from a central well through an agarose gel containing the substrates glycol chitin, a soluble modified form of chitin, used for assaying chitinase activity (Trudel *et al.*, 1989; Pan *et al.*, 1991). Gel plates for chitinase assays were prepared by melting 1.6% (W/V) agar in incubation buffer (0.1M citric acid, 0.2M sodium phosphate , pH5.0); the solution was cooled down to 60°c to 50°c and glycol chitin (0.5%) was added to the solution and mixed well. 30 ml was dispensed into a disposable Petri dish (diameter 9 cm). When the solution solidified, wells were punched with a Pasteur pipette. 10 µl of protein were added to

the well. After incubation at 28°C for 48 hours, the gel was stained with 20ml 0.1% calcofluor (Sigma ref18909) for 10min. The gel plate was washed with distilled water and chitinase activity visualized under a UV light. The chitinase diffuses from the well and catalyzes the cleavage of glycol chitin, leaving a clear non-fluorescent zone in the gel, the diameter of which is proportional to enzyme activity. The detection of chitinase activity is based on the affinity of calcofluor for chitin (Maeda *et al.,* 1967). Calcofluor fluoresces under UV light when it binds to undigested chitin, whereas the region where glycol chitin was digested appears dark under UV light, as calcofluor does not have affinity for the digested chitin.

2-3-4 Flocculation and coagulation test

The treatment of drinking water involves a number of combined processes based on the quality of the water source such as turbidity and amount of microbial load present in water. Coagulation/ flocculation are the most common processes used to purify surface water from impurities (Figure 2.15).
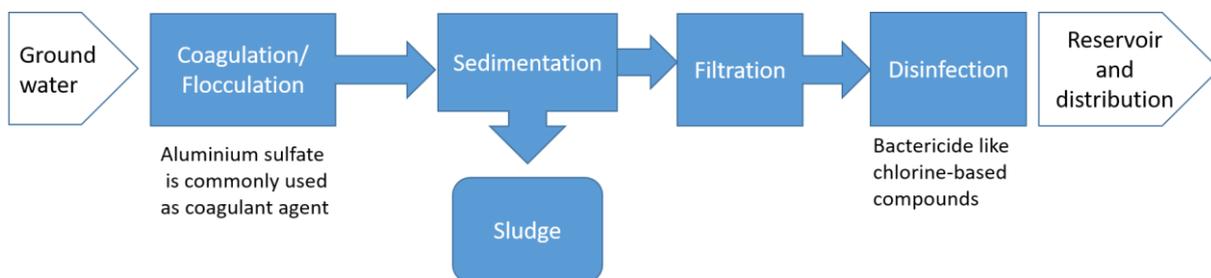
Figure 2.15: Scheme summarising conventional water treatment. It consist of different unit processes: coagulation-flocculation, sedimentation, filtration and followed by disinfection often done by chlorination.

These methods were adapted on a small scale in this thesis work to demonstrate the role of *Mo*-CBP3 protein in the water treatment by a flocculation activity on latex beads and the coagulation effect on living cells.

Latex beads

Polystyrene lattices were used for flocculation tests. These were synthesized in Uppsala University (Sweden). The hydrogenous latex chosen for flocculation is designated PS3 and the particle have a radius of 721 ± 5 Å. They were prepared following the procedure of Goodwin *et al.,* (1974).

Flocculation methods

Flocculation is a physical and chemical process which is used for the removal of the visible sediments and material that makes it a colloidal solution. It can be performed by agitation or by adding flocculating agents like latex beads, as in this study. In the flocculation process, the polymers are used as flocculating agent for the formation of bridges between the flocs (clumps of bacteria and impurities which form clusters). After slow addition of anionic flocculants or flocculating agents, these agents get absorbed on the particle by reacting with the positively charged suspension. It is essential to mix the flocculating agent gently at a slow speed so that small flocs can easily agglomerate into large particles. These large particles are known as *clumps* or *agglomerates*. They are formed by the aggregation of particles with the adsorption of the polymer chain segments on the particles (Figure 2.16)
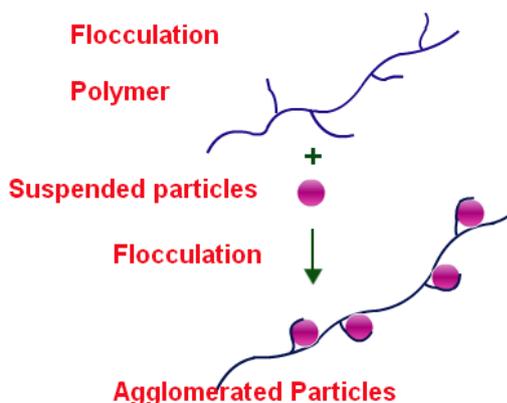
Figure 2.16: Scheme showing flocculation process taken from https://chemistry.tutorvista.com/physical-chemistry/flocculation.

After flocculating the suspended particles into larger particles, they are generally separated from the fluid by a sedimentation technique as there is a density difference between suspended matter and the liquid. The use of a polymer as a flocculating agent should be done with great care because an overdose will make the setting process difficult. As they are lighter than water, a high dose of anionic polymer will increase the floating ability of flocculation. The preparation procedure consisted of the addition of 20 µl of pure protein from a stock solution with concentration 4 mg/ml to 20 µl of the dispersion particles previously diluted with an equal volume of pure water. The solution was gently mixed for a few seconds and immediately observed under a microscope.

Coagulation methods applied on living cells

The purpose of coagulation is to make solid, suspended colloidal particles stick together . This is done by adding a sufficient amount of cationic coagulant which helps in the neutralisation of the suspended particles. In this thesis work, *Mo*-CBP3 isoforms played the role of cationic coagulant, and the living cells like *E. coli* or algae that of suspended particles. In flocculation, a flocculent is added to form clumps of impurities. These are called *flocs*. These flocs are heavy enough to settle down. The flocculation is done after coagulation process because it is easy for thick particles to make clumps of heavy particles.
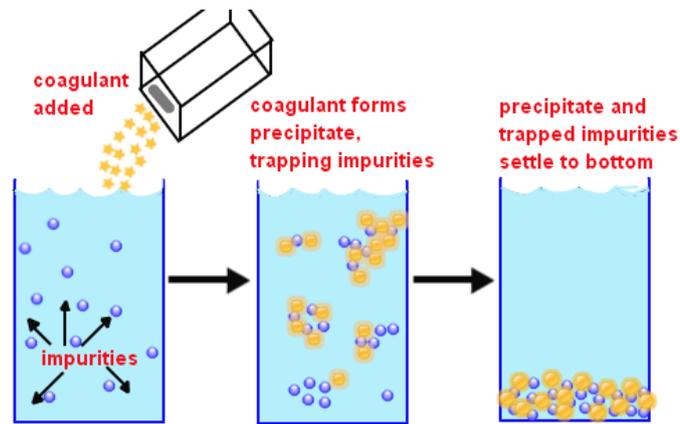
Figure 2.17: Scheme showing the general coagulation process in water treatment. In this work, impurities have been replaced by living cells such as bacteria or algae and the coagulant factor is *Mo*-CBP3 isoforms (taken from https://chemistry.tutorvista.com/physical-chemistry/flocculation.html).

2-4-Methods for Biophysical characterisation

2-4-1 Mass spectrometry (MS) and tandem mass spectrometry (MS/MS).

Mass spectrometry is an analytical technique that was used to determine the mass of different Mo-CBP3 isoforms and seed extract proteins. Electrospray ionisation (ESI) and matrix-assisted laser desorption/ionisation (MALDI) techniques were used. The MS experiments were carried out with the help of Luca Signor at Institut de Biologie Structurale (IBS), Grenoble. 10 µM of sample in 20µl of water was ionised and the mass to charge ratio measured.

Mass spectrometry analysis is achieved by ionising the sample and the mass/charge ratio is measured. The sample is 'cleaned' to remove any salt and buffer that would interfere with the ionisation, and is then redissolved into an organic solvent. ESI is a soft ionisation technique used to help prevent fragmentation of the components. The environment from the ionisation source to the detector is under vacuum to prevent interference from other gases present. The charged sample is passed through an electrically charged capillary which forms a spray of ions

to form lots of little droplets. These droplets are then passed through an environment where they are dispersed by a combination of heat and nitrogen gas (N_2). The molecule of interest is now in the gas phase without any solvent present; this is then passed through a time-of-flight (TOF) mass analyser. The analysis occurs according to the mass and charge of the molecule in question: the larger the ion, the longer its time-of-flight. As proteins can have many charge states, and a mass is calculated for each, the final mass is very accurate due to it being an average of all these states.

Tandem MS (or MS/MS) was performed by Syvie Kieffer-Jaquinod at Commissariat à l'Energie Atomique et aux energies Alternatives (CEA) in Grenoble. It was used to identify the different isoforms of *Mo*-CBP3, their modification states (pyroglutamate positions and disulphide bridges) and their abundance. In this technique, peptides obtained from trypsin in gel digestion of proteins are fragmented into daughter ions, which provides information regarding the amino acid sequence of the peptide.
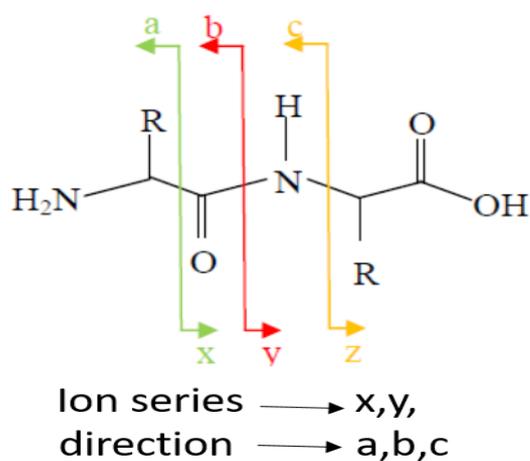
Figure 2.18: Schematic diagram of daughter ion nomenclature adapted from Roepstorff &Fohlmann (1984). A positively charged peptide (in black) is fragmented and the daughter ions are shown (a,b,c,x,y,z).

The MS fragmentation occurs in the mass spectrometer mass analyser or in a collision cell through the action of collision energy on gas phase ions generated in the mass spectrometer

ion source. Several parameters influence this fragmentation process, including amino acid

composition, size of peptide, excitation method, time scale of the instrument used, ion charge

state (Paiz & Suhai., 2005). After the mass spectrometer analysis, a file is created containing a

list of masses observed for the peptides, which could be used as precursor ions (Figure 2.19)
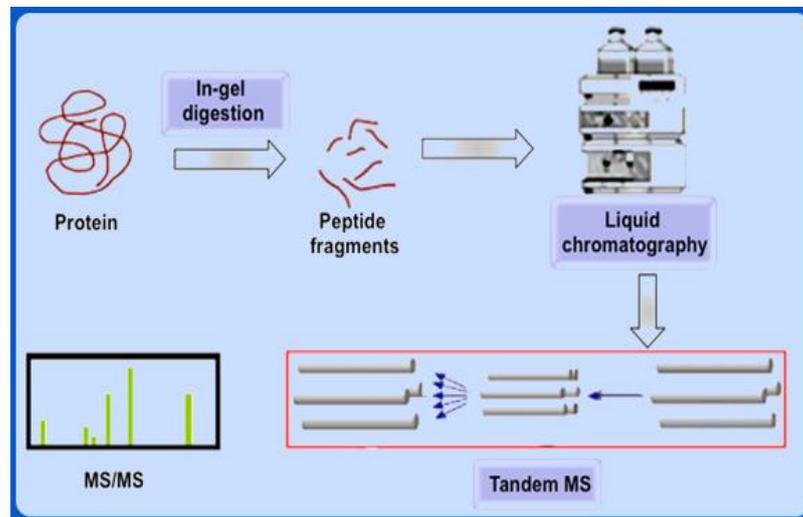
Figure 2.19: A typical LC-Ms/MS set up for proteomics applications. The protein of interest is pre-digested into small fragments and, after separation by Liquid chromatography, they are analysed by tandem MS. The spectrum generated provides a set of peaks whose masses represent each of the peptide present in the mixture (taken from http://nptel.ac.in/).

For each peptide selected for fragmentation there are, in this experimental file, associated

peptide fragments and all these data can be analysed in order to obtain information regarding

the amino acid sequence of the peptide and further used for protein identification or

characterization (Cottrell, 2011). There are basically two ways to analyse this type of data:

manually or submitting the data to search engines where they are compared to selected

protein sequence databases. The software, based on mathematical algorithms, make a

theoretical digestion, or *in silico* digestion, of all proteins present in a database and generate

theoretical fragments of these peptides. After this, a comparison is made between the virtual

and experimental data obtained in the mass spectrometer, and a score is attributed to the peptide or protein identified, where high scores indicate good confidence in the identification.

2-4-2 Circular Dichroism spectroscopy (CD)

Proteins are made of amino-acids residues which are chiral molecules and thus have optical activity: when illuminated with a polarised light a solution is able to rotate the polarisation direction. The intensity of rotated light depends on both the incident light wavelength and the relative orientation of amino-acids (as well as protein concentration). Hence, a CD spectrum will vary with the secondary elements present in the protein structure. Each type of secondary structure (α-helix, β-sheet turn or random coil) results in a characteristic spectrum (Figure 2.20).
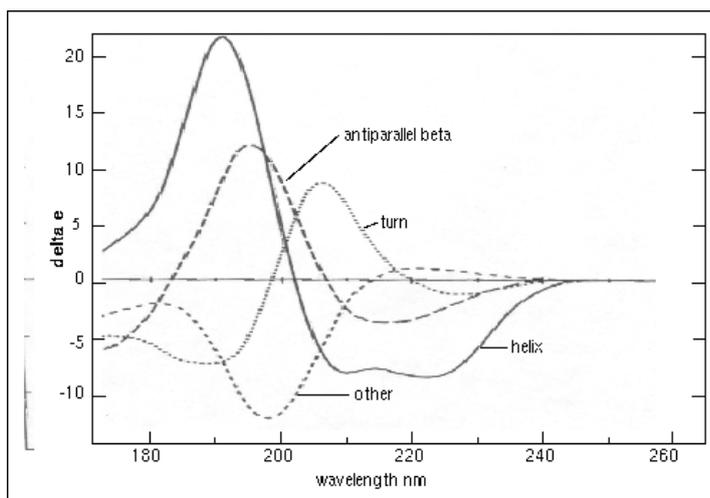
Figure 2.20: Circular dichroism spectra of a « pure » secondary structure (adapted from Brahms *et al.,* 1980)

The average composition of the sample in terms of secondary structure elements can be deduced from the resulting curve. In this project work, CD was used as a way of assessing changes in structure or stability of the *Mo*-CBP3 protein by comparison with that of the crude extract.

2-5 X-ray crystallography

X-ray crystallography is one of the few techniques that allows detailed structural information of biological macromolecules to be obtained. It is the most common technique used to get atomic level detail even for large macromolecular assemblies. The biggest limitation of X-ray crystallography is the requirement for suitably diffracting crystals. Crystals are assemblies of a precisely well-defined molecular unit repeated regularly in three dimensions. The crystals are mounted on a diffractometer at an X-ray source (*eg* such as at the European Synchrotron Radiation Facility (ESRF) in Grenoble) and diffraction patterns collected for a detailed range of crystal orientations in the beam, depending on the crystal symmetry. These patterns consist of many thousands of spots, each of which is measured in position and intensity. These data provide information on the molecular structure to a resolution that depends on the quality of the crystal.

2-5-1 Crystallisation of *Mo*-CBP3-4 protein

Initially, the crystallisation conditions of *Mo*-CBP3-4 were unknown. The parameters to be explored included a large range in pH, ionic strength, temperature, protein concentration, the presence of various salts, the type of precipitant, as well as the crystallisation methodology itself (hanging drop, sitting drop, dialysis, etc.).

High-throughput crystallisation screening

The initial crystallization screening was carried out at the High Throughput Crystallisation (HTX) Laboratory of the EMBL Grenoble outstation. The *Cartesian* robot of the HTX platform uses nanodrops (100 nl) in a sitting drop configuration - allowing automatic screening of 768 different crystallisation conditions from commercially available screens. A total of 5 imaging

inspections was applied to each plate over a period of 12 weeks as follows 1, 3, 7, 15, 33, 61, and 87 days. Promising conditions were then reproduced manually with the technique of hanging drop vapour diffusion to optimise the conditions and obtain the best possible quality and size of crystals.

Vapour diffusion: Hanging and sitting drop methods

Vapour diffusion can be performed using either hanging-drop or sitting-drop methods. The hanging drop method consists of suspending a pure and concentrated protein from a glass cover slip over a pool of precipitant. The drop slowly evaporates to equilibrate with the well below. The chamber must be sealed. The reservoir contains 1 ml of precipitant. For the hanging drops, the drop was typically 2 µl in volume. The sitting drop method is essentially the same as the hanging drop technique except for the fact that the drop can be larger and placed on the bridge which sits over the precipitant reservoir of 1ml. The drop sizes for sitting drops range from 2 µl to 10 µl.
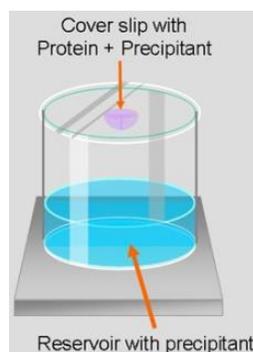
Figure 2.21: Schematic diagram showing a well reservoir, containing a precipitant solution, capped with a cover slip, as used in the hanging drop technique (taken from http://www.xtal.iqfr.csic.es).

In vapour diffusion methods, the equilibrium of the chemical potentials in the well is responsible for the evaporation of water in the drop towards the well, increasing gradually the protein concentration. If this balance crosses the nucleation zone, crystalline seeds will

form. The concentration of soluble protein will then decrease, allowing a crystalline growth phase (Figure 2.22)
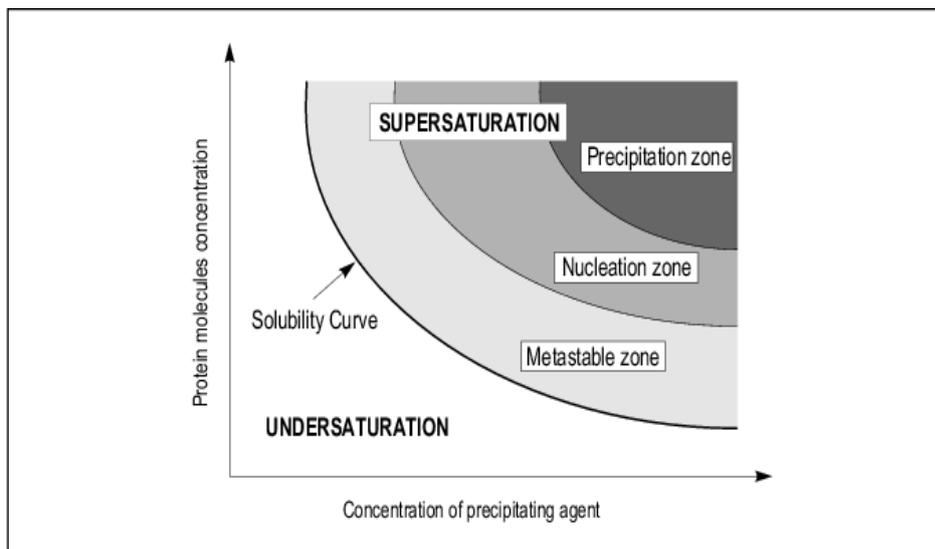
Figure 2.22:. Schematic representation of a phase diagram, showing the protein concentration plotted against the precipitating agent concentration. The concentration space is divided by the solubility curve into two areas corresponding to undersaturated and supersaturated state of a protein solution. The supersaturated area comprises of the metastable, nucleation and precipitation zones (taken from Ducruix, A. and Giege, R. (1992)).

2-5-2 Data acquisition and processing

Cryo-crystallography

In a protein crystal, the interface between adjacent units is usually made of weak interactions producing a fragile assembly which is sensitive to heat (through thermal motion). Synchrotron X-ray beam-lines produce a high intensity flux, part of which is absorbed by the crystal. The absorbed energy generates heat and free radicals which typically damage the crystal packing (radiation damage) and limit the quality of the available data. To enhance crystal lifetime in the X-ray beam, crystals can be frozen and kept at very low temperature (80 to 100 K). In this process the water ideally remains glass-like (versus crystalline) in character to avoid both

crystal breakage and strong diffraction due to ice crystals (problematic for data collection). To achieve this goal, cryoprotectants are used to enhance the vitreous character of the sample

X-ray diffraction on beamline ID 29 at ESRF

All the crystallographic data were collected at the European Synchrotron Radiation Facility (ESRF) situated in Grenoble (France) on the ID29-1 beam-line. The beamline uses an undulator to provide a high brilliance,monochromatic beam. ID29 is a fully automated macromolecular crystallography beamline intended for high-energy resolution anomalous dispersion phasing experiments and for high-resolution X-ray diffraction experiments. This beamline is equipped with a MicroDifractometer (MD2) which allows to tailor the beam sizes down to 75 (full beam), 50, 30, 20 and 10 microns in diameter. The beamline also operates an EMBL/ESRF sample changer. Diffraction data are recorded with a fast readout Pilatus 6 M pixel detector (Dectris LTD) that allows data collection with a maximum frame rate of 12 images/s.
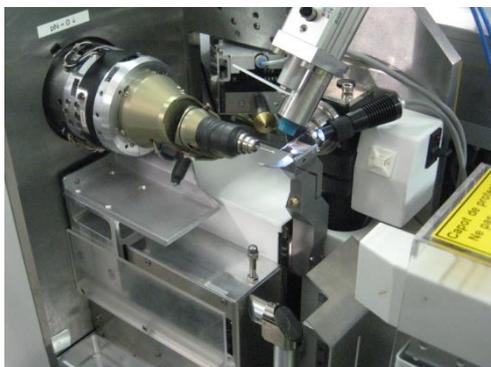
Figure 2.23: The ID29 diffractometer at the ESRF – the instrument is optimised for high resolution macromolecular crystallography.

Image processing

Diffraction data were processed using the XDS program package (Kabsch, 2010). The aim of this programme is to obtain accurate unit cell dimensions and experimental parameters and then use these parameters to index each spot (assign Miller indices) and integrate its intensity.

2-5-3 Phasing and structure solution

The structure solution was determined by sulphur SAD phasing method using *SHELX* program suite (Sheldrick, 2008).This method allows the determination of protein structures *de novo* without reference to derivatives such as selenomethionine. The positions of the anomalous scatterers were found using *SHELXD* and the initial phase information was determined using *SHELXE*. The final phases obtained were then used by the automated model-building routines in ARP/*w*ARP *7.3* (Langer *et al.,* 2008) to construct an atomic model, which was refined using REFMAC*5* (Murshudov *et al.,* 1997). Electron density maps were viewed using *COOT* (Emsley & Cowtan, 2004).

2-6 Neutron reflectometry

In this work, the main experimental technique used for the study of adsorption on flat solid substrates and at air-liquid interfaces was neutron reflectometry (NR). Its ability to resolve structure at the nanoscale level and the possibility to contrast match some parts of the system makes it an appropriate tool for interaction studies of adsorption on mineral surfaces. A reflectivity experiment consists in sending a neutron beam on a surface and varying the scattering wave vector Q. Neutrons are scattered in matter by nuclei, unlike X-rays which are scattered by electron clouds, and the difference in scattering cross section for hydrogen and deuterium is very high (Figure 2.24) .
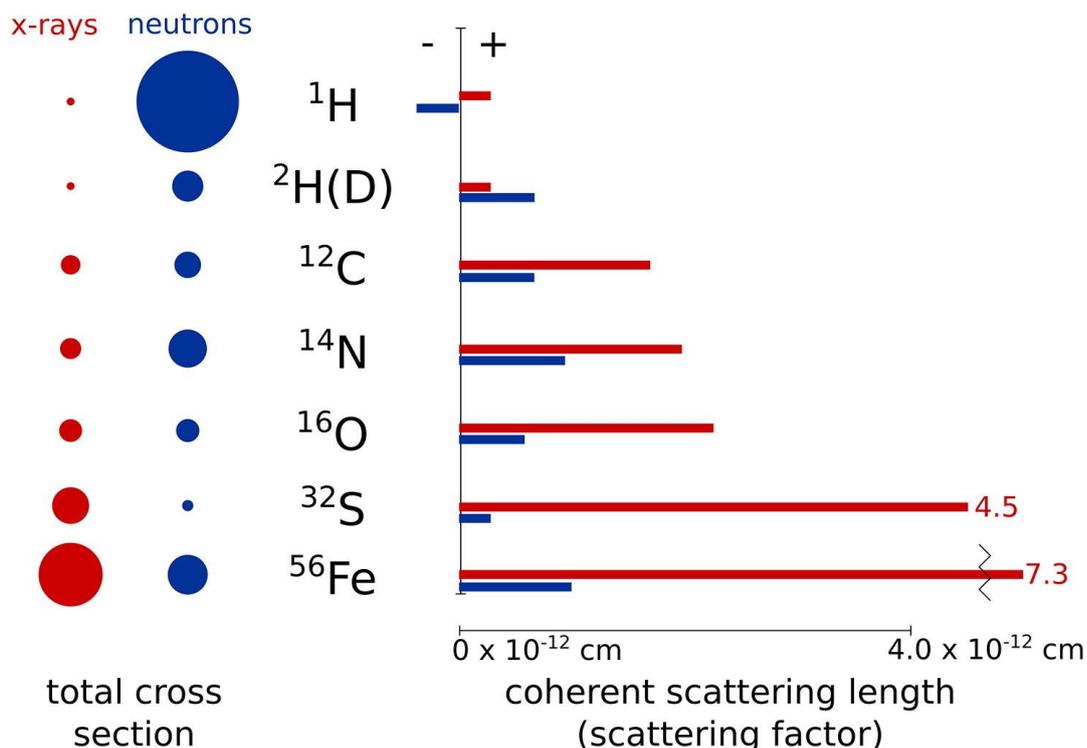
Figure 2.24: X-ray and neutron scattering cross sections and coherent scattering lengths (scattering factors) for different elements. Circle and bars are drawn to scale (taken from Castellanos *et al.*,2017).

Consequently, neutron techniques are very useful for studying structural properties of highly protonated materials such as biomolecules because it is possible to substitute the hydrogen atoms of the compound by deuterium in order to highlight it within a fully hydrogenated surrounding environment.
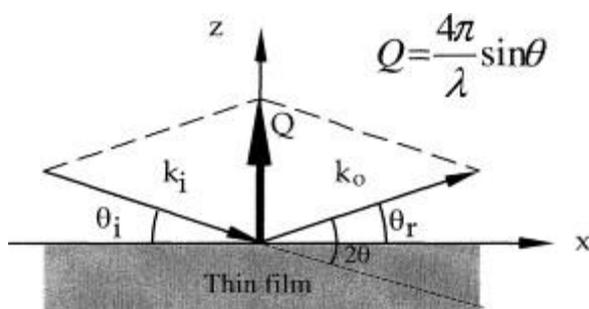
Figure 2.25:.Geometry of a specular reflectivity experiment. The scattering wave-vector Q is perpendicular to the plane of the thin film.(Ott *et al.*, 2004)

In a reflectivity experiment, the incident angle θ_i on the surface is small (typically ranging from 0.5 to 5°). The reflection angle θ_r is the same as the incidence angle θ_i. As the consequence, the scattering wave-vector **Q** is perpendicular to the surface. The typical neutron wavelengths are in the range of 2-20 Å. Thus, the range of accessible scattering wave-vector $\mathbf{Q} = k_0 - K_i$, is the range of 0.05-3 nm^{-1}. This corresponds in the real space to typical lengthscales between 2 nm and 100 nm. NR is a technique adapted for the study of thin films but does not probe structures at the atomic level. The interaction of the protein with mineral surfaces was investigated by measuring neutron reflectivity with the D17 reflectometer (Cubitt & Fragneto, 2002) at the Institut Laue Langevin (ILL) in Grenoble.

3-Biochemical and biophysical characterisation and activity assays of the *Mo*-CBP3 protein

This chapter focuses on the biochemical and biophysical characterisation of *Mo*-CBP3 proteins that have been purified and fractionnated from *Moringa* seed extract initially prepared by our collaborator Dr Kwaambwa (Namibia) and provided via Professor Rennie (Uppsala University). The work described was carried out on native proteins whose amino acid sequences were unknown at the beginning of this study. The subsequent development of a purification strategy allowed the isolation of a main protein fraction (named fraction C1) that eluted at high salt using ion exchange chromatography. The biochemical characterisation (amino acid composition, N-terminal sequencing, Epsilon determination and heat stability tests) focused on this fraction and has shown the presence of mainly small highly positively charged proteins having high heat stability and proteolysis resistance that rendered the N-terminal sequencing analysis extremely challenging. This structural stability was confirmed by circular dichroism spectroscopy showing that there was no effect of heat treatment on the secondary structure. The combination of mass spectrometry (MS) analysis and tandem MS allowed the identification of the isoforms (*Mo*-CBP3-3 and *Mo*-CBP3-4) in this specific fraction. The flocculant properties of these isoforms on different organisms were demonstrated, showing that*Mo*-CBP3 protein family plays a key role in water treatment.

3-1 Introduction

The overall strategy was to identify native proteins inside the *Moringa* seed extract responsible for the flocculation activity and to characterise them using biochemical and biophysical techniques. Section 3.1 describes the water soluble extraction and the

fractionation of the crude extract. A considerable amount of time was put into the characterisation and identification of the main peak (fraction C1) that elutes at the highest salt concentration during ion exchange chromatography using biochemical techniques such as amino acid composition, N-terminal sequencing, Epsilon determination and heat stability tests (section 3.2). The publication of the sequences of 4 isoforms of the *Mo*-CBP-3 protein family (Freire *et al.*, 2015) was crucial in facilitating the biophysical characterisation and the identification of the proteins in this fraction especially by the MS/MS technique described in the section 3.3. Section 3.4 focuses on the activity assays comprising antibacterial, antifungal activities and flocculation tests. The minimal inhibitory concentration (MIC) and the antifungal assays were mainly performed through a collaboration established with the Centre Hospitalier Universitaire (C.H.U) at Grenoble in the unité Médicale de Bactériologie-Hygiène Hospitalière of Professor Max Maurin, whereas flocculation tests were carried out in the laboratory. The Chapter finishes with a discussion/conclusion section that summarises the overall results and their implications for the structural and reflectometry studies which are described later in Chapters 4 and 5 of this thesis.

3-2 Purification of *Moringa* proteins

3-2-1 Extraction and analysis of the protein seed extract

The protein seed extract (crude extract (CE)) provided by Prof. A. Rennie (Uppsala University, Sweden) and Dr. Majority Kwaambwa (Polytecnic of Namibia, Windhoek) was prepared from seeds of Moringa *oleifera (Mo)*. These seeds were collected in Namibia. The initial extraction of *Moringa* protein involves treatment with petroleum ether to remove oil, extraction of the protein with water, precipitation of protein with ammonium sulphate and filtration. The

precipitate is then dissolved in water and dialysed to remove excess ammonium sulphate. This is followed by elution through a carboxymethyl (CM) cellulose column using 1 M NaCl. The purified protein is dialysed and finally freeze dried. This procedure has been described by Kwaambwa and Maikokera, 2008 (Figure 3.1).

Figure 3.1: Schematic summary of the procedure used for the extraction of water soluble protein from the crude extract (CE) of Moringa seeds. Approximately 10 g of seeds would provide 100 mg of CE.

Three different lyophilised seed extracts were studied during this PhD project. All batches were obtained from different trees of the same species. The first two batches (batches 1 and 2) allowed the purification protocol to be established and the first crystallisation trials (described in the next chapter) to be undertaken. Most of the characterisation results described in this chapter were conducted on the last batch (batch 3).
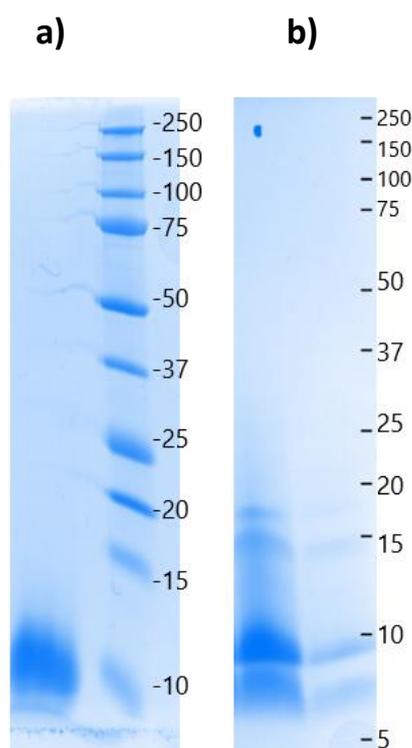
Figure 3.2 :Tris-Tricine polyacrylamide gel electrophoresis (PAGE) analysis of the crude extract (CE) content **a)** 12% Tris-Tricine gel **b)** Tris-Tricine gel with a gradient 4 to 16 %.

Figure 3.2 shows a polyacrylamide gel that illustrates the protein composition of the seed extract. 5 µl of CE at a concentration at 4 mg/ml were loaded onto two types of Tris-Tricine gel. The 12% Tris-Tricine gel shows the presence of a major protein with a molecular weight (MW) around 10 Kilodalton (kDa) (Figure 3.2) whereas the Tris-Tricine gel gradient analysis revealed the presence of at least 2 bands which migrate between 5 and 10 kDa. An upper band of 15 kDa is also visible which may reflect the presence of a dimer.

3-2-2 Purification of the crude extract of *Moringa* seed proteins

The protocol for purification used was described by Gebremichael *et al*.,(2005) and consisted of one step of carboxymethyl cellulose (CM) sepharose purification. In this project, peaks of proteins resulting from the CM step were further purified using size exclusion chromatography (SEC). This second step was added to improve the purity of the sample required for the biochemical and biophysical characterisation and also to perform crystallisation trials.

CM sepharose purification

20 mg of powdered CE was resuspended in 2 ml of CM sepharose buffer (1O mM ammonium acetate pH 6.7), and loaded onto 5 ml of a handmade sepharose CM column. A gradient from 0 to 60 % of CM sepharose buffer with 1 M NaCl for 8 column volumes (CV) was applied. The elution fractions were analysed on a 4-20% Tris-Tricine gradient gel (Bio-rad) and fractions corresponding to the main peak C were pooled. The purification was monitored using UV spectroscopy at 280 nm, which produced a chromatogram that indicated the presence and concentration of proteins.

a)
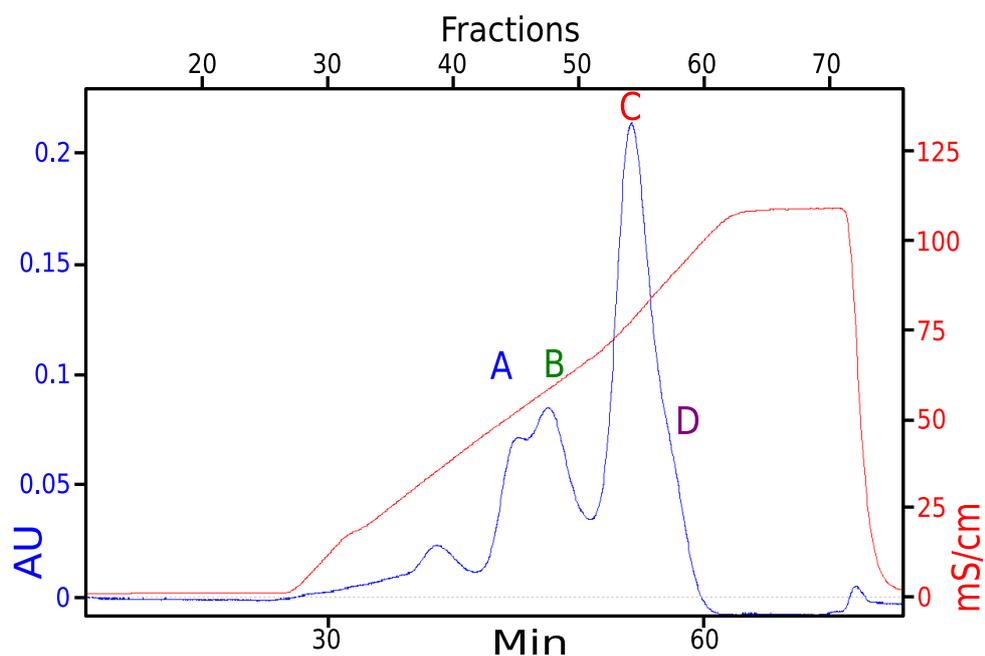
b)
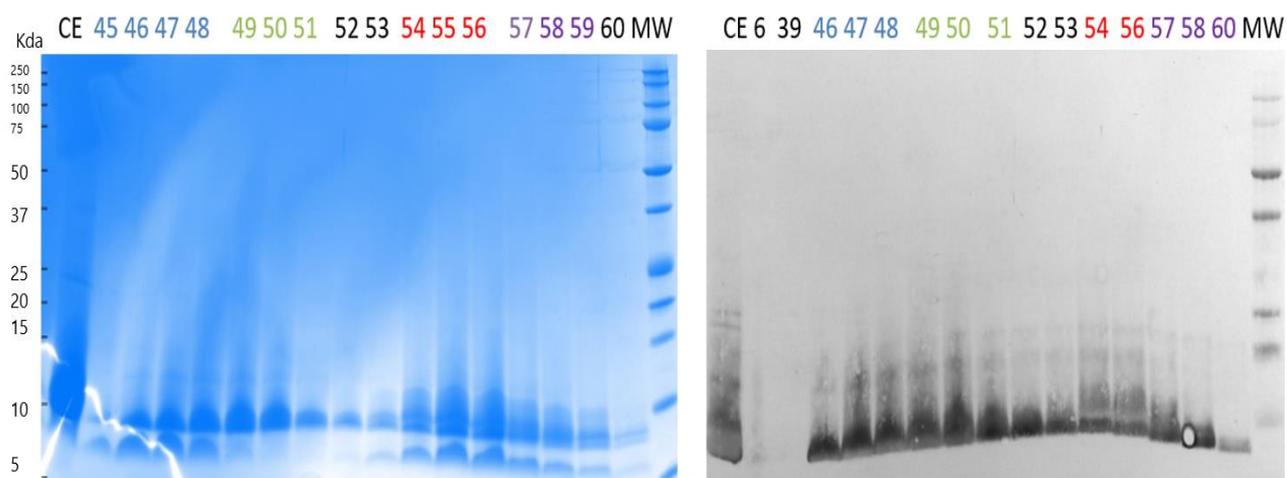
Figure 3.3: **a)** Cation exchange chromatography purification of the crude extract (CE) using a carboxymethyl cellulose (CM) sepharose column . The blue plot shows the absorption curve at 280 nm (in milli absorption unit (mAU)). The red curve shows the conductivity (in milliSiemens (mS)). **b)**Tris-Tricine gel analysis and western blot obtained for the collected fractions. The different pools of fractions are designated with different colors (peak A in blue, peak B in green, peak C in red and peak D in purple), corresponding to the peaks recorded for the CM chromatography. The western blot was performed with a polyclonal antibody against fraction C of the CM sepharose.

Following the UV chromatogram and the gel analysis, the fractions were combined into 4 pools. Pool A was composed of proteins migrating as a doublet on the gel between 5 and 10 kDa whereas proteins from pool B migrated as a single band about 10 kDa. Pool C eluted at a concentration of 0.6M NaCl, represented the main peak; 3 bands of proteins were visible on the gel. The last peak D corresponded to a small shoulder of peak C and showed the same pattern of migration as peak C. A western blot performed on the same fractions, was incubated with a polyclonal antibody (produced against fraction C), and shows the presence of a signal in all the different pools, suggesting a high level of similarity between the different proteins present in the CE.

Nevertheless, these CE batches were obtained from natural sources and their protein composition may depend in part on naturally variable growth conditions and extraction methods. Evidence of such variation is visible in Figure 3.4. This figure shows an overlay of the CM profiles obtained for the three different batches studied for this PhD thesis.
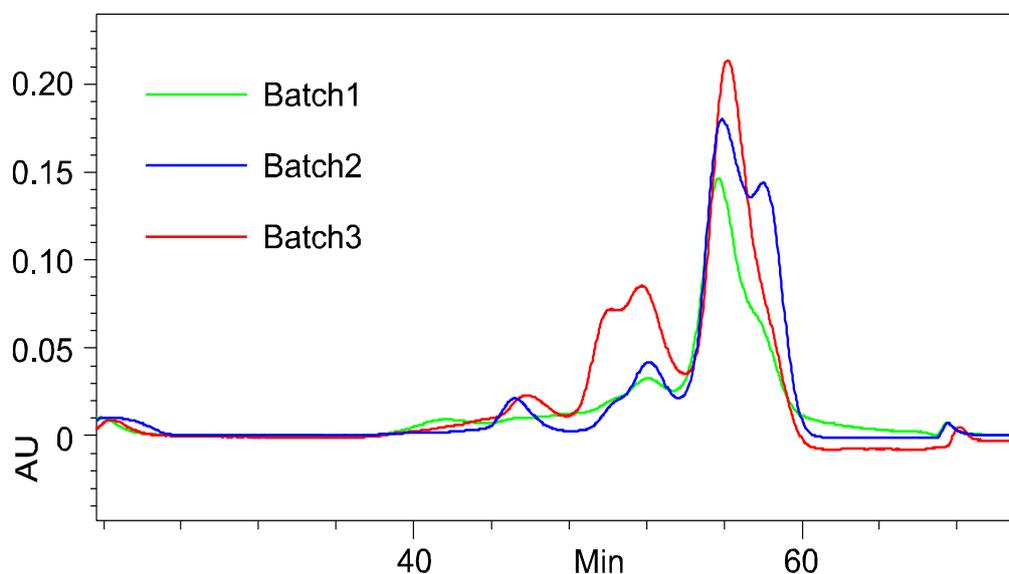
Figure 3.4: Overlays of cation exchange chromatography purification of three different crude extracts (CE) using carboxymethyl (CM) sepharose column. Batch 1 in green, batch 2 in blue and batch 3 in red.

The chromatogram recorded for batch 1 (in green) shows very small peaks following by a major peak having a small shoulder at the end, whereas the profile for batch 2 exhibits two main peaks (blue curve). Batch 3 (red) provided the largest overall quantity (≈100 mg) and has a profile similar to batch 1 and was suitable for a comprehensive investigation.

Size Exclusion Chromatography

Peak C, carrying the highest positive charge and corresponding to the main peak of the CM sepharose chromatography, was then purified using size exclusion chromatography (SEC). The specific fractions were concentrated using an Amicon ultracentrifugal unit (Millipore)(MWCO 3K) to a final volume of 1 ml, and then loaded on a Superdex 75 HR 10/ 30 column (GE Healthcare) which was equilibrated in distilled water running at 0.5 ml/min. The elution was followed using the same technique as mentioned above for CM sepharose step.

a)
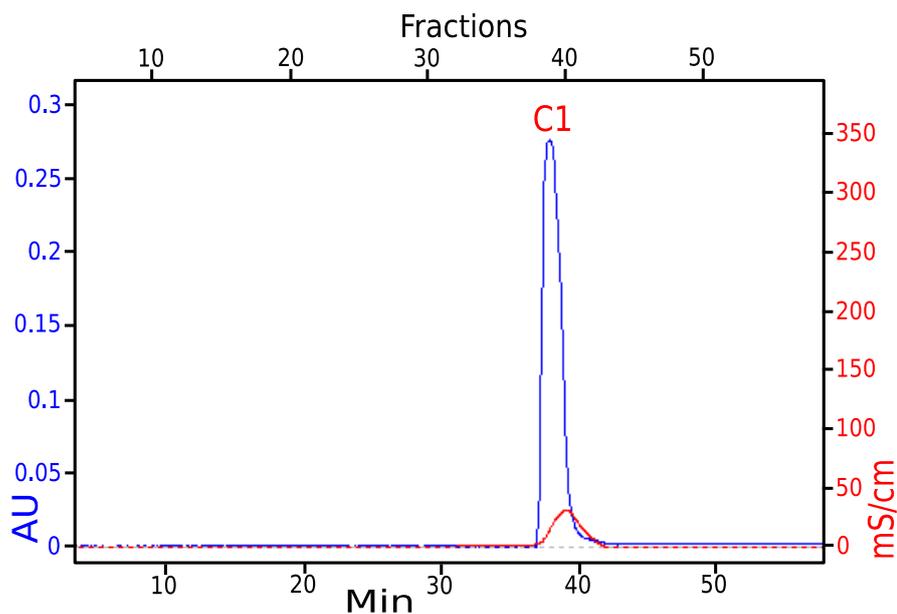
b)
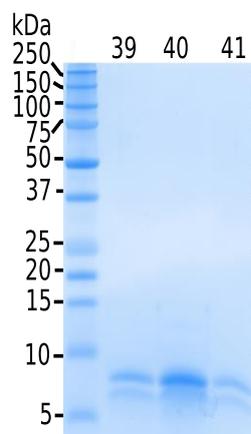
Figure 3.5: **a)** Purification of fraction C using size exclusion chromatography (SEC). The blue plot shows the absorption curve at 280 nm (in mAU). The red curve shows the conductivity (in mS). **b)** Tris-Tricine gradient gel analysis of the main peak.

Figure 3.5 shows the elution of 18 ml of fraction C from the SEC column. The eluted volumes (each of volume was 0.5ml) corresponding to the peak C1 were analysis on a Tris-Tricine gradient gel (Figure3.5b)), and shows the presence of a double band. This fraction C1 was considered highly pure for subsequent biochemical, biophysical analysis and activity assays.

The pooled fractions from peaks A, B and D were also concentrated and loaded on the same column. The different profiles were overlaid on the same chromatogram (Figure.3.6 a))

a)
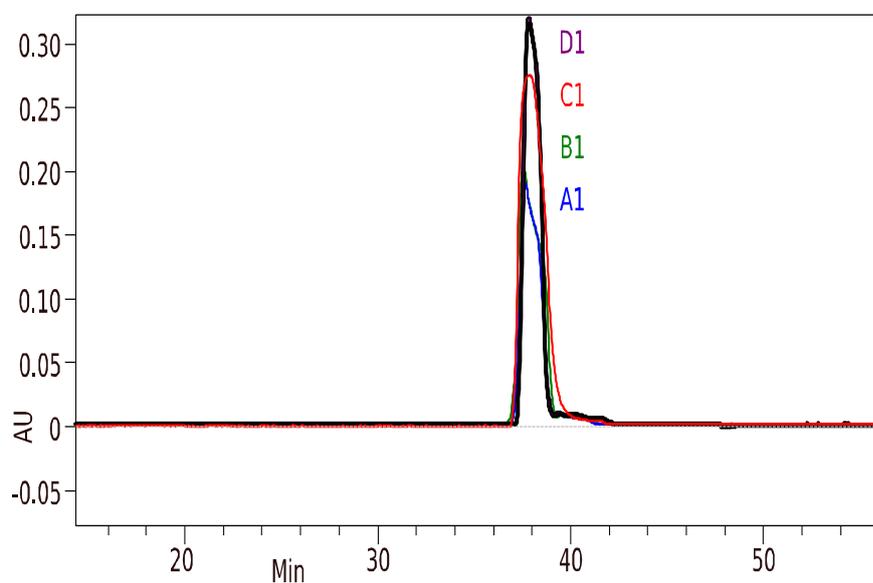
b)
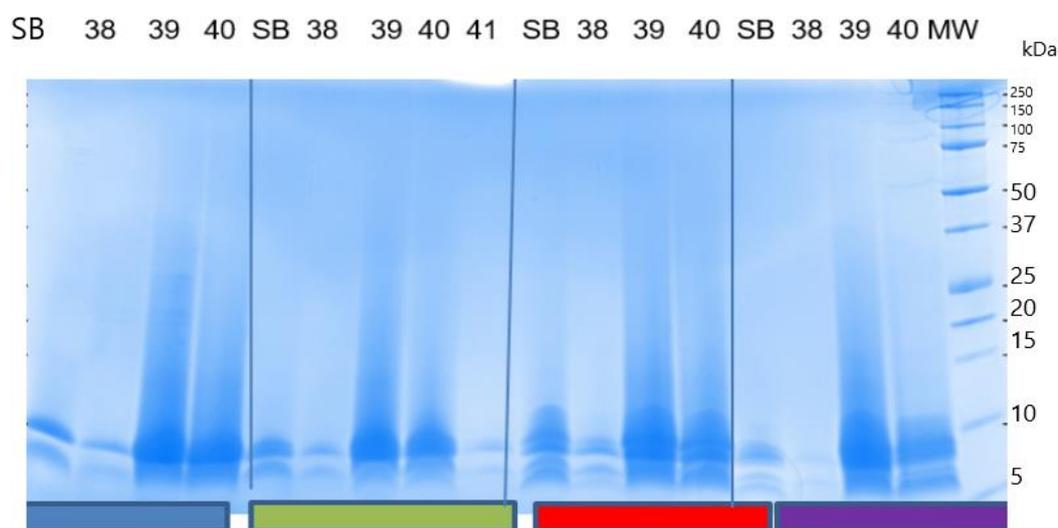
Figure 3.6: **a)** Overlay of size exclusion chromatography (SEC) results for fractions A (in blue), B (in green), C (in red) and D (in purple). **b)** Tris-tricine gradient gel analysis of fractions of each peak A1, B1, C1 and D1 obtained after SEC. (SB=sample before; MW=molecular weight)

The overlay of the different peaks resulting from the CM purification shows that all the pools are eluted in the same volume. However, peaks A1 and B1 have a shoulder whereas D1 and C1 correspond to sharp peaks suggesting a highly pure sample material.

Quantification of fractionation of the protein crude extract

The purification protocol was carried out on 20 mg of CE. After fractionation using two different columns, quantification was carried out using a colorimetric method, *the Bradford assay*, and the results are summarized in Table 3.1. 75% of proteins were recovered at the end of the purification procedure and shows that fraction C1 represents 33% of the extract.

Samples	Bradford quantification after SEC in mg	Percentage of each fraction
Fraction A1	2.75	21 %
Fraction B1	2.7	21%
Fraction C1	4.3	33%
Fraction D1	3.3	25 %
Total	13.05	100 %

Table 3.1: Summary of quantities (in mg) of different fractions obtained after two steps of purification.

The loss of 25 % of protein may correspond to that part of the CE that did not bind to the CM column (very small peak, data not shown) and to the non-specific binding of proteins on the membrane during the concentration steps.

This purified fraction C1 was used for all subsequent characterisation. However the crystal structure that was subsequently derived from the protein in this fraction revealed that the protein consisted of two chains linked by four disulphide bonds. This structural information led to the individual purification of both chains of the protein.

3-2-3 Separation and purification of the two chains of proteins component in fraction C1

The separation and the purification of both chains is reported in this subsection. To achieve the separation of both chains, which are linked by disulphide bridges (see Chapter 4), it was necessary to denature the protein, strongly reduce its disulphide bonds and block them in this configuration using an alkylation step. Once this separation was performed, the alkylated chains were purified by reverse phase chromatography (RPC) and the fractions analysed by gel electrophoresis.

Separation of chains by denaturation, reduction and carboxymethylation steps

For the denaturation and reduction step, 1mg of fraction C1 (in denaturing buffer - 6M guanidinium hydrochloride in 0.5 M of Tris pH8.0) was incubated in the presence of 20mM tris (2-carboxyethyl) phosphine (TCEP) for 1 hour at 55°C. Immediately before use, 9.3mg of iodoacetamide was dissolved in 132 µl of 200 mM ammonium bicarbonate (pH 8.0) to make 375 mM iodoacetamide. This solution was protected from light. 40 µl of freshly prepared 375 mM iodoacetamide was added to the denatured and reduced sample and incubated for 30 minutes (again protected from light). The alkylated sample obtained was then purified by RPC.

Purification on reverse phase chromatography

The C18 column volume (CV=3ml) was pre-equilibrated in 0.1% trifluoroacetic acid (TFA) (equilibration buffer) at 1ml/min. 200 µl of protein with alkylated cysteine was loaded on the column and after a wash of 3 CV with equilibration buffer, chain separation was achieved on a gradient of 0 to 70 % with 0.1% TFA in 100% acetonitrile for 8 CV (Figure 3.7 a)). The 0.25 ml fractions corresponding to the peak observed were collected and analysed on a 4-20% Tris-Tricine gradient gel (Figure3.7b))

a)
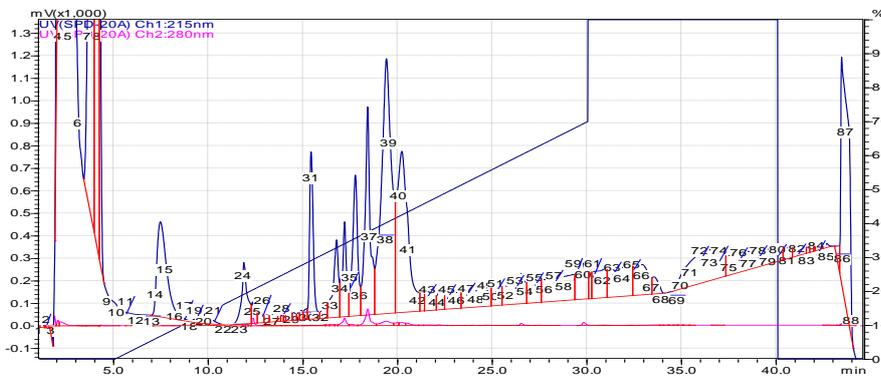
b)
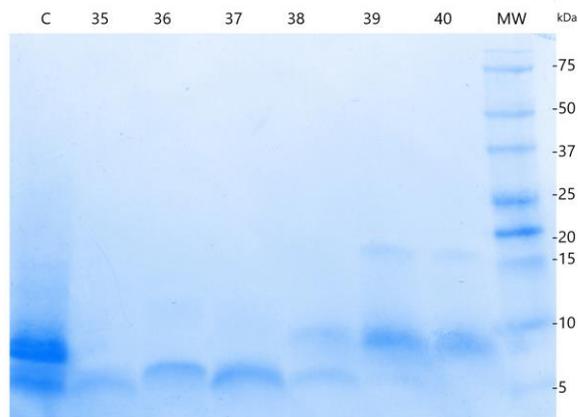
Figure 3.7: **a)** Purification of the alkylated chains of fraction C1 using the C18 column. The purification was monitored by UV spectroscopy at 280 nm (pink) and 215 nm (blue). The absorption at 280 nm is due to the presence of aromatic amino acids (tryptophan, tyrosine and phenylalanine) in the amino acid sequence and the absorption at 215 nm corresponds to the peptide bond. **b)** Gel analysis of fractions corresponding to the different peaks obtained. (C= control (input sample)).

The chromatogram and the gel analysis show a good separation of both chains on the C18 column. The small chain (less hydrophobic) is observed over three peaks, and migrates with a molecular weight close to 5000 Da. The long chain is eluted in two peaks and migrates around 8000 Da on the gel. It is noted that the small peptide eluted from two different peaks (fractions 35 and 36) have different patterns of migration. These RPC fractions and fraction C1 were analysed further by a combination of techniques including N-terminal sequencing and mass spectrometry (MS). The results are reported in the section 3.3 and 3.4. A summary of the main steps used in the purification of fraction C1 and the RPC purification is described in Figure 3.8.
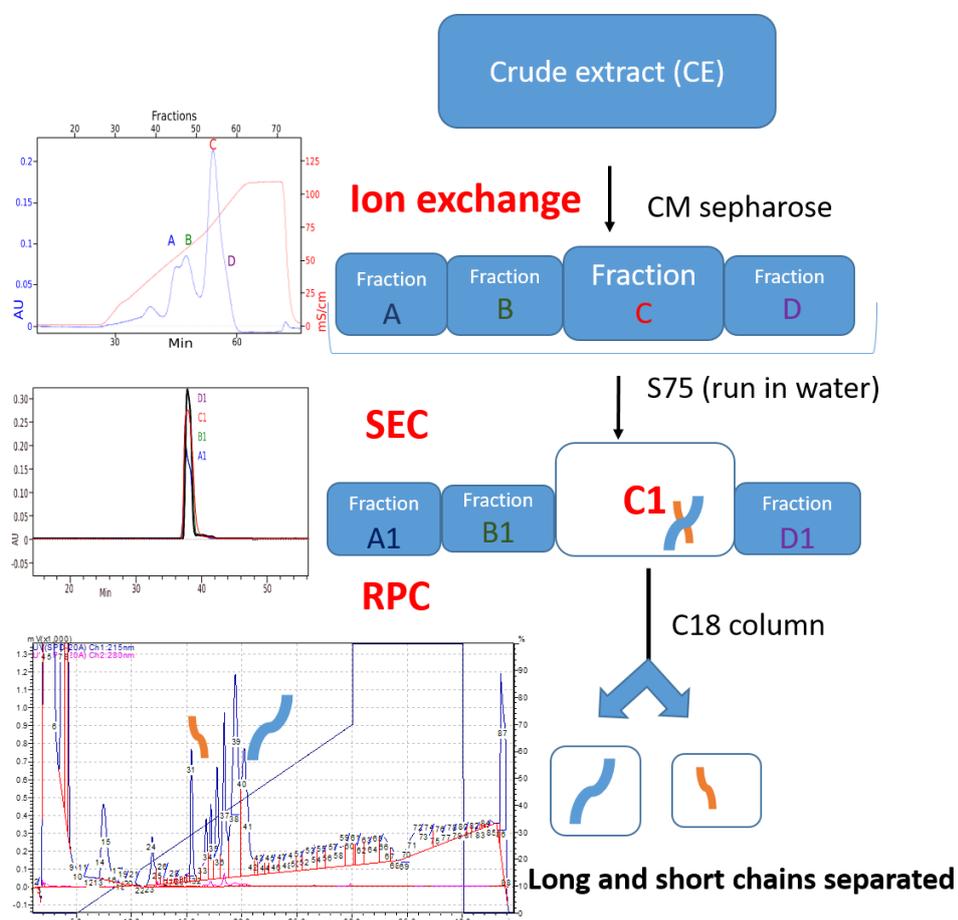
Figure 3.8: Summary of the main steps of the purification of proteins from the crude extract (CE). The fraction C1 is obtained after 2 steps of purification comprising an ion exchange and a size exclusion chromatography step (SEC). Reverse phase chromatography (RPC) fractions are obtained after purification on C18 column of the fraction C1.

3-3 Biochemical characterisation of *Moringa* proteins

This section describes the biochemical characterisation performed on the CE and on the purified fraction C1. Various techniques were applied that yielded information on the amino acid composition, the molar absorption coefficient ε (also called epsilon coefficient) which allows protein quantification, the amino acid sequence, the glycosylation state, and the thermostability of the component of this fraction of interest.

3-3-1 Amino acid composition and epsilon determination of *Moringa* proteins

The amino acid composition analysis was performed respectively on the CE (Figure 3.9) and on the fraction C1 (Figure 3.10), obtained from the three different batches studied. The conditions of the experiment described in the Material and Methods chapter (Chapter 2) describe conventional acidic hydrolysis. These rather drastic conditions have an effect on the stability of some amino acids and therefore on the analysis of these results. Asparagine and glutamine are completely hydrolysed to aspartic acid and glutamic acid respectively. The tryptophan is fully destroyed and cysteine cannot be directly determined from the acid-hydrolyzed samples. Tyrosine is also partially destroyed by traces of impurities present in the agent and serine and threonine are partially hydrolysed as well - losses of about 10 and 5 % respectively usually occur.
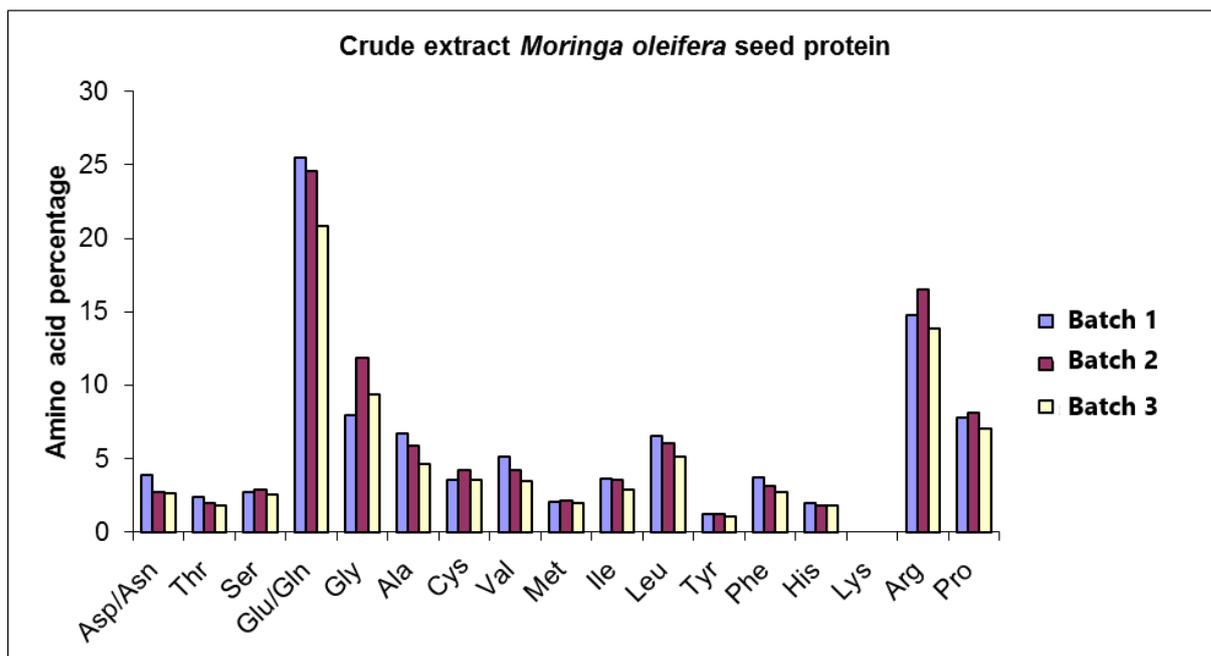
Figure 3.9: Bar graph representing the amino acid composition of the crude extract (CE) of the three different batches studied.

Figure 3.9 shows that all of the CE fractions contain a majority of glutamine-asparagine (about 25%). More than 15 % is arginine regardless of the different batches analysed. The absence of lysine in the amino acid composition is notable. In addition, the bar graph shows the heterogeneity of the different batches - which was previously highligthted by chromatography (Section 3-2-2, Figure 3.4).
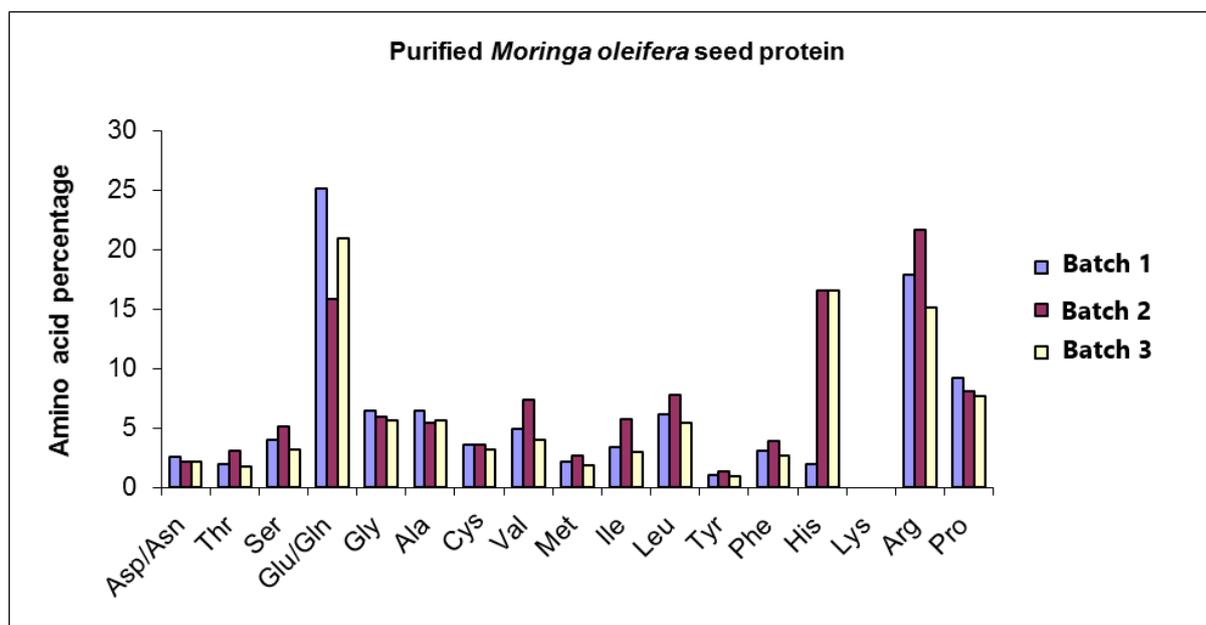
Figure 3.10 : Bar graph representing the amino acid composition of the fraction C1 purified from the three different batches studied.

By comparison with the analysis of the CE, the amino acid composition (Figure 3.10) obtained for the three purified C1 fractions showed the same dominance of Glu/Gln (≈25%) and arginine. The high percentage of histidine in batches 2 and 3 resulted from sample contamination. However, the amino acid composition differs slightly between each purified protein, showing a heterogeneity in the sample preparation from natural sources. Overall, the composition of the fraction reflected the composition of the CE, tending to confirm that the purified fraction was one of the main components of the CE. This amino acid analysis, combined with the dry weight data, enabled the determination of the molar absorption coefficient (ε), which is used in the Beer-Lambert law (section 2-1-4, Chapter 2). The determination of this coefficient allowed the quantification of protein concentration simply by measuring the absorption at 280 nm. The results are summarized in Table 3.2

	Fraction C1 Diluted by a factor 8
OD 280 nm	0.158
Nanomoles/g	29.6
MW (Da)	11800
Extinction coefficient (M^{-1} cm^{-1})	5448
Epsilon 0.1% (=1 g/l)	0.46

Table 3.2: Quantification of the protein of fraction C1 by amino acid analysis and determination of the molar extinction coefficients in M^{-1} cm^{-1}, at 280 nm or Epsilon for 0.1% solutions (=1 g/l) (Epsilon0.1% = Extinction coefficient/ Molecular weight (MW)).

The extinction coefficient obtained was 5448 M^{-1} cm^{-1} and this value was used for the calculation of the protein concentration throughout the thesis experimental work. The coefficient is very low, indicative of a low content of aromatic residues in the protein. The N-terminal sequencing described in the next section was the most suitable biochemical approach to gather further information regarding the amino acid sequence of this fraction.

3-3-2 Determination of amino acid sequence of protein of fraction C1 by N-terminal sequencing

Sequence analysis was performed on fraction C1, which was blotted on polyvinylidene fluoride (PVDF) membrane following a Tris-Tricine gel. A specific transfer protocol (described below) was used to avoid the presence of glycine in the transfer buffer that can interfere with data analysis.

Samples were boiled and loaded on a Tris-Tricine gel. When electrophoresis was complete, the gel was equilibrated in water for 1 min and 10 minutes in transfer buffer (10 mM CAPS (3-

[cyclohexamino]-1-propanesulfonic acid) pH 11). In parallel, the PVDF membrane was dipped in methanol followed by an equilibration in Transfer buffer for 20 minutes. The chamber was filled with the same buffer and a transblotting sandwich using Whatman filter paper. A Scotch-Brite pad was carefully assembled (see Chapter 2) to avoid bubbles between membrane and gel. The settings used for the transfer were 90 Volt for 1 h00 with cooling at 4°C. Once the transfer was complete, the PVDF membrane was removed, rinsed with water and saturated with methanol for a few seconds.

The detection was performed by staining the membrane with 0.1% Coomassie Blue R-250 in 40% methanol/1% acetic acid for 1 minute followed by a destaining with 50% methanol (several changes). The membrane was extensively rinsed with water and dried. The bands of interest were excised using a razor blade for protein/peptide sequencing. The N-Terminal amino acid sequence of fraction C1 protein was analysed on Applied Biosystems gas-phase sequencer model 492 (s/n: 9510287J) performing Edman degradation. Phenylthiohydantoin amino acid derivatives generated at each sequence cycle are identified on-line with an HPLC system using the data analysis system for protein sequencing from Applied Biosystems Model 610A (software version 2.1). The sequence obtained was submitted to automatic alignment software, performed using the NCBI_BLAST search system. A sequencing of 5 to 8 residues is usually sufficient to identify a protein. A lot of effort was put in the identification of the protein(s) of fraction C1 using this technique. The first sample analysed gave a weak signal and several sequences were obtained (Table 3.3 row2). The weak signal could be explained by the presence of pyroglutamate on the N-terminus of the protein that could block the Edman degradation process (as reported by Gassenschmidt *et al*, 1995), hence, a pre-digestion step using pyroglutamate aminopeptidase was systematically added. Moreover, the existence of

two chains in the fraction could complicate the analysis of results. Therefore, two approaches were developed: one consisted in sequencing proteolytic fragments of Fraction C1 and the second one, to sequence both chains separately. Both strategies are described in the following section. At this stage of the study, the possibility that these 2 chains originated from different proteins was not excluded.

Digestion by pyroglutamate aminopeptidase

10 milliunits (mU) of *Pyrococcus furiosus (Pfu)* pyroglutamate aminopeptidase was reconstituted in 50 µl of 50 mM sodium phosphate buffer (pH7.0) containing 10 mM DTT and 1 mM EDTA. The N-terminal pyroglutamyl residues was released from the protein following an overnight incubation in 50 mM sodium phosphate buffer (pH7.0) containing 10 mM DTT at 65 °C for a molar substrate-to-enzyme ratio of 20 to 1 without previous denaturation of the fraction C1. The sample pre-digested with *Pfu* Pyroglutamate was sequenced following the same procedure.

N-terminal sequencing on limited proteolysis fragment

Prior to limited proteolysis digestion, the fraction C1 was extensively dialyzed against digestion buffer (50 mM tris pH 8.0, 50 mM NaCl) in the presence of 2, 4 or 8 M urea, and concentrated to 1mg/ml. Digestions were performed on 30 µl quantities, corresponding to about 20 µg of protein sample. The reaction was started by the addition of 10 µl of a specific protease solution, freshly dissolved in digestion buffer. Proteolytic enzymes of sequencing grade quality (such as subtilisin, trypsin and chymotrypsin) were purchased from Sigma, dissolved in stock solutions at 5 mg/ml in 1mM chloric acid, and stored at -20°C. Digestion reactions were stopped by addition of 1 µl PMSF 10mM and 10 µl of protein sample buffer. Samples were boiled for 5 min and finally analysed on Tris-Tricine gel. The first digestion trials

were performed with trypsin, subtilisin and chymotrypsin at 30°C with an incubation time of 2 hours and a protease: protein ratio between 1:2 and 1:10. The protein was resistant to different proteases as described in Figure 3.11a) where no changes were visible between control (c) and trypsin cleavage (T). Digestion reactions were repeated overnight at 37°C with a trypsin: protein ratios of 1:20 to 1:50 and the results are shown in Figure 3.11(b).

a)   b)
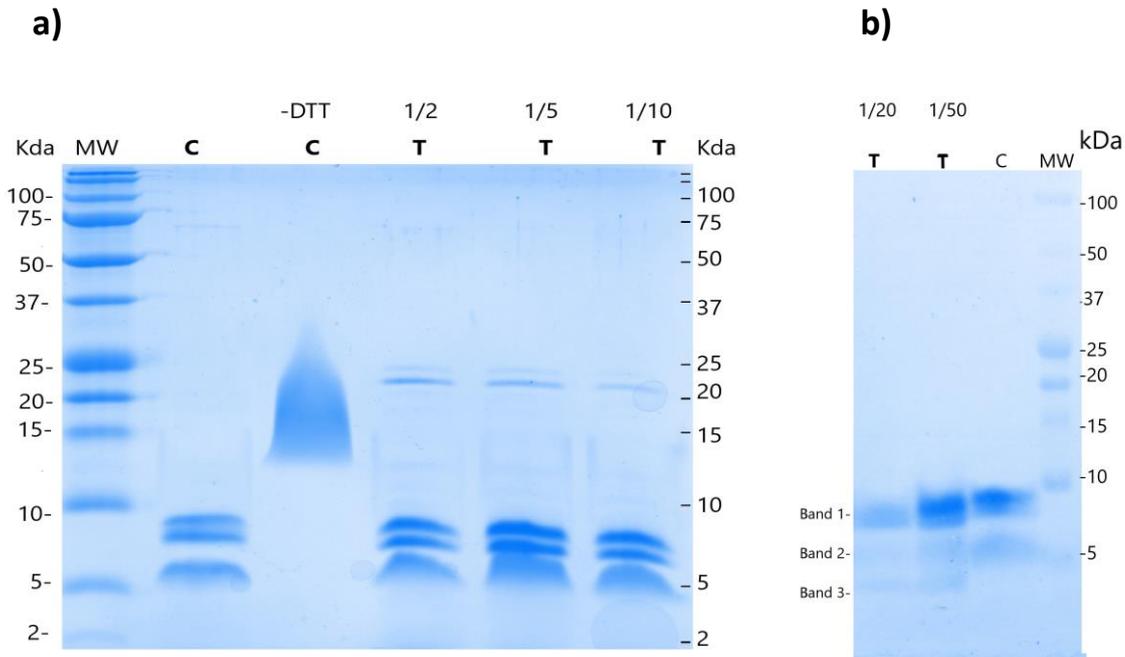
Figure 3.11 : Example of Tris-Tricine gradient gel of proteolysis digestion of fraction C1 a) and PVDF membrane of proteolytic fragment obtained b). In a) results of **trypsin (T)** digestion incubated with trypsin for 2 hours at 30 deg are compared with the control **(C)** (lane 2). b) PVDF membrane showing overnight digestion at 37 °C in presence of **trypsin**.

Figure 3.11 b) shows the presence of 3 weak bands obtained after an overnight digestion at a trypsin : protein ratio at 1:20. These bands were cut and analysed by N-terminal sequencing and the results are summarized in Table 3.3.

The first N-terminal sequencing results of the fraction C1 obtained at the beginning of this project were not interpretable mainly due to the lack of sequence data in the database of *Moringa oleifera*, but also due to the poor quality of the sequence obtained. The deposition

of the *Mo*-CBP3 isoform sequence in the proteomic data bank in 2015 improved the situation and allowed the identification of these sequences.

Number of amino acid	Uncleaved protein	Limited proteolysis using trypsin		
	Purified fraction	Band 1	Band 2	Band 3
1	R	R/H	S/G/E/D	Q/T/S/G
2	P	P	P	T/P
3	A	A	A	Q
4	I/T	I/T	I	Q
5	Q/L	Q/L	L	L
6	R	Q/R	C	L
7		R	Y	A
8			Q	A
9	Q		Q	Q
10	Q		F/L	Q
11	L		L	F
12			T	V/I
13	N		T	I
14				R
15				Q
16				T
17				S
18				Q
19				G
20				G
Sequence Identification on http://www.uniprot.org	**Long chain of *Mo*-CBP3-3** RRPAIQRCCQQLRNIQPRCRC **Long chain of *Mo*-CBP3-4** ARRPPTLQRCCRQLRNVSPFCRCP	**Long chain of *Mo*-CBP3-3** RRPAIQRCCQQLRNIQPRCRC **Long chain of *Mo*-CBP3-4** ARRPPTLQRCCRQLRNVSPFCRCP	Not identified as moringa protein	

Table 3.3 : Sequence identification by N-terminal sequencing of the native protein (row2) and peptides (3rd row: bands 1,2,3) obtained after limited proteolysis.

Table 3.3 (2nd row) indicates the presence of two sequences of the long chain of *Mo*-CBP3 isoforms : *Mo*-CBP3-3 and *Mo*-CBP3-4 are present in the fraction C1 corresponding to the uncleaved protein. It is notable that no short chain sequences were observed. The analysis of tryptic fragments (3rd row) confirms the presence of the same sequences as identified in the uncleaved fraction. Unfortunately, the tryptic digestion approach did not provide any additional information regarding the identification of the components of fraction C1.

N-terminal sequencing on separated chains using reverse phase chromatography

Another strategy for the identification of the amino acid sequence was the study of each chain individually after reduction, alkylation and reverse phase steps. The purification has already been described in section 3.2. The same fractions as those noted in Figure 3.7 were loaded and transferred on a PVDF membrane for N-terminal sequencing (Figure 3.12).
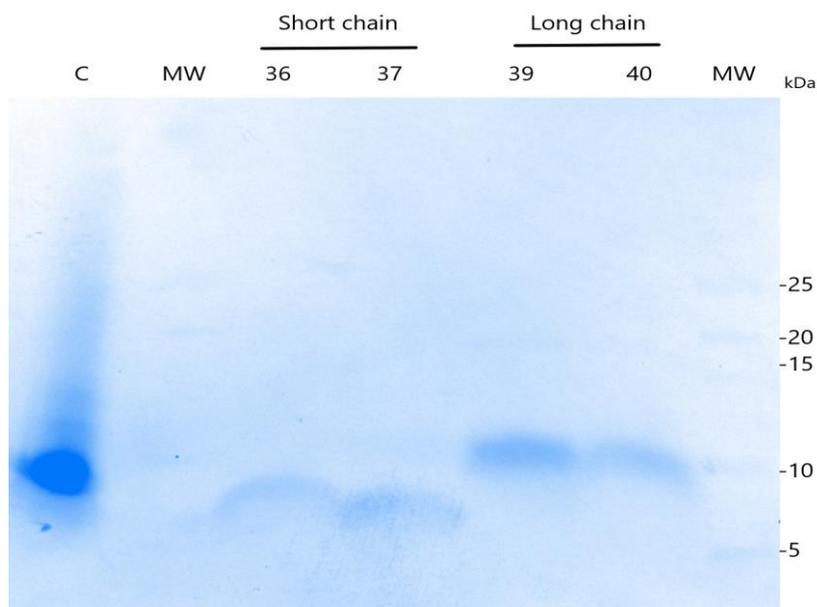
Figure 3.12: PVDF membrane obtained after transblotting of fractions of Reverse Phase chromatography (RPC). (C=control, MW=molecular weight standard).

Fraction 35 (not see on the PVDF membrane) was also analysed. Table 3.4 summarizes the results obtained for fractions 36-37 (short chain) and fractions 39-40 (long chain). Surprisingly, sequences of long chain were identified in the analysis of the fractions 35 to 37, where only short chains were expected. These sequences of long chains were the same as those identified from the native and tryptic samples (Table 3.3). However, five amino acids (in red in the table) of *Mo*-CBP3-3 short chain were detected in fraction 35 and 36. For fractions 39 and 40, the same amino acids of the same isoforms were identified as the long chain of *Mo*-CBP3-3 and *Mo*-CBP3-4, (green and pink color, respectively).

Number of amino acid	Short chain analysis				Long chain analysis	
	Fraction 35	Fraction 36	Fraction 37		Fraction 39	Fraction 40
1	Q/A/T/E/G/V	Q/T/G/A/V	V/A/Q/T/E/G			D
2	P/Q	Q/P/P	P/Q/L/P		P/P	P/P
3	A/Q	Q/A/P	Q/A		A	A/I
4	I/Q	Q/I	Q/T/L/I		T/I	I/T
5	Q/L	Q/L/L	L/Q/L		L/Q ou L	L/Q ou L
6	R	Q/Q	Q/R/Q		Q/R/Q	Q/R
7	R	Q/R	R/Q		R/Q	Q/R
8	Q/R	Q/R/Q	Q/R		Q/R	
9	Q/Q	Q/R/Q/Q	Q/Q		Q/L	Q/L/Q
10	F	Q/R/F/Q	Q/R/Q		Q/R	Q/R
11	Q/L/R/L	Q/L/R	Q/L/L		Q/L	Q/L
12	Q/R	R/L	R/L/V/I		L/R	R/L
13	N/R	R	R/Q/N		R/N	N/R
14	R/Q/I	R/N	R/N/Q/I		N/I	I/N/R/Q
15	Q/R	Q/R/A/V	V		V/Q	V/Q
16	R					
17	R/Q/I					
18	R/Q/A					
19						
20	R/Q					
Sequence identification	**Short chain of *Mo*-CBP3-3** QQQQCRQQFLTHQRLRACQRFIRR **Long chain of *Mo*-CBP3-3** RRPAIQRCCQQLRNIQPRCRC **Long chain of *Mo*-CBP3-4** ARRPPTLQRCCRQLRNVSPFCRCPS				**Long chain of *Mo*-CBP3-3** RRPAIQRCCQQLRNIQPRCRC **Long chain of *Mo*-CBP3-4** ARRPPTLQRCCRQLRNVSPFCRCP	

Table 3.4: Sequence identification of fractions (35 to 37 and 39-40) obtained after reverse phase chromatography (RPC).

The N-terminal sequencing results on fraction C1 and its tryptic fragments did not allow the determination of an unambiguous amino acid sequence required to complete the crystallographic analysis. However, in 2015,(Freire *et al.*), the deposition of new sequences of *Mo*-CBP3 in the data bank rendered the interpretation of these results possible and showed the presence of two isoforms of *Mo*-CBP3, corresponding to *Mo*-CBP3-3 and *Mo*-CBP3-4 in fraction C1. This study was combined with parallel mass spectrometry (MS) experiments. These results are described in section 3-4 below.

3-3-3 Glycosylation state and thermostability study of *Mo*-CBP3 proteins

The biochemical characterisation was complemented by a study on the glycosylation state of the different fractions A1 to D1 (obtained after SEC) and on the CE. Different amount of these samples were loaded on a Tris-tricine gel and the presence of carbohydrate was evaluated by specific staining after electrophoresis using the Glycoprotein Detection Kit (Sigma). The detection limit was found to be in the range of 25-100 ng for carbohydrates, depending on the nature and the degree of glycosylation of the protein. Horseradish peroxidase with a carbohydrate content of 16% was provided as a positive control in the kit used.
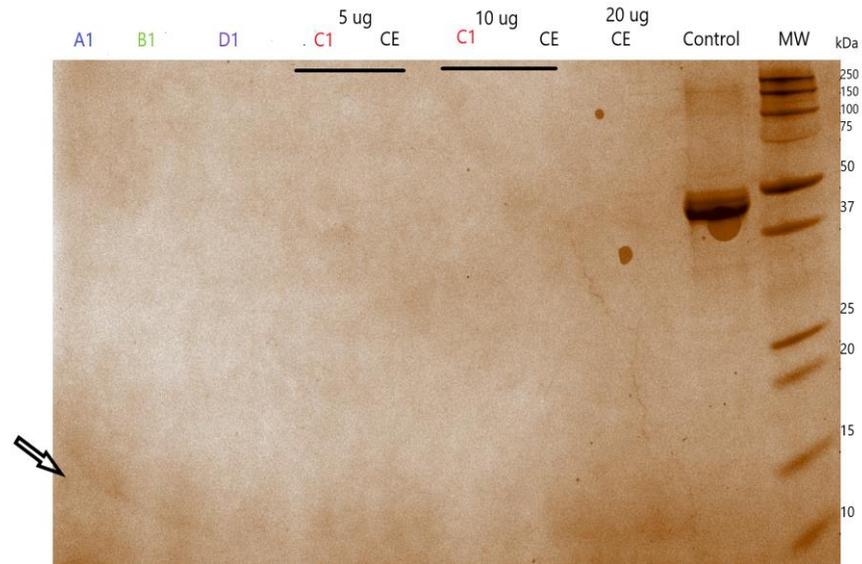
Figure 3.13: Tris-Tricine gel obtained after staining based on a modification of Periodic Acid-Schiff (PAS) method, yielding magenta bands with a light pink or colourless background as observed for the protein control (Horseradish peroxidase).The arrow shows a very weak signal visible for fraction A1 and B1 that might be background or a low level of carbohydrate.

The analysis of the gel (Figure 3.13) seems to indicate that the different purified fractions (A1, B1, C1 and D1) of the CE and the whole CE are not highly glycosylated

The last experiment carried out in the biochemical characterisation section was the study of the thermostability of *Mo*-CBP3 proteins contained in the fraction C1. These proteins were dialysed against different molarities of urea from 2 to 8 in presence or absence of DTT, and incubated at 95 degrees overnight. Samples were loaded and analysed on a Tris-Tricine gel (Figure 3.14)
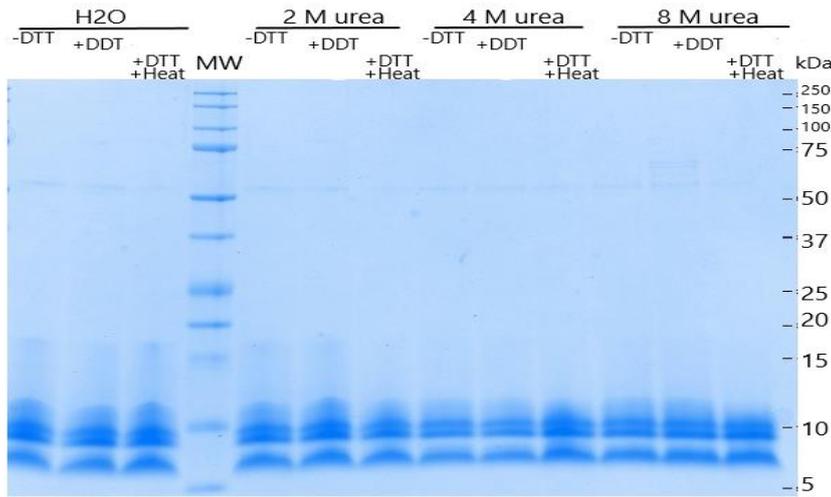
Figure 3.14: Tris-Tricine gel obtained after heat treatment of fraction C1 in presence or absence of different molarities of urea.

This gel shows the same pattern of migration for samples that were either heat-treated or not in the presence or absence of DTT. These results reveal that *Mo*-CBP3 proteins are thermostable, as confirmed in the next section by circular dichroism (CD) measurements. The different techniques used in this biochemical characterisation of the fraction C1 are summarized in Figure 3.15.
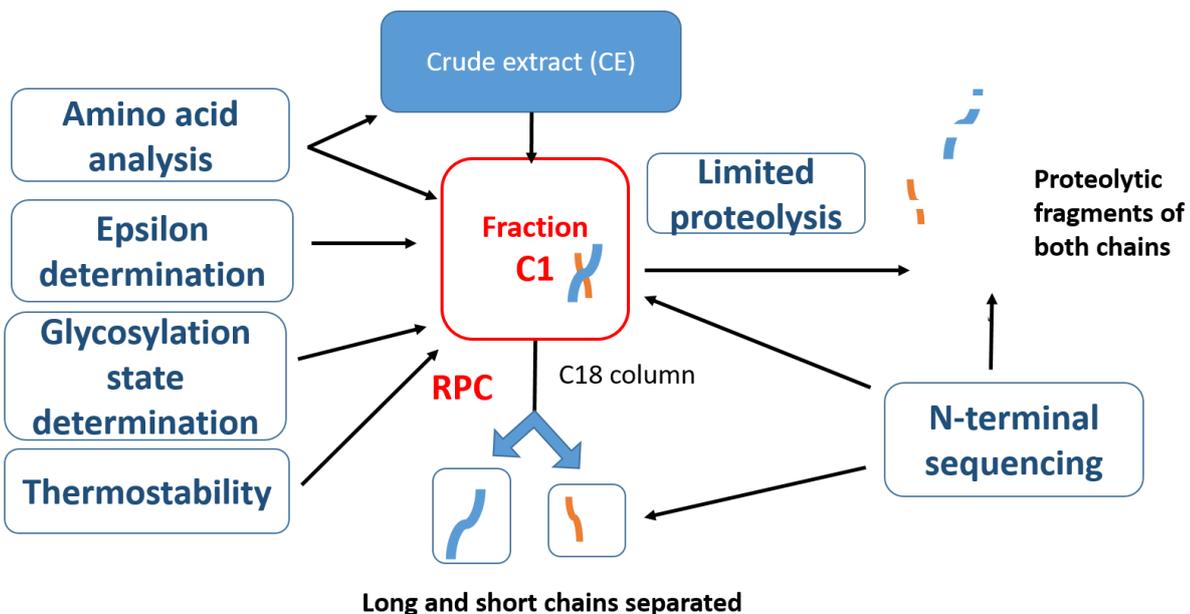
Figure 3.15: Scheme showing the various biochemical techniques carried out on the crude extract (CE), on fraction C1, and also on its chains separated and on its proteolytic fragments.

3-4 Biophysical characterisation of *Moringa* proteins

Biophysical characterisation techniques were mainly applied to determine the mass of the protein and to obtain the amino acid sequence which was incomplete with the N-terminal sequencing approaches described previously. This study was complemented by circular dichroism (CD) measurements that were used as a way of assessing changes in structure and stability of the *Mo*-CBP3 proteins by comparison with the CE.

3-4-1 Circular dichroism measurements

CD measurements were carried out by recording a spectrum in the far UV (from 180 to 250 nm) at a speed of 50 nm/min. For each protein sample, a blank sample which contains only distilled water was also measured – later to be subtracted from the final sample measurement. For this purpose, the crude extract was weighed and solubilised to a final concentration at 0.02 mg/ml in distilled water; fraction C1 was diluted to the same concentration. When recording CD spectra, a 0.1 cm thick quartz cuvette containing 1 ml of protein solution was placed in the sample holder of a Jasco-810 spectrometer. Melting experiments were also performed on the same samples. The CD signal was recorded at the minimum of the spectrum at 224 nm while the temperature was ramped from 10 to 95 °C at a rate of 2°C/min; the 222 nm minimum was used because absorption was low at this wavelength allowing a better signal-to-noise ratio to be obtained in the measurement of the circular polarisation of the sample.
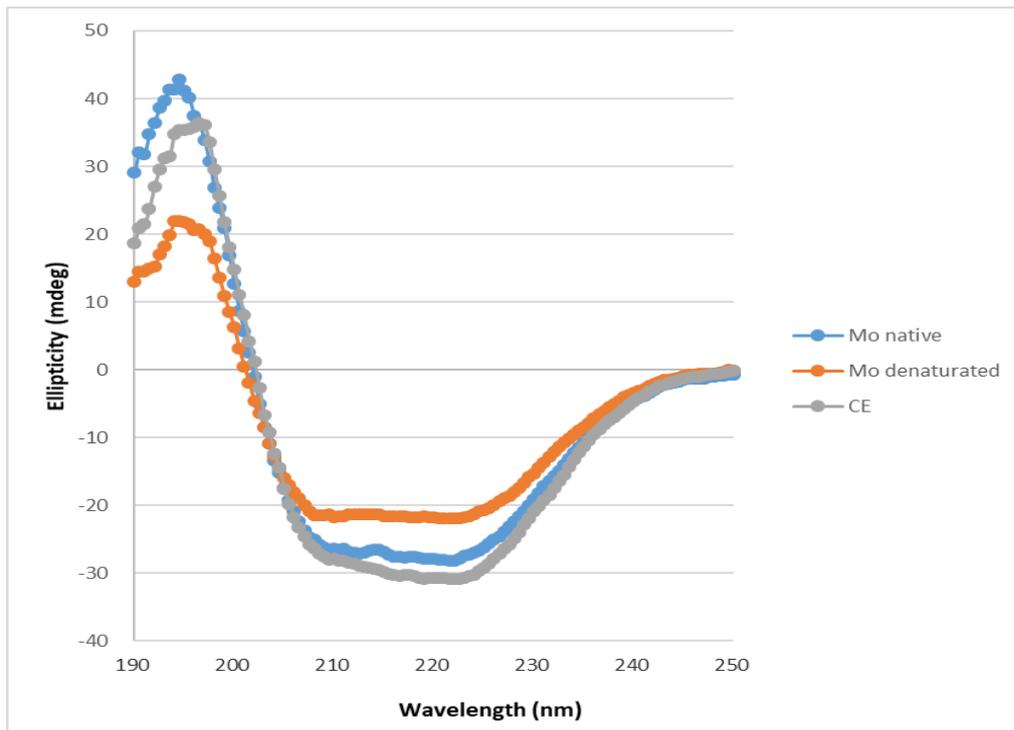
Figure 3.16: Circular Dichroism (CD) spectrum of *Mo*-CBP3 in native conditions (blue), *Mo*CBP3 after a step of denaturation from 10 to 95 degrees (orange).The grey plot shows the spectrum recorded from the crude extract (CE).

The CD measurements (Figure 3.16) on *Mo*-CBP3 show a flat trough between 230 and 210 nm, with a minimum at around 220 nm. This curve is interpretable as typical of alpha (α) helical conformations because α helices have negative bands at 222 nm and 208 and a positive one at 190 nm. The CE and *Mo*-CBP3 native curves are very similar, suggesting a similar structural composition. The curve corresponding to denatured *Mo*-CBP3 shows a discrete alteration of α-helical secondary structure following heat treatment from 10 °C to 95 °C, suggesting a very high protein thermostability.

3-4-2 Mass spectrometry (MS) studies

This section reports the different assays that were performed to characterise and identify the proteins present in the purified fractions and as well as in the CE, combining the MS and tandem MS (also called MS/MS) techniques. MS/MS involves multiple steps of mass analysers

that increase specifity and enhance the detection of targeted protein. The MS experiments were carried out by Luca Signor at Institut de Biologie Structurale (IBS), Grenoble. The different fractions (A1, B1, C1 and D1) obtained after CM/SEC fractionation were first characterized and then the purified components from C1 material purified from different batches were compared. Due to the ambiguity of the preliminary results, the same study was carried out on each of the chains of fraction C1, previously separated on RPC, and the measurements repeated after an "in gel digestion" step. The complexity of the results obtained made the interpretation challenging. Fortunately, the possibility of separating different isoforms on a SEC column facilitated the analysis and the combination with tandem MS was very helpful. Tandem MS (or MS/MS) was performed by Sylvie Kieffer-Jaquinod at Commissariat à l'Energie Atomique et aux Energies alternatives (CEA) at Grenoble and allowed the identification of the different isoforms of *Mo*-CBP3, their modification states (pyroglutamate positions and disulphide bridges) and their abundance.

MS on protein fractionated of the crude extract

The CM purification was resolved into 4 fractions according to the peak distribution (Figure 3.3). These four fractions were further purified on a S75 SEC column and analysed by electrospray ionisation time of flight (ESI-TOF) mass spectrometry. Each sample was diluted 1:2 or 1:3 with ACN 50%, H_2O 5%, FA 0.1% before analysis and were directly infused with a syringe pump at a flow rate of 10 µl/min. The value for the fragmentor was set at 300 V. Table 3.5 summarizes the identification of the main protein species observed.

samples	Fraction C1	Fraction B1	Fraction A1	Fraction D1
Observed mass * (Da)	Main specie: 11785 Other minor specie: 12084 11956 11883 11688	Several species observed: 11447 11758 11785 11899 11956 12029 12084 12126	Main specie: 11894 Other minor specie: 11740 11797	Main specie: 11785 11447 Other minor specie: 11956 11688 11350

Table 3.5: Identification of the main protein species observed in different fractions from batch 1 by electrospray ionisation time of flight (ESI-TOF) mass spectrometry (MS) analysis. In red the main peaks detected, in black minor peaks.

* Mass error tolerance (acceptability): protein ESI TOF MS: 10-50 ppm depending on protein size; peptide ESI TOF MS: 5-10 ppm depending on peptide size

Fraction A1 and C1 contain one main species whereas fraction D1 has two and fraction B1 none. The masses of the main species ranged from 11447 to 11894 Da and a multitude of masses very close to the main one were identified. These results show that each fraction does not contain one single protein but a mixture of proteins having very similar masses. The difference in the observed masses may correspond to one or two amino acids or to a post-translational modification such as the addition of a pyroglutamate. The table confirms that the protein components isolated from the CE with CM and SEC purification are mainly composed of small proteins having masses from 11350 to 12126 Da.

MS on purified fraction C1 from different batches

The purified fraction C1 (containing isoforms of *Mo*-CBP3) was the fraction chosen and studied for all the biochemical and biophysical characterisation, crystallisation and reflectometry measurements in this work. In section 3-2, the overlay of CM chromatograms for batches 1, 2

and 3 exhibited different profiles (Figure 3.4), suggesting a potential heterogeneity in the

fraction C1 content. The electrospray ionisation time of flight mass spectrometry (ESI-TOF MS)

analysis allowed this specific fraction to be characterised after purification. Each sample was

diluted 1:2 or 1:3 with ACN 50%, H_2O 5%, FA 0.1% before analysis and were directly infused

with a syringe pump at flow rate of 10ul/min. The fragmentor voltage was set at 300 V.
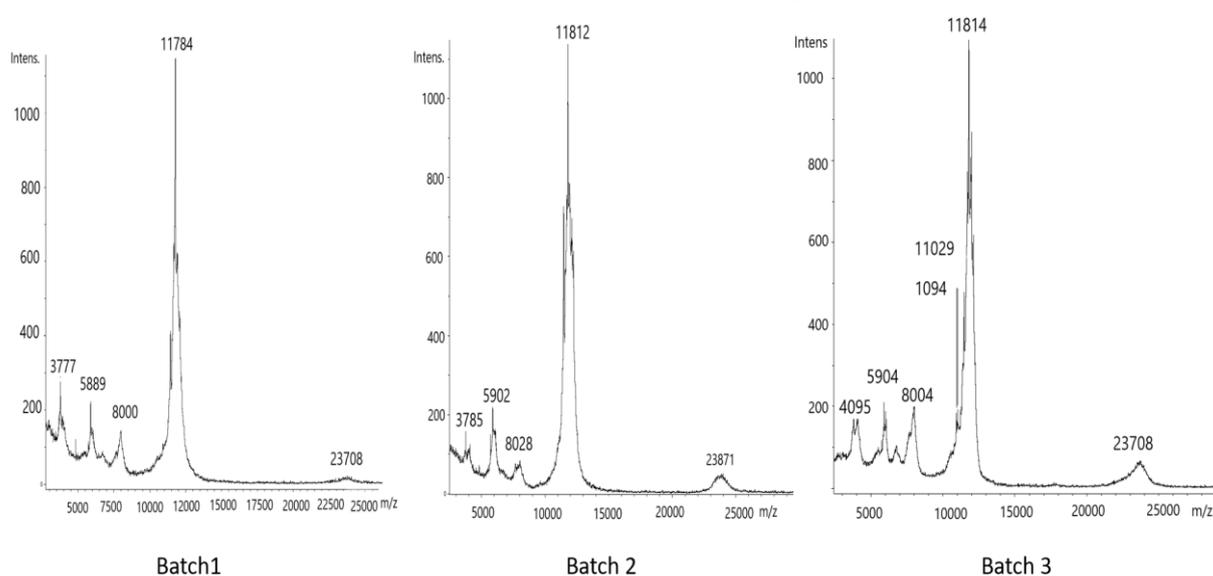
Figure 3.17: Electrospray ionisation time of flight mass spectrometry (ESI TOF MS analysis of fraction C1 sample for the 3 batches obtained

The spectrum for batch 1 (Figure 3.17) allow the identification of one main species - having a

mass of 11784 Da whereas the spectra for batches 2 and 3 gave a peak having similar values

of 11812 and 11814 Da respectively. However, the peaks are not very accurate and many

other species (summarized in the table 3.6) with values ranging from 11443 to 12244 Da were

also identified. The same observations were made from the previous analysis. A mass of

around 5900 Da was also identified for the 3 batches, in addition to two other values of around

8000 Da and 3700 Da. The sum of these masses lead to the value of the main species obtained,

suggesting the presence of two chains in the composition of the proteins.

Fraction C1	Batch1	Batch2	Batch 3
	11784	11812	11814
	5892	5902	5904
	11443	11471	11467
Observed	11682	11715	11714
mass*	11912	11906	11987
(Da)	12085	12117	12058
		12244	12115
	8000	8028	8004
	3777	3785	4095

Table 3.6 : Identification of the main protein species observed in fraction C for the different batches by mass spectrometry analysis. In red the main peaks are shown, in black minor peaks.
 * Mass error tolerance (acceptability): protein ESI TOF MS: 10-50 ppm depending on protein size; peptide ESI TOF MS: 5-10 ppm depending on peptide size

The N-terminal sequencing performed in parallel with this study confirmed the presence of

these two chains and the reverse phase purification described in section 3-2-3 allows the

separation of the chains. The measurements described below were carried out on separated

chains of fraction C1 and more accurate data were expected allowing the characterisation and

the identification of proteins.

MS on separated chains

6 different samples corresponding to different peaks were obtained from C18 purification

(section 3-2-3, Figure 3.7) and analysed by mass spectrometry using matrix-assisted laser

desorption/ionization time of flight mass spectrometry (MALDI TOF MS) (Autoflex, Brucker

Daltonics). Fraction numbers 32 to 34 and 36 corresponding to the small chain (Figure 3.18)

and 37 and 39 (long chain) at 0.1 mg/ml were diluted 1:2 or 1:5 in sinapinic acid matrix and 2

µl were deposited directly on the target.
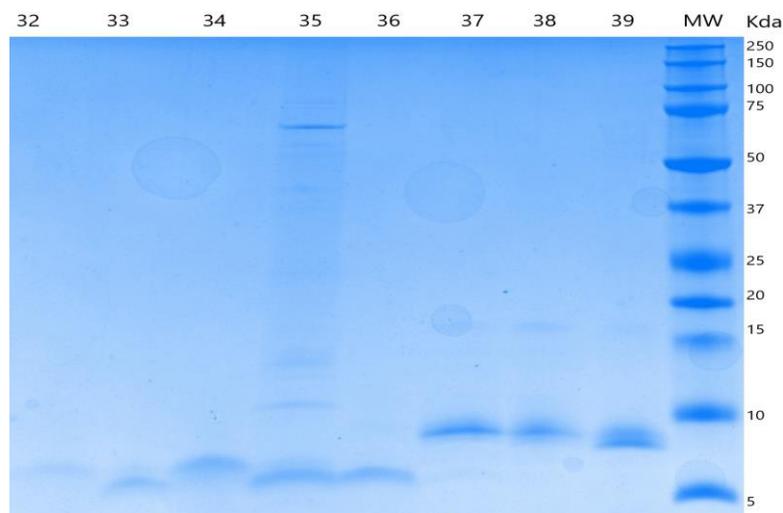
Figure 3.18 : Tris-Tricine gel showing different fractions of the C18 purification.

As this separation was achieved after a reducing and alkylation process on each cysteine, the

observed masses of the chains are heavier by 158 Da for each alkylated cysteine present in

the sequence. A summary of the MS data obtained is given in Table 3.7. The analysis of the

different fractions revealed a complex mixture of different masses for each fraction analyzed.

The main values ranged from 3781 to 4198 Da for the small chain and from 7704 to 8139 for

the long chain . This heterogeneity prevented the acquisition of an definitive mass and the

identification of peptides.

Samples	Fraction 32	Fraction 33	Fraction 34	Fraction 36	Fraction 37	Fraction 39
Observed mass* (Da)	**4198** 4141 3251 2562	**3897** 3833	**4079** 4018 3962 3905 4149 4172	**3781**	**7846** **8139**	**7704** 8012 7838 8134

Table 3.17 : Identification of the main protein species observed in different samples from C18 purification by mass spectrometry analysis. In red the main peaks detected, in black minor peaks

* <u>Mass error tolerance</u> (acceptability): protein MALDI TOF MS: 300-1000 ppm depending on protein size; peptide MALDI TOF MS: 50-300 ppm depending on peptide size

To go further in the composition of theses mixtures, a MS study after *trypsin in gel digestion* was performed on the same samples and is described below.

MS on separated chains after *tryspsin in gel digestion*

The samples analysed for this *trypsin in gel digestion* approach were the same as used in the section above. The choice of the fractions was based on the quality of the spectra obtained previously. Three samples of "the short chain", called S1 to S3 and four samples of "the long chain", called S4 to S7, were digested following the procedure described in Chapter 2 section 2-2-3. (Figure 3.19). This technique consists of 3 steps: the wash, the digestion and the peptide extraction.
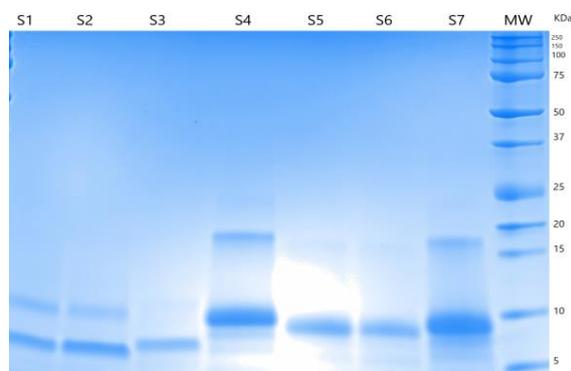
Figure 3.19 : Tris-Tricine gel showing fractions of short and long chains of fraction C1. Bands of each sample were cut and analysed by mass spectrometry (MS),(MW=molecular weight standard).

This technique consists of 3 steps: the wash, the digestion and the peptide extraction. The eluates obtained, were lyophilized to a volume of 10 µl volume and samples analyzed using MALDI-TOF MS. MS data were processed automatically using Mascot Distiller software (v. 2.5.1, Matrix Science; Perkins *et al*., 1999). The searches were carried out using a sequence database containing the *Moringa* sequences derived. The main tryptic fragments generated for each sample are summarized in the table 3.8 and were compared with the database.

Mo-CBP3 isoforms				
	Mo-CBP3-1	*Mo*-CBP3-2	*Mo*-CBP3-3	*Mo*-CBP3-4
Observed mass *m/z*	SHORT CHAIN			
S1				
1072,391		QQFQTHQR		HQFQTQQR
1388,535		CRQQFQTHQR		CRHQFQTQQR
S2				
1072.47		QQFQTHQR		HQFQTQQR
1348.549		CRQQFQTHQR		CRHQFQTQQR
S3				
1057.662	HQFQSQQR		QQFLTHQR	HQFQTQQR
1072.640		QQFQTHQR		
1388.795		CRQQFQTHQR		CRHQFQTQQR
2384.131				HQFQTQQRLRQCQRVIRR
	LONG CHAIN			
S4				
867.608	RPPTLQR			RPPTLQR
879.552	NVSPFCR			NVSPFCR
1276.783	QLRNVSPFCR			QLRNVSPFCR
1312.823				IPAICNLQPMR
2100.324			QAVQLAHQQQGQVGPQQVR	QAVQSAQQQQGQVGPQ QVGHMYR
2553.528	QAVQSAQQQQGQVG PQQVGHMYR			
S5				
740.13			RPAIQR	
1312.858				IPAICNLQPMR
2100.403			QAVQLAHQQQGQVGPQQVR	
2272.507	VASRIPAICNPQPMRC PFR			
2553.552	QAVQSAQQQQGQVG PQQVGHMYR			QAVQSAQQQQGQVGPQ QVGHMYR
S6				
740.508			RPAIQR	
1312.871				IPAICNLQPMR
2100.386			QAVQLAHQQQGQVGPQQVR	
2553.571	QAVQSAQQQQGQVG PQQVGHMYR			QAVQSAQQQQGQVGPQ QVGHMYR
S7				
867.627	RPPTLQR			RPPTLQR
879.546	NVSPFCR			NVSPFCR
1312.94				IPAICNLQPMR
2100.441			QAVQLAHQQQGQVGPQQVR	
2553.633	QAVQSAQQQQGQVG PQQVGHMYR			QAVQSAQQQQGQVGPQ QVGHMYR

Table 3.8: Identification of the main peptides obtained from "trypsin in gel digestion" approaches used for both chains. The error value in the measurement is between 54 and 964 in ppm or 0.06 and 1.1 in Da.

Table 3.8 shows only the main peptides identified from the trypsin digestion step. The observed masses for samples S1 to S3 (small chain) may correspond to peptides of *Mo*-CBP3-2 or *Mo*-CBP3-4 isoforms due to the palindromic sequence. An additional peptide was observed only in sample S3 that might fit to 3 different isoforms except *Mo*-CBP3-2. For samples S4 to S7 (long chain), no peptides corresponding to isoforms *Mo*-CBP3-2 were found, but the other peptides could belong to the three other isoforms; the high level of similarities in the amino acid sequence of these 4 isoforms restricted the interpretation of the data. On the one hand, a better separation of the isoforms from the fraction C1 may improve the data analysis, while on the other hand, the combination of MS and MS/MS could be helpful in distinguishing the different isoforms.

3-4-3 Combining studies of MS and Tandem MS on fractionation of Fraction C1 (*Mo*-CBP3 isoforms) and on the crude extract (CE)

Tandem MS on *Mo*-CBP3 isoforms

In the literature (Baptista *et al.*,2017), it has been reported that the properties of the CE in water treatment may vary depending how the extraction was done – *eg* in the presence or absence of salt. An experiment was performed by running SEC pre-equilbrated with different amounts of salt (0 to 150 mM NaCl) to see if the behaviour of fraction C1 was affected. Surprisingly, it was found that the sharp eluted peak (Figure 3.20 a)) could be fractionated in a series of several peaks (Figure 3.20 b)) with a concentration of salt at 50 mM (peak C2).

a1)
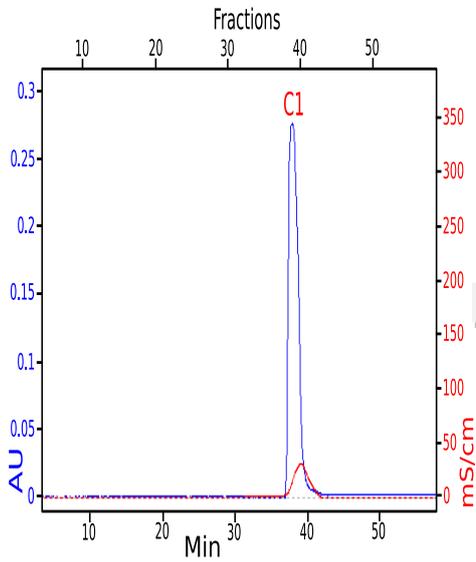
a2)
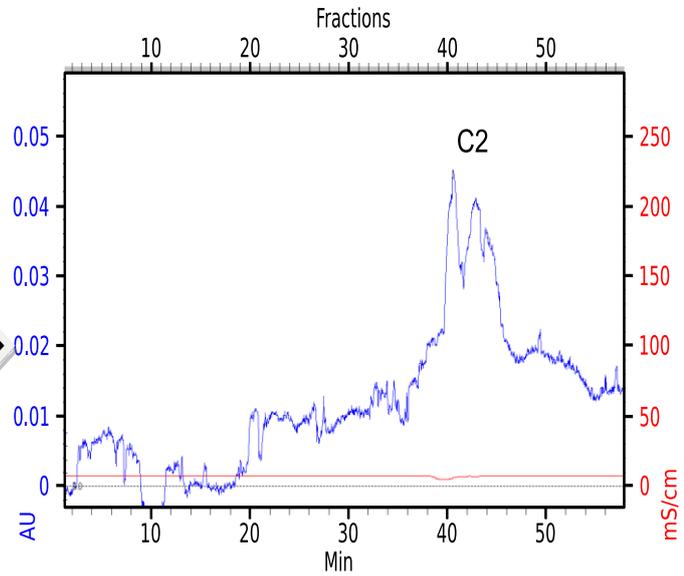
b)
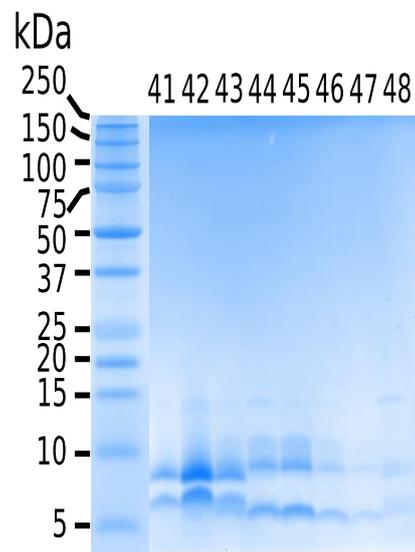
Figure 3.20: **a)** Purification of the fraction C1 using S75 column, **a1)** S75 equilibrated in water, **a2)** S75 equilibrated in 50 mM NaCl. Blue curve : absorption curve at 280 nm (in milliabsorption unit (mAU)); red curve: conductivity curve (in milliSiemens, mS). **b)**Tris-Tricine gel analysis of the 3 small peaks of C2.

This peak C2 was composed of three small peaks and the corresponding components were analysed by gel electrophoresis (Figure 3.20 b). This figure shows that the fractions 41 to 43 exhibit the same pattern of migration whereas the bands of fraction 44 to 46 migrate differently. These results show an unexpected change in the apparent size distribution of the protein components when run through the 50 mM NaCl equilibrated column. This difference in the two elution profiles (*i.e.* peaks C1 and C2) could be attributed to salt-dependent interactions between *Moringa* proteins and the dextran polysaccharide component of the size exclusion column. Fractions 42, 44 and 46 were analysed by MS, using MALDI TOF MS, nano LC-ESI MS, and MS/MS to characterize and identify the main proteins present in these extracts.

Peptide variable modifications allowed during the search were: acetylation (Protein N-ter), oxidation (M) and PyroGlu (Q) whereas carbamidomethyl (C) was set as a fixed modification. The peptide mass tolerance was set at 8 ppm and fragment tolerance to 25 mmu (0.025 Da). The searches were carried out using two sequence databases: one, containing the *Moringa* sequences derived using the FASTA program (Lipman *et al.*,1985), and the other from Uniprot Viridiplantae (7,849,010 sequences).

The samples were measured before and after reduction with tris (2-carboxyethyl)phosphine (TCEP). A summary of the mass spectrometric data obtained is given in Table 3.9.

Without reduction

Sample [1]	Experimental mass (Da) [2]	Theoretical mass (Da) [3]	Sequence assignment [4]	Modifications	Protein isoform
42	11959 *12088* *12074*	11956	S+L +2 pyroGlu + 3 or 4 SS bridges	Gln→Pyr (N-ter of S and L chains)	*Mo*-CBP-3-3
44	11785 11898 12010	11785	S+L +2 pyroGlu + 3 or 4 SS bridges	Gln→Pyr (N-ter of S and L chains)	*Mo*-CBP-3-4
46	11786 *11900* *12014*	11785	S+L +2 pyroGlu + 3 or 4 SS bridges	Gln→Pyr (N-ter of S and L chains)	*Mo*-CBP-3-4

With reduction[5]

Sample	Experimental mass (Da)	Theoretical mass (Da)	Sequence assignment	Modifications	Protein isoform
42	4083 7881	4082 7882	small chain (S) large chain (L)	Gln→Pyr (N-ter of S and L chains)	***Mo*-CBP3-3**
44/46	3777 8016	3777 8016	small chain (S) large chain (L)	Gln→Pyr (N-ter of S and L chains)	***Mo*-CBP3-4**

Table 3.9: Identification of the main protein species observed in different fractions by mass spectrometry MS and tandem MS analysis.

[1] Fractions 42 and 46 and crude extract were analyzed by nanoLC ESI MS and MS/MS; crystals were analyzed by MALDI TOF MS.

[2] Mass values: the mass of the highest specie of isotopic pattern for nanoLC ESI MS spectra and average mass for MALDI TOF MS spectra are reported respectively.

[3] Theoretical mass values calculated considering the formation of four disulfide bridges between the eight available cysteine residues present on the small (S) and large (L) chains.

[4] Mo-CBP3-3 isoform:

Small chain (S): Pyr-QQGQQQQQCRQQFLTHQRLRACQRFIRRQTQGGG

Large chain (L): Pyr-QARRPAIQRCCQQLRNIQPRCRCPSLRQAVQLAHQQQGQVGPQQVRQMYRLASNIPAICNLRPMSCPFG

Mo-CBP3-4 isoform:

Small chain (S): Pyr-QQQRCRHQFQTQQRLRACQRVIRRWSQGGGP

Large chain (L): Pyr-QARRPPTLQRCCRQLRNVSPFCRCPSLRQAVQSAQQQQGQVGPQQVGHMYRVASRIPAICNLQPMRCPFR

[5] Samples were measured without reduction and after reduction with TCEP.

The Tandem MS measurements on the reduced chains (S: small chain and L: large chain) allowed the identification of the abundant isoforms in each fraction (Table 3.9)

MS analysis demonstrates the separation of different species after the size exclusion purification, with a predominance of *Mo*-CBP3-3 component present in fraction 42 and of *Mo*-CBP3-4 component present in fractions 44 and 46. These measurements also reveal a complex mixture of N- and C-terminal processed species of various *Mo* seed protein isoforms (*Mo*-CBP-3). The next paragraph reports the MS/MS analysis performed on the CE that could be compared with the present results.

Tandem MS on the crude extract

For the deep proteomic analysis of the *Moringa* CE, 20 µg were deposited on a 1D gel (4-12 % NuPage Gel - Invitrogen) for a short migration, followed by Coomassie blue staining. The twelve 1D gel bands were treated and analyzed by liquid chromatography LC-MS/MS (Figure 3.21). The data were combined using the software mMass data miner (Strohalm *et al.*, 2008) and the searches were carried out using two sequence databases - one, containing the *Moringa* sequences derived using the FASTA program, and the other from Uniprot Viridiplantae (7,849,010 sequences).
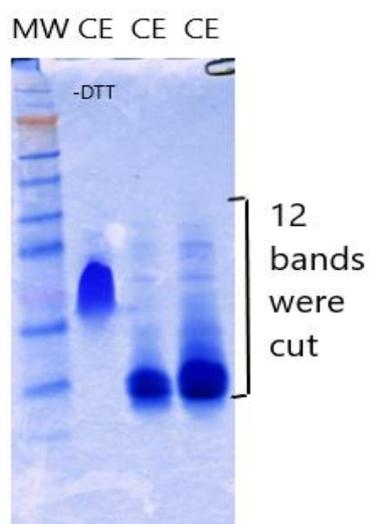
Figure 3.21 : Coomassie Gel staining showing different amount of CE loaded in absence (lane 2) or presence (lanes 3 and 4) of 1,4-Dithiothreitol (DTT) .

The MS/MS measurements allowed the identification of different *Mo*-CBP3 isoforms present in the CE. The 2D map Figure 3.22 a) shows a high abundance of proteins exhibiting a mass between 11000 and 12000 Da in the circle. Proteins with a mass of 8244/8301 Da are relatively abundant but could not be identified due to lack of database information. A multitude of small peptides from 1000 to 5000 Da are also present in the CE.

a)
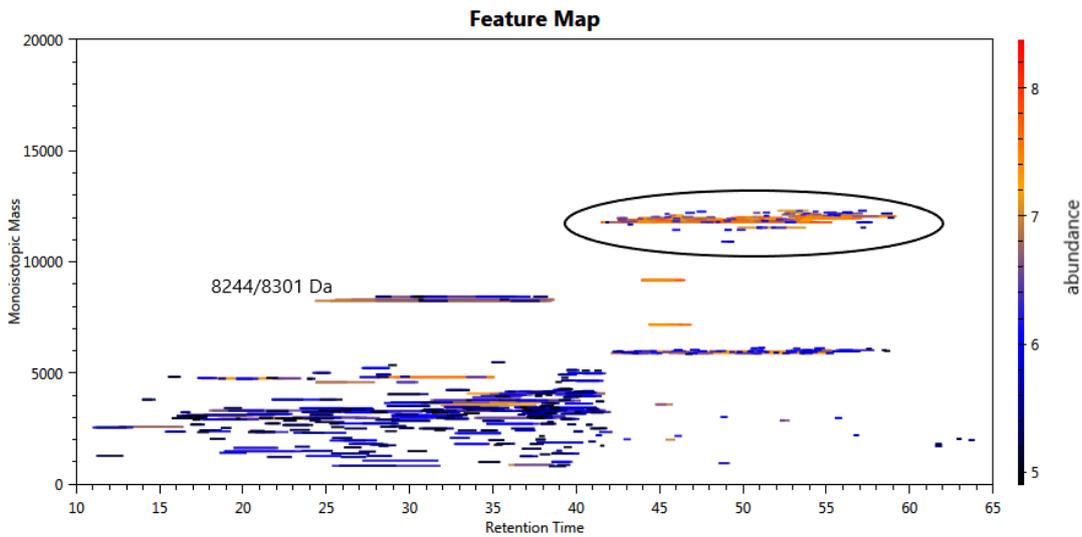
b)
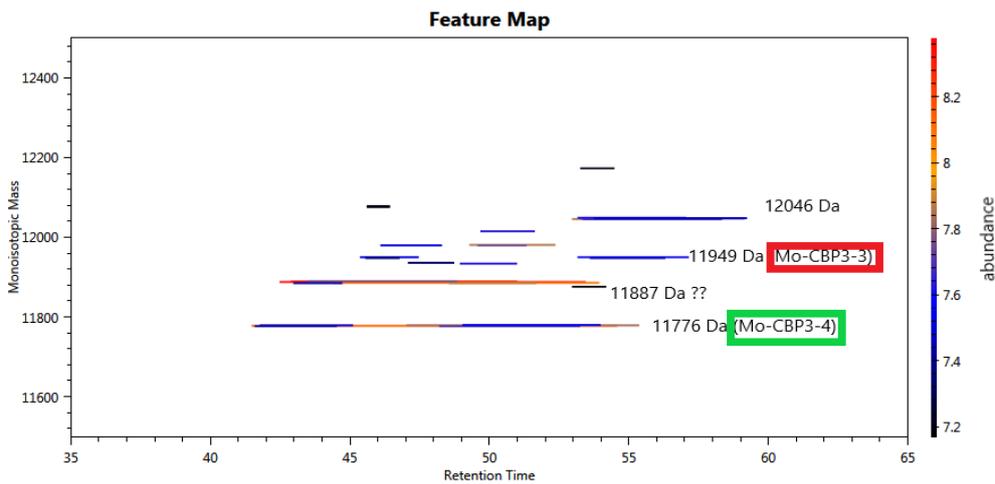
Figure 3.22: 2D features map of the crude extract (CE) of mono-isotopic masses **a)** overview of the different species present in the CE Circled are the masses corresponding to *Mo*-CBP3 proteins **b)** zoom on the isoforms visible in the CE. Orange represents the most abundant species and blue the low abundance species.

Figure 3.22b shows the 'zoomed' region of the 2D map of the CE and allows the clear

identification of two isoforms *Mo*-CBP3-3 at 11949 Da, and *Mo*-CBP3-4 at 11776 Da. The

identification of *Mo*-CBP3-2 was possible only under reducing conditions, as shown in Table

3.10.
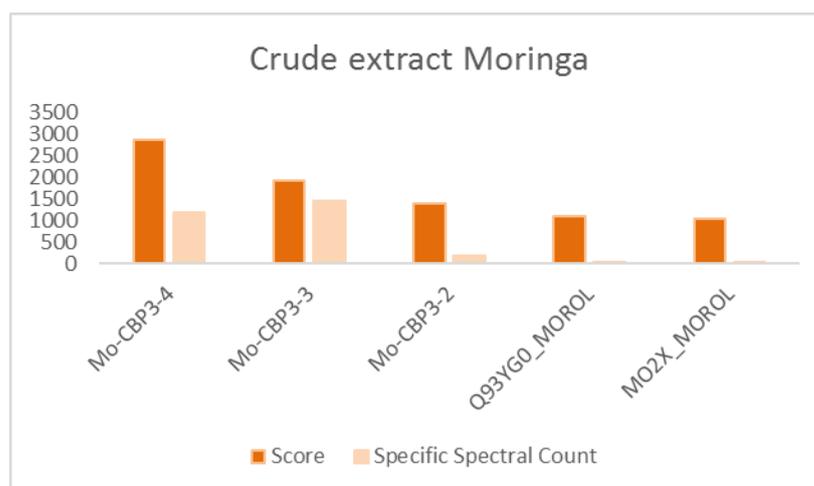
Figure 3.23: Bar Graph showing the abundance of Moringa species in the crude extract (CE) after analysis of the 12 bands digested by trypsin. Q39YGO_Morol and MO2X_MOROL represent both MO2.1 and MO2.2 flocculant proteins.

The bar graph in Figure 3.23 is a representation of the distribution of the different proteins

identified in the CE using the Moringa database. Due to the high aspecificity of the peptides,

scores do not reflect the proportion of the protein. The number of specific peptides was very

low for the two small proteins MO2.1 and MO2.2 (coded MO2X and Q93YG0 on the uniprot

website).

Without reduction

Sample [1]	Experimental mass (Da) [2]	Theoretical mass (Da) [3]	Sequence assignment [4]	Modifications	Protein isoform
Crude Extract	11785 11955	11785 11956	S+L +2 pyroGlu + 3 or 4 SS bridges	Gln→Pyr (N-ter of S and L chains)	Mo-CBP-3-4 Mo-CBP-3-3

With reduction

Sample [1]	Experimental mass (Da) [2]	Theoretical mass (Da) [3]	Sequence assignment [4]	Modifications	Protein isoform
Crude Extract	Trypsin digestion on 12 *1D gel bands Nano LC-MS/MS analysis [6] and identification by Mascot software				**Mo-CBP3-4** **Mo-CBP3-3** **Mo-CBP3-2**

Table 3.10: Identification of the main protein species observed in the crude extract (CE) from *Mo* by tandem mass spectrometry or MS/MS analysis.

[1] Fractions 42 and 46 and crude extract were analyzed by nanoLC ESI MS and MS/MS; crystals were analyzed by MALDI TOF MS.

[2] Mass values: the mass of the highest specie of isotopic pattern for nanoLC ESI MS spectra and average mass for MALDI TOF MS spectra are reported respectively.

[3] Theoretical mass values calculated considering the formation of four disulfide bridges between the eight available cysteine residues present on the small (S) and large (L) chains.

[4] Mo-CBP3-3 isoform:

Small chain (S): Pyr-QQGQQQQQCRQQFLTHQRLRACQRFIRRQTQGGG

Large chain (L): Pyr-QARRPAIQRCCQQLRNIQPRCRCPSLRQAVQLAHQQQGQVGPQQVRQMYRLASNIPAICNLRPMSCPFG

Mo-CBP3-4 isoform:

Small chain (S): Pyr-QQQRCRHQFQTQQRLRACQRVIRRWSQGGGP

Large chain (L): Pyr-QARRPPTLQRCCRQLRNVSPFCRCPSLRQAVQSAQQQQGQVGPQQVGHMYRVASRIPAICNLQPMRCPFR

[5] Samples were measured without reduction and after reduction with TCEP.

[6] Nano LC-ESI-MS/MS analysis were carried out on the crude extract sample after running it on a 1D-SDS gel; twelve bands were cut and submitted to in-gel reduction and trypsin digestion before mass spectrometry.

In *gel tryptic digestion* of the CE, the results for the identification of the three forms of *Mo*-CBP3 demonstrate that the concentration decreases in the following order: *Mo*-CBP3-4> *Mo*-CBP3-3>> *Mo*-CBP3-2. A pyroglutamate residue was identified at the N-terminal position of each individual chain for all *Mo*-CBP-3 isoforms. However, many other peptides with small

molecular weight were observed but could not be characterised in absence of genomic data of *Mo.*
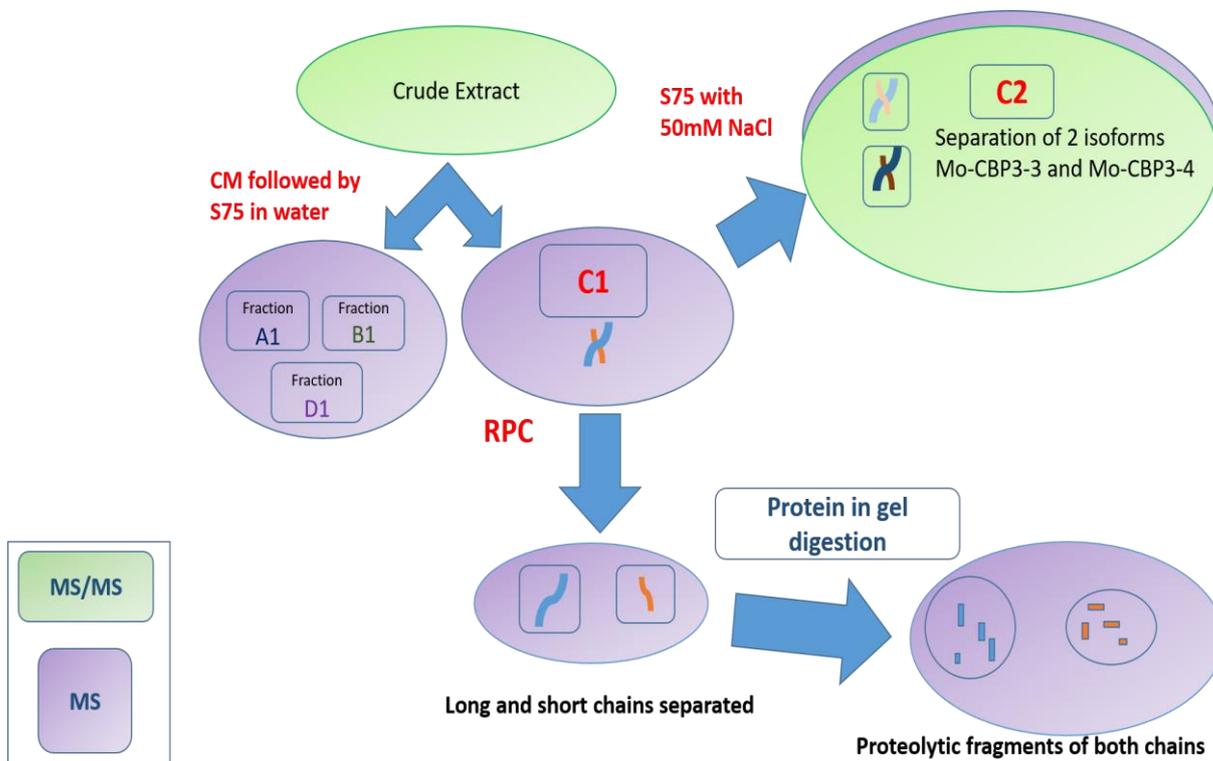
Figure 3.24: Summary of the different mass spectrometry (MS) (in purple) and MS/MS (in green) analyses carried out on the different samples obtained. The MS/MS results on both crude extract (CE) and isoforms separated allowed the estimation of the abundance of the different species present in each samples.

3-5 Activity assays

As described in the introduction chapter (Chapter 1), many studies have shown that crude seed extract are an active flocculating agent and can remove over 90% of the bacterial load of unprocessed water. The present study uses activity assays to relate the properties and behaviour of the purified fraction C1 containing essentially two isoforms of *Mo*-CBP3 (tandem MS analysis section 3-4-3) to those of the CE. The next section describes the different activity assays that were carried out such as the determination of the minimal inhibitory concentration (MIC), an antifungal study on human pathogenic fungi and flocculation tests on

different organisms. The final experiment focuses on the existence of a chitinolytic activity as demonstrated by Gifoni and co-workers (2012).

3-5-1 Determination of Minimal Inhibitory Concentration (MIC)

The minimal inhibitory concentration (MIC) is defined as the lowest concentration of an antimicrobial substance that will inhibit the visible growth of a microorganism after overnight incubation. This study was carried out in collaboration with the Centre Hospitalier Universitaire (C.H.U.) at Grenoble in the Unité Médicale de Bactériologie-Hygiène Hospitalière of Professor Max Maurin. Antimicrobial susceptibility testing of CE and purified fraction C1 was performed against two bacterial reference strains representative for gram positive and gram negative bacteria: *E. coli* ATCC25922 (gram negative) and *S. aureus* ATCC29213 (gram positive).The protocol is described in Chapter2 (section 2-3-1). MICs corresponded to the minimum concentration that allowed complete inhibition of visual growth of bacteria.

compound	strain	MIC (mg/ml)
Crude extract (CE)	*S. aureus* ATCC29213	>40
	E. coli ATCC25922	>40
Purified fraction C1 containing components *Mo*-CBP3-3 and *Mo*-CBP3-4.	*S. aureus* ATCC29213	>10
	E. coli ATCC25922	>10
Gentamicin	*S. aureus* ATCC29213	0.25
	E. coli ATCC25922	0.5

Table 3.11: Summary showing the minimal inhibitory concentrations (MIC) (mg/ml) of purified *Moringa* seed proteins and crude extract (CE) and compared with the antibiotic gentamicin.

The results, as shown in Table 3.11, reveal no intrinsic antibacterial activity of purified *Mo*-CBP3-3/*Mo*-CBP3-4 against the two strains. The MIC values are shown to be over the highest concentration tested (i.e. over 40 mg/ml for the CE and 10 mg/ml for the purified material). The method was validated by the control measurements for Gentamicin which were in the expected range (0.00012 to 0.001 mg/ml and 0.00025 to 0.001 mg/ml for *S. aureus* and *E. coli* respectively).

3-5-2 Antifungal tests

In 2012*, Gifoni et al* demonstrated that the *Mo*-CBP3 protein exhibits an antifungal activity against phytopathogens. The goal of this work was to investigate if this protein has the same biological properties against human pathogens such as the *Candida* species. This study was carried out in collaboration with the C.H.U at Grenoble in the Laboratoire de Parasitologie-Mycologie of Professor Muriel Cornet.

The antifungal tests were conducted based on the inhibition of the growth of different *Candida* species. To evaluate this effect, 5 species of *Candida* were chosen: *Candida albicans* (ATTCC90028), *Candida glabrata* (ATCC2001),*Candida tropicalis* (ATCC7349), *Candida parapsilosis* (ATCC22019) and *Candida krusei* (ATCC6258). 200 µl of *candida* suspensions ($1x10^5$/ml) were incubated with a *Mo*-CBP3 or CE concentrations at 50 µg/ml in 96 wells microplates. After 24h and 48 h of incubation, the turbidity of these suspensions was visually observed. These tests were repeated 3 times. The results reveal no inhibition of the growth of all *Candida* species tested either with *Mo*-CBP3 or with CE (data not shown). To complete this study, a synergistic test was also performed to determine if the protein could play a role as enhancer of antifungal drug activity.

The synergistic tests were performed with four antifungals used in human therapy belonging to different families Fluconazole, Voriconazole, Anidulafungin and Amphotericin B. They were mixed at a various concentration with Mo-CBP3 or the CE at two concentrations of 15 µg/ml or 7.5 µg/ml respectively in order to determine the MIC. The MIC value was found to be higher than 7.5 µg/ml and this concentration was chosen to carry out the synergistic experiment. In addition to the *candida species* already used, three other strains carrying out resistant to azole antifungal like DSY296 (*Candida albicans*) ; TG5 (*Candida glabrata*) and Guer (*Candida parapsilosis*) were tested.

The results (data not shown) indicate that no synergistic effects were observed on *Candida* species when the protein *Mo*-CBP3 or CE were associated to the antifungal drugs.

3-5-3 Gel diffusion assay for chitinase activity

The mechanism of action that relates to the antifungal activity against phytopathogenes is not well understood. Two mechanisms can be envisaged that lead to the disruption of the membrane, either the degradation of chitin (action on the chitinases enzymes) which is the main component of the membrane or the augmentation of chitin (action on chitin synthase enzymes). This gel-diffusion study, described in Chapter2 (section 2-3-3) was conducted to elucidate how the protein may act. The assay is based upon diffusion of the enzyme from a central well through agarose gel containing the substrate glycol chitin, a soluble modified form of chitin. Calcofluor fluoresces under UV light when it binds to undigested chitin, whereas the region where glycol chitin was digested appears dark under UV light, as calcofluor does not have affinity for the digested chitin. Purified chitotriosidase standard was used as a positive control.
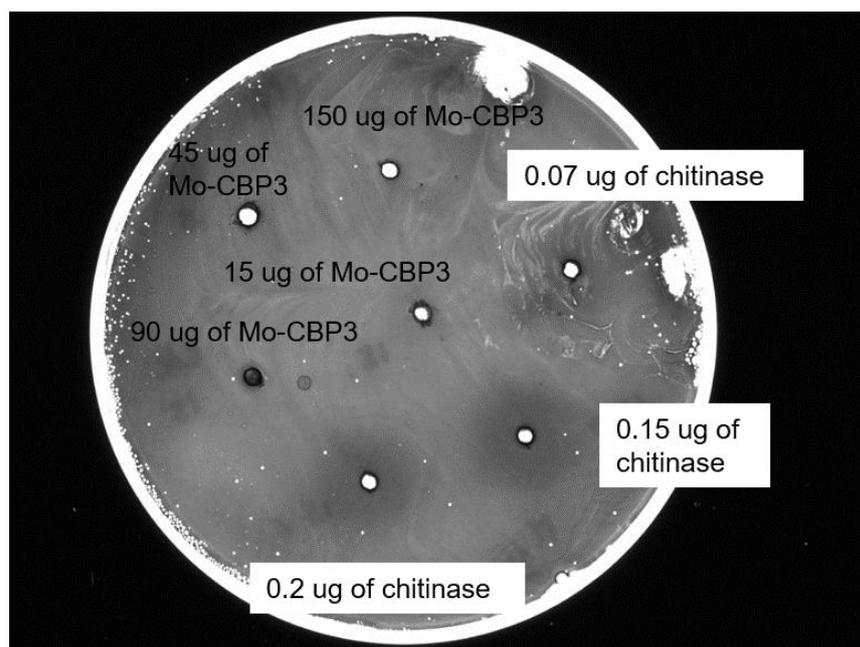
Figure 3.25: Gel diffusion assay of chitinase activity from fraction C1 (*Mo*-CBP3 isoforms). The row on the right contained serial dilutions of purified human chitotriosidase protein available in the laboratory.

As indicated by the enzyme control, the diffusion zone area was proportional to the amount of the enzyme applied. In the case, of *Mo*-CBP3 protein, no diffusion zone area was visible, indicating that the chitinase activity was not detected in *Mo*-CBP3 isoforms (Figure 3.25).

3-5-4 Determination of flocculation and coagulation activities

Simple tests of flocculation were made by adding *Moringa* protein from fraction C1 or CE to dispersions of sulfonated polystyrene latex, and observing cluster formation with an optical inverted microscope (Leica DMI1). The latex, PS3, was the same as that used in previous studies of flocculation using ultra small-angle neutron scattering made by Hellsing *et al.,* (2014). Latex has been extensively characterised using a variety of techniques (Rennie, 2013) and has a zeta potential of about -35 mV. The final concentration of fraction C1 of Moringa proteins was 2mg/ml and the concentration of particles was 4% wt.

To test the coagulation activity on *E.coli* and *Nannochloropsis gaditana* cells, stationary phase cultures were supplemented with 0.4mg/ml of fraction C1 of Moringa proteins and the coagulation observed under the microscope.
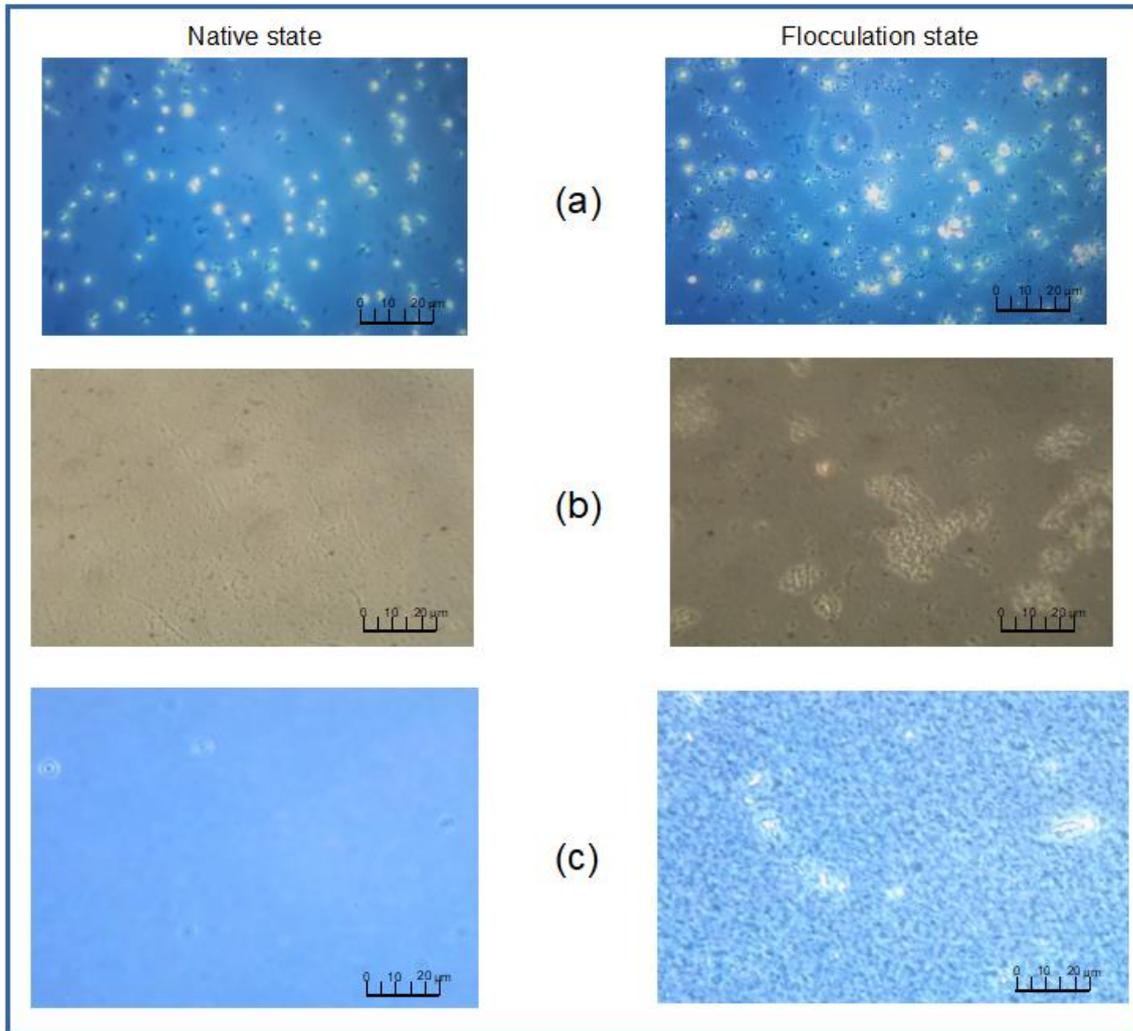
Figure 3.26: Microscopic images showing flocculation by the purified protein (peak C1 in Figure 3.5) for three different material : (a) the algae *Nannochloropsis gaditana*, (b) the bacterium *Escherichia coli* (BL21)*,* (c) Latex particles.

Flocculation activity was visualised by standard light microscopy. Figure 3.26 (a) shows that clear flocculation effects (large visible aggregates) are introduced by the addition of the purified seed proteins for the algae *N. gaditana*; the same effect is noted in Figure 3.26 (b) for

E. coli bacteria. Figure 3.26 (c) shows that, as with the CE, the purified protein causes flocculation of anionic polystyrene latex particles. Simple inspection with optical microscopy allows large aggregates to be identified. The particles are the same as those described in previous scattering studies (Hellsing *et al*., 2014) where compact flocs with a high fractal dimension were observed. These results demonstrate the flocculating activity in relation to a wide variety of materials.

3-6 Discussion and conclusion

The Moringa seed proteins of water soluble CE extraction have been purified and the main fraction (eluted by 0.6 M NaCl on CM) has been intensively studied using biochemical and biophysical techniques and activity assays. From the purification procedure of the CE, which includes size-exclusion chromatography, it was found that the proteins eluted as a sharp peak appearing later than for a small globular protein, suggesting an unexpected behaviour in aqueous conditions. The overlay of this same fraction purified from the three different batches confirms the heterogeneity of the sample extracted from natural sources. The batch heterogeneity is very likely to arise climatic variation, or from harvesting time and the storage conditions of seeds. The amino acid composition revealed the high content of glutamine/glutamic acid (25%) and arginine (15%), corresponding to the main features of the 2 S albumin Youle and Huang, 1981. This high content in basic amino acids make the protein highly positively charged with a high isolectric point (IEP). The determination of the IEP was not successful using 1D gel electrophoresis because the IEP value is higher than the range of efficiency of these types of gels. The IEP of the CE was estimated to be between 10 and 11 by Ndabigengesere *et al.,*(1995). The determination of the extinction coefficient (ε) suggested a relatively low quantity of aromatic residues in this fraction and allowed accurate

concentration measurements for all the experiments performed. The heat stability under drastic conditions was analysed by gel electrophoresis and confirmed by CD measurement. The CD analysis showed a typical curve of helical secondary structures which was further confirmed by the crystallographic structure described in Chapter 4. The CD curves obtained with or without heat treatment are very similar, suggesting a high thermostability - a property that has been previously demonstrated on the CE by Kwaambwa and Maikokera (2008). These different characterisations revealed that this fraction of interest contains proteins that belong to the 2 S albumin family. The proteins are characterised by a compact fold due to the presence of a small chain and a large chain, linked together by four disulfide bonds. This arrangement is the key to the extreme stability of the protein as well as its heat and proteolysis resistance. These properties did not facilitate the determination of tryptic fragments useful for the N-terminal sequencing analysis. Moreover, the observation of a weak signal from the N-terminal-sequencing technique was interpreted as the presence of a pyroglutamate on N-terminal and a pre-digestion step by an aminopeptidase was added. The first results obtained revealed the presence of different peptides, suggesting that the fraction might contain several proteins. The X-ray structure analysis carried out in parallel with this characterisation confirmed the presence of two chains. The RPC step therefore allowed the separation of the two chains but these samples were too complex to acquire accurate data for the sequence identification. At this stage of the study, the results of the N-terminal technique were not exploitable due to the lack of sequences in the *Moringa oleifera* database, and did not yield the amino sequence identification required to build the structure model. Since the publication of new *Moringa* sequences in 2015 by Freire and co-workers, the subsequent interpretation

of the data analysis became possible and revealed the presence of two isoforms *Mo*-CBP3-3 and *Mo*-CBP3-4 in the fraction of interest.

The MS analysis conducted in parallel with the biochemical characterisation provided a value around 11800 Da that varied according to the different CE batches studied. The same experiment in reducing conditions allowed the determination of mass of each chain - respectively 8000 Da and 3800 Da. These measurements also revealed the presence of peptides with masses very close to the main species. These observations show the presence of a complex mixture of N- and C-terminal processed species of various *Mo* seed protein isoforms (*Mo*-CBP-3) that may develop during the maturation of the 2S albumin. The change in the apparent size distribution of the protein isoforms when run through the 50 mM NaCl equilibrated SEC allowed their separation and their identification by MS/MS. As mentioned previously, the identification of four isoforms of *Mo*-CBP3 in 2015 by Freire *et al.,* was a major breakthrough in the data analysis. The MS/MS results revealed the presence of two main isoforms *Mo*-CBP3-3 and *Mo*-CBP3-4. The analysis of the CE was very challenging but allowed the identification of *Mo*-CBP3-4 as the most abundant component among the different *Mo*-CBP3 isoforms present. Moreover, MO2.1, which was the most characterised protein of *Moringa* seed in the literature was also observed, but in a tiny amount. The analysis revealed the presence of a multitude of small peptides possessing a molecular weight ranging from 1000 to 5000 Da that could not be identified due to the lack of database and were not observable on the Tris-tricine gel due to their small sizes. These small peptides could occur naturally or be the results of degradation of the proteins. In 2001, Okuda *et al.,* mentioned the presence of polyelectrolites exhibiting a mass of 3000 Da and having coagulation properties.

In terms of carbohydrate identification, neither the gel staining using glycoprotein detection kit, nor the MS analysis approaches demonstrated the presence of a glycosylation state on protein of the fraction C1; these results differ from those reported by Gifoni *et al.*, in 2012. In their paper, these authors characterised the *Mo*-CBP3 protein as a basic glycoprotein with 2.5 % sugar, having an affinity for the chitin (the major component of the fungus wall). Their fractionation was carried out using affinity chromatography for the chitin. They demonstrated the antifungal role of *Mo*-CBP3 against phytopathogenic fungi based essentially on the observation of the spore germination. In this work, the experiment was carried out on human pathogenic fungi which are not able to germinate and they develop as filaments. The study was performed on the basis of growth inhibition of *candida* species either in presence of CE or in presence of the two isoforms *Mo*-CBP3-4, *Mo*-CBP3-3 (Fraction C1) with a working concentration at 50 µg/ml (*eg* 4µM). At this molarity, no significant inhibition or synergistic effects were observed. However, a recent work published in 2017 by Neto *et al.*, reported that *Mo*-CBP3 possesses an anticandidal activity with a MIC_{50} (minimum concentration that inhibits 50% of fungal growth) at 300 µM. This MIC value is very high regarding the concentration of their control used (Nystatin 11µM). This low or absence of activity against *candida* species suggested that the wall composition for spores and filaments might differ and the interaction of protein with their membranes is weaker than with phytopathogenes. Indeed, it has been shown by Gifoni et al., (2012) that *Mo*-CBP3 did not inhibit some fungi in which cellulose predominates and chitin is minor component of their cell wall. Moreover, the purified material (*Mo*-CBP3-4, *Mo*-CBP3-3), did not exhibit a chitinase activity which is similar to the observation reported by the same research team on *Mo*-CBP3. This protein including the isoforms *Mo*-CBP3-4 and *Mo*-CBP3-3 which do not disrupt chitin, possesses probably one chitin-binding

domain and lack a chitin catalytic site. Moreover, Chaung *et al., (*2007) compared the activity

of different extraction methods of CE against dermatophytes. They demonstrated that the CE

ethyl acetate fractionation was the most active whereas the water soluble extraction of CE (as

prepared by collaborators) was poorly active. These different observations show that the

antifungal properties may depend on the combination of proteins present in the sample. This

composition of proteins can vary depending on the extraction methods of the CE from seeds

and the type of fractionation carried out.

The determination of MIC values was shown to be over the highest concentration tested (i.e.

over 40 mg/ml for the crude extract and 10 mg/ml for the purified material) meaning that no

antibacterial or bacteriostatic activity against pathogenic bacteria was observed for the CE or

the purified *Mo*-CBP3 fraction. Suarez and co-workers in 2003 described very low antibacterial

activity of CE. However, their extraction conditions were different from those used in the

present study. Despite the fact that neither the purified protein nor the CE have shown direct

antibacterial effects on their own in the present study, they may well reduce the bacterial load

in treated water by means of flocculation and coagulation. In 2012, Sanchez-Martin and co-

workers demonstrated that the charged fraction studied in this thesis work exhibited the

highest flocculation activity for kaolinite**.** Many publications demonstrated the flocculation

and coagulation properties by monitoring settling rate behaviour of different coagulants in

comparison with the CE (Ghebremichael *et al.*, 2005). These tests were carried out on a large

scale using jar tests composed of glass beakers of 500ml capacity. At the scale of this study,

the use of latex particles to characterise the coagulation effects was more relevant and

required a lower amount of purified material. The surface of the latex particles has a small

overall negative charge that arises from dissociation of sulfate groups: this explains why the

particles are stabilised in aqueous dispersion and how flocculation occurs following combination with the positively charged protein. Similar interactions may also play a role in the flocculation/coagulation properties of the purified fraction for bacteria and algae as demonstrated. In 2006, Katayon *et al.,* demonstrated that the coagulation efficiency of *Moringa* seeds decreased as storage duration increased whereas the temperature and container of storage did not have any effect. Presence of moisture content within the seeds may causes the deterioration of biomaterial. The cationic protein content appears to vary with the stage of seed maturity as shown by Abubakar *et al*., 2017. The mature dried seeds having the higher content of cationic protein, are the most effective in comparison with the green immature seeds; nevertheless the water purification capabilities are only slightly affected. A very recent study by Baptista *et al.,* in 2017, showed that the *Moringa* seeds are mainly composed of proteins of globulin and albumin type, which represent 53 and 44 % of total proteins of the seed and exhibit a high coagulant potential for treatment of low turbidity water. These observations imply that the valuable range of interactions that causes the *Moringa* seed proteins to flocculate a wide variety of materials may depend on a mixture of proteins being present.

4-X-ray crystallographic studies of *Mo*-CBP3-4

This chapter focuses on a crystallographic study of the mature form of *Moringa oleifera (Mo)* chitin binding protein isoform 3-4 (*Mo*-CBP3-4 protein), one isoform of *Mo*-CBP3. Extensive preliminary work was carried out on the determination of the crystallisation conditions of the protein, which were challenging due to the heterogeneity of the sample. When the conditions of the crystal growth were well defined, diffraction data were collected at ESRF synchrotron radiation source in Grenoble on beamline ID29. The crystal structure of the protein was determined at 1.6 Å resolution and the phase determination was solved using the single-wavelength anomalous diffraction (SAD) at the sulphur absorption edge. The compact structure consists of two chains linked by four disulphide bonds corresponds to the characteristic features of the 2 S albumin family to which the protein belonging. This provides structural information relating to this diverse family of albumins and provides structural insights for the observed flocculating properties of the protein.

4-1 Introduction

The crystallographic study of *Mo*-CBP3-4 protein, an isoform of *Mo*-CBP3 which belongs to the 2 S albumin family, is described in this chapter. As mentioned in the introductory chapter (Chapter1), this family displays a pattern of eight-cysteine residues (8CM) in a specific order of type …C…C…/…CC…CXC…C…C…, as described by José-Estanyol *et al*., (2004). Plant sequences containing this motif relate to proteins having different functions, ranging from

storage, protection, enzyme inhibition and lipid transfer to cell wall structure. Four member

representatives of the 8 CM family of proteins, soybean hydrophobic seed protein

(GmHSP)(Baud *et al.*, 1993), wheat α-amylase dimeric inhibitor (TaA10.19)(Oda *et al.*, 1997),

bifunctional corn Hageman factor inhibitor (CHFI)(Behnke *et al.*, 1998), bifunctional α-

amylase/trypsin inhibitor from ragi seeds (EcRATI)(Gourinath *et al.*, 2000 and Strobl *et al.*,

1998) have been crystallized and their tertiary structure solved. The 8CM motif appears to be

a structural scaffold of conserved helical regions connected by variable loops, as observed by

three-dimensional structure analysis. The same pattern of disulphide bridges has been found

by nuclear magnetic resonance (NMR) studies of three albumin proteins namely:

(i)   napin (BnIB) a 2S-albumin from *Brassica napus* purified from Brassica seeds (Rico

         et al., 1996), and its recombinant protein (Pantoja-Uceda *et al.*, 2003)) expressed

         in *Pichia pastoris* (Palomares *et al.*, 2002);

(ii)   an allergen 2S-albumin recombinant protein from *Ricinus communis* (RicC3)

         expressed using minimal medium cultures of *E.coli* (Fernandez-Tornero *et al.*,

         2002)

(iii)   SFA-8 albumin (HaSF8) purified from sunflower seeds (Pantoja-Uceda *et al.*, 2002)

         and from its cyanogen bromide cleavage (Egorov *et al.*, 1996).

However, the only crystal structure of a 2 S albumin available at the beginning of this thesis

work was that of the sweet protein Mabinlin II from *Capparis masaikai* (Li *et al.*, 2008).

The overall strategy for this part of the PhD thesis work was to study the X-ray crystallographic

structure of *Mo* chitin binding protein isoform 3-4 (*Mo*-CBP3-4 protein) which shows

flocculation properties (Chapter 3) and to understand the mechanism of action of this protein.

The crystallisation conditions, as described in section 4-2, were challenging to identify and

optimise because of the heterogeneity of the protein prepared from natural sources. High quality X-ray data were collected (section 4-3) that allowed 80 % of the model to be built *ab initio* in the absence of sequence information. The publication of Freire and co-workers in 2015 which provided the sequences of four isoforms of *Mo*-CBP3, helped to complete the remaining sequence information (section 4-4). A comparable analysis (albeit at slightly worse resolution - 1.68 Å, as opposed to 1.7 Å) was subsequently published by Ullah and co-workers in 2015 where it was mistakenly referred to *Mo*-CBP3-1. A comparison of both structures showed that they were essentially identical.

4-2 Crystallisation of *Mo*-CBP3-4

The protein extract from the 3 different batches provided by the collaborators (Dr Majority Kwaambwa and Prof. Adrian Rennie) was produced using the procedure described in Chapter 3. This material was sent to the High Throughput Crystallisation (HTX) Laboratory of the EMBL Grenoble outstation (a platform of the Grenoble Partnership for the structural Biology (PSB)).

4-2-1 High-throughput crystallisation screening

The initial crystallization screening was carried out at the HTX laboratory. The *Cartesian* robot of the HTX platform uses 100nl drops in a sitting drop configuration - allowing automatic screening of 768 different crystallisation conditions using commercially available screens. A total of 5 imaging inspections was applied to each plate at 1, 3, 7, 15, 33, 61, and 87 days. Promising conditions were then reproduced manually using the hanging drop vapour diffusion technique to optimise the conditions and obtain the best possible quality and size of crystals. The results obtained from three different hanging drop batches are shown in Table 4.1

	Batch 1	Batch 2	Batch 3
Growth time	2 months	2 months	2 months
Crystallisation conditions (identified from initial HTX studies)	0.1M citric acid pH 4 3.2 M sodium formate pH 4	0.1 M citric acid pH 5 2.4M sodium formate pH 5	0.1 M citric acid pH 5 2.4M sodium formate pH 5
Morphology of the crystal			
Size	130 x 80µm	70 x 80 µm	120x 60µm
Resolution	1.6 Å	No good diffraction	1.6 Å
Reproducibility	No	No	Yes
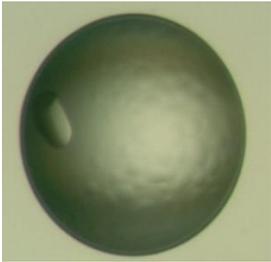
Table 4.1: Summary of crystallisation conditions obtained using high throughput screening for 3 different batches of crude extract (CE). The best crystals were obtained with batch 1 and 3 giving high resolution data (1.6 Å).

Crystallisation of the protein was challenging. The table above shows that the crystallisation conditions used in batch 1 were substantially different from the other batches. Batch 1 crystals appeared at pH 4.0 in the presence of 3.2 M sodium formate and 0.1 M acid citric whereas for the following batches the pH was pH 5.0, with a low concentration of sodium formate. It was observed that in all crystallisation conditions, the growth time was very long: 2 months - corresponding to the last HTX robot inspection. These observations made the crystallisation conditions challenging to determine and optimise. For the last batch, manual screening was carried out using hanging and sitting drop methods since a larger quantity of material was available, and the behaviour of the crystals was reproducible.

4-2-2 Crystallisation by hanging and sitting drop

The hanging and sitting drop vapour diffusion methods were used with solutions of 9 mg/ml and 18 mg/ml *Mo*-CBP3-4, with drops consisting of 2 µl of protein and 2 µl of precipitant solution, the drops were equilibrated at room temperature. Hand-made drops were screened around the following condition: **0.1 M acid citric pH5.0; 2.4M sodium formate pH 5.**

Sodium Formate (Molarity)	1 M Citric acid pH 5.0 (µl)	6 M Sodium formate (µl)	ddH2O (µl)
2.35	100	392.0	508.0
2.40	100	400.0	500.0
2.45	100	408.3	491.7
2.50	100	416.6	483.4
2.55	100	425.0	475.0
2.60	100	433.0	467.0
2.65	100	441.6	458.4
2.70	100	450.0	450.0
2.75	100	458.3	441.7
2.80	100	466.6	433.4
2.90	100	483.3	416.7
3.0	100	500.0	400.0

Table 4.2: Screening around *Mo*-CBP3-4 crystallisation conditions.

Table 4.2 summarises the crystallisation conditions tested where the pH is maintained constant and the sodium formate concentration ranged from 2.35 to 3.0 M. Other trials were performed in which the pH was varied from 4.7 to 5.5, while keeping the sodium formate concentration constant. Several assays were carried out both in hanging and sitting drops at room temperature. Crystal growth was reproducible using both approaches with crystals appearing in the range from 2.4 to 2.9 M of sodium formate, and in the pH range 5.0 - 5.4.
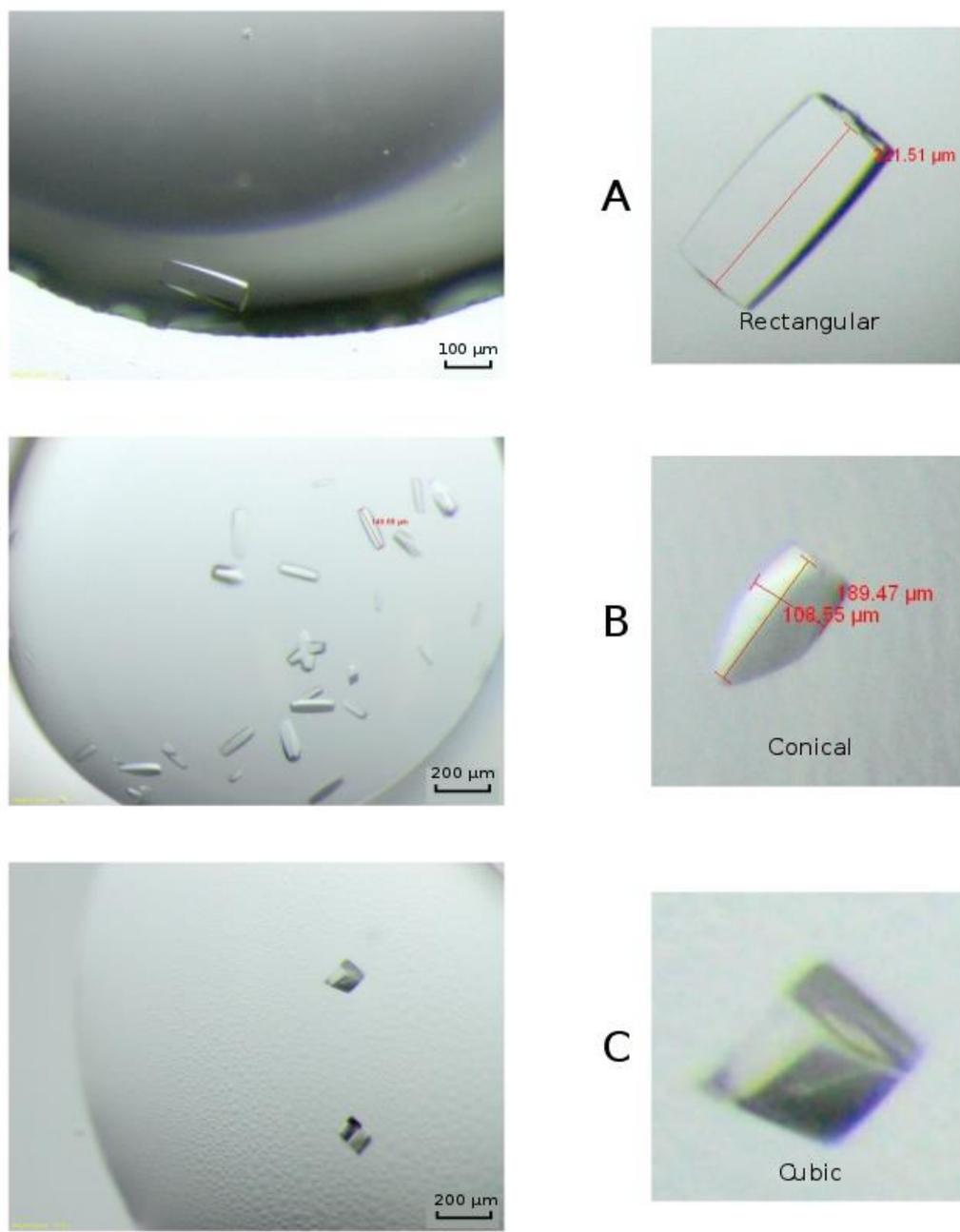
A: 2.45M Sodium formate; 0.1 M citric acid pH5.2 @ 22 mg/ml
B: 2.75M Sodium formate; 0.1 M citric acid pH5.0 @ 22 mg.ml
C: 2.55M Sodium formate; 0.1 M citric acid pH5.4 @ 34 mg/ml

Figure 4.1: Crystallisation behaviour of *Mo*-CBP3-4 in various conditions. Different crystal morphologies are visible – cubic, rectangular, and conic.

Figure 4.1 shows the different crystal morphologies obtained in very similar conditions, with cubic, rectangular and conic shaped crystal observed. Different crystal shapes were also visible in the same drop. Several crystals were exposed to the x-ray beam and based on the quality of the diffraction, the best was the one shown in Figure 4.1B.
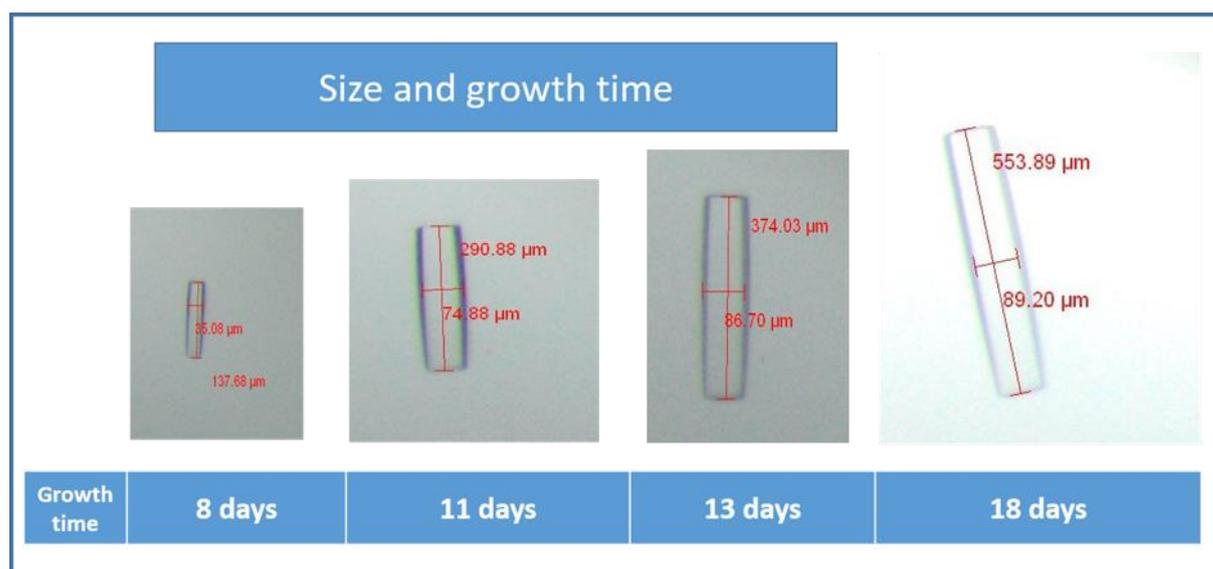
Figure 4.2: Example of growth time versus size of a crystal. The growth conditions were 2.45M sodium formate, 0.1M citric acid pH5.4.

The figure above demonstrates that the crystal grew to a relatively large size in a short period. After 18 days it was already measuring a ~0.5 mm in length. The crystal was then removed, and mounted for data collection on ESRF beamline. Considering the size of these crystals and the fact that these could be obtained readily, a neutron crystallographic study is likely to be possible in the future.

4-2-3 Mass spectrometry of the crystal

Mass spectrometry (MS) is often used as a quality control to measure the mass of a protein accurately prior to crystallisation. However this may not necessarily correspond to the mass of the protein which subsequently crystallizes. MS may be especially helpful if the mature form of a protein is obtained following a series of proteolytic events or if the initial sample contains more than one form of the same protein. In Chapter 3, it has been demonstrated that the purified fraction is composed of a complex mixture of N- and C-terminal processed species of various *Mo* seed protein isoforms. MS was used to determine an accurate mass for the protein present in the crystal; this aided structure determination and functional assignment. The first measurement was performed using crystals that had been previously fished and mounted on cryoloops. Despite several washes, the presence of PEG as cryo-protectant was an obstacle to satisfactory MS measurements from protein samples derived from crystals. The second attempt was performed by fishing crystals directly from crystallisation drops in the absence of either PEG or glycerol. The success of this procedure was aided by the fact that crystals were of reasonable size and as such more amenable to MS measurements.

The MS analysis were carried out at Institut de Biologie Structurale (IBS) with the help of Dr. Luca Signor. The samples were measured before and after reduction with Tris (2-carboxyethyl) phosphine (TCEP). The results of MS analysis are shown in Figure 4.3 and summarized in Table 4.3.
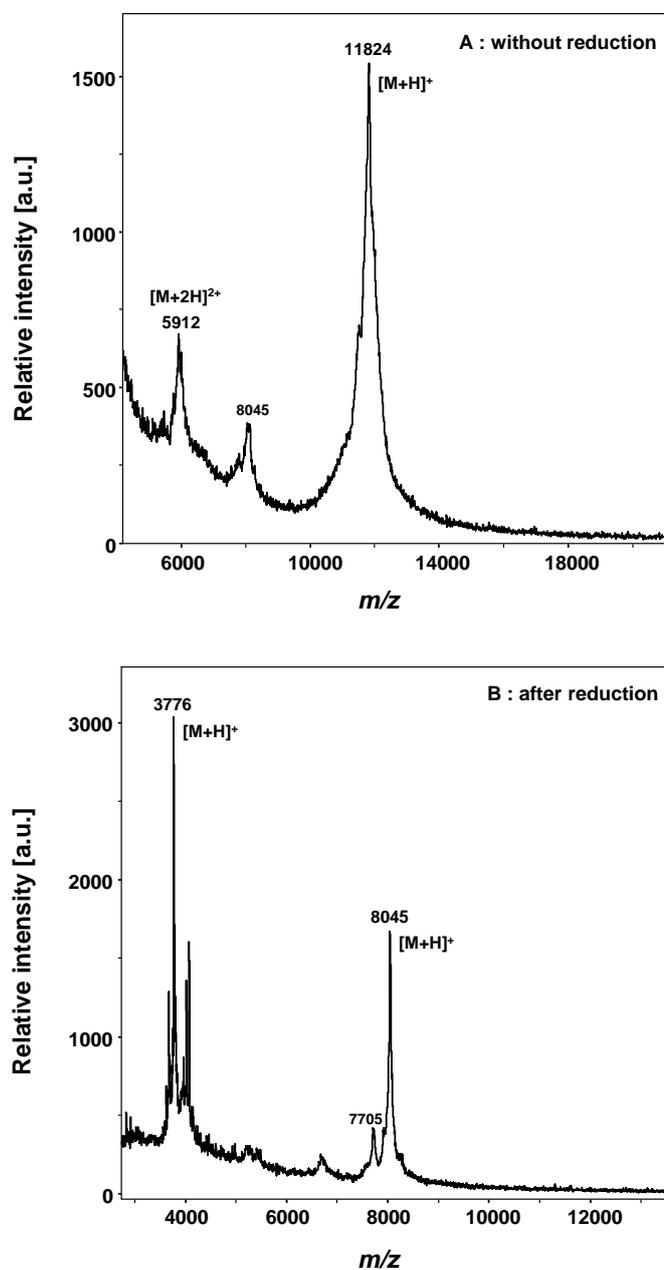
Figure 4.3: Matrix Assisted Laser Desorption Ionisation - Time of Flight (MALDI TOF) mass spectrometry (MS) spectra of protein crystal samples from *Moringa oleifera (Mo)* **A**-without reduction, **B**- after reduction with TCEP 50mM.

Sample [1]	Experimental mass (Da) [2]	Theoretical mass (Da) [3]	Sequence assignement [4]	Modifications	Protein isoform
without reduction [5]					
Crystals	11824	11785.63	S + L	Gln→Pyr (N-ter of S and L chains) 2 oxydations on L chain	*Mo*-CBP3-4
after reduction [5]					
Crystals	3776 8045	3777.28 8016.41	small chain (S) large chain (L)	Gln→Pyr (N-ter of S and L chains) 2 oxydation sites on L chain	*Mo*-CBP3-4

[1]; Crystal sample was analysed by MALDI TOF MS.

[2] Mass values: the mass of the highest specie of isotopic pattern for nanoLC ESI MS spectra and average mass for MALDI TOF MS spectra are reported respectively.

[3] Theoretical mass values calculated considering the formation of four disulfide bridges between the eight available cysteine residues present on the small (S) and large (L) chains.

[4] Mo-CBP3-3 isoform:

Small chain (S): Pyr-QQGQQQQQCRQQFLTHQRLRACQRFIRRQTQGGG

Large chain (L): Pyr-QARRPAIQRCCQQLRNIQPRCRCPSLRQAVQLAHQQQGQVGPQQVRQMYRLASNIPAICNLRPMSCPFG

Mo-CBP3-4 isoform:

Small chain (S): Pyr-QQQRCRHQFQTQQRLRACQRVIRRWSQGGGP

Large chain (L): Pyr-QARRPPTLQRCCRQLRNVSPFCRCPSLRQAVQSAQQQQGQVGPQQVGHMYRVASRIPAICNLQPMRCPFR

[5] Sample were measured without reduction and after reduction with TCEP

Table 4.3: Identification of the protein observed in crystal samples from *Moringa oleifera* (*Mo*) by mass spectrometry (MS).

Matrix Assisted Laser Desorption Ionisation - Time of Flight (MALDI TOF) MS measurements of the protein crystals from *Mo* are summarized in Figure 4.3 and Table 4.3. In Figure 4.3 (the MALDI spectrum without reduction), the main peak is seen at *m/z* 11824, corresponding to the singly charged [M+H]$^+$ molecular ion; the *m/z* 5912 species corresponds to the doubly-charged [M+2H]$^{2+}$ molecular ion. These mass values are in agreement with the intact mass of

isoform *Mo*-CBP3-4 (2S albumin precursor) where the two chains (S: small and L: large) are linked by four disulfide bridges and with the two N-terminal glutamine (Q) modified to pyroglutamic acid (Pyr). The amino acid sequences and modifications for S and L chains were further confirmed by liquid chromatography-electrospray ionisation (LC-ESI) MS/MS analysis. Under reducing conditions, after treatment with TCEP, the main species observed in the MALDI spectrum are at *m/z* 8045 and *m/z* 3776; these mass values correspond respectively to the sequences of the S chain and the L chain, with the L chain carrying two oxidation (+32 Da). Minor peaks could be also be observed in the MALDI spectrum of the crystal sample, for example the peak at *m/z* 7705, corresponding to a shorter form of the L chain, with loss of 3 amino acids glutamine-alanine-arginine (QAR) at C-terminus and carrying two oxidations

This MS analysis allowed the determination of the accurate mass of the *Mo*-CBP3-4 protein present in the crystal of 11824 Da in non-reduced conditions and 8045 and 3776 Da respectively for small and long chain after treatment with TCEP. These results also highlighted the presence of post-translational modifications such as pyroglutamate in N-terminal, and proteolytic processing in the C-terminus- for example the loss of QAR in the long chain. The oxidation observed could be the result of the relatively long storage (2 years) of the crystal at room temperature.

4-3 X-ray crystallographic data collection

Data collection was performed on the ID29 beamline at ESRF (De Sanctis *et al.*, 2012). All data were collected at 100 K on a single crystal using a 0.1° oscillation range. Figure 4.4 shows a crystal being collected on the beamline and Figure 4.5 shows one of the diffraction patterns recorded.
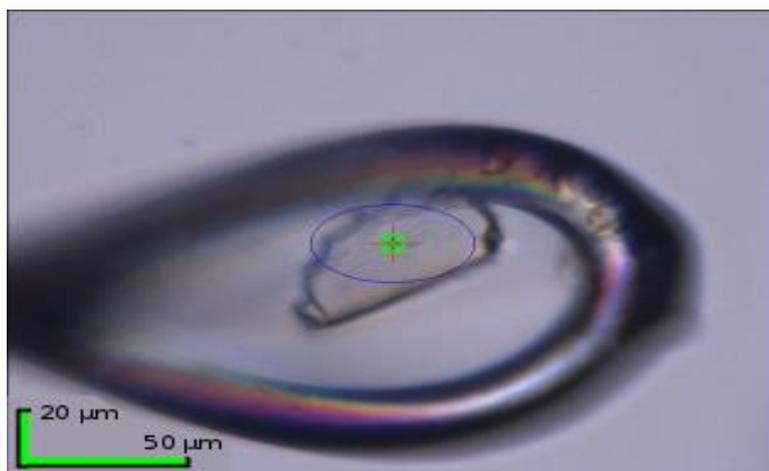
Figure 4.4: Photograph of the *Mo*-CBP3-4 crystal mounted in the cryostream on beamline ID29.

For cryoprotection individual crystals were briefly soaked in 2.75 M sodium formate, 0.1 M citric acid, pH 5.0 with 10% glycerol prior to mounting in the cryo-stream system. A helical data collection strategy was used to minimize the effects of radiation damage.
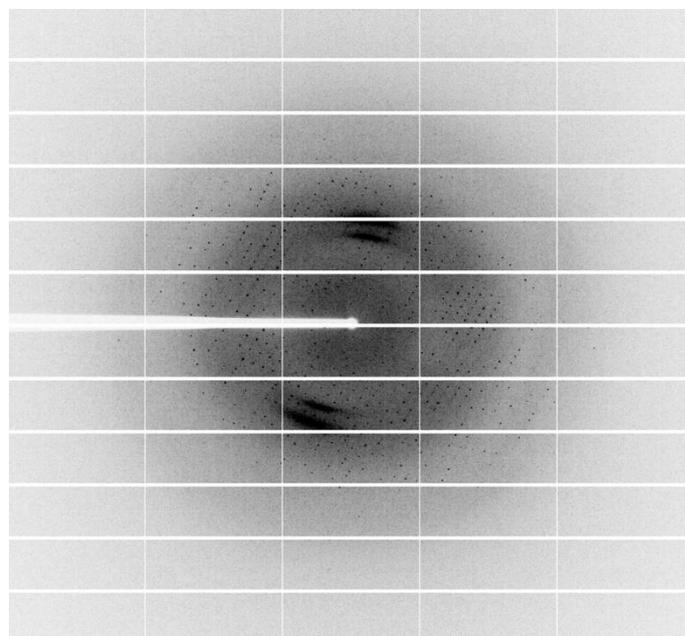
Figure 4.5: A diffraction pattern recorded from the *Mo*-CBP3-4 crystal using beamline ID19 at the ESRF.

Two sets of data were collected: a high resolution (HR) set and a sulphur SAD set. Both datasets obtained were of high quality and high resolution (1.68 Å for HR and to 2.13 Å for SAD). The data were indexed in space group I4$_1$22 with $a = b$ = 108.07 Å and c = 43.62 Å, consistent with the previously published crystal structure of Ullah *et al.,* (2015). The solvent content was 54%. Table 4.4 summarises the data collection statistics.

Data collection	High resolution	SAD
Type of radiation diffracted	Synchrotron X-ray	
Crystal volume (µm)	100 x 50 x 20	
Temperature (K)	100	
Space group	I4$_1$22	
Wavelength (Å)	0.99	1.91
Cell dimensions *a, b, c* (Å)	a=b=108.07 Å, c=43.62 Å	
Resolution (Å)	54.45-1.68	54.03-2.13
R$_{merge}$	0.035 (0.691)	0.054 (0.276)
R$_{pim}$	0.020 (0.411)	0.025 (0.235)
Mean $I/\sigma(I)$	24.7(2.8)	23.2 (2.2)
Wilson B-factor (Å^2)	27.4	29.0
Completeness (%)	99.6 (95.5)	96.4 (68.2)
Redundancy	7.1 (7.1)	9.0 (2.6)
Number observations	106878 (5225)	65531 (1092)
Number unique reflections	15156 (741)	7243 (413)
Anomalous completeness	98.5 (93.3)	88.3 (37.1)
Anomalous multiplicity	3.4 (3.4)	4.7 (1.7)

Table 4.4: Statistics for the data collected from the *Mo*-CBP3-4 crystal on ID29. These show the high quality of the data as seen from the low R values, high completeness, redundancy and resolution.

The XDS program package (Kabsch, 2010) was used to integrate and scale the data. The structure solution was determined by the sulphur SAD method using the *SHELX* program suite (Sheldrick, 2008). The positions of the anomalous scatterers were found using *SHELXD* and the initial phase information was determined using *SHELXE*. The final phases that were obtained were then used by the automated model-building routines in ARP/*w*ARP *7.3* (Langer *et al.,* 2008) to construct an atomic model.

4-4 Crystal structure

4-4-1 Refinement

Refinement was performed using REFMAC*5* (Vagin *et al.*, 2004) with the high resolution data set. Table 4.5 summarizes the data statistics of the refinement. The values of R_{merge} (0.035) completeness (99.6%), R_{free} and R_{work} (0.23/0.20) are good indicators of the high quality of the data.

Refinement	
No. of reflections	13557
R_{work}/R_{free}	0.20/0.23
R.M.S deviations bonds (Å)	0.0166
R.M.S deviations angles (°)	1.522

R_{merge} = (Σ(I - <I>) / Σ(I); where I is the intensity measured for a given reflection, <I> is the average intensity taken over the multiple measurements of this reflection.
R_{pim} = ($\Sigma[1/(N - 1)]^{1/2} \Sigma$|I - <I>|) / Σ(I).
R_{work} = Σ||F_{obs}|-|F_{calc}|| / Σ|F_{obs}|; where F_{obs} and F_{calc} are the observed and calculated structure factor amplitudes, respectively, for the 95 % of the reflection data used in refinement.
R_{free} = Σ||F_{obs}|-|F_{calc}|| / Σ|F_{obs}|; for the 5% of the reflection data excluded during the refinement.

Table 4.5: Refinement statistics for the *Mo*-CBP3-4 structure.

Water molecules were modelled using *COOT* (Emsley & Cowtan, 2004) resulting in a final model of 90 amino acid residues, 48 solvent water molecules, 1 glycerol, 1 formic acid and 1 chloride ion. The reflection data and PDB will be deposited in the PDB databank.

4-4-2 The structure analysis

Most of the backbone of the protein was readily visible from the experimentally determined electron density map, allowing unambiguous identification of 80 residues without resorting to any sequence information. In spite of the correlation with results from N-terminal sequencing in combination with trypsin digestion, it was not possible to fit the whole sequence as a result of to the heterogeneity of the protein preparation from natural sources. The publication of Freire and co-workers in 2015 reporting the *Mo*-CBP3 isoform sequences, was an important breakthrough in completing the whole model structure.
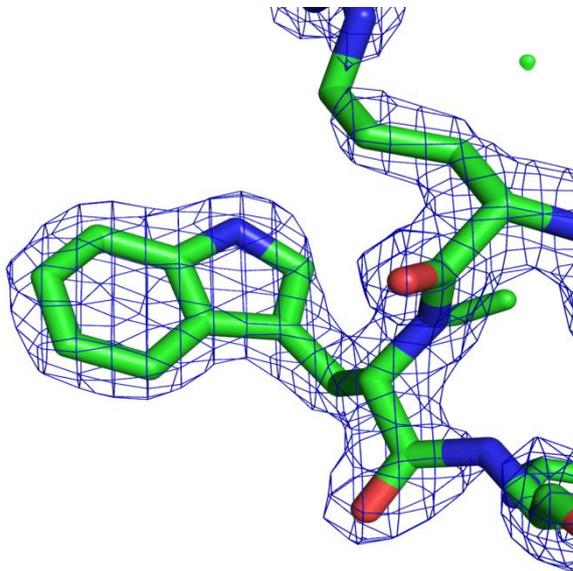
Figure 4.6: Electron density of *Mo*-CBP3-4 showing the quality of the map and the identification of the only tryptophan (residue 23) of the short chain present in the structure model.

Figure 4.6 shows the presence of a tryptophan residue which was easily identified due to the

high quality of the density map. Its presence was seen only in 2 isoforms of *Mo*-CBP3 protein

(Freire *et al.*, 2015): *Mo*-CBP3-1 and *Mo*-CBP3-4. The MS analysis of the protein in the crystal

allowed the identification of the isoform as *Mo*-CBP3-4
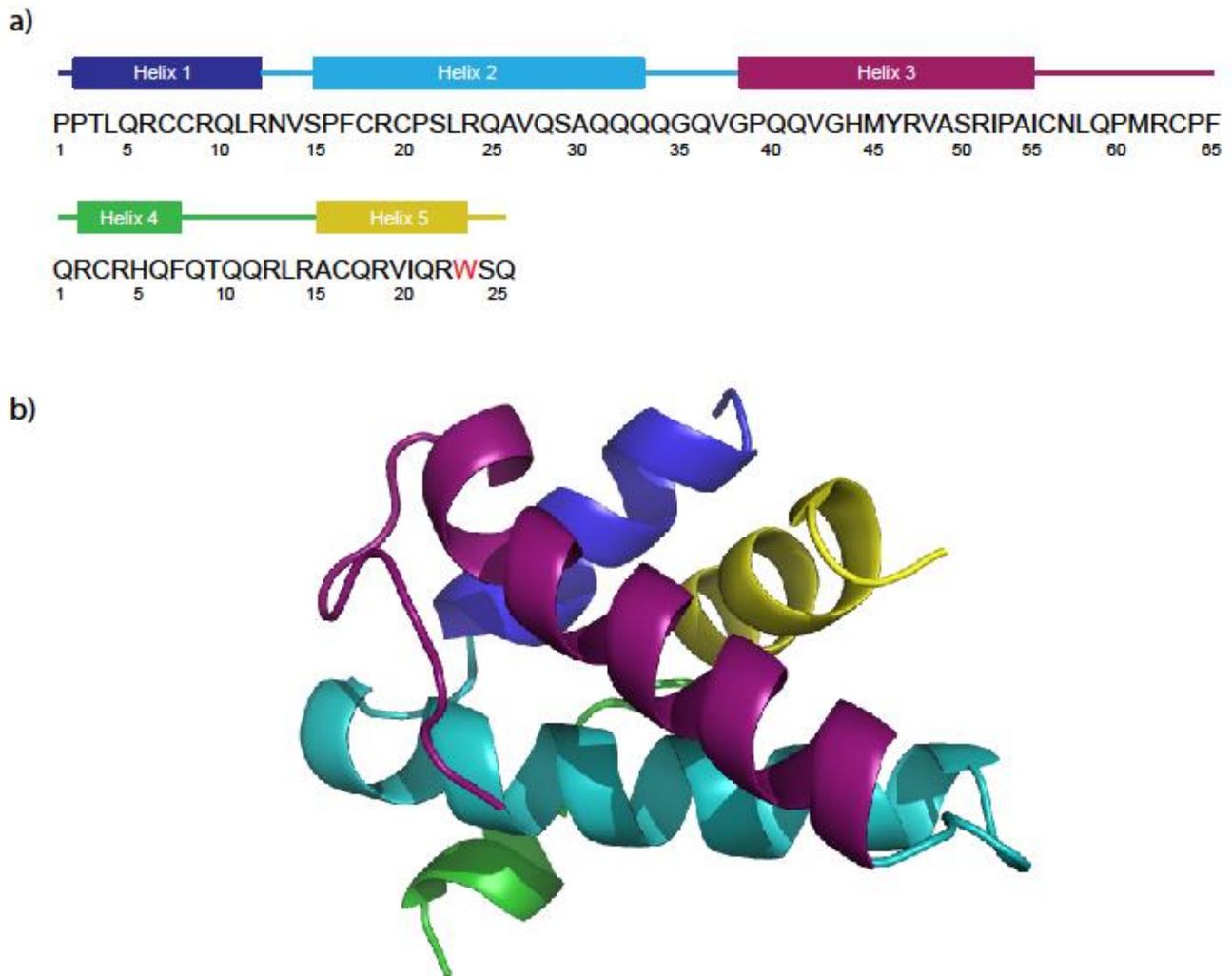
Figure 4.7: Overall structure of mature *Mo*-CBP3-4 **a)** Amino-acid sequence of *Mo*-CBP3-4 along with the secondary structure assignment based on the crystal structure highlighting the presence of the tryptophan. **b)** Cartoon representation of the crystallographic structure. Each of the 5 helices are represented with different colours.

The scheme a) in Figure 4.7 sheds light on the presence of the 8 CM spaced as follow

…C…C…/…CC…CXC…C…C… This is in common with other 2S albumin proteins. This typical

feature renders the structure very compact, heat stable and resistant to proteolysis.
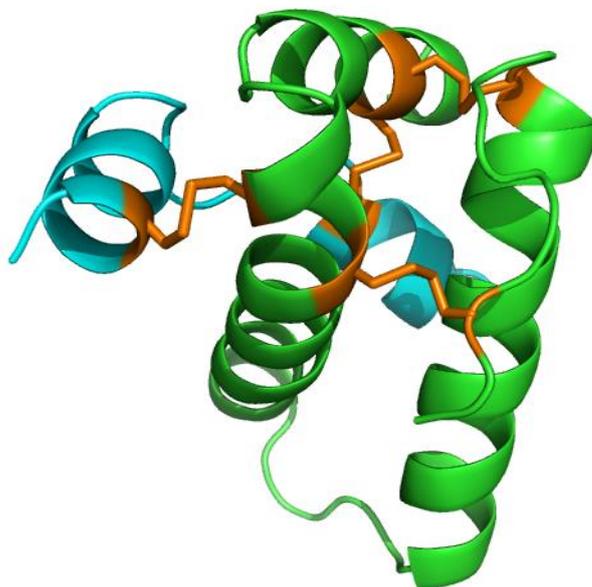
Figure 4.8: Representation of the crystallographic structure of *Mo*-CBP3-4. The long chain and the short chain are respectively represented in green and cyan. The disulphide bridges are highlighted in orange.

The molecular structure consists of 5 alpha helices (Figure 4.7 b)), forming two chains, each

having an intra-chain disulphide bond. In addition, there are two inter-chain disulphide bonds

(Figure 4.8). The heavy chain and light chains contain respectively 65 and 25 amino acids.

a)   **b)**
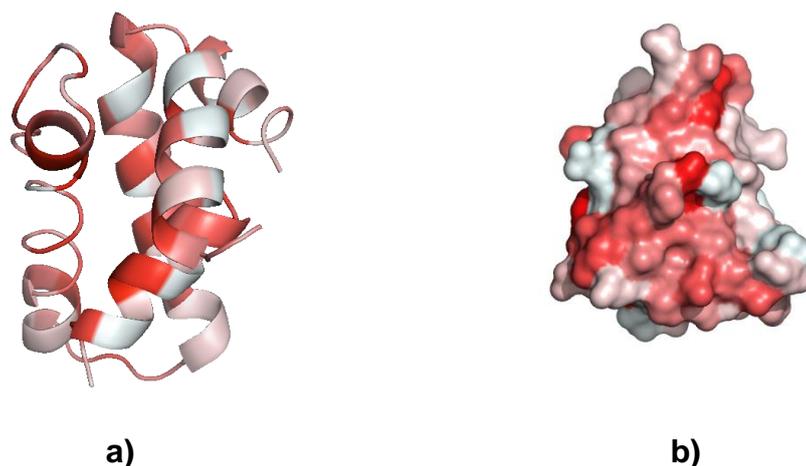
Figure 4.9: Diagram showing the crystal structure of *Mo*-CPB3-4 helix **a)** and surface representation **b)** showing the polar versus non-polar areas of the protein. Red represents the most hydrophobic and white the most hydrophilic regions according to the Eisenberg hydrophobicity scale (Eisenberg *et al.*, 1984)

Figure 4.9 represents the crystal structure, highlighting the hydrophobic residues (in red).

While the core of the protein appears to be mostly hydrophobic, the solvent-exposed region,

which contains a large number of arginine residues (10 out of 90 residues) is highly hydrophilic,

as seen in Figure 4.10.
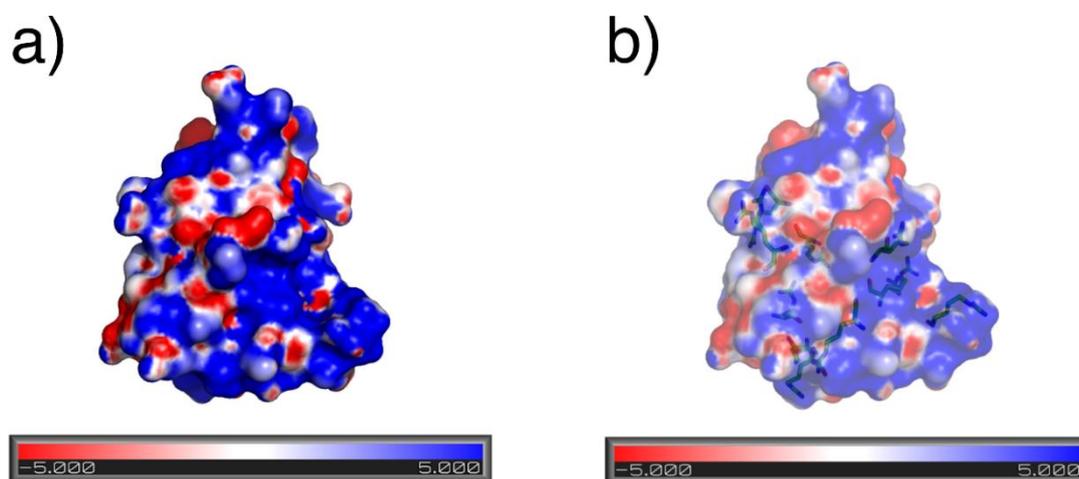
Figure 4.10: Diagrams showing the surface charge distribution of *Mo*-CBP3-4.
a) Surface charge distribution of *Mo*-CBP3-4 where red, white and blue represent respectively negative, neutral and positive charges.
b) Arginine residues highlighted

The surface charge distribution analysis indicates that *Mo*-CBP3-4 is positively charged and the arginines, considered important for the flocculating and antibacterial activities are exposed to the solvent. The surface charge distribution of the protein (Figure 4.10) confirms its high net positive charge, as well as the existence of both cationic and anionic residues. This is consistent with the high calculated isoelectric point of the protein (11.8). This surface charge distribution is similar as the cationic peptides that often displaying antimicrobial activity by its interaction with negatively charged microbial surfaces and its amphiphilic structure allowing the incorporation into cellular membranes

4-4-3 Comparison with other 2S albumin proteins structure and flocculating proteins.

This X-ray crystallographic structure of the mature form of *Mo*-CBP3-4 can be compared with the structure published by Ullah and collaborators (mistakenly designated *Mo*-CBP3-1), but also with the mabinlin II which was the only 2 S albumin crystal structure available since this study was carried out. Moreover, other proteins from *Mo* seeds display flocculating properties and the structure provides a model comparison for their surface charge distribution.

Comparison with the structure published by Ullah and co-authors

As mentioned previously, Ullah and co-workers in 2015 have published a "*Mo*-CBP3-1"structure; the isoforms *Mo*-CBP3-1 and *Mo*-CBP3-4 have a sequence difference of only two amino acids, as reported by Freire and co-workers (2015) who have characterised the 4 isoforms of *Mo*-CBP3 (*Mo*-CBP3-1, *Mo*-CBP3-2, *Mo*-CBP3-3, *Mo*-CBP3-4). The *Mo*-CBP3-4 sequence contains a threonine instead of a serine (small chain) and a leucine instead of proline (long chain).
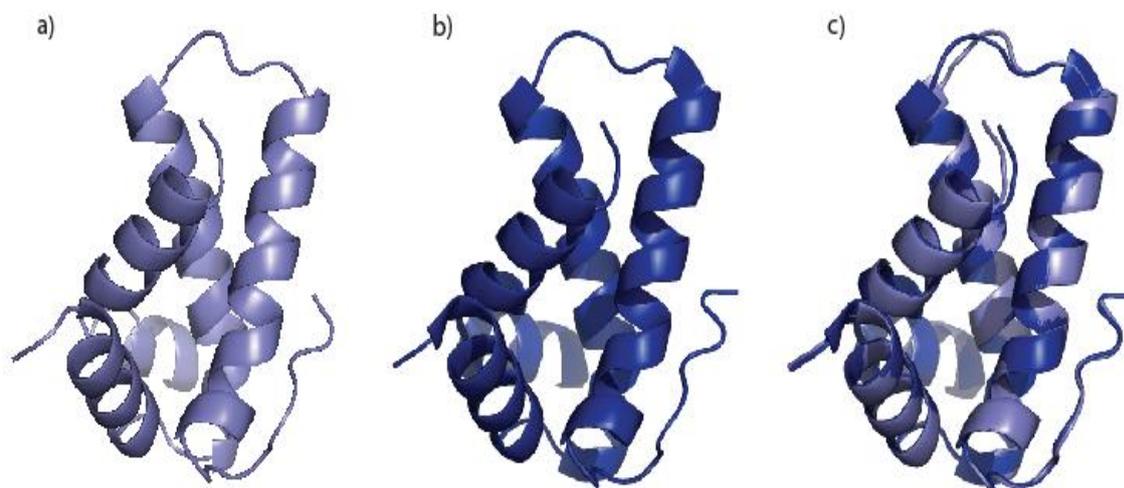
Figure 4.11: Ribbon diagrams showing the *Mo*-CBP3-4 structure a) determined in this thesis work (light purple) b) the structure published by Ullah *et al.*(2015) (dark purple) and c) overlay structure of both *Mo*-CBP3-4 structures. The Root Mean Square Deviation (RMSD) between the two structures was 0.18Å after rejection of outliers.

The comparison with the published structure (PDB code: 5DOM) showed a Root Mean Square

Deviation (RMSD) of 2.37Å using pymol align option with all atoms aligned. Rejection of the

outliers yielded a 0.182 Å RMSD with 384 out of 470 atoms. These results confirm the similarity

of the two structures. (The RMSD is the square root of the mean of the square of the distances

between the matched atoms).

Comparison of 2S albumin structure and with the *Mo* flocculating proteins

The sequence alignment among *Mo*-CBP3-4 and 2S albumins from other plants indicates an

average identity of 40%. *Mo*-CBP3-4 displays high sequence identity (47%) with the sweet

protein Mabinlin II from *C. masaikai*. Both proteins contain four isoforms (Mabinlin I, II, III, IV)

(Nirasawa *et al.*, 1994) and only one of the isoform structures has been solved. The eight

cysteine residues forming four disulphide bridges are fully conserved among these proteins;

additionally leucine at position 51 and alanine at position 141 are also conserved. In the

Mabinlin II crystal structure, a sequence motif NLPNICNIPNI is encountered; this recognizes

the sweet taste cells hT1R2/T1R3 and produces the sensation of sweetness. In *Mo*-CBP3-4, this sequence is not conserved and is RIPAICNLQPM. Despite this high homology, Ullah *et al.*, (2015) performed a surface charge analysis demonstrating a real difference in the distribution with *Mo*-CBP3-4.
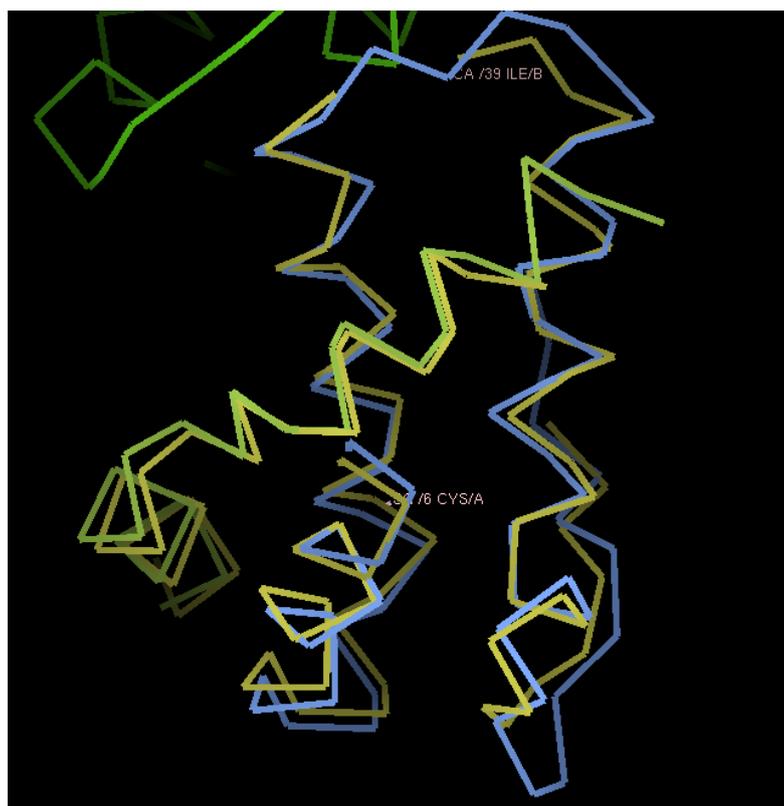
Figure 4.12: Comparison of Mabinlin II (blue) and *Mo*-CBP3-4 (yellow) models.

It should be noted that the first strategy envisaged to solve the phase problem of the unknown structure of *Mo*-CBP3-4 was to perform molecular replacement (MR) using these homology properties: the presence of two chains linked by 4 disulfide bridges and 8 cysteines, and the number of amino acid residues. Surprisingly, this approach was unsuccessful despite overall protein folds that were very similar (Figure 4.12). As noted above, the phases were eventually

determined experimentally by using the SAD method thanks to the presence of 8 cysteines in the protein (see Chapter 3).

The other solution structure known for 2S albumin includes proBnlb from rapeseed (PDB ID 1SM7), napin Bnlb (PDB ID 1PNB), Ber e1 from Brazil nut (PDB ID: 2LVF) and RicC3 from *Ricinus communis* (PDB ID 1PSY); these contain a linker peptide that was cleaved in *Mo*-CBP3-4 during the maturation by proteolysis. Ullah and co-workers reported that the surface charge of these 2S proteins from other plants are not the same as *Mo*-CBP3-4 because the structural motifs (arginines and glutamines) are not conserved.

In contrast, *Mo*-CBP-3-4 also shares high sequence identities (>72%) with MO2.1 and MO2.2 - other flocculating *Mo* proteins. Their proteins are small, basic, and have a single chain with molecular masses of approximately 6.5kDa. They contain cysteins residues but they are not stabilized in the same way as *Mo*-CBP3-4 that can form a total of four disulphide bridges. However, Ullah has shown that all the arginines and the glutamines considered important for flocculating activity are located on the surface and conserved among these proteins.

4-5 Discussion and conclusion

This chapter has described the crystal structure of *Mo*-CBP3-4, a mature form of a typical 2s albumin from *Mo* seeds determined at 1.68 Å. This structure provides a model for understanding the diversity of the structures of this large family of albumins, serves as a model for the structures of precursors, and indicates the structural basis for known flocculating behaviour. The structure of *Mo*-CBP3-4 is very similar to that described by Ullah and co-workers in 2015.

The establishment of crystallisation conditions was challenging for several reasons. The different crude extracts were a mixture of proteins with a composition that might depend in

part on naturally variable growth conditions and extraction methods. This heterogeneity of the three different batches tested may explain in part the various conditions of crystallisation obtained and the difficulty encountered in crystal reproducibility. On the other hand, the mechanism of maturation of the 2 S albumin storage proteins which are synthesised as precursors involved the cleavage into four fragments by the successive actions of different proteases such as endoproteases, endopeptidases and carboxypeptidase. This cleavage shows that the mature form of the protein is composed of two heterodimeric chains linked by disulphide bridges and also generates a complex mixture of N- and C-terminal processed species of various *Mo* seed protein isoforms (Chapter 3). Despite several steps of purification, these proteolytic species are difficult to separate and the presence of *Mo*-CBP3-3 (characterised by MS/MS as described in Chapter 3) in the purified fraction could prevent the crystallisation of *Mo*-CBP3-4. However, the addition of 50 mM NaCl in the equilibration buffer during the SEC chromatography allowed the separation of both isoforms *Mo*-CBP3-3 and *Mo*-CBP3-4 and should allow the crystallisation of a second isoform of *Mo*-CBP3.

The MS analysis of the crystal protein confirmed the presence of these N and C terminal processed species, and allowed the determination of the accurate mass of *Mo*-CBP3-4. The mass in the non-reduced conditions is 11824Da and 8045 and 3776 Da respectively for the small and long chains after treatment with TCEP. These data were unknown in the beginning of the building of the model structure.

The high resolution model of the structure allowed the determination of a substantial part of the amino acid sequence (approximately 80 %) and a lot of biochemical and biophysical characterisation had been achieved prior to the publication of the model by Ullah and co-workers. The protein structure shows a compact fold due to the presence of a small chain and

a large chain linked together by four disulphide bonds. This arrangement is the key to the extreme stability of the protein as well as its heat- and proteolysis resistance. This structural stability has been also been characterised by circular dichroism (Chapter 3), showing that there is no effect of heat treatment on the secondary structure. Moreover, the charge distribution showed a contrast between the core of the protein which is hydrophobic with the solvent-exposed region, which contains a large number of arginine residues (10 out of 90 residues) that is highly hydrophilic and could be responsible for colloidal interactions in the flocculation process. Despite the fact that *Mo*-CBP3-4 possesses a fold similar to other 2S albumins, it differs from this family by having a highly positive molecular surface charge, as observed in flocculating proteins. This characteristic surface charge has been noted for cationic peptides displaying antimicrobial activity by interacting with negatively charged microbial surfaces and its amphiphilic structure, allowing incorporation into cellular membranes. In the water treatment process, usually a combination of polymers and cationic coagulant are necessary. The polymers are used as flocculating agent for the formation of bridges between the flocs (clumps of bacteria and impurities which form clusters). A cationic coagulant helps in the neutralisation of the suspended particles. The positive charge distribution of *Mo*-CBP3 may act as a cationic coagulant. In contrast to polymers, its structure is globular. The organisation of this protein in solution is not well understood The development of a recombinant expression system with a large scale production of perdeuterated protein could be useful for further detailed characterisation using neutron diffraction. For instance, this could be used to produce *Mo*-CBP3-4 crystals suitable for a neutron crystallographic analysis aimed at understanding the nature of hydration interactions and protonation states in the protein structure. It also could be interesting to produce and

determine the structures of the different isoforms which were not easily separable from their

natural source using usual biochemical techniques.

5-Reflectometry studies on *Mo*-CBP3 isoforms

This chapter describes the adsorption of *Mo*-CBP3 isoforms to silica and sapphire interfaces as part of an initiative to elucidate the mechanism of its flocculation properties. Neutron reflection (NR) was used to determine the structure and composition of interfacial layers at the solid/solution interface. The results show a clear interaction of the protein with the silica interface and an essentially negligible interaction with the alumina surface. The interfacial behaviour at the silica surface occurs as a uniform protein monolayer having a thickness of 15.3 ± 1 Å, and a surface excess (1.3 ± 0.2 mg m^{-2}). The binding mechanism is linked to the positively charged nature of the purified fraction and its molecular arrangement. The results are consistent with the crystallographic structure described in Chapter 4 which exhibits positively charged groups on its surface. These reflectometry measurements have been related to those from the crude extract (CE) data reported previously by collaborators. It is believed that some of the binding mechanisms associated with the CE may depend on a mixture of proteins being present.

5-1 Introduction

Protein adsorption to surfaces is common in many biological and industrial processes. A knowledge of the mechanism of adsorption and the structure of the adsorbed protein is important in many areas relevant to biology, medicine, food processing and biotechnology. The adsorption phenomena for proteins include a number of interactions at solid-liquid interfaces. To understand the interaction between protein and solid surfaces, it is important to consider many factors such as the physical and chemical properties of solid surfaces (roughness and chemical dissociation) and proteins (folding, isoelectric point, and buffer

conditions). This chapter describes a study of the adsorption from dilute aqueous solution of

Mo-CBP3 isoforms (characterized in Chapter 3) to aluminium oxide (sapphire) and silicon oxide

surfaces. Numerous studies have reported the role of *Moringa* seeds or MO2.1 as an effective

flocculating agent in water purification (Ndabigensere *et al.*,1995; Okuda *et al.*, 2001; Sajidu

et al., 2006; Sarpong *et al.*, 2010; Madrona *et al.*, 2011; Broin *et al.*, 2002). Although the details

of the mechanism of the water purification properties are not yet well understood, there is

evidence that the role of proteins is directly related to adsorption, as described by Kwaambwa

et al., (2010) and Jerri *et al.*, (2012). Information on the amount of material that is adsorbed

at the surface, the structure of the adsorbed layer, and how this relates to concentration in

solution is very important in assessing the potential exploitation of this system for water

purification. To understand the adsorption mechanism, well defined interfaces are required

to avoid any ambiguity originating from the substrate surface. The binding of the *Moringa*

seed crude extract (CE) to silicon oxide (SiO_2) and to alumina has been investigated previously

using neutron reflection (NR) over a range of solution conditions with different concentrations

of proteins using an *in-situ* solid/liquid adsorption cell (Kwaambwa *et al*., 2010 and 2015). The

NR technique is widely used to study adsorption on flat solid substrates and at air–liquid

interfaces, and allows the determination of structural properties such as thickness, solvent

penetration, or roughness of adsorbed layers at interfaces at molecular length scales. The

previous results of Kwaambwa and co-workers have shown that the CE binds to silica and

alumina surfaces with a diffuse layer of protein that extends from a denser but hydrated layer

near the surface. However, on alumina, the CE is not bound strongly and can be displaced

whereas for the silica surface, the protein was found to be irreversibly bound. In this work,

the reflectometry study was designed to investigate the range of materials that will interact

with the *Mo*-CBP3 isoforms (fraction C1 as described in Chapter 3) to those of the CE previously studied. The reflection experiments carried out as part of this thesis used the D17 reflectometer (Cubitt & Fragneto, 2002) which is described in section 5-2 together with the sample cell used and the data analysis process. The results obtained for both mineral surfaces are reported in section 5-3 and are related to the CE data in section 5-4. An understanding as to why particular seeds or particular proteins are more effective than others as a flocculating agent will be important for the future development of applications from other plant varieties or from trees grown under different conditions.

5-2 Instruments used and experimental procedure and surfaces used

NR is widely used to study adsorption on flat solid substrates and at air–liquid interfaces; such experiments involve the determination of the reflectivity of an interface as a function of the wavelength or angle (Figure 5.1). The data allows quantitative structural and compositional information about the absorbed material to be obtained at molecular length scales. This technique is highly suitable for the study of the interaction of *Mo*-CBP3 isoforms with mineral surfaces.
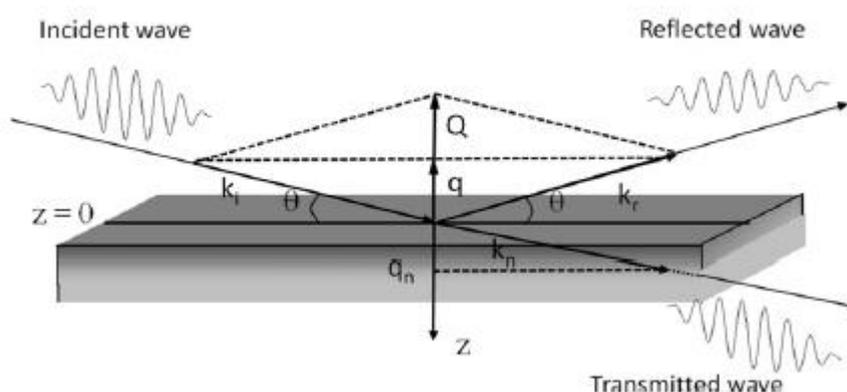
Figure 5.1: Reflection from an infinite planar surface (Cousin and Menelle, 2015). The scattering wave-vector Q is perpendicular to the plane of the thin film.

Neutron reflectivity, R(Q), is defined as the ratio of the intensity of the reflected beam to that of the incident beam and is usually presented as a function of the momentum transfer, Q, perpendicular to the reflecting interface given as Q $(=(4\pi/\lambda) \sin \theta)$ where θ is the incidence angle that is equal to the angle of reflection and λ the wavelength of the incident neutron beam. The scattering length density (ρ) (SLD) that governs the neutron refractive index is determined by the chemical composition of the sample (Table 5.1).

Coherent scattering length b (10^{-12} cm)							
^{1}H	D(^{2}H)	C	O	N	Si	P	S
-0.374	0.667	0.665	0.580	0.936	0.415	0.513	0.284
Scattering length density (Å^{-2})							
H_2O		D_2O	Sapphire	SiO_2		Si (crystal)	
-0.56 10^{-6}		6.38 10^{-6}	5.71 10^{-6}	3.41 10^{-6}		2.07 10^{-6}	

Table 5.1: Coherent scattering lengths of some atoms and scattering length densities of some molecules/substrates. (Adapted from Cousin and Menelle, 2015)

Because the proton (^{1}H) and the deuteron (^{2}H) have opposite signs of scattering lengths, deuterium substitution can be used in neutron reflection experiments to highlight different parts of the absorbed interfacial layers differently (Figure 5.2).
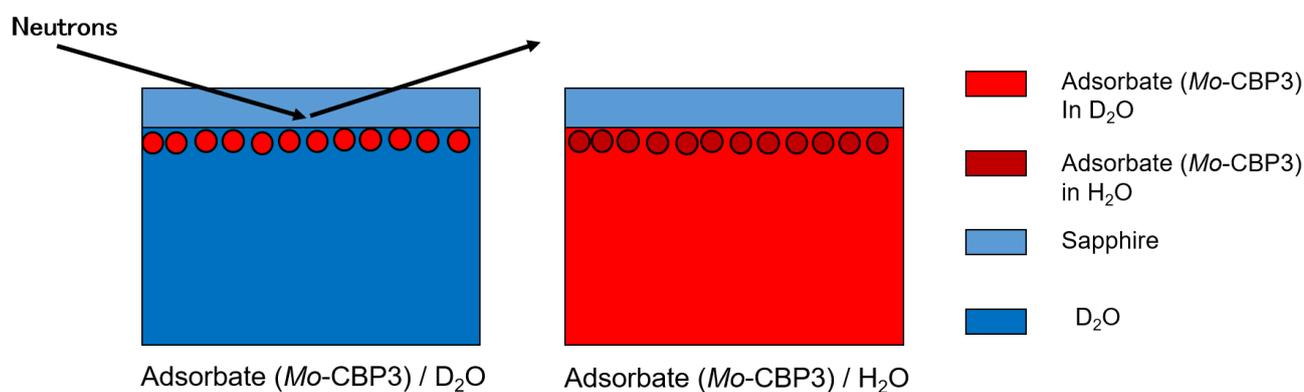
Figure 5.2: Representation of contrast matching in the solvent hydrogen/deuterium composition for a neutron reflection experiment.

Because D_2O has an SLD of 6.35×10^{-6} Å^{-2} and H_2O has an SLD of -0.56×10^{-6} Å^{-2}, adjustment of the ratio of D_2O to H_2O is often used to obtain different contrasts so that interfacial protein

layers are highlighted. Reflection experiments carried out as part of this thesis used the D17 reflectometer (Cubitt & Fragneto, 2002) at the Institut Laue-Langevin (ILL), Grenoble, France.

5-2-1 The D17 instrument at the Institut Laue-Langevin

D17 is a neutron reflectometer with horizontal scattering geometry (vertical surfaces) designed to be as flexible as possible in resolution and modes of operation. It is suitable for the study of the surface structures in solids and solid/liquid interfaces over a wide range of length scales.
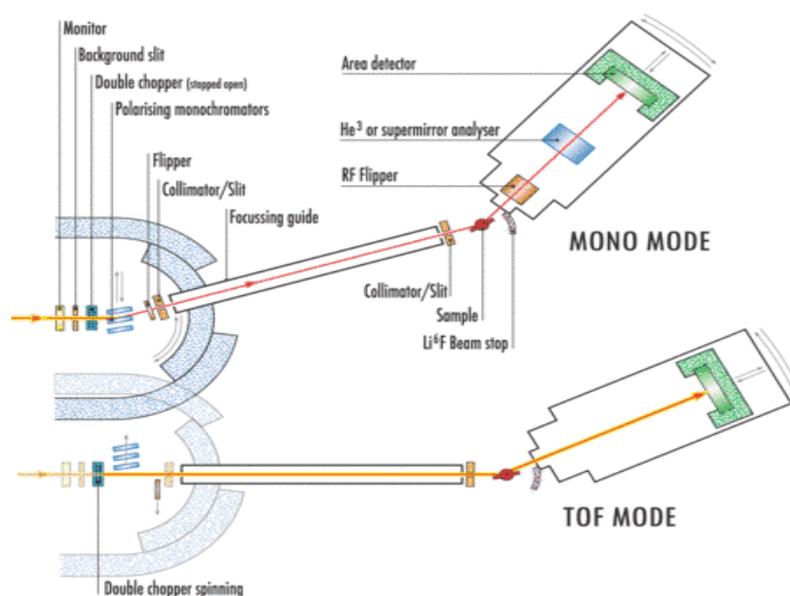
Figure 5.3: The D17 reflectometer at the ILL: Instrument layout for both monochromatic and time-of-flight modes of operation (https://www.ill.eu/instruments-support/instruments-groups/instruments/d17/description/instrument-layout/)

A reflectivity measurement consists in sending a neutron beam on a surface and varying the scattering wave-vector. This can be performed in two ways. The first possibility is to use a fixed wavelength λ (defined by a monochromator crystal as in the top diagram of Figure 5.3) and perform a standard θ-2θ scan, by rotating the sample by an angle θ and the detector by 2θ (Figure 5.3). The second possibility is to work at a fixed incident angle and to measure a range of incident wavelengths. The neutron wavelength can then be simply be measured by

the travel time between the chopper (a rotating disk with a small aperture allowing λ

selection) and the detection systems. This technique is called time-of-flight (ToF) (see lower

part of Figure 5.3). See Figure 5.4 for a photograph of the D17 instrument.

Figure 5.4: Photograph of the D17 instrument at the Institut Laue-Langevin (ILL) (Grenoble).

The time-of-flight mode on D17 was used to allow reflectivity data to be measured at just two

incident beam angles, θ, of 0.7 and 3.2 degrees, using wavelengths between 2.5 and 25 Å.

These parameters gave a useful range of momentum transfer, perpendicular to the interface,

from a Q (Q= $(4\pi/\lambda)$ sin θ) of 0.006 to a maximum of 0.25 Å^{-1}, depending on the signal-to-

noise. The neutron reflectivity curves R=f(Q) are traditionally represented on logarithmic-

logarithmic scales (Figure 5.5)
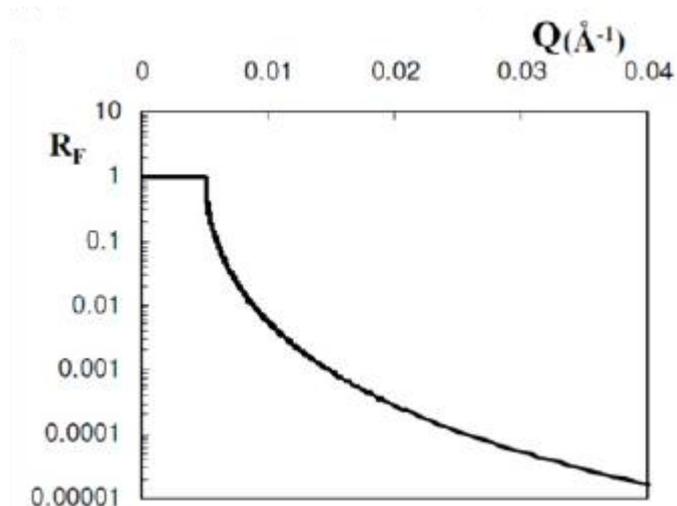
Figure 5.5: Logarithmic-logarithmic representation of the Fresnel reflectivity as function of Q. (Adapted from Cousin and Menelle, 2015).

Typical measurement times for each reflectivity curve were about 1 h. In order to make measurements with a small amount of sample at two different interfaces, a special sample cell was used that had a reservoir for solution formed by a PTFE gasket between two crystals of silicon and sapphire (Rennie *et al.*, 2015).

5-2-2 Sample cell unit

The sample cell used (Figure 5.6), was provided by Prof. A. Rennie (Uppsala) and was designed specifically to study solid/liquid interfaces by NR (Rennie *et al.*, 2015).
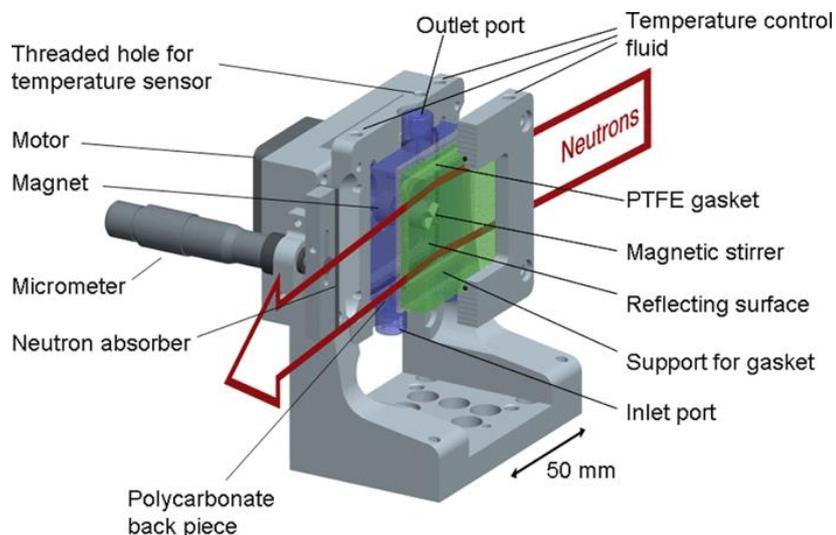
Figure 5.6: Diagram of the sample cell used on D17, showing a cutaway view of the holder and the substrate, liquids, and gasket. (Rennie *et al.,* 2015)

This cell has many advantages over other designs (see below) and allows rapid sample exchange either for measurements with different contrasts (H_2O/D_2O), the change of the concentration of the adsorbate, the modification of sample conditions/different chemicals, or pH. The sample holder could be rotated by 180 degrees and translated so that either the silicon/silicon oxide surface, or the alumina surface could be used for reflection measurements (Figure 5.7).
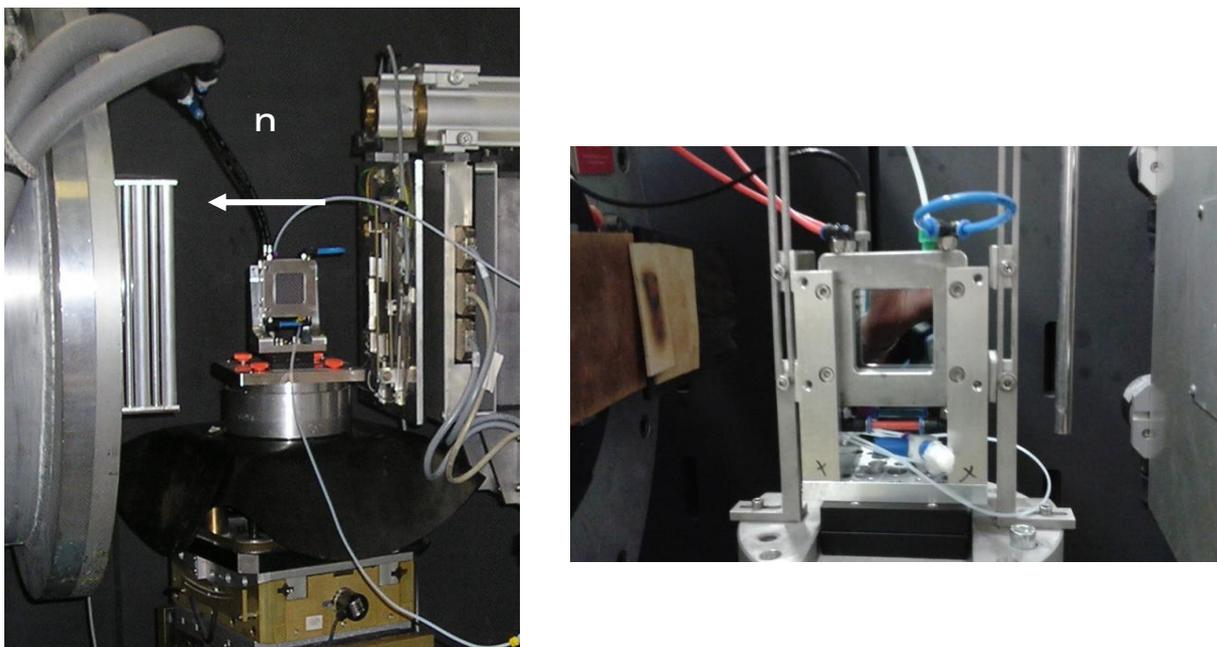
Figure 5.7: Photograph of the sample cell on the D17 instrument at the Institut Laue-Langevin (ILL).

The aqueous solvent can be exchanged automatically using a Knauer HPLC pump connected to the filling port. In order to minimize the volumes of protein solution, the sample solutions were injected manually using a syringe in conjunction with a three-way valve at the bottom of the sample cell. Initial measurements were made using pure D_2O and pure H_2O to characterize the oxide layer on the silicon crystal and any gel layer or roughness at the alumina surface. When changing contrast, at least 20 ml of solvent was flushed through the sample holder to an exit tube at the top of the cell. It was sufficient to inject 4 ml of purified protein at various concentrations to fill the sample cell or to increase the concentration.

Prior to use, the sapphire and silicon oxide surfaces were cleaned by addition of few drops of concentrated sulphuric acid until all the surfaces are covered. An equal volume of water is then added, allowing to the strong acid to heat. The surfaces are left under acid for 5 minutes and then rinsed copiously with pure water. The cleanliess is checked during the experiment

by measuring the interfaces in pure water, water matched to silicon and pure D_2O. The cleaning procedure is important and may affect the substrate surface chemistry. For the silicon oxide , three main types of surface group are present: silanol (Si-OH), siloxane (Si-O-Si) and silane (SiH). For instance, cleaning with strong oxidising acids seems to favour the formation of silanol groups whereas oxidation without an acid conduct mainly to the formation of siloxane group. These different groups give different properties of the surface. The other parts of the cell and connecting tubing were cleaned with Decon 90, followed by extensive rinsing with pure water.

5-2-3 Protein solutions

The measurements were made at 25°C. The sample solutions were obtained by directly dissolving the pure lyophilised protein in pure D_2O. The neutron reflectometry measurements were carried out on the *Moringa* fraction C1 as characterized in Chapter 3. The "purified fraction" contains mainly two isoforms of *Mo*-CBP3, *Mo*-CBP3-3, and *Mo*-CBP3-4. Stock solutions of 2 mg/ml of *Mo*-CBP3 isoforms in D_2O was used to make samples of 0.01, 0.05, 0.1, 0.25 and 0.5 mg/ml.

5-2-4 Surfaces

Measurements were made successively using, the sapphire, (Al_2O_3) and the silicon/silicon oxide (SiO_2) crystal surfaces. The sapphire crystal was cut so that measurements were made on a 001 face whereas the silicon crystal had a 111 face that was used. The latter has the advantage that the oxide growth is limited and so only a thin oxide layer is formed. This is less rough than surfaces that grow thicker oxides.

Crystals of sapphire (α-Al_2O_3) are convenient substrates for studies of adsorption to alumina because it has an advantageous combination of optical and mechanical properties. Sapphire

was chosen as a good model for a mineral surface that does not have strong negative charge in solution at a neutral pH. The isoelectric point (IEP) of alumina is between 6 to 8.5, and depends on the crystal plane (Franks *et al.,* 2007). This is close to neutral pH which makes it easy to achieve either a positively or negatively charged surface. SiO_2 layers on the surface of silicon crystals have been widely used with NR as a flat hydrophilic model substrate to assess the adsorption characteristics of surfactants, synthetic polymers or proteins with specific biological or medical interest such as those in membranes (Nylander *et al.*, 2008). The hydrophilic SiO_2 surface, covered with water, is a good mimic of common water impurity surfaces using NR because it is negatively charged (as suspended particles in raw water) due to the hydroxyl group. This substrate for surface adsorption studies has an isoelectric point of 2, making only a neutral or negatively charged surface available.

5-2-5 Model fitting and data analysis

Data were reduced to yield reflectivity curves by normalizing measurements of the incident spectra that were transmitted through the relevant crystal using the COSMOS software available at the instrument (Cosmos 2017, and Gutfreund *et al.,*2018). The program also subtracts the background that is observed on the instrument multidetector around the specularly reflected signal.

NR, in common with other scattering techniques such as small-angle scattering, cannot always yield a unique structure from direct inversion of the data. The reflectivity curve is instead fitted with a model scattering length density profile having appropriate layer compositions and thicknesses, along with the shape of interfacial mixing profiles between layers in the film. A model has first to be proposed and the parameters of this model must be varied until the reflectivity curve calculated from the model matches the experimental data. However, there

is generally more than one unique solution. One has to be very careful in modelling in excluding other plausible answers. All fits shown in this chapter were made using CPROF and the special software programs DRYDOC, LPROF and CPROF for polymers at interfaces by Rennie (http://www.reflectometry.net/fitprogs/cprof.htm).

Material	Chemical formula	Density (g cm^{-3})	Scattering length density (10^{-6} A^{-2})
Water	H_2O	0.997	-0.56
Heavy Water	D_2O	1.105	6.35
Alumina	Al_2O_3	3.98	5.71
Silicon	Si	2.33	2.07
Silicon oxide	SiO_2	1.88	3.41
Purified fraction C1 in H_2O		1.35	1.46
Purified fraction C1 in D_2O		1.36	2.6

Table 5.2: Properties of materials used in neutron reflection studies - chemical formulas, densities, and scattering length densities (SLD)

The amount of protein at the surface is calculated by summing the volume fraction in each layer. The volume fraction f is obtained from $f=(\rho_{lay}-\rho_{D2O})/(\rho_{prot}-\rho_{D2O})$ where ρ is the scattering length density (SLD). The calculations of SLD (summarized in the Table 5.2) were based on density of the CE reported by R. Maikokera (Ph.D. Thesis, Univ. Botswana, 2006) in aqueous solution as 1.35 g cm^{-3}, and on the amino acid analysis described in Chapter 3. For D_2O, the assumption was made that all exchangeable hydrogens would exchange with deuterium atoms of heavy water (*i.e.* OH and NH).

5-3 Results of reflection studies of *Mo*-CBP3 interfaces

To characterize each layer, reflectivity profiles were first measured in pure water with different isotopic contrasts (determined by the D_2O/ H_2O ratio), and refractive indices or (contrast) matched to silica or alumina in this order. D_2O was used to highlight the oxide layer so that its thickness and composition were well-determined. The results summarised in the following section are related to those obtained for the CE by collaborators few years ago (Kwaambwa *et al*., 2010 and 2015) and are important for comparison with the data recorded from the purified fraction of *Mo*-CBP3 isoforms.

5-3-1 Adsorption to silica surface

The data recorded for the purified fraction (*Mo*-CBP3 isoforms) are shown in Figure 5.8 and clearly demonstrate adsorption to the silica layer on the silicon substrate for a solution having a concentration as low as 0.025 mg/ml. These reflection data show a marked difference from those observed for the clean interface with a strong reduction in the reflectivity for the solution with protein in the Q range between 0.05 and 0.15 A^{-1}.
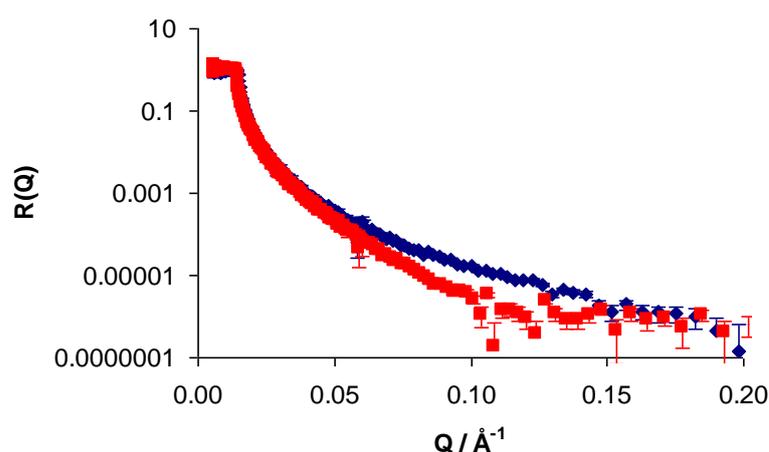
Figure 5.8: Neutron reflectivity data measured for the clean silicon/silica/D_2O interface (♦) and the surface in contact with 0.05 mg/ml solution of the *Mo*-CBP3 isoforms in D_2O (■).

In contrast to previous results on CE, there was no change in the reflectivity following increased protein concentration. This is illustrated in Figure 5.9 which shows little change in the reflectivity as the concentration is increased to 0.5 mg/ml. By comparison, the CE showed an increased adsorption until a plateau was reached at a concentration of 0.5 mg/ml.

a)CE   **b) *Mo*-CBP3 isoforms**
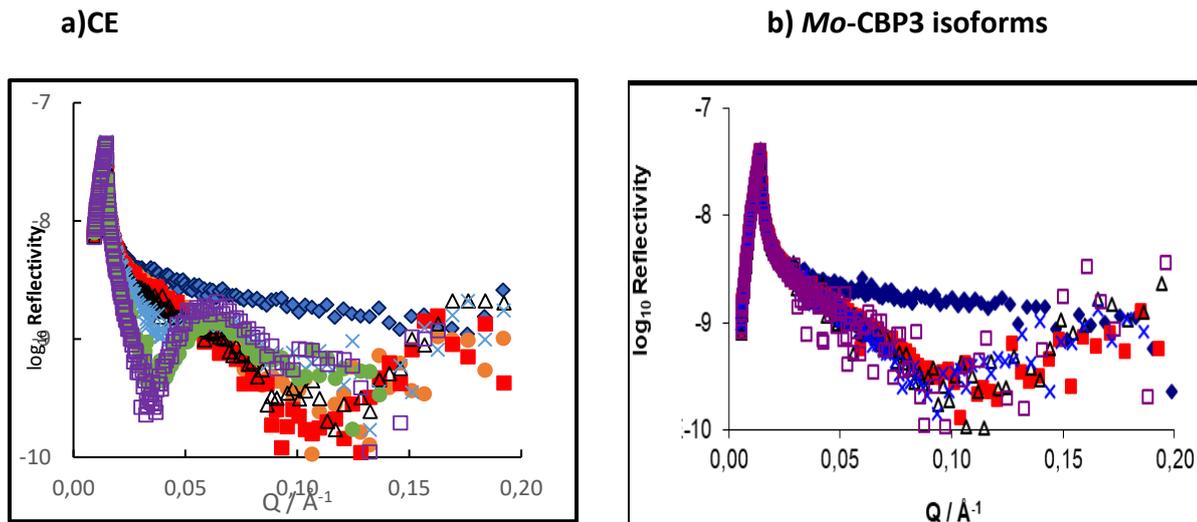
Figure 5.9: Comparison of reflectivity data between the CE and *Mo*-CBP3 isoforms on silica surface.
a) Reflectivity data obtained for different concentrations of crude extract (CE) from 0 to 0.5mg/ml) (● 0.01 mg/ml; ■ 0.025 mg/ml, Δ 0.05 mg/ml, × 0.1 mg/ml,● 0.25 mg/ml; □ 0.5 mg/ml); Data obtained from our collaborators (Kwaambwa *et al.*, 2010).
b) Reflectivity data obtained for different concentrations (0 to 0.5 mg/ml) of *Mo*-CBP3 isoforms (■ 0.025 mg/ml, Δ 0.05 mg/ml, × 0.1 mg/ml, □ 0.5 mg/ml) at the oxide layer on a silicon substrate. For comparison, the data for the clean silicon/silica substrate are also shown (♦).

The clear change in reflectivity on adsorption of protein that is apparent in the data shown in Figures 5.8 and 5.9 can be interpreted quantitatively using optical matrix calculations of the reflectivity as implemented in the CPROF program (Rennie, 2015). The plot of RQ^4 *vs.* Q allows some features in the data to be identified given that the effect of the large contrast difference between silicon and D_2O is diminished. The minimum in the reflectivity at about Q = 0.12 Å^{-1}

indicates a well-defined layer. Rinsing the surface with D_2O after adsorption did not change the reflectivity (Figure 5.10), showing that the adsorbed protein is not displaced with pure water.
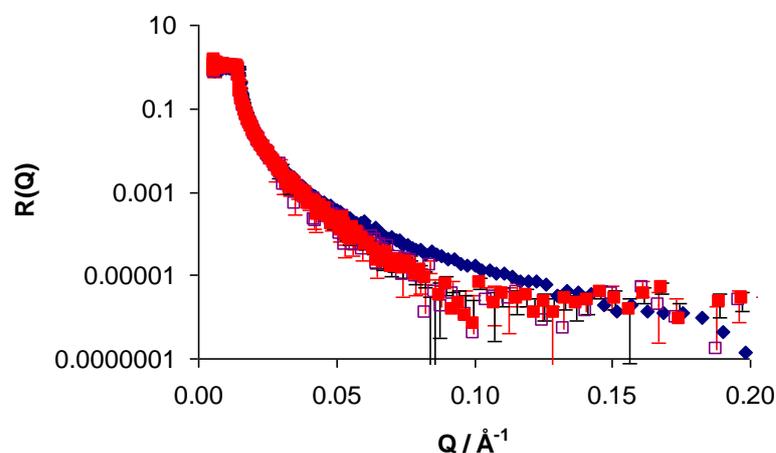
Figure 5.10: The reflectivity data of *Mo*-CBP3 isoforms showing that rinsing with 25 mL D_2O does not displace the adsorbed layer of protein at the silica interface. The data for the measurement with the solution in D_2O □, and after rinsing, ■, are within uncertainty the same, and are clearly different to the data measured with pure D_2O prior to adsorption, ♦.

Measurements were also made with a second solvent contrast, providing more details about the interfacial layer. The fit made to the combined data for the bound layer after rinsing first with D_2O and then H_2O is shown in Figure 5.11. The fitted curves (solid lines) correspond to a layer of 15.3 ± 1 Å that consists of 53 ± 3 % protein and 47 ± 3 % water. In the fit to the data, allowance was made for exchange of protons in the protein as well as the solvent contrast. The data are adequately modelled with a uniform protein layer having just 2 - 3 Å roughness at each of the interfaces.
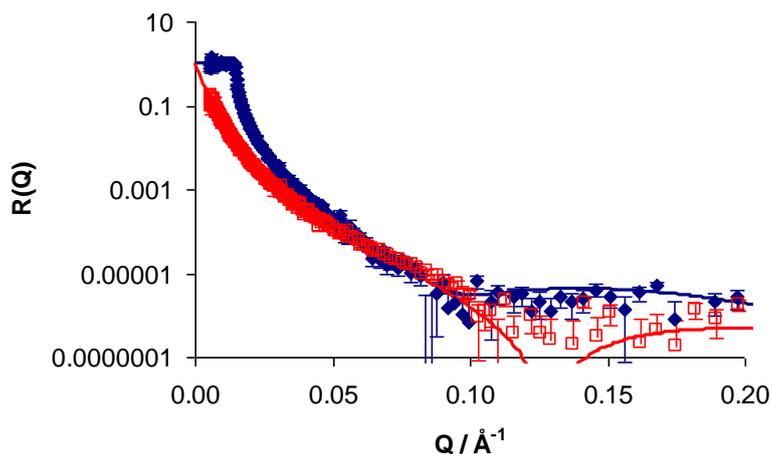
Figure 5.11: Reflection data for adsorbed Mo-CBP3 layer after rinsing, measured in D_2O (♦) and H_2O (□), shown with the curves for the model described in the text (1.3 mg m^{-2}, 47% water, thickness 15 Å).

The surface coverage of the protein can be calculated from the fit parameters as the product

of the volume fraction and the thickness of the layer, and corresponded to 1.3 ± 0.2 mg.m^{-2}

(Figure 5.11). The thin uniform layer that is hydrated could correspond to a single molecular

layer of protein, and contrasts with the multilayer adsorption that was found for the CE.

5-3-2 Adsorption to alumina interface

The design of the sample holder (Rennie *et al.*, 2015) allowed measurements to be made with

the same protein solution using the alumina/solution interface by simply rotating the cell so

as to reflect from a second solid/liquid interface.

a) CE **b) *Mo*-CBP3 isoforms**
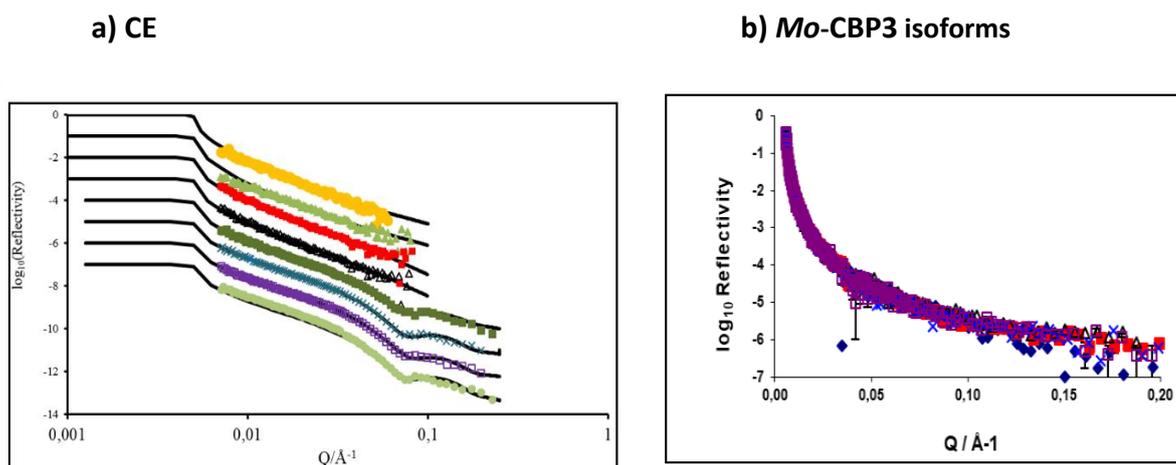
Figure 5.12: Comparison of reflectivity data between the crude extract (CE) and *Mo*-CBP3 isoform on alumina surface.
a) Reflectivity data for different concentrations of CE from 0 to 2mg/ml shows an interaction on the surface;(⬤ 0.01mg/ml;△ 0.025mg/ml; ■ 0.05 mg/ml, Δ 0.1 mg/ml; ▢ 0.25; × 0.5mg/ml, ▢ 1 mg/ml;⬤ 2 mg/ml). Data obtained from our collaborators (Kwaambwa *et al.*, 2015).
b) The neutron reflectivity data for the solution/alumina interface shows no significant adsorption as the protein of *Mo*-CBP3 isoforms concentration is increased. ◆ pure D$_2$O, ■ 0.05 mg/ml, Δ 0.1 mg/ml, × 0.5mg/ml, □ 1 mg/ml.

The data shown in Figure 5.12 indicate that there is no significant change in the reflectivity and this demonstrates that adsorption did not occur at this surface. This contrasts with the results found previously for the CE of *Moringa* seed protein (Kwaambwa *et al.*, 2015) that had shown adsorption to alumina, albeit with a lower plateau coverage than observed at the silica interface.

5-4 Discussion and conclusion

The reflectometry measurements have determined the layer structure at silica and alumina interfaces with water and they demonstrated significantly different interfacial properties between the CE and *Mo*-CBP3 isoforms (main components of the purified fraction). On the one hand, on the silicon oxide surface, the interfacial behaviour of the purified fraction occurs as a protein monolayer, whereas that for the CE it appears to incorporate additional less defined layers beyond the immediate surface (Figure 5.13).

a) Crude extract

b) *Mo*-CBP3 purified fraction

Thickness 60 Å

Thickness 15 Å
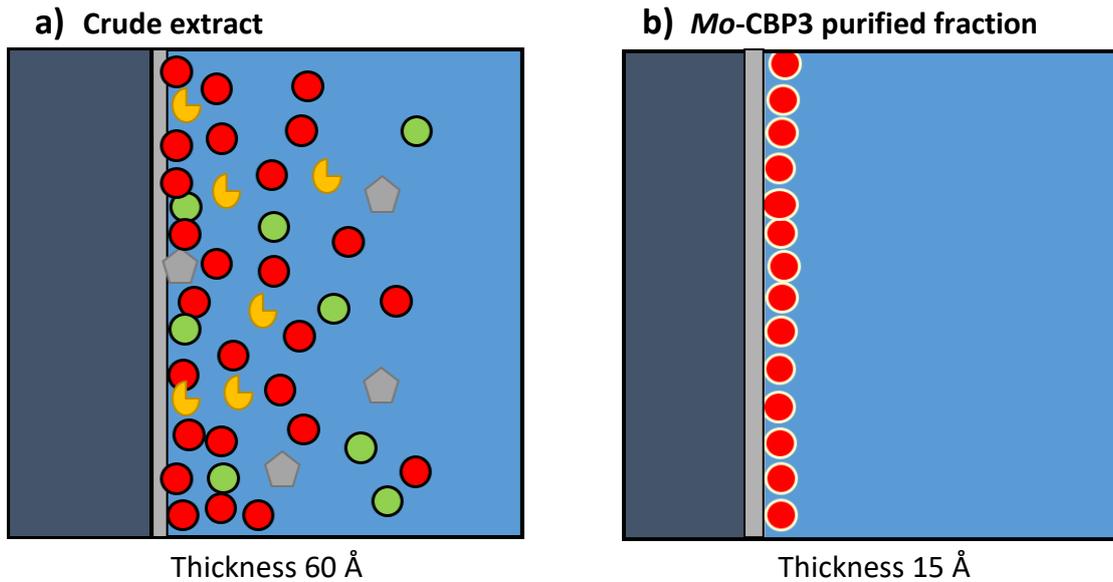
Figure 5.13: Model representation of the crude extract (CE) versus *Mo*-CBP3 isoforms adsorbed layer on Silica (SiO$_2$) Different seed proteins are represented as symbol (● red spot could be *Mo*-CBP3 isoforms,○ green spot,◖ yellow moon, and ⬠grey octagon as non-identified seed proteins).
a) For CE, an increase of adsorption (up to 0.5mg/ml) was observed as a diffuse layer having a thickness of 60 Å. The dense layer of the CE was about 5.5 mg m^{-2}
b) For *Mo*-CBP3 isoforms a clear adsorption was obtained as a well-defined layer having a thickness of 15 Å and a density of layer about 1.3 ± 0.2 mg m^{-2}. This adsorption was observed with a low concentration at 0.025 mg/ml.

For the sapphire surface, the data recorded for the purified protein indicate that there is no significant change in the reflectivity and no change in the absorption as the concentration is increased beyond 0.5 mg/ml. This means that adsorption was not occurring at this surface whereas the CE showed an increasing adsorption until a plateau amount was reached at a concentration of 0.5 mg/ml. These observations can be understood in terms of the positively charged nature of the purified fraction, and are consistent with the crystal structure of *Mo*-CBP3-4, the major component of the analyzed fraction. It is interesting to consider the crystal structure alongside the composition and arrangement of the bound protein on the solid interfaces. The hydration of the crystal structure described in Chapter 4, is about 54% very

similar to the 47% identified in the protein layer at the silica interface. The interfacial behavior of the purified fraction occurs as a protein monolayer, whereas that for the CE appears to incorporate additional diffuse layers beyond the immediate surface, and is apparently a consequence of the presence of a mixture of different molecules (Kwaambwa *et al.*, 2010 and 2015).

The molecular arrangement of the protein in the crystal, at neutral pH, suggests a fairly uniform distribution of positively charged groups (Figure 5.14 from Chapter 4); given the stability of the protein and its resistance to denaturation and degradation, the distribution of polar groups is likely to be similar in free solution.

a)   **b)**
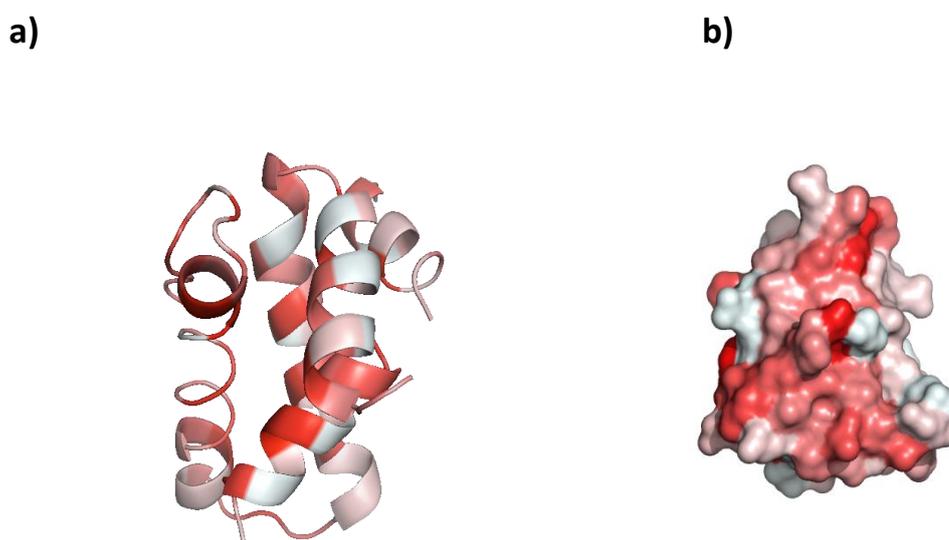
Figure 5.14: Schematic showing the crystal structure of *Mo*-CPB3-4 (helix) **a)** and surface representation **b))** showing the polar versus non-polar areas of the protein. Red represents the most hydrophobic and white the most hydrophilic regions according to the Eisenberg hydrophobicity scale (Eisenberg et *al*., 1984)

This may explain why the purified fraction does not associate as multilayers at surfaces, as observed for the CE. It may also explain the lack of adsorption on the uncharged alumina surface at neutral pH, in contrast to the clear binding to negatively charged silica. These observations imply that the valuable range of interactions that causes the *Moringa* seed

proteins to flocculate a wide variety of materials may depend on a mixture of proteins being present. The NR results may be of general significance for an understanding of binding mechanisms relevant to applications involving selective separation of different particles in aqueous dispersions. A NR experiment playing with different changes in ionic strength could be envisaged and would be very relevant. The knowledge of the amount of the different materials that are bound to the interfaces and to particles removed from water is important. The optimization of the amount of material used would depend on its composition. Figure 5.15 shows a model of mechanism of coagulation, based on the NR experiment, where CE might interact less selectively with particles present in dirty water in a) than the purified fraction in b). An excess of CE in solution could release organic matter which may facilitate bacterial growth especially in hot climates as observed with the crushed seeds. Therefore the use of a semi-pure or pure form of the protein may offer advantages in comparison with CE.
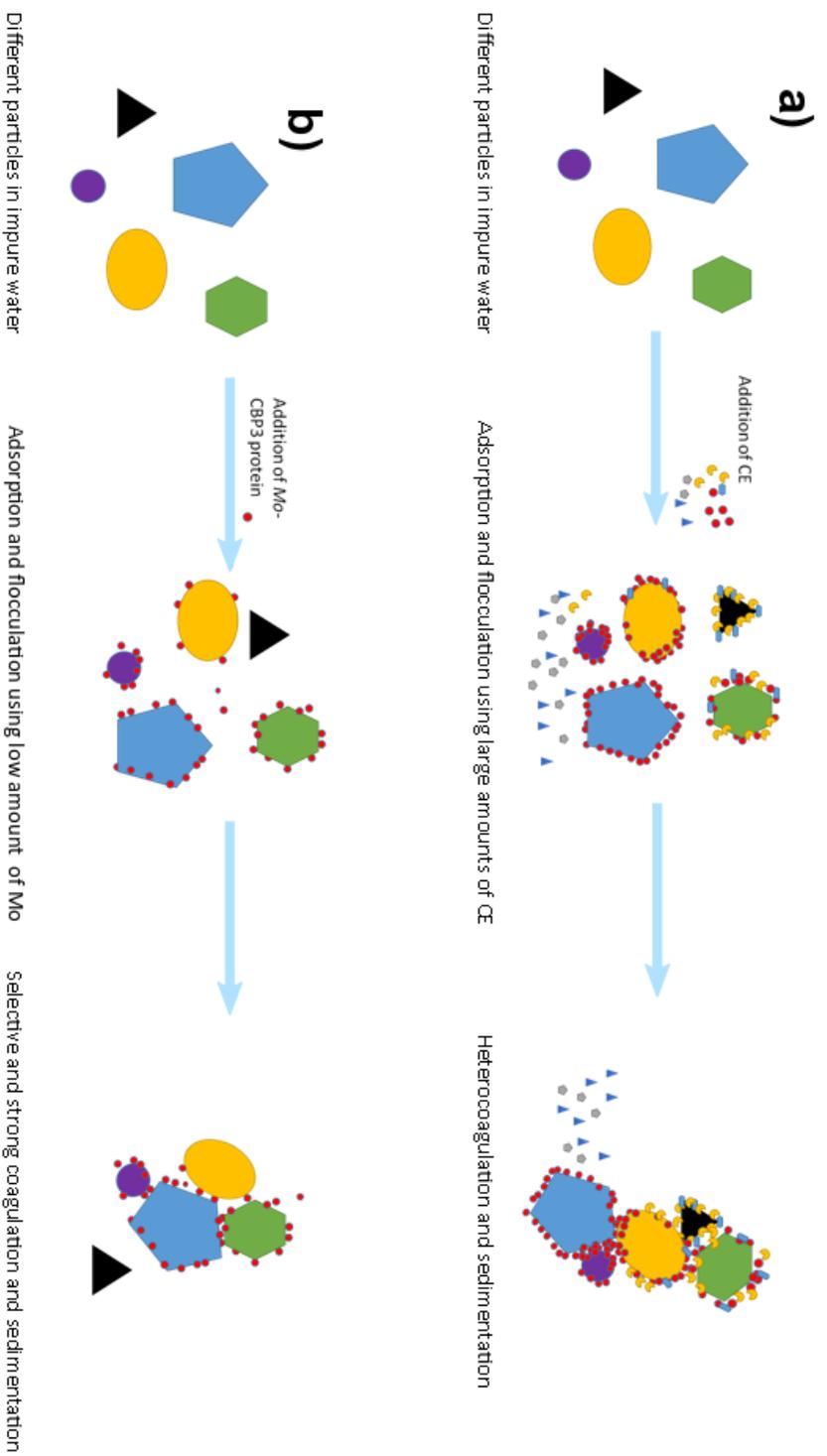
Figure 5.15: Model representing flocculation and coagulation mechanisms of impurities that could occur in dirty water - **a)** in presence of CE or -**b)** with purified fraction (Mo-CBP3 isoforms). According to neutron reflectivity (NR) data obtained from collaborators, CE material was able to interact with both negative and neutral surfaces (ie particles) but the NR data obtained for the purified fraction demonstrated that 20 times more CE was needed to cover the surface at the silica surface **a)**. These Mo–CBP3 isoforms were highly positively charged and interacted more specifically with the silica surface with low concentration of materials.

A future approach would be to investigate interfacial properties of proteins other than 2S albumin like MO2.1, particularly because it can be expressed in a recombinant expression system (Broin *et al.*, 2003). It would be a huge advantage to be able to obtain pure samples easily characterized and produced in sufficient amounts. In future, this study may be extended to different varieties of *Moringa* and to different growth conditions for *Mo* that can give rise to variations in the molecular composition in the seeds. Previous work has shown differences in the flocculation of particles in the presence of extracts from *M. stenopetala* and *M. oleifera* (Hellsing *et al.*, 2014). A future direction linking to the neutron reflection would be to investigate protein interaction with lipid bilayers so as to understand the antimicrobial function. Such studies would benefit from availability of specifically labelled and mutant proteins to determine how various parts interact with interfaces and model cell membranes (bilayers).

6-Conclusions and perspectives

This thesis work has produced new results relating to the characterisation of proteins from *Moringa oleifera* (*Mo*) seeds, their roles in traditional water treatment as well as their potential exploitation in contemporary water purification in developing countries. The key objective was to use a wide range of biochemical and biophysical techniques together to study the crude and purified extracts from these seeds and to establish a rational understanding of their properties in terms of the molecular structure and interaction behaviour.

One of the major challenges was the biochemical and biophysical characterisations of individual components of the crude extract (CE) obtained using water soluble extraction. In Chapter 3, a purified fraction determined as the main component of this CE was extensively studied and characterised. Its amino acid composition analysis revealed a high content in basic amino acids making the proteins highly positively charged with a high IEP. The difficulties encountered in the protein sequence identification by automated Edman degradation may be explained by the presence of glutamine at the N-terminus. Indeed, cyclization of the N-terminal glutamine to pyroglutamate leads to a blocked chain; this has been described for several 2S albumins (Moreno *et al*, 2005). The extremely high stability of this purified fraction as well as its heat and proteolysis resistance was also demonstrated in Chapter 3 by circular dichroism (CD) analysis. These different properties confirmed that the proteins of interest are member of the 2S albumin family. This family is characterised by a compact fold due to the presence of a small chain and a large chain linked together by four disulfide bonds. The presence of these two chains was demonstrated by MS analysis under denaturing and non-denaturing conditions. A main mass of about 11 800 Da for the native form was obtained

under non-reducing conditions whereas values of about 3800 Da and 8000 Da were measured for the short and long chain under reducing conditions respectively. The combination of MS and MS/MS allowed the identification of the fraction of interest as a mixture of two main components: 2 isoforms of *Mo*-CBP3, *Mo*-CBP3-3, and *Mo*-CBP3-4. Since these proteins resulted from the maturation of a precursor, the fractions obtained, also contained a complex mixture of N- and C-terminal processed species of various *Mo* seed protein isoforms (*Mo*-CBP-3). In parallel, the same study was conducted on the CE and revealed, in addition to these two isoforms *Mo*-CBP3-3 and *Mo*-CBP3-4, the presence of a third isoform : *Mo*-CBP3-2 as well as MO 2.1. The tandem MS studies enabled the identification of individual components already deposited in the database; however numerous peaks are still not characterized due to the lack of genomic data. An accurate quantification of all of these individual components present either in the purified fraction or in the CE was not possible on the timescale of this project work. In terms of linking the biophysical and biochemical characterization results to the quality of a sample preparation, the addition of an affinity chromatography step like a chitin resin (as described by Gifoni *et al.*, 2012) would allow the isolation of protein exhibiting chitin binding properties only. Nevertheless, the separation of individual isoforms that differ by two or three amino acids will be challenging without the development of a recombinant expression system. In addition, the natural variability of protein compositions in different CE batches (as described in Chapter 3), further widens the problem. In this same chapter, the activity assays carried out showed the coagulation and the flocculation properties of the 2 isoforms *Mo*-CBP3-3 and *Mo*-CBP3-4. However, the determination of minimum inhibitory concentration (MIC) values were over the highest concentration tested (i.e. over 40 mg/ml for the crude extract and 10 mg/ml for the purified material), demonstrating no significant antibacterial or bacteriostatic activity

against pathogenic bacteria for the CE or the purified *Mo*-CBP3 fraction. Despite this low MIC value, the CE or purified *Mo*-CBP3 fractions may well reduce the bacterial load in treated water by means of flocculation and coagulation. The mechanism of antimicrobial and or flocculation activities could be verified by culture tests. These consist of following bacterial growth in presence or absence of flocculating proteins. Two characteristics could be investigated 1) the ability of the extracts to aggreagate bacterial strains, 2) the effect of the extract on the growth of the bacterial strains. These tests were envisaged the quantity of protein required made the study impractical. These properties were visualized by standard light microscopy using a wide spectrum of organisms such as *E. coli* bacteria and *N. gaditana* algae. These observations are in agreement with recent work by Baptista and co-workers (2017) demonstrating that albumin proteins (which represent 44% of the protein fraction) exhibit a very high coagulant potential even with low turbidity water. Moreover, Suarez *et al.,* in 2005 described the antibacterial activity of *Moringa* cationic peptides as a mechanism of destabilization of membrane through its interaction with negatively charged surfaces and its amphiphilic structure allowing their incorporation into cellular membranes. This hypothesis is in agreement with the biochemical properties (IEP and amino acid composition) of the fraction of interest to this PhD thesis project. The mechanism of action could be identified as a transitory bacteriostatic effect that immobilized the cells and forming clusters for several hours as observed by microscopy. Future molecular-level work is needed to clarify this mode of action. A combination of cryo-electron microscopy (cryo-EM) and fluorescence assays to observe and study the kinetics of fusion membrane in liposomes as Shebek and co-workers (2015) did with cationic peptides, could be envisaged with *Mo*-CBP3 isoforms.

Chapter 4 described the crystal structure (to a resolution of 1.68 Å) of *Mo*-CBP3-4, a mature form of a typical 2S albumin from *Mo* seeds. The obtention and reproducibility of these crystals was challenging due to the variable compositions of the CE and the presence of a mixture of the two isoforms. Nevertheless, the size and the high quality of these crystals resulted in a detailed structural analysis in 80% of the structure was sequenced from the electron density map. The structure shows a pattern of eight-cysteine residues (8CM) in a specific order of type …C…C…/…CC…CXC…C…C…. The publication of Freire and co-workers (2015) which provided the 4 isoform sequences of *Mo*-CBP3, helped identify ambiguous residues that were required to complete the model. *Mo*-CBP3-4 possesses a fold similar to other 2S albumins, corresponding to a structural scaffold of conserved helical regions connected by variable loops. However it differs from this family by having a highly positive molecular surface charge, as observed in flocculating proteins. This crystal structure provides a model for understanding the diversity of the structures of this large family of albumins, serves as a model for the structures of precursors, and has helped identify the structural basis for the observed flocculating activities. Another angle would be to perform a neutron crystallographic study that would be focused on specific issues relating to hydration and protonation states. This is a feasible idea given that crystals of ~500 µm on edge could be obtained readily. Such an approach would required large scale production of perdeuterated protein that could be obtained only via the development of a recombinant expression system. A long term goal could be to express the different *Mo*-CBP3 isoforms and compare their structures.

In addition to the biochemical, biophysical and crystallographic studies, Chapter 5 describes a reflectometry study performed on silica and sapphire interfaces in order to elucidate the

mechanisms relating to flocculated properties. The results show a clear interaction of the protein with the silica interface and an essentially negligible interaction with the alumina surface. The binding mechanism is linked to the positively charged nature of the purified fraction and its molecular arrangement exhibiting positively charged groups on its surface, as observed in the crystal structure. Moreover, an other possible flocculation mechanism by adsorption and bridging of destabilised particles seems less likely due to the globular shape of the protein which differs from polymer structures that are random coil. These reflectometry measurements were related to those recorded for the CE data reported previously by collaborators, and demonstrated that the understanding of binding mechanisms may depend on a *mixture* of proteins being present. These neutron reflection results may be of general significance for an understanding of binding mechanisms relevant to applications involving selective separation of different particles in aqueous dispersions. A further direction linking to the neutron reflection studies would be to investigate protein interactions with lipid bilayers so as to understand antimicrobial properties. Such studies would be challenging and would benefit from availability of specifically labelled and mutant proteins to determine how various parts interact with interfaces and model cell membranes (bilayers). Investigations of interfacial properties of proteins other than the 2S albumin that has been crystallised could be envisaged. Further work may also benefit from molecular dynamics (MD) calculations that exploit the crystal structure information in identifying interfacial interactions at a molecular level. Recent MD studies of bovine serum albumin on silica surfaces have been described by Kubiak-Ossowska *et al* in 2017.

This study constitutes a first step in understanding the molecular basis of the water purification applications of *Moringa* protein. In characterizing this *Mo*-CBP3 protein, it is important to note that *Moringa* seeds contain several other constituents in addition to the protein, and the properties of the extract may depend on multiple biomolecular interactions. The development of a recombinant expression system would help avoid the natural variability of protein composition in the seeds and, as noted above, would allow the production of defined proteins in sufficient amounts for structural studies including neutron reflectometry, SANS and neutron crystallography. The recombinant expression of *Mo*-CBP3 isoforms consisting of two chains would require the production of a single precursor protein and the generation of the mature form by proteolytic processing. The sequences of precursor and mature forms are known from Freire *et al.*, (2015). Two different expression systems could allow the « secretion » process needed for correct disulfide-bond formation in a precursor *Mo*-CBP3. *Brevibacillus choshinensis* is an interesting candidate system to consider for this. It is a highly effective gram-positive bacterial protein expression system, particularly advantageous for the expression of secreted proteins. It lacks endogenous proteases, and is able to secrete large amounts of recombinant protein (Mizumakiet *et al.*, 2010). The secreted precursor would be designed so that the two chains are connected by a linker peptide that could subsequently be removed by a specific protease after purification and refolding.

Pichia pastoris, a methylotrophic yeast, is another option for the construction of such an expression system. It is a production system known for its growth to very high cell densities, for the availability of strong and tightly regulated promoters, and for options to produce gram quantities of recombinant protein per litre of culture in a secretory fashion. The

perdeuteration of yeast cells is possible and again, the secreted precursor would be designed with two chains connected by a linker peptide that could then be removed.

In addition, *Escherichia coli* would also be considered as expression host to produce *Mo*-CBP3 precursor ; individual chains would most likely form as inclusion bodies - as observed for related proteins (Gu *et al*., 2015) such as Mabinlin. The feasibility of refolding would then have to be tested.

Another option would be production by peptide synthesis. This should be reasonably straight forward for the production of the short chain of 25 amino acids. In the case of the long chain, two peptides could be synthetized that would then be linked chemically (Dealwis *et al*., 1998) A further direction linking to this thesis work would be to extend our understanding of coagulation mechanisms to other proteins from seed extracts. The process of water purification could then be optimised, generating significant potential for societal impact. The use of seeds in this way has numerous advantages given that the effectiveness does not depend on the pH of water, and the fact that seed extract reduces 92- 99 % of turbidity (especially when the turbidity is high), and (in contrast to aluminum treatment) does not induce toxicity. However, seed extracts in solution/suspension release organic matter that facilitates bacterial growth especially in hot climates. Moreover, the coagulation activity varies between the different types of seed material used, the extraction methods, the initial turbidity of water, the mixing of suspended particles, and the ionic constant of the water samples. This work has demonstrated the high selectivity and effectiveness of a pure/semi-pure form of an active coagulant in comparison with a crude extract. The interest in using such a coagulant to improve water purification treatment would be to reduce the total amount organic material present and to introduce a standardisation of the process that would improve reliability. The

Association for the Promotion and the Propagation of Plant Resources of Arid and Semi-arid Lands (PROPAGE), a French non-governmental organization reported that, , 20 kg of seeds per day is necessary to treat 100m³of dirty water corresponding to 7.3 tonnes of seed material per year. Such an effort would need space, water for the culture, energy to pump water, all factors that affect the cost-effectiveness. A Suiss company (OPTIMA S.A.) have developed a flocculating product based on *Mo* seeds ; this was patented in 2004. The product is named « Phytofloc » ; it is natural and can be used as an attractive alternative to chemical products. This coagulant is composed of low molecular weight proteins and played as synthetic cationic poly electrolytes. According to the patent this preparation is obtained from a ground presscake of Moringa seeds and mixed with salt water at 1 :5w/v ratio. The extract is filtered and heated at 75°C. After centrifugation the clarified liquid is concentrated. In a similar way, some studies have reported the developement of a Moringa seed filter that is able to retain impurities. These filters are made with sand which was modified by the inclusion of the Moringa peptide, making an effective coagulant (Xiong *et al.*, 2018).

The societal interest in these systems demonstrates the need for a deeper understanding of these seed proteins at a molecular level; the work undertaken as part of this PhD and by others in the field is a significant start but there is a great deal more to learn – as well as a social need for further progress, especially in developing countries.

References

Abdulkarim SM, Long K, Lai OM, Muhammad SKS, Ghazali HM. Some physico-chemical properties of Moringa oleifera seed oil extracted using solvent and aqueous enzymatic methods. *Food Chem.* 2005; *93* (2), 253–263.

Abubakar BY, Xiong B, Lauser K, Darrell V, Kumar M, Velegol S. Effects of *Moringa oleifera* lam.seed maturity and harvesting period on water clarification. *Nigeria Journal of Scientific Research.* 2017;16 (1).

Agizzio AP, Da Cunha M, Carvalho AO, Oliveira MA, Ribeiro SFF, Gomes VM. The antifungal properties of a 2S albumin-homologous protein from passion fruit seeds involve plasma membrane permeabilization and ultrastructural alterations in yeast cells. *Plant Sci.* 2006; *171* (4), 515–522.

Anwar F, Ashraf M, Bhanger MI. Interprovenance variation in the composition of *Moringa oleifera* oil seeds from Pakistan. *JAOCS, J. Am. Oil Chem. Soc.* 2005; *82* (1), 45–51.

Barth VH, Habs M, Klute R, Miiller S, Tauscher B. Trinkwasseraufbereitung mil Samen von *Moringa oleifera* Lam*. Chemiker-Zeitung*, 1982 ; 106, 75.

Baptista ATA, Silva MO, Gomes RG, Bergamasco R, Vieira MF, Vieira AMS. Protein fractionation of seeds of Moringa oleifera lam and its application in superficial water treatment. *Sep Purif Technol*. 2017;180:114-124.

Batista AB, Oliveira JTA, Gifoni JM, Pereira ML, Almeida MGG, Gomes VM, Da Cunha M, Ribeiro SFF, Dias GB, Beltramini LM, Lopes JLS, Grangero TB, Vasconcelos IM. New insights into the structure and mode of action of *Mo*-CBP3, an antifungal chitin-binding protein of *Moringa oleifera* seeds. *PLoS One* 2014; *9* (10), 1–9.

Baud F, Pebay-Peyroula E, Cohen-Addad C, Odani S, Lehmann MS. Crystal Structure of Hydrophobic Protein from Soybean; a Member of a New Cysteine-rich Family. *J Mol Biol*. 1993;231(3):877-887.

Baud, S, Boutin JP, Miquel M, Lepiniec L, Rochat C. An integrated overview of seed development in Arabidopsis thaliana ecotype WS. *Plant Physiol. Biochem.* 2002 ; *40* (2), 151–160.

Beccari**.** 1745 De Frumento .De Bononiensi Scientiarum et Artium Instituto atque Academia Commentarii, II. Part I., 122–127.

Beltrán-Heredia J, Sánchez-Martín, J. Improvement of water treatment pilot plant with *Moringa oleifera* extract as flocculant agent. *Environ. Technol.* 2009; *30* (6), 525–534.

Behnke CA, Yee VC, Trong I Le,Pedersen LC, Stenkamp RE, Kim SS, Reeck GR, Teller DC. Structural Determinants of the Bifunctional Corn Hageman Factor Inhibitor: X-ray Crystal Structure at 1.95 Å Resolution. *Biochemistry*. 1998;37(44):15277-15288.

Berger MR, Habs M, Jahn SAA, Schmahl, D. Toxicological assessment of seeds from *Moringa oleifera* and *Moringa stenopetala*, two highly efficient primary coagulants for domestic water treatment of tropical raw waters. *East Afr. Med. J.* 1984; *61* (9), 712–716.

Brahms S, Brahms J. Determination of protein secondary structure in solution by vacuum ultraviolet circular dichroism. *J. Mol. Biol.* 1980; *138* (2), 149–178.

Bonnerjera J, Oh S, Hoare M, Dunhill P .The right step at the right time. *Bio/Technology***,** 1986; 4, 954-958.

Bradford MM. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem*. 1976;72(1-2):248-254.

Broin M, Santaella C, Cuine S, Kokou K, Peltier G, Joët, T. Flocculent activity of a recombinant protein from Moringa oleifera Lam. seeds. *Appl. Microbiol. Biotechnol.* 2003; *60* (1–2), 114–119.

Castellanos MM, McAuley A, Curtis JE. Investigating Structure and Dynamics of Proteins in Amorphous Phases Using Neutron Scattering. *Comput Struct Biotechnol J*. 2017;15:117-130.

CLSI (The clinical and laboratory standard insitut) guidelines : CLSI; M07-A8 Vol. 29, N°2.

Chalker JM*,* Bernardes GJL*,* Lin YA*,* Davis BG*. Chemical Modification* of *Proteins at Cysteine*: *Opportunities* in *Chemistry* and *Biology*. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim. *Chem. Asian J*. 2009 ; (4), 630 – 640.

Chuang PH, Lee CW, Chou JY, Murugan M, Shieh BJ, Chen HM. Anti-fungal activity of crude extracts and essential oil of *Moringa oleifera* Lam. *Bioresour Technol*. 2007;98(1):232-236.

COSMOS, (2017). URL: https://www.ill.eu/instruments-support/instruments-groups/instruments/d17/more/documentation/d17-lamp-book/cosmos/

Cottrell JS. Protein identification using MS/MS data. *J Proteomics*. 2011;74(10):1842-1851.

Cousin F, Menelle A. Neutron reflectivity. *EPJ Web Conf*. 2015;104:1005.

Cubitt R, Fragnetto G. D17: The new reflectometer at the ILL. *Appl Phys A Mater Sci Process*. 2002;74:S329-S331.

Dealwis C, Fernandez EJ, Thompson DA, Simon RJ, Siani MA, Lolis E. Crystal structure of chemically synthesized [N33A] stromal cell-derived factor 1α, a potent ligand for the HIV-1 "fusin" coreceptor. *Proceedings of the National Academy of Sciences of the United States of America*. 1998;95(12):6941-6946.

De Sanctis D, Beteva A, Caserotto H, Dobias F, gabadinho J, Giraud T, Gobbo A, Giujarro M, Lentini M, Lavault B, Mairs T, McSweeney S, Petitdemange S, Rey-Bakaikore V, Surr J, Theveneau P, Leonard GA, Mueller-Dieckmann C. ID29: A high-intensity highly automated ESRF beamline for macromolecular crystallography experiments exploiting anomalous scattering. *J Synchrotron Radiat*. 2012;19(3):455-461.

Doerr B. *Moringa* water treatment: ECHO technical note.2005 Education concerns for Hunger organization (ECHO), North fort Myers,Fla.

Ducruix A, Giegé R. *Crystallization of nucleic acids and proteins: a practical approach*. 1992 Oxford [England]: IRL Press at Oxford University Press.

Duke J, Atchley A. ; *Proximate Analysis In The Handbook of Plant Science in Agricultural*. 1984 Christie,BR (ed). CRC Press inc, Boca Raton Florida USA.

Coordinators NR. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2017;45(D1):D12-D17.

Egorov TA, Odintsova TI, Musolyamov AK, Fido R, Tatham AS, Shewry PR. Disulphide structure of a sunflower seed albumin: conserved and variant disulphide bonds in the cereal prolamin superfamily. *FEBS Lett*. 1996;396(2):285-288.

Eisenberg D, Weiss RM, Terwilliger TC. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc Natl Acad Sci*. 1984;81(1):140-144.

Emsley P, Cowtan K. Coot: Model-building tools for molecular graphics. *Acta Crystallogr. Sect. D Biol. Crystallogr.* 2004; *60* (12 I), 2126–2132.

Fernandez-Tornero C, Ramon A, Navarro ML, Varela J, Gimenez-Gallego G. Synthesis of proteins with disulphide bonds in E. coli using defined culture media. *Biotechniques* 2002;(6), 1238-1242.

Franks G, Gan Y. Charging Behavior at the Alumina–Water Interface and Implications for Ceramic Processing. *J Am Ceram Soc*. 2007;90:3373-3388.

Freire JEC, Vasconcelos IM, Moreno FBMB, Batista AB, Lobo MDP, Pereira ML, Lima JPMS, Almeida RVM, Sousa AJS, Monteiro-Moreira ACO, Oliveira JTA, Grangeiro TB. *Mo*-CBP3, an antifungal chitin-binding protein from *Moringa oleifera* seeds, is a member of the 2S albumin family. *PLoS One* 2015 ; *10* (3), 1–24.

Gassenschmidt U, Jany KD, Bernhard T, Niebergall H. Isolation and characterization of a flocculating protein from *Moringa oleifera* Lam. *BBA - Gen. Subj.* 1995; *1243* (3), 477–481.

Ghebremichael KA, Gunaratna KR, Henriksson H, Brumer H, Dalhammar G. A simple purification and activity assay of the coagulant protein from Moringa oleifera seed. *Water Res*. 2005;39(11):2338-2344.

Gifoni JM, Oliveira JTA, Oliveira HPHD, Batista AB, Pereira ML, Gomes AS, Grangeiro TB, Vasconcelos IMA. novel chitin-binding protein from *Moringa oleifera* seed with potential for plant disease control. *Biopolymers* 2012; *98* (4), 406–415.

Goodwin JW, Hearn J, Ho CC, Ottewill RH. Studies on the preparation and characterisation of monodisperse polystyrene laticee. *Colloid Polym Sci*. 1974;252(6):464-471.

Gourinath S, Alam N, Srinivasan A, Betzel C, Singh TP. Structure of the bifunctional inhibitor of trypsin and {$α$}-amylase from ragi seeds at 2.2Åresolution. *Acta Crystallogr Sect D*. 2000;56(3):287-293.

Gu W, Xia Q, Yao J, Fu S, Guo J, Hu X. Recombinant expressions of sweet plant protein mabinlin II in *Escherichia coli* and food-grade *Lactococcus lactis*. *World J Microbiol Biotechnol*. 2015 ; 31 (4):557-567.

Gutfreund P, Gonzalez MA, Pellegrini E, Laver M, Dewhurst C, Cubitt R. 'Towards generalized data reduction on a time-of-flight neutron reflectometer *J. Appl. Cryst*. 2018;51,3.

Harada J.J. Seed Maturation and Control of Germination. In: Larkins B.A., Vasil I.K. (eds) Cellular and Molecular Biology of Plant Seed Development. Advances in Cellular and Molecular Biology of Plants,1997; vol 4. Springer, Dordrecht.

Hellsing MS, Kwaambwa HM, Nermark FM, Nkoane BBM, Jackson AJ, Wasbrough MJ, Berts I, Porcar L, Rennie A R. Structure of flocs of latex particles formed by addition of protein from Moringa seeds. *Colloids Surfaces A Physicochem. Eng. Asp.* 2014; *460*, 460–467.

Jahn SAA. Traditional methods of water purification in the riverain Sudan in relation to geographic and socioeconomic conditions.*Erdkunde* (Bonn) 1977 ; 31,120.

Jahn SAA, Dirar H. Studies on natural water coagulants in the Sudan with reference to Moringa oleifera seeds. *Water SA*. 1979; pp 90–97.

Jahn SAA. Proper use of African natural coagulants for rural water supplies: research in the Sudan and a guide for new projects. 1986 Eschborn, Fed. Rep. Germany, GTZ n°191, p 539.

Jahn SAA. Using *Moringa* seeds as coagulants in developing countries. *J. / Am. Water Work. Assoc.* 1988; *80* (6), 43–50.

Jerri HA, Adolfsen K J, McCullough L R, Velegol D, Velegol S B. Antimicrobial sand via adsorption of cationic *Moringa oleifera* protein. *Langmuir* 2012; *28* (4), 2262–2268.

José-Estanyol M, Gomis-Rüth F X, Puigdomènech P. The eight-cysteine motif, a versatile structure in plant proteins. *Plant Physiology and Biochemistry*. 2004; 355–365.

Kabsch W. XDS. *Acta Crystallogr Sect D Biol Crystallogr*. 2010;66(2):125-132.

Katayon S, Noor MJMM, Asma M, Ghani LAA, Thamer AM, Azni I, Ahmad J, Khor BC, Suleyman AM. Effects of storage conditions of *Moringa oleifera* seeds on its performance in coagulation. *Bioresour Technol*. 2006;97(13):1455-1460.

Katre UV, Suresh CG, Khan MI, Gaikwad SM. Structure-activity relationship of a hemagglutinin from *Moringa oleifera* seeds. *Int. J. Biol. Macromol.* 2008; *42* (2), 203–207.

Kubiak-Ossowska K, Tokarczyk K, Jachimska B, Mulheran PA. Bovine Serum Albumin Adsorption at a Silica Surface Explored by Simulation and Experiment. *J Phys Chem B*. 2017;121(16):3975-3986.

Kwaambwa HM, Maikokera R. Infrared and circular dichroism spectroscopic characterisation of secondary structure components of a water treatment coagulant protein extracted from Moringa oleifera seeds. *Colloids Surfaces B Biointerfaces*. 2008;64(1):118-125.

Kwaambwa HM, Hellsing M, Rennie AR. Adsorption of a water treatment protein from *Moringa oleifera* seeds to a silicon oxide surface studied by neutron reflection. *Langmuir* 2010; *26* (6), 3902–3910.

Kwaambwa HM, Rennie AR. Interactions of surfactants with a water treatment protein from *Moringa oleifera* seeds in solution studied by zeta-potential and light scattering measurements. *Biopolymers* 2012; *97* (4), 209–218.

Kwaambwa HM, Hellsing MS, Rennie A R, Barker R. Interaction of *Moringa oleifera* seed protein with a mineral surface and the influence of surfactants. *J. Colloid Interface Sci.* 2015; *448*, 339–346.

Langer G, Cohen S X, Lamzin VS, Perrakis A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat. Protoc.* 2008; *3* (7), 1171–1179.

Lea M. Bioremediation of Turbid Surface Water Using Seed Extract from *Moringa oleifera* Lam. (Drumstick) Tree. *Curr Protoc Microbiol*. 2010; Chapter 1:Unit1G.2.

Leone A, Spada A, Battezzati A, Schiraldi A, Aristil J, Bertoli S. *Moringa oleifera* seeds and oil: Characteristics and uses for human health. *Int J Mol Sci*. 2016;17(12).

Li D-F, Jiang P, Zhu D-Y, Hu Y, Max M, Wang D-C. Crystal structure of Mabinlin II: A novel structural type of sweet proteins and the main structural basis for its sweetness. *J Struct Biol*. 2008;162(1):50-62.

Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. *Science (80-)*. 1985;227(4693):1435 LP-1441.

Luz LA, Silva MCC, da Silva Ferreira R, Santana L A, Silva—Lucca RA, Mentele R, Oliva MLV, Paiva PMG, Coelho LCBB. Structural characterization of coagulant *Moringa oleifera* Lectin and its effect on hemostatic parameters. *Int J Biol Macromol*. 2013;58:31-36.

Madrona GS, Serpelloni GB, Salcedo Vieira AM, Nishi L, Cardoso KC, Bergamasco R. Study of the effect of Saline solution on the extraction of the *Moringa oleifera* seed's active component for water treatment. *Water. Air. Soil Pollut.* 2010; *211* (1–4), 409–415.

Madrona GS, Bergamasco R, Seolin VJ, Fagundes Klen MR. The Potential of Different Saline Solution on the Extraction of the *Moringa oleifera* Seed's Active Component for Water Treatment. *Int. J. Chem. React. Eng.* 2011; *9* (1).

Madsen M, Schlundt J, Omer EF. Effect of water coagulation by seeds of *Moringa oleifera on* bacterial contaminations. J. Trop. Med. Hyg., 1987 ; (90) 101-109.

Maeda H, Ishida N. Specificity of Binding of Hexopyranosyl Polysaccharides with Fluorescent Brightener. *J Biochem*. 1967;62(2):276-278.

Makkar HPS, Becker K. Nutrients and antiquality factors in different morphological parts of the *Moringa oleifera* tree. *J. Agric. Sci.* 1997; *128* (3), 311-322.

Maschke, O. Ueber den Bau und die Bestandtheile der Kleberblaschen in Bertholletia, deren Entwickelung in Ricinus, nebst einigen Bemerkungen über Amylonblaschen. Botanische Zeitung, 1859 ; 17,409-447.

Matilainen ET, Gjessing T, Lahtinen L, Hed A, Bhatnagar M. Sillanpää, Chemosphere an overview of the methods used in the characterisation of natural organic matter (NOM) in relation to drinking water treatment, Chemosphere 2011 ; (83) 1431–1442.

McConnachie GL, Folkard GK, Mtawali MA, Sutherland J P. Field trials of appropriate hydraulic flocculation processes. *Water Res.* 1999; *33* (6), 1425–1434.

Mizukami M, Hanagata H, Miyauchi A. Brevibacillus expression system : Host-vector system for efficient production of secretory proteins. *Curr. Pharm. Biotechnol.* 2010; 11(3):251-258.

Murshudov GN, Vagin AA, Dodson E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallographica Section D: Biological Crystallography*. 1997; 240–255.

Moreno-Risueno MÁ, González N, Díaz I, Parcy F, Carbonero P, Vicente-Carbajosa J. FUSCA3 from barley unveils a common transcriptional regulation of seed-specific genes between cereals and Arabidopsis. *Plant J*. 2008; 53 (6):882-894.

Muhl QE, du Toit ES, Steyn J M, Robbertse P J. The embryo, Endosperm and seed coat structure of developing *Moringa oleifera* seed. *South African J. Bot.* 2016; *106*, 60–66.

Mylne JS, Hara-Nishimura I, Rosengren KJ. Seed storage albumins: Biosynthesis, trafficking and structures. *Funct Plant Biol*. 2014; 41 (7):671-677.

Ndabigengesere A, Narasiah KS, Talbot BG. Active agents and mechanism of coagulation of turbid waters using Moringa oleifera. *Water Res.* 1995; *29* (2), 703–710.

Nirasawa S, Nishino T, Katahira M, Uesugi S, Hu Z, Kurihara Y. Structures of heat-stable and unstable homologues of the sweet protein mabinlin. *Eur J Biochem*.1994 ; 223(3):989-995.

Nylander T, Campbell RA, Vandoolaeghe P, Cárdenas M, Linse P, Rennie AR. Neutron reflectometry to investigate the delivery of lipids and DNA to interfaces (Review). *Biointerphases*. 2008;3(2):FB64--FB82.

Oda Y, Matsunaga T, Fukuyama K, Miyazaki T, Morimoto T. Tertiary and Quaternary Structures of 0.19 α-Amylase Inhibitor from Wheat Kernel Determined by X-ray Analysis at 2.06 Å Resolution,. *Biochemistry*. 1997;36(44):13503-13511.

Okuda T, Baes AU, Nishijima W, Okada M. Isolation and characterization of coagulant extracted from *Moringa oleifera* seed by salt solution. *Water Res*. 2001;35(2):405-410.

Olagbemide PT, Philip C. Proximate Analysis and Chemical Composition of Raw and Defatted *Moringa oleifera* Kernel.Advances in Life Science and Technology,2014**;** 24, 92-99.

Oliveira JTA, Silveira SB, Vasconcelos IM, Cavada BS, Moreira RA. Compositional and nutritional attributes of seeds from the multiple purpose tree *Moringa oleifera* Lamarck. *J Sci Food Agric***.** 1999; 79(6):815-820.

Osborne.T.B ' the vegetables proteins',Logmans,Green and Co., London 1924.

Paizs B, Suhai S. Fragmentation pathways of protonated peptides. *Mass Spectrom Rev*. 2005 ; 24(4):508-548.

Palomares O, Monsalve RI, Rodríguez R, Villalba M. Recombinant pronapin precursor produced in *Pichia pastoris* displays structural and immunologic equivalent properties to its mature product isolated from rapeseed. *European Journal of Biochemistry*,2002 ; 269: 2538-2545.

Pan SQ, Ye XS, Kué J. A technique for detection of chitinase, β-1,3-glucases, and protein patterns after a single separation using polyacrylamide gel electrophoresis or isoelectrofocusing *Phytopath*, 1991; (81) , 970-973.

Pantoja-Uceda D, Bruix M, Giménez-Gallego G, Rico M, Santoro J. Solution Structure of RicC3, a 2S Albumin Storage Protein from Ricinus communis,. *Biochemistry*. 2003;42(47):13839-13847.

Pereira, On the purification of drinking water. *Pharmaceutical Journal*, 1850 9,474.

Perkins DN, Pappin DJ, Creasy DM, Cottrell, JS. *Probability-based protein identification by searching sequence databases using mass spectrometry data*. Electrophoresis, 1999; 20(18) 3551-67.

Pritchard M, Craven T, Mkandawire T, Edmondson A S, O'Neill J G. A comparison between *Moringa oleifera* and chemical coagulants in the purification of drinking water - An alternative sustainable solution for developing countries. *Phys. Chem. Earth* 2010; *35* (13–14), 798–805.

Rosenfeld J, Capdevielle J, Guillemot JC, Ferrara P. In-gel digestion of proteins for internal sequence analysis after one- or two-dimensional gel electrophoresis. *Anal Biochem*. 1992;203(1):173-179.

Rennie AR, Hellsing MS, Wood K, Gilbert EP, Porcar L, Schweins R, Dewhurst CD, Lindner P, Heenan RK, Rogers SE, Butler PD, Krywon JR, Ghosh RE, Jackson AJ, Malfois M. Learning about SANS instruments and data reduction from round robin measurements on samples of polystyrene latex. *J Appl Crystallogr*. 2013;46(5):1289-1297.

Rennie AR, ; Hellsing M S, ; Lindholm E, ; Olsson A. Note: Sample cells to investigate solid/liquid interfaces with neutrons. *Rev Sci Instrum*. 2015;86(1).

Rico M, Bruix M, González C, Monsalve RI, Rodríguez R. 1H NMR Assignment and Global Fold of Napin BnIb, a Representative 2S Albumin Seed Protein. *Biochemistry*. 1996;35(49):15672-15682.

Rodriguez-Tudela JL, Alcazar-Fuoli L, Cuesta I, Alastruey-Izquierdo, A, Monzon A, Mellado E, Cuenca-Estrella M. Clinical relevance of resistance to antifungals. *Int. J. Antimicrob. Agents* 2008; *32* (SUPPL. 2).

Roepstorff P, Fohlman J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed Mass Spectrom*. 1984;11(11):601.

Sajidu SMI, Henry EMT, Persson I, Masamba WRL, Kayambazinthu D. pH dependence of sorption of Cd^{2+}, Zn^{2+}, Cu^{2+} and Cr^{3+} on crude water and sodium chloride extracts of Moringa stenopetala and Moringa oleifera. *African J Biotechnol*. 2006;5(23).

Sánchez-Martín J, Beltrán-Heredia J, Peres JA. Improvement of the flocculation process in water treatment by using *Moringa oleifera* seeds extract. *Brazilian J Chem Eng*. 2012;29(3):495-501.

Sarpong G, Richardson CP. Coagulation efficiency of *Moringa oleifera* for removal of turbidity and reduction of total coliform as compared to aluminum sulfate. *African J Agric Res*. 2010;5(21):2939-2944.

Schägger H, von Jagow G. Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 kDa. *Anal Biochem*. 1987;166(2):368-379.

Shebek K, Schantz AB, Sines I, Lauser K, Velegol S, Kumar, M. The flocculating cationic polypetide from *Moringa oleifera* seeds damages bacterial cell membranes by causing membrane fusion. *Langmuir* 2015; *31* (15), 4496–4502.

Sheldrick G M. SHELXL97. Program for the Refinement of Crystal Structures. *Acta. Cryst.* 2008 ; *A64*, 112–122.

Shewry PR. Plant Storage Proteins. *Biol. Rev.* 1995; *70* (3), 375–426.

Singh U, SinghB. Tropical grain legumes as important human foods*. Econ Bot*. **1992**, 46:310-321.

Strobl S, Maskos K, Betz M, et al. Crystal structure of yellow meal worm α-amylase at 1.64 Å resolution11Edited by I. A. Wilson. *J Mol Biol*. 1998;278(3):617-628.

Strohalm M, Hassman M, Kosata B, Kodicek M. mMass data miner: an open source alternative for mass spectrometric data analysis. *Rapid Commun Mass Spectrom*. 2008;22(6):905-908.

Suarez M, Entenza JM, Doerries C, Meyer F, Bourquin L, Sutherland J, Marison I, Moreillon P, Mermod N. Expression of a Plant-Derived Peptide Harboring Water-Cleaning and Antimicrobial Activities. *Biotechnol Bioeng*. 2003;81(1):13-20.

Suarez M, Haenni M, Canarelli S, Fisch F, Chodanowski P, Servis C, Michielin O, Freitag R, Moreillon P, Mermod N. Structure-function characterization and optimization of a plant-derived antibacterial peptide. *Antimicrob. Agents Chemother.* 2005; *49* (9), 3847–3857.

Trudel J, Asselin A. Detection of chitinase activity after polyacrylamide gel electrophoresis. *Anal Biochem*. 1989;178(2):362-366.

Ullah A, Mariutti RB, Masood R, Caruso IP, Costa GHGC, De Freita CM, Santos CR, Zanphorlin LM, Mutton MJR, Murakami MT, Arni RK. Crystal structure of mature 2S albumin from *Moringa oleifera* seeds. *Biochem Biophys Res Commun*. 2015;468(1-2):365-371.

Vagin AA, Steiner RA, Lebedev AA, Potterton L, McNicholas S, Long F, Murshudov GN. REFMAC5 dictionary: Organization of prior chemical knowledge and guidelines for its use. *Acta Crystallogr Sect D Biol Crystallogr*. 2004;60(Pt12Pt1):2184-2195.

Vicente-Carbajosa J, Carbonero P. Seed maturation: Developing an intrusive phase to accomplish a quiescent state. *International Journal of Developmental Biology*. 2005; pp 645–651.

Vijayaraghavan G, Sivakumar T, Vimal Kumar A. Application of plant based coagulants for waste water treatment*, Int. J. Adv. Eng. Res. Stud*. 1 2011; 88–92.

Wobus U, Weber H. Seed maturation: Genetic programmes and control signals. *Current Opinion in Plant Biology*. 1999; pp 33–38.

Xiong B, Piechowicz B, Wang Z, Marinaro R, Clement E, Carlin T, Uliana A, Kumar M, Velegol SB. *Moringa oleifera* f-sand Filters for Sustainable Water Purification. *Environ Sci Technol Lett*. 2018; 5(1):38-42.

Youle, RJ, Huang AHC. Occurrence of Low Molecular Weight and High Cysteine Containing Albumin Storage Proteins in Oilseeds of Diverse Species. *Am. J. Bot.* 1981; *68* (1), 44.