



**University of Brasília
Institute of Biological Sciences
Department of Cell Biology
Molecular Biology Postgraduate Program**

GUSTAVO TRAJANO DE MOURA

**CRYSTAL STRUCTURES AND BIOPHYSICAL STUDIES OF
PHYTOCYSTATINS**

BRASÍLIA, DF

2019

GUSTAVO TRAJANO DE MOURA

**CRYSTAL STRUCTURES AND BIOPHYSICAL STUDIES OF
PHYTOCYSTATINS**

Dissertation presented to the Molecular Biology
Postgraduate program of the University of Brasília

Research field: Molecular Biophysics

Advisor: Prof. Dr. Sônia Maria de Freitas

Co-advisor: Prof. Dr. Napoleão Fonseca Valadares

**BRASÍLIA, DF
2019**

BANCA EXAMINADORA

Prof^a. Dr^a. Sônia Maria de Freitas

Laboratório de Biofísica Molecular
Departamento de Biologia Celular
Universidade de Brasília
Presidente da Banca e Orientadora

Prof. Dr. João Alexandre Ribeiro Gonçalves Barbosa

Laboratório de Biofísica Molecular
Departamento de Biologia Celular
Universidade de Brasília
Membro Interno Vinculado ao Programa

Prof. Dr. Guilherme Martins Santos

Laboratório de Farmacologia Molecular
Universidade de Brasília
Membro Interno Não-Vinculado ao Programa

Prof. Dr. Aisel Valle Garay

Laboratório de Biofísica Molecular
Departamento de Biologia Celular
Universidade de Brasília
Membro Suplente

ACKNOWLEDGEMENTS

Though this journey to knowledge and self-growth started from the aspirations of a single individual, this work could not have been done without the support of several other people. Therefore, I would like to wholeheartedly express my gratitude:

To my family, for supporting me in all of my choices, even when uncertain of where it would lead me.

To professor Sônia de Freitas, for giving a chance to this random physicist that appeared so suddenly, for the words of advice and for pushing me to be a better professional.

To professor Napoleão Valadares, for believing in my potential, for all of the guidance and for always being present.

To professors João Alexandre and Aisel Valle, for all of the shared wisdom and for always inspiring me to be more curious.

To Paulo Jardim, for tutoring me during my very first steps in the laboratory, for being the best lab partner one could wish for, and for a friendship I deeply treasure.

To Azadeh Mehdad, for being an inspiration in my pursuit for good standards and culture. Meeting you has attached a special tenderness to the word "free".

To my colleagues and friends from the Laboratory of Molecular Biophysics: Amanda Souza, Cláudia Bravo, Diego Honda, Diogo de Sá, Gideane Mendes, João Paulo Fernandes, Jônatas Lima, Julia Freitas, Liem Canet, Luisa Meneghetti, Muhammad Faheem, Patrícia Alves and Wallace Vasconcelos. Thank you for the camaraderie, the pleasant conversations, the thought provoking debates over coffee and all of the support.

To Bruno Daniel, Carlos Eduardo, Darlan Lucio, Gustavo Ferreira, Gustavo Guissem, Juliano Pires, Lethícia Vieira, Luís Hagemann, Ricardo Leite and Valmir Filho, for all of the good times that helped me ease my mind.

To the CNPq, for the financial aid, facilitating the elaboration of this work.

To the LNLS for making the X-ray diffraction experiments possible.

ABSTRACT

Phycystatins are cysteine-protease inhibitors found in plants, in which they take part in various physiological processes such as storage protein deposition, seed germination and leaf senescence. The expression of phycystatins in plants has an important role in abiotic stress tolerance and resistance to herbivory by insects, which makes such proteins valuable study targets due to biotechnological applications in culture improvement. In this work, biophysical techniques were employed to characterize phycystatins from organisms of economic relevance, like *Hevea brasiliensis* (rubber tree), *Theobroma cacao* (cocoa tree), *Humulus lupulus* (hop) and *Cannabis sativa* (cannabis). The proteins were obtained through heterologous expression in *E. coli* Lemo21 (DE3), cultivated in auto-inducing medium ZYM-5052, and purified by immobilized Ni²⁺ ion affinity chromatography followed by size exclusion chromatography S-75. Thermal stability was monitored by circular dichroism measurements in unfolding assays at different pHs, and protein oligomerization was studied by analytical ultracentrifugation. Protein crystallization screenings were performed with the mosquito[®] robot and the crystallization conditions were manually refined prior to sending the crystals for the collection of X-ray diffraction data at LNLS. Thermal denaturation assays showed that the proteins exhibited high melting temperatures ranging from 61 °C to 84 °C and unfolding ΔG_{25} 's higher than 4 kcal/mol in the studied pHs, indicating moderate to high thermal stability. In the experimental conditions (20 °C and pH 7.6), dimeric conformations of most of the proteins seemed to be favored, though monomers and tetramers were also observed. The crystal structure of a phycystatin from hop was solved from crystals of orthorhombic space groups $P 2_1 2_1 2_1$ and $C 2_2 2_1$ at maximum resolutions of 1.80 Å and 1.68 Å, respectively, and the crystal structure of a cannabis phycystatin was solved at a 3.6 Å resolution from a hexagonal $P 6_5 2_2$ space group crystal. The three structures presented domain swapped dimers, where the main differences to the phycystatin structure with the highest identity deposited in the PDB are the angle and the distance between the lobes of the dimers.

Keywords: phycystatins, crystallography, domain-swapping, analytical ultracentrifugation, circular dichroism, protein stability.

RESUMO

Fitocistatinas são inibidores de cisteíno-proteases encontrados em plantas, onde participam de vários processos fisiológicos, dos quais se destacam a deposição de proteínas de armazenamento, germinação de sementes e senescência de folhas. A expressão de fitocistatinas em plantas tem um papel importante na tolerância a fatores de estresse abióticos e resistência a herbivoria por insetos, o que torna essas proteínas valiosos alvos de estudo, devido a aplicações biotecnológicas no melhoramento de culturas. Neste trabalho, técnicas biofísicas foram empregadas para caracterizar fitocistatinas de organismos economicamente relevantes, como *Hevea brasiliensis* (seringueira), *Theobroma cacao* (cacaueiro), *Humulus lupulus* (lúpulo) e *Cannabis sativa* (cannabis). As proteínas foram obtidas por expressão heteróloga em *E. coli* Lemo21 (DE3), cultivadas em meio auto-indutor ZYM-5052, e purificadas por cromatografia de afinidade ao níquel (Ni^{2+}) seguida de cromatografia de exclusão molecular S-75. A estabilidade térmica foi monitorada por dicroísmo circular em ensaios de desnaturação térmica em diferentes pHs, e a oligomerização das proteínas foi estudada por ultracentrifugação analítica. *Screenings* de cristalização das proteínas foram realizados com o robô mosquito[®], as condições de cristalização foram refinadas manualmente e os cristais foram submetidos à coleta de dados de difração de raios-X no LNLS. Ensaio de desnaturação térmica mostraram que as proteínas apresentam elevadas temperaturas de desdobramento (T_m), indo de 61 °C a 84 °C, e ΔG_{25} 's resultantes do desdobramento acima de 4 kcal/mol em todos os pHs estudados, indicando estabilidade térmica de moderada a alta. Nas condições experimentais (20°C e pH 7,6), conformações diméricas da maioria das proteínas foram favorecidas, embora monômeros e tetrâmeros também tenham sido observados. A estrutura cristalina de uma fitocistatina do lúpulo foi resolvida a partir de cristais dos grupos espaciais ortorrômbicos $P 2_1 2_1 2_1$ e $C 2_2 2_1$ a resoluções máximas de 1,80 Å e 1,68 Å, respectivamente, e a estrutura cristalina de uma fitocistatina da cannabis foi resolvida a 3,6 Å a partir de um cristal do grupo espacial hexagonal $P 6_5 2 2$. As três estruturas apresentaram dímeros *domain-swapped*, onde a principal diferença em relação à estrutura de fitocistatina de maior identidade depositada no PDB são o ângulo e a distância entre os lobos do dímero.

Palavras-chave: fitocistatinas, cristalografia, domain-swapping, ultracentrifugação analítica, dicroísmo circular, estabilidade de proteínas.

IMAGE LIST

Figure 1. Conserved tridimensional cystatin structure elements represented by oryzacystatin-I (PDB code 1EQK) (NAGATA et al, 2000)	16
Figure 2. Inhibition of papain by tarocystatin (PDB code 3IMA).....	17
Figure 3. Monomer (PDB code 1MDT) and domain swapped dimer (PDB code 1DDT) of DT (BENNETT & EISENBERG, 1994; BENNET et al, 1994).....	18
Figure 4. The domain swapping mechanism	19
Figure 5. Sequence alignment of the proteins in this study	20
Figure 6. Electrophoretic profile in SDS-PAGE 15 % of the expression and solubility of Late and Theo in LB medium.....	34
Figure 7. Electrophoretic profile in SDS-PAGE 15 % of the expression and solubility of Late in auto-inducing medium ZYM-5052	34
Figure 8. Electrophoretic profile in SDS-PAGE 15 % of the expression and solubility of Theo in auto-inducing medium ZYM-5052.....	35
Figure 9. Electrophoretic profile in SDS-PAGE 15 % of the expression and solubility of Hop1, Hop3 and Hop5 in auto-inducing medium ZYM-5052	35
Figure 10. Electrophoretic profile in SDS-PAGE 15 % of the expression and solubility of Hemp1 in auto-inducing medium ZYM-5052	36
Figure 11. IMAC chromatogram of Late	38
Figure 12. Superposition of two size exclusion chromatograms of Late samples	38
Figure 13. Electrophoretic profile in SDS-PAGE 15 % of the purification of Late	39
Figure 14. IMAC chromatogram of Theo	39
Figure 15. Superposition of two size exclusion chromatograms of Theo samples.....	40
Figure 16. Electrophoretic profile in SDS-PAGE 15 % of the purification samples of Theo.....	40
Figure 17. IMAC chromatogram of Hop1	41
Figure 18. Superposition of two size exclusion chromatograms of Hop1 samples.....	41
Figure 19. Electrophoretic profile in SDS-PAGE 15 % of the purification samples of Hop1	42
Figure 20. IMAC chromatogram of Hop5.....	42
Figure 21. Superposition of two size exclusion chromatograms of Hop5 samples.....	43
Figure 22. Electrophoretic profile in SDS-PAGE 15 % of the purification samples of Hop5	43
Figure 23. IMAC chromatogram of Hemp1	44
Figure 24. Superposition of two size exclusion chromatograms of Hemp1 samples.....	44
Figure 25. Electrophoretic profile in SDS-PAGE 15 % of the purification samples of Hemp1	45

Figure 26. Native 15 % polyacrylamide gel with Late samples	46
Figure 27. Sedimentation profile of 0.79 mg/mL of Late in Buffer A, monitored by the absorbance at 280 nm as a function of the radial distance to the rotation axis.....	48
Figure 28. Continuous distribution of the sedimentation coefficients $c(s)$ of the Late sample.....	48
Figure 29. Sedimentation profile of 0.93 mg/mL of Theo in Buffer B, monitored by the absorbance at 280 nm as a function of the radial distance to the rotation axis.....	49
Figure 30. Continuous distribution of the sedimentation coefficients $c(s)$ of the Theo sample.....	50
Figure 31. Sedimentation profile of 0.65 mg/mL of Hop1 in buffer A, monitored by the absorbance at 280 nm as a function of the radial distance to the rotation axis.....	51
Figure 32. Continuous distribution of the sedimentation coefficients $c(s)$ of the Hop1 sample	51
Figure 33. Sedimentation profile of 0.84 mg/mL of Hop5 in buffer C collected from the third SEC peak, monitored by the absorbance at 280 nm as a function of the radial distance to the rotation axis	52
Figure 34. Continuous distribution of the sedimentation coefficients $c(s)$ of the Hop5 sample from the third SEC peak	53
Figure 35. Sedimentation profile of 0.41 mg/mL of Hop5 in buffer A collected from the fourth SEC peak, monitored by the absorbance at 280 nm as a function of the radial distance to the rotation axis.....	53
Figure 36. Continuous distribution of the sedimentation coefficients $c(s)$ of the Hop5 sample from the fourth peak	54
Figure 37. Sedimentation profile of 1.43 mg/mL of Hemp1 in buffer A collected from the second SEC peak, monitored by the absorbance at 280 nm as a function of the radial distance to the rotation axis.....	55
Figure 38. Continuous distribution of the sedimentation coefficients $c(s)$ of the Hemp1 sample from the second SEC peak	55
Figure 39. Late's thermal unfolding curves at different pHs.....	58
Figure 40. Far-UV circular dichroism spectra of Late at pH 4.0, measured under thermal denaturation (25 °C to 95 °C) in intervals of 10 °C	58
Figure 41. Far-UV circular dichroism spectra of Late at pH 7.0, measured under thermal denaturation (25 °C to 95 °C) in intervals of 10 °C	59
Figure 42. Far-UV circular dichroism spectra of Late at pH 9.5, measured under thermal denaturation (25 °C to 95 °C) in intervals of 10 °C	59
Figure 43. Thermal unfolding curves of Hop1 at different pHs.....	60
Figure 44. Far-UV circular dichroism spectra of Hop1 at pH 4.0, measured under thermal denaturation (25 °C to 95 °C) in intervals of 10 °C	60
Figure 45. Far-UV circular dichroism spectra of Hop1 at pH 7.0, measured under thermal denaturation (25 °C to 95 °C) in intervals of 10 °C	61

Figure 46. Far-UV circular dichroism spectra of Hop1 at pH 9.5, measured under thermal denaturation (25 °C to 95 °C) in intervals of 10 °C	61
Figure 47. Thermal unfolding curves of Hop5 at different pHs.....	62
Figure 48. Far-UV circular dichroism spectra of Hop5 at pH 4.0, measured under thermal denaturation (25 °C to 95 °C) in intervals of 10 °C	62
Figure 49. Far-UV circular dichroism spectra of Hop5 at pH 7.0, measured under thermal denaturation (25 °C to 95 °C) in intervals of 10 °C	63
Figure 50. Far-UV circular dichroism spectra of Hop5 at pH 9.5, measured under thermal denaturation (25 °C to 95 °C) in intervals of 10 °C	63
Figure 51. Thermal unfolding curves of Hemp1 at different pHs.....	64
Figure 52. Far-UV circular dichroism spectra of Hemp1 at pH 4.0, measured under thermal denaturation (25 °C to 95 °C) in intervals of 10 °C	64
Figure 53. Far-UV circular dichroism spectra of Hemp1 at pH 7.0, measured under thermal denaturation (25 °C to 95 °C) in intervals of 10 °C	65
Figure 54. Far-UV circular dichroism spectra of Hemp1 at pH 9.5, measured under thermal denaturation (25 °C to 95 °C) in intervals of 10 °C	65
Figure 55. Late crystals obtained in the crystallization condition containing 20 % w/v PEG 6000, 100 mM sodium citrate pH 4.0, 1 M lithium chloride and ≈12.4 mg/mL Late	69
Figure 56. Hop1 crystals obtained in the automated screening (hanging drop) after three weeks using ≈11 mg/mL Hop1	70
Figure 57. Hop5 crystals seen 1 week after the automated screening (hanging drop) made with the crystallization kit JBScreen PACT++ and ≈12.35 mg/mL Hop5.....	70
Figure 58. Hemp1 crystals obtained in the crystallization condition 20 % w/v PEG 3350, 200 mM di-ammonium citrate pH 5.0 with ≈12.9 mg/mL Hemp1 ...	71
Figure 59. Similarities between interacting helices in the crystal lattice from Hop1 crystals of different space groups	75
Figure 60. Interaction between Hop1 symmetry mates in the space group $C 2 2 2_1$ involving residues close to the first inhibitory loop	76
Figure 61. The packing of Trp80 in the $C 2 2 2_1$ Hop1 crystal	76
Figure 62. The packing of Trp80 in the $P 2 2_1 2_1$ Hop1 crystal	77
Figure 63. Solvent channels in the Hemp1 $P 6_5 2 2$ crystal packing.....	78
Figure 64. Hop1 structure solved from the $C 2 2 2_1$ crystal	79
Figure 65. $2F_o - F_c$ electron density map for the first inhibitory loop of Hop1 from the structure in the $C 2 2 2_1$ space group	80
Figure 66. Superposition of Hop1 crystal structures	80
Figure 67. Rotamers of Trp76 and Trp80 in the $C 2 2 2_1$ and $P 2 2_1 2_1$ space group Hop1 structures.....	81
Figure 68. β bulges in the Hop1 $C 2 2 2_1$ structure.....	82
Figure 69. Similarities between Hop1 and a cowpea phytocystatin (PBD code 4TX4)	83
Figure 70. Hemp1 structure solved from a hexagonal $P 6_5 2 2$ crystal	84

TABLE LIST

Table 1. Buffer composition and parameters used for SV-AUC fitting and analysis	47
Table 2. Analytical ultracentrifugation parameters obtained from the sedimentation data fitting for each phytocystatin sample	56
Table 3. Thermodynamic parameters for the unfolding of Hop1, Hop5 and Hemp1	66
Table 4. Data collection and refinement statistics for two Hop1 data sets	73

ABBREVIATION LIST

[θ]	Molar ellipticity
°	Degrees
°C	Degrees Celsius
Å	Ångström
A _{280nm}	Light absorption at 280 nm
APS	Ammonium persulfate
Arg	Arginine
Asn	Asparagine
Asp	Aspartic acid
C	Concentration
c(s)	Continuous distribution of sedimentation coefficients
cal	Calories
CC	Correlation coefficient
CD	Circular dichroism
cm	Centimeter
Da	Dalton
deg	Degree
dmol	Decimol
DT	Diphtheria toxin
<i>E. coli</i>	<i>Escherichia coli</i>
f/f ₀	Frictional ratio
F ₀	Experimentally measured amplitude
FAO	Food and Agriculture Organization of the United Nations
F _c	Model-based calculated amplitude
f _u	Unfolded protein fraction
g	Grams/relative centrifugal force
Gln	Glutamine
Glu	Glutamic acid
Gly	Glycine
h	Hours
I	Signal intensity
IMAC	Immobilized metal affinity chromatography
IPTG	Isopropyl β -D-1-thiogalactopyranoside
J	Joules
K	Kelvin
kcal	Kilocalories
kDa	Kilodaltons
K _{eq}	Equilibrium constant
L	Optical length

LB	Lysogeny Broth
Leu	Leucine
LLG	Log-likelihood gain
LNLS	Laboratório Nacional de Luz Síncrotron
Lys	Lysine
M	Molar
mdeg	Millidegrees
Met	Methionine
mg	Milligrams
min	Minutes
mL	Milliliters
mm	Millimeters
mM	Millimolar
MM	Molecular mass
MRW	Mean residue weight
Native-PAGE	Native polyacrylamide gel electrophoresis
ng	Nanograms
nm	Nanometers
OD	Optical density
PDB	Protein Data Bank
PEG	Polyethylene glycol
Pro	Proline
R	Ideal gas constant
rmsd	Root mean square deviation
rpm	Revolutions per minute
s	seconds/sedimentation coefficient
S	Svedberg
S _{20,W}	Standardized sedimentation coefficient
SDS	Sodium dodecyl sulfate
SDS-PAGE	Sodium dodecyl sulfate-polyacrylamide gel electrophoresis
SEC	Size exclusion chromatography
Ser	Serine
T	Temperature
TEMED	Tetramethylethylenediamine
TFZ	Translational function Z-score
Thr	Threonine
TLS	Translation/libration/screw
T _m	Melting temperature
Trp	Tryptophan
Tyr	Tyrosine
UV	Ultraviolet
v/v	Volume to volume ratio

Val	Valine
v_{bar}	Partial specific volume
w/v	Weight to volume ratio
α	Alpha
β	Beta
ΔG	Gibbs free energy
ΔG_{25}	Standard Gibbs free energy
ΔH	Enthalpy variation
ΔS	Entropy variation
η	Viscosity
θ	Ellipticity
μL	Microliters
μm	Micrometers
ρ	Density
σ	Sigma/standard deviation

SUMMARY

1. INTRODUCTION	14
1.1. PROTEASES	14
1.2 THE CYSTATIN SUPERFAMILY	15
1.2.1 Phytocystatins.....	17
1.3. DOMAIN SWAPPING	18
1.4. SOURCE ORGANISMS	20
1.4.1. Rubber Tree (<i>Hevea brasiliensis</i>)	20
1.4.2. Cocoa Tree (<i>Theobroma cacao</i>).....	21
1.4.3. Hop (<i>Humulus lupulus</i>)	22
1.4.4. Cannabis (<i>Cannabis sativa</i>).....	22
1.5. PHYTOCYSTATINS AND BIOTECHNOLOGY	23
2. JUSTIFICATION	24
3. GOALS	25
3.1. GENERAL GOALS.....	25
3.2. SPECIFIC GOALS	25
4. METHODOLOGY	26
4.1. <i>E. COLI</i> LEMO21 (DE3) TRANSFORMATION	26
4.2. GENE EXPRESSION.....	26
4.2.1. Gene expression in LB medium	26
4.2.2. Gene expression in auto-induction medium (ZYM-5052).....	27
4.3. PROTEIN PURIFICATION.....	28
4.3.1. Immobilized metal affinity chromatography (IMAC).....	28
4.3.2. Size exclusion chromatography (SEC)	28
4.4. THERMAL STABILITY ANALYSIS BY CIRCULAR DICHROISM (CD)..	29
4.5. OLIGOMERIC PROFILE CHARACTERIZATION.....	30
4.5.1. Native polyacrylamide gel electrophoresis	30
4.5.2. Analytical ultracentrifugation assays	30
4.6. CRYSTALLOGRAPHY.....	31
4.6.1. Crystallization assays	31
4.6.2. Crystal diffraction and structure resolution	32
5. RESULTS AND DISCUSSION	33
5.1. PROTEIN EXPRESSION IN <i>E. COLI</i> LEMO 21 (DE3)	33

5.2. PROTEIN PURIFICATION.....	37
5.3. OLIGOMERIC PROFILE CHARACTERIZATION.....	46
5.3.1. Native PAGE.....	46
5.3.2. Sedimentation Velocity Analytical Ultracentrifugation (SV-AUC)	46
5.4. THERMAL DENATURATION ASSAYS	57
5.5. CRYSTALLOGRAPHY.....	68
5.5.1. Protein Crystallization	68
5.5.2. X-Ray Diffraction and Crystal Structure Resolution	71
5.5.3. Crystal Packing.....	73
5.5.3.1. Hop1	74
5.5.3.2. Hemp1	77
5.5.4. Crystal Structure	78
6. CONCLUSIONS.....	85
7. REFERENCES.....	87

1. INTRODUCTION

One of the most essential characteristics of life is the ability of living beings to orchestrate pathways to overcome time barriers. A clear example comes from the tridimensional folding of proteins. The structure of a protein is directly related with its activity, and when being synthesized the protein must assume its correct conformation from a vast number of possibilities. If proteins were to randomly assume each possible conformation until its correct fold was reached, the amount of time taken would be insurmountable. This was illustrated in a thought experiment by Cyrus Levinthal in 1969: considering a simplified model of a protein composed of 150 residues, the amount of conformations the protein would have to sample would be around 10^{300} . Even if the protein changed conformations in picoseconds, this process would take longer than the age of the universe. Then how do proteins achieve a correct fold? Proteins are finely and naturally crafted to be able to overcome these time limitations by establishing local interactions that guide the folding of the peptide during the synthesis. A group of proteins is of utmost importance to assure that reactions or processes like the aforementioned occur in an ideal time frame: the enzymes.

1.1. PROTEASES

Proteases or peptidases are enzymes that catalyze the hydrolysis of peptide bonds in proteins. The catalysis is achieved by using a molecule to perform a nucleophilic attack to the carbon of the carboxyl group of a peptide bond.

According to the residue or molecule involved in the catalysis, proteases may be classified as cysteine, serine, threonine, glutamic, aspartic, metallo or asparagine proteases. As it suggests, cysteine, serine and threonine proteases receive these names as these residues are activated in the catalytic triad of the enzyme to perform the nucleophilic attack. Aspartic, glutamic and metalloproteases on the other hand use these residues, or a metal in the case of metalloproteases, to turn a water molecule nucleophilic. Asparagine proteases constitute the exception to the catalysis by hydrolysis performed by the other

proteases, as they cleave themselves by using an asparagine residue as the nucleophile to perform an elimination reaction, not requiring a water molecule (RAWLINGS et al., 2011).

Due to their activity, proteases are able to modify and degrade proteins, playing a role in several physiological processes such as digestion, programmed cell death, zymogen activation, defense against pathogens, blood coagulation, etc. Aside from its physiological importance, many proteases are used in the industry for being an inexpensive alternative to hydrolyzing peptide bonds, being used as detergent additives, meat tenderizers, to coagulate milk in cheese production, etc (SUMANTHA et al., 2016).

Like other enzymes, the activity rate of proteases is regulated not only by environmental factors, such as pH and temperature, but also by a series of mechanisms that affect the amount of available proteases. These include, among others, gene regulation, protein degradation, covalent modifications and interaction with protein inhibitors.

1.2 THE CYSTATIN SUPERFAMILY

The cystatin superfamily comprises proteins that share a conserved tridimensional fold and often act as cysteine-protease inhibitors, even though cystatins that do not show any inhibitory activity have already been identified (BROWN & DZIEGIELEWSKA, 1997). Members of this superfamily are found in a variety of eukaryotic organisms, bacteria and archaea (KORDIŠ & TURK, 2009).

The conserved tridimensional structure is composed by an amino-terminal alpha-helix surrounded by five antiparallel beta-strands, as well as two loops which are responsible for the inhibition activity of these molecules (**Figure 1**) (TURK et al., 2008). The first inhibitory loop presents the highly conserved sequence comprised of QxVxG, while the second loop usually has a proline residue followed by a tryptophan. A glycine residue in the flexible amino-terminal end is also essential for the protein inhibition (TURK & BODE, 1991).

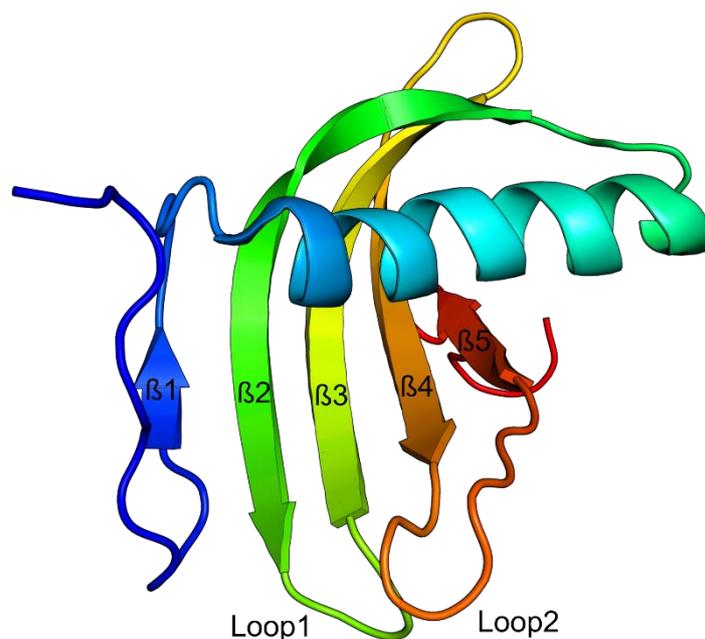


Figure 1. Conserved tridimensional cystatin structure elements represented by oryzacystatin-I (PDB code 1EQK) (NAGATA et al., 2000). The β strands and inhibitory loops are numerated according to the conventions.

Cystatins that are able to inhibit proteases bind competitively in a reversible process to papain-like or legumain-like cysteine-proteases, acting like a pseudo-substrate. In the case of papain-like proteases, this binding occurs with the introduction of both inhibitory loops and part of the N-terminal region in the active site of the target protease, as shown, for instance, in the structure of the tarocystatin-papain complex (**Figure 2**) or the stefin B-papain complex (STUBBS et al., 1990).

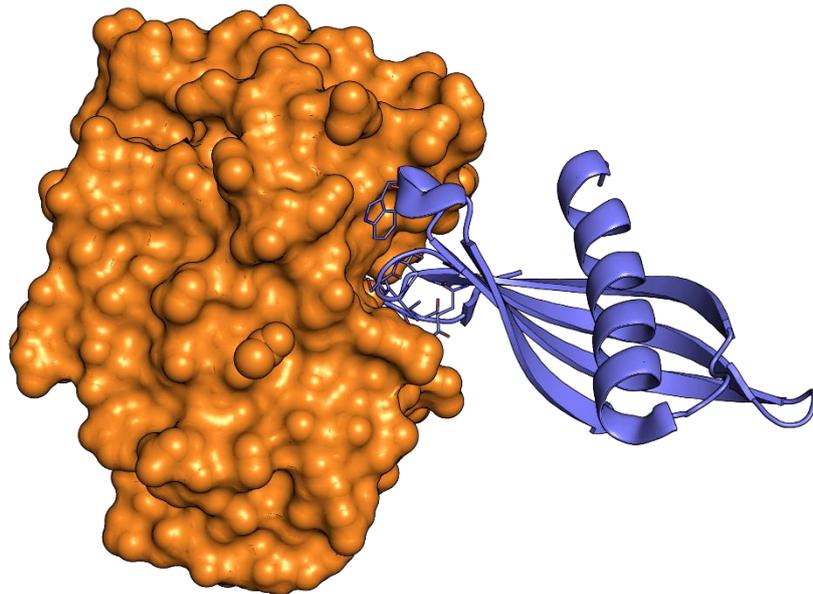


Figure 2. Inhibition of papain by tarocystatin (PDB code 3IMA). Papain is represented as the orange surface, while the purple cartoon represents tarocystatin. Residues essential to the protein inhibition are showed as sticks.

1.2.1 Phytocystatins

Regarded as another ramification in the superfamily, plant cystatins, also called phytocystatins, are identified by the consensus sequence [LVI]-[AGT]-[RKE]-[FY]-[AS]-[VI]-X-[EDQV]-[HYFQ]-N present in the amino-terminal region of the alpha-helix and usually by the absence of disulfide bonds or cysteine residues (MARGIS et al., 1998). Physiologically, phytocystatins take part in a variety of endogenous processes, such as programmed cell senescence (CARDOSO et al., 2015), nodulation (YUAN et al., 2016), seed germination (DIAZ-MENDOZA et al., 2016) and storage protein deposition (MUNGER et al., 2015).

Furthermore, these proteins act in the line of defense of plants, inhibiting proteases secreted by fungi and bacteria, as well as digestive proteases present in the midgut of herbivorous arthropods and insects, partaking in a coevolutionary process with their targets from other organisms (RYAN, 1990; MARTINEZ, 2016).

In solution, some cystatins are able to auto-associate, constituting various oligomers and eventually amyloid fibers (SANDERS et al., 2004). Some of these oligomers may display domain-swap, a structural feature in which secondary elements of a monomer take the place of the same structures in other monomer

of the protein. In this state, phytocystatins are unable to inhibit cysteine proteases, due to a change in the structure of the first inhibitory loop, as seen in the structure of the canecystatin-1 (VALADARES et al., 2013).

1.3. DOMAIN SWAPPING

One of the key characteristics of some phytocystatins is their ability to oligomerize into domain-swapped dimers. The term “domain-swapping” was first introduced by Bennett et al. in 1994 after studying the crystal structure of both monomeric and dimeric diphtheria toxin (DT). They noticed that the dimeric structure of DT strongly resembled two adjacent monomers, in which the same domains were present. However, the receptor binding domain of each monomer within the dimeric structure seemed to have switched places with its identical counterpart from the other molecule (**Figure 3**). This way, although some of the intramolecular interactions became intermolecular, the same interactions from the monomeric structure were present in the dimer (BENNETT et al., 1994).

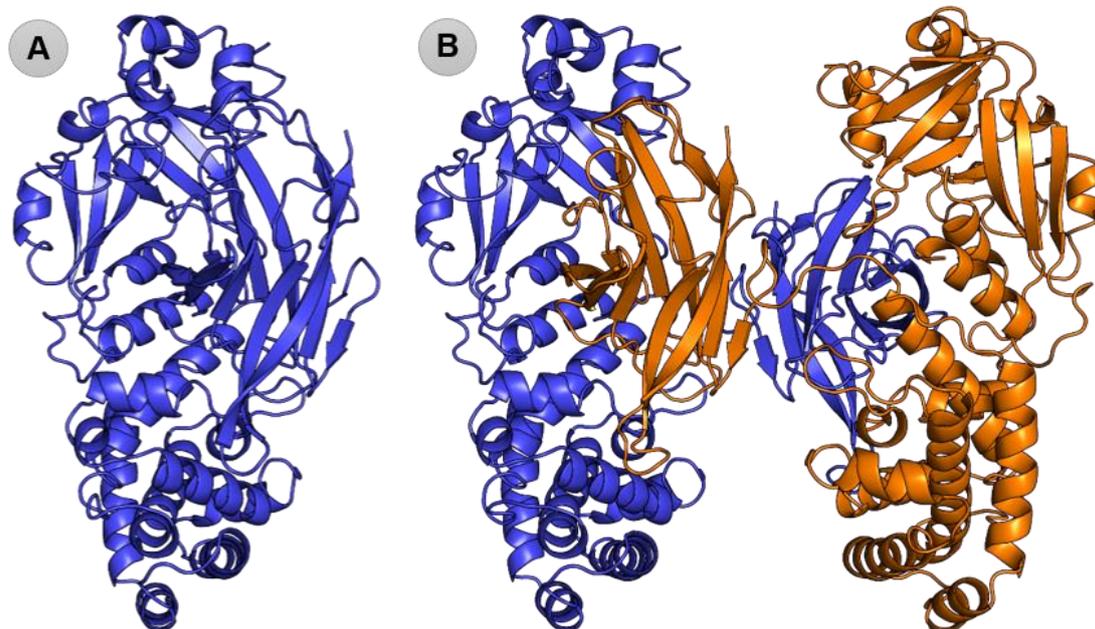


Figure 3. Monomer (PDB code 1MDT) (A) and domain swapped dimer (PDB code 1DDT) (B) of DT (BENNETT & EISENBERG, 1994; BENNETT et al, 1994). Each monomer is showed in a single color. The same interactions in the monomeric structure are also present in each lobe of the dimer

The mechanism of domain swapping involves favoring the partial unfold of monomers, by pre-denaturing conditions or mutations, and the subsequent interaction of these open monomers (**Figure 4**). A hinge loop in the interdomain interface connects the swapped domain to its native monomer and assumes a different conformation when the protein undergoes oligomerization, at times forming new interactions not present in the monomer. Bennett et al. also coined two terms: the closed interface and the open interface. Closed interfaces correspond to the interdomain interfaces present in the closed monomer (red region in **Figure 4A**). On the other hand, the open interface occurs only between domains of the open monomers in the domain swapped oligomer (red region in **Figure 4C**) (BENNETT et al., 1995).

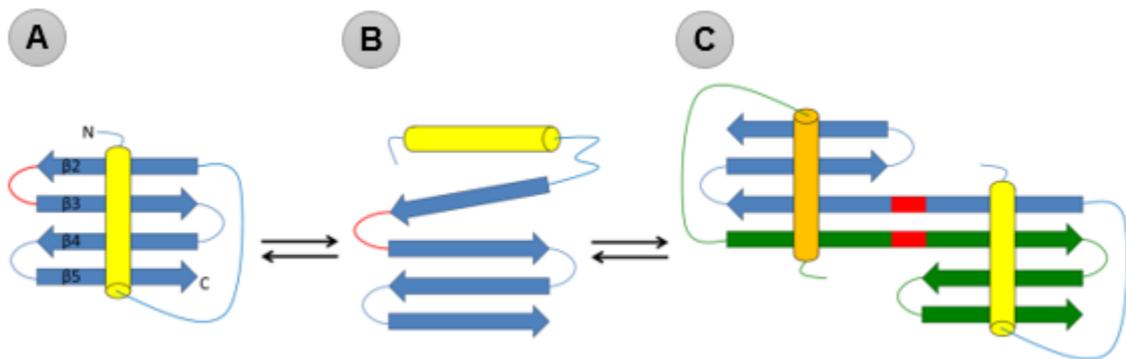


Figure 4. The domain swapping mechanism. **(A)** Closed monomer of a common cystatin, which may open due to pre-denaturing conditions **(B)**, and assembled into a domain-swapped dimer with another open monomer **(C)**. In **(A)** the red region represents the closed interface, while in **(C)** it represents the open interface. Image adapted from Valadares et al., 2013.

Domain swapping may have had a role in the evolution of some oligomers. Since domain swapped dimers maintain the structure of its composing monomers, it could serve as a shortcut for the formation of an initial interaction surface between the two proteins while still maintaining the activity of each monomer, avoiding the necessity of several random mutations which could negatively affect the protein activity (BENNETT & EISENBERG, 2004).

Although the underlying reasons for the formation of domain swapped phytocystatin dimers in plants are currently under research, this phenomenon is of utmost importance for various physiological processes, since one of the inhibitory loops of the monomer is also a hinge loop. As the protein dimerizes,

the closed interface becomes open, which, consequently, give rise to new interactions of the inhibitory loop with the other molecule. The constraints imposed by these new interactions compromise the inhibitory activity of the phytocystatin, as the conformation of the first inhibitory loop changes and it cannot interact with the protease the same way as the monomer (VALADARES et al., 2013).

1.4. SOURCE ORGANISMS

The proteins selected for the present studies were chosen based on the economic relevance of the source organisms to Brazil and other countries, as well as on their sequence homology (**Figure 5**). The analyzed proteins and their sources are: Late (*Hevea brasiliensis*), Theo (*Theobroma cacao*), Hop1 (*Humulus lupulus*), Hop5 (*Humulus lupulus*) and Hemp1 (*Cannabis sativa*).

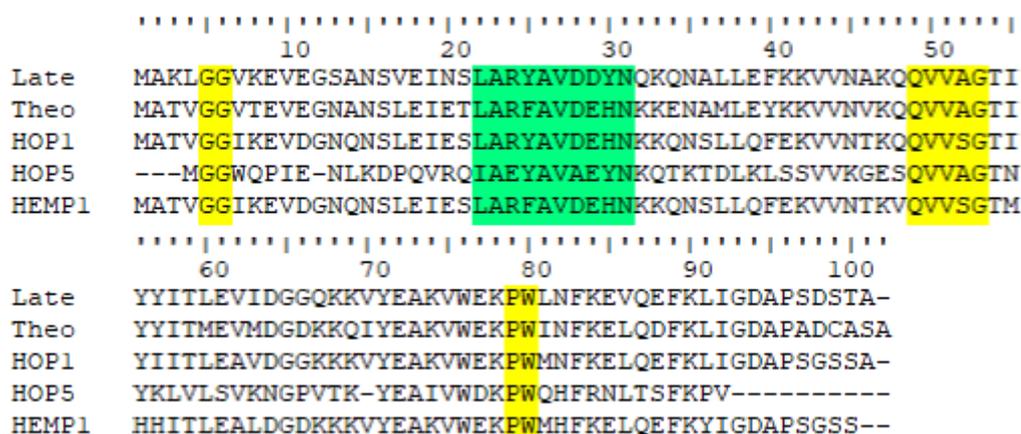


Figure 5. Sequence alignment of the proteins in this study. The consensus sequence that characterizes phytocystatins is shaded in green, while the conserved amino acids that take part in the protein inhibition are shaded in yellow

1.4.1. Rubber Tree (*Hevea brasiliensis*)

A member of the Euphorbiaceae family, *Hevea brasiliensis*, commonly known as rubber tree, is a species native to northern South America, although it is currently found in other tropical and subtropical countries. Its economic value comes primarily from the use of the latex extracted from the tree as a material for natural rubber production, but the tree can also be sawn for its wood, and the oils

contained in its seeds can be used in the production of soap and paints (BOER et al., 2000).

Since the 20th century, Brazil has been suffering a huge impact in the latex production due to the pathogenic fungus *Microcyclus ulei*, responsible for leaf blight in rubber tree crops in all of South America (GASPAROTTO et al., 1990). As a consequence, only countries whose rubber productions are not hampered by this disease, such as the ones located in South and Southeast Asia, are able to export rubber worldwide. Despite that, according to data from the Food and Agriculture Organization of the United Nations (FAO) regarding 2016, Brazil produced 189,377 tons of rubber, ranking as the twelfth biggest natural rubber producer in the world.

Efforts are being dedicated to protect new cultures from these fungi, especially cultures in places still not affected by these plagues, due to the devastating aftermath of these pathogenic agents.

1.4.2. Cocoa Tree (*Theobroma cacao*)

Among the selected organisms, *Theobroma cacao* (cocoa tree) is one of the most beloved worldwide. A member of the Malvaceae family and native to South America, its economic relevance comes from the use of its beans in the confection of innumerable delicacies and dishes, the most famous one being the chocolate.

Similarly to rubber trees, cocoa production in Brazil is also significantly hampered by a pathogenic fungus, namely, *Moniliophthora perniciosa*, which is responsible for witches' broom disease in cocoa trees (PEREIRA et al., 1996). The mechanism of infection involves surpassing the line of defense provided by phytocystatins by favoring their dimerization by modifications in the pH, consequently inactivating them (FREITAS et al., 2015). Moreover, recent studies have also shown that cocoa bean yield in Brazil may be severely affected by droughts such as the ones caused by El Niño-Southern Oscillations (GATEAU-REY et al., 2018). Nevertheless, cocoa trees are still of economic importance for the country: according to data from FAO for 2016, Brazil was the sixth biggest cocoa producer in the world (213,843 tons of cocoa beans produced).

1.4.3. Hop (*Humulus lupulus*)

Hop (*humulus lupulus*), a member of the Cannabaceae family, is best known for its use in beer brewing. As the basis of beer production consists of the fermentation of cereal starch, which is somewhat insipid, the flavor of beer is dependent on additive agents. Furthermore, the low alcohol concentration of beer leads to a shorter shelf-life when compared to other alcoholic beverages. The use of hop becomes handy as the plant adds a distinctive bitter taste while also bestowing antibacterial properties to the brew, delaying the spoilage of the product (BEHRE, 1999).

According to data from the Barth-Haas group for the year of 2017 (The Barth Reports 2017/2018), Brazil produced around 14 million kiloliters of beer, ranking as the third biggest beer producer in the world, losing only to China and the USA.

1.4.4. Cannabis (*Cannabis sativa*)

Another member of the Cannabaceae family, cannabis (*Cannabis sativa*) is surrounded by controversy. Despite its cultivation and/or use being illegal in several countries, cannabis is valued for the many psychoactive compounds found in the plant, known as cannabinoids, which are used for recreational and medicinal purposes.

Cannabinoids are being heavily researched as their use is evaluated as a complementary alternative medicine to neurodegenerative disorders such as Alzheimer's disease, multiple sclerosis and Parkinson's disease (RUSSO, 2018). Furthermore, the use of medicinal cannabis is starting to take the place of some prescription drugs among cannabis users in the USA, due to the lower cost, fewer side effects and better effectiveness (KRUGER & KRUGER, 2019).

It is rather difficult to define quantitatively the production of cannabis worldwide as in many countries its cultivation is still illegal and therefore lacks any effective regularization or market study. However, according to data from the World Drug Report, published by the United Nations Office on Drugs and Crime regarding 2016, cannabis had around 192 million users, earning the status of the world's most commonly used drug that year.

1.5. PHYTOCYSTATINS AND BIOTECHNOLOGY

Phytocystatins have been intensively studied not only for its physiological importance to plants but also for its possible biotechnological application as putative substitutes to pesticides in the protection of cultures against pests, and also in resistance improvement to abiotic stress factors, such as drought, temperature and wounds (KUNERT et al., 2015; HWANG et al., 2010).

Due to the low availability of the source organisms and the complexity of isolating proteins directly from the plants, the use of heterologous expression systems presents itself as a more viable option to obtain the recombinant proteins, both economically and time-wise. In the present work, the genes that code for the target proteins were used to transform competent *Escherichia coli* cells which in turn were used in large scale expression assays.

2. JUSTIFICATION

The ever increasing demand for agricultural products and their derivatives renders it necessary to increase their yield through the use of genetic manipulations. When improving a culture, one should always have in mind the improvement of its resistance to both biotic and abiotic stress factors, as crops are susceptible to pests and changing weather conditions, such as the ones caused by global warming, which have already caused great losses in the soybean production in the USA (MOURTZINIS et al., 2015).

The resolution of different phytocystatins structures is fundamental to the understanding of their mechanisms of action, especially regarding the effect of amino acid variability in the interaction of the protein with their substrates and to their structural stability. Therefore, techniques to create organisms which are more resistant to plagues and abiotic stress factors may be developed aimed at specific goals.

In this context, the present work aims at providing a base for the use and development, by genetic recombination, of phytocystatins with structural stability and that are more suitable to endure environmental stresses and microorganism attacks.

3. GOALS

3.1. GENERAL GOALS

Express and characterize biophysically the phytolectins Late (*Hevea brasiliensis*), Theo (*Theobroma cacao*), Hop1 (*Humulus lupulus*), Hop5 (*Humulus lupulus*) and Hemp1 (*Cannabis sativa*) and solve their structures by X-ray diffraction.

3.2. SPECIFIC GOALS

- Express the proteins using a heterologous expression system (*E. coli*)
- Purify the proteins by immobilized metal affinity and size exclusion chromatography
- Study their thermal stability by circular dichroism assays
- Identify oligomer populations found in the purified samples
- Identify and refine the conditions that favor the formation of protein crystals
- Collect X-ray diffraction data from the crystals
- Solve the structures using the molecular replacement method
- Analyze the solved structures

4. METHODOLOGY

4.1. *E. COLI* LEMO21 (DE3) TRANSFORMATION

The genes that codify the target proteins (Late, Theo, Hop1, Hop5 and Hemp1) were cloned in pET-24a vectors by the company GenOne Biotechnologies and provided by Prof. Dr. Valadares. These vectors were chosen for allowing the expression of the proteins with a C-terminal histidine tail (containing the residues LEHHHHHH). Approximately 200 ng of the plasmid were added to 100 μ L of *E. coli* Lemo21 (DE3) competent cell containing 100 mM CaCl_2 and 15 % v/v glycerol. The cells were incubated in ice for 40 min and were afterward brought to a thermal shock at 42 °C for 90 s. Then, the cells were supplemented with 900 μ L of LB medium (1 % w/v peptone, 0.5 % w/v yeast extract and 1 % w/v NaCl) and incubated at 37 °C under agitation (220 rpm) for 1 h. Subsequently, 500 μ L of this culture were added to Petri dishes prepared with agar 2 % w/v LB medium, 100 μ g/mL kanamycin and 34 μ g/mL chloramphenicol. Transformed cells were selected considering the resistance of the Lemo21 (DE3) strains and the vector pET-24a to chloramphenicol and to kanamycin, respectively.

4.2. GENE EXPRESSION

4.2.1. Gene expression in LB medium

An isolated colony with the transformed strains was inoculated in approximately 5 mL of LB medium supplemented with 100 μ g/mL kanamycin and 34 μ g/mL chloramphenicol and was incubated under agitation (220 rpm) at 37 °C for 15 h. Then, 1 mL of this culture was inoculated in 25 mL of LB medium containing 100 μ g/mL kanamycin and 34 μ g/mL chloramphenicol and kept under agitation (220 rpm) at 37 °C. When the 600 nm optical density of the culture reached 0.8, the expression was induced by isopropyl β -D-1-thiogalactopyranoside (IPTG) at a concentration of 0.5 mM. The expression assay lasted until 6 h after the induction.

The expression and solubility of the recombinant proteins were confirmed by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) assays. The separating gel was prepared with 15 % w/v polyacrylamide (14.5 % acrylamide and 0.5 % bis-acrylamide), 250 mM Tris-HCl pH 8.8, 0.3 % w/v SDS, 0.1 % w/v ammonium persulfate (APS) and 3 μ L tetramethylethylenediamine (TEMED). The stacking gel was prepared with 5 % w/v polyacrylamide (4.85 % acrylamide and 0.15 % bis-acrylamide), 140 mM Tris-HCl pH 6.8, 0.3 % w/v SDS, 0.1 % w/v APS and 3 μ L TEMED.

The cells were resuspended in 4 mL of standard buffer (150 mM NaCl and 75 mM tris pH 7.6) and lysed by sonication for 80 s in 10 s pulses of 20 J, with 30 s intervals. The lysate was centrifuged at 17 400 \cdot g and 4 $^{\circ}$ C for 30 min. The samples were then supplemented with Laemmli buffer (2 % w/v SDS, 2 mM 2-mercaptoethanol, 0.01 % w/v bromophenol blue, 40 % v/v glycerol and 40 mM Tris-HCl pH 6.8) and heated at 95 $^{\circ}$ C for 10 min.

The electrophoresis was performed using an EPS301 electrophoresis power supply (Amersham Pharmacia Biotech) and running buffer containing 25 mM Tris, 192 mM glycine, 0.1 % w/v SDS at pH \approx 8.3. The applied current and electric tension were limited to 35 mA and 220 V, respectively. The gels were stained with Coomassie Blue (0.1 % w/v Coomassie brilliant blue R-250, 50 % v/v methanol and 10 % v/v acetic acid) and destained with 20 % v/v acetic acid.

Higher amounts of soluble protein were obtained inoculating the isolated colony with the transformed strains in 500 mL of LB medium. After that, the culture was centrifuged at 4 340 \cdot g and 4 $^{\circ}$ C for 15 min and the resulting pellet was stored at -4 $^{\circ}$ C for posterior protein purification.

4.2.2. Gene expression in auto-induction medium (ZYM-5052)

An isolated colony containing the transformed strains was inoculated in approximately 5 mL of LB medium supplied with 100 μ g/mL kanamycin and 34 μ g/mL chloramphenicol and was incubated under agitation (220 rpm) at 37 $^{\circ}$ C for 15 hours. Then, 1 mL of this culture was inoculated in 25 mL of auto-induction medium ZYM-5052 containing 100 μ g/mL kanamycin and 34 μ g/mL chloramphenicol. The ZYM-5052 medium was prepared as described by Studier (2005), using 100 μ M FeCl₃ as trace metal. The cultures were incubated under

agitation at 37 °C and 220 rpm for 8 to 12 hours. The cultures were centrifuged at 4 °C and 4 340 ·g for 15 min, the supernatant was discarded and the remaining cell pellet was frozen at -4 °C. The same SDS-PAGE procedure detailed in section 4.2.1 was followed for the analysis of the protein expressed in ZYM-5052 medium.

Due to the super-expression indicated by SDS-PAGE of all five proteins in their soluble forms, the assay was repeated inoculating 500 ml of ZYM-5052 medium. The obtained cell pellets were stored at -4 °C for posterior protein purification.

4.3. PROTEIN PURIFICATION

4.3.1. Immobilized metal affinity chromatography (IMAC)

The pellets obtained in the large scale expression assays were resuspended in 45 mL of standard buffer and lysed by 25 min of sonication with output control of 3 and duty cycle of 30 %. The lysate was centrifuged at 17 400 ·g at 4 °C for 30 min in order to isolate the soluble fraction present in the supernatant. The supernatant was filtered through nitrocellulose membranes with pore sizes of 0.45 µm and supplemented with 15 mM imidazole.

The chromatography was performed using a 1 mL HisTrap™ HP (GE Healthcare) Ni²⁺ column connected to an ÄKTA Prime (GE Healthcare) equipment. The filtered samples applied into the column and the fractions were eluted at different concentrations of elution buffer (300 mM NaCl, 75 mM Tris pH 8.0 and 500 mM imidazole) and consequently, different concentrations of imidazole, in order to remove contaminants that interacted unspecifically with the resin. The target proteins were eluted at 500 mM imidazole. The light absorption at 280 nm of the eluted samples was monitored and the corresponding peaks were collected for SDS-PAGE analysis.

4.3.2. Size exclusion chromatography (SEC)

The samples obtained by IMAC were further purified by size exclusion chromatography. The samples were injected and eluted in a Superdex™ 75

Increase 10/300 GL (GE Healthcare) column previously equilibrated in standard buffer at a flow rate of 0.3 mL/min using an ÄKTA Purifier (GE Healthcare). Samples corresponding to the different absorption peaks seen in the chromatogram were collected for SDS-PAGE analysis.

4.4. THERMAL STABILITY ANALYSIS BY CIRCULAR DICHROISM (CD)

The thermal stability of the proteins was evaluated by circular dichroism (CD) spectroscopy using the Jasco J-815 spectropolarimeter (JASCO Inc.) and a Peltier heating system. Samples were prepared in different buffer solutions (2 mM sodium acetate pH 4.0, 2 mM Tris-HCl pH 7.0 and 2 mM glycine pH 9.5), in the concentrations of 0.080 mg/mL for Late, 0.091 mg/mL for Hop1, 0.071 mg/mL for Hop5 and 0.096 mg/mL for Hemp1. Thermal unfolding curves and Far-UV CD (195 to 260 nm) spectra were recorded from 25 °C to 95 °C at a rate of approximately 0.2 °C/min using a 1 mm quartz cuvette. The unfolding curves were recorded at 218 nm for Late, Hop1 and Hop5, and at 220 nm for Hemp1. For each Far-UV CD spectrum, three consecutive spectra were measured and their mean values were recorded. The displayed spectra and unfolding curves were smoothed with the Savitzky–Golay method using ORIGINPRO 8. The mean ellipticities were converted to molar ellipticity prior to data analysis according to equation 1:

$$[\theta] = \frac{\theta * MRW}{10 * L * C} \quad (1)$$

where MRW is the mean residue weight in g/mol, θ is the measured ellipticity in mdeg, L is the optical length of the cuvette in cm and C is the protein concentration in mg/mL.

An estimate of the proteins' melting temperatures (T_m), i.e. the temperature in which only half of the molecules is folded, was obtained by the sigmoidal fit of the unfolding curves performed by ORIGINPRO 8 using the Biphasic Dose Response function. The thermodynamic parameters were estimated from the unfolding curves taking into account the unfolding fractions of proteins (f_U) and Van't Hoff approximation, as following.

$$f_U = \frac{y_N - [\theta]}{y_N - y_U} \quad (2)$$

where y_N and y_U represent the characteristic values of $[\theta]$ of the folded and unfolded states, respectively. The equilibrium constant (K_{eq}) for the thermal unfolding was then calculated from:

$$K_{eq} = \frac{[U]}{[N]} = \frac{f_U}{1 - f_U} \quad (3)$$

Then, the unfolding enthalpy (ΔH) and entropy (ΔS) variations were estimated and the Gibbs free energy was calculated following a plot of $R \cdot \ln K_{eq}$ (where R is the universal gas constant) against the temperature (T) for each curve and using equation 4:

$$\Delta G = -R * T * \ln K_{eq} = \Delta H - T\Delta S \quad (4)$$

4.5. OLIGOMERIC PROFILE CHARACTERIZATION

4.5.1. Native polyacrylamide gel electrophoresis

Native PAGE gels were prepared according to the protocol described in section 4.2.1. for SDS-PAGE 15 % gels, with the exception of the use of SDS, which was replaced by distilled water. Late samples obtained from SEC were supplemented with native page sample buffer (62.5 mM Tris-HCl pH 6.8, 40 % v/v glycerol and 0.01 % w/v bromophenol blue) prior to the run. The electrophoresis procedure was executed using an EPS301 electrophoresis power supply (Amersham Pharmacia Biotech) inside a refrigerated 4 °C chamber, using a pH 7.5 running buffer (25 mM Tris and 192 mM glycine). The current and electric tension applied to the gel were limited to 35 mA and 220 V, respectively. The gel was stained in Coomassie Blue and destained in 20 % v/v acetic acid.

4.5.2. Analytical ultracentrifugation assays

Sedimentation velocity assays of the proteins were performed at 20 °C and 42 000 rpm (for putative dimer samples) and 45 000 rpm (for putative monomer samples) in a ProteomeLab XL-A (Beckman Coulter) analytical ultracentrifuge,

using an AN-60 Ti (Beckman Coulter) rotor. The protein samples purified by IMAC followed by SEC at concentrations corresponding to approximately 0.4, 0.75 and 1.0 OD_{280nm} were analyzed in aluminum cells with double sample channels and quartz windows. The absorbance of the sample at 280 nm was measured as a function of the radial distance do the rotation axis in intervals of 5 min.

The SEDFIT 15.01b software was employed to perform sedimentation data fittings of Lamm's equation based on the continuous distribution of sedimentation coefficients $c(s)$, using values estimated by the SEDNTERP 2 software for the proteins' partial specific volume (v_{bar}), buffer density (ρ) and buffer viscosity (η), leaving the frictional ratio (f/f_0) as the parameter to be determined by the program. The contribution of each species to the measured signal was determined by the area below the $c(s)$ curves.

4.6. CRYSTALLOGRAPHY

4.6.1. Crystallization assays

Crystallization assays were performed by vapor diffusion through the hanging drop method. Using the pipetting robot mosquito[®] HTS (TTP Labtech) and 96-well plates, volumes of 100 nL to 200 nL of pure protein (after SEC) in different concentrations were mixed to 100 nL to 200 nL of solutions from the sparse matrix crystallization screening kits Crystal Screen (Hampton Research), JBScreen JCSG++ (Jena Bioscience) and JBScreen PACT++ (Jena Bioscience) to a final volume of 300 nL or 400 nL per droplet (JANCARIK & KIM, 1991). For each crystallization solution, the screening was performed with 3 different protein concentrations. After the appearance of microcrystals in a specific crystallization condition, its refinement was carried out varying the pH and concentration of its components in sitting drop crystallization assays in 24-well plates in order to obtain crystals suitable for X-ray diffraction.

4.6.2. Crystal diffraction and structure resolution

X-ray diffraction data of putative protein crystals were collected in the MX2 beamline of the Laboratório Nacional de Luz Síncrotron (LNLS) in Campinas, São Paulo.

The XDS software (KABSCH, 2010) was employed to process the diffraction images, including defining the background of the images, indexing and integrating the diffraction spots, scaling the reflections intensities and assigning space groups to the crystals. 5 % of the indexed reflections were flagged to be left out of the structure refinement steps in order to be used as the R_{free} reflection set, an unbiased indicator of the refinement quality, and prevent data overfitting (BRÜNGER, 1992).

The number of protein molecules in the asymmetric unity of each crystal was estimated from the calculation of their Matthews coefficient, inputting the unit cell parameters, space group and protein molecular weight for each crystal in the MATTHEWS_COEF software of the CCP4 7.0 (WINN et al., 2011) software suite. The Matthews coefficient, i.e. the crystal volume per unit of protein molecular weight, is empirically shown to assume specific value ranges. Therefore, we are able to estimate the macromolecular content of the unit cell by calculating the integer amount of molecules in the asymmetric unit that give rise to the most probable Matthews coefficient values (MATTHEWS, 1968; KANTARDJIEFF & RUPP, 2013).

Molecular replacement (MR) was employed with PHASER (MCCOY et al., 2007) of the CCP4 software suite to retrieve phase information using a cowpea (*Vigna unguiculata*) cystatin with around 70 % sequence identity as a search model (PDB code 4TX4) (JÚNIOR et al., 2017). The structure refinement was performed in cycles, using PHENIX 1.13 (ADAMS et al., 2010) for reciprocal space refinements and then COOT 0.8.9 for real space refinements. In the first refinement cycle of each structure, PHENIX was employed to perform a simulated annealing as a means to remove search model biases.

5. RESULTS AND DISCUSSION

5.1. PROTEIN EXPRESSION IN *E. COLI* LEMO 21 (DE3)

The initial expression assays focused on the proteins Late, a phytocystatin from *Hevea brasiliensis*, and Theo, a phytocystatin from *Theobroma cacao*, using LB medium as described in the methodology. As showed in the **Figure 6**, the production of Late and Theo was qualitatively low when expressed in LB medium. In contrast, both proteins were super-expressed in an auto-inducing medium, namely, the medium ZYM-5052 (STUDIER, 2005) (**Figures 7 and 8**). These results were expected since it was well documented by Busso et al. (2008) that the protein yield when using auto-inducing media is higher in comparison to LB medium. Due to the early success in the expression assays and the unpredictability of protein crystallization, one of the core techniques in this study, we decided to include 4 other proteins in our studies: Hop1, Hop3 and Hop5, phytocystatins from *Humulus lupulus*, and Hemp1, a phytocystatin from *Cannabis sativa*. **Figures 7 to 10** indicate that all target proteins, with the exception of Hop3, were successfully expressed in the auto-inducing medium ZYM-5052. Furthermore, a comparison of the lysate with the lysis supernatant shows that these proteins were produced in their soluble forms, rendering renaturation steps after the protein purification unnecessary. As a means of standardizing, all subsequent expression assays were performed for 10 hours at 37 °C and 220 rpm using the auto-inducing medium ZYM-5052.

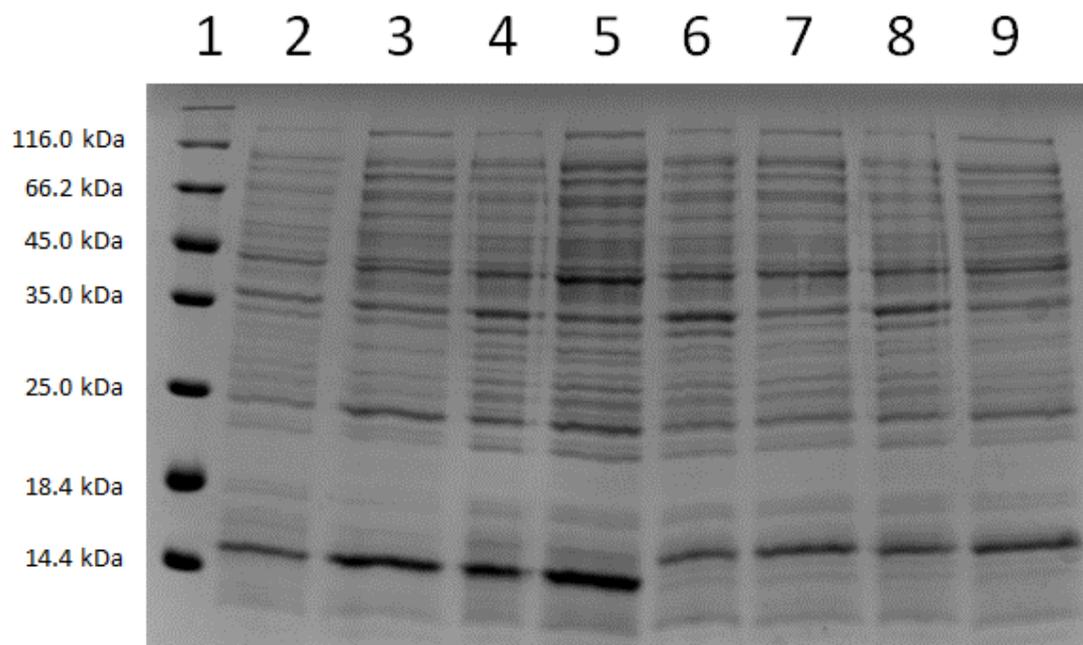


Figure 6. Electrophoretic profile in SDS-PAGE 15 % of the expression and solubility of Late (intense band around 14.4 kDa in lanes 2 to 5) and Theo (intense band around 14.4 kDa in lanes 6 to 9) in LB medium. 1- Molecular weight marker; 2- Lysate after 4 h of Late expression; 3- Lysis supernatant after 4 h of Late expression; 4- Lysate after 6 h of Late expression; 5- Lysis supernatant after 6 h of Late expression; 6- Lysate after 4 h of Theo expression; 7- Lysis supernatant after 4 h of Theo expression; 8- Lysate after 6 h of Theo expression; 9- Lysis supernatant after 6 h of Theo expression.

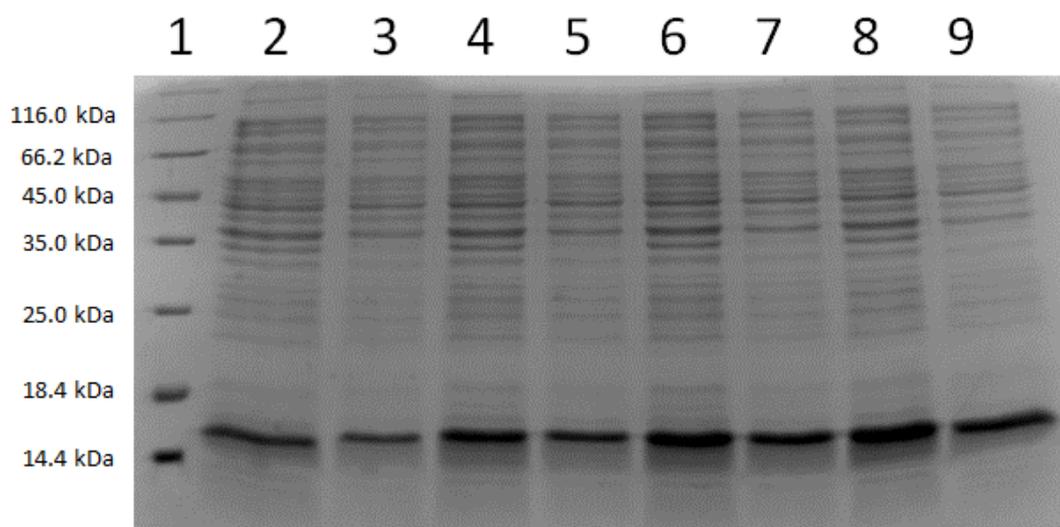


Figure 7. Electrophoretic profile in SDS-PAGE 15 % of the expression and solubility of Late (intense band around 14.4 kDa) in auto-inducing medium ZYM-5052. The molecular weight of Late is 12.3 kDa. 1- Molecular weight marker; 2- Lysate after 7 h of expression; 3- Lysis supernatant after 7 h of expression; 4- Lysate after 8 h of expression; 5- Lysis supernatant after 8 h of expression; 6- Lysate after 9 h of expression; 7- Lysis supernatant after 9 h of expression; 8- Lysate after 10 h of expression; 9- Lysis supernatant after 10 h of expression.

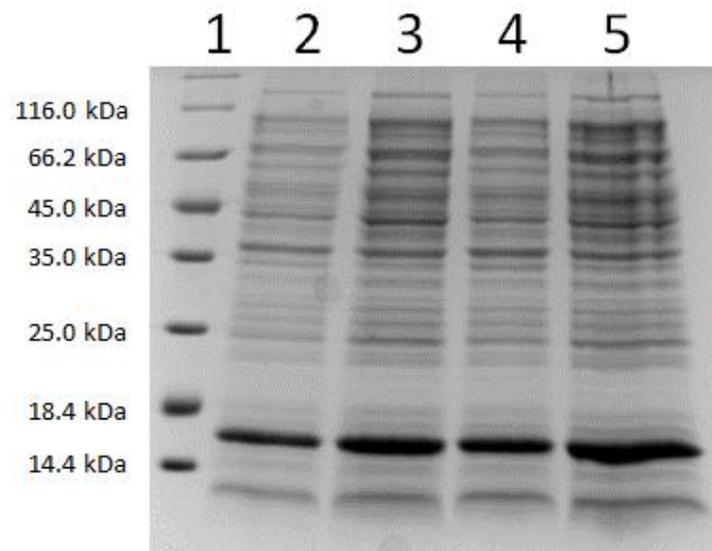


Figure 8. Electrophoretic profile in SDS-PAGE 15 % of the expression and solubility of Theo (intense band around 14.4 kDa) in auto-inducing medium ZYM-5052. The molecular weight of Theo is 12.5 kDa. 1- Molecular weight marker; 2- Lysate after 10 h of expression; 3- Lysis supernatant after 10 h of expression; 4- Lysate after 12 h of expression; 5- Lysis supernatant after 12 h of expression.

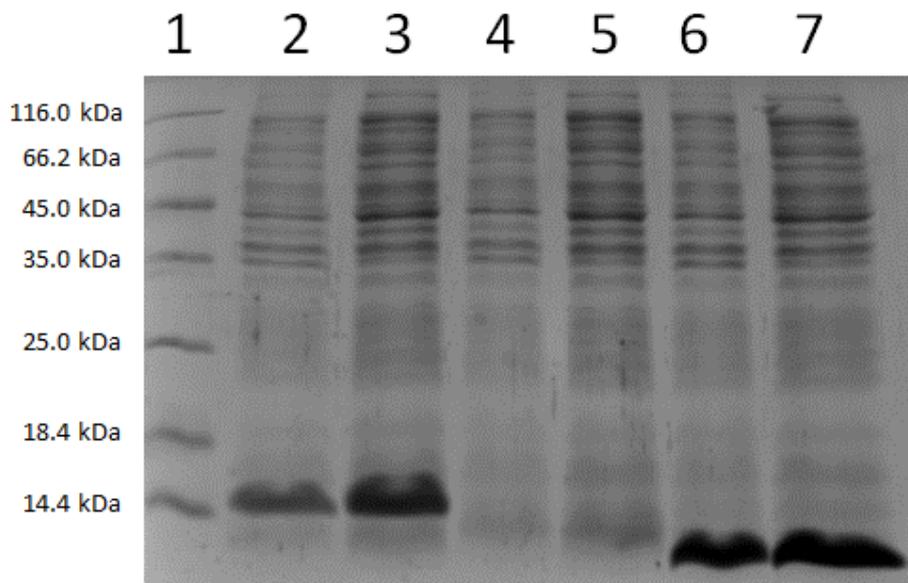


Figure 9. Electrophoretic profile in SDS-PAGE 15 % of the expression and solubility of Hop1 (intense band around 14.4 kDa in lanes 2 and 3), Hop3 and Hop5 (intense band below 14.4 kDa in lanes 6 and 7) in auto-inducing medium ZYM-5052. 1- Molecular weight marker; 2- Lysate after 10 h of Hop1 expression; 3- Lysis supernatant after 10 h of Hop1 expression; 4- Lysate after 10 h of Hop3 expression; 5- Lysis supernatant after 10 h of Hop3 expression; 6- Lysate after 10 h of Hop5 expression; 7- Lysis supernatant after 10 h of Hop5 expression.

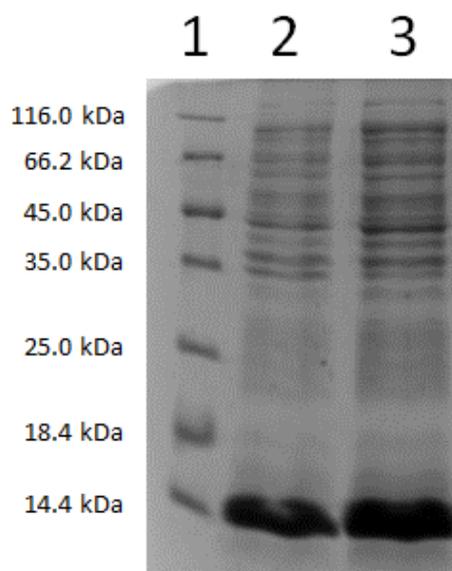


Figure 10. Electrophoretic profile in SDS-PAGE 15 % of the expression and solubility of Hemp1 (intense band around 14.4 kDa) in auto-inducing medium ZYM-5052. 1- Molecular weight marker; 2- Lysate after 10 h of expression; 3- Lysis supernatant after 10 h of expression.

5.2. PROTEIN PURIFICATION

All proteins were purified by immobilized metal affinity chromatography (IMAC) followed by size exclusion chromatography (SEC). The IMAC chromatograms are showed in **Figures 11** (Late), **14** (Theo), **17** (Hop1), **20** (Hop5) and **23** (Hemp1). In each IMAC, 2 pellets obtained from 500 mL of culture from the expression assays were resuspended in 45 mL of standard buffer each and injected in the HiTrap columns which were then washed with 50 mM imidazole to remove contaminants prior to the protein elution with 500 mM imidazole.

Size exclusion chromatography was then employed to improve the purity of samples collected from IMAC. In **Figures 12** (Late), **15** (Theo), **18** (Hop1), **21** (Hop5) and **24** (Hemp1) representative SEC chromatograms are showed.

The collected purification samples are seen in the SDS-PAGE 15 % gels from **Figures 13** (Late), **16** (Theo), **19** (Hop1), **22** (Hop5) and **25** (Hemp1). The SDS-PAGE revealed that at portion of all proteins was eluted in the flow through and in the column washing with 50 mM imidazole; nevertheless, most of the protein was eluted only at 500 mM imidazole. Late and Hop1 presented a qualitative purity after size exclusion chromatography, while Hemp1 and Hop5 exhibited only a small contaminant band in the vicinity of the 25.0 kDa molecular weight marker. The presence of several contaminants in the purified Theo samples rendered the realization of circular dichroism assays impossible and had a negative impact on the protein crystallization.

All purified proteins were eluted in more than one peak in the size exclusion chromatography, suggesting the presence of different protein oligomers formed by self-association, as already described for other phytocystatins (VALADARES et al., 2013)

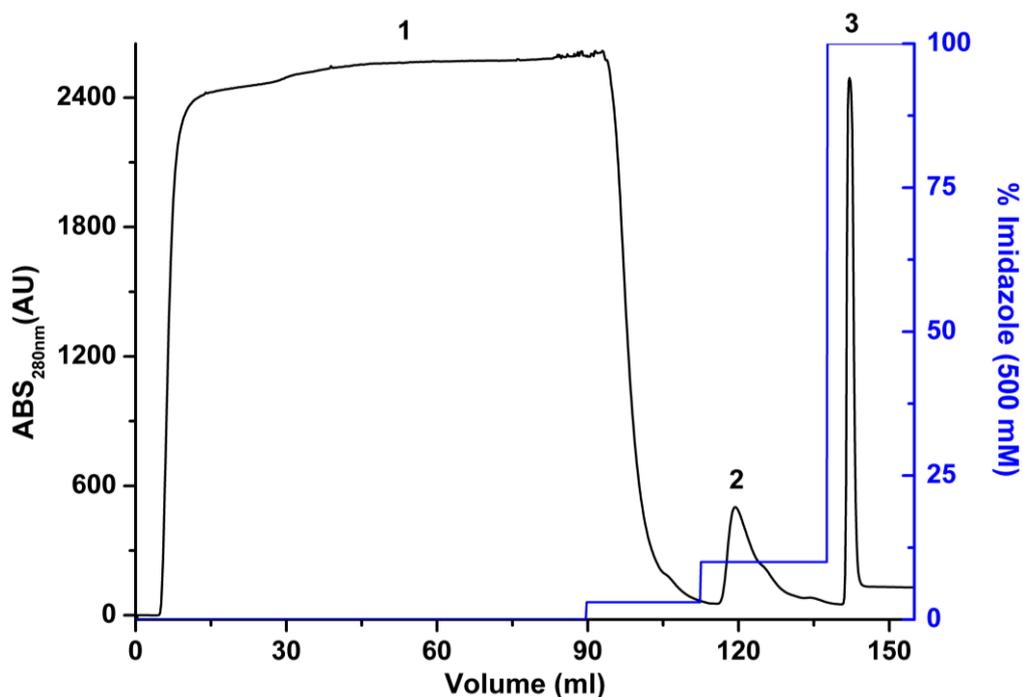


Figure 11. IMAC chromatogram of Late. The flow through is indicated by 1. The column was washed with 50 mM imidazole (peak 2) prior to protein elution with 500 mM imidazole (peak 3).

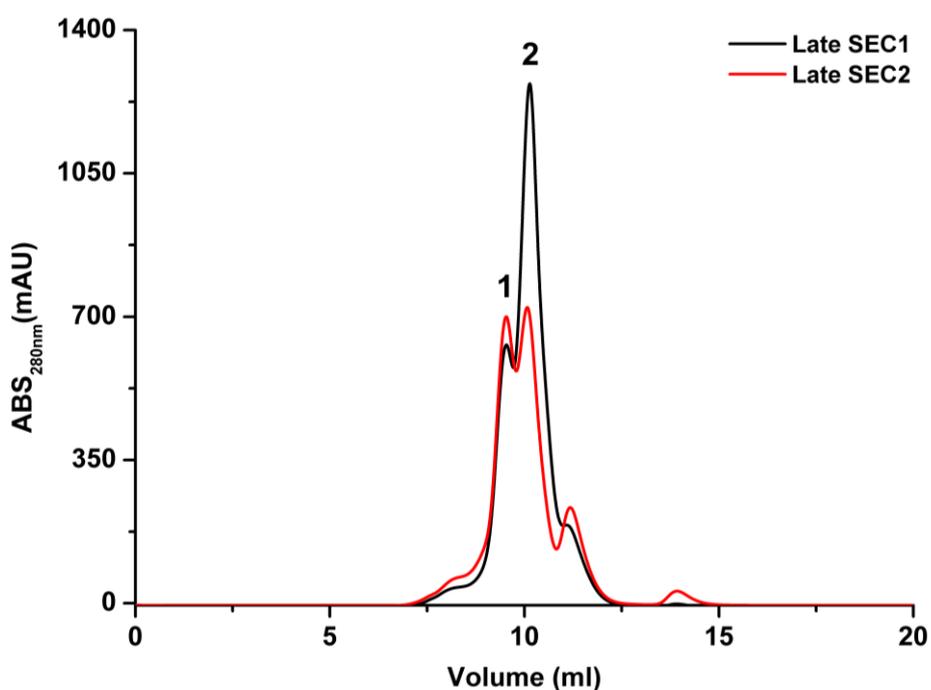


Figure 12. Superposition of two size exclusion chromatograms of Late samples. The 500 μ L samples injected in the Superdex 75 Increase 10/300 column (GE) were previously purified by IMAC, as shown in **Figure 11**.

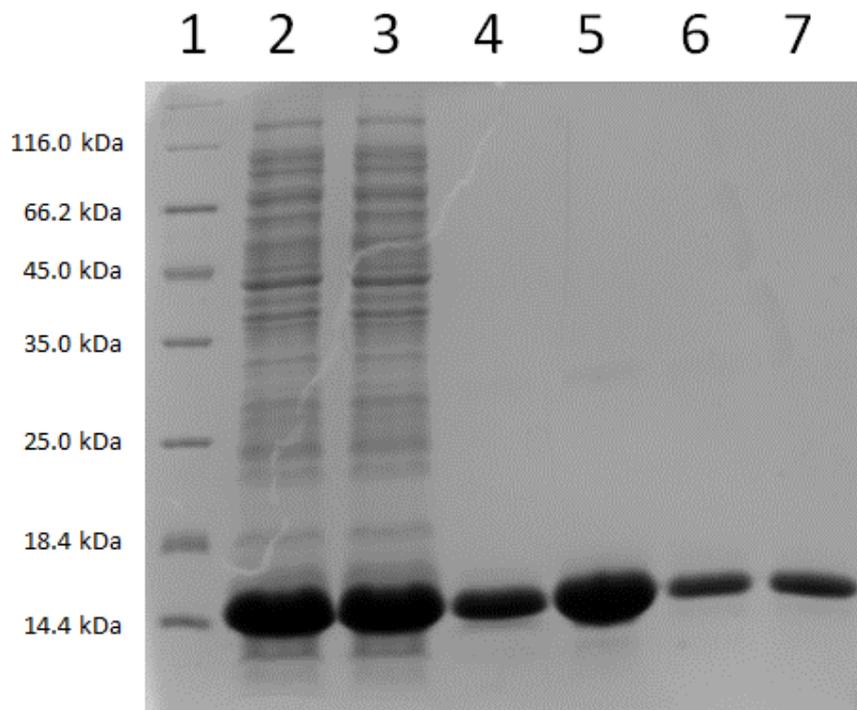


Figure 13. Electrophoretic profile in SDS-PAGE 15 % of the purification of Late. 1- Molecular weight marker; 2- Cell lysis supernatant; 3- Flow through; 4- Fraction eluted with 50 mM imidazole in the IMAC; 5- Fraction eluted with 500 mM imidazole in the IMAC; 6- SEC peak 1; 7-SEC peak 2.

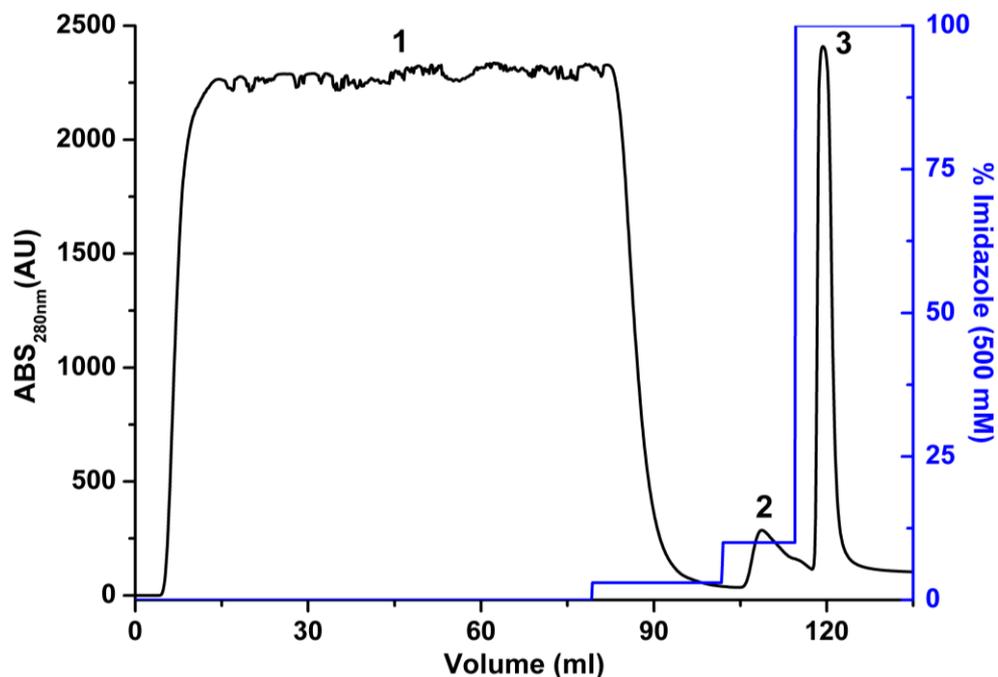


Figure 14. IMAC chromatogram of Theo. The flow through is indicated by 1. The column was washed with 50 mM imidazole (peak 2) prior to protein elution with 500 mM imidazole (peak 3).

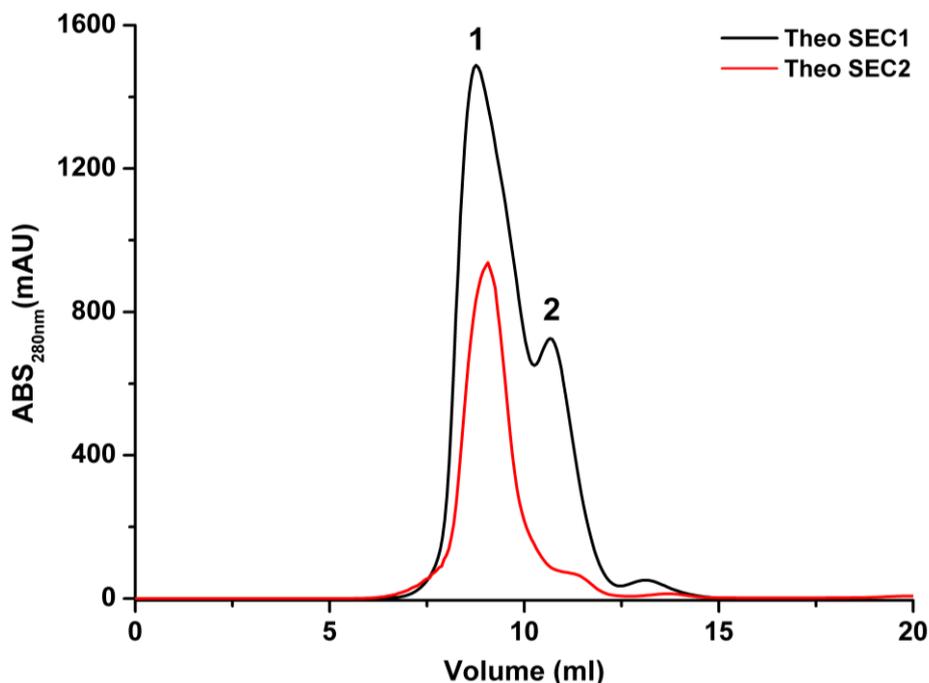


Figure 15. Superposition of two size exclusion chromatograms of Theo samples. The 500 μ L samples injected in the Superdex 75 Increase 10/300 column (GE) were previously purified by IMAC, as shown in **Figure 14**.

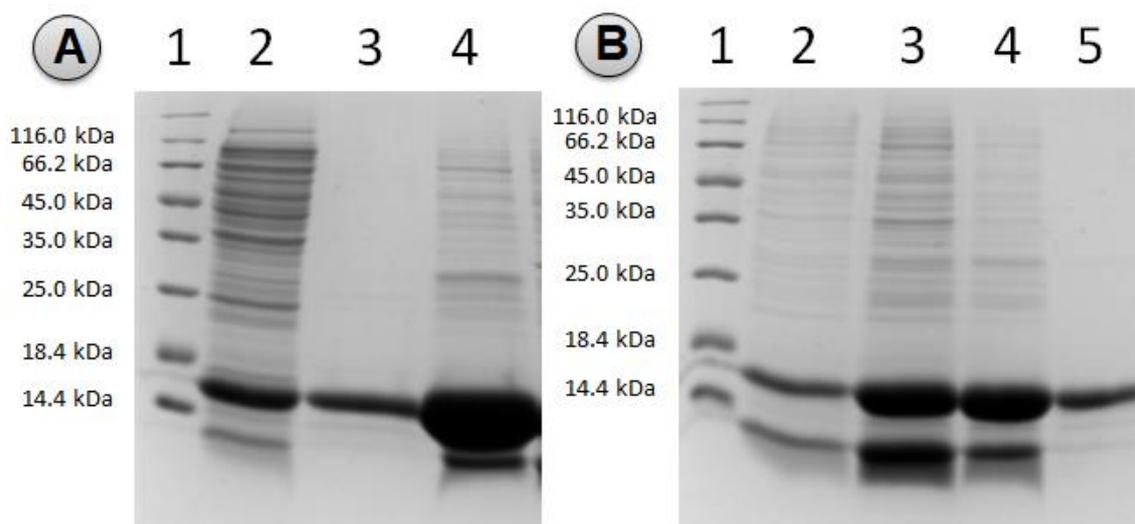


Figure 16. Electrophoretic profile in SDS-PAGE 15 % of the purification samples of Theo. **A.** IMAC samples: 1- Molecular weight marker; 2- Flow through; 3- Fraction eluted with 50 mM imidazole; 4- Fraction eluted with 500 mM imidazole. **B.** SEC samples: 1- Molecular weight marker; 2- Base of peak 1; 3- Peak 1; 4- Peak 2; 5- Base of peak 2.

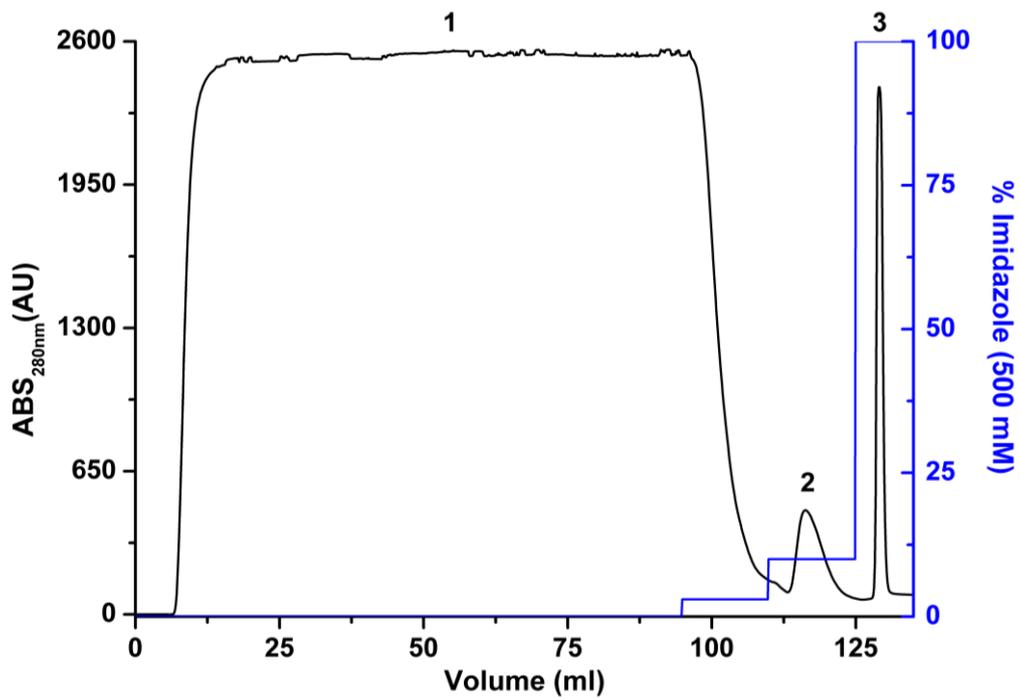


Figure 17. IMAC chromatogram of Hop1. The flow through is indicated by 1. The column was washed with 50 mM imidazole (peak 2) prior to protein elution with 500 mM imidazole (peak 3).

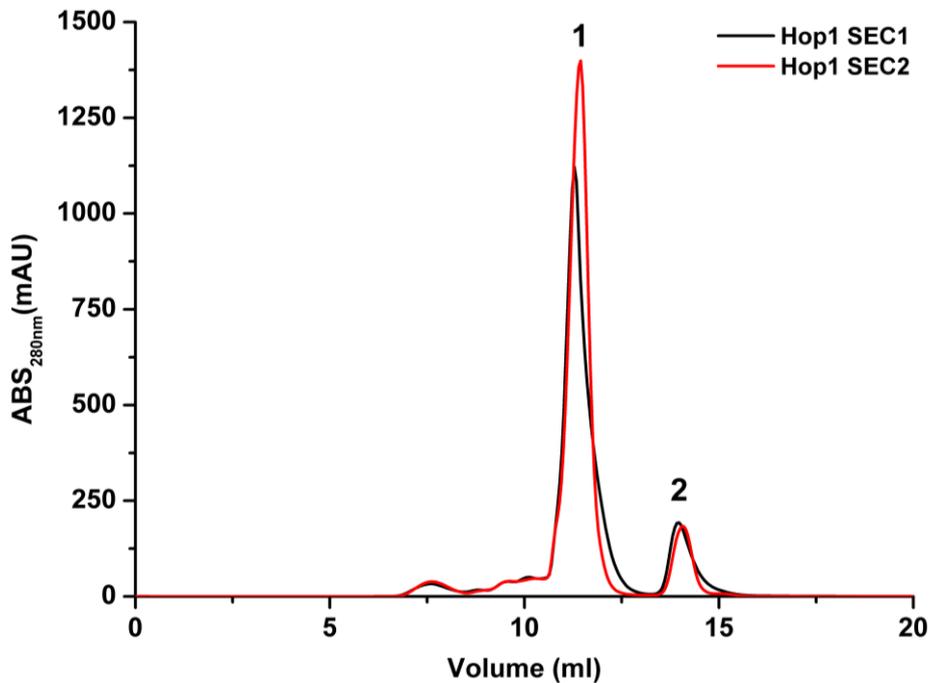


Figure 18. Superposition of two size exclusion chromatograms of Hop1 samples. The 500 μ L samples injected in the Superdex 75 Increase 10/300 column (GE) were previously purified by IMAC, as shown in **Figure 17**.

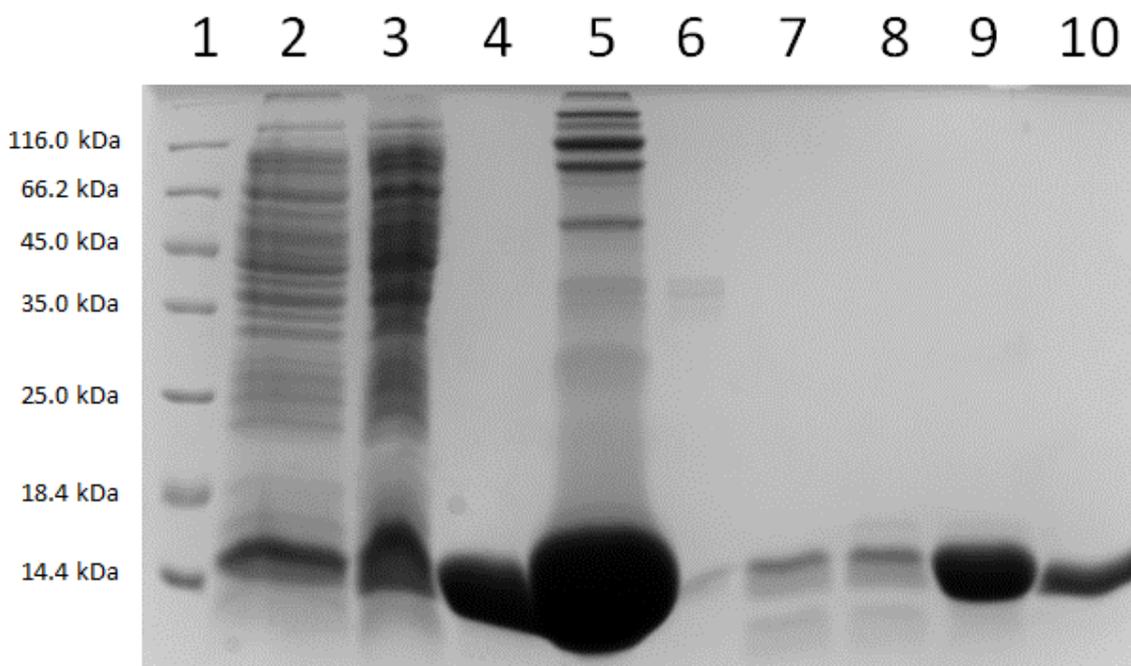


Figure 19. Electrophoretic profile in SDS-PAGE 15 % of the purification samples of Hop1. 1- Molecular weight marker; 2- Cell lysis supernatant; 3- Flow through; 4- Fraction eluted with 50 mM imidazole in the IMAC; 5- Fraction eluted with 500 mM imidazole in the IMAC; 6- SEC fraction eluted at 8.9 mL; 7- SEC fraction eluted at 9.5 mL; 8- SEC fraction eluted at 10.2 mL; 9- SEC peak 1; 10- SEC peak 2.

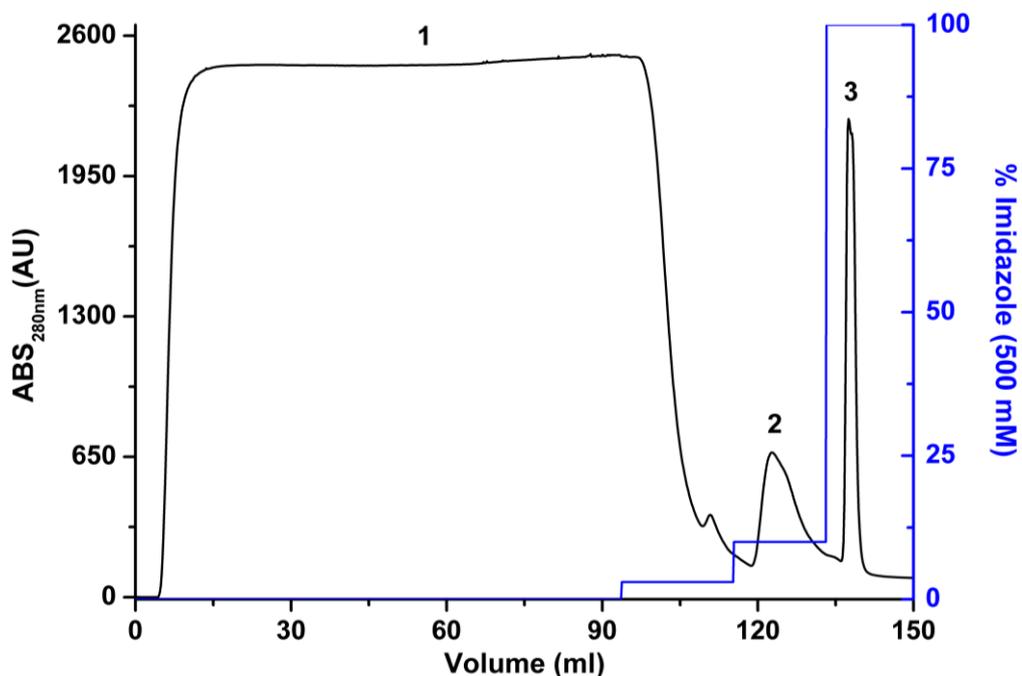


Figure 20. IMAC chromatogram of Hop5. The flow through is indicated by 1. The column was washed with 50 mM imidazole (peak 2) prior to protein elution with 500 mM imidazole (peak 3).

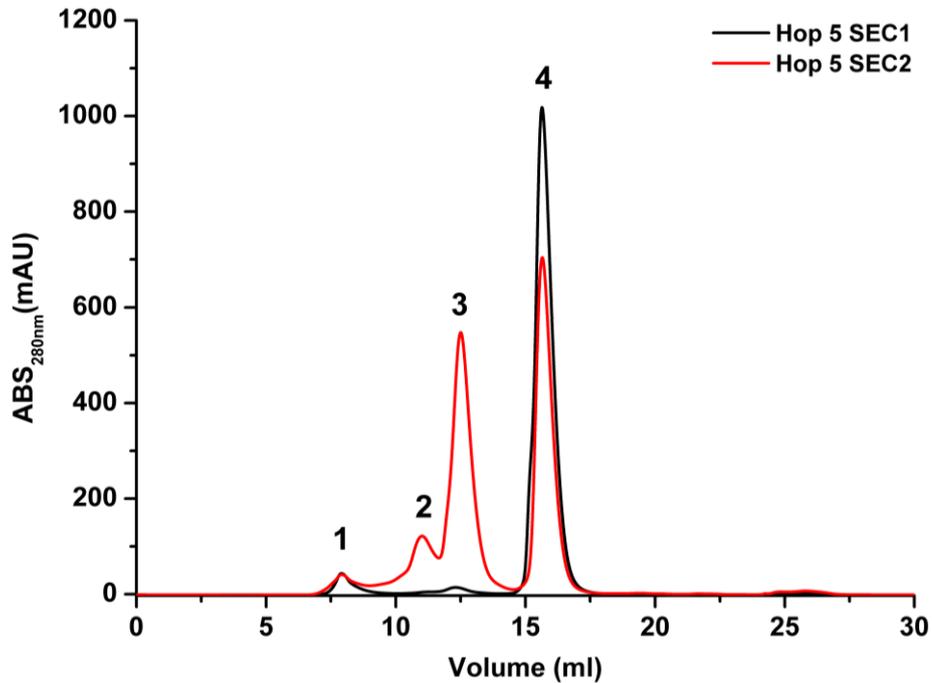


Figure 21. Superposition of two size exclusion chromatograms of Hop5 samples. The 500 μ L samples injected in the Superdex 75 Increase 10/300 column (GE) were previously purified by IMAC, as shown in **Figure 20**.

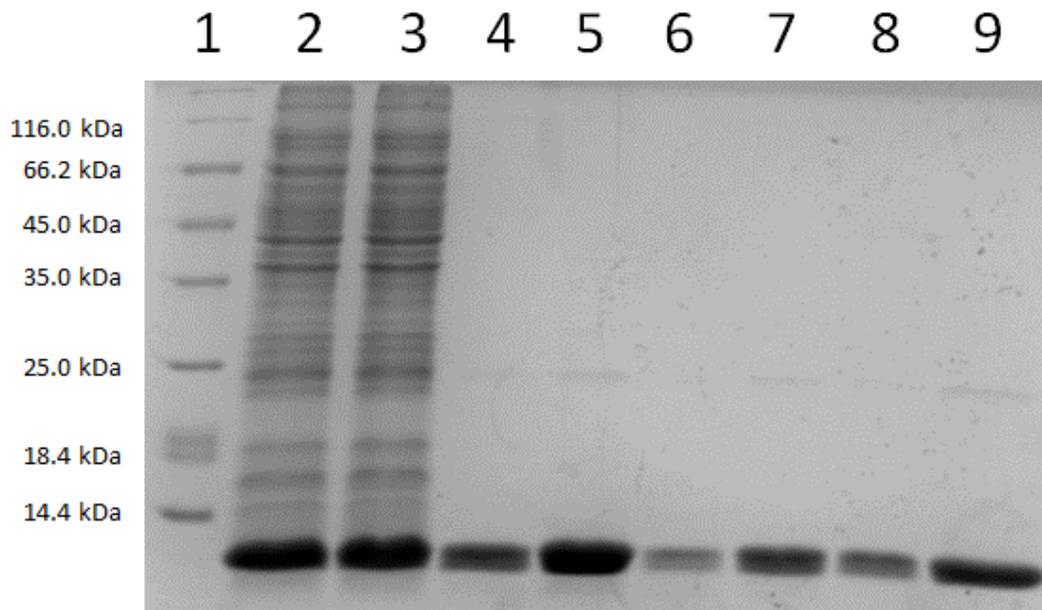


Figure 22. Electrophoretic profile in SDS-PAGE 15 % of the purification samples of Hop5. 1- Molecular weight marker; 2- Cell lysis supernatant; 3- Flow through; 4- Fraction eluted with 50 mM imidazole in the IMAC; 5- Fraction eluted with 500 mM imidazole in the IMAC; 6- SEC peak 1; 7- SEC peak 2; 8- SEC peak 3; 9- SEC peak 4.

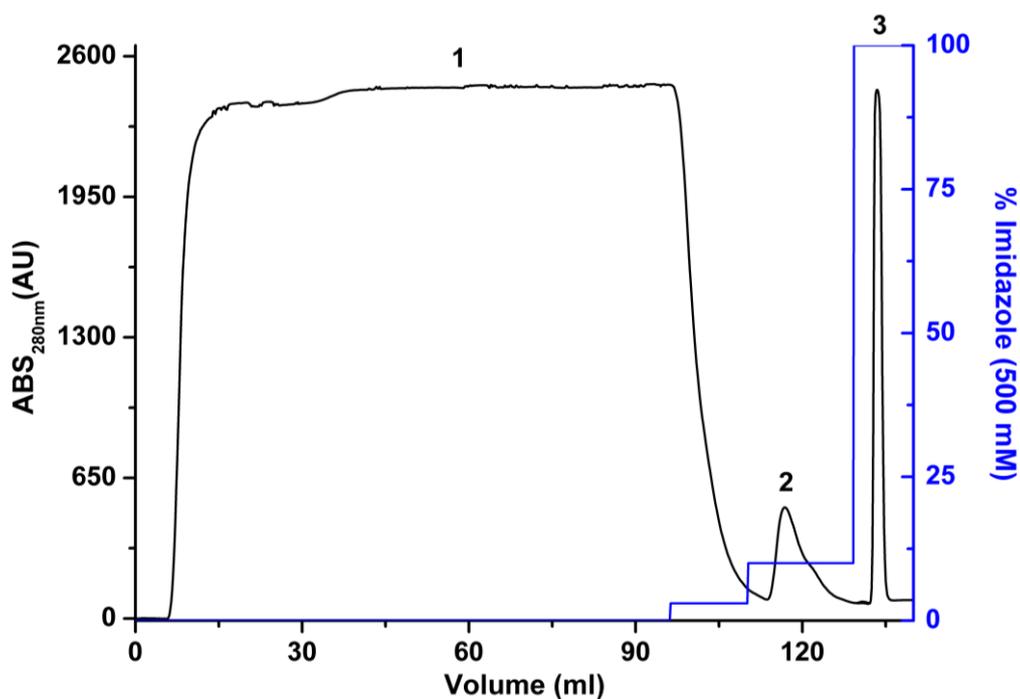


Figure 23. IMAC chromatogram of Hemp1. The flow through is indicated by 1. The column was washed with 50 mM imidazole (peak 2) prior to protein elution with 500 mM imidazole (peak 3).

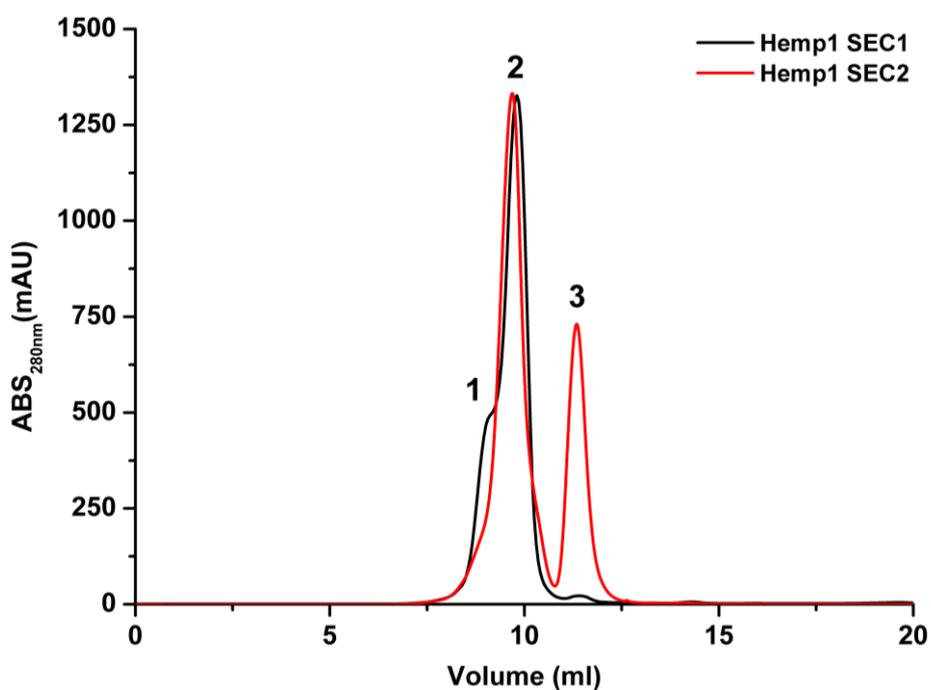


Figure 24. Superposition of two size exclusion chromatograms of Hemp1 samples. The 500 μ L samples injected in the Superdex 75 Increase 10/300 column (GE) were previously purified by IMAC, as shown in **Figure 23**.

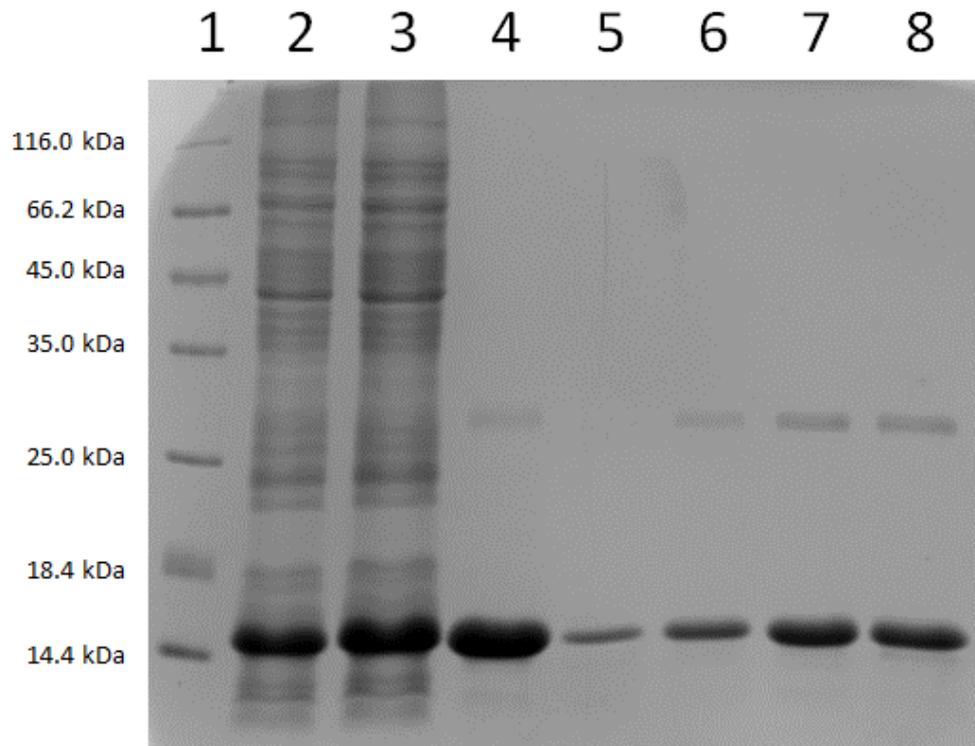


Figure 25. Electrophoretic profile in SDS-PAGE 15 % of the purification samples of Hemp1. 1- Molecular weight marker; 2- Cell lysis supernatant; 3- Flow through; 4- Fraction eluted with 50 mM imidazole in the IMAC; 5- Fraction eluted with 500 mM imidazole in the IMAC; 6- SEC peak 1; 7- SEC peak 2; 8- SEC peak 3.

5.3. OLIGOMERIC PROFILE CHARACTERIZATION

5.3.1. Native PAGE

Purified Late samples (after IMAC and SEC), with homogeneity indicated by SDS-PAGE (**Figure 13**) were analyzed in a native 15 % polyacrylamide gel as a means of observing possible oligomeric populations. The samples were incubated at different concentrations of NaCl, in order to investigate a possible dynamic equilibrium between the oligomeric populations. The gel, presented in **Figure 26**, shows that all samples comprise at least 3 defined oligomeric states and that the ionic strength bestowed by the salt had no effect in the equilibrium of such states.

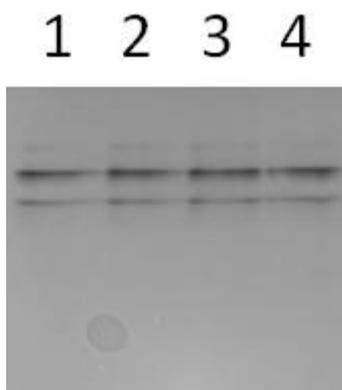


Figure 26. Native 15 % polyacrylamide gel with Late samples. 1- Sample without NaCl; 2- Sample with 300 mM NaCl; 3- Sample with 600 mM NaCl; 4- Sample with 900 mM NaCl.

5.3.2. Sedimentation Velocity Analytical Ultracentrifugation (SV-AUC)

Sedimentation velocity analytical ultracentrifugation (SV-AUC) assays were conducted in order to characterize the oligomeric states of the purified protein samples. SV-AUC experiments were performed with purified proteins in different buffers according to the crystallization conditions analyzed. The buffer composition of each sample and the parameters predicted by SEDNTERP are displayed in **Table 1**. Each SV-AUC data set (300 scans) consisted of the measured light absorption at 280 nm ($A_{280\text{nm}}$) as a function of the radial distance to the rotation axis. Representations of the sedimentation data with their best

Lamm equation fits are shown in **Figures 27, 29, 31, 33, 35** and **37** while the continuous sedimentation coefficient distribution are showed in **Figures 28, 30, 32, 34, 36** and **38**. The contribution of each molecular species to the total signal was calculated from the area of the peaks in the sedimentation coefficient distributions. All displayed sedimentation coefficients were adjusted to standard conditions, i.e. sedimentation coefficients of samples at 20 °C in water ($s_{20,W}$).

Table 1. Buffer composition and parameters used for SV-AUC fitting and analysis

Protein (SEC peak)	V_{bar} (mL/g)	Buffer	Tris (mM)	NaCl (mM)	ρ (g/mL)	η (Poise)
Late (1)	0.74093	A	75	300	1.0127	0.010532
Theo (1)	0.73733	B	75	150	1.0066	0.01039
Hop1 (1)	0.73871	A	75	300	1.0127	0.01053
Hop5 (3)	0.74053	C	20	150	1.0050	0.01002
Hop5 (4)	0.74053	A	75	300	1.0127	0.01053
Hemp1 (2)	0.73530	A	75	300	1.0127	0.01053

The Late SV-AUC experiment was performed with samples eluted in the first SEC peak (**Figure 12**) in buffer A (300 mM NaCl and 75 mM Tris-HCl pH 7.6) and its data are represented in **Figures 27** and **28**. The best Lamm equation fit was obtained with a frictional ratio of 1.267 and rmsd of 0.006. The sedimentation coefficient distribution curve (**Figure 28**) shows two sedimentation peaks with sedimentation coefficients of 1.888 S and 2.584 S. The smaller peak (2.81 % of the signal) presented an estimated molecular mass of 18.3 kDa, while the major one (80.49 % of the signal) presented an estimated molecular mass of 29.3 kDa. These results suggest the presence of two different Late oligomers, possibly a combination of monomers (12.3 kDa) and dimers (24.6 kDa). The presence of multiple oligomeric forms of the protein, as seen in the native-PAGE gel (**Figure 26**), may account for the inaccuracy in the determination of the molecular masses, as the frictional ratio is calculated as an average between all molecules.

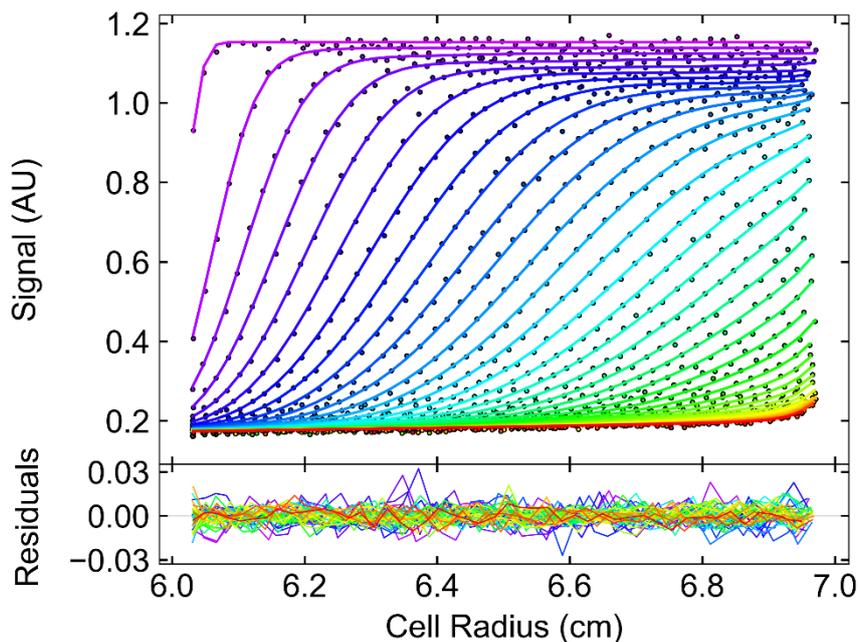


Figure 27. Sedimentation profile of 0.79 mg/mL of Late in Buffer A, monitored by the absorbance at 280 nm as a function of the radial distance to the rotation axis. The residuals of the fit, an indicative of its quality, are shown below.

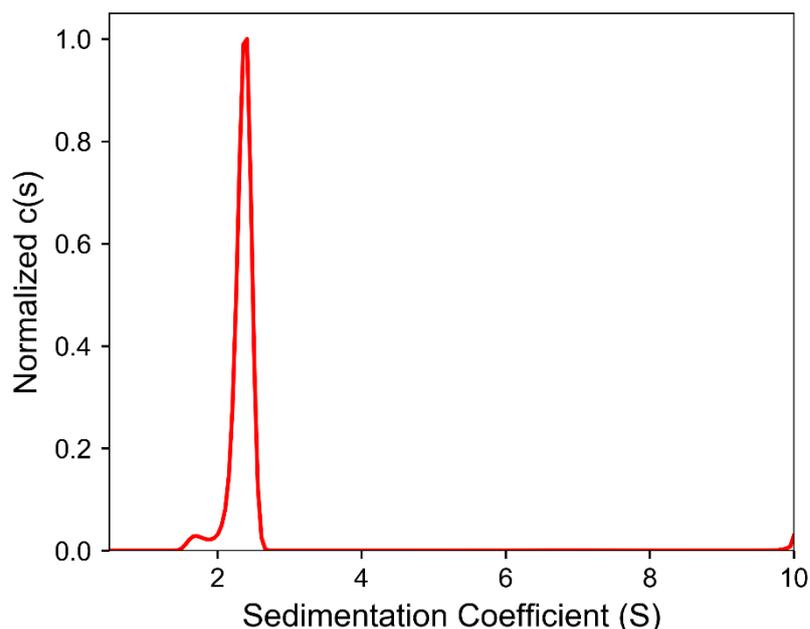


Figure 28. Continuous distribution of the sedimentation coefficients $c(s)$ of the Late sample. The distribution indicates the existence of two molecule populations: the first with sedimentation coefficient of 1.888 S and estimated molecular mass of 18.3 kDa, accounting for 2.81 % of the total; the second, with sedimentation coefficient of 2.584 S and estimated molecular mass of 29.3 kDa, composing 80.49 % of the total.

Sedimentation data of Theo samples eluted in the first SEC peak (**Figure 15**) in buffer B (150 mM NaCl and 75 mM Tris-HCl pH 7.6) are depicted in **Figures**

29 and **30**. The best Lamm equation fit was obtained with a frictional ratio value of 1.369 and rmsd of 0.012. The sedimentation coefficient distribution presented in **Figure 30** shows two peaks with sedimentation coefficients of 2.147 S and 3.714 S. The major peak (89.82 % of the signal) corresponds to a molecular weight of 24.4 kDa, which is comparable to the mass of a dimer of Theo (25.0 kDa). The smaller peak (4.04 %) corresponds to an estimated molecular mass of 55.4 kDa, which could be associated to a contaminant or a tetramer of Theo (50.1 kDa).

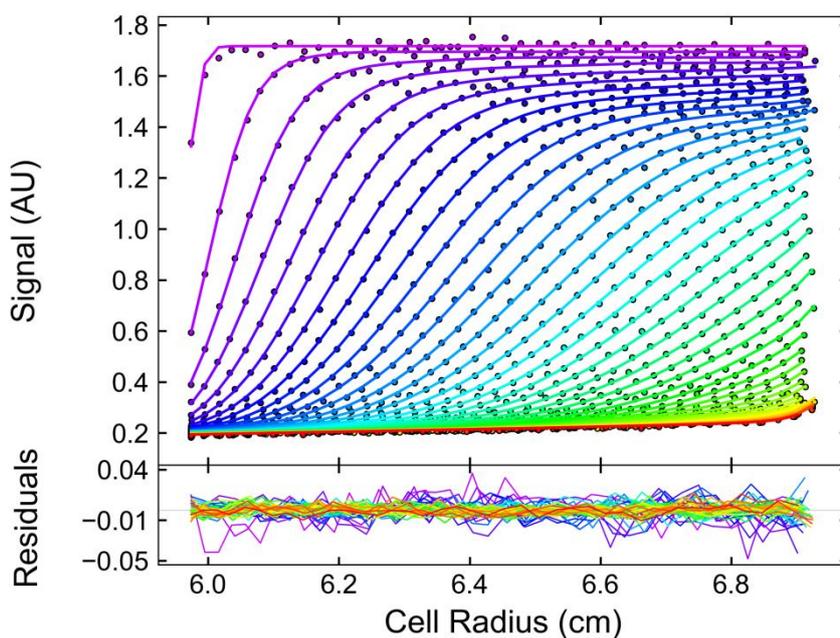


Figure 29. Sedimentation profile of 0.93 mg/mL of Theo in Buffer B, monitored by the absorbance at 280 nm as a function of the radial distance to the rotation axis. The residuals of the fit, an indicative of its quality, are shown below.

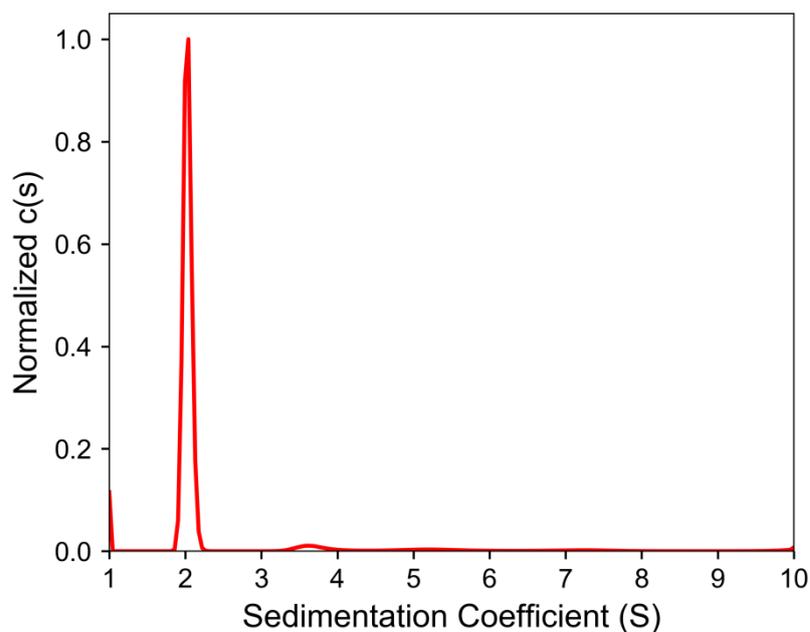


Figure 30. Continuous distribution of the sedimentation coefficients $c(s)$ of the Theo sample. The distribution indicates the existence of two molecule populations: the first with sedimentation coefficient of 2.147 S and estimated molecular mass of 24.4 kDa, composing 89.82 % of the total; the second, with sedimentation coefficient of 3.714 S and estimated molecular mass of 55.4 kDa, composing 4.04 % of the total.

Hop1 SV-AUC data were collected from the samples of the first SEC peak (**Figure 18**) in buffer A (300 mM NaCl and 75 mM Tris-HCl pH 7.6) and its sedimentation profile is shown in **Figure 31**. The best Lamm equation fit was obtained with a frictional ratio of 1.384 and rmsd of 0.011. The sedimentation coefficient distribution curve (**Figure 32**) shows a single peak (93.88 % of the signal) of sedimentation coefficient 2.073 S and estimated molecular weight of 23.7 kDa, suggesting the presence of Hop1 dimers (24.6 kDa).

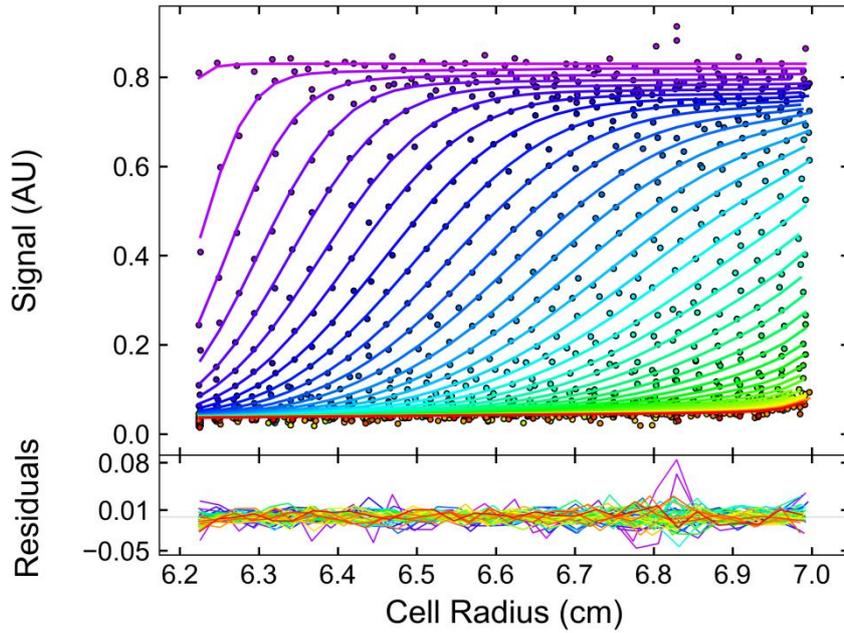


Figure 31. Sedimentation profile of 0.65 mg/mL of Hop1 in buffer A, monitored by the absorbance at 280 nm as a function of the radial distance to the rotation axis. The residuals of the fit, an indicative of its quality, are shown below.

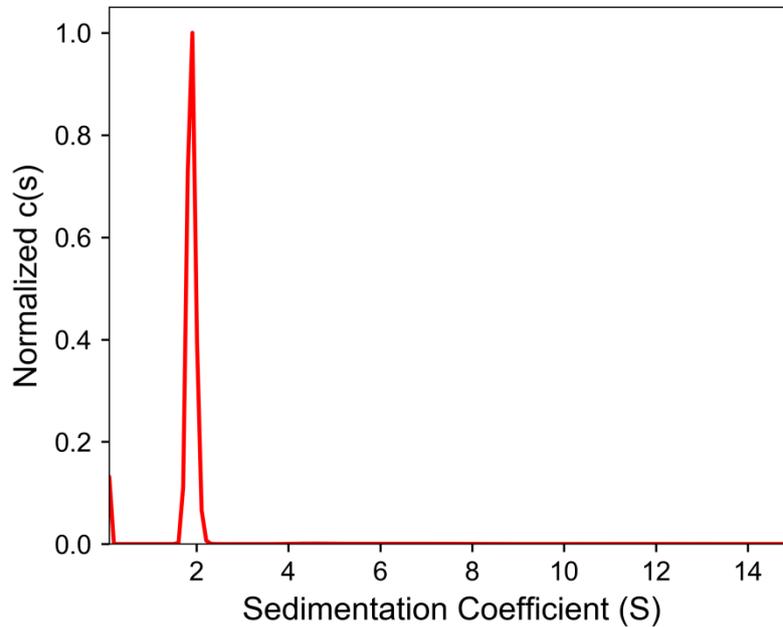


Figure 32. Continuous distribution of the sedimentation coefficients $c(s)$ of the Hop1 sample. The distribution indicates the existence of a single molecule population, with sedimentation coefficient of 2.073 S and estimated molecular mass of 23.7 kDa, composing 93.88 % of the total.

Hop5 was analyzed by SV-AUC using samples from the third and fourth SEC peaks (**Figure 21**) due to their high availability. Its sedimentation profiles are depicted in **Figures 33** and **35**, respectively. The best Lamm equation fits were obtained at frictional ratios of 1.367 and 1.325, with rmsd's of 0.008 and 0.007 for the third peak and fourth peaks, respectively. The sedimentation coefficient distribution curve for samples of the third SEC peak (**Figure 34**) showed a main sedimentation peak (33.66 % of the signal) with a sedimentation coefficient of 1.997 S and estimated molecular weight of 22.3 kDa, corresponding to dimers of Hop5 (22.0 kDa). **Figure 36** presents the sedimentation coefficient distribution curve for samples of the fourth SEC peak, which shows two sedimentation peaks with sedimentation coefficients of 1.352 S and 2.809 S, respectively. The first and main peak (81.98 %) has an estimated molecular weight of 11.8 kDa, corresponding to monomers of Hop5 (11.0 kDa). The second and smaller peak (2.74 %) presented an estimated molecular weight of 35.4 kDa, which could be explained by monomers of Hop5 associated into trimers (32.9 kDa), as a result of the protein concentration increase during sedimentation. Thus, we concluded that the third SEC peak is formed by dimers of Hop5, while the fourth peak is formed by monomers.

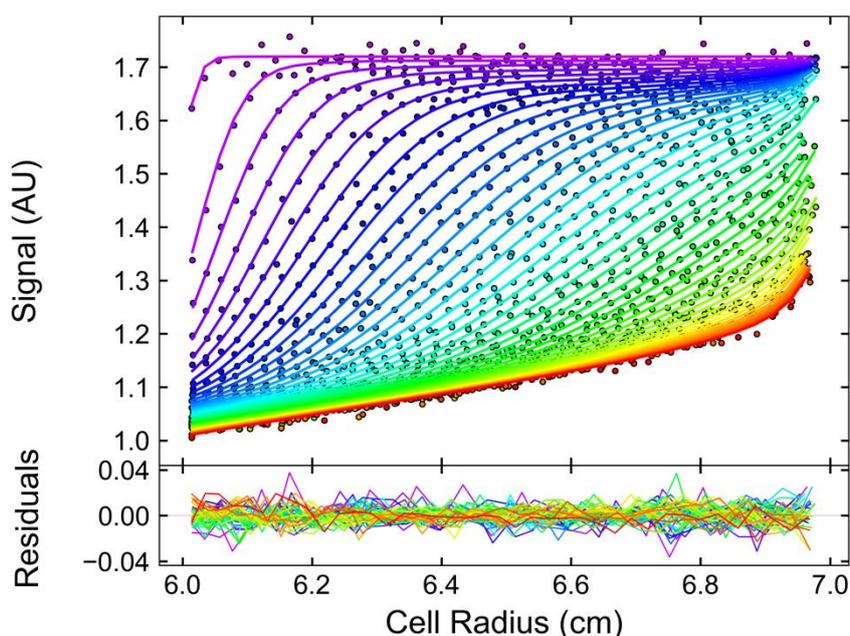


Figure 33. Sedimentation profile of 0.84 mg/mL of Hop5 in buffer C collected from the third SEC peak, monitored by the absorbance at 280 nm as a function of the radial distance to the rotation axis. The residuals of the fit, an indicative of its quality, are shown below.

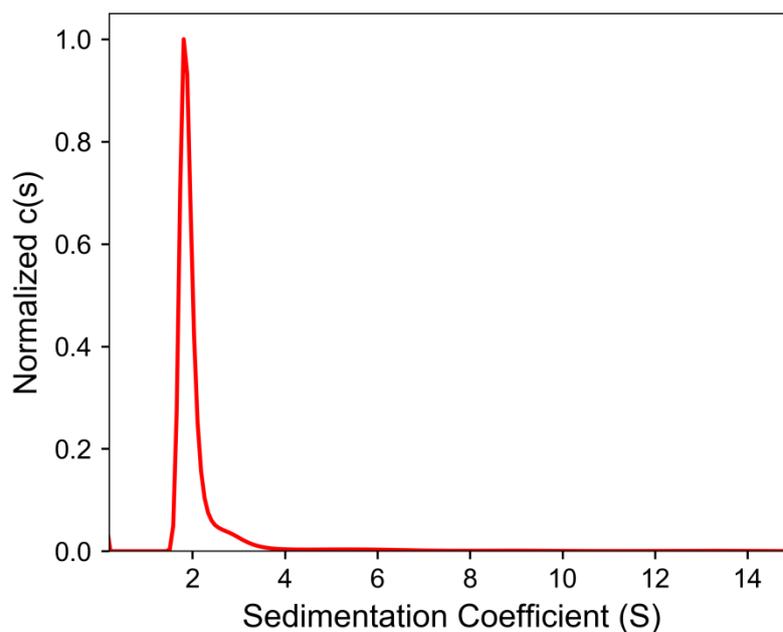


Figure 34. Continuous distribution of the sedimentation coefficients $c(s)$ of the Hop5 sample from the third SEC peak. The distribution indicates the existence of a single molecule population, with sedimentation coefficient of 1.997 S and estimated molecular mass of 22.3 kDa, composing 33.66 % of the total.

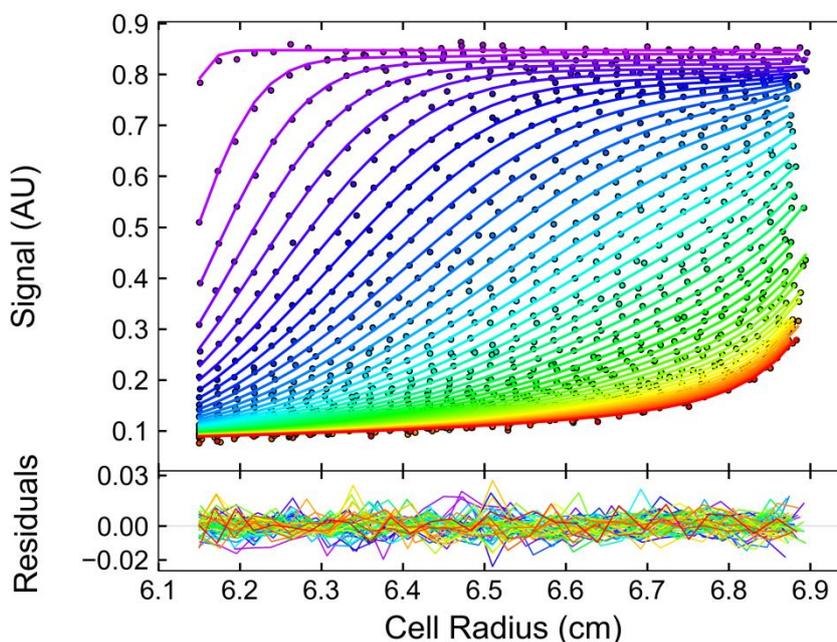


Figure 35. Sedimentation profile of 0.41 mg/mL of Hop5 in buffer A collected from the fourth SEC peak, monitored by the absorbance at 280 nm as a function of the radial distance to the rotation axis. The residuals of the fit, an indicative of its quality, are shown below.

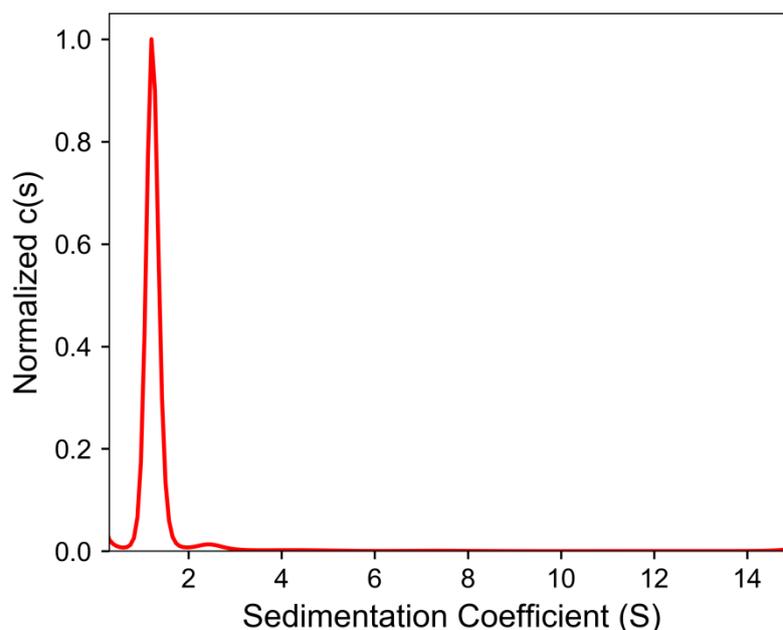


Figure 36. Continuous distribution of the sedimentation coefficients $c(s)$ of the Hop5 sample from the fourth peak. The distribution indicates the existence of two molecule populations: the first with sedimentation coefficient of 1.352 S and estimated molecular mass of 11.8 kDa, composing 81.98 % of the total; the second, with sedimentation coefficient of 2.809 S and estimated molecular mass of 35.4 kDa, composing 2.74 % of the total.

For Hemp1, samples eluted in buffer A (300 mM NaCl and 75 mM Tris-HCl pH 7.6) from the second SEC peak (**Figure 24**) were analyzed by SV-AUC as depicted in **Figure 37**. The best Lamm equation fit was obtained at a frictional ratio of 1.380, with an rmsd of 0.006. The sedimentation coefficient distribution curve (**Figure 38**) shows three peaks with sedimentation coefficients of 1.857 S, 3.240 S and 5.117. The estimated molecular mass for the species of the smaller first peak (2.63 %) was 19.6 kDa, probably associated to a contaminant, also seen in the Coomassie Blue stained SDS-PAGE gel (**Figure 25**), or a dimer of Hemp1 (24.7 kDa). The second peak has the most abundant species (80.24 %), with an estimated molecular mass of 45.1 kDa corresponding to a tetramer of Hemp1 (49.4 kDa). The third peak, with 5.60 % of the total signal, presented an estimated molecular mass of 89.5 kDa, probably associated to a contaminant or even Hemp1 multimers, as cystatin multimers are widely documented (SANDERS et al., 2004).

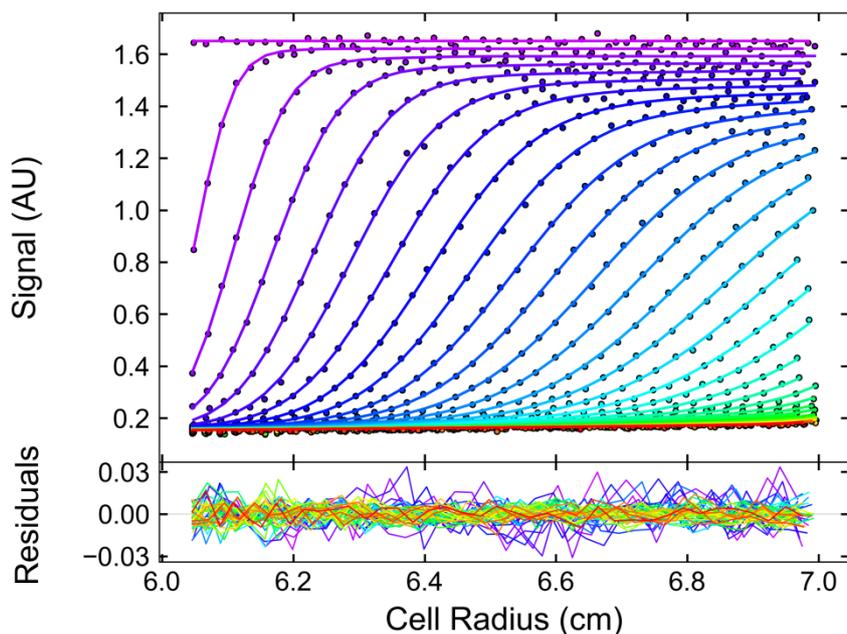


Figure 37. Sedimentation profile of 1.43 mg/mL of Hemp1 in buffer A collected from the second SEC peak, monitored by the absorbance at 280 nm as a function of the radial distance to the rotation axis. The residuals of the fit, an indicative of its quality, are shown below.

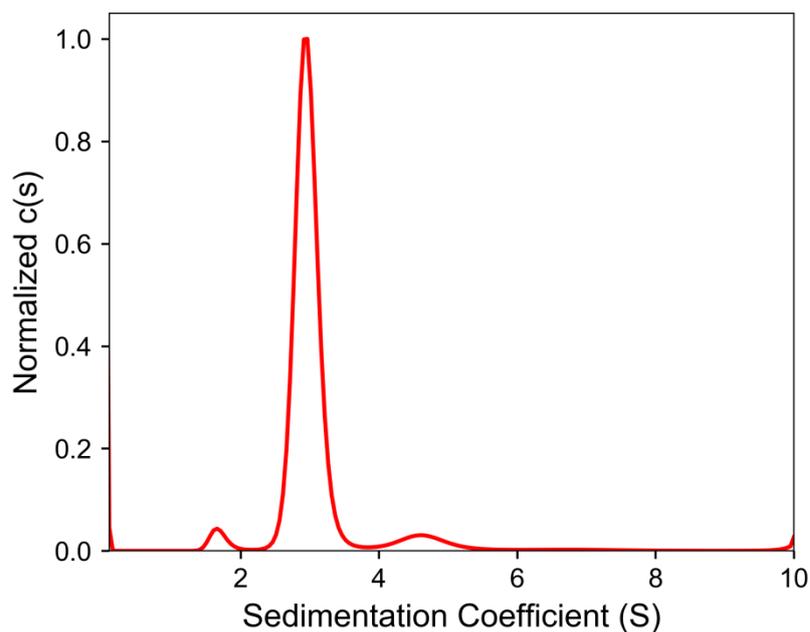


Figure 38. Continuous distribution of the sedimentation coefficients $c(s)$ of the Hemp1 sample from the second SEC peak. The distribution indicates the existence of three molecule populations: the first, with sedimentation coefficient of 1.857 S and estimated molecular mass of 19.6 kDa, composing 2.63 % of the total; the second, with sedimentation coefficient of 3.240 S and estimated molecular mass of 45.1 kDa, composing 80.24 % of the total; the third, with sedimentation coefficient of 5.117 and molecular mass of 89.5 kDa, composing 5.60 % of the signal.

All SV-AUC parameters calculated through the data fitting are presented in **Table 2**. All analyzed proteins presented similar frictional ratios which might be related to their conserved sequence and folding (TURK et al., 2008). The SV-AUC and SEC data indicate that these phytocystatins adopt different oligomeric states in solution, with Theo and Hop1 presenting themselves mostly as dimers, Hop5 as both monomers and dimers and Hemp1 as tetramers. These results are in agreement with the literature, where phytocystatins are described as being monomeric (NAGATA et al., 2000), dimeric or tetrameric (VALADARES et al., 2013). The imprecision in estimating the species' molecular weight in the more heterogeneous samples occurs because the frictional ratio, an essential parameter to the mass calculation, is fitted by SEDFIT as a weight-average between all molecules in solution. Thus, molecules with different frictional ratio in the solution may shift the value of the estimated molecular mass of all sedimentation peaks (CHATON et al., 2015).

Table 2. Analytical ultracentrifugation parameters obtained from the sedimentation data fitting for each phytocystatin sample.

Protein (SEC peak)	f/f₀	c(s) peak	s_{20,w} (S)	MM (kDa)	Signal %
Late (1)	1.267	1	1.888	18.3	2.81
		2	2.584	29.3	80.49
Theo (1)	1.369	1	2.147	24.4	89.82
		2	3.714	55.4	4.04
Hop1 (1)	1.384	1	2.073	23.7	93.88
Hop5 (3)	1.367	1	1.997	22.3	33.66
Hop5 (4)	1.325	1	1.352	11.8	81.98
		2	2.809	35.4	2.74
Hemp1 (2)	1.380	1	1.857	19.6	2.63
		2	3.240	45.1	80.24
		3	5.117	89.5	5.60

5.4. THERMAL DENATURATION ASSAYS

The effect of the temperature in the structure of Late, Hop1, Hop5 and Hemp1 can be observed in the thermal unfolding curves as following. The molar ellipticity measured at 218 nm for Late, Hop5 and Hop1, and at 222 nm for Hemp1 (**Figures 39, 43, 47 and 51**, respectively) showed standard two-states transitions from native to unfolded states, where the proteins kept their monitored secondary structures until the temperature reached around 72-75 °C for Late, 72-75 °C for Hop1, 58-63 °C for Hop5 and 49-62 °C for Hemp1, depending on the pH. The dichroic signal of the proteins decreased as a function of the temperature increase, presenting a disordered structure profile at around 95 °C. CD spectra obtained in the unfolding assays showed that the proteins presented seemingly unaltered secondary structures in the studied pHs until 74 °C for Late, 65 °C for Hop1, 55 °C for Hop5 and 45-65 °C for Hemp1 (**Figures 40-42, 44-46, 48-50 and 52-54**, respectively), with pronounced negative bands in 208 nm, 218 nm and 222 nm. However, in temperatures above these values, the signal intensity in these negative bands decreases and shifted to 200 nm, which is associated with disordered structures. After the denaturation assays, the samples were cooled back to 25 °C and a spectra was collected, indicated by the dashed lines (**Figures 40-42, 44-46, 48-50 and 52-54**), showing that Late, Hop1 and Hop5 presented a refolding profile with similar secondary structure to the native one. Despite not showing a conformation similar to the one in the beginning of the assay, the change in Hemp1's renaturation secondary structure profile may be attributed to a change in its oligomeric state. This is supported by the fact that the Hemp1 samples were the only ones comprising of protein tetramers, as shown by AUC (**Table 2**), and that its final spectra resemble the ones obtained for the other studied proteins, which consisted only on protein dimers.

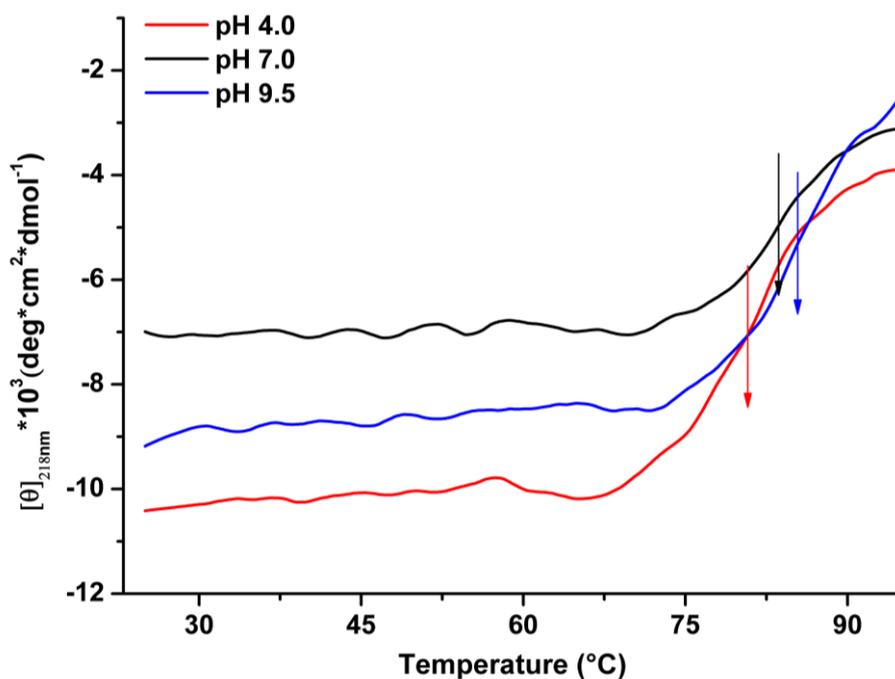


Figure 39. Late's thermal unfolding curves at different pHs. The vertical arrows point to the curves' estimated T_m values which are 80.8 °C, 83.6 °C and 85.4 °C for pHs 4.0, 7.0 and 9.5, respectively.

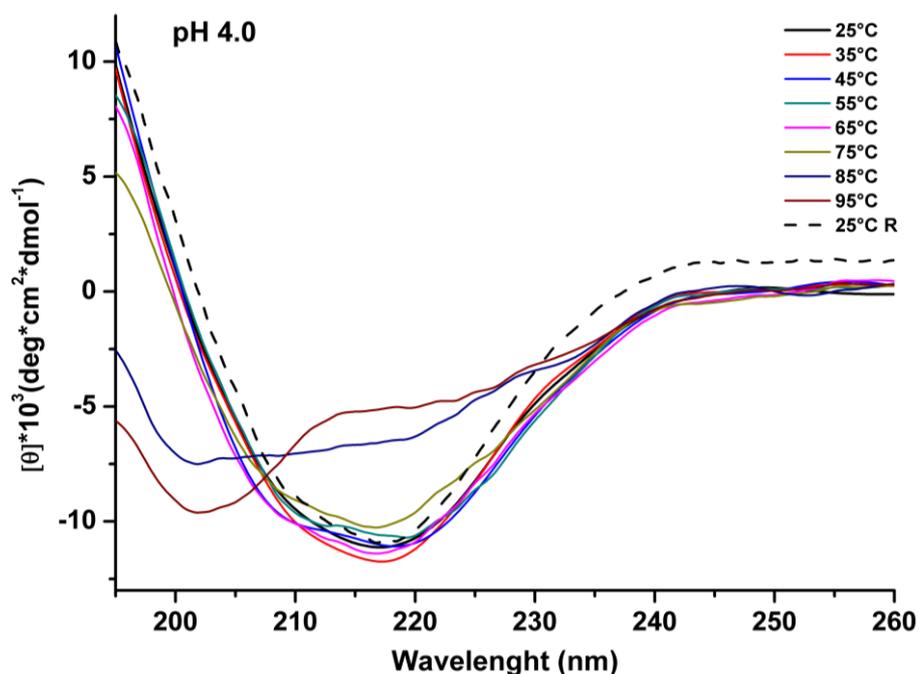


Figure 40. Far-UV circular dichroism spectra of Late at pH 4.0, measured under thermal denaturation (25 to 95°C) in intervals of 10 °C. The dashed line indicates the spectrum measured after the protein's temperature returned to 25 °C.

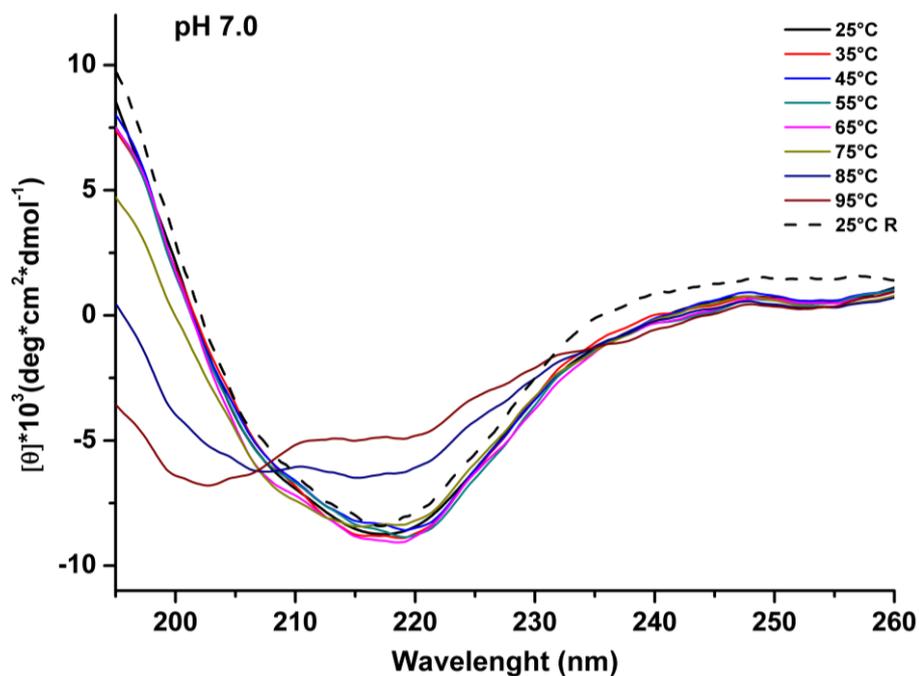


Figure 41. Far-UV circular dichroism spectra of Late at pH 7.0, measured under thermal denaturation (25 to 95 °C) in intervals of 10 °C. The dashed line indicates the spectrum measured after the protein's temperature returned to 25 °C.

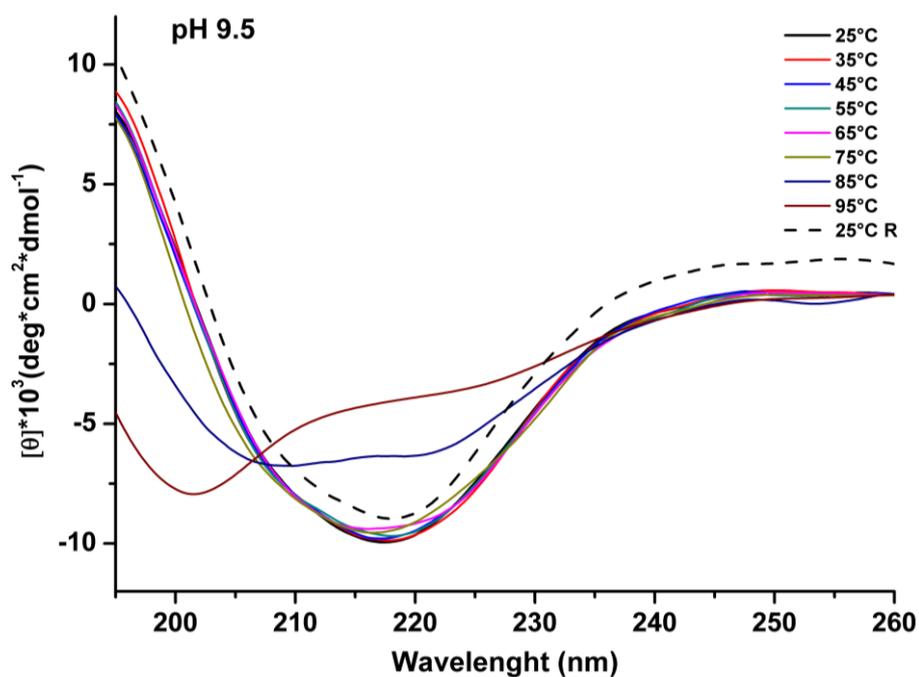


Figure 42. Far-UV circular dichroism spectra of Late at pH 9.5, measured under thermal denaturation (25 to 95 °C) in intervals of 10 °C. The dashed line indicates the spectrum measured after the protein's temperature returned to 25 °C.

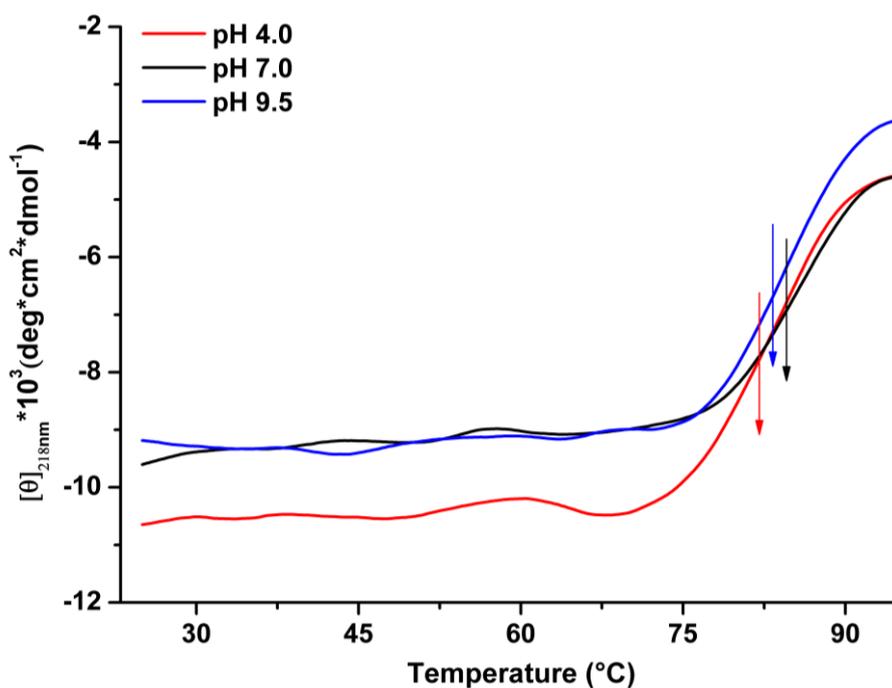


Figure 43. Thermal unfolding curves of Hop1 at different pHs. The vertical arrows point to the curves' calculated T_m values which are 82.3 °C, 84.8 °C and 83.7 °C for pHs 4.0, 7.0 and 9.5, respectively.

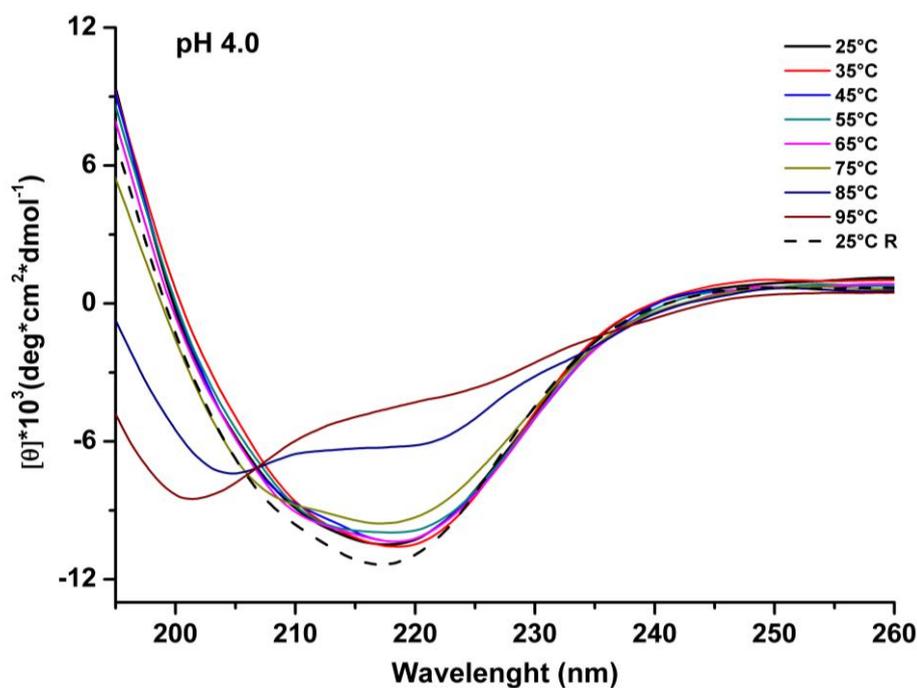


Figure 44. Far-UV circular dichroism spectra of Hop1 at pH 4.0, measured under thermal denaturation (25 to 95 °C) in intervals of 10 °C. The dashed line indicates the spectrum measured after the protein's temperature returned to 25 °C.

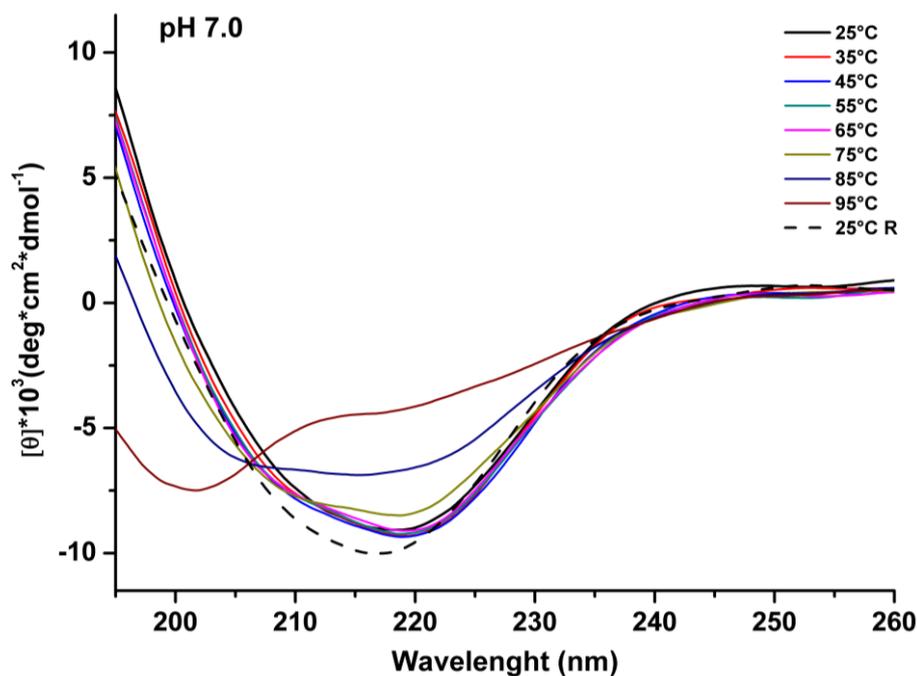


Figure 45. Far-UV circular dichroism spectra of Hop1 at pH 7.0, measured under thermal denaturation (25 to 95 °C) in intervals of 10 °C. The dashed line indicates the spectrum measured after the protein's temperature returned to 25 °C.

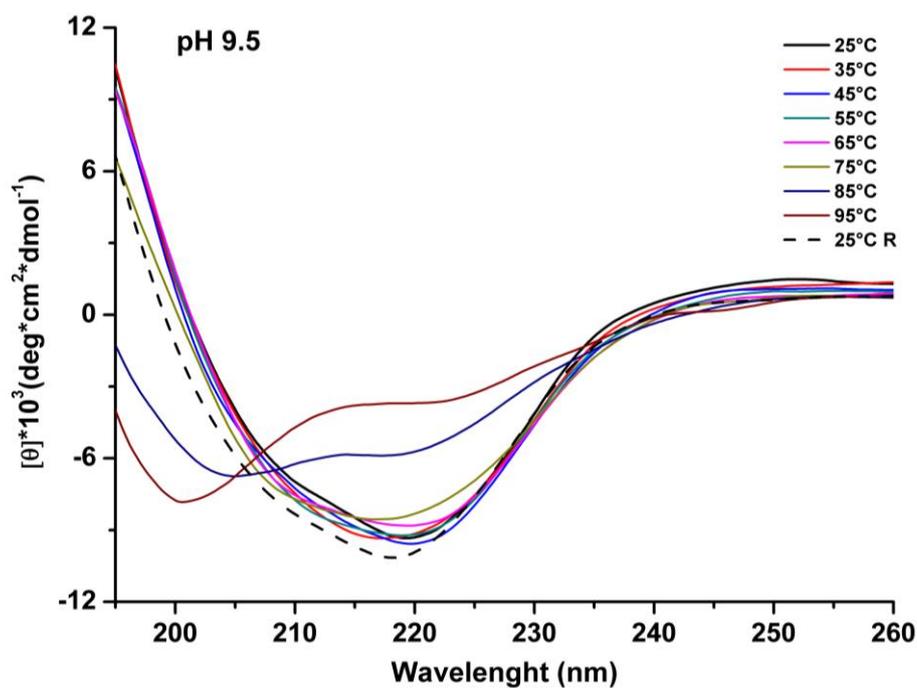


Figure 46. Far-UV circular dichroism spectra of Hop1 at pH 9.5, measured under thermal denaturation (25 to 95 °C) in intervals of 10 °C. The dashed line indicates the spectrum measured after the protein's temperature returned to 25 °C.

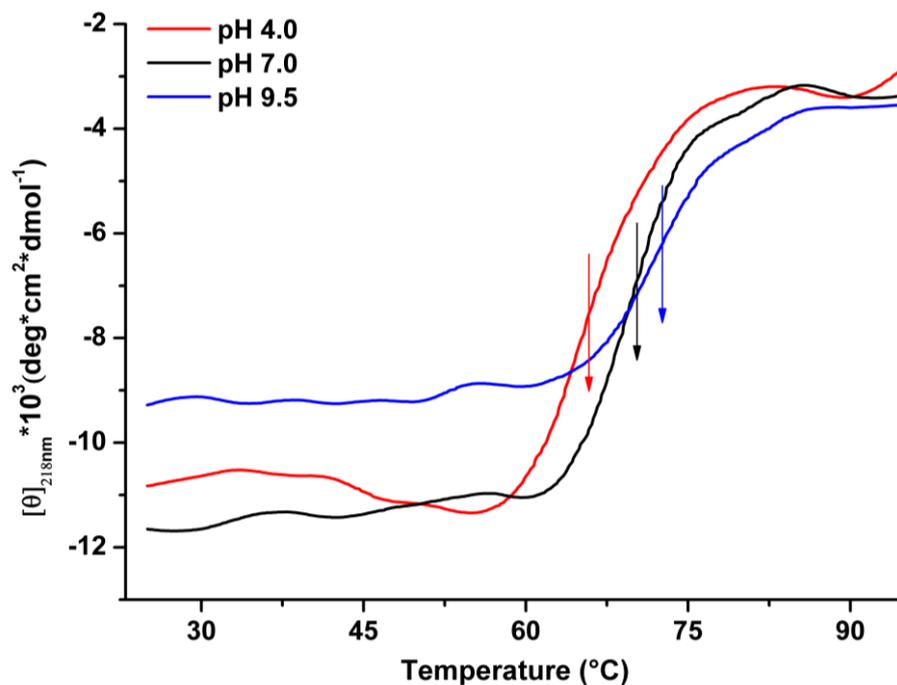


Figure 47. Thermal unfolding curves of Hop5 at different pHs. The vertical arrows point to the curves' calculated T_m values which are 65.8 °C, 70.2 °C and 72.4 °C for pHs 4.0, 7.0 and 9.5, respectively.

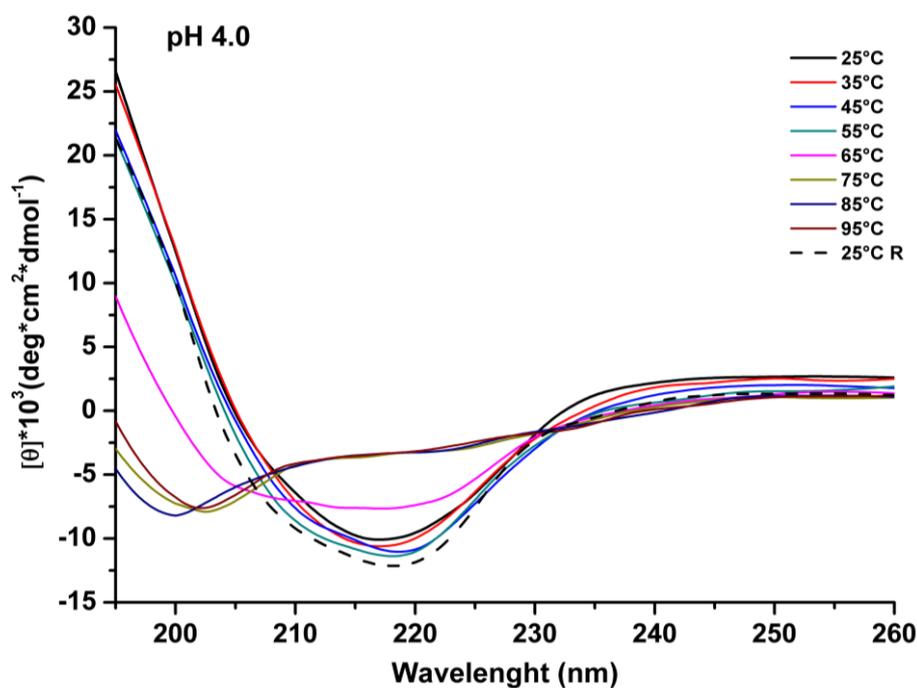


Figure 48. Far-UV circular dichroism spectra of Hop5 at pH 4.0, measured under thermal denaturation (25 to 95 °C) in intervals of 10 °C. The dashed line indicates the spectrum measured after the protein's temperature returned to 25 °C.

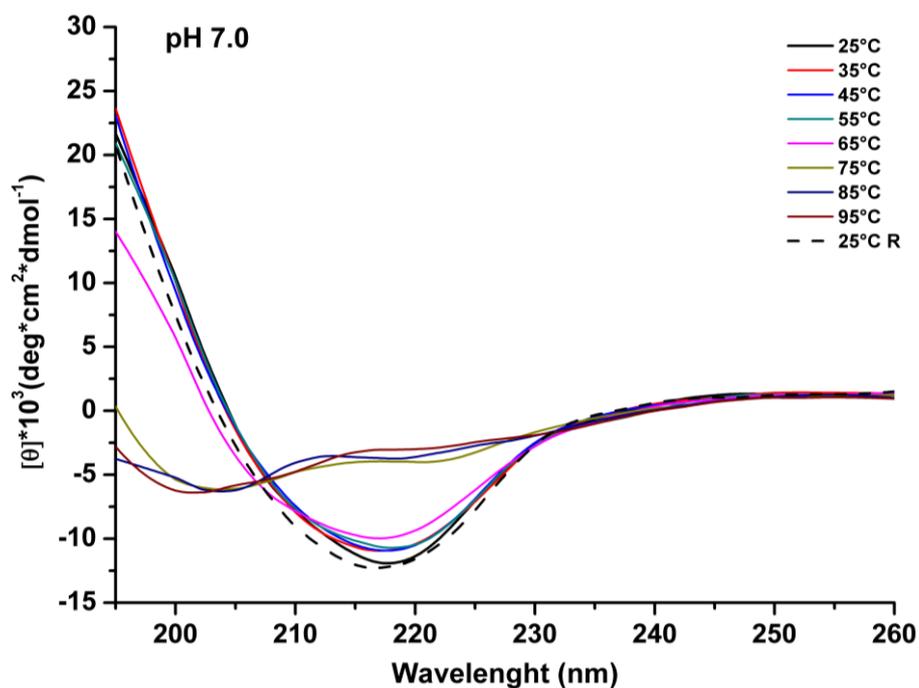


Figure 49. Far-UV circular dichroism spectra of Hop5 at pH 7.0, measured under thermal denaturation (25 to 95 °C) in intervals of 10 °C. The dashed line indicates the spectrum measured after the protein's temperature returned to 25 °C.

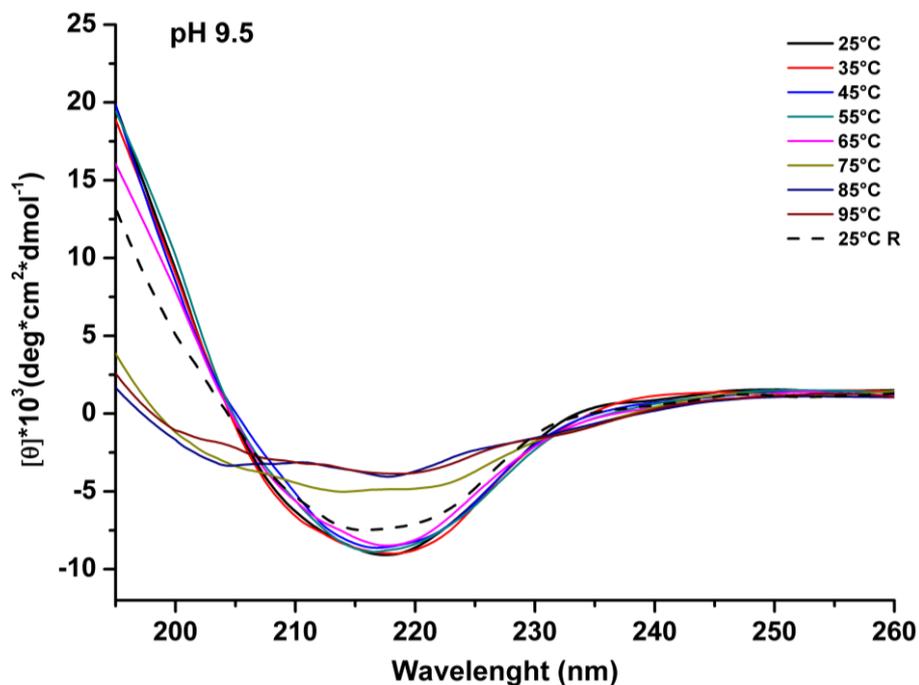


Figure 50. Far-UV circular dichroism spectra of Hop5 at pH 9.5, measured under thermal denaturation (25 to 95 °C) in intervals of 10 °C. The dashed line indicates the spectrum measured after the protein's temperature returned to 25 °C.

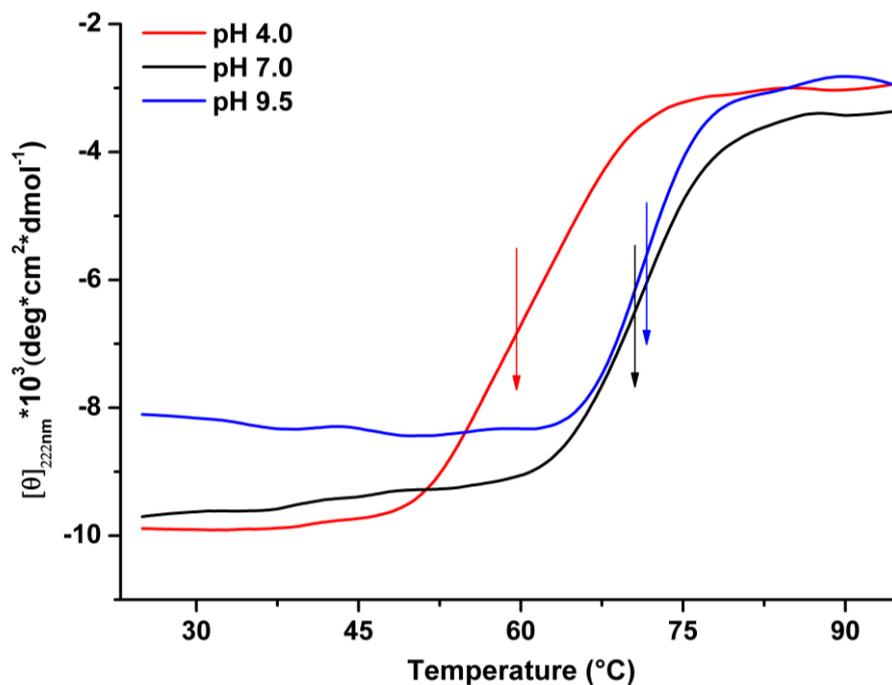


Figure 51. Thermal unfolding curves of Hemp1 at different pHs. The vertical arrows point to the curves' calculated T_m values which are 60.6 °C, 71.2 °C and 71.4 °C for pHs 4.0, 7.0 and 9.5, respectively.

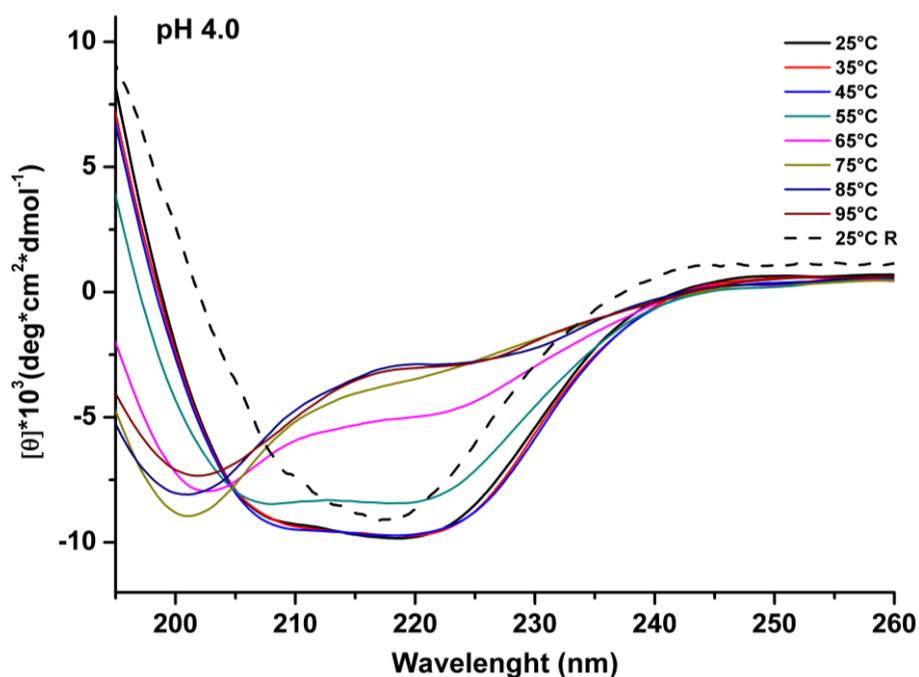


Figure 52. Far-UV circular dichroism spectra of Hemp1 at pH 4.0, measured under thermal denaturation (25 to 95 °C) in intervals of 10 °C. The dashed line indicates the spectrum measured after the protein's temperature returned to 25 °C.

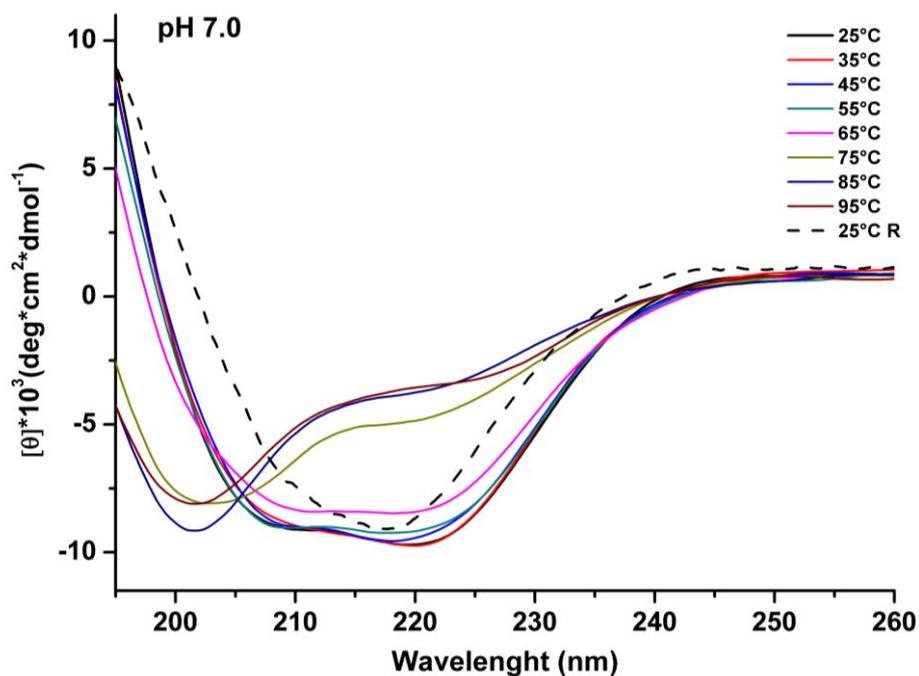


Figure 53. Far-UV circular dichroism spectra of Hemp1 at pH 7.0, measured under thermal denaturation (25 to 95 °C) in intervals of 10 °C. The dashed line indicates the spectrum measured after the protein's temperature returned to 25 °C.

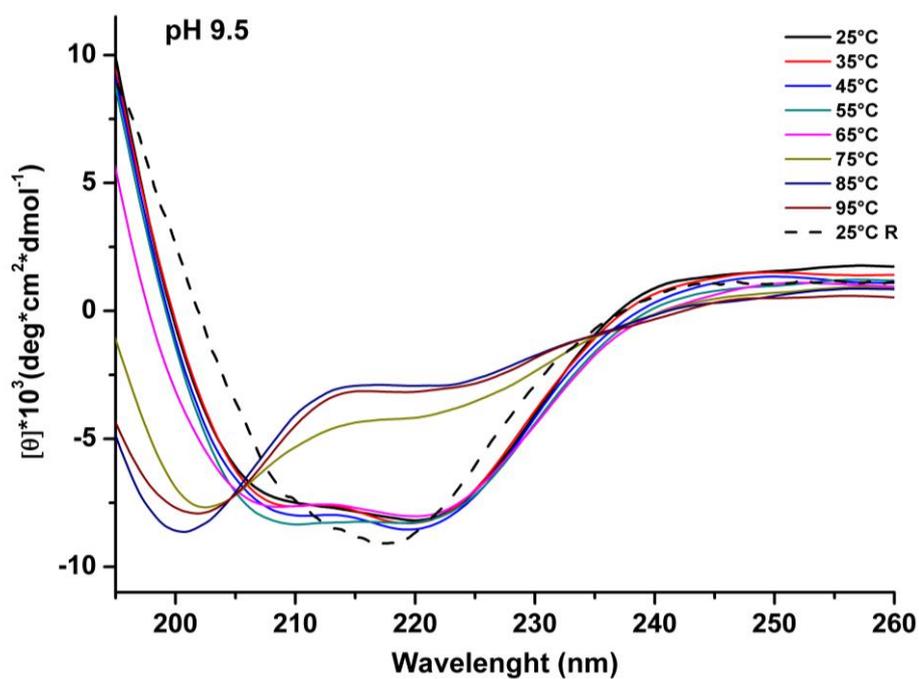


Figure 54. Far-UV circular dichroism spectra of Hemp1 at pH 9.5, measured under thermal denaturation (25 to 95 °C) in intervals of 10 °C. The dashed line indicates the spectrum measured after the protein's temperature returned to 25 °C.

The curves were normalized in order to estimate the proteins' unfolding thermodynamic parameters and T_m values through the Gibbs free energy equation (**Equation 4**). The parameters estimated as described in the methodology are presented in **Table 3**. A positive value of the unfolding's Gibbs free energy indicates that it's an endergonic process, with higher values meaning higher thermal stability. The Gibbs free energy calculation depends on an enthalpy variation, in this case mostly related to the proteins' non-covalent interactions, and the entropy variation, caused by a change in the hydration and degrees of freedom of the proteins' residues. All studied proteins presented decreased thermal stability in acidic pHs in comparison to neutral and basic pHs, as indicated by the unfolding standard Gibbs free energy (ΔG_{25}) values, with their minimum values found in pH 4.0. However, even in acidic pHs the proteins presented moderate to high thermostability as shown by the ΔG_{25} higher than 4 kcal/mol. Hop1 showed little ΔG_{25} variation when analyzed in different pHs (from 12.48 kcal/mol at pH 4.0 to 14.28 kcal/mol at pH 7.0) while Hemp1, despite its sequence similarity to Hop1, presented the highest ΔG_{25} variation, from 4.87 kcal/mol at pH 4.0 to 12.20 kcal/mol at pH 9.5. Late's unfolding thermodynamic parameters were not calculated as the protein did not unfold completely in the experiment's temperature range and Theo was not analyzed by circular dichroism due to the presence of impurities in the obtained samples.

Table 3. Thermodynamic parameters for the unfolding of Hop1, Hop5 and Hemp1. The melting temperatures estimated by the sigmoidal fit of the curves is indicated in parenthesis.

Protein	pH	T_m (curve fit) (°C)	ΔG_{25} (kcal/mol)	ΔS (cal/mol*K)	ΔH (kcal/mol)
Hop1	4.0	82.3 (84.3)	12.48	217.65	77.34
	7.0	84.8 (85.4)	14.28	239.02	85.51
	9.5	83.7 (84.4)	13.47	229.65	81.90
Hop5	4.0	65.8 (67.0)	9.08	222.26	75.31
	7.0	70.2 (69.9)	11.72	258.95	88.88
	9.5	72.4 (72.5)	12.18	257.04	88.78
Hemp1	4.0	60.6 (65.3)	4.87	137.06	45.72
	7.0	71.2 (70.9)	9.47	204.75	70.49

9.5	71.3 (71.6)	12.20	263.37	90.68
-----	-------------	-------	--------	-------

The dichroic spectra of Late, Hop1 and Hop5 exhibited a more intense signal at 218 nm, which is associated with the contribution of beta sheets to the signal, result that confirms the predominance of beta sheets in cystatins (TURK et al., 2008), while Hemp1's most intense signal was at 222 nm, wavelength associated with the contribution of alpha helices to the signal. In general, phytocystatins are considered thermostable, as seen for TcCYS3 and TcCYS4 (*Theobroma cacao* cystatins), which present 100 % of their inhibitory activity after being heated at 80 °C for 10 minutes (FREITAS et al., 2015) or for VuCys1 and VuCys2 (*Vigna unguiculata* cystatins), whose CD spectra only showed altered signals when the proteins were incubated at temperatures higher than 80°C (JÚNIOR et al., 2017). Despite showing similar amino acid contents, as shown in the sequence alignment (**Figure 5**), the difference in the thermal stability between the studied proteins may arise from the position in which these residues are found. According to the thermal denaturation curves (**Figures 39, 43, 47 and 51**), all studied proteins are more stable at basic pHs than at acidic pHs as indicated by the T_m 's and the unfolding standard Gibbs free energies which were higher than 12 kcal/mol for Hop1, Hop5 and Hemp1 at pH 9.5 in contrast to values as low as 9.08 kcal/mol for Hop5 and 4.87 kcal/mol for Hemp1 at pH 4.0. Charged and non-polar residues are responsible to the maintenance of the protein structure, due to intramolecular interactions and the hydrophobic effect. As a consequence, pH has a direct effect in the protein stability, as the charge of some sidechains may be altered according to the environment pH. At pH 4.0, the glutamic acid residues ($pK_a = 4.3$) are not completely ionized. As such residues become deprotonated, acquiring negative charge, the protein structure is stabilized via electrostatic interactions with positively charged residues. At pH 7.0, we speculate that the charge loss by the histidine residues ($pK_a = 6.0$) through deprotonation further increases the conformational stability due to the increase in the protein's hydrophobicity. Similarly, at pH 9.5, more histidine residues present no liquid charge, contributing to the structure stabilization through the hydrophobic effect.

5.5. CRYSTALLOGRAPHY

5.5.1. Protein Crystallization

The mosquito[®] HTS robot (TTP Labtech) was employed to perform automated crystallization screenings with Theo, Late, Hop1, Hop5 and Hemp1 by the hanging drop method, using solutions from the crystallization kits Crystal Screen (Hampton Research), MembFac (Hampton Research), JBScreen JCSG++ (Jena Bioscience) and JBScreen PACT++ (Jena Bioscience).

Despite the wide range of protein and buffer concentrations used in the screening assays, no crystallization solution favored the appearance of Theo crystals, possibly due to the considerable amount of contaminants in the purified samples, as one can see in **Figure 16**.

Late crystals grew in 20 % w/v polyethylene glycol 6000 (PEG 6000), 100 mM sodium citrate pH 4.0, 1 M lithium chloride and ≈ 12.4 mg/mL Late (**Figure 55A**) from the second SEC peak (**Figure 12**). We performed then manual condition refinement assays using the sitting drop method, by varying the concentration of PEG 6000 and the solution pH. Crystals like the ones seen in **Figure 55B**, were seen in most of the plate, especially when concentration of PEG 6000 was between 16 % and 20 %.

The crystallization screenings produced Hop1 crystals in two different conditions, the first, comprising 30 % w/v PEG 4000, 100 mM sodium acetate trihydrate pH 4.6, 200 mM ammonium acetate (**Figure 56A**) and the second comprising 30 % w/v PEG 8000, 100 mM sodium cacodylate pH 6.5 and 200 mM sodium acetate trihydrate (**Figure 56B**) with ≈ 11 mg/mL Hop1 from the first SEC peak (**Figure 18**). We were not able to reproduce any of the Hop1 crystals in manual refinement assays. Nevertheless, the Hop1 crystals obtained through the automated screening were suitable for X-ray diffraction.

Hop5 crystallized in a larger amount of conditions in the automated screening. Using ≈ 12.35 mg/mL Hop5 from the third SEC peak (**Figure 21**), the protein crystallized in a few conditions from the Crystal Screen and JBscreen JCSG++ kits and several conditions from the JBScreen PACT++ kit, including one comprising 20 % w/v PEG 3350, 100 mM BIS-TRIS propane pH 8.5, 200 mM sodium sulfate (**Figure 57A**) and another comprising 20 % w/v PEG 3350, 100

mM BIS-TRIS propane pH 8.5, 200 mM potassium sodium tartrate (**Figure 57B**). The conditions that gave rise to the most promising crystals were selected to be refined manually, but unfortunately we weren't able to reproduce the crystals in sitting drop assays.

Hemp1 crystals were obtained in the automated screenings using ≈ 13 mg/mL Hemp1 from the third SEC peak (**Figure 24**) in various conditions from the crystallization kits, including a condition comprising 20 % w/v PEG 3350 and 200 mM di-ammonium citrate pH 5.0 (**Figure 58A**), which was refined in manual sitting drops assays (**Figure 58B**). All crystals appeared in a time frame of 15 min to 24 hours, and although easily reproduced in manual refinement assays, most of them didn't diffract the synchrotron light with enough intensity for the collection of a useful x-ray diffraction data set.

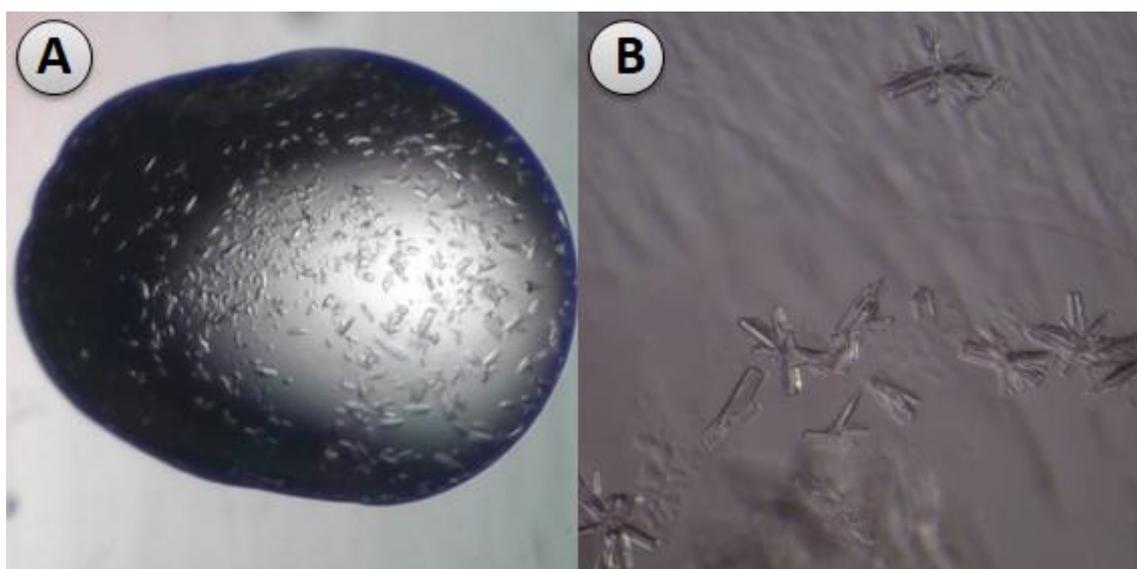


Figure 55. Late crystals obtained in the crystallization condition containing 20 % w/v PEG 6000, 100 mM sodium citrate pH 4.0, 1 M lithium chloride and ≈ 12.4 mg/mL Late. **A.** Automated screening (hanging drop). **B.** Manual refinement (sitting drop).

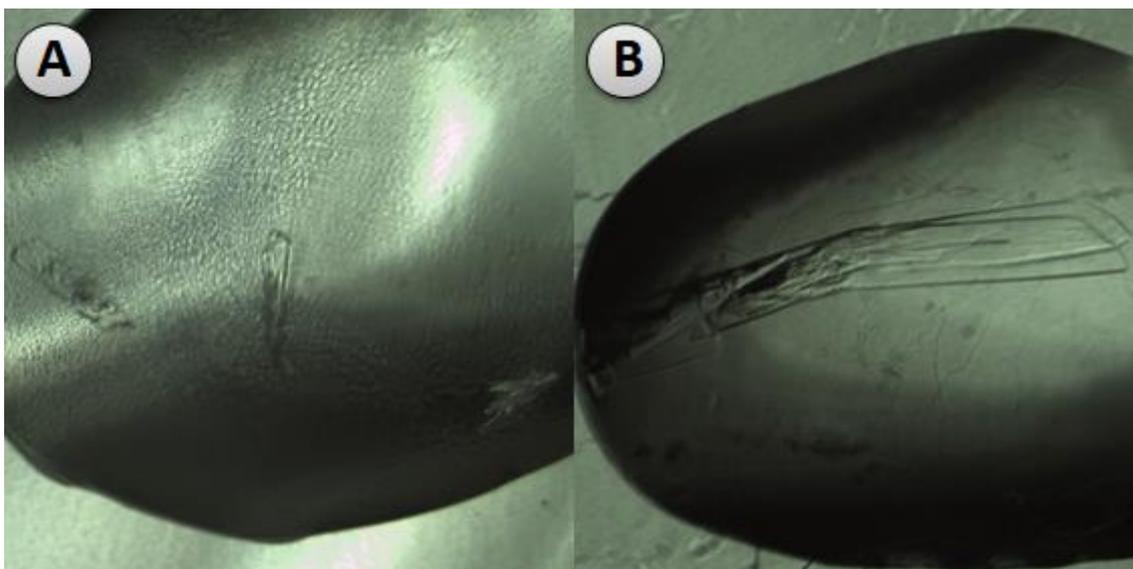


Figure 56. Hop1 crystals obtained in the automated screening (hanging drop) after three weeks using ≈ 11 mg/mL Hop1. **A.** 20 % w/v PEG 4000, 100 mM sodium acetate trihydrate pH 4.6 and 200 mM ammonium acetate. **B.** 30 % w/v PEG 8000, 100 mM sodium cacodylate pH 6.5 and 200 mM sodium acetate trihydrate.

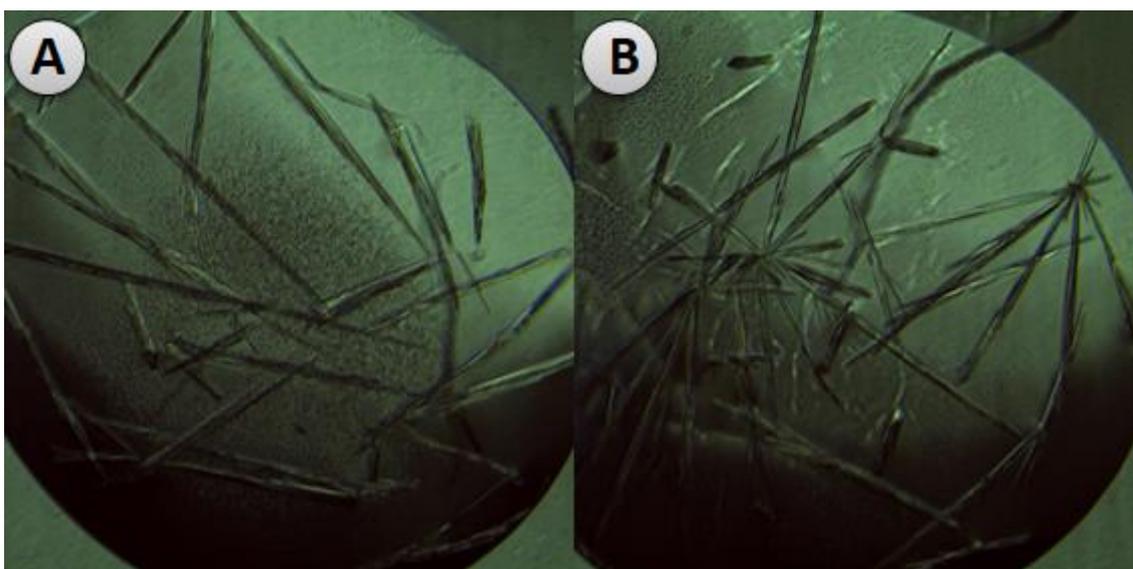


Figure 57. Hop5 crystals seen 1 week after the automated screening (hanging drop) made with the crystallization kit JBScreen PACT++ and ≈ 12.35 mg/mL Hop5. **A.** Crystallization condition containing 20 % w/v PEG 3350, 100 mM BIS-TRIS propane pH 8.5, 200 mM sodium sulfate. **B.** Crystallization condition containing 20 % w/v PEG 3350, 100 mM BIS-TRIS propane pH 8.5, 200 mM potassium sodium tartrate.

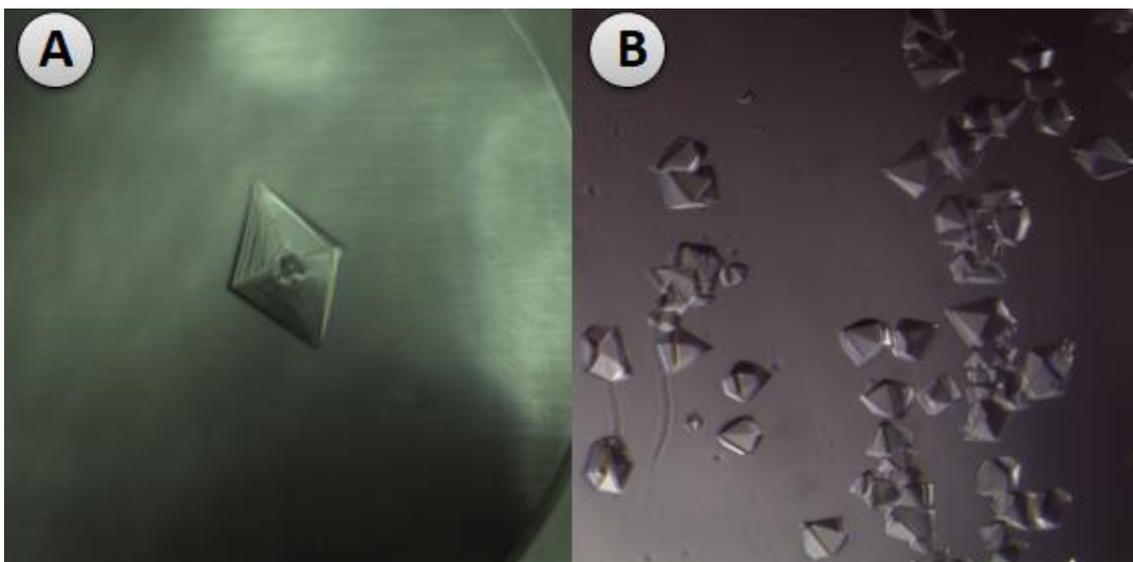


Figure 58. Hemp1 crystals obtained in the crystallization condition 20 % w/v PEG 3350, 200 mM di-ammonium citrate pH 5.0 with ≈ 12.9 mg/mL Hemp1. The crystals appeared 15 min after the beginning of the assay. **A.** Automated screening (hanging drop). **B.** Manual refinement (sitting drop).

5.5.2. X-Ray Diffraction and Crystal Structure Resolution

The crystals obtained in the crystallization assays were taken to LNLS in Campinas, São Paulo, for X-ray diffraction experiments. The data were collected in the MX2 beamline with the Pilatus M2 Detector and X-rays with a wavelength of 1.45 Å. Each collected image consisted in the X-ray diffraction pattern obtained when rotating the diffracting crystal by small amounts. Although dozens of crystals were tested in the beamline, most of them did not present appropriate diffraction. These observed diffraction patterns did not present visible spots, i.e. areas in the detector where the constructive interference of the diffracted waves results in an intense signal, indicating that the crystals did not have appropriate dimensions or that their crystal packing were not sufficiently ordered. Nonetheless, two data sets of Hop1 and one data set of Hemp1 were collected and solved.

Both Hop1 data sets were collected from crystals grown in 30 % w/v PEG 8000, 100 mM sodium cacodylate pH 6.5 and 200 mM sodium acetate trihydrate (**Figure 56**). One of the data sets, comprising 1800 images, was recorded with a distance to the detector of 12 cm, with a beam exposure time of 4 s, while rotating the crystal 0.2° per image. The other data set, comprising 3600

images, was recorded with a distance to the detector of 14 cm, with a beam exposure time of 1 s, while rotating the crystal 0.1° for each image.

The Hemp1 data set was collected from a crystal grown in 18 % w/v PEG 3350 and 200 mM di-ammonium citrate pH 5.2 (**Figure 58**). 720 images were collected, with a detector distance of 35 cm, and an exposure time of 10 s per image, while rotating the crystal 0.5° per image.

XDS readily determined that the two analyzed Hop1 crystals belong to the orthorhombic space groups $C 2 2 2_1$ and $P 2 2_1 2_1$ and the high resolution cutoff of their diffraction data were 1.68 Å and 1.80 Å, respectively. Meanwhile, the Hemp1 crystal space group estimates by XDS were inconclusive. Therefore, it was only determined during the phasing step, using the PHASER software of the CCP4 package and the data processed by XDS in each one of the most probable space groups. Following this laborious step, the hexagonal space group $P 6_5 2 2$ was assigned to the Hemp1 crystal and its diffraction data resolution was 3.6 Å. The programs ZANUDA (LEBEDEV & ISUPOV, 2012) and POINTLESS (CCP4) were used to assure the correct assignment of the space groups.

The high resolution data cutoff was defined taking into consideration the values of the completeness, diffraction intensity to noise ratio and the Pearson's correlation coefficient, which was shown to be a reliable data and processing quality indicator (KARPLUS & DIEDERICHS, 2012).

The PHASER software of CCP4 was employed to retrieve the phase information of the data sets and to solve the tridimensional structures using the crystal structure of 4TX4 (JÚNIOR et al., 2017), a phytocystatin from cowpea (*Vigna unguiculata*), as a search model. Prior to the phasing, the 4TX4 structure, which shares an identity of over 70% with both Hop1 and Hemp1, was prepared by the CHAINSAW software of the CCP4. The quality of the solutions provided by PHASER was estimated by the electron density map and the values of the translational function Z-score (TFZ) and log-likelihood gain (LLG) generated by the solution.

The data collection and refinement statistics for both Hop1 data sets are presented in **Table 4**. Hemp1 data is not shown as its structure is still undergoing refinement. In crystallographic structures, regions with higher flexibility such as exposed sidechains and glycine loops generate less defined electron density maps. This becomes a major concern when considering the Hemp1 data set due

to the high amount of solvent in its structure and its low resolution, which renders its refinement a challenging task.

Table 4. Data collection and refinement statistics for two Hop1 data sets. Statistics for the highest resolution shell are shown in parentheses.

Space group	<i>C</i> 2 2 2 ₁	<i>P</i> 2 2 ₁ 2 ₁
Resolution range (Å)	28.99 - 1.68 (1.74 - 1.68)	25.78 - 1.80 (1.86 - 1.80)
Unit cell dimensions a, b, c (Å)	42.55 63.28 72.38	28.92 43.82 56.83
Total reflections	138709 (10633)	83639 (5203)
Unique reflections	11170 (1011)	6715 (506)
Multiplicity	12.4 (10.5)	12.5 (10.3)
Completeness (%)	97.44 (89.62)	93.49 (69.05)
R_{meas}	0.08523 (1.643)	0.08546 (3.074)
Mean I/σ(I)	20.72 (1.42)	17.13 (0.67)
CC_{1/2}	0.999 (0.554)	0.999 (0.312)
Reflections used in refinement	11166 (1010)	6665 (473)
Reflections used for R-free	557 (50)	335 (25)
R_{work}	0.1911 (0.3086)	0.2200 (0.4087)
R_{free}	0.2092 (0.3072)	0.2666 (0.4311)
Number of protein atoms	651	627
Number of solvent atoms	70	14
Protein residues	80	79
RMS (bonds)	0.008	0.007
RMS (angles)	1.22	1.10
Ramachandran favored (%)	98.72	97.40
Ramachandran allowed (%)	1.28	2.60
Rotamer outliers (%)	0.00	0.00
Clashscore	2.23	1.61
Average B-factor	32.04	46.85
Number of TLS groups	3	3

5.5.3. Crystal Packing

5.5.3.1. Hop1

A crystal lattice analysis of the two different orthorhombic Hop1 crystals revealed a series of potentially relevant contact interfaces between symmetry related protein subunits. In the $C 2 2 2_1$ and $P 2 2_1 2_1$ crystals, 13 and 14 of such interfaces were identified, respectively. In both structures, domain-swapped dimers are formed by symmetry related molecules.

The amount of solvent in the unit cell of each crystal was estimated by calculating their Matthews coefficients, as described in the methodology. The unit cell of the Hop1 $C 2 2 2_1$ crystal presented 37.88 % of solvent while the unit cell of the Hop1 $P 2 2_1 2_1$ crystal presented only 15.96 % of solvent, showing that despite containing the same protein and crystallizing in the same condition and droplet, the space group and unit cell parameters difference led to changes in the crystal packing.

In both Hop1 crystal lattices, common interactions between symmetry related subunits were observed. The most prominent involves extensive contacts between the helices and the loop connecting the helix to the $\beta 2$ strand (**Figure 59**). In each Hop1 symmetry mate, Arg24 of one of the subunits hydrogen bonds Asp28 of the other subunit while Leu37 is in close contact with its identical counterpart.

In order to compare these contacts, the two subunits responsible for this helix interface were extracted from each crystal form and a superposition was performed between the two different space groups, resulting in an rmsd of 1.03 Å. The PYMOL script “angle_between_helices” was employed to calculate the angle between the interacting helices using four different methods. The calculated angles for the $C 2 2 2_1$ space group ranged from 50° to 61°, while for the $P 2 2_1 2_1$ space group, the values ranged from 47° to 58°. The superposition of the structures and the comparison of the angles between the helices shows that this interaction is very similar in both crystal lattices.

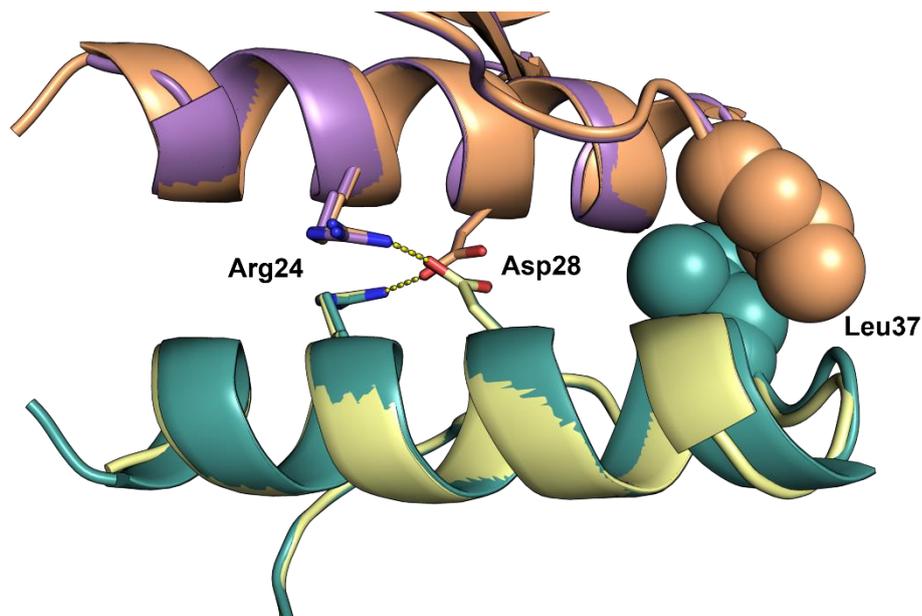


Figure 59. Similarities between interacting helices in the crystal lattice from Hop1 crystals of different space groups. The subunits of the $C 2 2 2_1$ space group are represented in orange and dark green, while the subunits of the $P 2 2_1 2_1$ space group are represented in purple and light green. Some of the interacting residues are shown as sticks, and Leu37 is represented as spheres.

Another contact between symmetry mates in the Hop1 $C 2 2 2_1$ crystal involves residues of different secondary structure elements such as $\beta 2$ to $\beta 4$ strands and residues close to the first inhibitory loop (**Figure 60**). These contacts consist mainly of a salt bridge between Glu61 and Lys42, a hydrogen bond between Asn45 and the main chain carboxyl group of Asp94 and a hydrogen bond between Lys47 and the main chain of Leu92. In addition, a C-terminal Pro96 residue is also in close contact to Val43, contributing to this interface.

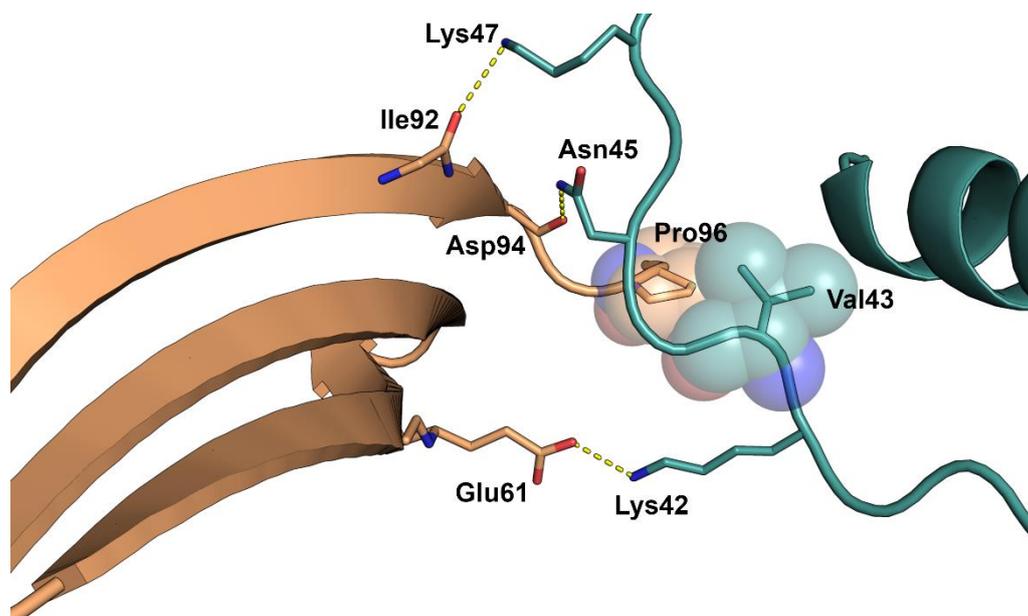


Figure 60. Interaction between Hop1 symmetry mates in the space group $C 2 2 2_1$ involving residues close to the first inhibitory loop. Residues containing polar atoms in close contact are represented as sticks, and hydrogen bonds are shown as yellow dashes.

It is also noteworthy in the $C 2 2 2_1$ crystal that the acidic group of Asp64 is in the vicinity of the side chain nitrogen of Trp80 (**Figure 61**). This tryptophan is part of the second inhibitory loop and this interaction may influence its conformation inside the crystal lattice.

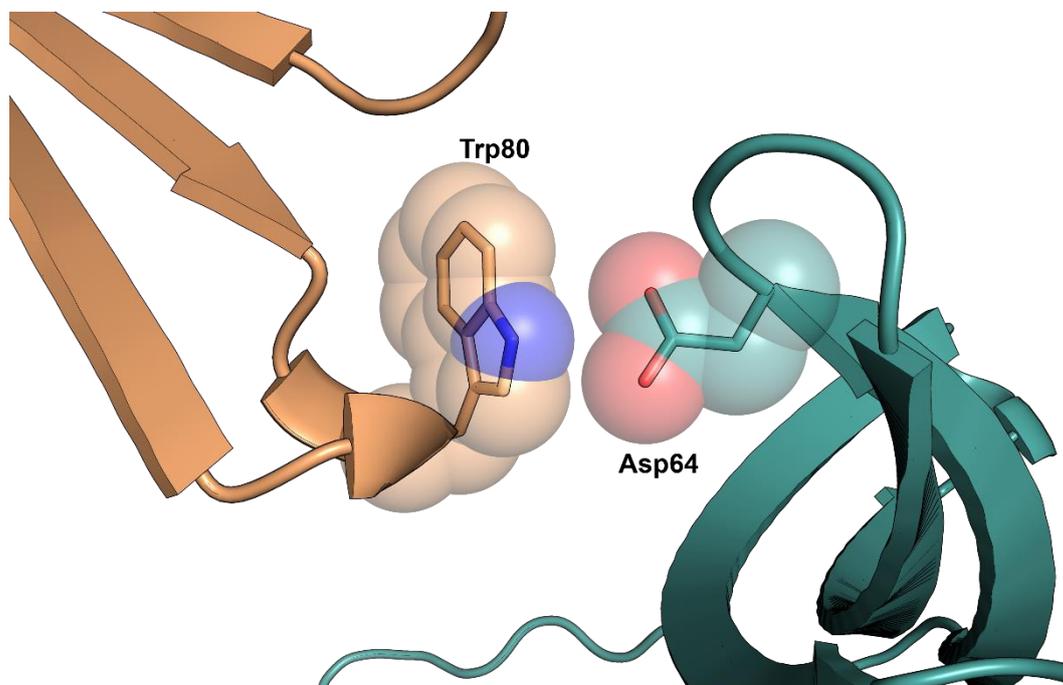


Figure 61. The packing of Trp80 in the $C 2 2 2_1$ Hop1 crystal. Trp80 and Asp64 are labeled and shown as sticks, and transparent spheres delimit their van der Waals radius.

In contrast, in the space group $P 2_1 2_1$, Trp80 is tightly packed between Met81 of the same Hop1 molecule and the C-terminal Pro96 of a symmetry mate (**Figure 62**). Additionally, a relevant intramolecular interaction comprises the side chain nitrogen atom of Trp80 in a polar contact with the carboxyl oxygen atom of Ser52, a residue of the open interface.

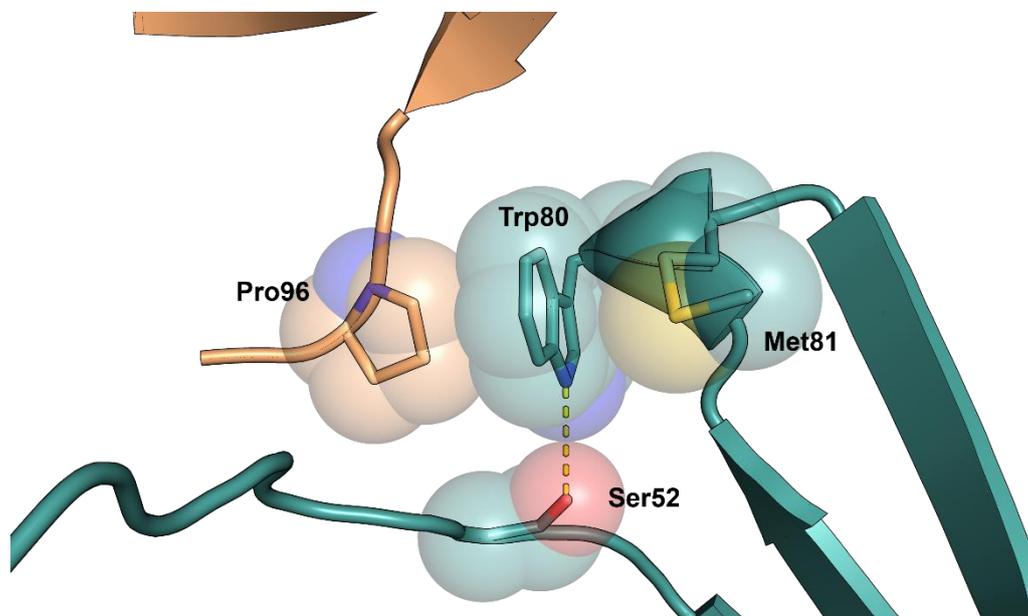


Figure 62. The packing of Trp80 in the $P 2_1 2_1$ Hop1 crystal. Ser52, Trp80 and Met81 are labeled and shown as sticks, and transparent spheres delimit their van der Waals radius. A symmetry mate is colored in salmon and Pro96 is labeled.

5.5.3.2. Hemp1

The Hemp1 crystal structure is still undergoing refinement, therefore a thorough analysis involving its side chains will be presented posteriorly. However, it is possible to study its crystal packing at this stage. An analysis of the Matthews coefficient of the Hemp1 crystal showed a much higher solvent content compared to the Hop1 crystals (68.66 %). Unlike Hop1, the Hemp1 structure presents a domain swapped dimer in its asymmetric unit. In the crystal packing, each domain-swapped dimer contacts only three other domain swapped dimers, leading to the formation of large solvent channels in the lattice (**Figure 63**). This contrasts with the orthorhombic structures of Hop1 in $C 2 2 2_1$ and in $P 2_1 2_1$ in which the domain-swapped dimers contact 8 and 11 other domain swapped dimers, respectively. This combination of high solvent content and the peculiar

arrangement of the domain-swapped dimers forming large solvent channels might explain the low resolution of the collected Hemp1 data sets.

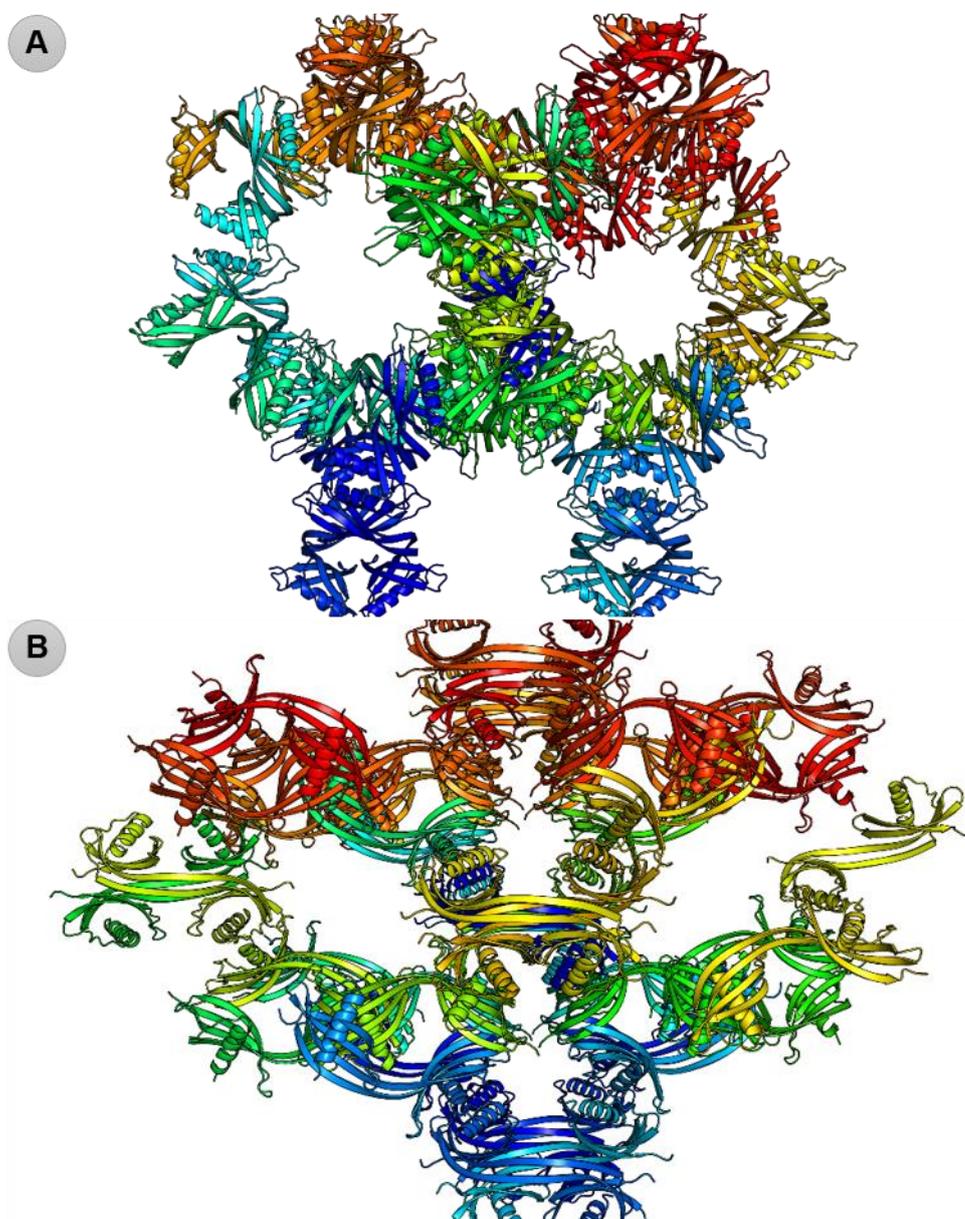


Figure 63. Solvent channels in the Hemp1 $P6_5 2 2$ crystal packing. The colors represent different Hemp1 dimers in the crystal lattice. **(B)** shows **(A)** after a 90° rotation.

5.5.4. Crystal Structure

Hop1 was solved in two different orthorhombic space groups, $C2 2 2_1$ and $P2 2_1 2_1$ at resolutions of 1.68 \AA and 1.80 \AA , respectively. In the $C2 2 2_1$ structure the observed electron density comprised amino acids 18 to 97, while in the $P2 2_1 2_1$ structure the electron density allowed a fit of amino acids 19 to 97. As

disclosed in the last sections, Hop1 presents itself as dimers in solution. In agreement, domain swapped dimers are observed in the crystal lattices of both space groups (**Figure 64**). The lobes of the Hop1 dimers present a small inclination towards each other, where the helices present angles ranging from 46° to 60° for the $C 2 2 2_1$ structure and 47° to 62° for the $P 2 2_1 2_1$, depending on the method used to measure the angle. The region of the first inhibitory loop, which corresponds to the hinge loop of the protein, and its electron density are shown in **Figure 65**.

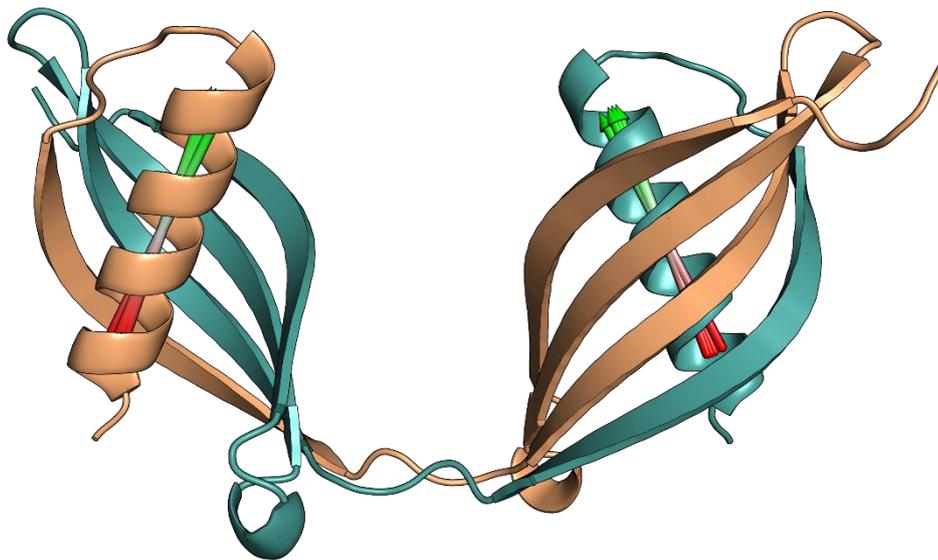


Figure 64. Hop1 structure solved from the $C 2 2 2_1$ crystal. Each subunit of the dimer is showed in a different color, and the arrows inside the helices show the axis fit made by PYMOL to calculate the angle between them.

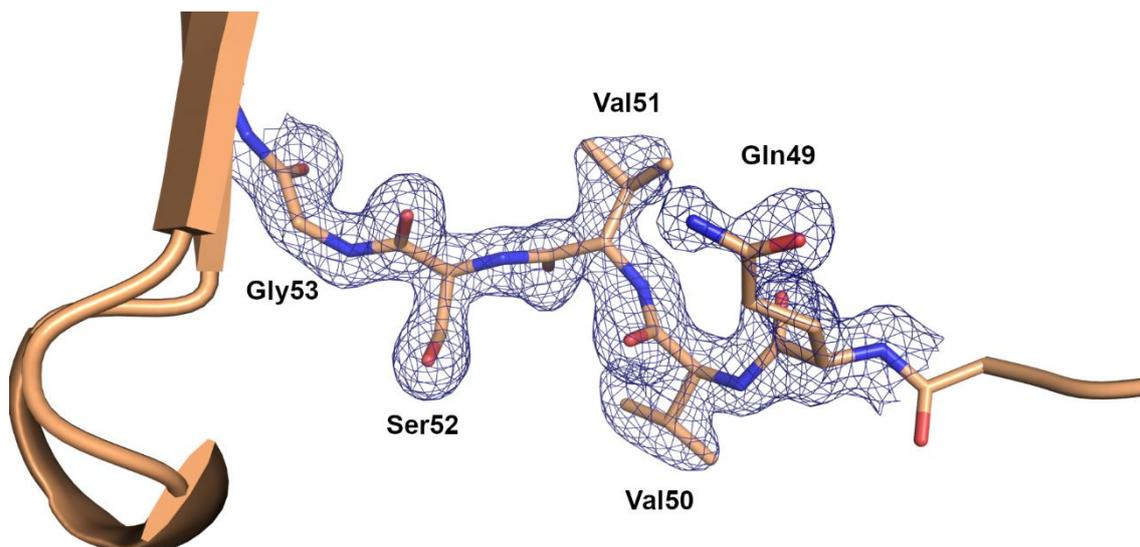


Figure 65. $2F_o-F_c$ electron density map for the first inhibitory loop of Hop1 from the structure in the $C222_1$ space group. The electron density is showed only for the conserved residues QVVSG, contoured at 1.0σ . These residues are presented as sticks and the protein is depicted in cartoon.

The orthorhombic Hop1 structures share many similarities regarding their secondary structure elements. The superposition of both structures was generated with PYMOL leading to an rmsd of 0.77 \AA (**Figure 66**).

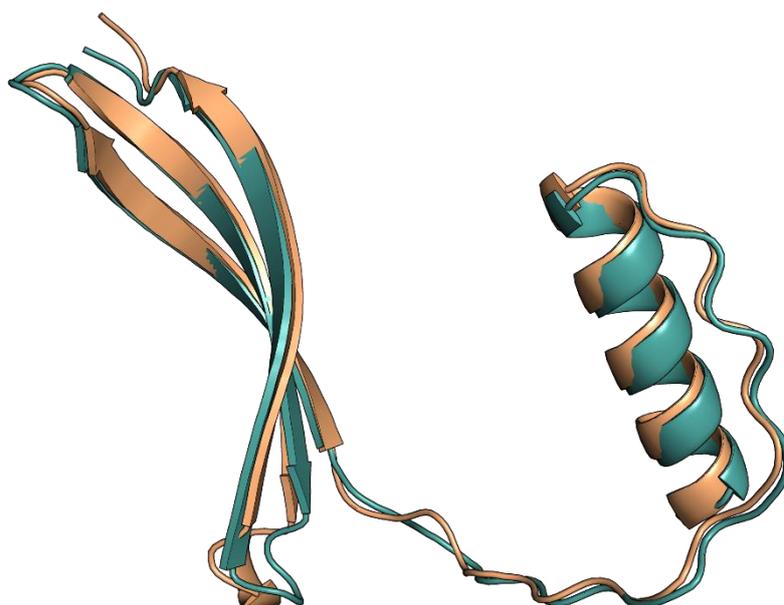


Figure 66. Superposition of Hop1 crystal structures. The $C222_1$ structure is shown in orange while the $P22_12_1$ structure is shown in dark green. The superposition yields an rmsd of 0.77 \AA .

In both Hop1 structures the main chain carboxyl group of Ser52 hydrogen bonds the nitrogen of a tryptophan in the same chain (**Figure 67**). The striking difference is that in the $C 2 2 2_1$ structure Ser52 contacts Trp76, while in the $P 2 2_1 2_1$ structure Ser52 contacts Trp80. This interaction may contribute to the rotamer changes between the tryptophan residues in the structures of different space groups.

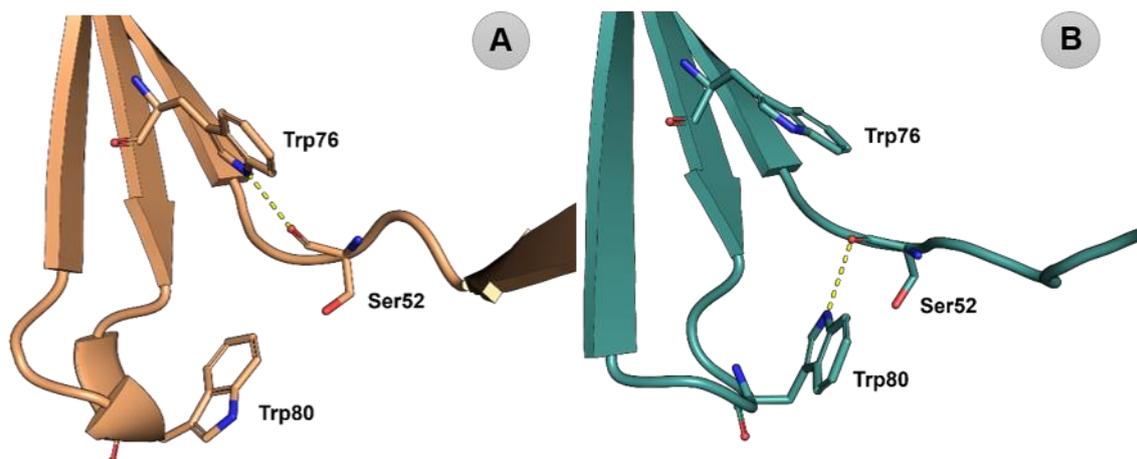


Figure 67. Rotamers of Trp76 and Trp80 in the $C 2 2 2_1$ (**A**) and $P 2 2_1 2_1$ (**B**) Hop1 structures. The tryptophan residues and Ser52 are represented as sticks, and the hydrogen bonds are shown as yellow dashes.

Like in other phytocystatins, an extensive hydrophobic core is formed between the α helix and the β sheet, including at least seven helix side chains and seven β sheet side chains. Curiously, only two polar contacts are observed between residues in this region, both involving the hydroxyl oxygen of Tyr25, the first with the main carbonyl of Lys84 and the second with the main chain nitrogen atom of Thr46. Residues in that region present a high degree of conservation, such as the ones located in the consensus sequence in the α helix that identifies phytocystatins [LVI]-[AGT]-[RKE]-[FY]-[AS]-[VI]-X-[EDQV]-[HYFQ]-N. In Hop1, this sequence is **LARYAVDEHN**, where the side chain of the residues in bold participate in this hydrophobic interface.

A recurrent feature in the tridimensional structure of phytocystatins is the presence of two consecutive β bulges in the $\beta 2$ strand. A β bulge might be described as an alteration of the hydrogen bond pattern in a β strand due the insertion of an additional residue. In the case of phytocystatins, these β bulges contribute to alleviate the torsional strain caused by the curvature of the

antiparallel β sheet (RICHARDSON et al., 1978). This feature allows all the regular main chain hydrogen bonds of the β 3 strand and also that the β sheet curves around the helix, in an arrangement in which all odd residues from the β 3 strand are hydrogen bonding with residues from the β 2 strand. In the Hop1 structure, Glu41 and Lys42 form the first β bulge while Val44 and Asn45 form the second one (**Figure 68**).

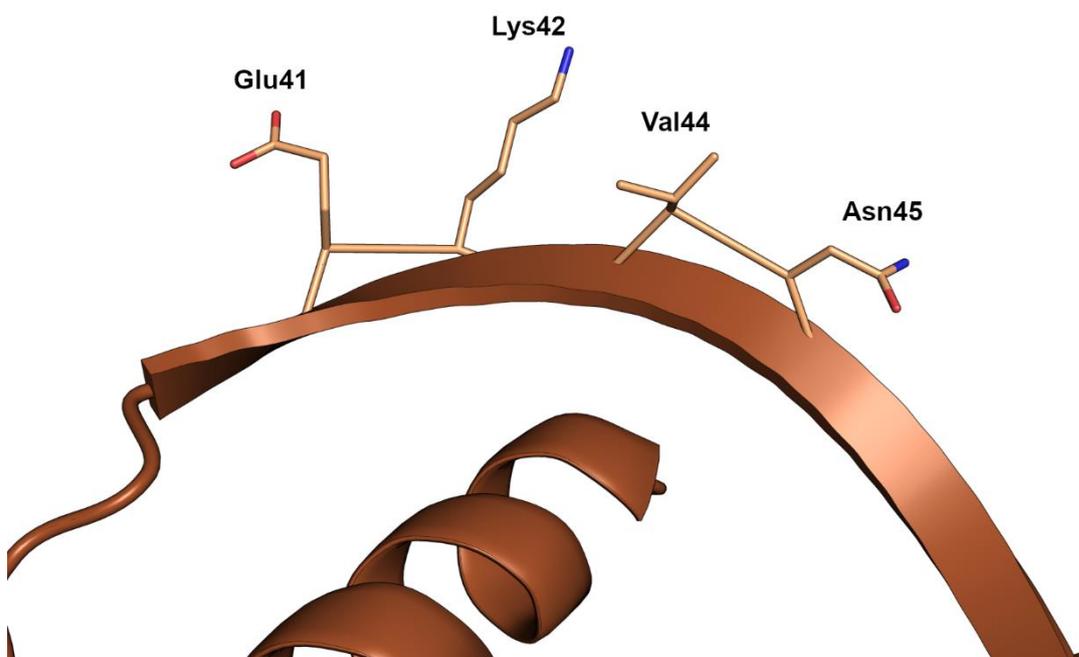


Figure 68. β bulges in the Hop1 C 2 2 2₁ structure. The residues that compose the β bulges are labeled and their side chains are represented as sticks.

Hop1 may be analyzed in parallel to the structure of a cowpea phytocystatin (PDB code 4TX4), the phytocystatin with the highest sequence identity whose crystal structure is available and was used as the search model for the molecular replacement (**Figure 69A**). The crystal structure 4TX4 comprises a domain swapped dimer, as expected for dimeric phytocystatins, with the particularity that the two lobes are positioned closer to each other than in any other phytocystatin deposited in the PDB, allowing direct interactions between side chains of both lobes. This effectively increases the amount of residues that contribute to the open interface, a phenomenon unique to this phytocystatin (JÚNIOR et al., 2017). The angle between the helices measured by PYMOL ranged from 72° to 85° in contrast to the maximum value of 62° to the Hop1. A superposition was performed between one of the lobes of the 4TX4 domain swapped dimer and a lobe of the C 2 2 2₁ Hop1 structure. Though this lobe may

be finely aligned, the structures start to diverge in the region of the first inhibitory loop, leading to a displacement between the second lobes (**Figure 69B**). An analysis of the main chain dihedral angles showed that this abrupt difference between the 4TX4 and the Hop1 structure happens in the vicinity of Thr46, next to the second β bulge.

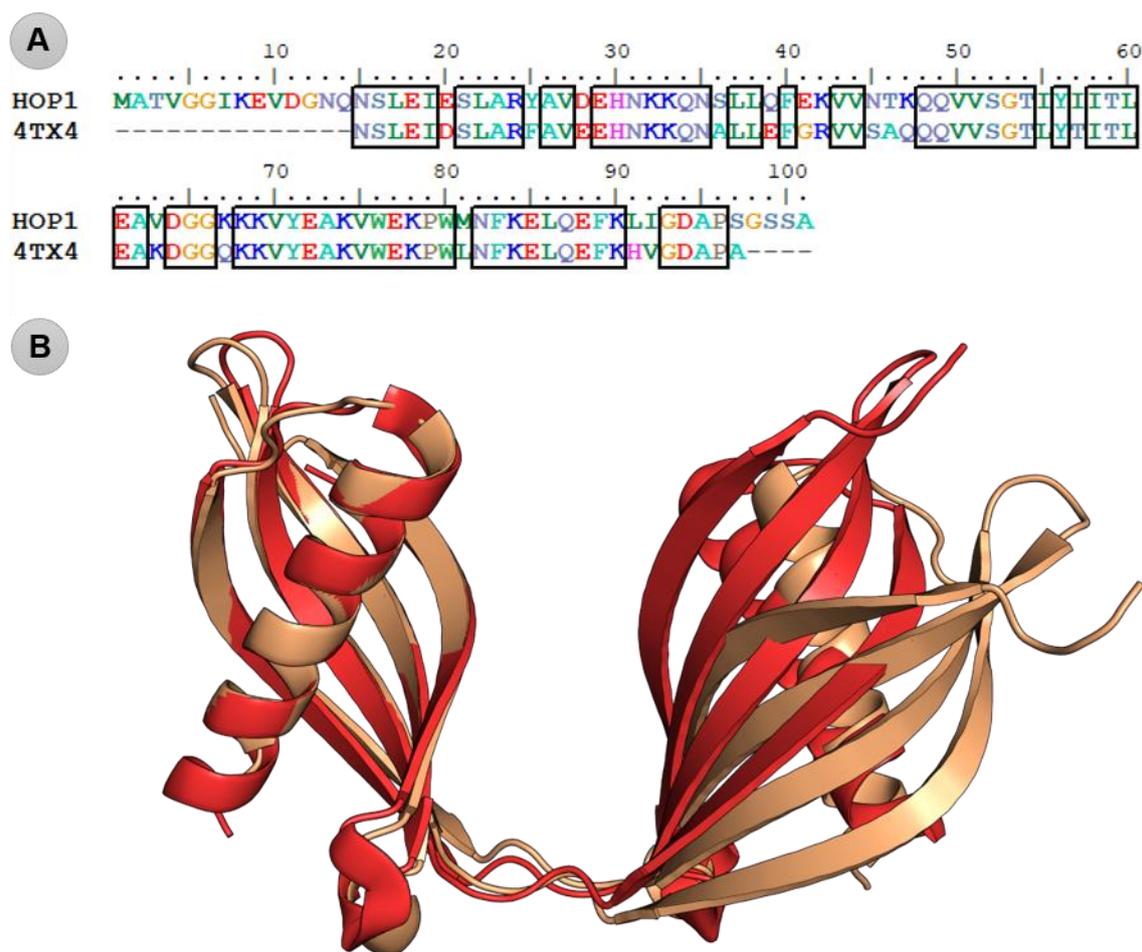


Figure 69. Similarities between Hop1 and a cowpea phycocystatin (PDB code 4TX4). **A.** Sequence alignment. The outline residues are present in both sequences **B.** 4TX4 crystal structure superposed with the $C 2 2 2_1$ Hop1 structure. 4TX4 is showed in red, while Hop1 is showed in orange. The lobes on the left were aligned with PYMOL, and due to the different angles between the lobes of these two phycocystatins, the lobes on the right are displaced in relation to each other.

As previously stated, currently the structural analysis of the Hemp1 can only be done at a superficial level. Nevertheless, the structure presents a domain swapped dimer, with each lobe further apart than in the Hop1 or 4TX4 structures, leaving plenty of space for solvent molecules (**Figure 70**). In addition, the lobes

have a pronounced inclination, corresponding to angles from 68° to 77° . It is noteworthy that although the Hemp1 structure diverges significantly from the solved Hop1 structures, the difference between the sequence of these two proteins is only of 10 amino acids (**Figure 5**)

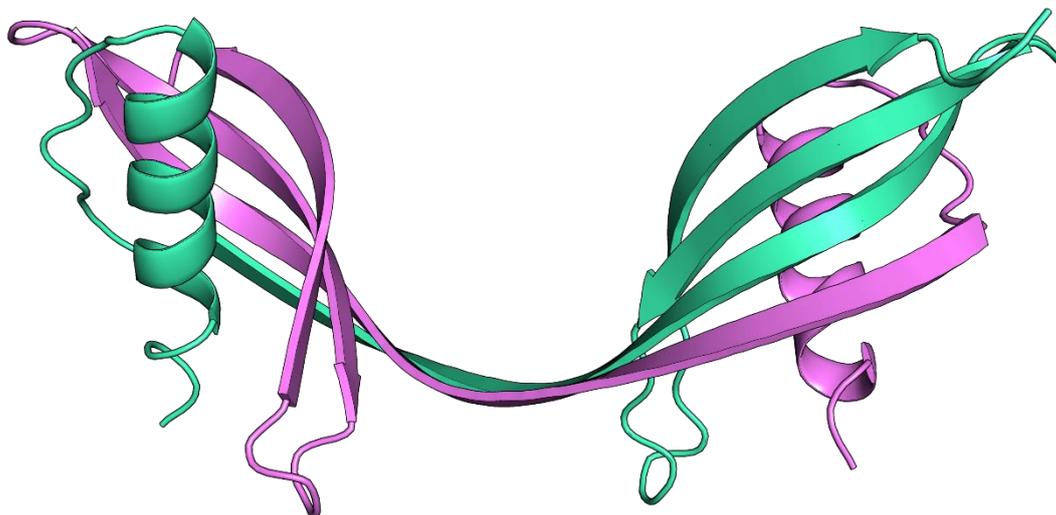


Figure 70. Hemp1 structure solved from a hexagonal $P 6_5 2 2$ crystal. Each color represents a single subunit.

6. CONCLUSIONS

In conclusion, the expression of the studied phytocystatins in the *E. coli* heterologous system using auto inducing medium ZYM-5052 was a success, especially when compared to results obtained using LB medium.

With the exception of Theo, the use of immobilized ion metal affinity chromatography followed by size exclusion chromatography sufficed to produce samples suitable for circular dichroism, analytical ultracentrifugation, and protein crystallization assays.

Ultracentrifugation assays revealed that the studied phytocystatins are mostly dimeric in solution, although Hop5 showed a considerable amount of monomers and Hemp1 presented tetramers. Assays to disclose the effects of pH, temperature and ionic strength on the oligomer dynamics are to be performed.

The studied phytocystatins presented moderate thermal stability, as indicated by the Gibbs free energy higher than 4 kcal/mol after the unfolding process in all studied pHs. All proteins showed more stability in neutral and basic pHs than acidic pHs and Hemp1 might have changed its oligomeric state when heated.

The crystallization assays gave rise to a wide variety of crystals from Late, Hop1, Hop5 and Hemp1, although many of them weren't suitable for X-ray diffraction.

The crystal structure of Hop1 in the two orthorhombic space groups $C 2 2 2_1$ and $P 2 2_1 2_1$ was solved at 1.68 Å and 1.80 Å, respectively. The protein is tightly packed in the crystal as showed by the amount of contacts a subunit makes with its symmetry mates. The protein crystallized as a domain-swapped dimer, where β bulges may be observed.

The Hemp1 crystal structure solved from a hexagonal space group $P 6_5 2 2$ crystal at a 3.6 Å resolution is undergoing refinement due to the challenges imposed by the resolution and amount of solvent in the crystal lattice. Like Hop1, Hemp1 also presented domain-swapped dimers, however, the dimer lobes are further apart.

During the course of this study the methodology to express, purify, characterize biophysically and crystallize phytocystatins was established in the Molecular Biophysics Laboratory of the University of Brasília, paving the way for

the study of other cystatins of interest. The biophysical and structural data obtained will contribute to a better understanding of the structural determinants of domain swapping in phytocystatins, shedding a light on the regulation of the activity of these protease inhibitors with potential biotechnological applications.

7. REFERENCES

- ADAMS, Paul D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. **Acta Crystallographica Section D: Biological Crystallography**, v. 66, n. 2, p. 213-221, 2010.
- BARTH-HAAS GROUP. The Barth Report: Hops 2017/2018. 2018.
- BEHRE, Karl-Ernst. The history of beer additives in Europe—a review. **Vegetation history and archaeobotany**, v. 8, n. 1-2, p. 35-48, 1999.
- BENNETT, M. J.; CHOE, S.; EISENBERG, David. Domain swapping: entangling alliances between proteins. **Proceedings of the National Academy of Sciences**, v. 91, n. 8, p. 3127-3131, 1994.
- BENNETT, M. J.; CHOE, S.; EISENBERG, David. Refined structure of dimeric diphtheria toxin at 2.0 Å resolution. **Protein Science**, v. 3, n. 9, p. 1444-1463, 1994.
- BENNETT, M. J.; EISENBERG, David. Refined structure of monomelic diphtheria toxin at 2.3 Å resolution. **Protein Science**, v. 3, n. 9, p. 1464-1475, 1994.
- BENNETT, Melanie J.; EISENBERG, David. The evolving role of 3D domain swapping in proteins. **Structure**, v. 12, n. 8, p. 1339-1341, 2004.
- BENNETT, Melanie J.; SCHLUNEGGER, Michael P.; EISENBERG, David. 3D domain swapping: a mechanism for oligomer assembly. **Protein Science**, v. 4, n. 12, p. 2455-2468, 1995.
- BOER, E.; BELLA, A. B. **Plant resources of South-East Asia**. Backhuys Publ., 2000.
- BROWN, W. M.; DZIEGIELEWSKA, K. M. Friends and relations of the cystatin superfamily—new members and their evolution. **Protein science**, v. 6, n. 1, p. 5-12, 1997.
- BRÜNGER, Axel T. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. **Nature**, v. 355, n. 6359, p. 472, 1992.
- BUSSO, Didier et al. A comparison of inoculation methods to simplify recombinant protein expression screening in *Escherichia coli*. **Biotechniques**, v. 44, n. 1, p. 101-106, 2008.
- CARDOSO, Thyago Hermylly Santana et al. TcCYPR04, a cacao papain-like cysteine-protease detected in senescent and necrotic tissues interacts with a cystatin TcCYS4. **PloS one**, v. 10, n. 12, p. e0144440, 2015.

- CHATON, Catherine T.; HERR, Andrew B. Elucidating complicated assembling systems in biology using size-and-shape analysis of sedimentation velocity data. In: **Methods in enzymology**. Academic Press, 2015. p. 187-204.
- DIAZ-MENDOZA, Mercedes et al. HvPap-1 C1A protease and HvCPI-2 cystatin contribute to barley grain filling and germination. **Plant physiology**, p. pp. 01944.2015, 2016.
- EKIEL, Irena et al. NMR structural studies of human cystatin C dimers and monomers1. **Journal of molecular biology**, v. 271, n. 2, p. 266-277, 1997.
- EMSLEY, Paul et al. Features and development of Coot. **Acta Crystallographica Section D: Biological Crystallography**, v. 66, n. 4, p. 486-501, 2010.
- FAOSTAT, F. A. O. Statistical databases. **Food and Agriculture Organization of the United Nations**, 2018.
- FREITAS, Ana Camila Oliveira et al. The Activity of TcCYS4 Modified by Variations in pH and Temperature Can Affect Symptoms of Witches' Broom Disease of Cocoa, Caused by the Fungus *Moniliophthora perniciosa*. **PloS one**, v. 10, n. 4, p. e0121519, 2015.
- GASPAROTTO, Luadir et al. **Enfermidades da seringueira no Brasil**. EMBRAPA-CPAA, 1990.
- GATEAU-REY, Lauranne et al. Climate change could threaten cocoa production: Effects of 2015-16 El Niño-related drought on cocoa agroforests in Bahia, Brazil. **PloS one**, v. 13, n. 7, p. e0200454, 2018.
- HWANG, Jung Eun et al. Distinct expression patterns of two Arabidopsis phytocystatin genes, AtCYS1 and AtCYS2, during development and abiotic stresses. **Plant cell reports**, v. 29, n. 8, p. 905-915, 2010.
- JANCARIK, J. A. K. S.; KIM, S.-H. Sparse matrix sampling: a screening method for crystallization of proteins. **Journal of applied crystallography**, v. 24, n. 4, p. 409-411, 1991.
- JÚNIOR, José Edvar Monteiro et al. Expression in Escherichia coli of cysteine protease inhibitors from cowpea (*Vigna unguiculata*): The crystal structure of a single-domain cystatin gives insights on its thermal and pH stability. **International journal of biological macromolecules**, v. 102, p. 29-41, 2017.

- KABSCH, Wolfgang. Xds. **Acta Crystallographica Section D: Biological Crystallography**, v. 66, n. 2, p. 125-132, 2010.
- KANTARDJIEFF, Katherine A.; RUPP, Bernhard. Matthews coefficient probabilities: improved estimates for unit cell contents of proteins, DNA, and protein–nucleic acid complex crystals. **Protein Science**, v. 12, n. 9, p. 1865-1871, 2003.
- KARPLUS, P. Andrew; DIEDERICHS, Kay. Linking crystallographic model and data quality. **Science**, v. 336, n. 6084, p. 1030-1033, 2012.
- KORDIŠ, Dušan; TURK, Vito. Phylogenomic analysis of the cystatin superfamily in eukaryotes and prokaryotes. **BMC evolutionary biology**, v. 9, n. 1, p. 266, 2009.
- KRUGER, Daniel J.; KRUGER, Jessica S. Medical Cannabis Users' Comparisons between Medical Cannabis and Mainstream Medicine. **Journal of psychoactive drugs**, p. 1-6, 2019.
- KUNERT, Karl J. et al. Potential use of phytocystatins in crop improvement, with a particular focus on legumes. **Journal of experimental botany**, v. 66, n. 12, p. 3559-3570, 2015.
- LEBEDEV, Andrey A.; ISUPOV, Michail N. Space group validation with Zanuda. **CCP4 Newsletter Prot Crystallogr**, v. 53, 2012.
- LEVINTHAL, C. How to Fold Graciously. Mossbauer Spectroscopy in Biological Systems, Allerton House. 1969.
- MARGIS, Rogério; REIS, Emerson M.; VILLERET, Vincent. Structural and phylogenetic relationships among plant and animal cystatins. **Archives of Biochemistry and Biophysics**, v. 359, n. 1, p. 24-30, 1998.
- MARTINEZ, Manuel et al. Phytocystatins: defense proteins against phytophagous insects and Acari. **International journal of molecular sciences**, v. 17, n. 10, p. 1747, 2016.
- MATTHEWS, Brian W. Solvent content of protein crystals. **Journal of molecular biology**, v. 33, n. 2, p. 491-497, 1968.
- MCCOY, Airlie J. et al. Phaser crystallographic software. **Journal of applied crystallography**, v. 40, n. 4, p. 658-674, 2007.
- MOURTZINIS, Spyridon et al. Climate-induced reduction in US-wide soybean yields underpinned by region-and in-season-specific responses. **Nature plants**, v. 1, n. 2, p. 14026, 2015.

- MUNGER, Aurélie et al. Cereal cystatins delay sprouting and nutrient loss in tubers of potato, *Solanum tuberosum*. **BMC plant biology**, v. 15, n. 1, p. 296, 2015.
- NAGATA, Koji et al. Three-dimensional solution structure of oryzacystatin-I, a cysteine proteinase inhibitor of the rice, *Oryza sativa* L. japonica. **Biochemistry**, v. 39, n. 48, p. 14753-14760, 2000.
- NEURATH, Hans; WALSH, Kenneth A. Role of proteolytic enzymes in biological regulation (a review). **Proceedings of the National Academy of Sciences**, v. 73, n. 11, p. 3825-3832, 1976.
- PEREIRA, J. L.; DE ALMEIDA, L. C. C.; SANTOS, S. M. Witches' broom disease of cocoa in Bahia: attempts at eradication and containment. **Crop Protection**, v. 15, n. 8, p. 743-752, 1996.
- RAWLINGS, Neil David; BARRETT, Alan John; BATEMAN, Alex. Asparagine peptide lyases: a seventh catalytic type of proteolytic enzymes. **Journal of Biological Chemistry**, p. jbc. M111. 260026, 2011.
- RUSSO, Ethan B. Cannabis Therapeutics and the Future of Neurology. **Frontiers in integrative neuroscience**, v. 12, p. 51, 2018.
- RYAN, Clarence A. Protease inhibitors in plants: genes for improving defenses against insects and pathogens. **Annual review of phytopathology**, v. 28, n. 1, p. 425-449, 1990.
- SANDERS, Anna et al. Cystatin forms a tetramer through structural rearrangement of domain-swapped dimers prior to amyloidogenesis. **Journal of molecular biology**, v. 336, n. 1, p. 165-178, 2004.
- STEIN, Norman. CHAINSAW: a program for mutating pdb files used as templates in molecular replacement. **Journal of Applied Crystallography**, v. 41, n. 3, p. 641-643, 2008.
- STUBBS, Milton T. et al. The refined 2.4 Å X-ray crystal structure of recombinant human stefin B in complex with the cysteine proteinase papain: a novel type of proteinase inhibitor interaction. **The EMBO journal**, v. 9, n. 6, p. 1939-1947, 1990.
- STUDIER, F. William. Protein production by auto-induction in high-density shaking cultures. **Protein expression and purification**, v. 41, n. 1, p. 207-234, 2005.

- SUMANTHA, Alagarsamy; LARROCHE, Christian; PANDEY, Ashok. Microbiology and industrial biotechnology of food-grade proteases: a perspective. **Food Technology and Biotechnology**, v. 44, n. 2, p. 211, 2006.
- TURK, Vito et al. Cystatins: biochemical and structural properties, and medical relevance. **Front Biosci**, v. 13, n. 1, p. 5406-5420, 2008.
- TURK, Vito; BODE, Wolfram. The cystatins: protein inhibitors of cysteine proteinases. **FEBS letters**, v. 285, n. 2, p. 213-219, 1991.
- UNITED NATIONS OFFICE ON DRUGS AND CRIME. **World drug report 2018**. United Nations Publications, 2018.
- VALADARES, Napoleão F. et al. X-ray crystallography and NMR studies of domain-swapped canecystatin-1. **The FEBS journal**, v. 280, n. 4, p. 1028-1038, 2013.
- WINN, Martyn D. et al. Overview of the CCP4 suite and current developments. **Acta Crystallographica Section D**, v. 67, n. 4, p. 235-242, 2011.
- YUAN, Songli et al. Search for nodulation and nodule development-related cystatin genes in the genome of Soybean (*Glycine max*). **Frontiers in plant science**, v. 7, p. 1595, 2016.