CHAPTER 3

Phylogenetic and structural comparisons of phytocystatins: A bioinformatics approach

Scientific Communications

Conference presentation and proceedings:

Kiggundu A., Kunert K. and Michaud D. 2005. The N-terminal trunk of plant cystatins determines their inhibitory specificity against cysteine proteases. Proceedings of Plant Canada 2005 Conference, June 15th -18th, Edmonton Canada.

Phylogenetic analysis and structural modelling from this study contributed to the publication:

Girard C., Rivard D., Kiggundu A., Kunert K., Gleddie S. C., Cloutier C., and Michaud D. 2007. A multicomponent, elicitor-inducible cystatin complex in tomato, *Solanum lycopersicum*. New Phytologist 173 (4), 841–851.

Manuscript in reparation:

Kiggundu A., Kunert K., Viljoen V., Van de Vyver C., and Michaud D. Phylogenetic and structural comparisons of phytocystatins: A bioinformatics approach.

3.1 Abstract

With the use of bioinformatics tools the phylogenetic relationships of phytocystatins based on amino acid sequence information was elucidated and their secondary and tertiary structures were investigated for structural comparisons. Sixty six distinct phytocystatins from 43 plant species and 5 different tissue types were investigated. Inhibition constants for inhibition of the model cysteine protease papain varied greatly from 0.00011nM for chelidocystatin to 19,000nM for a soybean cystatin. Phytocystatins could be divided into five distinct phylogenectic groups but their structural features were highly conserved. Amino acid sequence similarities ranged from 7 to 94%. A new highly conserved amino acid sequence motif, YEAKxKxWxKxF, in the C-terminal end being unique to phytocystatins was identified. The predicted 3D homology models showed a high conservation of the general central structure of the phytocystatins i.e. the 4-5 anti-parallel β-sheets, wrapping halfway round a single central α-helix, and particularly the three active site regions, the N-terminal, the 1st and 2nd hairpin loops. Any structural differences seem to be mainly in the length of the N and C terminal, the length of the 2nd hairpin loop and the 5th β-sheet. Via docking experiments, small heterogeneties were observed in the vicinity of the OC-I active sites that seemed to be influential in the binding process and stability of the resultant inhibitor-protease complex.

3.2 Introduction

Phytocystatins are proteinacious inhibitors of plant origin that inhibit specifically cysteine proteases by forming tight reversible bonds thus preventing the hydrolysis of proteins by proteases. The cystatin super family is subdivided into three families based mainly on the three criteria sequence homology, presence of disulfide bonds and on the molecular mass of the protein. These families are the stefins, cystatins and kininogens. Many different phytocystatins have been isolated from different plants and their gene sequences deposited on public databases.

Phylogenetic analysis provides an insight into the molecular evolution of proteins. Numerous bioinformatic and computational biology tools are now available online providing automated analysis of relationships of proteins at molecular and structural level. Public sequence databases have also provided a very useful and wide range of resources to perform such analyses. One of the key ideas in genomic bioinformatics is the concept of homology. This is used to predict the function of genes and proteins. This is followed by a next level where not only protein function can be predicted but also ere the primary, secondary and tertiary structures of a protein can be predicted. This is achieved through powerful computation methods referred to as *in-silico* analysis. Such analysis provides a better understanding of the microstructures on the protein surface that contribute or may even hinder its proper function.

Part of the aim of this study was therefore to analyse, based on available amino acid sequence information, the phylogenetic relationships of phytocystatins. This was carried out by a comparative study on the primary, predicted 2D and 3D structures of known phytocystatins. In particular the 3D positions of the amino acids involved in

binding, structures of active sites and the local structural variation among members of the proposed phytocystatin family were studied.

3.3 Materials and methods

3.3.1 Sequence analysis

Amino acid sequences of phytocystatins were obtained from various online databases (Table 5.1) using the sequence retrieval system (SRS) (http://srs.embl-heidelberg.de:8000/srs5/). The program BLAST (Altschul *et al*., 1990) was used against the GenBank database to further obtain recent submissions that may not have reached the more advanced databases like European Molecular Biology Laboratory (EMBL) sequence database and the Protein Information Resource (PIR) database.

Multiple alignments were performed using the program CLUSTALX (Thomson *et al*., 1997) with default settings and the alignment edited manually. Long sequences were truncated both at the N and C terminal to include only the domain region and the alignment was repeated. A consensus sequence and a PAM250 (Gonnet *et. al*., 1992) sequence similarly matrix were generated using BIOEDIT suite (Hall, 1999).

Phylogenetic inference was performed using the PHYLIP version 3.5 suite (Felsenstein, 1989). First a distance matrix was generated using the PRODIST program followed by the neighbour joining method using NEIGHBOR program. A consensus tree derived after 1000 bootstraps through the programs BOOTST and CONSES. An un-rooted phylogenetic tree was contracted using the TREEVIEW program (Page, 1996).

3.3.2 Protein structure modelling

The coordinate files (pdb) for OC-I and papain were obtained from the protein data bank (PDB) database. The OC-I pdb file was used to predict the 3D structures of selected representative from each of the phylogenetic groups. Structure modelling to predict the unknown structures was done using the program MODELLER (Sanchez and Sali, 2000) that determines structure using the satisfaction of spatial constraints. The input files consisted of the pdb file of OC-I and the amino acid sequence alignment between OC-I with the unknown sequence at greater that 30% sequence similarity with OC-I. Predicted models were evaluated for energy distribution. Stereo-chemical quality of the predicted structures was tested using the ENERGY command on MODELLER and PROCHECK (Laskowski, 1993) programs, respectively. Structures were visualised using both SWISS-PDB Viewer (Guex and Peitsch, 1997) and PYMOL (www.pymol.org).

3.3.3 Active site and docking

Based on the X-ray crystal structure of recombinant human stefin-B and papain (Studds *et al*., 1990), the structure of OC-I bound to papain was extensively modelled manually followed by refinement using MULTIDOCK program in the 3D-DOCK suite (Jackson *et. al*., 1998). Since the binding structural motifs in stefin-B an animal cystatin are present in OC-I and in other phytocystains, it was expected that OC-I would bind papain in the same manner (Nagata *et al*., 2000).

3.4 Results

To date, phytocystatins have been isolated from at least 43 different plant species (Table 3.1) but the rate at which new members are identified and isolated is rapid. In

this study a total of 66 phytocystatins have been collected either deposited in sequence databases or reported in the literature.

For some of the known phytocystatins characterisation studies including inhibition kinetics have been carried out either on wild-type proteins extracted directly from the plant or recombinant proteins expressed and purified in the laboratory. The known inhibition constants (K_i) for the model cysteine protease papain are included in Table 3.1. Papain K_i values of known phytocystatins ranged from 0.00011nM for *Chelidonium majus* L. (Celandine-chelidocystatin) to 19,000nM for soybean domain L1 cystain, respectively. The celandine plant, from which the most potent phytocystatin known was found, is traditionally used in China and Europe as a herb to treat bacterial and viral infections (Rogel *et al*., 1998) in humans.

Table 3.1 Known phytocystatins obtained from sequence databases: EMBL= European Molecular Biology Laboratory, PIR=Protein information Resource, SP=SwissProt, GB=GeneBank and NCBI=National Centre for Biotechnology Information.

Code	Common name	Specie name	Database[1] (Acc No.)	Papain K_i	Reference[2]
Apple	Apple	*Malus domestica*	EMBL:AY173139	0.2-0.3nM	Ryan *et. al.,* 1998
AraI	Arabidopsis	*Arabidopsis thaliana*	GB:AF315737	-	-
AraII	Arabidopsis	*Arabidopsis thaliana*	EMBL: BT002775	-	Yamada *et al.,* (unpub)
AraIII	Arabidopsis	*Arabidopsis thaliana*	EMBL: AAM64985	-	Haas *et al.,* (unpub)
Avo	Avocado	*Persea americana*	PIR: JH0269	-	Kimura *et al.,* 1995
Bar	Barley	*Hordeum vulgare*	EMBL:Y12068	0.02nM	Gaddour *et al.,* 2001
Bea	Bean	*Phaseolus vulgaris*	-	-	Santino *et. al.,* 1998
Bit	Bitter dock	*Rumex obtusifolius*	EMBL:AJ428415	-	Tinney et. al. (unpub.)
Broc	Broccoli	*Brassica oleracea*	EMBL:AY065838	-	Watson and Coupe 2001(unpub.)
CabI	Chinese cabbage	*Brassica rapa*	EMBL:L41355	-	Lim *et al.,* 1996
CabII	Chinese cabbage	*Brassica rapa*	EMBL:L42819	-	Kim and Chung 2000 (unpub.)
Car	Carnation (clove pink)	*Dianthus caryophyllus*	EMBL: AY028994	-	Sugawara *et al.,* (unpub.)
Carr	Carrot	*Daucus carrota*	PIR: T14323	-	Ojima *et al.,* 1997
Cass	Cassava	*Manihot esculenta*	EMBL:AF265551	-	Reilly *et al.,* (unpub.)
Cast	Castor	*Ricinus communis*	EMBL:Z49697	-	Szederkenyi and Schobert (unpub.)
Cau	Cauliflower	*Brassica oleracea*	TrEMBL:Q8VYX5	-	Watson and Coupe 2001(unpub.)
Chel	Celandine (Chelidocystatin)	*Chelidonium majus*	-	0.00011nM	Rogel *et al.,* 1998
ChesI	European chestnut (CsC)	*Castanea sativa*	EMBL: AJ224331	29nM	Pernas *et al.,* 1998
ChesII	American chestnut	*Castanea dentate*	EMBL:AF480168	-	Connors *et al.,* (unpub.)
Chrb	Christmas bells	*Sandersonia aurantiaca*	EMBL:AF469485		Eason 2002 (unpub.)
Cock	Cockscomb (Celosiacystatin)	*Celosia cristata*	EMBL:AJ535712		Gholizadeh *et al.,* 2005
CornI	Corn I (Maize)	*Zea mays*	EMBL:D10622	0.083nM	Abe *et al.,* 1992
CornII	Corn II (Maize)	*Zea mays*	EMBL:D38130	-	Abe *et al.,* 1995
Cow	Cowpea	*Vigna unguiculata*	EMBL:Z21954	-	Fernandes *et al.,* 1993
Cuc	Cucumber	*Cucumis sativus*	-	-	Yamakawa *et al.,* (unpub.)
Faba	Faba bean	*Vicia faba*	EMBL:AY237958	-	
Job	Job's tears	*Coix lacryma-jobi*	-	190nM	Yoza *et al.,* 2002
Kid	Kidney bean	*Phaseolus vulgaris L.*	-	0.08nM	Brzin *et al.,* 1998
Kiwi-I	Kiwi fruit	*Actinidia deliciosa*	GB:AY390353	0.16nM	Rassam and Laing 2004
Kiwi-II	Kiwi fruit	*Actinidia deliciosa*	GB:AY390354	-	Rassam and Laing 2004
Mugb	Mugbean	*Vigna radiata*	-	-	Kang *et al.,* (unpub.)
Mugw	Mugwort	*Artemisia vulgaris*	EMBL:AF143677	-	Hubinger *et al.,* 1999
Mus	Mustard	*Brassica campestris*	PIR:S65071	-	Lim *et al.,* 1996
RiceI	Rice (Oryzacystatin I)	*Oryza sativa*	EMBL:J03469	30nM	Abe *et al.,* 1987, Kondo *et al.,* 1990

[1]Entries without database accession number were obtained from the referred publication.
[2]Years on unpublished references indicate date sequences were deposited in the database.

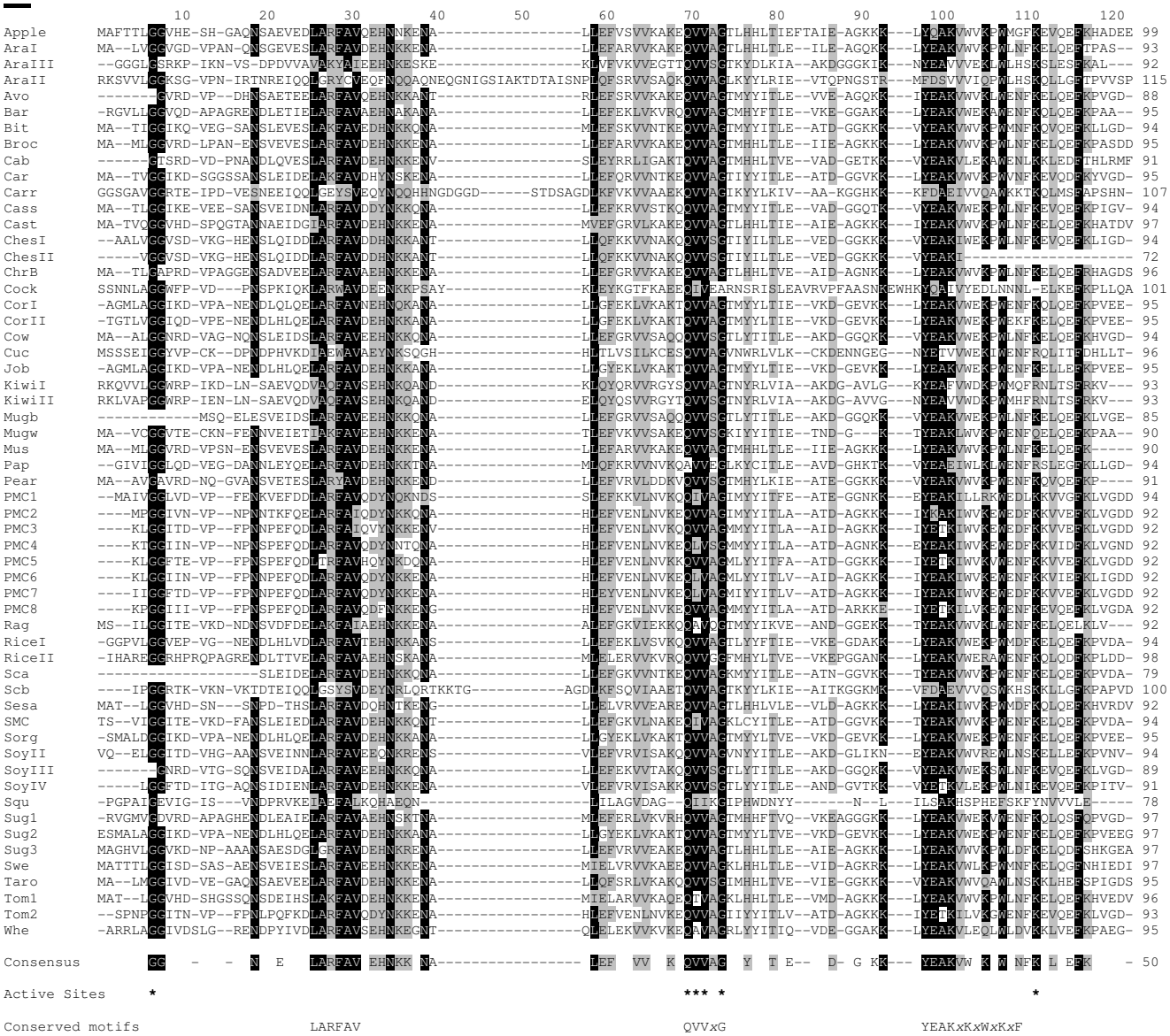
Table 4.1 continued

Code	Common name	Specie name	Database[1] (Acc No.)	Papain K_i	Reference[2]
RiceII	Rice (Oryzacystatin II)	*Oryza sativa*	EMBL:J05595	8.3nM	Kondo *et al.,* 1990
Pap	Papaya	*Carica papaya*	EMBL:X71124	0.75nM	Song *et al.,* 1995
Pear	Pear	*Pyrus communis*	-	-	Gauillard *et al.,* (unpub.)
Pot	Potato	*Solanum tuberosum*	PIR:PQ0469	-	Hildmann *et al.,* 1992.
PMC1	Potato multicystatin	*Solanum tuberosum*	-	-	Michaud, D. (Per. Comm.)
PMC2	Potato multicystatin	*Solanum tuberosum*	-	-	Michaud, D. (Per. Comm.)
PMC3	Potato multicystatin	*Solanum tuberosum*	-	-	Michaud, D. (Per. Comm.)
PMC4	Potato multicystatin	*Solanum tuberosum*	-	-	Michaud, D. (Per. Comm.)
PMC5	Potato multicystatin	*Solanum tuberosum*	-	-	Michaud, D. (Per. Comm.)
PMC6	Potato multicystatin	*Solanum tuberosum*	-	-	Michaud, D. (Per. Comm.)
PMC7	Potato multicystatin	*Solanum tuberosum*	-	-	Michaud, D. (Per. Comm.)
PMC8	Potato multicystatin	*Solanum tuberosum*	-	-	Michaud, D. (Per. Comm.)
PMC10-4	Potato multicystatin (10-4)	*Solanum tuberosum*	GB:AAB29661	0.5nM	Walsh *et al.,* 1993
PMC32	Potato multicystatin (32)	*Solanum tuberosum*	SPROT:P37842	0.7nM	Walsh *et al.,* 1993 Waldron *et al.,* 1993
Rag	Ragweed	*Ambrosia artemisiifolia*	PIR:JN0906	-	Rogers *et al.,* 1993
Sesa	Sesame	*Sesamum indicum*	-	-	Tai *et al.,* (unpub.)
Sorg	Sorghum	*Sorghum bicolor*	EMBL:X87168	-	Li *et al.,* 1996
SoyI	Soyabean	*Glycine max*	PIR:S10588	-	Brzin *et al.,* 1990
SoyII	Soyabean (N2)	*Glycine max*	EMBL:U51855	57nM	Zhao *et al.,* (unpub.); Botella et al.(unpu)
SoyII	Soyabean (L1)	*Glycine max*	-	19,000nM	Zhao *et al.,* 1996
SoyIV	Soyabean (R1)	*Glycine max*	-	21nM	Zhao *et al.,* 1996
Squ	Squash	*Cucurbita maxima*	-	-	Farley *et a.,* 1998
Sug1	Sugarcane	*Saccharum officinarum*	NCBI:AAM78598	-	Soares-Costa *et al.,* 2002
Sug2	Sugarcane	*Saccharum officinarum*	-	-	Reis and Margis 2001
Sug3	Sugarcane	*Saccharum officinarum*	-	-	Reis and Margis 2001
SMC-I	Sunflower (Sca)	*Helianthus annuus*	PIR:JC4791	0.005nM	Kouzuma *et al.,* 1996
SMC-II	Sunflower (Scb)	*Helianthus annuus*	PIR:JC4792	0.00017nM	Kouzuma *et al.,* 1996
SMC-III	Sunflower multicystatin	*Helianthus annuus*	PIR:JC7333	0.04nM	Kouzuma *et al.,* 2000
Swe	Sweet potato (Batate)	*Ipomoea batatas*	EMBL:AF117334	-	To *et al.,* 1999; Huang *et al.,* 2001
Taro	Taro (Cocoyam)	*Colocasia esculenta*	EMBL:AF525880	-	Yang *et al.,* (unpub.)
Tom1	Tomato	*Solanum lycopersicum*	PIR:A59155	4.7nM	Jacinto *et al.,* 1998
Tom2	Tomato	*Solanum lycopersicum*	EMBL AF198388	-	Girard and Michaud 1999 (unpub.)
Whe	Wheat	*Triticum aestivum*	EMBL:AB038393	-	Kuroda *et al.,* 2001
Wist	Wisteria	*Wisteria floribunda*	PIR:PX0039	-	Hirashiki *et al.,* 1990

[1]Entries without database accession number were obtained from the referred publication.
[2]Years on unpublished references indicate date sequences were deposited in the database

The multiple sequence analysis of the phytocystatins showed high levels of sequence homology and conservation especially around the regions involved in function and important structural features (Figure 3.1). The conserved glycine residue in the N-terminal region, known to be characteristic to this group of proteins and involved in N-terminal binding, was as expected present in all but three phytocystatins, mungbean (mugb), potato (pot) and sunflower multi-cystatin domain (sca). However this may have been due to incomplete sequences being deposited on the databases or intentional truncation of the gene by the research groups that provided the sequence. The QxVxG motif characteristic of all members of the cystatin super family and responsible for the second binding site (located in the 2nd hairpin loop) was clearly identified in the multiple alignments (Figure 3.1). Also found was the LARFAV motif in the N-terminal corresponds to the alpha-helix structure and is characteristic to phytocystatins only (Margis *et al*., 1998). A new YEAKxKxWxKxF was identified in the C-terminal of the phytocystatins. This motif being unique to phytocystatins further adds to their qualification for a separate sub-family. This region is not as highly conserved as in animal cystatins. It has been reposted to constitute the third binding region but with less binding capacity and probably more important in stabilising the complex with proteases. Since this region is characteristic only to phytocystatins, several workers have proposed that this group of proteins may constitute a separate sub-family within the cystatin family. From the multiple alignments, it is also clear that there is a very high correlation of conserved regions to important structural features used either for binding or structural conformity of the protein (Figure 3.1).

A



B

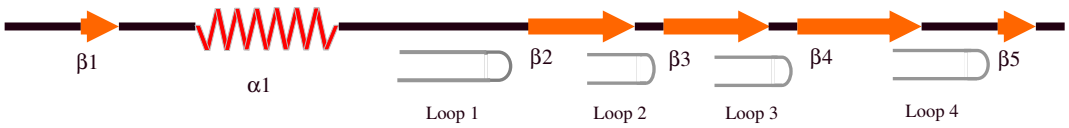


Figure 3.1 (A) Amino acid sequence alignment of known phytocystatins showing residue conservation across the different cystatins studied. A consensus sequence was also generated. Identical amino acids are highlighted in black while similar ones are in grey. (See Table 2.1 for a guide to abbreviated sequence titles). (B) Cartoon of the generalized secondary structural elements of phytocystatins. The orange arrows are the β-sheets (numbered from 1 to 5) and the red spiral representing the single α-helix. The positions were the loops occur are indicated with a gray paper clip mark and labelled 1 to 4.

Based on the neighbour joining phylogenetic tree that was generated, phytocystatins could be separated into five distinctive clades (Figure 2.2). Clades 5 and 6 seemed to be more primitive and may be progenitors of all the other groups of phytocystatins. The biggest clade, clade 1, could further be divided into two sub-clades, sub-clade 1 and 2 with the entire monocot cystatins grouping together in sub-clade 2. Sub-clade 1 included a rather more diverse group of phytocystatins. However, members of this group showed the lowest K_i values for papain (mean 0.37nM –data not shown) rendering them the most potent phytocystatins. They seem to have evolved from the monocot cystatins as deduced from a evolutionary distance tree (data not shown), which show the next lowest K_i values (mean 38.0nM - data not shown). Potency of phytocystatins seems to decrease down the tree with the exception of clade 5 (scb and SMC), which has a mean papain K_i of 0.2nM (data not shown).

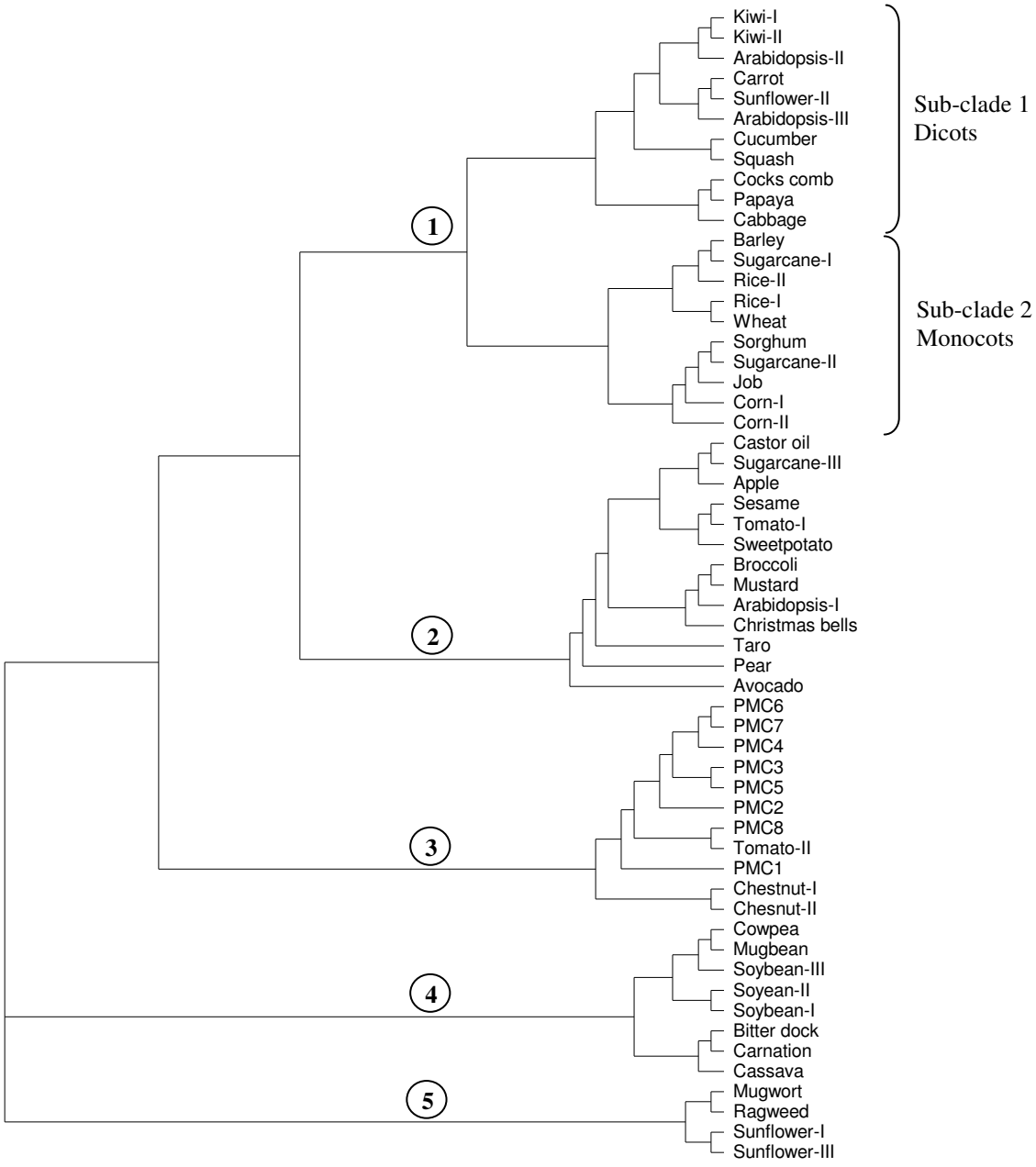


Figure 3.2 Phylogenetic tree for known phytocystatins based on the neighbour-joining method using PROTDIST and NEIGHBOR programs available in the PHYLIP (Phylogeny Inference Package) Version 3.57. Circled numbers indicate the 7 clades that were obtained.

Table 3.2 Percentage identity matrix of phytocystatins (codes detailed in Table 3.1). Identity percentages >50 are highlighted in grey.

	Code	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49
1	Apple	68	32	23	57	53	57	70	42	56	26	53	70	50	70	30	51	52	56	29	50	38	49	52	69	37	59	38	44	46	46	55	45	51	31	62	49	46	52	52	17	47	49	70	65	59	64	47	44
2	Aral	---	31	28	67	59	60	87	53	61	27	61	70	58	76	36	59	60	66	31	60	37	58	56	86	44	64	39	44	45	50	55	49	57	31	64	60	53	60	58	18	56	59	70	64	67	70	47	49
3	Arall	---	---	25	31	34	38	33	33	31	31	33	31	37	34	21	39	39	39	32	39	43	38	36	34	36	38	24	28	30	40	39	34	32	37	32	36	33	39	37	16	32	36	35	31	31	30	27	36
4	Arall	---	---	---	29	27	25	27	20	24	42	30	26	26	25	16	30	29	30	17	29	28	26	25	27	25	24	23	21	22	18	28	25	24	39	25	27	30	23	25	7	22	27	28	27	28	25	21	23
5	Avo	---	---	---	---	59	62	69	51	59	26	61	61	59	65	36	60	61	64	29	60	38	61	54	68	53	59	49	57	57	55	57	53	66	33	55	61	55	66	56	16	58	60	59	55	60	58	57	51
6	Bar	---	---	---	---	---	51	61	47	49	28	53	53	53	57	29	72	70	55	32	69	34	49	51	59	47	55	44	42	44	49	69	71	54	29	50	68	45	53	46	18	74	66	51	49	56	50	43	56
7	Bit	---	---	---	---	---	---	63	48	75	28	73	56	63	60	29	57	54	68	32	55	37	60	57	63	58	61	49	59	55	59	52	46	64	33	53	55	54	67	56	17	48	55	55	58	57	59	58	44
8	Broc	---	---	---	---	---	---	---	53	61	26	63	73	58	79	35	59	59	66	30	60	38	57	58	93	46	67	43	49	51	52	55	53	59	34	66	60	54	59	57	18	60	60	72	69	68	72	49	49
9	Cab	---	---	---	---	---	---	---	---	44	20	47	47	47	54	30	55	55	51	31	57	31	45	39	54	44	51	39	38	40	43	47	46	44	30	47	58	44	47	52	15	50	57	47	47	56	47	39	40
10	Car	---	---	---	---	---	---	---	---	---	28	72	59	61	61	28	57	57	66	27	57	37	57	55	61	52	58	49	53	52	56	51	46	65	31	55	57	56	63	57	14	51	56	61	56	52	63	57	44
11	Carr	---	---	---	---	---	---	---	---	---	---	27	25	27	25	16	31	31	27	19	28	25	24	31	26	24	30	29	27	27	25	26	25	27	51	26	27	29	25	26	13	27	27	27	25	26	22	28	27
12	Cass	---	---	---	---	---	---	---	---	---	---	---	55	66	58	34	57	58	72	30	58	35	66	56	63	53	58	48	52	52	57	56	46	62	34	50	58	62	69	65	14	47	58	57	57	58	56	52	46
13	Cast	---	---	---	---	---	---	---	---	---	---	---	---	53	72	30	51	53	62	30	52	36	55	57	71	47	61	40	46	48	48	50	48	54	33	61	51	49	55	54	16	52	51	72	65	60	70	49	45
14	Ches-I	---	---	---	---	---	---	---	---	---	---	---	---	---	55	31	60	62	67	28	60	36	61	52	59	57	54	53	48	51	50	58	47	55	31	53	62	49	69	60	16	48	61	53	53	57	54	55	47
15	ChrB	---	---	---	---	---	---	---	---	---	---	---	---	---	---	30	58	57	68	31	58	40	57	53	77	45	59	40	47	48	56	52	52	55	29	61	59	51	59	56	16	59	59	70	67	68	66	49	45
16	Cock	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	32	33	34	29	36	25	30	29	35	32	26	24	28	25	29	29	29	31	21	26	36	32	32	33	10	24	37	29	32	31	32	29	33
17	Corn-I	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	89	59	33	94	39	53	52	59	48	52	44	48	46	52	69	59	57	33	51	89	49	57	51	16	62	89	52	52	55	51	47	55
18	Corn-II	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	59	33	92	40	54	52	60	51	52	44	49	47	53	73	56	59	34	49	91	50	58	53	17	60	89	53	53	55	52	47	54
19	Cow	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	32	59	41	81	57	66	55	59	44	48	49	60	60	48	63	32	57	59	63	83	67	15	53	59	61	62	59	66	49	46
20	Cuc	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	34	37	27	31	31	30	29	22	30	32	28	31	33	26	20	31	35	31	28	27	18	32	35	29	34	28	30	35	28
21	Job	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	41	53	52	60	49	51	42	48	46	52	71	57	58	34	49	94	51	57	52	16	59	93	53	53	56	53	46	56
22	Kiwi-I	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	37	35	40	36	35	23	28	31	35	41	31	35	24	36	40	36	36	33	17	34	40	38	37	32	35	27	30
23	Mugb	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	53	58	49	53	41	44	44	54	54	44	68	30	48	54	56	80	62	13	47	53	54	53	50	55	46	42
24	Mugw	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	57	46	54	38	42	44	58	52	41	56	31	46	51	53	54	56	16	41	51	49	54	47	53	48	44
25	Mus	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	46	68	42	48	50	55	56	51	59	32	65	60	55	60	57	18	57	59	72	69	67	71	49	47
26	Pap	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	43	47	48	49	45	45	45	49	30	40	48	46	54	47	15	45	47	46	45	49	44	47	36
27	Pear	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	43	44	46	53	49	47	55	35	55	49	49	55	56	17	51	49	59	51	59	53	46	40
28	PMC-1	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	61	62	45	46	41	46	30	36	44	45	49	40	17	42	43	39	36	42	34	59	47
29	PMC-2	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	83	51	44	37	52	31	45	48	45	51	43	18	42	47	44	40	45	41	77	42
30	PMC-3	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	52	46	36	48	30	47	46	52	52	48	16	41	45	47	43	46	43	77	43
31	Rag	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	50	43	57	31	44	55	54	57	54	16	43	55	47	44	48	48	50	41
32	Rice-I	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	57	56	29	51	69	51	57	52	17	57	68	54	47	47	48	47	59
33	Rice-I	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	47	27	47	60	40	46	41	17	66	59	44	49	51	48	38	50
34	SF-I	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	32	47	59	52	64	55	13	48	58	53	48	49	53	52	45
35	SF-II	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	26	32	36	32	33	8	24	32	31	28	35	27	26	31
36	Sesa	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	49	44	48	47	18	47	49	64	65	50	73	46	42
37	Sorg	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	50	58	53	15	61	97	52	53	56	53	47	53
38	Soy-I	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	60	65	13	40	49	52	51	52	52	48	44
39	Soy-II	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	67	14	51	57	54	53	55	57	52	47
40	Soy-III	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	14	41	52	53	53	53	54	52	46
41	Squ	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	16	14	15	14	15	16	17	17
42	Sug-1	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	60	49	49	54	49	41	46
43	Sug-2	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	52	53	56	53	46	53
44	Sug-3	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	64	61	64	46	44
45	Swe	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	57	76	44	46
46	Taro	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	57	42	48
47	Tom-1	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	45	46
48	Tom-2	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	42
49	Whe	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

Other clades also showed high plant taxa relationships, for example in clade 3 are members of the tomato and potato multi-cystatin, both plants belong to Solanaceae family and both phytocystatins are characterised by multiple domains. Clade 4 includes mostly members of the Fabaceae (Leguminosae) family for example the soybean multi-cystatin domains and cowpea cystatin. This clade seems to be evolutionary primitive. However, one of the domains clustered in clade 3, shows significant difference from its other domain cousins.

In a similarity matrix drawn to compare the sequence similarity of phytocystatins, percentage similarly ranged from 7% to 94% (Table 4.3). Phytocystatins with the least similarities included Arabidopsis-II and III, corks comb, cucumber and squash. High similarities were observed in avocado (Avo), barley (Bar), bitter dick (Bit) and broccoli (Broc). The highest similarity percentage was found between corn-I and job cystatins. This suggests that these are orthologs, genes that have maintained sequence and functional similarity even after speciation. A few more examples were identified in this similarity matrix; Arabidopsis-I (AraI) and broccoli (Broc) with 87% similarity, broccoli and Christmas bells, 79% similarity, corn-I and sorghum with 89% similarity (Table 3.2).

Modelled 3D structures of phytocystatins did show a few variations in the secondary structure elements and their arrangement. OC-I, which is the only phytocystatin whose crystal structure has been determined so far, was the only template structure used for the comparative modelling to determine the 3D models of unknown phytocystatins. Despite the diversity of origin (plant species and tissue types), phytocystatins structures have many structural features in common. The major

differences in the 3D structures were length of the N-terminal trunk, length of the 2nd hairpin loop, length of the 5th β-strand and length of the C-terminal (Figure 3.3). Further, as found from experimental structures of chicken egg white (Rawlings and Barret, 1990; Turk and Bode, 1991) and OC-I, other phytocystatins display the same general structural features i.e. five (in some cases four as in the corn cystatin and sunflower multi-cystatins sca and scb) anti-parallel β-sheets, wrapping halfway round a single central α-helix structure and three hairpin loops (Figure 3.3).

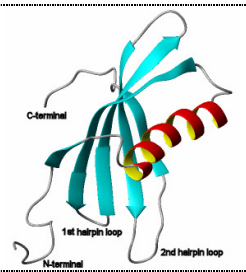
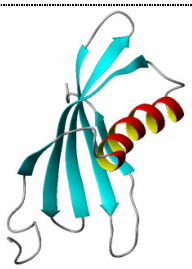
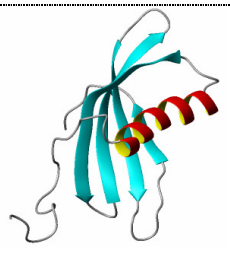
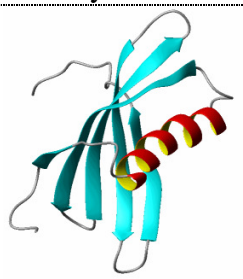
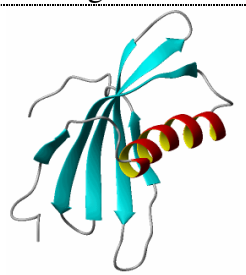
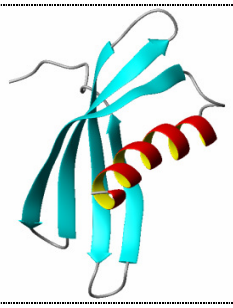
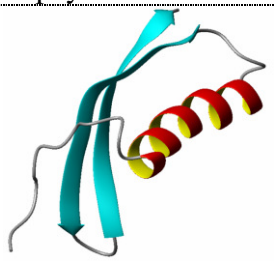
A			
	Oryzacystatin-I	Barley cystatin	Corn cystatin-I
B			
	Cabbage-II	Arabidopsis-I	Avocado
C			
	Bitter Dock	Chestnut	Papaya
D			
	SCA	SCB	PMC10-4

Figure 3.3 Predicted three-dimensional structures of selected phytocystatins representing the major phylogenetic groups (Figure 3.2; (A) group 1; (B) group 2; (C) group 3; (D) groups 4 and 5), and showing the secondary structural elements; five anti-parallel β-strands (blue), one α-helix (red), three hairpin loops, a long N-terminal trunk and a short C-terminal. The figures were made with MOLMOL program (Koradi *et al.*, 1996).

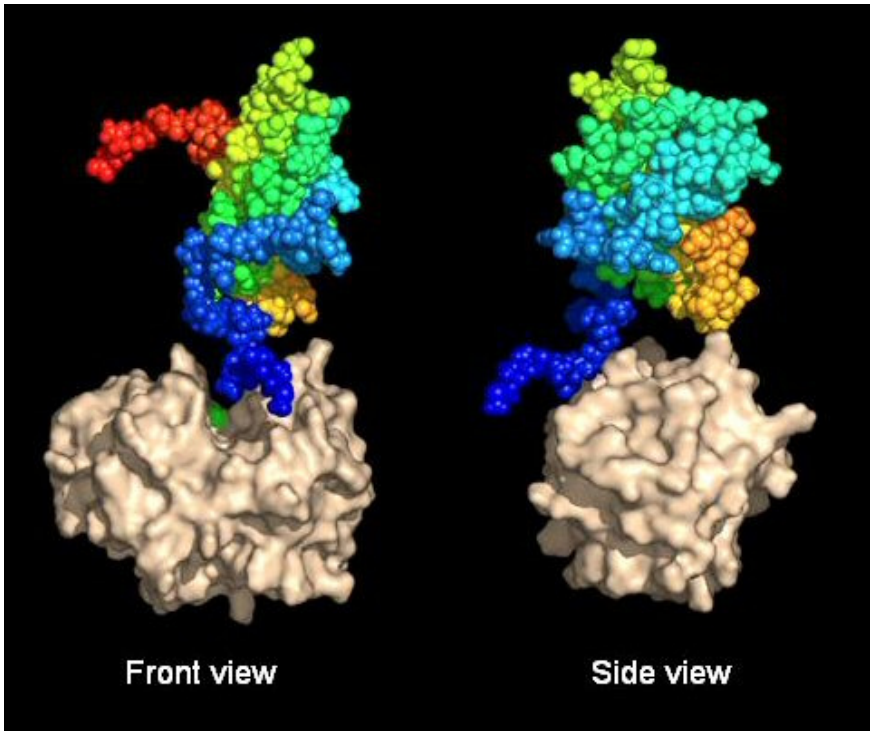


Figure 3.4 Modelled complex between OC-I (top) and papain (bottom) in front and side views. OC-I is shown in spheres and its structure coloured by rainbow. The N-terminal is blue towards the C-terminal red. The surface of papain is coloured light brown. The active site of papain appears as a trench extending from the front to the back of the molecule. The active cysteine residue of papain here coloured green (front view) occurs in the middle of this trench. The complex was initially modelled manually and the model refined using MULTIDOCK program based on minimisation algorithms. Visualisation and rendering graphics were done using PYMOL program.

Figure 3.4 shows the predicted binding of OC-I onto the papain active site cleft. *In-silico* docking experiments involving OC-I and papain revealed that OC-I attaches onto the active site cleft of papain and possibly in the same way with other cysteine

proteases (Figure 4.4). Due to electrostatic forces along these two molecules, an average distance of 1.8Å separates them from each other. During the docking process, one residue in the N-terminal of OC-I, aspartic acid (Asp4) prevented the inhibitor from docking closer into the papain active site.

3.5 Discussion

This study has provided new information about the structure of phytocystatins. Firstly, despite high structural similarity of phytocystatins, there is very wide variation in inhibition potential meaning that biochemical screening could yield selections of cystatins targeting a wide range of pests as well as other uses. It has been shown in this study that the experimentally determined structure of OC-I can effectively and successfully be used to predict structural conformations of unknown cystatins. This improves their analysis and evaluation as demonstrated by Girad *et al.,* (2007). Further, it has been elucidated through *in-silico* prediction of structure that the individual domains of multicystatins (e.g. tomato and potato), when separated, can fold into functional proteins individually. Docking and inhibitor- protease complex prediction was possible for the first time using phytocystatins. From the analyses of binding candicate residues to engineer for improved binding (and thus inhibition capacity) were inferred.

In some plant species, for example *Oryza sativa* (rice), Zea mays (corn) and *Solanum lycopersicum* (tomato), more than one highly homologus phytocystatin has been identified. In other plants, like *Glycine max* (soybean), *Helianthus annuus* (sunflower) and *Solanum tuberosum* (potato), multiple domain cystatins have been identified. It has been shown that when these domains, which occur in tandem, are cleaved by

enzyme digestion, the separated domains can fold into functional proteins retaining their inhibitory activity (Walsh *et al*., 1993). This suggests that these forms of multi-domain cystatins may have arisen as a result of gene duplication events. The potato multicystatin for example has eight domains while the sunflower multi-cystatin has 4 active domains.

Evolutionary relationships among phytocystatins were inferred using an unrooted phylogenetic tree. As expected, most phytocystatins grouped together to reflect the plant taxonomic groups. However, some members of the multidomain cystatins tended to occur in distinctly different groupings. This suggests that plants, such as tomato, soybean, sunflower and sugarcane, contain complete cystatin coding genes that may have distinctly different evolutionary origins.

It is still structurally unclear how the phytocysatins with longer N-terminal trunks are more potent than the shorter forms. From nuclear magnetic resonance (NMR) data of OC-I it is known that the N-terminal trunk is highly flexible and does not form any ordered structure. *In-silico* observations in this study have shown the possible formation of a bond between residues GLY6 and VAL8 forming a loop structure in long enough N-terminals. This probably stabilises the trunk allowing a more precise binding more and rendering the complex more stable. Cystatins bind to proteases with a 1:1 stoichiometry and with varying affinities. However, it is not clear weather this is true for the multi-domain cystatins. The whole phytocystatin molecule is wedge shaped with the N-terminal and the two hairpin loops forming the sharp edge in some cases the N-terminal protrudes out into a long arm extending outwards from the rest of the structure (Figure 3.3) forming what has been referred to as a trunk. This sharp

edge is highly hydrophilic and complimentary to the active cleft of cysteine proteases (Bode *et al*., 1988) forming the active site region. The active site itself is composed of a glycine residue in the N-terminal and this appears to be the most important binding site in many cystatins although its removal or absence does not seem to affect binding by other types of phytocystatins (Arai *et al*., 1991). The sca and scb cystatin domains of the sunflower multi-cystatin do not have N-terminal trunks (Figure 3.3) despite retaining high affinity for cysteine proteases (Kouzuma *et al*., 1996).

OC-I is still the only phytocystatin one whose tertiary structure has been experimentally determined. In this study, bioinformatics tools have been successfully used to predict the inhibitor-protease (OC-I and papain) complex. In general, the binding in the predicted complex was in agreement with that of the experimental complex structures previously reported between stefin-B and papain (Studds *et al*., 1990) and stefin-A with cathepsin-H (Jenko *et al*., 2003). In docking OC-I and papain, it was difficult to dock two residues of aspartic acid ASP4 in the N-terminal trunk and ASP86 in the 2nd binding hairpin loop of the C-terminal region. These two residues are close to the active sites and seemed to prevent closer binding of the active sited to the target papain. Therefore, these sites appear to be potential targets for site-directed mutagenesis directed at improving binding and therefore potency of OC-I to papain and probably to other cysteine proteases. These might be replaced by either asparagine (ASN) or glutamic acid (GLU) based on the Doolittle amino acid substitution matrix (Mark *et al*., 1993). This is one of the many database derived matrices showing evolutionally substitution of amino acids in many similar proteins. Through such a matrix amino acid substitutions can be done through site-directed

mutagenesis, to maintain the overall structure as much as possible but vary small parameters like bond distance and eventually levels of potency.

The wide variation in affinities found for phytocystatins in this study, and indeed also in other animal cystatins, is not explainable by a simple structural difference. For example, an inhibitor with highly similar structural features has a great difference in affinity. Nikawa *et al.,* (1989) reported that the 1st binding hairpin loop with the QVVAG highly conserved motif was not essential for cysteine protease inhibitory activity in cystatin-A. This study identified that PMC10-4, a potato multi-cystatin domain, retains high affinity despite having only the N-terminal and the 1st loop. Sca and scb retain high affinity (mean K_i is 0.003nM) despite not having an N-terminal. This means that the two rely on the 1st and 2nd hairpin loops for their activity. Therefore, it is possible that such functional differences may be explainable by small structural features at residue level that result in great differences in affinity of the inhibitor.

CHAPTER 4

Engineering of a papaya cystatin using site-directed mutagenesis to improve its activity against papain and weevil digestive cysteine proteases

Scientific Communication

Parts of this cahpter led to the publication:

Kiggundu, A., Goulet MC., Goulet C., Dubuc JF., Rivard D., Benchabane M., Pépin G., Van der Vyver C., Kunert K. and Michaud D. (2006). Modulating the protease inhibitory profile of a plant cystatin by single mutations at positively selected amino acid sites. The Plant Journal, 48, 403–413.

4.1 Abstract

The usefulness of native phytocystatins for pest control is limited by the co-evolution between the pest and host-plant. This has allowed insects to develop ways of overcoming the presence of inhibitors in plant tissues. This includes the production of insensitive proteases in the variable gut environment helping insects to elude the anti-nutritive effects of cystatins. Protein engineering was employed in this part of the study to attempt to produce variants of a papya cystatin with improved activity against a model protease papain and also against gut proteases of banana weevil and the black maize beetle *Heteronychus arator* Fabricius (Coleoptera: Scarabaeidae). Specific amino acids in the amino acid sequence of the papaya cystatin were changed using site-directed mutagenesis. An evolutionary and structural analysis strategy was applied to improve cystatin activity against cysteine proteases. The papaya cystatin was amendable to improvement and papaya cystatin mutants showed 1.5- to 6-fold improved inhibition of papain. Amino acid changes close to conserved regions of the protein provided the most improved inhibition against cysteine proteases. Improvement was not as high as for papain when banana weevil and black maize beetle gut extracts were tested. Improvements ranged from 1.5- to 2-fold in the mutant E52Q with a change from glutamic acid to glutamine. Novel cystatin mutants with increased inhibitory activity represent a first step in setting up a library of mutated phytocystatins with improved inhibition against both endogenous cysteine proteases and proteases derived from plant pests.

4.2 Introduction

The usefulness of native protease inhibitors for peat control is limited due to the fact that over evolutionary time insects have developed ways of overcoming the presence of inhibitors in plant tissues. This is mainly due to the production of insensitive proteases in insects and the variable gut environment which helps insects to elude the anti-nutritive effects of phytocystatins. Engineering of phytocystatins by changing amino acids in the amino acid sequence for better binding to proteases is one strategy to possible improve the efficiency of cystatins. In a first approach, using cystatin engineering, better protection against a plant pest has been found with transgenic plants expressing an engineered OC-I (Urwin *et al*., 1997, Irie *et al*., 1996). Site-directed mutagenesis is applied as a tool to alter the amino acid sequence through the replacement of single or several nucleotide bases to alter amino acid sequence of the respective protein.

There are mainly two general strategies for protein engineering (i) rationale design and (ii) directed evolution. In rationale design, detailed knowledge of the structure and function of the target protein is used to make changes in the sequences that through site-directed mutagenesis leads to the desired modulation of function and properties (Carter, 1986; Young and Dong, 2003). The second method known as directed evolution mimics natural evolution. This method is performed by application of random mutagenesis to a protein followed by a high throughput selection to identify variants that have the desired qualities. This method has been shown to successfully produce improved proteins. However, it requires large amounts of recombinant DNA which has to be mutated. Also, the products screened often require expensive robotic equipment for automated selection assays.

However, for the successful application of the rationale design approach, additional evolutionary analysis has to be carried out. This includes positive site selection at the amino acid level as a guide to mutagenesis. A number of studies have shown that proteins involved in host defence responses are subject to adaptive evolution. This results from direct selection pressure on amino acid resides that directly interacts with target molecules of invading or predatory organisms (Barbour *et al.,* 2002; Bishop *et al*., 2000; Sawyer *et al.,* 2005). Most genetic variation detected at the molecular level is assumed to result from randomly generated mutations so that mutations that confer a selective advantage to the host are maintained in evolutionary time (Yang and Bielawski, 2000). At the gene level, this process of positive selection (maintenance of mutations that confer advantage) can be detected by comparing the rate of non-synonymous codon substitutions (dN), where the original amino acid is substituted for an alternative residue, and the rate of synonymous substitutions (dS), where the original amino acids are preserved. In practice, the ratio of dN to dS, referred to as ω, is considered to be a reliable measure of the directional selection exerted on the protein (Yang, 2005). For amino acid sites with little or no impact on the activity of the protein, the ω ratio will be close to 1 as nonsynonymous mutations will be fixed at the same rate as synonymous mutations by neutral selection. Conserved amino acid sites, where any amino acid substitution would strongly compromise biological activity, will typically show a ω ratio close to 0 as a result of negative (or purifying) selection. In contrast, amino acid substitutions giving the organism a selective advantage will tend to be readily fixed in the population, resulting in calculated ω values greater than 1 for the corresponding amino acid site. Statistical methods based on maximum-likelihood models have been developed to detect positive selection by

the estimation of ω values (Yang and Bielawski, 2000). These methods allow the identification of specific codon and amino acid sites subject to positive selection (Bielawski *et al.,* 2004; Ivarsson *et al.*, 2003; Sawyer *et al.*, 2005; Yang *et al.*, 2000).

The objective of this part of the study was therefore to use an evolutionary guided rationale for engineering of a papaya cystatin for improved inhibition of a cysteine protease. Maximum-likelihood models were used to detect amino acid sites in Poaceae (monocots; seven species) and Solanaceae (dicots; potato and tomato) that have, over evolutionary time, been subjected to positive selection. Possible sites for mutations were selected that can improve the activity or inhibitory profile of phytocystatins to asses if actually positive selection has occurred in phytocystatins.

4.3 Materials and methods

4.3.1 Phylogenetic and structural model analysis

Phylogenetic analysis as well as the protein structural modelling analysis that was used to predict potential mutatable sites has been outlined in Chapter 3 of this thesis.

4.3.2 Detection of positive selection sites in PhyCys

Positive selection sites for phytocystatin genes were detected using maximum likelihood models M0, M1, M2, M3, M7, M8, R1 and R2, which are present in the software package Phylogenetic Analysis Maximum Likelihood (PAML) version 3.14 (http://abacus.gene.ucl.ac.uk/software/paml.html) (Yang, 1997). PAML includes a suite of codon-based models that can be used to estimate x, the ratio of the rate of non-synonymous substitutions per non-synonymous codon site (dN) to the rate of

synonymous substitutions per synonymous site (dS) as well as calculation of posterior Bayesian probabilities needed to identify positively selected sites in genes.

4.3.3 Construction of over-expression vector for papaya cystatin (PC)

The PC coding sequence was excised from the cloning vector p*BlCYS1* using the restriction enzymes *EcoR*I and *Pst*I. The *EcoR*I/*Pst*I fragment was then first cloned into the *EcoR*1/*Pst*I site of p*BlueScript* (Stratagene, USA) and then as a *BamH*I/*Kpn*I fragment from p*BlueScript* into the vector p*QE31* to achieve in-frame expression of a 6Xhis-tagged protein. This sub-cloning procedure created the plasmid p*QE31PC-I* (Figure 4.1). This plasmid was transformed into *E. coli* cells (strain JM109) for storage and into *E. coli* strain M15 for expression according to the QIAexpressionist kit user's manual (Qiagen, Germany). Site-directed mutagenesis to engineer PC was done directly in the expression vector p*QE30XaCYS*.

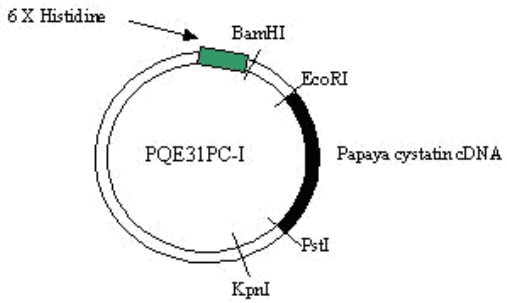


Figure 4.1 Schematic representation of recombinant protein expression vector *pQE31PC-I* created to express PC. In this vector site-directed mutagenesis was also performed.

4.3.5 Mutagenesis primer design

For each site to be mutated, two mutagenic oligonucleotide primers were designed. These primers contained the desired mutation with at least 10 to 13 bases flanking the

mutation were exactly complimentary to the template DNA. This was achieved by using PrimerX (http://bioinformatics.org/primerx/), a web-based program developed to automate the design of mutagenic PCR primers for application in site-directed mutagenesis. Based on input (DNA or amino acid sequence), the program compares a template sequence that already incorporates the desired mutation. It then generates several forward and reverse primer sequences by encoding the mutation and finally computes for other necessary primer information like melting temperature and GC content for each primer pair. The primers, which were used in this study, are outlined in Table 4.1.

Table 4.1 Sequence information of the mutagenic primer pairs used for the mutations. Mismatched bases are underlined. *Mutation 16 was an N-terminal truncation to remove seven amino acids. The primer pairs were created to cut out 21 bases and include part of the vector backbone.

Mutation number	**Mutant Code**	**Primer sequence** Forward and Reverse
1	CYSP03F	5′GAGGGAAGGATGGAGTTCGGAATTGTGATC 3′ 5′CCGATCACAATTCCGAACTCCATCCTTCCC 3′
2	CYSP03S	5′GGGAAGGATGGAGTCCGGAATTGTGATC 3′ 5′GATCACAATTCCGGACTCCATCCTTCCC 3′
3	CYSV06R	5′GAGCCCGGAATTCGGATCGGTGGTTTG 3′ 5′CAAACCACCGATCCGAATTCCGGGCTC 3′
4	CYSI07L	5' CCCGGAATTGTGCTCGGTGGTTTGC 3' 5' GCAAACCACCGAGCACAATTCCGGG 3'
5	CYSI07A	5' CCGGAATTGTGGCAGGTGGTTTGCAG 3' 5' CTGCAAACCACCTGCCACAATTCCGG 3'
6	CYS07V	5' CCCGGAATTGTGGTCGGTGGTTTGC 3' 5' GCAAACCACCGACCACAATTCCGGG 3'
7	CYSI07D	5′CCCGGAATTGTGGACGGTGGTTTGC 3′ 5′GCAAACCACCGTCCACAATTCCGGG 3′
8	CYSA32V	5′CGCCGTCGATGTGCCACAACAAAG 3′ 5′CTTTGTTGTGGCACATCGACGGCG 3′
9	CYSA52P	5' GTGAATGTAAAGCAGCCAGTGGTTGAAGGC 3' 5' GCCTTCAACCACTGGCTGCTTTACATTCAC 3'
10	CYSA52Q	5′GAATGTAAAGCAGCCAGTGGTTGAAGGC 3′ 5′GCCTTCAACCACTGGCTGCTTTACATTC 3′
11	CYSE55A	5′CAGGCAGTGGTTGCAGGCTTAAAGTAC 3′ 5′GTACTTTAAGCCTGCAACCACTGCCTG 3′
12	CYSC60T	5' GTTGAAGGCTTAAAGTACACCATCACTTTGGAGGCTG 3' 5' CAGCCTCCAAAGTGATGGTGTACTTTAAGCCTTCAAC 3'
13	CYSI78V	5′GTATATGAGGCCGAGGTCTGGGTGAAGCTC 3′ 5′GAGCTTCACCCAGACCTCGGCCTCATATAC 3′
14	CYSW79P	5′GAGGCCGAGATCCCGGTGAAGCTCTGG 3′ 5′CCAGAGCTTCACCGGGATCTCGGCCTC 3′
15	CYSE84A	5' GTGAAGCTCTGGGCGAATTTCAGGAGC 3' 5' GCTCCTGAAATTCGCCCAGAGCTTCAC 3'
16	CYSN85X	5' GAAGCTCTGGGAGTTCAGGAGCTTG 3' 5' CAAGCTCCTGAACTCCCAGAGCTTC 3'
15	CYSR87C	5′CTGGGAGAATTTCTGCAGCTTGGAGGGATT 3′ 5′GAATCCCTCCAAGCTGCAGAAATTCTCCCA 3′
16*	CYStNT	5′CTGGTATCGAGGGAAGGATGGGTTTGCAGG 3′ 5′CCCTCGACGTCCTGCAAACCCATCCTTCCC 3′

4.3.6 Site-directed mutagenesis

Site-directed mutagenesis was done following a modified *Quick Change* mutagenesis method (Stratagene, USA), which was an effective and simple method with which mutations can be carried out inside expression vectors (Fisher and Pei, 1997).

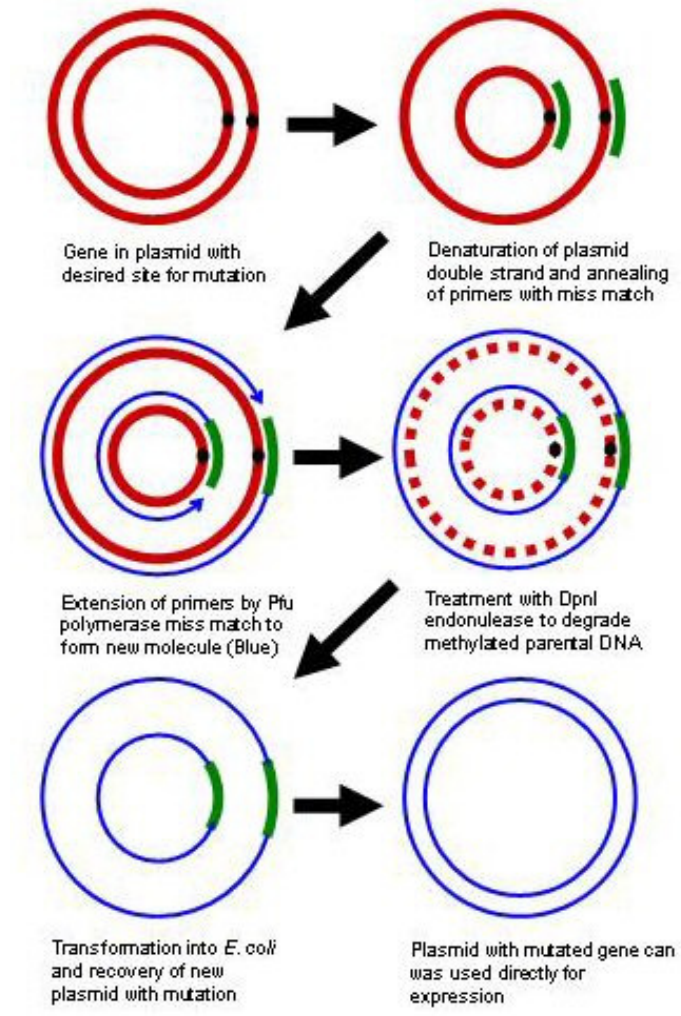


Figure 4.2 Schematic representation of the site-directed mutagenesis protocol used (modified from QuickChange® Site-Directed Mutagenesis Kit Manual #200518, Stratagene, USA)

The whole method is divided into three stages; amplification of mutant DNA, degradation of parental DNA (methylated) and transformation into *E. coli* cells. Briefly, primers containing miss-matched nucleotides at site of intended mutation

resulting in the desired modification anneal to complementary opposite strands of the template plasmid DNA. These are then extended in a PCR using *Pfu* DNA polymerase. The PCR reaction is then treated with *Dpn*I endonuclease enzyme, which specifically targets methylated DNA, in this case the parental plasmid DNA. The new double-stranded DNA containing the desired mutations is then transformed into *E. coli* competent cells and stored until needed for further use.

In this particular study, the amplification step was modified into a two-stage PCR as described by Wang and Malcom (1999). Two separate primer extention reactions one for each primer were set up as follows; 10-15ng template plasmid (5µl), 10X Pfu buffer (5µl), 25µM primer-1 (1µl), 10mM dNTPs (1µl), 2.5 units *Pfu* DNA polymerase (Fermentas #EP0571) (1µl) and sterile distilled water up to 50µl total reaction volume. Exactly the same reaction was set up for the second primer. The PCR conditions setup on an automated cycler (Palm Cycler, Corbett Life Science, Australia) were denaturation at 94°C for 30sec, 4 cycles of; 95°C 30sec, 55°C 1min, 68°C 7min (2 minutes/kb of plasmid length and p*QE31PC-1* is 3500bp). The reaction was held at 4°C on completion of cycling.

In the second PCR stage, 25µl from each of the separate reactions above were combined in a new tube, 1µl *Pfu* polymerase added, mixed and incubated as above except that the cycles were increased to 18. To degrade parental DNA, 10 units of *DpnI* enzyme were added to the cooled reaction, mixed well and incubated at 37°C for 1hr.

Finally, the reaction containing mutant DNA was used for transformation in competent *E. coli* cells. This was done by placing 200µl competent JM109 cells in a Falcon tube on ice plus 2µl of the digested PCR reaction and incubated on ice for 20min. Heat shock was performed by placing the cell/DNA mix at 42°C for 60sec and then returned on ice for 2min. LB (500µl) broth was added and then the cells were incubated at 37°C with shaking at 200rpm for 1hr after which 100µl of this culture was plated on solid LB containing 100ml/l ampicillin and incubated overnight 37°C. Three individual colonies were picked and inoculated into LB broth (50ml) containing 100ml/l ampicillin and again incubated at 37°C overnight with shaking at 180rpm. Minipreps were made and pure plasmid DNA sent for sequencing.

4.3.7 Protein expression

All the mutants were expressed directly in the p*QE31P-1* vector in which the mutations were done using the QIAexpressionist kit (Qiagen, Germany) as described in the manufactures manual and also described in Chapter 2 (Section 2.3.8) of this thesis. Briefly, LB medium (5ml) with antibiotics (50mg/l kanamycin and 100mg/l ampicillin) was inoculated with a single bacterial colony of *E. coli* (strain M15) cells containing p*QE31PC-1* mutants and grown overnight at 37°C with shaking at 200rpm. Pre-warmed LB medium (100ml) with antibiotics (as above) in a 250ml conical flask was inoculated with 5ml of the overnight culture and incubated at 37°C with shaking as above until the optical density at 600nm (OD600) reached 0.6. Isopropypyl-β-D-thiogalactopyranoside (IPTG) was then added to a final concentration of 1mM to induce expression and incubation continued for another 4hrs. Bacterial cells were harvested by centrifugation (13000rpm at 4°C) for 10min and stored frozen at -20°C until purification.

4.3.8 Protein purification

Purification was performed under native conditions to preserve the conformational integrity of the protein. Frozen cell pellets were thawed on ice for 30min, re-suspended in his-tag lysis buffer (50mM sodium di-hydrogen phosphate, pH 8.0; 300mM sodium chloride; 10mM imidazole) at a rate of 2ml per 1mg of cells and 1mg lysozyme was added. This was mixed gently and incubated on ice for 1hr. The cell suspension was then sonicated using a sonicator (Cell Disruptor B-30, Branson Sonic Power Co./SmithKline Co.) fitted with a standard micro-tip and set to 20% duty cycle, 2 output control and in pulse mode. The cells were sonicated using 10 bursts with 10sec cooling on ice between each burst, taking care not to create much frothing. The lysates thus obtained were centrifuged at 10,000rpm for 30min at 4^{o}C in a centrifuge and the clear supernatant transferred into fresh Eppendorf tubes to which 800µl of 50% Ni-NTA slurry (Qiagen, Germany) was added. The tubes were shaken at 200rpm for 30min at 4^{o}C after which the cell lysate mixture was poured into a short plastic column (made with a 2.5ml syringe and a glass wool plug at the bottom) with the bottom cover in place. The cover was removed after the slurry settled and the flow-through collected. Two-times 1ml wash buffer (50mM sodium di-hydrogen phosphate, pH 8.0; 300mM sodium chloride; 50mM imidazole) was carefully poured over the column and collected at the bottom. This was followed by pouring slowly 4-times 500µl elution buffer (50mM sodium di-hydrogen phosphate, pH 8.0; 300mM sodium chloride; 250mM imidazole) over the slurry. The elutions were collected separately in 500µl fractions. Five micro-liters of each fraction (flow-through, washes and elution fractions) were each added to 5.0µl SDS-PAGE sample buffer (6% β-mercaptoethanal, 6% SDS, 0.6% bromophenol blue, 20% glycerol) heated to 37 ^{o}C for 10min and loaded onto a 15% polyacrylamide gel for evaluation of the purification

process and detection of the recombinant proteins. The purity of the inhibitors was assessed using 15% (w/v) SDS-PAGE analysis as described in (Sambrook *et al.*, 1989). The protein concentration of the elution fractions was finally determined using the Bio-Rad protein assay kit (Bio-Rad, South Africa), and fractions were stored in aliquots at 4^{o}C until required.

4.3.7 Enzyme kinetics of mutants

Dissociation constants ($K_{i(app)}$) for the interaction and inhibition of a model cysteine protease, papain, by the papaya cystatin variants obtained were determined by the monitoring of substrate hydrolysis progress curves as described by Salvesen and Nagase (1989). Papain activity was measured in 50mM Tris-HCl, pH 6.0 containing 5mM L-cysteine as reducing agent using the synthetic substrate *N*-CBZ-Phe-Arg-7-amido-4-methylcoumarin. Hydrolysis was allowed to proceed at room temperature while monitoring progress on the spectro-fluorometer with excitation and emission filters at 360nm and 450nm, respectively. When the reaction reached a steady state, the inhibitors were added and monitoring continued until a new steady state was reached. The difference in the initial vs final reaction rates was used to compute the apparent K_i value of the inhibitor.

4.4 Results

4.4.1 Rationale of mutations

Table 4.2 below outlines the particular amino acid change. Some of these changes were prompted by literature reports. In particular the truncation of the N-terminal has been reported not being important in some cystatins. However the modelling study showed that it may be important in stabilising the protein at the active site. Figure 4.3

below illustrates that mutations at sites 52 and 55 flanking the major functional motif 'VV' gave the highest improvement in inhibition. An indication that activity differences in phytocystatins could be explained by the sequence variability close to the active sites.

Table 4.2 Mutations performed on native PC, the amino acid changes made and the respective rationale. The mutant code refers to the amino acid changes made, for example CYSC60T refers to a mutation were cysteine at position 60 was replaced with thereonine.

Mutation number	***Mutant code***	***Amino acid change***	***Rationale***
1	CYSP03F	Proline (position 3) to phenylalanine	Mutation in positive selection site in the N-terminal
2	CYSP03S	Proline (position 3) to serine	Mutation in positive selection site in the N-terminal
3	CYSV06R	Valine (position 6) to arginine	Random mutation in a less conserved region close to the N-terminal active site.
4	CYSI07L	Isoleucine (position 7) to leusine	Random mutation in a less conserved region close to the N-terminal active site.
5	CYSI07A	Isoleucine (position 7) to alanine	Random mutation in a less conserved region close to the N-terminal active site.
6	CYSI07V	Isoleucine (position 7) to valine	Random mutation in a less conserved region close to the N-terminal active site. Both isoleuince and valine are aliphatic and hydrophobic.
7	CYSI07D	Isoleucine (position 7) to aspertic acid	Random mutation in a less conserved region close to the N-terminal active site. Isoleuince and aspertic acid have very different chemical properties, however aspertic acid has the smallest side chain, than would less interfere with binding of the N-terminal.
8	CYSA32V	Alanine (position 32) to valine	Random mutation in a less conserved region, both are small and hydrophobic.
9	CYSA52P	Alanine (position 52) to proline	Random mutation in a positively selected site close to the 2nd loop active site. Proline substitution showed increased bond number in the 2nd loop and may improve structural strength.
10	CYSA52Q	Alanine (position 52) to Glutamine	Random mutation in a positively selected site close to the 2nd loop active site. Proline substitution showed increased bond number in the 2nd loop and may improve structural strength.
11	CYSE55A	Glutamic acid (position 55) to alanine	Random mutation in a positively selected site close to 2nd binding site.

Mutation number	*Mutant code*	*Amino acid change*	*Rationale*
12	CYSC60T	Cysteine (position 60) to threonine	Cysteine was found to be a very rare amino acid in phytocystatins so it was changed to threonine also a small and hydrophobic amino acid.
13	CYSI78V	Isoleucine in (position 78) to valine	Random mutation in a positively selected site close to the C-terminal active site.
14	CYSW79P	Tryptophan (position 79) to proline	Random mutation in a positively selected site close to the C-terminal active site.
15	CYSE84A	Glutamic acid (position 84) to alanine	Random mutation in a positively selected site close to the C-terminal active site.
16	CYSN85X	Deletion of asparagine in position 85	Random mutation in a less conserved region close to the C-terminal active site. Asparagine deleted from the sequence.
17	CYSR87C	Arginine (position 87) to cysteine	Random mutation in a less conserved region close to the C-terminal active site. Arginine's long side chain seemed to interfere with C-terminal binding.
18	CYStNT	Trucation of the first 7 amino acids of the N-terminal	Truncation of N-terminal to reduce interference in binding
19	CYSA52QE55A	Combined 9 and 10	Combining two improved mutations

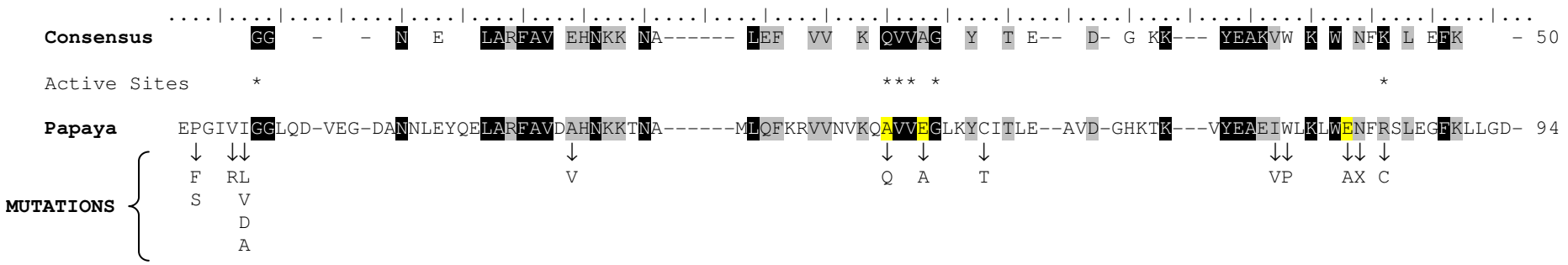


Figure 4.3 Consensus sequence from a multiple alignment (see also Chapter 3). The positions of the three active sites on the sequence are indicated by an asterix. The mutations performed are indicated on the amino acid sequence as in Table 4.1. The mutations that gave the highest improvement are highlighted in yellow.

4.4.2 Positive selection among plant cystatin genes

Maximum-likelihood tests were carried out to predict positive selection among codon sites in a combined dataset of Poaceae and Solanaceae cystatin coding sequences. This was based on the phylogenetic analysis outlined in Chapter 3 of this thesis according to the methods described by Yang *et al*. (2000). Based on codon substitution models M0, M1, M2, M3, M7 and M8 (Yang *et al.*, 2000), 18 codon sites showing Bayesian posterior probabilities greater than 60% were considered to have been subjected to positive section during the evolutionary advancement of these genes (Table 4.3; Figure 4.3). Three models M2, M3 and M8 allowing for positive selection fitted the data significantly better than M0, M1 and/or M7, with $p<0.01$ for all likelihood ratio tests (Table 4.3). Models M3 and M8, which included 5 and 4 parameters respectively, gave a ω value greater than 1 (1.27) for the codons 1, 2, 6, 10, 15, 16, 17, 25, 29, 31, 45, 47, 51, 57, 58, 60, 76, 84 (Table 4.3). When the ratio of the rate of non-synonymous codon substitutions to rate of synonymous substitutions is greater than 1, the substitution at this site has given the organism a selective advantage and is largely fixed in the population. As expected and previously reported with other data sets (Yang *et al.*, 2000), positive selection could not be detected under M2 ($\omega 2<1$), whereas M3 and M8, which are more powerful as they allow for heterogeneous distributions of ω ratios among codon sites (Yang and Bielawski, 2000), gave ω2 (ωfor M8) values greater than 1.

Posterior Bayesian probabilities were calculated to estimate the probability of each individual codon belonging to an alternate codon assuming that the codon is being subject to positive selection. Eight sites, showing posterior probabilities greater than 95%, were thus identified to be highly positively selected. This included codons 1, 2,

6, 10, 45, 47, 76 and 84 (Figure 4.4). Site 2, subjected to several mutations to investigate mutants at this positively selected site, would improve activity and inhibition of proteases *in-vitro*.

Table 4.3 Evidence for positive selection events among codon sites of Poaceae and Solanaceae cystatins (n=21)

Model	*p*[a]	*Ω*	*l*	*Positively selected sites*[b]
M2	***3***	***0.23***	***-2690.5***	
M3	***5***	***1.27***	***-2688.4***	***1, 2, 6, 10, 15, 16, 17, 25, 29, 31, 45, 47, 51, 57, 58, 60, 76, 84***
M8	***4***	***1.27***	***-2688.4***	***1, 2, 6, 10, 15, 16, 17, 25, 29, 31, 45, 47, 51, 57, 58, 60, 76, 84***
R1		***0.42***	***-2762.7***	
R2		***0.34, 0.65***[c]	***-2757.4***	

[a] *p*, number of parameters in the model.
[b] Codon numbering was based on the 2nd codon before the GG conserved motif in the N-terminal as number 1 ring to the last codon in the C-terminal
[c] For Poaceae and Solanaceae respectively

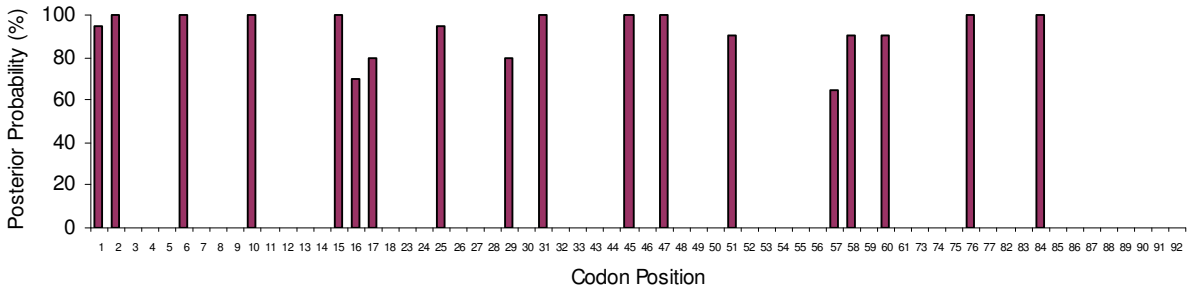


Figure 4.4 Location of positively selected codon sites (with Bayesian posterior probabilities greater than 60% under model M3) in Poaceae and Solanaceae cystatins. Codon 1 corresponds to the second codon before the conserved 'GG' motif. The dashed line indicates a posterior probability of 95%.

4.4.3 Mutation and expression of recombinant mutant papaya cystatins

Mutation success using the modified quick change protocol resulted in about 100 colonies on each plate after transformation of *E. coli*. Usually 2 out of 3 colonies picked for sequencing were positively mutated. Expression and purification under native conditions followed the recommended protocols in the QIAexpressionist kit manual. Figure 4.5 shows a 12% SDS-PAGE with successful expression and purification of mutants CYSI07D, CYSA53P CYSA32V and CYSW78P. Mutant proteins were highly pure resulting in a sinle band on the SDS-PAGE gel, but tended to precipitate after buffer exchange. This problem solved by adding Triton X100 up to the exchange buffer.

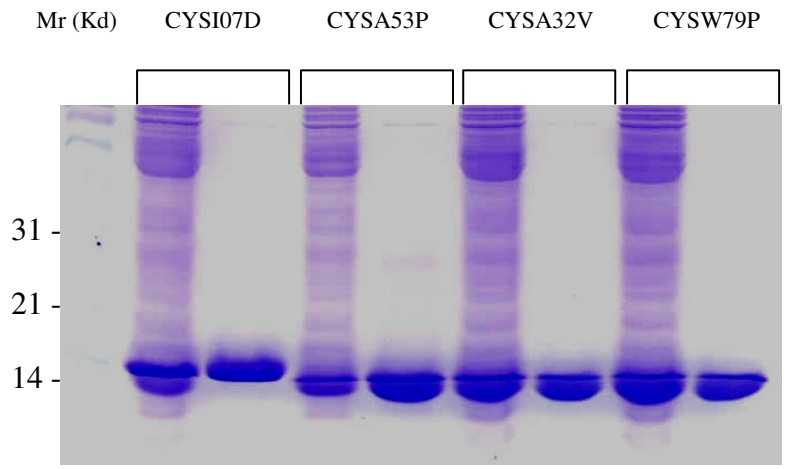


Figure 4.5 SDS-PAGE (12%) of the purified fractions of selected papaya cystatin mutants CYSI07D, CYSA53P CYSA32V and CYSW78P to establish purity during the purification process; lane 1: Molecular weight markers, lanes 2, 4, 6, and 8 crude extracts from an *E. coli* culture; lanes 3, 5, 7 and 9 the respective 1st elution from the purification column.

4.4.4 Inhibition activity of papaya cystatin mutants

To determine if the various mutations performed on the papaya cystatin resulted in any improvement inhibition, enzymatic assays were performed with all mutants and compared to the original native PC using papain, banana weevil and also black maize beetle gut extracts. As shown in Figure 4.5, 10 out of the 18 mutants showed a significant increase in inhibition of papain *in-vitro*. Mutant CYSA52P and CYSE55A further had the highest increase (6-fold) compared to the native PC. Mutant CYSE84A had a 5-fold, while CYStNT (truncation of the N-terminal trunk), CYSW79P, CYSI07D and CYSI07L all had a 2-fold increase. All increases greater that 5-fold were under positive selection pressure. Five mutants did not show any

improvement from the native PC, while 2 mutants, CYSC60T and CYSN85X, had a significant reduction in inhibitory activity compared to the native PC.

When the mutants were tested against banana weevil and black maize beetle gut enzymes, the increases in activity were less than for papain (Figure 4.6). Ten mutants showed significant increases ($p<0.05$) against banana weevil gut proteases with CYSE84A (2-fold) while CYSA52P and CYSI07L with a 1.5-fold increase. For the black maize beetle, only CYSA52P showed a 2.5-fold increase. Mutants CYSE84A, CYSN97P, CYSE52Q, CYSI07L and CYSP03F, had non-significant increases of less than 1.5-fold ($p<0.05$) in their inhibition capacity.

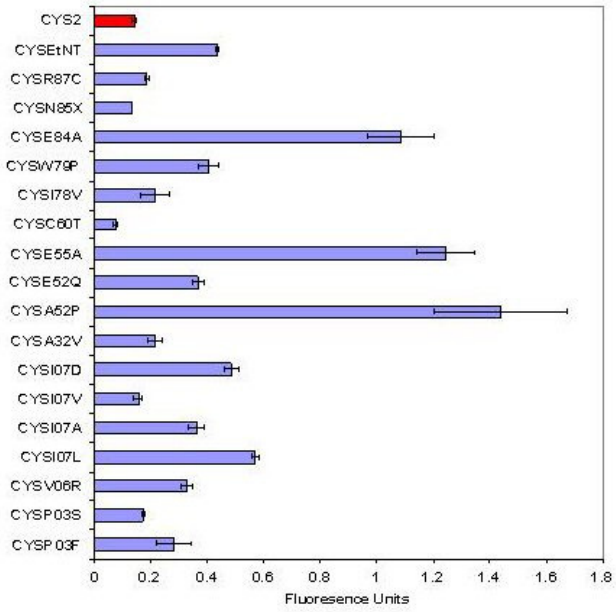
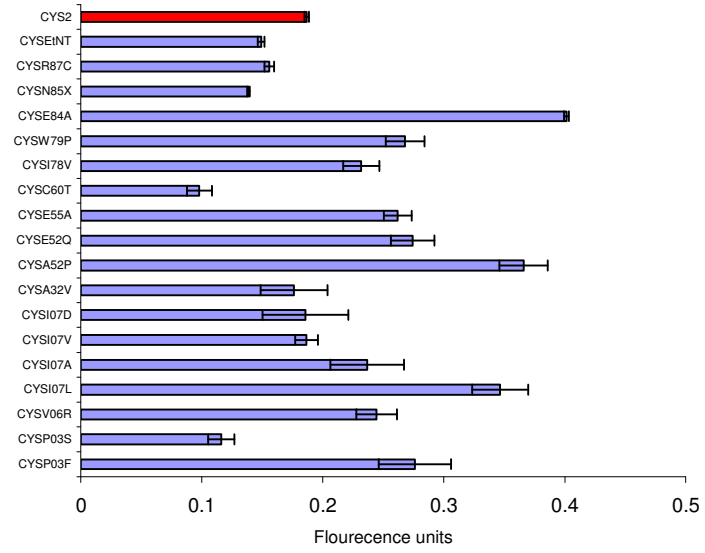


Figure 4.6 Comparison of inhibition activity between native PC (red bar) and 18 PC mutants. The inhibitors were tested by monitoring change in reaction rates of papain hydrolyzing Z-Phe-Arg-AMC a cathepsin-L specific substrate after addition of the inhibitor. Bars represent the mean ± SE of 3 replications of difference in reaction rate before and after addition of the inhibitor in fluorescence units.

A



B

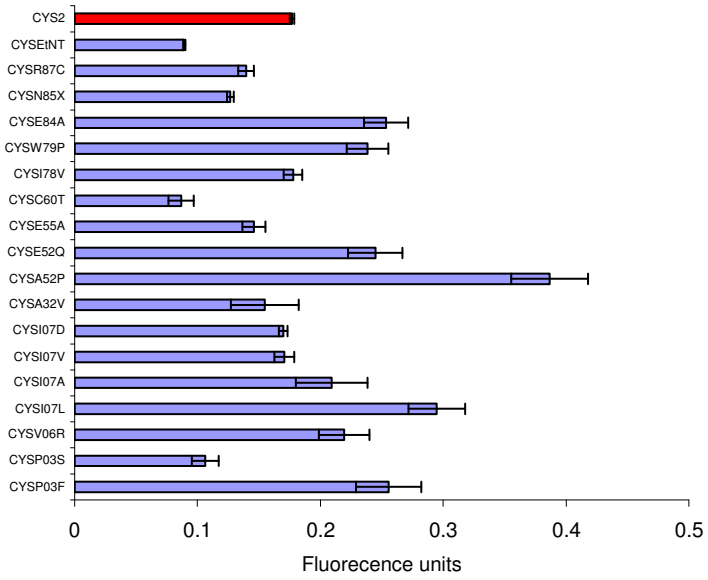


Figure 4.7 Inhibition activities between native PC (red bar) and 18 mutants of the papaya cystatin. Inhibitors were tested by monitoring change in reaction rates of banana weevil (A) and black maize beetle (B) gut extracts hydrolyzing Z-Phe-Arg-AMC substrate after addition of inhibitor. Bars represent mean ± SE of 3 replications of difference in reaction rate before and after addition of inhibitor in florescent units.

4.5 Discussion

This part of the study showed that while there is high sequence conservation particularly in areas that are important to protein structure and function, there were also areas with high variability in amino acid sites in close proximity to the active site e.g the first binding loop. Such variability close to inhibitory sites has been documented for serine-type inhibitors of animal origin. This has been further shown to generate inhibitor variants with significantly different affinities for serine proteases (Creighton and Darby, 1989).

The occurrence of hypervariable amino acid sites among plant protease inhibitors as well as the high variability in affinity supports the idea that these inhibitors have been under selective pressure to evolve in response to herbivorous insect pests and protease diversification in the insects (Lopes *et al.,* 2004). The use and diversity of digestive proteases in coleopteran insects has been well documented (Murdock *et al.,* 1987). Walter *et al.* (1998) showed that the more advanced insects had the highest diversity of cysteine proteases in their gut. This indicates that they have evolved to overcome inhibitors and are perhaps have a more polyphagous feeding habit. Other studies have found that herbivorous insects are able eluding the inhibitory effects of phytocystatins by the use of 'cystatin-insensitive' digestive cysteine proteases (Cloutier *et al.,* 2000; Girard *et al.,* 1998; Michaud *et al.,* 1996) or by breakdown of cystatins using non-target proteases (Girard *et al.,* 1998; Michaud, 1997; Zhu-Salzman *et al.,* 2003).

Some proof has been provided in this study that at molecular level positive selection is active in phytocystatins. This is most likely within the variable amino acid residues in the active site cleft. Differential sensitivity to cystatins was previously identified in

the coleopteran insect *Callosobruchus maculatus*, challenged with soyacystatin, a wound-inducible cystatin from soybean (Moon *et al.,* 2004; Zhu-Salzman *et al.,* 2003). The accumulation of cystatin-insensitive proteases following cystatin ingestion, also observed for the potato herbivore *Leptinotarsa decemlineata* (Cloutier *et al.,* 2000; Gruden *et al.,* 2004), clearly supports the hypothesis of a co-evolution process. This is possibly driven by positive selection explaing the long-term interactions of cystatins with digestive cysteine proteases in plant–insect systems. In this study, the PC mutants CYSA52P, CYSE55A, CYSE84A and CYSI07L showed the highest and consistent improvement in inhibition of papain and protease activity of both banana weevil and black maize beetle. These mutations were all either at positive selection sites or in variable regions close to the active site of the phytocystatin.

In practice, the search for positive selection events among insect digestive cysteine proteases and phytocystatins could be useful in forthcoming years in interpreting the complex structural interactions taking place naturally between these presumably co-evolving proteins. It could futher help in developing rationale strategies for the molecular improvement of phytocystatin variants with potential in plant protection. As a first step, the novel phytocyatatin variants from this study should therefore also be tested in transgenic plants to prove their improved activity *in planta*. The identification of positively selected sites in phytocystatins could further be of general interest in biotechnology. Accumulated data on the functional characteristics of phytocystatins over the last 10 years have made these proteins not only attractive genes for pest control in plants but also for the control of cysteine proteases in various industrial and medical systems (Arai *et al.,* 2002).